The members of the committee approve the dissertation of Leif Ellingson defended on May 26, 2011.

Vic Patrangenaru  
Professor Directing Thesis

Washington Mio  
University Representative

Jinfeng Zhang  
Committee Member

Xufeng Niu  
Committee Member

Approved:

Daniel McGee, Chair, Department of Statistics

Joseph Travis, Dean, College of Arts and Sciences

The Graduate School has verified and approved the above-named committee members.
For my parents. You instilled in me the character, strength, work ethic, and yearning to do what I love that, without which, none of this would have been possible.
ACKNOWLEDGMENTS

I would first like to give great thanks to my major professor, Dr. Victor Patrangenaru, for all of his advice, direction, encouragement, guidance, and patience. I am deeply grateful for all of the time, effort, and energy he has invested in working with me and preparing me for success both now and in the future. Little did I know at the time that enrolling in his Limit Theory class would so deeply and profoundly shape my future. Without Dr. Patrangenaru, I would not have become involved with these fascinating areas of research or have had the phenomenal opportunities with which I have been presented.

Next, I would like to thank the members of my committee, Dr. Washington Mio, Dr. Xufeng Niu, and Dr. Jinfeng Zhang. They have provided valuable advice and insight that have greatly enriched my knowledge and improved this document. I would especially like to acknowledge Dr. Zhang for all of his time, encouragement, energy, and guidance he has provided on our work for the protein binding site project.

In addition to those named above, I would also like to acknowledge my collaborators on these research projects: Dr. Rabi Bhattacharya, Dr. Michael Crane, Dr. Xiwen Liu, and Dr. Frits Ruymgaart. Furthermore, I would like to thank the Statistical and Applied Mathematical Sciences Institute for hosting and organizing the 2010-2011 Program on the Analysis of Object Data. In particular, I want to thank the members of the Data Analysis on Sample Spaces with a Manifold Stratification and Geometric Correspondence working groups, from whom I have learned a great amount and have been provided with invaluable insight on my areas of research and beyond.

I would now like to thank all of the members of the Department of Statistics for providing an extremely positive environment for fostering growth on both sides of the classroom and in research. I have learned a great deal from the faculty members and I am very grateful for all of the work they put in to educating students. I am extremely grateful to the staff for all of their assistance, whether with routine paperwork or with problems that arose. I would like to thank my fellow students in the department for their part in making my experiences here so overwhelmingly positive. I would especially like to thank my colleagues and good friends Sebastian Kurtek, Vernon Lawhern, and Daniel Osborne, who have served as a sounding board for many of my thoughts and ideas and provided much support and
levity through even the most trying times I encountered here.

Lastly, but certainly not least, I want to thank my family for all of their love and support throughout the years. They have been a source of great advice, continuous encouragement, and immense support that have helped provide me the strength and will to complete my graduate studies and finish this dissertation.
# TABLE OF CONTENTS

List of Tables .................................................................................. viii  
List of Figures ................................................................................ ix  
Abstract .......................................................................................... xii  

1 Introduction ................................................................................ 1  

2 Nonparametric Statistics on Manifolds .................................. 4  
  2.1 Statistical Manifolds ................................................................. 5  
  2.2 Statistics on Manifolds ............................................................... 5  
  2.3 Extrinsic Analysis on Manifolds ............................................. 7  
    2.3.1 Asymptotic Distributions of Extrinsic Sample Means ..... 8  
    2.3.2 Confidence Regions for Extrinsic Means ...................... 8  
  2.4 Intrinsic Analysis on Manifolds .............................................. 9  
    2.4.1 Asymptotic Distributions of Intrinsic Sample Means .... 10  
    2.4.2 Confidence Regions for Intrinsic Means ...................... 10  

3 Hypothesis Testing for Means on Hilbert Spaces .................. 11  
  3.1 Hilbert-Valued Random Objects ........................................... 11  
  3.2 The One-Sample Procedure ................................................. 12  
  3.3 The Multi-Sample Procedure ................................................. 13  

4 Extrinsic Analysis on Hilbert Manifolds ............................... 14  
  4.1 Fréchet Differentiation ......................................................... 14  
  4.2 Hilbert Manifolds ................................................................. 15  
  4.3 Extrinsic Analysis ................................................................. 15  

5 Shape Analysis of Planar Contours ........................................ 18  
  5.1 Kendall’s Shape Spaces of Finite Planar Configurations ....... 19  
    5.1.1 Veronese-Whitney Extrinsic Mean Shape ................. 20  
  5.2 Shape Spaces of Planar Curves as Continuous Functions ... 21  
    5.2.1 Bending-only Representation .................................. 22  
    5.2.2 Square-Root Elastic Representation ...................... 23  
  5.3 Kendall’s Shape Spaces of Planar Contours ....................... 27  
    5.3.1 Infinite-dimensional Planar Similarity Shape Spaces .... 28  
  5.4 The Veronese-Whitney Extrinsic Mean Shape ................. 30  
    5.4.1 Asymptotic Distribution of the Sample Extrinsic Mean Shape .... 31
5.5 Finite Representation of Planar Contours ........................................... 33
5.5.1 Random Landmark Selection ......................................................... 34
5.5.2 Behavior of $k$-ad when Adding Landmarks ................................. 36
5.5.3 Approximations for Analysis ....................................................... 40
5.5.4 Landmark Correspondence Within a Sample ................................. 41
5.6 Computations of Sample Means of Planar Contours ......................... 42
5.7 A One-Sample Test for Mean Shape ............................................... 45
5.7.1 Application of the One Sample Test for Mean Contour Shape ......... 47
5.8 Bootstrap Confidence Regions for the Sample Mean ......................... 49
5.9 Discussion ..................................................................................... 52

6 Size-and-Shape Analysis for 3D Structural Proteomics .......................... 54
6.1 Study of Relationship Between Protein Structure and Function ........... 55
6.2 Size-and-Shape Analysis for Configurations in 3-Dimensions ............... 57
6.2.1 Size-and-Shape Spaces ............................................................... 58
6.2.2 The Schoenberg Embedding ....................................................... 59
6.2.3 Extrinsic Means for Reflection Shape Spaces ............................... 61
6.2.4 Asymptotics for the Extrinsic Sample Mean Reflection Size-and-Shape 63
6.3 Landmark Selection and Registration ................................................. 66
6.3.1 Algorithm for Obtaining Corresponding Atoms ............................ 67
6.4 Verification of Algorithm .................................................................. 70
6.4.1 Benchmark Data set .................................................................. 70
6.4.2 Identification of Maximum Common Atom Set ............................. 71
6.4.3 Classification of Ligand Binding Sites ......................................... 73
6.4.4 Effect of Search Radius and Number of Nearest Neighbors .......... 78
6.4.5 Conclusion and Discussion ......................................................... 78
6.5 Mean Size-and-Shape of Protein Binding Sites .................................. 80

7 Current Directions in Statistics on Manifolds ........................................ 83
7.1 Computational Comparison of Intrinsic and Extrinsic Analysis ............ 84
7.1.1 Directional Data Analysis ........................................................... 84
7.1.2 Size-and-Shape Analysis for 3D Configurations ............................ 87
7.1.3 Direct Similarity Shape Analysis of Planar Contours ....................... 89
7.2 Data Analysis on Sample Spaces with Manifold Stratification ............. 93
7.2.1 The Reflection-Affine Shape Spaces ............................................ 93
7.2.2 The Space of Phylogenetic Trees ............................................... 94
7.3 Summary and Conclusions ................................................................ 95

Bibliography ....................................................................................... 97
Biographical Sketch ........................................................................... 106
LIST OF TABLES

5.1 Estimates for the effective coverage of nominal 95% bootstrap confidence regions for the extrinsic mean shape ........................................... 52
6.1 Atom types used for obtaining correspondence .................................. 68
6.2 The binding sites included in the Kahraman set, listed by ligand ............ 71
6.3 Atoms from ATP binding sites of 1ayl and 1c2q not found by both methods. 72
6.4 Summary of CAS of two methods for three pairs of binding sites .......... 72
6.5 Summary of CAS of two methods for three pairs of binding sites .......... 74
6.6 The misclassified active sites with classified ligands in parentheses. ..... 76
6.7 Classification error for various combinations of SR and $k$. ................. 78
6.8 Classification error for various combinations of SR and $k$. ................. 78
7.1 The times required to compute the mean size-and-shape for various sample sizes ................................................................. 90
LIST OF FIGURES

5.1 Two nearby closed curves in the Hilbert space that have a fairly large distance $d$ between them. ................................................................. 30

5.2 A dog represented by 1000 and, respectively, 200 randomly chosen landmarks ......................................................... 35

5.3 An image of the contour of a boot from the Kimia image database (a) as a silhouette, (b) as a contour, (c) as 200 randomly selected landmarks, and (d) as a contour constructed from the chosen landmarks ......................... 36

5.4 An image of the contour of a hand gesture from the Kimia image database (a) as a silhouette, (b) as a contour as obtained by an edge detector, (c) as randomly selected landmarks, and (d) as a contour constructed by a piecewise linear interpolation from the chosen landmarks ......................... 37

5.5 Correspondence of 6 landmarks for hand gestures of (a) the number '5' and (b) the letter 'L' ......................................................... 42

5.6 (a) Sample of 4 curves of 't' gestures and (b) the extrinsic mean shape of the sample of 't' gestures ................................. 43

5.7 (a) Sample of 9 curves of "dogs" and (b) the extrinsic mean shape of the sample of dogs ......................................................... 43

5.8 (a) Sample of 20 curves of "sting ray" fish and (b) the extrinsic mean shape of the sample of sting rays. ......................... 43

5.9 (a) Sample of 20 curves of worm fish and (b) the extrinsic mean shape of the sample of worm fish. ................................. 44

5.10 (a) Sample of 20 curves of red snapper fish and (b) the extrinsic mean shape of the sample of red snapper fish. ......... 44

5.11 (a) Sample of 20 curves of pears and (b) the extrinsic mean shape of the sample of pears. ......................................................... 44

5.12 The extrinsic sample mean shape of a sample of 10 sting rays and, respectively, the hypothesized extrinsic mean shape ................................. 48
5.13 The sample extrinsic mean shape of a sample of 87 pears and, respectively, the hypothesized extrinsic mean shape ................................. 48
5.14 The sample extrinsic mean shape of a sample of 83 pears and, respectively, the hypothesized extrinsic mean shape ................................. 49
5.15 Bootstrap 95% confidence region for the extrinsic mean shape of (a) the "t" hand gesture and (b) the dogs. .................................................. 50
5.16 Bootstrap 95% confidence region for the extrinsic mean shape of (a) the sting rays and (b) the wormfish. .................................................. 51
5.17 Bootstrap 95% confidence region for the extrinsic mean shape of (a) the red snapper fish and (b) the pears. .................................................. 51
5.18 (a) A grey-level image of an electronic circuit and (b) the edge-map of the circuit found using Canny edge detection. The original image is included as an example for image processing in MATLAB. ................................. 53
6.1 The similarity matrix for the IN-TI + Rg model. Lower-valued pixels represent greater similarity and higher-valued pixels correspond to less similarity. The steroid group consists of both the AND and EST ligand groups in accordance to Kahraman et al (2007). .................................................. 77
6.2 Active sites that bind to heme for proteins 1phf, 1phg, 2cpp, and 1m85 . . . 81
6.3 The Schoenberg extrinsic mean reflection size-and-shape of the atoms common to the binding sites .................................................. 82
7.1 The wind direction data used for the computations. .......................... 86
7.2 (a) The extrinsic and intrinsic means for the data (b) The Frechet function using arc distance as a function of colatitude and longitude .......................... 86
7.3 The extrinsic and intrinsic bootstrap means means for the data compared to the extrinsic and intrinsic means .................................................. 87
7.4 Active sites that bind to heme for proteins 1phf, 1phg, 2cpp, and 1m85 . . . 89
7.5 (a) The Schoenberg extrinsic mean reflection size-and-shape and (b) the intrinsic mean size-and-shape of atoms common to the active sites .......................... 89
7.6 4 observations of contours of the "l" hand gesture .......................... 91
7.7 Bootstrap 95% confidence regions using 400 resamples for (a) the extrinsic mean shape of the "l" hand gesture and (b) the intrinsic mean shape of the "l" hand gesture .................................................. 91
7.8 4 observations of contours of a side view of a dog .......................... 92
7.9 Bootstrap 95% confidence regions using 400 resamples for (a) the extrinsic mean shape of the 4 dogs and (b) the intrinsic mean shape of the 4 dogs.
ABSTRACT

The technological advances in recent years have produced a wealth of intricate digital imaging data that is analyzed effectively using the principles of shape analysis. Such data often lies on either high-dimensional or infinite-dimensional manifolds. With computing power also now strong enough to handle this data, it is necessary to develop theoretically-sound methodology to perform the analysis in a computationally efficient manner. In this dissertation, we propose approaches of doing so for planar contours and the three-dimensional atomic structures of protein binding sites.

First, we adapt Kendall’s definition of direct similarity shapes of finite planar configurations to shapes of planar contours under certain regularity conditions and utilize Ziezold’s nonparametric view of Fréchet mean shapes. The space of direct similarity shapes of regular planar contours is embedded in a space of Hilbert-Schmidt operators in order to obtain the Veronese-Whitney extrinsic mean shape. For computations, it is necessary to use discrete approximations of both the contours and the embedding. For cases when landmarks are not provided, we propose an automated, randomized landmark selection procedure that is useful for contour matching within a population and is consistent with the underlying asymptotic theory. For inference on the extrinsic mean direct similarity shape, we consider a one-sample neighborhood hypothesis test and the use of nonparametric bootstrap to approximate confidence regions.

Bandulasiri et al (2008) suggested using extrinsic reflection size-and-shape analysis to study the relationship between the structure and function of protein binding sites. In order to obtain meaningful results for this approach, it is necessary to identify the atoms common to a group of binding sites with similar functions and obtain proper correspondences for these atoms. We explore this problem in depth and propose an algorithm for simultaneously finding the common atoms and their respective correspondences based upon the Iterative Closest Point algorithm. For a benchmark data set, our classification results compare favorably with those of leading established methods. Finally, we discuss current directions in the field of statistics on manifolds, including a computational comparison of intrinsic and extrinsic analysis for various applications and a brief introduction of sample spaces with manifold stratification.
CHAPTER 1

INTRODUCTION

Statistical shape analysis is a growing field of study. This is, in part, due to the increasing amount of research involving shape-related data throughout many fields. With recent technological developments, digital image data is becoming increasingly abundant and can be studied using principles of shape analysis. The technological growth has also allowed researchers to explore microscopic objects with amazing detail and clarity, even allowing them to visualize the structure of molecules, such as proteins. However, while data is readily available, the increase in computing power also now allows for more effective and efficient analysis, as well.

Despite most people having their own general ideas about what “shape” is, when analyzing the shape of an object, it is, of course, crucial to have a formal definition of shape. However, there is not just one definition, as many different fields and applications require different ideas of what constitutes shape. Among the well-studied types are direct similarity shape, indirect similarity shape, direct similarity size-and-shape, indirect similarity size-and-shape, affine shape, and projective shape.

In general, the $G$-shape of a configuration $x \in \mathbb{R}^p$ is defined to be the orbit, or equivalence class, of $x$ under the action of group $G$. The groups considered in each of the above types of shape are as follows, where $\xi = x - \bar{x}$ is a centered configuration:

- **Direct similarity shape.** $SO(p)$, the special orthogonal group of $p \times p$ matrices, acting on $u = \frac{\xi}{\|\xi\|}$.
- **Indirect similarity shape.** $O(p)$, the orthogonal group of $p \times p$ matrices, acting on $u = \frac{\xi}{\|\xi\|}$.
- **Direct similarity size-and-shape.** $SO(p)$, the special orthogonal group of $p \times p$ matrices, acting on $\xi$.
- **Indirect similarity size-and-shape.** $O(p)$, the orthogonal group of $p \times p$ matrices, acting on $\xi$. 

1
• **Affine shape.** $GL(p)$, the general linear group of $p \times p$ matrices, acting on $\xi$.

• **Projective shape.** $PGL(p)$, the projective linear group of $p \times p$ matrices, acting on the projective coordinates of $x$.

Regardless of the definition, though, one common aspect is that shape data is non-Euclidean, so standard statistical analysis techniques cannot be directly applied. However, statistical analysis can be conducted by using techniques defined for manifolds. A manifold $M$ is a metric space that has a local, smooth mapping to a Euclidean (or Hilbert) space. A more rigorous definition will be provided in Chapter 2.

While standard statistical techniques used for Euclidean data cannot be applied directly to such data, methodology has been developed for statistical analysis of data on manifolds, including that on shape manifolds. For instance, when describing probability distributions, it is often important to have some notion of location, such as the mean or median. A natural index of location for a probability measure $Q$ on a metric space $M$ with the distance $\rho$ is the so-called Fréchet mean which is the minimizer of $F(p) = \int \rho^2(p,x) Q(dx)$, if the minimizer is unique. In general, though, the set of all such minimizers is called the Fréchet mean set.

There are multiple manners, though, in which to define the distance $\rho$ used to calculate the Fréchet mean, resulting in different definitions of the Fréchet mean, as shown in Bhat-tacharya and Patrangenaru (2003) [1]. If $(M, g)$ is a $d$-dimensional connected and complete Riemannian manifold, meaning that it is associated with a complete Riemannian metric $g$, then the Fréchet mean (set) with respect to the geodesic distance $d_g$ under $g$ is defined to be the intrinsic mean (set).

However, intrinsic means are often difficult to compute, so it is often more convenient to analyze the data using an alternative approach [1]. A manifold can be also looked at as being a submanifold of a Euclidean space, and a probability measure on it can be regarded as a probability measure in that ambient linear space. Such an approach has been employed for shape analysis by Kent (1992) [2], Dryden and Mardia (1993) [3], Le (1998) [4]. In this case, assuming that $M$ is a closed submanifold of the Euclidean space $E^k = (\mathbb{R}^k, d_0)$ where $d_0$ denotes the Euclidean distance, $d_0(x,y) = \|y-x\|$, let $Q$ be a probability measure on $M$. Let $G^c$ be the set of nonfocal points of $M$ in $E^k$. The projection map $P_M : G^c \rightarrow M$ is defined as $P_M(p) = x$ if $d_0(p,M) = d_0(p,x)$. In this case, the Fréchet function is defined on $M$ by

$$F_0(p) = \int_M \|p-x\|^2 Q(dx)$$

(1.1)

and is called the extrinsic mean (set). Chapter 2 discusses manifolds and both intrinsic and extrinsic analysis of data on general manifolds in more depth. Following that, this dissertation will focus primarily on extrinsic analysis of direct similarity shape and indirect similarity size-and-shape.
The two chapters that follow provide the foundation for much of the material presented in Chapter 5. Chapter 3 discusses methodology for performing hypothesis testing for means in Hilbert spaces. Chapter 4 discusses extrinsic analysis on Hilbert manifolds. Both of these chapters motivate the theory of Chapter 5, which presents methodology for performing extrinsic analysis for direct similarity shapes of planar contours. This chapter extends D. G. Kendall’s approach to direct similarity shape analysis for finite-dimensional configurations to the infinite-dimensional setting. However, since discretization is necessary when performing computations, an efficient, automated method for selecting landmarks is proposed that is consistent with the underlying theory. Two one-sample procedures are introduced for one-sample inference. First, a neighborhood hypothesis test based upon the methodology from Chapter 3 is presented. Finally, nonparametric nonpivotal bootstrap is used to obtain approximate confidence regions for the extrinsic mean shape.

Chapter 6 concerns using indirect size-and-shape analysis for configurations in $\mathbb{R}^3$ for the study of the atomic structure of the binding sites of proteins. After discussing the theory and methodology for the size-and-shape analysis, the chapter then presents an algorithm for obtaining corresponding landmarks for pairs of binding sites. This methodology is then validated by performing classification on a benchmark data set to compare the results to established methods.

Chapter 7 discusses current directions in the field of statistics on manifolds. This includes a comparison of computational aspects of extrinsic and intrinsic analysis. Methods for calculating both types of means are presented for directional data analysis with an example of wind data, reflection size-and-shape analysis with an example of protein binding site structures, and direct similarity shape analysis for planar contours. Additionally, the computational efficiency of both approaches is addressed. Finally, the chapter provides a brief overview of the analysis of data from sample spaces with manifold stratification.
The connection between the fields of statistics and differential geometry has a long and fruitful history. Today there are two important directions along these lines:

- Statistical Manifolds
- Statistics on Manifolds

A topological manifold of dimension $p$ is a metric space $(\mathcal{M}, \rho)$ that is locally homeomorphic to the Euclidean space $\mathbb{E}^p = (\mathbb{R}^p, \|\cdot\|_0)$. The set of all these local homeomorphisms $(U, \varphi), \varphi : U \to \varphi(U) \subseteq \mathbb{R}^p$, is a topological atlas of the manifold. As examples, a planet’s surface may be thought of as a 2-dimensional manifold and, in physics, space-time is often regarded as a 4-dimensional manifold.

Statisticians are interested in data on differentiable manifolds. These are manifolds that admit a tangent space at each of their points. The tangent space $T_p\mathcal{M}$ at a point $p$ of a manifold $\mathcal{M}$ is the set of all vectors tangent to curves in $\mathcal{M}, c : (-\varepsilon, \varepsilon) \to \mathcal{M}$ with $c(0) = p$. Observed measurements in statistics are often regarded as points on a manifold and are often called objects.

The most basic example of data on manifolds is multivariate data, which lies on flat manifolds. For such data, one can appeal to the usual statistical techniques developed for multivariate analysis. However, less elementary manifolds arise in, among other fields, directional data analysis, axial data analysis, medical imaging, astronomy, geology, pattern recognition, visual quality control for manufacturing processes, shape analysis, and proteomics. For such data, it is necessary to appeal to statistical methodology defined for these often abstract spaces. While this dissertation is primarily concerned with shape analysis and, to some extent, proteomics, this chapter presents methodology for performing nonparametric statistical analysis on general manifolds.
2.1 Statistical Manifolds

One of the most quoted references in the area of statistical manifolds is a textbook by S. Amari (1985) [5], which was considered by C. R. Rao to be a “timely and fascinating monograph which takes the reader into an intriguing realm of mathematical statistics”. The origin of the differential geometric methods in statistics may be found in the early work of the C. R. Rao [6], who considered the space of a parametric family of probability distributions as a Riemannian manifold with a natural metric provided by the Fisher information matrix. The revived interest in this area is motivated by the contributions of Efron [7], Chentsov [8], Barndorff-Nielsen et al (1986) [9], and others. Efron’s interpretation of Fisher’s second-order loss of information and C. R. Rao’s concept of second-order efficiency in estimation as the curvature of a statistical manifold led various workers to exploit differential-geometric concepts and methods in studying asymptotic statistical inference. C. R. Rao commented that ”differential geometry will provide the basic tools for research in statistical inference as matrix algebra has done in many areas of mathematical statistics.”

2.2 Statistics on Manifolds

Statistical inference for distributions on manifolds is now a broad discipline with wide ranging applications. Its study has gained momentum in recent years, especially due to applications in the biosciences, medicine, and in image analysis. Among the substantial body of literature in this field are the books Dryden and Mardia (1998) [10], Kendall et al (1999) [11], Mardia and Jupp (2000) [12], Small(1996) [13], Watson(1983) [14]. While much of this literature focuses on parametric or semiparametric models, we consider a general framework for nonparametric inference for location, as detailed by Bhattacharya and Patrangenaru (2002, 2003, 2005) [15, 1, 16] where general properties of extrinsic and intrinsic mean sets on general manifolds, the problem of consistency of the corresponding sample indices were explored. In addition, we wish to consider the asymptotic distributions of intrinsic and extrinsic sample means and confidence regions based on them. We provide classical CLT-based confidence regions as well as those based on Efron’s bootstrap (Efron(1982)[17]).

Measures of location and dispersion for distributions on a manifold $M$ were studied by Bhattacharya and Patrangenaru (2002 and 2003) [15, 1] as Fréchet parameters associated with two types of distances on $M$. If $j : M \rightarrow \mathbb{R}^k$ is an embedding, the Euclidean distance restricted to $j(M)$ yields the extrinsic mean set and the extrinsic total variance. On the other hand, a Riemannian distance on $M$ yields the intrinsic mean set and intrinsic total variance.
The Fréchet mean of a probability measure $Q$ on a complete metric space $(M, \rho)$ is the minimizer of the function $F(x) = \int \rho^2(x, y)Q(dy)$, when such a minimizer exists and is unique (Fréchet (1948)). However, there is no known condition for the existence of the Fréchet mean of an arbitrarily spread random object on a complete separable metric space.

The intrinsic mean $\mu_I(Q)$ is the Fréchet mean of a probability measure $Q$ on a complete $d$-dimensional Riemannian manifold $M$ endowed with the geodesic distance $d_g$ determined by the Riemannian structure $g$ on $M$. It is known that if $Q$ is sufficiently concentrated, then $\mu_I(Q)$ exists (see Theorem 2.2 (a) below).

The extrinsic mean $\mu_E(Q) = \mu_{j,E}(Q)$ of a probability measure $Q$ on a manifold $M$ with respect to an embedding $j: M \rightarrow \mathbb{R}^k$ is the Fréchet mean associated with the restriction to $j(M)$ of the Euclidean distance in $\mathbb{R}^k$. In [1], it was shown that the extrinsic mean of $Q$ exists if the ordinary mean of $j(Q)$ is a non-focal point of $j(M)$, i.e., if there is a unique point $x_0$ on $j(M)$ having the smallest distance from the mean of $j(Q)$. In this case $\mu_{j,E}(Q) = j^{-1}(x_0)$.

It is easier to compute the intrinsic mean if the Riemannian manifold has zero curvature in a neighborhood containing $\text{supp} Q$ (Patrangenaru (2001)). In particular this is the case of distributions on linear projective shape spaces (Mardia and Patrangenaru (2002)). If the manifold has nonzero curvature around $\text{supp} Q$, with respect to a metric of our choice, the intrinsic sample mean associated, can be approximated using a Newton-Raphson type of algorithm (see Pennec (1999) [18]). On the other hand it is straightforward to compute the extrinsic sample mean for a given embedding. It may be pointed out that if $Q$ is highly concentrated as in medical imaging examples in Bhattacharya and Patrangenaru (2003), the intrinsic and extrinsic means are virtually indistinguishable, should the arc and chord distances be close to each other around $\text{supp} Q$.

However, as with traditional Euclidean data, both the intrinsic and extrinsic means are unknown in practice. As such, it is often of primary interest to estimate these means using analogous statistics. If $X_1, \ldots, X_n$ are i.i.d. random objects from a probability distribution on $(\mathcal{M}, \rho)$, the sample Fréchet mean set is the Fréchet mean set of the empirical distribution

$$\hat{Q}_n = \frac{1}{n}(\delta_{X_1} + \cdots + \delta_{X_n}).$$

The consistency of the Fréchet sample mean set, as an estimator of the Fréchet mean set, is essentially due to H. Ziezold (1977) with a strengthening and a detailed proof due to Bhattacharya and Patrangenaru (2003).

Note that for arbitrary separable complete metric spaces, there are no results on asymptotic distributions of Fréchet sample means, since one requires the use of differential calculus on the sample space $(\mathcal{M}, \rho)$ for this purpose. The most general class of separable metric spaces on which one may differentiate, is the class of differentiable manifolds. A manifold
has a differentiable structure if it admits an atlas \( \mathcal{U} \), with the property that for any pair of charts \( ((U, \varphi) \in \mathcal{U}, (V, \psi) \in \mathcal{U}) \), the transition maps \( \varphi \circ \psi^{-1} : \psi(U \cap V) \to \varphi(U \cap V) \) are differentiable. For simplicity we assume that the transition maps are of class \( \mathcal{C}^\infty \).

Manifolds arising as sample spaces in statistics are smooth. These spaces include \( \mathbb{R}^p \) for multivariate analysis, the spheres \( S^{p-1} \) for directional data analysis, certain Lie groups, such as the special orthogonal groups for the analysis of tectonic plate data and the group of positive definite symmetric matrices for digital tensor image analysis, real or complex Grassman manifolds for affine shape analysis and Kendall’s similarity shape analysis, and products of real projective shape spaces for projective shape analysis. The asymptotic distribution of the Fréchet sample means for a random sample from a probability measure on a smooth manifold was derived by Bhattacharya and Patrangenaru (2005). Computations of Fréchet sample means for a given distance are, in general, based on iterative algorithms, making estimation of Fréchet means time consuming.

### 2.3 Extrinsic Analysis on Manifolds

The set of probability measures on a manifold embedded in an Euclidean space that have an extrinsic mean is generic (open and dense) in the space of all probability measures on that manifold. If \( j : \mathcal{M} \to \mathbb{R}^N \) is an embedding of a manifold and \( X \) is a \( j \)-nonfocal random object on \( \mathcal{M} \), then \( \mu_j = j^{-1}(P_j(E(j(X)))) \).

The extrinsic covariance matrix \( \Sigma_j \) of \( X \) is the restriction to the tangent space at the extrinsic mean at the range of \( j \), \( T_{\mu_j} j(\mathcal{M}) \) of

\[
\Sigma_j = d_{\mu_j} P_j \Sigma(d_{\mu_j} P_j)^T, \quad \text{where } \Sigma = \text{Cov}(j(X)).
\]

This can be stated more explicitly in the following manner. The tangential component \( \tan(v) \) of \( v \in \mathbb{R}^k \) with respect to the basis \( e_a(P_j(\mu)) \in T_{P_j(\mu)} j(\mathcal{M}) \), \( a = 1, \ldots, d \). It follows then that the random vector \( (d_{\mu_j} j^{-1}(\tan(P_j(j(X)))) - P_j(\mu))) \) has the following covariance matrix with respect to the basis \( e_a(P_j(\mu)) \in T_{P_j(\mu)} j(\mathcal{M}) \), \( a = 1, \ldots, d \) and is called the extrinsic covariance matrix:

\[
\Sigma_{j,E} = e_a(P_j(\mu)) \Sigma e_b(P_j(\mu))^T, \quad \text{for } 1 \leq a, b \leq d
\]

\[
= \left[ \sum_{a=1}^{d} d_{\mu} P_j(e_b) \cdot e_a(P_j(\mu)) \right] \Sigma \left[ \sum_{a=1}^{d} d_{\mu} P_j(e_b) \cdot e_a(P_j(\mu)) \right]^T,
\]

\[(2.1)\]

From these definitions, it follows that the extrinsic mean and extrinsic covariance are easy to compute.

Similarly, assume \( X = (X_1, \ldots, X_n) \) are i.i.d. \( M \)-valued random variables whose common
distribution is a nonfocal measure $Q$ on $(M,j)$. If the mean $\bar{j}(X)$ of the sample $j(X) = (j(X_1),...,j(X_n))$ is a nonfocal point, the extrinsic sample mean is

$$\overline{X}_E := j^{-1}\left(P_M(\bar{j}(X))\right) \equiv \mu^j_E(\hat{Q}_n),$$

(2.3)

where $\hat{Q}_n = n^{-1}\sum_{i=1}^n \delta_{X_i}$ is the empirical distribution.

A consistent estimator of $\Sigma_j$ is the extrinsic sample covariance matrix

$$G(j,X) = \left[\sum d_j^{-1} P_j(e_b) \cdot e_a(P_j(\overline{j(X)}))\right]_{a=1,...,m} \cdot S_{j,n}$$

(2.4)

$$S_{j,n} = n^{-1}\sum (j(X_r) - \bar{j}(X))(j(X_r) - \bar{j}(X))^t$$

is the sample covariance, and $(e_a(y),a = 1,...,N)$ is an adapted orthoframe field around $P_j(\overline{j(X)})$.

### 2.3.1 Asymptotic distributions of extrinsic sample means

The direct sum $\mathbb{R}^N = T_{\mu_j}(\mathcal{M}) \oplus T_{\mu_j}(\mathcal{M})^\perp$ yields a decomposition of any vector $u \in \mathbb{R}^N, u = u_{tan} + u_\perp$. The standardized X-mean is

$$\overline{Z}_{j,n} := \frac{n}{2} \Sigma_j^{-\frac{1}{2}} X_{tan}^T,$$

(2.5)

If $\{X_r\}_{r=1,...,n}$ are i.i.d. random objects from a $j$-nonfocal distribution $Q$ on $\mathcal{M}$ and $\Sigma_j$ of $Q$ is finite. Then (a) the extrinsic sample mean $\overline{X}_j$ has asymptotically a $N_m(0,n^{-1}\Sigma_j)$ in $T_{\mu_j}\mathcal{M}$, and (b) if $\Sigma_j$ is nonsingular, the $j$-standardized mean vector $\overline{Z}_{j,n}$ in (2.5) converges weakly to $N_m(0,I_m)$.

However, this result cannot be used in practice to construct confidence intervals since the population extrinsic covariance is unknown. For such purposes, instead of using $\Sigma_j$, the consistent estimator $G(j,X)$ is used. It follows that

$$n\|G(j,X)^{-\frac{1}{2}} tan_{P_j(\overline{j(X)})}(P_j(\overline{j(X)}) - P_j(\mu))\|^2$$

(2.6)

converges weakly to $\chi^2_m$.

### 2.3.2 Confidence regions for extrinsic means

Utilizing the above results, a large sample confidence region for $\mu_j$ of asymptotic level $1-\alpha$ is given by (a) $C_{n,\alpha} := j^{-1}(U_{n,\alpha})$, where $U_{n,\alpha} = \{ \mu \in j(\mathcal{M}) : n\|G(j,X)^{-\frac{1}{2}} tan_{P_j(\overline{j(X)})}(P_j(\overline{j(X)}) - P_j(\mu))\|^2 \leq \chi^2_{m,1-\alpha} \}$, or by
(b) \( D_{n,\alpha} := j^{-1}(V_{n,\alpha}) \), where \( V_{n,\alpha} = \{ \mu \in j(M) : n\|G(j,X)^{-\frac{1}{2}}tan_{P_j(j(X))}(P_j(j(X)) - P_j(\mu))\|^2 \leq \chi_{m,1-\alpha}^2 \} \).

However, for many circumstances a large sample is not available. In those cases, an alternative method to obtain a confidence region for \( \mu_j \) is needed. One such approach is the use of nonparametric bootstrap. A 100(1 - \alpha)\% nonparametric bootstrap confidence region for \( \mu_j \) is \( D^{*}_{n,\alpha} := j^{-1}(V^{*}_{n,\alpha}) \) with \( V^{*}_{n,\alpha} \) given by

\[
V^{*}_{n,\alpha} = \{ \mu \in j(M) : n\|G(j,X)^{-\frac{1}{2}}tan_{P_j(j(X))}(P_j(j(X)) - P_j(\mu))\|^2 \leq d^{*}_{1-\alpha} \}, \tag{2.7}
\]

where \( d^{*}_{1-\alpha} \) is the upper 100(1 - \alpha)\% point of the values

\[
n\|G(j,X)^{-\frac{1}{2}}tan_{P_j(j(X))}(P_j(j(X)) - P_j(\mu))\|^2 \tag{2.8}
\]

among the bootstrap resamples.

This region has coverage error \( O_p(n^{-2}) \), so \( n_{\text{bootstrap}}^{-\frac{1}{2}} \sqrt{n_{\text{asymptotic}}} \).

### 2.4 Intrinsic Analysis on Manifolds

Let \((M, g)\) be a \(d\)-dimensional connected and complete Riemannian manifold, i.e., \(M\) is a \(d\)-dimensional \(C^\infty\) connected manifold with a complete Riemannian metric \(g\). Denote by \(d_g\) the (geodesic) distance under \(g\). We consider \(M\)-valued random variables \(X\), i.e., measurable maps on a probability space \((\Omega, \mathcal{A}, P)\) into \((M, \mathcal{B})\), where \(\mathcal{B}\) denotes the Borel sigma-algebra of \(M\). All probability measures on \(M\) below are defined on \(\mathcal{B}\).

Let \(Q\) be a probability measure on the metric space \((M, d)\). The Fréchet mean set of \(Q\) is the set of all minimizers of the map \(F\) on \(M\) defined by

\[
F(p) = \int d^2(p, x)Q(dx), \quad p \in M. \tag{2.9}
\]

If \(M\) is a Riemannian manifold, the Fréchet mean (set) with respect to the geodesic distance \(d = d_g\) is defined to be the intrinsic mean (set) of \(Q\); if the minimizer is unique, the intrinsic mean will be labeled \(\mu_I\). For Riemannian manifolds, the points in the intrinsic mean set are points of local minima of \(F\) and are therefore also referred to as Karcher means (W. Kendall (1990) [19], Le (1998) [4]).

Let \(X_1, \ldots, X_n\) be independent random variables with a common distribution \(Q\) on a metric space \((M, d)\), and consider their empirical distribution \(Q_n = \frac{1}{n} \sum_{k=1}^{n} \delta_{X_k}\). The intrinsic sample mean is the minimizer \(m\) of \(p \rightarrow \frac{1}{n} \sum_{j=1}^{n} d_g^2(X_j, p)\), where \(d_g\) is the geodesic distance, and is a consistent estimator of the intrinsic mean.
2.4.1 Asymptotic distributions of intrinsic sample means

Given \( q \in M \), the exponential map \( \text{Exp}_q : U \rightarrow M \) is defined on an open neighborhood \( U \) of \( 0 \in T_q M \) by the correspondence \( v \rightarrow \gamma_v(1) \), where \( \gamma_v(t) \) is the unique geodesic satisfying \( \gamma(0) = q, \dot{\gamma}(0) = v \), provided \( \gamma(t) \) extends at least to \( t = 1 \).

Let \( \mu_n = \phi(\mu_{n,1}) \) and \( \mu = \phi(\mu_I) \) be the images of, respectively, the intrinsic sample mean \( \mu_{n,1} \) of \( \hat{Q}_n = \frac{1}{n} \sum_{i=1}^n \delta X_i \) and the intrinsic mean \( \mu_I \) of \( Q \) under the inverse \( \phi \) of the exponential map, \( \phi = (\text{Exp}_{x_o})^{-1} \). Define the distance \( \rho^\phi(u,v) := \rho(\phi^{-1}(u),\phi^{-1}(v)) \), \( u,v \in \phi(U) \).

Under certain conditions described in Bhattacharya and Patrangenaru (2005) [16],

\[
\sqrt{n}(\mu_n - \mu) \xrightarrow{L} \mathcal{N}(0, \Lambda^{-1}C(\Lambda^T)^{-1}), \tag{2.10}
\]

where \( C \) is the covariance matrix of \( \Psi(\hat{X}_i; \mu) \), the gradient with respect to \( \theta \) of \( (\rho^\phi)^2(u, \theta) \), and \( \Lambda = E((D_r\Psi(X_i; \mu)))^d_{r,r'=1} \).

2.4.2 Confidence regions for intrinsic means

In order to obtain a confidence region for \( \mu_I \) using the above asymptotic results in the traditional manner, one needs to estimate the covariance matrix \( \Gamma = \Lambda^{-1}C(\Lambda^T)^{-1} \). For this purpose, one may use proper estimates of \( \Lambda \) and \( C \), namely,

\[
\hat{\Lambda}(\theta) := \frac{1}{n} \sum_{i=1}^n (\text{Grad}\Psi)(\hat{X}_i; \mu_n), \quad \hat{C}_n = \text{Cov}\hat{Q}_n, \\
\hat{\Gamma} := \hat{\Lambda}^{-1}\hat{C}(\hat{\Lambda}^T)^{-1}, \quad \hat{\Gamma}^{-1} = \hat{\Lambda}^t\hat{C}^{-1}\hat{\Lambda} \tag{2.11}
\]

It then follows immediately that under the previous conditions and if \( \hat{C} \) is nonsingular, a large sample confidence region for \( \mu_I \) of asymptotic level \( 1-\alpha \) is given by \( U_{n,\alpha} := \phi^{-1}(D_{n,\alpha}) \), where \( D_{n,\alpha} = \{ v \in \phi(U) : n(\mu_n - v)^t\hat{\Gamma}^{-1}(\mu_n - v) \leq x_{\alpha/2}^2 \} \).

However, for small samples, a confidence region for \( \mu_I \) can instead be found using the bootstrap. Let \( X_{i,n}^* \) be i.i.d. with common distribution \( \hat{Q}_n \) (conditionally, given \{ \( X_i : 1 \leq i \leq n \) \}. Write \( \hat{X}_{i,n}^* = \phi(X_{i,n}^*) \), \( 1 \leq i \leq n \) and let \( \mu_{n}^* \) be a measurable selection from the Fréchet mean set of \( \hat{Q}_n^* \phi := \frac{1}{n} \sum_{i=1}^n \delta \hat{X}_{i,n}^* \). Let \( E_{n,\alpha} \) be a subset of \( \phi(U) \), such that \( P^*(\mu_n^* - \mu_n \in E_{n,\alpha}^*) \rightarrow 1 - \alpha \) in probability, where \( P^* \) denotes the probability under \( \hat{Q}_n \).

It follows that under the above conditions, \( \phi^{-1}((\mu_n^* - E_{n,\alpha}^* \cap \phi(U))) \) is a confidence region for \( \mu_I \) of asymptotic level \( (1 - \alpha) \) with coverage error of \( O_p(n^{-2}) \).
CHAPTER 3

HYPOTHESIS TESTING FOR MEANS ON HILBERT SPACES

The field of multivariate analysis is built upon the generalization and extension of theory, ideas, and methodology from the analysis of univariate data. For example, Hotelling’s $T^2$ statistic [20] is a multivariate extension of Student’s $t$ statistic, allowing for the implementation of inference procedures for mean vectors. Unfortunately, however, this is not the case for further generalization to data from infinite-dimensional Hilbert spaces. For inference about a mean function, the $T^2$ statistic cannot be extended because the rank of the sample covariance operator is limited by the size of the sample, and as such, will never be full. This makes proper studentization as in the $T^2$ statistic impossible.

However, a long-standing approach for performing inference for such data is replacing the usual null hypothesis of equality with approximate equality. This type of procedure is often called a neighborhood hypothesis test or precise hypothesis test and has been studied and utilized in a variety of applications by a number of researchers. Among these studies are Hodges and Lehmann (1954), Berger and Delampady (1987) [21], Dette and Munk (1998, 2003) [22, 23], Munk and Dette (1998) [24], Goutis and Robert (1998) [25], and Liu and Lindsay (2005) [26]. In the context of shape analysis, this type of procedure is essentially testing for a sufficient visual similarity between shapes. Unless otherwise noted, the methodology presented in this chapter follows from Munk et al. (2008) [27], which utilizes neighborhood tests for one-sample and multi-sample problems in 2-dimensional projective shape analysis.

3.1 Hilbert-Valued Random Objects

Let $\mathbb{H}$ denote a separable Hilbert space over $\mathbb{R}$ with inner product $\langle \cdot, \cdot \rangle$ and norm $\| \cdot \|$. If $(\Omega, \mathcal{F}, P)$ is the underlying probability space, then let $\mathcal{B}$ be the $\sigma$-field generated by open subsets of $\mathbb{H}$. The random object $X$ is a $(\mathcal{F}, \mathcal{B})$- measurable mapping from $\Omega$ to $\mathbb{H}$.
If $E\|X\|^2 < \infty$, then there exist a mean vector $\mu \in \mathbb{H}$ and a covariance operator $\Sigma : \mathbb{H} \to \mathbb{H}$ that are uniquely determined by:

$$
E\langle x, X \rangle = \langle x, \mu \rangle, \forall x \in \mathbb{H}
$$
$$
E\langle x, X - \mu \rangle \langle y, X - \mu \rangle = \langle x, \Sigma y \rangle \forall x, y \in \mathbb{H}
$$

The covariance operator $\Sigma$ is linear, semi-positive definite, Hermitian, and has finite trace.

Let $X_1, X_2, \ldots, X_n$ be i.i.d. random objects in $\mathbb{H}$ with $E\|X_1\|^4 < \infty$ and mean $\mu$ and covariance $\Sigma$. These parameters are estimated, respectively, as follows:

$$
\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i
$$
$$
S = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X}) \otimes (X_i - \overline{X})
$$

### 3.2 The One-Sample Procedure

Let $X_1, X_2, \ldots, X_n$ be i.i.d. random objects in $\mathbb{H}$ and $\delta$ be an arbitrary positive number. If $M$ is a linear subspace in $\mathbb{H}$ of dimension $m$, then the orthogonal projection onto $M$ is denoted by $\Pi$ and the orthogonal projection onto $M^\perp$ is $\Pi^\perp$. The squared distance of $x \in \mathbb{H}$ to $M$ is defined by the functional

$$
\varphi_M(x) = \|x - M\|^2 = \|\Pi^\perp x\|^2
$$

From here, the following neighborhood hypotheses can be tested:

$$
H_0 : \mu \in M_{\delta} \cup B_{\delta} \quad \text{for some } \delta > 0, \quad \text{vs.} \quad H_0 : \mu \in M_{\delta}^c \cap B_{\delta}^c,
$$

(3.1)

where $M_{\delta} = \{x \in \mathbb{H} : \varphi_M(x) < \delta^2\}$ and $B_{\delta} = \{x \in \mathbb{H} : \varphi_M(x) = \delta^2, \langle \Pi^\perp x, \Sigma \Pi^\perp x \rangle > 0\}$. The most important component of $H_0$ is $M_{\delta}$; $B_{\delta}$ is included so that the power of the test is $\alpha$. The test-statistic for these hypotheses is given by

$$
Z = \{\varphi_M(\overline{X}) - \delta^2\}/\hat{\nu}_n
$$

where

$$
\hat{\nu}_n^2 = \langle \Pi^\perp \overline{X}, S \Pi^\perp \overline{X} \rangle.
$$

Under $H_0$, Munk et al. (2007) proved that $Z \to N(0, 1)$ as $n \to \infty$. It follows, then, that $H_0$ should be rejected at asymptotic level $\alpha$ if $Z > \xi_{1-\alpha}$, where $\xi_{1-\alpha}$ is the $(1 - \alpha)$-th quantile
of the standard normal distribution for $0 < \alpha < 1$.

### 3.3 The Multi-Sample Procedure

For $j = 1, \ldots, p$, let $X_{j1}, X_{j2}, \ldots, X_{jn_j}$ be i.i.d. random objects in $\mathbb{H}$ with mean $\mu_j$ and covariance operator $\Sigma_j$ such that $E\|X_{j1}\|^4 < \infty \ \forall j$. Additionally, the $p$ samples are mutually independent and $\frac{n_j}{n} \to \lambda_j$ as $n \to \infty$, where $n = \sum_j n_j$. For $j = 1, \ldots, p$, define

$$
\overline{X}_j = \frac{1}{n_j} \sum_{i=1}^{n_j} X_{ji},
$$

$$
\overline{X} = \frac{1}{p} \sum_{j=1}^{p} \frac{n_j}{n} \overline{X}_j.
$$

The functional $\psi_n : \mathbb{H}^p \to \mathbb{R}$ is defined as follows:

$$
\psi_n(\overline{X}_1, \ldots, \overline{X}_p) = \sum_{j=1}^{p} \frac{n_j}{n} \| \overline{X}_j - \overline{X} \|^2.
$$

From here, the following neighborhood hypotheses can be tested:

$$
H_0 : (\mu_1, \ldots, \mu_p) \in M_{p,\delta} \cup B_{p,\delta} \text{ for some } \delta > 0, \text{ vs. } H_a : (\mu_1, \ldots, \mu_p) \in M_{p,\delta}^c \cap B_{p,\delta}^c,
$$

where $M_{p,\delta} = \{ x \in \mathbb{H}^p : \psi(x) < \delta^2 \}$ and $B_{p,\delta} = \{ x \in \mathbb{H} : \psi(x) = \delta^2, \sum_{j=1}^{p} \lambda_j \langle \lambda_j x_j - \overline{x}, \Sigma_j (\lambda_j x_j - \overline{x}) \rangle > 0 \}$. These hypotheses essentially test for approximate equality of the means in the $p$ samples. The test-statistic $Z_p$ for these hypotheses is

$$
Z_p = \sqrt{n} \{ \psi_n(\overline{X}_1, \ldots, \overline{X}_p) - \delta^2 \} / \hat{\tau}_{p,n},
$$

where

$$
\hat{\tau}_{p,n} = \sum_{j=1}^{p} \lambda_j \langle \lambda_j x_j - \overline{x}, S_j (\lambda_j x_j - \overline{x}) \rangle > 0
$$

and $S_j$ is the sample covariance operator for the $j$-th sample. As with the one-sample test, under $H_0$, Munk et al. (2007) proved that $Z_p \to N(0, 1)$ as $n \to \infty$. It follows, then, that $H_0$ should be rejected at asymptotic level $\alpha$ if $Z_p > \xi_{1-\alpha}$, where $\xi_{1-\alpha}$ is the $(1 - \alpha)$-th quantile of the standard normal distribution for $0 < \alpha < 1$. 

13
CHAPTER 4

EXTRINSIC ANALYSIS ON HILBERT MANIFOLDS

Much of the work in statistics on manifolds has been focused on data lying in a sample space that is a manifold $\mathcal{M}$ with finite dimension $p$. For such data, the sample space is locally homeomorphic to $\mathbb{R}^p$ and extrinsic analysis can be performed by embedding $\mathcal{M}$ in a Euclidean space of higher dimension.

However, for some types of infinite-dimensional data, such as direct similarity shapes of planar contours, the sample space is a manifold that is locally homeomorphic to a Hilbert space $\mathbb{H}$. This chapter discusses theory and methodology for performing extrinsic analysis on such manifolds.

4.1 Frechét Differentiation

In this chapter we assume that $\mathbb{H}$ is a (separable) Hilbert space over the reals. Any such space is isometric with $l_2$, the space of sequences $x = (x_n)_{n \in \mathbb{N}}$ of reals for which the series $\sum_{n=0}^{\infty} x_n^2$ is convergent, with the scalar product $\langle x, y \rangle = \sum_{n=0}^{\infty} x_n y_n$. A Hilbert space with the norm $\|v\| = \sqrt{\langle v, v \rangle}$, induced by the scalar product becomes a Banach space. Differentiability can be defined with respect to this norm.

**DEFINITION 4.1** A function $f$ defined on an open set $U$ of a Hilbert space $\mathbb{H}$ is Frechét differentiable at a point $x \in U$, if there is a linear operator $T : \mathbb{H} \rightarrow \mathbb{H}$, such that if we set

$$\omega_x(h) = f(x + h) - f(x) - T(h),$$

then

$$\lim_{h \rightarrow 0} \frac{\|\omega_x(h)\|}{\|h\|} = 0.$$  \hspace{1cm} (4.2)

Since $T$ in Definition 4.1 is unique, it is called the differential of $f$ at $x$ and is labeled $d_x f$. 

14
4.2 Hilbert Manifolds

DEFINITION 4.2 A chart on a separable metric space \((M, \rho)\) is a one to one homeomorphism \(\varphi : U \to \varphi(U)\) defined on an open subset of \(M\) to a Hilbert space \(\mathbb{H}\). A Hilbert manifold is a separable metric space \(M\), that admits an open covering by domain of charts, such that the transition maps \(\varphi_V \circ \varphi_U^{-1} : \varphi_U(U \cap V) \to \varphi_V(U \cap V)\) are differentiable.

Example 1 The projective space \(P(\mathbb{H})\) of a Hilbert space \(\mathbb{H}\), space of all one dimensional linear subspaces of \(\mathbb{H}\), has a natural structure of Hilbert manifold modeled over \(\mathbb{H}\). Define the distance between two vector lines as their angle, and, given a line \(L \subset \mathbb{H}\), show a neighborhood \(U_L\) of \(L\) can be mapped, via a homeomorphism \(\varphi_L\) onto an open neighborhood of the orthocomplement \(L^\perp\) by using the decomposition \(\mathbb{H} = L \oplus L^\perp\). Then for two perpendicular lines \(L_1\) and \(L_2\), show that the transition maps \(\varphi_{L_1} \circ \varphi_{L_2}^{-1}\) are differentiable as maps between open subsets in \(L_1^\perp\), respectively in \(L_2^\perp\). Use a countable orthobasis of \(\mathbb{H}\) and the lines \(L_n, n \in \mathbb{N}\) generated by the vectors in this orthobasis to cover \(P(\mathbb{H})\) with the open sets \(U_{L_n}, n \in \mathbb{N}\). Finally use the fact that for any line \(L^\perp\) and \(\mathbb{H}\) are isometric as Hilbert spaces. The line \(L\) spanned by a nonzero vector \(\gamma \in \mathbb{H}\) is usually denoted \([\gamma]\) when regarded as a projective point on \(P(\mathbb{H})\).

REMARK 4.1 Similarly, one may consider complex Hilbert manifolds, modeled on Hilbert spaces over \(\mathbb{C}\). A vector space over \(\mathbb{C}\) can be regarded as a vector space over the reals, by restricting the scalars to \(\mathbb{R}\); therefore any complex Hilbert manifold automatically inherits a structure of real Hilbert manifold.

4.3 Extrinsic Analysis

For our planar direct similarity shape analysis presented in Chapter 5, we consider a Hilbert space over the complex numbers and its associated projective space, which we have, respectively, denoted by \(\mathbb{H}\) and \(P(\mathbb{H})\).

DEFINITION 4.3 An embedding of a Hilbert manifold \(M\) in a Hilbert space \(\mathbb{H}\) is a one-to-one differentiable function \(j : M \to \mathbb{H}\), such that for each \(x \in M\), the differential \(d_x j\) is one to one, and the range \(j(M)\) is a closed subset of \(\mathbb{H}\) and the topology of \(M\) is induced via \(j\) by the topology of \(\mathbb{H}\).

Example 2 We embed \(P(\mathbb{H})\) in \(L_{HS} = \mathbb{H} \otimes \mathbb{H}\), the space of Hilbert-Schmidt operators of \(\mathbb{H}\) into itself, via the Veronese-Whitney embedding \(j\) given by

\[
j([\gamma]) = \frac{1}{\|\gamma\|^2} \gamma \otimes \gamma.
\]
If \( \|\gamma\| = 1 \), this definition can be reformulated as

\[
j([\gamma]) = \gamma \otimes \gamma.
\] (4.4)

Given that \( \gamma^*(\beta) = \langle \beta, \gamma > \) equation (4.4) is equivalent to

\[
j([\gamma])(\beta) = \langle \beta, \gamma > \gamma.
\] (4.5)

The range of this embedding is the submanifold \( M_1 \) of rank one Hilbert-Schmidt operators of \( H \).

**DEFINITION 4.4** If \( j : M \rightarrow H \) is an embedding of a Hilbert manifold in a Hilbert space, the chord distance \( \rho \) on \( M \) is given by

\[
\rho(x,y) = \|j(x) - j(y)\|,
\]

and given a random object \( X \) on \( M \), the associated Fréchet function is

\[
\mathcal{F}_j(x) = E(\|j(X) - j(x)\|^2).
\] (4.6)

The minimizer(s) of \( \mathcal{F}_j \) is called the extrinsic mean (extrinsic mean set) of \( X \) (see Bhattacharya and Patrangenaru (2003, 2005)). If the extrinsic mean exists (the mean set has one element only), it is labeled \( \mu_{E,j} \) or simply \( \mu_E \).

**PROPOSITION 4.1** Consider a random object \( X \) on \( M \), that has an extrinsic mean set. Then (i) \( j(X) \) has a mean vector \( \mu \) and (ii) the extrinsic mean set is the set of all points \( x \in M \), such that \( j(x) \) is at minimum distance from \( \mu \). (iii) In particular if \( \mu_E \) exists, then \( \mu_E = j^{-1}P_j(\mu) \).

**PROOF 4.1** Let \( Y = j(X) \). Note that the Hilbert space \( H \) is complete as a metric space, therefore \( \inf_{j(y)} E(\|Y - y\|^2) = \min_{j(y)} E(\|Y - y\|^2) \leq \min_{j(M)} E(\|Y - y\|^2) \leq \inf_{j(y)} E(\|Y - y\|^2) \) is finite, which proves (i). To prove (ii), assume for \( \nu \) is a point in the extrinsic mean set, and \( x \) is an arbitrary point on \( \in M \). From \( E(\|j(\nu) - Y\|^2) \leq E(\|j(x) - Y\|^2) \) and since \( j(\nu) - \mu \) and \( j(x) - \mu \), are constant vectors, it follows that

\[
\|j(\nu) - \mu\|^2 \leq \|j(x) - \mu\|^2 + 2E(\langle j(x) - j(\nu), \mu - Y >\).
\] (4.7)

It is obvious that the expected value on the extreme righthand side of equation (5.18) is zero.

**DEFINITION 4.5** A random object \( Q \) on a Hilbert manifold \( M \) embedded in a Hilbert space is \( j \)-nonfocal if there is a unique point \( p \) on \( j(M) \) at minimum distance from \( E(j(X)) \).
**PROPOSITION 4.2** Assume $X = [\Gamma]$ is a random object in $P(H)$, with $\|\Gamma\|^2$ finite. Then the Veronese-Whitney extrinsic mean of $X$ exists if and only if $E(\frac{1}{\|\Gamma\|^2} \Gamma \otimes \Gamma)$ has a simple largest eigenvalue, and in this case the extrinsic mean is $\mu_E = [\gamma]$, where $\gamma$ is an eigenvector for this eigenvalue.

**PROOF 4.2** We select an arbitrary point $[\gamma] \in P(H), \|\gamma\| = 1$. The spectral decomposition of $\Lambda = E(\frac{1}{\|\Gamma\|^2} \Gamma)$ is $\Lambda = \sum_{k=1}^{\infty} \delta_k^2 E_k, \delta_1 \geq \delta_2 \geq \ldots$ where for all $k \geq 1, E_k = e_k \otimes e_k, \|e_k\| = 1$, therefore if $\gamma = \sum_{k=1}^{\infty} x_k e_k, \sum_{k=1}^{\infty} x_k^2 < \infty$, then $\|j(\gamma) - \mu\|^2 = \|\gamma \otimes \gamma\|^2 + \sum_{k=1}^{\infty} \delta_k^2 - 2 < \Lambda, \gamma \otimes \gamma$. To minimize this distance it suffices to maximize the projection of the unit vector $\gamma \otimes \gamma$ on $\Lambda$. If $\delta_1 = \delta_2$ there are the vectors $\gamma_1 = e_1$ and $\gamma_2 = e_2$ are both maximizing this projection, therefore there is a unique point $j([\gamma])$ at minimum distance from $\Lambda$ if and only if $\delta_1 > \delta_2$. 
CHAPTER 5

SHAPE ANALYSIS OF PLANAR CONTOURS

Traditional methods of statistical shape analysis for planar configurations represent the configurations as a set of $k$ ordered, labeled points of interest. The set of points that represent the configuration are called a $k$-ad and the points of interest are called landmarks.

Based on these traditional methods, analysis of planar curves has relied on the selection of a small number of landmarks of interest along the curve. Because a small number of points are chosen, the analysis has depended greatly on the manner in which landmarks were selected, requiring the landmarks to be chosen with great care, as described in Dryden and Mardia (1998) [10]. In many cases, this has required an expert to choose landmarks that corresponded to specific points of interest, such as locations of homologous anatomical features of organisms when dealing with biological data. Another common method for landmark selection has been choosing points according to some mathematical property of the curve, such as at points of high curvature.

However, while both of these types of landmarks provide information about key characteristics of a curve, the details of the curve are not captured, resulting in an effective loss of data. Because of this, one can consider including additional landmarks, often called pseudolandmarks, to fill in the gaps created when using only the previously described landmarks. Commonly, these points are chosen systematically, such that they are equally spaced, either along the entire curve or between the carefully chosen landmarks. While doing so includes more information about the curve, if it is subsequently determined that the inclusion of additional landmarks is necessary, the selection process must be restarted in order to maintain the constant spacing. In section 5.1, the development of the now-standard techniques used to represent and perform analysis on finite planar configurations is presented.

Because of the issues that arise when using only a small amount of landmarks when studying curves, alternative approaches should be considered for the study of shapes of planar curves. This is especially desirable now that the computational power needed to perform the necessary calculations for more detailed representations of the curves is readily
available. In recent years, methodology has been developed to treat the curves as continuous functions. The development of such approaches and brief descriptions of such methods are provided in section 5.2.

5.1 Kendall’s Shape Spaces of Finite Planar Configurations

Recall that if \( V \) is a vector space over the commutative field \( F \), then the set of all one dimensional linear subspaces of \( V \) is the projective space of \( V \), and is labeled \( P(V) \). If \( V = F^{d+1} \) we use the standard notation \( P(F^{d+1}) = FP^d \).

D. G. Kendall (1984)[28] defined the planar direct similarity shape space \( \Sigma_2^k \) as a space of orbits of \( k \)-ad’s, labeled points in the Euclidean plane (which is identified with the complex plane) and showed that such a shape space is canonically diffeomorphic with a complex projective space \( P(C^{k-1}) = CP^{k-2} \). General nonparametric indices on shape spaces were first defined by Ziezold (1977[29], 1994[30], 1998[31]) (see also Huckemann and Ziezold (2006)[32]). Ziezold’s suggestion of using Fréchet means were used in statistical shape analysis by Le and Kume (2000)[33], by Kume and Le (2000[34], 2003[35]), and by Bhattcahrya and Patrangenaru (2003)[1]. To be specific, with the notation in Bandulasiri et. al. (2009)[7], or in Huckemann and Hotz (2009) [36], two \( k \)-ads \( z_1 = (z_1^1, \ldots, z_1^k) \), \( z_2 = (z_2^1, \ldots, z_2^k) \) \( \in C^k \) have the same direct similarity shape, if there is a direct similarity \( S \) of the complex plane, \( S(z) = wz + b, w \in C^*, b \in C \), such that \( S(z_1^j) = z_2^j, \forall j = 1, \ldots, k \). If \( \tilde{z}_a = \frac{1}{k}(z_1^1 + \ldots z_1^k) \), then \( S(\tilde{z}_1) = \tilde{z}_2 \), and since \( S \) is linear, if we set

\[
\zeta_a = (\zeta_a^1, \ldots, \zeta_a^k), \quad \zeta_a^j = z_1^j - \tilde{z}_a, \forall j = 1, \ldots, k,
\]

then two nontrivial \( k \)-ads \( z_1 \) and \( z_2 \) (for which the coordinates of \( z_a \) are not the same) have the same direct similarity shape, if and only if \( \zeta_1, \zeta_2 \) are not zero and differ by a nonzero scalar multiple. Note that the centered \( k \)-ads \( \zeta_1, \zeta_2 \) lie in the complex vector hyperplane \( L_k^2 \) of \( C^k \), \( L_k^2 = \{ \zeta = (\zeta^1, \ldots, \zeta^k) \in C^k : \zeta 1_k = 0 \} \) therefore the two nontrivial \( k \)-ads \( z_1 \) and \( z_2 \) have the same direct similarity shape, if and only if \( \zeta_1, \zeta_2 \in L_k^2 \setminus \{0\} \) are on the same 1-dimensional complex vector subspace of \( L_k^2 \), and thus the Kendall shape space \( \Sigma_2^k \) is identified with the complex projective space \( P(L_k^2) \) of the complex vector space \( L_k^2 \).

An example of map identification of \( P(L_k^2) \) with \( CP^{k-2} \) is given by Kendall (op. cit.)[28]. If \( \tilde{Q} \) is the \( (k - 1) \times k \) matrix of coefficients of the linear function from \( C^k \) to \( C^{k-1} \) given by

\[
\xi^j = \frac{1}{\sqrt{j^2 + j}}(j\zeta^{j+1} - (\zeta^1 + \cdots + \zeta^j)), j = 1, \ldots, k - 1,
\]

19
then the bijection
\[ F : P(L^2_k) \rightarrow \mathbb{C}P^{k-2}, [\xi] = F([\zeta]) = [\tilde{Q}\zeta] \]
is such in identification.

**REMARK 5.1** In addition to direct similarity shapes of k-ads (labeled finite configurations), one may define also the direct similarity shape of finite unlabeled configurations (subsets) of size \( k \). Two such subsets \( A \) and \( B \) have the same direct similarity shape if there is a direct similarity \( S \) such that, as sets, \( S(A) = B \). The corresponding shape space of unlabeled planar configurations of size \( k \) can be identified with a subset of \( \mathbb{C}P^{k-2} \). For example, Kendall (op. cit.) showed that, when \( k = 3 \), given that \( \mathbb{C}P^1 \) with the Fubini-Study metric is isometric with a round sphere, the corresponding shape space of unlabeled triads in the Euclidean plane is Kendall’s ‘spherical blackboard’: a spherical triangle, two sides of which are shapes of isosceles triangles (one for triangles with two angles larger than \( \frac{\pi}{3} \) and the other with one angle larger than \( \frac{\pi}{3} \)), and the third side degenerated triangles (collinear triads).

### 5.1.1 Veronese-Whitney Extrinsic Mean Shape

Kendall (op. cit.) considers probability distributions of direct similarity shapes, however his work does not include a definition for parameters of location and spread of such distributions. Kent (1992) [2] defines the full Procrustes estimate of a mean shape \( \hat{\mu} \), and shows that for a given sample \( [\zeta_1], \ldots, [\zeta_n] \) of points on \( \mathbb{C}P^{k-2} \), this estimate is \( \hat{\mu} = [m] \), where \( m \in \mathbb{C}^{k-1} \) is the eigenvector corresponding to the largest eigenvalue of \( \sum_{i=1}^{n} \frac{1}{||\zeta_i||} \zeta_i \zeta_i^* = \frac{1}{n} \sum_{i=1}^{n} j([\zeta_i]) \), assuming this eigenvalue is simple; however in Kent (op. cit.) there is no nonparametric definition for the mean of a random direct similarity shape. Since complex projective spaces do not have a linear structure, basic notions of mean and covariance had to be defined from a new perspective; the first paper to consider this crucial concept is due to Ziezold (1994) [30], who defines the mean set of a random direct similarity shape \( [U] \) of a k-ads, regarded as a random point of \( \mathbb{C}P^{k-2} \) as the Fréchet mean set of \( [U] \) with respect to the Fubini-Study (FS) metric, that was singled out by Kendall among all distances on \( \mathbb{C}P^{k-2} \). If the FS mean set has a unique point \( [\mu] \), from the consistency of the Fréchet sample mean sets due to Ziezold (1977) [29], it follows that the FS sample mean is a consistent estimate of \( [\mu] \). However, even if the spread of the random shape \( [U] \) is small enough to insure that both the FS mean and its sample counterpart are well defined, determination of the FS sample mean given by the iterative algorithm in Le (2001) [37] is computationally expensive. Patrangenaru (1998) [38] defines the Veronese Whitney (VW) extrinsic mean \( \mu_{j1} \) of a random point \( [X] \) on in terms of the VW embedding.
\( j_1 \) of \( \mathbb{C}P^{k-2} \) into the space \( S(k - 1, \mathbb{C}) \) of selfadjoint \((k - 1) \times (k - 1)\) matrices is given by

\[
\begin{align*}
  j_1(\xi) &= \frac{1}{||\xi||^2} \xi^*, \\
  \text{(5.3)}
\end{align*}
\]

as projection of the mean of \( j_1[X] \) on \( j_1(\mathbb{C}P^{k-2}) \), and shows that \( \mu_{j_1} \) exists if \( [X] \) the largest eigenvalue of \( E(XX^*) \) is simple, and in this case \( \mu_{j_1} = [\mu] \), where \( \mu \) is an eigenvector corresponding to the largest eigenvalue of \( E(XX^*) \), and Bhattacharya and Patrangenaru (2005) [16] note that the the VW extrinsic mean is the Fréchet mean of \( [X] \) for the distance on \( \mathbb{C}P^{k-2} \) induced by the Euclidean distance on \( S(k - 1, \mathbb{C}) \) via \( j_1 \). Note that the shape space \( \Sigma^k_2 = \mathbb{P}(L_k^2) \) can be embedded into the space \( S(k, \mathbb{C}) \) of selfadjoint \( k \times k \) matrices using a similar VW embedding \( j \) given by

\[
\begin{align*}
  j(\zeta) &= \frac{1}{||\zeta||^2} \zeta^*, \\
  \text{(5.4)}
\end{align*}
\]

and these VW-embeddings are related by

\[
\begin{align*}
  j_1(\xi) &= K j(\zeta) K^T. \\
  \text{(5.5)}
\end{align*}
\]

Therefore if \( [U] \) is a random point on \( P(L_k^2) \) with a \( j \)-n focal probability measure \( Q \), meaning that the largest eigenvalue of \( E(UU^*) \) is simple (Bandulsiri et. al. (2009) [?]), the extrinsic VW- mean shape \( \mu_j \) is given by the eigenvector corresponding to the largest eigenvalue of \( E(UU^*) \). From this, it follows that if \( \zeta_1, \ldots, \zeta_n \) is a random sample from \( Q \), then the extrinsic sample mean \( \hat{\mu}_{E,n} \) is the projective point of the eigenvector corresponding to the largest eigenvalue of \( \frac{1}{n} \sum_{i=1}^{n} \frac{1}{||\zeta_i||^2} \zeta_i \zeta_i^* = \frac{1}{n} \sum_{i=1}^{n} j([\zeta_i]) \), which coincides with Kent’s full Procrustes mean estimate, and shows that the Kent full Procrustes mean estimator is a consistent estimator of the VW extrinsic (direct similarity) mean shape. The asymptotic distribution of the VW extrinsic sample mean shape, and the resulting bootstrap distribution are given in Bhattacharya and Patrangenaru (2005) [39], Bandulasiri et. al (2009) [?] and in Amaral et at. (2010) [40].

5.2 Shape Spaces of Planar Curves as Continuous Functions

Motivated by applications in object recognition from digital images, Grenander (1993) [41] considered shapes as points on some infinite dimensional differentiable manifold. A manifold model for direct similarity shapes of planar closed curves, first suggested by Azencott (1994) [42], was pursued in Azencott et. al. (1996) [43], and detailed by Younes (1998 [44], 1999 [45]). This idea gained ground at the turn of the millennium, with more researchers studying shapes of closed planar curves (eg. Sebastian et. al. (2003) [46]). A simple
closed curve $\gamma$ can be regarded as a one to one piecewise differentiable function from the unit circle $S^1$ to $\mathbb{C}$. If we take Kendall’s viewpoint, each point $\gamma(z)$ on this curve can be regarded as a landmark, labeled by the point $z$ on $S^1$. Ignoring the labels in the particular case of simple closed curves is equivalent to identifying two such curves $\gamma_1, \gamma_2 : S^1 \to \mathbb{C}$, if there is an orientation preserving diffeomorphism $\phi : S^1 \to S^1$, such that $\gamma_2 = \gamma_1 \circ \phi$.

According to Klassen et. al. (2004) [47], two curves $\gamma_a : S^1 \to \mathbb{C}$, $a = 1, 2$ have the same direct similarity shape, if there is a direct similarity $S$, and a self diffeomorphism $\phi$ of $S^1$, such that $\gamma_2 = S \circ \gamma_1 \circ \phi$, meaning that, in fact, Klassen et. al. (2004) [47] consider direct similarity shapes of contours (ranges of simple closed curves).

Klassen et. al. (op. cit), Michor and Mumford (2004) [48] and Younes et. al. (2008) [49] follow Small (1996) [13] or Kendall’s choice, and define a Riemannian structure on the resulting shape manifold. While discretization is unavoidable when performing computations, the idea behind these approach is to treat the curves as continuous functions for as long as possible and only then discretizing when it is no longer possible to avoid doing so.

### 5.2.1 Bending-only Representation

Klassen et al (2004)[47] developed an initial framework for this approach. To avoid representing the curves using landmarks, they employed two alternative approaches, representing curves either as direction functions or as curvature functions. To do so, they used arclength parametrization $\alpha : \mathbb{R} \to \mathbb{R}^2$ constrained to having period $2\pi$ and $|\alpha'(s)| = 1$ for every $s$. For each $\alpha$, there is an associated velocity function $v : \mathbb{R} \to S^1$ such that $v(s) = \alpha'(s) = e^{i\theta(s)}$, where $\theta(s)$ is an angle function that describes the angle that the vector $\alpha'(s)$ makes with the positive $x$-axis. The curvature representation for $\alpha$ is defined as $\kappa(s) = \theta'(s)$, assuming that $\alpha$ is twice differentiable.

To perform analysis on such representations, they proposed to use a metric on $L^2$, the space of all real-valued functions $\mathbb{R} \to \mathbb{R}$ that have period $2\pi$ and are square integrable on $[0, 2\pi]$. In this environment, they define $\langle f_1, f_2 \rangle = \int_0^{2\pi} f_1(s)f_2(s)\,ds$. This metric was chosen because of its interpretation as being the ”bending” energy used in going from one function to another.

Using these notions, the authors defined their representations of shape. Due to the formulations for direction and curvature representations being similar to each other and because of subsequent developments in the methodology, we focus here on their direction function representation. For any closed curve for which $\theta(s + 2\pi) - \theta(s) = 2\pi$, the direction function is of the form $\theta = \theta_0 + f$, where $f \in L^2$ and $\theta_0$ is the direction function of a unit circle. It should be noted that $\theta_0 + L^2$ is an affine space, not a vector space. It follows that adding a constant to $\theta$ results in a rotation of the curve in the plane. To make shapes invariant to rotation, they remove such actions by restricting to those $\theta$ for which
\[
\frac{1}{\pi} \int_0^{2\pi} \theta(s) \, ds = \pi.
\]
To restrict to closed curves, the authors also imposed the condition that
\[
\int_0^{2\pi} e^{i\theta(s)} \, ds = 0.
\]
They defined the preshape space \( C_1 \) to be all elements of \( \theta_0 + L^2 \) that satisfy both of the above restrictions. To define a shape space \( S_1 \), they also remove the effect of reparametrization by allowing for different placements of \( s = 0 \) along the curve.

To describe the differences between shapes, the authors described an iterative algorithm with which to calculate geodesics between elements of \( S_1 \). A geodesic is the shortest path, with constant speed, between two points on a manifold with respect to a Riemannian metric. Here, this metric is given by the \( L^2 \) inner product on the tangent space. The length of such a geodesic provides an intrinsic metric on \( S_1 \).

Letting \( d(\theta_i, \theta_j) \) denote the length of the geodesic from \( \theta_i \) to \( \theta_j \) in \( S_1 \), the authors defined the Karcher mean shape to be any point \( \mu \in S_1 \) which is a local minimizer of \( \sum_{j=1}^n d(\theta, \theta_j)^2 \) with respect to \( \theta \). It should be noted that while the Karcher mean will always exist for a set of points on \( S_1 \), it is not necessarily unique. There is no analytic expression for the calculation of the Karcher mean, instead requiring the use of an iterative algorithm, described by the authors, to be found.

### 5.2.2 Square-Root Elastic Representation

While the framework established in Klassen et al (2004) \[47\] succeeded in achieving many of the goals of the authors, that metrics used only measured the "bending" energy and did not measure the use of stretching that could be used as a deformation when moving between shapes. Developing such a framework was explored in Mio and Srivastava (2004) \[50\], Mio et al (2005) \[51\], and Joshi et al (2007) \[52, 53\].

Using similar techniques and ideas as Klassen et al (2004) , Mio and Srivastava (2004) represented curves as elastic strings and developed methodology for defining and analyzing shape under such representation. However, this paper only considered open curves. Mio et al (2005) extended this framework further to also include closed curves, as well as addressing the calculation of shape statistics for both types of curves.

In both papers, curves are represented in the same manner. Letting \( I = [0, 1] \), the authors defined a smooth parametric curve \( \alpha : I \to \mathbb{R}^2 \), where \( \alpha'(t) \neq 0 \forall t \in I \). From here, they expressed the velocity vector using log-polar coordinates as

\[
\alpha'(t) = e^{\phi(t)} e^{i\theta t} \tag{5.6}
\]

Here, \( \phi : I \to \mathbb{R} \) is a smooth function that can be interpreted as the log-speed and \( \theta : I \to \mathbb{R} \) is a smooth function that represents the angle that the velocity vector makes with the \( x \)-axis. Alternatively, \( \phi \) can be interpreted as a measurement of the rate at which \( I \) is compressed or stretched to form the curve and \( \theta \) provides information about the bending needed, as
in Klassen et al (2004). Using this interpretation, \( \phi(t) < 0 \) denotes a local compression of \( I \) and \( \phi(t) > 0 \) denotes a local stretching. It should also be noted that the arc-length parametrized curves from Klassen et al (2004) can be represented in this notation using \( \phi \equiv 0 \). Using this expression for the velocity vector, the authors define the space \( \mathcal{H} \) to be the collection of all \( (\phi, \theta) \), where each pair represents a corresponding \( \alpha \).

For both papers, the authors consider parametric curves that differ by at most affine transformations of the plane or reparametrizations to be representing the same shape. Letting \( \gamma : I \rightarrow I \) be an orientation-preserving diffeomorphism, the authors defined \( \beta(t) = \alpha(\gamma(t)) \) to be a reparametrization of \( \alpha \) by \( \gamma \). Mio et al (2005) shows that, using this definition,

\[
\beta'(t) = \alpha'(\gamma(t))\gamma'(t) = e^{\phi(\gamma(t))} e^{i\theta(\gamma(t))} \gamma'(t) = e^{\phi(\gamma(t)) + \log(\gamma'(t))} e^{i\theta(\gamma(t))} \gamma'(t),
\]

which allows \( \beta \) to be represented by the pair \( (\phi \circ \gamma + \log(\gamma'), \theta \circ \gamma) \), where \( \gamma \) since \( \gamma \) is a diffeomorphism. Thus an action of the orientation-preserving reparametrization group \( D_I \) on \( \mathcal{H} (\phi, \theta) \cdot \gamma \) results in the preceding pair.

In order to quantify shape deformations in this setting, the authors defined a Riemannian metric on \( \mathcal{H} \) in the following manner. Because the space \( \mathcal{H} \) is linear, the tangent space at any point in \( \mathcal{H} \) is \( \mathcal{H} \), itself. Based on this, given \( (\phi, \theta) \in \mathcal{H} \), and two vectors \((h_1, f_1)\) and \((h_2, f_2)\) that are tangent to \( \mathcal{H} \) at \( (\phi, \theta) \), for \( a, b > 0 \), they defined the inner product to be

\[
\langle (h_1, f_1), (h_2, f_2) \rangle_{(\phi, \theta)} = a \int_0^1 h_1(t) h_2(t) e^{\phi(t)} \, dt + b \int_0^1 f_1(t) f_2(t) e^{\phi(t)} \, dt.
\]

(5.8)

Each of the integrals corresponds to the standard \( L^2 \) inner product with respect to the arc-length parameter \( s \), as used in Klassen et al (2004), because \( ds = e^{\phi(t)} \, dt \). This implies that (5.8) is invariant to any reparametrization \( \gamma \). The elastic nature of the curves is built into this model by way of \( a \) and \( b \), which can, respectively, be thought of as tension and rigidity coefficients, which are often restricted such that \( a + b = 1 \). Large values of the ratio \( a/b \) suggest that the curve is highly resistant to stretching and compression compared to bending.

Because curves are represented by \( (\phi, \theta) \), it follows that this representation of shapes of curves is invariant to translations in the plane. To impose uniform scaling and rotational invariance, Mio and Srivastava (2004) restrict shape representatives for open curves to those satisfying, respectively, \( \int_0^1 e^{\phi(t)} \, dt = 1 \) and \( \int_0^1 \theta(t) e^{\phi(t)} \, dt = \pi \). For closed curves, Mio et al (2005) impose additional closure constraints, such that \( \int_0^1 \cos(\theta(t)) e^{\phi(t)} \, dt = 0 \) and \( \int_0^1 \sin(\theta(t)) e^{\phi(t)} \, dt = 0 \). The sets of curves satisfying the above conditions comprise,
respectively, the preshape space $A$ of open curves and the preshape space $C$ of closed curves.

For both open and closed curves, the respective shape spaces are defined to be quotient space with reparametrization removed. That is to say that $S^a = A/D_T$ and $S^c = C/D_T$. The authors define a pseudo-metric on the desired shape space by

$$d(s_1, s_2) = \min_{(\phi_1, \theta_1), (\phi_2, \theta_2)} d((\phi_1, \theta_1), (\phi_2, \theta_2)),$$

where $(\phi_j, \theta_j)$ correspond to the elements of the preshape space that represent $s_j$ for $j = 1, 2$. Letting $(0, \theta^*_j)$ to be the representative of $s_j$ that is parametrized by arc-length, then any preshape representing $s_j$ can be expressed as $(0, \theta^*_j) \cdot \gamma$ for $\gamma \in D_I$. The authors show that if the representative of $s_1$ is taken to be $(0, \theta^*_1)$, then (5.9) can be reexpressed as

$$d(s_1, s_2) = \min_{\gamma \in D_I} d((0, \theta^*_1), (0, \theta^*_2) \cdot \gamma).$$

In order to calculate the geodesic, the authors show that in order to calculate the geodesic, it is necessary to find both the optimal reparametrization $\gamma^*$ and the appropriate shooting direction $(h, f)$ to go from $(0, \theta^*_1)$ to $(0, \theta^*_2) \cdot \gamma^*$. Both papers suggest using an iterative algorithm for solving this problem that alternates between the variables $(h, f)$ and $\gamma$. While both suggest using a gradient descent method to optimize for $(h, f)$, the papers suggest different methods for finding $\gamma^*$. Mio and Srivastava (2004) suggest also using a gradient descent method to solve this problem, as well, whereas Mio et al (2005) suggests using a dynamic programming algorithm to accomplish this goal. The latter paper acknowledges that while the optimization for the reparametrization is slow, it is a necessary aspect of finding geodesics in the shape spaces. Mio et al (2004) refers to the algorithm used in Klassen et al (2004) to calculate the Karcher mean shape, but adapts it to the constraints for the elastic shape spaces.

Subsequently, work was done in simplifying and improving the representation and calculations needed for analysis of elastic planar curves by Joshi et al (2007) [52, 53], while also extending the framework to curves in $n$ dimensions. However, here we consider only the case of $n = 2$. In [52], the authors redefined the representation using the square-root velocity function $\sqrt{r(s)} e^{i\theta(s)}$, where $r(s)$ is the velocity of the curve at $s$. The advantages to this approach include that it uses just single function to represent the curve and that it is the only known representation for which the elastic metric simplifies to the $L^2$ metric, which is the same for all points and is simpler to implement than under the previous representation. In addition, the preshape space is a unit-sphere inside a Hilbert space, which has a geometry for which the computations are simplified. Despite the changes to the methodology used, this representation has convenient, isometric mappings to the previous representations.
The main idea behind this algorithm is to iteratively straighten a path between shapes until used in Klassen et al (2004) [47] and Mio and Srivastava (2004) [50] because it is more stable.

The authors in [52] here let $I = [0, 2\pi]$ and define $\beta : I \rightarrow \mathbb{R}^2$ to be a planar curve with a non-vanishing derivative everywhere. They denote the shape of a curve by $q : I \rightarrow \mathbb{R}^2$ by

$$q(s) = \frac{\beta'(s)}{\sqrt{\|\beta'(s)\|}} \in \mathbb{R}^2, \forall s \in I,$$

(5.11)

where $\| \cdot \|$ denotes the standard Euclidean norm in $\mathbb{R}^2$. The square-root of the instantaneous speed of $\beta$ is represented by $\|q(s)\|$ and the unit tangent vector along $\beta$ for each $s \in [0, 2\pi)$ is given by $\frac{q(s)}{\|q(s)\|}$. The curve can be recovered from this representation by letting $\beta(s) = \beta(0) + \int_0^s \|q(t)\|q(t)dt$. The authors let $Q = \{q = (q_1, q_2) : I \rightarrow \mathbb{R}^2| \forall i, q_i \in \mathbb{L}^2, \forall s, q(s) \neq 0\}$ be the space of all vector-valued functions representing curves. It follows, then, that $B = \{q \in Q| \int_0^{2\pi} \langle q(s), q(s) \rangle_{\mathbb{R}^2} ds = 1\}$ is the space of all unit-length, elastic curves.

In the case of closed curves, the additional constraint that $\int_0^{2\pi} \beta'(s)ds = 0$ must also be imposed. In this representation, the closure condition can be reexpressed as $\int_0^{2\pi} \|q(t)\|q(t)ds = 0$. Defining the mapping $G : Q \rightarrow \mathbb{R}^2$ by $G_i = \int_0^{2\pi} \|q(t)\|q_i(t)dt$ for $i = 1, 2$, the authors define $A = G^{-1}(0, 0)$ to be the space of all closed, elastic curves. [52] shows that the tangent space of $Q$ at any point is $Q_i$, itself. The authors define the inner product for $u, v \in T_q(Q)$ to be $\langle u, v \rangle \equiv \int_0^{2\pi} \langle u(s), v(s) \rangle_{\mathbb{R}^2} ds$. It follows that $C = A \cap B \subset Q$ is the space of all unit-length, closed, elastic curves that are invariant to translation and scaling, which can be thought of as the preshape space.

In [53], the authors described that in addition to translation and scaling, shape should also be invariant to the placement of the starting point, rigid rotations, and reparametrization by speed. A change in the placement of the starting point can be considered to be an action of $\mathbb{S}^1$ on $q$. A rigid rotation is a group action of $SO(2)$, the special orthogonal group of 2 by 2 matrices, on $q$. Lastly, reparametrization by speed is defined to be a right-action of $D$, the group of orientation-preserving and origin-preserving diffeomorphisms. It then follows that the elastic shape space is the quotient space $S = C / (\mathbb{S}^1 \times SO(2) \times D)$.

The authors in [52] show that a geodesic path in $S$ is given by

$$d(q_0, q_1) = \min_{r \in \mathbb{S}^1, O_2 \in SO(2), \gamma \in D} d(q_0, (r \cdot O_2 q_1) \cdot \gamma).$$

(5.12)

In order to find the geodesic, an iterative algorithm must be used. For each iteration, a geodesic must first be found under all rotations in $C / (\mathbb{S}^1 \times SO(2))$, which is then used as an initial condition for computing the geodesic in $S$. This process is repeated until convergence.

However, in order to find the geodesic in $C / (\mathbb{S}^1 \times SO(2))$, [52] suggests using what is called the path-straightening method, which is proposed as an alternative to the shooting method used in Klassen et al (2004) [47] and Mio and Srivastava (2004) [50] because it is more stable. The main idea behind this algorithm is to iteratively straighten a path between shapes until
the path becomes a geodesic.

5.3 Kendall’s Shape Spaces of Planar Contours

The remainder of this chapter introduces novel methodology for the study of planar contours in a manner that is substantially less computationally complex than the approaches described in the previous sections. In Section 5.3.1, we extend Kendall’s approach to shape analysis, by considering regular contours, and defining the direct similarity shape of a regular contour as a point on the projective space of a Hilbert space $\mathbb{H}$. The simple geometric regularity condition is generic, and, more importantly, it is the key to transition from labeled contours (parametrized curves) to unlabeled contours, thus avoiding the issue of identifying parametrized curves in Klassen et. al. (op.cit.)[47].

In Section 5.4, we embed the projective space $P(\mathbb{H})$ into the space $L_{HS}$ of Hilbert-Schmidt operators of $\mathbb{H}$ into itself, via the infinite dimensional analogue of a Veronese-Whitney embedding introduced in Chapter 4. This allows us to define and compute the VW extrinsic mean shape of a random contour in Proposition 5.1. Section 5.4.1 is dedicated to asymptotics. The large sample distribution of the sample VW extrinsic mean contour is given in equation (5.27), along with its studentized version in equation (5.20). In practice, however, due to the infinite dimensionality, the extrinsic sample covariance is degenerate, therefore the asymptotics can not be used in practice.

To avoid this misgiving, in Section 5.5 we use discrete approximations of the contours and of the VW embedding, thus reducing the problem of estimating the VW mean shape of a contour as a limit of VW mean shapes of matched configurations of points on a sample of contours. As the number $k$ of selected points increases, care is taken to reorder these landmarks as if they had all been selected initially. The key is to randomly select additional landmarks from the list of remaining points on the contour and resorting the landmarks such that the indices are monotonically increasing. This algorithm is constructed such that the interpolation of the finite configuration of the extrinsic VW sample mean shape, yielding the shape of a piecewise linear contour, converges to the VW sample mean shape contour. This section also provides a fully automated algorithm for selecting the contours from a Black and White image, extracted from a digital camera image to represent a certain feature. Section 5.6 is then dedicated to computations of the VW-extrinsic sample mean shape of planar contours for selected contours from Ben Kimia’s contour database.

In Section 5.7, we develop tests for hypotheses for the VW-extrinsic mean similarity shape in general following a neighborhood methodology given by Munk and Dette (1998)[24] in the context of regression. The neighborhood hypothesis already proved its efficiency in similar hypothesis testing problems in image analysis, when applied to projective shapes of
planar curves (Munk et al. (2007) [24]), where it was shown that for large samples, the test statistic for the null hypothesis has a standard normal distribution, so that nonparametric bootstrap could be used in practical recognition applications. To prove asymptotic normality of the test statistic in our case, which is given in equation (5.42), we use an orthobasis in the space of Hilbert Schmidt operators, adapted to the VW embedding, by extending a similar choice for an adapted frame (see Bhattacharya and Patrangenaru (2005) [16] and in Bhattacharya and Bhattacharya (2008) [54]) to the infinite dimensional case. Three concrete examples of testing the one sample hypothesis for the VW mean contour for contour data from the Ben Kimia library are also given in this section.

A first example of computations of bootstrap distributions for VW means of finite approximations of contours can be found in Amaral et al. (2010) [40]; the algorithm used there is less flexible, using nonrandomized finite approximations of contour data. In Section ??, we consider nonparametric nonpivotal bootstrap as a methodology to obtain a confidence region for the mean direct similarity shape of contours, based on randomized finite approximations of the contours that imposes at the same time a reasonable within sample matching of the pseudo-landmarks as described in Section 4. We also present the results of a simulation study to explore the coverage error of the confidence regions.

The chapter ends with a discussion of potential future work, including a suggestion of the extension of this methodology for analysis of more complicated infinite dimensional features captured in digital images.

### 5.3.1 Infinite-dimensional Planar Similarity Shape Spaces

Here, we extend the notion of direct similarity shape from planar $k$-ads to infinite labelled configurations, including shapes of regular contours. While in general, the labels can be assigned in infinitely many ways, we will dwell here only on contours, boundaries of 2D topological disks in the plane. To keep things simple, we will make the additional assumption that there is a unique point $p_0$ on such a contour, at the maximum distance to its center of mass, so that the label of any other point $p$ on the contour is the counter-clockwise travel time at constant speed from $p_0$ to $p$ and that the total time to travel time needed to travel from $p_0$ to itself around the contour once is the length of the contour. Therefore we consider direct similarity shapes of nontrivial contours in the plane as described above.

A contour $\tilde{\gamma}$ is the range of a piecewise differentiable function, that is parameterized by arclength, i.e. $\gamma : [0, L] \rightarrow \mathbb{C}, \gamma(0) = \gamma(L)$, that is one-to-one on $[0, L]$; therefore a contour can also be regarded as the range of a one-to-one piecewise differentiable function from $S^1$ to $\mathbb{C}$. Recall that the length of a piecewise differentiable curve $\gamma : [a, b] \rightarrow \mathbb{R}^2$, is defined as
follows:
\[
 l(\tilde{\gamma}) = \int_a^b |\frac{d\gamma}{dt}(t)||dt, \tag{5.13}
\]
and its center of mass (mean of a uniform distribution on \(\tilde{\gamma}\)) is given by \(z_{\tilde{\gamma}} = (x_{\tilde{\gamma}}, y_{\tilde{\gamma}})\), where
\[
x_{\tilde{\gamma}} = \frac{1}{L} \int_\gamma xds, \quad y_{\tilde{\gamma}} = \frac{1}{L} \int_\gamma yds. \tag{5.14}
\]

The contour \(\hat{\gamma}\) is said to be regular if \(\gamma\) is a simple closed curve and there is a unique point \(z_0 = \arg\max_{z \in \tilde{\gamma}} \|z - z_{\tilde{\gamma}}\|\).

Two contours \(\hat{\gamma}_1, \hat{\gamma}_2\) have the same direct similarity shape if there is a direct similarity \(S: \mathbb{C} \rightarrow \mathbb{C}\), such that \(S(\hat{\gamma}_1) = \hat{\gamma}_2\).

**REMARK 5.2** The centered curve \(\tilde{\gamma}_0 = \hat{\gamma} - z_{\tilde{\gamma}} = \{z - z_{\tilde{\gamma}}, z \in \hat{\gamma}\}\) has the same direct similarity shape as \(\hat{\gamma}\). Two regular simple closed curves \(\hat{\gamma}_1, \hat{\gamma}_2\) have the same similarity shape if \(\gamma_{2,0} = \lambda \gamma_{1,0}\), where \(\lambda\) is a nonzero complex number.

**REMARK 5.3** A function \(\gamma: S^1 \rightarrow \mathbb{C}\) is centered if
\[
\int_{S^1} \gamma(z)ds = 0. \quad \text{We consider regular contours since the complex vector space spanned by centered functions \(\gamma\) yielding regular contours \(\tilde{\gamma}\) is a pre-Hilbert space. Henceforth we will be working with the closure of this space. This Hilbert space \( \mathbb{H} \), can and will be identified with the space of all measurable square integrable centered functions from \(S^1\) to \(\mathbb{C}\).}
\]

Let \(P(\mathbb{H})\) be the projective space corresponding to the Hilbert space \(\mathbb{H}\).

**REMARK 5.4** Let \(\Sigma_{2}^{\text{reg}}\) be the set of all direct similarity shapes of regular contours, which is the same as the space of all shapes of regular contours centered at zero. From Remarks 5.2 and 5.3, with any regular contour \(\hat{\gamma}\), we associate a unique piecewise differentiable curve \(\gamma\), such that \(\hat{\gamma} = \tilde{\gamma}\), by taking \(\gamma(0) = p_0\), the point at the maximum distance to the center of \(\hat{\gamma}\), and by arc length parameterizing \(\gamma\) in the counter clockwise direction. Therefore \(\Sigma_{2}^{\text{reg}}\) is a dense and open subset of \(P(\mathbb{H})\). Henceforth, to simplify the notation, we will omit the symbol \(\tilde{\gamma}\) in \(\tilde{\gamma}\) and identify a regular contour with the associated closed curve, without confusion.

Caution should be taken in dealing with the existence of regular contours with sections of the contour having extremely high curvature. An example of a contour with such curvature, but not regular, is provided in the left subplot of Fig. 5.1. Using the standard Hilbert metric, the distance between the two contours pictured is near 0, despite the fact that the first contour has locally a very high curvature. To adjust for such situations, we want to consider a neighborhood of regular curves according a modified distance function incorporating the
curvature in order to more properly indicate the difference in the shapes of the two regular contours. This distance function, with the identification of a regular contour with an arc length parameterized curve is defined as

\[ d(\gamma_1, \gamma_2) = \sup_s |\kappa_1(s) - \kappa_2(s)| + \sup_s |\gamma_1(s) - \gamma_2(s)| \]  \hspace{1cm} (5.15)

where \( \kappa_1 \) and \( \kappa_2 \) are, respectively, the curvatures of \( \gamma_1 \) and \( \gamma_2 \). It should be noted, however, that while we use this metric to determine the neighborhood of curves of interest, the Hilbert space metric is still used for all other purposes.

### 5.4 The Veronese-Whitney Extrinsic Mean Shape

Following from the methodology for extrinsic analysis described in Chapter 4, we embed \( P(H) \) in \( \mathcal{L}_{HS} = H \otimes H \), the space of Hilbert-Schmidt operators of \( H \) into itself, via the Veronese-Whitney embedding \( j \) given by

\[ j(\gamma) = \frac{1}{\|\gamma\|^2} \gamma \otimes \gamma. \]  \hspace{1cm} (5.16)
The range of this embedding is the submanifold $\mathcal{M}_1$ of rank one Hilbert-Schmidt operators of $\mathbf{H}$. Note that $P(\mathbf{H})$ is a Hilbert manifold that is embedded in the Hilbert space $\mathcal{L}_{HS}$, and, for any probability measure $Q$ on a Hilbert manifold $M$ embedded in a Hilbert space, we may define the extrinsic mean of $Q$ w.r.t. that embedding, similarly to the finite dimensional case. Just as in the finite dimensional case, we obtain the following result:

**PROPOSITION 5.1** Assume $\Gamma$ is a random object in $P(\mathbf{H})$, with $\|\Gamma\|^2$ finite. Then the Whitney-Veronese extrinsic mean of $\Gamma$ exists if and only if $E(\frac{1}{\|\Gamma\|^2}\Gamma \otimes \Gamma)$ has a simple largest eigenvalue, and in this case the extrinsic mean is $\mu_E = [\gamma]$, where $\gamma$ is an eigenvector for this eigenvalue.

If $\gamma_1, \ldots, \gamma_n$ is a random sample of size $n$ from $\Gamma$, then the extrinsic sample mean $\hat{\mu}_{E,n}$ is the projective point of the eigenvector corresponding to the largest eigenvalue of $\frac{1}{n} \sum_{i=1}^n \gamma_i \otimes \gamma_i$.

### 5.4.1 Asymptotic Distribution of the Sample Extrinsic Mean Shape

Recall that we embed $P(\mathbf{H})$ in $\mathcal{L}_{HS}$, the space of Hilbert-Schmidt operators of $\mathbf{H}$ into itself, via the Veronese-Whitney embedding $j$ given in (5.16). In our situation, as defined in Proposition 5.1, the extrinsic (Veronese-Whitney) sample mean $\hat{\mu}_{E,n}$ is equal to $e_1$, the eigenvector corresponding to the largest eigenvalue of $\frac{1}{n} \sum_{i=1}^n \gamma_i \otimes \gamma_i$, where $\gamma_1, \ldots, \gamma_n$ is a random sample from $\Gamma$, a random object in $P(\mathbf{H})$. The extrinsic mean $\mu_E$, also defined in Proposition 5.1, is $e_1$, where $e_1$ is the eigenvector corresponding to the largest eigenvalue of $\mu = E(\frac{1}{\|\Gamma\|^2}\Gamma \otimes \Gamma)$. The asymptotic distribution of $\overline{j(X)}_n$ is as follows:

\[
\sqrt{n}(\overline{j(X)}_n - \mu) \xrightarrow{d} N_{\mathcal{L}_{HS}}(0, \Sigma) \quad \text{as} \quad n \to \infty,
\]

where $N_{\mathcal{L}_{HS}}(0, \Sigma)$ denotes a zero mean Gaussian distribution on $\mathcal{L}_{HS}$ with covariance operator $\Sigma$.

From Proposition 5.1, it follows that the projection $P_j : \mathcal{L}_{HS} \to j(P(\mathbf{H})) \subset \mathcal{L}_{HS}$ is given by

\[
P_j(A) = \nu_A \otimes \nu_A,
\]

where $\nu_A$ is the eigenvector of norm 1 corresponding to the largest eigenvalue of $A$, and $P_j(\mu) = j(\mu_E)$. Applying the delta method to (5.17) yields

\[
\sqrt{n}(P_j(\overline{j(X)}_n) - j(\mu_E)) \xrightarrow{d} N_{\mathcal{L}_{HS}}(0, d_\mu P_j \Sigma (d_\mu P_j)^T),
\]

as $n \to \infty$, where $d_\mu P_j$ denotes the differential of the projection $P_j$, evaluated at $\mu$. 31
Again using the above definition of the projection to obtain $P_j(\overline{j(X)}_n) = j(\hat{\mu}_{E,n})$, this can be reexpressed as follows:

$$\sqrt{n} (j(\hat{\mu}_{E,n}) - j(\mu_E)) \xrightarrow{d} N_{LHS}(0, d_\mu P_j \Sigma(d_\mu P_j)^T), \quad (5.20)$$

as $n \to \infty$. It remains to find the expression for $d_\mu P_j$. To do so, we first consider a function $F : \mathbb{H}_1 \rightarrow \mathbb{H}_2$. $F$ is differentiable at a point $x \in \mathbb{H}_1$ if $F(x + h) - F(x) = Th + \omega_x(h)$, for $h \in \mathbb{H}_1$ where $\lim_{h \to 0} \frac{\|\omega_x(h)\|}{\|h\|} = 0$ and $T \in BL(\mathbb{H}_1, \mathbb{H}_2)$, the set of bounded linear operators from $\mathbb{H}_1$ to $\mathbb{H}_2$. The differential of $F$ at $x$ is defined as $d_x F = T$.

To determine the formula for the differential, we must consider the equivariance of the embedding $J$. Because of this, we may assume without loss of generality that $\mu = \text{diag}\{\delta_a^2\}_{a=1,2,3,...}$. As defined previously, the largest eigenvalue of $\mu$ is a simple root of the characteristic polynomial, with $e_1$ as the corresponding complex eigenvector of norm 1, where $\mu_E = [e_1]$. An orthonbasis for $T_{\mu_E}P(\mathbb{H})$ is formed by $e_a, ie_a$, for $a = 2, 3, \ldots$, where $e_a$ is the eigenvector over $\mathbb{R}$ that corresponds to the $a$-th eigenvalue.

For any $\gamma$ which is orthogonal to $e_1$ w.r.t the real scalar product, we define the path $\psi_{\gamma}(t) = [\cos(t)e_1 + \sin(t)\gamma]$. Then $T_{\mu_E}j(P(\mathbb{H}))$ is generated by the vectors tangent to such paths at $t = 0$. Such vectors have the form $\gamma \otimes e_1 + e_1 \otimes \gamma$. In particular, since the eigenvectors of $\mu$ are orthogonal w.r.t the complex scalar product, we may take $\gamma = e_a$, $a = 1, 2, 3, \ldots$, or $\gamma = ie_a$, $a = 1, 2, 3, \ldots$ to get an orthonbasis for $T_{\mu_E}j(P(\mathbb{H}))$. Normalizing these vectors to have unit lengths, we obtain the following orthonormal frame for $a = 1, 2, 3, \ldots$:

$$d_\mu j(e_a) = 2^{-1/2}(e_a \otimes e_1 + e_1 \otimes e_a), \quad (5.21)$$

$$d_\mu j(ie_a) = i2^{-1/2}(e_a \otimes e_1 + e_1 \otimes e_a), \quad (5.22)$$

As stated previously, since the map $j$ is equivariant, we may assume that $\overline{j(X)}_n$ is a diagonal operator $D$, with the eigenvalues $\delta_1^2 > \delta_2^2 \geq \ldots$. In this case,

$$d_{\mu_E} j(e_a) = 2^{-1/2} E_a^1 = F_a^1, \quad (5.23)$$

$$d_{\mu_E} j(ie_a) = i2^{-1/2} E_a^1 = iF_a^1, \quad (5.24)$$

where $E_a^b$ has all entries zero except those in the positions $(a,b)$ and $(b,a)$ that are all equal to 1. From these formulations and computations of the differential of $P_j$ in the finite dimensional case in Bhattacharya and Patrangenaru (2005), it follows that $d_D P_j(E_a^b) = 0$, for all values $a \leq b$, except for $a = 1 < b$. In this case

$$d_D P_j(F_a^1) = \frac{1}{\delta_1^2 - \delta_2^2} F_a^b, d_D P_j(iF_a^1) = \frac{1}{\delta_1^2 - \delta_2^2} iF_a^b. \quad (5.25)$$
Equation (5.25) implies that the differential of the projection $P_j$ at $\mu$ is the operator $Q_1$ given by

$$Q_1 = \sum_{k=2}^{\infty} \frac{1}{\delta_1^2 - \delta_k^2} E_k,$$

(5.26)

where $\delta_1, \delta_2, \ldots$ are the eigenvalues of $E(\frac{1}{\|\Gamma\|^2} \Gamma \otimes \Gamma)$ and $E_1, E_2, \ldots$ are the corresponding eigenprojections. Also, in this situation, $G$ is a normally distributed random element in $L_{HS}$. This results in the tangential component of difference between the VW - images extrinsic sample mean and of the extrinsic mean having an asymptotic normal distribution, albeit with a degenerate covariance operator. From these computations, the asymptotic distribution of this difference can be expressed more explicitly in the following manner.

$$\sqrt{n}(\tan(j(\hat{\mu}_{E,n}) - j(\mu_E))) \xrightarrow{d} Q_1 G$$

(5.27)

However, this result cannot be used directly because $Q_1$, which is calculated using the eigenvalues of $E(\frac{1}{\|\Gamma\|^2} \Gamma \otimes \Gamma)$, and $\mu_E$ are unknown. This problem is solved by estimating $\mu_E$ by $\hat{\mu}_{E,n}$ and $Q_1$ in the following manner.

$$\hat{Q}_1 = \sum_{k=2}^{\infty} \frac{1}{\hat{\delta}_1^2 - \hat{\delta}_k^2} \hat{E}_k,$$

(5.28)

where $\hat{\delta}_1, \hat{\delta}_2, \ldots$ are the eigenvalues of $\frac{1}{n} \sum_{i=1}^{n} \frac{1}{\|\gamma_i\|^2} \gamma_i \otimes \gamma_i$ and $\hat{E}_1, \hat{E}_2, \ldots$ are the corresponding eigenprojections. Using this estimation, the asymptotic distribution is as follows:

$$\sqrt{n}(\tan(j(\hat{\mu}_{E,n}) - j(\mu_E))) \xrightarrow{d} \hat{Q}_1 G$$

(5.29)

Applying this result to (5.20), we obtain the following

$$\sqrt{n}(j(\hat{\mu}_{E,n}) - j(\mu_E)) \xrightarrow{d} d_{\hat{\mu}_n} P_j G \text{ as } n \to \infty,$$

(5.30)

where $\tan$ is the tangential component relative to the tangent space of $j(P(H))$ at $j(\hat{\mu}_{E,n})$ and $\hat{\mu}_n = J(X)_n$ is a consistent estimator of $\mu$. However, it must be noted that because of the infinite dimensionality of $G$, in practice, a sample estimate for the covariance that is of full rank cannot be found. Because of this issue, this result cannot be properly studentized.

### 5.5 Finite Representation of Planar Contours

In order to perform computations involving regular planar contours, the contours must be represented by a finite number $k$ of landmarks. Since we are approximating curves
having infinitely many points, it follows that selecting more landmarks will result in better approximations of a contour. However, for digital imaging data, the number $k$ of landmarks is in fact bounded from above by the number of pixels of the image of the contour. Contour image data can be presented in multiple forms. The data can be provided as an ordered set of landmarks, each being chosen according to some rule, whether systematically selected or based on the requirements of a specific application. In such an instance, provided that all of the observations are made up of $k$ landmarks, the described analysis can be performed immediately.

5.5.1 Random Landmark Selection

If the data is not provided as a set of coordinates, but as an image with each pixel indicating whether that point is on the contour or not, then before analysis can begin, landmarks must be chosen. We propose to do this by selecting a random sample of the contour points by sampling from the uniform distribution on the contour. In order to do this, we first find the point $z_0$ at the largest distance from the center of the contour and choose that as one of the landmarks in order to maintain our methodology. We then randomly select $k-1$ landmarks from the remaining pixels on the curve, making sure to maintain the proper ordering. This is ensured by sorting the indices of the selected pixels in order from smallest to largest.

It is important to choose an appropriate number of landmarks for the given data set. The selected landmarks will be distributed fairly uniformly around the contour for large values of $k$, ensuring that the curve is accurately represented by the landmarks. However, doing so also increases the computational cost of performing calculations. This is most noticeable when finding the bootstrap confidence regions for the mean shape. Choosing too few landmarks, though, while keeping computational cost down, can be extremely detrimental as the landmarks may not be sufficiently uniform to provide adequate coverage of the contour. This can significantly distort the finite representation of the shape, as shown in Fig.5.2. In this particular instance, the shape of the dog, as represented with fewer landmarks, includes no information about one of the ears and very little detail about one of the feet. It should be noted that the extraction of the contour, the selection of landmarks, parametrization and subsequent relabeling of landmarks and the previously described eigenanalysis can be automated, or semi-automated, allowing for efficient execution of the methodology.

When working with digital imaging data, the length of a contour $L$ can be used to assist in determining an appropriate number of landmarks to be chosen for the given data. When a contour is extracted from the image, it will be represented by $K$ pixels. After selecting an initial set of $k$ landmarks as describe above, the length $kL$ of this $k$-ad can be calculated
Figure 5.2: A dog represented by 1000 and, respectively, 200 randomly chosen landmarks

as

\[ kL = \sum_{j=2}^{k+1} \|z_j - z_{j-1}\|, \quad (5.31) \]

where \( z_{k+1} = z_1 \). The \( K \)-pixel approximation of the contour has length \( KL \) and can be calculated similarly.

If \( k \ll K \), as in the case of \( k = 200 \) in Figure 5.2, then \( kL \) may not provide an adequate estimate of \( L \). An appropriate lower bound for the number of landmarks an be determined by randomly selecting landmarks for various values of \( k \). Compute \( kL \) for each of these \( k \)-ads using (5.31) and compute the relative error compared to \( KL \). This should be repeated many times to obtain a mean relative error and standard deviation of the relative error for each value of \( k \) used. To determine an appropriate number of landmarks to use, compare the mean relative error to a desired threshold. Additionally, the distributions of the relative errors could also be examined.

Because the landmarks are sampled from the uniform distribution, when \( k \) is large, we are able to also approximate the arclength parametrization of \( \gamma \) in the following manner:

\[ \gamma\left(\frac{j - 1}{k} kL\right) \approx z_j \quad \text{for} \quad j = 1, \ldots, k+1. \quad (5.32) \]

It follows that an estimate for the arclength of the contour from time \( \frac{j - 1}{k} kL \) to \( \frac{j'-1}{k} kL \),
where $1 \leq j < j' \leq k + 1$ can be calculated as

$$kL_j^{j'} = \sum_{i=j+1}^{j'} \|z_i - z_{i-1}\|.$$ (5.33)

Examples of this landmark selection process from start to finish are shown in Figs. 5.3 and 5.4.

![Figure 5.3: An image of the contour of a boot from the Kimia image database (a) as a silhouette, (b) as a contour, (c) as 200 randomly selected landmarks, and (d) as a contour constructed from the chosen landmarks](image)

5.5.2 Behavior of $k$-ad when Adding Landmarks

As stated previously, as the number of landmarks $k$ used increases, the better the representation of the contour will be. While this is intuitive, it is important to understand the behavior of key characteristics of the $k$-ad for increasingly large numbers of landmarks. These characteristics allow for an increased understanding of the data and assist in the ability to utilize the discrete approximations for analysis.

While in practice there is a maximum number of landmarks for use with digital image data, in theory, as $k \to \infty$, the finite approximation of the contour $\zeta$ will approach the functional representation $\gamma$. As such, since $[k] \in P(L_k)$ and $\Sigma^reg_2$ is a subset of $P(H)$, as
$k \to \infty$, $P(L_k) \to P(H)$. Furthermore, for large values of $k$, we are able to approximate the desired analysis for $\gamma$ using corresponding analysis for the finite approximation $\zeta$ in the following manner.

We can express any vector $x \in \mathbb{C}P^{k-2}$ as $(x_1, \ldots, x_{k-1})$ and any element $y \in \mathbf{H}$ by $(y_1, y_2, \ldots)$. That is to say that for a given orthonormal basis $e_1, e_2, \ldots$ for $\mathbf{H}$, we can write $y = \sum_{i=1}^{\infty} y_i e_i$, so that $\|y\|^2 = \sum_{i=1}^{\infty} y_i^2$. We may identify the Hilbert space $\mathbf{H}$ with $l_2$, the space of square-summable sequences, such that $\sum_{i=1}^{\infty} y_i^2 < \infty$ and define the mapping $\alpha : \mathbb{C}^{k-1} \to l_2$ such that $(x_1, \ldots, x_{k-1}) \overset{\alpha}{\to} (x_1, \ldots, x_{k-1}, 0, \ldots)$. This relation is maintained when considering the respective projective spaces of $\mathbb{C}^{k-1}$ and $l_2$ using the projective mapping $P(\alpha)$. However, as $k$ increases, it does not suffice to simply concatenate another landmark to the previous set. This is due to the fact that the points are ordered. Because of this, if a $k$th landmark were simply appended, it would be restricted to fall between $x_{k-1}$ and $x_1$, which would result in the landmarks no longer being selected from a uniform distribution as $k \to \infty$.

In order to correct this issue, as $k$ increases, the landmarks must be reordered as if they had all been selected initially. This is done by randomly selecting additional land-
marks from the list of remaining points on the contour and resorting the landmarks such that the indices are monotonically increasing. To illustrate this, denote the $k$ initially chosen ordered landmarks as $z_{j_1}, \ldots, z_{j_p}, z_{j_{p+1}}, \ldots, z_{j_k}$ and the newly-chosen landmark as $z_{j_{k+1}}$. Suppose that $j_p < j_{k+1} < j_{p+1}$. Then the landmarks should be reordered as $(z_{j_1}, \ldots, z_{j_p}, z_{j_{k+1}}, z_{j_{p+1}}, \ldots, z_{j_k})$ and relabeled as $(z_{m_1}, \ldots, z_{m_{k+1}})$. This results in the distance between successive landmarks going to 0 in probability as the number of landmarks increases. This can be stated more formally, as follows.

Lemma 1 Letting $kz_j$ denote the $j$-th landmark of the $k$-ad, define

$$||[kz]|| = \max_{j=1,k} ||kz_j - kz_{j+1}||, \text{ } kz_{k+1} = k z_1.$$  \hspace{1cm} (5.34) 

to be the maximum of the distances between successive landmarks. If landmarks are selected from a uniform distribution over $[0, L)$, then $||[kz]|| \overset{p}{\to} 0$.

PROOF 5.1 Recalling the arc-length parameterization for the contour discussed in (5.13),

$$P(||[kz]|| > \epsilon) = P(\text{All } k \text{ landmarks are within the remaining } L - \epsilon \text{ of the contour})$$

We can assume w.l.o.g. that the section of the contour for which the distance between landmarks is greater than $\epsilon$ is over the arc parameterized by the interval $(0, \epsilon)$. In addition, since the landmarks are independently chosen,

$$P(||[kz]|| > \epsilon) = (F(L) - F(\epsilon))^k = \left(\frac{L}{L} - \frac{\epsilon}{L}\right)^k = \left(1 - \frac{\epsilon}{L}\right)^k,$$

where $F$ is the cdf for the uniform distribution over the interval $[0, L)$. Taking the limit of this expression as $k \to \infty$ results in $||[kz]|| \overset{p}{\to} 0$ since

$$\lim_{k \to \infty} P(||[kz]|| > \epsilon) = \lim_{k \to \infty} (1 - \frac{\epsilon}{L})^k = 0$$

The center of mass of the contour will also be better approximated as $k$ increases. Recalling that we denote the center of mass for the uncentered finite approximation of the contour $z$ by $\bar{z}_k$, then the law of large numbers states that as $k \to \infty$, $\bar{z}_k \to \bar{z}_\gamma$. This can be seen by examining the formula for $\bar{z}_k$.

$$\bar{z}_k = \frac{1}{k}(z_1 + \cdots + z_k)$$
From this, we obtain the following:

\[
\tilde{z}_{k+1} = \frac{1}{k+1}(z_1 + \cdots + z_k + z_{k+1})
\]

\[
= \frac{1}{k+1}(k\tilde{z}_k + z_{k+1})
\]

\[
= \frac{k}{k+1}\tilde{z}_k + \frac{1}{k+1}z_{k+1}.
\] (5.35)

Based upon this formulation, \(\tilde{z}_{k+1}\) can be thought of as a weighted average of \(\tilde{z}_k\) and \(z_{k+1}\). For large values of \(k\), \(z_{k+1}\) has a very small contribution to the center of mass. Based upon this, it is apparent that \(\tilde{z}_k \to \tilde{z}_\gamma\) as \(k \to \infty\).

Recall that \(L_k = \{\zeta = (\zeta^1, \ldots, \zeta^k) \in \mathbb{C}^k : \zeta^1 = 0\}\) is the set of all centered \(k\)-ads. It is also important to understand the relationships of these vector subspaces as additional landmarks are added. Most notably, we are interested in examining how the first \(k\) landmarks are recentered when a \((k+1)\)-th landmark is added, which describes the relationship between \(L_k\) and the restriction of \(L^2_{k+1}\) to those \((k+1)\)-ads containing the original \(k\) landmarks. To do so, we consider the definition of the centered landmarks.

Recall that \(z_j\) is the \(j\)th uncentered landmark and allow \(\zeta_{j,k}\) and \(\zeta_{j,k+1}\) to be, respectively, the corresponding centered \(j\)th landmark for the \(k\)-ad and for the \(k+1\)-ad. Alternatively, these can be expressed as

\[
\zeta_{j,k} = z_j - \bar{z}_k, \quad \zeta_{j,k+1} = z_j - \bar{z}_{k+1}
\] (5.36)

Combining these equations results in

\[
\zeta_{j,k+1} = \zeta_{j,k} + \bar{z}_k - \bar{z}_{k+1}.
\] (5.37)

However, because the centers of mass for the \(k\)-ad and \(k+1\)-ad are related via (5.35), we can use utilize this information to express the centered landmarks in the following manner:

\[
\zeta_{j,k+1} = \zeta_{j,k} + \frac{1}{k+1}(\bar{z}_k - z_{k+1}), j = \frac{1}{k+1}.
\] (5.38)

In order to find a similar formula for the \(k+1\)th centered landmark, the fact that the centered landmarks add to 0 is utilized.

\[
\zeta_{k+1,k+1} = -\sum_{j=1}^{k} \zeta_{j,k+1}.
\]

However, the coordinates for \(\zeta_{k+1,k+1}\) can also be found using (5.36). This relationship between the elements of \(L_k\) and \(L_{k+1}\) provides the following result relating the centered
Lemma 2 The $j$th centered landmark $\zeta_{j,k}$ of a $k$-ad converges to $\gamma_0 \left( \frac{j-1}{k} L \right)$ as $k \to \infty$.

PROOF 5.2 It suffices to show that $\|\zeta_{j,k+1} - \zeta_{j,k}\| \to 0$ as $k \to \infty$. From formula (5.38), it follows that the distance $\|\zeta_{j,k+1} - \zeta_{j,k}\|$ is smaller than $\frac{1}{k+1}$ times the diameter $D$ of the convex hull of the contour. From this,

$$\|\zeta_{j,k+1} - \zeta_{j,k}\| = \frac{1}{k+1} \|\tilde{z}_k - z_{k+1}\|$$

$$= \frac{1}{k+1} \left\| \frac{1}{k} \sum_{j=1}^{k} z_j - z_{k+1} \right\|$$

$$= \frac{1}{k+1} \left\| \frac{1}{k} \sum_{j=1}^{k} (z_j - z_{k+1}) \right\|$$

$$\leq \frac{1}{k+1} \left\| \frac{1}{k} \sum_{j=1}^{k} D \right\|$$

$$= \frac{\|D\|}{k+1} \xrightarrow{k \to \infty} 0.$$

It is apparent that the coordinates for each landmark from the centered $k$-ad change in increasingly small amounts as additional landmarks are added. It then follows that, as $k \to \infty$, the coordinates of a landmark form the centered $k$-ad converge to the coordinates of the corresponding point on the centered contour.

5.5.3 Approximations for Analysis

We consider the inclusion $\iota_k$ of $C^k$ in $C^{k+1}$ by adding a 0 for the last coordinate. Now $L_k \subset \mathbb{C}^k$ and $L_{k+1} \subset \mathbb{C}^{k+1}$, so the effect of adding $z_{k+1}$ to the $k$-ad, is given by a function $A_k : L_k \to L_{k+1}$ that leads to a function $\alpha_k : P(L_k) \to P(L_{k+1})$ in the form $\alpha_k = \phi_k \circ \iota_k$, where $\iota_k : L_k \to L_{k+1}$ is the restriction of the inclusion map $\iota_k$ and $\phi_k$ is an automorphism close to the identity of $P(L_{k+1})$. This can be shown in the following manner.

Using (5.37) and (5.38), the formula for $\alpha_k$ can be expressed as

$$\alpha_k(\zeta_k) = [(\zeta_k, 0) + \frac{1}{k+1} (\tilde{z}_k - z_{k+1}) (1_k, 0) + (0_k, z_{k+1} - \tilde{z}_{k+1})].$$

Letting $\tilde{\zeta}_{k+1} = (\zeta_k, z_{k+1} - \tilde{z}_{k+1})$, then $[\zeta_{k+1}] = \alpha_k([\zeta_k]) = \phi_k([\tilde{\zeta}_{k+1}]) = [A \tilde{\zeta}_{k+1}]$. Setting $a_{i,j}$ to be the entries of $A$ for $i, j = 1, \ldots, k+1$, it follows that $a_{i,j} = 0$ for $i \neq j$ and $j \neq k+1$.
and that
\[
a_{i,i} \zeta_{i,k} + a_{i,k+1}(z_{k+1} - \bar{z}_{k+1}) = \zeta_{i,k} + \frac{1}{k+1}(\bar{z}_k - z_{k+1}) \quad \text{for } i = 1, \ldots, k
\]
\[
a_{k+1,k+1}(z_{k+1} - \bar{z}_{k+1}) = (z_{k+1} - \bar{z}_{k+1}).
\]

Solving this system of equations, it is apparent that \( a_{i,i} = 1 \) for \( i = 1, \ldots, k+1 \) and that \( a_{i,k+1} = -\frac{1}{k+1}(\frac{z_{k+1} - z_k}{z_{k+1} - \bar{z}_{k+1}}) \) for \( i = 1, \ldots, k \). It follows that for large values of \( k \), \( A \approx I_{k+1} \) and that \( |\zeta_{k+1}| \approx \phi_k(|\zeta_{k+1}|) \).

In order to perform the desired analysis, we also must consider the finite approximation of the embedding. If we could work in the infinite-dimensional setting, we would use the Veronese-Whitney embedding \( j(\gamma) \) of \( P(H) \) into \( L_{HS} \), as defined in (5.16). Since we must discretize, though, we shall instead consider the finite Veronese-Whitney embedding \( j_k(\zeta) \) of \( P(L_k) \) into \( S(k, \mathbb{C}) \), as defined in (5.4). The relationships between the complex projective spaces of interest and the embeddings can be summarized in the following diagram:

\[
\begin{align*}
\mathbb{C}P^{k-2} & \quad \xrightarrow{P(\alpha)} \quad P(l_2) & \quad P(H) \\
\downarrow {j_k} & \quad \downarrow {j} & \\
S(k-1, \mathbb{C}) & \quad \xrightarrow{k \to \infty} \quad L_{HS}
\end{align*}
\]

5.5.4 Landmark Correspondence Within a Sample

While it is important that the \( k \)-ad representing a contour converges to the contour as \( k \) increases, it is also crucial that, within a sample of \( k \)-ads, every selected landmark from a given \( k \)-ad corresponds geometrically to a landmark in each of the \( k \)-ads in the sample. Even with random selection of the landmarks, proper correspondence can be obtained. This is aided substantially by the regularity constraint and the inclusion of \( z_0 \) as a landmark.

For a sample of \( n \) digital images of contours, the number of pixels along each contour is \( K_i \) for \( i = 1, \ldots, n \). Letting \( K_{\min} = \min(K_1, \ldots, K_n) \), randomly select \( k \) landmarks from the contour such that the ordered indices are \((1, j_2, \ldots, j_k)\), where \( 1 < j_2 < \cdots < j_k \leq K_{\min} \). The indices of the corresponding landmarks for contour \( i \) are \((1, \text{round}(K_i/K_{\min}) j_2, \ldots, \text{round}(K_i/K_{\min}) j_k)\) for \( i = 1, \ldots, n \).

Fig. 5.5 shows examples of landmark correspondence for two samples. Each subfigure shows the piecewise linear interpolations of four contours of a hand gesture. Six of the landmarks for each contour are highlighted as red points to show the correspondence within the sample.

The previously described methodology for choosing an appropriate lower bound for \( k \) should be used as a guide and that, while choosing larger values of \( k \) is desired, adjustments
may need to be made if \( K_i \ll K_j \) for images \( i \) and \( j \) in the data set. It should also be noted that the extraction of the contour, the selection of landmarks, parametrization and subsequent relabelling of landmarks and the previously described eigenanalysis can be automated, or semi-automated, allowing for efficient execution of the methodology.

5.6 Computations of Sample Means of Planar Contours

Based upon the described procedures for calculating Veronese-Whitney extrinsic mean shapes of finite approximations of contours, we now present some computational results using selected data from Ben Kimia’s contour database. A few samples of contours and their corresponding sample extrinsic mean shapes are given below. These examples illustrate key characteristics of the discrete approximations of the extrinsic mean shapes.

For standard numeric data, the sample mean averages out variability to provide an estimate of the analogous parameter for the location of a probability distribution. Similarly, the VW extrinsic sample mean shape estimates the location parameter for a probability distribution on the shape space. This results in the extrinsic sample mean serving as a representative form for the given data.

As shown in Figs. 5.6 through 5.11, the extrinsic sample mean for each group does not necessarily have the same shape as any of the individual observations. This is perhaps best shown with the contours of the dogs in Fig. 5.7. For this data, there is a large amount of variability in the contours, especially in the legs and tail of the dogs. In some of the observations, the tail is clearly visible, while in others the tail is mostly obscured. As such, the tail in the extrinsic mean shape is not very well pronounced. This result is also very
Figure 5.6: (a) Sample of 4 curves of 't' gestures and (b) the extrinsic mean shape of the sample of 't' gestures.

Figure 5.7: (a) Sample of 9 curves of "dogs" and (b) the extrinsic mean shape of the sample of dogs.

Figure 5.8: (a) Sample of 20 curves of "sting ray" fish and (b) the extrinsic mean shape of the sample of sting rays.
Figure 5.9: (a) Sample of 20 curves of worm fish and (b) the extrinsic mean shape of the sample of worm fish.

Figure 5.10: (a) Sample of 20 curves of red snapper fish and (b) the extrinsic mean shape of the sample of red snapper fish.

Figure 5.11: (a) Sample of 20 curves of pears and (b) the extrinsic mean shape of the sample of pears.
noticeable with the worm fish (Fig. 5.9) and the pears (Fig. 5.11).

For the contours of the red snapper fish (Fig. 5.10), the dorsal and tail fins are sharply pronounced features. In the extrinsic sample mean shape, these features are still present, but the variability causes them to be rounded off. Similar results are found for the worm fish contours (Fig. 5.9), as well.

Also of note is the extrinsic mean shape (Fig. 5.8(b)) of the sample of contours of sting rays (Fig. 5.8(a)). In this case, 15 of the 20 sting rays have straight tails. The other 5 have tails that are curved to varying degrees and in different directions. Despite these differences in the tails of the observed contours, the extrinsic mean shape has a straight tail, but is wider in order to account for the curved tails.

5.7 A One-Sample Test for Mean Shape

As above, let \([\gamma_1], \ldots, [\gamma_n]\) be i.i.d.r. objects on \(P(H)\) and let \(X_i = j([\gamma_i])\) for \(i = 1, \ldots, n\). Following a neighborhood method in the context of regression by Munk and Dette (1998)[24], from Munk et al. (2008)[27] we can develop similar tests for hypotheses for the extrinsic mean \(\mu_E\) in general. Suppose that \(j : P(H) \to \mathcal{L}_{HS}\) is the embedding in Section 5.5, \(M\) is a compact submanifold of dimension \(d\) of \(P(H)\) and that \(\delta > 0\) is a given positive number. We now introduce the functional on \(P(H)\),

\[
\varphi_M([\gamma]) = \rho^2([\gamma], M) = \inf_{m \in M} \rho^2([\gamma], m), \tag{5.40}
\]

where the distance \(\rho\) is inherited from the Euclidean distance on \(\mathcal{L}_{HS}\) via the embedding \(j\).

From here, we can test the following hypotheses:

\[
H_0 : \mu_E \in M_\delta \cup B_\delta \quad \text{for some } \delta > 0, \quad \text{vs.} \quad H_a : \mu_E \in M_\delta^c \cap B_\delta^c, \tag{5.41}
\]

where \(M_\delta = \{[\gamma] \in P(H) : \varphi_M([\gamma]) < \delta^2\}\) and \(B_\delta = \{[\gamma] : \varphi_M([\gamma]) = \delta^2, \tilde{\Sigma}_E |_{v_{[\gamma]}(M_\delta)} > 0\}\), where \(\tilde{\Sigma}_E\) is the extrinsic covariance matrix (see Bhattacharya and Patrangenaru (2005)[16]). As shown in Munk et al. (2008) [27], the test statistic for these types of hypotheses has a standard normal distribution for large sample sizes.

REMARK 5.5 For \(0 < \alpha < 1\), let \(\xi_{1-\alpha}\) denote the \((1 - \alpha)\)-th quantile of the standard normal distribution. The neighborhood hypothesis testing is used in this section, when the submanifold \(M\) consists of a point \(m_0\) on \(P(H)\). In this particular situation \(T_{m_0}M = 0\), therefore the test statistic for the hypotheses specified in (5.41) is given by:

\[
Z = \sqrt{n} (\varphi_M(\hat{\mu}_E) - \delta^2) / s_n, \tag{5.42}
\]
where

\[ s_n^2 = 4 \langle \nu, S_{E,n} \nu \rangle \]  

(5.43)

and

\[
S_{E,n} = \frac{1}{n} \sum_{i=1}^{n} \left( \tan_{\hat{\mu}} d_{j_{(X_i)}} \right) \left( P_j (j(X_i) - j(X)_n) \right) \otimes \left( \tan_{\hat{\mu}} d_{j_{(X_i)}} \right) \left( P_j (j(X_i) - j(X)_n) \right)
\]

(5.44)

is the extrinsic sample covariance matrix for \( \{X_i\}_{i=1}^{n} \), and

\[
\hat{\nu} = (d_{\hat{\mu}_{E,n}})^{-1} \left( \tan_{j_{(\hat{\mu}_{E,n})}} (j(m_0) - j(\hat{\mu}_{E,n})) \right).
\]

(5.45)

Assume \( X_r = [\gamma_r], \|\gamma_r\| = 1, r = 1, \ldots, n \) is a random sample from a \( j \)-nonfocal probability measure \( Q \) with a nondegenerate \( j \)-extrinsic covariance matrix on \( P(H) \). Then equation (5.30) has a complex version of the matrix perturbation results in Watson (1983,p.216), extension to the infinite dimensional case, of the complex formulation in Amaral et. al. (2010) of the studentized limit theorem for VW-sample means on \( \mathbb{C}P^{k-2} \) in Bhattacharya and Patrangenaru (2005). This shows that asymptotically the tangential component of the VW-sample mean around the VW-population mean has a complex multivariate normal distribution. Note that such a distribution has a Hermitian covariance matrix (see Goodman (1963)), therefore in this setting, the extrinsic covariance matrix and its sample counterpart are Hermitian matrices. In particular, if we extend the CLT for VW-extrinsic sample mean Kendall shapes in Bhattacharya and Patrangenaru (2005), to the infinite dimensional case, the \( j \)-extrinsic sample covariance matrix \( S_{E,n} \), when regarded as an infinite Hermitian complex matrix has the following entries

\[
S_{E,n,ab} = n^{-1} (\hat{\delta}_a^2 - \hat{\delta}^2) (\hat{\delta}_b^2 - \hat{\delta}^2)^{-1}
\]

(5.46)

\[
\sum_{r=1}^{n} <m_a, \gamma_r><m_b, \gamma_r>^* \mid <m_1, \gamma_r>^2, a,b = 1,2,\ldots
\]

with respect to the complex orthobasis \( m_2, m_3, m_4, \ldots \) of unit eigenvectors in Section 3. Recall that this orthobasis (over \( \mathbb{C} \) in \( T_{\hat{\mu}_{E,n}} P(H) \)) corresponds via the differential \( d_{\hat{\mu}_{E,n}} \) with an orthobasis (over \( \mathbb{C} \)) in the tangent space \( T_{j_{(\hat{\mu}_{E,n})}} (j(P(H))) \), therefore one can compute the components \( \hat{\nu}^a \) of \( \hat{\nu} \) from equation (5.45) with respect to \( m_2, m_3, m_4, \ldots \), and derive for \( s_n^2 \) in (5.43) the following expression

\[
s_n^2 = 4 \sum_{a,b=2}^{\infty} S_{E,n,ab} \hat{\nu}^a \hat{\nu}^b,
\]

(5.47)

46
where $S_{E,n,ab}$ given in equation (5.46) are regarded as entries of a Hermitian matrix.

### 5.7.1 Application of the One Sample Test for Mean Contour Shape

This test could be performed for a variety of applications. The most likely applications involve having a known extrinsic mean shape determined from historical data. In such cases, the hypothesis test can be used to determine whether there is a significant deviation from the historical mean shape. An application in agriculture would be determining whether the use of a new fertilizer treatment results in the extrinsic mean shape of a crop significantly changing from the historical mean. Similarly, this test could be performed for quality control purposes to determine if there is a significant defect in the outline of an produced good.

In practice, $\delta$ will be determined by the application and the decision for a test would be reached in the standard fashion. However, for the examples presented here, there is no natural choice for $\delta$, so one can instead consider setting $Z = \xi_{1-\alpha}$ and solving for $\delta$ to show what decision would be reached for any value of $\delta$. To do so, it is important to understand the role of $\delta$. The size of the neighborhood around $m_0$ is completely determined by $\delta$. As such, it follows that smaller values of $\delta$ result in smaller neighborhoods. In terms of $H_0$, this places a greater restriction on $M_\delta$ and $B_\delta$, requiring $\mu_E$ to have a smaller distance to $m_0$.

For the examples presented here, the contours are approximated using $k = 300$ landmarks, so the shape space is embedded into $S(300, \mathbb{C})$ to conduct analysis. In this environment, consider having two $k$-ads that are identical except for at one landmark. If this exceptional point for the second $k$-ad differs from the corresponding point in the first $k$-ad by a difference of 0.01 units, then the distance between the shapes inherited from $S(300, \mathbb{C})$ is approximately 0.0141. For the hypothesis test, if $\delta = 0.0141$, then the neighborhood around $m_0$ would consist of distances between shapes similar in scope to the situation described above.

First, consider an example for which the one sample test for extrinsic mean shape is performed for sting ray contours. In this case, the sample extrinsic mean shape for a sample of contours of $n = 10$ sting rays is the shape shown on the left hand side in Fig. 5.12. After performing the calculations, it was determined that for an asymptotic level 0.05 test, the largest value of $\delta$ for which we would reject the null hypothesis is 0.0290. For perspective, this neighborhood has a radius roughly 2 times larger than the example with the nearly identical $k$-ads described above. This means that we would only reject the null hypothesis if we required the sample extrinsic mean to be nearly identical to the hypothesized mean. It should also be noted here that the sample size is small here, but that the conclusion agrees with intuition based upon a visual inspection of the contours.

Now consider two examples involving contours of pears. In this first case, the sample
Figure 5.12: The extrinsic sample mean shape of a sample of 10 sting rays and, respectively, the hypothesized extrinsic mean shape consists of \( n = 87 \) pears. The sample extrinsic mean shape and hypothesized extrinsic mean shape are shown in Fig 5.13. It was determined that for an asymptotic level 0.05

Figure 5.13: The sample extrinsic mean shape of a sample of 87 pears and, respectively, the hypothesized extrinsic mean shape
test, the maximum value of $\delta$ for which we would reject the null hypothesis is 1.2941. This value of $\delta$ is almost 92 times greater than the distance between the nearly identical $k$-ads. This suggests that even if we greatly relax the constraints for similarity, the null hypothesis would still be rejected. This again agrees with intuition.

In this last example, consider another sample of contours of pears. In this scenario, we consider a sample of $n = 83$ pears. The sample extrinsic mean shape and hypothesized extrinsic mean shape are shown in Fig 5.14. After performing the calculations, we determined

![Figure 5.14: The sample extrinsic mean shape of a sample of 83 pears and, respectively, the hypothesized extrinsic mean shape](image)

that for an asymptotic level 0.05 test, the largest value of $\delta$ for which we would reject the null hypothesis is 0.1969, meaning that our procedure does not reject the null hypothesis, unless $\delta$ is smaller than 0.1969. For perspective, this neighborhood has a radius nearly 14 times larger than the example with the nearly identical $k$-ads described above. Unlike in the previous two examples it is unclear whether the null hypothesis should be rejected in this case without having a specific application in mind. As such, this could be considered a borderline case without additional information.

### 5.8 Bootstrap Confidence Regions for the Sample Mean

As with most types of data, it may not be enough to obtain an estimate of the mean shape. Indeed, it is desirable to understand the variability of the estimate. The method
employed here to explore this is the nonparametric bootstrap. This is accomplished by repeatedly resampling from the available data. By doing so, we can obtain a confidence region for the mean shape (see Bandulasiri, et al. (2008) [55]). Below are examples of 95% bootstrap confidence regions for the same sets of contours as provided previously. These regions are based upon 400 resamples from the data. These computations reveal that these regions behave as would be expected. For instance, the confidence region for the 't' hand gesture in Fig. 5.15(a) is wider in the portions of the shape where there is more variability in the sample (Fig. 5.6), such as the knuckle and finger areas, but is narrower in the portions where there is less variability in the sample, such as the wrist. This is also evident in the confidence region for the sting rays 5.16(a) since the bands are thicker in the regions corresponding to the tail of the fish, where, as shown in Fig. 5.8, the variability is the greatest.

It is also noticeable that the samples with less variability have narrower confidence regions, as well. Comparing the samples of dogs, wormfish, and pears (Figs. 5.7, 5.9, & 5.11, respectively), it is easy to see that the contours of the dogs have more variability than the contours of the wormfish, which have more variability than the contours of the pears. This is reflected in the widths of the confidence regions for these three groups of contours, as seen in Figs. 5.15(b), 5.16(b) and 5.17(b).

In addition to being able to obtain sensible and fairly intuitive results, this approach is advantageous because it allows us to easily produce a confidence region for the VW mean shape without making any assumptions about the underlying distribution of the shapes. Furthermore, the processing time needed to compute the bootstrap confidence regions or the VW mean is small compared to employing an intrinsic analysis, as the algorithms for
computing intrinsic means are iterative, and consequently result in a higher computational cost. The computational efficiency of our approach will be discussed in greater detail in Chapter 7.

Because these are approximate 95% confidence regions, the effective coverage may differ from the nominal rate. To explore the coverage of these regions for increasing sample sizes, we performed a brief simulation study using the pear contours. Treating the 88 available observations as the population of contours of pears, we calculated $\mu_E$. After obtaining a sample of size $n$ from the population, we computed $\hat{\mu}_E$. We then computed the 95% nonparametric nonpivotal bootstrap confidence region using 200 resamples from the original sample and checked to see if $\mu_E$ was included. This was repeated 400 times to obtain an
Table 5.1: Estimates for the effective coverage of nominal 95% bootstrap confidence regions for the extrinsic mean shape

<table>
<thead>
<tr>
<th>n</th>
<th>Effective Coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>91%</td>
</tr>
<tr>
<td>10</td>
<td>95%</td>
</tr>
<tr>
<td>15</td>
<td>98.25%</td>
</tr>
<tr>
<td>20</td>
<td>98.50%</td>
</tr>
<tr>
<td>25</td>
<td>98.75%</td>
</tr>
<tr>
<td>30</td>
<td>98.25%</td>
</tr>
</tbody>
</table>

estimate for the effective coverage for samples of size $n$. This process was repeated for $n = 5, 10, 15, 20, 25,$ and $30$. The results of this study are shown below in Table 5.1.

These results show that for sufficiently large sample sizes, the confidence regions are conservative, including $\mu_E$ more frequently than is expected. For such cases, hypothesis tests associated with these confidence regions would tend to reject the null hypothesis less often. This data suggests that the confidence regions are conservative for $n > 10$. Conversely, the confidence regions are too narrow for sufficiently low sample sizes, as in the case here for $n = 5$.

### 5.9 Discussion

In this paper, we have only described the estimation of the extrinsic mean shape of planar closed curves, but this approach could be extended further to any infinite configurations in the Euclidean plane, including 1-dimensional CW-complexes and planar domains, given that the plane is separable. For example, one may consider shapes of edge maps obtained from gray-level images, such as that in Fig. 5.18. In these cases, the problem of properly matching becomes much more difficult. Using edge maps as an example, this is because, not only do points on a given edge from one image need to be matched to corresponding points on the corresponding edge in another image, but each edge in an image must be matched to the corresponding edge in another image, as well.
Figure 5.18: (a) A grey-level image of a electronic circuit and (b) the edge-map of the circuit found using Canny edge detection. The original image is included as an example for image processing in MATLAB.
Proteins are molecules the consist of chains of amino acids. These amino acids are molecules that contain chemicals called an amine and a carbonylic acid, as well as additional chains of chemicals that vary across different amino acids. Commonly, up to 20 types of amino acids are included in the formation of proteins.

The study of proteins is of great interest because they are key components of biological functionality. This is because proteins bind to certain chemicals, allowing for a variety of functions to be performed within an organism. These functions can range from catalyzing biochemical reactions to maintaining structural or mechanical components of cells. The function of a protein is related both to its chemical composition and its structure.

Proteins fold into 3-dimensional structures based upon the sequence amino acids that comprise the chain and the chemicals contained within. Because of the complexity of proteins, the structure of a protein is often discussed as consisting of four aspects, as described in Murray et al (2006) [56]. The primary structure of a protein is the sequence of amino acids that comprise the protein. The secondary structure of a protein describes local structures that may repeat regularly within a protein. The tertiary structure describes the actual 3-dimensional structure of the protein. The quaternary structure describes the arrangement of multiple proteins or smaller units that together function to perform a function.

With advances in technology over the last 50 years, researchers have been able to determine and visualize protein structures using techniques such as X-ray crystallography [57] and nuclear magnetic resonance spectroscopy [58]. With technology now being more readily available, researchers have been able to study the structures of tens of thousands of proteins. With the advent of the internet, these structures are able to be stored centrally in databases such as the RCSB Protein Data Bank [59]. The amount of structural data available is steadily increasing. On April 9, 2010, the RCSB Protein Data Bank contained entries for 59,705 proteins, while as of March 15, 2011, it the database contained 71,794
6.1 Study of Relationship Between Protein Structure and Function

As the availability of protein structure data has increased, so, too, has interest in studying the relationship between the structure and function of proteins. The functions of individual proteins and genes are essential for understanding the functions of cells or organisms as a whole. Since proteins functions are determined by their structures, structural genomic projects have been initialized with the aim to solve representative proteins in each protein families [60, 61, 62]. The solved protein structures can be used to predict the structures of those homologous proteins, whose functions can then be deduced from their structures. At the same time, structural genomic projects have produced a large number of protein structures, whose functions are still unknown. As many as 26% of all SG structures deposited to the protein data bank (PDB) [59] are described as proteins of unknown functions, or quite often their functions are referred to as putative [63]. Predicting the functions of proteins based on their structural information has become one of the major roadblocks towards the goal of well-annotated genomes.

For a time, research was focused on exploring only the information provided by the sequence of amino acids, that is the primary structure, and examining the relationship with the function of the protein. Such approaches were used by Altschul et al (1997) [64], Wu et al (2000) [65], and Hulo et al (2004) [66]. However, with the availability of additional structural information, increasingly more focus has been placed on the study of secondary and tertiary structures. Efforts were originally made to study this relationship using the overall fold structure for entire proteins, such as in Holm and Sander (1990) [67].

Jones and Thornton (2004) [68] show that while fold structure can help, in certain cases, identify the function of a protein, the relationship between the two is complicated and instead suggest studying localized structures, including those near the surface of a protein, such as binding sites, which are where binding activity occurs. Since proteins function by interacting with other molecules through the binding sites (active sites), analysis of the binding site provides a direct mean to infer the function of a protein. A common hypothesis is that proteins with similar functions should have binding sites with similar shape and chemical properties. Many studies have been conducted based on the idea of comparing the putative binding site of a target protein with unknown function with the binding sites of proteins with known functions to infer the function of the target protein. These previous studies can be roughly divided into two classes: those use only structure information [69, 70, 71, 72, 73, 74, 75, 76, 77, 78, 79, 80, 81] and those use both structure...
and sequence/evolutionary information [82, 83, 84, 85]. Among those using only structure information to match binding sites, they can be further divided into two classes: those based on point clouds [69, 70, 71, 72, 73] and those based on shapes of binding sites [79, 80, 81, 86] or shape-based descriptors or features [74, 75, 87, 76, 77, 78].

Several algorithms have been developed for matching binding sites represented by point clouds. SPASM and RIGOR [88] scan a structural database for occurrences of structure motifs using a combinatorial search with constraints. Jess [89] matched structure templates based on a constrained logic programming method. Cavbase and eF-site used clique-based method [71, 90] to match structure templates formed by surface patches. SiteEngine [70, 91], SitesBase [92, 93], MultiBind [72] and TESS [69] used a geometric hashing algorithm [69, 70, 72, 73, 94, 95] to match protein surfaces and binding sites. In a study by Westkamp et. al. [73], clique detection and geometric hashing are combined. The above methods produce the correspondences between the atoms/residues of two binding sites, which can be used to calculate similarity using rigid superposition, such as root-mean-square-deviation (RMSD).

Instead of matching binding sites represented by point clouds, other methods have extracted information related to the shape of the binding sites, which is then used for binding site comparison. Hoffmann et al introduced a similarity measure called sup-CK and utilized global information from binding sites to align the sites based upon their principal axes [86]. This method calculates similarity using a Gaussian convolution kernel, which does not require correspondences between atoms of the two binding sites. Sael et. al. developed 3D Zernike descriptors to characterize and compare protein surfaces [74, 77]. Xiong et al used feature vectors based on distance of groups of atoms on binding sites [78].

While many of the approaches focus primarily on the structure of the binding site, sequence and evolutionary information have also been used in assisting structure-based function prediction. Binkowski et al first identify possible binding sites from the shape of protein surfaces and used sequence alignment of binding site residues to detect the similarity of protein binding pockets [79, 82, 96]. Lichtarge et al use the evolutionary trace (ET) method to identify important amino acids from multiple sequence alignments, which are then used to construct local structure templates to compare protein surfaces. [85, 97, 98].

When similar local structure motifs or templates are identified, assessment of the statistical significance of the similarity also plays an important role in function inference. To avoid the drawback of RMSD as a measure based on rigid superposition, a modified RMSD, oRMSD was used by Binkowski et al to measure the similarity of local surface structures [96]. Other similarity measures such as the Tanimoto index (TI) [99, 100] and the Poisson index (PI) [101] have also been adopted in protein binding site comparison.

In a recent study by Kahraman and co-workers, it has been found that pockets binding
the same ligand show greater variation in their shapes than can be accounted for by the conformational variability of the ligand [87]. They suggest that geometrical complementarity in general is not sufficient to drive molecular recognition. The data set created for this study has since served as a benchmark to which other methods, including sup-CK, have compared.

In more recent years, it has been proposed to study the relationship between structure and function using shape analysis techniques. Bandulasiri et al (2008) [55] proposed using landmark-based reflection size-and-shape analysis to study this relationship for ligand binding sites. However, other researchers have begun to explore using shape analysis to study the overall fold structures of proteins. Liu et al (2010) [102] proposed an alternative approach using the square-root elastic framework and Schmidler (2006) [103] suggested applying Bayesian shape matching to the problem. For the purposes of studying binding site structures, the approach considered here follows from Bandulasiri et al (2008a, 2008b) [104, 55].

### 6.2 Size-and-Shape Analysis for Configurations in 3-Dimensions

In this section, we are concerned with nonparametric statistical analysis of landmark based size-and-shape data in which each observation \( x = (x^1, \ldots, x^k) \) consists of \( k > 3 \) points in 3 dimensions. Except where noted, the theory and methodology follows largely from Bandulasiri et al (2009) [55], which discusses the more general case of data in which each landmark is in \( p \) dimensions. However, since the coordinates for the atoms within a protein are given in just 3 dimensions, we restrict to that specific case and use the term \( k \)-ad in reference to this case.

For size-and shape analysis, we consider only \( k \)-ads in which the \( k \) points are not equal. For the use of this methodology on protein data, this assumption is valid since only a single atom can exist at each location. Translation is removed by centering the \( k \)-ad \( x = (x^1, \ldots, x^k) \) to

\[
\xi = (\xi^1, \ldots, \xi^k) \\
\xi^j = x^j - \bar{x}, \forall j = 1, \ldots, k.
\]  

(6.1)

It should be noted that the set of all centered \( k \)-ads lie in a vector subspace \( L^3_k \) in \((R^3)^k\) of dimension \( 3k - 3 \), where \( L^3_k = \{\xi = (\xi^1, \ldots, \xi^k) \in (R^3)^k : \xi^1 + \cdots + \xi^k = 0\} \), and, \( L^3_k^* = L^3_k \setminus \{0\} \).
If size were not relevant, its effect could be removed by scaling \( \xi \) to unit size as

\[
\mathbf{u} = \frac{\xi}{|\xi|}
\]

(6.2)

where \( \mathbf{u} \) is called a preshape. The set \( S(L^3_k) \) of all such preshapes comprises a manifold of dimension \( 3k - 4 \), which is, in fact, the unit sphere in \( L^3_k \) and is identified with \( S^{3k-4} \), which is the unit sphere centered at the origin in \( \mathbb{R}^{3k-3} \). However, because size is important when studying the structure of a protein, we will not consider removing the effect of scaling in the analysis, as we did for the shape analysis of planar curves.

### 6.2.1 Size-and-Shape Spaces

The size-and-shape \([x]_S\) of the \( k \)-ad \( \mathbf{x} \) is defined to be the orbit, or equivalence class, of \( \xi = (\xi^1, \ldots, \xi^k) \) under all rotations in \( \mathbb{R}^3 \), as in Dryden and Mardia (1998, pp. 57) [10]). Similarly, the shape \([x]\) of the \( k \)-ad \( \mathbf{x} \) is defined to be the orbit, or equivalence class, of \( \mathbf{u} = (u^1, \ldots, u^k) \) under all rotations in \( \mathbb{R}^3 \). A more concise way of expressing these notions is as follows:

\[
[x]_S = \{ A\xi = (A\xi^1, \ldots, A\xi^k) : A \in SO(p) \},
\]

\[
[x] = \{ Au = (Au^1, \ldots, Au^k) : A \in SO(p) \},
\]

(6.3)

where \( SO(3) \) is the special orthogonal group of all \( 3 \times 3 \) matrices \( A \) such that \( A^T A = I_3, \text{Det}(A) = 1 \). It follows then that the 3-dimensional size-and-shape-space is \( \Sigma^k_3 = L^3_k / SO(3) \sim (\mathbb{R}^{3k-3} \setminus \{0\}) / SO(3) \) and the Kendall shape space, or the similarity shape space, is \( \Sigma^k_3 = S(L^3_k) / SO(3) \sim S^{3k-4} / SO(3) \).

However, unlike planar direct similarity shape spaces, for higher dimensions, including \( p = 3 \), the spaces defined above have singularities. This is due to the fact that there exist \( k \)-ads \( \mathbf{x} \) for which there are different \( A, B \in SO(3) \) such that \( A\xi = B\xi \), so that the orbits of \( \xi \) are not, in general, one-to-one. For dimension \( p = 3 \), such singularities occur if all \( k \) landmarks in a \( k \)-ad are collinear. This results in \( \xi \) being left invariant to the non-trivial subgroup of rotations around this line. The existence of such singularities has also been noted and recognized as a problem by Small (1996) [13] and Kendall et al (1993) [11].

These singularities would impede the analysis if we were to stick to Kendall’s shape spaces. However, analysis can still be performed if other shape spaces are considered instead. The reflection size-and-shape \([x]_RS\) of a \( k \)-ad \( \mathbf{x} \) and the reflection shape \([x]_R\) of a \( k \)-ad, are, respectively, the \( O(3) \)-orbit of the centered configuration \( \xi \), and of the preshape \( \mathbf{u} \) under the action \( A\xi = (A\xi^1, \ldots, A\xi^k) \) of the orthogonal group \( O(3) \) on the set of all centered \( k \)-ads, as in Dryden and Mardia (1998, p. 57) [10], where \( O(3) \) is the set of \( 3 \times 3 \) matrices
A satisfying $A^T A = I_3$. Thus

$$[x]_{RS} = \{ A \xi : A \in O(p) \}, \quad [x]_R = \{ A u : A \in O(p) \}. \quad (6.4)$$

A $k$-ad is said to be in general position if $\{ \xi_1, \ldots, \xi_k \}$ spans $\mathbb{R}^3$. It then follows that the set of all reflection size-and-shapes of $k$-ads in general position $\xi$ is the reflection size-and-shape space $S \Sigma^k_{3,0}$. Similarly, the set of all reflection shapes of $k$-ads in general position $\xi$ is the reflection shape space $\Sigma^k_{3,0}$. Both these spaces are manifolds because the action of an orthogonal matrix on $\mathbb{R}^3$ is uniquely determined by its action on a basis of $\mathbb{R}^3$, and a centered $k$-ad in general position includes such a basis. The manifold approach to reflection-shape analysis that follows brought forth by Bandulasiri and Patrangenaru (2005) [39].

The size-and-reflection-shape space $S \Sigma^k_{3,0}$ is the set

$$S \Sigma^k_{3,0} = \{ [x]_{RS}, x \text{ in general position } \} = \{ [x]_{RS}, \text{rk} x = 3 \}, \quad (6.5)$$

where the reflection size-and-shape $[x]_{RS}$ of a $k$-ad $x$, which is regarded here as a $3 \times k$ matrix, in $\mathbb{R}^3$, is as defined in (6.4). If we set $L_{k,3,0} = \{ \xi \in I^3_k, \text{rk} \xi = 3 \}$, the manifold $S \Sigma^k_{3,0}$ can be represented as a quotient $L_{k,3,0}/O(3)$. The manifold dimension of $\Sigma^k_{3,0}$ is $3k - \frac{3(3+1)}{2} = 3k - 6$.

Similarly, the reflection shape space $\Sigma^k_{3,0}$, then, is the set

$$\Sigma^k_{3,0} = \{ [x]_R, x \text{ in general position } \} = \{ [x]_R, \text{rk} x = 3 \}, \quad (6.6)$$

where $\text{rk} x$ is the rank of $x$ and the reflection shape $[x]_R$ of a $k$-ad $x$ is as defined by (6.4). The manifold dimension, codimension of the $O(3)$-orbits in $(\mathbb{R}^3)^k$, is $\dim \Sigma^k_{3,0} = 3k - \frac{3(3+1)}{2} - 1 = 3k - 7$.

### 6.2.2 The Schoenberg Embedding

In order to perform an extrinsic analysis on $S \Sigma^k_{3,0}$ and $\Sigma^k_{3,0}$, Bandulasiri and Patrangenaru (2005) [39] introduced the Schoenberg embedding of similarity shapes in higher dimensions. This approach is based upon a result in multidimensional scaling (MDS), as described by Schoenberg (1935) [105], Gower (1966) [106] and Mardia et al (1979) [107].

To motivate this approach, let $D = (d_{rs})_{r,s=1,...,k}$ be a distance matrix between $k$ points and consider $A$ and $B$ defined by $A = (a_{rs})$, $a_{rs} = -\frac{1}{2} d_{rs}^2$, $\forall r, s = 1, \ldots, k$, $B = \tilde{H} A \tilde{H}^T$, where $\tilde{H} = I_k - k^{-1} 1_k 1_k^T$ is the centering matrix in $\mathbb{R}^k$. To further motivate the approach, we consider a version of Schoenberg’s theorem as presented by Mardia et al (1979, p. 397) [107].
Theorem 6.1  

(a) If $D$ is a matrix of Euclidean distances $d_{rs} = \|x^r - x^s\|$, $x^j \in \mathbb{R}^p$, and we set $X = (x^1 \ldots x^k)^T$, $k > p$, then $B$ is given by

$$b_{rs} = (x^r - \overline{x})^T (x^s - \overline{x}), \quad r, s = 1, \ldots, k$$

(6.7)

In matrix form, $B = (\tilde{H}X)(\tilde{H}X)^T$. Note that $B$ can be interpreted as the "centered inner product matrix" for the configuration $X$.

(b) Conversely, assume $B$ is a $k \times k$ symmetric positive semidefinite matrix $B$ of rank $p$, having zero row sums. Then $B$ is a centered inner product matrix for a configuration $X$ constructed as follows. Let $\lambda_1 \geq \cdots \geq \lambda_p$ denote the positive eigenvalues of $B$ with corresponding eigenvectors $x^{(1)}, \ldots, x^{(p)}$ normalized by

$$x^T_{(i)} x^{(i)} = \lambda_i.$$  

(6.8)

Then the points $P_r \in \mathbb{R}^p, r = 1, \ldots, k$, with coordinates $x^r = (x_{r1}, \ldots, x_{rp})^T$ with $x^r$ as the $r^{th}$ row of the $k \times p$ matrix $(x^{(1)} \ldots x^{(p)})$, have center $\overline{x} = 0$ and the inter-point distances given by $D$.

Again, because the analysis of protein structure incorporates scale, we restrict our attention to the analyses for the reflection-size-and-shape spaces, although the approach is very similar for the reflection-shape spaces. The Schonberg embedding of the reflection size-and-shape manifold is $J : S^{k} S_{3,0} \rightarrow S(k, \mathbb{R})$, given by

$$J([\xi]_{RS}) = \xi^T \xi.$$  

(6.9)

$J$ is an embedding because both it and its derivative are injective, as shown by the authors in [55]. The formulas for the extrinsic parameters and for their respective estimators that will be subsequently shown are derived using the following result.

Theorem 6.2  

The range of the Schoenberg embedding of $S^{k}_{3,0}$ is the subset $SM_{k,3}$ of $k \times k$ positive semidefinite symmetric matrices $A$ with $rkA = 3$, $A1_k = 0$.

It should be noted that if $H$ is the $(k - 1) \times k$ Helmert sub-matrix, as in Dryden and Mardia (1998, p. 34) [10], then we can get an equivalent isometric embedding of $S^{k}_{3,0}$ into the space $S(k - 1, \mathbb{R})$ that is given by

$$\psi([\xi]_{RS}) = H\xi^T \xi H^T,$$  

(6.10)

where the range of $\psi$ is the set $\tilde{N}_{3,k}$ of $(k - 1) \times (k - 1)$ symmetric matrices of rank 3.
### 6.2.3 Extrinsic Means for Reflection Shape Spaces

As defined in Bhattacharya and Patrangenaru (2003) [1], the extrinsic mean $\mu_{J,E}(Q)$ of a nonfocal probability measure $Q$ on a manifold $M$ with respect to an embedding $J : M \rightarrow \mathbb{R}^N$, when it exists, is given by $\mu_{J,E}(Q) = J^{-1}(P_J(\mu))$, where $\mu$ is the usual mean of $J(Q)$ as a probability measure on $\mathbb{R}^N$ and $P_J$ is its projection on $J(M)$. When the embedding $J$ is given, and the projection $P_J(\mu)$ is unique, we often identify $\mu_{J,E}$ with its image $P_J(\mu)$, and refer to the latter as the extrinsic mean. The term "nonfocal $Q$," as used above, means that the projection, or the minimizer of the distance from $\mu$ to points in $J(M)$, is unique.

We will denote the extrinsic mean by $\mu_E$. We define the corresponding sample mean in a similar fashion. Assuming that $(X_1, \ldots, X_n)$ are i.i.d. $M$-valued random objects whose common probability measure is $Q$, we let $\hat{X}_E := \mu_E(\hat{Q}_n)$ be the extrinsic sample mean. Here, $\hat{Q}_n = \frac{1}{n}\sum_{j=1}^{n} \delta_{X_j}$ is the empirical distribution.

We consider now the calculation of the extrinsic mean reflection size-and-shape using the Schoenberg embedding. Let $X$ be a random $k$-ad in general position which is centered as $X_0 = (X^1 - \overline{X}, \ldots, X^k - \overline{X}) \in (\mathbb{R}^3)^k \simeq M(3,k;\mathbb{R})$. Set

$$C = E(X_0^T X_0) \quad (6.11)$$

It follows that $C_{1k} = 0, C \succeq 0$.

The extrinsic mean reflection size-and-shape of $[X]_{RS}$ exists if $Tr(C - \xi^T \xi)^2$ has a unique solution $\xi \in M(3,k;\mathbb{R})$ up to an orthogonal transformation, with

$$\xi_{1k} = 0, \text{rk}\xi = 3. \quad (6.12)$$

That is the same as saying that, given $C$, $\xi$ is a classical solution in $\mathbb{R}^3$ to the MDS problem, as given in Mardia et al (1979)[107] in terms of the 3 largest eigenvalues of $C$. Bandulasiri et al (2008) [55] show that the extrinsic mean reflection size-and-shape is found in the following manner:

**Theorem 6.3** Assume $C = \sum_{i=1}^{k} \lambda_i e_i e_i^T$ is the spectral decomposition of $C = E(X_0^T X_0)$, then the extrinsic mean $\mu_E$ reflection size-and-shape exists if and only if $\lambda_p > \lambda_{p+1}$ and if this is the case, $\mu_E = [\xi]_{RS}$ where $\xi^T$ can be taken as the matrix

$$V = (v_1v_2v_3). \quad (6.13)$$

satisfying

$$v_j^T v_j = \lambda_j, \text{ for } j = 1, 2, 3. \quad (6.14)$$

Let $\{x_1, \ldots, x_n\}$, where $x_j = (x_{j1}^1, \ldots, x_{jk}^k), j = 1, \ldots, n$, be a sample of $k$-ads in general
position in $\mathbb{R}^3$. It follows from theorem 6.3 that the extrinsic sample mean reflection size- and shape is $[\mathbf{x}]_E = [\hat{\xi}]_{RS}$, where $\hat{\xi}$ is the classical solution in $\mathbb{R}^3$ to the MDS problem for the matrix

$$\hat{C} = \frac{1}{n} \sum_{j=1}^{n} \xi_j^T \xi_j$$

(6.15)

assuming that $\hat{\lambda}_3 > \hat{\lambda}_4$. Here $\xi_j$ is the matrix obtained from $\mathbf{x}_j$ after centering and $\hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_k$ are the eigenvalues of $\hat{C}$. As a result, the configuration of the extrinsic sample mean reflection size-and-shape is given by the eigenvectors corresponding to the 3 largest eigenvectors of $\hat{C}$, subject to the constraints of (6.14).

The extrinsic mean reflection shape is found in a similar manner. Let $\mathbf{U}$ be defined such that $\mathbf{U} = \mathbf{X}_0/\|\mathbf{X}_0\|$. If $B = E(\mathbf{U}^T \mathbf{U})$, then the mean is calculated in the following manner, as shown by A. Bhattacharya (2008)[108], adapted to 3-dimensional objects:

**Theorem 6.4** Assume $B = \sum_{i=1}^{k} \lambda_i e_i e_i^T$ is the spectral decomposition of $B$, then the extrinsic mean reflection shape exists if and only if $\lambda_3 > \lambda_4$. If this is this case, then $\mathbf{u}^T$ can be taken as the matrix

$$V = (v_1 v_2 v_3),$$

(6.16)

whose columns are orthogonal eigenvectors of $B$ corresponding to the largest eigenvalues $\lambda_1 \geq \cdots \geq \lambda_p$ of $B$, with

$$(a) \ v_j^T 1_k = 0$$

and

$$(b) \ v_j^T v_j = \lambda_j + \frac{1}{3}(\lambda_4 + \cdots + \lambda_k), \ \text{for} \ j = 1, 2, 3.$$  

(6.17)

Independently, Dryden et al (2008) [109] used an alternative approach involving MDS for objects in dimension $p$. In place of using the Schoenberg embedding, the authors defined a one-to-one function from $\Sigma_{p+1}^k$, the space of reflection shapes of $k$-ads of points that are not all identical, to $S(k - 1, \mathbb{R})$. The range of $S(k - 1, \mathbb{R})$ is the set $N_{p,k}$ of all positive definite matrices $B$ with $rk(B) \leq p, Tr(B) = 1$. The MDS mean reflection shape found using this approach is found similarly to the extrinsic sample mean reflection shape of A. Bhattacharya (2008) [108]. However, the scaling of the eigenvectors $v_1, \ldots, v_p$ corresponding to the $p$ largest eigenvalues of $B$ is given by

$$v_j^T v_j = \frac{\lambda_j}{\lambda_1 + \cdots + \lambda_p}, \ \forall j = 1, \ldots, p,$$

(6.18)

as opposed to the scaling given by (6.17).
6.2.4 Asymptotics for the Extrinsic Sample Mean Reflection Size-and-Shape

We now consider the asymptotic distributions for the extrinsic sample mean reflection size-and-shape, which allows us to conduct inference about the extrinsic mean reflection size-and-shape. Bandulasiri et al (2008) [55] used the general results for extrinsic means on manifolds from Bhattacharya and Patrangenaru (2005) [16] to derive the asymptotics when working in $S^k_{p,0}$. Here, we present those derivations as applied to our specific case of $S^k_{3,0}$.

Rather than using the Schoenberg embedding $J$, we consider the alternative embedding $\psi$ as defined in (6.10) from $S^k_{3,0}$ into the space $S(k-1, \mathbb{R})$ with a range of $\tilde{N}_{3,k}$. The tangent space of $\tilde{N}_{3,k}$ at $\psi([\xi]_{RS})$ is the range of the differential of $\psi$ at $[\xi]_{RS}$. Letting $\eta = \xi H^T$, then $\text{rank}(\eta) = 3$ and $\psi([\xi]_{RS}) = \eta^T \eta$. It also follows that $T_{\psi([\xi]_{RS})} N_{p,k} = T_{\eta^T \eta} \tilde{N}_{p,k} = \{ v \in S(k-1, \mathbb{R}) : v = y^T \eta + \eta^T y, y \in M(k-1, p, \mathbb{R}) \}$. (6.19)

Let $\Lambda$ be the diagonal matrix with diagonal elements $\lambda_1, \lambda_2, \lambda_3, 0, \ldots, 0$. Due to the equivariance of $\psi$, if $\eta^T \eta = A \Lambda A^T$, where $A \in O(k-1)$ then $T_{\eta^T \eta} \tilde{N}_{p,k} = A T_{\Lambda} \tilde{N}_{3,k} A^T$. (6.20)

From this, it follows that the tangent space at $\Lambda$ can be expressed as

$$T_{\Lambda} \tilde{N}_{3,k} = \{ v \in S(k-1 : \mathbb{R}) : v = \begin{pmatrix} V_3 & W \\ W^T & 0 \end{pmatrix}, V_3 \in S(3, \mathbb{R}), W \in M(k-4, 3; \mathbb{R}) \}. \quad (6.21)$$

A standard orthonormal basis for $S(k-1, \mathbb{R})$ that is described in Bhattacharya and Patrangenaru (2003) [1] is:

$$\tilde{E} = (E_1^1, \ldots, E_1^{k-1}, 2^{-\frac{1}{2}}(E_i^j + E_j^i), 1 \leq i < j \leq k-1), \quad (6.22)$$

where $E_i^j$ has all entries zero, except for the entry in the $i^{th}$ row and $j^{th}$ column, which equals 1. It follows from (6.21) that $T_{\Lambda} \tilde{N}_{3,k}$ is spanned by the orthobasis

$$e(\Lambda) = (E_1^1, \ldots, E_3^3, 2^{-\frac{1}{2}}(E_i^j + E_j^i), 1 \leq i < j \leq 3 \text{ or } 1 \leq i \leq 3 \leq j \leq k-1). \quad (6.23)$$

The authors show that, for any $A \in O(k-1)$, the map $v \rightarrow A v A^T$ is an isometry of
$S(k-1, \mathbb{R})$, an orthobasis for the space $T_{\eta r \eta} \tilde{N}_{3,k}$ is given by

$$e(\eta^T \eta) = (E_1^1(\eta), \ldots, E_3^3(\eta), 2^{-\frac{1}{2}}(E_i^1(\eta) + E_j^1(\eta)), 1 \leq i < j \leq 3 \text{ or } 1 \leq i \leq 3 < j \leq k-1,$$

where $E_i^j(\eta) = \lambda E_i^j A^T$. However, since the asymptotic results will be expressed using vector notation, a more convenient ordering for the orthobasis is

$$E = (E_1^1, \ldots, E_p^p, 2^{-\frac{1}{2}}(E_i^j + E_j^i), 1 \leq i < j \leq p \text{ or } 1 \leq i \leq p < j \leq k-1, E_{p+1}^p, \ldots, E_{k-1}^{k-1}, 2^{-\frac{1}{2}}(E_i^j + E_j^i), p + 1 \leq i < j \leq k-1).$$

As before, the extrinsic mean $\mu_E$ of a Schoenberg nonfocal probability measure $Q$ on $S\Sigma_{3,0}^k$ is given by $\mu_E = \psi^{-1}(P_\psi(\mu))$, where $\mu$ is the mean of $\psi(Q)$ in $S(k-1, \mathbb{R})$ and $P_\psi$ is the projection on $\tilde{N}_{p,k}$. As defined in Bhattacharya and Patrangenaru (2005) [16], the extrinsic covariance operator $\Sigma_E$ is the restriction of the self-adjoint linear operator $d_\mu P_\psi \Sigma d_\mu P_\psi^T$ to $T_{P_\psi(\mu)} \tilde{N}_{p,k}$. The extrinsic covariance matrix is the matrix associated to $\Sigma_E$ with respect to an orthobasis $e_1(P_\psi(\mu)), \ldots, e_d(P_\psi(\mu))$ of $T_{P_\psi(\mu)} \tilde{N}_{p,k}$, where $d = \frac{3}{2}(2k-4)$.

The following is a specific case of a result shown by Bandulasiri et al (2008) [55]:

**Proposition 6.1** If the spectral decomposition of the mean $\mu$ of $\psi(Q)$ is $\mu = \sum_{i=1}^{k-1} \lambda_i \tilde{e}_i \tilde{e}_i^T,$ with $\lambda_1 \geq \cdots \geq \lambda_p > \lambda_{p+1} \cdots \geq \lambda_{k-1}$, then

(i) the tangent space $T_{\psi(\mu)} N_{p,k} = T_1 \oplus T_2$, where $T_1$ has the orthobasis

$$(\tilde{e}_1 \tilde{e}_1^T, \ldots, \tilde{e}_p \tilde{e}_3^T, 2^{-\frac{1}{2}}(\tilde{e}_i \tilde{e}_j^T + \tilde{e}_j \tilde{e}_i^T), 1 \leq i < j \leq 3),$$

and $T_2$ has the orthobasis

$$(2^{-\frac{1}{2}}(\tilde{e}_j \tilde{e}_l^T + \tilde{e}_l \tilde{e}_j^T), 1 \leq j \leq 3 < l \leq k-1).$$

(ii) Let $N$ be the orthocomplement of $T_{\psi(\mu)} N_{3,k}$ Then if

$$d_\mu P_\psi | T_1 = Id_{T_1},$$

$$d_\mu P_\psi (\tilde{e}_j \tilde{e}_l^T + \tilde{e}_l \tilde{e}_j^T) = \frac{\lambda_j}{\lambda_j - \lambda_l}(\tilde{e}_j \tilde{e}_l^T + \tilde{e}_l \tilde{e}_j^T), \forall (j,l), 1 \leq j \leq 3 < l \leq k-1,$$

$$d_\mu P_\psi | N = 0.$$ (6.28)

An orthobasis of $N$ in Proposition 6.1 is

$$2^{-\frac{1}{2}}(\tilde{e}_j \tilde{e}_l^T + \tilde{e}_l \tilde{e}_j^T), 3 < j < l \leq k-1).$$

The three orthobases (6.26), (6.27) and (6.29) yield an orthobasis $\tilde{e}$ of $S(k-1, \mathbb{R})$. From
proposition 6.1 it follows that the matrix $D$ associated with the differential $d_{\mu}P_\psi$ relative to the orthobasis $\tilde{e}$ is diagonal:

$$D = \begin{pmatrix} I_3 & 0 & 0 \\ 0 & \Delta_{3(k-4)} & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

(6.30)

where

$$\Delta_{3(k-4)} = \begin{pmatrix} \frac{\lambda_1}{\lambda_1-\lambda_4} & \ldots & 0 \\ \ldots & \ldots & \ldots \\ 0 & \ldots & \frac{\lambda_3}{\lambda_3-\lambda_{k-1}} \end{pmatrix}. $$

(6.31)

The space of symmetric matrices $S(k-1, \mathbb{R})$ can be regarded as its own tangent space at $\mu$ and splits in three orthogonal subspaces

$$S(k-1, \mathbb{R}) = T_1 \oplus T_2 \oplus N.$$  

(6.32)

This leads to a decomposition of the covariance matrix $\Sigma$ of $\psi(Q)$, with respect to the orthobasis of $S(k-1, \mathbb{R})$. It is obtained by augmenting the orthobasis (6.23) by an orthonormal basis of $N$, as follows:

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \Sigma_{13} \\ \Sigma_{12}^T & \Sigma_{22} & \Sigma_{23} \\ \Sigma_{13}^T & \Sigma_{23}^T & \Sigma_{33} \end{pmatrix}. $$

(6.33)

If the orthonormal basis $\tilde{e}$, the eigenvectors $\tilde{e}_1, \ldots, \tilde{e}_{k-1}$ of $\mu$, then the coordinates of $\mathbb{R}^{k-1}$ can be changed such that the mean is a diagonal matrix $\Lambda$ and the matrix $\Sigma_\mu = D\Sigma D^T$ is

$$\Sigma_\mu = \begin{pmatrix} \Sigma_{11} & \Sigma_{12}\Delta & 0 \\ \Delta\Sigma_{12}^T & \Delta\Sigma_{22}\Delta & 0 \\ 0 & 0 & 0 \end{pmatrix}. $$

(6.34)

It follows that the extrinsic covariance matrix $\Sigma_E$ defined in Bhattacharya and Patrangenaru (2005), with respect to the basis $d_{\mu_\psi}^{-1}(e(\Lambda))$, where $e(\Lambda)$ is as defined in (6.23), is

$$\Sigma_E = \begin{pmatrix} \Sigma_{11} & \Sigma_{12}\Delta \\ \Delta\Sigma_{12}^T & \Delta\Sigma_{22}\Delta \end{pmatrix}. $$

(6.35)

Let $Y_1, \ldots, Y_n$ be independent identically distributed random reflection objects from a $\psi$-nonfocal probability distribution $Q$ on $S^{\Sigma^E_3,0}$, with $\lambda_3 > \lambda_4$. Furthermore, let $[\mu]_{RS}$ be the mean reflection size-and-shape of $\psi(Q)$ and $\Sigma$ the covariance matrix of $\psi(Q)$ with respect to the orthobasis $\tilde{e}$ defined above. Let $\tilde{W}$ be the vectorized form of a matrix $W \in S(k-1, \mathbb{R})$...
with respect to the basis \( \tilde{V} \). Assume \( \tan \tilde{W} \) denote the component of \( \tilde{W} \) tangent to \( \tilde{N}_{3,k} \) at \( \tilde{\psi}(\mu_\psi) \).

**Theorem 6.5** (a) The random vector \( n \tan(\tilde{\psi}(\tilde{Y}_E) - \tilde{\psi}(\mu_E)) \) converges weakly to a random vector having a \( N(0, \Sigma_E) \) distribution, where \( \Sigma_E \) is given in (6.35).

(b) If \( \Sigma_E \) is nonsingular, then \( n \tan(\tilde{\psi}(\tilde{Y}_E) - \tilde{\psi}(\mu_E))^{T} \Sigma_{E}^{-1} \tan(\tilde{\psi}(\tilde{Y}_E) - \tilde{\psi}(\mu_E))^{T} \) converges weakly to a \( \chi^2_{3k-6} \) distribution.

It follows that the extrinsic mean reflection size-and-shape can be easily estimated using non-pivotal bootstrap. Let \( \{x_1, \ldots, x_n\} \) be a random sample of configurations \( x_j = (x_{1j}, \ldots, x_{kj}), j = 1, \ldots, n \). If we resample at random with repetition \( N \) times from this sample, where \( N \) is a reasonably large number, say \( N \geq 500 \), then for each such resample \( x_1^*, \ldots, x_n^* \), we can compute the extrinsic sample mean \( \overline{x}_{RS_E}^* \). We then use a local parametrization of \( S_{\Sigma_{3,0}}^k \) and find \( (1 - \alpha)100\% \) Bonferoni simultaneous confidence intervals for the corresponding \( 3k - 6 \) local coordinates.

### 6.3 Landmark Selection and Registration

The selection of landmarks is an important aspect of shape analysis. For some data, such as the planar contours presented in 5, it is not obvious which points should be selected as landmarks. However, once they are chosen, it is straightforward to determine the proper ordering of the landmarks so that they accurately represent the contour. If care was taken during the selection process, obtaining a notion of proper geometric correspondence should follow.

However, this is not the case for the 3-D structures of protein binding sites. Because the sites are comprised of a set of atoms, it is natural to use these atoms as landmarks. Unlike the planar contour data, though, there is no natural ordering of these landmarks. Furthermore, while each site is represented by an obvious set of landmarks, in order to use size-and-shape information to analyze similarity, it is necessary that the observations in a data set consist of the same number \( k \) of landmarks ordered such that correspondence within the sample is preserved.

The problems of landmark selection and registration are ultimately solved by finding the maximum common atom set (MCAS), which is defined to be the maximum number of superposable atoms between two binding sites where the distance between any pair of matched atoms in the rigid superposition of the binding sites is smaller than a given threshold value, and the matching of atoms in this set so that correspondence is obtained. To do so, we developed a method for comparing ligand binding sites of proteins using atom-level representation based upon the Iterative Closest Point (ICP) algorithm (Besl and
McKay, 1992) [110]. Compared to the original ICP algorithm, our algorithm starts from a multitude of initial local alignments derived from 3D Delaunay triangulations and the iterative procedure uses the Hungarian algorithm to find additional matched atoms.

6.3.1 Algorithm for Obtaining Corresponding Atoms

Our approach is to treat the atoms of ligand binding sites as point clouds with corresponding labels specifying the chemical properties of the atoms. We compare the binding sites represented by the point clouds to find the MCAS between pairs of binding sites. We refer to a subset of atoms that can be matched between two binding sites as a common atom set (CAS), and the largest of such sets as the maximum common atom set (MCAS). Many past studies based on point clouds have used similar criteria. Although we are not aware of any literature on the time complexity of this problem, there is probably no polynomial time algorithm for its solution. One has to use heuristic methods to find sub-optimal solutions, as in previous methods reviewed in the Introduction. In this study, we aim to answer the following questions: (1) Can our algorithm find solutions comparable to or better than other existing methods in terms of finding MCAS? (2) How can the CAS found by our method be used to predict the binding ligands of proteins and what kind of accuracy can be achieved?

A standard technique used for alignment and registration for point clouds is known as the Iterative Closest Point (ICP) algorithm, which was introduced by Chen and Medioni [111] and Besl and McKay [110]. The ICP algorithm aligns and registers an unlabeled set of points \( p \) to a model set \( X \) by iteratively alternating between registration and alignment steps. Registration is obtained by finding the closest point \( y \in X \) to each point \( p_i \in p \), resulting in the corresponding set \( Y \). An alignment is then obtained by finding the optimal rotation matrix \( R \) and translation vector \( v \) such that \( p \) is superposed onto \( Y \). These two steps are repeated until the change in mean square error between \( p \) and \( Y \) falls beneath a desired threshold.

However, ICP cannot be directly applied for matching ligand binding sites for two reasons. Firstly, since the algorithm is deterministic in nature, the results depend greatly on the initial alignment used and the algorithm may land on a solution, which is far from optimal. Besl and McKay suggest solving this problem by considering a large number of initial rotation states while superposing the centers of mass of two objects. However, superposing the centers of two binding sites, in many cases, may not provide good initial matching. Secondly, ICP does not guarantee unique correspondence between atoms, as registration is performed one point at a time, not jointly. Additionally, this approach to registration does not take labels on the points into consideration. For the purposes of this application, unique correspondence and making use of the chemical labels is necessary. We propose a different approach to solve the initial alignment problem. For the registration problem, we
Table 6.1: Atom types used for obtaining correspondence

<table>
<thead>
<tr>
<th>Label</th>
<th>Atom Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>carbonyl C</td>
</tr>
<tr>
<td>2</td>
<td>aliphatic C CA, Other sp3 C, Disulfide bond S, Met S, Cys S</td>
</tr>
<tr>
<td>3</td>
<td>aromatic C</td>
</tr>
<tr>
<td>4</td>
<td>O, acceptor Backbone and carbonyl O in Asn and Gln, carboxyl O in Asp and Glu</td>
</tr>
<tr>
<td>5</td>
<td>O, donor and acceptor hydroxyl O in Ser, Thr and Tyr</td>
</tr>
<tr>
<td>6</td>
<td>N, donor backbone N except proline N, TRP side chain NE1, GLN NE2, ASN ND2, ARG NE NE1 NE2, LYS NZ</td>
</tr>
<tr>
<td>7</td>
<td>N, donor and acceptor HIS side chain NE1 NE2</td>
</tr>
<tr>
<td>8</td>
<td>H, polar H and backbone H</td>
</tr>
</tbody>
</table>

use the Hungarian algorithm to find the optimal correspondences between atoms in two binding sites given the rotational and translation matrices. During the alignment, we also take the atom types into account. The list of atom types is shown in Table 6.1.

To address the dependence of global alignment on the initial state, we propose to solve a problem of local alignment first and build to a set of global solutions. The procedure for our method is described as follows:

1. **Delaunay triangulation.** For each protein, we compute the 3-dimensional Delaunay triangulation to obtain a set of tetrahedra with labeled atoms as vertices. The two sets of tetrahedra are compared pair-wise in order to obtain similar pairs that act as seeds, which are used to obtain potential initial alignments for the matching process.

2. **Comparison of tetrahedra from two binding sites.** These inter-protein tetrahedral pairs are first checked for identical chemical composition. For those pairs with matching chemical compositions, the structural similarity of the tetrahedra is checked using the Distance Root Mean Square Deviation (dRMSD), which is calculated as follows:

\[
dRMSD(A, B) \propto \sqrt{\frac{1}{4(3)} \sum_{i=1}^{4} \sum_{j=1}^{4} \| l_{ij}^A - l_{ij}^B \|}
\]

where \( l_{ij}^A \) is the length of the edge from atom \( i \) to atom \( j \) of the tetrahedron from the first protein and \( l_{ij}^B \) is the length of the edge from atom \( i \) to atom \( j \) of the tetrahedron from the second protein. The proportional equivalent shown in the second line is used to obtain
simpler relationships to other quantities. This formulation can be used only because every tetrahedron has six edges, resulting in the two formulae differing only by the same constant for all pairs.

At this stage, dRMSD is used in place of Root Mean Square Deviation (RMSD), which is utilized subsequently, in order to save on computational cost. The only pairs considered further are those with dRMSD values less than a 1.5 times a chosen RMSD cutoff value (1.25 Å). This cutoff for dRMSD was chosen based upon the relationship between pairs of RMSD values and corresponding dRMSD values for a number of superpositions.

In many cases, the chemical composition for a tetrahedral pairing may lead to the possibility of multiple potential alignments. This occurs if there are multiple atoms of the same type within a tetrahedron. For example, if the tetrahedron consists of three carbon atoms and one oxygen atom, there are 6 possible alignments of the tetrahedra. In such instances, all possible alignments must be initially considered. All the tetrahedra in the two binding sites are compared and their dRMSDs are sorted.

3. Iterative alignment. Once all pairs of tetrahedral seeds are obtained and sorted, the process of checking for additional matched atoms begins. For each seed pairing, one tetrahedron is held in a fixed position and the other is superposed onto it, yielding an optimal translation vector \( v \) that aligns the centers of mass for the seeds, and rotation matrix \( R \), both of which are then applied to the moving protein, resulting in a rigid transformation that aligns the proteins at the location of the seed pairing. The translation vector is given by

\[
v = B_C - A_C
\]

where \( B_C \) and \( A_C \) are, respectively, the centers of mass for the coordinates \( A \) and \( B \). The optimal rotation matrix \( R \) is calculated in the following manner:

\[
C = B_C^T A_C = U \Sigma V, \quad R = UV^T
\]

For this superposition, the RMSD is then calculated using the formula

\[
RMSD(A, B) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} \|a_i - (b_i - v) \cdot R\|^2}
\]

where \( a_i \) and \( b_i \) are, respectively, the coordinates for the \( i \)th atom in \( A \) and \( B \). If the RMSD for this configuration is less than the chosen cutoff value of 1.25 Å and \( \det R = 1 \), then additional matched atoms are searched for. Despite considering only tetrahedral pairs with low dRMSD at this point, it is still necessary to calculate the RMSD so as to solve the multiple solution problem discussed above, resulting in the removal of improper seed pairs.
The restriction on the determinant of R is used to ensure that it is a true rotation matrix and not a rotation-reflection matrix.

Once the translation and rotation matrices are applied to the moving protein, we want to search for each atom of the moving protein an atom with the same type from the fixed protein with a distance smaller than a cutoff value, called the search radius (SR). To determine the matches for each atom, it does not suffice to consider each atom from the fixed protein separately due to the fact that doing so could lead to multiple fixed atoms sharing matches. For a given alignment, the solution to this matching problem, once the restrictions on labels and locality are imposed, is provided by the Hungarian algorithm [112, 113, 114], which finds at most one unique match for each atom in the fixed protein. To find the optimal search radius we tested several reasonable values on a benchmark dataset (see Results).

After the matched atoms are found for a pair of tetrahedral seeds, additional matched atoms are searched for by refining the optimal configuration of the proteins by expanding the seed for each pairing to encompass all of the matched atoms for that superposition. To increase computational efficiency, those configurations that resulted in few atoms being matched are excluded from further consideration. For a given expanded seed, the optimal translation vector and rotation matrix are recalculated as above. For this refined superposition, additional matched atoms are searched for. This process of refining and searching is repeated for a given seed pair until no additional matched atoms are found.

After using the iterative procedure to find the maximum number of matched atoms, the number of matched atoms is recorded. This process of superposing and searching for additional matches is repeated for all pairs of tetrahedral seeds. However, it can often be the case that multiple tetrahedral pairings will result in the same superposition of the moving protein onto the fixed protein. In order to avoid needlessly repeating the process in such cases, if the list of matched atoms includes all of the atoms from a remaining tetrahedral pair, then that pair is removed from consideration as a possible seed.

Upon completion of the above procedure, the optimal superposition is taken to be the configuration that results in the largest CAS. In the case that multiple configurations produce the largest number of matched atoms, the optimal configuration is taken to be that with the smallest RMSD. Accordingly, a list of the matched atoms in corresponding order is also obtained.

6.4 Verification of Algorithm

6.4.1 Benchmark Data set

In order to assess the performance of the algorithm, we perform classification of binding ligand using the benchmark data set compiled by Kahraman et al (2007) [87]. The Kahra-
Table 6.2: The binding sites included in the Kahraman set, listed by ligand

<table>
<thead>
<tr>
<th>Ligand</th>
<th>PDB IDs for the Proteins</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMP</td>
<td>12as, 1amu, 1c0a, 1ct9, 1jp4, 1kht, 1qb8, 1tb7, 8gpb</td>
</tr>
<tr>
<td>ATP</td>
<td>1a0i, 1a49, 1ayl, 1b8a, 1dv2, 1dy3, 1e2q, 1e8x, 1esq, 1gn8, 1kvk, 1o9t, 1rdq, 1tid</td>
</tr>
<tr>
<td>FAD</td>
<td>1cqx, 1e8g, 1evi, 1h69, 1hsk, 1jqi, 1jr8, 1k87, 1pox, 3grs</td>
</tr>
<tr>
<td>FMN</td>
<td>1dnl, 1f5v, 1ja1, 1mvl, 1p4c, 1p4m</td>
</tr>
<tr>
<td>GLC</td>
<td>1bdg, 1cq1, 1k1w, 1nf5, 2gbp</td>
</tr>
<tr>
<td>HEM</td>
<td>1d0c, 1d7c, 1dk0, 1eqg, 1ew0, 1gwe, 1iqc, 1naz, 1np4, 1po5, 1pp9, 1qhu, 1qla, 1qpa, 1sox, 2cpo</td>
</tr>
<tr>
<td>NAD</td>
<td>1ej2, 1hex, 1ib0, 1jq5, 1mew, 1mi3, 1o04, 1og3, 1qax, 1rlz, 1s7g, 1t2d, 1tox, 2a5f, 2npx</td>
</tr>
<tr>
<td>PO4</td>
<td>1a6q, 1b8o, 1brw, 1eqj, 1d1q, 1dak, 1e9g, 1ejd, 1euc, 1ew2, 1f7t, 1gyp, 1h6l, 1ho5, 1l5w, 1l7m, 1lby, 1lyv, 1qf5, 1tco</td>
</tr>
<tr>
<td>AND</td>
<td>1e3r, 1j99</td>
</tr>
<tr>
<td>EST</td>
<td>1fds, 1llu, 1qkt</td>
</tr>
</tbody>
</table>

man data set consists of 100 active sites that are grouped according to which of 10 ligands they bind to. These ligands have varying amounts of flexibility. AMP, AND, EST, GLC, and PO4 are rigid; ATP, FMN, and HEM are moderately flexible; FAD and NAD are highly flexible. The ligands are also of varying sizes, with PO4 being the smallest and FAD being the largest. A summary of the active sites included in the set is provided in Table 6.2. Our goal is to correctly determine which ligand a given active site binds to. For the purposes of comparison to previous studies, an active site is taken to consist of those atoms within 5.3 Å of the specified ligand (Hoffmann et al, 2010) [86].

6.4.2 Identification of Maximum Common Atom Set

We first verify whether the method can successfully find good solutions for the common atom set (CAS) from two binding sites. Although a proof that the obtained solution is optimal is difficult to provide, we have compared our matching result with another method, SitesBase, which is based on geometric hashing, on several pairs of binding sites. Here, we present a detailed example of one such case, comparing our algorithm to SitesBase and then provide a quick summary of comparisons for a few examples.

We consider the ATP binding sites of 1ayl and 1c2q, which are found in both the Kahraman set and the SitesBase set. For the purposes of this example, we utilize the definition of an active site used for SitesBase; the sites consist of all atoms within 5 Å of the ligand molecule. SitesBase found a common atom set of size 38, whereas our method found a common atom set of size 44. Table 6.3 displays those atom correspondences for which
Table 6.3: Atoms from ATP binding sites of 1ayl and 1c2q not found by both methods.

<table>
<thead>
<tr>
<th>SitesBase</th>
<th>Us</th>
</tr>
</thead>
<tbody>
<tr>
<td>37</td>
<td>CA 254</td>
</tr>
<tr>
<td>41</td>
<td>CD 254</td>
</tr>
<tr>
<td>58</td>
<td>C 256</td>
</tr>
<tr>
<td>59</td>
<td>CB 256</td>
</tr>
<tr>
<td>98</td>
<td>CA 441</td>
</tr>
<tr>
<td>108</td>
<td>NE 449</td>
</tr>
<tr>
<td>110</td>
<td>NH1 449</td>
</tr>
<tr>
<td>111</td>
<td>NH2 449</td>
</tr>
<tr>
<td>123</td>
<td>CG2 450</td>
</tr>
<tr>
<td>142</td>
<td>CG2 455</td>
</tr>
</tbody>
</table>

Table 6.4: Summary of CAS of two methods for three pairs of binding sites

<table>
<thead>
<tr>
<th>Binding Site Pair</th>
<th>Size of CAS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SitesBase</td>
</tr>
<tr>
<td>1 1ayl.ATP and 1e2q.ATP</td>
<td>38</td>
</tr>
<tr>
<td>2 1ct9.AMP and 1q19.APC</td>
<td>46</td>
</tr>
<tr>
<td>3 1qpa.HEM and 1tox.NAD</td>
<td>19</td>
</tr>
</tbody>
</table>

the methods did not agree. The methods agreed on all but 10 atoms. Our method found 7 matched atoms in the ATP site of 1e2q that SitesBase did not, while it found only 1 that ours did not. Atom NE 143 of the ATP site of 1e2q was found as a match by both methods, but it was found to match to different atoms from the ATP site of 1ayl due to differences in matching procedure.

One such difference is our use of the Hungarian algorithm, which finds the optimal MCAS for a given superposition, and can result in atom correspondence appearing to shift compared to other methodologies. The likely reason our algorithm did not find a match for the CA 441 atom of 1ayl while SitesBase did is due to the iterative aligning of our algorithm. While CA 441 may have been included in the common atom set based on the initial superposition of the tetrahedra, it could have been dropped from the set because we allow for the common atom set to change freely during the iterative alignment process. At the end of the procedure, there is no atom within the search radius around CA 441.

To further show how the maximum common atom sets from our algorithm compare to those of SitesBase, we display a summary of results for a few examples in Table 6.4. Within
pairs 1 and 2, the binding ligands are either identical or are structurally similar. For such cases, it should be expected that alignment methods would find very similar CAS if they are indeed finding common structures. For pair 2, our method finds identical correspondence for 37 of the atom pairs, while finding additional common atoms. Similarly to pair 1, a number of shifts are also present and SitesBase finds one pair that our method does not. Despite these differences, our method found similar sets of common atoms to SitesBase for pairs 1 and 2, but was able to identify additional matches, suggesting that our approach is able to find more atoms in the MCAS for similar binding sites. We also examined the additional atom matches we found for the second pair and they are indeed meaningful matches, which should be included in the CAS. In performing these comparisons, we used a search radius of 2.5 Å, which puts an upper bound for the resulting RMSD calculated from the obtained CAS at 2.5 Å.

In pair 3, the binding sites are for different ligands, which can be considered as a more challenging case. Our algorithm found 34 matched atoms, whereas SitesBase found only 19. Unlike the previous cases, the methods identified different common atom sets. However, since the ligands are not the same, it is reasonable to expect that there is no meaningful set of matched atoms. Instead, the atoms that are identified are essentially random matches, although still statistically significant. It is desirable to compare with more methods in terms their performance in detecting MCAS. However, neither the data nor the programs are available to us. Both the iterative procedure and Hungarian algorithm may contribute to the better performance of our method.

6.4.3 Classification of Ligand Binding Sites

We then compare our results to those of the sup-CK approach and MultiBind, both of which were tested on Kahraman data set by Hoffmann et al (2010). The Sup-CK and Sup-CKL methods do not consider atom correspondence, but rather use a family of Gaussian convolution kernels to determine similarity. However, the Sup-TI (there referred to as Sup-PI) method determines atom correspondence at the completion of their correspondence-free alignment and uses the Tanimoto Index (TI) to determine similarity. The TI for sites A and B is defined as follows:

\[
TI(A, B) = \frac{n_{AB}}{n_A + n_B - n_{AB}},
\]

where \(n_A\) and \(n_B\) are, respectively, the number of atoms in site A and the number of atoms in site B and \(n_{AB}\) is the number of atoms common to sites A and B.

Hoffmann et al (2010) also considers \(Vol(A, B) = |Vol_A - Vol_B|\) as a measure of the difference in the sizes of the active sites and present classification results using it alone.
Table 6.5: Summary of CAS of two methods for three pairs of binding sites

<table>
<thead>
<tr>
<th>Method</th>
<th>Classification Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>IN-TI</td>
<td>0.43</td>
</tr>
<tr>
<td>IN-TI + Gyr</td>
<td>0.29</td>
</tr>
<tr>
<td>IN-TI + RMSD4</td>
<td>0.43</td>
</tr>
<tr>
<td>IN-TI + HydProp</td>
<td>0.36</td>
</tr>
<tr>
<td>IN-TI + Gyr + HydProp</td>
<td>0.28</td>
</tr>
<tr>
<td>Gyr</td>
<td>0.54</td>
</tr>
<tr>
<td>RMSD4</td>
<td>0.71</td>
</tr>
<tr>
<td>HydProp</td>
<td>0.64</td>
</tr>
<tr>
<td>Sup-CK</td>
<td>0.36</td>
</tr>
<tr>
<td>Sup-CK + Vol</td>
<td>0.34</td>
</tr>
<tr>
<td>Sup-CK_L</td>
<td><strong>0.27</strong></td>
</tr>
<tr>
<td>Sup-CK_L + Vol</td>
<td>0.26</td>
</tr>
<tr>
<td>Vol</td>
<td>0.39</td>
</tr>
<tr>
<td>Sup-TI</td>
<td>0.42</td>
</tr>
<tr>
<td>MultiBind</td>
<td>0.42</td>
</tr>
</tbody>
</table>

and in combination with their method. The authors also provide classification results for MultiBind, which utilizes geometric hashing and uses a scoring method based upon the number of matched atoms.

For the purposes of comparison to the above methods, performance is measured using classification error (CE). The results are obtained using the double leave-one-out cross validation method for k-nearest neighbor classification described in Hoffmann et al (2010). Using this scheme, a classification is considered to be correct only if the predicted ligand exactly matches the actual binding ligand. The similarity of some of the ligands is not taken into consideration. A summary of CE for the methods examined is provided in Table 6.5. Our method with TI as a similarity measure, or IN-TI, has a CE of 0.43. There is a negligible difference between this method and the previous methods Sup-TI and MultiBind, which both result in CE of 0.42, suggesting that IN-TI compares well to these approaches when only a subset of matched atoms are considered in classification. While these CEs are far better than random, performances achieved based on solely the common atom sets identified by these methods are still not very satisfactory compared to Sup-CK (CE=0.36) and Sup-CK_L (CE=0.27) method, which also use information from other non-matched atoms.

Hoffmann et al (2010) also examined the volumes of the binding sites in the classification. The Vol measure alone performs fairly well, resulting in a CE of 0.39. However, using a linear combination of their Sup-CK and Sup-CK_L scores with Vol results in decreases in
CE of, respectively, 0.02 and 0.01. The reason for such a small decrease is likely due to the
fact that the Sup-CK and Sup-CK_L scores implicitly consider the sizes of the sites.

Inspired by the idea that global geometric information of the binding site could bolster
classification based on only matched atoms, we consider a number of additional measures.
RMSD is commonly used to measure structural similarity, but, just as the TI standardizes
\( n_{AB} \), it is desirable to standardize this, as well. To do so, we utilize a version of the
normalized RMSD of Carugo and Pongor (2001) [115], calculated as
\[
RMSD_4(A, B) = \frac{RMSD(A, B)}{1 + \ln \sqrt{n_{AB}}},
\]
which can be interpreted as being the RMSD value that would be observed for a pair of
binding sites containing 4 atoms which exhibit the same amount of similarity as the active
sites that were actually compared. It is natural to use 4 atoms for the normalization since
the initial step involves calculating the similarity of pairs of tetrahedra.

As shown in Hoffmann et al (2010) and Kahraman et al (2007), the size of a binding
site is useful for classification. While the TI only reduces (Davies et al, 2007) the effect of
the sizes of the binding sites for examining the number of matched atoms, an additional
measure, such as Vol, is needed to include size information. The calculation of the volume of
an active site is not trivial, though, so we consider an alternative method for incorporating
size information that is easy to calculate; the radius of gyration, which is calculated as
follows:
\[
Rg_A = \sqrt{\frac{1}{n_A} \sum_{i=1}^{n_A} (x_i^A - \bar{x}^A)^2},
\]
where \( x_i^A \) is the coordinate of the \( i \)th atom from site A and \( \bar{x}^A \) is the center of mass of site
A. This provides a measure of the average distance between the atoms and center of mass
of an active site. To utilize this information, we define the following similarity measure:
\[
Gyr(A, B) = |Rg_A - Rg_B|
\]
An additional measure we consider provide information about the chemical composition
of the binding sites. We consider the proportion of hydrophobic atoms present in the
active sites, and define the measure HydProp(A, B) to be the square difference between the
proportions for sites A and B.

Table 6.5 shows that, while each of these measures performs better than chance, they all
fall short of the previously explored methods. The best linear combination, calculated by
searching over a fine grid of weights, of each of these scores with IN-TI shows some amount
of improvement. HydProp improves classification slightly. The optimal linear combination
Table 6.6: The misclassified active sites with classified ligands in parentheses.

<table>
<thead>
<tr>
<th>Ligand</th>
<th>PDB IDs for the Misclassified Active Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>AMP</td>
<td>12as (ATP), 1amu (EST), 1c0a (GLC), 1jp4 (ATP), 1kht (EST), 1tb7 (AND), 8gpb (PO4)</td>
</tr>
<tr>
<td>ATP</td>
<td>1b8a (AMP), 1dy3 (FMN), 1esq (AMP), 1gn8 (FMN), 1o9t (NAD), 1tid (EST)</td>
</tr>
<tr>
<td>FAD</td>
<td>1jr8 (HEM)</td>
</tr>
<tr>
<td>FMN</td>
<td>1f5v (ATP), 1ja1 EST), 1mvl (GLC), 1p4m (NAD)</td>
</tr>
<tr>
<td>GLC</td>
<td>none</td>
</tr>
<tr>
<td>HEM</td>
<td>1qpa (NAD)</td>
</tr>
<tr>
<td>NAD</td>
<td>1ej2 (PO4), 1ib0 (AMP), 1o04 (HEM), 1tox (AND)</td>
</tr>
<tr>
<td>PO4</td>
<td>1e9g (GLC), 1gyp (ATP), 1l7m (NAD)</td>
</tr>
<tr>
<td>AND</td>
<td>1e3r (EST), 1j99 (NAD)</td>
</tr>
<tr>
<td>EST</td>
<td>1fds (AND)</td>
</tr>
</tbody>
</table>

places roughly 80% of the weight on HydProp. The CE for the optimal linear combination of RMSD<sub>4</sub> and IN-TI is 0.43, resulting in a reduction of less than 0.001, suggesting that this feature does not warrant being included.

Using a linear combination of Rg with IN-TI produces a CE of 0.29, which is comparable to Sup-CK<sub>L</sub>. However, our alignment method also results in obtaining atom correspondence, achieving the goal of classification, while also identifying the important atoms. The optimal combination places 52% of the weight on Rg, indicating both features contribute roughly the same amount towards the classification.

Since separate linear combinations with both Rg and HydProp improved classification results, we considered utilizing all three features to perform classification. Using a linear combination of IN-TI, Rg, and HydProp with respective weights of 37.74%, 41.51% and 20.75% produces a CE of 0.28. This reduction in CE is marginal and does not necessarily warrant the inclusion of HydProp. However, since some amount of improvement is shown, further investigation is warranted in the future.

Selecting the IN-TI + Rg model as optimal, we explored the erroneous classifications to gain a better understanding of the results. Using the cross validation procedure, the 29 binding sites listed in Table 6.6 were all classified correctly at most once while all of the other binding sites were *misclassified* at most once. This means that these 29 sites account for nearly all of the errors made in classification. The similarity matrix for the IN-TI + Rg model is shown in Figure 6.1. The ligand groups that produced the highest proportion of missed classifications are AMP, ATP, FMN, EST, and AND. From Figure 1, we see that the AMP, ATP, and FMN sites are troublesome due to the high amount of similarity with other
ligand groups. The EST and AND sites are misclassified largely due to the small number of sites binding to these ligands in the data. For the purposes of comparing results with Hoffmann et al (2010), these two ligands were considered separately. However, as presented in Kahraman et al (2007), these two groups can be combined and thought of as a steroid group, as displayed in Figure 6.1.

Kahraman et al (2007) examined the structural similarity of the ligands of these binding sites using spherical harmonic coefficients. The ligand similarity matrix shown in the paper has a similar overall pattern to what is shown in Figure 6.1. Among the 29 misclassified binding site pairs (Table 6.6), 16 of them bind to similar ligands. This brings the unexplained classification error to 0.13.
Table 6.7: Classification error for various combinations of SR and $k$.

<table>
<thead>
<tr>
<th>$k$-Nearest Neighbor</th>
<th>Search Radius</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1.0</td>
</tr>
<tr>
<td>1</td>
<td>0.56</td>
</tr>
<tr>
<td>3</td>
<td>0.58</td>
</tr>
<tr>
<td>5</td>
<td>0.58</td>
</tr>
</tbody>
</table>

Table 6.8: Classification error for various combinations of SR and $k$.

<table>
<thead>
<tr>
<th>$k$</th>
<th>1</th>
<th>3 and 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>IN-TI + Rg</td>
<td>0.29</td>
<td>0.33</td>
</tr>
<tr>
<td>IN-TI + RMSD4</td>
<td>0.43</td>
<td>0.41</td>
</tr>
<tr>
<td>IN-TI + HydProp</td>
<td>0.36</td>
<td>0.43</td>
</tr>
</tbody>
</table>

6.4.4 Effect of Search Radius and Number of Nearest Neighbors

In order to select the optimal search radius for the algorithm, we performed all pairwise comparisons for the Kaharaman data for SR of 1.0 Å, 1.5 Å, 2.0 Å, 2.5 Å and 3.0 Å and performed the double leave-one-out cross validation procedure. Table 6.7 shows the CE for these values of the search radius and for the nearest neighbor, 3-nearest neighbor, and 5-nearest neighbor classifiers. From Table 6.7, it is apparent that the optimal search radius is 2.5 Å. It appears that using a larger search radius defines similarity too loosely, resulting in dissimilar atoms being considered as matched. Using a smaller search radius appears to be too restrictive, not allowing for flexibility. The 3- and 5-nearest neighbor classifiers perform marginally better compared to the nearest neighbor when using the TI alone, but not enough to rule out using $k = 1$.

To more closely examine the optimal choice for $k$, we consider the linear combinations with the other similarity measures, as shown in Table 6.8. The CE are equal for $k = 3$ and $k = 5$. For the two linear combinations that perform better than IN-TI alone, the nearest neighbor works better than the other classifiers. For RMSD$_4$, the $k=3$ and $k=5$ perform marginally better than $k = 1$. Based upon these results, it appears that the nearest neighbor classifier provides the best results.

6.4.5 Conclusion and Discussion

This algorithm for landmark selection and registration is based on the iterative closest point (ICP) algorithm originally designed in computer vision for matching objects represented as point clouds. We addressed the starting-point problem by using similar tetrahedra...
from two binding sites being compared, which allows us to efficiently find good solutions. We applied the Hungarian algorithm in finding the optimal matched atoms at each iteration step and found that it significantly improved the matching results.

While our algorithm is similar in general to ICP, the aspects that differ allow us to better address the issues that arise in the problem of aligning protein binding sites. To solve the dependence on the initial alignment state, Besl and McKay (1992) suggest using a dense set of rotations to initialize the procedure. This approach works well if there is no additional information available to narrow down the initial alignments. However, by utilizing the tetrahedra from Delaunay triangulation to determine initial alignments, we are able to make use of the data, itself, to focus the alignment search to include only those that have the most promising outcomes.

By incorporating atom labels, we are able to further reduce the number of initial alignments to only those that are chemically feasible. The labels also greatly aid in the matching of atoms by considering correspondences only between atoms of the same type. This aids in the alignment process by using only these atoms of interest in the calculation of the alignment operators. Without using the labels, ICP considers all of the atoms from the moving binding site in the alignment step. This is problematic since the non-matched atoms can be thought of as noise that obscures the alignment.

In addition to considering only chemically identical atoms as potential matches, our algorithm is able to determine the optimal set of unique correspondences for a given alignment by utilizing the Hungarian algorithm. Furthermore, we restrict the set of potential matches to consist only of those atoms from the moving site that are near to the query atom from the stationary site. This censoring prevents an inflation of the number of matched pairs by ensuring that matches are local. The method for obtaining correspondences using ICP, while avoiding the locality problem, obtains a correspondence for every atom in the fixed binding site. This fails to solve the primary problem of identifying the common atom set.

The methodology (Brakoulias and Jackson, 2004) employed in the construction of Sites-Base is similar to our algorithm. However, there are a number of key distinctions. First, while it finds a set of one-to-one correspondences, the Hungarian algorithm is not used to do so. Thus, the optimal set of correspondences is not necessarily found. Secondly, the alignment process is not iterative. While this maintains the optimal alignment for the triplet pairs, it does not account for the information provided by the remaining atoms, including those that are determined to be matching.

Due in part to the flexibility and size of a number of the ligands, it is possible that our algorithm finds only a local alignment for a given pair of binding sites. For such pairs, there may be multiple regions of similarity that cannot all be captured by the same alignment. To
account for these cases, it may be advantageous to consider multiple alignments beyond that which produced the largest set of matched atoms. This may help increase the understanding of the relationship between the structure and function of a binding site and aid in obtaining better classification results.

As both a preliminary study and as validation of the methodology, we used the algorithm to examine the similarity of binding sites and compared the results to methods developed by other researchers. Using the Tanimoto Index alone as a similarity measure, the classification results for our algorithm perform comparably to the established sup-TI and MultiBind methods. This suggests that our algorithm properly aligns pairs of binding sites. Using MATLAB on a machine running Windows XP on an Intel Core 2 Duo processor running at 2.33 GHz, our approach required 16.4 hours to consecutively perform all 4950 pairwise alignments for the Kahraman dataset with a search radius of 2.5, or roughly 12 seconds per alignment.

However, upon further study, we found the including global geometric information in the form of the radii of gyration for binding sites allowed us to achieve a performance comparable to the best performance in previous studies. While the previous best performing method (Hoffmann, et al., 2010) does not obtain the common atom set between binding sites and the correspondences among the atoms, our method does. This means that while both methods perform comparably at determining similarity, the Sup-CKL method is not usable for achieving the goal of obtaining a common atom set with correspondences. Furthermore, the linear combination of scores from the factors we investigated allows a greater understanding of the roles the factors play in classification. Most notably, it is evident that the identification of common atom sets is not sufficient for adequately predicting the function of a binding site. The effectiveness of incorporating the radius of gyration in the classification shows that structural information is key, as well. This suggests that size-and-shape analysis should, indeed, be useful for the prediction of protein function.

### 6.5 Mean Size-and-Shape of Protein Binding Sites

The ultimate goal for this area of research is to combine the results of the alignment algorithm with the reflection size-and-shape analysis to aid in studying the relationship between the structure and function or binding sites. An important first step in doing so is calculating Schoenberg extrinsic sample mean reflection size-and-shapes. To demonstrate, we performed the above computations for a data set consisting of 4 protein binding sites. These sites, obtained from the RCSB Protein Data Bank and shown in Figure 7.4, are found in the proteins 1phf, 1phg, 2cpp, and 1m85 and bind to the ligand heme. The sample extrinsic mean size-and-shape is shown in 7.5.
While it is straightforward to calculate the extrinsic mean size-and-shape for a given sample, great care must be taken in selection of the sample to ensure that the binding sites are appropriately similar in structure. If this is not done, the mean size-and-shape for the resulting sample can be greatly altered. This can manifest as some combination of structural distortion and a great reductions in the number of atoms common to all of the binding sites in the sample. Both of these effects can thus result in the extrinsic mean size-and-shape to not truly be a representative form for the data.

In essence, the problematic binding sites can be considered to be outliers and the selection of proper samples involves an outlier detection step. As a result, it is crucial to develop a method for appropriately grouping sites in order to conduct meaningful size-and-shape analysis. It is likely not feasible to perform this in a fully automated fashion due to the complexity of the data. However, it is also infeasible to group the binding sites manually due to the amount of data available. Doing so would be difficult even with just the 100 sites in the Kahraman dataset. Instead, it is desirable to develop an approach that uses statistical methodology, such as clustering, as a means for forming initial groups and using these as a starting point to greatly aid in the final manual grouping.
Figure 6.3: The Schoenberg extrinsic mean reflection size-and-shape of the atoms common to the binding sites
CHAPTER 7

CURRENT DIRECTIONS IN STATISTICS ON MANIFOLDS

In our technological era, non-Euclidean data abounds, especially due to advances in digital imaging. Patrangenaru (1998) [38], following some definitions in the habilitation of Riemann and Gauss Egregium theorem, introduced extrinsic and intrinsic means on manifolds, as location parameters for non-Euclidean data, as mentioned earlier in this chapter. Once large sample and parametric bootstrap analysis was set in place around the year 2000 by Bhattacharya and Patrangenaru [1], [16], a flurry of papers in computer vision, statistical learning, pattern recognition, medical imaging and other computational intensive applied areas using these concepts followed. When pursuing such location parameters in varying instances of data analysis on manifolds, scientists are using intrinsic means, almost without exception. On the other hand, John Nash’s celebrated isometric embedding theorem [116] and its equivariant version due to J. D. Moore (1976)[117], show that any Riemannian manifold \((M,g)\), can be embedded in an Euclidean space isometrically, in such a way that the group of isometries of \((M,g)\) is made of Euclidean distance preserving transformations, restricted to the image of such an embedding.

The intrinsic mean \(\mu_I(Q)\) is the Fréchet mean of a probability measure \(Q\) on a complete \(d\)-dimensional Riemannian manifold \(M\) endowed with the geodesic distance \(d_g\) determined by the Riemannian structure \(g\) on \(M\).

It is known that if a probability distribution \(Q\) on a Riemannian manifold \((M,g)\) is sufficiently concentrated, then the intrinsic mean \(\mu_I(Q)\) exists. However there are no known conditions for the existence of the intrinsic mean of a distribution that is spread of even a slightly large region on a manifold [118].

On the other hand, the extrinsic mean \(\mu_E(Q) = \mu_{j,E}(Q)\) of a probability measure \(Q\) on a manifold \(M\) with respect to an embedding via \(j : M \to \mathbb{R}^k\) into an Euclidean space, always exists and is unique if the ordinary mean of \(j(Q)\) is a nonfocal point of \(j(M)\), i.e., if there is a unique point \(x_0\) on \(j(M)\) having the smallest distance from the mean of \(j(Q)\).
In practical applications, it is known that it is easier to compute the extrinsic mean in the particular case of distributions on Grassmann manifolds (real or complex). It should also be noted that if $Q$ is highly concentrated, as in medical imaging data examples (Bhattacharya and Patrangenaru (2003)[1]) the intrinsic and extrinsic means are virtually indistinguishable. An example of such a case will also be provided for protein structure data from Chapter 6.

It is also known that extrinsic analysis is computationally faster than intrinsic analysis. Computational examples will be shown for shape data from Chapters 5 and 6. In addition, there are as many Riemannian structures on an abstract manifold as there are embeddings. Due to these advantages, the current trend of using intrinsic analysis in computationally intensive applications may soon shift towards an extrinsic approach. This chapter provides computational examples illustrating some of the advantages of extrinsic analysis for a number of types of data.

However, in addition to the philosophical and computational debate between intrinsic and extrinsic analysis, this is an exciting time in the continuing development of the field. Recently, work has begun in the studying of the analysis of data sampled from more complicated spaces, particularly spaces with manifold stratification. To analyze such data, it is necessary to further generalize and adapt methodology to deal with the inherent complications, such as strata of varying dimensions and singularities. Following the computational examples showcasing computational differences between intrinsic and extrinsic analysis in Section 7.1, some examples of data sampled from spaces with manifold stratification will be briefly discussed.

### 7.1 Computational Comparison of Intrinsic and Extrinsic Analysis

In this section, we present examples of various types of data analysis on manifolds. For each type of data analysis, methodology is described for the calculation of an extrinsic mean and an intrinsic mean for that type of data. Example computations are then performed for a given sample and presented with the required computational time. For timing purposes, all computations were performed using MATLAB on a machine running Windows XP on an Intel Core 2 Duo processor running at 2.33 GHz.

#### 7.1.1 Directional Data Analysis

We first consider data analysis for directional data. In general, given a random object $X$ on a manifold $\mathcal{M}$ and an embedding $J : \mathcal{M} \to \mathbb{R}^N$, if the mean vector $\mu = E(J(X))$ is
a nonfocal point of $J(M)$, then the extrinsic mean is $\mu_E = J^{-1}(P_J(\mu))$, where $P_J$ is the projection on $J(M)$. In particular, given a random vector $X, X^TX = 1$ on the unit sphere $S^p \subset \mathbb{R}^{p+1}$, the extrinsic mean is $\mu_E = \frac{1}{\|E(X)\|}E(X)$. The extrinsic mean is the projection of the Euclidean mean onto $S^p$. For observations $x_1, \ldots, x_n \in S^p$, the extrinsic sample mean $\hat{\mu}_E$ is calculated similarly using the formula:

$$\hat{\mu}_E = \frac{\bar{x}}{\|x\|},$$

(7.1)

$$\hat{\mu}_E = \frac{\bar{x}}{\|\bar{x}\|},$$

(7.2)

where $\bar{x}$ is the usual Euclidean mean and $\| \cdot \|$ is the standard norm. This mean is based upon the chord distance between points.

For observations $x_1, \ldots, x_n \in S^p$, the intrinsic sample mean $\hat{\mu}_I$ is the minimizer of the Fréchet function:

$$F(p) = \sum_{i=1}^{n} d^2(x_i, p),$$

(7.3)

where $d(x, y)$ is the arc distance between points $x$ and $y$. Unlike the extrinsic mean, there is no closed form solution for the intrinsic mean. Instead, an iterative algorithm must be used. Such an algorithm is given as follows:

1. Make an initial estimate $\hat{\mu}_I$ of $\hat{\mu}_E$; i.e. Use $\hat{\mu}_E$

2. For each $x_i$, compute

$$v_i = u_i \ast \arccos \frac{x_i \hat{\mu}_I^T}{\sqrt{u_i u_i^T}},$$

where $u_i = x_i - (x_i \hat{\mu}_I^T)\hat{\mu}_I^T$

3. Compute $\bar{v}$ and update $\hat{\mu}_I$ in the following manner:

$$\hat{\mu}_{I,new} = \cos (\epsilon \sqrt{\bar{v}\bar{v}^T})\hat{\mu}_I + \sin (\epsilon \sqrt{\bar{v}\bar{v}^T}) \frac{\bar{v}}{\sqrt{\bar{v}\bar{v}^T}},$$

where $\epsilon > 0$.

4. Repeat until $\sqrt{\bar{v}\bar{v}^T} < \epsilon$.

To illustrate these computations and the time required to perform such tasks, we consider a set of wind direction data in $S^2$ from Fisher, Lewis, and Embleton (1987, p.308). The data consist of 29 observations and is provided as pairs of colatitude and longitude. Figure 7.1 displays this data on $S^2$ in four views.
The extrinsic and intrinsic means were calculated using the methods described above and are shown in Figure 7.2, though in order to calculate the extrinsic mean, it was first necessary to convert to Cartesian coordinates. Additionally, a heat map for the Fréchet function of which the intrinsic mean is the minimizer is displayed for a grid of values of colatitude and longitude in Figure 7.2. The amount of time required to compute $\hat{\mu}_E$ was $9.531 \times 10^{-5}$ seconds. The amount of time required to compute $\hat{\mu}_I$ was 10.88 seconds.

![Figure 7.1](image1.png)

Figure 7.1: The wind direction data used for the computations.

![Figure 7.2](image2.png)

Figure 7.2: (a) The extrinsic and intrinsic means for the data (b) The Frechet function using arc distance as a function of colatitude and longitude
To further illustrate the disparity in computational cost, we consider the calculation of bootstrap means. Using the same methodology as previously, 200 resamples were taken and both types of means were obtained for each. To obtain the bootstrap means, as displayed in Figure ??, the methodology was applied again on the sample of means. For the extrinsic mean, the computational time required was 0.016762 seconds. For the intrinsic mean, the computational time required was 1572 seconds.

Figure 7.3: The extrinsic and intrinsic bootstrap means means for the data compared to the extrinsic and intrinsic means

7.1.2 Size-and-Shape Analysis for 3D Configurations

We now consider data for which we are interested in analyzing both the shape and size of objects. For such data, each observation is represented as a k-ad, an ordered set of k points. A k-ad in \( \mathbb{R}^p \) is said to be in general position if the k-ad spans \( \mathbb{R}^p \). For our purposes, we consider the case that \( p=3 \).

Let \( \{x_1, \ldots, x_n\} \), where \( x_j = (x_{1j}, \ldots, x_{kj}), j = 1, \ldots, n, \) be a sample of k-ads in general position in \( \mathbb{R}^3 \). Recall from Chapter 6 that the **extrinsic sample mean reflection size-and-shape** is \( [\bar{x}]_E = [\bar{\xi}]_RS \), where \( \bar{\xi} \) is given by the eigenvectors corresponding to the 3 largest eigenvalues of

\[
\hat{C} = \frac{1}{n} \sum_{j=1}^{n} \xi_j^T \xi_j
\]

assuming that \( \hat{\lambda}_3 > \hat{\lambda}_4 \), where \( \hat{\lambda}_1 \geq \cdots \geq \hat{\lambda}_k \) are the eigenvalues of \( \hat{C} \), where \( \xi_j = x - \bar{x} \) (Bandulasiri et al (2009)).
The intrinsic sample mean size-and-shape $\hat{\mu}_I$ is the minimizer of the Frechet function:

$$F(p) = \sum_{i=1}^{n} d^2(x_i, p) = \sum_{i=1}^{n} \inf_{\Gamma_i \in SO(3)} \|p - x_i \Gamma_i\|^2,$$

(7.4)

where $\| \cdot \|$ is the standard norm in $\mathbb{R}^3$ and $\Gamma_i$ is a special orthogonal matrix. As with the spherical data, there is no closed form solution for the intrinsic mean size-and-shape. Instead, the following iterative algorithm is used (Dryden and Mardia (1998)):

1. Make an initial estimate $\tilde{\mu}_I$ of $\hat{\mu}_I$; i.e. Use $\hat{\mu}_E$

2. For each $x_i$, find the optimal "rotation" matrix $\Gamma_i$ using Procrustes alignment and compute

$$V_i = \xi_i \Gamma_i - \tilde{\mu}_I,$$

(7.5)

3. Compute $\bar{V}$ and update $\tilde{\mu}_I$ in the following manner:

$$\tilde{\mu}_{I,\text{new}} = \tilde{\mu}_I + \epsilon \bar{V},$$

(7.6)

where $\epsilon > 0$.

4. Repeat until $\|\bar{V}\| < \epsilon$.

To demonstrate, we performed the above computations for a data set consisting of 4 protein active sites. The active sites, obtained from the RCSB Protein Data Bank and shown in Figure 7.4, are found in the proteins 1phf, 1phg, 2cpp, and 1m85 and bind to the ligand heme. As shown in figure 7.5, the extrinsic mean size-and-shape is visually indistinguishable from the intrinsic mean-size-and-shape.

To detail the computational speeds of the two types of analysis, bootstrap means are computed, similarly to the wind direction data. To examine the effect of sample size on the
computational cost, these calculations were performed for samples of size 4, 5, 6, 8, 12, and 16. For the samples of size greater than 4, the observed data was simulated based upon the original sample. The times, in seconds, required for these computations are shown in Table 7.1. Increasing the sample size has no significant effect on the computational cost for calculating the mean-size and-shape. However, increasing the sample size has a large effect on the computational cost for calculating the intrinsic mean-size-and-shape.

7.1.3 Direct Similarity Shape Analysis of Planar Contours

In this section, we return to the data analysis of planar contours described in detail in Chapter 5. Recall that for our approach based on an extension of Kendall’s shape space, the data is discretized and approximated using landmark-based shape analysis. In this approach, such data are discretized and each curve is represented as a $k$-ad, where each landmark on the curve is a point in $\mathbb{C}$. Let $\zeta_1, \ldots, \zeta_n$ be a sample of centered $k$-ads. The Veronese-Whitney extrinsic sample mean shape is the unit eigenvector corresponding to the
The recently developed elastic framework for shape analysis of planar curves instead utilizes functional representation for such data and utilizes intrinsic analysis. As with the previously discussed types of data, there is no closed form for the intrinsic mean shape, so an iterative algorithm similar in concept, but including adaptations for rotation and reparametrization, to that used for the wind direction data must be used to perform computations.

Computations for both approaches were performed on a sample of 4 observations of contours of the "l" hand gesture, shown in Figure 7.6. To illustrate the difference in computational cost, 95% bootstrap confidence intervals were computed for both the extrinsic mean shape and the intrinsic mean shape using 400 resamples and 300 randomly chosen landmarks. These confidence regions are shown in Figure 7.7. For the extrinsic mean shape, these calculations required 48 seconds to complete. However, for the intrinsic mean shape, these calculations required 47.9 hours.

As a second example, these methods were also performed on a sample of 4 observations of contours of dogs, which is shown in 7.8. Again, 95% bootstrap confidence regions were computed for both approaches, using 300 resamples, where each contour is provided as 100 evenly spaced landmarks. These confidence regions are shown in Figure 7.7. For the extrinsic mean shape, these calculations required 5.6 seconds to complete. However, for the intrinsic mean shape, these calculations required 8.9 hours.

The higher computational cost is due to a combination of the intrinsic analysis and the elastic representation. The calculation of an intrinsic mean requires the use of an iterative algorithm. Klassen et al. (2004) provides an algorithm for doing so for arc-length parametrized curves. The square-root elastic framework of Joshi et al. (2007) [119] adapts this algorithm by inserting a reparametrization step at each iteration. This reparametrization step requires the use of either a dynamic programming algorithm or a gradient descent approach. Due to the nature of the algorithm, these time-consuming steps are repeated

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>Extrinsic</th>
<th>Intrinsic</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>15.9</td>
<td>29.3</td>
<td>1.84</td>
</tr>
<tr>
<td>5</td>
<td>15.3</td>
<td>32.0</td>
<td>2.09</td>
</tr>
<tr>
<td>6</td>
<td>15.2</td>
<td>36.2</td>
<td>2.38</td>
</tr>
<tr>
<td>8</td>
<td>14.8</td>
<td>60.2</td>
<td>4.07</td>
</tr>
<tr>
<td>12</td>
<td>15.3</td>
<td>92.1</td>
<td>6.01</td>
</tr>
<tr>
<td>16</td>
<td>16.5</td>
<td>123.6</td>
<td>7.49</td>
</tr>
</tbody>
</table>

Table 7.1: The times required to compute the mean size-and-shape for various sample sizes.
Figure 7.6: 4 observations of contours of the "l" hand gesture

Figure 7.7: Bootstrap 95% confidence regions using 400 resamples for (a) the extrinsic mean shape of the "l" hand gesture and (b) the intrinsic mean shape of the "l" hand gesture

a number of times during the calculation of the intrinsic mean. When obtaining a bootstrap confidence region, this computational cost is further compounded, resulting in a far
greater run time. It should also be noted that the differences in appearance are due to the differences in the representation of the contours. For example, the difference in the size of the contours displayed in Fig. 7.7 is due to the approaches using different methods for normalization.
7.2 Data Analysis on Sample Spaces with Manifold Stratification

While the theory and methodology described throughout this dissertation has been based upon manifold models, not all shape spaces are manifolds. As discussed in Section 6.2.1, Kendall’s direct similarity shape spaces have singularities for configurations of landmarks in 3-dimensions and higher. In order to work with such data, Bandulasiri et al (2008) worked around this problem by introducing the indirect similarity shape spaces for configurations in general position. Unlike the former case, these spaces are manifolds. However, there are additional sample spaces that are not themselves in entirety manifolds, but rather have manifold stratification.

7.2.1 The Reflection-Affine Shape Spaces

One such type of sample space with manifold stratification is the space of reflection-affine shapes. The reflection-affine shape $\tilde{\alpha}(x)$ of the $k$-ad $x$ under the diagonal action $\alpha_k$ of the affine group on $(\mathbb{R}^m)^k$, where $\alpha_k : Aff(m) \times (\mathbb{R}^m)^k \to (\mathbb{R}^m)^k$ such that $\alpha_k(A,B)(x_1,\ldots,x_k) = (y_1,\ldots,y_k)$ where $y_j = Ax_j + b, \forall j = 1,\ldots,k$ Summing up over $j$ results in $\bar{y}_k = A\bar{x}_k + b$. If $x_j - \bar{x}_k = \xi_j$ and $y_j - \bar{y}_k = \zeta_j$, then $\zeta_j = A\xi_j \forall j = 1,\ldots,k$.

The reflection-affine shape of $(x_1,\ldots,x_k)$ is the orbit of $(x_1,\ldots,x_k)$ under $\alpha_k$. In other words, $(x_1,\ldots,x_k)$ and $(y_1,\ldots,y_k)$ have the same reflection-affine shape if $(y_1,\ldots,y_k) = \alpha_k((A,b)(x_1,\ldots,x_k))$. To each $k$-ad $(x_1,\ldots,x_k)$, we associate the $k$-ad $(\xi_1,\ldots,\xi_k)$ which is centered at $0 \in \mathbb{R}^m (\bar{\xi}_k = 0)$. This filters out translation, which means if we consider an action $\beta_k$ of $GL(m,\mathbb{R})$ on $L^m_k = \{\xi \in (\mathbb{R}^k)^m : \xi 1_k = 0_m\}$, given by $\beta(A,\xi) = A\xi = \zeta$, there is a one-to-one correspondence between reflection-affine shapes and orbits of the action $\beta_k$.

The reflection-affine shape space $A\Sigma^k_m$ of $k$-ads in $m$-dimensions can be stratified as follows: $A\Sigma^k_m = (A\Sigma^k_m)_m \cup (A\Sigma^k_m)_{m-1} \cup \cdots \cup (A\Sigma^k_m)_0$, where $(A\Sigma^k_m)_m$ is the space of reflection-affine shapes if the configuration is in general position in $\mathbb{R}^m$, and $(A\Sigma^k_m)_j$ is the space of reflection-affine shapes if the configuration spans a subspace of dimension $j \forall j = 0,\ldots,m-1$. Each of these stratifications can be shown to be a Grassman manifold so that the stratification of the reflection-affine shape space can be re-expressed as $A\Sigma^k_m = GL_m(L_k) \cup GL_{m-1}(L_k) \cup \ldots$, where $L_k \subset \mathbb{R}^{k-1}$. The details of the geometric construction of this space are presented in detail by Groisser and Tagare (2009) [120]. Despite the stratified structure, however, for data analysis it is much more feasible to work solely with the stratum $(A\Sigma^k_m)_m$ in the usual manner due to the complicated manner in which the strata are attached.
7.2.2 The Space of Phylogenetic Trees

While the non-manifold structures of the sample spaces mentioned above can be avoided by considering, respectively, orbits under a different group and only one of the strata, there are other sample spaces for which, currently, no known work-around exists, so the full stratified space must be considered. One such example is the space of phylogenetic trees.

A metric $n$-tree is defined to be a connected graph with no circuits that has a distinguished vertex called a root and $n$ sequentially-labelled vertices of degree 1, which are called leaves, such that all interior edges have lengths greater than 0. Interior edges are defined to be those edges that do not connect to a leaf. In this construction, a tree is specified by the lengths of it's interior edges.

A binary $n$-tree with interior edges $e_1, \ldots, e_{n-2}$ with respective lengths of $l_1, \ldots, l_{n-2}$ is represented by the vector $(l_1, \ldots, l_{n-2}) \in (0, \infty)^{n-2}$. Every $n$-tree in the same orthant is combinatorially the same, but differ in the lengths of their interior edges. A point on the boundary of a given orthant has at least one interior edge with zero length and the corresponding tree can also be represented as a combinatorially different tree with the same interior edges having zero length.

The tree space $T_n$ is constructed by attaching the $(2n-3)!! = \frac{(2n-2)!}{2^{n-1}(n-1)!}$ combinatorially different orthants along their common faces. This construction leads to the stratification of the space such that each orthant is a flat $(n-2)$-dimensional manifold and those faces along which orthants are attached are lower dimensional flat manifolds. The orthants are all attached at the origin $0_p$, which is the point at which all interior edges have zero length, and is a singularity. This point corresponds to the so-called star tree, which is named for it's star-like appearance.

For high-dimensional tree spaces, it is difficult to visualize the stratification due to the combinatorics involved, but the spaces can be described easily for $n = 3$ and $n = 4$. $T_3$ is the space of binary trees with 3 leaves and 1 interior edge and so consists of 3 rays that are attached at the origin. The stratification is thus given by 3 strata of dimension 1 and 1 stratum of dimension 0.

$T_4$ is the space of binary trees with 4 leaves and 2 interior edges and consists of 15 planar quadrants that are all attached at the origin. The stratification of this space is somewhat more complicated than $T_3$. It consists of 15 strata of dimension 2, 10 strata of dimension 1, and a stratum of dimension 0, which is again at the origin.

Tree data has so far been analyzed entirely using intrinsic analysis with geodesic distances defined as follows. For two points in $T_n$ that either lie in the same orthant or share a common face, the distance between them is given by the Euclidean distance. For all other points, the distance is the sum of the distances between each point and the origin. Billera
et al (2001) [121] proved that $\mathcal{T}_n$ has non-positive curvature, which results in each geodesic being unique. Owen and Proven (2011) [122] developed a polynomial-time algorithm for computing geodesics.

The notion of Fréchet means applies to tree spaces. However, while it is simple to understand the problem of finding a mean tree, calculating a sample mean is more difficult. While the intrinsic mean is guaranteed to exist by the non-positive curvature of the space, there is currently no known algorithm for computing the intrinsic sample mean for observations on general $\mathcal{T}_n$ and there is no guarantee of computational efficiency, especially for large $n$. If $\mathcal{T}_n$ is canonically embedded in $\mathbb{R}^d$, where $d$ is the number of 1-dimensional strata, then the sample mean vector in the ambient space is easy to compute. However, finding the projection onto $\mathcal{T}_n$ may be computationally difficult for large $n$ due to the combinatorics of the space. One of our future goals is to develop a fast algorithm for computing extrinsic means, as this would be an important step forward for the analysis of phylogenetic tree data.

7.3 Summary and Conclusions

The computational cost of performing extrinsic analysis on manifold data is substantially less than the computational cost of performing intrinsic analysis on the same data. This is especially noticeable when working with large data sets and/or performing analysis requiring large numbers of repetitions, as with nonparametric bootstrap techniques. As shown with the protein data, in many cases, the extrinsic and intrinsic means are indistinguishable from each other despite the difference in computational time, providing strong support for the use of extrinsic analysis in such situations.

In other scenarios, one must look to the requirements for the application at hand. Certainly, if efficiency is key, one should prefer extrinsic analysis. However, one must also consider other properties, as well. For instance, extrinsic means exist with probability 1, whereas intrinsic means are not, in general, guaranteed to exist. Furthermore, for a given embedding $J$, the extrinsic mean is the projection on $J(M)$, but there is no natural description of the intrinsic mean. Extrinsic covariance is known to be related to the 2nd fundamental form, but there is no such known relationship for intrinsic covariance. Due to these considerations, extrinsic analysis is often preferable for statistical analyses.

The notions of intrinsic and extrinsic analysis also apply to more general sample spaces that have manifold stratification. While some such spaces are more easily, but just as properly, analyzed when restricting data to just one stratum, other spaces cannot be simplified so. One example is the space of phylogenetic trees, which has diverse applications in various fields of biology. Due to the current predominance of intrinsic analysis among scientists,
tree data has so far been analyzed almost entirely using the intrinsic approach. However, the same computational restrictions exist for such data as they do for manifold data, and, as such, it is of great interest to apply extrinsic methodology, as it may greatly simplify computations and allow for more efficient analysis.
BIBLIOGRAPHY


BIOGRAPHICAL SKETCH

Leif Ellingson was born in 1985 in Columbia, Maryland. He graduated from River Hill High School in 2003 with his high school diploma. He then attended the University of Maryland and graduated in Spring 2007 with a Bachelor of Science degree in Mathematics and an Honors Citation. In the Fall of 2007, he enrolled in the Department of Statistics at Florida State University. He graduated with a Master of Science in Statistics in Spring 2009. Leif began his doctoral work under the supervision of Dr. Victor Patrangenaru in January 2009 and is expecting to graduate in the Summer 2011 term. During his graduate studies, Leif worked as a graduate teaching assistant, serving at various times as a grader, recitation leader, and solo instructor. While he taught one semester each of Fundamental Business Statistics and Introduction to Applied Statistics, Leif was most often involved in the instruction of Statistics through Example, having served in all three of the above roles for the course. His research interests include statistical shape analysis, nonparametric statistics on manifolds, asymptotic theory, structural proteomics, and the analysis of data from sample spaces with manifold stratification, including the space of phylogenetic trees.