Visualization of Uncertain Multivariate 3D Scalar Fields

David Feng

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Computer Science.

Chapel Hill
2010

Approved by:

Russell M. Taylor II

Christopher Healey

Yueh Z. Lee

Mary C. Whitton

Marc Niethammer

ABSTRACT

DAVID FENG: Visualization of Uncertain Multivariate 3D Scalar Fields
(Under the direction of Russell M. Taylor II)

This dissertation presents a visualization system that enables the exploration of relationships in multivariate scalar volume data with statistically uncertain values. The n-Dimensional Volume Explorer (nDive) creates multiple visualizations that emphasize different features of the data. nDive provides data selection mechanisms that are linked between views to enable the comparison of separate data populations in the presence of uncertainty.

Incorporating representations of uncertainty into data visualizations is crucial for preventing viewers from drawing conclusions based on untrustworthy data points. The challenge of visualizing uncertain data increases with the complexity of the data set. nDive separates the visualization into multiple views: a glyph-based spatial visualization and abstract multivariate density plots that incorporate uncertainty. The spatial visualization, Scaled Data-Driven Spheres (SDDS), distributes (for each variable) a set of scaled, colored spheres in the sample space. A user study demonstrates that viewers are faster and more accurate using SDDS to identify spatial values and relationships as compared to superquadric glyphs, an alternative technique. The abstract plots complement SDDS by preattentively focusing the viewer on trustworthy data points using the probability density function of the data. These views, coupled with novel interaction techniques that utilize data value uncertainty, enable the identification of relationships while avoiding false conclusions based on uncertain data.

The primary application of nDive is aiding radiologists who use magnetic resonance spectroscopy (MRS) to better understand the composition of abnormalities such as brain tumors and improve diagnostic accuracy. This work demonstrates how nDive has been successfully used to identify metabolic signatures for multiple types of tumors.

ACKNOWLEDGMENTS

I would like to thank first my advisor, Russell Taylor, for not only his guidance throughout this work but also for his consistently honest, encouraging, and selfless attitude. I continually count myself lucky for the opportunity to have him as an advisor. My entire committee helped me to polish my work (and this document) and ensure its veracity, so I also thank Chris, Mary, Yueh, and Marc. Likewise, thanks go to my colleagues in my research group at UNC, Computer Integrated Systems for Microscopy and Manipulation, for their support.

Yueh and his colleague Lester Kwock were the force that kept my work honest and practical. They provided the driving problem and pragmatic perspective that ensured that success in this project would be defined by the usefulness of what I produced. My collaboration with Yueh and Lester ensured that I frequently walked that halls of the UNC hospitals. To all of the anonymous hospital staff and patrons who helped me to remember what is important: thank you.

In my experience, the staff of the UNC computer science department are unrivaled in their helpfulness and positive attitudes. My work in this department would certainly not have been the same without their consistent and timely aid. I can only hope that the spirit of collegiality and productivity in this department will follow me on from here.

None of this would have been possible without the unfailing support of my wife, Lauren. Helping someone persevere throughout a process such as this is no simple task, yet she managed despite all of her other obligations. Humble thanks are not enough, but will have to suffice for now.

TABLE OF CONTENTS

LIST OF TABLES **x**

LIST OF FIGURES **xi**

LIST OF ABBREVIATIONS **xiv**

1 Introduction **1**

1.1 Problem Description 2

1.2 Thesis Statement 4

1.3 Approach 5

1.4 Definitions 8

1.5 Summary of Results 8

2 Scaled Data-Driven Spheres **10**

2.1 Background 11

2.1.1 Surfaces 11

2.1.2 Direct Volume Rendering 13

2.1.3 Correlation Fields 16

2.1.4 Glyphs 18

2.2 SDDS Visualization Design 21

2.3 SDDS Implementation 24

2.4 Comparison of Superquadrics to Sparse Glyphs 25

2.4.1 Task 1: Value Estimation 29

2.4.2 Task 2: Correlation Identification 30

2.4.3 Equipment and Materials 31
2.4.4 Procedure 31
2.5 User Study Results 32
2.5.1 Task 1: Value Estimation 32
2.5.2 Task 2: Correlation Identification 34
2.5.3 Analysis of Learning 36
2.5.4 Follow-up Questionnaire 37
2.6 Attempts at Uncertainty Visualization 38
2.7 Discussion 40
3 Multivariate Uncertainty Visualization in Abstract Plots 42
3.1 Background 45
3.2 Information Visualization 45
3.2.1 Scatter Plots 46
3.2.2 Parallel Coordinates 48
3.2.3 Sources and Classifications of Uncertainty 50
3.2.4 Uncertainty Visualization 51
3.2.5 Animated Plots 53
3.2.6 Image Structure and Preattentive Vision 54
3.2.7 Density Estimation 55
3.3 Uncertain Plots 56
3.3.1 From Scatter Plots to Density Plots 58
3.3.2 Mean Emphasis 60
3.3.3 PDFs in Parallel Coordinates 61
3.3.4 Focus and Context 63
3.3.5 Scalability and Accelerated Rendering 64
3.3.6 Probabilistic Plots 65

3.4 Discussion 67

4 User Interaction with Uncertain Plots 70

4.1 Background 71

4.1.1 Linked Visualization Systems 71

4.1.2 User Interaction with Abstract Plots 72

4.1.3 Variable Ordering 72

4.1.4 The Point-Line Duality 73

4.2 Linear Function Brushing 73

4.3 Lasso Brushing 79

4.4 Implementing Brushing in Rescaled Screen Space 80

4.5 New Axis Construction 82

4.6 Brushing in Uncertain Plots 87

4.7 Discussion 91

5 Magnetic Resonance Spectroscopy 93

5.1 Nuclear Magnetic Resonance 94

5.2 Magnetic Resonance Imaging 97

5.3 Spectroscopic Imaging 101

5.4 LCModel: Estimating Absolute Concentrations 102

6 nDive: n-Dimensional Volume Explorer 105

6.1 Goals 105

6.2 MRS Visualization 106

6.3 nDive 106

6.4 Tumor Analysis 108

6.5 nDive Evaluation 113

6.5.1 Neuroradiology 116

6.5.2 Neurology 121

6.5.3 Spectroscopy . . . 122
6.6 Neurosurgery . . . 124
6.7 Discussion . . . 125
7 Conclusion **129**
7.1 Results . . . 129
7.2 Limitations . . . 130
7.3 Future Work . . . 130
A Appendix: nDive User Manual **133**
A.1 Loading Data . . . 134
A.1.1 Handling Raw LCModel Data . . . 135
A.1.2 Loading Spectroscopy Data . . . 135
A.1.3 Loading Anatomical Data . . . 136
A.2 Visualization Modes . . . 136
A.2.1 3D Visualization Window . . . 136
A.2.2 2D Visualization Window . . . 139
A.2.3 Scatter Plot Window . . . 140
A.2.4 Parallel Coordinates Window . . . 142
A.3 Voxel Selection . . . 144
A.3.1 Exporting Selected Voxels . . . 144
A.3.2 Normalization by Selection . . . 144
A.4 Miscellaneous Tools . . . 145
A.4.1 Threshold Classification . . . 145
A.4.2 Load Default Data . . . 145
A.4.3 Compute Ratio . . . 146
A.4.4 Combine Variables . . . 146
A.5 Compilation . . . 146

A.5.1 Qt 147
A.5.2 VTK 147
A.5.3 GDCM 147
A.5.4 ITK 147
A.5.5 nDive 148
Bibliography **149**

LIST OF TABLES

2.1 Significant Results for Variable-specific Estimation Error 34

2.2 Significant Results for Variable-specific Correlation Misidentification 35

3.1 Example Table of Election Data 42

3.2 Example Table with Spatial Coordinates 45

6.1 MRS Sphere Color Code 107

LIST OF FIGURES

1.1 Examples of Multivariate 3D Visualization 2

1.2 Visualization System Components 5

2.1 Multiple Opaque and Transparent Isosurfaces 12

2.2 Textured Isosurfaces 12

2.3 Direct Volume Rendering and Isosurfaces 13

2.4 Multidimensional Transfer Functions 15

2.5 PCA-based Transfer Functions 16

2.6 Set Operations and Multi-field Graphs 17

2.7 Superquadric Glyphs and Surface Glyphs 19

2.8 Data-driven Spots 20

2.9 SDDS Applied to a Brian Tumor 22

2.10 SDDS with Five Variables 24

2.11 SDDS in Cross-eyed Stereo 25

2.12 Superquadrics Used in the User Study 27

2.13 Example of the Value Estimation Task 29

2.14 Example of the Correlation Identification Task 30

2.15 Value Estimation Error and Timing Results 32

2.16 Per-variable Value Estimation Error 33

2.17 Correlation Identification Error and Timing Results 35

2.18 Per-variable Correlation Identification Errors 36

2.19 Learning Analysis 37

2.20 Uncertain SDDS via Opacity 39

2.21 Uncertain SDDS vis Texture 39

3.1 Example Bar Graph with Error Bars 44
3.2 Example Scatter Plot 46
3.3 Scatter Plot Matrix 47
3.4 Continuous Scatter Plots 48
3.5 Example Parallel Coordinates Plot 49
3.6 Clustered Parallel Coordinates 49
3.7 Pang's Uncertainty Visualizations 52
3.8 Probabilistic Surfaces 53
3.9 Preattentive Visual Properties 54
3.10 Kernel Density Estimation 55
3.11 A False Negative in Parallel Coordinates 56
3.12 A False Positive in Parallel Coordinates 57
3.13 Choline-to-Creatine Scatter Plot 58
3.14 Normal Distributions in Parallel Coordinates 60
3.15 Parallel Coordinates Plot of 4 MRS Metabolites 61
3.16 Probabilistic Plots 64
3.17 Monte Carlo Integration in Parallel Coordinates 65

4.1 The Point-Line Duality 74
4.2 Example Scatter Plot and Parallel Coordinates Plot 74
4.3 Cluster Selection 75
4.4 Linear Function Selection 76
4.5 Linear Patterns in Parallel Coordinates 76
4.6 Lasso Selection 78
4.7 Brushing in a Transformed Space 80

4.8 Example 4-Variable Parallel Coordinates 82
4.9 Adding a New Column - Sum 84
4.10 Adding a New Column via Sum and Ratio 85
4.11 New MRS Axis 86
4.12 Integrating a Normal Distribution within a Boundary 89

5.1 MRS Spectra Overlaid on a T1 MRI 93
5.2 Larmor Precession 94
5.3 Circular Polarization 95
5.4 Sinc RF Pulse 96
5.5 MRI - After the Slice Selection Gradient 97
5.6 MRI - After the Phase Encoding Gradient 98
5.7 MRI - During the Read-out Gradient 99

6.1 Use Case: Stereo SDDS 108
6.2 Use Case: Axis Reordering 109
6.3 Use Case: Scatter Plot Selection 110
6.4 Use Case: Parallel Coordinate Analysis 110
6.5 Use Case: Parallel Coordinates Selection 111
6.6 Use Case: Probabilistic Plot 111
6.7 Potential Central Nervous System Lymphoma 112
6.8 Potential Gliomatosis Cerebri 114

A.1 nDive Screenshot 134
A.2 3D Visualization Window 137
A.3 2D Visualization Window 139
A.4 Scatter Plot Window 141
A.5 Parallel Coordinates Window 142

LIST OF ABBREVIATIONS

1D	One-dimensional
2D	Two-dimensional
3D	Three-dimentionsal
Cho	Choline
Cr	Creatine
CSF	Cerebrospinal Fluid
DDS	Data-driven Spots
DVR	Direct Volume Rendering
Glu	Glutamine
KDE	Kernel Density Estimation
MALDI	Matrix-assisted Laser Desorption/Ionization
MR	Magnetic Resonance
MRI	Magnetic Resonance Imaging
MRS	Magnetic Resonance Spectroscopy
MRSI	Magnetic Resonance Spectroscopic Imaging
NAA	N-acetylaspartate
NAAG	N-acetylaspartylglutamic Acid
ND	N-dimensional
nDive	n-Dimensional Volume Explorer
NMR	Nuclear Magnetic Resonance
PC	Parallel Coordinates
PCA	Principal Component Analysis
PDF	Probability Density Function
SDDS	Sparse Data-driven Spheres

CHAPTER 1

Introduction

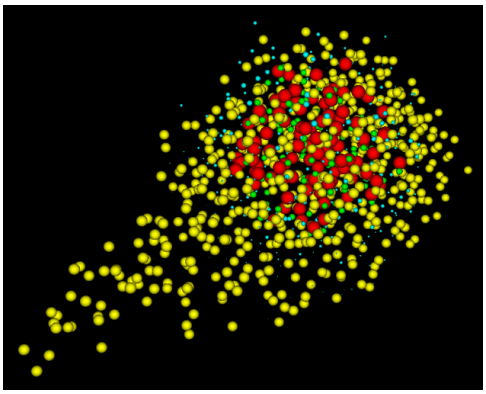
Richard Hamming, one of the pioneers of computing, eloquently states the following:

> *The purpose of computing is insight, not numbers.*

Scientific visualization has the highest bandwidth of the many paths between a computer and insight. Its appeal and utility stem from the elegant and powerful machine that is the human visual system. There are few other abilities that humans come by so naturally and devote so much time to as learning to interpret the environment visually. When scientists make a series of measurements, it is only natural that for many the first instinct is to draw a picture.

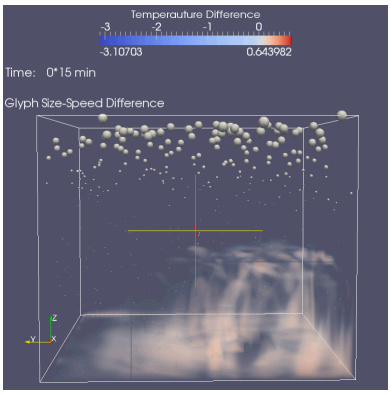
For many problems, the desired picture is obvious (although how to produce that picture may not be obvious). When an explorer records the positions of encountered landmarks, an annotated map is often the first technique that comes to mind. This is because the question of where to draw the data elements has been answered implicitly. Facing north, the lake is to the right of the village; we don't need to think hard about where to draw the house and village in that picture.

Visual representations of geographical and other data sets are called visualizations, and more complex data often require more complex visualizations. It is the job of the visualization designer to understand *what* the visual system does well and *how* to leverage its subtleties to help the viewer best understand the underlying data. The basic question is simple: "What

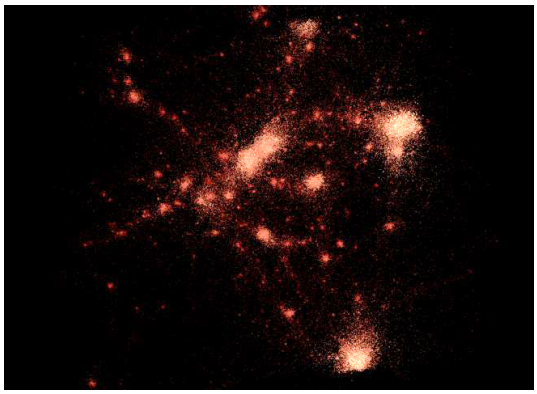


(a) Virtual Cell Simulation

(b) Matrix-assisted Laser Desorption/Ionization (MALDI)



(c) Weather Simulation



(d) Galaxy Simulation

Figure 1.1: Examples of Multivariate 3D Visualization. a) Visualization of multiple chemicals in a cellular chemical reaction. b) Specotroscopy based analysis of brain tissue decomposed into different biomolecule concentrations. c) Weather simulation of a single time-point in a time-varying temperature and wind speed simulation. d) Galaxy simulation of multiple properties of a large number of galaxies.

is the best mapping from data to visual representation?" To answer this question, the visualization designer must understand both the strengths and limitations of human vision and understand the driving problem that produced the data in detail.

1.1 Problem Description

One challenging problem in visualization is how to represent data sets in which multiple scalar measurements are taken throughout three-dimensional space (a.k.a. multivariate 3D data). Such data sets come from diverse research areas. For example, biological simulations of cellular

processes estimate the concentrations of many chemicals over time with the goal of understanding the driving chemical interactions (Figure 1.1a). Biomolecule spectroscopy techniques like matrix-assisted laser desorption/ionization (MALDI) attempt to correlate concentrations of different molecules with disease (Figure 1.1b). At a larger scale, weather simulations produce 3D volumes of pressure, temperature, and wind speed estimates ((Figure 1.1c). Larger still, astrophysics simulations of galaxy formation involve many different galaxy properties (Figure 1.1d). These multivariate scalar 3D data sets are often large and dense, which makes effective visualization of them a challenge. Visually representing a single variable throughout 3D space is a difficult problem still under active research; representing many variables at once is even more difficult.

To make matters more difficult, nearly all scientific measurements are made with some degree of uncertainty or error. Mechanical sensors are noisy, numerical simulations have error, and some values can only be estimated probabilistically. Each measurement is paired with a quantification of its trustworthiness, which essentially doubles the size of the data. Without any sort of visual encoding of uncertainty, a visualization runs the risk of leading viewers to draw incorrect conclusions. Because of this, uncertainty visualization has been recognized as one of the top problems in the visualization literature (Johnson, 2004).

This dissertation describes a novel system for visualizing uncertain multivariate scalar 3D data sets that attempts to guide the viewer's attention to trustworthy values and prevent uncertain values from leading to false conclusions. The approach taken is to address the information density problem by displaying spatial information and uncertainty information in separate visualizations. These visualizations must be effectively linked together such that the user understands the correspondence between value representations in all views. The results of user interactions in one view are immediately represented in the other visualizations. Both views leverage knowledge of the human visual system's preattentive feature discrimination to make the visualization easy to understand.

The problem that drives the design of this system comes from magnetic resonance spectroscopy (MRS). Radiologists use MRS to understand brain abnormalities such as tumors. MRS imaging measures the concentrations of multiple metabolites; the relationships among

metabolites and anatomy can clarify tumor extents that are not always obvious using anatomical magnetic resonance imaging (MRI). For example, tumors can extend beyond the boundaries visible using traditional MRI techniques. Also, radiation necrosis (dead tissue resulting from radiation treatment) is difficult to distinguish from recurring tumors after surgery. MRS reveals the chemical composition of tissue, which enables radiologists to more easily distinguish between malignant and benign tissue.

The radiologists using MRS have two primary goals for their visualization:

1. **Identify Metabolite Relationships:** Radiologists use relationships among metabolites as indicators of disease.

2. **Extract Metabolite Concentrations:** Once radiologists identify a tumor, clinicians and surgeons need to extract absolute metabolite concentrations at precise 3D locations to accurately plan procedures. This goal requires viewers to have positional awareness within the data space and direct access to raw data values.

MRS was developed based on the foundation of nuclear magnetic resonance (NMR). The raw data consists of per-voxel metabolite spectra, in which spectral peaks roughly correspond to different metabolites. The signal of the spectra is fairly weak when compare to the many sources of noise in a clinical magnetic resonance imager, so radiologists use an optimization technique to estimate both the absolute concentrations of individual metabolites and the certainty of those estimations. The magnitude of the uncertainty is critical in deciding whether any particular metabolite concentration is useful during analysis. This dissertation presents a system for visualizing MRS data that attempts to guide the focus of the viewer to confident values and prevent uncertain values from distracting the viewer.

1.2 Thesis Statement

A scaled data-driven spheres visualization of multivariate scalar volume data enables simultaneous value estimation and relationship identification faster and more accurately than previous methods. Adding density-based scatter and parallel coordinates plots coupled with functional brushing enables relationship discovery while

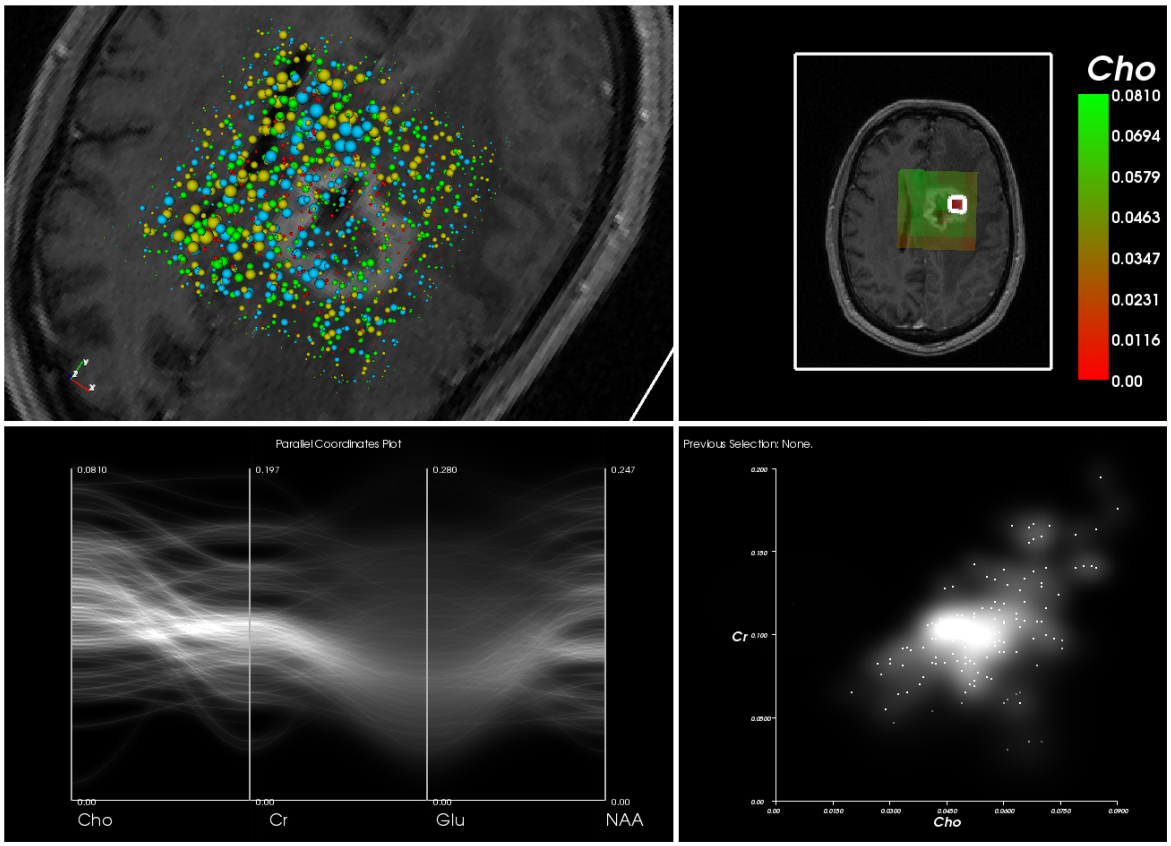


Figure 1.2: Visualization System Components. Upper left: a 3D view of the visualization system including an anatomical slice plane and SDDS glyph visualization of the spectroscopy data. Upper right: orthogonal view of the anatomical slice plane with a red-to-green Choline concentration pseudocoloring. Lower-right: a density-based scatter plot of Choline and Creatine. Lower-left: a density-based parallel coordinates plot of Choline, Creatine, Glutamine, and NAA.

de-emphasizing false relationships based on uncertain data.

1.3 Approach

The proposed visualization system, called the n-Dimensional Volume Explorer (nDive), has four linked components, as shown in Figure 1.2. First, a 3D spatial visualization technique called Scaled Data-Driven Spheres (SDDS) that is coupled with an anatomical slice plane helps viewers see global relationships among metabolites and anatomical structure. Using SDDS, a sphere scaled by concentration and colored uniquely for each metabolite represents

each data sample as described in Chapter 2.

Two more visualizations, described in Chapter 3, help radiologists confirm and explore hypotheses generated via the SDDS visualization. The second visualization is a scatter plot that highlights the relationships between two metabolites. Basic scatter plots display data points in Cartesian space as small dots or other simple glyphs, however such a plot does not account for uncertainty. This work describes how to encode statistical uncertainty into scatter plots using kernel density estimation (KDE). KDE superimposes the distributions of the individual data values to produce a density plot. Values with small standard deviations appear as small, high contrast features that the visual system is able to preattentively distinguish from large, low contrast values with large standard deviations. This uncertainty encoding helps to prevent viewers from making false conclusions based on uncertainty values by drawing the viewer's attention to the trustworthy values.

The third visualization is a parallel coordinates (PC) plot, which extends the scatter plot into more than two dimensions. It is an abstract data representation that plots data samples (for MRS, voxels) as lines traveling through multiple parallel variable axes. PC is most useful for identifying patterns, such as clusters of values or relationships. Because a single data point is now a line that crosses the entire plot, the PC plot is invariably more visibly dense than a scatter plot. The parallel coordinates plot is useful for examining data points selected in the scatter plot in the context of more than two variables. Uncertainty is encoded into the PC plot using the same KDE-based approach used for the scatter plot.

The strengths of the spatial and abstract visualizations complement each other. The density of information in an SDDS visualization can result in clutter and occlusion problems, but the PC and scatter plot visualizations compensate by discarding spatial information. When a viewer finds patterns in the PC or scatter plot visualization, they can refer to the SDDS visualization to understand the positions of interesting voxels and their relation to anatomy. User interactions such as selections are linked in all visualizations so that data points and relationships selected in one view are reflected in the others, as described in Chapter 4.

This work describes several novel user interaction techniques for both PC and scatter

plots that help users identify particular classes of relationships among variables. Radiologists commonly use ratios of variable pairs as abnormality indicators, so linear relationship discovery is a useful feature. Linear function brushing is a class of techniques for selecting data points in both PC and scatter plots from user-drawn lines. These techniques incorporate the uncertainty of data points by efficiently computing the probability that a data point should be included in a selection. These techniques are described in Chapter 4.

The fourth visualization component satisfies the surgical planning visualization goal. Once users have identified a meaningful relationship or set of voxels, our colleagues need an interface for precisely extracting MRS data values in anatomical context. An interactive pseudocolored slice plane with isovalue contours shows values of a particular metabolite overlaid on a grayscale coloring of anatomy. While simple, this visualization is a critical last step before information discovered in the exploratory visualizations can be used in practice. All four components of the visualization and analysis system are presented in Figure 1.2. All visualization components have been integrated into a single application, called nDive (n-dimensional volume explorer). Expert users participated in a focus group study evaluating nDive's usability, as described in more detail in Chapter 6.

While MRS drove the design of this visualization system, it is more generally applicable to other multivariate 3D data sets because the visualization goals of MRS are similar to those of other multivariate data sets. A common theme with multivariate data is understanding relationships between variables, with or without spatial information. Researchers from the Virtual Cell project from the University of Connecticut (`http://www.nrcam.uchc.edu/`) wish to understand the relationships and interactions among chemicals in a simulated cellular process. Microarray expression visualizations have noisy measurements that can benefit from the uncertainty-encoded abstract visualization design. Correlating spectroscopy data to anatomical features is also a driving problem for researchers using Matrix-assisted Laser Desorption/Ionization (MALDI). The diversity of these problems shows that the techniques presented in this work are broadly applicable outside of MRS.

1.4 Definitions

The visualization system as a whole described in this work applies to statistically uncertain multivariate 3D scalar data. For reference, I formally define the relevant terms as follows:

- **Dimension**: this will generally be used to refer specifically to the *spatial* dimensionality of a data set. For example, a 3D data set is composed of a samples on a grid structure with positional coordinates that have three spatial components (x,y,z).
- **Data Value**: a single measurement. In this context, each data value is the concentration of a metabolite at a single sample location.
- **Data Point**: a position in space, which may contain multiple data values. For MRS data, this is a single spectroscopy voxel.
- **Scalar**: each data value is a magnitude, as opposed to a direction vector or more complex tensor.
- **Multivariate**: each spatial position has multiple data values. Here, many metabolite concentrations are sampled at the same spatial locations. This is also called multi-field data, but for consistency this term will not be used.
- **Uncertain Data**: shorthand for data sets that have quantified uncertainty measured or estimated for each data value.
- **Statistically Uncertain**: the quality of a data value is defined by a statistical distribution (e.g. the normal distribution). Different classes of uncertainty are discussed in more detail in Section 3.2.3.

1.5 Summary of Results

The results of the research presented here include:

- A multivariate 3D scalar visualization technique, called Sparse Data-driven Spheres, that uses a sparse-glyph methodology to display variable values using small, colored

spheres. SDDS enables viewers to identify relationships among variables for low resolution 3D data.

- Results of a user study showing that viewers are faster and more accurate at value estimation and correlation identification with SDDS as compared to superquadric glyphs (Feng et al., 2009).

- A density-based augmentation to scatter plots and parallel coordinates plots for visualizing uncertain multivariate data that preattentively highlights data points with low uncertainty and draws the viewer's attention away from data points with high uncertainty (Feng et al., 2010a; Feng et al., 2010b).

- Novel techniques for interacting with uncertain scatter plots and parallel coordinates plots. These techniques utilize data value uncertainty to help users interact predominantly with trustworthy data vales by making uncertain data values more difficult to select (Feng et al., 2010b).

- Expert commentary and evaluation by users in a focus group study of nDive, the application that links these visualizations together. Experts said that nDive's visualizations and interaction techniques were useful, although complex, ways of distinguishing between different voxel populations.

These results have been presented in three publications:

- Evaluation of glyph-based multivariate scalar volume visualization technique (Feng et al., 2009)

- Linked exploratory visualizations for uncertain MR spectroscopy data (Feng et al., 2010a)

- Matching visual saliency to confidence in plots of uncertain data (Feng et al., 2010b)

CHAPTER 2

Scaled Data-Driven Spheres

(The contents of this chapter were presented at the ACM Symposium on Applied Perception in Graphics and Visualization 2009 (Feng et al., 2009))

Two fundamental challenges for the display of spatially positioned multivariate data are occlusion and information density. As humans we deal with visual occlusion so frequently that our visual system uses occlusion to determine properties of what we see (Palmer, 1999). For example, binocular disparities between the edges produced by occluding objects help us determine the relative distances of objects in our visual field. Occlusion cues are both a blessing and curse in visualization: they help us understand what we see, but by definition occlusion prevents all of the data from being visible at the same time.

Visualizations of multivariate data push the limits of information density. Such visualizations are also called multi-field visualizations because they attempt to visualization multiple 3D fields of data at once. A visualization of a bivariate data set (e.g. wind speed and air pressure) must represent twice as much information as a corresponding univariate data set within the same volume of space. As the number of variables increases and screen size does not, eventually the information will be too dense to visualize reasonably. The visualization designer must be very careful when choosing the display technique for such data sets.

This chapter describes a novel visualization technique called Scaled Data-Driven Spheres (SDDS) that displays multiple 3D scalar fields at once. This technique extends Bokinsky's 2D Data-Driven Spots (DDS) into 3D with modifications to account for occlusion. The

fundamental idea behind both SDDS and DDS is to use multiple sets of sparse, similar, and simple glyphs to encode the values of each variable. The following sections will motivate the decision to use a glyph-based technique over alternatives, describe how to produce SDDS visualizations, and report the results of a study comparing SDDS to another glyph-based technique.

2.1 Background

There are four broad classes of techniques that are potentially useful for visualization multivariate 3D data. These include surfaces, direct volume rendering, correlation fields, and glyphs. SDDS was ultimately developed because alternative techniques either did not directly meet the visualization goals of MRS or did not meet them well enough.

2.1.1 Surfaces

The most common technique for visualizing a single scalar field is to draw a surface at a boundary of interest within the data. Surface representations are useful for applications where object segmentation ("Where is the cell in this image?") and shape understanding ("Is the cell bumpy?") are the goals.

Identification of such boundaries in images is called segmentation. For simple images, a shape of interest may be bounded by a single intensity value, and one need only connect those values together to form a polygonal surface (Lorensen and Cline, 1987). These are called isosurfaces. While more complex images may require more sophisticated boundary detection techniques, isosurfaces are commonly used for visualization because they can be computed efficiently and can be effective for visualizing a single 3D scalar field.

To extend surface-based techniques to multivariate data, the simplest option is to create separate isosurfaces for each variable and render them simultaneously. However, a surface closer to the viewpoint will occlude those behind it, thereby hiding information that may be useful to the viewer. Making occluding surfaces partially transparent reveals hidden surfaces at the cost of important depth and shape cues. Interrante et al. showed that the human visual

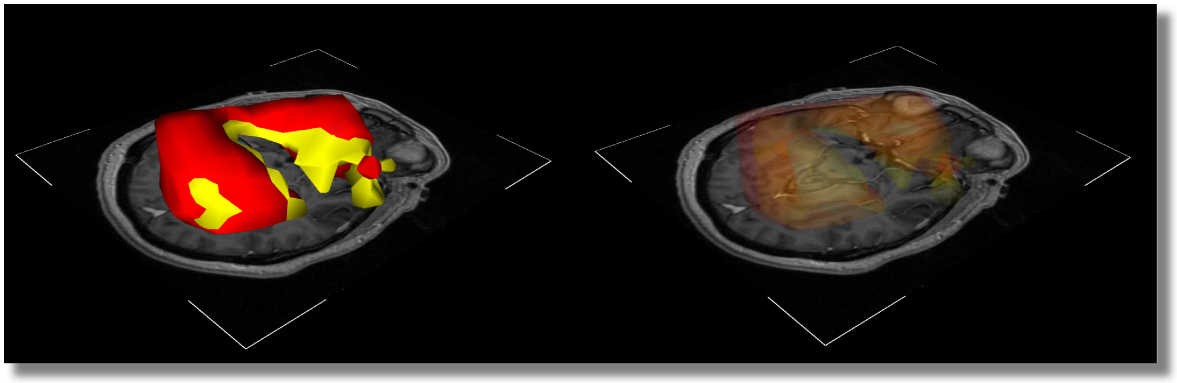


Figure 2.1: Multiple Opaque and Transparent Isosurfaces. Both visualizations depict isosurfaces of three metabolites in an MRS data set. On the left, the yellow and red opaque surfaces almost completely occlude the cyan surface. On the right, transparency reveals the hidden surface but makes shape and depth determination more difficult.

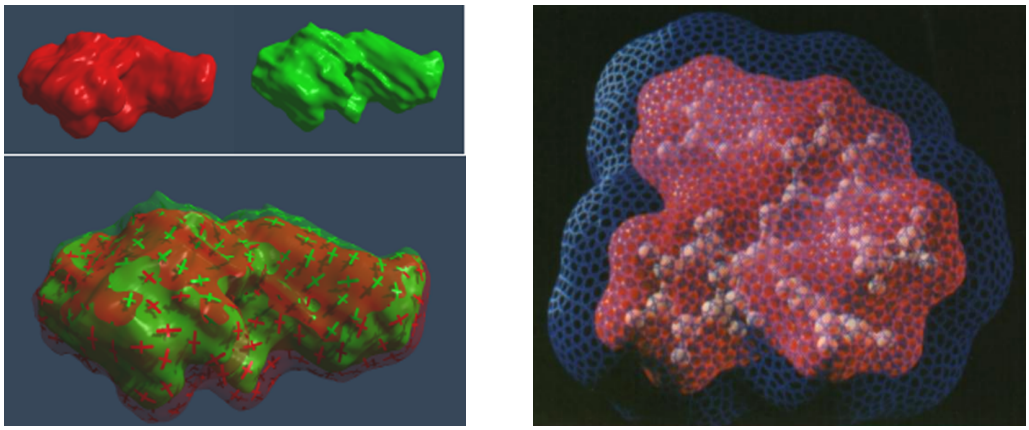


Figure 2.2: Two different techniques for displaying textured isosurfaces. Left: Weigle's nested surface visualization. Right: Rheingen's mesh surfaces.

system does not interpret depth accurately behind transparent surfaces (Interrante et al., 1997). The issues with opaque and transparent isosurfaces as applied to three metabolites for MRS data are shown in Figure 2.1. Notice how the opaque cyan surface is almost entirely occluded (left), but still difficult to interpret with added transparency (right).

Interrante goes on to discuss how partially transparent textures can offer limited improvements to this situation. She proposes using surface textures based on line integral convolution along the first principal curvature direction of the surface (Interrante, 1997). Weigle and Taylor proceed further to use both principal curvature directions and discuss the use of projective shadows for two nested surfaces (Weigle and Taylor, 2005). They show that surface glyphs

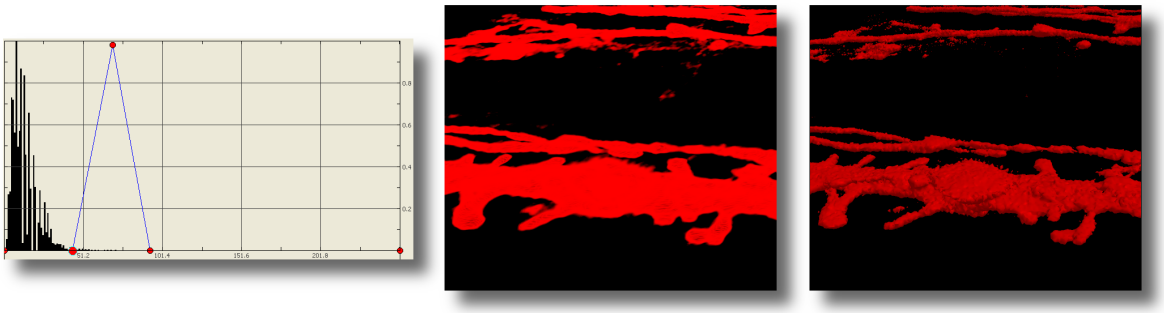


Figure 2.3: Example of Direct Volume Rendering and comparison to isosurface rendering. Left: A transfer function that emphasizes values around 75 with a histogram in the background. Center: the corresponding DVR. Right: isosurface rendering with isovalue 75.

in combination with translucent surfaces enhance a viewer's ability to perceive the shape of exterior surfaces for the display of two nested surfaces. Rheingans proposes an alternative mesh-like transparent texture for displaying nested isosurfaces (Rheingans, 1996).

While surface-based techniques such as these can be useful for visualizing two or three variables at once, ultimately they do not address the visualization goals this application. By definition, isosurfaces only display a single data value, so viewers must imiplicitely infer information about other data values. A technique that preserves data value visibility and shows variable relationships is required.

2.1.2 Direct Volume Rendering

Direct Volume Rendering (DVR) is a standard technique for displaying a single scalar volume (Levoy, 1988). This involves projecting all of the data values in the volume onto the image plane and compositing overlapping values. A transfer function that maps data values to opacity and color is required for compositing.

While complex, the transfer function is a versatile means of controlling the resulting image. Standard transfer functions are simple 1D graphs with the range of possible data values on the x-axis and the compositing parameter, such as opacity or color, on the y-axis. Figure 2.3 (left) depicts a transfer function as line that gives the user control over both opacity and color at the same time. Individual nodes can be colored, and their opacity is controlled by y-axis position. The transfer function shown emphasizes a range of values with its peak at

75. The resulting image, shown in Figure 2.3 (center) is similar to an isosurface in that it emphasizes a single narrow range of values. The DVR image looks foggy because of the low data values near 50 that are rendered with low opacity values. It also contains no diffuse shading, which is possible to estimate using DVR extensions. The right frame of Figure 2.3 shows an isosurface of the same data with an isovalue of 75.

The simplest way to extend DVR to multiple variables is to separately define opacity transfer functions for each data variable and combine the resulting intensity images using different color components (e.g. red, green, and blue) (Cai and Sakas, 1999; Rösler et al., 2006). Images generated via this type of color mixing are difficult to interpret, especially when more than two colors are being combined. The reasons for this are similar to why overlapping transparent isosurfaces are hard to interpret (Rheingans, 1992).

Rather than define multiple transfer functions, a single DVR image can be modified to highlight features in multiple variables. Stompel et al. use techniques inspired from non-photorealistic rendering (NPR) to emphasize interesting image characteristics such as high spatial or temporal gradients (Stompel et al., 2002). It is not apparent how this technique can be extended to handle more than a few variables at once.

Alternatively, the standard one-dimensional transfer-function can be extended to multiple dimensions. Kniss et al. describe a widget for exploring features in 2D gradient-value histograms (Kniss et al., 2001). The user can control the opacity and color of bins selected with the widget, thereby creating a DVR transfer function (Figure 2.4). In a sense, the multi-dimensional transfer function is an exploratory visualization tool that links spatial and abstract data representations together via specialized user interaction techniques. This work describes a similar class of tool that has been customized for multivariate MRS data sets with particular attention paid to uncertainty. Examples of such tools are described in more detail in Section 3.1.

Because of the complexity of useful transfer function specification, Kindlmann and others have attempted to automatically initialize transfer functions to identify potentially useful features (Kindlmann and Weinstein, 1999). With the aid of creative user interfaces, the process of finding the right transfer function conceptually becomes user-guided identification

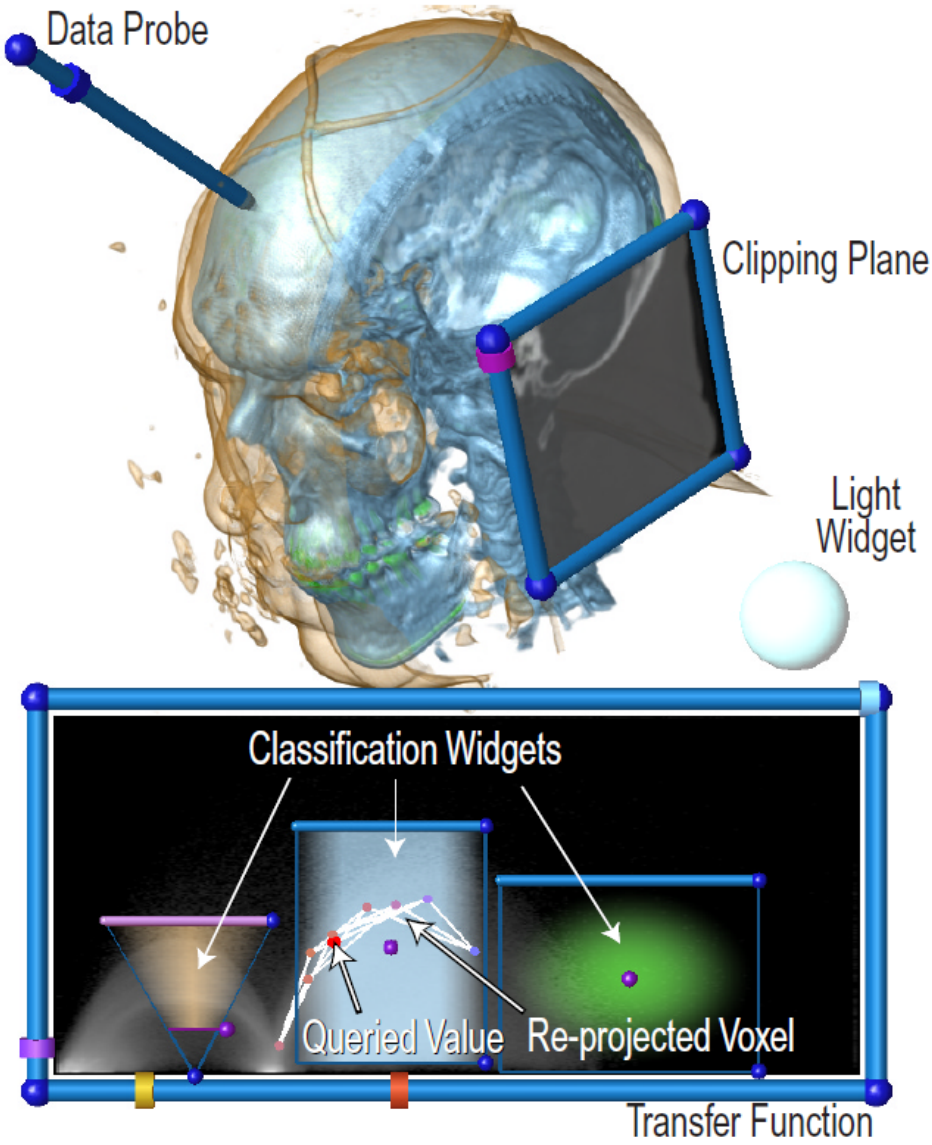


Figure 2.4: Multidimensional transfer functions for feature identification (Kniss et al., 2001). The user is given a widget to identify features in an abstract space which are used to automatically create a transfer function.

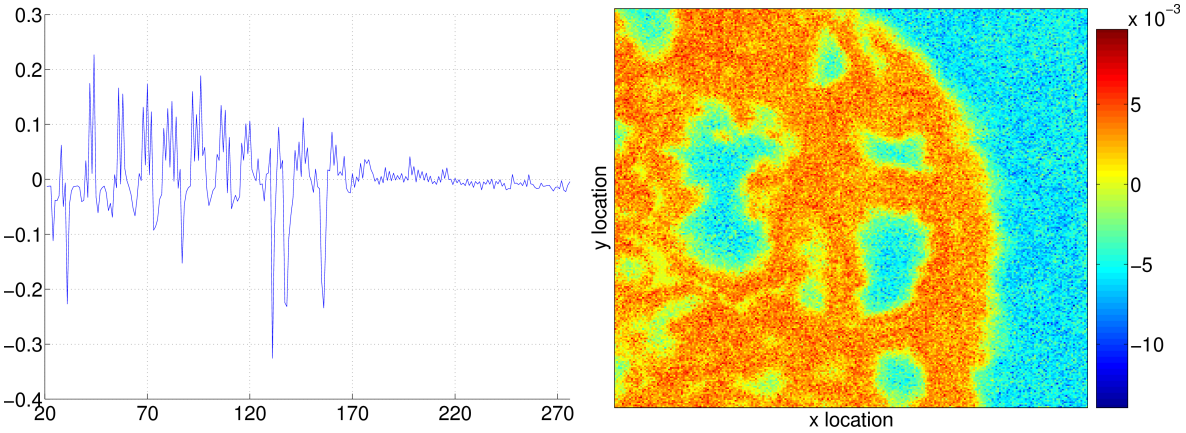


Figure 2.5: PCA-based Transfer Functions. Broersen et al. use PCA of spectroscopy data to identify features that can be used to generate transfer functions for DVR (Broersen and van Liere, 2005). Left: a measured spectrum at a single location. Right: a pseudocolored representation of an entire image with colors computed automatically.

of correlations and patterns among variables, as described in the next section.

2.1.3 Correlation Fields

There are many techniques designed to highlight potentially meaningful correlations in multivariate data to users. Nattkemper reviewed the application of several of these techniques to biomedicine, some of which I describe below (Nattkemper, 2004). Once correlations have been computed (e.g. the ratio of choline to creatine), they can be rendered via single variable DVR or isosurfaces. Principal component analysis and other dimensionality reduction approaches attempt to project higher dimensional data spaces down to two or three orthogonal dimensions that represent the greatest variation in the data. Broersen and van Liere apply PCA to raw spectroscopy data in order to find images that represent the greatest variation in image space and spectrum space (Broersen and van Liere, 2005). They subsequently auto-generate transfer functions for the computed eigenvectors and display the images using standard DVR (Figure 2.5).

This technique is useful for finding the images that explain the most variance in the data, but these images are not necessarily easy to interpret nor is "greatest variance" always the feature of interest. The work reported here proposes an interactive visualization tool that lets the user visually explore different patterns of interest to identify and distinguish different

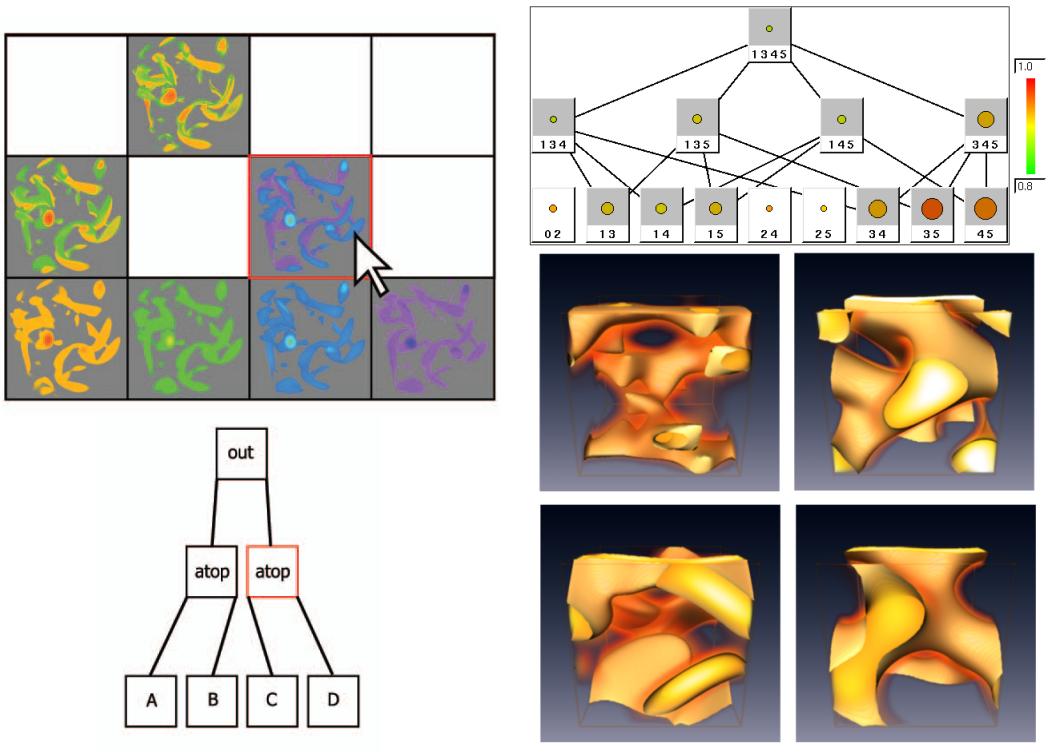


Figure 2.6: Set Operations and Multi-field Graphs. Left: a tool for combining multiple variables into a single rendering using set-like operations (Woodring and Shen, 2006). Right: an exploratory interface for identifying correlated variables and visualizing interesting variable combinations.

classes of multivariate values.

Automatic discrimination techniques (e.g. between tumor and healthy tissue) do not currently apply to MR spectroscopy because radiologists cannot score results. One reason for this is the difficulty in distinguishing dead tissue from necrotic tissue (tissue that appears dead) after surgery. Another is that staining techniques for labeling tumors are known to be inaccurate. Our collaborators require an exploratory view that they can use to generate hypotheses about which relationships matter. Crouzil et al. describe such an interface that uses gradient alignment of different scalar fields to display multiple correlations (Crouzil et al., 1996). The right frame of Figure 2.6 depicts multi-field graphs, which provide an exploratory interface to the large number of potential correlations and use DVR to display a volume of interest in those correlations (Sauber et al., 2006). As shown in the left frame of Figure 2.6, Woodring and Shen have designed a system in which users can combine an

arbitrary number of scalar fields using various set operations (e.g. AND, OR, XOR, etc.) to generate an interactive expression tree (Woodring and Shen, 2006). These correlation detection routines turn the multivariate visualization problem into a guided search through O($mn!$) possible relationships, where n is the number of variables and m is the number of relationship types. With large numbers of variables, correlation computation time can be prohibitive. Additionally, rapidly estimating values of individual variables in the display, the second visualization goal described in Chapter 1, is not possible.

Correlation computation techniques result in a single 3D data field that can be easily viewed using DVR. This work complements these techniques by giving users a sense of the raw data values that is lost due to the dimensionality reduction inherent in correlation field computation. SDDS enables the radiologists to explore correlations among five to ten fields, whereas correlation fields allow radiologists to see specific, arbitrarily complex combinations of fields. Also, SDDS can be used to visualize higher dimensional reductions, such as 5-dimensional PCA. This chapter focuses on visualization techniques that display raw data and evaluates their relative abilities to convey relationships to help radiologists during the hypothesis formation stage of analysis.

2.1.4 Glyphs

In sparse glyph visualization techniques, data values are represented via the properties of geometrical shapes (glyphs). The glyphs must be large enough that multiple data values can be represented at once by mapping multiple properties such as shape, size, color, and opacity to variable data values.

The use of sparse volume glyphs for visualization has been actively studied in the field of tensor visualization. Kindlmann et al. describe the use of superquadrics and other shapes for glyph-based tensor visualization (Kindlmann and Westin, 2006). Superquadrics like those shown in the left of Figure 2.7 have two independently variable shape parameters (α and β). Kindlmann uses a subset of the glyphs (shadowed in the figure) to visualize the anisotropy of flow in diffusion tensor images. Completely isotropic flow is represented as a spherical glyph which is deformed as the flow becomes more directional. Kindlmann's work is primarily based

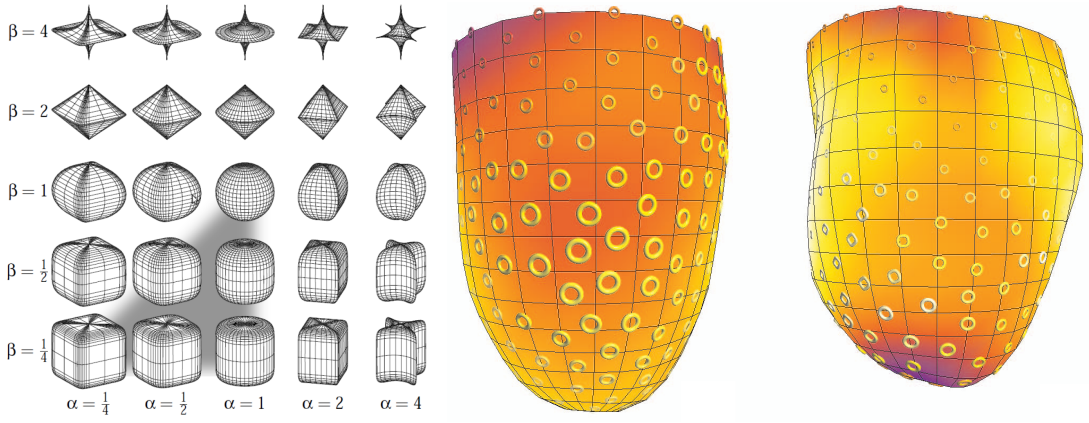


Figure 2.7: Two examples of 3D glyphs. Left: Kindlmann's superquadric glyphs (Kindlmann and Westin, 2006). The shadowed subset were used for diffusion tensor visualization. Right: toroidal superquadric glyphs applied to a time-varying scalar field sampled on a single surface at two time steps (Meyer-Spradow et al., 2008).

in 2D tensor field visualization in which glyph shape varies according to tensor components, although he does show an example of glyphs applied on 3D data sets.

Forsell et al. used 3D glyphs on multivariate data in which they vary properties like specularity and concavity of a 2D surface representing a 2D scalar field (Forsell et al., 2005). An extension of such surfaces to 3D would be to alter surface properties of an isosurface, however there is no single surface to texture in MR spectroscopy data. Meyer-Spradow et al. use superquadric toroidal glyphs to visualize multiple time-varying variables sampled on a single 3D surface (Meyer-Spradow et al., 2008). The data is restricted to a single surface, so issues with depth occlusion and clutter are mediated to an extent. Section 2.4 discusses a comparison of superquadric toroidal glyphs applied to SDDS for MRS data.

Ebert et al. have studied glyph usage for multi-dimensional data visualization, primarily by discussing the different ways of varying shape to convey different scalar values (Ebert et al., 2000). They propose varying color, size, shape, and opacity along separate scalar components. Such an encoding is problematic for the MR spectroscopy data set because using two different encodings for two semantically similar metabolite fields makes relative magnitude estimation difficult. Also, varying shape for one variable and varying size for another does not convey the impression that all scalar fields are of the same modality. More importantly, this scheme ensures that some variables will be harder to interpret than others. To address these issues,

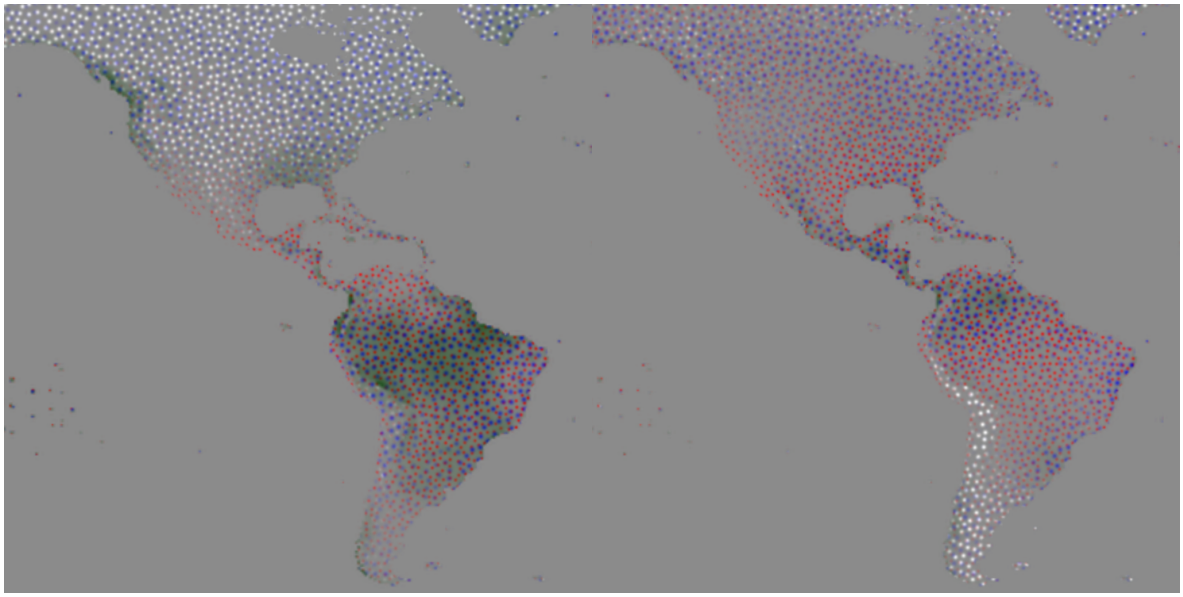


Figure 2.8: Two examples of Bokinsky's Data-Driven Spots applied to geographical classification data (Bokinsky, 2003).

SDDS uses only color to differentiate metabolite volumes. The perceptual differences among colors are smaller than those among shape encodings, as is confirmed in the user study described in Section 2.4.

SDDS is a volume glyph technique that is a 3D extension of Bokinsky's 2D Data-Driven Spots (DDS), a 2D multivariate visualization technique (Bokinsky, 2003). In her work, Bokinsky displays multiple scalar fields using color-encoded Gaussian splats placed on a jittered sample grid and shows that multiple layers of differently-colored spots were as effective for the display of the shape of overlapping 2D scalar fields as direct display of the computed intersection. Examples of DDS appear in Figure 2.8.

BrainExplorer, developed by Lau et al., uses a sphere-based glyph technique similar to SDDS to visualize gene expression in mouse brains (Lau et al., 2008). This technique was developed concurrently to SDDS and uses a similar sphere-based glyph visualization, mapping glyph size to expression level and glyph color to anatomical annotation, gene type, or gene expression level redundantly. SDDS is similar to BrainExplorer when glyph size is mapped to expression level and glyph color separates different genes. The results of the user study described in Section 2.4 should apply to the BrainExplorer work as well SDDS.

2.2 SDDS Visualization Design

SDDS uses the sparse similar glyph visualization methodology developed by Bokinksy and advocated by Taylor (Taylor, 2002; Bokinsky, 2003). In theory, a variable can be encoded into any of a number of glyph properties, including size, shape, color, opacity, sharpness, orientation, packing density, and others. The question is therefore which are the most effective in combination. The principal difficulty with assigning separate variables to perceptually distinct channels like shape and texture is that different variables are will be perceived with different degrees of accuracy. When different visual cues are present in a single image, one will tend to dominate the other (Ware, 2000). Different perceptual mechanisms are used when the visual system searches for color patterns as opposed to texture patterns. Sparse similar glyphs take a different approach. Each variable of a sample gets its own simple glyph, and different variables are nominally encoded into different values of the same visual channel, usually color. The values of variables are encoded into a second channel (e.g., size). For example, rather than two variables getting mapped to the color and size of a single glyph, sparse similar glyphs split that glyph into two simpler glyphs: the color of the glyphs distinguishes between variables and their sizes indicate data values. The result is a more dense set of glyphs for which the different variables are as distinct as their associated colors.

SDDS is a 3D extension of Bokinsky's 2D Data-Driven Spots, a 2D multivariate visualization technique described previously that uses the sparse similar glyph methodology(Bokinsky, 2003). DDS varies the opacity of Gaussian splats based on the magnitude of the value at a position. The 3D extension of SDDS modifies the 2D version by replacing the Gaussian splat with a sphere that varies in size rather than opacity. Figure 2.9 contains an SDDS visualization in front of an anatomical slice plane, showing the negative correlation between choline concentration (yellow spheres) and lesion extent (gray anatomy).

The decision to use an opaque shape was perceptually motivated: Interrante's work on nested surface visualization shows (and informal usability testing confirmed) that partially transparent objects significantly inhibit the human visual system's ability to perceive the depth of objects behind them (Interrante et al., 1997). This was not an issue in Bokinsky's

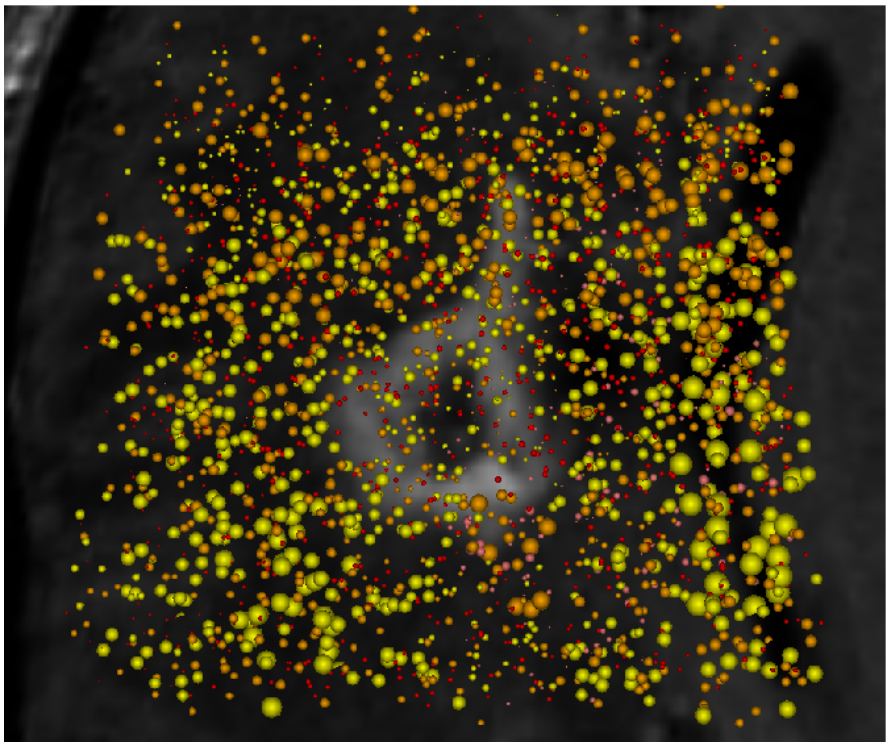


Figure 2.9: SDDS applied to a brian tumor. The yellow spheres show a negative correlation with the lesion extent.

2D visualization, but because depth perception is critical in complex 3D visualizations, opaque shapes are necessary.

The simplest shape to use in place of the 2D Gaussian is the sphere. While cubes and tetrahedra may be easier to draw in graphics hardware, their silhouettes change with different viewing angles. What the viewer perceives in these shapes will depend on their viewing angle. The sphere, on the other hand, is completely symmetric and therefore is equally perceptible at different viewing angles. Encoding data value into sphere size achieves the same goal as varied opacity in 2D DDS. Intuitively, smaller spheres indicate smaller values.

The SDDS technique distributes shaded 3D glyphs throughout the data volume, as shown in Figure 2.9. If the spheres are placed on a regular grid, they are guaranteed to maximize sphere occlusion at viewing angles aligned with any of the grid axes. This regularity can also potentially induce strong aliasing effects, so SDDS resamples the data on a jittered version of the original sample grid. Each scalar variable uses a separately generated jittered grid. The

sphere radius is determined as follows:

$$r_{max} = k\frac{s}{2n} \tag{2.1}$$

$$r = r_{max}\frac{v - v_{min}}{v_{max} - v_{min}} \tag{2.2}$$

where r_{max} is maximum glyph radius, r is the glyph radius for a particular data value v in the range $[v_{min}, v_{max}]$, s is the original sample spacing, n is the number of scalar volumes visible and k is a user-adjustable parameter. When $k = 1$, the glyphs for all scalar volumes at a single sample point can fit within a voxel without overlapping. When sphere glyphs get too large, they begin to intersect with each other, potentially occluding each other unnecessarily. While more advanced techniques for distributing glyphs such as glyph packing are possible to ensure that glyphs do not overlap (Kindlmann and Westin, 2006), in practice this has not been necessary for SDDS.

As with DDS, the viewer distinguishes between different variables in the data set by looking at the color of the sphere. Because the human visual system perceives color preattentively, viewers can easily distinguish the nominally colored variables provided that the colors are sufficiently distinct. Preattentive vision theory, which describes the tasks that the visual system can perform quickly enough to not require focused attention, is discussed in more detail in Section 3.1. Bokinsky found that DDS viewers could visually attend to a single field in the presence of at least 8 other fields (Bokinsky, 2003). Both Ware and Healey have shown that humans can consistently name and differentiate 12 color values (Ware, 2000; Healey, 1996). Discarding one color for use as the background color, 11 therefore serves as a theoretical upper limit of the number of simultaneously renderable variables based on color alone.

One potential source of error is that different colors are perceived differently, primarily based on their luminance. Selecting isoluminant colors for the glyphs would prevent one variable from being easier to locate than another, but results from the user study described in Section 2.4 show that the measured difference is not statistically significant for the analyzed tasks.

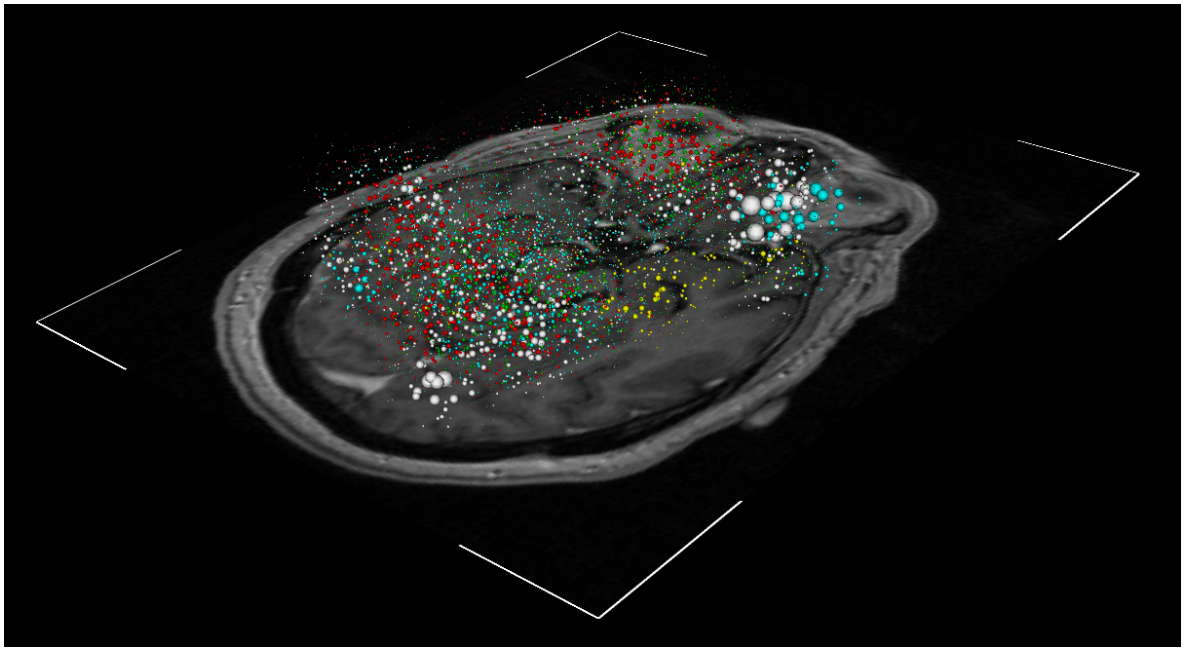


Figure 2.10: SDDS with five variable MRS data (white, yellow, red, cyan, green).

The inevitable increase in visual occlusion resulting from adding a new variable restricts the number of renderable variables more than color. Depending on the desired glyph size, sample density, and data sparsity, rendering more than five to six scalar fields simultaneously begins to cause over-occlusion. The result is that the spheres closest to the viewpoint entirely occlude the spheres behind them. Figure 2.10 shows SDDS with five MRS variables. Notice how it is difficult to pay attention to all of the variables at once, but it is possible to attend on a single color.

2.3 SDDS Implementation

Displaying large numbers of spheres can be accomplished quickly and accurately using an shader-based technique proposed by Taylor (Taylor, 2004). The simplest sphere-drawing technique is to create a polygonal approximation of a sphere, however this can require a large number of triangles per sphere. Taylor's approach is to create a single-polygon stand-in (or impostor) for the sphere that will be used for shading. Rendering a pixel perfect sphere can be done quite simply in a per-pixel fragment shader.

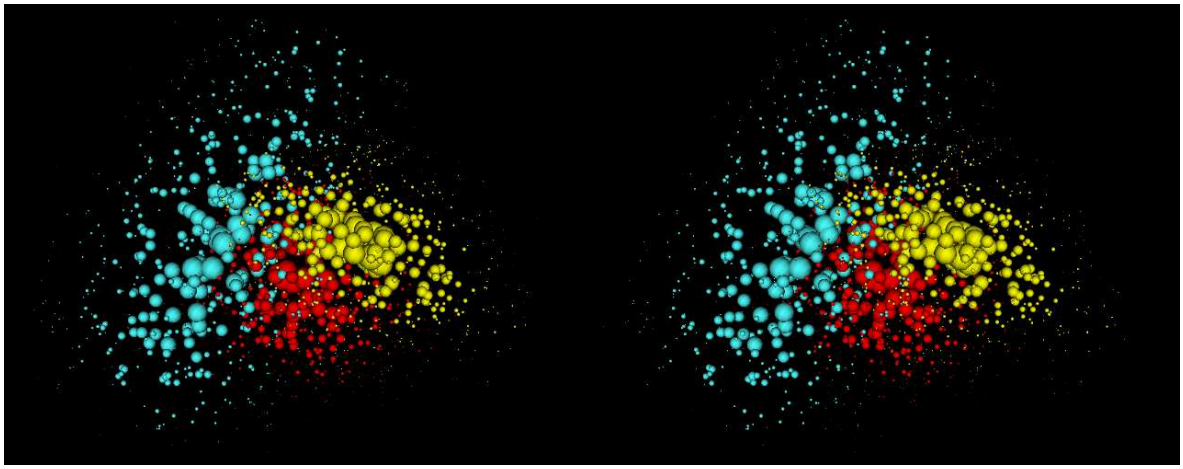


Figure 2.11: A cross-eyed stereo pair of SDDS applied to an artificial data set. Notice how the three fields are easier to separate in depth when viewed in stereo.

The fragment shader essentially computes the distance between a sphere's center and the view ray passing through the current fragment. This distance can be used to compute whether the ray has intersected the sphere. If there is an intersection, the shader then computes the intersection point and the surface normal at the intersection. Taylor advocates a rectangular box for the impostor, however this is only necessary for more complex, asymmetric glyph shapes. A simple screen-oriented quad suffices in the case of the sphere because of its perfect symmetry.

Stereo viewing is an important part of complex 3D visualizations such as SDDS. While stereo is not the strongest depth cue used by the human visualization system, it does contribute to the viewer's ability to distinguish the depths of different spheres in the visualization (Palmer, 1999). Consider the cross-eyed stereo images in Figure 2.11. Viewed individually, it not immediately apparent that the different colored spheres are at approximately different depths. Viewed in stereo, it is immediately clear that the yellow spheres are in front of the blue spheres, which are in front of the red spheres.

2.4 Comparison of Superquadrics to Sparse Glyphs

To evaluate the effectiveness of SDDS, it is necessary to first decide how "effectiveness" can be quantified. SDDS was designed for visualizing MRS data, so the metric is to see how well

SDDS satisfies the original design goals, renumerated below:

1. **Estimate values** of multiple variables in local data regions.

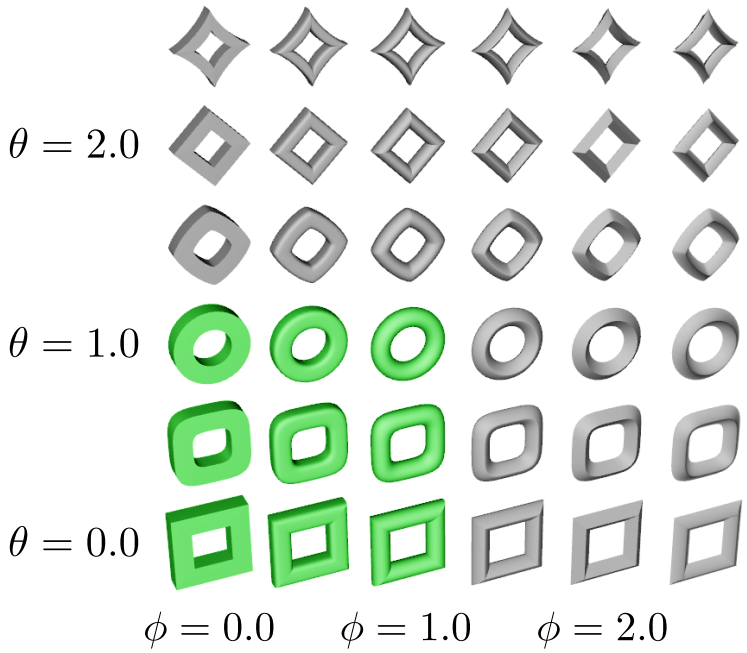
2. **Identify relationships** between data variables.

All participants performed two separate tasks to evaluate the effectiveness of the visualizations for both goals.

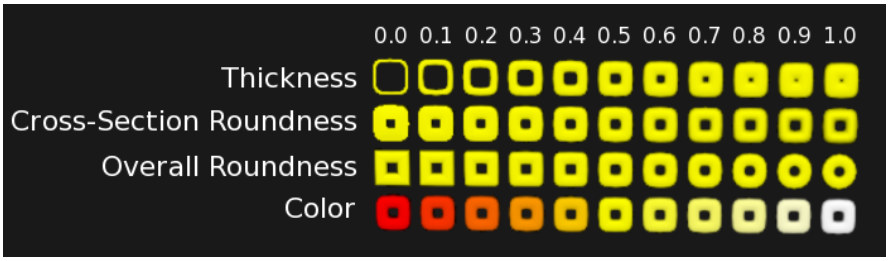
While absolute performance measures are useful, more insight can be derived by comparing SDDS to other, similar visualization techniques. The study therefore also evaluated superquadric glyphs. The study compared SDDS to superquadric glyphs alone because no other current visualization techniques, in our opinion, seem likely to display effectively the raw values of four or more 3D fields. DVR with multi-dimensional transfer functions (Section 2.1.2) and guided exploration techniques (Section 2.1.3) reveal relationships, but they are generally used to visualize a single field representing a single computed relationship of interest. Surface-based techniques described in Section 2.1 can be used to display multiple fields, but surfaces do not display raw data values in regions away from the surface. Compositing multiple monochrome direct volume renderings results in color blending that makes it impossible to distinguish the values of more than two or three variables at once. 3D glyph-based techniques are the strongest candidates for this type of visualization. Superquadric glyphs exemplify the class of techniques that represent multiple variables in a single, more complex glyph.

The study tests both the SDDS and superquadric visualizations on four-variable data sets. This number was chosen for two reasons. First, the radiologists whose needs drove the design of SDDS are primarily interested in four metabolites: choline, creatine, NAA, and lipids. Second, toroidal superquadric glyphs have four natural variable properties: thickness, overall roundness, cross-section roundness, and color. SDDS can potentially accommodate more than four variables, but additional properties for the superquadric glyphs (e.g., size) would interfere with the other properties. For SDDS, the four variables were given four sphere colors: red, yellow, green, and cyan. These approximately match the basic color opponency channels in the human visual system. Cyan was chosen instead of blue because the dark background

color has maximimum contrast with high luminance colors. Pure blue contributes only a small amount to the luminance channel of our visual system, so a bright blue color is more appropriate.



(a) Superquadric Tori



(b) Study Stimuli

Figure 2.12: Examples of superquadric tori used in the study. a) Superquadric tori for different values of ϕ and θ. The highlighted set in green was used in the study. b) The legend shown to participants during the study, where the channels aside from the varying channel were held fixed at .5.

The range of glyphs for each variable property of the superquadric glyphs was chosen to maximize dynamic range and minimize potentially confusing perceptual artifacts. Toroidal thickness ranged from thin but visible to thick but open. The color variable used a truncated black-body radiation color map with black removed so as to not conflict with the dark

background. This resulted in a red-yellow-white color map, which is perceptually ordered and has controlled luminance variation. There is a clear trade-off between dynamic range and intuitiveness in the roundness variables ϕ and θ, and there are no fixed rules for deciding upon a correct set of glyph shapes. The set of superquadric tori with different values of ϕ and θ are shown in Figure 2.12a. Even though both variables can continue beyond completely round shapes to produce concave shapes, this range was not included because the transition from round to concave is a perceptual discontinuity that does not correspond to features in the data. The selected range of ϕ and θ values is highlighted in green in Figure 2.12a, and the complete shape legend presented to participants is shown in Figure 2.12b.

The SDDS and superquadric visualizations used in the study were designed to contain exactly the same amount of information. Each SDDS sphere displays the value of a single variable, whereas the superquadric torus displays four variables at once in a single glyph. The SDDS visualizations therefore had four times as many glyphs as the superquadric visualizations per unit volume. Similarly, the superquadrics were scaled so that they were approximately the size of the voxel they represented. This meant that the superquadric glyphs were approximately four times as large as individual spheres by volume. This highlights an important trade-off in glyph design: breaking the variables into separate glyphs means that simpler, smaller glyphs can be used, but more glyphs are required.

The data used in the study was computer-generated with properties similar to MRS data, rather than actual MRS data. There are two reasons for this: first, it is much simpler to ensure that the necessary properties to be tested exist in computer-generated data; second, there were few real data sets available at the time of the study, and it would be difficult to generalize results based on only a few data sets. In order to mimic the MRS data as closely as possible, the generated data contained slow spatial variation and similar resolution. Each data set was a composite of several Gaussian distributions with randomized centers, orientations, and standard deviations. Each trial consisted of four random pulls from a set of 50 such data sets. The randomized order of data sets was different for each participant. After randomly selecting either SDDS or superquadrics as the initial visualization type, visualizations alternated between SDDS and superquadrics.

Stereo is a very useful depth cue for complex 3D visualizations such as these, so user study participants wore shutter stereo goggles to see a 3D visualization. An equally useful depth cue (if not more useful) is motion parallax, so viewers were also given the ability to move and rotate the visualization camera along a fixed rocking path and with fixed speed. Users could not directly manipulate the camera (for example, using the standard mouse/trackball interaction style) because a participant's previous experience with the mouse and user interface could potentially confound results. Instead, they toggled camera motion on and off by pressing a button.

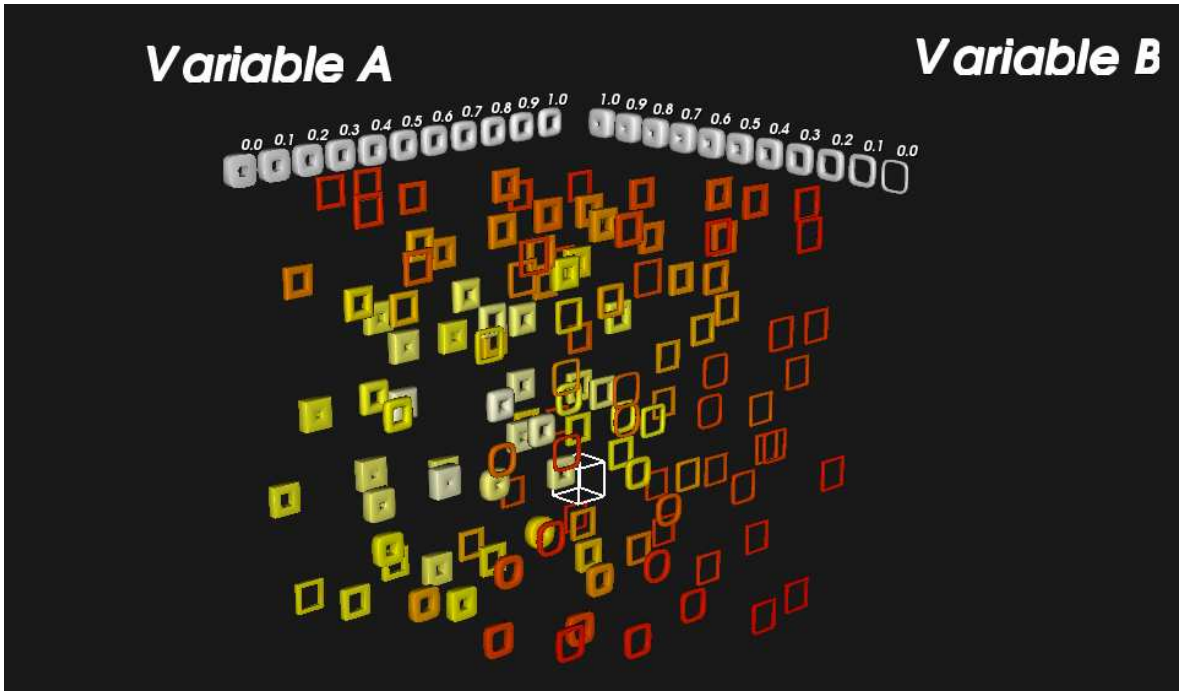
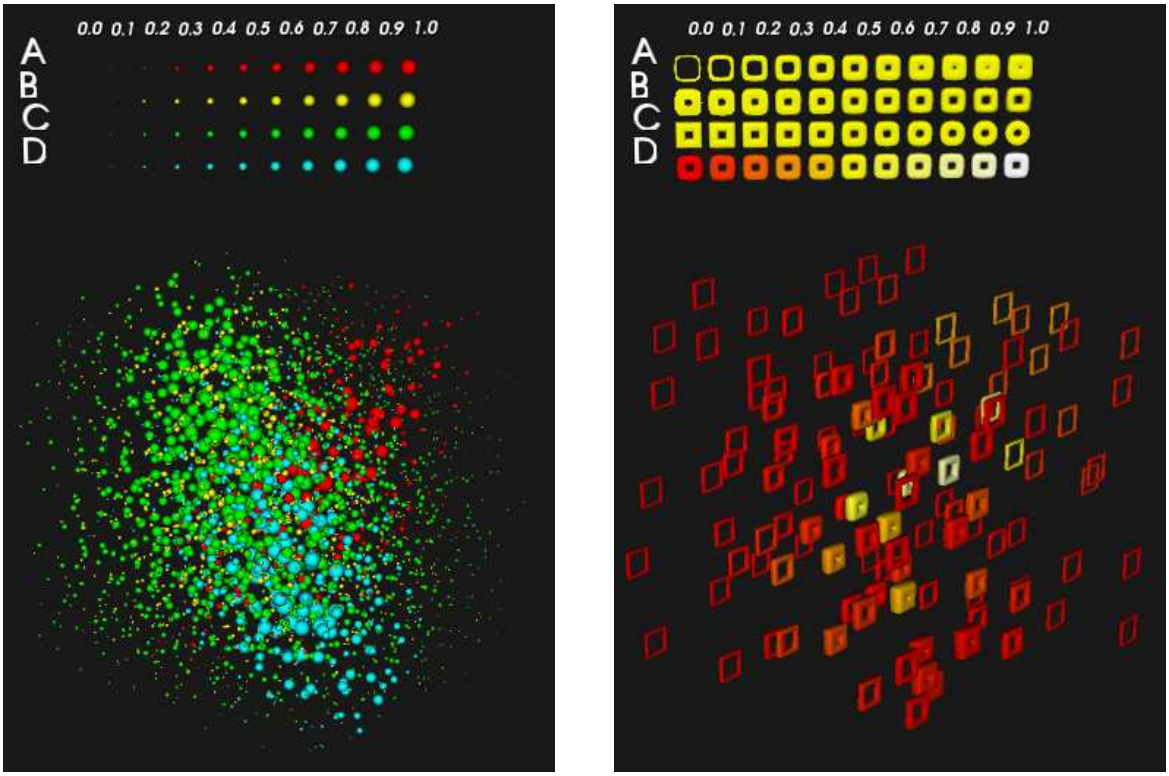


Figure 2.13: An example visualization of the type participants saw during the value estimation task. The participant first estimated the value in the cube for variable A, as designated by the labeled 3D legend, then variable B.

2.4.1 Task 1: Value Estimation

For the first task, participants viewed alternating SDDS and superquadric glyph visualizations and were asked for each to estimate the value of two of the four visible variables at a particular region in space. Participants were allowed to enter discrete multiples of .1 ranging from 0 to 1. This implies that the mean expected error for either visualization technique is 0.025. The

region of interest was labeled using a white wireframe cube. Because this is an inherently 3D task, the visualizations also contained a 3D legend for the two variables of interest that rotated along with the glyphs, as shown in Figure 2.13. In the superquadric legend, all variables except for the variable of interest were set at their middle value (0.5). For this task, response accuracy and response time were recorded.



(a) SDDS (b) Superquadrics

Figure 2.14: Example visualizations of the type participants saw during the correlation identification task. The participant estimated which two variables, demonstrated in the 3D legend above the data, were positively correlated with each other. The same data is used in both visualizations, where variables B and C have a strong positive correlation. When viewed in stereo and in motion, the images are sparser and clearer.

2.4.2 Task 2: Correlation Identification

The second task consisted of participants identifying the two variables present in the data that contained a strong positive correlation. The data for three of the variables were randomly selected from a data set generated as described above. One of those three variables was

subsequently selected at random to be uniformly scaled down, which then became the fourth data set. This ensured that any accidental positive correlations of the other variables would not be as strong as the perfect correlation. As with Task 1, a 3D legend was presented to the users, however this legend displayed all four variables at once, as shown in Figure 2.14. For this task, the participant's answer and response time were recorded.

2.4.3 Equipment and Materials

The user study ran on a single computer containing a nVIDIA Quadro FX5600 GPU with 3-pin stereo output. Participants wore CrystalEyes goggles throughout the experiment synchronized with a 21" flat screen CRT refreshing at 120 Hz (60 Hz per two-eye frame). All interaction with the visualization was through a 40-key X-Keys programmable keypad customized for this study. We chose this input device to avoid potential error caused by varying levels of typing experience in the participants. The study was performed in a darkened room with a small desk lamp illuminating the keypad.

2.4.4 Procedure

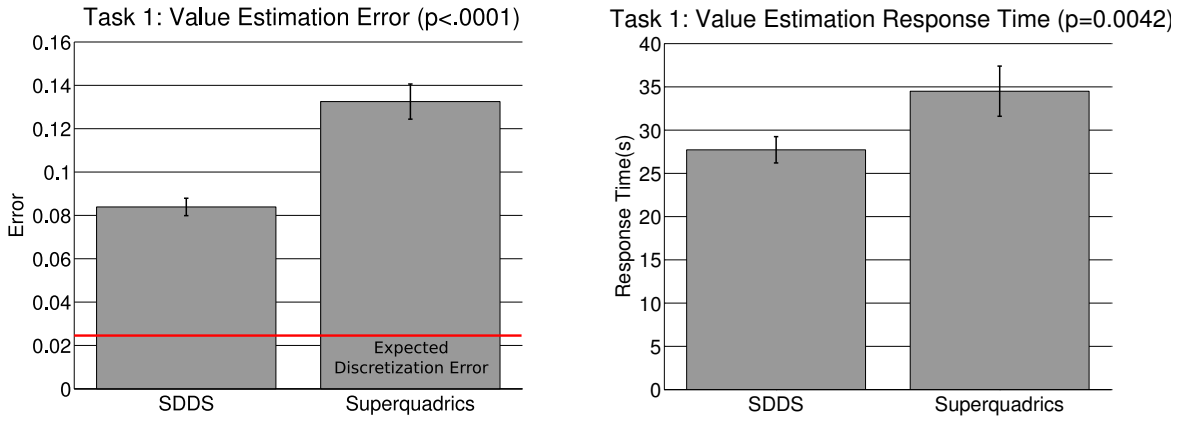
The study included 17 participants (14 male, 3 female, predominantly between 20 and 30 years old). All participants were first led to a private room where they read a short document explaining the data and visualization techniques. They then read about the first task (value estimation) and began an 8 trial training session with no data recorded, mimicking the upcoming recorded session. Participants could ask clarifying questions throughout the introduction and training session. Once the training session was complete, the tester left the room and participants began a 60 trial recorded session on the first task.

Participants could take a short break following their completion of the first task, after which they read a short introduction to the second task (correlation identification). A second 8 trial training session then began with the tester present, followed by 40 timed trials with the tester absent. Participants had fewer trials for this task both to avoid fatigue and because a pilot study indicated that 40 trials would be sufficient to achieve statistical significance. Once participants finished the second session, they filled out a short follow-up questionnaire asking

them for visualization preference, ratings for each task, and general comments.

2.5 User Study Results

The results of the study show that participants were more accurate and responded faster when viewing the SDDS visualization. Response error for Task 1 and response times for both tasks were analyzed using generalized linear models with normal distributions. Because Task 2 error is a binary value (correct vs. incorrect), we used a generalized linear model with a binary distribution. All error bars in the figures correspond to standard error. Participants did not exhibit a significant learning effect in either the value estimation task or the correlation identification task, as is described in more detail in Section 2.5.3.



(a) (b)

Figure 2.15: Value Estimation Error and Timing Results. a) Average value estimation error for the first user study task. Users were 37% more accurate with the SDDS visualization as compared to the superquadric glyph visualization. Expected error due to discretization of possible responses is .025. b) Average response time for the first user study task. Users were 20% faster with the SDDS visualization as compared to the superquadric glyph visualization.

2.5.1 Task 1: Value Estimation

Participants viewing the SDDS condition responded with a mean error of .0839 units, or 8.39% of the value range. Because participants could only enter discrete multiples of .1, perfect accuracy would involve an expected mean error of 0.025, or 2.5% of the value range.

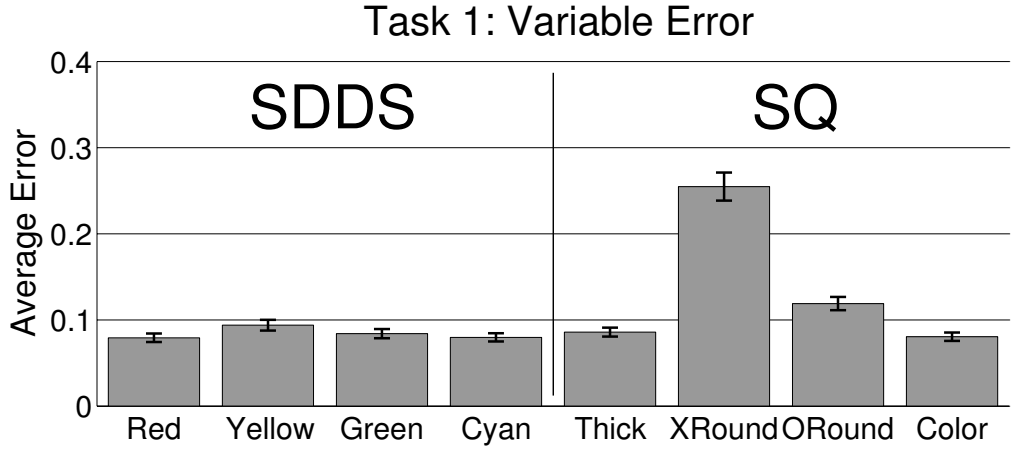


Figure 2.16: Average error all glyph properties. Cross-section roundness and to a lesser extent overall roundness accounted for most of the error for the superquadric glyphs.

The mean response time for SDDS was 27.7 seconds.

Response error was on average lower for the SDDS condition. The mean value estimation error for superquadrics was .1325 units (13.35% of the value range), meaning that participants were 37% more accurate with the SDDS condition. The probability that these two means represent the same distribution (the p value) is less than .0001. Figure 2.15a presents this data along with standard error bars.

Response time was also lower for the SDDS condition. The mean response time for superquadrics was 34.5 seconds, a 20% mean improvement from SDDS to superquadrics. This is the measured time it took a participant to estimate the values of two variables, so we infer than the mean response time for estimating a single value is 13.4 seconds for SDDS and 17.3 seconds for superquadric glyphs.

Whereas the error distributed across the different SDDS colors was not statistically different, the value estimation error for the superquadric glyphs was weighted heavily to the cross-section roundness variable and to a lesser extent the overall roundness variable. The differences between these conditions were statistically significant, as the ANOVA reported in Table 2.1. The mean error for the SDDS colors confirms that these four colors, despite having different luminance, do not significantly differ in terms of viewer performance. The mean error per variable/glyph property is shown in Figure 2.16.

Repr. I	Repr. J	Mean Diff. (I-J)	Std. Error	Sig	95% CI for Diff. Lower	95% CI for Diff. Upper
Red	XRound	-.175	.016	<.001	-.225	-.124
	ORound	-.039	.010	.003	-.070	-.008
Yellow	XRound	-.161	.016	<.001	-.210	-.111
Green	XRound	-.169	.016	<.001	-.219	-.119
	ORound	-.033	.010	.028	-.064	-.002
Blue	XRound	-.173	.015	<.001	-.222	-.124
	ORound	-.037	.010	.005	-.067	-.006
Thick	XRound	-.166	.016	<.001	-.216	-.116
XRound	ORound	.136	.017	<.001	.083	.189
	Color	.175	.016	<.001	.123	.227
ORound	Color	.039	.011	.007	.006	.072

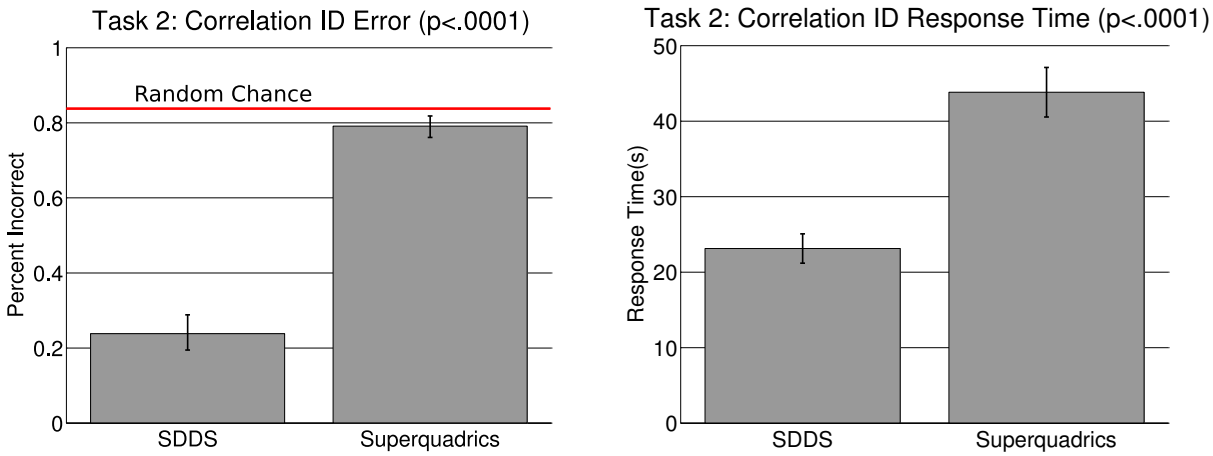
Table 2.1: Significant Results for Variable-specific Estimation Error

Mauchly's test indicated that the assumption of sphericity had been violated for this test, $\chi^2(27) = 411.24, p < .05$, therefore degrees of freedom were corrected using Greenhouse-Geiser estimates of sphericity ($\epsilon = .56$). The results show that value estimation error was significantly affected by the glyph representation, $F(3.92, 890) = 56.72, p < .05$. Bonferroni post hoc tests revealed a significant difference between multiple conditions, enumerated in Table 2.1. All unreported comparisons did not show a significant difference.

2.5.2 Task 2: Correlation Identification

For the SDDS condition, participants incorrectly identified the correlated pair of variables on 23.8% of the trials with a mean response time of 23.1 seconds. Participants viewing the superquadric condition incorrectly identified the correlated pair 79.1% of the time with a mean response time of 42.8 seconds. Comparatively, participants were 70% more accurate and 47% faster under the SDDS condition, both with $p < .0001$. This comparison is shown in Figure 2.17a. The error bars on Figure 2.17a are asymmetric because binary distribution analysis is done in log space.

Participants viewing the superquadric glyph condition correctly identified the correlated pair only slightly more often than chance. For four variables, the probability of randomly selecting an incorrect pair is 84.4%; participants incorrectly identified the correlated pair 79.2% the time. Figure 2.18 shows the mean frequency with which participants selected a



(a) (b)

Figure 2.17: Correlation Identification Error and Timing Results. a) Average correlation identification error for the second user study task (lower is better). Users correctly identified the correlated pair 70% more often with the SDDS visualization as compared to the superquadric glyph visualization. b) Average response time for the second user study task. Users were 47% faster with the SDDS visualization as compared to the superquadric glyph visualization.

variable that did not belong to the correlated pair. The four SDDS color variables were chosen incorrectly at a uniform rate, whereas the error for the superquadric properties is unevenly distributed. Participants tended to erroneously select data sets represented by color for the superquadric glyphs far more often than they should have. Responses from the follow-up questionnaire confirm the fact that participants often began looking at color first because it was the easiest property to see. This user-reported inequality of perceptibility in superquadric shape properties has a significant effect on what variable relationships viewers perceive.

		Mean Diff.			95% CI for Diff.	
Repr. I	Repr. J	(I-J)	Std. Error	Sig	Lower	Upper
Red	XRound	-1.412	.354	.030	-2.736	-.087
	Color	-3.059	.627	.005	-5.404	-.714
Yellow	Color	-3.235	.621	.002	-5.560	-.911
Green	Color	-3.059	.661	.008	-5.532	-.586
Blue	XRound	-1.588	.384	.022	-3.024	-.152
	Color	-3.235	.694	.007	-5.831	-.639

Table 2.2: Significant Results for Variable-specific Correlation Misidentification

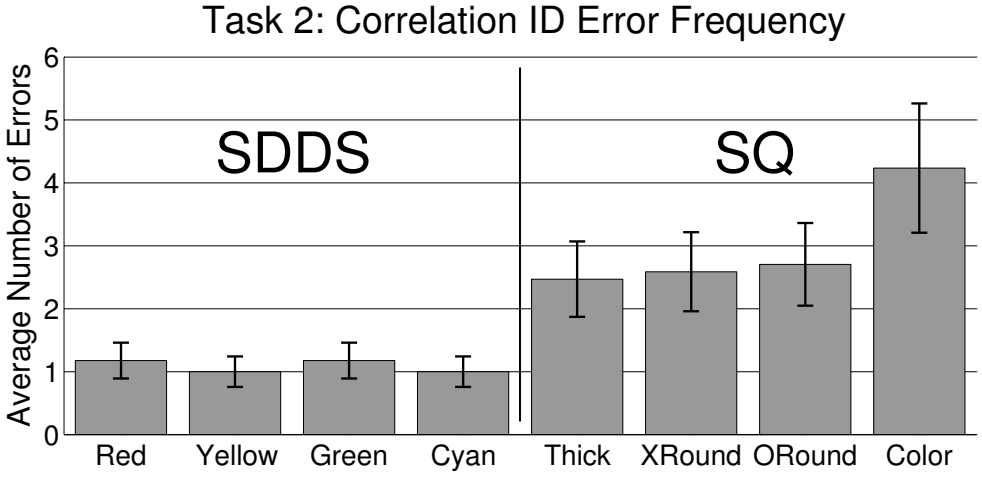


Figure 2.18: Average number of incorrect identifications per glyph property for the second user study task. The sphere glyph properties all exhibit uniform error distributions. The color property of the superquadric glyphs was chosen incorrectly more often than the other shape properties.

Mauchly's test indicated that the assumption of sphericity had been violated, $\chi^2(27) = 68.04, p < .05$, therefore degrees of freedom were corrected using Greenhouse-Geiser estimates of sphericity ($\epsilon = .55$). The results show that value estimation error was significantly affected by the glyph representation, $F(3.84, 61.4) = 7.5, p < .05$. Bonferroni post hoc tests revealed a significant difference between multiple conditions, enumerated in Table 2.2. All unreported comparisons did not show a significant difference. Color did not show a significant difference relative to the other superquadric variable representations. Adding more participants to the study may have added enough statistical power to differentiate those means.

One potential concern with SDDS is that the sphere colors have different luminance and thus are perceived differently. For the correlation identification task, the differences between mean identification errors for the four SDDS colors do not appear to vary with color luminance, nor do they appear to vary significantly.

2.5.3 Analysis of Learning

Analysis of the average error over time, shown in Figure 2.19, indicates that error did not increase or decrease noticeably over time during the study, although this has not be proven

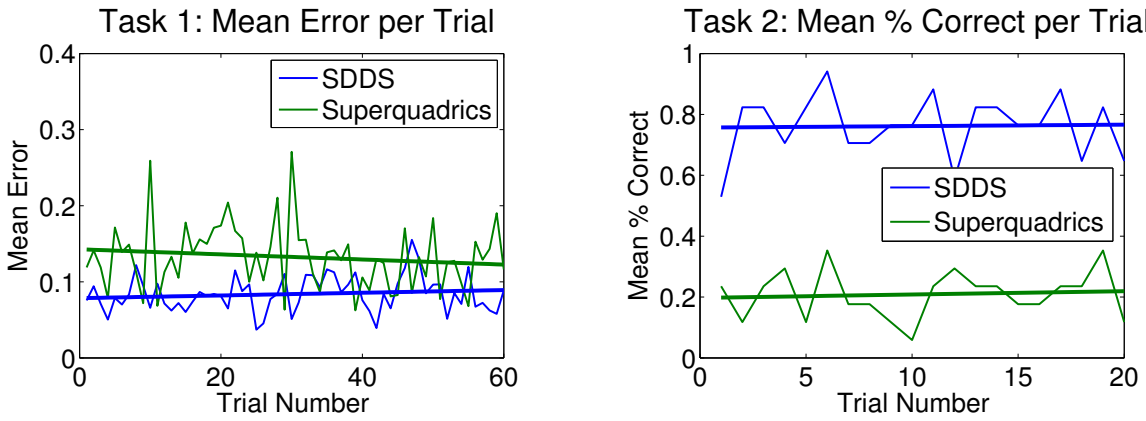


Figure 2.19: Graphs showing average responses per trial to the value estimation task (left) and correlation identification task (right). Neither task exhibited learning by participants during the study, as shown by the linear fits applied to the responses for both spheres and superquadrics.

statistically. This is an indicator that the training sessions successfully mitigated any confounding learning effect.

2.5.4 Follow-up Questionnaire

After completing Task 2, participants filled out a short questionnaire in which they evaluated the two visualization techniques they saw. All participants but two (15/17) preferred the SDDS visualizations over the superquadric glyph visualizations. Reasons for these responses included the simplicity of the color map and value range of the spheres and the complexity and variability of the superquadric properties. One participant preferred the superquadric glyphs for Task 1 because the glyphs were less densely packed, but found SDDS simpler for Task 2. The data for the two participants who state preferences for the superquadric visualization did not support their stated preferences, as both performed more accurately and faster when viewing SDDS visualizations.

Participants also ranked the four superquadric variable properties in order of ease of interpretation. The ranks for each variable are as follows: color ranked 1, thickness ranked 2.23, overall roundness ranked 2.77, and cross-section roundness ranked 4. Participants were in unanimous agreement about the rankings of color as easiest and cross-section roundness as most difficult. Several participants made the comment that the color and thickness variables

"stood out" to them, whereas the roundness variables did not. This agrees with results from studies indicating that color is perceived preattentively (Ware, 2000). If thickness was also perceived preattentively, it may be because thickness increases visual density of glyphs, which is known to be perceived preattentively. Most participants commented on the difficulty of interpreting the cross-section roundness variable, which indicates that these more subtle shape changes may not be perceived preattentively and instead require a serial search. These self-reported perceived preferences support the error measurements taken in Task 1 and Task 2.

Participants also made several recurring general comments about the different glyphs. Many participants required repeated explanations to understand the cross-section variation, and many also said that this property was difficult to understand because of insufficient shape variability. This is reflected by the individual variable errors shown in Figures 2.16 and 2.18. Increasing this channel's dynamic range by including concave shapes could ameliorate this discrepancy.

2.6 Attempts at Uncertainty Visualization

The SDDS visualization technique i a spatial multivariate visualization technique, but it does not take into account the uncertainty associated with the MRS data. Incorporating uncertainty in an already complex and cluttered visualization is challenging. The most obvious option is to map uncertainty to opacity; uncertain values will be less visible and the visualization will be less cluttered. However, as discussed previously, a partially transparent object affects the depth perception of any objects behind it. These spheres would not have been visible before, but arguably they are not more understandable when placed behind partially transparent objects. In practice, the result is a very confusing image, as shown in Figure 2.20.

Issues with transparent surfaces are usually mitigated by introducing texture to the surface. Figure 2.21 contains two images that add opaque contours to the partially transparent spheres. The left image may seem promising: spheres with low uncertainty are easier to see because completely uncertain spheres are invisible. Compare this image to the same type

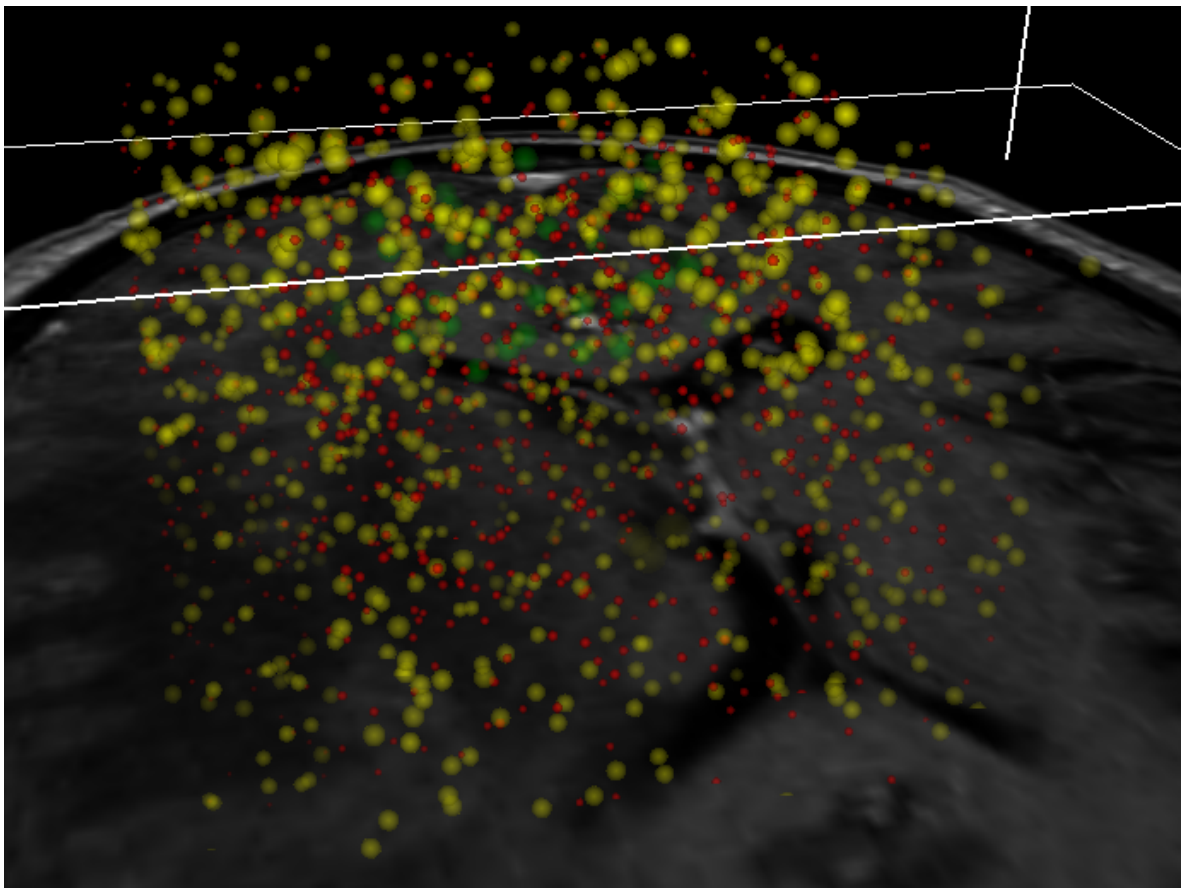
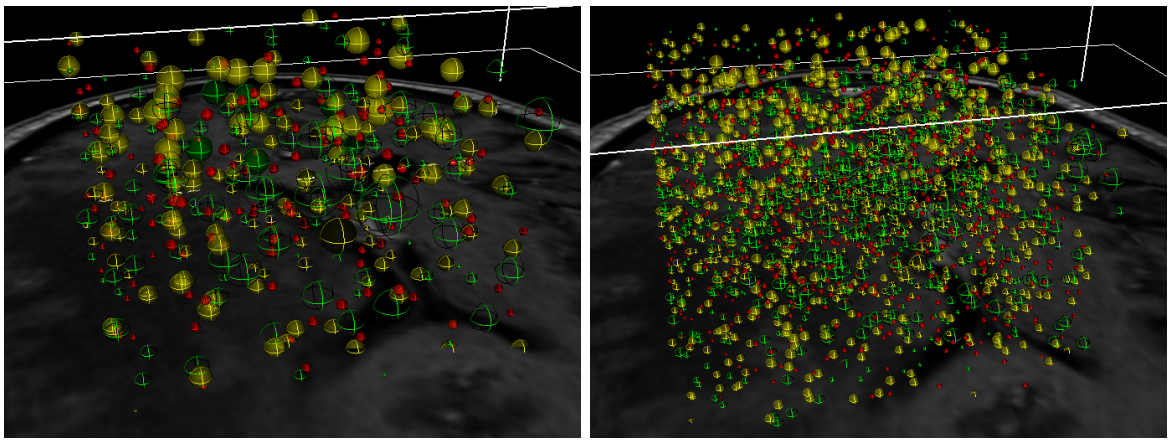


Figure 2.20: An SDDS visualization where data value uncertainty is mapped to sphere opacity.



(a) (b)

Figure 2.21: An SDDS visualization where data value uncertainty is mapped to opacity in a partially transparent texture. a) Low density, larger spheres. b) High density, smaller spheres.

visualization with a more dense set of glyphs to the right. This illustrates a fundamental challenge with glyphs: the more complex the glyph, the larger the glyph must be.

The clear difficulty of uncertainty visualization in spatial multivariate scalar visualizations indicates that a separate view may be more useful for visualizing the uncertainty in the data set. Chapter 3 explores an abstract plotting technique that is meant to complement SDDS and visualize the uncertainty in MRS data.

2.7 Discussion

The user study presented in this chapter evaluates the effectiveness of SDDS for tasks performed on multivariate 3D data sets similar to those used by radiologists studying MR spectroscopy. Effectiveness has been defined as how well viewers could perform two tasks designed to address the original visualization goals: value estimation and relationship identification. The study compared SDDS to superquadric glyphs, an alternative spatial multivariate 3D visualization technique, for context and comparison. Compared to the superquadric condition, participants were significantly more accurate and faster estimating data values and identifying positive correlations with SDDS visualizations. The contrast between the two visualization techniques was particularly dramatic for the correlation identification task, with participants responding nearly twice as quickly and more than twice as accurately.

The study did not necessarily test the ideal set of shape-varying glyph properties or the perfect dynamic range for those properties, nor does it prove that SDDS is always better for these tasks. However, the results do highlight an important effect of choosing variable channels that are not perceptually equivalent. The most common explanation given by participants for why they preferred SDDS to superquadrics was that the superquadric roundness properties were much more difficult to interpret than thickness or color, which was the property that participants understood the most accurately. SDDS uses color in a more perceptually uniform nominal color encoding for different variables. Because humans perceive color preattentively, distinguishing between two colors is generally easier than distinguishing between color variation and shape variation.

From a visualization design perspective, this difference in perceptibility between variable channels results in an emphasis on one variable over another. The difference is unavoidable in the case of techniques like superquadrics: separate channels such shape and color have different perceptual saliences. Having perceptually different channels in a single glyph is not necessarily a bad decision in general. The freedom to manipulate the dynamic range of particular variables that results from using perceptually different channels is useful if one variable is more important than another for the viewer. That said, a study of the parameter space of shape-varying glyphs is necessary to customize dynamic range accurately. The need to perform such a study may prohibit the use of significantly different variable channels in a visualization like the superquadrics described in this work. Note also that for the glyphs chosen for this study SDDS always worked at least as well as the strongest channels in the superquadric glyphs.

One point of minor concern is how exactly viewers perceive the size of spherical glyphs. Acevedo and Laidlaw have studied this problem for circular glyphs and indicate they perceive size by the square root of the circle's radius rather than the radius itself (Acevedo and Laidlaw, 2006). Along the same lines, the perception of size in glyphs is also affected by local size contrast. If a glyph is surrounded by many larger glyphs, it appears to be larger than a similar glyph surrounded by smaller glyphs (Ware, 2000). Studying how size is perceived for spherical glyphs, which now have a volume component, is a potentially interesting avenue of future research.

Glyph-based techniques like SDDS apply most best to data sets with slow spatial variation, as is the case with the MR spectroscopy data set. Applying SDDS to high-frequency data sets will reveal low frequency trends in properly filtered data. SDDS should also work well when visualizing sub-regions of high frequency data.

The SDDS visualization enables viewers to explore and analyze relationships between multivariate volume scalar fields. SDDS is the first known multivariate scalar volume visualization technique that can potentially scale to 11 simultaneous display channels. The complexity of the SDDS visualization makes adding uncertainty extremely difficult. The next chapter will discuss a complementary view that attempts to solve this problem.

CHAPTER 3

Multivariate Uncertainty Visualization in Abstract Plots

(The contents of this chapter were presented in the proceedings at the Conference on Information Visualization 2010 and the Conference on Visualization and Data Analysis 2010 (Feng et al., 2010a; Feng et al., 2010b))

Data uncertainty can have a critical impact on what can be properly inferred from a data set. As an example, consider a simple election data set consisting of the percentage of votes cast for a set of candidates in a region:

Name	East	West	North	South
Betsy	60	30	15	50
Frank	30	10	15	30
John	10	20	20	10
Nancy	10	40	50	10

Table 3.1: An example data of election data (percentage of votes cast), used to illustrate the problems of hidden uncertainty.

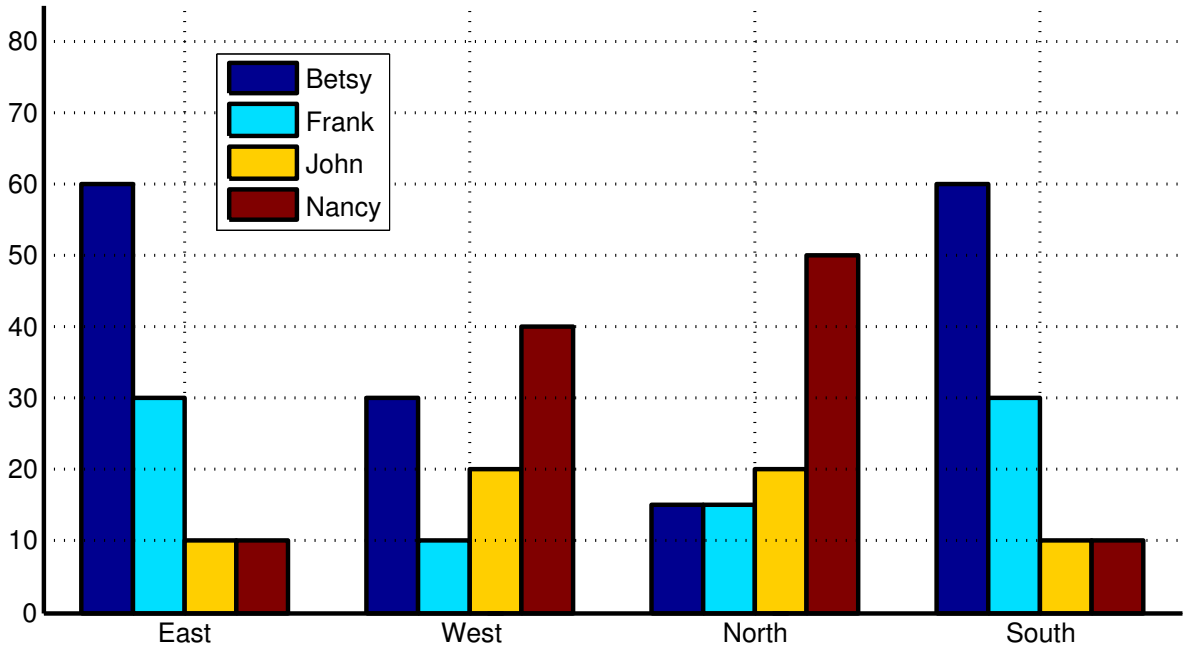
By looking at this table, the casual observer might notice a few interesting facts. First, Betsy carried the east and south regions, while Nancy carried the west and north regions. Second, candidate scores in the east and south are negatively correlated with candidate scores in the north and west. However, what if these are only preliminary results, with only 50% of the votes tallied? In this case, these are only projected results. They will be reported with

error bars, for example ± 20%, since so few of the votes have been counted. In this context the problem is clear: with such large error error bars, conclusions about trends and winners are unreliable. This example is based on simple statistics: if two distributions overlap significantly, one cannot confidently argue that they represent different populations. Bar graphs with and without errors bars of the data in Table 3.1 are shown in Figure 3.1. From a visualization perspective, overlapping error bars indicate the difference between the two means, but as data sets grow in size and complexity, encoding uncertainty into visualizations becomes more challenging.

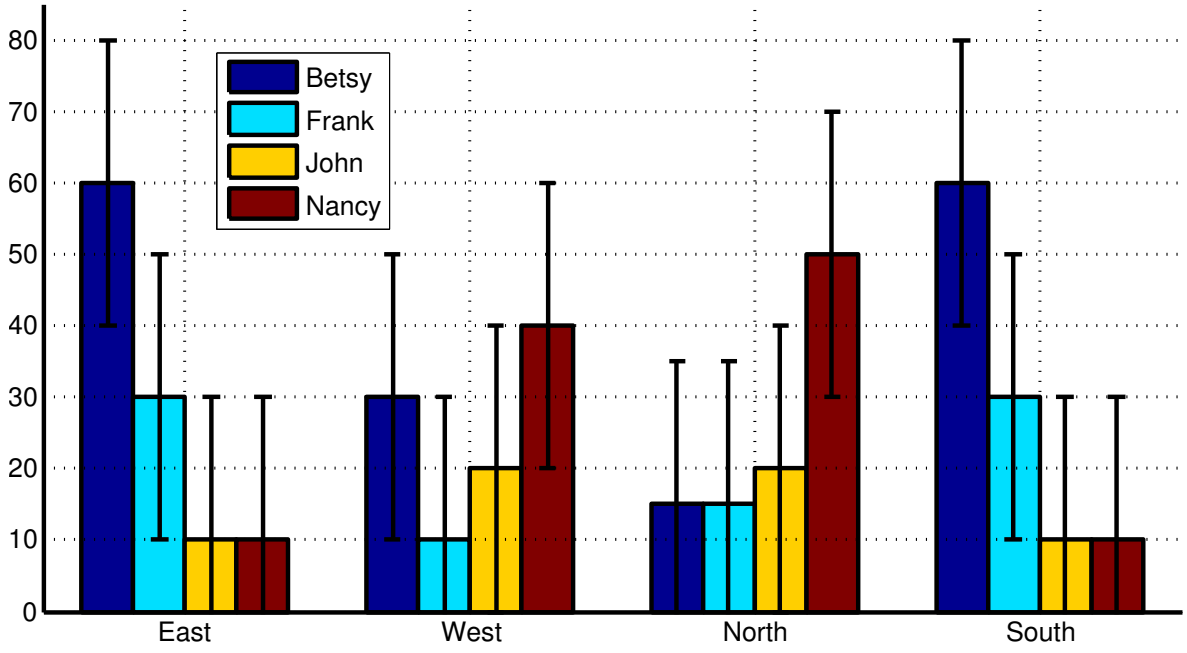
This chapter describes a set of visualization techniques for the display of multivariate data with statistical estimations of data value error. The techniques leverage known characteristics of visual perception so that viewers can preattentively identify patterns in trustworthy values while discouraging viewers from making false inferences based on uncertain values. The goal is not simply to let viewers distinguish certain values from uncertain values, but to disengage the visual system's preattentive feature detection mechanisms for uncertain values. This ensures that a data value's visual saliency grows with its certainty.

The primary contributions of this chapter are novel augmentations of the scatter plot and parallel coordinates plot. These plots are augmented to incorporate uncertainty using density plots. More formally, the plots use kernel density estimation (KDE) to approximate the probability density function (PDF) of the underlying data. The PDF describes the likelihood of each value, so it naturally de-emphasizes unreliable data points in the original data set. One prerequisite to using KDE is that data point uncertainty must be quantified using statistical distributions. For example if the data value can fall anywhere within a range of values, this uncertainty can be modeled using uniform distributions.

Density plots are useful tools for summarizing extremely large data sets. Whereas normal plots become over-plotted with glyphs overlapping each other, PDFs always highlight regions of high density. This chapter demonstrates that density-based plots are useful for visualizing uncertain data sets, large or small. However, density plots have two noticeable problems: data values are no longer always individually identifiable and outliers are de-emphasized. I make two modifications to density plots to address these problems: first, scaling the distribution



(a) Without Error Bars



(b) With Error Bars

Figure 3.1: A bar graph of the data in Table 3.1. Without error bars (above), the viewer may attempt to look for correlations in the results. Overlapping error bars (below) indicate that very little can be learned from the data.

mean to a brighter intensity introduces a discrete, identifiable feature that fades in proportion to its uncertainty; second, a novel animated plot, called the probabilistic plot, cycles through PDF samples so that outliers draw the viewer's eye by intermittently flickering in and out. Confident regions remain stable over time. When summed, these random samples aggregate into a histogram approximation of the fully integrated PDF.

3.1 Background

The design of density plots for uncertain data builds on previous work in multivariate visualization, uncertainty visualization, scatter plots, parallel coordinates plots, and visual perception.

3.2 Information Visualization

The MR spectroscopy data set that drove the design of the techniques reported in this dissertation has 3D spatial coordinates. In contrast to techniques discussed in Chapter 2, the plots in this chapter do not treat spatial data differently from other kinds data values. When visualizing spatial data in abstract plots, it is often simplest to think of the data not as arranged in its spatial coordinate frame, but rather as a simple table of values. If required, the spatial coordinates can be additional columns in the table of variable values:

v_1	v_2	...	v_n	x	y	z
.1	.5	...	.9	0	0	0
.4	.4	...	.5	1	0	0
.8	.8	...	.4	2	0	0
.9	.8	...	.1	3	0	0

Table 3.2: An example multivariate data set with n variables. For information visualizations, position coordinates are often treated like any other variable (if they aren't ignored entirely).

Visualizations that operate on data without spatial coordinates or that ignore spatial coordinates are commonly called *information visualizations*. This chapter describes methodology for incorporating uncertainty into two commonly used information visualizations.

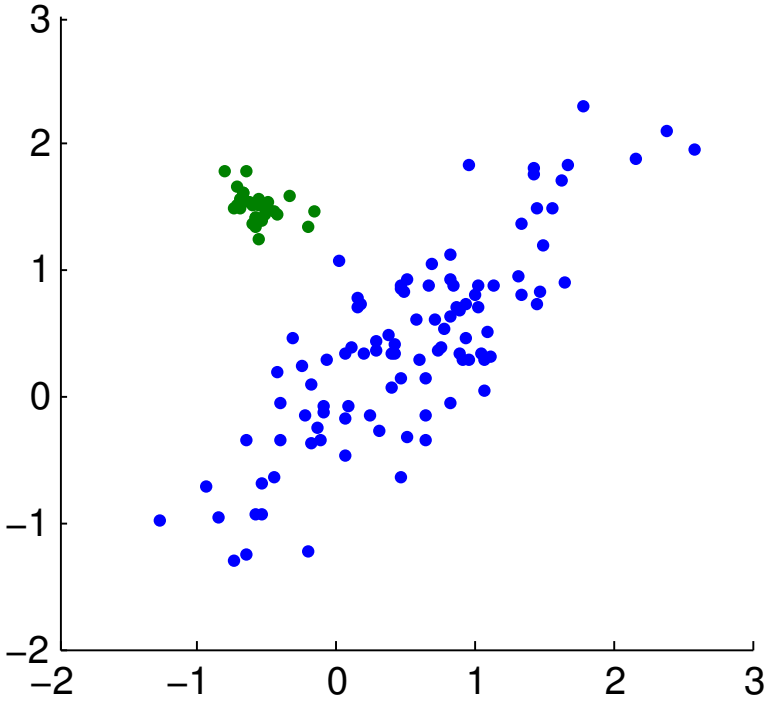


Figure 3.2: An example scatter plot with two interesting features. The green points correspond to a cluster of values. The blue points have a strong positive linear correlation.

3.2.1 Scatter Plots

The scatter plot is a standard technique for graphically representing a bivariate data set that places discrete glyphs on a Cartesian grid. It is commonly used to identify value clusters and trends such as linear relationships. In a scatter plot, the Cartesian coordinates of glyphs correspond to the value of the two (or three, for 3D plots) variables for a single sample. In this way, glyph position encodes the value of the variables. The most common glyph used is an opaque circle. Clustered sets of glyphs are often indicators of interesting sets of values. The shape of the cluster is also important; line-shaped clusters indicate a linear dependence between the two variables for points within the cluster. Figure 3.2 depicts a data set with two classes of points. The green set is a tight cluster of values. The blue set is a set of values that exhibit a linear dependency.

The most straightforward and scalable way to incorporate more variables is to use a scatter plot matrix (Cleveland and McGill, 1988). The scatter plot matrix sacrifices the resolution of a single plot to display more plots comparing other pairs of variables, as shown

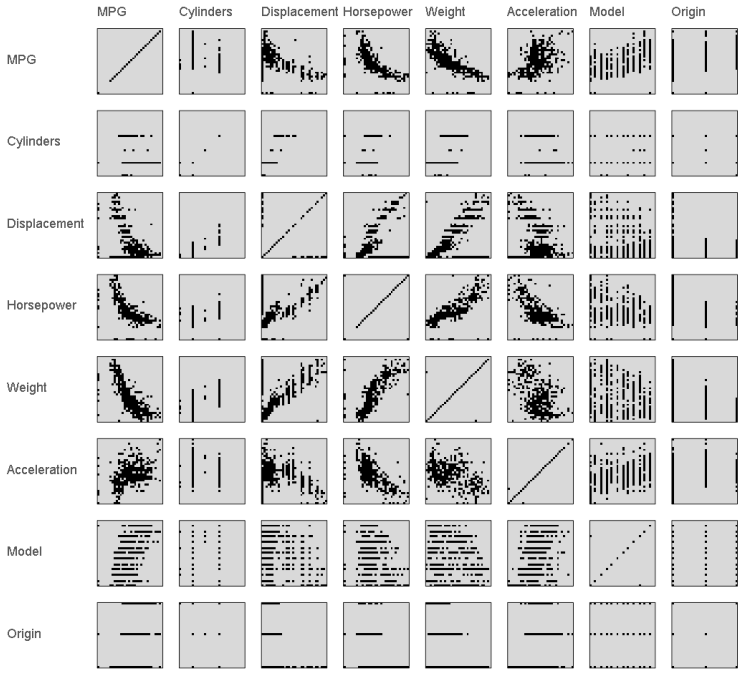


Figure 3.3: An example of a scatter plot matrix, taken with permission from Elmqvist et al. (Elmqvist et al., 2008).

in Figure 3.3. In this instance, both navigating through the plots (Elmqvist et al., 2008) and choosing display order become interesting problems (Seo and Shneiderman, 2004). Because this dissertation describes the use of KDEs in a single plot, it can be applied generally to any scatter plot technique.

Bachthaler and Weiskopf propose a modification to the standard scatter plot for use with data sets that have a spatial coordinate frame (Bachthaler and Weiskopf, 2008). The technique, called continuous scatter plots, leverages knowledge of the sample footprint (e.g. voxel, tetrahedron, etc.) to replace discrete glyphs with a composition of all values contained therein. A glyph from a single voxel becomes a shape representing all of the interpolated values within that voxel. The primary contribution of their work is the transformation of complex geometry from the spatial domain into data distributions, as shown in Figure 3.4.

The techniques presented in this chapter use similar distribution-based scatter plots for visualizing uncertain data. The distributions used in the novel plots come from the statistical uncertainty of the samples themselves rather than spatial interpolation.

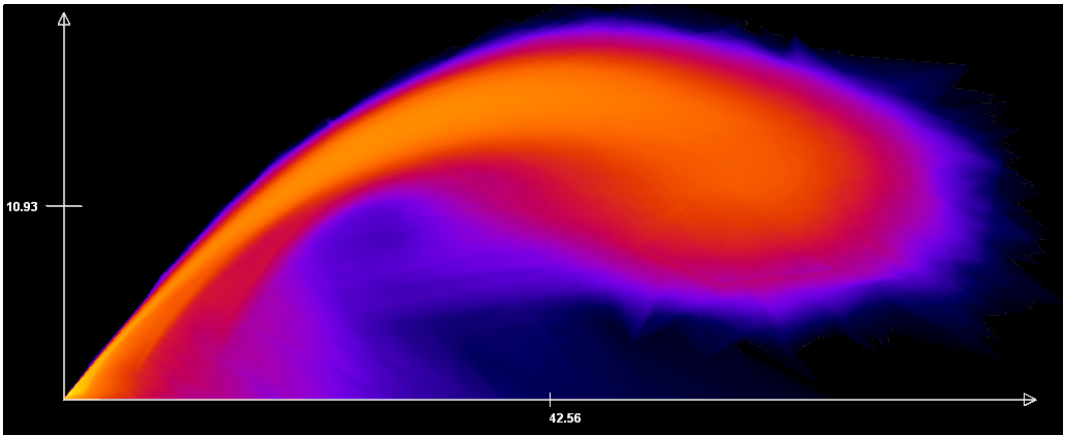


Figure 3.4: A continuous scatter plot. In this image (made by Bachtaler et al. (Bachthaler and Weiskopf, 2008)), data sampled on a spatial grid have been transformed from discrete scatter plot points into shapes that represent the entire footprint of the interpolated values in the voxels.

3.2.2 Parallel Coordinates

The parallel coordinates plot is a multivariate visualization technique that bypasses the two-variable limit of the scatter plot. Invented by d'Ocagne and popularized by Inselberg, parallel coordinates plots arrange individual variable axes parallel to each other and represent individual samples as a line passing through all of the axes (d'Ocagne, 1885; Inselberg, 1985). The entire data set with all variable values is represented in a single plot. These visualizations are useful for identifying clusters and trends between pairs of variables and observing how a collection of lines behaves for all variables. An example of the parallel coordinates plot is shown in Figure 3.5. The X and Y axes in this plot show the same data as the scatter plot in Figure 3.2. The addition of the Z axis shows how parallel coordinates plots extend beyond two variables.

One common concern with parallel coordinates is how it handles large data sets. Over-plotting becomes a problem when there are millions of lines overlapping in the plot. One attempt to address over-plotting in large data sets is to only plot lines representing clusters of values (Fua et al., 1999). The line intensity falls off linearly according to the variance of the cluster. By navigating the hierarchy of clusters, the viewer can see different levels of detail, as shown in Figure 3.6.

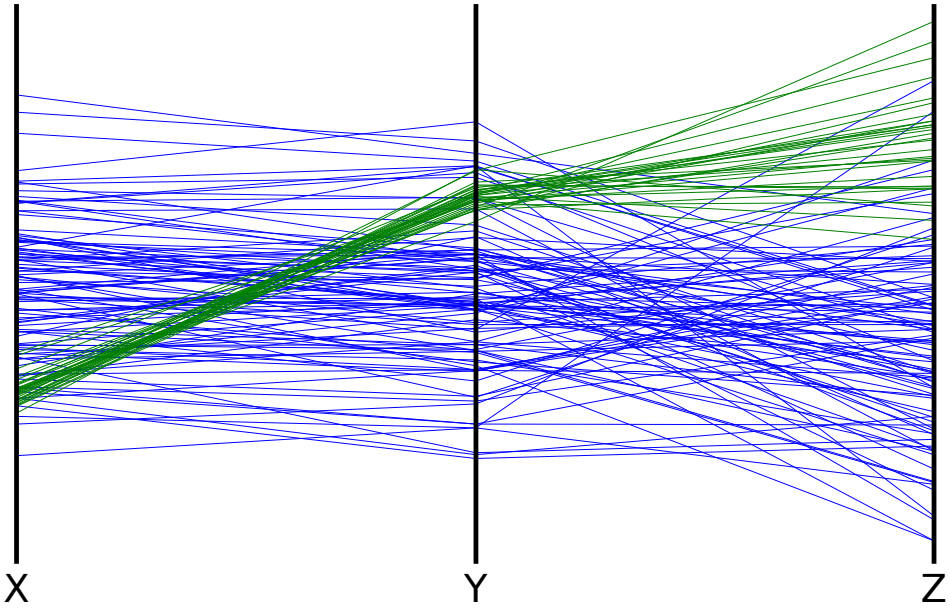


Figure 3.5: An example parallel coordinates plot. This is a plot of the same data shown in Figure 3.2, with one additional variable.

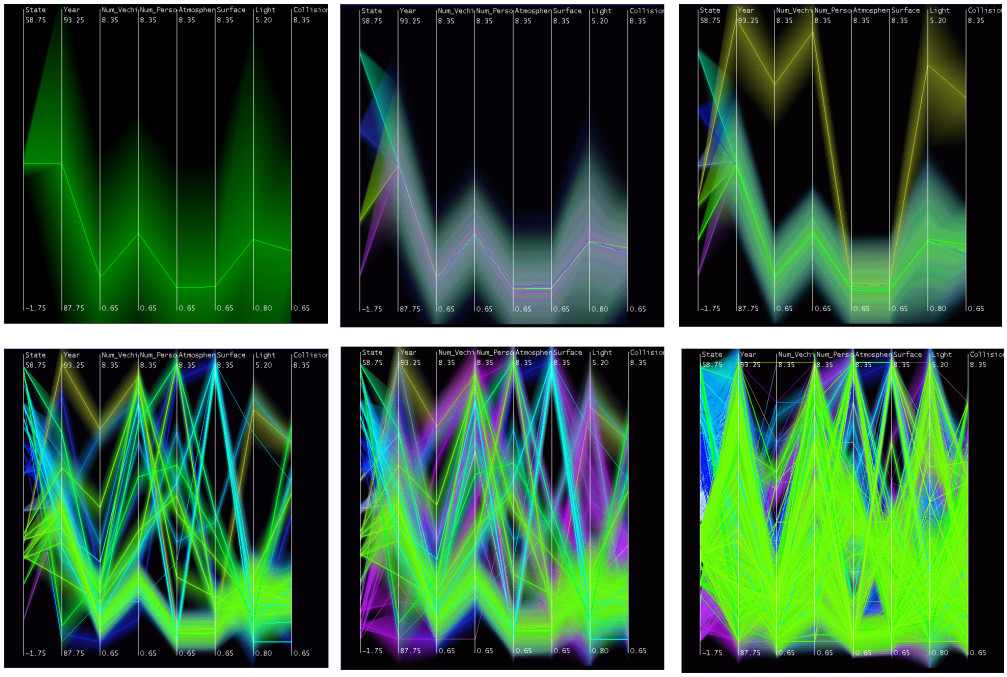


Figure 3.6: Clustered parallel coordinates. For extremely large data sets where drawing individual lines is prohibitively slow, it is possible to perform clustering on the data set and draw shaded bars for the individual clusters. This image was made by Fua et al (Fua et al., 1999).

Novotny and Hauser perform this clustering in 2D histograms for adjacent variable pairs and also identify outliers through histogram analysis (Novotny and Hauser, 2006). Muigg et al. and Blaas et al. describe how to extend this system for use in large, time-varying data sets (Muigg et al., 2008; Blaas et al., 2008). Uncertainty visualization shares many characteristics with such large cluster visualizations. In the case of statistically uncertain data, each data value has an associated distribution, similar to how clusters have means and variances. Cluster visualizations like these inspired the uncertainty visualization techniques described in this chapter.

Heinrich and Weiskopf show how parallel coordinates plots can be made continuous in the same manner as scatter plots (Heinrich and Weiskopf, 2009), described previously. This chapter shows how to use data distributions to visualize statistically uncertain data, complementing their spatial distribution plots. Miller and Wegman describe the analytical solution to evaluating normal distributions in parallel coordinates plots (Miller and Wegman, 1991). The data set that drove the design of the techniques described in this chapter are normally distributed, so Miller and Wegman's analysis is directly applicable. More general data distributions can use Heinrich and Weiskopf's transformation.

3.2.3 Sources and Classifications of Uncertainty

In the context of scientific data, uncertainty can broadly be considered as a lack of confidence in the measured, estimated, or simulated value. Such uncertainty can arise from a number of sources, including:

- **Hardware Limitations:** if the data value is measured by a hardware system, that system may have limited accuracy. For example, a microscope system has a physical resolution limit defined by its components (lenses, wavelength of light, etc.).

- **Simplifying Assumptions:** if the data value comes from a numerical simulation, the algorithm designer may use a mathematical model that does not perfectly reflect the simulated phenomenon. This is often done for tractability: simplifying the mathematical model may make it feasible to simulate an impossible problem.

- **Numerical Error:** repeated mathematical operations can gradually accumulate error over time. Poorly conditioned/unstable numerical systems can produce such errors quickly.
- **Statistical Error:** values estimated statistically often characterize their own error via statistical distributions.
- **Visualization Error:** the chosen visualization will introduce value estimation error. For example, 2D scalar fields visualized with a grayscale color map are known to incur up to 20% estimation error (Ware, 1988). One solution to this error is to choose a more perceptually balanced color map.

The sources of uncertainty are described in more detail by Wittenbrink et al. and Thomson et al. (Wittenbrink et al., 1996; Thomson et al., 2005).

From a visualization perspective, it is just as important to understand how the uncertainty should be interpreted. Pang et al. classify uncertainties into three categories: statistical, error, and range (Pang et al., 1996). Statistical uncertainty can be represented as a statistical distribution. Error uncertainty is a measured difference with respect to a correct value. Range uncertainty prescribes an interval in which a data value may fall, which is similar to a uniform statistical distribution. This chapter describes techniques for visualizing data with uncertainty from any source provided it can be represented using a statistical distribution, which includes uncertainty falling into the statistical and range categories.

3.2.4 Uncertainty Visualization

There are a wide array of proposed techniques for visualization of spatial uncertain scalar fields. For 2D surfaces, Pang et al. survey many techniques for visualizing error uncertainty (Pang et al., 1996). These include modifying surface color, surface orientation, surface specularity, superposition of scaled 2D glyphs, superposition of scaled 3D glyphs, and others to show the difference between one surface and another (Figure 3.7). Grigoryan and Rheingens suggest several techniques for perturbing isosurfaces to show surface position uncertainty (Grigoryan and Rheingans, 2004). Some of these techniques are shown in Figure 3.8.

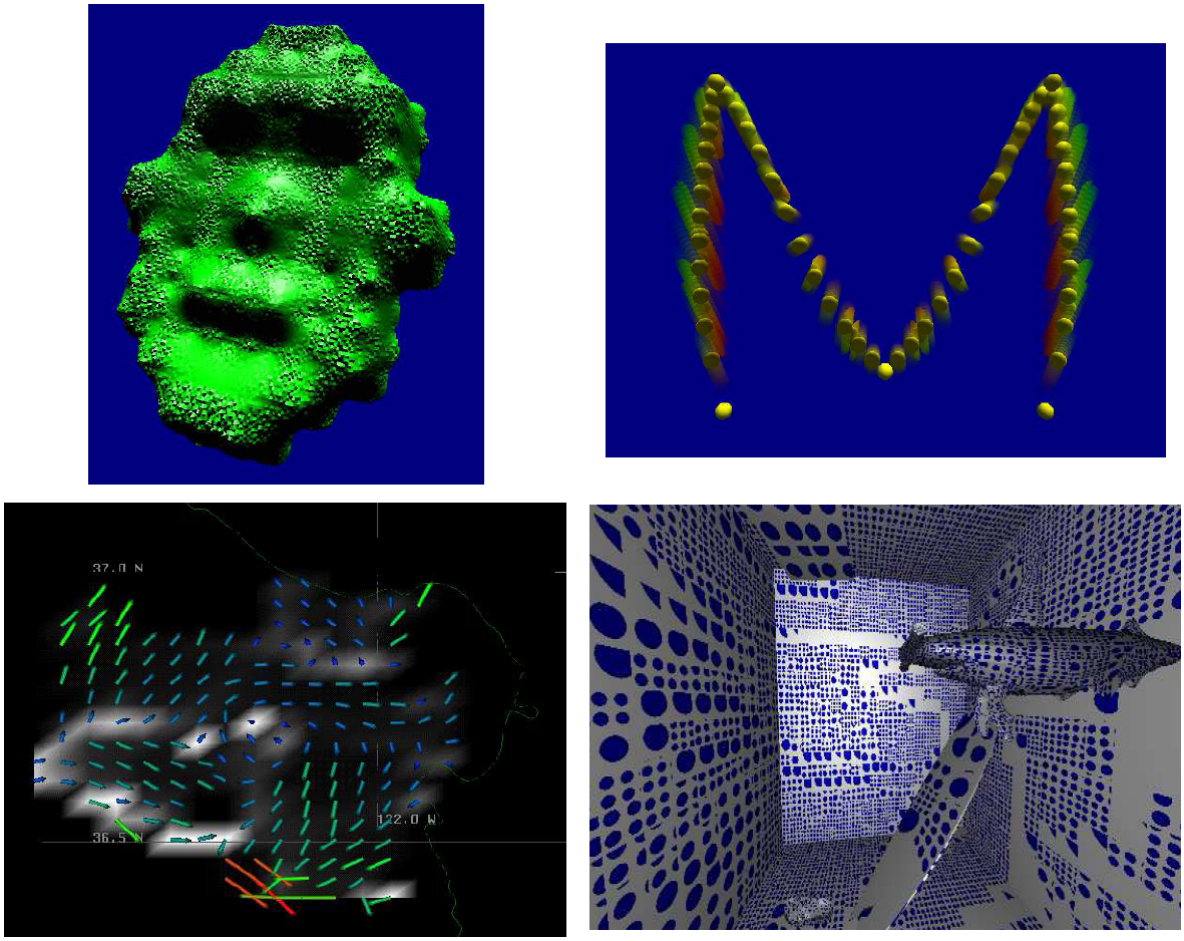


Figure 3.7: Several uncertainty visualizations documented by Pang et al (Pang et al., 1996). Upper-left: surface uncertainty encoded by textured bump-mapping. Upper-right: positional uncertainty encoded with motion blur. Lower-right: shading uncertainty encoded by circle glyphs. Lower-left: vector field uncertainty encoded by glyph color.

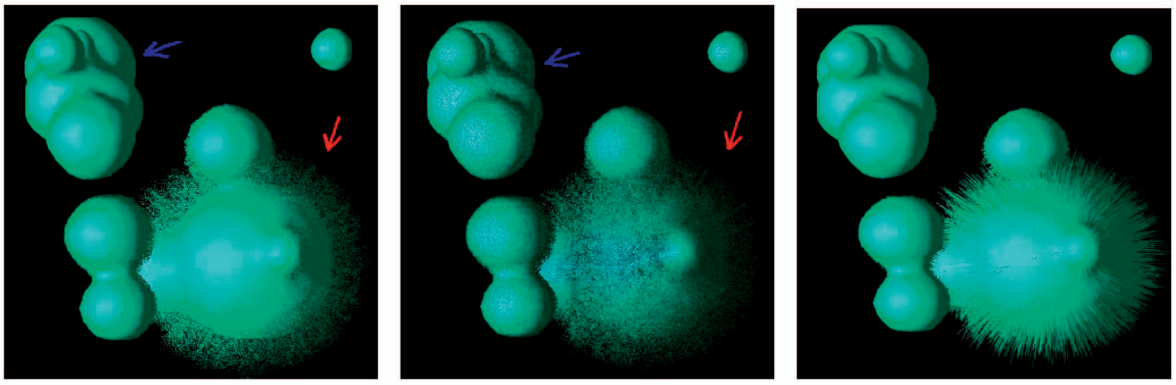


Figure 3.8: Probabilistic surfaces, generated by Grigoryan and Rheingens (Grigoryan and Rheingans, 2004). Surfaces with uncertainty stored at each vertex are replaced by probabilistic points that are displaced along surface normals (center). These points can be mixed with the original surface (left), or be used to displace the original surface (right).

These techniques do not seem to be easily applicable to multivariate data, as the clutter and confusion added by the uncertainty for a single variable would make adding more variables difficult.

Research in information visualization techniques for statistically uncertain data has discussed the use of error bars (Potter, 2006), glyphs (Potter et al., 2008), scale modulation (Sanyal et al., 2009), and ambiguation (Olston and Mackinlay, 2002). While many of these techniques are useful for one-dimensional (1D) data sets, uncertainty annotations can overlap in multivariate plots like the scatter plot and parallel coordinates plot. Over-plotting becomes more problematic as the number of data points grows. Density plots are a scalable solution for visualizing uncertainty in multivariate plots.

3.2.5 Animated Plots

This chapter describes an approach that is closely related to previous work in animated plots. PixelPlexing is an animation-based approach to handling extremely large data sets by emphasizing different sets of randomly selected elements of a visualization over time (Shearer et al., 2008). The probabilistic plots described in this chapter use this basic idea to visualize the distributions present in statistically uncertain data. Fisher's soil map uncertainty visualizations similarly use animation. The data consists of terrain that can have multiple classifications

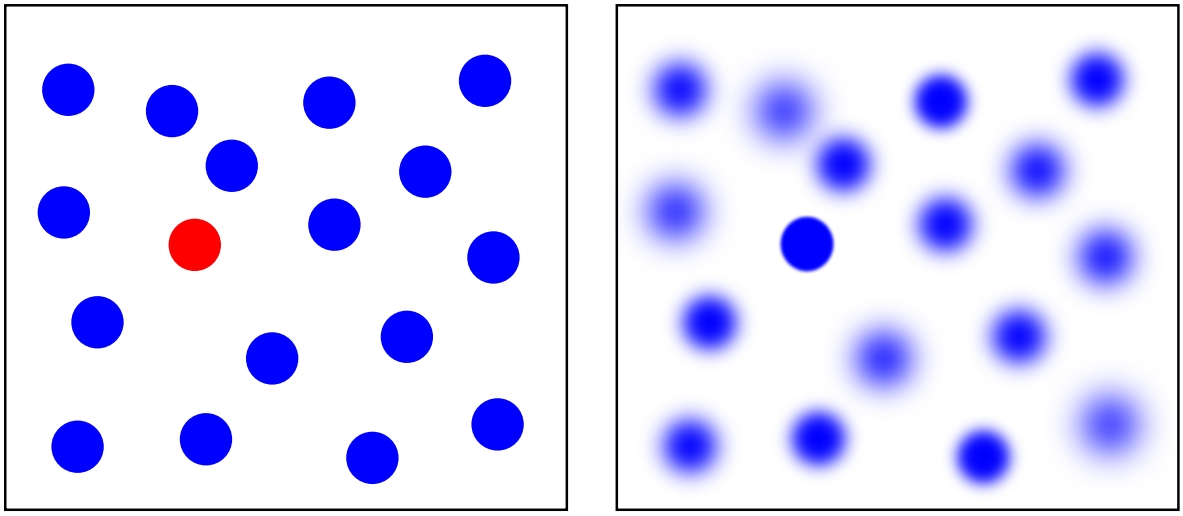


Figure 3.9: Examples of two visual properties that humans can perceive preattentively (hue, blur).

with associated likelihoods. Fisher's technique is to randomly select a classification for visualization at different points in time based on those likelihoods. Probabilistic plots use a similar sampling technique, in this case for data sets in which uncertainty is defined using statistical distributions (Fisher, 1993).

This chapter describes an animated plot related to Fisher's soil map visualizations and PixelPlexing. In Fisher's maps, each pixel can have one of several classifications that has its own likelihood. Over time, a classification is chosen for each pixel based on those likelihoods. Pixelplexing emphasizes different randomized subsets of a visualization over time. The probabilistic plots described in Section 3.3.6 quickly sample random data values from the data PDF in a similar manner and then aggregate the samples into an approximation of the PDF of the data.

3.2.6 Image Structure and Preattentive Vision

Visualizations that use simple primitives like lines and points rely on the fact that the human visual system has low-level physiological structures dedicated to the perception of such features (Palmer, 1999). For example, on-center/off-surround cells identify bright points at different scales, and more complex cells identify edges, lines, and bars at different scales and

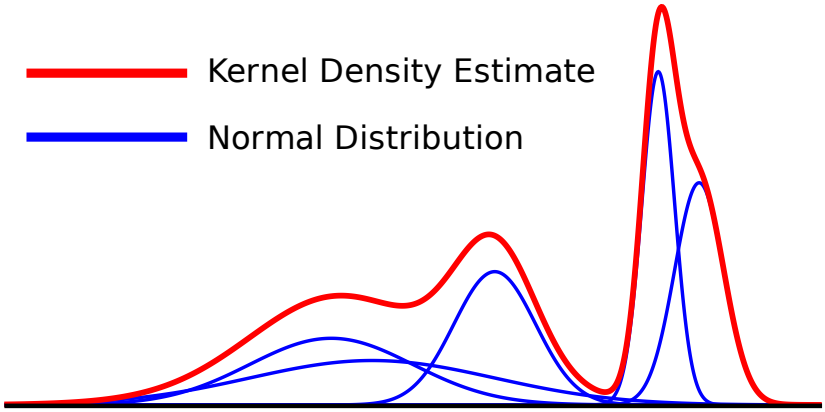


Figure 3.10: An illustration of 1D kernel density estimation. The normal distributions in blue are superimposed to produce the red density estimate.

orientations. These structures contribute to preattentive visual processing, a phenomenon by which humans perceive certain visual properties so quickly (∼200-250ms) that they do not seem to require conscious attention. The list of properties includes hue, size, luminance, and (most relevantly) blur, among many others (Ware, 2000). Visual discrimination tasks for hue and blur are shown in Figure 3.9.

Previous work by Kosara et al. has shown that preattentive perception of blur can be used to obscure irrelevant image features using a technique called semantic depth of field (Kosara et al., 2001). In essence, our proposed methodology for producing density plots uses the statistical distributions of uncertain data as the blurring kernel. Uncertain data values with large distributions are blurrier, so the viewer's attention is drawn to high contrast image features.

3.2.7 Density Estimation

The visualization techniques described in this work use kernel density estimation (also called Parzen Windowing), which is a class of techniques for estimating the probability density function of a data set (Parzen, 1962). For scalar-valued data, KDE superimposes a distribution for each value (Figure 3.10). The distribution width is a user-controlled parameter that determines the feature size in the density estimate.

Histograms are another technique for density estimation that separate the range of data

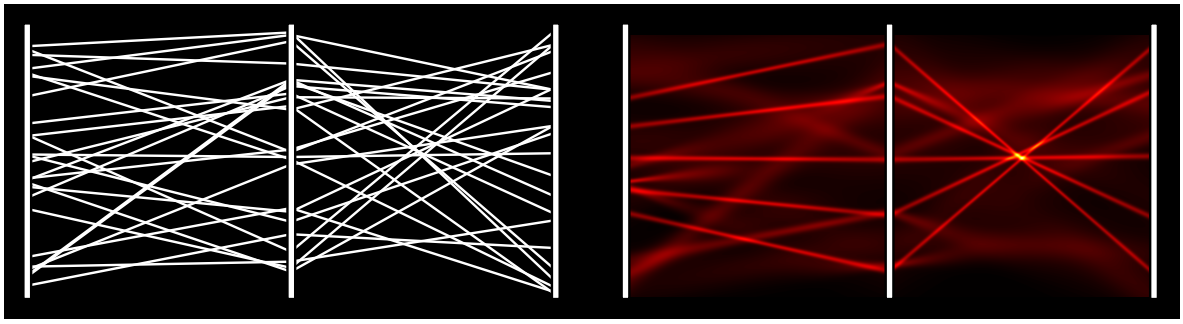


Figure 3.11: A constructed example of a false negative in parallel coordinates. Left: a parallel coordinates plot of three variables. Right: visualization of the PDF using the distributions of values preattentively highlights the more certain values.

values into regularly spaced bins. Each bin stores how many data points fit into its range of values. After normalizing to a total bin area of one, the histogram approximates the probability density function of the data set. KDE and histograms will be very similar when the histogram bin width is close to the KDE kernel width.

3.3 Uncertain Plots

Uncertainty visualizations should prevent viewers from making incorrect observations based on unreliable data. More specifically, they should prevent uncertainty from leading to false positives, in which viewers mistakenly identify a feature, and false negatives, in which viewers fail to identify existing patterns. Figure 3.11 presents a contrived example of a parallel coordinates plot that results in a false negative. The viewer may incorrectly assume that the data is uncorrelated for all variables. The right panel of Figure 3.11 presents a density plot that incorporates uncertainty. Viewed in this way, the remaining lines with high certainty show strong correlation between variables. Figure 3.12 shows a real-world example of a false positive in real MR spectroscopy data: the Glutamine column appears to contain a cluster of values, but the corresponding density plot shows that the cluster is not statistically distinguishable from the rest of the Glutamine values. Visualizations that do not incorporate such uncertainty display put their viewers at risk of making false conclusions. This section presents a method to generate plots that help viewers to preattentively identify trustworthy values while avoiding uncertain values.

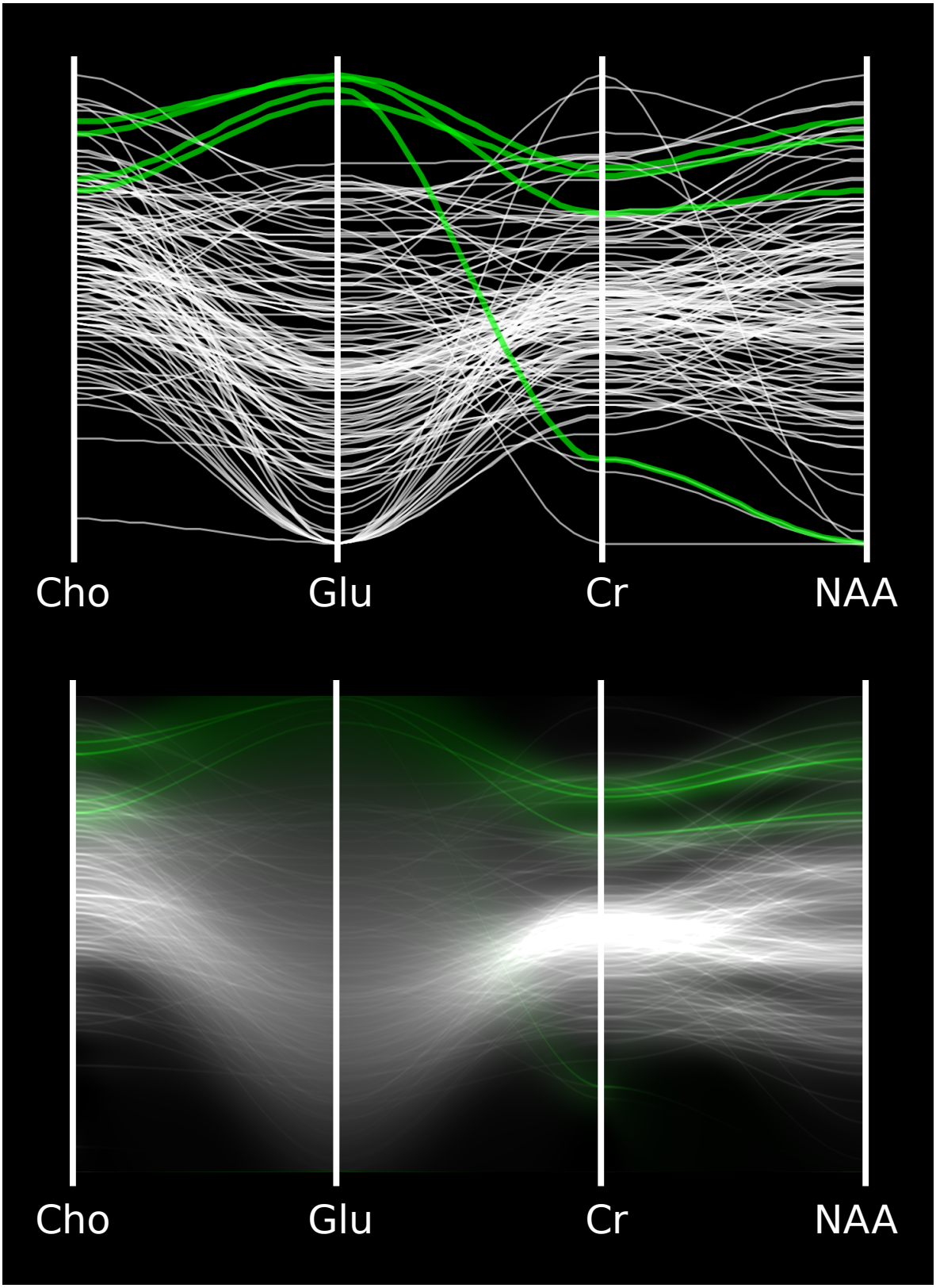


Figure 3.12: A false positive in parallel coordinates. Above: a standard parallel coordinates plot reveals a potential cluster of interesting values on the Glu variable. Below: density plots with mean emphasis reveal that the selection actually is not a cluster when uncertainty is taken into account.

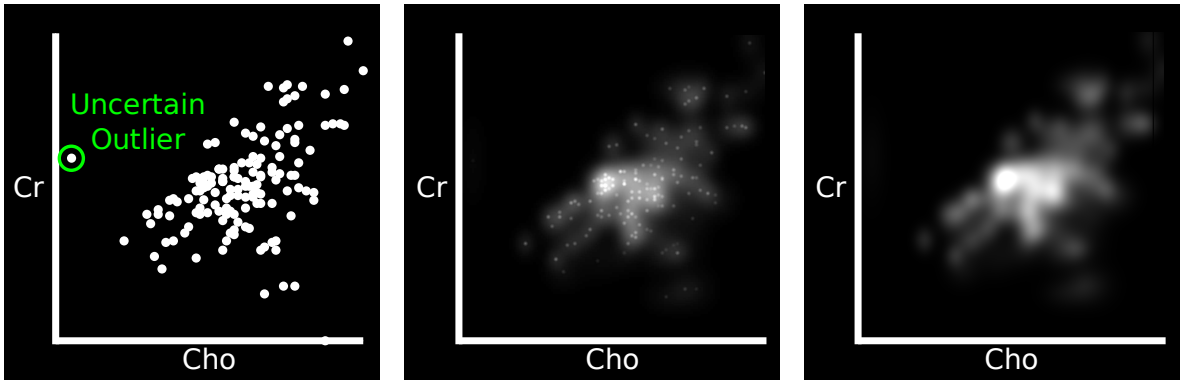


Figure 3.13: Choline-to-creating scatter plot of two MR spectroscopy metabolites. From left to right, means are emphasized by varying amounts. Left: a standard scatter plot of MR spectroscopy data, with choline concentration on the x-axis and creatine concentration on the y-axis. Middle: the PDF of the data with emphasized means, computed using KDE over normal distributions assuming $\rho = 0$. Right: direct rendering of the PDF.

3.3.1 From Scatter Plots to Density Plots

The standard scatter plot overlays a set of glyphs at Cartesian coordinates corresponding to all the (x, y) value pairs in the data set. When the data set is sufficiently large, regions with high densities become over-plotted. The simplest solution to this problem is to make the glyphs partially transparent: as glyphs accumulate, bright regions indicate higher glyph density. Such a plot is similar to a form of KDE with a glyph-shaped kernel. In statistics, the most commonly used kernel is a normal distribution with standard deviation used to control the feature size of the resulting PDF.

For statistically uncertain data, the appropriate kernel is the statistical distribution of each individual sample. In this way, large overlapping distributions become difficult to distinguish and small, high-valued distributions are easy to locate. For the sake of example, the following demonstration uses KDE to compute the PDF with normal distributions, which is one of the most common statistical distributions:

$$PDF(x, y) = \frac{1}{N} \sum_{i=0}^{N} D_i(x, y) \tag{3.1}$$

$$D_i(x, y) = \frac{1}{2\pi\sigma_{xi}\sigma_{yi}} \exp\left[\frac{-(x-\mu_{xi})^2}{2\sigma_{xi}^2} + \frac{(y-\mu_{yi})^2}{2\sigma_{yi}^2}\right] \tag{3.2}$$

where PDF is the probability density function, which is the average of the N individual distributions. D_i is an arbitrary distribution, here demonstrated as a normal distribution where σ_{xi} and σ_{yi} are the standard deviations of the μ_{xi} and μ_{yi} means for the i^{th} sample. This form of the normal distribution assumes that x and y are uncorrelated; a more complex form of D_i includes the correlation coefficient ρ. The data sets that drove the development of these techniques assume uncorrelated data, so Equation 3.2 applies. However, D_i can easily be replaced with any distribution in Cartesian space. The distributions as described must be discretized for display. The simplest way to do so is to subdivide the domain into pixel-sized bins.

Using the PDF as a basis for uncertainty visualization highlights regions of high point probability density, which is useful for discovering clustered values and trends. It also emphasizes points with high certainty (small, bright spots) while de-emphasizing points with low certainty (large, dim spots) by leveraging the human visual system's ability to preattentively separate high and low contrast image features.

As described in Section 3.2.6, differences in blur of image features are perceived preattentively. As distributions with high variance tend to overlap, they become harder to distinguish from each other and easier to distinguish from small, high density values. The result makes intuitive sense from a statistical perspective as well: the viewer should not be able to distinguish two distributions that overlap significantly.

Direct visualization of the PDF scales with data set size more effectively than standard opaque glyphs. Whereas traditional opaque glyph scatter plots become easily over-plotted, a one-time cost of sampling the PDF yields a density image that can be displayed at both high and low resolutions. Figure 3.13 demonstrates the transition from a normal scatter plot to direct PDF visualization.

Encoding uncertainty magnitude with glyph hue may seem like a reasonable alternative to blur. Color is also perceived preattentively, so the viewer will be able to quickly distinguish between certain and uncertain values. However, the uncertainty color scale must be chosen arbitrarily. The viewer will have to refer to the legend to interpret the hue differences they perceive. Worse, discrete representations of uncertain points can mislead the viewer into

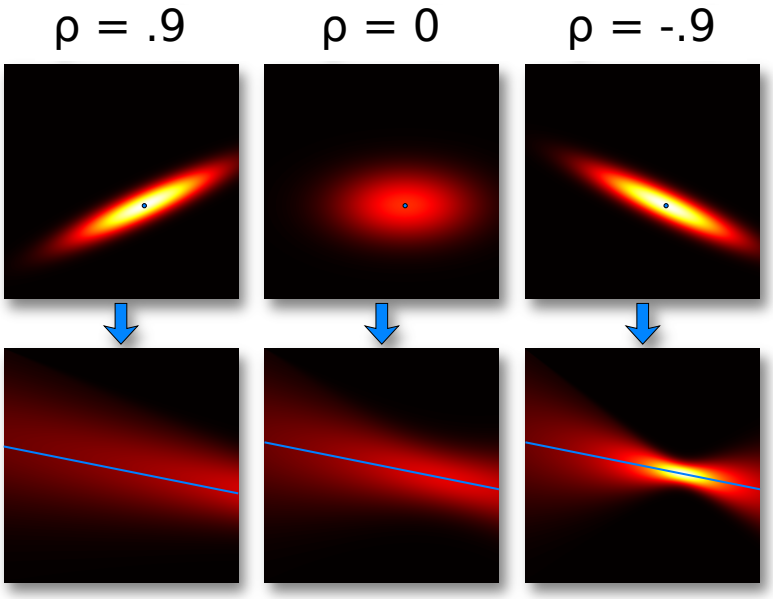


Figure 3.14: Three normal distributions with different values of ρ, the correlation coefficient, mapped from Cartesian space into parallel coordinates space. Notice how the $\rho = .9$ and $\rho = 0$ look different in the Cartesian plot, but they look similar in parallel coordinates.

identifying unreliable clusters or patterns (false positive), which they must then attentively disregard after consulting the legend. Density plots represent uncertainty magnitude directly in data space, leaving less room for confusion. Finally, using color for uncertainty also uses up a visual channel that is often reserved for distinguishing among different selections of data values.

3.3.2 Mean Emphasis

For data sets that are small enough to avoid over-plotting, emphasizing distribution means helps viewers see the locations of distributions contributing the density plot. The degree of emphasis should scale with the certainty of the point. For normal distributions, the maximum value is at the mean, and smaller values of σ increase the data value of the mean. Therefore, scaling the mean of the distribution by a constant factor (e.g. doubling its value) emphasizes the mean more for confident values than it does for uncertain values. The center of Figure 3.13 depicts a density-based scatter plot with mean emphasis.

Mean emphasis is similar to overlaying points on the PDF with transparency scaled by the

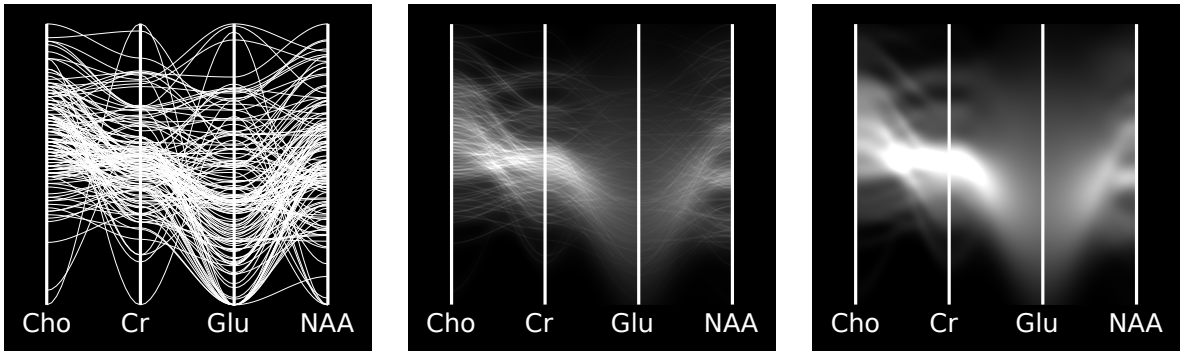


Figure 3.15: Parallel coordinates plots of four MR spectroscopy metabolites. From left to right, means are decreasingly emphasized. Left: a sigmoidal parallel coordinates plot of the same data shown in Figure 3.13 with two additional variables (Glutamine and n-Acetylaspartate). Center: the estimated PDF mapped into parallel coordinates space, with means emphasized according to their uncertainty. Right: direct visualization of the PDF.

height of the distribution mean. When combined with the density plot, as shown in the center of Figure 3.15, such points enable viewers to see directly both the scale of the distributions and the location of their means. If the data set is large enough that many of the means overlap, the viewer can look directly at the density plot for a summary of the data set.

3.3.3 PDFs in Parallel Coordinates

This section demonstrates how to compute the PDF in parallel coordinates space as a technique for visualizing uncertain multivariate data. Heinrich and Weiskopf describe how to transform an arbitrary distribution into parallel coordinates space (Heinrich and Weiskopf, 2009), and Miller and Wegman provide analytical solutions for bivariate normal distributions and uniform distributions (Miller and Wegman, 1991). Conceptually, the idea is to use the well-known point-line duality between scatter plots and parallel coordinates plots to transform samples from one space to the other while maintaining the formal properties of distributions (e.g., a unit integral). For the bivariate normal distribution discussed so far, the analytical form of the distribution is as follows:

$$\mu_a = (1-a)\mu_1 + a\mu_2 \tag{3.3}$$

$$\sigma_a^2 = (1-a)^2\sigma_1^2 + a^2\sigma_2^2 \tag{3.4}$$

$$PC(a,b) = \frac{1}{2\pi\sigma_a^2}\exp\left[\frac{-(b-\mu_a)^2}{2\sigma_a^2}\right] \tag{3.5}$$

$$a \in [0,1]$$

where a and b are the horizontal and vertical axes of the parallel coordinates plot: b is in the space of the data values and a is in a normalized space where the left axis is at $a = 0$ and the right axis is at $a = 1$. μ_a and σ_a are the mean and variance of the interpolated distribution. This definition of σ_a assumes that the two variables are uncorrelated. A more complex definition includes the correlation coefficient ρ.

Figure 3.14 demonstrates the transition of individual distributions into parallel coordinates plots, in this case differing by the value of their correlation coefficient ρ. The more linear the distribution, the stronger the positive or negative correlation. Notice how significant differences in the scatter plot PDF do not necessarily produce equally noticeable differences when transformed into parallel coordinates space.

Many applications use curved parallel coordinates plots, for example cardinal splines (Graham and Kennedy, 2003; Johansson and Johansson, 2009) or sigmoid curves (Moustafa and Wegman, 2006). The KDE representation of the parallel coordinates plot can accommodate both representations, but for brevity this section only demonstrates only the latter case. The premise is to warp the sample grid of the KDE to match the shape of the curve. Sigmoid curves can be represented by a number of functions, including the logistic function, sinusoid, and cubic polynomial. The following warps a parallel coordinates plot with a cubic polynomial with a change of variables:

$$a' \rightarrow -2a^3 + 3a^2 \tag{3.6}$$

The constants above produce a cubic polynomial that passes approximately through $(0,0)$

and $(1, 1)$. It should be noted that such a nonlinear change in variables results in a function which is no longer formally a PDF. This introduces a trade-off between the perceptual benefits of sigmoid curves and mathematical rigor.

For plots with a manageable number of lines, the density representation may remove a seemingly useful feature of the parallel coordinates plot: the viewer's ability to follow individual lines through the plot across multiple axes. However, this is in fact a benefit. The density plot discourages the user from following lines that may lead them to incorrect conclusions. When there are a large number of lines, density plots are a reasonable solution to over-plotting.

This work demonstrates how to compute density plots for statistically uncertain data. Figure 3.15 demonstrates the transition from a traditional line-based parallel coordinates plot to a density-based parallel coordinates plot. The center of the figure demonstrates a parallel coordinates density plot with mean emphasis, computed in a similar fashion to the mean-emphasized density plot described in Section 3.3.2.

3.3.4 Focus and Context

One difficulty with PDF-based representations is that they emphasize the most likely values while hiding potentially interesting outliers. In the PDF, outliers will manifest as a small range of values with high probability density. A clustered set of distributions with large variances will not be outliers, as their distributions cover a large range of values. The left frame of Figure 3.15 contains an apparent outlier, but its variances are so large that it cannot confidently be considered interesting. Large data sets exacerbate this problem: as the data set grows and values cluster in the most likely regions, small sets of outliers are de-emphasized even more.

Novotny and Hauser suggest several methods for identifying outliers in multivariate data sets by analyzing pairwise 2D histograms of variables. Their algorithm analyzes the connectivity of the histogram and attempts to discover small islands of histogram bins that contain a small number of values. Because histograms are closely related to PDFs, their techniques can be applied without modification to the PDFs computed above for statistically uncertain

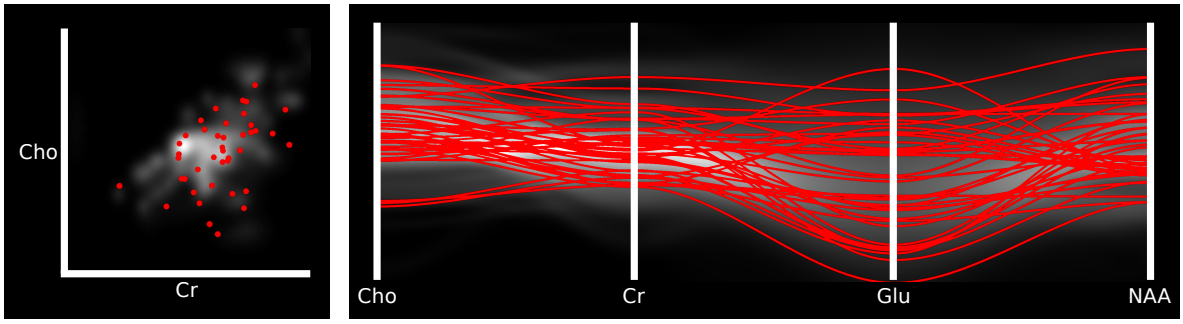


Figure 3.16: Probabilistic plots. Left: a demonstration of probabilistic scatter plots (here for Choline and Creatine). The gray-scale image in the background is the PDF of the two variables. Red dots are positions randomly sampled from the PDF. Right: the same random samples from the scatter plot extended to multivariate lines in a parallel coordinates plot. The gray-scale image in the background is the PDF in parallel coordinates space.

values. All distributions that predominantly fit within small, isolated regions can be labeled as outliers and drawn separately. This naturally handles uncertainty: a small isolated cluster of values that would otherwise be considered outliers will be ignored if their distributions are too large.

Integrating outliers with density plots combines both focus (the outliers) and context (the PDF). The outliers should be drawn differently from the rest of the plot. Representing outliers as discrete glyphs does this naturally for direct PDF visualization.

3.3.5 Scalability and Accelerated Rendering

It is important to consider the scalability of PDF computation, as it is the basis for the techniques described above. Because computational cost grows with grid resolution and data size, PDF computation may be expensive for extremely large data sets. However, the computation is also embarrassingly parallel: if the data is distributed across many nodes, each node can compute a local PDF, which can then all be averaged on a single node. In a multi-core shared-memory environment, the PDF bins/pixels also can be computed in parallel in separate threads.

For parallel coordinates plots, the PDF computation time has an important effect on plot interactivity. Incurring the precomputation cost of PDF computation may be acceptable, however if the user wishes to interactively rearrange the axes in the plot, it is no longer a

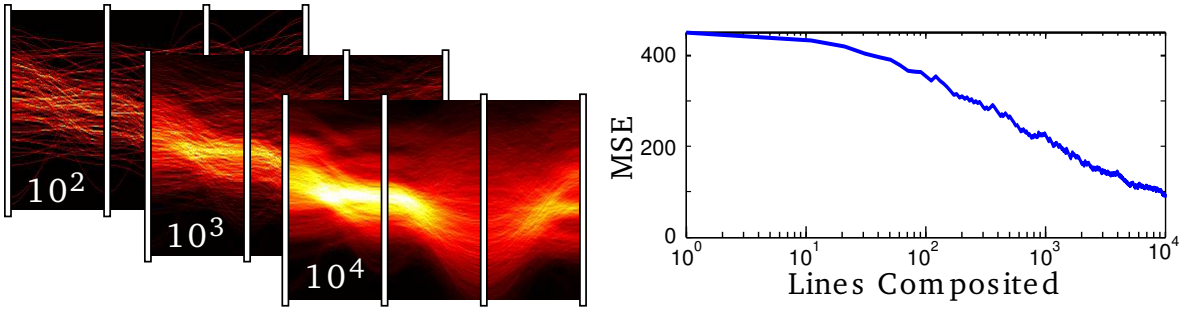


Figure 3.17: Monte-Carlo integration in parallel coordinates. Left: randomly sampled lines accumulating into a histogram PDF approximation, labeled with the number of contributing lines. Right: a log plot of mean square error as compared to the correct solution as computed in section 3.3.3.

one-time cost. In this case, spawning a background process to compute pair-wise PDFs for potential plot arrangements will be useful.

For data sets that have analytical solutions to PDF computation and are sufficiently small (like the MR spectroscopy data set), massively parallel graphics processing units (GPUs) are extremely efficient and accurate. A scatter-based algorithm simply renders each distribution into a texture with additive blending and then renders the texture to the screen. The fragment shader quickly and accurately samples the distributions on a per-pixel basis.

3.3.6 Probabilistic Plots

For extremely large data sets that require significant PDF computation time, the probabilistic plot is one way to quickly summarize the data point distribution. The probabilistic plot is composed of a set of random samples that have the same distribution as the underlying PDF. For a simple PDF composed of a single normal distribution, the random samples will be clustered around the mean and have the same variance as the distribution. If multiple distributions contribute to the PDF, points will similarly cluster around the other distribution means as well. Figure 3.16 illustrates a discrete scatter plot and parallel coordinates plot with random points drawn from an underlying PDF.

A single set of random samples will not accurately represent the total variability of the PDF and may contain false patterns. Therefore, the plot is animated continually replacing old samples with new samples from the distribution. There are two benefits to this type of

animated display. First, regions of high density will remain fairly stable over time whereas unlikely values will only appear briefly. Second, any outliers will flicker in and out of existence in regions with local density spikes. Intermittent flickering signals the viewer to look at the outlier region. This animated plot naturally provides a summary of the overall data while also highlighting potentially interesting outliers.

The probabilistic plot depends on the ability to acquire random samples with the same distribution as the PDF quickly without explicitly computing it. Note that in this case PDF refers to the N-dimensional (ND) PDF, where N is the number of variables, rather than the simple bivariate PDFs computed in the previous plots. There are several ways to sample a distribution, including inverse transform sampling, rejection sampling, and others (Robert and Casella, 2005). Of course, all of these techniques require that the ND PDF be known *a priori*, and storing discretized ND PDFs is intractable for even modestly sized data sets.

Fortunately, properties of the MR spectroscopy data set that drove the design of this technique make random sampling of the ND PDF extremely efficient. First, the individual data points have equal weight (1/N) in the kernel density estimate of the PDF. Second, the density at each point is composed of N independent normal distributions. Therefore, the PDF can be randomly sampled in two steps:

1. Pick a random data point p
2. Randomly sample all N of p's independent normal distributions.

Step 1 is trivial, requiring only a single uniform random integer. Step 2 uses the well-known Box-Muller transform to quickly generate N normally distributed values with with $\mu = 0$ and $\sigma = 1$ that can be converted to an arbitrary μ and σ with a simple shift and scale (Box and Muller, 1958). To summarize the entire data set, the process can simply be repeated M times to acquire M random data points. Over time, new sets of samples can be acquired and replace the old ones. The result is an animated plot in which regions of high probability density are stable and outliers intermittently flicker in and out of existence.

As new random samples are acquired over time, they can also be accumulated to approximate the underlying distribution, performing Monte Carlo integration of the PDF. As each

new set of samples gets added into a floating point buffer, the more likely points will overlap to produce brighter intensities. The result is a line density histogram. Each pixel is a bin, and the sampled lines vote in those bins. The convergence rate of the computation depends on the number of data set samples and the overall magnitude of uncertainty. When the area of the histogram is normalized to one, the histogram closely approximates the PDF. Figure 3.17 demonstrates the expected $\sim \sqrt{N}$ convergence of a parallel coordinates plot with $\sim$250 distributions.

A probabilistic parallel coordinates plot has the benefit that it is composed of discrete lines. Not only are the lines easy to draw, but the viewer can follow cords of lines passing through stable regions of high density. In regions of low density, lines only appear briefly and sporadically, making them difficult to follow. The viewer's ability to follow lines is directly proportional to probability density.

Finally, in order to link interactions in the probabilistic plot to other plots, as described in Chapter 4, it is important to record the row of the data table that produced each randomly sampled point or line. Selections can then be passed between data views as a set of table row indices.

3.4 Discussion

Incorporating KDEs into plots of statistically uncertain multivariate data using the visualization and interaction techniques described in this work can lead viewers to identify useful and interesting variable relationships. Just as important, they inhibit preattentive perception for uncertain values and therefore prevent viewers from forming false hypotheses. Clusters of very uncertain values with wide distributions combine into a large, low contrast shape, which correctly discourages viewers from distinguishing between the means. Likewise, uncertain values are spread across a wide area and do not distract the viewer's attention from certain values.

Normally distributed uncertainty enables simple, parallelizable PDF computation in both scatter plots and parallel coordinates plots. For extremely large data sets, probabilistic plots

are a way of summarizing the data using a fast random sampling technique. Animated over time, these plots help draw viewer attention to outliers while approximating the underlying PDF. Once the PDF has been reasonably approximated, more rigorous density-based multivariate analysis tools for outlier and pattern identification integrate naturally. The probabilistic plot has many potential future areas of exploration. Given that the plot changes over time, a time-aware interface for navigating the space of random plots would be interesting. The interaction techniques described in this work apply to probabilistic plots, however interaction techniques with dynamic plots deserve further consideration.

While the availability of per-sample statistical distributions makes KDE accurate and simple, it is not a prerequisite to using the visualization techniques described here. Traditional Parzen windowing lets the user manually control the shape of the statistical distribution used for each data point. The selected shape controls the emphasized feature size in the resulting PDF. Ensemble data sets can be summarized with the mean and variance of each sample. Clusters in large multivariate visualizations can be visualized similarly. Any PDF can be used as input for these techniques, regardless of the how it is estimated.

The techniques presented in this chapter were designed to explore MRS data, which will be described in more detail in Chapter 5. MRS is also used for a wide array of other disease processes, including multiple sclerosis and many cancers throughout the body. Outside of MR, another source of uncertain spectroscopy data comes from an optical inspection technique called Matrix-Assisted Laser Desorption/Ionization that decomposes spatially localized tissue samples into their protein spectra. Each point sample is often the average of many local spectral signals, which leads to normally distributed spectral peaks. Continued interdisciplinary collaboration between visualization researchers and domain scientists will hopefully help improve multivariate uncertainty visualization to address complex problems such as these.

Density-based plots apply directly to any multivariate data set that includes statistically quantified uncertainty. They have been demonstrated in both scatter plot and parallel coordinates displays of such data in an interactive, linked-views display. Chapter 5 shows how these interactive explorations have been used in MRS data. Abstract density plots are a useful

technique for exploring this and other uncertain data sets while preventing false conclusions based on untrustworthy values.

CHAPTER 4

User Interaction with Uncertain Plots

(The contents of this chapter were presented in the proceedings at the Conference on Information Visualization 2010 and the Conference on Visualization and Data Analysis 2010 (Feng et al., 2010a; Feng et al., 2010b))

Just as a visualization should draw the viewer's eye to features of interest in a data set, the system of interactions that a visualization tool provides should help the viewer to explore those features and answer their questions. For example, consider the design goals of radiologists studying magnetic resonance spectroscopy (MRS). They wish to understand the relationships among metabolites and anatomical features, so the interaction techniques provided by the visualization tool should satisfy those goals.

This chapter describes a set of techniques for interacting with the spatial and abstract visualizations described in previous chapters. These interaction tools were designed specifically to identify potentially meaningful relationships among metabolites. The goal is to enable the user to select a population of interesting data points and compare it to other populations (i.e. tumor versus non-tumor). The metaphor often associated with such techniques is a paint brush: the viewer selects a color and paints the data points they wish to select.

Brushing is a particularly powerful tool when the viewer is presented with multiple visualizations of their data that highlight different kinds of features. For example, consider the scatter plot and SDDS. SDDS is a spatial visualization of the data set, but problems of occlusion and clutter in 3D visualizations prevent uncertainty from being represented usefully.

The scatter plot removes spatial position information but adds uncertainty information, and is particularly useful for identifying clusters of values and linear trends. If the viewer sees two clusters of data points in the scatter plot, they can brush the scatter plot in two different colors and then look to see how those selections are represented in the spatial view. One way to do so is simply to draw the corresponding voxels and see how their spatial positions differ from each other and from features in the reference image slice included in the 3D SDDS view described in Chapter 2. This chapter will describe brushing techniques that link together the visualizations described in previous chapters and help viewers identify useful features in multivariate data.

This chapter describes four novel interaction techniques:

1. **Linear Function Brushing**: a technique for selecting linear patterns in parallel coordinates plots.

2. **Lasso Brushing**: a technique for selecting parallel coordinates lines that pass through a user-drawn curve.

3. **New Axis Construction**: a technique for creating new parallel coordinates axes composed of multiple variables that is useful for discovering more complex relationships.

4. **Uncertain Brushing**: techniques for extending existing brushing techniques to account for statistical uncertainty in data values.

4.1 Background

4.1.1 Linked Visualization Systems

Many systems coordinate multiple techniques into a single visualization application. XmdvTool (Ward, 1994) and GGobi (Swayne et al., 2003) combine multiple information visualization views (parallel coordinates plots, scatter plots, hierarchical views, etc.) of multivariate data sets. Akiba and Ma combine parallel coordinates, high dimensional histograms, and DVR into a single interface (Akiba and Ma, 2007). Users can brush over interesting clusters of lines

in the parallel coordinates view or features in the histogram view and see the resulting DVR update interactively. Systems like SimVis (Doleisch et al., 2003) and WEAVE (Gresh et al., 2000) combine information visualization techniques (statistical representations, feature analysis, ND projections) with scientific visualizations (Blaas et al., 2007). This chapter extends such techniques by introducing novel linked interactions that take data value uncertainty into account.

4.1.2 User Interaction with Abstract Plots

Multivariate exploration of scatter plots has a long research history. PRIM-9, one of the early interactive multivariate exploration systems, introduced plot rotation and projection followed by data point masking/filtering (Tukey et al., 1988). Such masking is often done by mouse painting/lasso interaction styles. A user highlights a region either by outlining and points within the region are selected. Martin and Ward extend this by creating multiple brush groups created using logical operators (AND, OR, etc.) (Martin and Ward, 1995). Elmqvist et al. describe the process of query sculpting by which selections are iteratively refined by looking at the selection in scatter plots of other variables (Elmqvist et al., 2008). To aid the sculpting process, they provide an animated tour of scatter plots as a user moves from plot to plot.

Angular brushing is a parallel coordinates interaction technique that enables the user to select lines that all have similar slope between two parallel axes (Hauser et al., 2002). This technique will be discussed in more detail in more detail in Section 4.2.

4.1.3 Variable Ordering

Although viewers can follow parallel coordinates lines across a plot, some variable relationships only appear between adjacent pairs of variables. A great deal of work already describes potentially useful orderings and variable subsets. For example, orderings can prefer variables pairs that are similar to each other on several metrics (Ankerst et al., 1998), minimize visual clutter (Peng et al., 2004), or have high ranking based on user-selected properties (Seo and Shneiderman, 2004). These variable ordering techniques are useful for determining which

parallel coordinates plot or scatter plot matrix configuration will show the most interesting patterns. That work is complementary to the techniques described in this chapter.

4.1.4 The Point-Line Duality

Parallel coordinates plots and scatter plots are both representations of multivariate data, and as such there are many relationships between their data representations. For example, consider a point in a scatter plot. It represents two variable values by its position in Cartesian space. Those same two values are represented as a line in the parallel coordinates plot connecting two axes. Now consider a point in the parallel coordinates plot. There are an infinite number of data values whose lines intersect at that point position. When represented in a scatter plot, those data values all lie on a single line. This duality between points and lines is illustrated in Figure 4.1.

Because a scatter plot point maps to a line in the parallel coordinates plot and the parallel coordinates point *generates* values that lie on a line in the scatter plot, the scatter plot and parallel coordinates plot are often said to have a *point-line duality*. This relationships is useful from an interaction design perspective. When the viewer sees an interesting visual pattern in the parallel coordinates plot, the point-line duality can be used to mathematically describe that pattern. This chapter describes the mathematical foundations of both existing and novel user interaction techniques combined with the uncertain plots described in the previous chapter.

4.2 Linear Function Brushing

The fundamental idea of brushing is to enable the viewer to easily isolate a group of data points that have some interesting visual pattern in a visualization. Once that pattern has been detected, the mathematical and statistical properties of that pattern can be described to the user. Such properties may include the mean and variance of a cluster of values or the slope and correlation coefficient of a linear regression. The scatter plot is commonly used for portraying clusters of values and linear trends, as shown in Chapter 3. The relevant figures

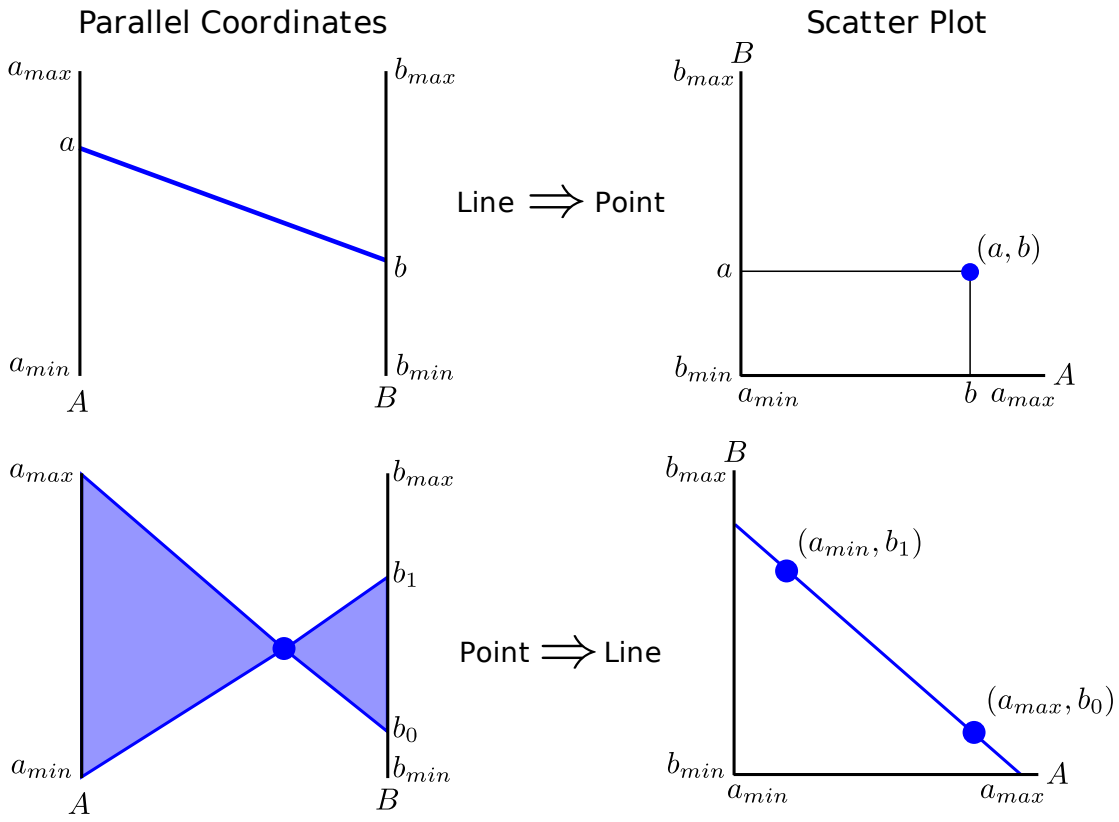
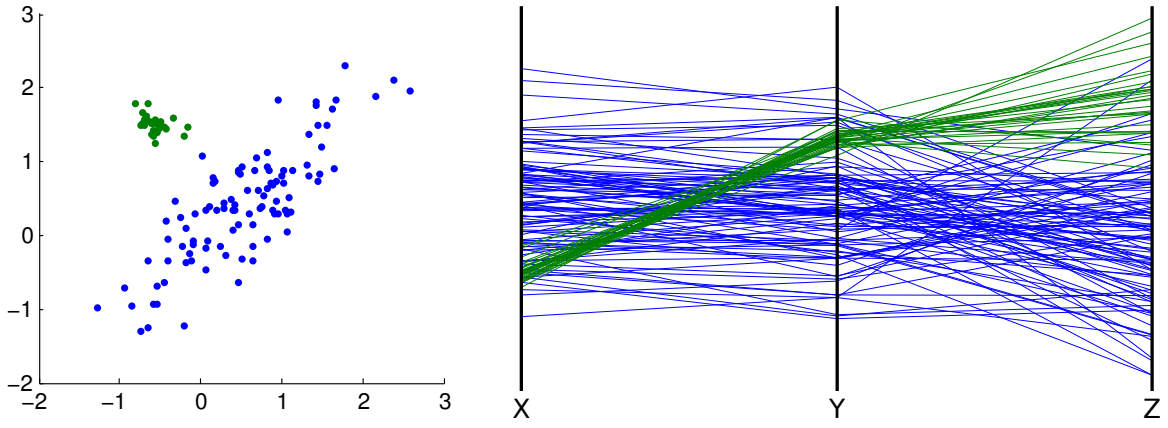


Figure 4.1: The point-line duality. Top: a segment between two axes in a parallel coordinates plot represents the same information as a point in a Cartesian scatter plot. Bottom: all line segments that intersect at a point between two parallel coordinates plot axes represent a line in a Cartesian scatter plot.



(a) scatter plot

(b) parallel coordinates plot

Figure 4.2: The same example scatter plot and parallel coordinates plot shown in chapter 3, shown here again for reference.

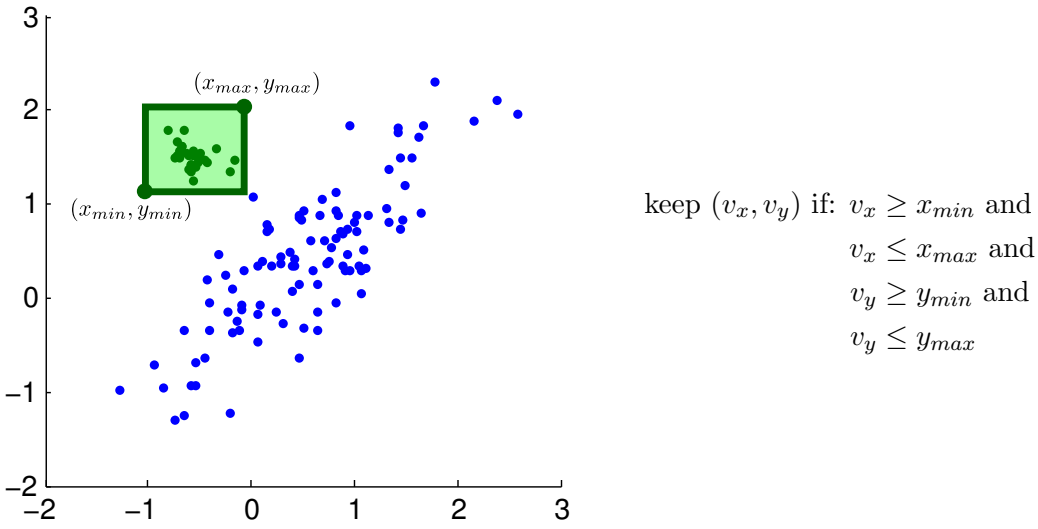


Figure 4.3: Selecting a cluster of values in a scatter plot. The user clicks and drags open a box defining a minimum and maximum range of x and y values.

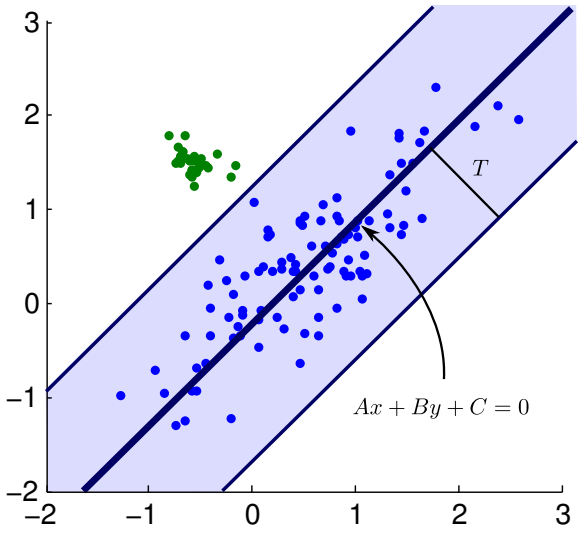
are reproduced here, in Figure 4.2.

When viewers see a tight cluster of values, such as the green dots in Figure 4.2a, they can select them simply by dragging open a bounding box around them and selecting all of the points that fall inside the box. This is nothing but a per-point bounding box evaluation, shown in Figure 4.3. The bounding box is defined by a minimum and maximum on the x and y axes.

When the viewer sees a strong linear trend, as is the case in the blue dots in Figure 4.2a, the standard brushing technique is for the viewer to draw a line and select all points that fall within a specified distance tolerance to that line. This requires the evaluation of each point's distance to the line, shown in Figure 4.4.

These scatter plot brushes both have equivalents in the parallel coordinates plot. Range selection for any variable is straightforward: the user draws a line or bar representing the selected range of values on any axes. Linear selection is only slightly more complex. When the user brushes a linear function on a scatter plot, they are specifying two endpoints of a line segment. The point-line duality shows that these two endpoints correspond to two lines in the parallel coordinates plot.

Before continuing, it is important to consider what linear relationships look like in the



keep (v_x, v_y) if: $\frac{|Av_x + Bv_y + C|}{\sqrt{A^2 + B^2}} < T$

Figure 4.4: Selecting a linear function in a scatter plot. The user clicks and drags a line, and all data points within a user-specified distance threshold T to that line are selected.

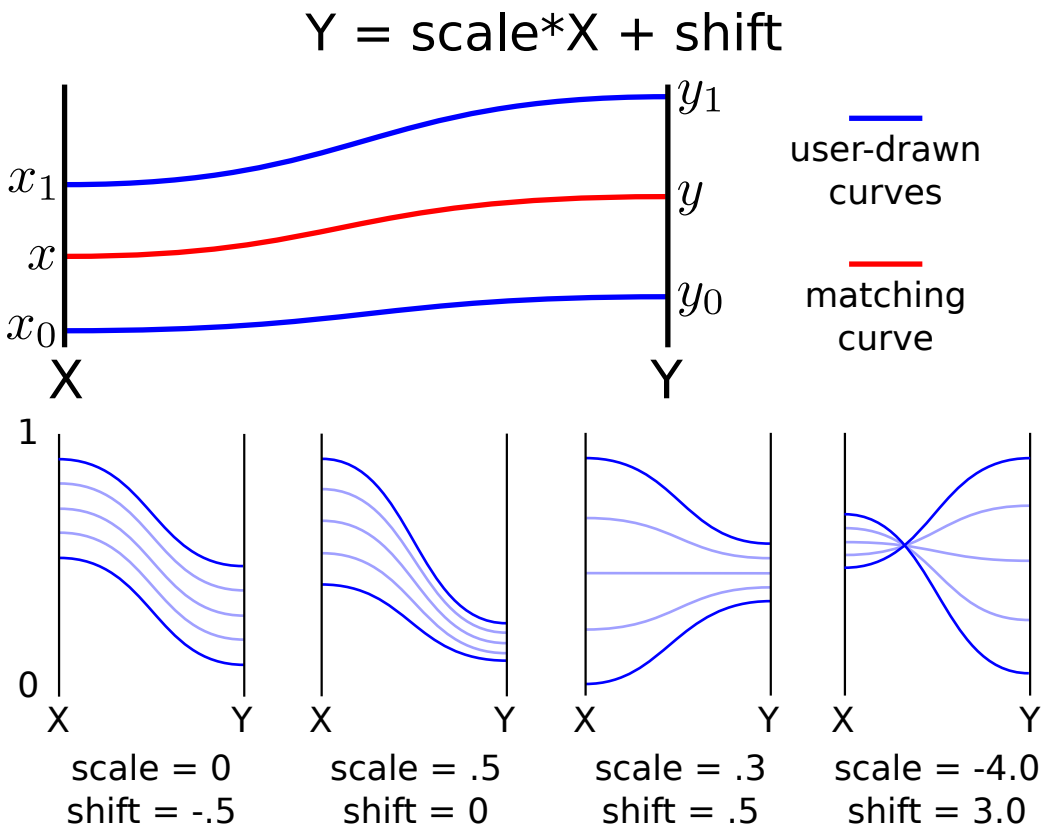


Figure 4.5: Linear patterns seen in parallel coordinates plots. Sets of lines that all intersect at the same point (between axes or outside of axes) indicate a linear relationship. Sets of lines that are parallel to each other indicate a constant relationship.

parallel coordinates plot. The bottom of Figure 4.5 depicts some of the patterns visible in parallel coordinates plots when the data values are linearly related. To select lines matching a linear pattern, the user needs only to draw two representative lines on the parallel coordinates plot. The two lines represent two values pairs (x_0, y_0) and (x_1, y_1), which can be trivially converted to an implicit line equation. With this equation, the mathematics of determining whether a data point is near enough to the line is identical to what was used for the scatter plot calculation. This equation can then be relayed back to the viewer. Note that if all of the values on one axis have the same value, these patterns will also appear. Consider an axis X for which all values equal 0. In this case, all other variables will reveal a pattern similar to the $\text{scale} = .5, \text{shift} = 0$ case in Figure 4.5, except that in this case $\text{scale} = 0$.

This ability of the brush tool to report the selected linear function is particularly important with the parallel coordinates plot. A complexity that can be initially confusing is whether the range of values represented on the different axes are all equal or are rescaled on an individual basis. The former is the case in examples shown in Figure 4.5, where all of the axes have a range of $[0, 1]$. When the axes have different ranges, the visual patterns will be shifted and scaled by a linear transformation. This means that the linear pattern on individually rescaled axes will still be a linear pattern, but it will not look the same. Mentally factoring this additional linear rescaling into an estimate of the correlation between two variables is challenging and unnecessary for the viewer when the brush itself can compute the mathematical basis of the visual pattern.

Angular brushing is a technique described by Hauser et al. for selecting parallel coordinates lines that all have a similar slope (Hauser et al., 2002). This is the left-most example in Figure 4.5, which means that angular brushing is a special case of the linear function brushing described in this chapter. The user need only brush a single line, because the second line they might otherwise draw would be superfluous.

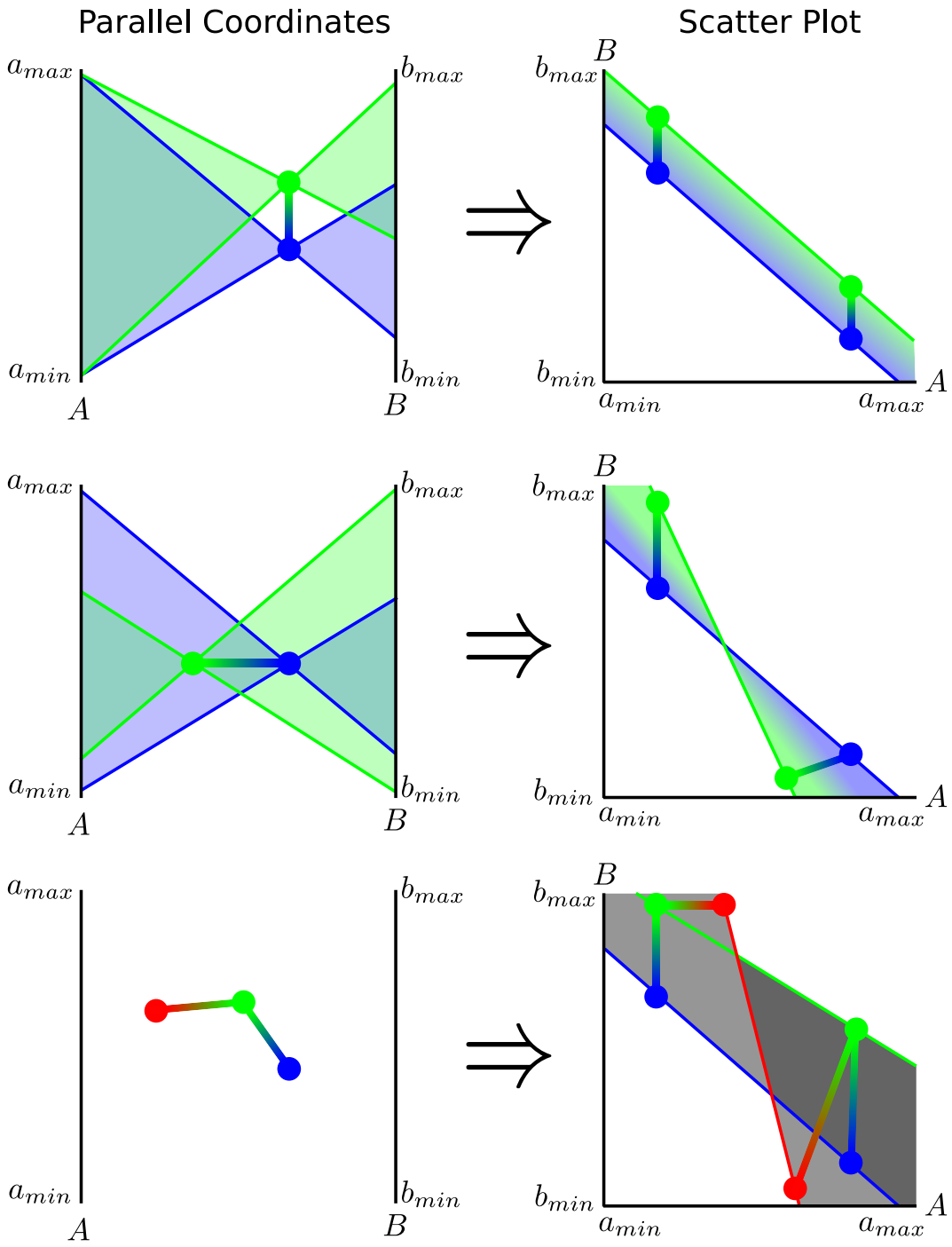


Figure 4.6: Top and middle: A line segment drawn in a parallel coordinates plot can be thought of as the region defined by the interpolation between the two Cartesian lines corresponding to the line segment's endpoints. Bottom: multiple line segments produce a more complex region. All points that fall within the union of these Cartesian regions are parallel coordinates lines that pass through the drawn line segments.

4.3 Lasso Brushing

The second novel parallel coordinates interaction technique developed in this work is a lasso that enables the viewer to select any line at at any location on the plot. This brush is a curve that the user draws anywhere to select all lines passing through it. This method contrasts with standard range-based selection, in which the user specifies a range of values on the axes themselves (Akiba and Ma, 2007), as it enables the user to select lines anywhere between axes as well as on the axes. Deciding which lines pass through a curve in parallel coordinates space is done using the point-line duality. First consider the simpler case of a single line segment. As shown in the top of Figure 4.6, the endpoints of that line segment correspond to two different lines in the scatter plot. Interpolating between those two lines defines a region in the plot; all scatter plot points within that region correspond to parallel coordinates lines that intersect with the user-drawn segment. To decide if a point (v_x, v_y) falls within that Cartesian region can be done by determining if the point is between the two lines $L_1(A_1, B_1, C_1)$ and $L_2(A_2, B_2, C_2)$:

$$\begin{aligned} s_1 =& A_1 v_x + B_1 v_y + C_1 \\ s_2 =& A_2 v_x + B_2 v_y + C_2 \\ & \text{keep } v \text{ if: } sign(s_1) \neq sign(s_2) \end{aligned}$$

A curve can be decomposed into a set of line segments. For a data value to be selected, its parallel coordinates line must pass through at least one of the curve's line segments, and thus it must be in at least one of the corresponding regions in Cartesian space. If one segment produces a single region between two lines in scatter plot space, multiple segments simply grow that region (see the bottom of Figure 4.6. Consider a coarse curve approximation consisting of two line segments. This curve has three points in it, which correspond to three different lines in scatter plot space ($L1$, $L2$, and $L3$). For a data value to pass this test, it must either fall within the region between $L1$ and $L2$ or the region between $L2$ and $L3$. For finer curve

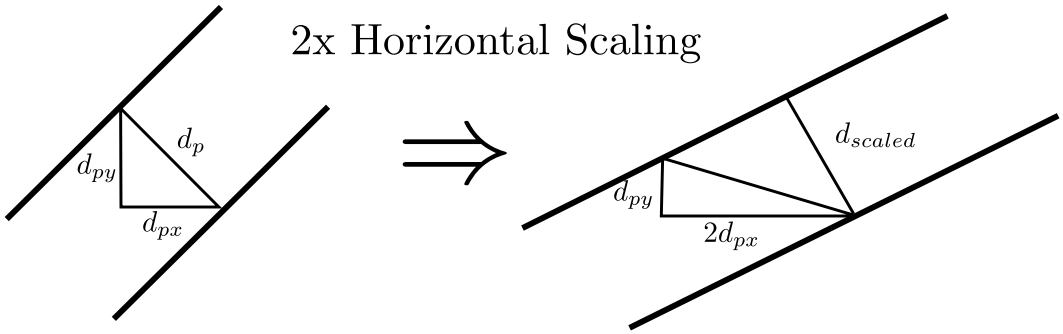


Figure 4.7: An example of how a perpendicular distance (d_p) is transformed into a separate perpendicular distance (d_{scaled}) after a transformation has been applied.

approximations with N points, the test is as follows:

$$
\begin{aligned}
s_1 =& A_1 v_x + B_1 v_y + C_1 \\
&\vdots \\
s_N =& A_N v_x + B_N v_y + C_N \\
&\text{keep } v \text{ if: } sign(s_1) \neq sign(s_2) \text{ or } \ldots \text{ or } sign(s_{N-1}) \neq sign(s_N)
\end{aligned}
$$

4.4 Implementing Brushing in Rescaled Screen Space

Up to now, the discussion of brushing techniques has focused exclusively on data-space operations. The range brush creates a bounding box defined by a lower-left (x_{min}, y_{min}) and upper-right (x_{max}, y_{max}) data point. The linear function brush selects points that fall within a distance threshold to a line ($Ax + By + C = 0$). However, there has been no discussion of how that data space gets mapped into an image, which is what the user actually interacts with. The scatter plots and parallel coordinates plots often have their axes rescaled to the range of contained data values. This means that two axes that have the same length on the screen (in pixels) can represent a dramatically different range of values in data space.

Consider a scatter with the lower-left point at pixel $(x_{min,p}, x_{max,p})$ corresponding to (x_{min}, y_{min}) in data space and the upper-right point at pixel $(x_{max,p}, y_{max,p})$ corresponding to (x_{max}, y_{max}) in data space. If the user attempts to select all points on a scatter plot within a certain number of pixels (d_p) to a line ($Ax+By+C = 0$), d_p can be decomposed into separate

x and y components that can represent drastically different distances once transformed in to data space. Figure 4.7 illustrates this with an example. If the transformation from pixel coordinates to data-space coordinates is a horizontal stretch, the perpendicular distance d_p gets transformed into a new distance d_{scaled}, which in this case is somewhat smaller than the original distance.

If the user draws a line in pixel coordinates with the equation $Ax + By + C = 0$ and wants to select all points within a distance d_p to that line, the user essentially is defining two bounding lines around the original equation:

$$Ax + By + C + d\sqrt{A^2 + B^2} = 0$$
$$Ax + By + C - d\sqrt{A^2 + B^2} = 0$$

The transformation from pixel coordinates to data-space coordinates is a transformation consisting of a shift and a scale. The scale factors are as follows:

$$s_x = \frac{x_{max} - x_{min}}{x_{max,p} - x_{min,p}}$$
$$s_y = \frac{y_{max} - y_{min}}{y_{max,p} - y_{min,p}}$$

The original linear equation and its bounding lines can be scaled as follows:

$$A\frac{x}{s_x} + B\frac{y}{s_y} + C = 0 \qquad \leftarrow \text{drawn line}$$
$$A\frac{x}{s_x} + B\frac{y}{s_y} + C + d_p\sqrt{A^2 + B^2} = 0 \qquad \leftarrow \text{bounding line above}$$
$$A\frac{x}{s_x} + B\frac{y}{s_y} + C - d_p\sqrt{A^2 + B^2} = 0 \qquad \leftarrow \text{bounding line below}$$

To compute the distance between the drawn line and a bounding line in transformed coordinates, it is only necessary to compute the distance between a point on the drawn line (for

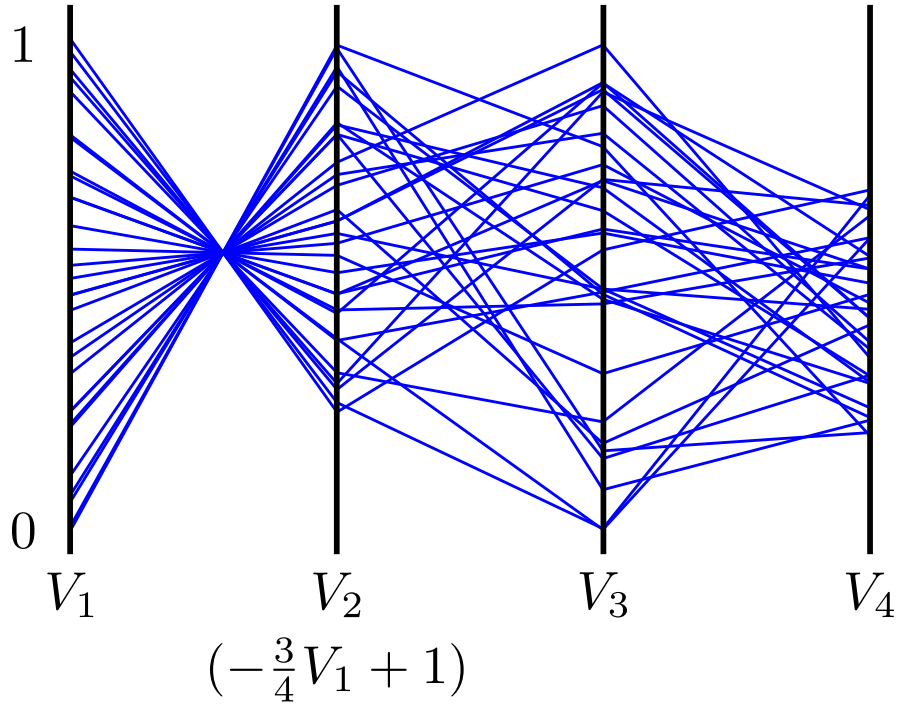


Figure 4.8: A parallel coordinates example with four variables (V_{1-4}). V_2 is linearly related to V_1.

example, the x-intercept) and a bounding line:

$$p = (0, \frac{-Cs_y}{B})$$

$$d_{scaled} = \frac{\frac{A}{s_x}(0) + \frac{B}{s_y}(\frac{-Cs_y}{B}) + C + d_p\sqrt{A^2 + B^2}}{\sqrt{\frac{A}{s_x}^2 + \frac{B}{s_y}^2}}$$

$$d_{scaled} = d_p\sqrt{\frac{A^2 + B^2}{\frac{A}{s_x}^2 + \frac{B}{s_y}^2}}$$

Using the computed value of d_{scaled}, it is possible to decide whether a point is within the distance d_{scaled} to the line.

4.5 New Axis Construction

Techniques like linear function brushing are useful for helping users to identify linear trends in data sets based on visual patterns; however, those patterns have so far been limited to two

variables. This is primarily because scatter plots by construction only display two variables and linear trends only manifest in adjacent parallel axes.

I extend the parallel coordinates plot to enable multivariate comparisons by adding new axes that combine one or more variables. The viewer can then compare the combined axis to other variables to build up more complex relationships. For example, consider the four parallel coordinates axes (V_{1-4}) displayed in Figure 4.8. Several relationships are visible in this plot:

1. V_1 is negatively correlated with V_2.
2. V_2 is uncorrelated with V_3.
3. V_3 is uncorrelated with V_4.

Based on 1 and 2, transitivity implies that V_1 is uncorrelated with V_3. Otherwise, the following would be true:

$$V_1 = AV_2 + B$$
$$V_1 = CV_3 + D$$
$$V_2 = \frac{C}{A}V_3 + \frac{B-D}{A} \qquad (!)$$

This would mean that V_2 is correlated with V_3, which is already shown to be false. Hence, V_1 cannot be correlated with V_3 via proof by contradiction. However, nothing can be said about how V_1 and V_2 are related to V_4. There are two options: either V_1 and V_2 are uncorrelated with V_4, or they are both correlated with V_4. If V_2 is correlated with V_4, then V_1 must be as well (and vice versa) because:

$$V_1 = AV_2 + B$$
$$V_2 = CV_4 + D$$
$$V_1 = ACV_4 + B + D$$

So to answer the question of how V_1 and V_2 are related to V_4, V_1 and V_2 can be mathematically

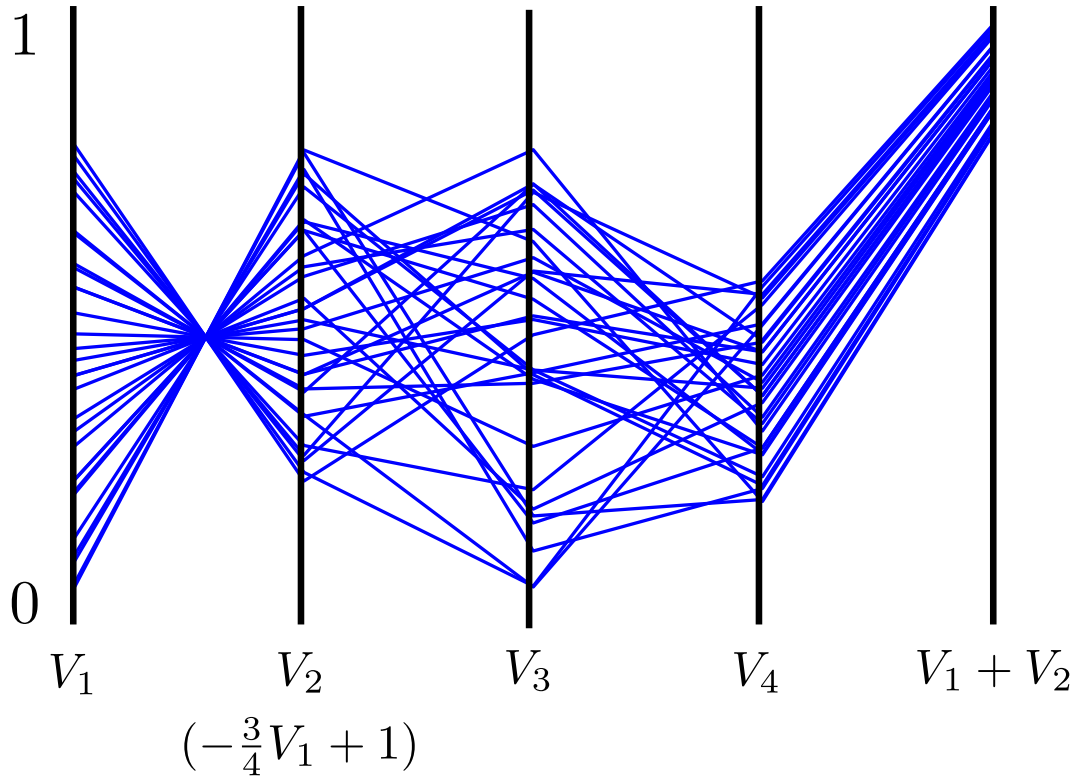
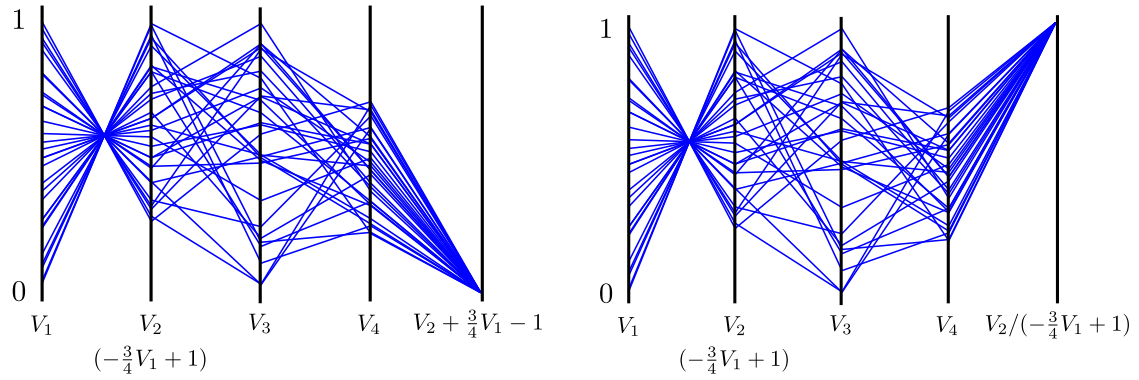


Figure 4.9: V_1 and V_2 can both be compared to V_4 simultaneously by adding the two together into a single variable $V_1 + V_2$. The new variable is positively correlated to V_4, which indicates that both V_1 and V_2 are correlated with V_4 individually.



(a) New Difference Column (b) New Ratio Column

Figure 4.10: A new column can be constructed using specific knowledge of a linear relationship between two axes. The new axis can represent this relationship as a difference (a) or a ratio (b), for example.

combined into a single, new column and place the new column next to V_4. One combination would be $V_1 + V_2$, as shown in Figure 4.9. If a linear pattern appears between the new column and V_4, a linear function brush can extract the linear relationship, which contains three variables, and report it to the user.

When the data contains more than one population of data points and only some match a visible pattern, a potentially useful column is one that evaluates how well a data point matches the linear function. There are two ways of doing so for a linear function defined using the standard line equation ($y = mx + b$):

1. **Difference**: for each data point, evaluate $y - mx - b$.

2. **Ratio**: for each data point, evaluate $\frac{y}{mx+b}$.

For method 1, data points that evaluate to (or near) zero match the linear relationship well. For method 2, data points that evaluate to (or near) 1 match the linear relationship well. The second form contains a division by a data value, which could potentially result in undefined result for values near zero. The difference between the two forms is shown in Figures 4.10a and 4.10b.

This methodology of combining multiple columns in parallel coordinates plots has been used by radiologists studying MRS to build up multivariate relationships among metabolites

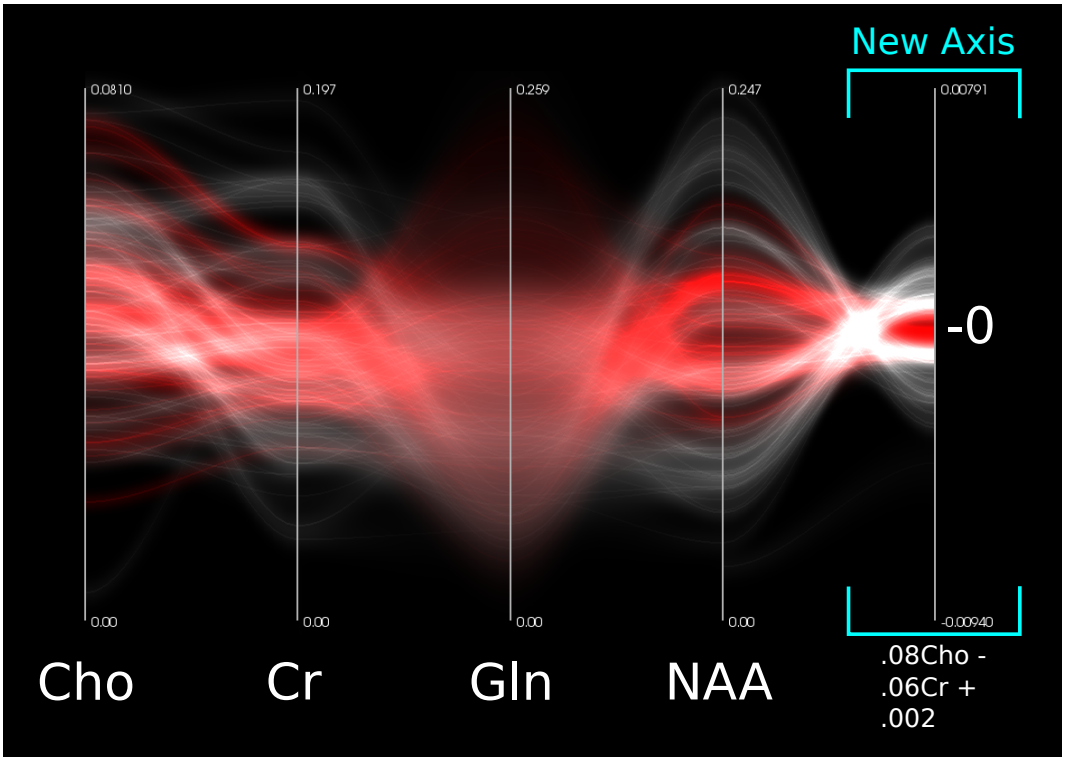


Figure 4.11: New axis construction applied to MRS data. The choline and creatine axes are combined into a new axis using a function identified with a linear function brush and then compared simultaneously to the NAA axis.

that describe the composition of tumors. An example of this is shown in Figure 4.11, in which the choline and creatine axes are combined and compared to NAA to produce a function that isolates tumor voxels. In general, this method of combining axes can be used to develop complex relationships that define segmentation classifiers.

The data used in MRS is statistically uncertain, meaning that each metabolite concentration measure is normally distributed. Combining two statistically uncertain variables requires careful consideration of the variance of those samples. The variance of a weighted combination of two uncorrelated variables is as follows:

$$\mathrm{Var}(aX + bY + c) = a^2\mathrm{Var}(X) + b^2\mathrm{Var}(Y) \tag{4.1}$$

$$\mathrm{Var}\left(\frac{aX}{bY}\right) = \frac{a^2}{b^2}\mathrm{Var}\left(\frac{X}{Y}\right) = \frac{a^2}{b^2}\left(\frac{\mu_1}{\mu_2}\right)^2\left(\frac{\sigma_1^2}{\mu_1^2} + \frac{\sigma_2^2}{\mu_2^2}\right) \tag{4.2}$$

Note that Equation 4.2 is the Pearsonian approximation to a more complex expansion, which Hinkley describes in more detail (Hinkley, 1969).

4.6 Brushing in Uncertain Plots

Just as density plots and probabilistic plots emphasize certain values and draw attention away from uncertain values, interaction with the plot should likewise favor certain points over uncertain ones. This section describes how to incorporate knowledge of uncertainty into three plot interaction techniques: interval queries, angular brushing, and linear function brushing. While the interaction primitives (lines, boxes, etc.) differ between scatter plots and parallel coordinates plots, the mathematics of selection is the same. The following discussion assumes a Cartesian space.

Interval queries select all values that fall within a prescribed range of values for a set of variables. In Cartesian space, the user will either draw a box to define that range or click a point to select values within a distance threshold. The analog in a parallel coordinates plot is to either manually specify a range on one or more axes or to draw a representative line segment. When the data values are defined as statistical distributions, the decision of

whether or not a point falls within the range of values is no longer an inside/outside test. In this case, the likelihood that a point drawn from that distribution will fall within the interval must be estimated.

The problem now becomes estimating a definite integral of a statistical distribution. For a general distribution, this will require analytic or numerical integration within the user-specified interval. The data sets that drove the design of these techniques have uncorrelated bivariate normal distributions, for which there is a fast, analytical integral solution using the error function:

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\ dt \tag{4.3}$$

$\mathrm{erf}(x)$ is available in many mathematical libraries, and is usually implemented as a table lookup. erf can be used to compute an area under a normal distribution $N(\mu, \sigma)$ in the range $[\mu - a, \mu + a]$ as follows:

$$\begin{aligned} A(N, [\mu - a, \mu + a]) &= \int_{\mu-a}^{\mu+a} N(\mu, \sigma)\ dx \\ &= \mathrm{erf}\left(\frac{a}{\sigma\sqrt{2}}\right) \end{aligned} \tag{4.4}$$

The integral with arbitrary boundary conditions is:

$$\begin{aligned} A(N, [a, b]) &= \int_a^b N(\mu, \sigma)\ dx \\ &= \frac{1}{2}\left[\mathrm{erf}\left(\frac{b-\mu}{\sigma\sqrt{2}}\right) - \mathrm{erf}\left(\frac{a-\mu}{\sigma\sqrt{2}}\right)\right] \end{aligned} \tag{4.5}$$

This extends to 2D for the uncorrelated bivariate normal distribution $N_{xy}(\mu_x, \mu_y, \sigma_x, \sigma_y)$ by evaluating the area within a box as follows:

$$\begin{aligned} A(N_{xy}, [a, b], [c, d]) &= \int_a^b \int_c^d N_{xy}(\mu_x, \mu_y, \sigma_x, \sigma_y)\ dy\ dx \\ &= \int_a^b N_x(\mu_x, \mu_y)\ dx \int_c^d N_y(\mu_y, \sigma_y)\ dy \\ &= A(N_x, [a, b]) \cdot A(N_y, [c, d]) \end{aligned} \tag{4.6}$$

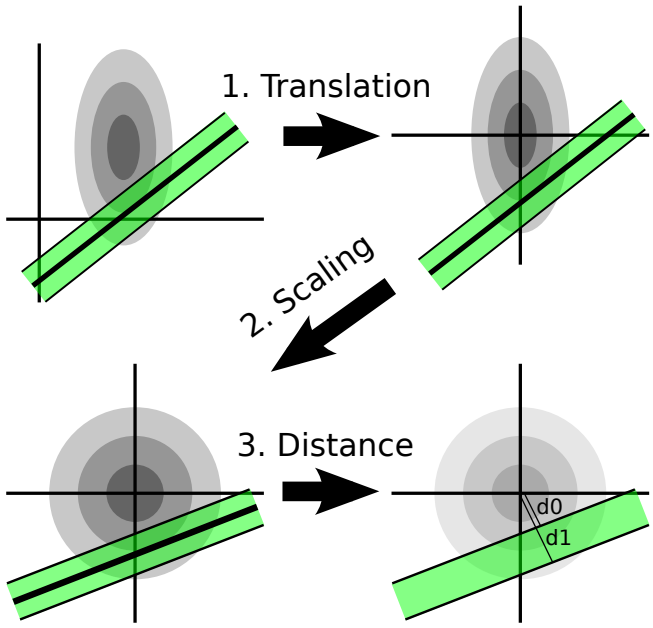


Figure 4.12: Computing the integral of a bivariate normal distribution within a distance threshold to an infinite line. Step 1: translate the line and distribution to the origin. Step 2: rescale the distribution to unit variances. Step 3: estimate the integral from the near and far bounds of the line (d_0 and d_1).

This uses the separability of the uncorrelated distributions to simplify the computation. Because probability distributions integrate to 1, a threshold test can decide if the range has selected enough of the distribution. For example, requiring that $A > .95$ means that the distribution will be selected only if the brush contains at least 95% of the distribution.

Angular brushes in parallel coordinates plots can use a very similar set of techniques as interval queries. Also, angular brushing is a subset of the more general linear function brushing, as the latter enables the user to select points using a wider set of linear functions (Feng et al., 2010a). Angular brushing is equivalent to selecting points within a distance threshold to a line in Cartesian space, which is to say within two bounding lines L_1 and L_2 around a drawn line L. Just as with interval queries, it is necessary to evaluate how much of the distribution is contained between L_1 and L_2 via integration.

As before, it is possible to evaluate this analytically for bivariate normal distributions using the error function. In this case the solution is more complex because $\text{erf}(x)$ is axis-aligned and selection lines may not be. Therefore, the coordinate space must be transformed

so that the distribution is zero-centered and radially symmetric, at which point $\text{erf}(x)$ can be applied to the distance from the line to the origin. In this transformed space, the distance from a point to the origin is called the Mahalanobis distance. The transformed distance of a line to the distribution mean can be estimated using the steps illustrated in Figure 4.12 and enumerated below:

1. Translate (μ_x, μ_y) to the origin.
2. Scale by $(1/\sigma_x, 1/\sigma_y)$ so the distribution has unit variances.
3. Compute the distance from the line to the origin.

Rotation of the lines is unnecessary because an isotropic normal distribution is radially symmetric. To evaluate the integral, it is only necessary to know how far away L_1 and L_2 are from the origin. The distance of an arbitrary implicit line $(Ax + By + C = 0)$ from the origin in transformed coordinates is:

$$d(A, B, C, N_{xy}) = \frac{-A\mu_x - B\mu_y - C}{\sqrt{(A\sigma_x)^2 + (B\sigma_y)^2}} \tag{4.7}$$

The area of the distribution within distance t to the line can be computed by constructing two lines and taking the difference of their error function results. L_1 and L_2 will have the same orientation (same A and B), but different values of C:

$$C1 = C - t\sqrt{A^2 + B^2} \tag{4.8}$$

$$C2 = C + t\sqrt{A^2 + B^2} \tag{4.9}$$

Using the implicit line equations with coefficients $(A, B, C1)$ and $(A, B, C2)$, the area between the two lines is as follows:

$$A = \frac{1}{2} \left| \text{erf}\left[d(A, B, C1, N_{xy})\right] - \text{erf}\left[d(A, B, C2, N_{xy})\right] \right| \tag{4.10}$$

If the user wishes to select all points near a line segment instead of an infinite line, one would only need to repeat the process for two more perpendicular lines that intersect the

endpoints of the line segment. Note that this process also applies for normal distributions where $\rho \neq 0$, with the additional step of rotating the oriented distribution to be axis aligned before applying the scaling in step 2. The angle of rotation (θ) required to rotate a correlated bivariate normal distribution so that it becomes uncorrelated is as follows:

$$\theta = \frac{1}{2} \arctan\left(\frac{2\rho\sigma_x\sigma_y}{\sigma_x^2 - \sigma_y^2}\right) \tag{4.11}$$

The new standard deviations after rotating about the origin can be determined by applying a rotation matrix (R) to the covariance matrix (Σ) of the distribution:

$$R = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix} \tag{4.12}$$

$$\Sigma = \begin{bmatrix} \sigma_x^2 & \rho\sigma_x\sigma_y \\ \rho\sigma_x\sigma_y & \sigma_y^2 \end{bmatrix} \tag{4.13}$$

$$R\ \Sigma\ R^{-1} = \begin{bmatrix} \sigma_{xnew}^2 & 0 \\ 0 & \sigma_{ynew}^2 \end{bmatrix} \tag{4.14}$$

When dealing with large data sets distributed across multiple computers, all user selections should occur in data space rather than in the space of the visualization. For example, selecting all points within a box on a scatter plot is equivalent to a filter on data values that fall within values ranges in x and y. In this manner, the selection operations can occur concurrently on each machine and the results can be combined on a single node and sent to the client. Direct PDF visualization only requires easily parallelizable queries to all machines and subsequent visualization of the combined results.

4.7 Discussion

This chapter has presented a set of techniques for interacting with scatter plots and parallel coordinates plots. The motivation for developing these techniques was to enable users to select visual patterns without requiring them to understand the mathematical relationships

the patterns represent. Once selected, it is trivial for the software to report the mathematics behind the selection.

Linear function brushing was designed with the principle of visual pattern matching in mind. Determining a relationship between variables based on a visual pattern in parallel coordinates is extremely difficult, especially when the axes have different ranges of values. Rather than requiring users to do this in their head, linear function brushes enable the user to draw two lines, and then the tool generates a mathematical definition for the user that can be used subsequently for selection.

Lasso brushing follows the principle of visual pattern matching as well. Selecting a set of lines that pass through at arbitrary location on a parallel coordinates plot may seem natural and straightforward, but doing so with range queries or a function brush requires careful derivation. Lasso brushing shifts the mental burden away from the user by enabling them to simply gesture over the plot to select lines of interest.

Once a selection is complete, the mathematical definition of the relationship represented by the selection can be used to create a new axis, a composite variable, that describes values that match the selection. Variables can be combined in multiple ways, depending on the desired visual representation. A new variable created by summing two variables can potentially reveal any existing multivariate linear relationships with other variables. Alternatively, the new column can describe how well two variables match a particular relationship. The values that match this relation will cluster together in a scatter plot or parallel coordinates plot. Combining variables in this way enables the user to interrogate multivariate relationships of more than two variables which can be used as classifiers for the data points.

Adding statistical uncertainty to the scatter plot and parallel coordinates plot complicates brushing. Rather than selecting discrete points in data space, brushes estimate whether it is *probable* that the user intended to select a point. This requires integration of the probability density function within the shape of the brush for each data point to see how much of its distribution was selected. This can be done efficiently for normally distributed data using the error function, which can even be used for complex brushes like the linear function brush.

CHAPTER 5

Magnetic Resonance Spectroscopy

The visualization techniques discussed in previous chapters were all designed to meet the needs of radiologists using magnetic resonance spectroscopy (MRS) to understand the metabolic description of disease processes. One prominent application is for brain tumors, one of which is depicted as a gray mass in Figure 5.1. This tumor mass that demonstrates enhancement in magnetic resonance imaging (MRI) due to an administered contrast agent. Contrast agents are useful for highlighting where brain tumors have damaged brain tissue, however they do not show the exact boundaries of brain tumors. The white curves overlaid on top of that image slice are metabolic spectra. Radiologists use MRS to study the metabolic makeup of brain tumors so that they can more accurately diagnose and treat patients with brain tumors and other diseases. This chapter gives an overview of how the MRS data is generated.

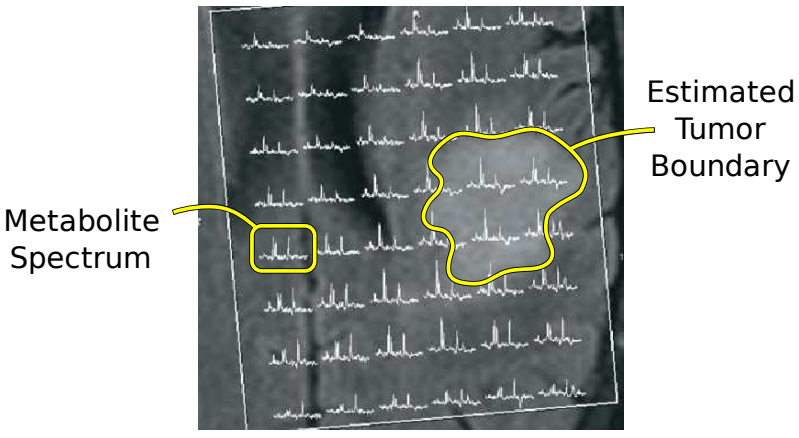


Figure 5.1: MRS data overlaid on a single slice of a T1 MRI.

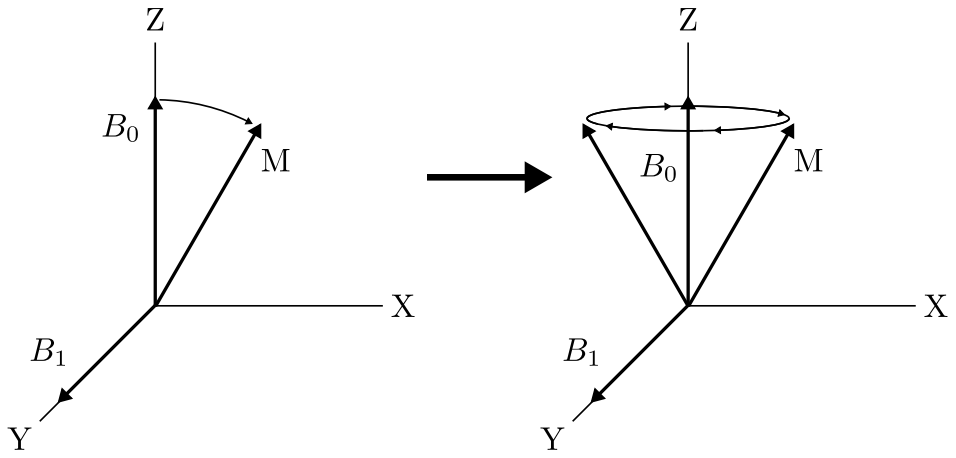


Figure 5.2: An illustration of Larmor precession. The bulk magnetization vector M is originally aligned with the dominant magnetic field B_0. When a secondary magnetic field B_1 is applied, M tips away and begins to precess around B_0.

5.1 Nuclear Magnetic Resonance

Nuclear magnetic resonance (NMR) was developed as a technique to interrogate the composition of a material sample via its response to magnetic fields. A portion of the sample's atoms placed under a strong magnetic field (B0) will align themselves with that field. The vector describing the mean magnetic alignment of a sample's atoms is called the *bulk magnetization vector* (M). When a brief secondary magnetic field (B1) is applied, M is perturbed (or tipped) away from B0 and begins to precess–or change the orientatio of the axis of rotation–around it. This is called Larmor Precession, and the frequency of rotation is called the Larmor Frequency. This tipping is illustrated in Figure 5.2.

When all of the sample's atoms precess in phase, they interfere constructively and produce a measurable electromagnetic signal. This is the fundamental signal in NMR. The magnitude of the signal is directly proportional to magnitude of the transverse (XY) component of the precessing vector M. This has several implications. First, the total angle that B1 tips M away from B0 (called the *tip angle*) influences the resulting signal. A tip angle of 90 degrees produces a fully transverse precession and therefore the strongest possible signal for a given B0. Second, a stronger B0 will produce a larger M and therefore a stronger signal.

Achieving a 90 degree tip angle is slightly more complex than simply applying a static B1.

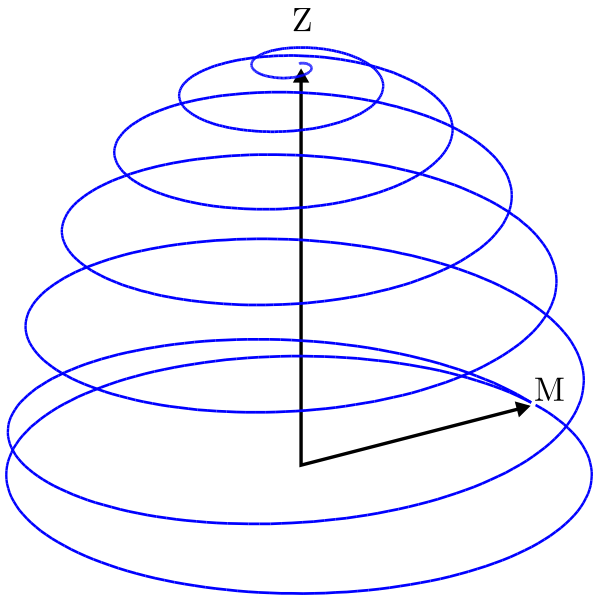


Figure 5.3: An illustration of circular polarization. When the second magnetic field is rotated at the precession frequency, it will always tip M toward the $X - Y$ plane. This is called circular polarization.

For the sake of discussion, assume that B0 is pointed in the $+Z$ direction and $B1$ is pointed in the $+Y$ direction. A static B1 will initially tip M away from B0 in the $+X$ direction. However, after M has precessed 180 degrees around B0, B1 will in fact tip M back *toward* B0. To prevent this, B1 can change periodically at the Larmor Frequency, sinusoidally alternating between a $+Y$ and $-Y$ orientation as M precesses. This is called linear polarization. Of course, it is not necessary to restrict the secondary magnetic field to a single direction. If another field B2 is added such that B1 is modulated sinusoidally in both Y direction and B2 is modulated sinusoidally in the X direction at the Larmor Frequency, M will tip away from B0 throughout the precession. This is called circular polarization, which is illustrated in Figure 5.3.

Once the secondary magnetic field is turned off, the precessing vector will gradually return (or relax) to its initial position (B0). The amount of time this "spin-lattice" or "longitudinal" relaxation takes (T1) varies from material to material and is one way to differentiate between materials.

The signal also begins to decay due to *dephasing*, which refers to the fact that the individual magnetic moments of the atoms contributing to M begin to fall out of alignment

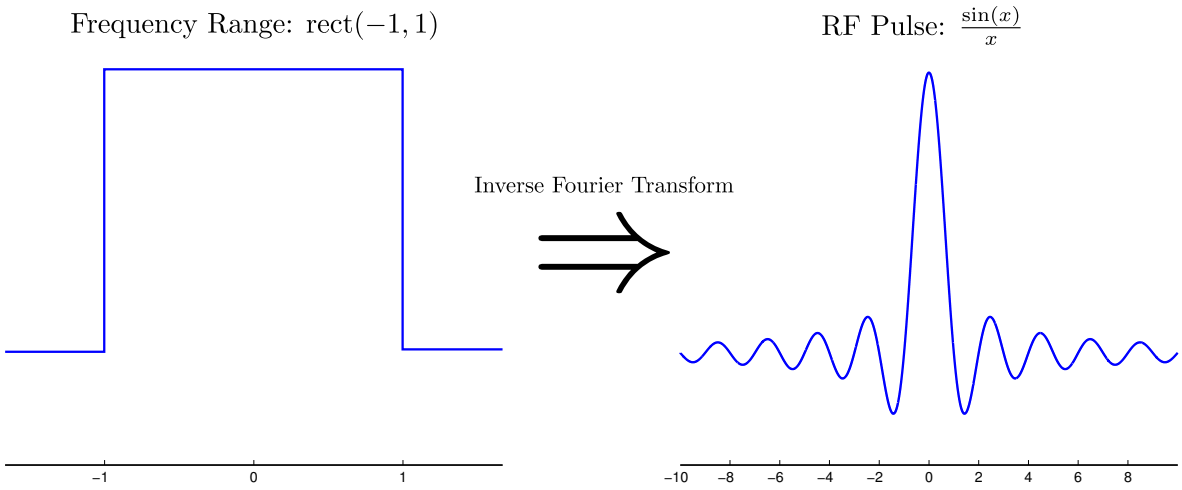


Figure 5.4: A truncated sinc function can be used to excite a range of resonance frequencies.

with each other. There are two reasons that this happens. First, inhomogeneities within the sample cause the spins of nearby atoms to perturb each other. Second, inhomogeneities in the main magnetic field (B0) similarly perturb atomic spins nonuniformly. The amount of time that this "spin-spin" or "transverse" relaxation takes is another source of tissue contrast. Measured transverse relaxation, called T2* relaxation, includes both the dephasing due to intrinsic material properties and extrinsic magnetic field inhomogeneities. T2 relaxation refers specifically to dephasing due to intrinsic material properties. T2 an T2* are both much smaller than T1.

Different atoms precess with different Larmor Frequencies. This is related to the number of electrons that the atom possesses. If a 90 degree pulse could be applied at a single frequency, the magnitude of the resulting signal would be proportional to the number of one type of atom in the sample. Of course, a signal containing only a single frequency has infinite extent in time, so it is not possible to selectively excite a single type of atom. However, if the pulse contains multiple frequencies, the signal will contain all of the oscillations of all resonating atoms. To equally excite a range of frequencies (a rectangle in frequency space), the pulse to use is a sinc function, shown in Figure 5.4. Sinc functions also have infinite extent in time, however the magnitude of oscillation decreases quickly, so they can be truncated without introducing too many aliasing artifacts. Shorter sinc functions excite a wider range of frequencies. Because different atoms resonate at different frequencies, the Fourier Transform of the resulting signal

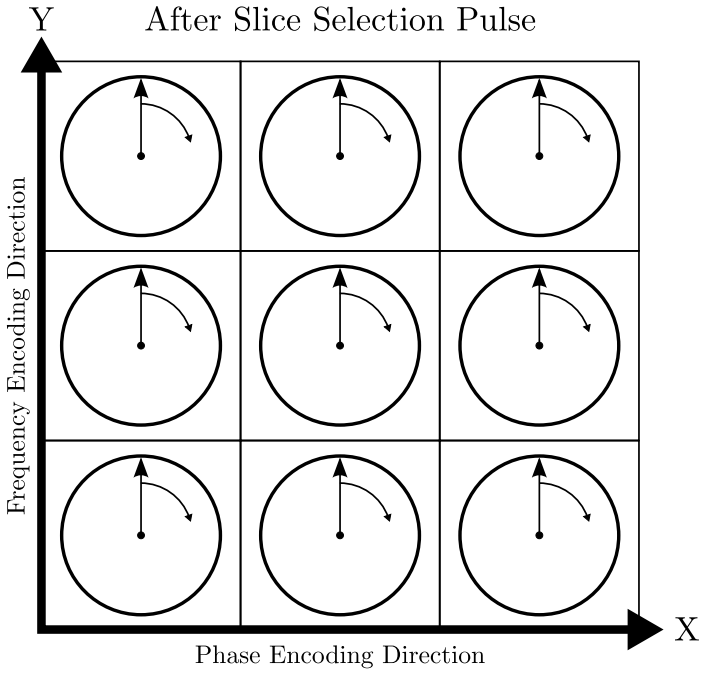


Figure 5.5: Applying an RF pulse while a gradient in the Z direction is active selectively tips the vectors of atoms within a range of Z values. The vectors for one atom type will all have roughly the same phase and precession frequency.

displays a spectrum with peaks corresponding to different atoms in the sample. To excite atoms with a range of precession frequencies, B1 can be applied during a radio frequency (RF) pulse that isolates those frequencies. This technique, called Fourier Transform Spectroscopy, is one of many ways to capture such spectra. Because it uses short pulses, it is relatively fast.

5.2 Magnetic Resonance Imaging

MRI adds spatial localization to the principles of NMR (**?**; Castillo, 2002). The fundamental principle that makes MRI possible is the fact that the Larmor Frequency of an atom is proportional to the magnitude of its surrounding magnetic field. The discussion of NMR assumed that B0 was uniform. Indeed, for the signal to be trustworthy a great deal of engineering is required to ensure a highly uniform B0. However, if B0 can be controlled spatially, it is possible to have the Larmor Frequency of a particular atom vary spatially. The most common usage is to apply a secondary gradient that causes B0 to vary linearly in the Z direction, which causes the Larmor Frequency to vary linearly in that same direction. This

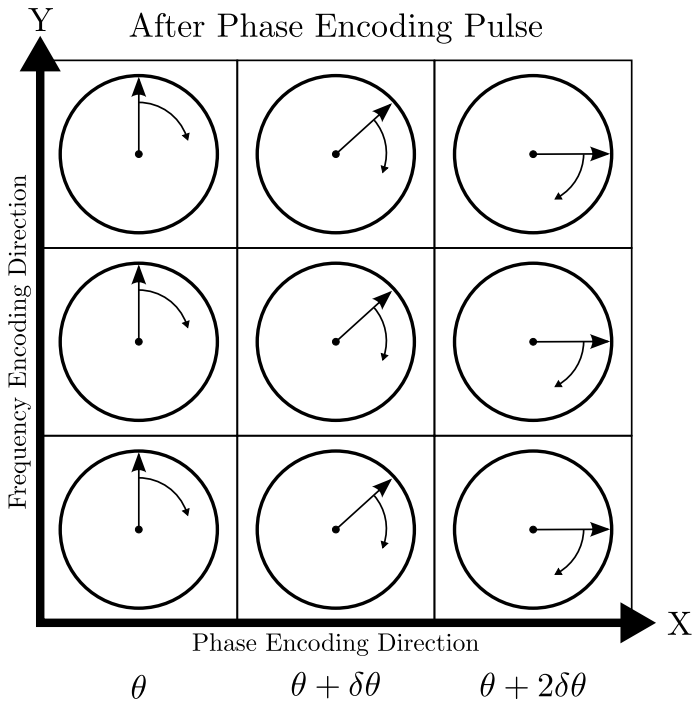


Figure 5.6: Applying a second linearly varying gradient in the X direction drives the precessions at different frequencies. When that gradient is turned off the precessions revert to their original frequencies, but they are now out of phase.

is called a *slice selection gradient*. By applying a slice selection gradient at the same time as the RF pulse, only those atoms with resonant frequencies contained within the pulse have their magnetic moments completely tipped over. As stated before, an RF pulse cannot excite only a single frequency, as this would require an infinitely long sinusoidal signal. In reality, the RF pulse will excite a range of frequencies. Because the slice selection gradient causes B0 to vary linearly, a range of tissue in Z will be tipped simultaneously. The thickness of the slice is directly proportional to the range of frequencies present in the RF pulse. Once a slice has been selected with the appropriate gradient, all of the spins will be precessing with approximately the same frequency and phase, as shown in Figure 5.5. This means that the resulting signal is the combination of all of the signals produced within that slice. It is possible to isolate a particular position in that slice by similarly applying gradients in the X and Y directions. Consider what happens when a gradient in X is applied for a short period of time after the slice selection gradient is turned off. The moments of atoms in the excited slice will now precess at spatially varying Larmor Frequencies. Once the X gradient is turned

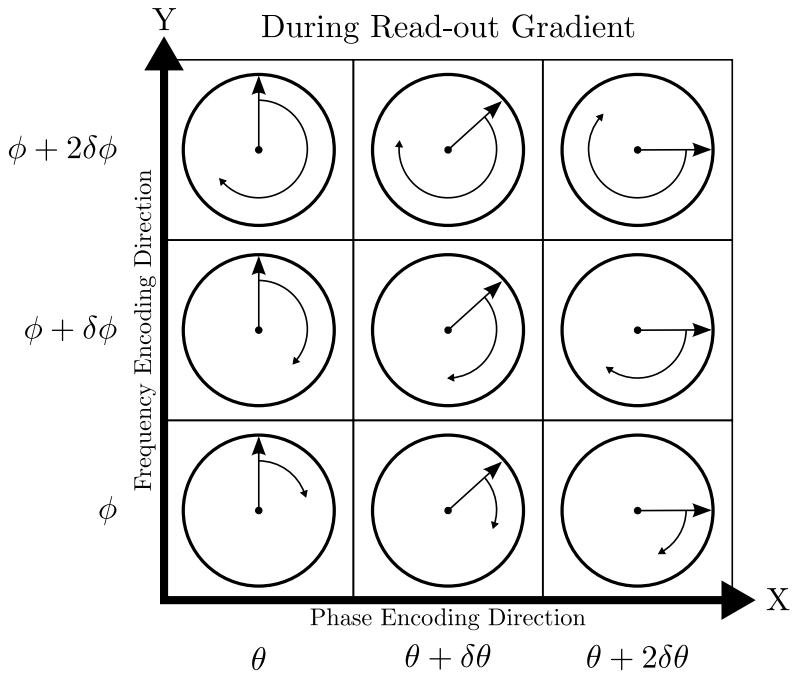


Figure 5.7: A gradient in the Y direction causes precession frequencies of the atoms to vary spatially in the Y direction. A strong response on the top of the image will correspond to a strong, quickly varying component to the signal. A strong response on the bottom of the image will correspond to a strong slowly varying component to the signal.

off, the moments will return to their original precession frequencies but will have spatially varying phase differences, as illustrated in Figure 5.6. Such a gradient is therefore called a *phase encoding gradient*, G_x. Applying another orthogonal gradient in the Y direction again causes a spatially varying precession frequency that retains the phase differences induced by G_x. This second gradient is called the *frequency encoding gradient*, G_y. While G_y is active, each spatial position has a unique combination of phase and precession frequency, as shown in Figure 5.7. In the frequency encoding direction (Y), different positions oscillate with different frequencies depending on the magnitude of G_y. Similarly, phase differences in the phase encoding direction (X) will vary sinusoidally with frequency depending on the magnitude of G_x. Reading out the signal while G_y is active (and G_x has been turned off) results in a time signal that shows different spatial frequencies in the Y direction for one phase variation frequency. For example, if the initial magnitude of G_x is small, the phase angles will vary only slightly from position to position, which essentially isolates the portion of the signal that has slow spatial variation. If G_x has a large magnitude, the phase angles will vary

quickly, which isolates the portion of the signal that has fast spatial variation. If this process is repeated for range of phase angle variations (different magnitudes or durations of G_x), we acquire a 2D frequency space representation of the data. Applying a 2D inverse Fourier Transform therefore recovers a spatial signal in which the magnitude of a value corresponds directly to the number of responding atoms at each that location.

A simple way to think about the phase and frequency encoding gradients is that they are vectors that control read out position in frequency space (often called k-space). A normal sequence consists of driving the read out position to one side of the frequency image, then reading out a column using G_y. G_x then moves the read out position to a new column and G_y proceeds again. Note, however, that the signal decays over time due to T1 relaxation, so frequency space may not necessarily be filled in one pulse sequence. The safest (and slowest) way to ensure good signal-to-noise is to read out one column per pulse, allowing the moments to recover to B0 in between. This is extremely slow ($\approx$ 5T1). Many of the more complex MRI sequences are simply more efficient traversals through frequency space using G_x and G_y that require fewer pulse sequences at the cost of worse signal-to-noise ratio.

Depending on the parameters of the gradient sequence, the resulting images will show tissue contrast based on local values of T1 and T2. This is the most useful for highlighting changes in soft tissue, such as gray matter, white matter, and cerebrospinal fluid (CSF). One common way to amplify contrast in regions of damaged tissue is to administer a contrast agent to the patient. These agents are commonly heavy metals that respond powerfully to magnetic fields, so they appear very bright in standard MRI sequences. The contrast agent response depends on the magnetic properties of the contrast agent. Paramagnetic materials have magnetic properties only in the presence of an external magnetic field; they amplify the local magnetic field properties to produce brighter spots in MRI. Diamagnetic materials create a magnetic field that opposes an external magnetic field, and therefore results in dark spots in MRI. Ferromagnetic materials are essentially permanently magnetic (no external field is required), and as a result also perturb the local relaxation times. These agents are useful for brain imaging because they tend to aggregate in regions where the blood-brain barrier

(BBB) has been compromised. The BBB normally acts to protect the brain from blood-borne diseases, but it gets compromised by brain tumors and other degenerative diseases. The contrast agents then pass through the compromised BBB and aggregate in damaged regions.

5.3 Spectroscopic Imaging

MRS refers to the capture of entire atomic spectra at particular spatial locations rather than just the hydrogen atom density. This is exactly the same as NMR, except that it is restricted to a voxel of tissue rather than all of the tissue, as described below. Once the response has been isolated to the voxel, the resulting time signal will contain different oscillation frequencies that correspond to the Larmor frequencies of different molecules. A Fourier transform takes the time signal and converts it into frequency space in which peaks correspond roughly to different metabolites (in the context of brain imaging). The position of a peak along the frequency domain axis is often called its "chemical shift." As a result, this type of imaging is also called chemical shift imaging.

The first step of MRS imaging is to apply multiple slice-selective gradients to isolate the magnetic response to a single voxel. There are many sequences that perform this selection (Klose, 2008). This section describes the point-resolved spectroscopy sequence (PRESS) to illustrate the basic concept. There are three steps to the PRESS sequence:

1. 90° Z slice selection pulse
2. 180° Y refocusing pulse
3. 180° X refocusing pulse

The first step is identical to the slice selection pulse used in MRI. The second step takes the spins precessing from the previous pulse and tips them to 180° away from B0. Doing so while a gradient in the Y direction is being applied restricts the response to a range of Y positions. As they recover longitudinally, the spins produce a spike in the signal as they pass through pure transverse precession. This is called an echo, and the duration required to

recover from 180° to 90° is called an echo time (TE). This is followed by a second 180° pulse, this time using an X gradient to isolate a range of X positions. The resulting echo contains frequencies that correspond to molecules in a single voxel of tissue.

For biological tissue like the brain, the resolvable peaks correspond to different metabolites, which are markers for various types of brain functionality (Soares and Law, 2009). To capture a full 3D image requires that this sequence be repeated for each voxel. While different sequences can capture this information more quickly (often sacrificing frequency resolution as a result), MRS sequences are invariably much slower than standard MRI sequences.

Metabolic information is particularly useful because standard techniques for identifying tumors and other disease processes can be inaccurate with hydrogen-based imaging. For example, a contrast-enhanced tumor may show clear boundaries, but radiologists know that these boundaries only highlight a compromised blood brain barrier; tumor cells may still be invading tissues at other locations. The goal of this work is to help radiologists correlate the features in metabolic data sets with tumors so as to better identify potentially cancerous areas of the brain.

The raw spectroscopy data consists of per-voxel metabolite spectra in which the heights of different spectral peaks correspond to different metabolite concentrations. The spectra for atoms larger than hydrogen are relatively weak and therefore noisy, so extensive expertise is necessary to properly understand peak correspondence and metabolite interactions between voxels. Additionally, raw spectra are known to exhibit several characteristics that make interpretation problematic. These include a background baseline noise spectrum, which can be difficult to differentiate from local information, and peak broadening due to eddy currents and other magnetic field inhomogeneities.

5.4 LCModel: Estimating Absolute Concentrations

An offline processing system called LCModel estimates absolute metabolite concentrations from the measured *in vivo* spectra (Provencher, 1993). LCModel uses high quality spectra captured from pure samples of single metabolites as the basis set for a linear least-squares

fit to *in vivo* spectra. By matching the sequence parameters used to capture the basis set to the sequence parameters for new sequences, the estimation process automatically takes into account many sequence irregularities.

The regression process automatically optimizes several parameters, including:

- peak scale (concentration)
- peak position and shift
- peak broadening
- baseline shape
- noise

After fitting the linearly combined set of optimized basis spectra to the *in vivo* spectrum, the final output is an absolute concentration and a standard deviation for each metabolite. The standard deviation of the noise is essentially a quality metric describing how well the optimization matches the *in vivo* spectrum. For example, if the peak actually represents a metabolite not contained within the basis set, it is likely that the standard deviation of the noise parameter will be very high.

The standard deviations are estimations of how the noise in the original spectra is propagated in the linear least squares fit. More formally, they are called standard error estimates or Cramer-Rao lower bounds (Provencher, 1982), and they define a lower bound to the variance in the concentration estimates. One way to understand this conceptually is in terms of error propagation. Linear regression essentially applies a linear matrix operator to a set of samples (in this case a measured spectrum), which produces a set of parameter estimates (in this case metabolite concentrations). The operator magnifies any noise present in the basis spectra. This means that different concentration estimates will have different, location-specific sensitivities to changes in the measured spectrum. The Cramer-Rao lower bounds combine the basis spectrum noise estimates with analysis of the least squares operator to determine the variance of the resulting concentration estimates. In this case, the lower bounds are calculated by scaling basis spectrum noise estimates by the diagonal elements of the inverse of the

linear operator. Conceptually, the resulting values describe how initial noise estimates are magnified by the operator, thereby giving an indication of the minimum amount of variance to be expected in any estimated parameter values.

Depending on the imaging sequence, LCModel estimates the concentrations of at least 20 different metabolite concentrations per spectrum. The data sets commonly associated with brain spectra, including choline (cho), creatine (cr), inositol (ins), glutamine (glu), and N-acetylaspartate (NAA), are selected for further processing and visualization. While it may be difficult to differentiate two overlapping peaks in a spectrum, LCModel is capable of fitting combinations of spectra. For example, if n-acetyleasparate (NAA) and n-acetylaspartylglutamic acid (NAAG) are too close together, LCModel can estimate the concentration of NAA+NAAG instead.

A Siemens Allegra 1.5T MRI scanner captures both brain MRI and MRS images. Each acquisition session lasts approximately six minutes. The metabolite spectra are sampled at a voxel size of $\sim 1cm^3$, with resolution on the order of $20 \times 20 \times 10$. Because both anatomical MRI and MRS data sets are captured simultaneously for a single patient, no additional registration is necessary. Standard T1 sequences produce anatomical images with $\sim 1mm^3$ voxel sizes.

CHAPTER 6

nDive: n-Dimensional Volume Explorer

(The contents of this chapter were presented at the ACM Symposium on Applied Perception in Graphics and Visualization 2009, the Conference on Information Visualization 2010, and the Conference on Visualization and Data Analysis 2010 (Feng et al., 2009; Feng et al., 2010a; Feng et al., 2010b))

6.1 Goals

Exploratory visualization is the first step in this process of understanding the MRS data set. As stated in Chapter 1, the radiologists have two primary goals:

1. **Identify Metabolite Relationships:** Radiologists use relationships among metabolites as indicators of disease.

2. **Extract Metabolite Concentrations:** Once radiologists identify a tumor, clinicians and surgeons need to extract absolute metabolite concentrations at precise 3D locations to accurately plan procedures. This goal requires viewers to have positional awareness within the data space and direct access to raw data values.

6.2 MRS Visualization

The current state of the art in medicine for visualizing MRS data does not display multiple metabolite concentration fields at once. The visualization for UNC radiologists overlays metabolite spectra over each voxel of a slice through the anatomical data, as shown in Figure 5.1. Maudsley et al. use a spectrum pseudocoloring to encode a single metabolite's concentration on top of a gray scale anatomical image (Maudsley et al., 2006). Cliniviewer displays multiple metabolites in an array of 2D slice planes (Uttecht and Thulborn, 2002). Chang et al. use the computed ratio of choline to creatine to diagnose gliomas, a particular type of brain tumor (Chang et al., 2004); they superimpose a grayscale, pseudocolored computed field over a grayscale anatomical slice plane.

UNC radiologists requested a technique for displaying multiple metabolite fields at once while supporting relationship identification rather than the display of a single metabolite relationship. This chapter describes a novel visualization system that was designed to address these specific visualization goals.

6.3 nDive

nDive is a system that links together the visualization techniques described in previous chapters. It has been used successfully to identify metabolic patterns that successfully locate tumor tissue. Additionally, several doctors familiar with spectroscopy data were interviewed to gain a qualitative evaluation of the merits of nDive and how it could be integrated into their workflow.

nDive presents four views of the MRS data, two spatial and two abstract. The first spatial view is an SDDS visualization that conveys the general spatial trends among metabolites (see Figure 6.1). The default color coding for the spheres is to map the chemical shift spectrum to a visible spectrum. Some of the more important metabolites are colored as described in Table 6.1.

The size and density of the spheres are initially set so that all spheres for all variables can fit into a single voxel without needing to overlap as described in Section 2.2, although random

Metabolite	Color
Choline	Red
Creatine	Orange
Glutamine	Yellow
n-acetylasparate	Green

Table 6.1: The colors assigned to some of the more important MR spectroscopy metabolites.

perturbations of position may produce overlapping spheres. The user can change sphere size, density, and color interactively.

The second spatial view is a pseudocolored anatomical slice plane that depicts a single metabolite field (or composition of multiple metabolite fields). The anatomy is shown in grayscale, which is then overlaid with an isoluminant color map for the spectroscopy image. When the user clicks on a pixel in the slice plane, text appears indicating both anatomical intensity and metabolite concentration(s) at the clicked location. This view targets the surgical planning visualization goal by enabling users to both visually and manually explore the data values in a single image. Both spatial views are interactive; the user can select spectroscopy voxels in either view and analyze them. For example, the user can select a set of voxels that are predominantly located in the white matter region of the brain and nDive will present simple summary statistics of values in the selection including the mean and variance. Metabolite voxel selections in the spatial views are propagated to the two remaining views.

The scatter plot and parallel coordinates plot are augmented to incorporate data value uncertainty using the techniques discussed in Chapter 3. In these views the user can look for patterns such as linear relationships and clusters among the metabolite data values. The scatter plot can be rearranged to display any pair of metabolites. The parallel coordinates plot shows all loaded metabolites at once, and the axes can be rearranged by click and dragging an axis to its desired location. See Section 6.4 for an illustration of these techniques.

Both the scatter plot and parallel coordinates plot support interval selection and linear selection of uncertain data. The parallel coordinates plot also supports lasso selection, as described in Chapter 4. nDive supports the creation of multiple selection groups, which are distinguished by their color in the plots. As selections are made in these plots, selected voxels appear as wireframe boxes in the spatial views. The selected values and summary statistics

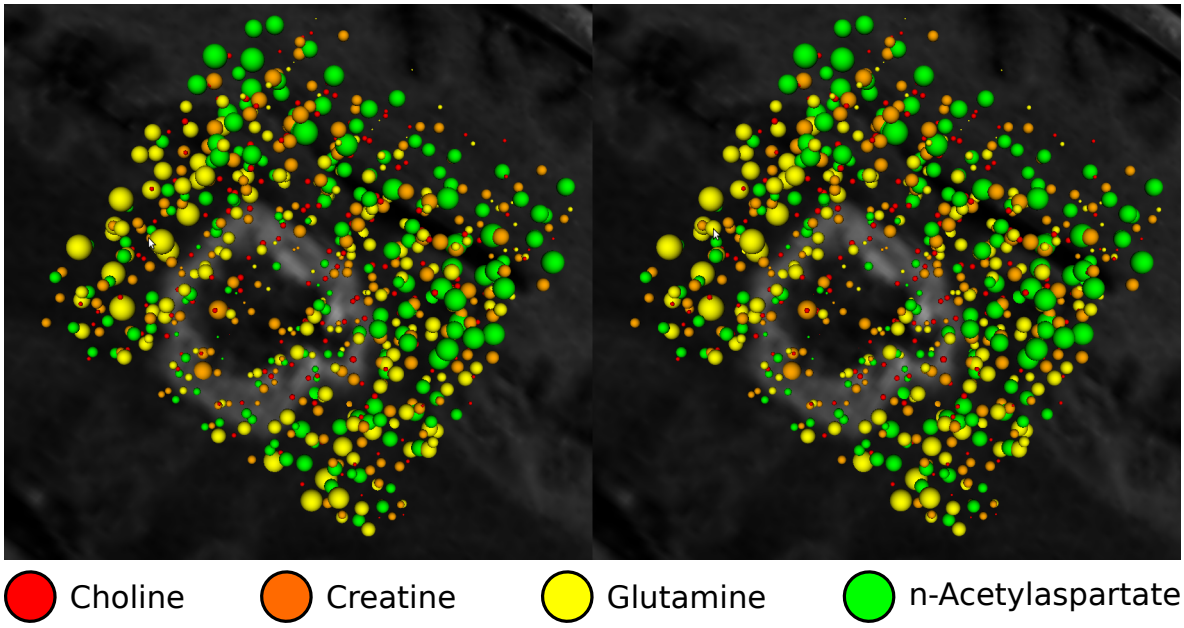


Figure 6.1: A cross-eyed stereo SDDS view of an MRS data set for a patient with a probable brain tumor. Stereo display is important for all multivariate 3D techniques.

for those selections can be exported to a CSV file for further processing.

Refer to the Appendix for a more thorough discussion of nDive's features, how to acquire it, and how to compile it.

6.4 Tumor Analysis

This section demonstrates a notional usage scenario of a radiologist using nDive to analyze MR spectroscopy from a patient with a brain tumor. The SDDS visualization shown in Figure 6.1 also contains an anatomical T1 slice plane. The tumor is visible in the slice plane because the patient has been administered an intravenous contrast agent. In this data set there are four metabolites: choline (Cho), creatine (Cr), glutamine (Gln), and N-acetylaspartate (NAA). First, the user views the SDDS visualization to get a global sense of the spatial relationships between variables. In this image, several relationships are apparent. All of the metabolites appear depressed inside of the tumor, except for a small voxel in orange (Cr). The green (NAA) and orange (Cr) metabolites appear to be positively correlated outside of the tumor. Also, green (NAA) and orange (Cr) appear to be negatively correlated with yellow (Gln)

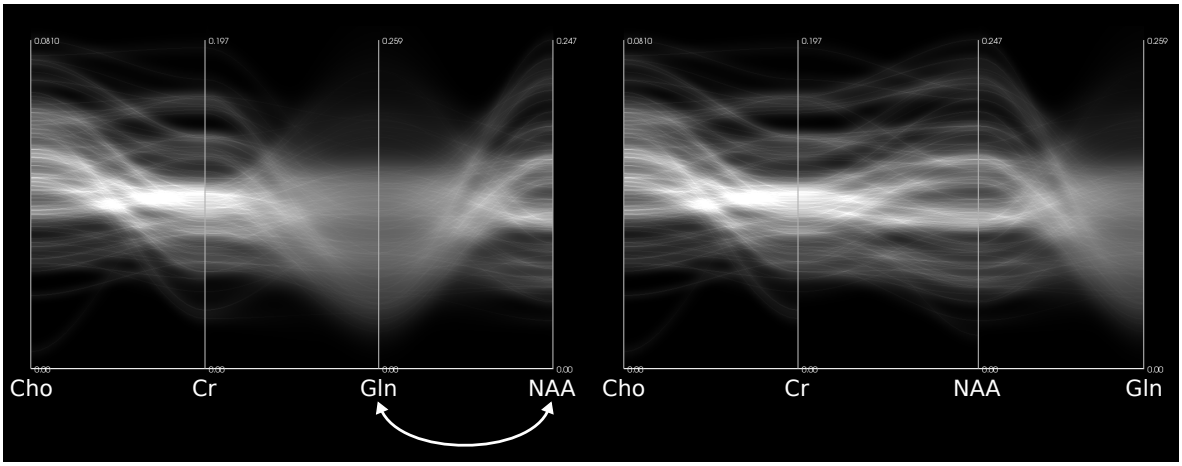


Figure 6.2: The Glutamine axis is sufficiently uncertain that it offers little useful information. The viewer can reorder the axes so that Glutamine is off on the side of the plot.

outside of the tumor.

Figure 6.2 illustrates how a density-based PC plot of MRS data with mean emphasis prevents the viewer from noticing an erroneous cluster of values in the Glutamine column. In fact, the density plot clearly reveals that the Glutamine values are all sufficiently uncertain that Glutamine should be discounted entirely for investigation. The viewer therefore reorders the axes via click-and-drag to place Glutamine off to the side of the plot.

Radiologists are aware that the ratio of Choline to Creatine is a useful tumor indicator. The viewer therefore examines a density-based scatter plot comparing Choline and Creatine, as shown in Figure 6.3. The density plot reveals a bulk of values with an apparent positive linear relationship, and mean emphasis highlights the sub-population of values clustered below the large group. The selection of these values using a linear function brush shows that this relationship between Choline and Creatine isolates tumor voxels reasonably well. The parallel coordinates plot shown in Figure 6.4 shows the values of the selected voxels in other variables. The selection (red lines) appears to have relatively low values of NAA.

Figure 6.5 depicts a second selection, in which a separate population of values is selected via an angular brush on the PC plot. This defines a second relationship between Choline and Creatine (higher this time), which tends to select voxels outside of the tumor. This new selection also tends to have higher values of n-Acetylaspartate (NAA), which supports the

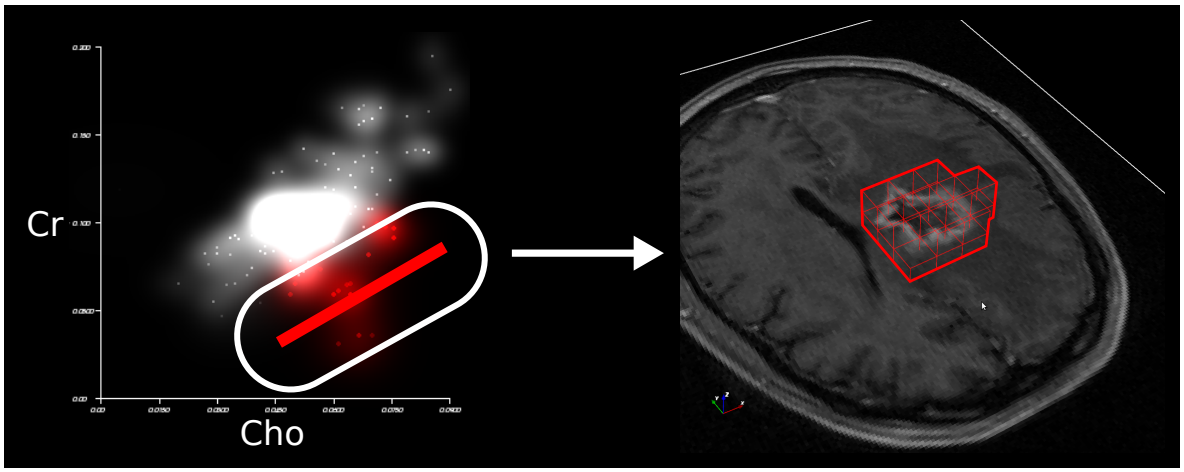


Figure 6.3: Left: this Choline-Creatine scatter plot reveals a sub-population of voxels that are somewhat different from the majority of the voxels that have a positive linear correlation. Right: the selected set of voxels (red) tend to be within the tumor.

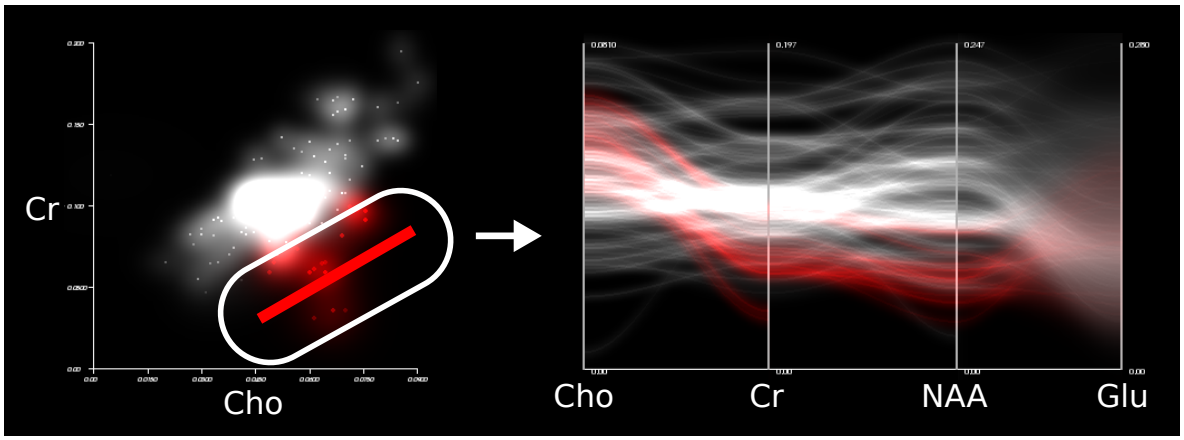


Figure 6.4: The selection (red) isolated in Figure 6.3 can be analyzed in other variables by looking at this parallel coordinates plot. This plot shows that the selected voxels have relatively low NAA.

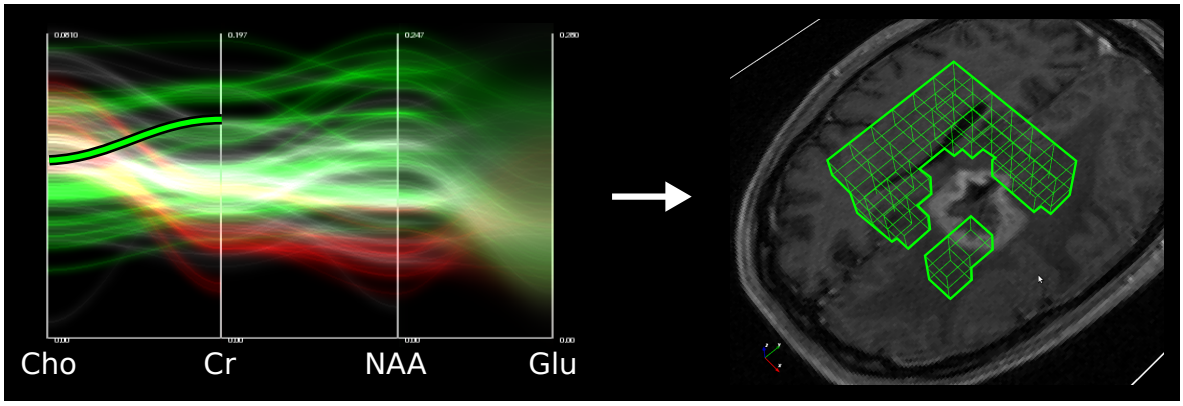


Figure 6.5: Selection of a second set of voxels (green) in the parallel coordinates plot with an angular brush highlights voxels that appear to be outside of the tumor. These values also appear to have higher values of NAA

.

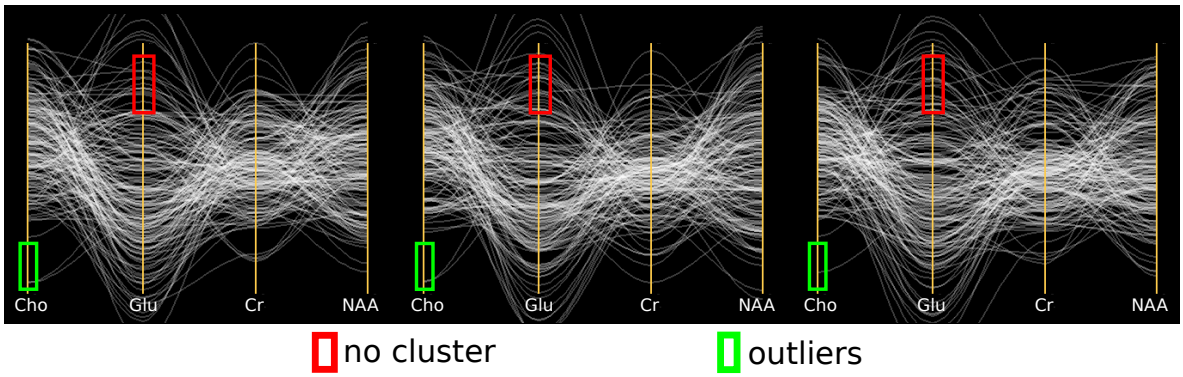
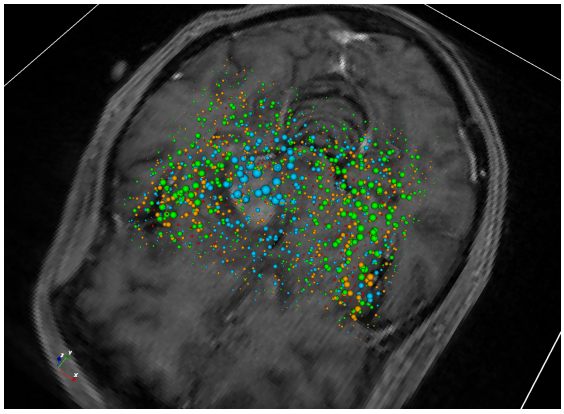


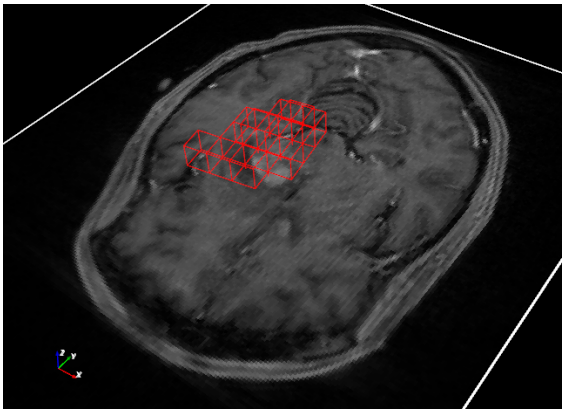
Figure 6.6: Several frames from the animated probabilistic plot revealing outliers in the PDF of the data.

hypothesis drawn after the first selection.

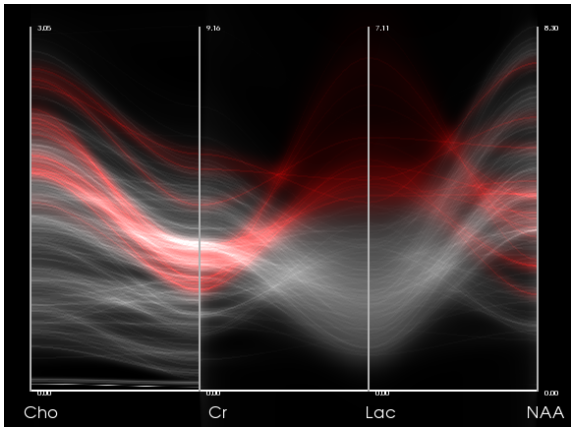
Looking at the MRS data in a probabilistic plot is also useful. As shown in Figure 6.6, the animated plot reveals outlier values in the PDF. Selection reveals that those voxels are in the center of the tumor despite having a Choline-to-Creatine ratio that differs from the other tumor voxels. This discovery warrants further scrutiny, as it may indicate that the center of a tumor may have a different signature than invasive tumor tissue. Also, a comparison of the Glutamine column in Figure 6.6 to the Glutamine column in the standard PC plot in Figure 6.2 shows that random sampling has produced no clusters in this frame. While a specific set of random samples may reproduce the cluster, plot animation fills in the region over time on average.



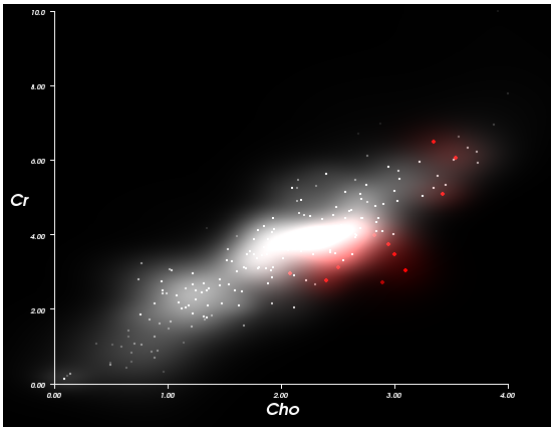
(a) SDDS



(b) Anatomy + Selection



(c) Parallel Coordinates Plot



(d) Scatter Plot

Figure 6.7: nDive applied to a potential central nervous system lymphoma. Lactate spheres are blue, NAA spheres are green, choline spheres are red, and creatine spheres are orange. Notice how lactate is much higher within the contrast-enhanced region in this type of tumor.

It is important to remember that while these types of visual explorations are possible with standard scatter plots and parallel coordinates plots, uncertain density plots enable the viewer to preattentively focus on the more useful information. They guide the viewer's search away from unreliable data points and thereby reduce the rate of false conclusions.

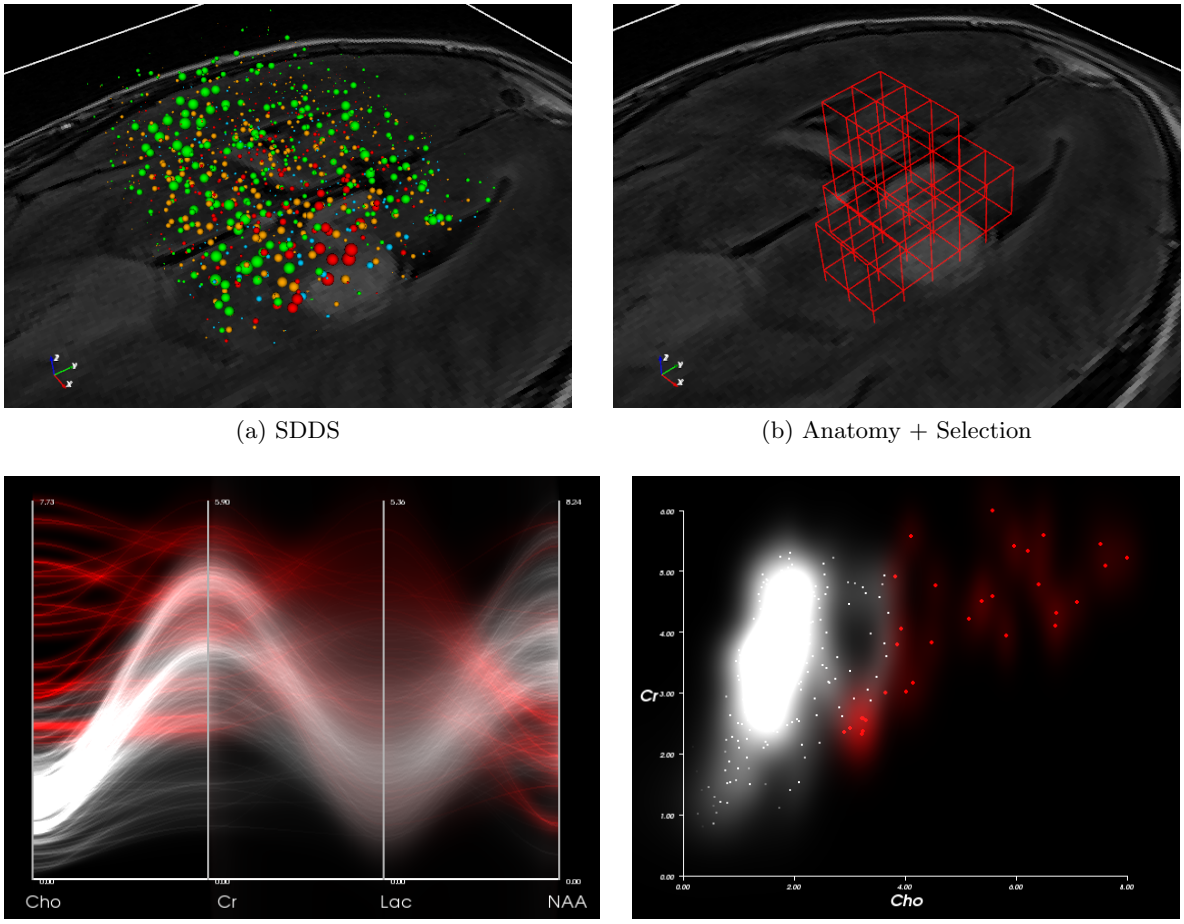
Applying the same type of analysis to other types of brain tumors has shown that different types of tumors have significantly different metabolic signatures. Figures 6.7 and 6.8 show two potentially different types of brain tumors: a central nervous system lymphoma and a gliomatosis cerebri, respectively. The former case shows higher lactate in the contrast-enhanced region and lower choline. The latter case shows higher choline in the contrast-enhanced region and lower lactate. Such differences indicate that no single function of metabolite concentrations exists for identifying tumors, but rather different tumors will have their own functions. This indicates that nDive may in fact be useful beyond the hypothesis-generation phase because the clinicians need to tune specific functions for each tumor. At the very least, having multiple pre-loaded choices would be useful for evaluating tumor type.

6.5 nDive Evaluation

nDive was originally designed to satisfy the needs of those who use and research MRS. To evaluate how well nDive achieves this goal, a qualitative usability study of doctors and researchers familiar with the data and methodology was performed. Six doctors doctors were interviewed, including neuroradiologists, a spectroscopist, a neurologist, and a neurosurgeon. The goal of the study was to determine how well trained professionals for whom the software was designed could learn to interpret the visualizations and whether they found the combined linked interaction system useful for how they study MRS.

The doctor interviews all followed the same general structure but were allowed to deviate when the participants had their own questions. The structure was as follows:

- **Introductory Presentation**: participants were given a 10-15 minute presentation introducing nDive and its intended usage.



(a) SDDS

(b) Anatomy + Selection

(c) Parallel Coordinates Plot

(d) Scatter Plot

Figure 6.8: nDive applied to a potential gliomatosis cerebri. Lactate spheres are blue, NAA spheres are green, choline spheres are red, and creatine spheres are orange. Notice how choline is much higher within the contrast-enhanced region in this type of tumor.

- **Live Demonstration**: participants were guided through a demonstration of nDive operating on a de-identified image of a brain containing a brain tumor.

- **Guided Discussion**: participants were asked a set of questions designed to elicit their qualitative impression of effectiveness of nDive for their work. The conversion was allowed to deviate as interesting topics presented themselves.

Throughout the discussion participants were all asked the same set of questions, although the direction of the conversation was allowed to deviate to subjects that participants wanted to discuss. Questions common to all discussions were as follows:

- Are you able to see metabolite relationships in the SDDS view?

- Is the SDDS view useful for your work?

- How do you envision using nDive for your work?

- Are you able to see metabolite relationships in the scatter plot?

- Is the scatter plot view useful for your work?

- Are you able to see metabolite relationships in the parallel coordinates plot?

- How hard is it to interpret/understand the parallel coordinates plot?

- Is the parallel coordinates plot useful for your work?

- Does the blurring give you a sense of the uncertainty in the data?

- Which view or views do you prefer for your work?

- How well would nDive fit into your workflow?

- What can be changed to improve workflow integration?

6.5.1 Neuroradiology

Neuroradiologists are doctors that focus on diagnosing and studying abnormalities centered in the head, spine and neck. MRI is one of the most important modalities used in neuroradiology because it is non-ionizing (insufficient energy to ionize atoms or molecules) and therefore relatively safe for patients. Because neuroradiologists tend to spend a majority of their time diagnosing their illnesses by reading patient scans, nDive's primary utility in this context is clinical in nature. As such, many of the comments gathered from the three neuroradiologists interviewed tended to focus specifically on how nDive could be used in the clinic and how it could be improved from the standpoint of diagnostic practicality. However, despite these similarities, the reviews from the three neuroradiologists were remarkably different. The following sections therefore discuss the comments of each of neuroradiologists separately. To preserve anonymity they will be referred to as NRA, NRB, and NRC. NRA and NRB are both board-certified neuroradiology attending physicians, both having completed neuroradiology fellowships. NRC is a neuroradiology fellow currently pursuing certification in neuroradiology.

Neuroradiologist A (NRA)

Because neuroradiologists traditionally spend the majority of their time looking at 2D images of the brain, NRA showed a strong preference for interacting with the 2D pseudocolored anatomical slice plane over the SDDS view. This is partially because the 2D view is useful for 2D data, which NRA handles more commonly. The ability to select voxels in the 2D view and see how they appear in the parallel coordinates plots and scatter plots seemed particularly useful for NRA as a way distinguish between different types of tissue. Specifically, NRA predicted that doing a simple visual comparison of two different colors on the scatter plot would be a useful way of distinguishing between tumor tissue and healthy tissue. In this case, NRA would specifically use this feature to compare contrast-enhanced voxels to voxels near the enhanced region.

NRA did not prefer the 3D view for understanding the 3D structure of the data for three reasons:

1. NRA is not used to looking at things in 3D.

2. The clutter and density of the visualization made it hard to understand.

3. Full 3D scans are use not infrequently, but not often. Usually they are only used around a particular region of interest, for example around a tumor.

NRA found the the scatter plot and parallel coordinates plots interesting primarily because of the ability they afford to select voxels and see their locations with an anatomical reference. NRA particularly liked the brushing techniques that enable selection of voxels with correlations among metabolites (angular brushing, linear function brushing). Initially the parallel coordinates density plot was difficult to understand, but after one or two explanations NRA was able to understand the uncertain density plot. This was confirmed by NRA's ability to find correlations in the data visually without prompting.

NRA did not find new axis construction in the parallel coordinates plot particularly useful, although NRA envisioned creating an axis that is composed of all of the metabolites associated with tumors (e.g., choline and lactate) being potentially useful. By dragging this new axis across the plot, it would possible to search for linear correlations between known dangerous metabolites and other metabolites of interest.

NRA's most prominent comment was excitement at the ability to look at absolute metabolite concentrations. Previously NRA has only been able to confidently look at ratios of metabolites. NRA also made the interesting point that that visualizing both absolute concentrations and ratios within the same plot would be a good way to transition from thinking in ratios to thinking in absolute concentrations.

For ideal workflow integration, NRA said that nDive should interface directly with the local image database. Although specialized software existing on a separate workstation has some precedent, nDive would be the most useful if it were integrated directly into the image reading software currently in use. NRA indicated that nDive would be used primarily for comparing populations of voxels with respect to particular metabolites (choline in particular).

NRA suggested three new features that would make nDive more useful from the clinical perspective:

- **Ratio Thresholding**: it would be useful to be able to select a cutoff ratio of two metabolites and show all matching voxels (e.g. above).

- **Raw Spectrum Access**: NRA would like to be able to "drill down" to see the spectra to evaluate the quality of fit for individual spectra. That should be integrated into the system.

- **Perfusion**: it would be good to look at perfusion data (blood flow volume measurements) in nDive. Perfusion is a useful discriminator for tumors because tumors have higher blood volume levels. Looking at both perfusion and spectroscopy on the same plots would be very useful.

Neuroradiologist B (NRB)

The second neuroradiologist interviewed has background and training similar to NRA. NRB also directs an MRI research laboratory and as such has a specific understanding of MRS in both clinical and research settings. NRB has also been able to observe MRS technology develop over time, and as such has an interesting perspective on the different ways to interpret spectra. For NRB, MRS has two primary functions: diagnosis confirmation and comparison of tissue in and around contrast-enhanced regions of the brain.

One of NRB's consistent comments was that some of the visualizations presented in nDive are overly complex for standard clinical use, although with minor modification that may not necessarily be true. Clinically, neuroradiologists are not necessarily interested in exploration of the data; they have a hypothesis concerning a specific location and want a simple, intuitive way of confirming that hypothesis. nDive was designed with data exploration in mind and focuses on presenting all of the data at once, which leads to more complex visualizations.

NRB suggested that one way to reduce the complexity of the visualization is in fact to reduce the complexity of the data itself. Rather than visualizing absolute metabolite concentrations, NRB advocated visualizing instead the difference of the patient data from healthy patient data. Similarly, NRB suggested that discovery of new, interesting spectroscopy patterns is less interesting than identifying whether or not known patterns are present in the

data. Those patterns could then be presented to the user, for example as pre-defined filters for selection. NRB suggested that nDive's present level of complexity is more suited to research applications, for which data exploration is a more common goal.

In an interesting deviation from NRA's comments, NRB indicated that looking at ratios of metabolites is more useful than looking at absolute concentrations, especially in the SDDS visualization. This may be because NRB is most familiar with how to interpret ratios of metabolites, or it may be because ratios of metabolites are more useful than absolute concentrations. NRB discussed several interesting ways of interpreting the raw metabolite spectra which involve searching for particular visual patterns. This indicates that having direct access to the raw spectra for each voxel is also useful. NRB agreed with NRA that the SDDS visualization is overly complex, especially for radiologists that don't have a strong background in MRS research. Looking at smaller number of variables specifically designed to highlight features in the data rather than metabolites would be more useful.

NRB initially interpreted the blurriness of the scatter plot as an indication of deviation from a pre-computed linear regression. This was not an anticipated interpretation, however it indicates that the source of the blur in the density plots is not necessarily obvious initially.

The parallel coordinates plot was initially confusing, but after a demonstration of nDive being used for tumor segmentation NRB was able to understand how to interpret the lines. Selection of voxels and seeing their representation in an anatomical reference frame seemed to greatly improve NRB's understanding of the parallel coordinates plot. This was also true for other participants in the study. NRB suggested that learning the plots was relatively easy because of deep understanding of the data. This indicates that the ability to interpret plots of MRS data is related to how well the viewer understands the data set.

The ability to create new axes was interesting to NRB for the generation of a tumor classification variable. NRB suggested that rather than building up this classifier through visual analysis of the data plots, plugging in a formula from the MRS literature would be more useful.

Neuroradiologist C (NRC)

NRC had a markedly different opinion of nDive compared to NRA and NRB. The primary difference stemmed from NRC's caution concerning the use of MRS for diagnostic purposes. In NRC's opinion, MRS is not yet a fully mature modality and therefore an MRS visualization tool is less interesting from a clinical standpoint.

NRC's first comment on the 3D SDDS view was to indicate that while the view is not directly applicable for neuroradiology, it may be useful for tasks that require better spatial orientation. For example, biopsy planning requires picking a 3D orientation, and as such a 3D view of the data may be more useful. For diagnostic purposes, however, NRC said that SDDS is too cluttered and busy to be useful. NRC did not see SDDS in stereo, so adding stereo viewing may have improved NRC's impression.

Another concern was that the ability to select more than a few voxels at once was extraneous; NRC spends most of his time comparing one voxel to another, and did not foresee the need to compare larger voxel populations.

NRC immediately understood and liked the density plot technique for representing uncertainty, especially in the scatter plot, which came naturally. However, the parallel coordinates plot was difficult understand. NRC did indicate that the idea of looking at more that two variables in one plot for multivariate analysis would be useful, but the parallel coordinates was not compelling. On the other hand, NRC said that pre-loading useful ratios and relationships as selection filters would be a good way to help get the analysis started and learn how to interpret the plots.

NRC's final comment was that MRS as an imaging modality is not sufficiently useful to warrant learning a complex analysis tool like nDive. The primary concerns were resolution (MRS tends to have large voxels) and noise, which make MRS challenging to interpret. For NRC, taking the time to understand parallel coordinates plots will only be useful once these characteristics are improved.

6.5.2 Neurology

Neurology and neuroradiology are similar disciplines with similar training. The primary difference is that neurologists are the doctors that patients typically consult about head problems (headaches, strokes, etc.); they are responsible for communicating with the patient and ordering tests. For the neurologist, being able to read spectra is less important since they mostly interact with patients. In contrast, neuroradiologists focus on interpreting brain images and diagnosing problems. I interviewed a single neurologist for this study.

When looking at the 3D view, the neurologist made an interesting comment regarding viewing angles. When the view direction is looking straight down on (or orthogonal to) to the slice plane, the view looks similar to a 2D view. This may be a bit confounding, as the neurologist's expectation was that the spheres present would correspond only to data in that plane, rather than actually showing all of the data in front of the plane. The neurologist said that looking at the spheres only for the visible slice, similar to a data-driven spots visualization (Bokinsky, 2003), would be a useful way of simplifying the 3D visualization. This also highlights the importance of stereo vision, because the relative depth's of the spheres would be much more apparent if the view was shown in stereo. The neurologist pointed out that moving the view direction away from the slice plane orientation helps convey depth better. Once it was clear how the spheres should be interpreted, the neurologist also said that using the anatomical slice plane to clip away sets of the spheres is also a good way to simplify the visualization.

The neurologist commented multiples times that having a 3D representation of the tumor mass (for example, an isosurface) could improve the visualization. The suggestion was that it might improve sphere interpretation and that it would definitely make comparing voxel selections to the tumor mass simpler. Acquiring such surfaces is unfortunately non-trivial, but it may be useful as suggested.

The neurologist indicated that nDive's ability to report precise correlations might be more useful for neurosurgeons, who are responsible for defining tumor margins when planning resection or radiation treatments. The neurologist similarly said that the data point selection

techniques would be equally useful for neurosurgeons, as this gives them the ability to careful filter out the voxels they want. Finally, the neurologist said that neurosurgeons would also appreciate the ability to create new variables composed of multiple metabolites, especially if given the ability to define the functions themselves. The neurologist's function as a diagnostician does not require this level of specificity.

The neurologist immediately understood and appreciated the ability to perform selections in any of the data views and have the views update themselves. The ability to modify selections back and forth between anatomical space and the plots seemed to be a useful way to evaluate the quality of selections. The neurologist also pointed out that selections should always stand out from other data points, so special care should be taken when deciding how bright to make them.

According to the neurologist, nDive's major strength is that it is easier to understand the spectroscopy data than studying the raw spectra. This contrasts with comments made by the second neuroradiologist (NRB), who thought that looking at raw spectra was still a useful way of interpreting MRS data. This difference in opinion probably stems from the difference in their backgrounds and training. The neurologist has not had extensive training on how to interpret spectra and therefore finds them complex.

Additionally, the neurologist said that another major benefit to nDive is that it provides the new ability to identify quantitative relationships by eye. Visual patterns on the plots are not hard to select with the provided brushes and they ultimately provide information about the mathematical relationships they represent. Looking at the relative heights of peaks in the spectra for individual voxels is less quantitative and is extremely challenging when comparing multiple spectra to each other. Similarly, the neurologist also was pleased that nDive applied to MRS has the potential to help clinicians rely less on unreliable contrast agents and focus more on quantitative tumor measures.

6.5.3 Spectroscopy

The fifth participant has a significantly different background from the other participants, who are all physicians. Trained as a biochemist in spectroscopy, this participant is one of the few

people without an MD who is authorized to make medical decisions in a clinical setting. Of all the participants, the spectroscopist probably has the most detailed understanding of how to interpret MRS data, the limitations of the technique, and how the data itself is generated.

At the beginning of the interview, the spectroscopist immediately pointed out that, rather than looking at absolute concentrations in the SDDS view, the data would be more useful once it has been normalized relative to values in healthy brain areas. The spectroscopist indicated that normalizing to values within the same patient will be useful until there are a large number of age-matched control data sets available. The result for SDDS would be small, uniformly sized spheres in healthy regions in theory, with large spheres in unhealthy regions. The desire to see normalized values was requested repeatedly throughout the interview while discussing the other views of the data. One way to implement this in nDive would be to allow the user to create a selection using any of the provided selection techniques then label that selection as "normal" and create a new set of variables consisting of the original concentrations divided by the mean normal concentration. Care would have to be taken regarding the standard deviations, in this case.

In every interview, I presented a metabolite (glutamate) that had such large standard deviations for all voxels that it was not useful. This was done to motivate the decision to use density plots rather than discrete scatter plot and parallel coordinates plots. The spectroscopist clarified why glutamate is a challenging metabolite to study. Primarily this is because it is a large multiplet composed of multiple peaks and other metabolites with strong peaks in brain tissue (including NAA) tend to overlap with those peaks. Additionally, the curve-fitting routines make the assumption that the curves for individual metabolites are shaped as Gaussians, and multiplets violate this assumption. This means that LCModel's curve-fitting routines cannot confidently estimate any absolute glutamate concentrations.

The spectroscopist initially found the parallel coordinates plot confusing and difficult to interpret, even after multiple explanations. I selected a single set of voxels on the parallel coordinates plot that corresponded to tumor voxels, but interpreting the relationships for that set was still challenging. The spectroscopist made several comments on how this plot would not be useful to him in this state. However, once a second set of voxels was selected

(with the opposite ratio used for the first selection), the plot immediately made sense and the spectroscopist indicated that he liked the ability to compare two different populations.

6.6 Neurosurgery

Neurosurgeons are responsible for diagnosis and treatment of brain disorders, with an emphasis on performing surgeries. Spectroscopy is potentially useful for neurosurgeons in that they can use spectroscopic information both to diagnose disorders and to plan their surgeries. Biopsies can require neurosurgeons to plan needle paths through the brain that minimize damage to important functional tissue. The interviewed neurosurgeon had used spectroscopy for patients in the past, but did not have extensive MRS experience at the time of the interview.

Of all those interviewed, the neurosurgeon was the most emphatic in supporting nDive and MR spectroscopy as useful tools. The SDDS visualization, which was one of the more confusing visualizations for other interviewed physicians, made sense to the neurosurgeon immediately. The neurosurgeon also clearly supported nDive's user interaction techniques for selecting populations of voxels. For example, after seeing nDive's selection capabilities, the neurosurgeon asked if nDive could isolate voxels along the outer, invasive ring of a tumor for comparison with normal tissue. After performing such a selection and seeing the resulting scatter and parallel coordinates plots, the neurosurgeon was sufficiently interested in using nDive and spectroscopy that spectroscopy sequences were immediately ordered for two patients with upcoming surgeries.

After viewing a demonstration of the software, the neurosurgeon began to plan future spectroscopy research within the neurosurgery domain. While this is not necessarily a direct indicator of nDive's usefulness, it does show that nDive can be an effective way of explaining spectroscopy and its potential to experienced physician's without extensive training with MRS. The neurosurgeon planned to order spectroscopy sequences to correlate metabolite relationships discovered using nDive with surgical biopsy results.

As with the other interviewees, the neurosurgeon found the parallel coordinates plot to be confusing. Interestingly, the neurosurgeon announced a much stronger understanding

once the point-line duality between scatter plots and parallel coordinates plots was explained (see Section 4.1.4 for details). The neurosurgeon indicated that with practice the parallel coordinates plot's additional variables (compared to the scatter plot) would be potentially useful.

6.7 Discussion

The variety of responses in the interviews was extremely useful in planning future directions of nDive. The diversity and occasional conflict of the desired changes indicates that, regardless of any changes made, a single interface to nDive will not be optimal for all people who study MRS. For example, NRA was excited by the ability to look at absolute concentrations, but several other participants complained that absolute concentrations are confusing, perhaps because they are already trained to look at metabolite ratios.

Multiple neuroradiologists and the neurologist indicated that it would be useful to pre-compute several patterns known to be tumor and show them as selections immediately. This has two potential uses: first, it simplifies the workflow and enables potentially faster interpretation of the data; second, if the users already understand how to interpret the pre-computed patterns, this could facilitate learning to interpret the parallel coordinates plot, which multiple participants found difficult. All of the participants indicated that the parallel coordinates plots are fairly challenging to learn, and only some of them found the ability to see multiple variables at once compelling enough to warrant the effort. Pre-computed patterns would be a way to reduce that effort to an extent.

NRB mentioned an interesting and unforeseen potential use of nDive: education. Because the selection techniques make the data fairly simple to manipulate and the visualizations provide multiple interpretations, users seeking to learn about spectroscopy data are performing a more exploratory analysis, as opposed to clinical analysis. Clinical use of spectroscopy seems more aligned with hypothesis confirmation than exploratory hypothesis generation. Even so, the visualizations of nDive applied to different types of brain tumors in Figures 6.7 and 6.8 show that nDive might still be useful in a clinical setting for distinguishing among different

types of tumors.

The following are a list of paraphrased comments about the software or spectroscopy that were clearly stated and interesting:

Paraphrased comments from Neuroradiologist A:

- I would definitely use this. Being able to see absolute concentrations like this without having to look at squiggly lines is great.
- I can't do anything like this with what I have now.
- While the 3D view provides a nice general overview, I think that the 2D images are most helpful from a clinical standpoint.
- I didn't find the spheres to be particularly useful. They just didn't do anything for me or show me anything that the other viewing modalities didn't. They just looked a little too busy.

Paraphrased comments from Neuroradiologist B:

- You can start nDive with three different categories that are known to be fairly reliable. From how easy it is for you to select an area inside the tumor, I feel like that would be useful.
- One of the things that I don't like as much about the 3D spheres is that because you're showing absolute values, even within the tumor area the fact that choline is high isn't necessarily obvious because its spheres are so much smaller.
- You've seen the visualizations that we have with simple color maps of a ratio. One reason that they're so popular is that they give the "red is bad" impression, which is the level of simplicity we need for clinical use.

Paraphrased comments from Neuroradiologist C:

- I usually look at the worst part of what I see is a tumor or demylenating region and I compare it to the most normal part that I can find. I'm trying to decide if this is the most characteristic spectroscopy pattern and if it fits with the ratios that we know that we ought to be expecting.

- If it ever gets to the point that spectroscopy finds the things that we aren't seeing, then we're really going to need something like this, because we don't have the obvious enhancing ring mass with the adjacent edema to tip us off.

- It's hard to understand what you're talking about in the parallel coordinates plot. I like the fact that there's different patterns with different colors, but it's difficult to conceptualize in my mind what that actually means.

Paraphrased comments from the Neurologist:

- As a person who doesn't see this much, conceptually the 3D view is good except that you have to take a step out and understand that it's a 3D image so some of these squares don't necessarily correlate.

- I think this is very powerful because it very quickly allows you to conceptualize where the associations are. You can just select the red outliers where the ratio is higher and it looks like it's all in the middle of this mass.

- I think it's going to be exciting for someone like oncologists just to be able to manipulate this data and provide the radiologists clinical, anatomical, and pathological correlation. I think these people are going to find it very exciting to get a better idea of what their resection margins should be.

Paraphrased comments from the Spectroscopist:

- The parallel coordinates plot doesn't help me, to be honest, because it's very confusing.

When I look at the scatter plot I can see the data point itself without having to go to a graph like this.

- You need both populations selected – you can't just have one. Once you showed me that, the parallel coordinates plot is more interesting.
- Right now I like a number of features that I'm seeing here that will allow us to visualize the changes and map the tumor – not only the spectroscopy data, but any data you're representing.

Paraphrased comments from the Neurosurgeon:

- If you come up with a pattern for specific tumors, and it's reproducible, then you may be able to save everyone a biopsy.
- Where this is going to be helpful is when you have something too difficult to biopsy because it's in a dangerous location.
- I haven't seen anything like this before, to date.

These comments highlight the most common statements and requests made by the interviewed participants. They reveal a degree of ambivalence regarding the SDDS view and parallel coordinates plots due to their complexity. For SDDS, this may be remedied by focusing on problems that require more specific 3D position information, like biopsy planning. For the parallel coordinates plot, comments indicate that the plot is potentially useful with proper instruction. On the other hand, several participants indicated that the interaction techniques provided are compelling, especially in their ability to select and compare different populations of voxels.

CHAPTER 7

Conclusion

7.1 Results

This dissertation presented a visualization system for analyzing statistically uncertain multivariate 3D data. The system, called nDive, was designed to support radiologists studying MRS data to better classify biological tissue such as brain tumors. It has also been applied to biological simulation of multiple chemicals within a cell. The contributions of this work can summarized as follows:

- A multivariate 3D scalar visualization technique, called Sparse Data-Driven Spheres, that uses a sparse-glyph methodology to display variable values using small, colored spheres. SDDS enables viewers to identify relationships among variables for low resolution 3D data.

- Results of a user study showing that viewers are faster and more accurate at value estimation and correlation identification with SDDS as compared to superquadric glyphs (Feng et al., 2009).

- A density-based augmentation to scatter plots and parallel coordinates plots for visualizing uncertain multivariate data that preattentively highlights data points with low uncertainty and draws the viewer's attention away from data points with high uncertainty (Feng et al., 2010a; Feng et al., 2010b).

- Novel techniques for interacting with uncertain scatter plots and parallel coordinates plots. These techniques utilize data value uncertainty to help users interact predominantly with trustworthy data values by making uncertain data values more difficult to select (Feng et al., 2010b).

- Expert commentary and evaluation by users in a focus group study of nDive, the application that links these visualizations together. Experts said that nDive's visualizations and interaction techniques were useful, although complex, ways of distinguishing between different voxel populations.

7.2 Limitations

The SDDS visualization technique described in Chapter 2 produces images that are densely populated with colored spheres. SDDS visualizations can become difficult to interpret when the density of spheres is too high, which may make SDDS unsuitable for looking at data sets with large spatial extents. However, SDDS could be useful for looking at a smaller spatial subset of such data.

The uncertainty visualization techniques described in Chapter 3 use density plots to represent statistically uncertain data points. One drawback of these techniques is that it is not always possible to visually distinguish a dense cluster of trustworthy points from a similarly dense cluster of uncertain data points. Mean emphasis can help with this situation, in that certain points will be brighter than uncertain points. However, mean emphasis is only useful when over-plotting of individual data points is not a problem. In general, density-based plots can only guarantee the ability to see general trends, not all of the individual data points.

7.3 Future Work

The uncertainty visualization techniques described in Chapter 3 are theoretically generalizable to any type of statistically uncertain data, but they have only been demonstrated with normal distributions. While normal distributions are quite common, it would be useful to see if the

visualizations are still useful with other, more unusually shaped distributions.

Similarly, the uncertainty visualization interaction techniques described in Chapter 4 rely heavily on an efficient implementation of the `erf` operator, which is only useful for normal distributions. Integrating more complex distributions may require numerical estimation, which will be significantly slower and may make smooth interactivity hard to achieve. When this interactivity is important, it may be possible to approximate the data point distributions with normal distributions so that `erf` may be used for the purpose of selection only.

Experts said that nDive would be useful for research applications, but nDive's major limitation for clinical applications is its complexity. The experts generally indicated that the SDDS and parallel coordinates visualizations are too complex to be used in direct clinical settings. This is primarily because clinical diagnosis is not an exploratory process, but rather a hypothesis confirmation process. Rather than looking for new patterns in the data, clinicians would prefer that nDive clearly highlight known patterns for different disease processes over presenting all of the data at once.

The SDDS view is a sphere-based multivariate visualization technique. The sphere radius is currently chosen based on the resolution of the data, the number of variables in the data, the range of values in the data, and the data value at the sphere location. However, this is not necessarily the most perceptually useful choice. For example, viewers looking at circles tend to see the size of the object in terms of its area rather than its radius (Acevedo and Laidlaw, 2006). If this is also true for 3D spheres, then it may be that SDDS sphere scaling should be modified accordingly. Similarly it would be useful to know exactly what size and density of spheres is optimal for SDDS visualizations. A user study would be required to answer either of these questions.

The human visual system is able to preattentively distinguish different fields of coherent motion, so motion may be an interesting cue to add to the SDDS visualization. Each variable's set of spheres could have its own type of motion, for example different linear trajectories, and the spheres can be resized as they move. This may help visually distinguish the different variables from each other, although it does risk increasing the load on preattentive processing to the point of viewer distraction.

nDive was designed for radiologists studying MRS. Unfortunately, high quality statistically uncertain data is hard to acquire, so the nDive visualizations have only been explicitly evaluated in the context of MRS. Also, only SDDS has been evaluated with a quantitative user study. Quantitative evaluations of linked visualization systems is challenging, however it would be useful to quantitatively evaluate how well viewers can identify relationships and values in statistically uncertain multivariate data in uncertain parallel coordinates plots and scatter plots. However, the qualitative evaluations did reveal interesting future directions. There was near unanimous agreement that nDive should incorporate known metabolic patterns that define tumors and other processes for simplified voxel selection.

nDive has so far been applied to statistically uncertain data, but similar visualization techniques could be used for ensemble data sets. Ensembles are sets of numerical simulations in which the parameters of the simulation have been varied. The variance of a data point across all parameter settings can be considered a measure of uncertainty. The uncertain plotting techniques would be useful in this case for preattentively highlighting the data points with low variance. This may be useful for examining parameter sensitivity, as sensitive parameters will result in a wide variance in the data and therefore appear blurry in the plots. It may also be possible to visually distinguish populations of data points based on their parameter sensitivity.

APPENDIX A

Appendix: nDive User Manual

nDive is a software tool designed for use by radiologists that study Magnetic Resonance Imaging (MRI). It specializes in visualizing multivariate volumetric data, which is to say multiple 3D images located within the same space. nDive was originally designed to handle MR spectroscopy data. Whereas traditional T1/T2-weighted MRI produces anatomical images in which each voxel contains a single intensity value, MRSI produces full metabolite spectra at every volume element. Separate software is able to convert these spectra into multiple metabolite volume scalar fields. The resulting data is thus a series of multiple metabolite volume scalar fields and an anatomical volume scalar field.

The principal goal of this visualization software is to enable radiologists to explore the relationships between the different MR data fields with respect to the anatomical data. This type of goal lends itself well to 3D visualization. The primary technique used to visualize the metabolite fields is a sparse, similar glyph-based technique in which we distribute small spheres throughout the volume and scale the spheres according to the interpolated scalar value at each sphere's location in the volume. The spheres are nominally colored according to which metabolite field they represent. Finally, the anatomy is displayed using color on a slice plane representation that can be interactively moved in depth.

The secondary goal of this visualization software is to help clinicians and surgeons plan biopsies and surgeries. As a result, a quantitative visualization technique is required. We address this goal by resampling the spectroscopy data into the anatomical image space and displaying a single data field transparently on top of a grayscale anatomy slice plane. We color the metabolite slice plane using an isoluminant gray-to-red color scale in order to avoid confusion between the anatomy and metabolite data sets, and overlay yellow isovalue contours of the metabolite data to emphasize the metabolite structure. Mouse interaction enables

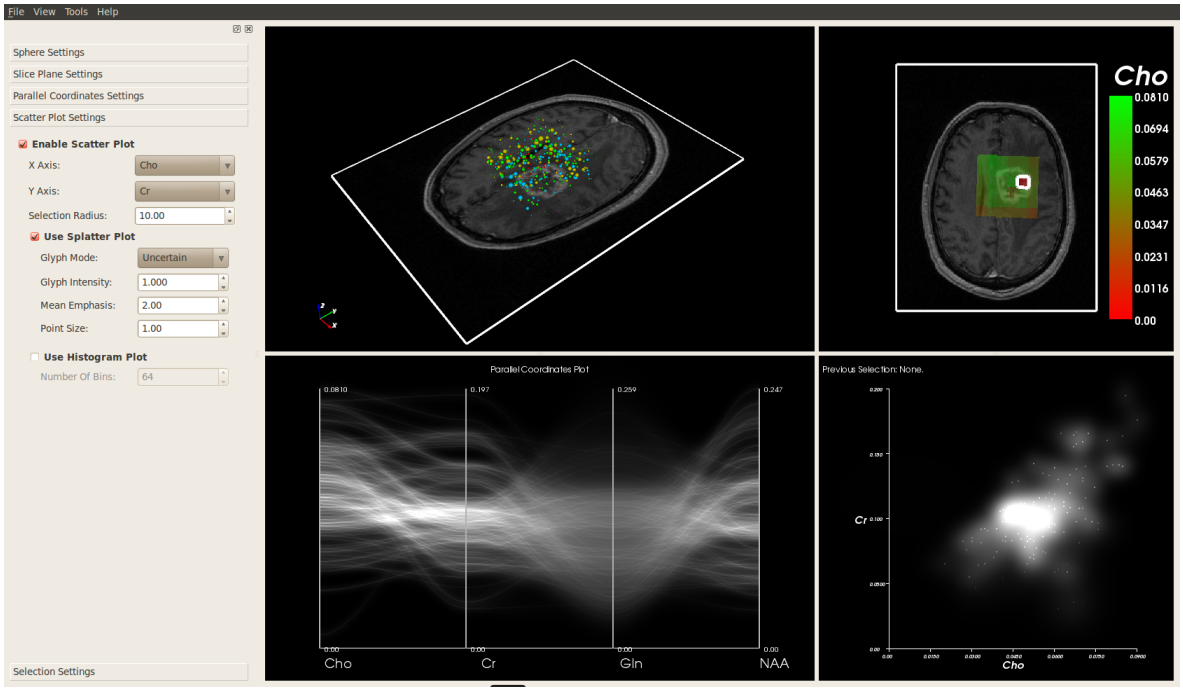


Figure A.1: A screenshot of the entire nDive system, including the settings panel.

viewers to see data values at particular locations on the plane.

The two spatial views described above are linked to a scatter plot and parallel coordinates plot of the data. These plots enable the direct exploration of data values absent their spatial coordinates. This is useful for looking for patterns in metabolite concentrations that define different types of tissues. Users can select voxels in the plots and subsequently the spatial locations of those voxels in the 3D and 2D plots.

A.1 Loading Data

nDive expects as input several different files in order to properly register the different incoming data components:

- Anatomical reference data: a directory filled with a set of 2D DICOM images.
- RDA spectroscopy description file: a single file containing registration information for spectroscopy data.
- Spectroscopy data files: either in MHA or DICOM format.

A.1.1 Handling Raw LCModel Data

nDive can convert Raw LCModel data into either the MHA or DICOM format for later use in the program. To do so, click `Tools | Convert LCModel Data`. A dialog box will appear which requires you to supply three important file locations: the directory containing the raw LCModel output, the RDA spectroscopy description file, and an output directory. nDive expects the LCModel directory to be organized as follows:

- The root directory must contain a set of directories labeled `slice<number>`, where each directory contains the information for the `<number>`th slice of data.
- Each slice directory must contain a file called `spreadsheet.csv`. This file contains the numerical concentrations and variances for every metabolite at every voxel in that slice.

When all of this data has been supplied, click OK to start the conversion. If the MHA file format was selected, a single MHA file for each metabolite concentration and each variance will be placed in the output directory. They will be labeled according to the column labels in `spreadsheet.csv` files of the LCModel output. MHA files consist of a single file containing either the concentrations or variances for a single metabolite. For example, if the choline concentration data is in Cho.mha, then the choline variance data will be in Cho%SD.mha. The %SD label is an important part of the filename, as nDive looks for that label when loading.

If the DICOM file format was selected, a single directory containing a series of slice files will be placed in the output directory for each metabolite concentration and variance. For each metabolite concentration or variance, a set of DICOM files are stored within their own directory.

A.1.2 Loading Spectroscopy Data

To load the spectroscopy data, click on `File | Load | Load Sphere Data`. A dialog will appear asking for the locations of different files containing the spectroscopy data. nDive can load the spectroscopy data converted using the conversion tool described in the previous

section by pressing the `Browse` button and selecting all of the files. When loading MHA files, variances will be automatically paired with their corresponding concentration files.

If DICOM was chosen as the output format for the LCModel Converted, be sure to check the `Load whole directories` box before browsing. In this case, selecting a concentration directory will automatically add the variance directory as well.

A.1.3 Loading Anatomical Data

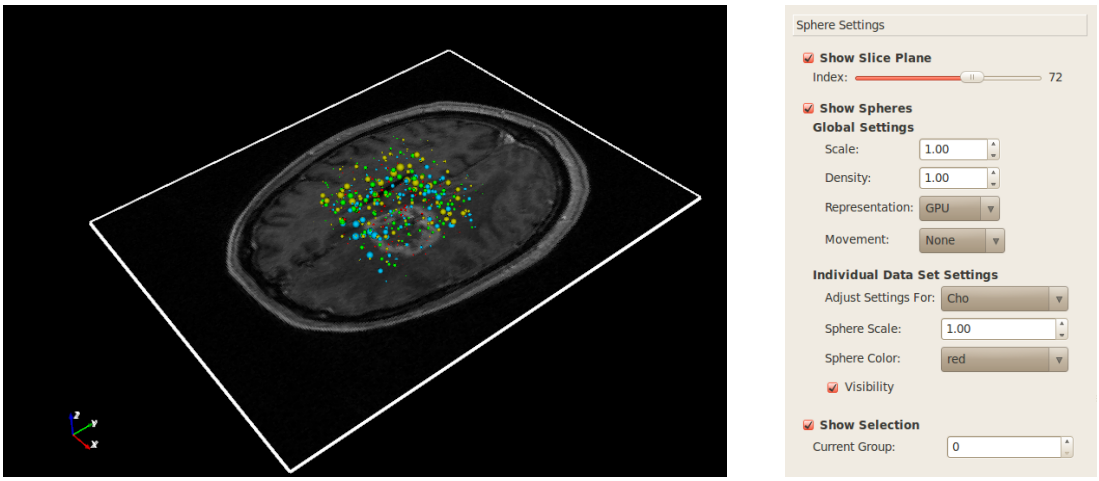
On the menu, click `File | Load | Load Slice Plane Data`. When loading MHA files (or any other single file format), simply click `Browse` button and select the desired file. When loading DICOM images, click the `Load whole directories` check box and select the desired directory containing only the desired DICOM images.

A.2 Visualization Modes

nDive has four display modes, a 3D visualization mode geared toward data exploration, a 2D visualization mode geared toward data analysis, and two plots for in depth data analysis. These modes each live in their own window frame, all of which are visible when the program begins. Each window has a corresponding settings panel on the left side of the window.

A.2.1 3D Visualization Window

The principal task for the 3D visualization window is to display as many data sets relative to the anatomical reference data set as possible. In this window there are two main components: an interactive anatomical slice plane that displays a single z-slice of the reference data and a collection of volume glyphs that create a 3D representation of all of the different overlapping data sets. Because this is a 3D visualization, you also have full control over the position and orientation of your view camera. An example view and settings panel for the 3D visualization window are shown in Figure A.2.



(a) 3D Visualization Window

(b) 3D Settings Panel

Figure A.2: The 3D visualization window containing an SDDS visualization and anatomical slice plane and the settings panel for adjusting the SDDS view.

Camera Interaction

All camera adjustments are made by clicking a mouse button and dragging. The camera actions associated with each mouse button are as follows:

- Left mouse button: rotate camera
- Middle mouse button: pan camera
- Right mouse button: zoom camera

Note that clicking on the data will have different results, as will be described later. To properly adjust the camera, be sure to click on empty space in the 3D visualization window.

Anatomy Slice Plane

This plane shows a single Z-slice of the loaded slice plane data. The anatomy plane can be moved up and down in Z by rotating the mouse wheel. Alternatively, it can also be moved by adjusting the slice index slider in the 3D settings panel.

Scaled Data-driven Spheres

In visualization, a glyph is a small symbol used to convey one or more data values to the viewer. nDive uses small spherical glyphs to convey the magnitude of the data values at different points in space. After the data is loaded, individual colors are assigned to each data set. Small colored spheres are then dispersed randomly within the data volume that are scaled according to the interpolated value of the data at each sphere's location. To summarize:

- Sphere color = which variable (metabolite)
- Sphere size = magnitude (concentration) of the variables at the sphere location

To adjust the properties of this window, click on the `Sphere Settings` panel of the settings frame. The settings are summarized here:

- Show Slice Plane: enable or disable slice plane visibility.

  - Index: slider that controls the slice plane Z index.

- Show Spheres: enable or disable the sphere visualization.

  - Global Scale: scale factor that affects all spheres.

  - Global Density: increase or decrease the density of spheres.

  - Representation: change how the spheres are rendered. GPU is recommended, but not functional on all graphics cards. If the spheres do not render properly under GPU mode, select either high resolution or low resolution for a polygonal representation.

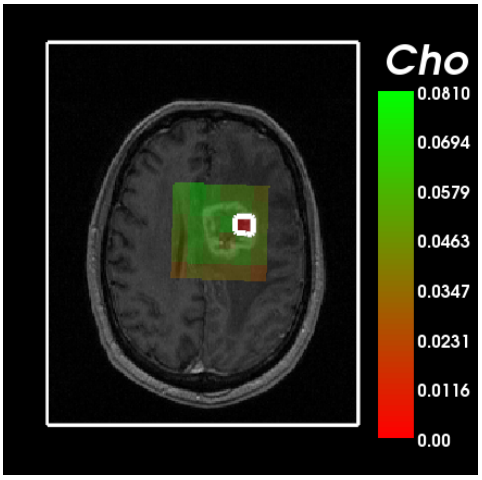
  - Movement: an experimental feature that causes glyphs to move linearly in different directions for each variable.

  - Adjust settings for: specify which variable the local glyph modifications will adjust.

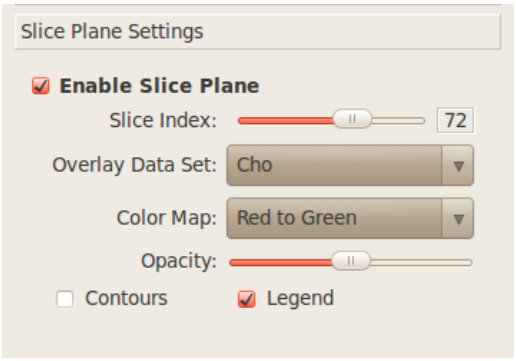
  - Local Sphere Scale: adjust the size of the spheres for a single variable.

  - Local Sphere Color: change the color of the spheres for a single variable.

  - Local Visibility: show or hide the spheres for a single variable.



(a) 2D Visualization Window



(b) 2D Settings Panel

Figure A.3: The 2D visualization window containing an anatomical slice place pseudocolored with a single metabolite field.

- Show Selection: show or hide the selected voxels.

  - Current Group: which selected voxel group is currently visible.

A.2.2 2D Visualization Window

This window contains a visualization meant to help users analyze and extract values from their data. It uses color to display a single data set overlaid on top of the grayscale anatomical slice plane. Both the settings panel and view for the 2D visualization window are shown in Figure A.3.

Anatomy Slice Plane

This slice plane is the same plane visible in the 3D visualization window. The only difference between the two is that they have separate window and level values. For interaction details, see the description in Section A.2.1.

Colored Data Overlay

The anatomy data is displayed on the slice plane in black and white because this high contrast color map does a good job of showing the detail of the higher resolution anatomical image.

We then overlay the currently selected data set, as selected in the Anatomy tab of the Settings window, using a gray-to-red color map that does not change in brightness. In this way, the two color maps do not conflict.

Isovalue Contours

In order to emphasize the structure of the overlaid data set, we also overlay yellow isovalue contours. These contours essentially are the lines that correspond to a particular data value within that data set.

The settings for this view are in the `Slice Plane Settings` panel:

- Enable Slice Plane: hide or show the slice plane.

  - Slice Index: move the slice plane up or down in Z. This is synchronized with the slice shown in the 3D visualization window.

  - Overlay Data Set: which variable to pseudocolor on top of the slice plane.

  - Color Map: which color map to use for pseudocoloring.

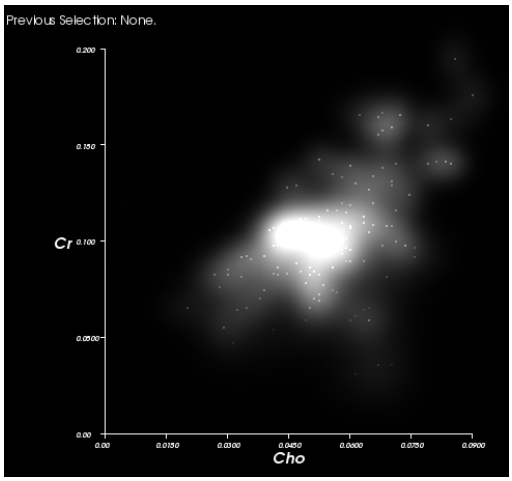
  - Opacity: the opacity of the pseudocoloring.

  - Contours: hide or show isovalue contours.

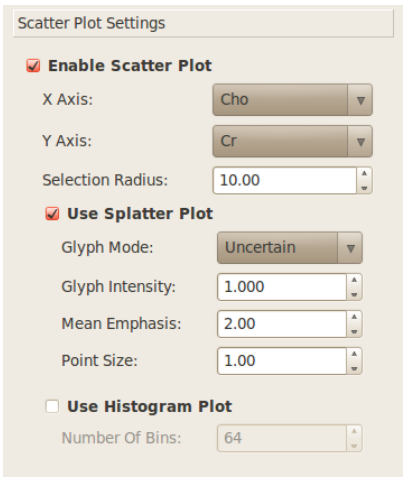
  - Legend: hide or show a legend for the pseudocoloring.

A.2.3 Scatter Plot Window

The scatter plot view plots voxels of the spectroscopy data into a two variable plot in which the concentration values for two metabolites are used a Cartesian grid coordinates. There are three ways to view the scatter plot: uncertainty mode, point mode, and histogram mode. In point mode, voxels are drawn as discrete points. In uncertainty mode, discrete points are replaced by bivariate normal distributions with standard deviations coming from the variance data. For histogram mode, a histogram of the data is drawn instead of the voxels themselves. The scatter plot frame and settings panel are shown in Figure A.4.



(a) Scatter Plot Window



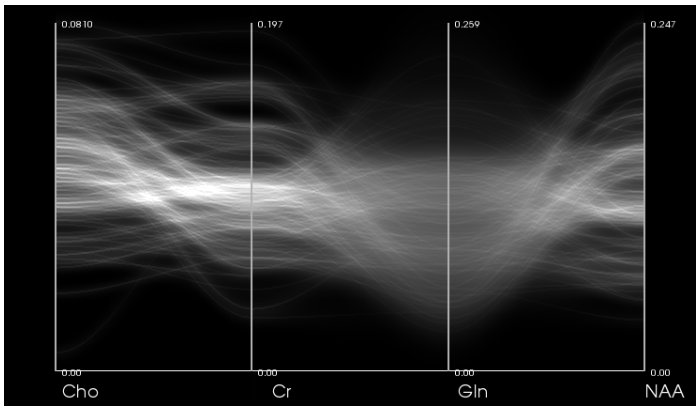
(b) Scatter Plot Settings Panel

Figure A.4: The scatter plot visualization window containing an uncertain scatter plot of two metabolites.

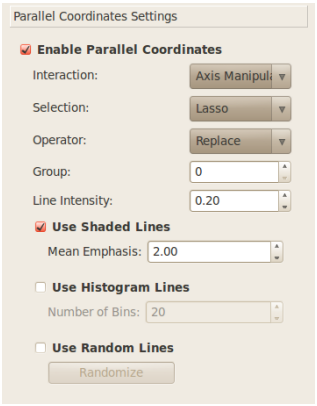
Voxel selection is possible in all three modes. The user can select all values near the mouse cursor by left-clicking on the plot. Similarly, values can be selected near a line by clicking and dragging out a line. These selections are represented as cubes in the 3D and 2D spatial views. Different groups of voxels can be selected and compared by changing the selection group in the settings panel. To deselect voxels, use the right mouse button instead.

The settings for the scatter plot view are in the `Scatter Plot Settings` panel of the settings frame:

- Enable Scatter Plot: Hide or show the scatter plot.
  - X Axis: select the variable to show on the X axis of the scatter plot.
  - Y Axis: select the variable to show on the Y axis of the scatter plot.
  - Selection Radius: increase or decrease the radius of the selection brush.
- Use Splatter Plot: toggle between the point/splat rendering mode and the histogram rendering mode
  - Glyph Mode: choose between drawing discrete points and uncertain splats.
  - Glyph Intensity: increase or decrease the intensity of an individual distribution.



(a) Parallel Coordinates Window



(b) Parallel Coordinates Settings Panel

Figure A.5: The parallel coordinates window containing a parallel coordinates plot of four metabolites.

- Mean Emphasis: how much to emphasize the mean of the distribution over the rest of the distribution.

- Point Size: the size of the points/means in the scatter plot.

- Use Histogram Plot: toggle between the point/splat rendering mode and the histogram rendering mode

  - Number of Bins: number of bins per axis for histogram computation.

A.2.4 Parallel Coordinates Window

Parallel coordinates plots can be thought of as a multivariate version of the scatter plot. Multiple variables axes are placed next to each other in parallel and lines passing through them represent voxels in the spectroscopy data. There are two categories of interactions with the parallel coordinates plot: axis manipulation and selection. The interaction mode can be chosen in the `Parallel Coordinates Settings` panel in the settings frame. The parallel coordinates frame and settings panel are shown in Figure A.5.

In axis manipulation, the left mouse button can be used to reposition and reorder axes via click-and-drag. The right mouse button zooms in on the plot. The middle mouse button pans across the plot

In selection mode, there are four different brushes to choose from: lasso, angle, function, and query. The lasso brush enables the user to select all of the parallel coordinates lines that pass through a curve drawn by the user using the left mouse button. The angle brush enables the user to select all lines that have the same angle as a user-drawn line (via left mouse button) between two axes. The function brush enables the user to draw two lines between the same two axes and select all parallel coordinates lines that match the linear function defined by those two lines. The query brush is the same as the linear brush except that it simply reports back to the user the function represented by the brush, rather than actually performing a selection. Different selection operators define how subsequent selections will affect the current selection.

The following settings are available in the `Parallel Coordinates Settings` panel in the settings frame:

- Enable Parallel Coordinates: hide or show the parallel coordinates plot.

  - Iteraction: choose between axis manipulation and selection modes.

  - Selection: choose the brush (lasso, angle, function, query).

  - Operator: choose how the next selection will affect the current selection (add, subtract, intersect, replace).

  - Group: choose the selection group (same as the scatter plot).

  - Line Intensity: how bright the lines will appear in the parallel coordinates plot.

- Use Shaded Lines: choose between shaded, uncertain lines, histogram lines, and random lines.

  - Mean Emphasis: how much brighter the mean of each distribution will be than the distribution itself.

- Use Histogram Lines: render the parallel coordinates plot in histogram mode.

  - Number of Bins: how many bins to use for histogram computation. In this mode, each bin is represented by a bar in the parallel coordinates plot whose brightness corresponds to how many lines fit in that bin.

- Use Random Lines: render the parallel coordinates plot in random line mode.

  - Randomize: generate a new set of random samples. This plot renders a representative sample of lines from the underlying data distributions rather than the distributions themselves.

A.3 Voxel Selection

For further analysis, nDive allows users to export values and statistics of selected subsets of the data volume. You can select particular voxels of the data by holding Control on the keyboard and left-clicking on the anatomy plane. A small box will appear where you clicked. To remove that box, hold Control and right click on the box. You can create multiple selection "groups" in the Selection Grid tab of the Settings window simply by changing the value in the current group spinner. The displayed boxes are colored differently depending on the current group.

The lower portion of the `Selection Settings` panel of the settings frame contains tables that show the selected values and average value and standard deviation for the current selection group for each data set.

A.3.1 Exporting Selected Voxels

You can export all of the data for selected voxels into a comma-separated value format (CSV) by clicking the `Export Data` button in the `Selection Settings` panel. You will be prompted to select a directory for output, and several files will subsequently be created. First, nDive will create statistics.csv, a file that organizes the averages and standard deviations of each selection group into a different row. nDive will also create a separate CSV file for each group containing the raw data values and error values (if applicable) for each selected voxel.

A.3.2 Normalization by Selection

The `Normalize` button in the selection panel will take the first (red) selection of voxels, and normalize all existing loaded variables by the mean of the selection. Any standard deviations

are unchanged in the new data sets, which are loaded immediately after clicking `Okay`.

A.4 Miscellaneous Tools

The `Tools` menu has several entries for various types of data processing. These include conversion of raw LCModel data into standard image formats, discussed in Section A.1.1, simple mathematical operations on the data, and others.

A.4.1 Threshold Classification

The `Tools | Threshold` menu option enables the user to compute a new data classification data set. The user selects a data set to threshold then supplies one or more value ranges. If a data value is within the range, the corresponding value in the output data set will contain a user-specified pass value. For example, to perform a binary classification, the user can supply two ranges: one above a certain value with a pass value of 1 and another below that value with a pass value of zero. When the user clicks the `Okay` button, they will be asked to supply a file name. Currently only MHA output is supported.

A.4.2 Load Default Data

The `Tools | Load Default Data` menu option is a short cut for loading a single data set. This data set can be specified by the user by editing `ndiveConfig.txt`, a file that should be located in the ndive install directory. The configuration file will have the following format:

```
root=[root directory]
defaultReference=[anatomical directory or file]
defaultSphere=[file or directory containing a metabolite data set]
defaultSphere=[file or directory containing a metabolite data set]
...
defaultError=[file or directory containing a metabolite SD data set]
defaultError=[file or directory containing a metabolite SD data set]
...
```

The root directory is the directory from which nDive will be executed. It should be the parent directory that contains `Shaders` installed with nDive. The metabolite and standard deviation data sets must be listed in the same order.

A.4.3 Compute Ratio

The `Tools | Compute Ratio` menu option lets the user compute the ratio of two loaded data sets and save the result to an MHA file. If either of the two data sets has associated percent standard deviations, the output will contain the larger standard deviation of the two on a per-voxel basis.

A.4.4 Combine Variables

The `Tools | Combine Variables` is an experimental feature for compute new variables in place, without writing them out to a file. The user specifies two data sets and an operator to apply to them, for example `Cho + Cr`. After clicking `Okay`, the new variable is computed and automatically loaded for comparison to other variables.

A.5 Compilation

The full list of required libraries to build the nDive C++ source code are as follows:

- Qt 4.6.x
- VTK 5.8
- GDCM 2.0.12
- ITK 3.18

All of these libraries should be compiled from source using CMake as the project file generator. Once complete, nDive can also be compiled using CMake. Instructions for how to use CMake are beyond the scope of this appendix, but documentation is available online.

A.5.1 Qt

Qt is a C++ graphical user interface (GUI) library that integrates well with VTK. Build Qt 4.6.2 from the LGPL source code. Along with the source libraries, the build will also produce an application called qmake. The location of this executable will be used in subsequent projects.

A.5.2 VTK

The Visualization Toolkit is a C++ visualization library for generating various types of interactive visualizations. It serves as the visualization backbone for nDive. As of the publication of this document (7/26/2010), nDive is compiled against the latest Git version of VTK. nDive should also work with VTK release 5.8. When building the project file, ensure that the following CMake flags are set properly:

- BUILD_SHARED_LIBS = ON
- VTK_USE_QT = ON
- QT_QMAKE_EXECUTABLE = (path to qmake)

A.5.3 GDCM

The Grassroots DICOM toolkit is a C++ library for importing DICOM images. DICOM (Digital Imaging and Communications in Medicine) is one of the most common standards for storing medical image data (and metadata). GDCM is used by nDive solely for image loading. Ensure that the following CMake flag is set properly:

- GDCM_BUILD_SHARED_LIBS = ON
- GDCM_USE_VTK = OFF

A.5.4 ITK

The Insight Toolkit is a templated C++ library for image analysis. ITK serves as the bridge for transferring data between GDCM and VTK. nDive currently builds from ITK release 3.18.

Be sure that the following CMake flags are set properly:

- BUILD_SHARED_LIBS = ON
- ITK_USE_REVIEW = ON
- ITK_USE_SYSTEM_GDCM = ON
- GDCM_DIR = (GDCM project build directory)

A.5.5 nDive

Once all of the prerequisites have been compiled, nDive itself can be compiled using CMake as the project generator. Be sure that the following CMake flags are set properly:

- VTK_DIR = (VTK project build directory)
- ITK_DIR = (ITK project build directory)
- QT_QMAKE_EXECUTABLE = (path to qmake)

Bibliography

Acevedo, D. and Laidlaw, D. (2006). Subjective quantification of perceptual interactions among some 2d scientific visualization methods. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):1133–1140. 41, 131

Akiba, H. and Ma, K.-L. (2007). A tri-space visualization interface for analyzing time-varying multivariate volume data. In *EuroVis07 - Eurographics / IEEE VGTC Symposium on Visualization*, pages 115–122. 71, 79

Ankerst, M., Berchtold, S., and Keim, D. (1998). Similarity clustering of dimensions for an enhanced visualization of multidimensional data. pages 52–60, 153. 72

Bachthaler, S. and Weiskopf, D. (2008). Continuous scatterplots. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1428–1435. 47, 48

Blaas, J., Botha, C., and Post, F. (2008). Extensions of parallel coordinates for interactive exploration of large multi-timepoint data sets. *Visualization and Computer Graphics, IEEE Transactions on*, 14(6):1436–1451. 50

Blaas, J., Botha, C. P., and Post, F. H. (2007). Interactive visualization of multi-field medical data using linked physical and feature-space views. In Museth, K., Mller, T., and Ynnerman, A., editors, *EuroVis*, pages 123–130. Eurographics Association. 72

Bokinsky, A. (2003). *Multivariate Data Visualization with Data-Driven Spots*. PhD thesis, UNC - Chapel Hill. 20, 21, 23, 121

Box, G. E. P. and Muller, M. E. (1958). A note on the generation of random normal deviates. *The Annals of Mathematical Statistics*, 29(2):610–611. 66

Broersen, A. and van Liere, R. (2005). Transfer functions for imaging spectroscopy data using principal component analysis. In *EuroVis*, pages 117–123. Eurographics Association. 16

Cai, W. and Sakas, G. (1999). Data intermixing and multi-volume rendering. *Computer Graphics Forum*, 18:359–368. 14

Castillo, M. (2002). *Neuroradiology*. Lippincott Williams & Jenkins. 97

Chang, J., Thakur, S., Perera, G., Kowalski, A., Huang, W., Karimi, S., Hunt, M., Koutcher, J., Leibel, S., Amols, H., and Narayana, A. (2004). Image-fusion of mr spectroscopic images for treatment planning of gliomas. *International Journal of Radiation Oncology*Biology*Physics*, 60(Supplement 1):S223 – S224. 106

Cleveland, W. C. and McGill, M. E. (1988). *Dynamic Graphics for Statistics*. CRC Press, Inc., Boca Raton, FL, USA. 46

Crouzil, A., Massip-Pailhes, L., and Castan, S. (1996). A new correlation criterion based on gradient fields similarity. In *Proceedings of the 13th International Conference on Pattern Recognition*, volume 1, pages 632–636. 17

d'Ocagne, M. (1885). *Coordonnees parallels et axiale*. Gauthier-Villars. 48

Doleisch, H., Gasser, M., and Hauser, H. (2003). Interactive feature specification for focus+context visualization of complex simulation data. In *VISSYM '03: Proceedings of the symposium on Data visualisation 2003*, pages 239–248, Aire-la-Ville, Switzerland, Switzerland. Eurographics Association. 72

Ebert, D. S., Rohrer, R. M., Shaw, C. D., Panda, P., Kukla, J. M., and Roberts, D. A. (2000). Procedural shape generation for multi-dimensional data visualization. *Computers & Graphics*, 24(3):375–384. 19

Elmqvist, N., Dragicevic, P., and Fekete, J.-D. (2008). Rolling the dice: Multidimensional visual exploration using scatterplot matrix navigation. *IEEE Transactions on Visualization and Computer Graphics*, 14(6):1141–1148. 47, 72

Feng, D., Kwock, L., Lee, Y., and Taylor, II, R. M. (2010a). Linked exploratory visualizations for uncertain MR spectroscopy data. volume 7530, pages 753004–1–753004–12. SPIE. 9, 42, 70, 89, 105, 129

Feng, D., Kwock, L., Lee, Y. Z., and Taylor, II, R. M. (2010b). Matching visual saliency to confidence in plots of uncertain data. *IEEE Transactions on Visualization and Computer Graphics (Proceedings Visualization / Information Visualization 2010)*, 16(6):sss–eee. 9, 42, 70, 105, 129, 130

Feng, D., Lee, Y., Kwock, L., and Taylor, II, R. M. (2009). Evaluation of glyph-based multivariate scalar volume visualization techniques. In *APGV '09: Proceedings of the 6th Symposium on Applied Perception in Graphics and Visualization*, pages 61–68, New York, NY, USA. ACM. 9, 10, 105, 129

Fisher, P. (1993). Visualizing uncertainty in soil maps by animation. *Cartographica: The International Journal for Geographic Information and Geovisualization*, 30:20–27. 54

Forsell, C., Seipel, S., and Lind, M. (2005). Simple 3d glyphs for spatial multivariate data. In *IEEE Symposium on Information Visualization*, pages 119–124. 19

Fua, Y.-H., Ward, M. O., and Rundensteiner, E. A. (1999). Hierarchical parallel coordinates for exploration of large datasets. In *VIS '99: Proceedings of the conference on Visualization '99*, pages 43–50, Los Alamitos, CA, USA. IEEE Computer Society Press. 48, 49

Graham, M. and Kennedy, J. (2003). Using curves to enhance parallel coordinate visualisations. *Proceedings of the 7th International Conference on Information Visualization 2003*, pages 10–16. 62

Gresh, D. L., Rogowitz, B. E., Winslow, R. L., Scollan, D. F., and Yung, C. K. (2000). Weave: a system for visually linking 3-d and statistical visualizations, applied to cardiac simulation and measurement data. In *VIS '00: Proceedings of the conference on Visualization '00*, pages 489–492, Los Alamitos, CA, USA. IEEE Computer Society Press. 72

Grigoryan, G. and Rheingans, P. (2004). Point-based probabilistic surfaces to show surface uncertainty. *IEEE Transactions on Visualization and Computer Graphics*, 10(5):564–573. 51, 53

Hauser, H., Ledermann, F., and Doleisch, H. (2002). Angular brushing of extended parallel coordinates. In *Proceedings of IEEE Symposium on Information Visualization*, pages 127–130. IEEE Computer Society Press. 72, 77

Healey, C. H. (1996). Choosing effective colours for data visualization. In *Proceedings of IEEE Visualization '96*, pages 263–270. 23

Heinrich, J. and Weiskopf, D. (2009). Continuous parallel coordinates. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1531–1538. 50, 61

Hinkley, D. V. (1969). On the ratio of two correlated normal random variables. *Biometrika*, 56:645–639. 87

Inselberg, A. (1985). The plane with parallel coordinates. *The Visual Computer*, 1(4):69–91. 48

Interrante, V. (1997). Illustrating surface shape in volume data via principal direction-driven 3D line integral convolution. In *SIGGRAPH '97: Proceedings of the 24th annual conference on Computer Graphics and Interactive Techniques*, pages 109–116, New York, NY, USA. ACM Press/Addison-Wesley Publishing Co. 12

Interrante, V., Fuchs, H., and Pizer, S. (1997). Conveying the 3d shape of smoothly curving transparent surfaces via texture. In *IEEE Transactions on Visualization and Computer Graphics*, volume 3, pages 98–117. 12, 21

Johansson, S. and Johansson, J. (2009). Scattering points in parallel coordinates. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1001–1008. 62

Johnson, C. (2004). Top scientific visualization research problems. *IEEE Comput. Graph. Appl.*, 24(4):13–17. 3

Kindlmann, G. and Weinstein, D. (1999). Hue-balls and lit-tensors for direct volume rendering of diffusion tensor fields. In *Proceedings of IEEE Visualization '99*, pages 183–189, Los Alamitos, CA, USA. IEEE Computer Society Press. 14

Kindlmann, G. and Westin, C.-F. (2006). Diffusion tensor visualization with glyph packing. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):1329–1335. 18, 19, 23

Klose, U. (2008). Measurement sequences for single voxel proton mr spectroscopy. *European Journal of Radiology*, 67(2):194 – 201. Clinical 1H MR Spectroscopy. 101

Kniss, J., Kindlmann, G., and Hansen, C. (2001). Interactive volume rendering using multi-dimensional transfer functions and direct manipulation widgets. In *VIS '01: Proceedings of the conference on Visualization '01*, pages 255–262, Washington, DC, USA. IEEE Computer Society. 14, 15

Kosara, R., Miksch, S., and Hauser, H. (2001). Semantic depth of field. In *INFOVIS '01: Proceedings of the IEEE Symposium on Information Visualization 2001 (INFOVIS'01)*, page 97, Washington, DC, USA. IEEE Computer Society. 55

Lau, C., Ng, L., Thompson, C., Pathak, S., Kuan, L., Jones, A., and Hawrylycz, M. (2008). Exploration and visualization of gene expression with neuroanatomy in the adult mouse brain. *BMC Bioinformatics*, 9:153. 20

Levoy, M. (1988). Display of surfaces from volume data. *IEEE Computer Graphics and Applications*, 8(3):29–37. 13

Lorensen, W. E. and Cline, H. E. (1987). Marching cubes: A high resolution 3d surface construction algorithm. In *SIGGRAPH '87: Proceedings of the 14th Annual Conference on Computer Graphics and Interactive Techniques*, volume 21, pages 163–169, New York, NY, USA. ACM Press. 11

Martin, A. R. and Ward, M. O. (1995). High dimensional brushing for interactive exploration of multivariate data. In *VIS '95: Proceedings of the 6th conference on Visualization '95*, page 271, Washington, DC, USA. IEEE Computer Society. 72

Maudsley, A. A., Darkazanli, A., Alger, J. R., Hall, L. O., Schuff, N., Studholme, C., Yu, Y., Ebel, A., Frew, A., Goldgof, D., Gu, Y., Pagare, R., Rousseau, F., Sivasankaran, K., Soher, B. J., Weber, P., Young, K., and Zhu, X. (2006). Comprehensive processing, display and analysis for in vivo MR spectroscopic imaging. *NMR in Biomedicine*, 19:492–503. 106

Meyer-Spradow, J., Stegger, L., Döring, C., Ropinski, T., and Hinrichs, K. H. (2008). Glyph based spect visualization for the diagnosis of coronary artery disease. *IEEE Transactions on Visualization and Computer Graphics (TVCG) (Vis Conference Issue)*, pages 1499–1506. 19

Miller, J. J. and Wegman, E. J. (1991). *Construction of line densities for parallel coordinate plots*, pages 107–123. Springer-Verlag New York, Inc., New York, NY, USA. 50, 61

Moustafa, R. and Wegman, E. (2006). *Multivariate Continuous Data - Parallel Coordinates*, pages 143–155. Statistics and Computing. Springer New York. 62

Muigg, P., Kehrer, J., Oeltze, S., Piringer, H., Doleisch, H., Preim, B., and Hauser, H. (2008). A four-level focus+context approach to interactive visual analysis of temporal features in large scientific data. *Computer Graphics Forum*, 27(3):775–782. 50

Nattkemper, T. W. (2004). Multivariate image analysis in biomedicine. *J. of Biomedical Informatics*, 37(5):380–391. 16

Novotny, M. and Hauser, H. (2006). Outlier-preserving focus+context visualization in parallel coordinates. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):893–900. 50

Olston, C. and Mackinlay, J. (2002). Visualizing data with bounded uncertainty. In *Proceedings of the IEEE Symposium on Information Visualization*, pages 37–40, Boston, Massachusetts. 53

Palmer, S. (1999). *Vision Science: Photons to Phenomenology.* MIT Press. 10, 25, 54

Pang, A. T., Wittenbrink, C. M., and Lodh, S. K. (1996). Approaches to uncertainty visualization. *The Visual Computer*, 13:370–390. 51, 52

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3):1065–1076. 55

Peng, W., Ward, M., and Rundensteiner, E. (2004). Clutter reduction in multi-dimensional data visualization using dimension reordering. pages 89–96. 72

Potter, K. (2006). Methods for presenting statistical information: The box plot. *Hans Hagen, Andreas Kerren, and Peter Dannenmann (Eds.), Visualization of Large and Unstructured Data Sets, GI-Edition Lecture Notes in Informatics (LNI)*, S-4:97–106. 53

Potter, K., Krueger, J., and Johnson, C. (2008). Towards the visualization of multidimentional stochastic distribution data. In *Proceedings of The International Conference on Computer Graphics and Visualization (IADIS) 2008.* 53

Provencher, S. (1993). Estimation of metabolite concentrations from localized in vivo proton NMR spectra. *Magn Reson Med*, 30:672–679. 102

Provencher, S. W. (1982). Contin: A general purpose constrained regularization program for inverting noisy linear algebraic and integral equations. *Computer Physics Communications*, 27(3):229 – 242. 103

Rheingans, P. (1992). Color, change, and control for quantitative data display. In *Proceedings of IEEE Visualization '92*, pages 252–259, Los Alamitos, CA, USA. IEEE Computer Society. 14

Rheingans, P. (1996). Opacity-modulating triangular textures for irregular surfaces. In Yagel, R. and Nielson, G. M., editors, *IEEE Visualization '96*, pages 219–226. 13

Robert, C. P. and Casella, G. (2005). *Monte Carlo Statistical Methods.* Springer-Verlag New York, Inc., Secaucus, NJ, USA. 66

Rösler, F., Tejada, E., Fangmeier, T., Ertl, T., and Knauff, M. (2006). GPU-based multivolume rendering for the visualization of functional brain images. In *Proceedings of SimVis 2006*, pages 305–318. 14

Sanyal, J., Zhang, S., Bhattacharya, G., Amburn, P., and Moorhead, R. (2009). A user study to compare four uncertainty visualization methods for 1d and 2d datasets. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1209–1218. 53

Sauber, N., Theisel, H., and Seidel, H.-P. (2006). Multifield-graphs: An approach to visualizing correlations in multifield scalar data. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):917–924. 17

Seo, J. and Shneiderman, B. (2004). A rank-by-feature framework for unsupervised multidimensional data exploration using low dimensional projections. In *INFOVIS '04: Proceedings of the IEEE Symposium on Information Visualization*, pages 65–72, Washington, DC, USA. IEEE Computer Society. 47, 72

Shearer, J., Ogawa, M., Ma, K.-L., and Kohlenberg, T. (2008). Pixelplexing: Gaining display resolution through time. In *IEEE Pacific Visualization Symposium 2008.*, pages 159 –166. 53

Soares, D. P. and Law, M. (2009). Magnetic resonance spectroscopy of the brain: review of metabolites and clinical applications. *Clin Radiol*, 64:12–21. 102

Stompel, A., Lum, E., and Ma, K.-L. (2002). Feature-enhanced visualization of multidimensional, multivariate volume data using non-photorealistic rendering techniques. In *Proceedings of Pacific Graphics 2002*, pages 1–8. IEEE. 14

Swayne, D. F., Temple Lang, D., Buja, A., and Cook, D. (2003). GGobi: evolving from XGobi into an extensible framework for interactive data visualization. *Computational Statistics & Data Analysis*, 43:423–444. 71

Taylor, II, R. M. (2002). Visualizing multiple fields on the same surface. *Computer Graphics and Applications, IEEE*, 22(3):6–10. 21

Taylor, II, R. M. (2004). Directly rendering non-polygonal objects on graphics hardware using vertex and fragment programs. Technical Report TR04-023, University of North Carolina at Chapel Hill. 24

Thomson, J., Hetzler, E., MacEachren, A., Gahegan, M., and Pavel, M. (2005). A typology for visualizing uncertainty. volume 5669, pages 146–157. SPIE. 51

Tukey, J., Fisherkeller, M., and Friedman, J. (1988). *PRIM9: An interacive multidimensional data display and analysis system.* Wadsworth Inc. 72

Uttecht, S. and Thulborn, K. R. (2002). Software for efficient visualization and analysis of multiple, large, multi-dimensional data sets from magnetic resonance imaging. *Computerized Medical Imaging and Graphics*, 26:73–89. 106

Ward, M. O. (1994). Xmdvtool: integrating multiple methods for visualizing multivariate data. In *VIS '94: Proceedings of the conference on Visualization '94*, pages 326–333, Los Alamitos, CA, USA. IEEE Computer Society Press. 71

Ware, C. (1988). Color sequences for univariate maps: Theory, experiments and principles. *IEEE Comput. Graph. Appl.*, 8(5):41–49. 51

Ware, C. (2000). *Information visualization: perception for design.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. 21, 23, 38, 41, 55

Weigle, C. and Taylor, II, R. M. (2005). Visualizing intersecting surfaces with nested-surface techniques. In *IEEE Visualization*, page 64. 12

Wittenbrink, C. M., Pang, A., and Lodha, S. K. (1996). Glyphs for visualizing uncertainty in vector fields. *IEEE Trans. Vis. Comput. Graph.*, 2(3):266–279. 51

Woodring, J. and Shen, H.-W. (2006). Multi-variate, time varying, and comparative visualization with contextual cues. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):909–916. 17, 18