Statistical Variability in Nonlinear Spaces: Application to Shape Analysis and DT-MRI

by
P. Thomas Fletcher

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
2004

Approved by:

Stephen M. Pizer, Advisor

Sarang Joshi, Advisor

Guido Gerig, Reader

J. S. Marron, Reader

Michael Kerckhove, Reader

TABLE OF CONTENTS

1 Introduction **x**

1.1 Motivation 1

1.1.1 Shape Analysis 3

1.1.2 Diffusion Tensor Imaging 4

1.2 Thesis and Claims 4

1.3 Overview of Chapters 6

2 Mathematical Background **7**

2.1 Topology 8

2.1.1 Basics 8

2.1.2 Metric spaces 9

2.1.3 Continuity 10

2.1.4 Various Topological Properties 10

2.2 Differentiable Manifolds 11

2.2.1 Topological Manifolds 11

2.2.2 Differentiable Structures on Manifolds 12

2.2.3 Tangent Spaces 14

2.3 Riemannian Geometry 15

2.3.1 Riemannian Metrics 15

2.3.2 Geodesics 16

2.4 Lie Groups 20

2.4.1 Lie Group Exponential and Log Maps 22

2.4.2 Bi-invariant Metrics 24

2.5 Symmetric Spaces 25

2.5.1 Lie Group Actions 26

2.5.2 Symmetric Spaces as Lie Group Quotients 27

3 Image Analysis Background **29**
3.1 Statistical Shape Theory . . . 29
3.1.1 Point Set Shape Spaces . . . 30
3.1.2 Procrustes Distance and Alignment . . . 32
3.1.3 Shape Variability . . . 34
3.1.4 Nonlinear Statistical Analysis . . . 38
3.2 Deformable Models . . . 39
3.2.1 Active Contours . . . 39
3.2.2 Probabilistic Deformable Models . . . 41
3.3 Medial Representations . . . 43
3.3.1 The Medial Locus . . . 44
3.3.2 M-reps . . . 48
3.4 Diffusion Tensor Imaging . . . 57

4 Manifold Statistics **61**
4.1 Means on Manifolds . . . 61
4.1.1 Intrinsic vs. Extrinsic Means . . . 62
4.1.2 Computing the Intrinsic Mean . . . 63
4.2 Principal Geodesic Analysis . . . 64
4.2.1 Variance . . . 65
4.2.2 Geodesic Submanifolds . . . 66
4.2.3 Projection . . . 67
4.2.4 Defining Principal Geodesic Analysis . . . 67
4.2.5 An Alternative Definition of PGA . . . 68
4.2.6 Approximating Principal Geodesic Analysis . . . 69
4.3 Conclusions . . . 71

5 Statistics of M-reps **73**
5.1 M-reps as Elements of a Symmetric Space . . . 73
5.1.1 The Exponential and Log Maps for M-reps . . . 75
5.1.2 The Hippocampus Data Set . . . 76
5.2 M-rep Alignment . . . 77
5.3 M-rep Averages . . . 79
5.4 M-rep PGA . . . 81
5.5 PGA in Deformable M-reps Segmentation . . . 83
5.5.1 Principal Geodesic Deformations . . . 84

5.5.2 PGA-Based Geometric Prior 84
5.6 Conclusions 86

6 Statistics of Diffusion Tensors **88**
6.1 The Space of Diffusion Tensors 90
6.2 The Geometry of $PD(n)$ 91
6.2.1 The Lie Group Action on $PD(n)$ 92
6.2.2 The Invariant Metric on $PD(n)$ 93
6.2.3 Computing Geodesics 94
6.3 Statistics of Diffusion Tensors 96
6.3.1 Averages of Diffusion Tensors 96
6.3.2 Principal Geodesic Analysis of Diffusion Tensors 97
6.4 Properties of PGA on $PD(n)$ 98
6.5 New Methods: Comparison Metric, Interpolation, and Anisotropy 100
6.5.1 Comparison Metric 101
6.5.2 Diffusion Tensor Interpolation 101
6.5.3 Geodesic Anisotropy Measure 105
6.6 Conclusions 107

7 Discussion and Future Work **109**
7.1 Summary of Contributions 109
7.2 Future Work 113
7.2.1 Theoretical Questions 113
7.2.2 M-rep Extensions 115
7.2.3 Future Diffusion Tensor Work 116
7.2.4 Other Application Areas 117

BIBLIOGRAPHY **119**

ABSTRACT
P. THOMAS FLETCHER: Statistical Variability in Nonlinear Spaces: Application to Shape Analysis and DT-MRI.
(Under the direction of Stephen M. Pizer and Sarang Joshi.)

Statistical descriptions of anatomical geometry play an important role in many medical image analysis applications. For instance, geometry statistics are useful in understanding the structural changes in anatomy that are caused by growth and disease. Classical statistical techniques can be applied to study geometric data that are elements of a linear space. However, the geometric entities relevant to medical image analysis are often elements of a nonlinear manifold, in which case linear multivariate statistics are not applicable. This dissertation presents a new technique called principal geodesic analysis for describing the variability of data in nonlinear spaces. Principal geodesic analysis is a generalization of a classical technique in linear statistics called principal component analysis, which is a method for computing an efficient parameterization of the variability of linear data. A key feature of principal geodesic analysis is that it is based solely on intrinsic properties, such as the notion of distance, of the underlying data space.

The principal geodesic analysis framework is applied to two driving problems in this dissertation: (1) statistical shape analysis using medial representations of geometry, which is applied within an image segmentation framework via posterior optimization of deformable medial models, and (2) statistical analysis of diffusion tensor data intended as a tool for studying white matter fiber connection structures within the brain imaged by magnetic resonance diffusion tensor imaging. It is shown that both medial representations and diffusion tensor data are best parameterized as Riemannian symmetric spaces, which are a class of nonlinear manifolds that are particularly well-suited for principal geodesic analysis. While the applications presented in this dissertation are in the field of medical image analysis, the methods and theory should be widely applicable to many scientific fields, including robotics, computer vision, and molecular biology.

Chapter 1

Introduction

1.1 Motivation

Advances in medical imaging technology have provided the ability to acquire high-resolution 3D images of the human body. Imaging technologies such as CT and MR are non-invasive means for obtaining potentially life-saving information. The goal of medical image analysis is to maximize the potential benefit of this data, expanding its use beyond simple visualization of the raw data. Image analysis techniques provide more advanced visualizations and aid in disease diagnosis, radiotherapy treatment, surgery planning, and tracking of anatomic growth. For example, automatic extraction of anatomical geometry from a medical image is useful in planning radiation beam therapy to apply maximum radiation dose to a tumor while minimizing the exposure to surrounding organs. Image analysis techniques have shown promise in diagnosing brain disorders such as schizophrenia by distinguishing healthy brain structures from those with disease. Analysis of diffusion tensor magnetic resonance images of neonatal brains can give information about the early stages of development in brain connectivity.

These examples benefit from a particular tool from medical image analysis known as statistical shape analysis, which describes the geometric variability of anatomy. A probability distribution of the possible geometric configurations of an organ can be used as prior information to help guide the process of automatic extraction of anatomy geometry from medical images. Probability distributions of normal organ shape and of diseased organ shape can be used to assign a probability that a patient has a particular disease based on the shape of their anatomy. Statistics of diffusion tensor image data can be used to explore what common changes occur in the connectivity of the brain during development.

Previous approaches to these problems have used linear models of anatomic shape, and thus, linear statistical techniques to analyze the shape variability. Most models of shape currently in use are based on linear representations of the boundary that can only undergo linear variations in shape. However, richer models of shape and richer variations of shape can be achieved with nonlinear models. For example, medial representations of geometry, or m-reps, have shown promise in representing the interior of anatomic structures and describing shape changes in intuitive terms such as local thickening, bending, and twisting. The parameters of m-rep models are inherently nonlinear. Also, previous statistical models of diffusion tensor data have been linear models. As shown in this dissertation, diffusion tensors are more naturally modeled as elements of a nonlinear space. The drawback of nonlinear models is that classical linear statistical methods cannot be applied.

This dissertation presents a new technique called *principal geodesic analysis* for describing the variability of data in nonlinear spaces. Principal geodesic analysis is a generalization of a classical technique in linear statistics called principal component analysis, which is a method for computing an efficient parameterization of the variability of linear data. Principal component analysis also allows the dimensionality of the data to be reduced to only the true variables of change. This dissertation extends these concepts to nonlinear spaces known as manifolds. The driving problems in this dissertation are two: (1) statistical shape analysis using medial representations of geometry, which is applied within an image segmentation framework via posterior optimization of deformable medial models, and (2) statistical analysis of diffusion tensor data intended as a tool for studying white matter fiber connection structures within the brain imaged by magnetic resonance diffusion tensor imaging.

While the applications presented in this dissertation are in the field of medical image analysis, the methods and theory should be widely applicable to many scientific fields, including mechanical engineering, robotics, computer vision, and molecular biology. Many common geometric entities are elements of nonlinear spaces. These include transformations such as rotations, scalings, and affine transformations, and primitives such as lines, planes, and unit vectors. The statistical methods developed in this work can be applied to all of these spaces. These other possible applications are mentioned in the future work section in Chapter 7.

The remainder of this section continues the motivation for the two driving applications of this work: shape analysis and diffusion tensor imaging.

1.1.1 Shape Analysis

Shape analysis concerns the statistical study of the geometry of objects that is invariant to position, size, and orientation. An important aspect of shape theory is in studying the geometric variability of objects. Anatomical shape analysis plays an important role in several medical image analysis applications.

One motivation for statistical shape analysis is its use in segmentation of anatomical structures in medical images. Segmentation is the process of distinguishing important structures in an image from background. This is a fundamental task in medical image analysis that is often a prerequisite for further analysis, visualization, disease diagnosis, or planning of medical treatment. Knowledge of the geometric variability of the anatomy can be used as prior information to help guide the segmentation process. This geometric prior helps overcome difficulties inherent to segmentation, such as image noise, sampling artifacts, and low contrast.

Statistical shape analysis may also be useful in educational atlases of anatomy. Current anatomical atlases only present a single instance of the normal anatomy. A statistical shape atlas can present a full range of geometric variabilities that occur in normal anatomy.

Another application of statistical shape analysis is its potential to serve as a tool in understanding and diagnosing disease. For instance, brain disorders such as Alzheimer's and schizophrenia are often accompanied by structural changes in the brain. Understanding the changes in organ shape that occur could be fundamental in furthering our understanding of such diseases. Also, shape analysis can help in diagnosing disease by detecting differences in the shape of an organ affected by disease.

The goal of this work is to provide new methods for analyzing nonlinear shape variability. The medial representation of object geometry provides a powerful framework for describing shape variability in intuitive terms such as local thickness, bending, and widening. However, the medial parameters are not elements of a Euclidean space. Therefore, the standard linear techniques of shape analysis, namely linear averaging and principal component analysis, do not apply. This work describes how the medial parameters are in fact elements of a certain type of manifold known as a Riemannian symmetric space. The theory of principal geodesic analysis developed in this dissertation is applied to study the variability of medial representations of object shape. This dissertation develops a segmentation strategy for 3D medical images using m-rep models with a geometric prior based on principal geodesic analysis.

1.1.2 Diffusion Tensor Imaging

Diffusion tensor magnetic resonance imaging (DT-MRI) is an imaging technique that is used to obtain information about fiber structures such as the white matter fiber in the brain or the fiber structure of muscles. It produces at every voxel in an imaging volume a diffusion tensor, which is a model of the diffusivity properties of water in that voxel. In fibrous structures, such as the white matter fiber in the brain, water tends to diffuse more in the direction parallel to the fibers. This gives the ability to determine the local direction of white matter fibers from diffusion tensor images. Furthermore, algorithms for tracking these local fiber directions leads to global information about the connectivity of various regions in the brain. In addition to connectivity information, the diffusivity information has shown promise in understanding certain brain disorders. Studies have shown that pathologies such as multiple sclerosis and stroke can affect the diffusivity properties of the brain matter.

Statistical analysis of diffusion tensor images has the potential to further our understanding of the connectivity properties of the brain and the effects of disease on the brain fiber structure. Probability distributions on diffusion tensors could be used to generate statistical atlases of diffusion tensor data, describing the normal variation in brain connectivity across individuals. Statistical methods on diffusion tensor data might be used to quantify the variability of the brain matter that is caused by disease and to study the normal variability in the geometry of white matter fiber bundles. In addition, more fundamental processing of diffusion tensor data, such as smoothing, could benefit from statistics.

Previous approaches to statistical analysis of diffusion tensor data have used linear statistical models. However, such methods can assign nonzero probability to "illegal" instances of diffusion tensors, i.e., tensors that do not model any possible diffusion of water. In this dissertation it is shown that diffusion tensors are more naturally modeled by a nonlinear space. Principal geodesic analysis is applied to the space of diffusion tensors to parameterize the statistical variability of diffusion tensor data. It is shown that, unlike linear techniques, principal geodesic analysis preserves the important properties of diffusion tensors, including producing only legal models of water diffusion.

1.2 Thesis and Claims

Thesis: *Principal geodesic analysis is a natural generalization of principal component analysis for describing the statistical variability of geometric data that are parameter-*

ized as curved manifolds. Such manifolds include medial representations of shape and diffusion tensors. Principal geodesic analysis can be used to parameterize the shape variability of a population of m-rep models. The resulting probabilities can be used effectively as a statistical geometric prior in a deformable m-rep model segmentation of 3D medical images.

The contributions of this dissertation are

1. A novel theory called principal geodesic analysis has been developed as a natural generalization of principal component analysis for describing the statistical variability of geometric data that are parameterized as curved manifolds. This generalization is natural in the sense that it uses only intrinsic distances and geodesics in the data space.

2. It has been shown that medial representations of shape, or m-reps, can be formulated as elements of a Riemannian symmetric space and that the variability of a population of m-rep objects can be efficiently computed using principal geodesic analysis.

3. A new method for aligning m-reps to a common position, orientation and scale has been developed and demonstrated. It generalizes the Procrustes alignment method for aligning linear representations of shape. It proceeds by minimizing the sum-of-square geodesic distances between corresponding atoms in medial models.

4. A method for maximum posterior segmentation of 3D medical images via deformable m-reps models using principal geodesic analysis has been developed. The optimization of the objective function in the segmentation uses the principal geodesic modes of variation as a parameter space. A geometric prior based on principal geodesic analysis has been developed and incorporated into a Bayesian objective function.

5. It has been shown that diffusion tensors can be treated as data in a Riemannian symmetric space and that the variability of diffusion tensor data can be described using principal geodesic analysis.

6. New methods for interpolating diffusion tensors, comparing the similarity of diffusion tensor images, and measuring the anisotropy of diffusion tensors have been developed using the symmetric space formulation of the space of diffusion tensors.

1.3 Overview of Chapters

The remainder of this dissertation is organized in the following chapters:

Chapter 2 provides an overview of the required mathematics used in this dissertation. This includes differential geometry concepts such as Riemannian manifolds, Lie groups and symmetric spaces.

Chapter 3 presents the background topics in medical image analysis that are related to the applications treated in this dissertation. The topics include deformable models, shape analysis, m-rep models and segmentation, and diffusion tensor imaging.

Chapter 4 presents the main theoretical contribution of this work, principal geodesic analysis, which is a method for describing the statistical variability of data on a curved manifold. A discussion of existing methods for computing averages on manifolds is also included.

Chapter 5 applies the statistical methods from Chapter 4 to the space of 3D m-rep models. This provides a method for describing the statistical variability of m-rep models of anatomic shape. A generalization of the Procrustes alignment method is given for m-rep models. Principal geodesic analysis is applied to a collection of m-rep models of hippocampi, demonstrating the average and modes of variation. A geometric prior using principal geodesic analysis is developed for deformable m-rep model segmentation.

Chapter 6 applies the statistical methods from Chapter 4 to the space of diffusion tensors. Principal geodesic analysis is applied to synthetic diffusion tensor data to show that the average and modes of variation produce legal instances of diffusion tensors, while naive linear statistical analysis does not. A natural method for interpolating diffusion tensors and describing their anisotropy is also discussed.

Chapter 7 concludes with a discussion of the contributions of this dissertation and possible future work.

Chapter 2

Mathematical Background

The geometric entities studied in this thesis, namely m-rep shape models and diffusion tensors, are elements of high-dimensional, curved manifolds. More precisely, they are Riemannian symmetric spaces. It is useful to think of a point in a symmetric space as a transformation from a fixed base point. For example, when constructing the space of diffusion tensors, the base point is chosen to be the identity matrix, and any diffusion tensor is treated as a transformation from the identity. The transformation spaces that are being used are known as Lie groups, which are smooth manifolds themselves. It is useful to study these Lie group transformations of symmetric spaces because they tend to be algebraic in nature, and, therefore, certain computations on symmetric spaces, such as distances and shortest paths between two points, often have closed-form solutions. The same computations can require solving differential equations if the manifold in question is not a symmetric space. Since distances and shortest paths will be essential in the definitions of statistics for manifolds, symmetric spaces are particularly nice spaces for doing statistics.

Many geometric entities are representable as Lie groups or symmetric spaces. Transformations of Euclidean spaces such as translations, rotations, scalings, and affine transformations all arise as elements of Lie groups. Geometric primitives such as unit vectors, oriented planes, and symmetric, positive-definite matrices can be seen as points in symmetric spaces. This chapter is a review of the basic mathematical theory of Lie groups and symmetric spaces. The study of such spaces first requires some background in basic topology and manifold theory, which is provided in the first three sections. The reader already familiar with these topics may skip the appropriate sections.

The various spaces that are described throughout this chapter are all generalizations, in one way or the other, of Euclidean space, $\mathbb{R}^n$. Euclidean space is a topological space, a Riemannian manifold, a Lie group, and a symmetric space. Therefore, each section

will use $\mathbb{R}^n$ as a motivating example. Also, since the study of geometric transformations is stressed, the reader is encouraged to keep in mind that $\mathbb{R}^n$ can also be thought of as a transformation space, that is, as the set of translations on $\mathbb{R}^n$ itself.

2.1 Topology

The study of a topological spaces arose from the desire to generalize the notion of continuity on Euclidean spaces to more general spaces. Topology is a fundamental building block for the theory of manifolds and function spaces. This section is a review of the basic concepts needed for the study of differentiable manifolds. For a more thorough introduction see [91]. For several examples of topological spaces, along with a concise reference for definitions, see [118].

2.1.1 Basics

Remember that continuity of a function on the real line is phrased in terms of open intervals, i.e., the usual ϵ-δ definition. A topology defines which subsets of a set X are "open", much in the same way an interval is open. As will be seen at the end of this subsection, open sets in $\mathbb{R}^n$ are made up of unions of open balls of the form $B(x,r) = \{y \in \mathbb{R}^n : \|x - y\| < r\}$. For a general set X this concept of open sets can be formalized by the following set of axioms.

Definition 2.1. A **topology** on a set X is a collection $\mathcal{T}$ of subsets of X such that
(1) $\varnothing$ and X are in $\mathcal{T}$.
(2) The union of an arbitrary collection of elements of $\mathcal{T}$ is in $\mathcal{T}$.
(3) The intersection of a finite collection of elements of $\mathcal{T}$ is in $\mathcal{T}$.

The pair $(X, \mathcal{T})$ is called a **topological space**. However, it is a standard abuse of notation to leave out the topology $\mathcal{T}$ and simply refer to the topological space X. Elements of $\mathcal{T}$ are called **open sets**. A set $C \subset X$ is a **closed set** if it's complement, $X - C$, is open. Unlike doors, a set can be both open and closed, and there can be sets that are neither open nor closed. Notice that the sets $\varnothing$ and X are both open and closed.

Example 2.1. Any set X can be given a topology consisting of only $\varnothing$ and X being open sets. This topology is called the **trivial topology** on X. Another simple topology is the **discrete topology** on X, where any subset of X is an open set.

Definition 2.2. A **basis** for a topology on a set X is a collection $\mathcal{B}$ of subsets of X such that

(1) For each $x \in X$ there exists a $B \in \mathcal{B}$ containing x.

(2) If $B_1, B_2 \in \mathcal{B}$ and $x \in B_1 \cap B_2$, then there exists a $B_3 \subset B_1 \cap B_2$ such that $x \in B_3$.

The basis $\mathcal{B}$ generates a topology $\mathcal{T}$ by defining a set $U \subset X$ to be open if for each $x \in U$ there exists a basis element $B \in \mathcal{B}$ with $x \in B \subset O$. The reader can check that this does indeed define a topology. Also, the reader should check that the generated topology $\mathcal{T}$ consists of all unions of elements of $\mathcal{B}$.

Example 2.2. The motivating example of a topological space is Euclidean space $\mathbb{R}^n$. It is typically given the standard topological structure generated by the basis of open balls $B(x, r) = \{y \in \mathbb{R}^n : \|x - y\| < r\}$ for all $x \in \mathbb{R}^n, r \in \mathbb{R}$. Therefore, a set in $\mathbb{R}^n$ is open if and only if it is the union of a collection of open balls. Examples of closed sets in $\mathbb{R}^n$ include sets of discrete points, vector subspaces, and closed balls, i.e., sets of the form $\bar{B}(x, r) = \{y \in \mathbb{R}^n : \|x - y\| \leq r\}$.

2.1.2 Metric spaces

Notice that the topology on $\mathbb{R}^n$ is defined entirely by the Euclidean distance between points. This method for defining a topology can be generalized to any space where a distance is defined.

Definition 2.3. A **metric space** is a set X with a function $d : X \times X \to \mathbb{R}$ that satisfies

(1) $d(x, y) \geq 0$, and $d(x, y) = 0$ if and only if $x = y$.

(2) $d(x, y) = d(y, x)$.

(3) $d(x, y) + d(y, z) \geq d(x, z)$.

The function d above is called a **metric** or **distance function**. Using the distance function of a metric space, a basis for a topology on X can be defined as the collection of open balls $B(x, r) = \{y \in X : d(x, y) < r\}$ for all $x \in X, r \in \mathbb{R}$. From now on when a metric space is discussed, it is assumed that it is given this topology.

One special property of metric spaces will be important in the review of manifold theory.

Definition 2.4. A metric d on a set X is called **complete** if every Cauchy sequence converges in X. A Cauchy sequence is a sequence $x_1, x_2, \ldots \in X$ such that for any $\epsilon > 0$ there exists an integer N such that $d(x_i, x_j) < \epsilon$ for all $i, j > N$.

2.1.3 Continuity

As was mentioned at the beginning of this section, topology developed from the desire to generalize the notion of continuity of mappings of Euclidean spaces. That generalization is phrased as follows:

Definition 2.5. Let X and Y be topological spaces. A mapping $f : X \to Y$ is **continuous** if for each open set $U \subset Y$, the set $f^{-1}(U)$ is open in X.

It is easy to check that for a mapping $f : \mathbb{R}^d \to \mathbb{R}^n$ the above definition is equivalent to the standard ϵ-δ definition.

Definition 2.6. Again let X and Y be topological spaces. A mapping $f : X \to Y$ is a **homeomorphism** if it is bijective and both f and f^{-1} are continuous. In this case X and Y are said to be **homeomorphic**.

When X and Y are homeomorphic, there is a bijective correspondence between both the points and the open sets of X and Y. Therefore, as topological spaces, X and Y are indistinguishable. This means that any property or theorem that holds for the space X that is based only on the topology of X also holds for Y.

2.1.4 Various Topological Properties

This section is a discussion of some special properties that a topological space may possess. The particular properties that are of interest are the ones that are important for the study of manifolds.

Definition 2.7. A topological space X is said to be **Hausdorff** if for any two distinct points $x, y \in X$ there exist disjoint open sets U and V with $x \in U$ and $y \in V$.

Notice that any metric space is a Hausdorff space. Given any two distinct points x, y in a metric space X, we have $d(x, y) > 0$. Then the two open balls $B(x, r)$ and $B(y, r)$, where $r = \frac{1}{2}d(x, y)$, are disjoint open sets containing x and y, respectively. However, not all topological spaces are Hausdorff. For example, take any set X with more than one point and give it the trivial topology, i.e., $\varnothing$ and X as the only open sets.

Definition 2.8. Let X be a topological space. A collection $\mathcal{O}$ of open subsets of X is said to be an **open cover** if $X = \bigcup_{U \in \mathcal{O}} U$. A topological space X is said to be **compact** if for any open cover $\mathcal{O}$ of X there exists a finite subcollection of sets from $\mathcal{O}$ that covers X.

The Heine-Borel theorem (see [106], Theorem 2.41) gives intuitive criteria for a subset of $\mathbb{R}^n$ to be compact. It states that any closed and bounded subset of $\mathbb{R}^n$ is compact. Thus, for example, a closed ball $\bar{B}(x,r)$ is compact as is the unit sphere $S^{n-1} = \{x \in \mathbb{R}^n : \|x\| = 1\}$. The sphere, like Euclidean space, will be an important example throughout this chapter. This is partly because it is a simple example of a symmetric space, but also because it is an integral part of the medial representation used later.

Definition 2.9. A **separation** of a topological space X is a pair of disjoint open sets U, V such that $X = U \cup V$. If no separation of X exists, it is said to be **connected**.

2.2 Differentiable Manifolds

Differentiable manifolds are spaces that locally behave like Euclidean space. Much in the same way that topological spaces are natural for talking about continuity, differentiable manifolds are a natural setting for calculus. Notions such as differentiation, integration, vector fields, and differential equations make sense on differentiable manifolds. This section gives a review of the basic formulations that will be needed later. A good introduction to the subject may be found in [15]. For a comprehensive overview of differential geometry see [111,112,113,114,115]. Other good references include [2,87,58].

2.2.1 Topological Manifolds

A manifold is a topological space that is locally equivalent to Euclidean space. More precisely,

Definition 2.10. A **manifold** is a Hausdorff space M with a countable basis such that for each point $p \in M$ there is a neighborhood U of p that is homeomorphic to $\mathbb{R}^n$ for some integer n.

At each point $p \in M$ the dimension n of the $\mathbb{R}^n$ in Definition 2.10 is unique. If the integer n is the same for every point in M, then M is called a ***n*-dimensional** manifold. The simplest example of a manifold is $\mathbb{R}^n$, since it is trivially homeomorphic to itself. Likewise, any open set of $\mathbb{R}^n$ is also a manifold.

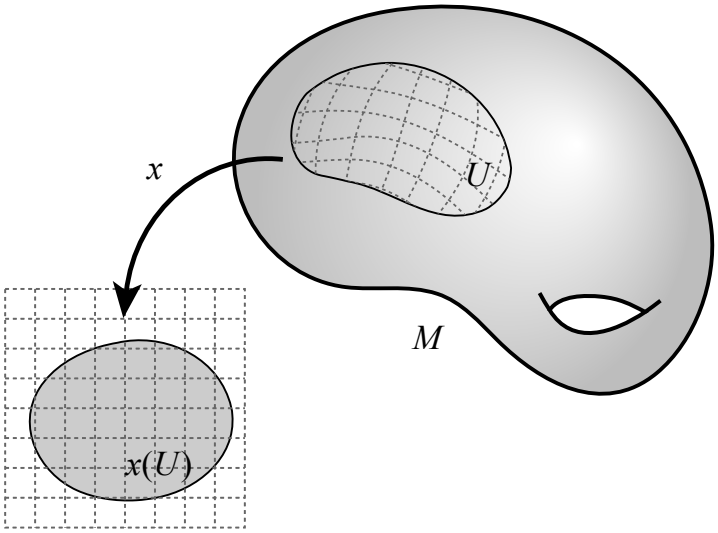

Figure 2.1: A local coordinate system (x, U) on a manifold M.

2.2.2 Differentiable Structures on Manifolds

The next step in the development of the theory of manifolds is to define a notion of differentiation of manifold mappings. Differentiation of mappings in Euclidean space is defined as a local property. Although a manifold is locally homeomorphic to Euclidean space, more structure is required to make differentiation possible. First, recall that a function on Euclidean space $f : \mathbb{R}^n \to \mathbb{R}$ is **smooth** or $\boldsymbol{C^\infty}$ if all of its partial derivatives exist. A mapping of Euclidean spaces $f : \mathbb{R}^m \to \mathbb{R}^n$ can be thought of as a n-tuple of real-valued functions on $\mathbb{R}^m$, $f = (f^1, \ldots, f^n)$, and f is smooth if each f^i is smooth.

Given two neighborhoods U, V in a manifold M, two homeomorphisms $x : U \to \mathbb{R}^n$ and $y : V \to \mathbb{R}^n$ are said to be $\boldsymbol{C^\infty}$**-related** if the mapping $x \circ y^{-1} : y(U \cap V) \to x(U \cap V)$ is C^∞. The pair (x, U) is called a **chart** or **coordinate system**, and can be thought of as assigning a set of coordinates to points in the neighborhood U (see Figure 2.1). That is, any point $p \in U$ is assigned the coordinates $x^1(p), \ldots, x^n(p)$. As will become apparent later, coordinate charts are important for writing local expressions for derivatives, tangent vectors, and Riemannian metrics on a manifold. A collection of charts whose domains cover M is called an **atlas**.

Definition 2.11. An atlas $\mathcal{A}$ on a manifold M is said to be **maximal** if for any any other atlas $\mathcal{A}'$ on M any coordinate chart $(x, U) \in \mathcal{A}'$ is also a member of $\mathcal{A}$.

Definition 2.12. A **smooth structure** on a manifold M is a maximal atlas $\mathcal{A}$ on M.

The manifold M along with such an atlas is termed a **smooth manifold**.

The next theorem demonstrates that it is not necessary to define every coordinate chart in a maximal atlas, but rather, one can define enough compatible coordinate charts to cover the manifold.

Theorem 2.1. *Given a manifold M with an atlas $\mathcal{A}$, there is a unique maximal atlas $\mathcal{A}'$ such that $\mathcal{A} \subset \mathcal{A}'$.*

Example 2.3. Consider the sphere S^2 as a subset of $\mathbb{R}^3$. The upper hemisphere $U = \{(x, y, z) \in S^2 : z > 0\}$ is an open neighborhood in S^2. Now consider the homeomorphism $\phi : S^2 \to \mathbb{R}^2$ given by

$$\phi : (x, y, z) \mapsto (x, y).$$

This gives a coordinate chart (ϕ, U). Similar charts can be produced for the lower hemisphere, and for hemispheres in the x and y dimensions. The reader may check that these charts are C^∞-related and cover S^2. Therefore, these charts make up an atlas on S^2 and by Theorem 2.1 there is a unique maximal atlas containing these charts that makes S^2 a smooth manifold. A similar argument can be used to show that the n-dimensional sphere, S^n, for any $n \geq 1$ is also a smooth manifold.

Now consider a function $f : M \to \mathbb{R}$ on the smooth manifold M. This function is said to be a **smooth function** if for every coordinate chart (x, U) on M the function $f \circ x^{-1} : U \to \mathbb{R}$ is smooth. More generally, a mapping $f : M \to N$ of smooth manifolds is said to be a **smooth mapping** if for each coordinate chart (x, U) on M and each coordinate chart (y, V) on N the mapping $y \circ f \circ x^{-1} : x(U) \to y(V)$ is a smooth mapping. Notice that the mapping of manifolds was converted locally to a mapping of Euclidean spaces, where differentiability is easily defined.

As in the case of topological spaces, there is a desire to know when two smooth manifolds are equivalent. This should mean that they are homeomorphic as topological spaces and also that they have equivalent smooth structures. This notion of equivalence is given by

Definition 2.13. Given two smooth manifolds M, N, a bijective mapping $f : M \to N$ is called a **diffeomorphism** if both f and f^{-1} are smooth mappings.

2.2.3 Tangent Spaces

Given a manifold $M \subset \mathbb{R}^d$, it is possible to associate a linear subspace of $\mathbb{R}^d$ to each point $p \in M$ called the **tangent space** at p. This space is denoted T_pM and is intuitively thought of as the linear subspace that best approximates M in a neighborhood of the point p. Vectors in this space are called **tangent vectors** at p.

Tangent vectors can be thought of as directional derivatives. Consider a smooth curve $\gamma : (-\epsilon, \epsilon) \to M$ with $\gamma(0) = p$. Then given any smooth function[1] $f : M \to \mathbb{R}$, the composition $f \circ \gamma$ is a smooth function, and the following derivative exists:

$$\frac{d}{dt}(f \circ \gamma)(0).$$

This leads to an equivalence relation $\sim$ between smooth curves passing through p. Namely, if γ_1 and γ_2 are two smooth curves passing through the point p at $t = 0$, then $\gamma_1 \sim \gamma_2$ if

$$\frac{d}{dt}(f \circ \gamma_1)(0) = \frac{d}{dt}(f \circ \gamma_2)(0),$$

for any smooth function $f : M \to \mathbb{R}$. A tangent vector is now defined as one of these equivalence classes of curves. It can be shown (see [2]) that these equivalence classes form a vector space, i.e., the tangent space T_pM, which has the same dimension as M. Given a local coordinate system (x, U) containing p, a basis for the tangent space T_pM is given by the partial derivative operators $\partial/\partial x^i$, which are the tangent vectors associated with the coordinate curves of x.

Example 2.4. Again, consider the sphere S^2 as a subset of $\mathbb{R}^3$. The tangent space at a point $p \in S^2$ is the set of all vectors in $\mathbb{R}^3$ perpendicular to p, i.e., $T_pS^2 = \{v \in \mathbb{R}^3 : \langle v, p\rangle = 0\}$. This is of course a two-dimensional vector space, and it is the space of all tangent vectors at the point p for smooth curves lying on the sphere and passing through the point p.

A **vector field** on a manifold M is a function that smoothly assigns to each point $p \in M$ a tangent vector $X_p \in T_pM$. This mapping is smooth in the sense that the components of the vectors may be written as smooth functions in any local coordinate system. A vector field may be seen as an operator $X : C^\infty(M) \to C^\infty(M)$ that maps a smooth function $f \in C^\infty(M)$ to the smooth function $Xf : p \mapsto X_pf$. In other words, the directional derivative is applied at each point on M.

[1] Strictly speaking, the tangent vectors at p are defined as directional derivatives of smooth **germs** of functions at p, which are equivalence classes of functions that agree in some neighborhood of p.

For two manifolds M and N a smooth mapping $\phi : M \to N$ induces a linear mapping of the tangent spaces $\phi_* : T_pM \to T_{\phi(p)}N$ called the **differential** of ϕ. It is given by $\phi_*(X_p)f = X_p(f \circ \phi)$ for any vector $X_p \in T_pM$ and any smooth function $f \in C^\infty(M)$. A smooth mapping of manifolds does not always induce a mapping of vector fields (for instance, when the mapping is not onto). However, a related concept is given in the following definition.

Definition 2.14. Given a mapping of smooth manifolds $\phi : M \to N$, a vector field X on M and a vector field Y on N are said to be **ϕ-related** if $\phi_*(X(p)) = Y(q)$ holds for each $q \in N$ and each $p \in \phi^{-1}(q)$.

2.3 Riemannian Geometry

As mentioned at the beginning of this chapter, the idea of distances on a manifold will be important in the definition of manifold statistics. The notion of distances on a manifold falls into the realm of Riemannian geometry. This section briefly reviews the concepts needed. A good crash course in Riemannian geometry can be found in [86]. Also, see the books [15, 111, 112, 77].

Recall the definition of length for a smooth curve in Euclidean space. Let $\gamma : [a, b] \to \mathbb{R}^d$ be a smooth curve segment. Then at any point $t_0 \in [a, b]$ the derivative of the curve $\gamma'(t_0)$ gives the velocity of the curve at time t_0. The length of the curve segment γ is given by integrating the speed of the curve, i.e.,

$$L(\gamma) = \int_a^b \|\gamma'(t)\| dt.$$

The definition of the length functional thus requires the ability to take the norm of tangent vectors. On manifolds this is handled by the definition of a Riemannian metric.

2.3.1 Riemannian Metrics

Definition 2.15. A **Riemannian metric** on a manifold M is a function that smoothly assigns to each point $p \in M$ an inner product $\langle \cdot, \cdot \rangle$ on the tangent space T_pM. A **Riemannian manifold** is a smooth manifold equipped with such a Riemannian metric.

Now the norm of a tangent vector $v \in T_pM$ is defined as $\|v\| = \langle v, v \rangle^{\frac{1}{2}}$. Given local coordinates $x^1, \ldots, x^n$ in a neighborhood of p, the coordinate vectors $v^i = \partial/\partial x^i$ at p

form a basis for the tangent space T_pM. The Riemannian metric may be expressed in this basis as an $n \times n$ matrix g, called the metric tensor, with entries given by

$$g_{ij} = \langle v^i, v^j \rangle.$$

The g_{ij} are smooth functions of the coordinates $x^1, \ldots, x^n$.

Given a smooth curve segment $\gamma : [a, b] \to M$, the length of γ can be defined just as in the Euclidean case as

$$L(\gamma) = \int_a^b \|\gamma'(t)\| dt, \tag{2.1}$$

where now the tangent vector $\gamma'(t)$ is a vector in $T_{\gamma(t)}M$, and the norm is given by the Riemannian metric at $\gamma(t)$.

Given a manifolds M and a manifold N with Riemannian metric $\langle \cdot, \cdot \rangle$, a mapping $\phi : M \to N$ induces a metric $\phi^* \langle \cdot, \cdot \rangle$ on M defined as

$$\phi^* \langle X_p, Y_p \rangle = \langle \phi_*(X_p), \phi_*(Y_p) \rangle.$$

This metric is called the **pull-back** metric induced by ϕ, as it maps the metric in the opposite direction of the mapping ϕ.

2.3.2 Geodesics

In Euclidean space the shortest path between two points is a straight line, and the distance between the points is measured as the length of that straight line segment. This notion of shortest paths can be extended to Riemannian manifolds by considering the problem of finding the shortest smooth curve segment between two points on the manifold. If $\gamma : [a, b] \to M$ is a smooth curve on a Riemannian manifold M with endpoints $\gamma(a) = x$ and $\gamma(b) = y$, a **variation of γ keeping endpoints fixed** is a family α of smooth curves:

$$\alpha : (-\epsilon, \epsilon) \times [a, b] \to M,$$

such that

1. $\alpha(0, t) = \gamma(t)$,
2. $\tilde{\alpha}(s_0) : t \mapsto \alpha(s_0, t)$ is a smooth curve segment for fixed $s_0 \in (-\epsilon, \epsilon)$,
3. $\alpha(s, a) = x$, and $\alpha(s, b) = y$ for all $s \in (-\epsilon, \epsilon)$.

Now the shortest smooth path between the points $x, y \in M$ can be seen as finding a critical point for the length functional (2.1), where the length of $\tilde{\alpha}$ is considered as a

function of s. The path $\gamma = \tilde{\alpha}(0)$ is a critical path for L if

$$\frac{dL(\tilde{\alpha}(s))}{ds}\Big|_{s=0} = 0.$$

It turns out to be easier to work with the critical paths of the **energy functional**, which is given by

$$E(\gamma) = \int_a^b \|\gamma'(t)\|^2 dt.$$

It can be shown (see [111]) that a critical path for E is also a critical path for L. Conversely, a critical path for L, once reparameterized proportional to arclength, is a critical path for E. Thus, assuming curves are parameterized proportional to arclength, there is no distinction between curves with minimal length and those with minimal energy. A critical path of the functional E is called a **geodesic**.

Given a chart (x, U) a geodesic curve $\gamma \subset U$ can be written in local coordinates as $\gamma(t) = (\gamma^1(t), \ldots, \gamma^n(t))$. Using any such coordinate system, γ satisfies the following differential equation (see [111] for details):

$$\frac{d^2\gamma^k}{dt^2} = -\sum_{i,j=1}^{n} \Gamma_{ij}^k(\gamma(t)) \frac{d\gamma^i}{dt}\frac{d\gamma^j}{dt}. \tag{2.2}$$

The symbols Γ_{ij}^k are called the **Christoffel symbols** and are defined as

$$\Gamma_{ij}^k = \frac{1}{2}\sum_{l=1}^{n} g^{kl}\left(\frac{\partial g_{jl}}{\partial x^i} + \frac{\partial g_{il}}{\partial x^j} - \frac{\partial g_{ij}}{\partial x^l}\right),$$

where g^{ij} denotes the entries of the inverse matrix g^{-1} of the Riemannian metric.

Example 2.5. In Euclidean space $\mathbb{R}^n$ the Riemannian metric is given by the identity matrix at each point $p \in \mathbb{R}^n$. Since the metric is constant, the Christoffel symbols are zero. Therefore, the geodesic equation (2.2) reduces to

$$\frac{d^2\gamma^k}{dt^2} = 0.$$

The only solutions to this equation are straight lines, so geodesics in $\mathbb{R}^n$ must be straight lines.

Given two points on a Riemannian manifold, there is no guarantee that a geodesic exists between them. There may also be multiple geodesics connecting the two points,

i.e., geodesics are not guaranteed to be unique. Moreover, a geodesic does not have to be a *global* minimum of the length functional, i.e., there may exist geodesics of different lengths between the same two points. The next two examples demonstrate these issues.

Example 2.6. Consider the plane with the origin removed, $\mathbb{R}^2 - \{0\}$, with the same metric as $\mathbb{R}^2$. Geodesics are still given by straight lines. There does not exist a geodesic between the two points $(1, 0)$ and $(-1, 0)$.

Example 2.7. Geodesics on the sphere S^2 are given by great circles, i.e., circles on the sphere with maximal diameter. This fact will be shown later in the section on symmetric spaces. There are an infinite number of equal-length geodesics between the north and south poles, i.e., the meridians. Also, given any two points on S^2 that are not antipodal, there is a unique great circle between them. This great circle is separated into two geodesic segments between the two points. One geodesic segment is longer than the other.

The idea of a global minimum of length leads to a definition of a distance metric $d : M \times M \to \mathbb{R}$ (not to be confused with the Riemannian metric). It is defined as

$$d(p, q) = \inf\{L(\gamma) : \gamma \text{ a smooth curve between } p \text{ and } q\}.$$

If there is a geodesic γ between the points p and q that realizes this distance, i.e., if $L(\gamma) = d(p, q)$, then γ is called a **minimal geodesic**. Minimal geodesics are guaranteed to exist under certain conditions, as described by the following definition and the Hopf-Rinow Theorem below.

Definition 2.16. A Riemannian manifold M is said to be **complete** if every geodesic segment $\gamma : [a, b] \to M$ can be extended to a geodesic from all of $\mathbb{R}$ to M.

The reason such manifolds are called "complete" is revealed in the next theorem.

Theorem 2.2 (Hopf-Rinow). *If M is a complete, connected Riemannian manifold, then the distance metric $d(\cdot, \cdot)$ induced on M is complete. Furthermore, between any two points on M there exists a minimal geodesic.*

Example 2.8. Both Euclidean space $\mathbb{R}^n$ and the sphere S^2 are complete. A straight line in $\mathbb{R}^n$ can extend in both directions indefinitely. Also, a great circle in S^2 extends indefinitely in both directions (even though it wraps around itself). As guaranteed by the Hopf-Rinow Theorem, there is a minimal geodesic between any two points in $\mathbb{R}^n$, i.e.,

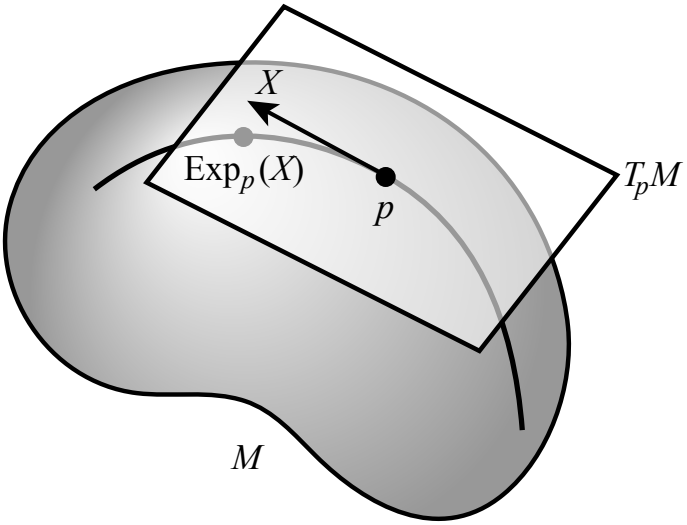

Figure 2.2: The Riemannian exponential map.

the unique straight line segment between the points. Also, between any two points on the sphere there is a minimal geodesic, i.e., the shorter of the two great circle segments between the two points. Of course, for antipodal points on S^2 the minimal geodesic is not unique.

Given initial conditions $\gamma(0) = p$ and $\gamma'(0) = v$, the theory of second-order partial differential equations guarantees the existence of a unique solution to the defining equation for γ (2.2) at least locally. Thus, there is a unique geodesic γ with $\gamma(0) = p$ and $\gamma'(0) = v$ defined in some interval $(-\epsilon, \epsilon)$. When the geodesic γ exists in the interval $[0, 1]$, the **Riemannian exponential map** at the point p (see Figure 2.2), denoted $\mathrm{Exp}_p : T_pM \to M$, is defined as

$$\mathrm{Exp}_p(v) = \gamma(1).$$

If M is a complete manifold, the exponential map is defined for all vectors $v \in T_pM$.

Theorem 2.3. *Given a Riemannian manifold M and a point $p \in M$, the mapping* Exp_p *is a diffeomorphism in some neighborhood $U \subset T_pM$ containing* 0.

This theorem implies that the Exp_p has an inverse defined at least in the neighborhood $\mathrm{Exp}_p(U)$ of p, where U is the same as in Theorem 2.3. Not surprisingly, this inverse is called the **Riemannian log map** and denoted by $\mathrm{Log}_p : \mathrm{Exp}_p(U) \to T_pM$.

Definition 2.17. An **isometry** is a diffeomorphism $\phi : M \to N$ of Riemannian manifolds that preserves the Riemannian metric. That is, if $\langle \cdot, \cdot \rangle_M$ and $\langle \cdot, \cdot \rangle_N$ are the metrics for M and N, respectively, then $\phi^* \langle \cdot, \cdot \rangle_N = \langle \cdot, \cdot \rangle_M$.

It follows from the definitions that an isometry preserves the length of curves. That is, if c is a smooth curve on M, then the curve $\phi \circ c$ is a curve of the same length on N. Also, the image of a geodesic under an isometry is again a geodesic.

2.4 Lie Groups

The set of all possible translations of Euclidean space $\mathbb{R}^n$ is again the space $\mathbb{R}^n$. A point $p \in \mathbb{R}^n$ is transformed by the vector $v \in \mathbb{R}^n$ by vector addition, $p+v$. This transformation has a unique inverse transformation, namely, translation by the negated vector, $-v$. The operation of translation is a smooth mapping of the space $\mathbb{R}^n$. Composing two translations (i.e., addition in $\mathbb{R}^n$) and inverting a translation (i.e., negation in $\mathbb{R}^n$) are also smooth mappings. A set of transformations with these properties, i.e., a smooth manifold with smooth group operations, is known as a Lie group. Many other interesting transformations of Euclidean space are Lie groups, including rotations, reflections, and magnifications. However, Lie groups also arise more generally as smooth transformations of manifolds. This section is a brief introduction to Lie groups. More detailed treatments may be found in [15, 36, 54, 58, 69, 111].

It is assumed that the reader knows the basics of group theory (see [59] for an introduction), but the definition of a group is listed here for reference.

Definition 2.18. A **group** is a set G with a binary operation, denoted here by concatenation, such that

1. $(xy)z = x(yz)$, for all $x, y, z \in G$,
2. there is an **identity**, $e \in G$, satisfying $xe = ex = x$, for all $x \in G$,
3. each $x \in G$ has an **inverse**, $x^{-1} \in G$, satisfying $xx^{-1} = x^{-1}x = e$.

As stated at the beginning of this section, a Lie group adds a smooth manifold structure to a group.

Definition 2.19. A **Lie group** G is a smooth manifold that also forms a group, where the two group operations,

$(x, y) \mapsto xy$	:	$G \times G \to G$	*Multiplication*
$x \mapsto x^{-1}$	:	$G \to G$	*Inverse*

are smooth mappings of manifolds.

Example 2.9. The space of all $n \times n$ non-singular matrices forms a Lie group called the **general linear group**, denoted $GL(n)$. The group operation is matrix multiplication, and $GL(n)$ can be given a smooth manifold structure as an open subset of $\mathbb{R}^{n^2}$. The equations for matrix multiplication and inverse are smooth operations in the entries of the matrices. Thus, $GL(n)$ satisfies the requirements of a Lie group in Definition 2.19. A **matrix group** is any closed subgroup of $GL(n)$. Matrix groups inherit the smooth structure of $GL(n)$ as a subset of $\mathbb{R}^{n^2}$ and are thus also Lie groups. The books [30, 54] focus on the theory of matrix groups.

Example 2.10. The $n \times n$ rotation matrices are a closed matrix subgroup of $GL(n)$ and thus form a Lie group. This group is called the **special orthogonal group** and is defined as $SO(n) = \{R \in GL(n) : R^T R = I \text{ and } \det(R) = 1\}$. This space is a closed and bounded subset of R^{n^2}, so it is compact by the Heine-Borel theorem.

Given a point y in a Lie group G, it is possible to define the following two diffeomorphisms:

$$L_y : x \mapsto yx \qquad (\textit{Left multiplication})$$
$$R_y : x \mapsto xy \qquad (\textit{Right multiplication})$$

A vector field X on a Lie group G is called **left-invariant** if it is invariant under left multiplication, i.e., $L_{y*}X = X$ for every $y \in G$. **Right-invariant** vector fields are defined similarly. A left-invariant (or right-invariant) vector field is uniquely defined by its value on the tangent space at the identity, T_eG.

Recall that vector fields on G can be seen as operators on the space of smooth functions, $C^\infty(G)$. Thus two vector fields X and Y can be composed to form another operator XY on $C^\infty(G)$. However, the operator XY is not necessarily vector field. Surprisingly, however, the operator $XY - YX$ is a vector field on G. This leads to a definition of the **Lie bracket** of vector fields X, Y on G, defined as

$$[X, Y] = XY - YX. \tag{2.3}$$

Definition 2.20. A **Lie algebra** is a vector space V equipped with a bilinear product $[\cdot, \cdot] : V \times V \to V$, called a **Lie bracket**, that satisfies

(1) $[X, Y] = -[Y, X]$,

(2) $[[X,Y],Z]+[[Y,Z],X]+[[Z,X],Y]=0$,
for all $X,Y,Z \in V$.

The tangent space of a Lie group G, typically denoted $\mathfrak{g}$ (a German Fraktur font), forms a Lie algebra. The Lie bracket on $\mathfrak{g}$ is induced by the Lie bracket on the corresponding left-invariant vector fields. If X,Y are two vectors in $\mathfrak{g}$, then let $\tilde{X},\tilde{Y}$ be the corresponding unique left-invariant vector fields on G. Then the Lie bracket on $\mathfrak{g}$ is given by

$$[X,Y]=[\tilde{X},\tilde{Y}](e).$$

The Lie bracket provides a test for whether the Lie group G is commutative. A Lie group G is commutative if and only if the Lie bracket on the corresponding Lie algebra $\mathfrak{g}$ is zero, i.e., $[X,Y]=0$ for all $X,Y \in \mathfrak{g}$.

Example 2.11. The Lie algebra for Euclidean space $\mathbb{R}^n$ is again $\mathbb{R}^n$. The Lie bracket is zero, i.e., $[X,Y]=0$ for all $X,Y \in \mathbb{R}^n$. In fact, the Lie bracket for the Lie algebra of any commutative Lie group is always zero.

Example 2.12. The Lie algebra for $GL(n)$ is $\mathfrak{gl}(n)$, the space of all real $n \times n$ matrices. The Lie bracket operation for $X,Y \in \mathfrak{gl}(n)$ is given by

$$[X,Y]=XY-YX.$$

Here the product XY denotes actual matrix multiplication, which turns out to be the same as composition of the vector field operators (compare to (2.3)). All Lie algebras corresponding to matrix groups are subalgebras of $\mathfrak{gl}(n)$.

Example 2.13. The Lie algebra for the rotation group $SO(n)$ is $\mathfrak{so}(n)$, the space of skew-symmetric matrices. A matrix A is skew-symmetric if $A=-A^T$.

The following theorem will be important later.

Theorem 2.4. *A direct product $G_1 \times \cdots \times G_n$ of Lie groups is also a Lie group.*

2.4.1 Lie Group Exponential and Log Maps

Definition 2.21. A mapping of Lie groups $\phi : G_1 \to G_2$ is called a **Lie group homomorphism** if it is a smooth mapping and a homomorphism of groups, i.e., $\phi(e_1)=e_2$, where e_1,e_2 are the respective identity elements of G_1,G_2, and $\phi(gh)=\phi(g)\phi(h)$ for all $g,h \in G_1$.

The image of a Lie group homomorphism $h : \mathbb{R} \to G$ is called a **one-parameter subgroup**. A one-parameter subgroup is both a smooth curve and a subgroup of G. This does not mean, however, that any one-parameter subgroup is a Lie subgroup of G (it can fail to be an imbedded submanifold of G, which is required to be a Lie subgroup of G). As the next theorem shows, there is a bijective correspondence between the Lie algebra and the one-parameter subgroups.

Theorem 2.5. *Let $\mathfrak{g}$ be the Lie algebra of a Lie group G. Given any vector $X \in \mathfrak{g}$ there is a unique Lie group homomorphism $h_X : \mathbb{R} \to G$ such that $h'_X(0) = X$.*

The **Lie group exponential map**, $\exp : \mathfrak{g} \to G$, not to be confused with the Riemannian exponential map, is defined by

$$\exp(X) = h_X(1).$$

Example 2.14. For the Lie group $\mathbb{R}^n$ the unique Lie group homomorphism $h_X : \mathbb{R} \to \mathbb{R}^n$ in Theorem 2.5 is given by $h_X(t) = tX$. Therefore, one-parameter subgroups are given by straight lines at the origin. The Lie group exponential map is the identity. In this case the Lie group exponential map is the same as the Riemannian exponential map at the origin. This is not always the case, however, as will be shown later.

For matrix groups the Lie group exponential map of a matrix $X \in \mathfrak{gl}(n)$ is computed by the formula

$$\exp(X) = \sum_{k=0}^{\infty} \frac{1}{k!} X^k. \tag{2.4}$$

This series converges absolutely for all $X \in \mathfrak{gl}(n)$.

Example 2.15. For the Lie group of 3D rotations, $SO(3)$, the matrix exponential map takes a simpler form. For a matrix $X \in \mathfrak{so}(3)$ the following identity holds:

$$X^3 = -\theta X, \quad \text{where } \theta = \sqrt{\frac{1}{2}\operatorname{tr}(X^T X)}.$$

Substituting this identity into the infinite series (2.4), the exponential map for $\mathfrak{so}(3)$ can now be reduced to

$$\exp(X) = \begin{cases} I, & \theta = 0, \\ I + \dfrac{\sin\theta}{\theta} X + \dfrac{1 - \cos\theta}{\theta^2} X^2, & \theta \in (0, \pi). \end{cases}$$

The Lie group log map for a rotation matrix $R \in SO(3)$ is given by

$$\log(R) = \begin{cases} I, & \theta = 0, \\ \dfrac{\theta}{2\sin\theta}(R - R^T), & |\theta| \in (0, \pi), \end{cases}$$

where $\operatorname{tr}(R) = 2\cos\theta + 1$.

The exponential map for 3D rotations has an intuitive meaning. Any vector $X \in \mathfrak{so}(3)$, i.e., a skew-symmetric matrix, may be written in the form

$$X = \begin{pmatrix} 0 & -z & y \\ z & 0 & -x \\ -y & x & 0 \end{pmatrix}.$$

If $v = (x, y, z) \in \mathbb{R}^3$, then the rotation matrix given by the exponential map $\exp(X)$ is a 3D rotation by angle $\theta = \|v\|$ about the unit axis $v/\|v\|$.

2.4.2 Bi-invariant Metrics

Definition 2.22. A Riemannian metric $\langle \cdot, \cdot \rangle$ on a Lie group G is said to be a **bi-invariant metric** if it is invariant under both right and left multiplication, that is, $R_g^* \langle \cdot, \cdot \rangle = L_g^* \langle \cdot, \cdot \rangle = \langle \cdot, \cdot \rangle$ for all $g \in G$.

Theorem 2.6. *For a Lie group G with bi-invariant metric the Lie group exponential map agrees with the Riemannian exponential map at the identity, that is, for any tangent vector $X \in \mathfrak{g}$*

$$\exp(X) = \operatorname{Exp}_e(X).$$

Using the left-invariance of the Riemannian metric, any geodesic at a point $g \in G$ may be written as the left multiplication of a geodesic at the identity. That is, the geodesic γ with initial conditions $\gamma(0) = g$ and $\gamma'(0) = L_{g*}(X)$ is given by

$$\gamma(t) = g \exp(tX).$$

Theorem 2.7. *A compact Lie group G has a unique bi-invariant metric (up to scale).*

2.5 Symmetric Spaces

Briefly, a Riemannian symmetric space is a connected manifold M such that at each point the mapping that reverses geodesics through that point is an isometry. For a detailed treatment of symmetric spaces see the standard texts [15, 58]. Common examples of symmetric spaces are Euclidean spaces, $\mathbb{R}^n$, spheres, S^n, and hyperbolic spaces, H^n. Symmetric spaces, and the methods for computing geodesics and distances on them, arise naturally from certain Lie group actions on manifolds.

A few preliminary definitions about mappings of sets are needed before symmetric spaces can be defined. Let X be a set and ϕ be any mapping of X into itself. A point $x \in X$ is called a **fixed point** of ϕ if $\phi(x) = x$. The mapping ϕ is called **involutive** if ϕ is not the identity mapping, but it's square is, i.e., $\phi \circ \phi = \mathrm{id}$.

Definition 2.23. A **symmetric space** is a connected Riemannian manifold M such that at each point $p \in M$ there is an involutive isometry $\phi_p : M \to M$ that has p as an isolated fixed point.

The term **isolated** means that there is a neighborhood U of p such that p is the only point in U that is a fixed point of ϕ_p. This definition is somewhat illusive in that it is hard to get an intuitive feel for what kinds of manifolds are symmetric spaces. Fortunately, this definition is sufficient to imply very nice properties of symmetric spaces. These properties are explained below, and the interested reader is referred to the appropriate references for derivations.

The next theorem (see [15], Lemma 8.2 and Theorem 8.4) shows that the involutive isometry ϕ_p in Definition 2.23 is more easily seen as the map that reverses geodesics through the point p.

Theorem 2.8. *A Riemannian symmetric space is complete, and if ϕ_p is an involutive isometry of M, then ϕ_{p*} is a reflection of the tangent space T_pM, i.e., $\phi_{p*}(X) = -X$, and ϕ_p reverses geodesics through p, i.e., $\phi_p(\mathrm{Exp}_p(X)) = \mathrm{Exp}_p(-X)$ for all $X \in T_pM$ such that those geodesics exist.*

As will be shown later, symmetric spaces arise naturally from certain Lie group transformations of a manifold M. This formulation requires a background to Lie group actions.

2.5.1 Lie Group Actions

Definition 2.24. Given a smooth manifold M and a Lie group G, a **smooth group action** of G on M is a smooth mapping $G \times M \to M$, written $(g, p) \mapsto g \cdot p$, such that for all $g, h \in G$ and all $p \in M$

1. $e \cdot p = p$,
2. $(gh) \cdot p = (g \cdot (h \cdot p))$.

The group action should be thought of as a transformation of the manifold M, just as matrices are transformations of Euclidean space.

The **orbit** of a point $p \in M$ is defined as $G(p) = \{g \cdot p : g \in G\}$. In the case that M consists of a single orbit, we call M a **homogeneous space** and say that the group action is **transitive**. The **isotropy subgroup** of p is defined as $G_p = \{g \in G : g \cdot p = p\}$, i.e., G_p is the subgroup of G that leaves the point p fixed.

Let H be a closed Lie subgroup of the Lie group G. Then the **left coset** of an element $g \in G$ is defined as $gH = \{gh : h \in H\}$. The space of all such cosets is denoted G/H and is a smooth manifold. There is a natural bijection $G(p) \cong G/G_p$ given by the mapping $g \cdot p \mapsto gG_p$. Now let M be a symmetric space and choose an arbitrary base point $p \in M$. We can always write M as a homogeneous space $M = G/G_p$, where G is a connected group of isometries of M, and the isotropy subgroup G_p is compact. The fact that G is a group of isometries means that $d(p, q) = d(g \cdot p, g \cdot q)$, for all $p, q \in M$, $g \in G$.

An element $g \in G$ induces a smooth mapping $\phi_g : M \to M$ via the group action, defined as $\phi_g(p) = g \cdot p$. Also, this mapping has a smooth inverse, namely $\phi_{g^{-1}}$. Therefore, ϕ_g is a diffeomorphism.

Definition 2.25. Given a Lie group action of G on a manifold M, a ***G*-invariant** Riemannian metric $\langle \cdot, \cdot \rangle$ on M is a metric such that the mapping ϕ_g is an isometry for all $g \in G$, i.e., $\phi_g^* \langle \cdot, \cdot \rangle$.

Example 2.16. The standard Euclidean metric on R^n is invariant under the $SO(n)$ group action. In other words, a rotation of Euclidean space is an isometry. The action of $\mathbb{R}^n$ on itself by translations is another example of a group of isometries. These two groups can be combined to form the **special Euclidean group**, $SE(n) = SO(n) \ltimes \mathbb{R}^n$. The semi-direct product $\ltimes$ means that $SE(n)$ as a set is the direct product of $SO(n)$ and $\mathbb{R}^n$, but multiplication is given by the formula

$$(R_1, v_1) * (R_2, v_2) = (R_1 R_2, R_1 \cdot v_2 + v_1).$$

2.5.2 Symmetric Spaces as Lie Group Quotients

The following theorem (see [15], Theorem 9.1) provides criteria for a manifold to possess a G-invariant metric.

Theorem 2.9. *Consider a Lie group G acting transitively on a manifold M. If for some point $p \in M$ the isotropy subgroup G_p is a connected, compact Lie subgroup of G, then M has a G-invariant metric.*

Symmetric spaces arise naturally from homogeneous spaces with G-invariant metrics, as the next theorem shows (see [15], Theorem 9.2 and Corollary 9.3).

Theorem 2.10. *Suppose that G, M, and p satisfy the conditions of Theorem 2.9. If $\alpha : G \to G$ is an involutive automorphism[2] with fixed set G_p, then M is a symmetric space.*

The converse to Theorem 2.10 is also true, as shown in the next theorem (see [58], Theorem 3.3).

Theorem 2.11. *If M is a symmetric space and p any point in M, then M is diffeomorphic to the Lie group quotient G/G_p, where $G = I_0(M)$ is the connected component of the Lie group of isometries of M and G_p is the compact Lie subgroup of G that leaves the point p fixed. Furthermore, there is an involutive automorphism $\alpha : G \to G$ that leaves G_p fixed.*

Theorem 2.12. *A connected Lie group G with bi-invariant metric is a symmetric space.*

Example 2.17. Euclidean space $\mathbb{R}^n$ is a symmetric space, as can be seen by Theorem 2.12. The involutive isometry ϕ_p is given by reflection about p, i.e., ϕ_p reverses lines through p by the equation

$$\phi_p(q) = 2p - q.$$

Geodesics on a symmetric space $M = G/G_p$ are computed through the group action. Since G is a group of isometries acting transitively on M, it suffices to consider only geodesics starting at the base point p. For an arbitrary point $q \in M$, geodesics starting at q are of the form $g \cdot \gamma$, where $q = g \cdot p$ and γ is a geodesic with $\gamma(0) = p$. Geodesics are the image of the action of a one-parameter subgroup of G acting on the base point p, as the next theorem shows.

[2]Recall that an automorphism of a group G is an isomorphism of G onto itself.

Theorem 2.13. *If M is a symmetric space with G-invariant metric, as in Theorem 2.10, then a geodesic γ starting at the point $p \in M$ is of the form*

$$\gamma(t) = \exp(tX) \cdot p,$$

where X is a vector in the Lie algebra $\mathfrak{g}$.

Example 2.18. The sphere S^2 is a symmetric space. The rotation group $SO(3)$ acts transitively on S^2, that is, for any two unit vectors x, y there is a rotation R such that $Rx = y$. The north pole $p = (0, 0, 1)$ is left fixed by any rotation of the x-y plane. Therefore, the isotropy subgroup for p is equivalent to $SO(2)$. The sphere can thus be written as the homogeneous space $S^2 = SO(3)/SO(2)$. The involutive isometry ϕ_p is given by reflection about p, i.e., a rotation of the sphere about the axis p by an angle of π.

The geodesics at the base point $p = (0, 0, 1)$ are the great circles through p, i.e., the meridians. Geodesics at an arbitrary point in S^2 are also great circles, i.e., rotated versions of the meridians. As Theorem 2.13 shows, these geodesics are realized by the group action of a one-parameter subgroup of $SO(3)$. Such a subgroup consists of all rotations about a fixed axis in $\mathbb{R}^3$ perpendicular to p. We consider a tangent vector in T_pS^2 as a vector $v = (v_1, v_2, 0)$ in the x-y plane. Then the exponential map is given by

$$\mathrm{Exp}_p(v) = \left(v_1 \cdot \frac{\sin \|v\|}{\|v\|},\ v_2 \cdot \frac{\sin \|v\|}{\|v\|},\ \cos \|v\| \right), \tag{2.5}$$

where $\|v\| = \sqrt{v_1^2 + v_2^2}$. This equation can be derived as a sequence of two rotations that rotate the base point $p = (0, 0, 1)$ to the point $\mathrm{Exp}_p(v)$. The first is a rotation about the y-axis by an angle of $\phi_y = \|v\|$. The second, aligning the geodesic with the tangent vector v, is a rotation about the z-axis by an angle of ϕ_z, where $\cos(\phi_z) = v_1/\|v\|$ and $\sin(\phi_z) = v_2/\|v\|$.

The corresponding log map for a point $x = (x_1, x_2, x_3) \in S^2$ is given by

$$\mathrm{Log}_p(x) = \left(x_1 \cdot \frac{\theta}{\sin \theta},\ x_2 \cdot \frac{\theta}{\sin \theta} \right), \tag{2.6}$$

where $\theta = \arccos(x_3)$ is the spherical distance from the base point p to the point x. Notice that the antipodal point $-p$ is not in the domain of the log map.

Chapter 3

Image Analysis Background

This chapter provides the necessary background to the aspects of image analysis that are relevant to this dissertation. It begins in Section 3.1 with an overview of the statistical theory of shape. This theory is an important tool in deformable models, which are discussed in Section 3.2. Medial representations, the particular type of deformable model used in this work, are presented in Section 3.3.2. Finally, in Section 3.4 the necessary background for diffusion tensor imaging is covered.

3.1 Statistical Shape Theory

Statistical shape analysis is emerging as an important tool for understanding anatomical structures from medical images. Given a set of training images, the goal is to model the geometric variability of the anatomical structures within a class of images. Statistical models give an efficient parameterization of the geometric variability of anatomy. These models can provide shape constraints during image segmentation [27]. Also, statistical descriptions of shape are useful in understanding the processes behind growth and disease [29]. The study of anatomical shape and its relation to biological growth and function dates back to the landmark work of D'Arcy W. Thompson in 1917 [125]. This section is a review of several key concepts in statistical shape theory. Subsection 3.1.1 is a review of the shape theory of point sets introduced by David G. Kendall in a brief note in 1977 [71] and detailed in his 1984 paper [72]. Similar ideas in the theory of shape were independently developed by Fred L. Bookstein [12, 13]. In the Kendall and Bookstein theories of shape an object is represented by a finite set of points in Euclidean space $\mathbb{R}^n$. In medical image analysis these points may represent a sampling of the boundary of an organ or important landmarks in a medical image. Subsection 3.1.2 is an overview

of methods for aligning geometric objects to a common position, orientation, and scale. Alignment is an important preprocessing step that is necessary for analyzing the shape differences between objects. Subsection 3.1.3 discusses the common linear methods used to analyze the statistical variability of shape. Subsection 3.1.4 reviews several methods that have been proposed for the statistical analysis of nonlinear geometric data. For a more in-depth overview of shape theory, including applications beyond the realm of medical image analysis, see the books [14, 35, 109] and the review article [73]. Of these references the book by Dryden and Mardia [35] is the easiest to digest. Readers interested in more of the mathematical details of shape theory are encouraged to read Small's book [109].

3.1.1 Point Set Shape Spaces

Shape is often defined as the geometry of objects that is invariant under translation, rotation, and scaling. This definition of shape provides an equivalence relation between objects, that is, two objects have the same shape if one can be transformed into the other by only a translation, rotation, and scaling (see Fig. 3.1). A shape space is a space in which each point represents an entire equivalence class of objects under this relation. This subsection is a review of Kendall's shape spaces of point sets in $\mathbb{R}^n$ [72]. The first step in the construction of these shape spaces is to define the transformation group of combined translations, rotations, and scalings. This group leads to an action on point sets in $\mathbb{R}^n$. Shape spaces will be defined as the orbit spaces under this action; that is, point sets that can be transformed into each other under this action will be associated with the same point in shape space.

A **similarity transform** of $\mathbb{R}^n$ is a combined translation, rotation, and scale of $\mathbb{R}^n$. The space of all such transformations can be written as the Lie group $\mathrm{Sim}(n) = (SO(n) \times \mathbb{R}^+) \ltimes \mathbb{R}^n$ (recall the definition of the special Euclidean group in Example 2.16). An element $S \in \mathrm{Sim}(n)$ can be written as an $(n+1) \times (n+1)$ matrix in the block form

$$S = \begin{bmatrix} sR & v \\ 0^T & 1 \end{bmatrix},$$

where $v \in \mathbb{R}^n$, $s \in \mathbb{R}^+$, and $R \in SO(3)$. Composition of two similarity transformations is now achieved by multiplying their representative matrices. Writing a vector $x \in \mathbb{R}^n$ in homogeneous coordinates, i.e., as the column vector $(x, 1)^T$, a similarity transform

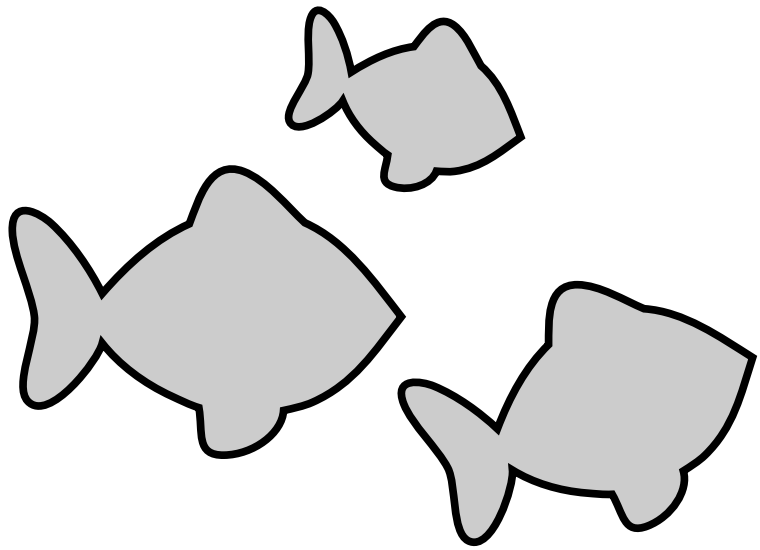

Figure 3.1: Three objects that have the same shape, yet have different positions, orientations, and scales.

matrix acts on x by

$$S \cdot x = \begin{bmatrix} sR & v \\ 0^T & 1 \end{bmatrix} \begin{bmatrix} x \\ 1 \end{bmatrix}.$$

Now consider a collection of k points $x_1, \ldots, x_k$ in n-dimensional Euclidean space. This collection should be thought of as the vector $\mathbf{x} = (x_1, \ldots, x_n)$ in $\mathbb{R}^{nk}$. The configuration where all points are equal is not allowed. That is, the n-dimensional linear subspace $V = \{(x_1, \ldots, x_k) \in \mathbb{R}^{nk} : x_1 = x_2 = \ldots = x_k\}$ is subtracted from the set $\mathbb{R}^{nk}$ to form the space of legal point configurations, $\mathbb{R}^{nk} \backslash V$. A similarity transformation $S \in \mathrm{Sim}(n)$ acts on $\mathbb{R}^{nk} \backslash V$ by applying S to each of the points in the collection:

$$S \cdot \mathbf{x} = (S \cdot x_1, \ldots, S \cdot x_k).$$

The **shape space** Σ_n^k of k points in $\mathbb{R}^n$ is defined as the set of orbits under this action. Recall that two points $\mathbf{x}$ and $\mathbf{y}$ are in the same orbit if there exists a similarity transformation S such that $S \cdot \mathbf{x} = \mathbf{y}$.

The shape space Σ_n^k can be constructed by removing the translation, scale, and rotation effects from the space $\mathbb{R}^{nk} \backslash V$. Consider the point set $\mathbf{x} = (x_1, \ldots, x_k)$. The center of mass of these points is given by $\bar{x} = (1/k) \sum_{i=1}^{k} x_k$. Then $\mathbf{x}$ is determined up to translation by the point set $\tilde{\mathbf{x}} = (x_1 - \bar{x}, \ldots, x_k - \bar{x})$. The space of all such point sets

with zero centroid can be identified with the space $\mathbb{R}^{n(k-1)}\backslash\{0\}$. The scale effects can be removed by dividing the point $\tilde{x}$ by its Euclidean norm in $\mathbb{R}^{nk}$. Therefore, the space of all scale-normalized point sets with zero centroid is a sphere of dimension $n(k-1)-1$. This is called the **preshape space** S_n^k. Finally, the shape space Σ_n^k is the set of orbits under the action of the rotation group $SO(n)$, that is, $\Sigma_n^k \cong S_n^k/SO(n)$. Recall that this quotient of spaces means that points in S_n^k that can be transformed into each other by a rotation in $SO(n)$ are associated with the same point in the quotient space $S_n^k/SO(n)$.

The topology of the shape space Σ_n^k is somewhat harder to understand. For data on the real line, i.e., $n = 1$, the rotation group $SO(1)$ consists of only the identity transformation. Thus, Σ_1^k is identical to the preshape space, which is the sphere S^{k-2}. For planar data, $n = 2$, it helps to consider the plane $\mathbb{R}^2$ as the set of complex numbers $\mathbb{C}$. In this case it can be shown (see [109]) that the shape space Σ_2^k is equivalent to the complex projective space $\mathbb{C}\mathrm{P}(k-2)$. Complex projective space $\mathbb{C}\mathrm{P}(n)$ is the manifold of all one-dimensional complex subspaces of $\mathbb{C}^{n+1}$. The picture gets more difficult for higher dimensions, $n \geq 3$. Here the shape space Σ_n^k, with $k > n$, is a singular manifold. The singularities arise from the configurations of points that are invariant to certain rotations. For example, given a preshape of points in $\mathbb{R}^3$ that are collinear, the rotations about that common line leave all the points fixed. This is a failure of $SO(3)$ to act **freely** on preshape space S_3^k. In the smooth parts of Σ_n^k, that is, where $SO(n)$ acts freely, the dimension will be reduced by the dimension of $SO(n)$, which is $\frac{1}{2}n(n-1)$. Therefore, the dimension of the smooth parts of Σ_n^k is $n(k-1) - \frac{1}{2}n(n-1) - 1$. The dimension of the singularities of Σ_n^k is the same as the dimension of the lower-dimensional shape space Σ_k^{n-2}. For example, in the shape space of 3D objects, Σ_3^k, the singularities have dimension $k-2$. For more information on the structure of the shape space manifolds, including the Riemannian metric and curvature properties, see the paper [76].

3.1.2 Procrustes Distance and Alignment

Consider the problem of defining a distance metric on the shape space Σ_n^k. Given two sets of landmarks, $\mathbf{x} = (x_1, \ldots, x_k)$ and $\mathbf{y} = (y_1, \ldots, y_k)$, whose points are in one-to-one correspondence with each other, the problem is to define a distance $d(\mathbf{x}, \mathbf{y})$ that is invariant to translation, rotation, and scaling of either $\mathbf{x}$ or $\mathbf{y}$. The **Procrustes distance** [47] is one such metric. It is based on a sum-of-squares Euclidean distance

between the corresponding points,

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^{k} \|x_i - y_i\|^2 \right)^{\frac{1}{2}}. \tag{3.1}$$

This is of course equivalent to the Euclidean norm $\|\mathbf{x}-\mathbf{y}\|$ if the point sets are considered as elements of $\mathbb{R}^{nk}$. Now the Procrustes distance is the distance induced on Σ_n^k by this Euclidean distance. This is typically approximated by using the distance in (3.1) after first aligning the two point sets to a common position, orientation, and scale in the following manner:

1. Translate each point set so that its centroid is located at zero.
2. Scale both point sets to norm one (considering them as points in $\mathbb{R}^{nk}$).
3. Rotate one point set to minimize the sum-of-square distances given in (3.1).

This alignment process is known as **ordinary Procrustes analysis** (OPA). The rotation necessary in the last step may be computed using a singular value decomposition (SVD) of the $n \times n$ matrix $\mathbf{x}^T\mathbf{y}$, where $\mathbf{x}$ and $\mathbf{y}$ are considered as $k \times n$ matrices, i.e., matrices with the landmarks as rows. Let $U\Lambda V$ be the corresponding SVD. Then the rotation matrix needed in step 3 to rotate $\mathbf{y}$ in alignment with $\mathbf{x}$ is UV^T. This is the rotation matrix that maximizes the correlation between the two point sets.

Alignment of more than two objects is achieved by a process called **generalized Procrustes analysis** (GPA) [48]. The GPA algorithm for a collection of objects $\mathbf{x}_1, \ldots, \mathbf{x}_N$ is given by

1. Translate each object to a centroid at zero.
2. Compute the linear average of the objects, i.e., $\mu = \sum_{i=1}^{N} \mathbf{x}_i$. Normalize the mean to norm one.
3. Align each object $\mathbf{x}_i$ to the mean μ with respect to orientation and scale using OPA.
4. Repeat steps 2 and 3 until the mean does not change.

In addition to aligning all objects into a common coordinate system, generalized Procrustes analysis also results in the production of a mean shape μ. In essence GPA

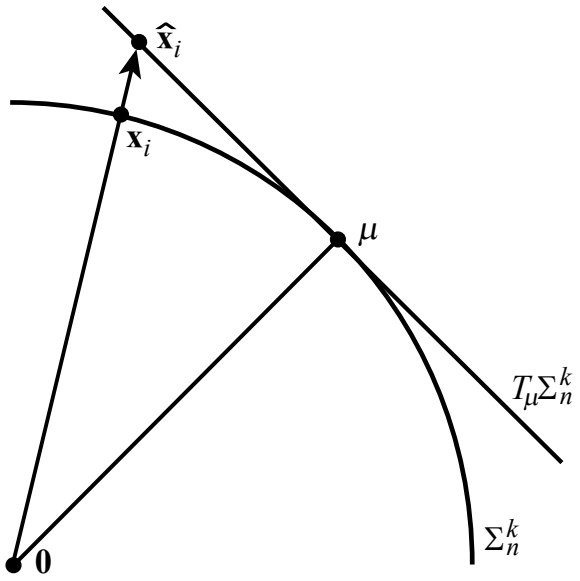

Figure 3.2: Projection of an object onto the tangent space of the shape space Σ_n^k.

maps each object onto the shape space Σ_n^k, which is a curved manifold. However, one would like to use linear statistics to analyze the variability of these shapes. Therefore, it is necessary to linearize the data in some fashion. This is achieved by projecting the shapes onto the tangent space at the mean, i.e., $T\mu\Sigma_n^k$. Since the shape space Σ_n^k has spherical curvature, the tangent space at μ is the set of vectors perpendicular to μ. The projection is accomplished by scaling the vector $\mathbf{x}_i$, producing a $\hat{\mathbf{x}}$ such that the difference $\hat{\mathbf{x}} - \mu$ is perpendicular to μ (see Figure 3.2). Given one of the objects $\mathbf{x}_i$ aligned using GPA, its projection onto the tangent space $T_\mu\Sigma_n^k$ is given by

$$\hat{\mathbf{x}}_i = \frac{1}{\langle \mathbf{x}_i, \mu \rangle} \mathbf{x}_i,$$

where μ is again the normalized mean, i.e., $\|\mu\| = 1$, resulting from GPA.

3.1.3 Shape Variability

The standard technique for describing the variability of linear shape data is principal component analysis (PCA), a method whose origins go back to Pearson [99] and Hotelling [60]. Its use in shape analysis and deformable models was introduced by Cootes and Taylor [26]. See the book [63] for a comprehensive review of PCA. The objectives of principal component analysis are (1) to efficiently parameterize the variability

of data and (2) to decrease the dimensionality of the data parameters. This section describes PCA of multivariate data $x_1, \ldots, x_N \in \mathbb{R}^n$ with mean μ. The reader may think of this data as a set of shapes represented as projections onto the linear tangent space $T_\mu \Sigma_n^k$.

The goal of PCA is to find a sequence of linear subspaces, $V_1, \ldots, V_n$, through the mean that best approximate the data. This may be formulated in two ways, both resulting in the same answer. The first is a least-squares approach, where the objective is to find the linear subspaces such that the sum-of-squares of the residuals to the data are minimized. More precisely, the linear subspace V_k is defined by a basis of orthonormal vectors, i.e., $V_k = \text{span}(\{v_1, \ldots, v_k\})$, which are given by

$$v_k = \arg\min_{\|v\|=1} \sum_{i=1}^{N} \|x_i^k - \langle x_i^k, v \rangle \, v\|^2, \tag{3.2}$$

where the x_i^k are defined recursively by

$$\begin{aligned} x_i^1 &= x_i - \mu, \\ x_i^k &= x_i^{k-1} - \langle x_i^{k-1}, v_{k-1} \rangle \, v_{k-1} \end{aligned}$$

Simply put, the point x_i^k is obtained by removing from $(x_i - \mu)$ the contributions of the previous directions, $v_1, \ldots, v_{k-1}$. In other words, the point x_i^k is the projection of $(x_i - \mu)$ onto the subspace perpendicular to V_{k-1}.

The other way of defining principal component analysis is as the subspaces through the mean that maximize the total variance of the projected data. The total variance for a set of points $y_1, \ldots, y_N$ is defined as

$$\sigma^2 = \frac{1}{N} \sum_{i=1}^{N} \|y_i - \mu\|^2.$$

Then the linear subspaces $V_k = \text{span}(\{v_1, \ldots, v_k\})$ are given by the vectors

$$v_k = \arg\max_{\|v\|=1} \sum_{i=1}^{N} \langle x_i^k, v \rangle^2, \tag{3.3}$$

where the x_i^k are defined as above. It can be shown (see [63]) that both definitions of PCA, i.e., (3.2) and (3.3), give the same results thanks to the Pythagorean theorem.

The computation of the spanning vectors v_k proceeds as follows. First, the linear

average of the data is computed as

$$\hat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i.$$

Next, the sample covariance matrix of the data is computed as

$$S = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \hat{\mu})(x_i - \hat{\mu})^T.$$

This is the unbiased estimate of the covariance matrix, that is, $N-1$ is used in the denominator instead of N. The covariance matrix is a symmetric, positive-semidefinite quadratic form, that is, $S = S^T$, and for any $x \in \mathbb{R}^n$ the inequality $x^T S x \geq 0$ holds. Therefore, the eigenvalues of S are all real and nonnegative. Let $\lambda_1, \ldots, \lambda_n$ be the eigenvalues of S ordered so that $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$, and let $v_1, \ldots, v_n$ be the correspondingly ordered eigenvectors [1]. These directions are the solutions to the defining PCA equations, (3.2) and (3.3), and are called the **principal directions** or **modes of variation**.

Any data point x_i can be decomposed as

$$x_i = \mu + \sum_{k=1}^{n} \alpha_{ik} v_k,$$

for real coefficients $\alpha_{ik} = \langle x_i - \mu, v_k \rangle$. The α_{ik} for fixed i are called the **principal components** of x_i. The total variation of the data is given by the sum of the eigenvalues, $\sigma^2 = \sum_{k=1}^{n} \lambda_k$. The dimensionality of the data can be reduced by discarding the principal directions that contribute little to the variation, that is, choosing an $m < n$ and projecting the data onto V_m, giving the approximation

$$\tilde{x}_i = \mu + \sum_{k=1}^{m} \alpha_{ik} v_k.$$

Typically the cut-off value m is chosen based on the percentage of total variation that should be preserved.

The mean μ and covariance matrix S can be considered as the maximum likelihood estimates of the parameters of a Gaussian probability distribution. The resulting

[1] When repeated eigenvalues occur, there is an ambiguity in the corresponding eigenvectors, i.e., there is a hyperplane from which to choose the corresponding eigenvectors. This does not present a problem as any orthonormal set of eigenvectors may be chosen.

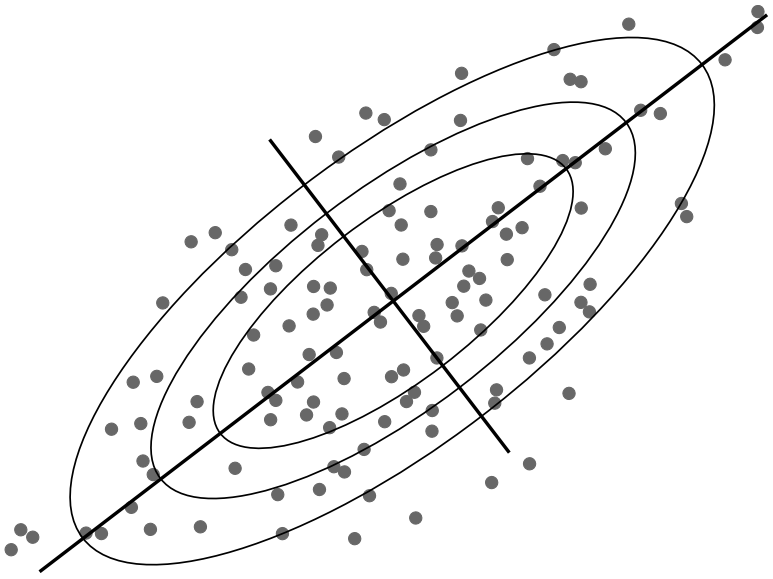

Figure 3.3: A set of points in $\mathbb{R}^n$ showing the resulting principal directions weighted by the corresponding variances and the level sets of Mahalanobis distance d.

Gaussian distribution is given by the density

$$p(x) = \frac{1}{(2\pi)^{\frac{n}{2}}|S|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T S^{-1}(x-\mu)\right).$$

This defines a probability distribution on the space of shapes that can be used as a geometric prior in a deformable models framework (described in the next section). However, PCA is a valid operation even if the data cannot be assumed to come from a Gaussian process. It can give reasonable resulting modes of variation for data that is "Gaussian-like", i.e., densities that are unimodal and fall off rapidly away from the mean. As an alternative to using the full Gaussian probability density, a useful measure of the geometric typicality of an object is given by the squared **Mahalanobis distance** from the mean,

$$d(x,\mu)^2 = (x-\mu)^T S^{-1}(x-\mu).$$

The Mahalanobis distance function uses the covariance matrix as a quadratic form to create an inner product on $\mathbb{R}^n$. This gives hyperelliptical level sets of distance emanating from the mean, where the axes of the hyperellipsoids are the principal directions

from PCA (see Fig. 3.3). The Mahalanobis distance skews Euclidean distance so that directions with higher variance become closer to the mean. Points nearer to the mean in Mahalanobis distance represent more probable shapes.

3.1.4 Nonlinear Statistical Analysis

While most work on the statistical analysis of shape has focused on linear methods, there has been some work on statistical methods for nonlinear geometric data. Hunt [61] describes probability measures on Lie groups that satisfy the semigroup property under convolution. This leads to a natural definition of a Gaussian distribution on a Lie group as a fundamental solution to the heat equation

$$\begin{aligned}\frac{\partial f}{\partial t} &= \Delta f = \mathrm{div}(\mathrm{grad} f) \\ &= g^{ij}(\frac{\partial^2 f}{\partial x^i \partial x^j} - \Gamma_{ij}^k \frac{\partial f}{\partial x^k}),\end{aligned}$$

where g^{ij} are the components of the inverse of the Riemannian metric, and Γ_{ij}^k are the Christoffel symbols. Wehn [128, 129] shows that such distributions satisfy a law of large numbers as in the Euclidean Gaussian case. Grenander's book [49] on probabilities on algebraic structures includes a review of these works on Gaussian distributions on Lie groups.

Pennec [100] defines Gaussian distributions on a manifold as probability densities that minimize information. Bhattacharya [5] develops nonparametric statistics of the mean and dispersion values for data on a manifold. Mardia [81] describes several methods for the statistical analysis of directional data, i.e., data on spheres and projective spaces. Kendall [72] and also Mardia and Dryden [82] have studied the probability distributions induced on shape space Σ_n^k by independent identically distributed Gaussian distributions on the landmarks. Olsen [96,95]and Swann [121] describe Lie group actions on shape space Σ_n^k that result in nonlinear variations of shape. Klassen et al. [75] develop an infinite-dimensional shape space representing smooth curves in the plane. The space of diffeomorphisms is an infinite dimensional and curved Lie group, and statistical analysis of diffeomorphisms has found interest recently. Davis et al. [32] describe a method for estimating a minimum mean squared error diffeomorphism from a set of images. Nielsen et al. [94] and Markussen [83] use Brownian motion warps as a least-committed prior on the space of diffeomorphisms.

3.2 Deformable Models

Segmentation is the process of distinguishing important structures in an image from background. This is a fundamental task in medical image analysis that is often a prerequisite for further analysis, visualization, disease diagnosis, or planning of medical treatment. Medical images can be very large, especially 3D images, time sequence images, or images with multi-dimensional values, such as diffusion tensor images. Finding complex geometric objects in such a vast amount of data can be a challenge. Segmentation is further complicated by the wide range of variability in the geometry and image intensities of the anatomy. Image noise, sampling artifacts, and the confusion of other nearby structures add to the difficulty of the task. Deformable models is a powerful image analysis method that overcomes many of the difficulties of segmentation by incorporating prior information of the objects to be segmented. A survey of deformable models methods can be found in [85].

The deformable models approach to segmentation involves the deformation of a geometric model into an image by optimizing an objective function. Deformable model approaches differ in the way they represent object geometry, how they deform objects, and in the objective functions they use to fit into an image.

3.2.1 Active Contours

The first deformable models to gain popularity in image analysis were the active contours or snakes [68]. Snakes represent an object in a 2D image as a parametric contour $c(s) = (x(s), y(s))$, for $s \in [0, 1]$. In the early papers on deformable models such a contour is fit to an object in an image $I(x, y)$ by minimizing the following energy functional:

$$E_{\text{snake}}(c) = \int_0^1 \alpha \|c'(s)\|^2 + \beta \|c''(s)\|^2 ds + \int_0^1 P(c(s)) ds. \tag{3.4}$$

The first integrand in the above equation is called the **internal energy**. The weights α and β specify the **elasticity** and **stiffness** of the contour. The second integrand in the above equation is known as the **external energy**. It measures how well the contour fits the image data. The function $P(x, y)$ is a potential function in the image plane that typically measures the desired image features, such as specific intensities or edges. A

common potential function is the edge-based potential

$$P(x, y) = \lambda g(\|\nabla I(x, y)\|)^2,$$

which attracts the contour to edges in the image, i.e., places with high gradient values. Here $g : [0, \infty) \to \mathbb{R}^+$ is a monotonic decreasing function, and the weight λ is chosen to balance the strength of the image attraction with the internal energy constraints. It is also common in this approach to first convolve the image with a Gaussian kernel to remove noise and extend the capture range of the local minima.

One drawback of the active contour method is that the snakes energy functional (3.4) is not intrinsic in the sense that it contains the term

$$E(c) = \int_0^1 \|c'(s)\|^2 ds,$$

which depends on the parameterization of the curve. This can be remedied by choosing an arclength parameterization for c. Caselles et al. [21] go a step further and phrase the active contour minimization as finding a geodesic curve under a particular Riemannian metric, resulting in a method called **geodesic active contours**. The metric is chosen in the image plane so that the resulting geodesics minimize the snakes energy functional (3.4) with the term $\beta = 0$. Instead of using a parameterized curve model for the snake, geodesic active contours use a level-set approach based on the work of Osher and Sethian [97, 107]. This can be phrased as finding a smooth function $f : \mathbb{R}^2 \to \mathbb{R}$ in the image plane whose zero set is the desired contour. Following a steepest-descent approach, the geodesic active contours solve the evolution equation

$$\frac{dc}{dt} = g(I)\kappa N - \langle \nabla g, N \rangle N,$$

where N is the curve normal, κ is the curvature, and g is a monotonic decreasing function as above.

A distinguishing characteristic of active contours is that they are inherently a *local* approach. The components of the energy functional (3.4) are all local properties: first and second derivatives of the contour and local image properties. In addition the search methods for active contours proceed by local searches along the normal direction at a point on the curve. This locality is an advantage in the sense that it is typically very fast because computations need only be made in local neighborhoods. However, it is also a disadvantage in the sense that active contours are unable to describe global aspects of

shape change and they do not take into account any correlations of image information at different points on an object. As a result, active contours have a tendency to be attracted to local spurious features in an image and can "leak", i.e., continue searching for an edge when the desired boundary in an image has low contrast.

3.2.2 Probabilistic Deformable Models

An alternative to the energy minimization formulation of deformable models is based on a probabilistic point of view.[2] The probabilistic framework is based on Grenander's **pattern theory** [50, 51, 52, 90]. Pattern theory encompasses a wide variety of methods for analyzing signals generated by the world. These signals may include images, audio, DNA sequences, or weather measurements. Pattern theory holds that the real world cannot be modeled deterministically because it is too complex and the sensors used to observe it are too limited. Therefore, observations must be modeled partly stochastically and partly deterministically in order to make analysis of the observations computationally practical.

In a stochastic model of observations, the world may be in one of many different states, and each state w in the set Ω of possible states occurs with probability $p(w)$. The probability $p(w)$ is called a **prior** and must be learned from past experience. For example, a radiologist uses prior training in anatomy when segmenting a new CT image. An observation f of the world has conditional probability $p(f|w)$, which is the **likelihood** of the observation f given that the world is in state w. The likelihood is often computed by generating a synthetic signal f_w from a given model w and comparing it to the signal f. The goal is to infer the true state w given an observation f. This may be done by maximizing the **a posteriori** probability $p(w|f)$ with respect to w, that is, find the state w that has maximum probability given the observation f. The posterior probability is computed using Bayes' formula

$$p(w|f) = \frac{p(f|w)p(w)}{p(f)}. \tag{3.5}$$

Now the most probable estimate for the state w, called the **maximum a posteriori**

[2]The energy minimization snakes can actually be phrased in the probabilistic setting as well. This approach involves using Gibbs probability distributions for a smoothness prior term and an image likelihood term. Such a formulation can be shown to be equivalent to the energy minimization (3.4) from the previous section. See [85] for more details.

(MAP) estimate, is given by

$$\hat{w} = \arg\max_{w \in \Omega} \; p(w|f) = \arg\max_{w \in \Omega} \; p(f|w)p(w). \tag{3.6}$$

The term $p(f)$ in the denominator of Bayes' formula (3.5) is dropped from the MAP equation because it does not alter which state maximizes the posterior probability.

In the deformable models setting, the observations f are images and the states w are geometric models of the objects in the images. Thus, the MAP estimate can be phrased as the most probable configuration of a geometric model with respect to a given particular image. In practice the posterior maximization cannot be solved analytically due to the large number of variables and the complexity of the deformable model problem. Therefore, the following procedure is used:

1. Begin with an initial estimate for the model w.
2. Generate a synthetic image f_w from w.
3. Evaluate the posterior probability, comparing f_w to f.
4. Update (deform) the model w according to the method used to search for the optimum w.
5. Repeat steps 2 through 5 until maximum is achieved.

The search method used in step 4 may be one of several optimization strategies (see [104], Chp. 10). Local methods such as the simplex method, gradient descent, or conjugate gradient can find optimum solutions quickly but can get stuck in local optima. Thus, local methods work well for problems where the initial estimate (step 1) can be placed near the correct answer. Global optimization strategies such as simulated annealing and genetic algorithms do a better job of avoiding local optima but are also much slower, even when near the correct solution.

Several different geometric representations have been used to model anatomy in deformable models approaches. The active shape model (ASM) of Cootes and Taylor [26, 27] represents an object's geometry as a dense collection of boundary points. Cootes et. al. [25] have augmented their models to include the variability of the image information as well as shape. Delingette [33, 34] uses a simplex mesh to represent the boundary of objects. Staib and Duncan [117] use Fourier decompositions of contours. Székely, Kelemen, et. al. [70, 122] also use Fourier representations in 2D and use a spherical

harmonic (SPHARM) decomposition of the object geometry in 3D. Joshi [64, 65] and Christensen [24] use volumetric representations of anatomy along with diffeomorphic transformations of that anatomy.

In all of these approaches the underlying geometry is parameterized as a Euclidean vector space. The prior probability density must be inferred from a training sample of known instances of the object. The training data is given as a set of vectors $x_1, \ldots, x_N$ in a vector space V. For active shape models each vector is constructed by concatenation of the boundary points in an object. This is followed by a general Procrustes analysis and a tangent space projection as described in the previous section. For Fourier and spherical harmonics each vector is constructed as the concatenation of the coefficients of a harmonic representation of the object boundary. Although diffeomorphisms themselves are not a vector space, prior probability models can be based on the velocity vector fields of the deformations, which do form a vector space. Therefore, in each of these approaches the parameters of the prior density can be inferred using the linear statistical techniques mentioned in the previous section, namely linear averaging and PCA.

In contrast to active contours, probabilistic deformable models are a more *global* approach. Methods for describing the statistical variability of shape, such as PCA, take into account the global variations in shape. That is, changes in the components of variation cause changes across the entire object. This is a result of the fact that PCA models the correlations of geometric changes in different parts of the object. Also, statistical models of the image variability use correlations of image values at different points on the object. This global approach has the advantage that the models of the geometry and the image values stay consistent across the entire object. Search methods can take steps in the model parameters, which result in global changes to the model geometry and image intensities. Thus, probabilistic models are less likely to be attracted to spurious image features and are more robust under low contrast or missing data. This added power comes at the cost of more complex models and longer run times.

3.3 Medial Representations

Medial representations of objects, or m-reps, are the foundation of the deformable models approach taken in this work. This section is a review of the necessary background in medial geometry representations and segmentation via deformable m-rep models. The first subsection (Section 3.3.1) is an overview of the medial locus and some of its mathematical properties. The next subsection (Section 3.3.2) describes m-reps and the

deformable models approach based on them. The article by Pizer et al. [103] provides an overview of the properties of the medial locus and methods for extracting the medial locus from an object. The deformable m-reps approach to image segmentation is described by Pizer et al. [102]. A fine overview of medial techniques that goes beyond the material covered in this section can be found in the Ph.D. dissertation of Yushkevich [131].

3.3.1 The Medial Locus

The medial locus is a means of representing the "middle" or "skeleton" of a geometric object. Such representations have found wide use in computer vision, image analysis, graphics, and computer aided design [8, 9, 62, 108, 119]. Psychophysical and neurophysiological studies have shown evidence that medial relationships play an important role in the human visual system [6, 17, 78, 79, 84]. The medial locus was first proposed by Blum in 1967 [10], and its properties were later studied in 2D by Blum and Nagel [11] and in 3D by Nackman [92]. Arising from the medial locus definition is a surprisingly rich mathematical theory that incorporates many aspects from differential geometry and singularity theory (see, for instance, [31, 46, 45]).

The definition of the medial locus of a set $A \subset \mathbb{R}^n$ is based on the concept of a maximal inscribed ball.

Definition 3.1. A **maximal inscribed ball** of a set $A \subset \mathbb{R}^n$ is an open ball $B_r(x) = \{y \in \mathbb{R}^n : \|x - y\| < r\}$ such that $B_r(x) \subset A$, and there does not exist another ball $B' \neq B_r(x)$ such that $B_r(x) \subset B' \subset A$.

Definition 3.2. The **medial locus** of a set $A \subset \mathbb{R}^n$ is the closure of the set of all pairs $(x, r) \in \mathbb{R}^n \times \mathbb{R}^+$ such that $B_r(x)$ is a maximal inscribed ball in A. The **medial axis** refers to the set of positions $x \in \mathbb{R}^n$ that are centers of maximal inscribed balls in A, i.e., the medial axis is the image of the medial locus under the projection $\pi : \mathbb{R}^n \times \mathbb{R}^+ \to \mathbb{R}^n$.[3]

Several authors have used the terms medial locus, medial axis, symmetry axis, and skeleton to mean either the medial positions or the medial position and radius tuples. The word "axis" is somewhat misleading (but it has stuck) since it connotes a straight line. However, as discussed below, the medial axis can have higher dimensions than a line and can be curved. The above definition of the medial locus is valid for any set $A \subset \mathbb{R}^n$. However, for the real-world objects that are found in images it is convenient to narrow the possible sets that can be considered. This leads to the following definition:

[3]The term medial locus is also used for other related skeletal structures. Here the terms "medial locus" and "medial axis" will always refer to this definition given by Blum.

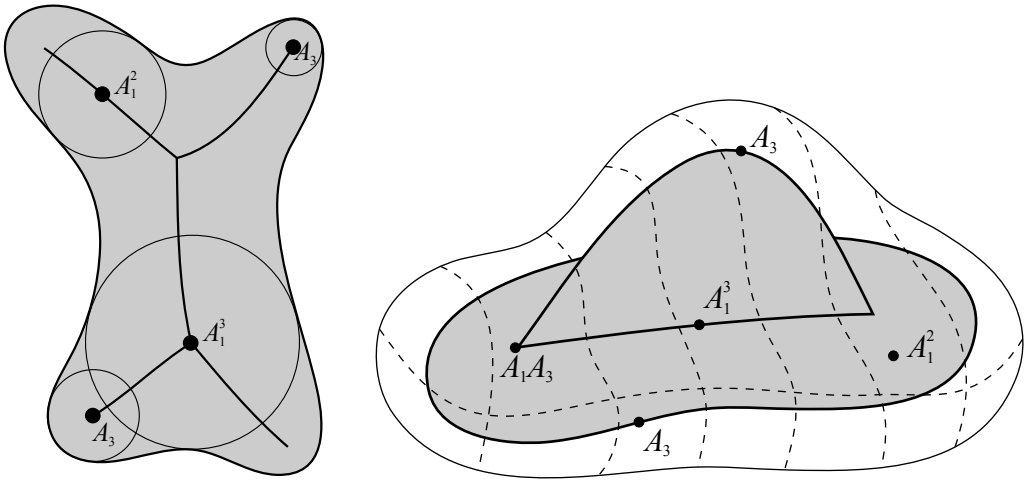

Figure 3.4: A 2D object and its medial axis (left). The medial axis of a 3D object (right). The different generic medial points are labelled (all types except the A_1^4 are present).

Definition 3.3. An **object** in $\mathbb{R}^n$ is a connected, compact, imbedded, n-dimensional manifold with boundary.

The compactness assures that an object is a bounded region, and the fact that an object is imbedded means that it cannot intersect itself in any way. Of course objects from images will be either 2D or 3D. An object does not have to have a smooth boundary, so, for example, a 2D region bounded by a polygon or a 3D region bounded by a closed polygonal surface is an object. Also, an object does not have to be simply connected, that is, it can have holes like an annulus in 2D or a solid doughnut in 3D.

The medial axis forms a structure known as a **stratified space**. It consists of a collection of smooth manifolds of different dimensions known as **strata**. Medial points that are tangent to the boundary in exactly two points make up a codimension 1 stratum. This is referred to as the **smooth part** of the medial axis. For example, in 2D the smooth parts of the medial axis are smooth curves, and in 3D the smooth parts of the medial axis are smooth surfaces. Each connected piece of the smooth part of the medial axis is called a **branch**. The remaining parts of the medial axis are the **singular parts**, which form lower-dimensional strata. They consist of boundaries of the branches and places where the branches connect. The types of generic points[4] on the medial axis

[4]The term "generic" refers to points or features of geometry that are stable under perturbations of that geometry. In this case a medial point is generic if it does not disappear under a small perturbation of the object's boundary.

have been classified in 3D by Giblin and Kimia [45]. A point on the medial axis is classified as an A_k^m point when the resulting inscribed sphere is tangent at m distinct points and the sphere has order k contact with the object boundary. No superscript indicates the sphere has contact at a single point. The types of medial points that can occur generically in 2D and 3D are (see Fig. 3.4)

1. A_1^2 points are the smooth parts of the medial axis, where the sphere is tangent at two distinct points. Points of this type form curves in 2D and surfaces in 3D.

2. A_1^3 points are where two branches of the medial axis meet, and the sphere is tangent at three distinct points on the boundary. Points of this type form points in 2D and curves in 3D.

3. A_3 points are the edges of the medial branches, where the sphere is tangent at a single point and has the same radius of curvature as the boundary at that point. The boundary also has a maximum of curvature at that point. Points of this type form points in 2D and curves in 3D.

4. A_1A_3 points are where a branch curve (A_1^3) meets an edge curve (A_3) in 3D. These points do not occur in 2D, and they form points in 3D.

5. A_1^4 points are where four branch curves (A_1^3) meet in 3D. These points do not occur in 2D, and they form points in 3D.

In contrast to boundary representations, which sample the boundary of an object, medial representations sample the medial locus of an object. These medial samples, called **medial atoms**, come in two different varieties.

Definition 3.4. An n-dimensional **order 0 medial atom** is a pair $(x, r) \in \mathbb{R}^n \times \mathbb{R}^+$.

An order 0 medial atoms represents the position and radius of a maximal inscribed ball at a location on the medial axis of an object. The phrase "order 0" is to distinguish these atoms from atoms with higher order information, i.e., derivatives of position and radius. The disadvantage of the order 0 medial atom is that it does not give enough information to reconstruct the corresponding boundary points, i.e., the points tangent to the sphere defined by the medial atom. This is remedied by adding first order information to the medial atom.

Definition 3.5. An n-dimensional **order 1 medial atom** is a tuple $(x, r, n_0, n_1) \in \mathbb{R}^n \times \mathbb{R}^+ \times S^{n-1} \times S^{n-1}$.

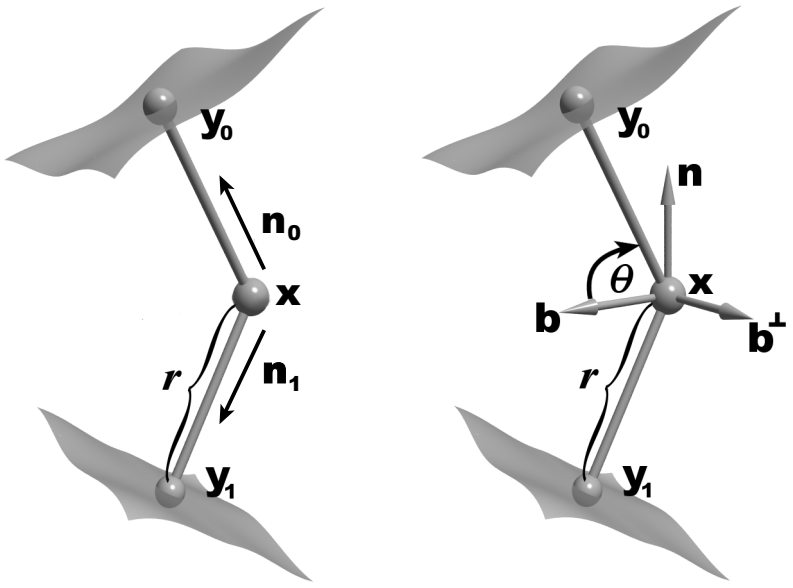

Figure 3.5: Two representations of a 3D order 1 medial atom and the portions of the implied boundaries associated with them. A medial atom as in Definition 3.5 as a position, radius, and two spoke directions (left). The same medial atom as a position, radius, frame, and object angle (right).

The order 1 medial atom (see Fig. 3.5) adds two unit length vectors, n_0, n_1, thought of as two points on the unit sphere S^{n-1}. These two points represent the tangency points of the boundary with the inscribed sphere. The vectors pointing from the medial locus position to the object boundary, called **spokes**, are given by rn_0 and rn_1. Therefore, order 1 medial atoms give enough information to reconstruct the corresponding boundary points on the object, y_0, y_1, given by the formulas

$$y_0 = x + rn_0, \qquad y_1 = x + rn_1. \tag{3.7}$$

The order 1 medial atom assumes that the inscribed sphere is bitangent to the object boundary. Thus, they are valid as samples of the smooth parts of the medial locus. Order 1 medial atoms encode the derivative information of the medial locus in a non-obvious way. Given an object in $\mathbb{R}^n$, let M be a branch of the object's medial axis, i.e., a smooth manifold in $\mathbb{R}^n$ of codimension 1, and let $r : M \to \mathbb{R}^+$ be the radius function on that branch. An order 1 medial atom at a point $x \in M$ is given by the tuple $(x, r, n_0, n_1) \in \mathbb{R}^n \times \mathbb{R}^+ \times S^{n-1} \times S^{n-1}$. Then the positional derivative information at $x \in M$ is given by the tangent space T_xM. This tangent space also has codimension 1, and as such it is uniquely determined by a single normal vector. This (unit) normal

vector is given by

$$n = \frac{n_0 - n_1}{\|n_0 - n_1\|}.$$

The derivative information of the radius function comes in the form of the gradient vector $\nabla r \in T_x M$. This gradient is given by the projection of the n_0 vector (or, equivalently, the n_1 vector) into the tangent space:

$$\nabla r = n_0 - \langle n_0, n \rangle \, n = n_1 - \langle n_1, n \rangle \, n.$$

Medial loci have the property that the gradient of the radius function in the smooth strata satisfies the inequality $\|\nabla r\| \leq 1$. With this restriction it can be seen that an order 1 medial atom, defined as the tuple (x, r, n_0, n_1), uniquely encodes the derivative information of the medial locus at the point x.

One contribution of this dissertation is the specification of an order 1 medial atom as defined above (Definition 3.5). In previous papers [66, 101, 102] order 1 medial atoms were given as a tuple $(x, r, F, \theta) \in \mathbb{R}^3 \times \mathbb{R}^+ \times SO(n) \times [0, \pi/2)$ (see Fig. 3.5). In other words, the two unit vectors, n_0, n_1, were replaced with a frame $F \in SO(n)$ and an angle $\theta \in [0, \pi/2)$, known as the **object angle**. In 3D the frame is given by three orthonormal vectors $\{\mathbf{b}, \mathbf{n}, \mathbf{b}^\perp\}$, where $\mathbf{b}$ is the unit bisector of the spokes, $\mathbf{n}$ is the unit normal to the medial axis, and $\mathbf{b}^\perp = \mathbf{b} \times \mathbf{n}$. The unit spoke directions can be derived from this representation as

$$n_0 = \cos(\theta)\mathbf{b} + \sin(\theta)\mathbf{n} \qquad n_1 = \cos(\theta)\mathbf{b} - \sin(\theta)\mathbf{n}.$$

The advantages of the order 1 medial atom as defined in this dissertation over the previous frame and object angle representation will be discussed in Chapter 5.

3.3.2 M-reps

Two major contributions of this dissertation are 1) a new method for studying statistical shape variability using medial representations and 2) application of this method to a deformable models approach to 3D medical image segmentation. Both the medial representation of object geometry and the resulting deformable models framework that are used in this dissertation are due to Pizer et al. [101, 102]. These medial representations, or **m-reps**, are described in this section. After an introduction to the benefits of m-reps as models of object shape, a description of the medial representation and data structure is given. Methods for interpolating smooth boundaries and figural-based coordinate sys-

tems from m-rep models are outlined. Then a different twist on medial representations called spline-based m-reps is reviewed. The section wraps up with a presentation of the deformable m-reps approach to 3D image segmentation.

The main selling points of the m-rep method as an approach to object geometry and deformable models are

1. *M-reps are multiscale.* They decompose the geometry of object collections in a coarse-to-fine manner. Multiscale approaches to deformable models have proven to be more robust to image problems such as noise, aliasing, and missing data. Also, multiscale methods are capable of extending the capture range of optimization procedures and increasing the rate of convergence.

2. *M-reps have a fuzzy boundary.* That is, they have a built-in boundary tolerance that allows them to extract fine-scale boundary perturbations without creating extra branches in the medial locus.

3. *M-reps provide a figural coordinate system.* This coordinate system measures distances along the medial directions of a figure and through the figure.

4. *M-reps are a solid representation.* Instead of just modeling the boundary, or shell, of an object, m-reps also model the interior (and just outside the object). This gives a means for indexing image values inside an object and also for modeling the physical properties of the object interior.

5. *M-reps directly model the medial locus of an object.* Methods that extract the medial locus from a boundary can be time-consuming and sensitive to perturbations in the boundary. Also, comparing the medial structure of two similar objects is easier when the medial locus is modeled directly because the medial branching can be kept consistent between objects.

The medial representation decomposes complex objects into a set of **figures**, which are slabs with unbranching medial locus. Each figure consists of a sheet of order 1 medial atoms. Single figures in 2D consist of 1D curve segment of atoms, and single figures in 3D consist of a 2D surface-with-boundary of atoms. The objects considered in this dissertation are all 3D single figure models (see Fig. 3.6), and they are the focus of this review. However, more complex models, i.e., models consisting of multiple figures and models consisting of collections of objects, are briefly reviewed.

Recall that an atom on the edge of the continuous medial locus has a single spoke with third order contact with the object boundary, i.e., an A_3 point. However, a single

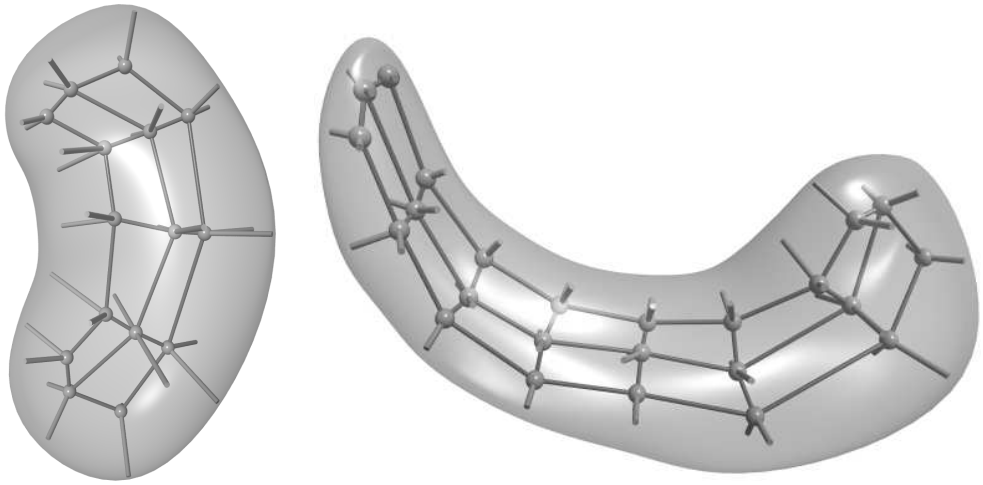

Figure 3.6: Two single figure m-rep models: a kidney (left) and a hippocampus (right).

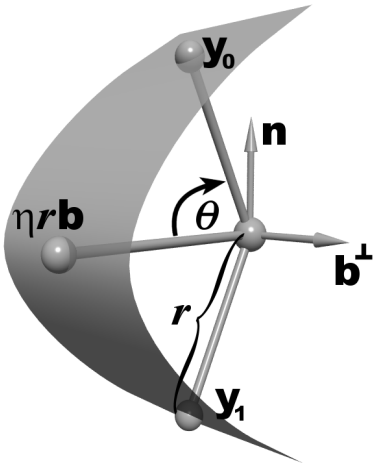

Figure 3.7: A 3D medial end atom, showing the portion of the boundary crest implied by the atom.

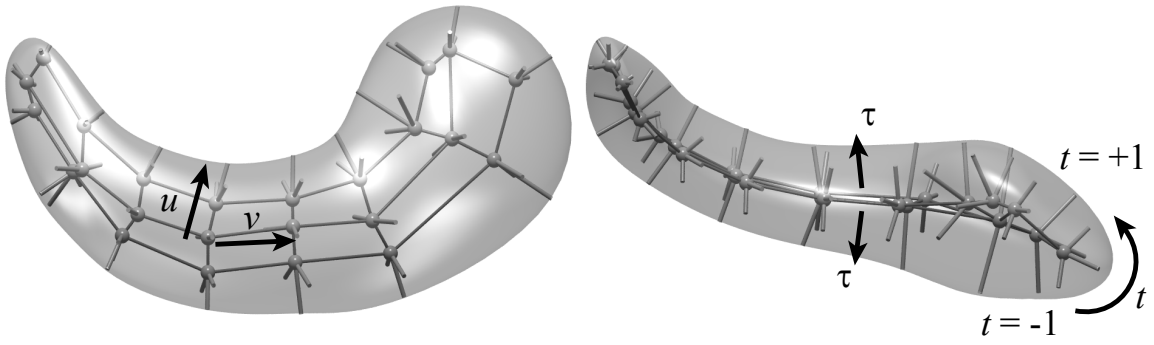

Figure 3.8: The figural coordinate directions (u, v, t, τ) demonstrated on an m-rep model of the hippocampus. Sample order 1 medial atoms on the sheet are also shown.

point of contact can be an unstable feature for image analysis tasks. The representation can be stabilized by restricting each edge point of the medial locus to lie along the **b** vector of a medial atom shifted back from the edge curve. An **end atom** (see Fig. 3.7) is a special type of order 1 medial atom that models this atom shifted back from the edge of the medial locus. It has an extra spoke in the bisector direction, **b**, along which the true edge of the medial locus lies. This extra spoke points to the crest of the implied boundary and has length ηr, where η is a parameter in the interval $[1, 1/\cos(\theta)]$. A value of $\eta = 1$ gives a circular end cap, while at the other extreme a value of $\eta = 1/\cos(\theta)$ gives a sharp corner.

Figural Coordinates

An m-rep sheet should be thought of as representing a continuous branch of medial atoms with associated continuous implied boundary. This continuous sheet of medial atoms can be parameterized by two real parameters (u, v). The choice of this parameterization may depend on the need to make comparisons at corresponding points between similar objects. In this case parameterizations that are in one-to-one correspondence are chosen. This correspondence can be based on geometric properties of the objects, or it can be chosen with the desire to build optimal statistical models of a population. A full discussion of these correspondence issues is beyond the scope of this review, and it is assumed from here on that some (u, v) parameterization for the medial locus is given. Since each internal medial atom in a single figure implies two boundary points, an extra parameter $t \in \{-1, 1\}$ can be added to extend the medial coordinates to a parameterization (u, v, t) of the implied boundary.

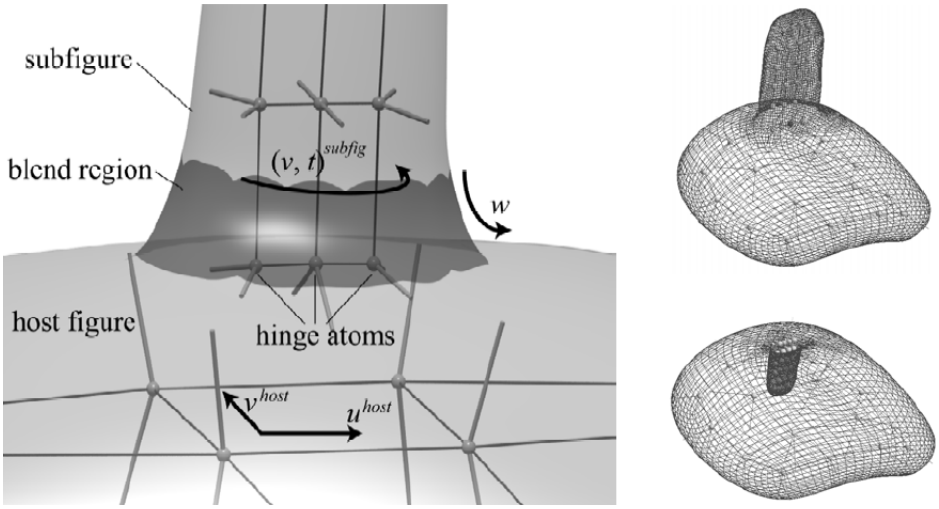

Figure 3.9: Left: the hinge arrangement of a subfigure with the subfigure on top and parent figure on bottom. Right top: a protrusion subfigure. Right bottom: an indentation subfigure. (This figure is courtesy of Qiong Han and appears in [55].)

The figural coordinates further extend the implied boundary coordinates to a parameterization of the space inside and just outside the m-rep figure. A **figural coordinate** (see Fig. 3.8) is a tuple (u, v, t, τ), where the τ parameter gives the r-proportional signed distance of the point in question from the surface point at (u, v, t). That is, τ is given as the signed distance along the normal to the surface at (u, v, t) divided by the r value at (u, v, t). This coordinate system is valid inside the entire solid represented by the m-rep figure (i.e., each point has a unique coordinate). It is also valid outside the figure's boundary up to the exterior shock set of the distance function to the boundary. Therefore, it is valid for all of $\mathbb{R}^3$ when the figure is convex. An important feature of the figural coordinate system is that the coordinates are invariant under translations, rotations, and scalings of the object. This makes the figural coordinate system an ideal parameterization when dealing with shape properties. Also, if one object is represented as a deformation of the medial representation of another object, the figural coordinates of the two objects are in one-to-one correspondence. This is useful, for example, during segmentation in comparing image intensity values of the target image with the image intensity values of a training image with respect to the current m-rep object.

Multi-figure Objects and Multi-object Complexes

For objects consisting of multiple figures (see Fig. 3.9), the object's figures are arranged in a hierarchical fashion, i.e., a tree or directed acyclic graph (DAG). A parent figure in the tree represents a more substantial part of the object, and a child figure represents a protrusion or indentation of its parent figure. Protrusions are figures that add material to an object, while indentations are figures that subtract material from an object. To make the representation focus on the portions of the object with the major internal substance, a child figure, also called a subfigure, is attached to its parent by a curve segment of its edge curve, called a **hinge**. The resulting implied boundary subdivision surfaces are blended using a method described in [55]. The figural coordinates of the subfigure arrangement include the figural coordinates of the parent, the figural coordinates of the child, and an extra blend coordinate w that parameterizes the blend area between the parent and child. More details about the geometry and segmentation process for multi-figure models can be found in [55].

Sometimes one wishes to represent and then segment multiple disconnected objects at the same time. An example is the cerebral ventricles, hippocampus, and caudate in which the structures are related but one is not a protrusion or an indentation on another. Another example are the pair of kidneys and the liver. In the m-reps system these can be connected by one or more links between the representations of the respective objects, allowing the position of one figure to predict boundary positions of the other. This matter is explained in detail in a paper covering the segmentation of multi-object complexes [43].

Mesh-Based M-reps

In 3D a single figure object can be represented by a quadrilateral mesh m_{ij} of order 1 medial atoms (see Fig. 3.6). Atoms on the edge of the mesh are represented by end atoms with three spokes as described above. The atoms in an m-rep mesh can be thought of as control points implying a full continuous sheet of order 1 medial atoms. The continuous medial locus extends beyond the end atoms to the curve of A_3 atoms osculating the crest of the implied object's boundary. In multi-figure models the hinge curve of a subfigure is represented as one edge of the subfigure's quadrilateral mesh of medial atoms.

The implied boundary of an m-rep figure is interpolated from the boundary points (y_0, y_1) and corresponding normals (n_0, n_1) implied by the medial atoms. This also

includes the crest points implied by the third spokes of the end atoms. The surface interpolation used is due to Thall [123, 124] and is an extension of the Catmull-Clark subdivision method [22]. As a result of this interpolation, each boundary point can be associated a boundary figural coordinate (u, v, t), where the parameter $t \in [-1, 1]$ is equal to either -1 or 1 for internal atoms to distinguish between the two spoke ends. At end atoms the t parameter transitions continuously from -1 to 1 around the crest.

Spline-Based M-reps

Yushkevich *et al.* [131,132] describe a medial representation built on a continuous spline model of the medial locus of an object. This in turn implies a continuous boundary of the object as well as a parameterization of the interior of the object. Although spline-based m-reps have been defined in both 2D and 3D, this review concentrates on the 3D case. The medial locus is parameterized as a pair of continuous functions $(x(u, v), r(u, v))$, where x is the medial position and r is the associated radius field. These functions are represented as b-spline surfaces. They are determined by control points $(x_{ij}, r_{ij}) : 0 \leq i \leq d_1,\ 0 \leq j \leq d_2$ and given by the b-spline formulas

$$x(u, v) = \sum_{i=0}^{d_1} \sum_{j=0}^{d_2} N_i^3(u) N_j^3(v) x_i j,$$
$$r(u, v) = \sum_{i=0}^{d_1} \sum_{j=0}^{d_2} N_i^3(u) N_j^3(v) r_i j,$$

where N_i^3 are third-order b-spline basis functions (see [38]).

The control points (x_{ij}, r_{ij}) can be thought of as order 0 medial atoms. However, order 1 medial atoms can be produced at each point in the continuous medial locus by using the first partial derivatives x_u, x_v, r_u, r_v of the b-spline functions. The spoke directions of the order 1 medial atoms are given by the functions

$$n_0 = -\nabla r + \sqrt{1 - \|\nabla r\|^2}\, n, \qquad n_1 = -\nabla r - \sqrt{1 - \|\nabla r\|^2}\, n,$$

where $n = x_u \times x_v / \|x_u \times x_v\|$ is the unit surface normal to x. The gradient of the radius is given by the formula

$$\nabla r = [x_u x_v] I^{-1} \begin{bmatrix} r_u \\ r_v \end{bmatrix},$$

where I is the metric tensor on the surface x, i.e.,

$$I = \begin{bmatrix} \langle x_u, x_u \rangle & \langle x_u, x_v \rangle \\ \langle x_u, x_v \rangle & \langle x_v, x_v \rangle \end{bmatrix}.$$

The b-spline medial locus must satisfy several constraints to ensure that it implies a valid, non-folding boundary surfaces. First, it must satisfy the constraint that $\|\nabla r\| < 1$ in the interior of the medial sheet. Second, the implied boundary must be constrained to not crease or fold by ensuring that the Jacobian of the medial-to-boundary mapping remains positive. Finally, the medial locus must be a manifold with boundary, where the edge curve of A_3 medial atoms satisfies the constraint $\|\nabla r\| = 1$. This last condition is achieved by setting the edge of the control point grid to large negative radii, while the interior control points all have positive radii. This causes the level curve of $\|\nabla r\| = 1$ to lie within the b-spline surface. The surfaces is then trimmed along this curve, resulting in the desired edge for the medial sheet.

Segmentation via Deformable M-reps

Following the deformable models paradigm, a 3D m-rep model $\mathbf{M}$ is deformed into an image $I(x, y, z)$ by optimizing an objective function, which is defined as

$$F(\mathbf{M}, I) = L(\mathbf{M}, I) + \alpha\, G(\mathbf{M}).$$

The function L, the *image match*, measures how well the model matches the image information, while G, the *geometric typicality*, gives a prior on the possible variation of the geometry of the model. The relative importance of the two terms is weighted by the non-negative real parameter α. The segmentation strategy described in this review is from Pizer *et al.* [102] and also developed in previous papers [101, 66]. One of the main contribution of this thesis, described later in Chapter 5, builds on this segmentation strategy by incorporating geometric statistics. This includes using a statistical geometric prior for the geometric typicality and using figural deformations based on the statistical modes of variation.

This objective function is optimized in a multiscale fashion. That is, it is optimized over a sequence of transformations that are successively finer in scale. In this review only segmentation of single figures is considered, which includes three levels of scale: the figural level, the medial atom level, and the dense boundary sampling level. At each scale level the deformations are defined as transformations of the current primitives,

either figures, medial atoms, or boundary points. The model is first initialized by the user placing a template model into the image I using a global translation, rotation, and scaling. At the figural level the transformation used is a similarity transformation plus an elongation of the entire figure. More generally, the figural stage transformation can be any operation that acts globally on the figure, such as an atom transformation applied to each atom in the figure. At the atom level each medial atom is independently transformed by a translation of the medial position, a 3D rotation of the frame, a scaling of the radius, and a rotation of the object angle. In the boundary stage each boundary point is displaced along its corresponding normal direction.

The computation of the image match term in the objective function is based on a a template model $\hat{\mathbf{M}}$. Image values in a template image $\hat{I}$ at a particular figural coordinate of the template model are compared to image values in the target image I at the corresponding figural coordinate of the candidate model. The image match term of the objective function is computed as a correlation over a collar ($\pm\epsilon$ in the normal direction) about the object boundary:

$$L(\mathbf{M}, I) = \int_{\mathcal{B}} \int_{-\epsilon}^{\epsilon} G(t)\hat{I}\left(\hat{\mathbf{s}} + (t/\hat{r})\hat{\mathbf{n}}\right)\, I\left(\mathbf{s} + (t/r)\mathbf{n}\right)\, dt\, dw.$$

In this equation a hat ($\hat{}$) always denotes an entity associated with the template model $\hat{\mathbf{M}}$, and the same entities without a hat are associated with the candidate model $\mathbf{M}$. The parameter $w = (u, v, t)$ is a figural boundary coordinate, $\mathcal{B}$ is the parameter domain of the boundary coordinates. The following are functions of the boundary figural coordinate w: $\mathbf{s}, \hat{\mathbf{s}}$ are parameterizations of the boundaries, $r, \hat{r}$ are the radius functions, and $\mathbf{n}, \hat{\mathbf{n}}$ are the boundary normals. The function G_σ is the Gaussian kernel with standard deviation σ. The Gaussian kernel is used to weight the importance of the image match so that features closer to the boundary are given higher weight. The values for the collar width and Gaussian standard deviation have been set by experience to $\epsilon = 0.3$ and $\sigma = 0.15$.

The geometric typicality term G consists of two terms. Each term is computed using r-proportional squared distances. The first term, denoted by P, measures the total amount of change in the object boundary during the current stage. The second term, denoted by N, measures the difference between the boundary of the candidate and the boundary of the candidate replacing the current primitive with the prediction of its neighbors. This neighbor term enforces a local consistency between model primitives.

The geometric typicality term is defined as

$$G(\mathbf{M}) = (1-\beta)\, P(\mathbf{M}) + \beta\, N(\mathbf{M}),$$

where $\beta \in [0,1]$ is a weighting term. The function P measures the change in the boundary from the previous level of scale in r-proportional terms:

$$P(\mathbf{M}) = -\int_{\mathcal{B}(\mathbf{M})} \frac{||\mathbf{s}-\mathbf{s}_0||^2}{r^2} d\mathbf{s},$$

where $\mathbf{s}_0$ is the initial position of the boundary at this scale level. The function N seeks to keep primitives in the same relationship with their neighboring primitives. It is defined as

$$N(\mathbf{M}) = -\int_{\mathcal{B}(\mathbf{M})} \frac{||\mathbf{s}-\mathbf{s}'||^2}{r^2} d\mathbf{s},$$

where now $\mathbf{s}'$ is the boundary surface of the model in which the current primitive is in the position predicted by its neighbors. For single-figure models there is no neighbor primitive at the figural stage. Therefore, the neighbor term is zero, i.e., $\beta = 0$, during the figural level. For the atom scale level each medial atom's neighbors are the adjacent atoms in the grid (4 neighbors for internal atoms, 3 for edge atoms, and 2 for the corner atoms). The neighbor term at the boundary scale level comes from comparing a boundary point to the prediction by its neighbors in the boundary mesh. This prediction is given by an average of the neighboring points.

3.4 Diffusion Tensor Imaging

The statistical methods introduced in this dissertation are shown in Chapter 6 to have application in the statistical analysis of diffusion tensor images. This section is a review of diffusion tensor imaging. It begins with a description of diffusion tensor imaging and Brownian motion. Several quantitative measures derived from diffusion tensors are then discussed. Finally, the clinical applicability of diffusion tensor imaging is reviewed, as well as research issues such as regularization, fiber tracking, and statistical studies.

Diffusion tensor magnetic resonance imaging (DT-MRI), developed at NIH by Peter Basser *et al.* [3], measures the random 3D motion of water molecules, i.e., the diffusion of water. It produces a 3D diffusion tensor, that is, a 3×3, symmetric, positive-definite matrix, at each voxel of an 3D imaging volume. Recall that a matrix A is symmetric if $A = A^T$, and it is positive-definite if $x^T A x > 0$ for all nonzero vectors x. This tensor

is the covariance in a Brownian motion model of the diffusion of water at that voxel. In homogeneous materials water tends to diffuse isotropically, that is, equally in all directions. For example, if a drop of dye is placed in water, it will slowly spread out equally in all directions. However, in fibrous materials, such as skeletal muscle, cardiac muscle, and brain white matter, water molecules tend to diffuse faster in the directions parallel to the fibers and slower in the directions perpendicular to the fibers. Therefore, DT-MRI can give important information about the microstructure of fibrous tissues in the body. In brain imaging DT-MRI is used to track white matter fibers and establish connectivity properties of the brain.

Brownian motion was first discovered by the biologist Robert Brown in 1827 [16]. He noticed that small pollen particles demonstrated a seemingly random, jittery, motion when suspended in water. However, a mathematical model of Brownian motion was not developed until 1905 by Albert Einstein [37]. Einstein showed that Wiener processes provide a reasonable model of Brownian motion (although more accurate and more complicated physical models appeared later). A Wiener process is a random process $w(t)$ of time t that satisfies the following two axioms:

1. $w(t) - w(s) \sim N(0, (t-s)\Sigma)$.

2. $w(t) - w(s)$ and $w(v) - w(u)$ are independent random variables for $0 \leq s \leq t \leq u \leq v$.

For DT-MRI the Wiener process $w(t)$ is a function giving the 3D position of a molecule under diffusion. The axioms state that the incremental motion of a particle is governed by a normal distribution at each point in space, and that the motion is independent of any previous movement of the particle. The covariance Σ is a 3×3 symmetric, positive-definite matrix, which can take different values at different locations in space. This covariance matrix is what is measured by DT-MRI at each voxel of an image.

The quantitative measurements derived from diffusion tensors that people have used can be roughly broken down into two categories: size and shape measurements. An important aspect of these measurements is that they should be independent of the laboratory coordinate system. That is, a derived measurement should be invariant to translation and rotation of the diffusion tensor. First of all, diffusion tensors themselves are invariant to translation (a translated diffusion tensor is the same tensor, just at a different point). Therefore, it suffices to consider only rotational invariance. If the coordinate system is rotated by a matrix $R \in SO(3)$, a diffusion tensor D will be

transformed to the tensor D' by the equation

$$D' = RDR^T.$$

The eigenvalues $\lambda_1, \lambda_2, \lambda_3$ of D are left invariant under this operation. Therefore, any combination of the eigenvalues is an invariant measurement of the tensor D. This leads to two measurements of the size of a diffusion tensor. The first size measurement, the **mean diffusivity** is given by the average of the eigenvalues,

$$\langle D \rangle = (1/3)(\lambda_1 + \lambda_2 + \lambda_3).$$

The second measurement of size is the determinant of the diffusion tensor, given by the product of its eigenvalues,

$$\det(D) = \lambda_1 \lambda_2 \lambda_3.$$

Measurements of the shape of a diffusion tensor depend on the relative magnitudes of the eigenvalues. There are two common anisotropy measures, which measure how far the diffusion tensor is from being isotropic. They are both based on the standard deviation of the eigenvalues of D, given by

$$\sigma(D) = (1/\sqrt{3})\sqrt{(\lambda_1 - \langle D \rangle)^2 + (\lambda_2 - \langle D \rangle)^2 + (\lambda_2 - \langle D \rangle)^2}.$$

The first anisotropy measure, known as the **relative anisotropy** (RA), is given by the ratio of the standard deviation of the eigenvalues of D with the average of the eigenvalues, that is,

$$RA(D) = \frac{\sigma(D)}{\langle D \rangle}.$$

The second anisotropy measure, known as the **fractional anisotropy** (FA), is similar to the RA, except the denominator is the magnitude of the tensor under the Frobenius matrix norm, $\|D\| = \sqrt{\lambda_1^2 + \lambda_2^2 + \lambda_3^2}$. The FA is given by

$$FA(D) = \frac{\sqrt{6}}{2} \frac{\sigma(D)}{\|D\|}.$$

Diffusion tensor imaging has shown promise in several clinical studies of the brain. See the paper by Le Bihan *et al.* [7] for a review. Ischemic areas of the brain, that is, areas with decreased blood supply, in stroke patients have demonstrated lower diffusivity [89, 110, 127]. Thus, DT-MRI could help doctors understand which areas of the brain

have been damaged and might be salvageable in the first hours after a stroke. DT-MRI has shown promise in diagnosing diseases such as multiple sclerosis [126, 130] and Alzheimer's [57, 56]. The health of the brain white matter can be assessed using the derived measures of the diffusion tensor. The diffusivity measures tell the overall content of water in the tissue, and the anisotropy measures indicate the health of the myelin fibers. Also, studies have shown that DT-MRI could be used to assess the growth and maturity of white matter in the newborn brain [93, 134].

From the perspective of the image analyst, diffusion tensor images present many new and interesting issues including visualization, regularization, fiber tracking, and statistical analysis of diffusion tensor data. A major challenge in these applications is that diffusion tensor images contain 6-dimensional tensors at each voxel rather than a single value per voxel found in other modalities. Several authors have addressed the problem of estimation and smoothing within a DT image [23, 28, 133].

Further insights might be had from the use of diffusion tensor imaging in intersubject studies. Statistical brain atlases have been used in the case of scalar images to quantify anatomical variability across patients. However, relatively little work has been done towards constructing statistical brain atlases from diffusion tensor images. Alexander *et al.* [1] describe a method for the registration of multiple DT images into a common coordinate frame, however, they do not include a statistical analysis of the diffusion tensor data. Previous attempts [4, 98] at statistical analysis of diffusion tensors within a DT image are based on a Gaussian model of the linear tensor coefficients.

Chapter 4

Manifold Statistics

This chapter[1] presents a novel framework for computing the statistical variability of data on general manifolds. Principal component analysis is a standard technique for describing the statistical variability of data in Euclidean space $\mathbb{R}^n$. The method presented in this chapter, called principal geodesic analysis (PGA), is a natural extension of principal component analysis to manifold-valued data.

In Section 4.1 we review existing definitions for the mean of manifold-valued data. The definition of the mean used in this work is intrinsic to the geometry of the manifold. In Section 4.2 we present principal geodesic analysis for describing the variability of data on manifolds. This is based on generalizing the definition of principal component analysis, using either the variance-maximizing or least-squares definition. We give an algorithm for computing principal geodesic analysis as well as an algorithm for efficiently approximating it. Finally, we demonstrate implementations of both the PGA and approximation to PGA algorithms on the sphere S^2.

4.1 Means on Manifolds

The first step in extending statistical methods to manifolds is to define the notion of a mean value. In this section we describe two different notions of means on manifolds called intrinsic and extrinsic means, and we argue that the intrinsic mean is a preferable definition. We then present a method for computing the intrinsic mean of a collection of data on a manifold. Throughout this section we consider only manifolds that are connected and have a complete Riemannian metric.

[1] The work presented in this chapter was done in collaboration with Dr. Sarang Joshi, Dr. Conglin Lu, and Dr. Stephen Pizer at the University of North Carolina. This chapter contains parts of the paper [42] and is also based on the previous papers [40, 41].

4.1.1 Intrinsic vs. Extrinsic Means

Given a set of points $x_1, \ldots, x_N \in \mathbb{R}^n$, the arithmetic mean $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ is the point that minimizes the sum-of-squared Euclidean distances to the given points, i.e.,

$$\bar{x} = \arg\min_{x \in \mathbb{R}^n} \sum_{i=1}^{N} ||x - x_i||^2.$$

Since a general manifold M may not form a vector space, the notion of an additive mean is not necessarily valid. However, like the Euclidean case, the mean of a set of points on M can be formulated as the point which minimizes the sum-of-squared distances to the given points. This formulation depends on the definition of distance. One way to define distance on M is to embed it in a Euclidean space and use the Euclidean distance between points. This notion of distance is extrinsic to M, that is, it depends on the ambient space and the choice of embedding. Given an embedding $\Phi : M \to \mathbb{R}^n$, define the *extrinsic mean* [53] of a collection of points $x_1, \ldots, x_N \in M$ as

$$\mu_\Phi = \arg\min_{x \in M} \sum_{i=1}^{N} ||\Phi(x) - \Phi(x_i)||^2.$$

Given the above embedding of M, we can also compute the arithmetic (Euclidean) mean of the embedded points and then project this mean onto the manifold M. This projected mean is equivalent to the above definition of the extrinsic mean (see [116]). Define a projection mapping $\pi : \mathbb{R}^n \to G$ as

$$\pi(x) = \arg\min_{y \in M} ||\Phi(y) - x||^2.$$

Then the extrinsic mean is given by

$$\mu_\Phi = \pi\Big(\frac{1}{N}\sum_{i=1}^{N} \Phi(x_i)\Big).$$

A more natural choice of distance is the Riemannian distance on M. This definition of distance depends only on the intrinsic geometry of M. We now define the *intrinsic mean* of a collection of points $x_1, \ldots, x_N \in M$ as the minimizer in M of the sum-of-

squared Riemannian distances to each point. Thus the intrinsic mean is

$$\mu = \arg\min_{x \in M} \sum_{i=1}^{N} d(x, x_i)^2, \tag{4.1}$$

where $d(\cdot, \cdot)$ denotes Riemannian distance on M. This is the definition of a mean value that we use in this paper.

The idea of an intrinsic mean goes back to Fréchet [44], who defines it for a general metric space. The properties of the intrinsic mean on a Riemannian manifold have been studied by Karcher [67]. Moakher [88] compares the properties of the intrinsic and extrinsic mean for the group of 3D rotations. Since the intrinsic mean is defined in (4.1) as a minimization problem, its existence and uniqueness are not ensured. However, Kendall [74] shows that the intrinsic mean exists and is unique if the data is well-localized.

We argue that the intrinsic mean definition is preferable over the extrinsic mean. The intrinsic mean is defined using only the intrinsic geometry of the manifold in question, that is, distances that are dependent only on the Riemannian metric of the manifold. The extrinsic mean depends on the geometry of the ambient space and the imbedding Φ. Also, the projection of the Euclidean average back onto the manifold may not be unique if the manifold has negative sectional curvatures.

4.1.2 Computing the Intrinsic Mean

Computation of the intrinsic mean involves solving the minimization problem in (4.1). We will assume that our data $x_1, \ldots, x_n \in M$ lies in a sufficiently small neighborhood so that a unique solution is guaranteed. We must minimize the sum-of-squared distance function

$$f(x) = \frac{1}{2N} \sum_{i=1}^{N} d(x, x_i)^2.$$

We now describe a gradient descent algorithm, first proposed by Pennec [100], for minimizing f. Using the assumption that the x_i lie in a strongly convex neighborhood, i.e., a neighborhood U such that any two points in U are connected by a unique geodesic contained completely within U, Karcher [67] shows that the gradient of f is

$$\nabla f(x) = -\frac{1}{N} \sum_{i=1}^{N} \mathrm{Log}_x(x_i).$$

The gradient descent algorithm takes successive steps in the negative gradient direction. Given a current estimate μ_j for the intrinsic mean, the equation for updating the mean by taking a step in the negative gradient direction is

$$\mu_{j+1} = \mathrm{Exp}_{\mu_j}\left(\frac{\tau}{N}\sum_{i=1}^{N}\mathrm{Log}_{\mu_j}(x_i)\right),$$

where τ is the step size.

Because the gradient descent algorithm only converges locally, care must be taken in the choices of the initial estimate of the mean μ_0 and the step size τ. Since the data is assumed to be well-localized, a reasonable choice for the initial estimate μ_0 is one of the data points, say x_1. The choice of τ is somewhat harder and depends on the manifold M. Buss and Fillmore [18] prove for data on spheres, a value of $\tau = 1$ is sufficient. Notice that if M is a vector space, the gradient descent algorithm with $\tau = 1$ is equivalent to linear averaging and thus converges in a single step. If $M = \mathbb{R}^+$, the Lie group of positive reals under multiplication, the algorithm with $\tau = 1$ is equivalent to the geometric average and again converges in a single step.

In summary we have the following algorithm for computing the intrinsic mean of manifold data:

Algorithm 4.1: Intrinsic Mean
Input: $x_1, \ldots, x_N \in M$
Output: $\mu \in M$, the intrinsic mean
$\mu_0 = x_1$
Do
$\Delta\mu = \frac{\tau}{N}\sum_{i=1}^{N}\mathrm{Log}_{\mu_j} x_i$
$\mu_{j+1} = \mathrm{Exp}_{\mu_j}(\Delta\mu)$
While $||\Delta\mu|| > \epsilon$.

4.2 Principal Geodesic Analysis

Although averaging methods on manifolds have previously been studied, principal component analysis has not been developed for manifolds. We present a new method called **principal geodesic analysis** (PGA), a generalization of principal component analysis to manifolds. We start with a review of PCA in Euclidean space. Consider a set of points $x_1, \ldots, x_N \in \mathbb{R}^n$ with zero mean. Principal component analysis seeks a sequence of linear subspaces that best represent the variability of the data. To be more precise,

the intent is to find an orthonormal basis $\{v_1, \ldots, v_n\}$ of $\mathbb{R}^n$, which satisfies the recursive relationship

$$v_1 = \arg\max_{||v||=1} \sum_{i=1}^{N} (v \cdot x_i)^2, \tag{4.2}$$

$$v_k = \arg\max_{||v||=1} \sum_{i=1}^{N} \sum_{j=1}^{k-1} (v_j \cdot x_i)^2 + (v \cdot x_i)^2. \tag{4.3}$$

In other words, the subspace $V_k = \text{span}(\{v_1, \ldots, v_k\})$ is the k-dimensional subspace that maximizes the variance of the data projected to that subspace. The basis $\{v_k\}$ is computed as the set of ordered eigenvectors of the sample covariance matrix of the data.

Now turning to manifolds, consider a set of points $x_1, \ldots, x_N$ on a manifold M. Our goal is to describe the variability of the x_i in a way that is analogous to PCA. Thus we will project the data onto lower-dimensional subspaces that best represent the variability of the data. This requires first extending three important concepts of PCA into the manifold setting:

- **Variance.** Following the work of Fréchet, we define the sample variance of the data as the expected value of the squared Riemannian distance from the mean.
- **Geodesic subspaces.** The lower-dimensional subspaces in PCA are linear subspaces. For general manifolds we extend the concept of a linear subspace to that of a **geodesic submanifold**.
- **Projection.** In PCA the data is projected onto linear subspaces. We define a projection operator for geodesic submanifolds, and show how it may be efficiently approximated.

We now develop each of these concepts in detail.

4.2.1 Variance

The variance σ^2 of a real-valued random variable x with mean μ is given by the formula

$$\sigma^2 = \mathcal{E}[(x - \mu)^2],$$

where $\mathcal{E}$ denotes expectation. It measures the expected localization of the variable x about the mean. When dealing with a vector-valued random variable $\mathbf{x}$ in $\mathbb{R}^n$ with mean

μ, the variance is replaced by a covariance matrix

$$\Sigma = \mathcal{E}[(\mathbf{x} - \mu)(\mathbf{x} - \mu)^T].$$

However, this definition is not valid for general manifolds again since vector space operations do not exist for such spaces.

The definition of variance we use comes from Fréchet [44], who defines the variance of a random variable in a metric space as the expected value of the squared distance from the mean. That is, for a random variable x in a metric space with intrinsic mean μ, the variance is given by

$$\sigma^2 = \mathcal{E}[d(\mu, x)^2].$$

Thus given data points $x_1, \ldots, x_N$ on a complete, connected manifold M, we define the sample variance of the data as

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} d(\mu, x_i)^2 = \frac{1}{N}\sum_{i=1}^{N} ||\operatorname{Log}_\mu(x_i)||^2, \tag{4.4}$$

where μ is the intrinsic mean of the x_i.

If M is a vector space, the variance definition in (4.4) is given by the trace of the sample covariance matrix, i.e., the sum of its eigenvalues. It is in this sense that this definition captures the total variation of the data.

4.2.2 Geodesic Submanifolds

The next step in generalizing PCA to manifolds is to generalize the notion of a linear subspace. A geodesic is a curve that is locally the shortest path between points. In this way a geodesic is the generalization of a straight line. Thus it is natural to use a geodesic curve as the one-dimensional subspace that provides the analog of the first principal direction in PCA.

In general if N is a submanifold of a manifold M, geodesics of N are not necessarily geodesics of M. For instance the sphere S^2 is a submanifold of $\mathbb{R}^3$, but its geodesics are great circles, while geodesics of $\mathbb{R}^3$ are straight lines. A submanifold H of M is said to be geodesic at $x \in H$ if all geodesics of H passing through x are also geodesics of M. For example a linear subspace of $\mathbb{R}^n$ is a submanifold geodesic at 0. Submanifolds geodesic at x preserve distances to x. This is an essential property for PGA because variance is defined as the average squared distance to the mean. Thus submanifolds geodesic at

the mean will be the generalization of the linear subspaces of PCA.

4.2.3 Projection

The projection of a point $x \in M$ onto a geodesic submanifold H of M is defined as the point on H that is nearest to x in Riemannian distance. Thus we define the projection operator $\pi_H : M \to H$ as

$$\pi_H(x) = \arg\min_{y \in H} d(x, y)^2. \tag{4.5}$$

Since projection is defined by a minimization, there is no guarantee that the projection of a point exists or that it is unique. However, by restricting to a small enough neighborhood about the mean, we can be assured that projection is unique for any submanifold geodesic at the mean.

4.2.4 Defining Principal Geodesic Analysis

We are now ready to define principal geodesic analysis for data $x_1, \ldots, x_N$ on a connected, complete manifold M. Our goal, analogous to PCA, is to find a sequence of nested geodesic submanifolds that maximize the projected variance of the data. These submanifolds are called the **principal geodesic submanifolds**.

Let $T_\mu M$ denote the tangent space of M at the intrinsic mean μ of the x_i. Let $U \subset T_\mu M$ be a neighborhood of 0 such that projection is well-defined for all geodesic submanifolds of $\mathrm{Exp}_\mu(U)$. We assume that the data is localized enough to lie within such a neighborhood. The principal geodesic submanifolds are defined by first constructing an orthonormal basis of tangent vectors $v_1, \ldots, v_n \in T_\mu M$ that span the tangent space $T_\mu M$. These vectors are then used to form a sequence of nested subspaces $V_k = \mathrm{span}(\{v_1, \ldots, v_k\}) \cap U$. The principal geodesic submanifolds are the images of the V_k under the exponential map: $H_k = \mathrm{Exp}_\mu(V_k)$. The first principal direction is chosen to maximize the projected variance along the corresponding geodesic:

$$v_1 = \arg\max_{||v||=1} \sum_{i=1}^{N} ||\,\mathrm{Log}_\mu(\pi_H(x_i))||^2, \tag{4.6}$$
$$\text{where} \quad H = \mathrm{Exp}_\mu(\mathrm{span}(\{v\}) \cap U).$$

The remaining principal directions are defined recursively as

$$v_k = \arg\max_{||v||=1} \sum_{i=1}^{N} ||\operatorname{Log}_\mu(\pi_H(x_i))||^2, \tag{4.7}$$
$$\text{where} \quad H = \operatorname{Exp}_\mu(\operatorname{span}(\{v_1, \ldots, v_{k-1}, v\}) \cap U).$$

4.2.5 An Alternative Definition of PGA

Recall from Section 3.1.3 that principal component analysis may be defined in two different ways, both giving the same end result. Thus far, we have based the definition of principal geodesic analysis on generalizing the variance maximization approach to PCA. In this section we describe an alternative definition of PGA based on generalizing the other approach to PCA, namely, the least-squares approach.

The least-squares approach to PCA of a collection of data $x_1, \ldots, x_N \in \mathbb{R}^n$ seeks a sequence of linear subspaces that are closest to the data in a least-squares sense. These subspaces are generated from an orthonormal basis $\{v_1, \ldots, v_n\}$ of $\mathbb{R}^n$, which satisfies the recursive relationship

$$v_1 = \arg\min_{\|v\|=1} \sum_{i=1}^{N} \|x_i - \langle v, x_i \rangle\, v\|^2,$$
$$v_k = \arg\min_{\|v\|=1} \sum_{i=1}^{N} \left\| x_i - \sum_{j=1}^{k-1} \langle v_j, x_i \rangle\, v_j + \langle v, x_i \rangle\, v \right\|^2.$$

In other words, the subspace $V_k = \operatorname{span}(\{v_1, \ldots, v_k\})$ is the k-dimensional subspace that minimizes the sum-of-squared distances to the data.

We now want to define principal geodesic analysis of data $x_1, \ldots, x_N$ on a manifold M by generalizing this least-squares approach. The least-squares distance is defined using geodesic distances on the manifold. Using the same notation as in the previous subsection, we define principal geodesic submanifolds via subspaces $V_k = \operatorname{span}(\{v_1, \ldots, v_k\})$ of the tangent space $T_\mu M$. The principal geodesic submanifolds are again given by $H_k = \operatorname{Exp}_\mu(V_k)$. The first principal direction is now chosen to minimize the sum-of-squared distance of the data to the corresponding geodesic:

$$v_1 = \arg\max_{||v||=1} \sum_{i=1}^{N} ||\operatorname{Log}_{x_i}(\pi_H(x_i))||^2,$$
$$\text{where} \quad H = \operatorname{Exp}_\mu(\operatorname{span}(\{v\}) \cap U).$$

The remaining principal directions are defined recursively as

$$v_k = \arg\max_{||v||=1} \sum_{i=1}^{N} ||\operatorname{Log}_{x_i}(\pi_H(x_i))||^2,$$
$$\text{where} \quad H = \operatorname{Exp}_\mu(\operatorname{span}(\{v_1, \ldots, v_{k-1}, v\}) \cap U).$$

The only difference in these equations from the variance approach given in (4.6) and (4.7) is that the base point for the Log is x_i rather than μ.

The question immediately arises: is the least-squares approach to PGA equivalent to the maximum variance definition? For data in $\mathbb{R}^n$ the two definitions are equivalent since PGA reduces to PCA in the linear case. The question remains unsolved for more general manifolds. This issue is discussed further in the future work section in Chapter 7.

4.2.6 Approximating Principal Geodesic Analysis

Exact computation of PGA, that is, solution of the minimizations (4.6) and (4.7), requires computation of the projection operator π_H. However, the projection operator does not have a closed-form solution for general manifolds. Projection onto a geodesic submanifold can be approximated linearly in the tangent space of M. Let $H \subset M$ be a geodesic submanifold at a point $p \in M$ and $x \in M$ a point to be projected onto H. Then the projection operator is approximated by

$$\begin{aligned}\pi_H(x) &= \arg\min_{y \in H} ||\operatorname{Log}_x(y)||^2 \\ &\approx \arg\min_{y \in H} ||\operatorname{Log}_p(x) - \operatorname{Log}_p(y)||^2.\end{aligned}$$

Notice that $\operatorname{Log}_p(y)$ is simply a vector in T_pH. Thus we may rewrite the approximation in terms of tangent vectors as

$$\operatorname{Log}_p(\pi_H(x)) \approx \arg\min_{v \in T_pH} ||\operatorname{Log}_p(x) - v||^2.$$

But this is simply the minimization formula for linear projection of $\operatorname{Log}_p(x)$ onto the linear subspace T_pH. So, if $v_1, \ldots, v_k$ is an orthonormal basis for T_pH, the projection

operator can be approximated by the formula

$$\mathrm{Log}_p\left(\pi_H(x)\right) \approx \sum_{i=1}^{k} \langle v_i, \mathrm{Log}_p(x)\rangle. \tag{4.8}$$

Analyzing the quality of the approximation to the projection formula (4.8) is difficult for general manifolds. It is obviously gives the exact projection in the case of $\mathbb{R}^n$. For other manifolds of constant curvature, such as spheres, S^n, and hyperbolic spaces, H^n, the projection formula can be computed exactly in closed form. This makes it possible to get an idea of how well the linear approximation does in these cases. The error computations for the sphere S^2 are carried out at the end of this subsection as an example.

If we use (4.8) to approximate the projection operator π_H in (4.6) and (4.7), we get

$$v_1 \approx \arg\max_{\|v\|=1} \sum_{i=1}^{N} \langle v, \mathrm{Log}_\mu(x_i)\rangle^2,$$
$$v_k \approx \arg\max_{\|v\|=1} \sum_{i=1}^{N} \sum_{j=1}^{k-1} \langle v_j, \mathrm{Log}_\mu(x_i)\rangle^2 + \langle v, \mathrm{Log}_\mu(x_i)\rangle^2.$$

The above minimization problem is simply the standard principal component analysis in $T_\mu M$ of the vectors $\mathrm{Log}_\mu(x_i)$, which can be seen by comparing the approximations above to the PCA equations, (4.2) and (4.3). Thus an algorithm for approximating the PGA of data on a manifold is given by

Algorithm 4.2: Principal Geodesic Analysis
Input: $x_1, \ldots, x_N \in M$
Output: Principal directions, $v_k \in T_\mu M$
Variances, $\lambda_k \in \mathbb{R}$
μ = intrinsic mean of $\{x_i\}$ (Algorithm 4.1)
$u_i = \mathrm{Log}_\mu(x_i)$
$\mathbf{S} = \frac{1}{N}\sum_{i=1}^{N} u_i u_i^T$
$\{v_k, \lambda_k\}$ = eigenvectors/eigenvalues of $\mathbf{S}$.

Now we demonstrate the error computations for the projection operator in the special case of the sphere S^2. Let H be a geodesic (i.e., a great circle) through a point $p \in S^2$. Given a point $x \in S^2$, we wish to compute its true projection onto H and compare that with the approximation in the tangent space T_pS^2. Thus we have the spherical right triangle as shown in Fig. 4.1. We know the hypotenuse length $c = d(p, x)$ and the angle

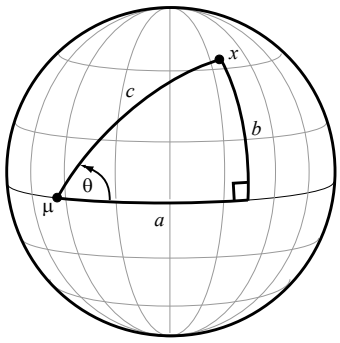

Figure 4.1: The spherical triangle used in the calculation of the projection operator for S^2.

θ, and we want to derive the true projection, which is given by the side length a. We use the following two relations from the laws of spherical trigonometry:

$$\cos c = (\cos a)(\cos b),$$
$$\frac{\sin b}{\sin \theta} = \sin c.$$

Solving for a in terms of the hypotenuse c and the angle θ, we have

$$a = \arccos\left(\frac{\cos c}{\sqrt{1-(\sin\theta \sin b)^2}}\right).$$

The tangent-space approximation in (4.8) is equivalent to solving for the corresponding right triangle in $\mathbb{R}^2$. Using standard Euclidean trigonometry, the tangent-space approximation (4.8) gives

$$a \approx c\cos\theta.$$

For nearby data, i.e., small values for c, this gives a good approximation. For example, for $c < \frac{\pi}{4}$ the maximum absolute error is 0.07rad. However, the error can be significant for far away points, i.e., as c approaches $\frac{\pi}{2}$.

4.3 Conclusions

In this chapter we have presented principal geodesic analysis, a new methodology for analyzing the statistical variability of data on a manifold. We reviewed two definitions

of a mean value on a manifold, the intrinsic and extrinsic mean, and we argued that the intrinsic mean was preferable. Principal geodesic analysis is defined as a direct generalization of principal component analysis using either a variance maximizing or least-squares approach. We gave an algorithm for computing an approximation to PGA in the tangent space to the mean.

The methods is this chapter will be applied to medial representations in Chapter 5, with the application of using the statistics as a geometric prior in a Bayesian deformable models segmentation method. In Chapter 6 we will apply PGA to the study of diffusion tensor data, with the driving application to study the statistical variability diffusion tensor images across patient populations.

There are several theoretical questions about principal geodesic analysis that remain to be solved. They are

1. Are the two definitions for PGA, variance maximization and least-squares, equivalent?

2. Is the greedy, i.e., recursive, approach the finding the v_k in (4.6) and (4.7) equivalent to finding each subspace V_k independently?

3. If the V_k are found independently, are they even subsets of one another, i.e., do they satisfy $V_k \subset V_{k+1}$?

Another open problem is an algorithm for computing principal geodesic analysis exactly when the projection operator is known in closed form. These issues will be discussed further in the future work section of Chapter 7.

Chapter 5

Statistics of M-reps

In this chapter[1] we apply the statistical framework presented in the previous chapter for general manifolds to the statistical analysis of m-rep models of anatomical objects. Throughout this chapter we use the mesh-based medial representation with order 1 medial atoms as described in Section 3.3.2. Thus the term "m-rep model" will always refer to this type of medial representation. We first show in Section 5.1 that the space of m-rep models containing n medial atoms is a symmetric space that we denote by $\mathcal{M}(n)$. As is the case with other shape analysis methods, since we are interested in studying the variability of shape alone, we must first align the models to a common position, orientation, and scale. In Section 5.2 we present an m-rep alignment algorithm that minimizes the sum-of-squared geodesic distances between models, i.e., has the desirable property that it minimizes the same metric as is used in the definition of the mean and principal geodesics, but over the global similarity transformations of alignment. Next the mean and PGA algorithms are adapted to the specific case of m-rep models in Sections 5.3 and 5.4. The initial data is a set of m-rep models that have been fit to a particular class of objects in a training set of images. Finally, in Section 5.5 we describe how the statistical methods developed in this chapter give both an optimization parameter space and a geometric prior in the Bayesian deformable m-reps segmentation method.

5.1 M-reps as Elements of a Symmetric Space

In this section it is shown that m-rep models can be parameterized as a symmetric space. This formulation will open up m-reps to the statistical methods introduced in

[1]The work presented in this chapter was done in collaboration with Dr. Sarang Joshi, Dr. Conglin Lu, and Dr. Stephen Pizer at the University of North Carolina. This chapter contains parts of the paper [42] and is also based on the previous papers [40, 41].

the previous chapter, namely, manifold means and PGA. We then give formulas for the Riemannian log and exponential maps that are required to be able to compute the statistics described in the previous chapter. Finally, we describe the data set of 86 hippocampus m-rep models that are used to demonstrate the methods that are presented in later sections. Let $\mathcal{M}_d(n)$ denote the space of all d-dimensional m-rep models containing n (order 1) medial atoms. The focus in this chapter will be on the case $d = 3$, so $\mathcal{M}(n)$ without a subscript will be used to denote the three-dimensional case. Recall from Definition 3.5 that a three-dimensional order 1 medial atom is defined as a tuple $(x, r, n_0, n_1) \in \mathbb{R}^3 \times \mathbb{R}^+ \times S^2 \times S^2$. Recall that such an atom represents a position x and two equal length vectors emanating from this position with length r and directions n_0, n_1. Therefore, $\mathcal{M}(1) = \mathbb{R}^3 \times \mathbb{R}^+ \times S^2 \times S^2$ is the space of all possible order 1 medial atoms. In general an m-rep model with n medial atoms is a point in the space $\mathcal{M}(n) = \mathcal{M}(1)^n = (\mathbb{R}^3 \times \mathbb{R}^+ \times S^2 \times S^2)^n$, i.e., the direct product of n copies of $\mathcal{M}(1)$. As was shown in the background section on symmetric spaces (Section 2.5), each of the space $\mathbb{R}^3$, $\mathbb{R}^+$, and S^2 are symmetric spaces. Therefore, to show that $\mathcal{M}(n)$ is a symmetric space it suffices to show that the direct product of symmetric spaces is also a symmetric space. (This is a well-known fact of symmetric spaces, but a quick derivation is given here all the same.)

Recall from Theorem 2.4 that the direct product of Lie groups is again a Lie group. The direct product operation is also defined for mappings, as the next definition shows.

Definition 5.1. The **direct product** of a collection of maps $f_i : X_i \to Y_i,\ (1 \leq i \leq n)$, where X_i, Y_i are sets, is defined as the map $(f_1 \times \cdots \times f_n) : X_1 \times \cdots \times X_n \to Y_1 \times \cdots \times Y_n$ given by

$$(f_1 \times \cdots \times f_n)(x_1, \ldots, x_n) = (f_1(x_1), \ldots, f_n(x_n)).$$

Now, if $G_i,\ (1 \leq i \leq n)$ are Lie groups with automorphisms $\phi_i : G_i \to G_i$, it is easy to see that the product map $\phi_1 \times \cdots \times \phi_n$ is an automorphism of $G_1 \times \cdots \times G_n$. The next theorem now follows easily from Theorems 2.4 and 2.10.

Theorem 5.1. *If $M_i : 1 \leq i \leq n$ are symmetric spaces, then the direct product manifold $M = M_1 \times \cdots \times M_n$ is also a symmetric space.*

Proof. Since M_i is a symmetric space, by Theorem 2.11it can be written as the quotient space $M_i = G_i/H_i$, where G_i is a connected Lie group, and H_i is a connected, compact subgroup of G_i. Also, from Theorem 2.11 there is an involutive automorphism $\alpha_i : G_i \to G_i$ that has fixed set H_i. The direct product $G = (G_1 \times \cdots \times G_n)$ is a connected Lie group

by Theorem 2.4, and the direct product $H = (H_1 \times \cdots \times H_n)$ is a connected, compact Lie subgroup of G. Also, the product map $\alpha = (\alpha_1 \times \cdots \times \alpha_n)$ is an involutive automorphism of G with fixed set H. Now, M is diffeomorphic to the Lie group quotient space G/H because of the equivalence $(G_1/H_1) \times \cdots \times (G_n/H_n) = (G_1 \times \cdots \times G_n)/(H_1 \times \cdots \times H_n)$. Therefore, M satisfies the conditions to be a symmetric space given in Theorem 2.10. □

5.1.1 The Exponential and Log Maps for M-reps

Before we can apply the statistical techniques for manifolds developed in the previous chapter, we must define the exponential and log maps for the symmetric space $\mathcal{M}(n)$, the space of m-rep models with n atoms. We begin with a discussion of the medial atom space $\mathcal{M}(1) = \mathbb{R}^3 \times \mathbb{R}^+ \times S^2 \times S^2$. Let $p = (0, 1, p_0, p_1) \in \mathcal{M}(1)$ be the base point, where $p_0 = p_1 = (0, 0, 1)$ are the base points for the spherical components. The tangent space for $\mathcal{M}(1)$ at the base point p can be identified with $\mathbb{R}^8$. We write a tangent vector $u \in T_p\mathcal{M}(1)$ as $u = (\mathbf{x}, \rho, v_0, v_1)$, where $\mathbf{x} \in \mathbb{R}^3$ is the positional tangent component, $\rho \in \mathbb{R}$ is the radius tangent component, and $v_0, v_1 \in \mathbb{R}^2$ are the spherical tangent components. The exponential map for $\mathcal{M}(1)$ is now the direct product of the exponential map for each component. The exponential map for $\mathbb{R}^3$ is simply the identity map, for $\mathbb{R}$ it is the standard real exponential function, and for S^2 it is the spherical exponential map given in (2.5). Thus for $\mathcal{M}(1)$ we have

$$\mathrm{Exp}_p(u) = (\mathbf{x}, e^\rho, \mathrm{Exp}_{p_0}(v_0), \mathrm{Exp}_{p_1}(v_1)),$$

where the two Exp maps on the right-hand side are the spherical exponential maps. Likewise, the log map of a point $\mathbf{m} = (\mathbf{x}, r, \mathbf{n}_0, \mathbf{n}_1)$ is the direct product map

$$\mathrm{Log}_p(\mathbf{m}) = (\mathbf{x}, \log r, \mathrm{Log}_{p_0}(\mathbf{n}_0), \mathrm{Log}_{p_1}(\mathbf{n}_1)),$$

where the two Log maps on the right-hand side are the spherical log maps given by (2.6). Finally, the exponential and log maps for the m-rep model space $\mathcal{M}(n)$ are just the direct products of n copies of the corresponding maps for the medial atom space $\mathcal{M}(1)$. For end atoms there is an extra parameter η representing the elongation of the bisector spoke that points to the crest (see Section 3.3.2). This is treated as another positive real number under multiplication. Therefore, end atoms are represented as the symmetric space $\mathbb{R}^3 \times \mathbb{R}^+ \times S^2 \times S^2 \times \mathbb{R}^+$. The exponential and log maps for these atoms are just augmented with another copy of the corresponding map for $\mathbb{R}^+$.

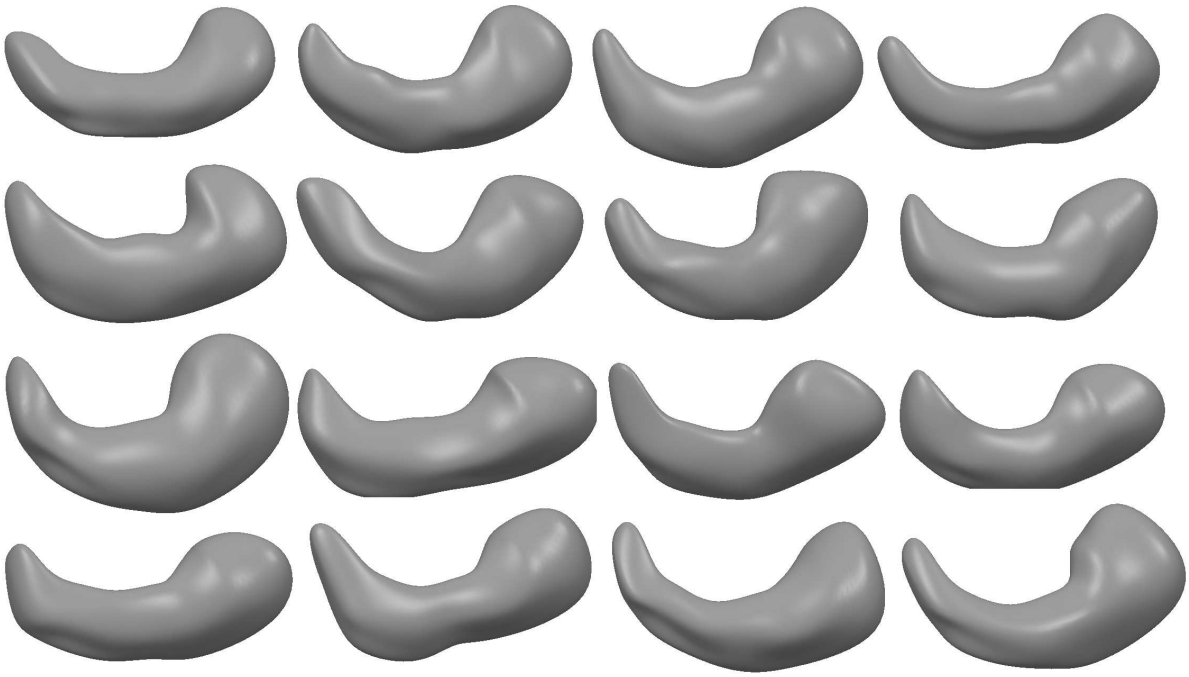

Figure 5.1: The surfaces of 16 of the 86 original hippocampus m-rep models.

Notice that the position, radius, and orientations are not in the same units. For the PGA calculations in Section 4.2 we scale the radius and sphere components (and η for end atoms) in the Riemannian metric to be commensurate with the positional components. The scaling factor for both components is the average radius over all corresponding medial atoms in the population. Thus the norm of the vector $u = T_p\mathcal{M}(1)$ becomes

$$||u|| = \left(||\mathbf{x}||^2 + \bar{r}^2(\rho^2 + ||v_1||^2 + ||v_2||^2)\right)^{\frac{1}{2}},$$

where $\bar{r}$ is the average radius over all corresponding medial atoms. Using this norm and the formula for Riemannian distance, the distance between two atoms $\mathbf{m}_1, \mathbf{m}_2 \in \mathcal{M}(1)$ is given by

$$d(\mathbf{m}_1, \mathbf{m}_2) = ||\operatorname{Log}_{\mathbf{m}_1}(\mathbf{m}_2)||. \tag{5.1}$$

5.1.2 The Hippocampus Data Set

The results of these techniques are demonstrated on a set of 86 m-rep models of hippocampi from a schizophrenia study. A subset of 16 of these models are displayed as surfaces in Fig. 5.1. The m-rep models were automatically generated by the method described in [120], which chooses the medial topology and sampling that is sufficient

to represent the population of objects. The models were fit to expert segmentations of the hippocampi from MRI data. The average distance error from the m-rep boundary to the original segmentation boundary ranged from 0.14mm and 0.27mm with a mean error of 0.17mm. This is well within the original MRI voxel size (0.9375mm x 0.9375mm 1.5mm). The sampling on each m-rep was 3×8, making each model a point on the symmetric space $\mathcal{M}(24)$. Since the dimensionality of $\mathcal{M}(1)$ is 8, the total number of dimensions required to represent the hippocampus models is 192.

5.2 M-rep Alignment

To globally align objects described by boundary points to a common position, orientation, and scale, the standard method is the Procrustes method [47]. Procrustes alignment minimizes the sum-of-squared distances between corresponding boundary points, the same metric used in defining the mean and principal components. We now develop an analogous alignment procedure based on minimizing sum-of-squared geodesic distances on $\mathcal{M}(n)$, the symmetric space of m-rep objects with n atoms.

Let $\mathbf{S} = (s, \mathbf{R}, \mathbf{w})$ denote a similarity transformation in $\mathbb{R}^3$ consisting of a scaling by $s \in \mathbb{R}^+$, a rotation by $\mathbf{R} \in SO(3)$, and a translation by $\mathbf{w} \in \mathbb{R}^3$. We define the action of $\mathbf{S}$ on a medial atom $\mathbf{m} = (\mathbf{x}, r, \mathbf{n}_0, \mathbf{n}_1)$ by

$$\mathbf{S} \cdot \mathbf{m} = \mathbf{S} \cdot (\mathbf{x}, r, \mathbf{n}_0, \mathbf{n}_1) = (s\mathbf{R} \cdot \mathbf{x} + \mathbf{w}, sr, \mathbf{R} \cdot \mathbf{n}_0, \mathbf{R} \cdot \mathbf{n}_1). \tag{5.2}$$

This action is the standard similarity transform of the position x, and the scaling and rotation of the spokes are transformations about the medial position x. Now the action of $\mathbf{S}$ on an m-rep object $\mathbf{M} = \{\mathbf{m}_i : i = 1, \ldots, n\}$ is simply the application of $\mathbf{S}$ to each of $\mathbf{M}$'s medial atoms:

$$\mathbf{S} \cdot \mathbf{M} = \{\mathbf{S} \cdot \mathbf{m}_i : i = 1, \ldots, n\}. \tag{5.3}$$

It is easy to check from the equation for the implied boundary points (3.7) that this action of $\mathbf{S}$ on $\mathbf{M}$ also transforms the implied boundary points of $\mathbf{M}$ by the similarity transformation $\mathbf{S}$.

Consider a collection $\mathbf{M}_1, \ldots, \mathbf{M}_N \in \mathcal{M}(n)$ of m-rep objects to be aligned, each consisting of n medial atoms. We write $\mathbf{m}_{\alpha i}$ to denote the ith medial atom in the αth m-rep object. Notice that the m-rep parameters, which are positions, orientations, and scalings, are in different units. Before we apply PGA to the m-reps, it is necessary to make the various parameters commensurate. This is done by scaling the log rotations

and log radii by the average radius value of the corresponding medial atoms. The squared-distance metric between two m-rep models $\mathbf{M}_i$ and $\mathbf{M}_j$ becomes

$$d(\mathbf{M}_i, \mathbf{M}_j)^2 = \sum_{\alpha=1}^{n} d(\mathbf{m}_{\alpha i}, \mathbf{m}_{\alpha j})^2, \tag{5.4}$$

where the $d(\cdot, \cdot)$ for medial atoms on the right-hand side is given by (5.1).

The m-rep alignment algorithm finds the set of similarity transforms $\mathbf{S}_1, \ldots, \mathbf{S}_N$ that minimize the total sum-of-squared distances between the m-rep figures:

$$d(\mathbf{S}_1, \ldots, \mathbf{S}_N; \mathbf{M}_1, \ldots, \mathbf{M}_N) = \sum_{i=1}^{N} \sum_{j=1}^{i} d(\mathbf{S}_i \cdot \mathbf{M}_i, \mathbf{S}_j \cdot \mathbf{M}_j)^2. \tag{5.5}$$

Following the algorithm for generalized Procrustes analysis for objects in $\mathbb{R}^3$, minimization of (5.5) proceeds in stages:

Algorithm 5.1: M-rep Alignment

1. *Translations.* First, the translational part of each $\mathbf{S}_i$ in (5.5) is minimized once and for all by centering each m-rep model. That is, each model is translated so that the average of it's medial atoms' positions is the origin.

2. *Rotations and Scalings.* The ith model, $\mathbf{M}_i$, is aligned to the mean of the remaining models, denoted μ_i. The alignment is accomplished by a gradient descent algorithm on $SO(3) \times \mathbb{R}^+$ to minimize $d(\mu_i, \mathbf{S}_i \cdot \mathbf{M}_i)^2$. The gradient is approximated numerically by a central differences scheme. This is done for each of the N models.

3. *Iterate.* Step 2 is repeated until the metric (5.5) cannot be further minimized.

The result of applying the m-rep alignment algorithm to the 86 hippocampus m-rep models is shown in Fig. 5.2. The resulting aligned figures are displayed as overlaid medial atom centers. Since the rotation and scaling step of the alignment algorithm is a gradient descent algorithm, it is important to find a good starting position. Thus the alignment was initialized by first aligning the m-rep models with the Procrustes method applied to the implied boundary points of the m-rep models.

Figure 5.2: The 86 aligned hippocampus m-reps, shown as overlayed medial atom centers.

5.3 M-rep Averages

Algorithm 4.1 can be adapted for computing means of m-rep models by taking the manifold to be the symmetric space $\mathcal{M}(n)$. Recall that the gradient descent algorithm for the mean, Algorithm 4.1, has a parameter τ, which is the step size taken in the downhill gradient direction. For m-reps a step size of $\tau = 1$ is used. Since $\mathcal{M}(n)$ is a direct product space, the algorithm will converge if each of the components converge. Notice that each of the $\mathbb{R}^3$ and $\mathbb{R}^+$ components in $\mathcal{M}(n)$ converge in a single iteration since they are commutative Lie groups. The step size of $\tau = 1$ is sufficient to ensure that the S^2 components converge as well. Also, care must be taken to ensure that the data is contained in a small enough neighborhood that the minimum in (4.1) is unique. For the $\mathbb{R}^3$ and $\mathbb{R}^+$ components there is no restriction on the spread of the data. However, for the S^2 components the data must lie within a neighborhood of radius $\frac{\pi}{2}$ (see [18]), i.e., within an open hemisphere. This is a reasonable assumption for the aligned m-rep models, whose spoke directions for corresponding atoms are fairly localized, and we have not experienced in practice any models that do not fall within such constraints. We now have the following algorithm for computing the intrinsic mean of a collection of m-rep models:

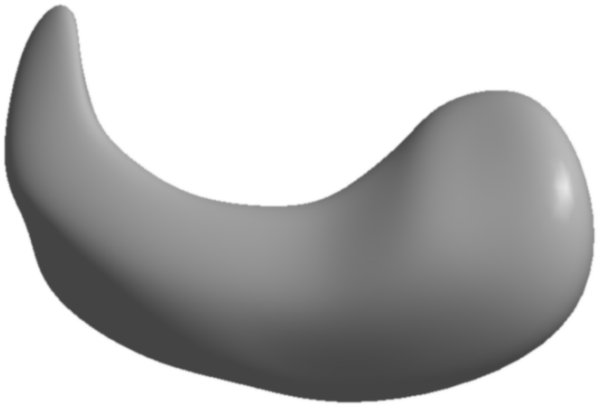

Figure 5.3: The surface of the mean hippocampus m-rep.

Algorithm 5.2: M-rep Mean

Input: $\mathbf{M}_1, \ldots, \mathbf{M}_N \in \mathcal{M}(n)$, m-rep models

Output: $\mu \in \mathcal{M}(n)$, the intrinsic mean

$\mu_0 = \mathbf{M}_1$

Do

$\Delta\mu = \frac{1}{N} \sum_{i=1}^{N} \mathrm{Log}_{\mu_j} \mathbf{M}_i$

$\mu_{j+1} = \mathrm{Exp}_{\mu_j}(\Delta\mu)$

While $||\Delta\mu|| > \epsilon$.

Fig. 5.3 shows the surface of the resulting intrinsic mean of the 86 aligned hippocampus m-rep models computed by Algorithm 5.2. The maximum difference in the rotation angle from the mean in either of the S^2 components was 0.1276 for the entire data set. Thus the data falls well within a neighborhood of radius $\frac{\pi}{2}$ as required.

One might be tempted to simplify the statistical computations by treating a medial atom as three points in $\mathbb{R}^3$: the center point $\mathbf{x}$, and the two implied boundary points $\mathbf{y}_0, \mathbf{y}_1$. With this linear representation, the symmetric space mean algorithm involving geodesic computations is replaced by a simpler linear average. However, linear averaging produces invalid medial atoms. To demonstrate this, we computed a linear average of the atoms at a corresponding location in the hippocampus mesh across the population. This average was compared to the symmetric space average described in this paper. The resulting two medial atoms are shown in Fig. 5.4. The symmetric space mean is a valid medial atom, while the linear average is not because the two spoke vectors do not have equal length. The ratio of the two spoke lengths in the linear average is 1.2 to 1.

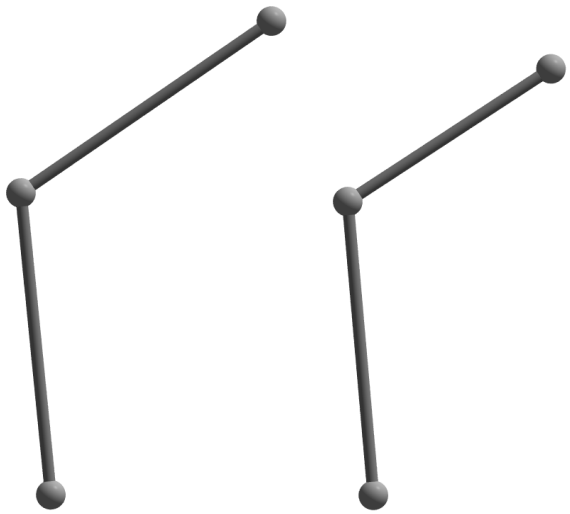

Figure 5.4: The resulting average of corresponding medial atoms in the hippocampus models using (a) symmetric space averaging and (b) linear averaging. Notice that the linear average is not a valid medial atom as the two spokes do not have equal length.

5.4 M-rep PGA

The PGA algorithm for m-rep models is a direct adaptation of Algorithm 4.2. The only concern is to check that the data is localized enough for the projection operator to be unique. That is, we must determine the neighborhood U used in (4.6) and (4.7). Again there is no restriction on the $\mathbb{R}^3$ and $\mathbb{R}^+$ components. For S^2 components it is also sufficient to consider a neighborhood with radius $\frac{\pi}{2}$. Therefore, there are no further constraints on the data than those discussed for the mean. Also, we can expect the projection operator to be well-approximated in the tangent space, given the discussion of the error in Section 4.2.3 and the fact that the data lie within 0.1276 rad. from the mean. Finally, the computation of the PGA of a collection of m-rep models is given by

Algorithm 5.3: M-rep PGA
Input: M-rep models, $\mathbf{M}_1, \ldots, \mathbf{M}_N \in \mathcal{M}(n)$
Output: Principal directions, $v_k \in T_\mu\mathcal{M}(n)$
Variances, $\lambda_k \in \mathbb{R}$
μ = intrinsic mean of $\{\mathbf{M}_i\}$ (Algorithm 5.2)
$\mathbf{u}_i = \mathrm{Log}_\mu(\mathbf{M}_i)$
$\mathbf{S} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{u}_i\mathbf{u}_i^T$
$\{v_k, \lambda_k\}$ = eigenvectors/eigenvalues of $\mathbf{S}$.

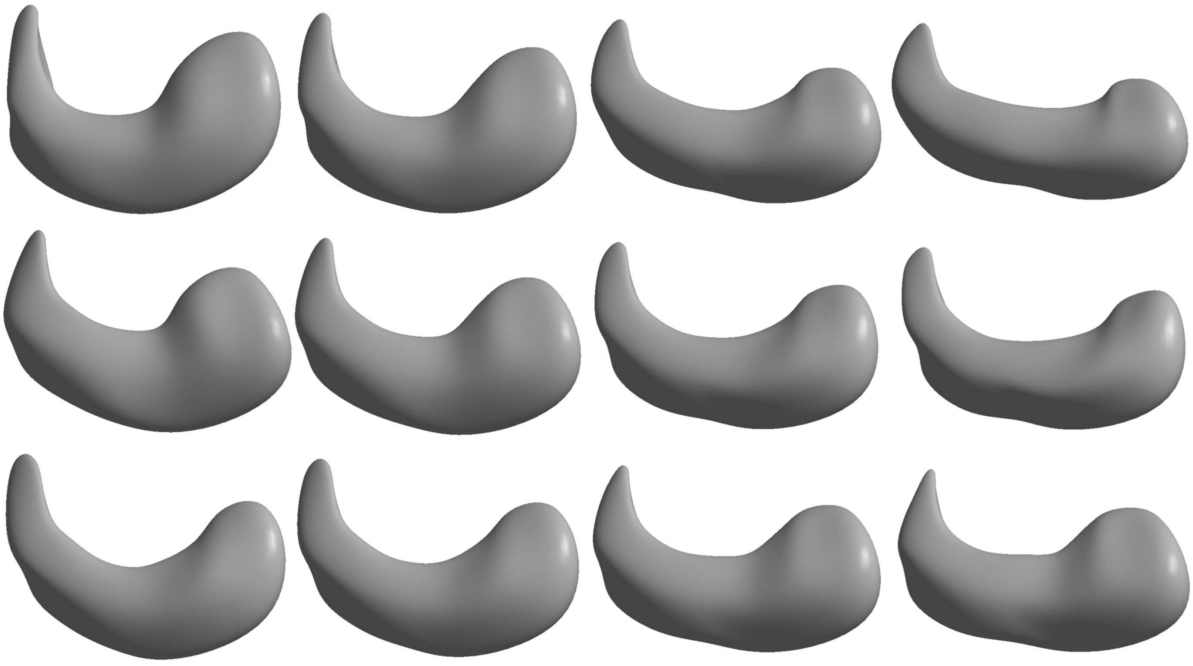

Figure 5.5: The first three PGA modes of variation for the hippocampus m-reps. From left to right are the PGA deformations for -3, -1.5, 1.5, and 3 times $\sqrt{\lambda_i}$.

Analogous to linear PCA models, we may choose a subset of the principal directions v_k that is sufficient to describe the variability of the m-rep shape space. New m-rep models may be generated within this subspace of typical objects. Given a set of real coefficients $\alpha = (\alpha_1, \ldots, \alpha_d)$, we generate a new m-rep model by

$$\mathbf{M}(\alpha) = \mathrm{Exp}_\mu \Big(\sum_{k=1}^{d} \alpha_k v_k \Big), \tag{5.6}$$

where α_k is chosen to be within $[-3\sqrt{\lambda_k}, 3\sqrt{\lambda_k}]$.

The m-rep PGA algorithm was applied to the aligned hippocampus data set. Fig. 5.5 displays the first three modes of variation as the implied boundaries of the m-reps generated from PGA coefficients $\alpha_k = -3\sqrt{\lambda_k}, -1.5\sqrt{\lambda_k}, 0, 1.5\sqrt{\lambda_k}, 3\sqrt{\lambda_k}$. A plot of the eigenvalues and their cumulative sums is given in Fig. 5.6. The first 30 modes capture 95 percent of the total variability, which is a significant reduction from the original 192 dimensions of the hippocampus m-rep model.

In this statistical analysis of the hippocampus, the resulting mean model (Fig. 5.3) and the models generated from the PGA (Fig. 5.5) qualitatively look like hippocampi. Also, the generated models are legal m-reps, that is, they produce valid meshes of medial

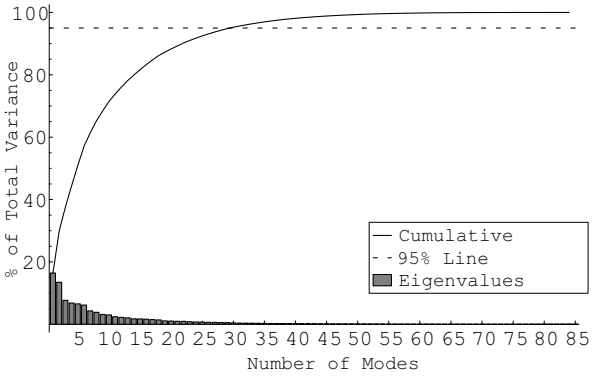

Figure 5.6: A plot of the eigenvalues from the modes of variation and their cumulative sums.

atoms and smooth, non-folding implied boundaries. The mean and PGA algorithms have also been applied to populations of m-rep models of the kidney, prostate, heart, and liver. In our experience so far, we have found that the mean and PGA methods described in this chapter generate legal m-rep models when the input models are legal. While we do not have quantitative results to say that these methods produce legal models, our experiments indicate that they produce valid results for real-world data.

5.5 PGA in Deformable M-reps Segmentation

In this section we describe how the method of principal geodesic analysis on m-reps that has been developed in this chapter can be used in a Bayesian deformable models segmentation method based on m-reps. Recall from Section 3.3.2 that m-reps segmentation proceeds in several stages corresponding to different levels of scale. In this section we focus on the figure stage of the optimization of a single figure model. Principal geodesic analysis will be used in two aspects of the segmentation process:

1. The principal geodesic components are used as a parameter space generating global deformations of the m-rep figure.

2. The geodesic Mahalanobis distance is used as the geometric prior term in the Bayesian objective function.

In the segmentation problem we are given an image I, and we want to fit an m-rep model to a particular object in the image. A statistical m-rep model is trained on a population of known objects of the same class. The training proceeds by fitting

a set of m-rep models to binary segmentations of objects from similar images. Next a mean m-rep model μ and a principal geodesic analysis are computed as described above. The principal geodesic analysis results in a set of principal directions $v_k \in T_\mu \mathcal{M}(n)$ and variances λ_k. The first d principal directions are chosen depending on the desired amount of variation that is desired.

5.5.1 Principal Geodesic Deformations

The mean model μ is used as the initial model in the optimization. It is placed within the image by the user applying a translation, rotation, and scale. As described in the background section on m-reps (Section 3.3.2), the figure stage proceeds by deforming the model by global transformations to optimize the objective function. The difference is that we now use the principal geodesics as the global deformations of the model. This is achieved by optimizing over parameters $c = (c_1, \ldots, c_d)$ that generate deformed versions of the mean model given by

$$\mathbf{M}(c) = S \cdot \mathrm{Exp}_\mu \left(\sum_{i=1}^{d} c_k \; v_k \right).$$

Here S represents the user-defined similarity transform used to place the mean model into the image. Care must be taken in the order that the similarity transform is applied with respect to the PGA transformations. The two operations do not commute, and since the principal directions are defined as tangent vectors to the mean model, it does not make sense to apply them to a transformed version of the mean. Therefore, the similarity transform must be applied *after* the principal geodesic deformation. An alternative would be to apply the similarity transform to the mean and also apply the derivative mapping of the similarity transform to the principal directions (since they are after all tangent vectors). Then the v_k can be replaced by the transformed vectors, and the similarity transform need not be applied during the optimization.

5.5.2 PGA-Based Geometric Prior

The next part of using principal geodesic analysis in the deformable m-reps segmentation is to use the geodesic Mahalanobis distance as a geometric prior in the objective function. Recall from Section 3.3.2 that the posterior objective function used for m-reps

segmentation is given by

$$F(\mathbf{M}(c), I) = L(\mathbf{M}(c), I) + \alpha\, G(\mathbf{M}(c)),$$

where L is the image match term and G is the geometric typicality. In the Bayesian setting this objective function F with $\alpha = 1$ can be seen as a log posterior probability density, where the image match L is a log likelihood probability and the geometric typicality G is a log prior probability.

We focus on the geometric typicality term G. We define this term to be the squared geodesic Mahalanobis distance, which is proportional to the log prior probability

$$G(\mathbf{M}(c)) = \sum_{i=1}^{d} \frac{c_k^2}{\lambda_k} \propto \log(p(\mathbf{M}(c)).$$

The probability distribution p can be constructed as a truncated Gaussian distribution in the tangent space to the intrinsic mean, $\mu \in \mathcal{M}(n)$. If $U \subset T_\mu\mathcal{M}(n)$ is the neighborhood in which PGA is well-defined (recall Section 4.2.4), then p is given by

$$p(\mathbf{M}) = \frac{1}{V(U)(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}\operatorname{Log}_\mu(\mathbf{M})^T\Sigma^{-1}\operatorname{Log}_\mu(\mathbf{M})\right),$$

where $V(U)$ denotes the normalization factor based on the neighborhood U to make p integrate to 1 and Σ is the covariance matrix approximated in Algorithm 5.3. It should be stressed that this distribution is not a Gaussian distribution on the manifold $\mathcal{M}(n)$ as defined in [61] (recall Section 3.1.4). That is, the density p is not a fundamental solution to the heat equation on $\mathcal{M}(n)$. Also, the intrinsic mean and the covariance matrix that are derived from the training data are not maximum-likelihood estimates of the density parameters. It is not clear that a Gaussian distribution is the correct model, and further research is required to investigate possible probability models and the estimation of their parameters.

The statistical segmentation method presented in this section has been implemented as a part of Pablo [102], the deformable m-reps segmentation tool developed at UNC. A study carried out by Rao *et al.* [105] compared deformable m-rep and human segmentations of kidneys from CT. The m-rep segmentation process used was the one presented in this section. The training set for the geometry statistics included 53 models of the right kidney and 51 models of the left kidney (left and right kidneys were trained as two separate groups). The target images to be segmented were 12 CT images of the

kidneys (left and right). Human segmentations were carried out by manual slice-by-slice contour outlining by two different raters. The statistical m-rep segmentation gave reasonable results that compared favorably with the human segmentations. The mean surface separations between the human and m-rep segmentations were sub-voxel. The differences between the human and m-rep segmentations were slightly larger than the differences between the two human segmentations. However, the experiment was biased towards this result since the humans used a slice-based segmentation while the m-reps segmentation was a smooth 3D model.

5.6 Conclusions

In this chapter we demonstrated how statistical m-rep shape models can be built using the mean and PGA methods presented in the previous chapter. We first showed that m-rep models are elements of a symmetric space and gave formulas for the Riemannian log and exponential map. We then developed an alignment method for m-rep models analogous to the Procrustes alignment method for point set shape models, except that the m-rep alignment method is based on a least-squares approach using geodesic distances on $\mathcal{M}(n)$. Finally, we adapted the mean and PGA algorithms to the m-rep case. These methods were demonstrated on a set of 86 hippocampus m-rep models fit from expert binary segmentations.

The work of this chapter brings up several questions that remain to be answered:

1. Following an analogous construction (see Section 3.1) that is used to build Kendall's shape spaces, $\Sigma_n^k = (\mathbb{R}^{nk} - \{0\})/\mathrm{Sim}(n)$, a medial shape space can be constructed as the quotient $\mathcal{M}(n)/\mathrm{Sim}(3)$. In other words, the medial shape space is the space created by identifying m-rep models that are different by only a similarity transform. An open problem is to classify the topology and Riemannian structure of these medial shape spaces.

2. The work in this chapter used the tangent space approximation to principal geodesic analysis. It would be preferable to solve for the projection operator explicitly and find the true principal geodesic analysis. In addition, averages and principal geodesic analysis could be computed on the medial shape space $\mathcal{M}(n)/\mathrm{Sim}(3)$ rather than on aligned models in the space $\mathcal{M}(n)$.

3. A thorough validation of the segmentation method using PGA is ongoing. It is designed to test whether the statistical segmentation method has advantages over

a method without geometry statistics. This is measured in both advantages in segmentation accuracy and in the speed of the segmentation.

4. The methods presented in this chapter were for single-figure objects. This could be extended to multi-figure models and multi-object complexes. The challenge for multi-figure models is to allow only variations that preserve the hinge relationship of a child figure with its parent. Care must be taken in the multi-object situation to prevent adjacent objects from intersecting in the PGA deformations.

These issues will be discussed further in the future work section in Chapter 7.

Chapter 6

Statistics of Diffusion Tensors

As discussed in the background chapter (Section 3.4) diffusion tensor magnetic resonance imaging (DT-MRI) is emerging as an important tool in medical image analysis of the brain. However, relatively little work has been done on producing statistics of diffusion tensors. A main difficulty is that the space of diffusion tensors, i.e., the space of symmetric, positive-definite matrices, does not form a vector space. Therefore, standard linear statistical techniques do not apply. This chapter[1] presents new methods for the statistical analysis of diffusion tensors.

We demonstrate that the space of diffusion tensors is more naturally described as a Riemannian symmetric space, rather than a linear space. Applying the ideas presented in Chapter 4 to this space, we develop new methods for averaging and describing the variability of diffusion tensor data. It is shown that these statistics preserve natural properties of the diffusion tensors, most importantly the positive-definiteness, that are not preserved by linear statistics. The framework presented in this chapter should be useful in the registration of diffusion tensor images, the smoothing of diffusion tensor images, the production of statistical atlases from diffusion tensor data, and the quantification of the anatomical variability caused by disease.

In Section 6.1 we show why the space of diffusion tensors is not a linear space and how linear statistical methods such as PCA break down in this space. In Section 6.2 we formulate the space of diffusion tensors as a Riemannian symmetric space. Section 6.3 presents the methods for averaging and principal geodesic analysis of diffusion tensors. Finally, Section 6.5 develops several new methods based on the symmetric space formulation of diffusion tensors that are essential for building statistical atlases of diffusion

[1]The work presented in this chapter was done in collaboration with Dr. Sarang Joshi at the University of North Carolina. This chapter is an expanded version of the paper [39].

tensor images. These methods include (1) a new similarity measure for comparing diffusion tensors, (2) a method for interpolating diffusion tensors, and (3) a new anisotropy measure.

6.1 The Space of Diffusion Tensors

Recall that a real $n \times n$ matrix A is symmetric if $A = A^T$ and positive-definite if $x^T Ax > 0$ for all nonzero $x \in \mathbb{R}^n$. We denote the space of all $n \times n$ symmetric, positive-definite matrices as $PD(n)$. The tensors in DT-MRI are thus elements of $PD(3)$. The space $PD(n)$ forms a convex subset of $\mathbb{R}^{n^2}$. One can define a linear average of N positive-definite, symmetric matrices $A_1, \ldots, A_N$ as $\mu = \frac{1}{N} \sum_{i=1}^{N} A_i$. This definition minimizes the Euclidean metric on $\mathbb{R}^{n^2}$. Since $PD(n)$ is convex, μ lies within $PD(n)$. However, linear averages do not interpolate natural properties. The linear average of matrices of the same determinant can result in a matrix with a larger determinant. Second order statistics are even more problematic. The standard principal component analysis is invalid because the straight lines defined by the modes of variation do not stay within the space $PD(n)$. In other words, linear PCA does not preserve the positive-definiteness of diffusion tensors. The reason for such difficulties is that space $PD(n)$, although a subset of a vector space, is not a vector space; for example, the negation of a positive-definite matrix is not positive-definite.

In this chapter we derive a more natural metric on the space of diffusion tensors, $PD(n)$, by viewing it not simply as a subset of $\mathbb{R}^{n^2}$, but rather as a Riemannian symmetric space. Following Fréchet [44], we define the average as the minimum mean squared error estimator under this metric. We apply the method of principal geodesic analysis developed in Chapter 4 to describe the variability of diffusion tensor data. In this framework the modes of variation are represented as flows along geodesic curves, i.e., shortest paths under the Riemannian metric. These geodesic curves, unlike the straight lines of $\mathbb{R}^{n^2}$, are completely contained within $PD(n)$, so they preserve the positive-definiteness.

To illustrate these issues, consider the space $PD(2)$, the 2×2 symmetric, positive-definite matrices. A matrix $A \in PD(2)$ is of the form

$$A = \begin{pmatrix} a & b \\ b & c \end{pmatrix}, \quad ac - b^2 > 0, \quad a > 0.$$

If we consider the matrix A as a point $(a, b, c) \in \mathbb{R}^3$, then the above conditions describe the interior of a cone as shown in Fig. 6.1. The two labeled points are $p_0 = (1, 0, 7), p_1 =$

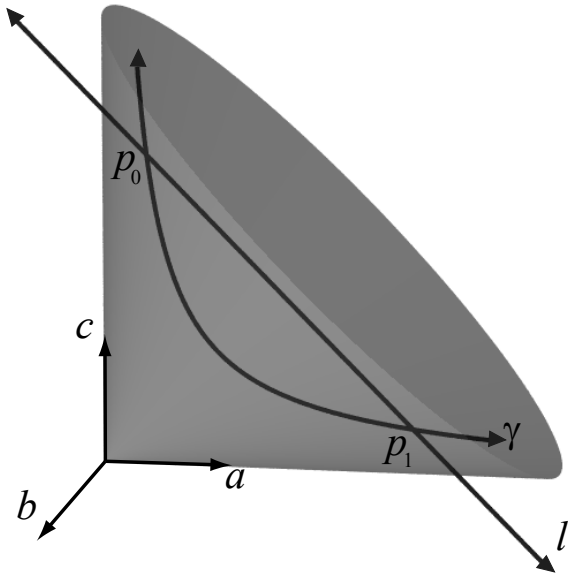

Figure 6.1: The space $PD(2)$, showing the geodesic γ and the straight line l between the two points p_0 and p_1.

$(7, 0, 1)$. The straight line l between the two points, i.e., the geodesic in $\mathbb{R}^{n^2}$, does not remain contained within the space $PD(2)$. The curve γ is the geodesic between the two points when $PD(2)$ is considered as a Riemannian symmetric space. This geodesic lies completely within $PD(2)$. We chose $PD(2)$ as an example since it can be easily visualized, but the same phenomenon occurs for general $PD(n)$, i.e., $n > 2$.

6.2 The Geometry of $PD(n)$

In this section we show that the space of diffusion tensors, $PD(n)$, can be formulated as a Riemannian symmetric space. This leads to equations for computing geodesics that will be essential in defining the statistical methods for diffusion tensors. The differential geometry of diffusion tensors has also been used in [23], where the diffusion tensor smoothing was constrained along geodesic curves. The fact that $PD(n)$ is a symmetric space has been known for some time. In fact, Cartan accomplished a complete classification of the possible symmetric spaces in two papers in 1926 and 1927 [19, 20]. A review of symmetric spaces can be found in [15, 58].

Recall from Theorem 2.8 that a symmetric space is a connected Riemannian manifold M such that for each $x \in M$ there is an isometry σ_x that reverses all geodesics through the point x. We will show that the space $PD(n)$ is a Riemannian symmetric space by constructing a transitive Lie group action on it. This leads to a natural Riemannian

metric on $PD(n)$ that is invariant under the group action. The equations for computing geodesics are then derived from the action of one-parameter subgroups.

6.2.1 The Lie Group Action on $PD(n)$

Consider the Lie group of all $n \times n$ real matrices with positive determinant, denoted $GL^+(n)$. This group acts on $PD(n)$ via

$$\begin{aligned} \phi : GL^+(n) \times PD(n) &\to PD(n) \\ \phi(g, p) &= gpg^T. \end{aligned} \tag{6.1}$$

We will sometimes write the group action as $g \cdot p = \phi(g, p)$.

We will show that this action satisfies the conditions in Theorem 2.10 for $PD(n)$ to be a symmetric space, namely,

1. The action ϕ is transitive.
2. The Lie group $GL^+(n)$ is connected.
3. The resulting isotropy subgroup is compact.
4. There is an involutive automorphism of $GL^+(n)$ leaving the isotropy subgroup fixed.

We prove each of these conditions in turn.

(1) Recall that the group action ϕ is transitive if for any two points $p, q \in PD(n)$, there exists an element $g \in GL^+(n)$ such that $q = \phi(g, p)$. Let I_n denote the $n \times n$ identity matrix. Given a matrix $p \in PD(n)$ let $p = U\Lambda U^T$ be the SVD of p. Then p can be written as the product $p = gg^T$, where $g \in GL^+(n)$ is given by $g = U\Lambda^{(1/2)}$. Therefore, $I_n = \phi(g^{-1}, p)$. Now write $q = hh^T$ for $h \in GL^+(n)$. Then $q = h\phi(g^{-1}, p)h^T = \phi(h, \phi(g^{-1}, p)) = \phi(hg^{-1}, p)$, which shows that ϕ is transitive. In other words, the space $PD(n)$ is a homogeneous space.

(2) Let g_1, g_2 be two matrices in $GL^+(n)$. To show that $GL^+(n)$ is connected, we show that there is a continuous path in $GL^+(n)$ connecting g_1 and g_2. Let $g_1 = U_1\Lambda_1 V_1$ and $g_2 = U_2\Lambda_2 V_2$ be singular value decompositions with $U_i, V_i \in SO(n)$. We can safely assume that the matrices Λ_i have positive diagonal entries. If Λ_i has negative entries,

we can write $g_i = U_i \tilde{\Lambda}_i \tilde{V}_i$, where $\tilde{\Lambda}_i = \Lambda_i R_i$, $\tilde{V}_i = R_i V_i$, and R_i is the diagonal rotation matrix with -1 in the entries corresponding to negative values in the Λ_i and $+1$ in the other entries. Consider the paths $c_i : [0,1] \to GL^+(n)$ for $i = 1, 2$ given by

$$c_i(s) = \exp(s \log U_i)\ \Lambda_i^s\ \exp(s \log V_i).$$

These are continuous paths from $c_i(0) = I_n$ to $c_i(1) = g_i$. Now consider the path $c : [0,1] \to GL^+(n)$ given by

$$c(s) = c_1(1-s)\ c_2(s).$$

This path is continuous with $c(0) = g_1$ and $c(1) = g_2$. Furthermore, for any $s \in [0,1]$ we have

$$\det(c(s)) = \det(\Lambda_1^{(1-s)} \Lambda_2^s) = \det(g_1)^{(1-s)} \det(g_2)^s > 0.$$

This shows that the path c lies completely within $GL^+(n)$, so $GL^+(n)$ is connected.

(3) The isotropy subgroup of I_n under ϕ is the set of all matrices $g \in GL^+(n)$ that satisfy $\phi(g, I_n) = I_n$. Thus, the isotropy subgroup is given by $SO(n) = \{g \in GL^+(n) : gg^T = I_n\}$, the space of $n \times n$ rotation matrices. This is a compact Lie subgroup of $GL^+(n)$ as was mentioned in the background section on Lie groups (Section 2.4).

(4) The mapping $\alpha : GL^+(n) \to GL^+(n)$ given by $\alpha(g) = (g^{-1})^T$ is an involutive automorphism that leaves $SO(n)$ fixed.

Therefore, the space of diffusion tensors, $PD(n)$, is a symmetric space and equivalent to the quotient space $GL^+(n)/SO(n)$. An intuitive way to view this quotient is to think of the polar decomposition, which decomposes a matrix $g \in GL^+(n)$ as $g = pu$, where $p \in PD(n)$ and $u \in SO(n)$. Thus, the diffusion tensor space $PD(n) \cong GL^+(n)/SO(n)$ comes from "dividing out" the rotational component in the polar decomposition of $GL^+(n)$.

6.2.2 The Invariant Metric on $PD(n)$

The space of diffusion tensors, $PD(n)$, has a Riemannian metric that is invariant under the $GL^+(n)$ action, which follows from the fact that the isotropy subgroup $SO(n)$ is connected and compact (recall Theorem 2.9).

The tangent space of $PD(n)$ at the identity matrix can be identified with the space of $n \times n$ symmetric matrices, $\mathrm{Sym}(n)$. Since the group action $\phi_g : s \mapsto gsg^T$ is linear, its derivative map, denoted $d\phi_g$, is given by $d\phi_g(X) = gXg^T$. If $X \in \mathrm{Sym}(n)$, it is easy to see that $d\phi_g(X)$ is again a symmetric matrix. Thus the tangent space at any point $p \in PD(n)$ is also identifiable with $\mathrm{Sym}(n)$. If $X, Y \in \mathrm{Sym}(n)$ represent two tangent vectors at $p \in PD(n)$, where $p = gg^T, g \in GL^+(n)$, then the Riemannian metric at p is given by the inner product

$$\langle X, Y \rangle_p = \mathrm{tr}(g^{-1}Xp^{-1}Y(g^{-1})^T).$$

Finally, the mapping $\sigma_{I_n}(p) = p^{-1}$ is an isometry that reverses geodesics of $PD(n)$ at the identity, and this turns $PD(n)$ into a symmetric space.

6.2.3 Computing Geodesics

Geodesics on a symmetric space are easily derived via the group action (see [58] for details). Let p be a point on $PD(n)$ and X a tangent vector at p. There is a unique geodesic, γ, with initial point $\gamma(0) = p$ and tangent vector $\gamma'(0) = X$. To derive an equation for such a geodesic, we begin with the special case where the initial point p is the $n \times n$ identity matrix, I_n, and the tangent vector X is diagonal. Then the geodesic is given by

$$\gamma(t) = \exp(tX),$$

where exp is the matrix exponential map given by the infinite series

$$\exp(X) = \sum_{k=0}^{\infty} \frac{1}{k!} X^k.$$

For the diagonal matrix X with entries x_i, the matrix exponential is simply the diagonal matrix with entries e^{x_i}.

Now for the general case consider the geodesic γ starting at an arbitrary point $p \in PD(n)$ with arbitrary tangent vector $X \in \mathrm{Sym}(n)$. We will use the group action to map this configuration into the special case described above, i.e., with initial point at the identity and a diagonal tangent vector. Since the group action is an isometry, geodesics and distances are preserved. Let $p = gg^T$, where $g \in GL^+(n)$. Then the action $\phi_{g^{-1}}$ maps p to I_n. The tangent vector is mapped via the corresponding tangent map to $Y = d\phi_{g^{-1}}(X) = g^{-1}X(g^{-1})^T$. Now we may write $Y = v\Sigma v^T$, where v is a rotation

matrix and Σ is diagonal. The group action $\phi_{v^{-1}}$ diagonalizes the tangent vector while leaving I_n fixed. We can now use the procedure above to compute the geodesic $\tilde{\gamma}$ with initial point $\tilde{\gamma}(0) = I_n$ and tangent vector $\tilde{\gamma}'(0) = \Sigma$. Finally, the result is mapped back to the original configuration by the inverse group action, ϕ_{gv}. That is,

$$\gamma(t) = \phi_{gv}(\tilde{\gamma}(t)) = (gv)\exp(t\Sigma)(gv)^T.$$

If we flow to $t = 1$ along the geodesic γ we get the Riemannian exponential map at p. That is,

$$\mathrm{Exp}_p(X) = \gamma(1).$$

In summary we have

Algorithm 6.1: Riemannian Exponential Map
Input: Initial point $p \in PD(n)$.
Tangent vector $X \in \mathrm{Sym}(n)$.
Output: $\mathrm{Exp}_p(X)$
Let $p = u\Lambda u^T$ ($u \in SO(n)$, Λ diagonal)
$g = u\sqrt{\Lambda}$
$Y = g^{-1}X(g^{-1})^T$
Let $Y = v\Sigma v^T$ ($v \in SO(n)$, Σ diagonal)
$\mathrm{Exp}_p(X) = (gv)\exp(\Sigma)(gv)^T$

An important property of the geodesics in $PD(n)$ under this metric is that they are infinitely extendible, i.e., the geodesic $\gamma(t)$ is defined for $-\infty < t < \infty$. This follows from the fact that all symmetric spaces are complete (Theorem 2.8). Again, Fig. 6.1 demonstrates that the symmetric space geodesic γ remains within $PD(2)$ for all t. In contrast the straight line l quickly leaves the space $PD(2)$.

The map Exp_p has an inverse, called the Riemannian log map and denoted Log_p. It maps a point $x \in PD(n)$ to the unique tangent vector at p that is the initial velocity of the unique geodesic γ with $\gamma(0) = p$ and $\gamma(1) = x$. Using a similar diagonalization procedure, the log map is computed by

Algorithm 6.2: Riemannian Log Map
Input: Initial point $p \in PD(n)$.
End point $x \in PD(n)$.
Output: $\text{Log}_p(x)$
Let $p = u\Lambda u^T$ ($u \in SO(n)$, Λ diagonal)
$g = u\sqrt{\Lambda}$
$y = g^{-1}x(g^{-1})^T$
Let $y = v\Sigma v^T$ ($v \in SO(n)$, Σ diagonal)
$\text{Log}_p(x) = (gv)\log(\Sigma)(gv)^T$

Using the notation of Algorithm 6.2, geodesic distance between the diffusion tensors $p, x \in PD(n)$ is computed by

$$d(p, x) = ||\operatorname{Log}_p(x)||_p = \operatorname{tr}(\log(\Sigma)^2). \tag{6.2}$$

6.3 Statistics of Diffusion Tensors

Having formulated the geometry of diffusion tensors as a symmetric space, we now develop methods for computing statistics in this nonlinear space. The algorithms for computing the mean and PGA will be direct adaptations of the algorithms described in Chapter 4 to the space $PD(n)$. The computations for the log and exponent maps described in the previous section will be instrumental in these statistical methods.

6.3.1 Averages of Diffusion Tensors

Again we define the intrinsic mean of a set of diffusion tensors $p_1, \ldots, p_N \in PD(n)$ as the diffusion tensor that minimizes the sum-of-squared distance to the p_i. That is, the intrinsic mean is given by

$$\mu = \arg\min_{p \in PD(N)} \sum_{i=1}^{N} d(p, p_i)^2. \tag{6.3}$$

Again let ρ_A denote the sum-of-squared distance function for the set of points $A = \{p_1, \ldots, p_N\}$, that is,

$$\rho_A(p) = \frac{1}{2N} \sum_{i=1}^{N} d(p, p_i)^2.$$

Recall that the gradient of ρ can be computed as

$$\nabla \rho_A(p) = -\frac{1}{N} \sum_{i=1}^{N} \mathrm{Log}_p(p_i)$$

when the data lie in a strongly convex neighborhood. In fact, the entire space $PD(n)$ is strongly convex; that is, any two points can be connected by a unique geodesic curve. This fact follows from the curvature properties of $PD(n)$ (it has everywhere non-positive sectional curvature). Therefore, the minimization problem for the intrinsic mean (6.3) has a unique solution. The gradient descent algorithm for computing the mean of diffusion tensors, which is a direct adaptation of Algorithm 4.1, is given by

Algorithm 6.3: Intrinsic Mean of Diffusion Tensors
Input: $p_1, \ldots, p_N \in PD(n)$
Output: $\mu \in PD(n)$, the intrinsic mean
$\mu_0 = I$
Do
$X_i = \frac{1}{N} \sum_{k=1}^{N} \mathrm{Log}_{\mu_i}(p_k)$
$\mu_{i+1} = \mathrm{Exp}_{\mu_i}(X_i)$
While $||X_i|| > \epsilon$.

6.3.2 Principal Geodesic Analysis of Diffusion Tensors

We are now ready to define principal geodesic analysis for diffusion tensor data $p_1, \ldots, p_N \in PD(n)$. Our goal, analogous to PCA, is to find a sequence of nested geodesic submanifolds that maximize the projected variance of the data, according to the defining equations (4.6) and (4.7) presented in Chapter 4. We must again determine under what conditions the projection operator (4.5) for the space $PD(n)$ is unique. That is, we must determine the neighborhood U used in the PGA equations (4.6) and (4.7). Actually, since $PD(n)$ has non-positive sectional curvature, the projection operator is well-defined for the entire space $U = PD(n)$.

We will use the tangent space approximation to the projection operator. That is, we will adapt Algorithm 4.2 for the tangent space approximation to PGA, giving the following algorithm on $PD(n)$:

Algorithm 6.4: PGA of Diffusion Tensors
Input: $p_1, \ldots, p_N \in PD(n)$
Output: Principal directions, $v_k \in \mathrm{Sym}(n)$
Variances, $\lambda_k \in \mathbb{R}$
$\mu =$ intrinsic mean of $\{p_i\}$ (Algorithm 6.3)
$x_i = \mathrm{Log}_\mu(p_i)$
$\mathbf{S} = \frac{1}{N}\sum_{i=1}^{N} x_i x_i^T$ (treating the x_i as column vectors)
$\{v_k, \lambda_k\} =$ eigenvectors/eigenvalues of $\mathbf{S}$.

A new diffusion tensor p can now be generated from the PGA by the formula $p = \mathrm{Exp}_\mu\left(\sum_{k=1}^{d} \alpha_k v_k\right)$, where the $\alpha_k \in \mathbb{R}$ are the coefficients of the modes of variation.

6.4 Properties of PGA on $PD(n)$

We now demonstrate that PGA on the symmetric space $PD(n)$ preserves certain important properties of the diffusion tensor data, namely the properties of positive-definiteness, determinant, and orientation. This makes the symmetric space formulation an attractive approach for the statistical analysis of diffusion tensor images. We have already mentioned that, in contrast to linear PCA, symmetric space PGA preserves positive-definiteness. That is, the principal geodesics are completely contained within $PD(n)$, and any matrix generated by the principal geodesics will be positive-definite.

The next two properties we consider are the determinant and orientation. Consider a collection of diffusion tensors that all have the same determinant D. We wish to show that the resulting average and any tensor generated by the principal geodesic analysis will also have determinant D. To show this we first look at the subset of $PD(n)$ of matrices with determinant D, that is, the subset $P_D = \{p \in PD(n) : \det(p) = D\}$. This subset is a **totally geodesic submanifold**, meaning that any geodesic within P_D is a geodesic of the full space $PD(n)$. Recall in Chapter 4 we discussed submanifolds geodesic at a point p, i.e., submanifolds whose geodesics passing through p were also geodesics of the ambient manifold. This is different from totally geodesic submanifolds, which are submanifolds geodesic at *every* point. Now, the fact that P_D is totally geodesic implies that the averaging process in Algorithm 6.3 will remain in P_D if all the data lies in P_D. Also, the principal directions v_k in the PGA will lie in the tangent subspace $T_\mu P_D$. Thus any diffusion tensor generated by the principal geodesics will remain in the space P_D.

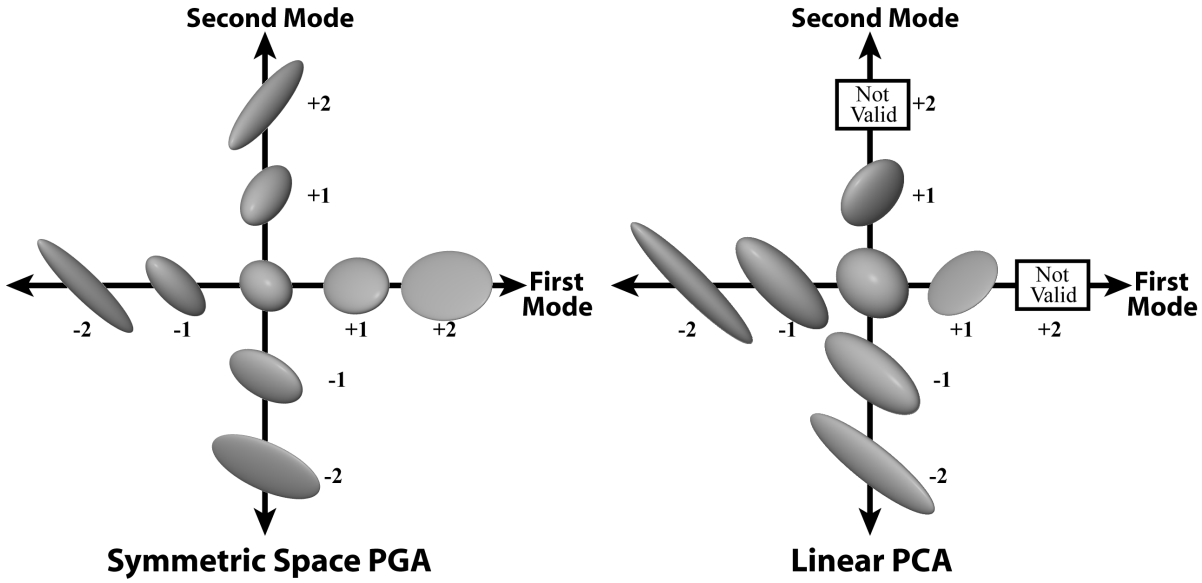

Figure 6.2: The first two modes of variation of the simulated data: (left) using the symmetric space PGA, and (right) using linear PCA. Units are in standard deviations. The boxes labelled "Not Valid" indicate that the tensor was not positive-definite, i.e., it had negative eigenvalues.

The same argument may be applied to show that symmetric space averaging and PGA preserve the orientation of diffusion tensors. In fact, the subset of all diffusion tensors having the same orientation is also a totally geodesic submanifold, and the same reasoning applies. Unlike the positive-definiteness and determinant, orientations are also preserved by linear averaging and PCA.

To demonstrate these properties, we simulated random 3D diffusion tensors and computed both their linear and symmetric space statistics. We first tested the determinant preservation by generating 100 random 3D diffusion tensors with determinant 1. To do this we first generated 100 random 3×3 symmetric matrices, with entries distributed according to a normal distribution, $N(0, \frac{1}{2})$. Then, we took the matrix exponential of these random symmetric matrices, thus making them positive-definite diffusion tensors. Finally, we normalized the random diffusion tensors to have determinant 1 by dividing each tensor by the cube root of its determinant. We then computed the linear average and PCA and symmetric space average and PGA of the simulated tensors. The results are shown in Fig. 6.2 as the diffusion tensors generated by the first two modes of variation. The linear PCA generated invalid diffusion tensors, i.e., tensors with negative eigenvalues, at +2 standard deviations in both the first and second modes. All of the diffusion tensors generated by the symmetric space PGA have determinant 1. The linear

mean demonstrates the "swelling" effect of linear averaging. It has determinant 2.70, and the linear PCA tensors within ± 2 standard deviations have determinants ranging from -2.80 to 2.82. The negative determinants came from the tensors that were not positive-definite. Therefore, we see that the symmetric space PGA has preserved the positive-definiteness and the determinant, while the linear PCA has preserved neither.

Next we tested the orientation preservation by generating 100 random, axis-aligned, 3D diffusion tensors. This was done by generating 3 random eigenvalues for each matrix, corresponding to the x, y, and z axes. The eigenvalues were chosen from a log-normal distribution with log mean 0 and log standard deviation 0.5. Next we generated a random orientation $u \in SO(3)$ and applied it to all of the axis-aligned matrices by the map $p \mapsto upu^T$. Thus each of the diffusion tensors in our test set had eigenvectors equal to the columns of the rotation matrix u. We computed both the symmetric space and linear statistics of the data. As was expected, both methods preserved the orientations. However, the linear PCA again generated tensors that were not positive-definite.

6.5 New Methods: Comparison Metric, Interpolation, and Anisotropy

In this section we present several novel methods for the analysis of diffusion tensors based on the symmetric space formulation of $PD(n)$ presented earlier. As mentioned at the beginning of this chapter, a primary application of the statistical methods presented above is for inter-subject studies of diffusion tensor data. To make such studies possible, images of different patients need to be registered into a common coordinate system for direct comparison. The first two methods in this section are intended to be used as part of a registration method for diffusion tensor images. The third method is a new anisotropy measure based on the differential geometry of the symmetric space $PD(n)$.

An algorithm for diffusion tensor image registration requires three important tools: (1) a method for warping a diffusion tensor image, (2) a method for resampling the warped diffusion tensor image, i.e., an interpolation method for diffusion tensors, and (3) a comparison metric for computing how close two diffusion tensor images are to one another. Warping diffusion tensor images is nontrivial because it is not obvious how a warp of space should affect the shape and orientation of a tensor. Alexander *et al.* [1] describe several strategies for warping diffusion tensor images. We focus on items (2) and (3) and present new methods for interpolating and comparing diffusion tensors based on the symmetric space formulation of $PD(n)$. We begin with the comparison

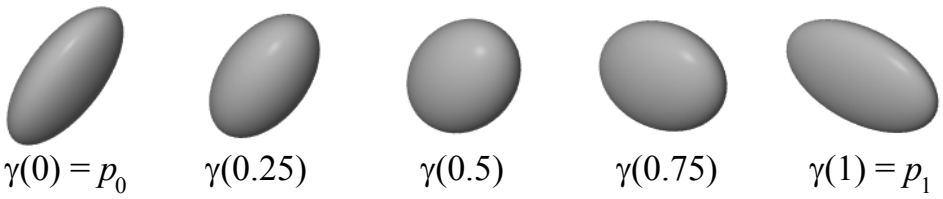

Figure 6.3: An example of geodesic interpolation of two diffusion tensors. First, two diffusion tensors p_0 and p_1 were chosen randomly. The diffusion tensors at times $0.25, 0.5$, and 0.75 were generated along the unique geodesic segment γ between p_0 and p_1.

metric.

6.5.1 Comparison Metric

We propose a new comparison metric for diffusion tensors defined as the geodesic distance between two tensors. That is, given two diffusion tensors $p_1, p_2 \in PD(3)$ the error metric between them is given by

$$E(p_1, p_2) = d(p_1, p_2),$$

where the distance d is the geodesic distance given by (6.2). For two diffusion tensor images $I_1, I_2 : \Omega \to PD(3)$, where $\Omega \subset \mathbb{R}^3$ is the image domain, the error metric is given by

$$E(I_1, I_2) = \left(\int_\Omega d(I_1(x), I_2(x))^2 dx \right)^{\frac{1}{2}}.$$

Alexander *et al.* [1] have proposed comparing diffusion tensors based on the angular difference between their principal directions. This difference is then weighted by the relative anisotropy to give higher weight to the more anisotropic tensors. We argue that the geodesic error metric presented here takes into account both the orientation and the anisotropy. In addition, the geodesic error metric is consistent with the proposed statistical methods; that is, it is based on geodesic distance on $PD(n)$ as are the mean and PGA methods presented above.

6.5.2 Diffusion Tensor Interpolation

The most basic method for resampling a warped image is a nearest neighbor approach. Another possibility is to use trilinear interpolation of the linear tensor coefficients. The

tensor interpolation method that we propose is based on the symmetric space averaging method developed in Section 6.3. First, consider the case of two diffusion tensors $p_1, p_2 \in PD(n)$. We would like an interpolation method given by a continuous curve $c : [0,1] \to PD(n)$ satisfying $c(0) = p_1$ and $c(1) = p_2$. Given the symmetric space formulation for $PD(n)$ presented above, an obvious choice for c is the unique geodesic curve segment between p_1 and p_2. This geodesic interpolation is demonstrated between two randomly chosen diffusion tensors in Fig. 6.3. Geodesic interpolation can be seen as a direct generalization of linear interpolation for scalar or vector data.

Now for 3D images of diffusion tensors an interpolation method can be thought of as a smooth function in a cube, where the tensor values to be interpolated are given at the corners of the cube. In other words, we want a smooth function $f : [0,1]^3 \to PD(n)$, where the values $f(i,j,k) : i,j,k \in \{0,1\}$ are specified. It is tempting to first create f using "tri-geodesic" interpolation, that is, by repeated geodesic interpolation in the three coordinate directions. However, unlike linear interpolation, geodesic interpolation of diffusion tensors does not commute. Therefore, a "tri-geodesic" interpolation would be dependent on the order in which the coordinate interpolations were made. A better method for interpolating diffusion tensors in three dimensions is using a weighted geodesic average.

Weighted averaging of data on an sphere S^n has been studied by Buss and Fillmore [18]. We follow their approach, extending the definition of weighted averages to diffusion tensors. Given a set of diffusion tensors $p_1, \ldots, p_N \in PD(n)$ and a set of weights $w_1, \ldots, w_N \in \mathbb{R}$, consider the weighted sum-of-squared distances function

$$\rho(p; p_1, \ldots, p_N; w_1, \ldots, w_N) = \frac{1}{N} \sum_{i=1}^{N} w_i d(p, p_i)^2.$$

Given a set of non-negative real weights $w_1, \ldots, w_N$ with sum equal to 1, the **weighted average** of the p_i with respect to the weights w_i is defined as a minimum of the weighted sum-of-squared distances function, i.e.,

$$\text{Avg}(p_1, \ldots, p_N; w_1, \ldots, w_N) = \arg \min_{p \in PD(n)} \rho(p; p_1, \ldots, p_N; w_1, \ldots, w_N). \tag{6.4}$$

The intrinsic mean definition given in Chapter 4 is equivalent to weighted average definition with all weights set to $w_i = (1/N)$. For vector-valued data $v_1, \ldots, v_N \in \mathbb{R}^n$ the

weighted average is given by the weighted sum

$$\text{Avg}(v_1, \ldots, v_N; w_1, \ldots, w_N) = \sum_{i=1}^{N} w_i v_i.$$

For diffusion tensor data the weighted average can be computed using a generalization of the intrinsic mean algorithm (Algorithm 6.3). The gradient of the sum-of-squared distances function is given by

$$\nabla \rho(p; p_1, \ldots, p_N; w_1, \ldots, w_N) = -\sum_{i=1}^{N} w_i \, \text{Log}_p(p_i).$$

Therefore, the gradient descent algorithm for finding the weighted average of a set of diffusion tensors is given by

Algorithm 6.5: Weighted Average of Diffusion Tensors
Input: $p_1, \ldots, p_N \in PD(n)$ and weights $w_1, \ldots, w_N \in \mathbb{R}$
Output: $\mu \in PD(n)$, the weighted average
$\mu_0 = I$
Do
$X_i = \frac{1}{N} \sum_{k=1}^{N} w_i \, \text{Log}_{\mu_i}(p_k)$
$\mu_{i+1} = \text{Exp}_{\mu_i}(X_i)$
While $||X_i|| > \epsilon$.

We will replace the cumbersome Avg with the more convenient notation

$$\bigodot_{i=1}^{N} w_i \cdot p_i = \text{Avg}(p_1, \ldots, p_N; w_1, \ldots, w_N).$$

The circle around the summation sign is intended to remind the reader that this is a weighted average of nonlinear data and not a linear sum.

Returning to the problem of finding an interpolating function for diffusion tensors in a volume image, we want to define our interpolating function $f : [0,1]^3 \to PD(n)$, where the values at the corners are given. Let $A = \{0,1\}^3$, and let $\alpha = (\alpha_1, \alpha_2, \alpha_3) \in A$ be a multi-index for the eight corners of the unit cube. Let $p_\alpha \in PD(n)$ be a set of diffusion tensors given at the corners of the unit cube. We define the **geodesic weighted**

interpolation of the p_α as the function $f : [0,1]^3 \to PD(n)$ via a weighted average

$$f(x_1, x_2, x_3) = \bigodot_{\alpha \in A} w_\alpha(x_1, x_2, x_3) \cdot p_\alpha, \tag{6.5}$$

where the $w_\alpha : [0,1]^3 \to \mathbb{R}$ are weight functions on the unit cube. These weight functions should satisfy the following properties:

1. For a corner index β the weights should give $w_\alpha(\beta) = 1$ if $\alpha = \beta$ and zero otherwise. This ensures that the function f indeed interpolates the corner values.

2. The weights should be chosen so that along any edge of the unit cube the interpolation function f gives a one-dimensional geodesic interpolation.

3. The interpolation function should be the same on each face of the unit cube, and it should depend only on the four corners of that face.

4. The weight function should be chosen so that the interpolation function $f : [0,1]^3 \to PD(n)$ given by (6.5) is a continuous function.

We define a set of weight functions by the polynomials

$$w_\alpha(x_1, x_2, x_3) = \prod_{i=1}^{3} \left(1 - \alpha_i + (-1)^{1-\alpha_i} x_i\right).$$

It is straightforward to check that these weights satisfy properties 1-3. The reader can check that these polynomials in fact give the standard tri-linear interpolation of scalar data $v_\alpha : \alpha \in A$ on the unit cube via the interpolating function

$$f(x_1, x_2, x_3) = \sum_{\alpha \in A} w_\alpha(x_1, x_2, x_3) v_\alpha.$$

Therefore, the diffusion tensor interpolation function (6.5) is a direct generalization of tri-linear interpolation for scalar data.

The continuity of the interpolation function (property 4) follows as a corollary of the next theorem. This theorem follows the analogous theorem for weighted averages of spherical data shown by Buss and Fillmore [18] (see Theorem 6).

Theorem 6.1. *The weighted average function* $\mathrm{Avg} : PD(n)^N \times \mathbb{R}^N \to PD(n)$ *given by* (6.4) *is a* C^∞ *function.*

Proof. This theorem is a direct application of the Implicit Function Theorem. The weighted average function Avg maps a set of diffusion tensors and weights to a root of the gradient of the sum-of-squared distance function, $\nabla\rho$. Let $p_1, \ldots, p_N \in PD(n)$ and $w_1, \ldots, w_N \in \mathbb{R}$ be a set of diffusion tensors and non-negative weights. The function $\nabla\rho$ is a C^∞ function of the points $p_1, \ldots, p_N$ and the weights $w_1, \ldots, w_N$, and its Jacobian matrix is given by the Hessian matrix H of ρ. According to Karcher [67], the Hessian H is positive-definite because $PD(n)$ has nonpositive sectional curvatures. Therefore, it is nonsingular, and $\nabla\rho$ satisfies the conditions of the Implicit Function Theorem, which now says that there must be an open neighborhood of $(p_1, \ldots, p_N; w_1, \ldots, w_N)$ in which Avg is a C^∞ mapping. □

Corollary 6.2. *The interpolation function $f : [0,1]^3 \to PD(n)$ given by (6.5) is a C^∞ function.*

Proof. Since the function f is a composition of the weight functions w_α and the weighted average function Avg, the fact that f is C^∞ follows from Theorem 6.1 and the fact that the w_α are polynomials. □

The weighted geodesic interpolation function is well-defined for any initial diffusion tensor values p_α, and it does not depend on any arbitrary choice of ordering as did the "tri-geodesic" method. Another important property of weighted geodesic interpolation is that it preserves determinants and orientations of the initial data. That is, if the p_α all have the same determinant (respectively, orientation), then any tensor interpolated by 6.5 will also have the same determinant (orientation). This follows from the same argument given in the previous section to show that the intrinsic mean preserves these properties. That is, if the data lie in the same totally geodesic submanifold (the submanifold representing diffusion tensors with the same determinant or the same orientation), the weighted average of the data will lie in the same submanifold. Since the weighted geodesic interpolation is defined via weighted averages, it follows that it also preserves determinants and orientations.

6.5.3 Geodesic Anisotropy Measure

We now develop a new anisotropy measure for diffusion tensors based on the geodesic distance on the symmetric space $PD(3)$. Anisotropy is intuitively a measure of how far away a diffusion tensor is from being isotropic. Therefore, a natural measurement of the anisotropy of a diffusion tensor $p \in PD(3)$ is the geodesic distance between p and the

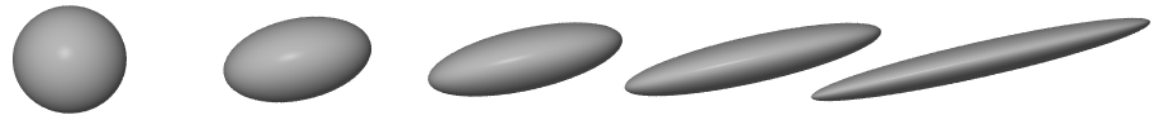

Figure 6.4: The geodesic anisotropy values for a sequence of diffusion tensors. The GA values are displayed as a function of t: $GA(t) = GA(\gamma(t))$.

closest isotropic diffusion tensor. It turns out that the nearest isotropic diffusion tensor to p is the one with the same determinant as p, i.e., the matrix $\det(p)^{\frac{1}{3}} \cdot I_3$. Thus we define the **geodesic anisotropy** as

$$GA(p) = d(\det(p)^{\frac{1}{3}} \cdot I_3, p). \tag{6.6}$$

To better understand the meaning of the geodesic anisotropy, it helps to write an explicit equation for it. Let λ_i denote the eigenvalues of p, and let $\overline{\log \lambda}$ denote the average of the logs of the eigenvalues. The geodesic anisotropy of p can be written as

$$\begin{aligned} GA(p) &= d(\det(p)^{\frac{1}{3}} \cdot I_3, p) \\ &= d(I_3, \frac{1}{\det(p)^{\frac{1}{3}}} \cdot p) \\ &= \left(\sum_{i=1}^{3} \left\| \log \left(\frac{\lambda_i}{(\lambda_1 \lambda_2 \lambda_3)^{\frac{1}{3}}} \right) \right\|^2 \right)^{\frac{1}{2}} \\ &= \left(\sum_{i=1}^{3} \| \log(\lambda_i) - \overline{\log \lambda} \|^2 \right)^{\frac{1}{2}}. \end{aligned}$$

The second line follows from the invariance of distance on $PD(n)$ under the group action of $GL^+(n)$. This shows that the geodesic anisotropy is equivalent to the standard deviation of the log of the eigenvalues (times a scale factor). This is similar to how the fractional anisotropy is defined via the standard deviation of the eigenvalues, which are treated as linear entities. The GA is consistent with the thinking of $PD(n)$ as a symmetric space, where the eigenvalues are treated as multiplicative entities rather than linear ones.

An example of the GA values for a sequence of diffusion tensors is shown in Fig 6.4. The diffusion tensors were generated along a geodesic starting at the identity matrix.

The sequence is given by the formula

$$\gamma(t) = \mathrm{Exp}_{I_3} tX, \qquad \text{where} \quad X = \begin{pmatrix} 1 & 0.5 & 0 \\ 0.5 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

The particular direction X for the geodesic was chosen for simplicity and so that the tensors would get increasingly more anisotropic.

6.6 Conclusions

In this chapter we showed how the statistical methods of averaging and principal geodesic analysis can be applied to study diffusion tensor data. We began by showing that the space of diffusion tensors $PD(n)$ is a symmetric space, and we gave the formulas for the Riemannian log and exponential map. We then adapted the mean and principal geodesic analysis algorithms from Chapter 4 to the space $PD(n)$. We showed that these statistical operations preserve natural properties of the diffusion tensor including positive-definiteness, determinant, and orientation. This is in contrast to linear statistical methods, i.e., linear averages and PCA, which do not preserve the positive-definiteness or the determinant. We presented several new methods for analyzing diffusion tensor data based on the symmetric space formulation. These include (1) a new comparison metric that could be used as a metric to optimize in a registration method for diffusion tensor images, (2) a new interpolation method based on weighted averaging that is a natural extension of trilinear interpolation of scalar images, and (3) a new anisotropy measure for diffusion tensors that is given by the distance of a diffusion tensor from the nearest isotropic tensor.

The work in this chapter provides tools for analyzing diffusion tensor data, but there is more work to be done in order to bring these methods into use:

1. A registration procedure for diffusion tensor images can be developed using the interpolation method and the comparison metric presented in this chapter. The transformations of the diffusion tensor images could be based on previous registration methods such as that of Alexander *et al.* [1], or new methods could be developed for transforming one diffusion tensor image into another.

2. The principal geodesic analysis of diffusion tensor data presented in this chapter uses the tangent space approximation to PGA. Perhaps better statistics would

result from a method that computes the exact PGA, which would require that the projection operator on $PD(n)$ be solved explicitly.

3. The driving problem of this work is to make possible the statistical studies of diffusion tensor images across patient populations. This first requires that a registration method for diffusion tensor images be developed as described above. Then the statistical methods developed in this chapter could be used to describe the variability of the diffusion tensor images.

This issues will be discussed further in the future work section of Chapter 7.

Chapter 7

Discussion and Future Work

This chapter reviews and discusses the contributions of this dissertation in Section 7.1. This is followed by a discussion in Section 7.2 of future work, including unsolved theoretical questions, goals for future research, and possible new areas of application of the presented methods.

7.1 Summary of Contributions

This section revisits the thesis and claims laid out in Chapter 1 and presented in Chapters 4,5, and 6. Each contribution is restated along with a discussion of how it was accomplished in this dissertation.

1. *A novel theory called principal geodesic analysis has been developed as a natural generalization of principal component analysis for describing the statistical variability of geometric data that are parameterized as curved manifolds. This generalization is natural in the sense that it uses only intrinsic distances and geodesics in the data space.*

   Principal geodesic analysis was introduced in Chapter 4. The underlying philosophy in this theory is that statistics of manifold-valued data should be *intrinsic* measurements. In other words, statistics should rely only on the intrinsic geometry of the manifold, namely, geometry that can be derived from the Riemannian metric, including geodesic curves and distances. The definition of a mean value that is used is from Fréchet [44] and is defined as the point that minimizes the expected value of the sum-of-squared distance function. Again, the distances used are intrinsic to the manifold, i.e., geodesic distances.

The definition of principal geodesic analysis was a direct generalization of principal component analysis to the manifold case. In other words, if the data lies in a linear manifold, i.e., $\mathbb{R}^n$, principal geodesic analysis reduces to principal component analysis. Principal geodesic analysis was defined via a sequence of nested submanifolds generated by the Riemannian exponential map at the mean. These geodesic submanifolds were chosen to maximize the variance of the data projected onto the submanifold. The important property of the geodesic submanifolds was that they preserve the intrinsic notion of geodesic distance from the mean. An alternative definition of principal geodesic analysis was given that instead minimized the sum-of-squared geodesic distances from the submanifolds to the data.

An algorithm for the computation of a tangent space approximation to principal geodesic analysis was given. This algorithm used the Riemannian log map to map the data into the tangent space to the mean. Then a principal component analysis was computed in the tangent space, and the resulting principal directions were mapped back into the manifold to give the approximate principal geodesic submanifolds. Though this approximation algorithm does its computations in the tangent space, the modes of variation are mapped back to valid points on the manifold.

2. *It has been shown that medial representations of shape, or m-reps, can be formulated as elements of a Riemannian symmetric space and that the variability of a population of m-rep objects can be efficiently computed using principal geodesic analysis.*

   In Chapter 5 it was shown that an order 1 medial atom can be parameterized as a point on a Riemannian symmetric space $\mathcal{M}(1)$. It then followed that m-rep meshes containing n order 1 medial atoms can be represented as the symmetric space $\mathcal{M}(n)$, which is the direct product of n copies of $\mathcal{M}(1)$. The intrinsic definitions for the mean and principal geodesic analysis were applied to the space $\mathcal{M}(n)$. This led to the development of algorithms for computing the mean and the principal geodesic analysis of a collection of m-rep models.

3. *A new method for aligning m-reps to a common position, orientation and scale has been developed and demonstrated. It generalizes the Procrustes alignment method for aligning linear representations of shape. It proceeds by minimizing the sum-of-square geodesic distances between corresponding atoms in medial models.*

The m-rep alignment method presented in Chapter 5 was developed as a generalization of the Procrustes alignment algorithm for point set shape models. Keeping with the philosophy of this dissertation, the m-rep alignment was defined via intrinsic distances on the m-rep symmetric space $\mathcal{M}(n)$. Alignment was achieved by minimizing sum-of-squared geodesic distances between m-rep models with respect to translation, rotation, and scale of the models.

4. *A method for maximum posterior segmentation of 3D medical images via deformable m-reps models using principal geodesic analysis has been developed. The optimization of the objective function in the segmentation uses the principal geodesic modes of variation as a parameter space. A geometric prior based on principal geodesic analysis has been developed and incorporated into a Bayesian objective function.*

   Principal geodesic analysis was incorporated into a deformable m-reps model segmentation method in Chapter 5. The first aspect of this approach was to use the geodesic modes of variation as a method for deforming the initial mean model. The optimization of the posterior objective function can thus proceed by optimizing over the components in the principal geodesic analysis. The second aspect of this approach was to use the geodesic Mahalanobis distance as the geometric prior term in the objective function.

5. *It has been shown that diffusion tensors can be treated as data in a Riemannian symmetric space and that the variability of diffusion tensor data can be described using principal geodesic analysis.*

   It was shown in Chapter 6 that diffusion tensors, i.e., symmetric, positive-definite matrices, are elements of a Riemannian symmetric space. It was argued that the symmetric space formulation of diffusion tensors is preferred to treating them as a linear space. This is because the Riemannian metric is complete in the symmetric space formulation and not in the linear case, which causes geodesics in the symmetric space to extend indefinitely while the geodesics in the linear case, i.e., straight lines, "fall off" the space.

   The intrinsic definitions for the mean and principal geodesic analysis were applied the symmetric space $PD(n)$ of $n \times n$ diffusion tensors, with the case $n = 3$ representing the tensors of DT-MRI. It was shown that the mean and the principal geodesic analysis preserved three important properties of the diffusion tensor: the

positive-definiteness, the determinant, and the orientation. A linear PCA, treating the diffusion tensor space as a subset of the linear vector space $\mathbb{R}^{n^2}$, was shown to not preserve the positive-definiteness or the determinant (although it does preserve the orientation).

6. *New methods for comparing the similarity of diffusion tensor images, interpolating diffusion tensors, and measuring the anisotropy of diffusion tensors have been developed using the symmetric space formulation of the space of diffusion tensors.*

   Chapter 6 developed several new methods for analyzing diffusion tensor data based on the geometry of the symmetric space of diffusion tensors. These methods were designed for the applications of registration of diffusion tensor images and intersubject statistical studies of DT-MRI. The first method was a comparison metric for two diffusion tensor images based on the geodesic distances between corresponding diffusion tensors. This metric treated the difference between two images as an image of tangent vectors to $PD(n)$, and the metric was just the L^2 norm of this difference.

   The second method was a new interpolation scheme for 3D diffusion tensor images. It was shown that a naive generalization of trilinear interpolation does not work for geodesic interpolation because it matters what order the interpolation in x, y, and z is carried out. An algorithm treating interpolation as a weighted averaging was developed that was not dependent on order. This algorithm was shown to be a natural generalization of trilinear interpolation of scalar data.

   The third method was a new anisotropy measure for diffusion tensors defined as the geodesic distance of a tensor from the closest isotropic tensor. It was shown that this measurement could be computed in closed form as the standard deviation of the log of the eigenvalues of the diffusion tensor. This approach to anisotropy keeps with the idea of using intrinsic geometric measurements to analyze manifold data.

Finally, the thesis statement presented in Chapter 1 is revisited.

Thesis: *Principal geodesic analysis is a natural generalization of principal component analysis for describing the statistical variability of geometric data that are parameterized as curved manifolds. Such manifolds include medial representations of shape and diffusion tensors. Principal geodesic analysis can be used to parameterize the shape variability of a population of m-rep models. The resulting probabilities can be effectively used*

as a statistical geometric prior in a deformable m-rep model segmentation of 3D medical images.

The first claim showed that principal geodesic analysis is a natural way to analyze the variability of manifold data. The important property that was stressed is that all computations involve intrinsic geometric properties of the underlying manifold. Also, the statistical definitions are all natural generalizations of linear methods, so linear averages and PCA are special cases of manifold averages and PGA, with $\mathbb{R}^n$ being the particular manifold. Claims 2 and 3 showed that shape analysis of medial representations is possible using principal geodesic analysis, and claim 4 applied the statistical methods as a geometric prior in deformable m-rep segmentation. It was shown in claims 5 and 6 that diffusion tensors are preferably treated as a nonlinear Riemannian symmetric space and that statistical analysis of diffusion tensors is made possible with the averaging and PGA methods presented in this dissertation. The final conclusion is that certain geometric entities, including m-reps and diffusion tensors, are best represented as points on a nonlinear manifold and that principal geodesic analysis is an effective way of describing the variability of these entities.

7.2 Future Work

This section proposes several extensions to the current work and possibilities for future research. Some of these ideas were alluded to in Chapters 4, 5, and 6. It is divided into four sections: Section 7.2.1 describes open theoretical questions about PGA, Sections 7.2.2 and 7.2.3 proposes several extensions and future research involving the statistical analysis of m-reps and DT-MRI, and Section 7.2.4 discusses other application areas outside of m-rep and DT-MRI that may benefit from the statistical methods presented in this dissertation.

7.2.1 Theoretical Questions

Three major questions in the theory of principal geodesic analysis remain to be answered. These questions were alluded to in Chapter 4.

1. *Is the maximum variance definition of PGA equivalent to the least-squares definition?* Principal component analysis can be defined as the linear subspaces that either maximize the variance of the projected data or minimize the sum-of-squared distance to the data (see Section 3.1.3). These two definitions result in the same linear subspaces

because of the Pythagorean theorem. The Pythagorean law does not hold for right triangles on general manifolds. Therefore, it does not seem promising that the maximum variance and the least-squares definitions will coincide for general manifolds.

There is some hope in special cases such as spheres S^n, which have constant positive curvature, and hyperbolic spaces H^n, which have constant negative curvature. These spaces come with a variant of the Pythagorean theorem for a right triangle with side lengths a, b, and hypotenuse c. On spheres the trigonometry for right triangles leads to the identity

$$(\cos a)(\cos b) = \cos c,$$

and on hyperbolic spaces the identity becomes

$$(\cosh a)(\cosh b) = \cosh c.$$

A Taylor series expansion for cos and cosh in the two formulas above shows that up to the second-order terms the above equations are just the standard Pythagorean formula $a^2 + b^2 = c^2$. In other words, the Pythagorean theorem holds infinitesimally. It might be possible using these trigonometric laws to show that the two definitions for PGA are equivalent, or at least show that the difference between the two answers is very small.

2. *Is the recursive approach equivalent to finding each principal geodesic submanifold independently?* Recall that PGA was defined as a recursive procedure finding a sequence of orthonormal vectors $v_1, \ldots, v_d$ in the tangent space to the mean value, $T_\mu M$. These vectors defined subspaces of the tangent space $V_k = \text{span}(\{v_1, \ldots, v_k\})$, which in turn defined the principal geodesic submanifolds $H_k = \text{Exp}_\mu(V_k \cap U)$. However, one could imagine rather defining each V_k independently as a minimization problem over all k-dimensional subspaces of $T_\mu M$. Either definition leads to the same result for PCA in linear spaces. Again, the proof relies on the Pythagorean property of Euclidean space. Therefore, the same comments from the previous question apply. It might be possible to use the constant curvature trigonometric laws to show that these two conditions are equal (or close to equal), but it is unlikely that this is true for general manifolds.

3. *If each principal geodesic submanifold is found independently, are they nested?* This is equivalent to asking if the generating subspaces of the tangent space are nested, i.e., $V_k \subset V_{k+1}$. Of course if the answer to question 2 is "yes", the geodesic submanifolds will be nested. However, it is possible for the answer to question 2 to be "no" and for the geodesic submanifolds to still be nested. If the principal geodesic submanifolds are not

nested, then a full k-frame would need to be generated and saved to define each subspace V_k. Things become easier with the nesting property because only one additional vector needs to be added to build V_k from the previous subspace V_{k-1}; i.e., only one vector needs to be saved per subspace.

7.2.2 M-rep Extensions

For m-rep models the PGA algorithm used was a tangent space approximation (see Chapter 5). It should be possible to compute the PGA exactly according to the recursive definition given in Chapter 4. This requires that the projection operator be solved either explicitly or as a minimization problem. If it can be solved explicitly and its derivatives computed, then a gradient descent algorithm could be developed. For the first principal direction this would be a minimization over all possible unit vectors in the tangent space $T_\mu M$, which is the sphere S^n. For the remaining principal directions, v_k, the minimization is over all unit vectors orthogonal to the previous ones, which is a minimization over the lower-dimensional spheres S^{n-k+1}.

Another area for future work is extending the statistical analysis of m-reps beyond just the global figure stage. M-reps are multiscale representations of geometry, and the scale levels of multiple objects, multiple figures, individual atoms, and dense boundary points are not handled in this dissertation. The work of Lu *et al.* [80] is a first step defining the geometry statistics for these scale levels. The basic idea is to treat the geometric statistics in a Markov random field framework where the deformation at each scale level is a residual from the deformation at the previous (next coarsest) scale level. The statistical model is based on statistics of residues in the shape across scales and between neighbor primitives defined at the current scale level.

The effectiveness of PGA in the deformable m-reps segmentation process needs to be further validated. An experiment testing the difference in segmentations using PGA versus segmentations without PGA is currently being designed. It is expected that PGA in the figure stage will bring the model closer to the correct segmentation, making the changes necessary in the atom stage and boundary displacement stage much smaller. This should both improve the quality of the segmentation and reduce the time spent in the more expensive atom stage. There is also more work to be done to make the segmentation process a true posterior optimization. For this to be the case, the image match measure must be a true likelihood probability $p(I|\mathbf{M})$.

As mentioned at the end of Chapter 5, a medial shape space can be constructed similar to how the Kendall point set shape space is constructed. This would be the quotient

space identifying all m-rep models that are equivalent up to a similarity transformation, i.e., the space $\mathcal{M}(n)/\mathrm{Sim}(n)$. This space, like Kendall's shape space Σ_n^k discussed in Section 3.1, is a stratified set (a manifold with singularities). The smooth parts of this space correspond to medial meshes that are not degenerate. It is future research to characterize this space fully, both the topology of the space and the Riemannian metric in the smooth part. This would then allow the mean and PGA computations to be carried out directly in the medial shape space, rather than on aligned m-rep models in the space $\mathcal{M}(n)$.

7.2.3 Future Diffusion Tensor Work

Several methods that were presented in Chapter 6, including the DTI comparison metric and the interpolation method, were meant as pieces to a DT-MRI registration method. A missing piece of a registration method for DT-MRI is a method for warping diffusion tensor images. This is not a trivial matter, as a warp of the space also should change the underlying diffusion tensors. It is not clear exactly how this should be done. The best approach to the warping problem so far has been proposed by Alexander [1]. It should be interesting to see if the symmetric space formulation leads to further insights on how diffusion tensors should be warped and registered.

The geodesic anisotropy measure needs to be investigated further to see whether it is useful in a clinical setting. It can be compared with other anisotropy metrics such as the fractional anisotropy, relative anisotropy, and volume ratio. While it seems that the geodesic anisotropy has more intrinsic geometric meaning than these other measures, it remains to be seen whether that translates into any practical use.

Finally, the statistical methods for diffusion tensor imaging need to be tested on real data. After a working registration method has been produced, this will allow comparisons of diffusion tensor images across patients. This might help in understanding the normal variability in brain connectivity and in fiber microstructure that is inherent in populations. It also could lead to a better understanding of how brain connectivity changes during brain development and how diseases affect the fiber structure. Also, statistical analysis of diffusion tensor data within the same patient may be useful in highlighting possible problem areas in brain connectivity due to pathology.

7.2.4 Other Application Areas

The statistical methods presented in Chapter 4 for computing the statistical variability of manifold-valued data should have far-reaching applications in several scientific fields. Manifold data, especially geometric entities such as Lie groups and symmetric spaces, can be found everywhere. As mentioned earlier, translations, rotations, scalings, and affine transformations are all examples of Lie groups. Examples of symmetric spaces include directional data, hyperplanes, frames, and positive-definite symmetric matrices. The statistical framework presented in this dissertation can handle all of these cases. Also, principal geodesic analysis can be used for non-statistical applications such as dimensionality reduction of large data sets.

There are several different fields that might benefit from PGA. Computer vision applications typically require analysis of geometric transformations (i.e., Lie groups). These include camera transformations which might be as simple as rotations or more complex affine transformations or projectivities. Also, analysis of geometric primitives in an image, such as points, direction fields, and frames could benefit from PGA. Robotics and control theory are often concerned with geometric transformations such as rigid motions of parts, representable as the Lie group $SE(3)$. Molecular biology is often concerned with the geometric configuration of DNA and protein molecules (again using multiple copies of $SE(3)$).

This dissertation shows the power of analyzing geometric entities that lie on a curved manifold using the intrinsic geometry of that manifold. Principal geodesic analysis is a new method for computing statistical variability on manifolds that uses intrinsic notions of distance to naturally generalize PCA of linear data. Principal geodesic analysis is an exciting approach to statistical analysis on manifolds because it has many practical applications to excite the engineer as well as many unanswered questions to excite the theoretician.

BIBLIOGRAPHY

[1] D. C. Alexander, C. Pierpaoli, P. J. Basser, and J. C. Gee. Spatial transformations of diffusion tensor MR images. *IEEE Transactions on Medical Imaging*, 20(11):1131–1139, 2001.

[2] L. Auslander and R. E. MacKenzie. *Introduction to Differentiable Manifolds*. Dover, 1977.

[3] P. J. Basser, J. Mattiello, and D. Le Bihan. MR diffusion tensor spectroscopy and imaging. *Biophysics Journal*, 66:259–267, 1994.

[4] P. J. Basser and S. Pajevic. A normal distribution for tensor-valued random variables: applications to diffusion tensor MRI. *IEEE Transactions on Medical Imaging*, 22(7):785–794, 2003.

[5] R. Bhattacharya and V. Patrangenaru. Nonparametric estimation of location and dispersion on Riemannian manifolds. *Journal for Statistical Planning and Inference*, 108:23–36, 2002.

[6] I. Biederman. Recognition-by-components: a theory of human image understanding. *Psychological Review*, 94(2):115–147, 1987.

[7] D. Le Bihan, J.-F. Mangin, C. Poupon, C. A. Clark, S. Pappata, N. Molko, and H. Chabriat. Diffusion tensor imaging: concepts and applications. *Journal of Magnetic Resonance Imaging*, 13:534–546, 2001.

[8] R. Blanding, C. Brooking, M. Ganter, and D. Storti. A skeletal-based solid editor. In W. F. Bronsvroot and D. C. Anderson, editors, *Proceedings of the Fifth Symposium on Solid Modeling and Applications (SSMA '99)*, pages 141–150. ACM Press, New York, 1999.

[9] J. Bloomenthal and K. Shoemake. Convolution surfaces. *Computer Graphics (SIGGRAPH '91 Proceedings)*, 25(4):251–256, 1991.

[10] H. Blum. A transformation for extracting new descriptors of shape. In W. Wathen-Dunn, editor, *Models for the Perception of Speech and Visual Form*, pages 363–380. MIT Press, Cambridge MA, 1967.

[11] H. Blum and R. Nagel. Shape description using weighted symmetric axis features. *Pattern Recognition*, 10(3):167–180, 1978.

[12] F. L. Bookstein. *The measurement of biological shape and shape change*. Number 24 in Lecture Notes in Biomathematics. Springer-Verlag, 1978.

[13] F. L. Bookstein. Size and shape spaces for landmark data in two dimensions (with discussion). *Statistical Science*, 1(2):181–242, 1986.

[14] F. L. Bookstein. *Morphometric Tools for Landmark Data: Geometry and Biology*. Cambridge University Press, 1991.

[15] W. M. Boothby. *An Introduction to Differentiable Manifolds and Riemannian Geometry*. Academic Press, 2nd edition, 1986.

[16] R. Brown. A brief account of microscopical observations made in the months on June, July, and August, 1827, on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies. *Philosophical Magazine*, 4:161–173, 1828.

[17] C. A. Burbeck, S. M. Pizer, B. S. Morse, D. Ariely, G. Zauberman, and J. Rolland. Linking object boundaries at scale: a common measurement for size and shape judgements. *Vision Research*, 36(3):361–372, 1996.

[18] S. R. Buss and J. P. Fillmore. Spherical averages and applications to spherical splines and interpolation. *ACM Transactions on Graphics*, 20(2):95–126, 2001.

[19] E. Cartan. Sur une classe remarquable d'espaces de Riemann. *Bull. Soc. Math. France*, 54:214–264, 1926.

[20] E. Cartan. Sur une classe remarquable d'espaces de Riemann. *Bull. Soc. Math. France*, 55:114–134, 1927.

[21] V. Caselles, R. Kimmel, and G. Sapiro. Geodesic active contours. *International Journal of Computer Vision*, 22(1):61–79, 1997.

[22] E. Catmull and J. Clark. Recursively generated b-spline surfaces on arbitrary topological meshes. *Computer Aided Design*, 10:183–188, 1978.

[23] C. Chefd'hotel, D. Tschumperlé, R. Deriche, and O. Faugeras. Constrained flows of matrix-valued functions: Application to diffusion tensor regularization. In *European Conference on Computer Vision*, pages 251–265, 2002.

[24] G. Christensen, S. Joshi, and M. Miller. Volumetric transformation of brain anatomy. *IEEE Transactions on Medical Imaging*, 16:864–877, 1997.

[25] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In *Fifth European Conference on Computer Vision*, pages 484–498, 1998.

[26] T. F. Cootes, A. Hill, C. J. Taylor, and J. Haslam. 1993. In H. H. Barrett and A. F. Gmitro, editors, *Proceedings of Information Processing in Medical Imaging*, volume 687 of *Lecture Notes in Computer Science*, pages 33–47. Springer-Verlag, 1993.

[27] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models – their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, 1995.

[28] O. Coulon, D. C. Alexander, and S. Arridge. Diffusion tensor magnetic resonance image regularization. *Medical Image Analysis*, 8(1):47–68, 2004.

[29] J. Csernansky, S. Joshi, L. Wang, J. Haller, M. Gado, J. Miller, U. Grenander, and M. Miller. Hippocampal morphometry in schizophrenia via high dimensional brain mapping. In *Proceedings National Academy of Sciences*, pages 11406–11411, 1998.

[30] M. L. Curtis. *Matrix Groups*. Springer-Verlag, 1984.

[31] J. Damon. On the smoothness and geometry of boundaries associated to skeletal structures i: Sufficient conditions for smoothness. *Annales de l'Institut Fourier*, 53:1941–1985, 2003.

[32] B. Davis, P. Lorenzen, and S. Joshi. Large deformation minimum mean squared error template estimation for computational anatomy. In *Proceedings of the International Symposium on Biomedical Imaging (ISBI)*, pages 173–176, 2004.

[33] H. Delingette. Simplex meshes: a general representation for 3d shape reconstruction. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, pages 856–857, 1994.

[34] H. Delingette. General object reconstruction based on simplex meshes. *International Journal of Computer Vision*, 32(2):111–146, 1999.

[35] I. Dryden and K. Mardia. *Statistical Shape Analysis*. John Wiley and Sons, 1998.

[36] J. J. Duistermaat and J. A. C. Kolk. *Lie Groups*. Springer, 2000.

[37] A. Einstein. Über die von der molekularkinetischen theorie der wärme geforderte bewegung von in ruhenden flüssigkeiten suspendierten teilchen. *Annalen der Physik*, 17:549, 1905.

[38] G. Farin. *Curves and Surfaces for CAGD: A Practical Guide*. Academic Press, 4th edition, 1997.

[39] P. T. Fletcher and S. Joshi. Principal geodesic analysis on symmetric spaces: statistics of diffusion tensors. In *Proceedings of Workshop on Computer Vision Approaches to Medical Image Analysis (CVAMIA)*, 2004.

[40] P. T. Fletcher, S. Joshi, C. Lu, and S. M. Pizer. Gaussian distributions on Lie groups and their application to statistical shape analysis. In *Information Processing in Medical Imaging*, volume LNCS 2732, pages 450–462. Springer-Verlag, 2003.

[41] P. T. Fletcher, C. Lu, and S. Joshi. Statistics of shape via principal geodesic analysis on Lie groups. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, pages 95–101, 2003.

[42] P. T. Fletcher, C. Lu, S. M. Pizer, and S. Joshi. Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE Transactions on Medical Imaging (to appear)*, 2004.

[43] P. T. Fletcher, S. M. Pizer, A. G. Gash, and S. Joshi. Deformable m-rep segmentation of object complexes. In *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI) (CD Proceedings)*, 2002.

[44] M. Fréchet. Les éléments aléatoires de nature quelconque dans un espace distancié. *Ann. Inst. H. Poincaré*, (10):215–310, 1948.

[45] P. Giblin and B. Kimia. A formal classifcation of 3D medial axis points and their local geometry. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 566–573, 2000.

[46] P. J. Giblin and S. A. Brassett. Local symmetries of plane curves. *American Mathematical Monthly*, 92:689–707, 1985.

[47] C. Goodall. Procrustes methods in the statistical analysis of shape. *Journal of the Royal Statistical Society*, 53(2):285–339, 1991.

[48] J. C. Gower. Generalized procrustes analysis. *Psychometrika*, 40:33–51, 1975.

[49] U. Grenander. *Probabilities on Algebraic Structures*. John Wiley and Sons, 1963.

[50] U. Grenander. *Pattern Synthesis: Lectures in Pattern Theory*, volume I. Springer-Verlag, 1976.

[51] U. Grenander. *Pattern Synthesis: Lectures in Pattern Theory*, volume II. Springer-Verlag, 1978.

[52] U. Grenander. *Regular Structures: Lectures in Pattern Theory*, volume III. Springer-Verlag, 1981.

[53] U. Grenander, M. I. Miller, and A. Srivastava. Hilbert-Schmidt lower bounds for estimators on matrix Lie groups for ATR. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):790–802, 1998.

[54] B. C. Hall. *Lie groups, Lie algebras, and representations: an elementary introduction*. Springer-Verlag, 2003.

[55] Q. Han, C. Lu, G. Liu, S. M. Pizer, S. Joshi, and A. Thall. Representing multi-figure anatomical objects. In *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1251–1254, 2004.

[56] H. Hanyu, H. Sakurai, T. Iwamoto, and et al. Diffusion-weighted MR imaging of the hippocampus and temporal white matter in Alzheimers disease. *Journal of Neurological Science*, 156:195–200, 1998.

[57] H. Hanyu, H. Shindo, D. Kakizaki, and et al. Increased water diffusion in cerebral white matter in Alzheimers disease. *Gerontology*, 43:343–351, 1997.

[58] S. Helgason. *Differential Geometry, Lie Groups, and Symmetric Spaces*. Academic Press, 1978.

[59] I. N. Herstein. *Topics in Algebra*. John Wiley and Sons, 2nd edition, 1975.

[60] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24:417–441, 498–520, 1933.

[61] G. A. Hunt. Semi-groups of measures on Lie groups. *Transactions of the American Mathematical Society*, 81:264–293, 1956.

[62] T. Igarashi, S. Matsuoka, and H. Tanaka. Teddy: a sketching interface for 3D freeform design. In *Proceedings SIGGRAPH '99*, pages 409–416, 1999.

[63] I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, 1986.

[64] S. Joshi. *Large Deformation Diffeomorphisms and Gaussian Random Fields for Statistical Characterization of Brain Sub-manifolds*. PhD thesis, Washington University, St. Louis, MO, 1997.

[65] S. Joshi, U. Grenander, and M. Miller. On the geometry and shape of brain submanifolds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11:1317–1343, 1997.

[66] S. Joshi, S. Pizer, P. T. Fletcher, P. Yushkevich, A. Thall, and J. S. Marron. Multiscale deformable model segmentation and statistical shape analysis using medial descriptions. *Transactions on Medical Imaging*, 21(5), 2002.

[67] H. Karcher. Riemannian center of mass and mollifier smoothing. *Communications on Pure and Applied Math*, 30(5):509–541, 1977.

[68] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: active contour models. *International Journal of Computer Vision*, 60(1):321–331, 1988.

[69] K. Kawakubo. *The Theory of Transformation Groups*. Oxford University Press, 1991.

[70] A. Kelemen, G. Székely, and G. Gerig. Elastic model-based segmentation of 3-D neuroradiological data sets. *IEEE Transactions on Medical Imaging*, 18(10):828–839, 1999.

[71] D. G. Kendall. The diffusion of shape. *Advances in Applied Probability*, (9):428–430, 1977.

[72] D. G. Kendall. Shape manifolds, Procrustean metrics, and complex projective spaces. *Bulletin of the London Mathematical Society*, 16:18–121, 1984.

[73] D. G. Kendall. A survey of the statistical theory of shape. *Statistical Science*, 4(2):87–120, 1989.

[74] W. S. Kendall. Probability, convexity, and harmonic maps with small image I: uniqueness and fine existence. *Proceedings of the London Mathematical Society*, 3(61):371–406, 1990.

[75] E. Klassen, A. Srivastava, W. Mio, and S. Joshi. Analysis of planar shapes using geodesic paths on shape space. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(3):372–383, 2004.

[76] H. Le and D. G. Kendall. The Riemannian structure of Euclidean shape space: a novel environment for statistics. *The Annals of Statistics*, 21(3):1225–1271, 1993.

[77] J. M. Lee. *Riemannian Manifolds: An Introduction to Curvature*. Springer, 1997.

[78] T. S. Lee, D. Mumford, and P. H. Schiller. Neuronal correlates of boundary and medial axis representations in primate striate cortex. In *Investigative Ophthalmology and Visual Science*, volume 36, page 477, 1995.

[79] M. Leyton. *Symmetry, Causality, Mind*. MIT Press, Cambridge, MA, 1992.

[80] C. Lu, S. M. Pizer, and S. Joshi. A Markov random field approach to multi-scale shape analysis. In *Proceedings of Scale Space Methods in Computer Vision*, volume LNCS 2695, pages 416–431, 2003.

[81] K. V. Mardia. *Directional Statistics*. John Wiley and Sons, 1999.

[82] K. V. Mardia and I. L. Dryden. Shape distributions for landmark data. *Advances in Applied Probability*, 21:742–755, 1989.

[83] B. Markussen. A statistical approach to large deformation diffeomorphisms. In *Proceedings of Workshop on Generative Model-Based Vision (GMBV)*, CD Proceedings, 2004.

[84] D. Marr and H. K. Nishihara. Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society, Series B*, 200:269–294, 1978.

[85] T. McInerney and D. Terzopoulos. Deformable models in medical image analysis: a survey. *Medical Image Analysis*, 1(2):91–108, 1996.

[86] J. W. Milnor. *Morse Theory*. Princeton University Press, 1963.

[87] J. W. Milnor. *Topology from the Differentiable Viewpoint*. Princeton University Press, 1997.

[88] M. Moakher. Means and averaging in the group of rotations. *SIAM Journal on Matrix Analysis and Applications*, 24(1):1–16, 2002.

[89] M. E. Moseley, J. Kucharczyk, J. Mintorovitch, and et al. Diffusion weighted MR imaging of acute stroke: correlation with T2-weighted and magnetic susceptibility-enhanced MR imaging in cats. *AJNR*, 11:423–429, 1990.

[90] D. Mumford. Pattern theory: a unifying perspective. In D. C. Knill and W. Richards, editors, *Perception as Bayesian Inference*, pages 25–62. Cambridge University Press, 1996.

[91] J. R. Munkres. *Topology: A First Course*. Prentice-Hall, 1975.

[92] L. R. Nackman and S. M. Pizer. Three-dimensional shape description using the symmetric axis transform, I: theory. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 7(2):187–202, 1985.

[93] J. J. Neil, S. I. Shiran, R. C. McKinstry, and et al. Normal brain in human newborns: apparent diffusion coefficient and diffusion anisotropy measured by using diffusion tensor MR imaging. *Radiology*, 209:57–66, 1998.

[94] M. Nielsen, P. Johansen, A. D. Jackson, and B. Lautrup. Brownian warps: a least committed prior for non-rigid registration. In *Proceedings of the Fifth International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, number 2489 in Lecture Notes in Computer Science, pages 557–564, 2002.

[95] N. H. Olsen. *Morphology and Optics of Human Embryos from Light Microscopy*. PhD thesis, University of Copenhagen, Denmark, 2003.

[96] N. H. Olsen and M. Nielsen. Lie group modeling of nonlinear point set shape variability. In Gerald Sommer and Yehoshua Y. Zeevi, editors, *Proceedings of the Second International Workshop on Algebraic Frames for the Perception-Action Cycle (AFPAC)*, volume 1888 of *Lecture Notes in Computer Science*. Springer, 2000.

[97] S. J. Osher and J. A. Sethian. Fronts propogation with curvature dependent speed: algorithms based on Hamilton-Jacobi formulations. *Journal of Computational Physics*, 79:12–49, 1988.

[98] S. Pajevic and P. J. Basser. Parametric and non-parametric statistical analysis of DT-MRI. *Journal of Magnetic Resonance*, 161(1):1–14, 2003.

[99] K. Pearson. On lines and planes of closest fit to points in space. *Philosophical Magazine*, 2:609–629, 1901.

[100] X. Pennec. Probabilities and statistics on Riemannian manifolds: basic tools for geometric measurements. In *IEEE Workshop on Nonlinear Signal and Image Processing*, 1999.

[101] S. Pizer, D. Fritsch, P. Yushkevich, V. Johnson, and E. Chaney. Segmentation, registration, and measurement of shape variation via image object shape. *IEEE Transactions on Medical Image Analysis*, 18:851–865, 1999.

[102] S. M. Pizer, P. T. Fletcher, S. Joshi, A. Thall, J. Z. Chen, Y. Fridman, D. S. Fritsch, A. G. Gash, J. M. Glotzer, M. R. Jiroutek, C. Lu, K. E. Muller, G. Tracton, P. Yushkevich, and E. L. Chaney. Deformable m-reps for 3D medical image segmentation. *International Journal of Computer Vision*, 55(2–3):85–106, 2003.

[103] S. M. Pizer, K. Siddiqi, G. Sz'ekely, J. N. Damon, and S. W. Zucker. Multiscale medial loci and their properties. *International Journal of Computer Vision*, 55(2–3):155–179, 2003.

[104] W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery. *Numerical Recipes in C++: The Art of Scientific Computing*. Cambridge University Press, 2nd edition, 2002.

[105] M. Rao, J. Stough, Y.-Y. Chi, K. Muller, G. S. Tracton, S. M. Pizer, and E. L. Chaney. Comparison of human and automatic segmentations of kidneys from CT images. submitted to *International Journal of Radiation Oncology, Biology, Physics*, June 2004.

[106] W. Rudin. *Principles of Mathematical Analysis*. McGraw-Hill, 1976.

[107] J. A. Sethian. A review of recent numerical algorithms for hypersurfaces moving with curvature dependent flows. *Journal of Differential Geometry*, 31:131–161, 1989.

[108] A. Sherstyuk. Shape design using convolution surfaces. In *Proceedings of Shape Modeling International '99*, 1999.

[109] C. G. Small. *The statistical theory of shape*. Springer, 1996.

[110] A. G. Sorensen, F. S. Buonanno, R. G. Gonzalez, and et al. Hyperacute stroke: evaluation with combined multisection diffusion-weighted and hemodynamically weighted echo-planar MR imaging. *Radiology*, 199:391–401, 1996.

[111] M. Spivak. *A Comprehensive Introduction to Differential Geometry*, volume 1. Publish or Perish, 3rd edition, 1999.

[112] M. Spivak. *A Comprehensive Introduction to Differential Geometry*, volume 2. Publish or Perish, 3rd edition, 1999.

[113] M. Spivak. *A Comprehensive Introduction to Differential Geometry*, volume 3. Publish or Perish, 3rd edition, 1999.

[114] M. Spivak. *A Comprehensive Introduction to Differential Geometry*, volume 4. Publish or Perish, 3rd edition, 1999.

[115] M. Spivak. *A Comprehensive Introduction to Differential Geometry*, volume 5. Publish or Perish, 3rd edition, 1999.

[116] A. Srivastava and E. Klassen. Monte-Carlo extrinsic estimators of manifold-valued parameters. *IEEE Transactions on Signal Processing*, 50(2):299–308, 2001.

[117] L. Staib and J. Duncan. Boundary finding with parametrically deformable models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(11):1061–1075, 1992.

[118] L. A. Steen and J. A. Seebach. *Counterexamples in Topology*. Dover, 1995.

[119] D. W. Storti, G. M. Turkiyyah, M. A. Ganter, C. T. Lim, and D. M. Stat. Skeleton-based modeling operations on solids. In *Proceedings of the Fourth Symposium on Solid Modeling and Applications (SSMA '97)*, pages 141–154. ACM, 1997.

[120] M. Styner and G. Gerig. Medial models incorporating object variability for 3D shape analysis. In *Information Processing in Medical Imaging*, pages 502–516, 2001.

[121] A. Swann and N. H. Olsen. Linear transformation groups and shape space. *Journal of Mathematical Imaging and Vision*, 19(1):49–62, 2003.

[122] G. Székely, A. Kelemen, C. Brechbühler, and G. Gerig. Segmentation of 2-D and 3-D objects from MRI volume data using constrained elastic deformation of flexible Fourier contour and surface models. *Medical Image Analysis*, 1(1):19–34, 1996.

[123] A. Thall. Fast C^2 interpolating subdivision surfaces using iterative inversion of stationary subdivision rules. Technical report, University of North Carolina Department of Computer Science, 2002. http://midag.cs.unc.edu/pub/papers/Thall_TR02-001.pdf.

[124] A. Thall. *Deformable Solid Modeling via Medial Sampling and Displacement Subdivision*. PhD thesis, University of North Carolina at Chapel Hill, 2003.

[125] D'Arcy W. Thompson. *On Growth and Form*. Cambridge University Press, 1917.

[126] A. L. Tievsky, T. Ptak, and J. Farkas. Investigation of apparent diffusion coefficient and diffusion tensor anisotropy in acute and chronic multiple sclerosis lesions. *AJNR*, 20:1491–1499, 1999.

[127] S. Warach, M. Boska, and K. M. A. Welch. Pitfalls and potential of clinical diffusion-weighted MR imaging in acute stroke. *Stroke*, 28:481–482, 1997.

[128] D. Wehn. *Limit distributions on Lie groups*. PhD thesis, Yale University, 1959.

[129] D. Wehn. Probabilities on Lie groups. *Proceedings of the National Academy of Sciences of the United States of America*, 48(5):791–795, 1962.

[130] D. J. Werring, C. A. Clark, G. J. Barker, A. J. Thompson, and D. H. Miller. Diffusion tensor imaging of lesions and normal-appearing white matter in multiple sclerosis. *Neurology*, 52:1626–1632, 1999.

[131] P. Yushkevich. *Statistical Shape Characterization Using the Medial Representation*. PhD thesis, The University of North Carolina at Chapel Hill, 2003.

[132] P. Yushkevich, P. T. Fletcher, S. Joshi, A. Thall, and S. M. Pizer. Continuous medial representations for geometric object modeling in 2D and 3D. *Image and Vision Computing*, 21(1):17–28, 2003.

[133] Y. Chen Z. Wang, B. C. Vemuri and T. Mareci. A constrained variational principle for direct estimation and smoothing of the diffusion tensor field from DWI. In *Information Processing in Medical Imaging*, pages 660–671, 2003.

[134] G. Zhai, W. Lin, K. Wilber, G. Gerig, and J. Gilmore. Comparison of regional white matter diffusion in healthy neonate and adults using a 3T head-only MR scanner. *Radiology*, 229:673–681, 2003.