IMAGE AND SHAPE ANALYSIS FOR SPATIOTEMPORAL DATA

Yi Hong

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

Chapel Hill
2016

Approved by:

Marc Niethammer

Stephen M. Pizer

J. S. Marron

Alexander C. Berg

Roland Kwitt

ABSTRACT

YI HONG: IMAGE AND SHAPE ANALYSIS FOR SPATIOTEMPORAL DATA.
(Under the direction of Marc Niethammer.)

In analyzing brain development or identifying disease it is important to understand anatomical age-related changes and shape differences. Data for these studies is frequently spatiotemporal and collected from normal and/or abnormal subjects. However, images and shapes over time often have complex structures and are best treated as elements of non-Euclidean spaces. This dissertation tackles problems of uncovering time-varying changes and statistical group differences in image or shape time-series.

There are three major contributions: 1) a framework of parametric regression models on manifolds to capture time-varying changes. These include a metamorphic geodesic regression approach for image time-series and standard geodesic regression, time-warped geodesic regression, and cubic spline regression on the Grassmann manifold; 2) a spatiotemporal statistical atlas approach, which augments a commonly used atlas such as the median with measures of data variance via a weighted functional boxplot; 3) hypothesis testing for shape analysis to detect group differences between populations. The proposed method for cross-sectional data uses shape ordering and hence does not require dense shape correspondences or strong distributional assumptions on the data. For longitudinal data, hypothesis testing is performed on shape trajectories which are estimated from individual subjects.

Applications of these methods include 1) capturing brain development and degeneration; 2) revealing growth patterns in pediatric upper airways and the scoring of airway abnormalities; 3) detecting group differences in longitudinal corpus callosum shapes of subjects with dementia versus normal controls.

ACKNOWLEDGMENTS

First of all, I would like to thank my advisor Dr. Marc Niethammer for his guidance and support during my PhD journey. His vision and wisdom pointed me into the direction of my thesis work. His optimism and enthusiasm encouraged me when I faced unknowns and challenges in research. He taught me how to think critically, write clearly, and more. I was educated by his broad knowledge of the field and his deep understanding of the theory and practice. Marc is a very generous and helpful mentor. He was always there when I had a question or when I got confused or lost. I am fortunate to have him as my advisor and without any doubt I will continue to benefit from his advising throughout my career.

It was a great pleasure to have worked with Dr. Roland Kwitt over these years. We had fantastic discussion and brain storming. I learnt various technical skills from him, even including tricks to generate beautiful figures. Also, I would like to thank other professors on my thesis committee, Dr. Stephen M. Pizer, Dr. J. S. Marron, and Dr. Alexander C. Berg, for great help with their expertise in image analysis, statistics, and computer vision. Dr. Stephen M. Pizer has always been patient with me and taught me why and how I should write or tackle a problem in a particular way. I appreciate that he held special hours for teaching me visual solid shapes. Dr. J. S. Marron is a walking library to me, from whom I always got my answers. Dr. Alexander C. Berg also gave me valuable feedback on research.

I am grateful to all my collaborators, including Dr. Brad Davis, Dr. Nikhil Singh, Dr. Sylvain Bouix, Dr. Martin Styner, Dr. Sarang Joshi, MD Carlton Zdanski, and Dr. Yi Gao.

This work was made possible with their help. I want to thank all fellow students (past and present) in our lab, including Dr. Liang Shan, Dr. Tian Cao, Istvan Csapo, Yang Huang, Heather Couture, Xiao Yang, and Xu Han. I always had highly useful discussions with them. I also thank my friends at UNC, including Qianwen Yin, Yaozong Gao, Zhishan Guo, Shan Yang, Dinghuang Ji, Qingyu Zhao, Wei Chen, and Enliang Zheng. They brought me a lot of fun.

I thank NIH and NSF for funding my work. Also, I thank the UNC Graduate school for providing the Dissertation Completion Fellowship, which allows me to focus on my dissertation in the last year. I am grateful to the faculty and staff in the UNC Computer Science Department and to all professors who taught me at UNC.

Finally, I am thankful for my beloved family. I am so lucky to share this experience with a wonderful person, my husband Dr. Qiang Cao. We originally came from China and managed to reunite in the triangle area of North Carolina. Over these years we learnt and grew together. I thank him for being in an important part of my life. To my parents, Shancheng Hong and Youxian Chen, my old sister Li Hong and old brother Lin Hong, no word can express how thankful I am for their constant support and unconditional love.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 : INTRODUCTION

1.1 Motivation

Spatiotemporal data analysis frequently arises in medical research and computer vision problems. Examples include MRI (magnetic resonance imaging) time-series collected to explore brain development [Evans et al., 2006], corpus callosum shapes at varying ages obtained to analyze the degeneration process of a single brain structure [Fletcher, 2013], and traffic video clips acquired by a surveillance system monitoring highway traffic to classify congestion in traffic sequences [Chan and Vasconcelos, 2005]. Analysis of this spatiotemporal data, e.g., image time-series, shape sequences, and videos, is an important topic in the field of computer vision and medical image analysis. In this dissertation, this topic will be explored from three aspects: 1) capturing time-varying changes within a subject or a population, e.g., studying brain development or a disease process; 2) estimating a spatiotemporal atlas while retaining population variation information, e.g., a pediatric airway atlas with a confidence region; and 3) identifying shape differences between populations, e.g., to differentiate normal control subjects from subjects with disease.

While methods to analyze time-varying scalar-valued data are advanced and well-studied [Kedem and Fokianos, 2005], the theory and methodology is much less developed but of equal importance, for spatiotemporal data, e.g., for image or shape sequences. Analysis of such data is challenging because it frequently no longer lives in Euclidean space, but instead in some non-Euclidean space which is a smooth manifold [Lee, 2012], e.g., the manifold of diffeomorphisms [Banyaga, 1997], Kendall shape space [Kendall, 1984], or the Grassmannian [Edelman et al., 1998]. Although a manifold resembles Euclidean space in the neighborhood of each point, globally it may not. As a result, methods developed in Euclidean space cannot be directly applied to the spatiotemporal data explored in this dissertation.

1.1.1 Estimation of Time-Varying Changes

Time-varying changes occur in spatiotemporal data, which is collected to study, for example, aging [Scahill et al., 2003], disease progression [Kogure et al., 2000], and brain development [Evans et al., 2006]. To summarize these changes within a subject or a population, regression analysis is popular, because it is a powerful tool to model the relationship between data objects [Marron and Alonso, 2014] and their associated descriptive variables, e.g., age.

Recently, regression models [Niethammer et al., 2011, Fletcher, 2013] have been proposed to estimate changes in shape or image time-series by generalizing linear regression in Euclidean space to Riemannian manifolds [Lee, 2012] and the manifold of diffeomorphisms respectively. An interesting problem is to develop a *uniform* framework of parametric regression, which arises when changes in different types of spatiotemporal data are to be captured using the equivalent of linear or higher-order fitting curves in non-Euclidean spaces. For instance, the collected data could be shapes or videos. While a regression model equivalent to linear curve fitting [Niethammer et al., 2011, Fletcher, 2013] is relatively simple, sometimes it may be too restrictive for data exhibiting more complex changes. In such cases, higher-order polynomial or spline fitting curves [Hinkle et al., 2014, Singh and Niethammer, 2014] can be attractive, but their formulations are typically complicated.

This dissertation develops regression models of increasing order for different types of spatiotemporal data and an approach for model criticism [Lloyd and Ghahramani, 2015] to check models' underlying assumptions. To achieve this, I first represent data objects, e.g., shapes or videos, as elements on the Grassmannian, using singular value decomposition (SVD) [Begelfor and Werman, 2006, Sepiashvili et al., 2003] and a dynamic texture model [Doretto et al., 2003] respectively. Then optimal control approaches are applied to develop regression models of increasing order on the Grassmannian. In existing work, two groups of solutions are presented to solve parametric regression problems in the spaces of shapes or images: 1) geodesic shooting based strategies that address the problem using adjoint methods from an optimal-control point of view [Niethammer et al., 2011, Hong et al., 2012a, Singh et al., 2013b], and 2) approaches that compute the required gradients using Jacobi fields for optimization [Rentmeesters, 2011, Fletcher, 2013]. Unlike Jacobi field approaches, solutions using optimal control methods do not require the computation of curvatures explicitly and can be easily extended to higher-order models, e.g., polynomials [Hinkle et al., 2014] or splines [Singh and Niethammer, 2014]. Hence, the strategy based on geodesic shooting is adopted to develop extensible solutions on the Grassmannian, i.e., extending the basic model to time-warped and cubic-spline variants. Overall, a uniform framework for *parametric regression on the Grassmannian* is proposed to solve fitting problems with models of increasing order for different types of spatiotemporal data that can be represented as elements on the Grassmann manifold.

Besides, in existing regression models for image time-series [Niethammer et al., 2011], image intensities are generally not explicitly captured and instead accounted for by using image similarity measures, which do not appropriately model some changes. Because typically image intensity changes occur jointly with spatial deformations, for example, in MRI studies of the developing brain due to myelination. While approaches accounting for intensity changes after image registration exist [Rohlfing et al., 2009], this dissertation develops a regression model that captures spatial deformations and intensity changes *simultaneously*. This can be achieved by using a metamorphic regression formulation, which combines the dynamical system formulation for geodesic regression on images [Niethammer et al., 2011] with image metamorphosis [Holm et al., 2009, Miller and Younes, 2001] for the large displacement diffeomorphic metric mapping (LDDMM) registration model [Beg et al., 2005]. The resulting proposed model is called *metamorphic geodesic regression* on image time-series.

1.1.2 Estimation of A Spatiotemporal Statistical Atlas

Atlas-building from population data has become an important task in medical imaging to provide templates for data analysis. Numerous methods for atlas-building exist, ranging from methods designed for cross-sectional, longitudinal, and random design data [Joshi et al., 2004, Fletcher et al., 2009, Hart et al., 2010]. These approaches typically estimate a representative data object (such as a shape, a surface, or an image) for a population, e.g., a mean [Joshi et al., 2004] (or a time-adjusted population mean [Hart et al., 2010]) or a median [Fletcher et al., 2009] with respect to spatial deformations and appearance.

A limitation of all these methods is that they result in a single summary representer and discard much of a population for subsequent analysis. For instance, a single point is used to summarize the entire population on the manifold, when one summarizes it using a mean or a median image or shape. For regression a single curve summarizes a population without carrying forward any information from the local distribution of data around the curve. These are restrictive representations that limit the capability of a model to present confidence bounds, quantile measurements or to identify outliers. In the literature, the limitation of the single summary representers has also been acknowledged. For instance, [Aljabar et al., 2009] suggest a multi-atlas approach to estimate multiple representers of the population. In another study, [Gerber et al., 2010] propose to learn a low-dimensional representation driven entirely by the population of images.

Another strategy to retain population variation is to represent additional aspects of the full data distribution, such as percentiles, the minimum and maximum, variance, confidence regions and outliers as captured by a boxplot for scalar-valued data. The functional boxplot [Sun and Genton, 2011] is an effective tool to represent such statistics for functions. Generalizing the notion of functional boxplots to summarize variabilities within a population of entities such as shapes and images provides a simple and generic method to augment a single representer with additional population information. Besides, as subject data typically has associated individual characteristics (e.g., age, weight, and gender), a method is expected to be able to compute the statistical information parameterized by these characteristics. For example, given a subject at a particular age the goal is to compute subject age-specific confidence regions to assess similarity with respect to the full data population.

Hence, a weighted variant of the functional boxplot is developed in this dissertation to enable the use of kernel regression for estimating a regressed curve with local distributional information. If each data object on the regressed curve, with its additional statistical information, is treated as a statistical atlas, this non-parametric regression model can be regarded as a model to build a spatiotemporal statistical atlas, which is referred to as *statistical atlas construction via weighted functional boxplot*.

1.1.3 Statistics of Group Differences

Apart from capturing variations within a subject or a population, differentiating shape differences between populations is another important topic in the analysis of spatiotemporal data. For example, in the studies of Alzheimer's disease (AD), which accounts for 60% to 70% of cases of dementia [Burns and Iliffe, 2009], researchers are interested in whether brain shapes of normal control subjects are significantly different from those of subjects with disease and where shape differences are located. To answer these questions, analysis approaches have been proposed to assess object properties and are used to characterize shape variations across subjects and between subject populations [Nitzken et al., 2014]. Most of these shape analysis methods are based on the classical point distribution model (PDM) [Cootes et al., 2004], and the PDM requires some form of point-to-point correspondences between shapes to allow precise local shape analysis. However, establishing these correspondences is highly non-trivial and arguably one of the main sources of inaccuracy. Because any misregistration may create artifacts in the final shape analysis results. Recently, shape characterizations have been explored based on concepts of order statistics [Whitaker et al., 2013, Hong et al., 2014a]. These methods utilize depth-ordering of shapes, for example, to generalize the median and the inter-quartile range (IQR) to shapes, effectively obtaining the equivalent of a boxplot for shapes. Using shape-descriptions based on depth-ordering makes it possible to perform shape analysis with very limited (e.g., rigid or affine) spatial alignment of shapes. Furthermore, it can avoid having strong distributional assumptions of the data population, e.g., an assumption of Gaussian distribution. Therefore, to address the problem of differentiating subject populations, a statistical testing method based on depth-ordering on shapes is proposed to detect potential global and local shape differences, which is referred to as *depth-ordering-based shape analysis*.

Like many shape analysis methods, the ordering-based statistical testing does not take into account the temporal dependencies of measurements. Hence, it will be inappropriate for some types of spatiotemporal data, e.g., the longitudinal data. Thanks to the parametric regression models [Niethammer et al., 2011, Fletcher, 2013], the longitudinal data of a subject can be summarized as a smooth path, i.e., a trajectory that is compactly represented by its initial conditions. Based on these regression models, Riemannian approaches for computing averages of trajectories have been proposed [Muralidharan and Fletcher, 2012, Singh et al., 2013a]. In general, statistical methods for longitudinal manifold-valued data focus on the first-order statistics, such as computing the mean, which uses limited information of the data distribution. Capturing higher-order statistics of trajectories themselves would be useful for a more comprehensive description of the underlying distributions and for designing test-statistics going beyond the simple comparison of means, for example testing differences in variances. Hence, an approach that leverages *both the first- and second-order statistics* of shape trajectories for group testing is proposed. It extends the principal geodesic analysis (PGA) [Fletcher et al., 2004] to the tangent bundle of a shape, i.e., the space of trajectories, to estimate both variance and principal directions of shape trajectories. Furthermore, the Bhattacharyya distance [Bhattacharyya, 1946] is generalized to manifold-valued data, which enables assessment of statistical differences between different groups of trajectories. With the generalized Bhattacharyya distance as the test-statistic, a permutation test is performed under the null hypothesis that the two distributions of trajectories are the same. Compared to an existing method, this *hypothesis testing for longitudinal data* shows stronger evidence in distinguishing groups with different distributions of trajectories, especially for cases with similar means but different variances.

1.2 Thesis Statement

Thesis: *Advanced regression models or a time-varying statistical atlas can efficiently capture individual or population changes in spatiotemporal image and shape data. Statistical differences between shape populations can be detected using depth-ordering and statistics on shape trajectories.*

The contributions of this dissertation are:

1. A uniform framework of *parametric regression on the Grassmannian* has been proposed to capture both linear and non-linear changes. It handles regression formulations with different orders, including standard geodesic regression, time-warped geodesic regression, and cubic spline regression.

2. A model of *metamorphic geodesic regression* has been proposed to simultaneously capture spatial deformations and intensity changes in image time-series. It is efficiently solved using a simple, approximate algorithm via pairwise shooting metamorphosis.

3. A *model criticism* for regression models on manifolds, e.g., the Grassmannian manifold, has been proposed to check if model assumptions hold, using kernel two sample tests.

4. A model for building *a spatiotemporal statistical atlas* has been proposed based on kernel regression and a weighted variant of the functional boxplot. It can construct a series of time-varying atlases, augmented by the local distribution of spatiotemporal data, e.g., confidence bounds or outliers. It has been applied to time-varying functions, shapes, and images.

5. A new method of *shape analysis based on depth-ordering* has been proposed to identify global and local shape differences without computing explicit dense correspondences or

making strong distributional assumptions. It has been applied to analyze and compare populations, e.g., normal controls and subjects with disease.

6. A model of *hypothesis testing for longitudinal data* has been proposed by leveraging shape trajectories and their second-order statistics, i.e., variances of trajectory distributions, to identify group differences of shapes.

1.3 Overview of Chapters

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

Chapter 2 provides an overview of the required background in this dissertation, including the necessary background for image and shape analysis and the mathematical background for some manifolds discussed in this dissertation.

Chapter 3 presents parametric regression models on two types of smooth manifolds, i.e., the Grassmann manifold and the manifold of diffeomorphisms.

Chapter 4 presents the model for building a spatiotemporal statistical atlas.

Chapter 5 presents two hypothesis testing approaches to identify shape differences between populations for cross-sectional and longitudinal data respectively.

Chapter 6 concludes the dissertation with a discussion of its contributions and some potential future work.

CHAPTER 2 : BACKGROUD

This chapter presents some necessary background material required in this dissertation. In particular, the mathematical background is briefly reviewed in Section 2.1, including some concepts in smooth manifolds and the Riemannian structure of the Grassmann manifold, as well as how to represent data objects as elements on the Grassmannian. The background for image and shape analysis is presented in Section 2.2. It starts with an overview of fundamental problems in image analysis, e.g., image registration and shape representations. The review of regression models, atlas construction, and statistical hypothesis testing aims to help promote a better understanding of the following chapters of this dissertation.

2.1 Mathematical Background

2.1.1 Smooth Manifolds

In general, smooth manifolds [Lee, 2012] are spaces that locally look like Euclidean space $\mathbb{R}^n$, but globally they may not, e.g., spheres. To be a smooth manifold, a space should first be a *topological manifold*, the most basic type of manifold. In particular, a topological space M is a topological manifold of dimension n or a topological n-manifold, if it has the following properties:

- M is a *Hausdorff space*: for every pair of points $p, q \in M$, there are disjoint open subsets $U, V \subset M$ such that $p \in U$ and $q \in V$.
- M is *countable*: there exits a countable basis for the topology of M.

- M is *locally Euclidean of dimension n*: every point of M has a neighborhood that is homeomorphic to an open subset of $\mathbb{R}^n$.

Apart from the topology, a smooth manifold also needs some extra structure, which allows to define calculus for computation of smooth functions on the manifold. For two open subsets U and V from Euclidean spaces $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively, a function/map $F : U \to V$ is smooth (or C^∞, or *infinitely differentiable*) if each of its component functions has continuous partial derivatives of all orders. Furthermore, if F is bijective and has a smooth inverse map, it is called a *diffeomorphism*. In particular, a diffeomorphism is a homeomorphism. Given two neighborhoods U, V in a manifold M, two homeomorphisms $x : U \to \mathbb{R}^n$ and $y : V \to \mathbb{R}^n$ are C^∞-related if the composite maps $x \circ y^{-1} : y(U \cap V) \to x(U \cap V)$ and $y \circ x^{-1} : x(U \cap V) \to y(U \cap V)$ are C^∞. Here, the pair (x, U) is called a *chart* or *coordinate system*. A collection of charts whose domains cover M is called an *atlas* for M. If any two charts in an atlas are smoothly compatible with each other, this atlas is called a *smooth atlas*. And a *smooth structure* on a manifold M is a maximal smooth atlas on M. The manifold M along with such an atlas is defined as a *smooth manifold* [Lee, 2012].

Tangent spaces. Let M be a smooth manifold, and let p be a point of M, associated with smooth real-valued functions $f : M \to \mathbb{R}$. The set of all derivatives of $C^\infty(M)$ at p is a vector space called the *tangent space* to M at p, which is denoted by T_pM. An element of T_pM is called a *tangent vector* at p [Lee, 2012].

Riemannian geometry. In manifold statistics, the definition of distances on manifolds is related to Riemannian geometry. When measuring the lengths of vectors, a natural way is to compute the inner product of these vectors. Similarly, a *Riemannian metric* or *Riemannian structure* on a smooth manifold M is defined for each point p of M as an inner product $\langle \cdot, \cdot \rangle$

on its tangent space T_pM [Do Carmo, 1992].

Geodesics. Using the Riemannian metric, we can compute the length of a curve on a smooth manifold. In Euclidean space, the length of a straight line connecting two points is the shortest path between them. Accordingly, the shortest smooth curve segment between two points on a manifold is a *geodesic*.

2.1.2 Riemannian Structure of the Grassmannian

The Grassmannian is an example of a smooth manifold with a Riemannian structure. The *Grassmann* manifold $\mathcal{G}(p, n)$ is defined as the set of p-dimensional linear subspaces of $\mathbb{R}^n$, typically represented by an orthonormal matrix $\mathbf{Y} \in \mathbb{R}^{n\times p}$, such that $\mathcal{Y} = \text{span}(\mathbf{Y})$ for $\mathcal{Y} \in \mathcal{G}(p, n)$. It can equivalently be defined as a quotient space within the special orthogonal group $\mathcal{SO}(n)$ as $\mathcal{G}(p, n) := \mathcal{SO}(n)/(\mathcal{SO}(n-p) \times \mathcal{SO}(p))$. The *canonical metric* $g_{\mathcal{Y}} : \mathcal{T}_{\mathcal{Y}}\mathcal{G}(p, n) \times \mathcal{T}_{\mathcal{Y}}\mathcal{G}(p, n) \to \mathbb{R}$ on $\mathcal{G}(p, n)$ is given by

$$g_{\mathcal{Y}}(\mathbf{\Delta}_{\mathcal{Y}}, \mathbf{\Delta}_{\mathcal{Y}}) = \text{tr}\ \mathbf{\Delta}_{\mathcal{Y}}^{\top}\mathbf{\Delta}_{\mathcal{Y}} = \text{tr}\ \mathbf{C}^{\top}(\mathbf{I}_n - \mathbf{Y}\mathbf{Y}^{\top})\mathbf{C}\ , \tag{2.1}$$

where $\mathbf{I}_n$ denotes the $n \times n$ identity matrix, $\mathcal{T}_{\mathcal{Y}}\mathcal{G}(p, n)$ is the tangent space at $\mathcal{Y}$, $\mathbf{\Delta}_{\mathcal{Y}}$ is the tangent vector in $\mathcal{T}_{\mathcal{Y}}\mathcal{G}(p, n)$, and $\mathbf{C} \in \mathbb{R}^{n\times p}$ is arbitrary. For the Grassmann manifold this essentially computes the distance between subspaces [Edelman et al., 1998]. Typically, the principal angles between two subspaces are used to measure their distance. This measure can be understood as an arc length distance. Under this choice of metric, the arc-length of the geodesic connecting two subspaces $\mathcal{Y}, \mathcal{Z} \in \mathcal{G}(p, n)$ is related to the *canonical angles* $\boldsymbol{\phi} = \{\phi_1, \ldots \phi_p\} \in [0, \pi/2]$ between $\mathcal{Y}$ and $\mathcal{Z}$ as $d_g^2(\mathcal{Y}, \mathcal{Z}) = ||\boldsymbol{\phi}||_2^2$. Next, the notation is slightly changed and $d_g^2(\mathbf{Y}, \mathbf{Z})$ is used with $\mathcal{Y} = \text{span}(\mathbf{Y})$ and $\mathcal{Z} = \text{span}(\mathbf{Z})$. In fact, the (squared) geodesic distance can be computed via SVD, i.e., $\mathbf{U}(\cos\mathbf{\Sigma})\mathbf{V}^\top = \mathbf{Y}^\top\mathbf{Z}$ as $d_g^2(\mathbf{Y}, \mathbf{Z}) = ||\operatorname{diag}\mathbf{\Sigma}||^2$, where $\mathbf{\Sigma}$ is diagonal with principal angles ϕ_i, or $\mathbf{U}\mathbf{\Sigma}'\mathbf{V}^\top = \mathbf{Y}^\top\mathbf{Z}$ as $d_g^2(\mathbf{Y}, \mathbf{Z}) = ||\cos^{-1}(\operatorname{diag}\mathbf{\Sigma}')||^2$ (cf. [Begelfor and Werman, 2006]). There are other definitions of the distance between two subspaces, e.g., the chordal 2-norm and Frobenious-norm distances. They are defined by embedding the Grassmann manifold in the vector space $\mathbb{R}^{n\times p}$ [Edelman et al., 1998]. Since the arc length distance is derived from the intrinsic geometry of the Grassmann manifold, it is chosen as the geodesic distance of the Grassmannian in this dissertation.

Now, consider a curve $\gamma : [0, 1] \to \mathcal{G}(p, n), r \mapsto \gamma(r)$ such that $\gamma(0) = \mathcal{Y}_0$ and $\gamma(1) = \mathcal{Y}_1$, with $\mathcal{Y}_0$ represented by $\mathbf{Y}_0$ and $\mathcal{Y}_1$ represented by $\mathbf{Y}_1$. The *geodesic equation* for such a curve, given that $\dot{\mathbf{Y}} = {}^{d}\!/_{dr}\mathbf{Y}(r) \doteq (\mathbf{I}_n - \mathbf{Y}\mathbf{Y}^\top)\mathbf{C}$, on $\mathcal{G}(p, n)$ is given by

$$\ddot{\mathbf{Y}}(r) + \mathbf{Y}(r)[\dot{\mathbf{Y}}(r)^\top\dot{\mathbf{Y}}(r)] = \mathbf{0} \ , \tag{2.2}$$

which also defines the Riemannian exponential map on the Grassmannian as an ODE (ordinary differential equation) for convenient numerical computations. This can be derived by solving an optimization problem from the calculus of variations. It minimizes the distance between two points on the Grassmann manifold, i.e., Eq.(2.1), under the constraints of the above definition for a tangent vector and keeping to be an orthonormal matrix along the geodesic, i.e., $\mathbf{Y}^\top\mathbf{Y} = \mathbf{I}$. Integrating Eq. (2.2), starting with initial conditions we can shoot the geodesic forward in time.

The following parts present the differential geometric “tools” on the Grassmann manifold. In particular, they include formulae/derivations for the Riemannian *exponential map (Exp-map)*, the *inverse exponential map (a.k.a. Log-map)* and *parallel transport* on $\mathcal{G}(p, n)$.

Exponential map. The exponential map (Exp-map) maps a point $\mathcal{Y} = \text{span}(\mathbf{Y})$ with a direction $\mathbf{D}$ in the tangent space $\mathcal{T}_{\mathcal{Y}}\mathcal{G}(p, n)$ to a point $\mathcal{Z} = \text{span}(\mathbf{Z})$ on the manifold $\mathcal{G}(p, n)$, *i.e.*,

$$\text{Exp}_{\mathbf{Y}}(t\mathbf{D}) = \mathbf{Z}, \quad t \in [0, 1]$$

along the geodesic that connects $\mathcal{Y}$ and $\mathcal{Z}$. By letting $\mathbf{D} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ denote the compact SVD of $\mathbf{D}$, the Exp-map on $\mathcal{G}(p, n)$, in terms of representers $\mathbf{Y}$ and $\mathbf{Z}$, can be written as (cf. [Begelfor and Werman, 2006])

$$\mathbf{Z} = \mathbf{Y}\mathbf{V}\cos(t\mathbf{\Sigma})\mathbf{V}^\top + \mathbf{U}\sin(t\mathbf{\Sigma})\mathbf{V}^\top \ . \tag{2.3}$$

Here, because $\mathbf{\Sigma}$ is a diagonal matrix, $cos(t\mathbf{\Sigma})$ and $sin(t\mathbf{\Sigma})$ are also diagonal matrices. They can be computed by simply taking the sines or cosines on the matrices' diagonal components. The proof of this formulation can be found in [Edelman et al., 1998].

Inverse exponential map. The inverse exponential map (Log-map) computes the mapping from a neighborhood $U \subset \mathcal{G}(p, n)$ of $\mathcal{Y}$ to $\mathcal{T}_{\mathcal{Y}}\mathcal{G}(p, n)$. In terms of representers $\mathbf{Y}, \mathbf{Z}$ for the subspaces $\mathcal{Y} = \text{span}(\mathbf{Y}), \mathcal{Z} = \text{span}(\mathbf{Z})$, the Log-map can be written as $\mathbf{H} = \text{Log}_{\mathbf{Y}}(\mathbf{Z})$. In other words, $\mathbf{H}$ is the direction matrix which allows starting at $\mathcal{Y}$ and walking along the geodesic in the direction of $\mathbf{H}$ to reach $\mathcal{Z}$ in unit time ($t = 1$), *i.e.*, $\mathbf{Z} = \text{Exp}_{\mathbf{Y}}(\mathbf{H})$. Let $\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$. Multiplying Eq. (2.3) with $\mathbf{Y}^\top$ on the left-hand side (and $t = 1$) results in

$$\mathbf{Y}^\top\mathbf{Z} = \underbrace{\mathbf{Y}^\top\mathbf{Y}}_{=\mathbf{I}_p}\mathbf{V}\cos(\mathbf{\Sigma})\mathbf{V}^\top + \underbrace{\mathbf{Y}^\top\mathbf{U}}_{=\mathbf{0}}\sin(\mathbf{\Sigma})\mathbf{V}^\top = \mathbf{V}\cos(\mathbf{\Sigma})\mathbf{V}^\top \ .$$

Furthermore, replacing $\mathbf{V}\cos(\mathbf{\Sigma})\mathbf{V}^\top$ with $\mathbf{Y}^\top\mathbf{Z}$ in Eq. (2.3) yields $\mathbf{U}\sin(\mathbf{\Sigma})\mathbf{V}^\top = \mathbf{Z} -$

$\mathbf{Y}\mathbf{Y}^\top\mathbf{Z}$. Thus

$$\mathbf{U}\sin(\mathbf{\Sigma})\mathbf{V}^\top(\mathbf{V}\cos(\mathbf{\Sigma})\mathbf{V}^\top)^{-1} = (\mathbf{Z} - \mathbf{Y}\mathbf{Y}^\top\mathbf{Z})(\mathbf{Y}^\top\mathbf{Z})^{-1},$$

which – upon noting that (1) $(\mathbf{V}\cos(\mathbf{\Sigma})\mathbf{V}^\top)^{-1} = \mathbf{V}\cos(\mathbf{\Sigma})^{-1}\mathbf{V}^\top$ and (2) $\mathbf{V}^\top\mathbf{V} = \mathbf{I}_p$ – reduces to

$$\mathbf{U}\tan(\mathbf{\Sigma})\mathbf{V}^\top = (\mathbf{Z} - \mathbf{Y}\mathbf{Y}^\top\mathbf{Z})(\mathbf{Y}^\top\mathbf{Z})^{-1} \ .$$

This yields $\mathbf{H}$ via the SVD of $(\mathbf{Z} - \mathbf{Y}\mathbf{Y}^\top\mathbf{Z})(\mathbf{Y}^\top\mathbf{Z})^{-1}$ (or $\mathbf{Z}(\mathbf{Y}^\top\mathbf{Z})^{-1} - \mathbf{Y}$) as

$$\mathbf{H} = \mathrm{Log}_{\mathbf{Y}}(\mathbf{Z}) = \mathbf{U}\arctan(\mathbf{\Sigma})\mathbf{V}^\top \ .$$

Parallel transport. Given two subspaces $\mathcal{X}, \mathcal{Y}$, represented via $\mathbf{X}, \mathbf{Y}$ and a direction matrix $\mathbf{H} \in \mathcal{T}_\mathcal{X}\mathcal{G}(p,n)$ such that $\mathbf{Y} = \mathrm{Exp}_\mathbf{X}(\mathbf{H})$, the objective is to transport an arbitrary tangent vector $\mathbf{\Delta}$ at $\mathcal{T}_\mathcal{X}\mathcal{G}(p,n)$ to $\mathcal{T}_\mathcal{Y}\mathcal{G}(p,n)$ along the geodesic connecting $\mathcal{X}$ and $\mathcal{Y}$. Letting $\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, parallel transport (denoted by τ) can be computed via [Edelman et al., 1998]

$$\tau\mathbf{\Delta}(t) = -\mathbf{Y}\mathbf{V}\sin(t\mathbf{\Sigma})\mathbf{U}^\top\mathbf{\Delta} + \mathbf{U}\cos(t\mathbf{\Sigma})\mathbf{U}^\top + (\mathbf{I}_n - \mathbf{U}\mathbf{U}^\top)\mathbf{\Delta}, \quad t \in [0,1] \ . \tag{2.4}$$

2.1.3 Representation on the Grassmannian

This dissertation describes two types of data that can be represented on $\mathcal{G}(p,n)$: linear dynamical systems (LDS) and shapes (see shape representation in Section 2.2.2).

Linear dynamical systems. In the computer vision literature, *dynamic texture* models [Doretto et al., 2003] are commonly applied to model videos as realizations of linear dynamical systems (LDS). For a video, represented by a collection of vectorized frames $\mathbf{y}_1, \ldots, \mathbf{y}_\tau$

with $\mathbf{y}_i \in \mathbb{R}^n$, the standard dynamic texture model with p states has the form

$$\begin{aligned} \mathbf{x}_{k+1} &= \mathbf{A}\mathbf{x}_k + \mathbf{w}_k, & \mathbf{w}_k &\sim \mathcal{N}(0, \mathbf{W}), \\ \mathbf{y}_k &= \mathbf{C}\mathbf{x}_k + \mathbf{v}_k, & \mathbf{v}_k &\sim \mathcal{N}(0, \mathbf{R}) \ , \end{aligned} \tag{2.5}$$

with $\mathbf{x}_k \in \mathbb{R}^p, \mathbf{A} \in \mathbb{R}^{p \times p}$, and $\mathbf{C} \in \mathbb{R}^{n \times p}$. When relying on the prevalent system identification of [Doretto et al., 2003], the matrix $\mathbf{C}$ is, by design, of (full) rank p (*i.e.*, the number of states) and by construction an *observable* system [Kalman, 1959] is obtained, where a full rank *observability* matrix $\mathbf{O} \in \mathbb{R}^{np \times p}$ is defined as $\mathbf{O} = [\mathbf{C}\ (\mathbf{CA})\ (\mathbf{CA}^2)\ \cdots\ (\mathbf{CA}^{p-1})]^\top$. This system identification using the dynamic texture model is not unique. Because systems $(\mathbf{A}, \mathbf{C})$ and $(\mathbf{TAT}^{-1}, \mathbf{CT}^{-1})$ with $\mathbf{T} \in \mathcal{GL}(p)$ have the same transfer function, i.e., with the same input they have the same output. Hence, the realization subspace spanned by $\mathbf{O}$ is a point on the Grassmannian and the observability matrix is a representer of this subspace. An LDS model is identified for a video by its $np \times p$ orthonormalized observability matrix.

To support a non-uniform weighting of samples during system identification, a *temporally* localized variant of [Doretto et al., 2003] is proposed in this dissertation. This is beneficial in a situation where a considerable number of frames are needed for stable system identification, yet not all samples should contribute equally to the LDS parameter estimates. Specifically, given the measurement matrix $\mathbf{M} = [\mathbf{y}_1, \cdots, \mathbf{y}_\tau]$ and a set of weights $\mathbf{w} = [w_1, \cdots, w_\tau]$ such that $\sum_i w_i = \tau$, a weighted SVD of $\mathbf{M}$ is computed as

$$\mathbf{U\Sigma V}^\top = \mathbf{M}\ \mathrm{diag}(\sqrt{\mathbf{w}}) \ . \tag{2.6}$$

Then, as in [Doretto et al., 2003], $\mathbf{C} = \mathbf{U}$ and $\mathbf{X} = \mathbf{\Sigma V}^\top$. Once the state matrix $\mathbf{X}$ has been determined, $\mathbf{A}$ can be computed as $\mathbf{A} = \mathbf{X}_2^{\tau}\mathbf{W}^{\frac{1}{2}}(\mathbf{X}_1^{\tau-1}\mathbf{W}^{\frac{1}{2}})^{\dagger}$, where † denotes the pseudoinverse, $\mathbf{X}_2^{\tau} = [\mathbf{x}_2, \cdots, \mathbf{x}_{\tau}]$, $\mathbf{X}_1^{\tau-1} = [\mathbf{x}_1, \cdots, \mathbf{x}_{\tau-1}]$ and $\mathbf{W}^{\frac{1}{2}}$ is a diagonal matrix with $W_{ii}^{\frac{1}{2}} = [\frac{1}{2}(w_i + w_{i+1})]^{1/2}$.

Shapes. Let a shape be represented by a collection of m points. A *shape matrix* is constructed from its m points as $\mathbf{L} = [(x_1, y_1, ...); (x_2, y_2, ...); \ldots; (x_m, y_m, ...)]$. By applying SVD on this matrix, *i.e.*, $\mathbf{L} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{\top}$, an affine-invariant shape representation is obtained by using the left-singular vectors $\mathbf{U}$ [Begelfor and Werman, 2006, Sepiashvili et al., 2003]. This establishes a mapping from the shape matrix to a point on the Grassmannian (with $\mathbf{U}$ as the representative). Such a representation has been used for facial aging regression, for instance [Turaga et al., 2010].

2.2 Image and Shape Analysis Background

2.2.1 Image Registration

Image registration is one of the fundamental research topics in image analysis. It is the process of placing different images into geometrical and/or anatomical agreement. Typically, images of the same scene are taken at different times, from different viewpoints, and/or different modalities. The regression problem presented in Chapter 3 can be reduced to an image registration problem if there are only two images. Hence, some necessary knowledge about image registration is presented here before the discussion of regression models. A detailed review of image registration in computer vision and medical image analysis can be found in [Zitova and Flusser, 2003] and [Oliveira and Tavares, 2014], respectively.

The high-level idea of image registration is to estimate a geometric transformation, which is applied to the moving image and optimizes its similarity with respect to the reference image, that is, the cost function. In addition, the cost function usually includes another term, a regularization term, to enforce smoothness of the geometric transformation. In general, there are three components needed for formulating an image registration problem: the geometric transformation, the similarity measure, and the regularization term discussed in the following.

Usually, the geometric transformations can be divided into two categories: rigid and non-rigid transformations. Rigid transformations are the simplest cases, typically estimating the parameters of translation and rotation. Non-rigid transformations include the similarity transform, affine, projective, and curved (elastic or fluid) transformations. In this dissertation, the image regression is built upon the large deformation diffeomorphic metric mapping (LDDMM) model [Beg et al., 2005], which is a representative of the fluid-based registration.

The similarity measures are mostly divided into two types of methods: intensity and feature-based approaches. The commonly-used measures in the category of intensity-based similarity include the sum of squared differences (SSD) and the mutual information (MI). The underlying assumption of SSD is that the structures of interest in both images should have identical intensities. Hence, a lower SSD value indicates a better registration result. In contrast, MI captures how well one image explains the other. Hence, for this measurement a higher MI value indicates a better registration result. On the other hand, in the category of the feature-based similarity measures, the similarity measures focus on computing the differences of structures extracted from images. The feature could be smaller image patches or volumes compared via their intensity differences or corresponding points compared via their Euclidean distances.

The last component in the formulation of image registration is the regularization term.

A commonly-used one is related to the second-order derivations of the transformation or the Jacobian of the transformation. In the following the LDDMM will be treated as a concrete example to discuss the non-rigid image registration, in particular the fluid flow registration.

Fluid flow registration and LDDMM [Beg et al., 2005]. As discussed before, the objective of image registration is to deform a source image I_0 (or the moving image) to match a target image I_1 (or the static/template image), as accurate as possible. In fluid flow registration, the deformation (i.e., the transformation used in the previous part) is parametrized through a spatiotemporal velocity field v. It "flows" the image I_0 to another image I_1 under a transport equation

$$I_t + \nabla I^\top v = 0, \quad I(0) = I_0 \ . \tag{2.7}$$

To ensure the smoothness of the velocity filed, it is meaningful to combine the intensity-based similarity term with a regularity term, resulting in a cost function of the form

$$E(I, v) \quad = \quad \underbrace{\frac{1}{2}\int_0^1 \|v\|_L^2 \, dt}_{\text{Regularity}} \quad + \quad \frac{1}{\sigma^2}\underbrace{\|I(1) - I_1\|^2}_{\text{Similarity}}, \tag{2.8}$$

where $\|v\|_L^2 = \langle Lv, Lv\rangle$ and L is a differential operator to encourage smoothness of the velocity field. Typically, $L = -\alpha\nabla^2 + \gamma$, and α and γ are constants. This cost function should be minimized with respect to the dynamic constraint in Eq. (2.7).

To solve this optimization problem, the dynamic constraint is added as an equality constraint through a Lagrangian multiplier λ, which has the same size with the image I, i.e.,

$$E(I, v, \lambda) = \frac{1}{2}\int_0^1 \|v\|_L^2 \, dt + \frac{1}{\sigma^2}\|I(1) - I_1\|^2 + \langle \lambda, I_t + \nabla^\top v\rangle \, dt, \tag{2.9}$$

then optimality conditions can be obtained through variational calculus, resulting in

$$\begin{cases} I_t + \nabla I^\top v & = 0, \quad I(0) = I_0, \\ -\lambda_t - div(\lambda v) & = 0, \quad \lambda(1) = \frac{2}{\sigma^2}(I_1 - I(1)), \\ v + (L^\dagger L)^{-1}\lambda \nabla I & = 0 \ . \end{cases} \tag{2.10}$$

Here, $L^\dagger$ denotes the adjoint of L. There are two commonly-used optimization methods for LDDMM to obtain a numerical solution, the relaxation optimization and the shooting optimization.

In relaxation optimization the velocity field at every time point is updated in each iteration. It proceeds as follows

- Given an estimation for the velocity field v, flow the intensity I forward according to the first equation in Eqs. (2.10).
- Compute the solution for the adjoint λ backward in time using the second equation in Eqs. (2.10).
- Compute the gradient for the velocity v at every point in time using the equation $\nabla_v E = v + (L^\dagger L)^{-1}(\lambda \nabla I)$.
- Update the estimate for the velocity v using a gradient descent step.
- Repeat the previous steps until convergence.

The relaxation strategy needs to store the intensity I, its adjoint λ, and the velocity field v at every discretized time. This is memory consuming. In the shooting strategy [Ashburner and Friston, 2011, Vialard et al., 2012], only the variables at time 0, i.e., $\lambda(0)$

and $v(0)$ are of the interest. Actually, given $\lambda(0)$ (the so-called initial momentum) we can compute the initial velocity $v(0)$ using the image I_0 and the third equation in Eqs. (2.10). This can save memory resources during the computation. In particular, the shooting optimization is derived according to the fact that the energy is conserved along the geodesic. Then, the regularity term can be rewritten using initial conditions only, i.e., the initial image and the initial velocity. From now on, the adjoint λ will be replaced with the notation p to represent the momentum, because new Lagrangian multipliers will be introduced in the shooting formulation. Since the velocity field $v = -(L^{\dagger}L)^{-1}p\nabla I$, we can derive

$$\|v\|_L^2 = \langle Lv, Lv\rangle = \langle v, L^{\dagger}Lv\rangle = \langle (L^{\dagger}L)^{-1}p\nabla I, p\nabla I\rangle \ . \tag{2.11}$$

With $K = (L^{\dagger}L)^{-1}$, we obtain the new cost function in form of

$$E(I(0), p(0)) = \frac{1}{2}\langle p(0)\nabla I(0), K * p(0)\nabla I(0)\rangle + \frac{1}{\sigma^2}\|I(1) - I_1\|^2 \ . \tag{2.12}$$

This cost function is minimized under the constraints of the relaxation optimality conditions:

$$\begin{cases} I_t + \nabla I^{\top} v &= 0, \quad I(0) = I_0, \\ p_t + div(pv) &= 0, \\ v + K * p\nabla I &= 0, \end{cases} \tag{2.13}$$

which are also called the shooting equations. Given an initial momentum, these equations are used to shoot an initial image forward to a target image. As the initial momentum is unknown, we solve it by adding Eqs. (2.13) through Lagrangian multipliers, λ^I, λ^p, and λ^v;

then we compute the adjoint model through variational calculus. As a result, we obtain

$$\begin{cases} \lambda^I_t + div(\lambda^I v + pK * \lambda^v) & = 0, \lambda^I(1) = \frac{2}{\sigma^2}(I_1 - I(1)), \\ \lambda^p_t + (\nabla\lambda^p)^\top v - \nabla I^\top K * \lambda^v & = 0, \lambda^p(1) = 0, \\ \lambda^v + \lambda^I \nabla I - \nabla\lambda^p p & = 0\ \ . \end{cases} \tag{2.14}$$

According to Eq (2.14), we can shoot backward to update the initial momentum using

$$\nabla_{p(t_0)} E = -\lambda^p(0) + \nabla I_0^\top K * (p(0)\nabla I(0)), \tag{2.15}$$

which is the gradient obtained through the variational calculus for the Lagrangian function. Note that, the initial image is not required to be updated, because it is fixed. Besides, solving transport equations can be non-trivial for non-smooth functions. In practice, instead of solving $I_t + \nabla I^\top v = 0$ we solve

$$\Phi_t + D\Phi v = 0, \quad \Phi(0) = id, \tag{2.16}$$

where Φ is the deformation map for the coordinate mapping and id is the identity map. Then, the intensity I at each time point can be estimated by $I(t) = I_0 \circ \Phi(t)$. Similarly, given a backward solution to

$$-\Phi^b_t - D\Phi^b v = 0, \quad \Phi^b(1) = id, \tag{2.17}$$

the corresponding adjoint can be computed as $\lambda(t) = |D\Phi^b(t)|\lambda(1) \circ \Phi^b(t)$. Because the velocity field is regularized by the L operator, this map-based approach is expected to produce smooth maps, which are much easier to propagate numerically.

2.2.2 Shape Representation

In [Dryden and Mardia, 1998], shape is defined as *all the geometrical information that remains when location, scale, and rotational effects are filtered out from an object.* This definition has been generalized to mean "a shape is the spatial information after an equivalence group of geometric transformations are filtered out, for example, the group of similarity transformations". There are different shape representations. In this dissertation the shape is represented in two ways: 1) point-based representation, i.e., describing a shape with a finite number of points on its boundary; and 2) binary representation with 1 indicating the inside of a shape and 0 outside. Besides, other shape representations exist, e.g., the surface-based representations [Dale et al., 1999], and the medial representations (e.g., m-reps [Pizer et al., 2003]).

In particular, a point-based shape can be represented as an element on the Grassmannian, as discussed in Section 2.1.3. This representation is designed for shapes after removing the effects of translation, rotation, and non-uniform scaling. In some other scenarios where uniform scaling is considered instead of non-uniform scaling, the Kendall shape space [Kendall, 1984] is one possible choice for the shape representation. For example, in Chapter 5, geodesic regression on Kendall shape space from [Fletcher, 2013] will be used to summarize a shape trajectory for each subject with longitudinal shape data. Therefore, some basic concepts in the Kendall shape space is presented here, including the required computation of Exponential/Log maps for geodesic regression. For more detailed explanations about the Kendall shape space, please refer to [Kendall, 1984] and [Fletcher, 2013].

Kendall Shape Space. A collection of 2D points can be considered as a complex k-vector,

$z \in \mathbb{C}^k$. After removing translation, rotation and scaling, a shape can be treated as a point in the complex projective space $\mathbb{CP}^{k-2}$. Given a centered shape x and the initial velocity v, which is a tangent vector at x, such that $\langle x, v \rangle = 0$, the exponential map is given by

$$\mathrm{Exp}_x(v) = \cos\theta \cdot x + \frac{\|x\| \sin\theta}{\theta} \cdot v, \quad \theta = \|v\| \ . \tag{2.18}$$

The Log-map computes the initial velocity between two shapes x and y and is given by

$$\mathrm{Log}_x(y) = \frac{\theta \cdot (y - \pi_x(y))}{\|y - \pi_x(y)\|}, \quad \theta = \arccos \frac{|\langle x, y \rangle|}{\|x\| \|y\|}, \tag{2.19}$$

where $\pi_x(y) = x \cdot \langle x, y \rangle / \|x\|^2$ denotes the projection of the vector y onto x.

2.2.3 Regression Models

In statistics, regression analysis is a powerful statistical tool to estimate the relationships among variables. At the coarsest level, we distinguish between two categories of regression approaches: *parametric* and *non-parametric* strategies, with trade-offs on both sides [Moussa and Cheema, 1998]. In computer vision and medical image analysis non-parametric regression on *nonflat* manifolds has gained considerable attention over the last years. Strategies range from kernel regression [Davis et al., 2007] on the manifold of diffeomorphic transformations to gradient-descent [Samir et al., 2012] approaches on manifolds commonly encountered in computer vision, such as the group of rotations $\mathcal{SO}(3)$ or Kendall's shape space. In other works, discretizations of the curve fitting problem have been explored [Boumal and Absil, 2011a, Boumal and Absil, 2011b, Su et al., 2012] which, in some cases, even allow employing second-order optimization methods [Boumal, 2013].

Regression models presented in Chapter 3 are representatives of the *parametric* category, in particular, on smooth manifolds. Although differential geometric concepts, e.g., geodesics and intrinsic higher-order curves, have been well-studied in the literature [Noakes et al., 1989, Camarinha et al., 1995, Crouch and Leite, 1995, Machado et al., 2010], their use for parametric regression, *i.e.*, finding parametric relationships between the manifold-valued variable and an independent scalar-valued variable, has only recently gained interest. A variety of methods extending concepts of regression in Euclidean space to nonflat manifolds have been proposed. [Rentmeesters, 2011, Fletcher, 2013] and [Hinkle et al., 2014] address the problem of geodesic fitting on Riemannian manifolds. They primarily focus on symmetric spaces, to which the Grassmann manifold belongs. [Batzies et al., 2015] study a theoretical characterization of fitting geodesics on the Grassmannian. And [Niethammer et al., 2011] generalize linear regression to the manifold of diffeomorphisms to model image time-series data, followed by work extending this concept [Hong et al., 2012a, Singh et al., 2013b] and enabling the use of higher-order models [Singh and Niethammer, 2014].

From a conceptual point of view, there are two groups of strategies to solve parametric regression problems on manifolds: first, *geodesic shooting* using adjoint methods from an optimal-control point of view [Niethammer et al., 2011, Hong et al., 2012a, Singh et al., 2013b, Singh and Niethammer, 2014, Hinkle et al., 2014]; second, strategies that leverage *Jacobi fields* to compute the required gradients [Rentmeesters, 2011, Fletcher, 2013]. The approaches in this dissertation are representatives of the first category. Unlike Jacobi field approaches, the presented approaches in the following chapter do not require computation of the curvature explicitly and easily extend to higher-order models, such as the cubic splines extension.

Regression on the Grassmann manifold. In the context of computer-vision problems, [Lui, 2012] recently adapted the known Euclidean least-squares solution to the Grassmannian. While this strategy works remarkably well for the presented gesture recognition tasks, the formulation does not guarantee the minimization of the sum-of-squared geodesic distances within the manifold, which would be the natural extension of least-squares to Riemannian manifolds. Hence, the geometric and variational interpretation of [Lui, 2012] remains unclear. In contrast, this dissertation addresses the problem from the aforementioned energy-minimization point of view which guarantees, by design, consistency with the geometry of the manifold.

To the best of my knowledge, the most related work to the regression models in this dissertation is [Rentmeesters, 2011, Fletcher, 2013] and [Batzies et al., 2015] in the context of fitting *geodesics*, as well as [Machado et al., 2010] (and to some extent [Hinkle et al., 2014]) in the context of fitting *higher-order curves*.

[Batzies et al., 2015] present a theoretical study of fitting geodesics (*i.e.*, first-order curves) on the Grassmannian and derive a set of optimality criteria. However, the work is purely theoretical and, as mentioned in [Batzies et al., 2015, Sect. 1], the objective is *not* to provide a numerical solution scheme. [Rentmeesters, 2011] and [Fletcher, 2013] propose optimization based on Jacobi fields to fit geodesics on general Riemannian manifolds. Contrary to the regression approaches presented in this dissertation, it does not follow trivially how to generalize [Rentmeesters, 2011] (or [Fletcher, 2013]) to higher-order models.

[Machado et al., 2010] specifically aim to address the problem of fitting higher-order curves on Riemannian manifolds. Based on earlier work presented in [Noakes et al., 1989, Camarinha et al., 1995, Crouch and Leite, 1995], they introduce a different variational formulation of the problem and derive optimality criteria from a theoretical point of view. From a practical perspective, it remains unclear (as with [Batzies et al., 2015] in case of geodesics) how these optimality criteria translate into a numerical optimization scheme. In other work, [Hinkle et al., 2014] address the problem of fitting polynomials but mostly focus on manifolds with a Lie group structure[1]. In that case, adjoint optimization is greatly simplified. For a general case curvature computations are required and can be tedious.

In comparison to prior work, in this dissertation I derive *alternative* optimality criteria for geodesics and cubic splines using principles from optimal-control. These conditions not only form the basis for the shooting approach but also naturally lead to convenient iterative algorithms. By construction, the obtained solutions are guaranteed to be the sought-for curves (*i.e.*, geodesics, splines) on the manifold. In addition, the derived formulation for cubic splines does *not* require computation of the Riemannian curvature tensor.

2.2.4 Atlas Construction

The problem of atlas construction is to build a template for a collection of data points, e.g., images or shapes. The commonly used strategy is to estimate an average representation of a data population, i.e., the mean for the population. For a collection of data samples $\{\mathbf{x}_i\}_{i=1}^N$ in a general metric space $\mathbf{M}$, the intrinsic mean (i.e., the Fréchet mean) is defined as the minimizer of the sum-of-squared distances to each of the data points, that is

$$\mu = \operatorname{argmin}_{\mathbf{x}\in\mathbf{M}} \sum_{i=1}^{N} d(\mathbf{x}, \mathbf{x}_i)^2, \tag{2.20}$$

[1] $\mathcal{G}(p, n)$ does not possess such a group structure.

where d denotes the distance metric on $\mathbf{M}$, i.e., $d : \mathbf{M} \times \mathbf{M} \to \mathbf{R}$. Typically, in the space of images or shapes this distance metric is the geodesic distance. For example, [Joshi et al., 2004] propose a diffeomorphic atlas construction for a collection of images. If we replace the squared distance with the absolute distance in Eq. (2.20), the resulting minimizer is the median of the data population [Fletcher et al., 2009]. If time information is further considered during the atlas construction for three dimensional data objects, a longitudinal/4D atlas [Kuklisova-Murgasova et al., 2011] can be obtained.

2.2.5 Statistical Hypothesis Testing

In data analysis, statistical hypothesis testing is a component that typically involves tests of the relationship among data samples drawn from different populations. Given a hypothesis of the statistical relationship between two data sets, hypothesis testing aims to determine the probability that the hypothesis is rejected. In particular, we first formulate the null hypothesis H_0, i.e., a general statement or default position that there is no relationship between two data populations or no differences among groups; and the alternative hypothesis H_1, which is contrary to the null hypothesis. A *type I error* is the incorrect rejection of a true null hypothesis, i.e., a false positive; and a *type II error* is the failure to reject a false null hypothesis, i.e., a false negative. In a hypothesis test, we need a test statistic to determine when to reject the null hypothesis. Typically, it measures some attribute of a data sample and compares to the distribution of the chosen attribute, resulting in a p-value to reject or accept a null hypothesis at some significance level.

Permutation test. The method of *permutation* (or called *randomization*) is a general approach in statistical hypothesis testing. The basic idea of a permutation test is comparing the test statistic without permuting the labels of data samples to the distribution of the test statistic with many permutations under the null hypothesis. For example, there are two groups with n and m subjects, respectively, i.e., $\{X_i\}_{i=1}^{n}$, and $\{Y_i\}_{i=1}^{m}$. To compute the difference between the two groups, we can measure their mean difference, i.e., $T_0 = \bar{X} - \bar{Y}$. This is the test statistic. The null hypothesis is that there is no difference between these two groups. Under this null hypothesis, we can permute the data in these two groups to estimate the distribution of the test statistic. This can be achieved by computing the mean difference $T = \bar{X} - \bar{Y}$ as many times as possible, e.g., repeating all the combinations, $\binom{N}{n}$ where $N = n + m$. Assume the mean of group X is expected to be smaller than the group Y. Because all the permutations are equally likely under the null hypothesis H_0, the p-value is

$$p = P(T \leq T_0 | H_0) = \frac{\sum_{i=1}^{\binom{N}{n}} I(T_i \leq T_0)}{\binom{N}{n}}, \tag{2.21}$$

where T_i is the value of the test statistic at the ith randomization and $I(\cdot)$ is the indicator function. For more details, please refer to [Ernst et al., 2004].

CHAPTER 3 : ESTIMATION OF TIME-VARYING CHANGES

This chapter[1] presents a general framework with an extensible set of regression formulations, including standard geodesic regression, time-warped geodesic regression, and cubic-splines on Riemannian manifolds, e.g., the Grassmann manifold, and the manifold of diffeomorphisms. This regression framework addresses the problem of fitting parametric curves on Riemannian manifolds for the purpose of intrinsic parametric regression. It starts from the energy minimization formulations of classical regression models, e.g., linear least-squares and cubic splines in Euclidean space, and then generalizes these concepts to general non-flat Riemannian manifolds, following an optimal-control point of view. This idea is specialized to the Grassmann manifold and it yields a simple, extensible, and easy-to-implement solution to the parametric regression problem. The idea is also extended to the manifold of diffeomorphisms to jointly capture spatial and intensity changes in image time-series.

In particular, Section 3.1 revisits regression models in Euclidean space from the optimal-control point of view. In Section 3.2 these classical models are generalized to general Riemannian manifolds. Section 3.3 presents the algorithms for computing parametric regression on the Grassmann manifold. It demonstrates the utility of the proposed solution on different vision problems, such as shape regression as a function of age, traffic-speed estimation and crowd-counting from surveillance video clips. Most notably, these problems can be con-

[1] The work presented in this chapter is based on previous papers [Hong et al., 2012a, Hong et al., 2012b, Hong et al., 2014c, Hong et al., 2014d, Hong et al., 2015b, Hong et al., 2016].

veniently solved within the same framework without any specifically-tailored steps along the processing pipeline. Furthermore, Section 3.4 presents the geodesic regression on image time-series to simultaneously capture spatial deformations and appearance changes. Finally, in Section 3.5 a model criticism approach is presented to evaluate regression models on manifolds, in particular for the Grassmann manifold.

3.1 Regression in $\mathbb{R}^n$ via Optimal-Control

This section starts with a review of linear regression in $\mathbb{R}^n$ and discusses its solution via optimal-control. While regression is a well studied statistical technique and several solutions exist for univariate and multivariate models [Freedman, 2009], the presented optimal-control perspective not only allows easily generalizing regression to manifolds but also defining more complex models on these manifolds.

3.1.1 Linear Regression

A straight line in $\mathbb{R}^n$ can be defined as an acceleration-free curve with parameter t, represented by states, $(x_1(t), x_2(t))$, such that $\dot{x}_1 = x_2,$ and $\dot{x}_2 = 0$, where $x_1(t) \in \mathbb{R}^n$ is the *position* of a particle at time t and $x_2(t) \in \mathbb{R}^n$ represents its *velocity* at t. Let $\{y_i\}_{i=0}^{N-1} \in \mathbb{R}^n$ denote a collection of N measurements at time instances $\{t_i\}_{i=0}^{N-1}$ with $t_i \in [0, 1]$. The linear regression problem is defined as that of estimating a parametrized linear motion of the particle $x_1(t)$, such that the path of its trajectory best fits the measurements in the least-squares sense. The constrained optimization problem, from an optimal-control perspective, is

$$\min_{\mathbf{\Theta}} \quad E(\mathbf{\Theta}) = \underbrace{\sum_{i=0}^{N-1} \|x_1(t_i) - y_i\|^2,}_{\text{Data-matching}} \quad s.t. \quad \underbrace{\dot{x}_1 - x_2 = 0, \quad and \quad \dot{x}_2 = 0}_{\text{“Dynamics” constraints}} \;, \tag{3.1}$$

with $\mathbf{\Theta} = \{x_i(0)\}_{i=1}^2$, *i.e.*, the *initial conditions.* Adding the dynamics constraints through time-dependent Lagrangian multipliers, $\mathbf{\Lambda} = \{\lambda_1, \lambda_2\} \in \mathbb{R}^n$, results in

$$E(\mathbf{\Theta}, \mathbf{\Lambda}) = \quad \sum_{i=0}^{N-1} \|x_1(t_i) - y_i\|^2 \quad + \quad \int_0^1 \lambda_1^\top(\dot{x}_1 - x_2) + \lambda_2^\top(\dot{x}_2)\ dt\ . \tag{3.2}$$

For readability the argument t has been omitted for $\lambda_1(t)$ and $\lambda_2(t)$. They are referred to as *adjoint* variables, enforcing the dynamical "straight-line" constraints. Equation (3.2) is the Lagrangian associated to the original constrained optimization problem. An optimal solution is a saddle point of the Lagrangian, and the Karush-Kuhn-Tucker (KKT) conditions are necessary conditions for an optimum [Boyd and Vandenberghe, 2004]. In particular, evaluating the gradients with respect to the state variables results in the *adjoint system* as $\dot{\lambda}_1 = 0,$ and $-\dot{\lambda}_2 = \lambda_1$, with jumps in λ_1 as $\lambda_1(t_i^+) - \lambda_1(t_i^-) = 2(x_1(t_i) - y_i)$, at measurements t_i. The optimality conditions on the gradients also result in the boundary conditions $\lambda_1(1) = 0$ and $\lambda_2(1) = 0$. Finally, the gradients with respect to the initial conditions are $\nabla_{x_1(0)} E = -\lambda_1(0),$ and $\nabla_{x_2(0)} E = -\lambda_2(0)$. These gradients are evaluated by integrating backward the adjoint system to $t = 0$ starting from $t = 1$. The gradients are used to update the initial conditions using a line-search [Nocedal and Wright, 2006]. Such a method, performing a gradient descent only on the initial conditions, is known as a *shooting method.* This is different from a *relaxation method*, which starts with a guess for the exact solution at each interior point and iteratively adjusts the guess to approximate the optimal solution. Since different from a relaxation method a shooting method works on the initial conditions, it may save memory especially when we deal with images and shapes later.

This optimal-control perspective constitutes a general method for estimating first-order curves which allows generalizing the notion of straight lines to manifolds (geodesics), as long

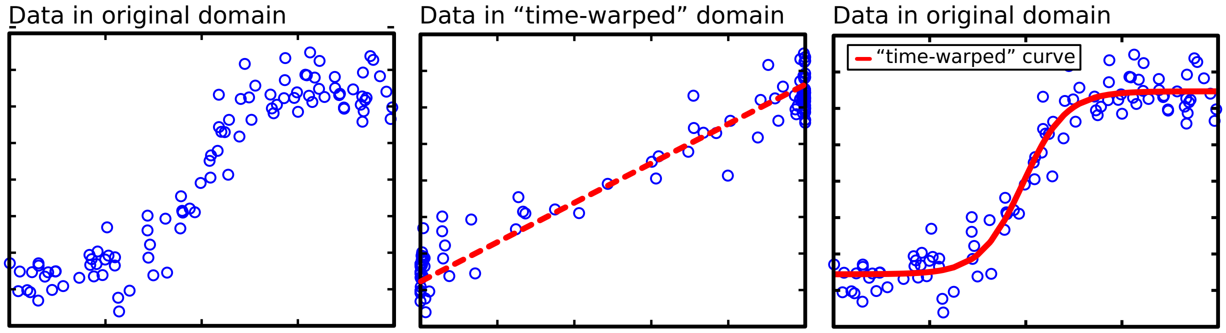


Figure 3.1: Illustration of time-warped regression in $\mathbb{R}$. The dashed straight-line (middle) shows the fitting result in the warped time coordinates, and the solid curve (right) demonstrates the fitting result to the original data points (left).

as the forward system (dynamics), the gradient computations, as well as the gradient steps all respect the geometry of the underlying space.

3.1.2 Time-Warped Regression

Fitting straight lines is too restrictive for some data. Hence, the idea of time-warped regression is to use a simple model to warp the time-points, or more generally the independent variable, when comparison to data is performed, *e.g.*, as in the data matching term of Eq. (3.1). The *time-warp* should maintain the order of the data, and hence needs to be diffeomorphic. This is conceptually similar to an *error-in-variables* model where uncertainties in the independent variables are modeled. However, in the concept of time-warping, we are not directly concerned with modeling such uncertainties, but instead in obtaining a somewhat richer model based on a known and easy-to-estimate linear regression model.

In principle, the mapping of the time points could be described by a general diffeomorphism. In fact, such an approach is followed in [Durrleman et al., 2013] for spatiotemporal atlas-building in the context of shape analysis. The motivation for proposing an approach to linear regression with *parametric* time-warps is to keep the model simple while gaining more flexibility. Extensions to non-parametric approaches can easily be obtained. A representative of a simple parametric regression model is *logistic regression*[2] which is typically used to model saturation effects. Under this model, points that are close in time for the linear fit may be mapped to points far apart in time, thereby it allows modeling saturations for instance (*cf*. Fig. 3.1). Other possibilities of parametric time-warps include those derived from families of quadratic, logarithmic and exponential functions.

Formally, let $f : \mathbb{R} \to \mathbb{R}$, $t \mapsto \bar{t} = f(t; \boldsymbol{\theta})$ denote a parametrized (by $\boldsymbol{\theta}$) time-warping function and let $x_1(\bar{t})$ denote the particle on the regression line in the warped time coordinates $\bar{t}$. Following this notation, the states are denoted as $(x_1(\bar{t}), x_2(\bar{t}))$ and represent position and slope in re-parametrized time $\bar{t}$. In *time-warped regression*, the *data matching* term in the sum of Eq. (3.1) then becomes $\|x_1(f(t_i;\boldsymbol{\theta}))-y_i\|^2$ and the objective (as before) is to optimize $x_1(\bar{t}_0)$ and $x_2(\bar{t}_0)$ as well as the parameter $\boldsymbol{\theta}$.

A convenient way to minimize the energy functional in Eq. (3.1) with the adjusted data matching term is to use an alternating optimization strategy. That is, we first fix $\boldsymbol{\theta}$ to update the initial conditions, and then fix the initial conditions to update $\boldsymbol{\theta}$. This requires the derivative of the energy with respect to $\boldsymbol{\theta}$ for fixed $x_1(\bar{t})$. Using the chain rule, we obtain the gradient as

$$\nabla_{\boldsymbol{\theta}} E = 2 \sum_{i=0}^{N-1} (x_1(f(t_i;\boldsymbol{\theta})) - y_i)^\top \dot{x}_1(f(t_i;\boldsymbol{\theta})) \nabla_{\boldsymbol{\theta}} f(t_i;\boldsymbol{\theta}) \ . \tag{3.3}$$

Given a numerical solution to the regression problem of Section 3.1.1, the time-warped extension alternatingly updates (1) the initial conditions $(x_1(\bar{t}_0), x_2(\bar{t}_0))$ in the warped time domain using the gradients in Eq. (3.1.1) and (2) $\boldsymbol{\theta}$ using the gradient in Eq. (3.3). Fig. 3.1

[2] Not to be confused with the statistical classification method.

visualizes the principle of time-warped linear regression on a collection of artificially generated data points. While the new model only slightly increases the overall complexity, it notably increases modeling flexibility by using a curve instead of a straight line.

3.1.3 Cubic Spline Regression

To further increase the flexibility of a regression model, cubic splines are another commonly used technique. This section revisits cubic spline regression from the optimal-control perspective. This will facilitate the transition to general Riemannian manifolds.

Variational formulation. An acceleration-controlled curve with time-dependent states (x_1, x_2, x_3) such that $\dot{x}_1 = x_2$ and $\dot{x}_2 = x_3$, defines a cubic curve in $\mathbb{R}^n$. Such a curve is a solution to the energy minimization problem, *cf.* [Ahlberg et al., 1967],

$$\min_{\mathbf{\Theta}} E(\mathbf{\Theta}) = \frac{1}{2}\int_0^1 \|x_3\|^2 \, dt, \quad \text{subject to} \quad \dot{x}_1 = x_2(t) \text{ and } \dot{x}_2 = x_3(t) \ , \tag{3.4}$$

with $\mathbf{\Theta} = \{x_i(t)\}_{i=1}^3$. Here, x_3 is referred to as the *control variable* that describes the acceleration of the dynamics in this system. Similar to the strategy for fitting straight lines, we can get a relaxation solution to Eq. (3.4) by adding adjoint variables which results in the system of adjoint equations $\dot{\lambda}_1 = 0$ and $\dot{x}_3 = -\lambda_1$.

Shooting solution. To obtain the shooting formulation, we explicitly add the evolution of x_3, *i.e.*, $\dot{x}_3 = -\lambda_1$, as another dynamical constraint; this increases the order of the dynamics. Setting $x_4 = -\lambda_1$ results in the classical system of equations for shooting cubic curves

$$\dot{x}_1 = x_2(t), \quad \dot{x}_2 = x_3(t), \quad \dot{x}_3 = x_4(t), \quad \dot{x}_4 = 0 \ . \tag{3.5}$$

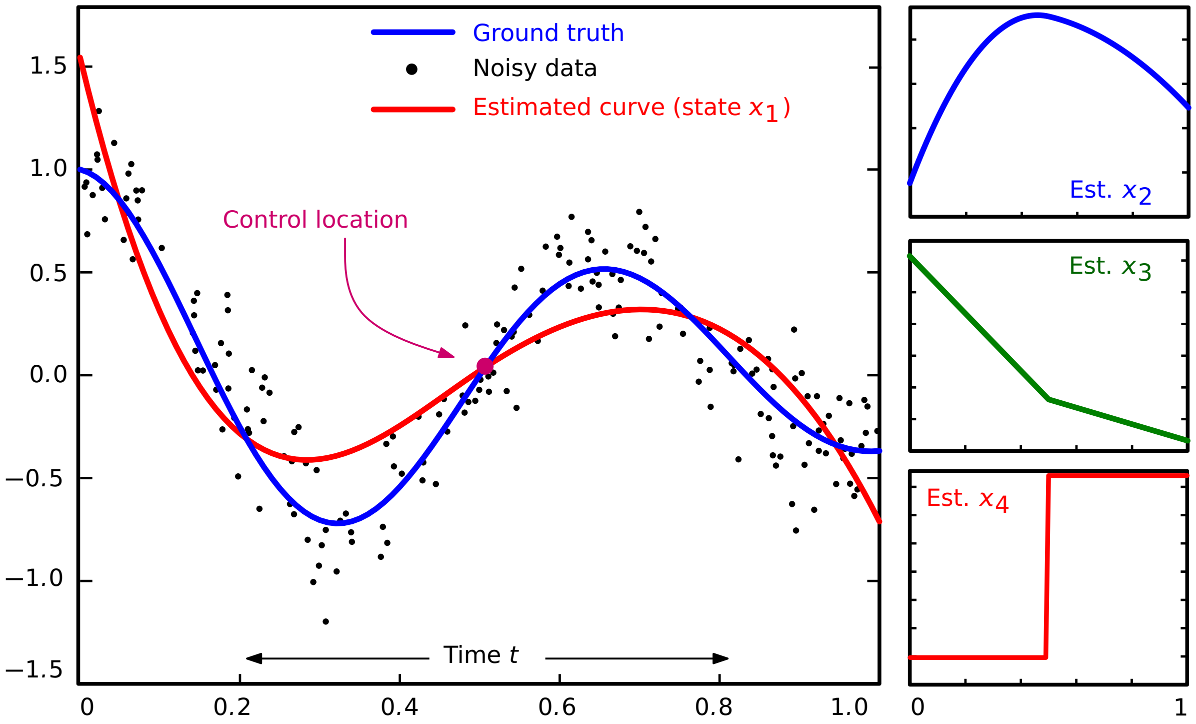


Figure 3.2: Cubic-spline regression in $\mathbb{R}$. The left side shows the regression result, and the remaining plots show the other states.

The states (x_1, x_2, x_3, x_4), at all times, are entirely determined by their initial values $\{x_i(0)\}_{i=1}^4$ and, in particular we have

$$x_1(t) = x_1(0) + x_2(0)t + \frac{1}{2}x_3(0)t^2 + \frac{1}{6}x_4(0)t^3 \quad . \tag{3.6}$$

Data-independent controls. Using the shooting equations of Eq. (3.5) for cubic splines, we can define a *smooth* curve that best fits the data in the least-squares sense. Since a cubic polynomial by itself is restricted to only fit "cubic-like" data, we add flexibility by gluing piecewise cubic polynomials together. Typically, we define controls at pre-defined locations and only allow the state x_4 to jump at those locations.

Now, let $\{t_c\}_{c=1}^C, t_c \in (0, 1)$ denote C data-independent fixed control points, which implicitly define $C+1$ intervals in $[0, 1]$, denoted as $\{\mathcal{I}_c\}_{c=1}^{C+1}$. The constrained energy minimization problem corresponding to the regression task, in this setting, can be written as

$$\begin{aligned}\min_{\boldsymbol{\Theta}} \quad & E(\boldsymbol{\Theta}) = \sum_{c=1}^{C+1} \sum_{t_i \in \mathcal{I}_c} \|x_1(t_i) - y_i\|^2, \\ \text{subject to} \quad & \dot{x}_1 = x_2(t),\ \dot{x}_2 = x_3(t),\ \dot{x}_3 = x_4(t),\ \dot{x}_4 = 0 \text{ within } \mathcal{I}_c \\ & \text{and } x_1, x_2, x_3 \text{ are continuous across } t_c \ \ ,\end{aligned} \tag{3.7}$$

with parameters $\boldsymbol{\Theta} = \{\{x_i(0)\}_{i=1}^4, \{x_4(t_c)\}_{c=1}^C\}$. Using time-dependent adjoint states $\{\lambda_i\}_{i=1}^4$ for the dynamics constraints, and (time-independent) duals $\nu_{c,i}$ for the continuity constraints, we can derive the adjoint system of equations from the unconstrained Lagrangian as

$$\dot{\lambda}_1 = 0, \quad \dot{\lambda}_2 = -\lambda_1, \quad \dot{\lambda}_3 = -\lambda_2, \quad \dot{\lambda}_4 = -\lambda_3 \ \ . \tag{3.8}$$

The gradients w.r.t. the initial conditions $\{x_i(0)\}_{i=1}^4$ are

$$\nabla_{x_1(0)} E = -\lambda_1(0), \quad \nabla_{x_2(0)} E = -\lambda_2(0), \quad \nabla_{x_3(0)} E = -\lambda_3(0), \quad \nabla_{x_4(0)} E = -\lambda_4(0) \ \ . \tag{3.9}$$

The *jerks* (*i.e.*, rate of acceleration change) at $x_4(t_c)$ are updated using $\nabla_{x_4(t_c)} E = -\lambda_4(t_c)$. The values of the adjoint variables at 0 are computed by integrating backward the adjoint system starting from $\forall i : \lambda_i(1) = 0$. Note that λ_1, λ_2 and λ_3 are continuous at joints, but λ_1 jumps at the data-point location as per $\lambda_1(t_i^+) - \lambda_1(t_i^-) = 2(x_1(t_i) - y_i)$. During backward integration, λ_4 starts with zero at each t_{c+1} and the accumulated value at t_c is used for the gradient update of $x_4(t_c)$.

It is critical to note that, along the time t, such a formulation guarantees that $x_4(t)$ is piecewise constant, $x_3(t)$ is piecewise linear, $x_2(t)$ is piecewise quadratic, and $x_1(t)$ is piecewise cubic; this results in a cubic spline curve. Fig. 3.2 demonstrates this shooting-based spline fitting method on data in $\mathbb{R}$. While it is difficult to explain this data with one simple cubic curve, it suffices to add one control point to recover the underlying trend. The state x_4 experiences a jump at the control location that integrates up three-times to give a C^2-continuous evolution for the state x_1.

3.2 Regression on Riemannian Manifolds

In this section the optimal-control perspective of Section 3.1 is adopted and the regression problems are generalized to nonflat, smooth Riemannian manifolds. In the literature this generalization of linear regression is typically referred to as *geodesic regression.* For a thorough treatment of Riemannian manifolds, please refer to [Boothby, 1986]. Note that the term geodesic regression here does not refer to the model that is fitted but rather to the fact that the Euclidean distance in the data matching term of the energies is replaced by the geodesic distance on the manifold. In particular, the measurements $\{y_i\}_{i=0}^{N-1}$ in Euclidean space now become elements $\{Y_i\}_{i=0}^{N-1}$ on some Riemannian manifold $\mathcal{M}$ with Riemannian metric $\langle \cdot, \cdot \rangle_p$ at $p \in \mathcal{M}$[3]. The geodesic distance, induced by this metric, will be denoted as d_g. Besides, the variable t_i is replaced with r_i, indicating that the independent value does not have to be *time*, but can also represent other entities, such as counts or speed.

The first objective is to estimate a geodesic $\gamma : \mathbb{R} \to \mathcal{M}$, represented by initial point $\gamma(r_0)$ and initial velocity $\dot{\gamma}(r_0)$ at the tangent space $\mathcal{T}_{\gamma(r_0)}\mathcal{M}$, *i.e.*,

$$\min_{\mathbf{\Theta}} E(\mathbf{\Theta}) = \alpha \underbrace{\int_0^1 \langle \dot{\gamma}, \dot{\gamma} \rangle_{\gamma(r)} \, dr}_{\text{Regularity}} + \underbrace{\frac{1}{\sigma^2} \sum_{i=0}^{N-1} d_g^2(\gamma(r_i), Y_i)}_{\text{Data-matching}} \tag{3.10}$$

$$\text{subject to} \quad \nabla_{\dot{\gamma}} \dot{\gamma} = 0 \text{ (geodesic equation)} \ ,$$

[3] The subscript p will be omitted when it is clear from the context.

with $\mathbf{\Theta} = \{\gamma(0), \dot{\gamma}(0)\}$ and ∇ denoting the Levi-Civita connection on $\mathcal{M}$. The covariant derivative $\nabla_{\dot{\gamma}}\dot{\gamma}$ of value 0 ensures that the curve is a geodesic. The parameters $\alpha \geq 0$ and $\sigma > 0$ balance the regularity and the data-matching term. In the Euclidean case, there is typically no regularity term because we usually do not have prior knowledge about the slope. Similarly, on Riemannian manifolds we may penalize the initial velocity by choosing $\alpha > 0$; but typically, α is also set to 0. The regularity term on the velocity can be further reduced to a smoothness penalty at r_0, *i.e.*, $\int_0^1 \langle \dot{\gamma}, \dot{\gamma} \rangle dr = \langle \dot{\gamma}(r_0), \dot{\gamma}(r_0) \rangle$, because of the energy conservation along the geodesic. Also, since the geodesic is represented by the initial conditions $(\gamma(r_0), \dot{\gamma}(r_0))$, we can move along the geodesic and estimate the point $\gamma(r_i)$ that corresponds to Y_i.

3.2.1 Optimization via Geodesic Shooting

Taking the optimal-control point of view, the second-order problem of Eq. (3.10) can be written as a first-order system, upon the introduction of auxiliary states $X_1(r) = \gamma(r)$ and $X_2(r) = \dot{\gamma}(r)$. Here, X_1 corresponds to the *intercept* and X_2 corresponds to the *slope* in classical linear regression. Considering the simplified smoothness penalty of the previous section, the constrained minimization of Eq. (3.10) reduces to

$$\begin{aligned} \min_{\mathbf{\Theta}} \quad & E(\mathbf{\Theta}) = \alpha \langle X_2(r_0), X_2(r_0) \rangle \; + \frac{1}{\sigma^2} \sum_{i=0}^{N-1} d_g^2(X_1(r_i), Y_i) \\ & \text{subject to} \quad \nabla_{X_2} X_2 = 0 \;\; , \end{aligned} \tag{3.11}$$

with $\mathbf{\Theta} = \{X_i(r_0)\}_{i=1}^2$. $X_1(r_i)$ is the estimated point on the geodesic at r_i, obtained by shooting forward with $X_1(r_0)$ and $X_2(r_0)$. Analogously to the elaborations of previous sections, we convert Eq. (3.11) to the Lagrangian function via time-dependent adjoint variables, then take variations with respect to its arguments resulting in the KKT conditions, and eventually obtain (1) dynamical systems of states and adjoint variables, (2) boundary conditions on the adjoint variables, and (3) gradients with respect to initial conditions. By shooting forward/backward and updating the initial states via the gradients, we can obtain a numerical solution to the problem.

3.2.2 Time-Warped Regression

The time-warping strategy of Section 3.1.2 can also be adapted to Riemannian manifolds, because it focuses on warping the axis of the independent scalar-valued variable, not the axis of the dependent manifold-valued variable. In other words, the time-warped model is independent of the underlying type of space. Formally, given a warping function f (*cf*. Section 3.1.2), all instances of the form $X_j(r_i)$ in Eq. (3.11) are replaced by $X_j(f(r_i;\boldsymbol{\theta}))$ for $j = 1, 2$. While the model retains its simplicity, *i.e.*, we still fit geodesic curves, the warping function allows for increased modeling flexibility.

Since we have an existing solution to the problem of fitting geodesic curves, the easiest way to minimize the resulting energy is by alternating optimization, similar to Section 3.1.2. This requires the derivative of the energy with respect to $\boldsymbol{\theta}$ for fixed $X_i(\overline{r})$. Application of the chain rule and [Samir et al., 2012, Appendix A] yields

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} E = \; & 2\alpha\langle \dot{X}_2(f(r_0;\boldsymbol{\theta})), X_2(f(r_0;\boldsymbol{\theta}))\rangle \nabla_{\boldsymbol{\theta}} f(r_0;\boldsymbol{\theta}) \\ & - \frac{2}{\sigma^2}\sum_{i=0}^{N-1}\langle \operatorname{Log}_{X_1(f(r_i;\boldsymbol{\theta}))} Y_i, \dot{X}_1(f(r_i;\boldsymbol{\theta}))\rangle \nabla_{\boldsymbol{\theta}} f(r_i;\boldsymbol{\theta}) \;\; ,\end{aligned} \tag{3.12}$$

where $\operatorname{Log}_{X_1(f(r_i;\boldsymbol{\theta}))} Y_i$ denotes the Riemannian log-map, *i.e.*, the initial velocity of the geodesic connecting $X_1(f(r_i;\boldsymbol{\theta}))$ and Y_i in unit time and $\dot{X}_1(f(r_i;\boldsymbol{\theta}))$ is the velocity of the regression geodesic at the warped-time point. This leaves to choose a good parametric model for $f(r;\boldsymbol{\theta})$. As the time warp is required to be diffeomorphic, we choose a parametric model that is diffeomorphic by design. One choice is the generalized logistic function [Fekedulegn et al., 1999], *e.g.*, with asymptotes 0 for $r \to -\infty$ and 1 for $r \to \infty$, given by

$$f(r;\boldsymbol{\theta}) = \frac{1}{(1+\beta e^{-k(r-M)})^{1/m}} \ , \tag{3.13}$$

with $\boldsymbol{\theta} = (k, M, \beta, m)$. The parameter k controls the growth rate, M is the time of maximum growth if $\beta = m$, β and m define the value of f at $t = M$, and $m > 0$ affects the asymptote of maximum growth. This function maps the original infinite time interval to a warped time-range from 0 to 1. In summary, the alternating optimization algorithm is as follows:

0) Initialize $\boldsymbol{\theta}$ such that the warped time is evenly distributed within (0, 1).

1) Update time-points $\{\overline{r}_i = f(r_i;\boldsymbol{\theta})\}_{i=0}^{N-1}$ and perform standard geodesic regression.

2) Update $\boldsymbol{\theta}$ by numerical optimization using the gradient given in Eq. (3.12).

3) Check convergence. If not converged goto 1).

3.2.3 Cubic Spline Regression

As in Section 3.1.3, cubic curves on a Riemannian manifold $\mathcal{M}$ can be defined as solutions to the variational problem of minimizing an acceleration-based energy. The notion of acceleration is defined using the covariant derivatives on Riemannian manifolds [Noakes et al., 1989, Camarinha et al., 1995]. In particular, we define a *time-dependent control*, *i.e.*, a forcing variable $X_3(r)$, as

$$X_3(r) = \nabla_{X_2(r)} X_2(r) = \nabla_{\dot{X}_1(r)} \dot{X}_1(r) \ . \tag{3.14}$$

We can interpret $X_3(r)$ as a control that forces the curve $X_1(r)$ to deviate from being a geodesic [Trouvé and Vialard, 2012] (which is the case if $X_3(r) = 0$). As an end-point problem, a Riemannian cubic curve is thus defined by the curve $X_1(r)$ such that it minimizes an energy of the form

$$E(X_1) = \frac{1}{2}\int_0^1 \|\nabla_{\dot{X}_1}\dot{X}_1\|^2 dt, \tag{3.15}$$

where the norm $\|\cdot\|$ is induced by the metric on $\mathcal{M}$ at X_1. In Section 3.3.3, this concept will be adapted to the Grassmannian to enable regression with cubic splines.

3.3 Regression on the Grassmannian

The Grassmannian is a type of Riemannian manifold where the geodesic distance, parallel transport, as well as the Riemannian Log-/Exp-map are relatively simple to compute (see [Gallivan et al., 2003] and Section 2.1.2). According to the Riemannian structure of the Grassmannian discussed in Section 2.1.2 (see [Absil et al., 2004] for details), those three regression models are specialized to this manifold.

3.3.1 Standard Geodesic Regression

First, the inner-product and the squared geodesic distance in Eq. (3.10) are adapted to $\mathcal{G}(p, n)$ by following Section 2.1.2. Next, given the auxiliary states, which are denoted as matrices $\mathbf{X}_1$ (initial point) and $\mathbf{X}_2$ (velocity), we can write the geodesic equation of Eq. (2.2) as a system of first-order dynamics:

$$\dot{\mathbf{X}}_1 = \mathbf{X}_2, \quad \text{and} \quad \dot{\mathbf{X}}_2 = -\mathbf{X}_1(\mathbf{X}_2^\top \mathbf{X}_2) \ . \tag{3.16}$$

For a point on $\mathcal{G}(p, n)$ it should further hold that (1) $\mathbf{X}_1(r)^\top \mathbf{X}_1(r) = \mathbf{I}_p$ and (2) the velocity at $\mathbf{X}_1(r)$ needs to be orthogonal to that point, *i.e.*, $\mathbf{X}_1(r)^\top \mathbf{X}_2(r) = \mathbf{0}$. If we enforce these

Algorithm 1 Standard Grassmannian geodesic regression (Std-GGR)

```
Input: $\{(r_i, \mathbf{Y}_i)\}_{i=0}^{N-1}$, $\alpha$ and $\sigma^2$
Output: $\mathbf{X}_1(r_0)$, $\mathbf{X}_2(r_0)$
Set initial $\mathbf{X}_1(r_0)$ and $\mathbf{X}_2(r_0)$, e.g., $\mathbf{X}_1(r_0) = \mathbf{Y}_0$, and $\mathbf{X}_2(r_0) = \mathbf{0}$.
while not converged do
    Solve Eqs. (3.16) with $\mathbf{X}_1(r_0)$ and $\mathbf{X}_2(r_0)$ forward for $r \in [r_0, r_{N-1}]$.
    Solve $\begin{cases} \dot{\lambda}_1 = \lambda_2 \mathbf{X}_2^\top \mathbf{X}_2, \ \lambda_1(r_{N-1}+) = 0, \\ \dot{\lambda}_2 = -\lambda_1 + \mathbf{X}_2(\lambda_2^\top \mathbf{X}_1 + \mathbf{X}_1^\top \lambda_2), \ \lambda_2(r_{N-1}) = 0 \end{cases}$ backward with jump conditions
    $\lambda_1(r_i-) = \lambda_1(r_i+) - \frac{1}{\sigma^2}\nabla_{\mathbf{X}_1(r_i)} d_g^2(\mathbf{X}_1(r_i), \mathbf{Y}_i)$, and $\nabla_{\mathbf{X}_1(r_i)} d_g^2(\mathbf{X}_1(r_i), \mathbf{Y}_i)$ computed as
    $-2\,\mathrm{Log}_{\mathbf{X}_1(r_i)} \mathbf{Y}_i$. For multiple measurements at a given $r_i$, the jump conditions for each measurement are added up.
    Compute gradients with respect to initial conditions:
        $\nabla_{\mathbf{X}_1(r_0)} E = -(\mathbf{I}_n - \mathbf{X}_1(r_0)\mathbf{X}_1(r_0)^\top)\lambda_1(r_0-) + \mathbf{X}_2(r_0)\lambda_2(r_0)^\top \mathbf{X}_1(r_0)$,
        $\nabla_{\mathbf{X}_2(r_0)} E = 2\alpha \mathbf{X}_2(r_0) - (\mathbf{I}_n - \mathbf{X}_1(r_0)\mathbf{X}_1(r_0)^\top)\lambda_2(r_0)$.
    Use a line search with these gradients to update $\mathbf{X}_1(r_0)$ and $\mathbf{X}_2(r_0)$ (see Algorithm 2).
end while
```

two constraints at the starting point r_0, they will remain satisfied along the geodesic. This yields

$$\begin{aligned} \min_{\mathbf{\Theta}} E(\mathbf{\Theta}) = \quad & \alpha \ \mathrm{tr}\ \mathbf{X}_2(r_0)^\top \mathbf{X}_2(r_0) \ + \ \frac{1}{\sigma^2}\sum_{i=0}^{N-1} d_g^2(\mathbf{X}_1(r_i), \mathbf{Y}_i) \\ \text{subject to} \quad & \mathbf{X}_1(r_0)^\top \mathbf{X}_1(r_0) = \mathbf{I}_p, \ \mathbf{X}_1(r_0)^\top \mathbf{X}_2(r_0) = \mathbf{0} \text{ and Eq. (3.16)} \ , \end{aligned} \tag{3.17}$$

with $\mathbf{\Theta} = \{\mathbf{X}_i(r_0)\}_{i=1}^2$. Based on the adjoint method, we obtain the shooting solution to Eq. (3.17), listed in Alg. 1. Note that the jump conditions on λ_1 involve the gradient of the residual term $d_g^2(\mathbf{X}_1(r_i), \mathbf{Y}_i)$ with respect to $\mathbf{X}_1(r_i)$, *i.e.*, the base point of the residual on the fitted geodesic; this gradient is $-2\,\mathrm{Log}_{\mathbf{X}_1(r_i)} \mathbf{Y}_i$ (See next subsection). This problem of fitting a geodesic is referred to as *standard Grassmannian geodesic regression (Std-GGR).*

Residuals to curves on the Grassmannian. In Algorithm 1 we need the gradient of the residuals $d_g^2(\mathbf{X}_1(r_i), \mathbf{Y}_i)$ with respect to the base point $\mathbf{X}_1(r_i)$ in order to compute the jump conditions for the adjoint variable λ_1. The residual measures the squared geodesic distance

Algorithm 2 Grassmannian equivalent of $x^{k+1} = x^k - \Delta t g$, where Δt is the time step and g is the gradient.

```
Input: $\mathbf{X}_1(r_0)$, $\mathbf{X}_2(r_0)$, $\nabla_{\mathbf{X}_1(r_0)} E$, $\nabla_{\mathbf{X}_2(r_0)} E$, $\Delta t$
Output: Updated $\mathbf{X}_1^u(r_0)$ and $\mathbf{X}_2^u(r_0)$
Compute $\overline{\mathbf{X}}_2^u(r_0) = \mathbf{X}_2(r_0) - \Delta t \nabla_{\mathbf{X}_2(r_0)} E$
Compute $\mathbf{X}_1^u(r_0)$ by flowing for $\Delta t$ along geodesic with initial condition $(\mathbf{X}_1(r_0), -\nabla_{\mathbf{X}_1(r_0)} E)$
Transport $\overline{\mathbf{X}}_2^u(r_0)$ along the geodesic connecting $\mathbf{X}_1(r_0)$ to $\mathbf{X}_1^u(r_0)$, using (3.23), resulting in $\overline{\mathbf{X}}_2^{uT}(r_0)$
Project updated initial velocity onto the tangent space (for consistency): $\mathbf{X}_2^u(r_0) \leftarrow (\mathbf{I}_n - \mathbf{X}_1^u(r_0)\mathbf{X}_1^u(r_0)^\top)\overline{\mathbf{X}}_2^{uT}(r_0)$.
```

between the point $\mathbf{X}_1(r_i)$ on the fitted curve and the corresponding measurement $\mathbf{Y}_i$. To derive this gradient, we consider the constrained minimization problem of two points with exact matching:

$$\begin{aligned} E(\mathbf{X}_1(r)) \quad &= \int_{r_0}^{r_1} \operatorname{tr} \mathbf{X}_2(r)^\top \mathbf{X}_2(r) \, dr, \\ \text{s.t.} \quad & \mathbf{X}_1(r_0) = \mathbf{Y}_0, \; \mathbf{X}_1(r_1) = \mathbf{Y}_1, \text{ and } \mathbf{X}_1(r_0)^\top \mathbf{X}_1(r_0) = \mathbf{I}_p \;, \end{aligned} \tag{3.18}$$

with $\mathbf{X}_2(r) = (\mathbf{I}_n - \mathbf{X}_1(r)\mathbf{X}_1(r)^\top)\mathbf{C}$. We know that the squared distance can be formulated as $d_g^2(\mathbf{Y}_0, \mathbf{Y}_1) = \min_{\mathbf{X}_1(r)} E(\mathbf{X}_1(r))$ for $r_0 = 0$ and $r_1 = 1$. After adding the constraint on the form of $\mathbf{X}_2(r)$ via the time-dependent adjoint variable λ and the constraint $\mathbf{X}_1(0)^\top \mathbf{X}_1(0) = \mathbf{I}_p$ via λ_c, we obtain (by taking variations), among other terms, the optimality condition

$$(2\mathbf{C}^\top - \lambda^\top)(\mathbf{I}_n - \mathbf{X}_1(r)\mathbf{X}_1(r)^\top) = \mathbf{0} \;, \tag{3.19}$$

and another optimality condition for a free initial condition[4] $\mathbf{X}_1(0)$

$$\nabla_{\mathbf{X}_1(0)} E = -\lambda(0) + \mathbf{X}_1(0)(\lambda_c^\top + \lambda_c) = \mathbf{0} \;. \tag{3.20}$$

[4] Note that technically we started with $\mathbf{X}_1(r_0) = \mathbf{Y}_0$, i.e., this condition would not be free and we would not need to consider variations of $\mathbf{X}_1(r_0)$. However, the goal is to compute the energy gradient with respect to the initial condition $\mathbf{X}_1(r_0)$. Consequentially, the variation of the energy with respect to it allows computing $\nabla_{\mathbf{X}_1(0)} E$.

Left-multiplication by $\mathbf{X}_1(0)^\top$ yields $\lambda_c + \lambda_c^\top = \mathbf{X}_1^\top(0)\lambda(0)$ which we can use to obtain, upon back-substitution into Eq. (3.20),

$$-(\mathbf{I}_n - \mathbf{X}_1(0)\mathbf{X}_1(0)^\top)\lambda(0) = \mathbf{0} \ . \tag{3.21}$$

Using Eq. (3.19) and the above expression for $\mathbf{X}_2(r)$, we can obtain

$$\nabla_{\mathbf{X}_1(0)} E = -(\mathbf{I}_n - \mathbf{X}_1(0)\mathbf{X}_1(0)^\top)\lambda(0) = -2\mathbf{X}_2(0) \ , \tag{3.22}$$

with $\mathbf{X}_2(0)$ being the tangent vector at $\mathbf{X}_1(0)$ such that $\mathrm{Exp}_{\mathbf{X}_1(0)}(\mathbf{X}_2(0)) = \mathbf{X}_1(1)$. This tangent vector can be computed via the Log-map presented in Section 2.1.2, i.e., $\mathbf{X}_2(0) = \mathrm{Log}_{\mathbf{X}_1(0)}(\mathbf{X}_1(1))$.

Line search on the Grassmannian. Performing a line search on the Grassmann manifold in Algorithm 1 is not as straightforward as in Euclidean space since we need to assure that the constraints for $\mathbf{X}_1(r_0)$ and $\mathbf{X}_2(r_0)$ are fulfilled for any given step. In particular, changing $\mathbf{X}_1(r_0)$ will change the associated tangent vector $\mathbf{X}_2(r_0)$. Once, we have updated $\mathbf{X}_1(r_0)$ to $\mathbf{X}_1^u(r_0)$ by moving along the geodesic defined by $\mathbf{X}_1(r_0)$ and the gradient of the energy with respect to this initial point, *i.e.*, $\nabla_{\mathbf{X}_1(r_0)} E$, we can transport the tangent $\mathbf{X}_2(r_0)$ to $\mathbf{X}_1^u(r_0)$ using the closed form solution for *parallel transport* of [Edelman et al., 1998] (*cf*. Section 2.1.2). In particular,

$$\mathbf{X}_2^u(r_0) = \ [\mathbf{X}_1(r_0)\mathbf{V} \ \ \mathbf{U}] \begin{pmatrix} -\sin t\mathbf{\Sigma} \\ \cos t\mathbf{\Sigma} \end{pmatrix} \mathbf{U}^\top + \ (\mathbf{I}_n - \mathbf{U}\mathbf{U}^\top)\mathbf{X}_2(r_0) \tag{3.23}$$

where $\mathbf{H} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ is the compact SVD of the tangent vector at $\mathbf{X}_1(r_0)$ along the geodesic connecting $\mathbf{X}_1(r_0)$ and $\mathbf{X}_1^u(r_0)$. Note that $t = 1$ in Eq. (3.23). Algorithm 2 lists the line search procedure in full technical detail.

Note: when implementing Algorithm 1 (the same holds for Algorithm 3) it is important to pay attention to the ordering of the matrix multiplications, as performing them in an appropriate order will reduce time and memory complexity.

3.3.2 Time-Warped Regression

Since the concept of time-warped geodesic regression is generic for Riemannian manifolds, specialization to the Grassmannian is straightforward. The Std-GGR solution is used during the alternating optimization steps. By choosing the generalized logistic function of Eq. (3.13) to account for saturations of scalar-valued outputs, the time-warped model on $\mathcal{G}(p, n)$ can be used to capture saturation effects for which standard geodesic regression is not sensible. This strategy is referred to as *time-warped Grassmannian geodesic regression (TW-GGR).*

3.3.3 Cubic Spline Regression

To enable cubic spline regression on the Grassmannian, we follow Section 3.2.3 and add the external force $\mathbf{X}_3$. In other words, we represent an acceleration-controlled curve $\mathbf{X}_1(r)$ on $\mathcal{G}(p, n)$ using a dynamic system with states $(\mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3)$ such that

$$\mathbf{X}_2 = \dot{\mathbf{X}}_1, \quad \text{and} \quad \mathbf{X}_3 = \dot{\mathbf{X}}_2 + \mathbf{X}_1(\mathbf{X}_2^\top \mathbf{X}_2) \ . \tag{3.24}$$

If $\mathbf{X}_3 = \mathbf{0}$, the second equation is reduced to the geodesic equation of Eq. (2.2), i.e., the geodesic is *acceleration-free.* To obtain an acceleration-controlled curve, we need to solve

$$\begin{aligned} &\min_{\mathbf{\Theta}} E(\mathbf{\Theta}) = \frac{1}{2} \int_0^1 \text{tr } \mathbf{X}_3^\top \mathbf{X}_3 \ dr \\ &\text{s.t.} \quad \mathbf{X}_1^\top \mathbf{X}_1 = \mathbf{I}_p, \ \mathbf{X}_1^\top \mathbf{X}_2 = \mathbf{0}, \text{ and Eq. (3.24)} \end{aligned} \tag{3.25}$$

with $\mathbf{\Theta} = \{\mathbf{X}_i(r_0)\}_{i=1}^3$. First, we show that by enforcing $\mathbf{X}_1^\top \mathbf{X}_2 = \mathbf{0}$ at all time points, we only need to enforce $\mathbf{X}_1^\top \mathbf{X}_1 = \mathbf{I}_p$ initially. This can be seen by taking the derivative of $\mathbf{X}_1^\top \mathbf{X}_1$ with respect to r, *i.e.*,

$$\frac{d}{dr}\mathbf{X}_1^\top \mathbf{X}_1 = \dot{\mathbf{X}}_1^\top \mathbf{X}_1 + \mathbf{X}_1^\top \dot{\mathbf{X}}_1 = \mathbf{X}_2^\top \mathbf{X}_1 + \mathbf{X}_1^\top \mathbf{X}_2 = \mathbf{0} \ . \tag{3.26}$$

In other words, if $\mathbf{X}_1(0)^\top \mathbf{X}_1(0) = \mathbf{I}_p$ holds, then $\mathbf{X}_1^\top \mathbf{X}_1 = \mathbf{I}_p$ holds for all r by enforcing $\mathbf{X}_1^\top \mathbf{X}_2 = \mathbf{0}$ at all time points. Hence, Equation (3.25) can be rewritten as

$$\begin{aligned} &\min_{\mathbf{\Theta}} E(\mathbf{\Theta}) = \frac{1}{2}\int_0^1 \text{tr } \mathbf{X}_3^\top \mathbf{X}_3 \; dr \\ &\text{s.t. } \mathbf{X}_1^\top \mathbf{X}_2 = \mathbf{0}, \text{Eq. (3.24), and } \mathbf{X}_1(0)^\top \mathbf{X}_1(0) = \mathbf{I}_p \ . \end{aligned} \tag{3.27}$$

For the first three time-evolving constraints, we introduce three time-dependent adjoint variables, λ_1, λ_2, and λ_3, to convert the constrained problem to an unconstrained one. We then take the variations with respect to the state variables which results in the following system of adjoint equations:

$$\begin{cases} -\dot{\lambda}_1^\top + (\mathbf{X}_2^\top \mathbf{X}_2)\lambda_2^\top + \lambda_3 \mathbf{X}_2^\top = \mathbf{0}, \\ -\dot{\lambda}_2^\top - \lambda_1^\top + \mathbf{X}_1^\top \lambda_2 \mathbf{X}_2^\top + \lambda_2^\top \mathbf{X}_1 \mathbf{X}_2^\top + \lambda_3^\top \mathbf{X}_1^\top = \mathbf{0}, \\ \mathbf{X}_3^\top - \lambda_2^\top = \mathbf{0} \ . \end{cases} \tag{3.28}$$

Using $\mathbf{X}_3 = \lambda_2$ and setting $\mathbf{X}_4 = \lambda_1$, we can rewrite the adjoint equations (3.28) as

$$\dot{\mathbf{X}}_4^\top = (\mathbf{X}_2^\top \mathbf{X}_2)\mathbf{X}_3^\top + \lambda_3 \mathbf{X}_2^\top, \tag{3.29}$$

$$\dot{\mathbf{X}}_3^\top = -\mathbf{X}_4^\top + \mathbf{X}_1^\top \mathbf{X}_3 \mathbf{X}_2^\top + \mathbf{X}_3^\top \mathbf{X}_1 \mathbf{X}_2^\top + \lambda_3^\top \mathbf{X}_1^\top \ . \tag{3.30}$$

Next, we derive the form of $\mathbf{X}_3$ using the dynamic constraints. By taking the time-derivative of $\mathbf{X}_1^\top \mathbf{X}_2$ we get

$$\begin{aligned}\frac{d}{dr}\mathbf{X}_1^\top \mathbf{X}_2 &= \dot{\mathbf{X}}_1^\top \mathbf{X}_2 + \mathbf{X}_1^\top \dot{\mathbf{X}}_2 = \dot{\mathbf{X}}_1^\top \mathbf{X}_2 + \mathbf{X}_1^\top (\mathbf{X}_3 - \mathbf{X}_1 \mathbf{X}_2^\top \mathbf{X}_2) \\ &= \mathbf{X}_2^\top \mathbf{X}_2 + \mathbf{X}_1^\top \mathbf{X}_3 - \mathbf{X}_2^\top \mathbf{X}_2 = \mathbf{X}_1^\top \mathbf{X}_3 \ .\end{aligned} \tag{3.31}$$

However, the constraint $\mathbf{X}_1^\top \mathbf{X}_2 = \mathbf{0}$ implies $\frac{d}{dr}\mathbf{X}_1^\top \mathbf{X}_2 = \mathbf{0}$ which yields $\mathbf{X}_1^\top \mathbf{X}_3 = \mathbf{0}$. It further implies $\frac{d}{dr}\mathbf{X}_1^\top \mathbf{X}_3 = \mathbf{0}$ and we get, as a side result,

$$\mathbf{X}_2^\top \mathbf{X}_3 + \mathbf{X}_1^\top \dot{\mathbf{X}}_3 = \mathbf{0} \ . \tag{3.32}$$

We can then use $\mathbf{X}_1^\top \mathbf{X}_3 = \mathbf{0}$ to simplify Eq. (3.30) to

$$-\dot{\mathbf{X}}_3^\top - \mathbf{X}_4^\top + \lambda_3^\top \mathbf{X}_1^\top = \mathbf{0} \quad \Leftrightarrow \quad -\dot{\mathbf{X}}_3 - \mathbf{X}_4 + \mathbf{X}_1 \lambda_3 = \mathbf{0} \ . \tag{3.33}$$

Upon left-multiplication of Eq. (3.33) by $\mathbf{X}_1$ we obtain the expression for λ_3 as

$$\lambda_3 = \mathbf{X}_1^\top \dot{\mathbf{X}}_3 + \mathbf{X}_1^\top \mathbf{X}_4 \overset{(3.32)}{=} \mathbf{X}_1^\top \mathbf{X}_4 - \mathbf{X}_2^\top \mathbf{X}_3 \ . \tag{3.34}$$

Substituting λ_3 into Eq. (3.33) yields the evolution equation for $\dot{\mathbf{X}}_3$ as

$$\dot{\mathbf{X}}_3 = -\mathbf{X}_4 + \mathbf{X}_1 \mathbf{X}_1^\top \mathbf{X}_4 - \mathbf{X}_1 \mathbf{X}_2^\top \mathbf{X}_3 \tag{3.35}$$

and substituting λ_3 in Eq. (3.29) yields the evolution equation for $\dot{\mathbf{X}}_4$

$$\dot{\mathbf{X}}_4 = \mathbf{X}_3 \mathbf{X}_2^\top \mathbf{X}_2 + \mathbf{X}_2 \mathbf{X}_4^\top \mathbf{X}_1 - \mathbf{X}_2 \mathbf{X}_3^\top \mathbf{X}_2 \ . \tag{3.36}$$

In summary, the system of equations for shooting cubic curves on $\mathcal{G}(p,n)$ are

$$\begin{aligned}
&\dot{\mathbf{X}}_1 = \mathbf{X}_2, \quad \dot{\mathbf{X}}_2 = \mathbf{X}_3 - \mathbf{X}_1\mathbf{X}_2^\top\mathbf{X}_2, \\
&\dot{\mathbf{X}}_3 = -\mathbf{X}_4 + \mathbf{X}_1\mathbf{X}_1^\top\mathbf{X}_4 - \mathbf{X}_1\mathbf{X}_2^\top\mathbf{X}_3, \\
&\dot{\mathbf{X}}_4 = \mathbf{X}_3\mathbf{X}_2^\top\mathbf{X}_2 + \mathbf{X}_2\mathbf{X}_4^\top\mathbf{X}_1 - \mathbf{X}_2\mathbf{X}_3^\top\mathbf{X}_2 \ .
\end{aligned} \tag{3.37}$$

It is important to note that $\mathbf{X}_1$ does not follow a geodesic path under non-zero force $\mathbf{X}_3$. Hence, the constraints $\mathbf{X}_1(r)^\top\mathbf{X}_1(r) = \mathbf{I}_p$ and $\mathbf{X}_1(r)^\top\mathbf{X}_2(r) = \mathbf{0}$ should be enforced at *every* instance of r to keep the path on the manifold. However, we showed that enforcing $\mathbf{X}_1(r)^\top\mathbf{X}_2(r) = \mathbf{0}$ at all times already guarantees that $\mathbf{X}_1(r)^\top\mathbf{X}_1(r) = \mathbf{I}_p$ if this holds initially at $r = 0$. Also, $\mathbf{X}_1(r)^\top\mathbf{X}_2(r) = \mathbf{0}$ implies that $\mathbf{X}_1(r)^\top\mathbf{X}_3(r) = \mathbf{0}$. By using this fact during relaxation, the constraints are already implicitly captured in Eqs. (3.37). Subsequently, for shooting we only need to guarantee that all these constraints hold initially. To get a cubic spline curve, we follow Section 3.1.3 and introduce control points $\{r_c\}_{c=1}^C$, which divide the support of the independent variable into several intervals $\mathcal{I}_c$. The first three states should be continuous at the control points, but the state $\mathbf{X}_4$ is allowed to jump. Hence, the spline regression problem on $\mathcal{G}(p,n)$ becomes, *cf*. Eq. (3.7),

$$\begin{aligned}
&\min_{\mathbf{\Theta}} E(\mathbf{\Theta}) = \alpha \int_{r_0}^{r_{N-1}} \operatorname{tr}\, \mathbf{X}_3^\top\mathbf{X}_3 \; dr \ + \ \frac{1}{\sigma^2}\sum_{i=0}^{N-1} d_g^2(\mathbf{X}_1(r_i), \mathbf{Y}_i) \\
&\text{subject to} \quad \mathbf{X}_1(r_0)^\top\mathbf{X}_1(r_0) = \mathbf{I}_p, \ \mathbf{X}_1(r_0)^\top\mathbf{X}_2(r_0) = \mathbf{0}, \ \mathbf{X}_1(r_0)^\top\mathbf{X}_3(r_0) = \mathbf{0}, \\
&\qquad \mathbf{X}_1, \mathbf{X}_2, \mathbf{X}_3 \text{ are continuous at } \{r_c\}_{c=1}^C, \text{ and Eqs. (3.37) hold in each } \mathcal{I}_c,
\end{aligned} \tag{3.38}$$

with $\mathbf{\Theta} = \{\{\mathbf{X}_i(r_0)\}_{i=1}^4, \{\mathbf{X}_4(r_c^+)\}_{c=1}^C\}$. Alg. 3 lists the shooting solution to Eq. (3.38), referred to as *cubic-spline Grassmannian geodesic regression (CS-GGR).*

Algorithm 3 Cubic-spline Grassmannian geodesic regression (CS-GGR)

```
Input: $\{(r_i, \mathbf{Y}_i)\}_{i=0}^{N-1}$, $\{r_c\}_{c=1}^{C}$, $\alpha$ and $\sigma^2$
Output: $\mathbf{X}_1(r_0)$, $\mathbf{X}_2(r_0)$, $\mathbf{X}_3(r_0)$, $\mathbf{X}_4(r_0)$, $\{\mathbf{X}_4(r_c^+)\}_{c=1}^{C}$
Set initial $\mathbf{X}_1(r_0)$ as $\mathbf{Y}_0$ for example, and $\mathbf{X}_2(r_0)$, $\mathbf{X}_3(r_0)$, $\mathbf{X}_4(r_0)$, $\{\mathbf{X}_4(r_c^+)\}_{c=1}^{C}$ as zero matrices.
while not converged do
    Solve Eq. (3.37) forward in each interval with $\mathbf{X}_1(r_0)$, $\mathbf{X}_2(r_0)$, $\mathbf{X}_3(r_0)$, $\mathbf{X}_4(r_0)$, $\{\mathbf{X}_4(r_c^+)\}_{c=1}^{C}$, and $\{\mathbf{X}_1(r_c^+) = \mathbf{X}_1(r_c^-), \mathbf{X}_2(r_c^+) = \mathbf{X}_2(r_c^-), \mathbf{X}_3(r_c^+) = \mathbf{X}_3(r_c^-)\}_{c=1}^{C}$.
    Solve $\begin{cases} \dot{\lambda}_1 &= \lambda_2 \mathbf{X}_2^\top \mathbf{X}_2 - \lambda_3(\mathbf{X}_4^\top \mathbf{X}_1 - \mathbf{X}_3^\top \mathbf{X}_2) - \mathbf{X}_4(\lambda_3^\top \mathbf{X}_1 + \mathbf{X}_2^\top \lambda_4), \\ \dot{\lambda}_2 &= -\lambda_1 + \mathbf{X}_2(\lambda_2^\top \mathbf{X}_1 + \mathbf{X}_1^\top \lambda_2 - \lambda_4^\top \mathbf{X}_3 - \mathbf{X}_3^\top \lambda_4) + \mathbf{X}_3(\lambda_3^\top \mathbf{X}_1 + \mathbf{X}_2^\top \lambda_4) \\ & \quad +\lambda_4(-\mathbf{X}_1^\top \mathbf{X}_4 + \mathbf{X}_2^\top \mathbf{X}_3), \\ \dot{\lambda}_3 &= -\lambda_2 - \lambda_4 \mathbf{X}_2^\top \mathbf{X}_2 + \mathbf{X}_2(\mathbf{X}_1^\top \lambda_3 + \lambda_4^\top \mathbf{X}_2) + 2\alpha \mathbf{X}_3, \\ \dot{\lambda}_4 &= \lambda_3 - \mathbf{X}_1(\mathbf{X}_1^\top \lambda_3 + \lambda_4^\top \mathbf{X}_2) \end{cases}$ backward
    with $\lambda_1(r_{N-1}) = \lambda_2(r_{N-1}) = \lambda_3(r_{N-1}) = \lambda_4(r_{N-1}) = \lambda_4(r_c^-) = 0$, and $\{\lambda_1(r_c^-) = \lambda_1(r_c^+), \lambda_2(r_c^-) = \lambda_2(r_c^+), \lambda_3(r_c^-) = \lambda_3(r_c^+)\}_{c=1}^{C}$, as well as jump conditions $\lambda_1(r_i^-) = \lambda_1(r_i^+) - \frac{1}{\sigma^2} \nabla_{\mathbf{X}_1(r_i)} d_g^2(\mathbf{X}_1(r_i), \mathbf{Y}_i)$, and $\nabla_{\mathbf{X}_1(r_i)} d_g^2(\mathbf{X}_1(r_i), \mathbf{Y}_i)$ computed as $-2\,\mathrm{Log}_{\mathbf{X}_1(r_i)} \mathbf{Y}_i$. For multiple measurements at a given $r_i$, the jump conditions for each measurement are added up.
    Compute gradients with respect to initial conditions and the fourth state at control points:
        $\nabla_{\mathbf{X}_1(r_0)} E = -(\mathbf{I}_n - \mathbf{X}_1(r_0)\mathbf{X}_1(r_0)^\top)\lambda_1(r_0^-) + \mathbf{X}_2(r_0)\lambda_2(r_0)^\top \mathbf{X}_1(r_0) + \mathbf{X}_3(r_0)\lambda_3(r_0)^\top \mathbf{X}_1(r_0)$,
        $\nabla_{\mathbf{X}_2(r_0)} E = -(\mathbf{I}_n - \mathbf{X}_1(r_0)\mathbf{X}_1(r_0)^\top)\lambda_2(r_0)$,
        $\nabla_{\mathbf{X}_3(r_0)} E = -(\mathbf{I}_n - \mathbf{X}_1(r_0)\mathbf{X}_1(r_0)^\top)\lambda_3(r_0)$,
        $\nabla_{\mathbf{X}_4(r_0)} E = -\lambda_4(r_0), \quad \nabla_{\mathbf{X}_4(r_c^+)} E = -\lambda_4(r_c^+),\ c = 1 \ldots C$ .
    Use a line search with these gradients to update $\mathbf{X}_1(r_0)$, $\mathbf{X}_2(r_0)$, $\mathbf{X}_3(r_0)$, $\mathbf{X}_4(r_0)$, and $\{\mathbf{X}_4(r_c^+)\}_{c=1}^{C}$.
end while
```

3.3.4 Experimental Results

Synthetic data. We first demonstrate Std-GGR, TW-GGR and CS-GGR on synthetic data and compare against two approaches from the literature [Rentmeesters, 2011, Su et al., 2012] (see Section 2.2.3 for detailed discussions about the literature).

Each data point in the following experiment represents a 2D sine/cosine signal, sampled at 630 evenly-spaced locations in $[0, 10\pi]$. These signals $\mathbf{s} \in \mathbb{R}^{2\times 630}$ are then linearly projected into $\mathbb{R}^{24}$ via $\bar{\mathbf{s}} = \mathbf{U}\mathbf{s}$, where $\mathbf{W} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{24})$ and $\mathbf{W} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$. White Gaussian noise with $\sigma = 0.1$ is added to $\bar{\mathbf{s}}$. For each embedded signal $\bar{\mathbf{s}} \in \mathbb{R}^{24\times 630}$, we estimate a two-state (*i.e.*, $p = 2$) LDS as discussed in Section 2.1.3, and use the corresponding observability matrix to represent it as a point on $\mathcal{G}(2, 48)$. Besides, each data point has an associated scalar value; this independent variable is uniformly distributed within $(0, 10)$ and controls the *signal frequency* of the data point. For Std-GGR, we directly use this value as the signal frequency to generate 2D signals, while for TW-GGR and CS-GGR, a generalized logistic function or a sine function is adopted to convert the values to a signal frequency for data generation. It is important to note that the *largest* eigenvalue of the state-transition matrix $\mathbf{A}$ in LDS (see Section 2.1.3) reflects the frequency of the sine/cosine signal, according to the experiments.

To quantitatively assess the quality of the fitting results, we design a “denoising” experiment. The data to be used for denoising is generated as follows: First, we use each regression method to estimate a model from the (clean) data points we just generated. In the second step, we take the initial conditions of each model, shoot forward and record the points along the regression curve at fixed values of the independent variable (*i.e.*, the signal frequency). These points serve as the *ground truth (GT)*. In a final step, we take each point on the ground truth curve, generate a random tangent vector at each location and shoot forward along that vector for a small time (*e.g.*, 0.03). The newly generated points then serve as the “noisy” measurements of the original points.

To obtain fitting results on the noisy data, we initialize the first state $\mathbf{X}_1$ with the first data point, and all other initial conditions with $\mathbf{0}$. Table 3.1 lists the differences between our estimated regression curves (Est.) and the corresponding ground truth using two strategies: (1) comparison of the initial conditions as well as the parameters of the warping function in TW-GGR; (2) comparison of the full curves (sampled at the values of the independent variable) and the data points. The numbers indicate that all three models captured different

Method	$\mathbf{X}_1(r_0)$	$\mathbf{X}_2(r_0)$	$\mathbf{X}_3(r_0)$	$\mathbf{X}_4(r_0)$	k	M
Std-GGR	0.02	0.16	–	–	–	–
[Rentmeesters, 2011]	0.02	0.16	–	–	–	–
TW-GGR	0.02	0.16	–	–	0.05	0.6e-2
CS-GGR	0.07	0.54	0.36	0.97	–	–
[Su et al., 2012]	–	–	–	–	–	–

Method	GT *vs.* Data	Data *vs.* Est.	**GT *vs.* Est.**
Std-GGR	0.7e-2	0.7e-2	0.3e-3
[Rentmeesters, 2011]	0.7e-2	0.6e-2	0.3e-3
TW-GGR	6.9e-3	6.6e-3	0.3e-3
CS-GGR	6.8e-3	5.8e-3	1.1e-3
[Su et al., 2012]	6.8e-3	8.2e-3	3.5e-3

Table 3.1: Comparison of the regression results on synthetic data. *First*, we report differences in the initial conditions $\mathbf{X}_i(r_0)$: for $\mathbf{X}_1$, we report the geodesic distance on the Grassmannian; for $\mathbf{X}_2$, $\mathbf{X}_3$ and $\mathbf{X}_4$, we report $\|\mathbf{X}_i^{Est.} - \mathbf{X}_i^{GT}\|_F/\|\mathbf{X}_i^{GT}\|_F$. For multiple $\mathbf{X}_4s$, we take the average. For TW-GGR, we also report the difference in the parameters of the time-warp function (k, M). *Second*, we report the mean squared (geodesic) distance (MSD) between two curves. In particular, we compute (1) the MSD between the data points and the corresponding points on the ground truth (GT) curves (GT *vs.* Data); (2) the MSD between the data points and the points on the estimated regression curves (Data *vs.* Est.) and (3) the MSD between the points on the ground truth curves and the data points on the estimated regression curves (Data *vs.* Est.). The second row shows a comparison to [Rentmeesters, 2011] (conceptually similar to [Fletcher, 2013]). The last row lists the (best) MSDs for the approach of Su *et al.* [Su et al., 2012] on the data used to test CS-GGR (for $\lambda_1/\lambda_2 = 10$).

types of relationships on $\mathcal{G}(2, 48)$. We compare to [Rentmeesters, 2011] which is a representative for Jacobi field based parametric regression (see also [Fletcher, 2013]). Since this approach fits a geodesic and returns an initial point and a velocity vector (as in Std-GGR), the same quantitative measures are reported in Table 3.1. As expected, we essentially obtain the same solution, since the same energy is minimized.

In the context of fitting cubic splines, we compare CS-GGR against the discrete curve fitting approach of [Su et al., 2012], adapted to $\mathcal{G}(p, n)$. Since, [Su et al., 2012] does not output the fitted curve in parametric form, but as a collection of points sampled along the sought-for curve, Table 3.1 only reports performance measures computed from sample points. Additionally, we assess performance by adopting a different evaluation protocol. In particular,

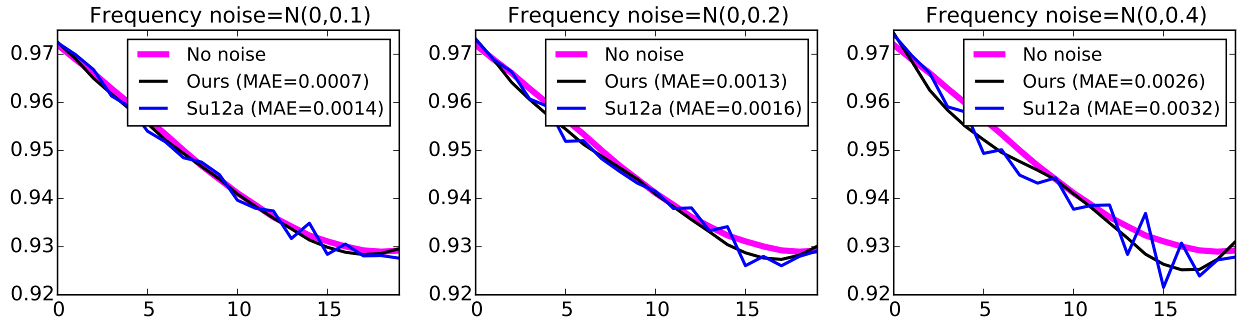


Figure 3.3: CS-GGR (1 control point) *vs.* [Su et al., 2012] ($\lambda_1/\lambda_2 = 10$) in terms of the the largest eigenvalue of the state-transition matrix $\mathbf{A}$ of Eq. (2.5) (reconstructed from the observability matrices that we obtain along each path) to the ground truth.

we take the observability matrices of the linear dynamical systems estimated from each $\bar{\mathbf{s}}$ as the ground truth. We then perturb the signal frequency with Gaussian noise, estimate new dynamical systems and eventually run [Su et al., 2012] and CS-GGR on the observability matrices of these systems. For evaluation, we report the *mean absolute error (MAE)* in the largest eigenvalue of the state-transition matrix $\mathbf{A}$ to the ground truth. Fig. 3.3 shows a visualization of the (real-part) of the largest eigenvalue for different levels of noise. The data matching / smoothing balance for [Su et al., 2012] was set to $(\lambda_1, \lambda_2) = (1, 0.1)$. As we see from Fig. 3.3, the numeric results are fairly similar between both strategies. However, CS-GGR is guaranteed to return a curve with a smooth change in momentum, whereas controlling data-matching *vs.* smoothness in [Su et al., 2012] can lead to instantaneous momentum changes at the sampling locations. Further, storage complexity of our approach scales with the number of control-points, whereas storage complexity of [Su et al., 2012] scales with the the number of sampled points, highlighting one advantage of fitting parametric models with respect to storage requirements. Finally, we remark that we can generate arbitrarily many points along our parametric curves *after* fitting. In contrast, discrete curve fitting strategies would require re-estimation of the curve once the number of desired samples increases.

Applications. To demonstrate Std-GGR, TW-GGR and CS-GGR on actual vision data,

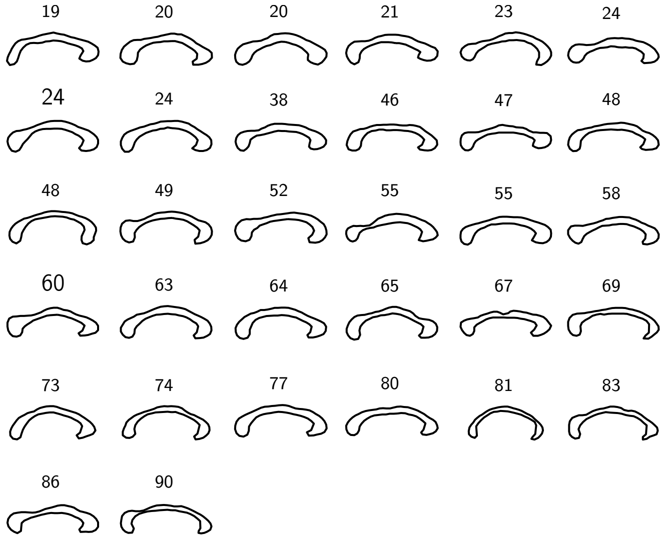


Figure 3.4: Corpora callosa (with the subject's age) [Fletcher, 2013].

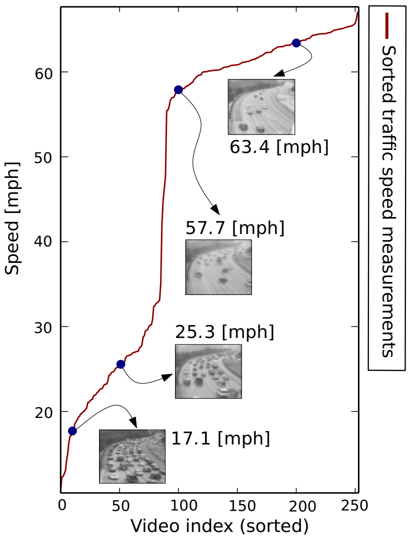


Figure 3.5: UCSD traffic dataset [Chan and Vasconcelos, 2005].

we present four applications: in the first two applications, we regress the manifold-valued variable, *i.e.*, landmark-based shapes; in the last two applications, we predict the independent variable based on the regression curve fitted to the manifold-valued data, *i.e.*, LDS representations of surveillance videos.

(1) Corpus callosum shapes [Fletcher, 2013]. We use a collection of 32 corpus callosum shapes with ages varying from 19 to 90 years, see Fig. 3.4. Each shape is represented by $m = 64$ 2D boundary landmarks and is projected to a point on the Grassmannian using the representation of Section 2.1.3.

(2) Rat calvarium landmarks [Bookstein, 1991]. We use 18 individuals with 8 time points from the Vilmann rat data, each in the age range of 7 to 150 days. Each shape is represented by a set of 8 landmarks. Fig. 3.8 (left) shows a selection of the landmarks projected onto the Grassmann manifold, using the same representation as the corpus callosum data.

(3) UCSD traffic dataset [Chan and Vasconcelos, 2005]. This dataset was introduced in the context of clustering traffic flow patterns with LDS models. It contains a collection

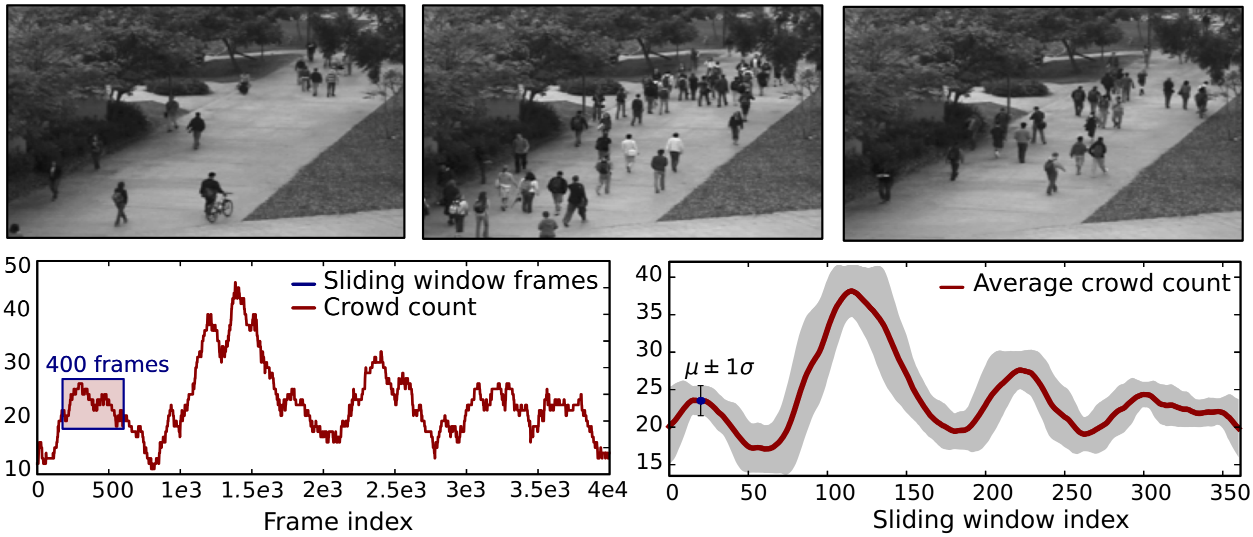


Figure 3.6: Illustration of the dataset for crowd counting. *Top*: Example frames from the UCSD pedestrian dataset [Chan and Vasconcelos, 2012]. *Bottom*: Total crowd count over all frames (left), and average people count over a 400-frame sliding window (right).

of short traffic video clips, acquired by a surveillance system monitoring highway traffic. There are 253 videos in total and each video is roughly matched to the speed measurements from a highway-mounted speed sensor (see Fig. 3.5). We use the pre-processed video clips introduced in [Chan and Vasconcelos, 2005] which were converted to grayscale and spatially normalized to 48×48 pixels with zero mean and unit variance. The rationale for using an LDS representation for speed prediction is the fact that clustering and categorization experiments in [Chan and Vasconcelos, 2005] showed compelling evidence that dynamics are indicative of the traffic class. We argue that the notion of speed of an object (*e.g.*, a car) could be considered a property that humans infer from its visual dynamics.

(4) UCSD pedestrian dataset [Chan and Vasconcelos, 2012]. We use the `Peds1` subset which contains 4000 frames with a ground-truth people count associated with each frame, see Fig. 3.6. Similar to [Chan and Vasconcelos, 2012] we ask the question whether we can infer the number of people in a scene (or clip) without actually detecting the people. While this problem has been addressed by resorting to crowd / motion segmentation and Gaussian process regression on low-level features extracted from the segmentation regions, we go one step further and try to avoid any preprocessing at all. In fact, our objective is to infer an *average* people count from an LDS representation of short video segments (*i.e.*, within a temporal sliding window). This is plausible because the visual dynamics of a scene change as people appear in it. In fact, it could be considered as another form of "traffic". Further, an LDS does not only model the dynamics but also the appearance of videos; both aspects are represented in the observability matrix. However, such a strategy does not allow for fine-grain frame-by-frame predictions as in [Chan and Vasconcelos, 2012]. Yet, it has the advantages of not requiring any pre-selection of features or potentially unstable preprocessing steps such as the aforementioned crowd segmentation. In our setup, we split the 4000 frames into 37 video clips via a sliding window of size 400, shifted by 100 frames (see Fig. 3.6), and associate an *average* people count with each clip. The clips are spatially down-sampled to 60×40 pixel (original: 238×158) to keep the observability matrices at a reasonable size. Since the overlap between the clips potentially biases the experiments, we introduce a weighted variant of system identification (see Section 2.1.3) with weights based on a Gaussian function centered at the middle of the sliding window and a standard deviation of 100. While this ensures stable system identification, by still using 400 frames, it reduces the impact of the overlapping frames on the parameter estimates. With this strategy, the average crowd count is localized to a smaller region.

General settings. In all experiments, α in the energy function is set to 0, σ to 1, the initial point is set to be the first data point, and all other initial conditions are set to zero. For the parameter(s) $\boldsymbol{\theta}$ of TW-GGR, we fix $(\beta, m) = (1, 1)$ so that M is the time of the maximal growth. Usually, one control point is used in CS-GGR.

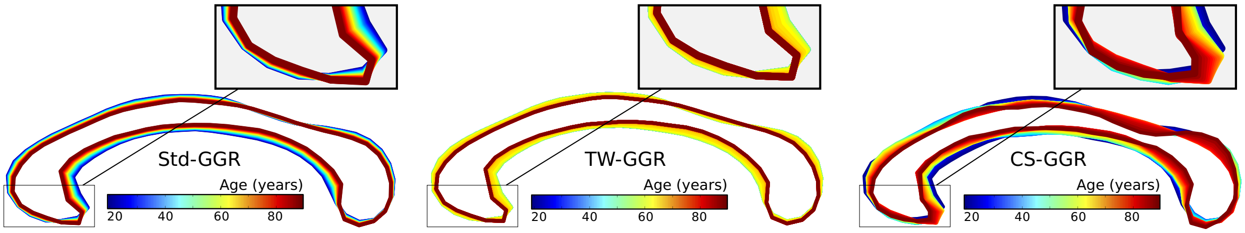


Figure 3.7: Comparison between Std-GGR, TW-GGR and CS-GGR (with one control point) on the corpus callosum data [Fletcher, 2013]. The shapes are generated along the fitted curves and are colored by age (best viewed in color).

Regressing the manifold-valued variable

The first category of applications leverages the regressed relationship between the independent variable, *i.e.*, age, and the manifold-valued dependent variable, *i.e.*, shapes. *The objective is to estimate the shape for a given age.* We demonstrate Std-GGR, TW-GGR and CS-GGR on both corpus callosum and rat calvarium data. The control point for CS-GGR is set to the mean age of the subjects. Three measures are used to quantitatively compare the regression results: (1) the regression *energy*, *i.e.*, the data matching error over all observations; (2) the R^2 statistic on the Grassmannian, which is between 0 and 1, with 1 indicating a perfect fit and 0 indicating a fit no better than the Fréchet mean (see [Fletcher, 2013] for more details); and (3) the *mean squared error (MSE)* on the testing data, reported by means of (leave-one-subject-out) cross-validation (CV). On both datasets, we compare against the approaches of Rentmeesters [Rentmeesters, 2011] and Su *et al.* [Su et al., 2012]. In case of the latter approach, the data *vs.* smoothness weighting (*i.e.*, λ_1/λ_2) is chosen to achieve an MSE as close as possible (or better) to the best result of our approaches.

Corpus callosum aging. Fig. 3.7 shows the corpus callosum shapes along the fitted curves for the time points in the data. The shapes are recovered from the points along the curve through scaling by the mean singular values of the SVD results. Table 3.2 lists the quantita-

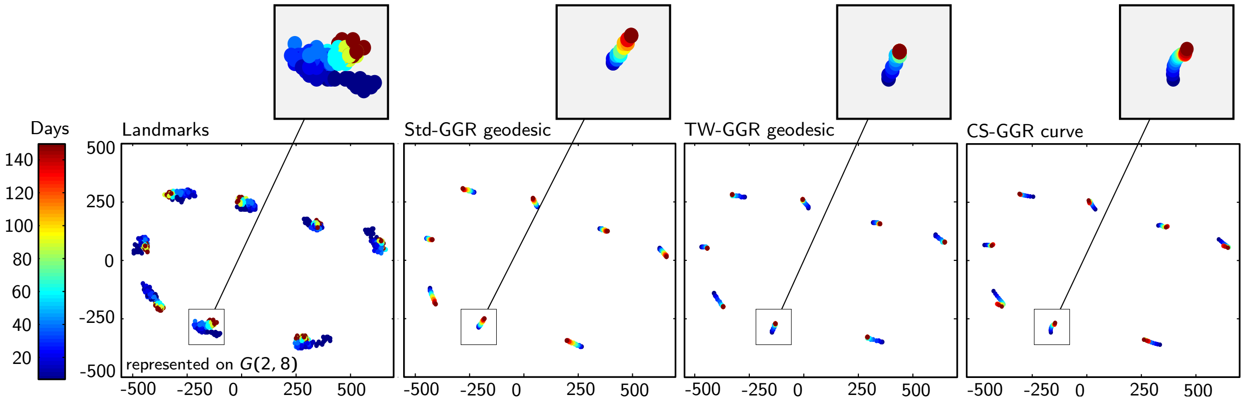


Figure 3.8: Comparison between Std-GGR, TW-GGR and CS-GGR (with one control point) on the rat calvarium data [Bookstein, 1991]. The shapes are generated along the fitted curves and the landmarks are colored by age in days (best-viewed in color).

	Corpus callosum [Fletcher, 2013]					
	[Rentmeesters, 2011]	Std-GGR	TW-GGR	(1)CS-GGR	(2)CS-GGR	[Su et al., 2012]
Energy	0.35	0.35	0.34	0.32	**0.31**	–
R^2	0.12	0.12	0.15	0.21	**0.23**	0.15
MSE(e-2)	1.25	1.25	**1.22**	1.36	1.43	1.25
	Rat calvarium [Bookstein, 1991]					
	[Rentmeesters, 2011]	Std-GGR	TW-GGR	(1)CS-GGR	(2)CS-GGR	[Su et al., 2012]
Energy	0.32	0.32	0.18	**0.16**	**0.16**	–
R^2	0.61	0.61	0.79	0.81	0.81	**0.89**†
MSE(e-3)	2.3	2.3	1.3	**1.2**	**1.2**	4.1†

Table 3.2: Comparison of Std-GGR, TW-GGR and CS-GGR with one (1) and two (2) control points to the approaches of [Rentmeesters, 2011] and [Su et al., 2012] (for $\lambda_1/\lambda_2 = 1/10$). For *Energy* and *MSE* smaller values are better, for R^2 larger values are better. In case of [Su et al., 2012], we fit *one* curve to each individual in the rat calvarium data; MSE and R^2 are then averaged.

tive measurements. With Std-GGR, the corpus callosum starts to shrink from the minimum age 19, which is consistent with the regression results in [Fletcher, 2013, Hinkle et al., 2014]. However, according to biological studies [Hopper et al., 1994, Johnson et al., 1994], the corpus callosum size remains stable during the most active years of the lifespan, which is consistent with our TW-GGR result. As we can see from the optimized logistic function in

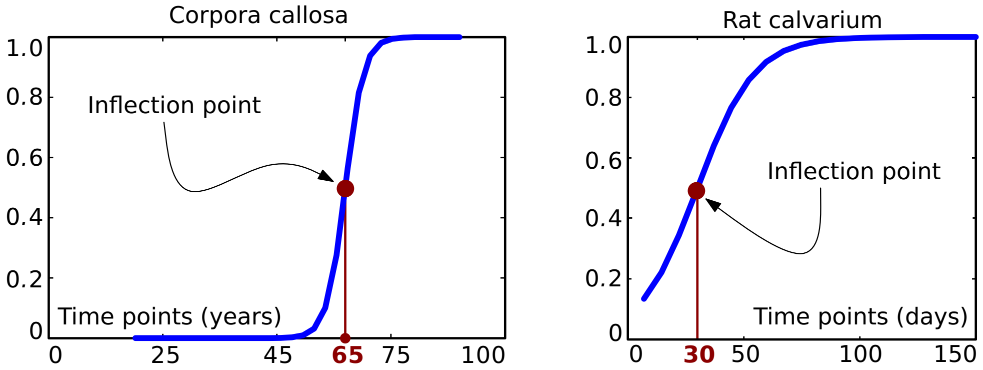


Figure 3.9: Estimated time-warp functions for TW-GGR.

Fig. 3.9 (left), TW-GGR estimates that thinning starts at ≈ 50 years, and at the age of 65, the shrinking rate reaches its peak. From the CS-GGR results reported in Table 3.2, we first observe that the R^2 value increases notably to 0.21/0.23, compared to 0.12 for Std-GGR. While this suggests a better fit to the data, it is not a fair comparison, since the number of parameters for CS-GGR increases as well and a higher R^2 value is expected. Secondly, the more interesting observation is that, qualitatively, we observe higher-order shape changes in the anterior and posterior regions of the corpus callosum, shown in the zoomed-in regions of Fig. 3.7; this is similar to what is reported in [Hinkle et al., 2014] for polynomial regression in 2D Kendall shape space. However, our shape representation, by design, easily extends to point configurations in $\mathbb{R}^3$. This is in contrast to 3D Kendall shape space which has a substantially more complex structure than its 2D variant [Dryden and Mardia, 1998]. Additionally, we notice that the result of [Rentmeesters, 2011] equals the result obtained via Std-GGR (as expected). For [Su et al., 2012], the result is comparable to TW-GGR.

Rat calvarium growth. Fig. 3.8 (leftmost) shows the projection of the original data on $\mathcal{G}(2, 8)$, as well as (part of) the data samples generated along the fitted curves. Table 3.2 lists the performance measures. From the zoomed-in regions in Fig. 3.8, we observe that the rat calvarium grows at an approximately constant speed during the first 150 days if the relationship is modeled by Std-GGR. However, the estimated logistic curve for TW-GGR, shown in Fig. 3.9 (right), indicates that the rat calvarium only grows fast in the first few weeks, reaching its peak at 30 days; then, the rate of growth gradually levels off and becomes steady after around 14 weeks. In fact, similar growth curves for the rat skull were reported in [Hughes et al., 1978]. Based on their study, the growth velocities of viscerocranium length and nurocranium width rose to the peak in the $26 - 32$ days period. Comparing the R^2 values for TW-GGR and CS-GGR, we see an interesting effect: although, we have more parameters in CS-GGR, the R^2 score only marginally improves. This indicates that TW-GGR already sufficiently captures the relationship between age and shape. It further confirms, to a large extent, a hypothesis from [Hinkle et al., 2014], where the authors noted that the cubic polynomial in 2D Kendall shape space degrades to a geodesic under polynomial time re-parametrization. Since TW-GGR re-parametrizes time (not via a cubic polynomial, but via a logistic function), it is not surprising that this relatively simple model exhibits similar performance to the more complex CS-GGR model. Similar to the corpus callosum data (and the synthetic data), [Rentmeesters, 2011] gives the same results as Std-GGR. For [Su et al., 2012], we record an MSE of 4.1e-3, however, the corresponding R^2 score is higher. This can be explained, in part, by the fact that we fit *one* model per individual (as opposed to one model for all individuals) and then average the MSE and R^2 scores. This is done because [Su et al., 2012] cannot handle multiple data instances per time point.

Predicting the independent variable

In the second category of applications the *objective is to predict the independent variable using its regressed relationship with the manifold-valued dependent variable.* Specifically, given a point on $\mathcal{G}(p, n)$, *e.g.*, an LDS representation of a video clip, we search along the

	Traffic speed			
	Baseline	Std-GGR	Std-GGR (PW†)	CS-GGR
Mean energy	–	2554.88	**2461.95**	2670.84
Train-MAE	–	2.98 ± 0.33	**1.48 ± 0.07**	2.42 ± 0.35
Test-MAE	4.14 ± 0.36	4.44 ± 0.16	**3.46 ± 0.64**	6.32 ± 1.62
	Crowd counting			
	Baseline	Std-GGR	Std-GGR (PW†)	CS-GGR
Mean energy	–	273.81	**224.87**	244.02
Train-MAE	–	0.97 ± 0.07	**0.59 ± 0.13**	0.63 ± 0.19
Test-MAE	2.40 ± 0.53	**1.88 ± 0.75**	2.14 ± 1.03	2.11 ± 0.76

Table 3.3: Mean energy and mean absolute errors (MAE) over all CV-folds $\pm 1\sigma$ on training and testing data. Comparisons to [Rentmeesters, 2011] and [Su et al., 2012] were left-out, because [Rentmeesters, 2011] did not converge appropriately and [Su et al., 2012] did not scale to the size of these problems. †PW means piecewise.

regressed curve (with a step size of 0.05, in the same unit of the independent variable in our experiments) to find its closest point, and then we take the corresponding independent variable of this closest point as its predicted value. This could be considered a variant of nearest-neighbor regression where the search space is restricted to the sampled curve. The case when the search space is *not* restricted but contains all data points, will be referred to as our *baseline*. Note that in our case, search complexity is controlled via the step-size, while the search complexity for the *baseline* scales linearly with the sample size.

Furthermore, we remark that in this category of applications, TW-GGR is not appropriate for predicting the independent variable for the following reasons: First, in case of the traffic speed measurement, the generalized logistic function tends to degenerate to almost a step-function, due to the limited number of measurement points in the central regions. In other words, two greatly different independent variables would correspond to two very close data points, even the same one, which would result in a large prediction error. Second, in case of crowd-counting, there is absolutely no prior knowledge about any saturation

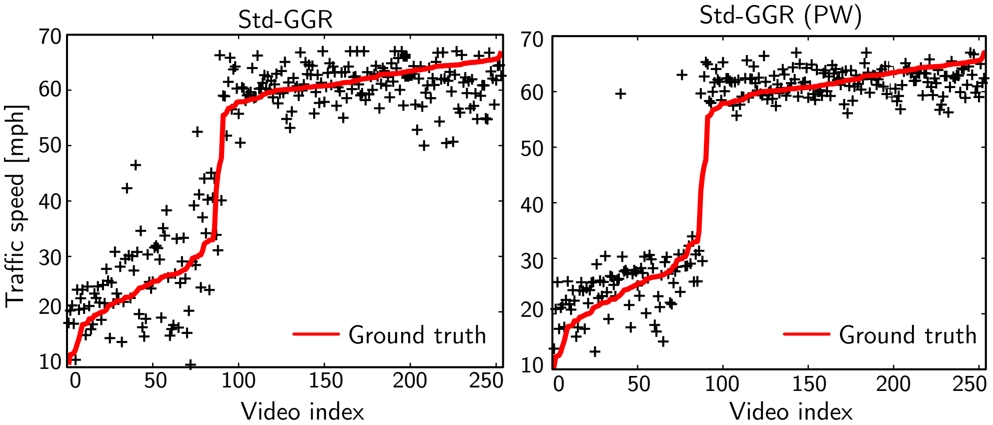


Figure 3.10: Traffic speed predictions via 5-fold CV. The red solid curve shows the ground truth (best-viewed in color).

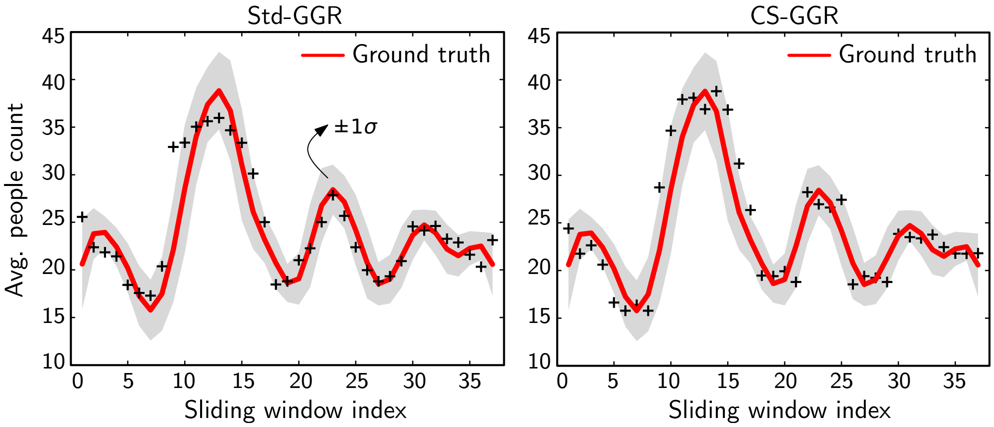


Figure 3.11: Crowd counting results via 4-fold CV. Predictions are shown as a function of the sliding window index. The gray envelope indicates the weighted standard deviation ($\pm 1\sigma$) around the average crowd size in a sliding window (best-viewed in color).

or growth effect which could be modeled via a logistic function. Consequently, we only demonstrate Std-GGR, piecewise Std-GGR, and CS-GGR on the two datasets. Note that prediction based on nearest neighbors could be problematic in case of CS-GGR, since the model does not guarantee a monotonic curve. We report the mean regression *energy* and the *mean absolute error (MAE)*, computed over all folds in a cross-validation setup with a dataset-dependent number of folds.

Speed prediction. For each video clip, we estimate LDS models with $p = 10$ states. The control point of CS-GGR and the breakpoint for piecewise Std-GGR are set at 50 [mph].

Results are reported for 5-fold CV, see Fig. 3.10. The quantitative comparison to the baseline in Table 3.3 shows that piecewise Std-GGR has the best performance.

Crowd counting. For each video clip, we estimate LDS models with $p = 10$ states using weighted system identification. The control point of CS-GGR and the breakpoint for piecewise Std-GGR are set to a count of 23 people which separates the 37 videos into two groups of roughly equal size. Quantitative results for 4-fold CV are reported in Table 3.3. Fig. 3.11 shows the predictions *vs.* the ground truth; and both Std-GGR and CS-GGR output predictions "close" to the ground truth, mostly within 1σ (shaded region) of the average crowd count. However, a closer look at Table 3.3 reveals a typical overfitting effect for CS-GGR: while the training MAE is quite low, the testing MAE is higher than for the simpler Std-GGR approach. Even though both models exhibit comparable performance (considering the standard deviations), Std-GGR is preferable, due to fewer parameters and its guaranteed monotonic regression curve.

3.4 Regression on Image Time-Series

The manifold of diffeomorphisms is another type of Riemannian manifold, which is commonly used to analyze images. In this section, we generalize linear regression to the manifold of diffeomorphisms, capturing spatial and intensity changes *simultaneously* in image time-series. This is achieved by combining the dynamical systems formulation for geodesic regression for images [Niethammer et al., 2011] with image metamorphosis [Holm et al., 2009, Miller and Younes, 2001]. Next, we start with the dynamical systems formulation for metamorphism to obtain the optimality conditions, that is, the dynamic constraints, for regression. And then we generalize the concept of a regression line to image time-series.

3.4.1 Metamorphosis

Starting from the dynamical systems formulation for LDDMM image registration

$$E(v) = \frac{1}{2}\int_0^1 \|v\|_L^2 \, dt + \frac{1}{\sigma^2}\|I(1) - I_1\|^2, \quad \text{s.t.} \;\; I_t + \nabla I^T v = 0, \; I(0) = I_0, \tag{3.39}$$

image metamorphosis allows exact matching of a target image I_1 by a warped and intensity-adjusted source image $I(1)$ by adding a control variable, q, which smoothly adjusts image intensities along streamlines. Here, $\sigma > 0$, v is a spatiotemporal velocity field and $\|v\|_L^2 = \langle Lv, Lv\rangle$, where L is a differential operator penalizing non-smooth velocities. The optimization problem changes to [Miller and Younes, 2001, Holm et al., 2009]

$$E(v, q) = \frac{1}{2}\int_0^1 \|v\|_L^2 + \rho\|q\|_Q^2 \, dt, \quad \text{s.t.} \;\; I_t + \nabla I^T v = q, \; I(0) = I_0, \; I(1) = I_1. \tag{3.40}$$

The inexact match of the final image is replaced by an exact matching, hence the energy value depends on the images only implicitly through the initial and final constraints; $\rho > 0$ controls the balance between intensity blending and spatial deformation. The solution to both minimization problems Eq. (3.39) and Eq. (3.40) is given by a geodesic, which is specified by its initial conditions. The initial conditions are numerically computed through a shooting strategy.

Optimality conditions for shooting metamorphosis. To derive the second order dynamical system required for the shooting method, we add the dynamic constraint through the momentum variable, p. Eq. (3.40) becomes

$$E(v, q, I, p) = \int_0^1 \frac{1}{2}\|v\|_L^2 + \frac{1}{2}\rho\|q\|_Q^2 + \langle p, I_t + \nabla I^T v - q\rangle \, dt, \quad \text{s.t.} \;\; I(0) = I_0, \; I(1) = I_1. \tag{3.41}$$

To simplify the numerical implementation of the problem, we use an augmented Lagrangian approach [Nocedal and Wright, 2006] converting the optimization problem (3.41) to

$$E(v,q,I,p) = \int_0^1 \frac{1}{2}\|v\|_L^2 + \frac{1}{2}\rho\|q\|_Q^2 + \langle p, I_t + \nabla I^T v - q\rangle \; dt \\ - \langle r, I(1) - I_1\rangle + \frac{\mu}{2}\|I(1) - I_1\|^2, \;\; \text{s.t.} \;\; I(0) = I_0, \quad (3.42)$$

where $\mu > 0$ and r is the Lagrangian multiplier function for the final image-match constraint. The variation of Eq. (3.42) results in the optimality conditions

$$\begin{cases} I_t + \nabla I^T v &= \frac{1}{\rho}(Q^\dagger Q)^{-1}p, \; I(0) = I_0, \\ -p_t - div(pv) &= 0, \; p(1) = r - \mu(I(1) - I_1), \\ L^\dagger L v + p\nabla I &= 0 \;\; . \end{cases} \quad (3.43)$$

The optimality conditions do not depend on q, since by optimality $q = \frac{1}{\rho}(Q^\dagger Q)^{-1}p$. Hence, the state for metamorphosis is identical to the state of LDDMM, i.e., (I,p), highlighting the tight coupling in metamorphosis between image deformation and intensity changes. The final state constraint $I(1) = I_1$ has been replaced by an augmented Lagrangian penalty function.

Shooting for metamorphosis

The metamorphosis problem Eq. (3.40) has so far been addressed as a boundary value problem by relaxation approaches [Garcin and Younes, 2005, Miller and Younes, 2001]. This approach hinders the formulation of the regression problem and assures geodesics at convergence only. We propose a shooting method instead. Since the final constraint has been successfully eliminated through the augmented Lagrangian approach, $\nabla_{p(0)}E$ can be computed using a first- or second-order adjoint method similarly as for LDDMM registration [Vialard et al., 2012, Ashburner and Friston, 2011]. The numerical solution alternates between a descent step for $p(0)$ for fixed r, μ and (upon reasonable convergence) an update step $r^{(k+1)} = r^{(k)} - \mu^{(k)}(I(1) - I_1)$. The penalty parameter μ is increased as desired such that $\mu^{(k+1)} > \mu^{(k)}$. Numerically, we solve all equations by discretizing time, assuming v and p to be piece-wise constant in a time-interval. We solve transport equations and scalar conservation laws by propagating maps [Beg et al., 2005] to limit numerical dissipation.

First-order adjoint method. Following [Ashburner and Friston, 2011], we can compute $\nabla_{v(0)}E$ by realizing that the Hilbert gradient is $\nabla_{v(0)}E = v(0) + K * (p(0)\nabla I(0))$, where $K = (L^\dagger L)^{-1}$. Therefore, based on the adjoint solution method [Beg et al., 2005, Hart et al., 2009]

$$\nabla_{v(0)}E = v(0) + K * (\hat{p}(0)\nabla I(0)) = v(0) + K * (|D\Phi|\hat{p}(1) \circ \Phi\nabla I(0)), \tag{3.44}$$

where Φ is the map from $t = 1$ to $t = 0$ given the current estimate of the velocity field $v(x,t)$ and $\hat{p}(1) = r - \mu(I(1) - I_1)$ with $I(1) = I_0 \circ \Phi^{-1}$. Storage of the time-dependent velocity fields is not required as both Φ and Φ^{-1} can be computed and stored during a forward (shooting) sweep. Instead of performing the gradient descent on $v(0)$ it is beneficial to compute it directly with respect to $p(0)$ since this avoids unnecessary matrix computation. Since at $t = 0$: $-(L^\dagger L)\delta v(0) = \delta p(0)\nabla I(0)$, it follows from Eq. (3.44) that

$$\nabla_{p(0)}E = p(0) - \hat{p}(0) = p(0) - |D\Phi|(r - \mu(I(1) - I_1)) \circ \Phi. \tag{3.45}$$

Second-order adjoint method. The energy can be rewritten in initial value form (w.r.t.

$(I(0), p(0)))$ as

$$\begin{aligned} E \ = \ & \frac{1}{2}\langle p(0)\nabla I(0), K * (p(0)\nabla I(0))\rangle + \frac{1}{2\rho}\langle (Q^{\dagger}Q)^{-1}p(0), p(0)\rangle \\ & -\langle r, I(1) - I_1\rangle + \frac{\mu}{2}\|I(1) - I_1\|^2, \ \ \text{s.t. Eq. (3.43) holds.} \end{aligned} \tag{3.46}$$

At optimality, the state equations (3.43) and

$$\begin{cases} -\lambda^I_t - div(v\lambda^I) & = div(pK * \lambda^v), \\ -\lambda^p_t - v^T\nabla\lambda^p & = -\nabla I^T K * \lambda^v + \frac{1}{\rho}(Q^{\dagger}Q)^{-1}\lambda^I, \\ \lambda^I\nabla I - p\nabla\lambda^p + \lambda^v & = 0, \end{cases} \tag{3.47}$$

hold, with final conditions: $\lambda^p(1) = 0$; $\lambda^I(1) = r - \mu(I(1) - I_1)$. The gradient is

$$\nabla_{p(0)}E = -\lambda^p(0) + \nabla I(0)^T K * (p(0)\nabla I(0)) + \frac{1}{\rho}(Q^{\dagger}Q)^{-1}p(0). \tag{3.48}$$

The dynamic equations and the gradient are only slightly changed from the LDDMM registration [Vialard et al., 2012] when following the augmented Lagrangian approach.

3.4.2 Metamorphic Geodesic Regression

Our goal is the estimation of a regression geodesic (under the geodesic equations for metamorphosis) w.r.t. a set of measurement images $\{I_i\}$ by minimizing

$$E = \frac{1}{2}\langle m(t_0), K * m(t_0)\rangle + \frac{1}{2\rho}\langle (Q^{\dagger}Q)^{-1}p(t_0), p(t_0)\rangle + \frac{1}{\sigma^2}\sum_{i=1}^{N} \text{Sim}(I(t_i), I_i) \tag{3.49}$$

such that Eq. (3.43) holds. Here, $\sigma > 0$ balances the influence of the change of the regression geodesic with respect to the measurements, $m(t_0) = p(t_0)\nabla I(t_0)$ and Sim denotes an image similarity measure. A solution scheme with respect to $(I(t_0), p(t_0))$ can be obtained following the derivations for geodesic regression [Niethammer et al., 2011]. Such a solution requires the integration of the state equation as well as the second-order adjoint. Further, for metamorphosis it is sensible to also define $\mathrm{Sim}(I(t_i), I_i)$ based on the squared distance induced by the solution of the metamorphosis problem between $I(t_i)$ and I_i. Since no closed-form solutions for these distances are computable in the image-valued case an iterative solution method is required which would in turn require the underlying solution of metamorphosis problems for each measurement at each iteration. This is costly.

Approximated metamorphic geodesic regression

To simplify the solution of metamorphic geodesic regression (3.49), we approximate the distance between two images I_1, I_2 w.r.t. a base image I_b at time t as

$$\begin{aligned} Sim(I_1, I_2) = d^2(I_1, I_2) \approx t^2 \frac{1}{2} \langle m_1(0) - m_2(0), K * (m_1(0) - m_2(0)) \rangle \\ + t^2 \frac{1}{2\rho} \langle (Q^{\dagger}Q)^{-1}(p_1(0) - p_2(0)), p_1(0) - p_2(0) \rangle, \quad (3.50) \end{aligned}$$

where $p_1(0)$ and $p_2(0)$ are the initial momenta for I_1 and I_2 w.r.t. the base image I_b (i.e., the initial momenta obtained by solving the metamorphosis problem between I_b and I_1 as well as for I_b and I_2 respectively) and $m_1(0) = p_1(0)\nabla I_b$, $m_2(0) = p_2(0)\nabla I_b$. This can be seen as a tangent space approximation for metamorphosis. The squared time-dependence emerges because the individual difference terms are linear in time.

We assume that the initial image $I(t_0)$ on the regression geodesic is known. This simplifying assumption is meaningful, for example, for growth modeling w.r.t. a given base image[5]. Substituting into Eq. (3.49) yields

$$E(p(t_0)) = \frac{1}{2}\langle m(t_0), K * m(t_0)\rangle + \frac{1}{2\rho}\langle (Q^\dagger Q)^{-1}p(t_0), p(t_0)\rangle$$
$$+ \frac{1}{\sigma^2}\sum_{i=1}^{N}\frac{1}{2}(\Delta t_i)^2\langle m(t_0) - m_i, K * (m(t_0) - m_i)\rangle + \frac{1}{2\rho}(\Delta t_i)^2\langle (Q^\dagger Q)^{-1}(p(t_0) - p_i), p(t_0) - p_i\rangle. \tag{3.51}$$

Here, $m(t_0) = p(t_0)\nabla I(t_0)$, $\Delta t_i = t_i - t_0$, $m_i = p_i \nabla I(t_0)$ and p_i is the initial momentum for the metamorphosis solution between $I(t_0)$ and I_i. For a given $I(t_0)$, the p_i can be independently computed. The approximated energy only depends on the initial momentum $p(t_0)$. The energy variation yields the condition

$$R[(1 + \frac{1}{\sigma^2}\sum_{i=1}^{N}(\Delta t_i)^2)p(t_0)] = R[\frac{1}{\sigma^2}\sum_{i=1}^{N}(\Delta t_i)^2 p_i], \tag{3.52}$$

where the operator R is $R[p] := \nabla I(t_0)^T K * (\nabla I(t_0)p) + \frac{1}{\rho}(Q^\dagger Q)^{-1}p$. Since $K = (L^\dagger L)^{-1}$ and $\rho > 0$ this operator is invertible and therefore

$$p(t_0) = \frac{\frac{1}{\sigma^2}\sum_{i=1}^{N}(\Delta t_i)^2 p_i}{1 + \frac{1}{\sigma^2}\sum_{i=1}^{N}(\Delta t_i)^2} \overset{\sigma \to 0}{\approx} \frac{\sum_{i=1}^{N}(\Delta t_i)^2 p_i}{\sum_{i=1}^{N}(\Delta t_i)^2}. \tag{3.53}$$

The last approximation is sensible since typically $\sigma << 1$. It recovers the metamorphosis solution if there is only one measurement image and the base image.

[5] Ideally one would like to construct an image on the geodesic given all the measurement images and then perform all computations with respect to it. For the linear regression model the point defined by the mean in time and the measurements, $(\bar{t}, \bar{y})$, is on the regression line. If such a relation exists for metamorphic geodesic regression, e.g., some form of unbiased mean with similar properties, remains to be determined.

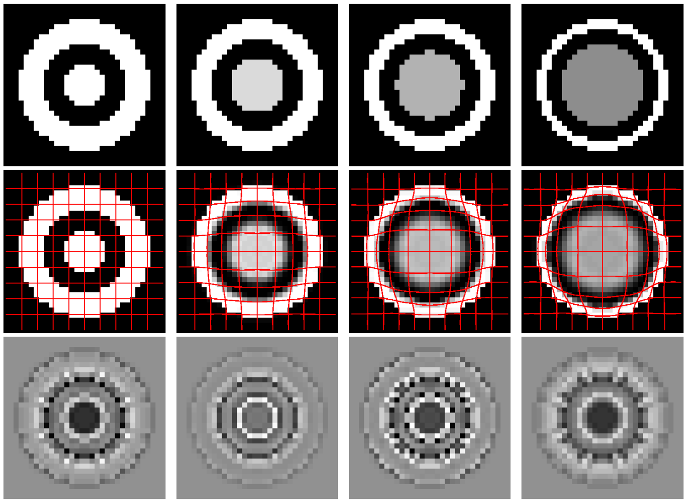


Figure 3.12: Bull's eye metamorphic regression experiment. Measurement images (top row). Metamorphic regression result (middle row) and momenta (bottom row). The first image is chosen as the base image. Momenta images: left: time-weighted average of the initial momenta; right: momenta of the measurement images with respect to the base image.

3.4.3 Experimental Results

Simulated Examples. In Fig. 3.12, four images (32 × 32, spacing 0.04) are synthesized to simulate the movement of a bull's eye. The outside white loop of the eye shrinks with no intensity changes, while the inside circle grows at a constant speed and its intensity changes from white to gray. The images are at time instants 0, 10, 20, 30 and we chose the first one as the base image. Eight Gaussian kernels [Risser et al., 2011] are used for K: $\{K_{0.5}, K_{0.4}, K_{0.3}, K_{0.25}, K_{0.2}, K_{0.15}, K_{0.1}, K_{0.05}\}$; $\rho = 0.75$. The result confirms that the spatial transformation and intensity changes are captured simultaneously. The dark solid circle at the center of the average momentum of Fig. 3.12 indicates that the intensity of the inside circle will decrease gradually. The white loop outside of the dark area captures the growth of the inside circle.

Fig. 3.13 shows a square (64 × 64, spacing 0.02) moving from left to right at a uniform speed with gradually decreasing intensity. Measurements are at 0, 10, 20, 30, 40. We used

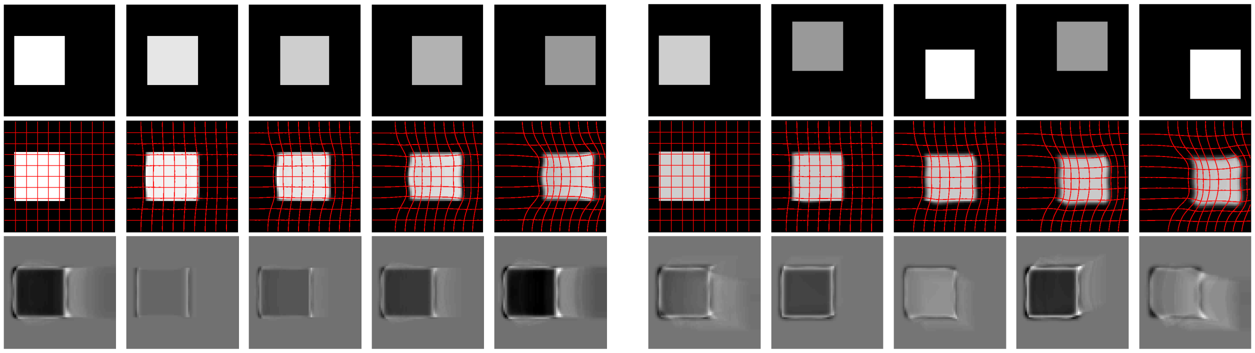


Figure 3.13: Square metamorphic regression experiment. *Left*: moving square with decreasing intensities and no oscillations during movement; *Right*: moving and oscillating square with alternating intensities. In both cases the base image is the first one. Top row: measurement images, middle row: metamorphic regression results, bottom row: momenta images (left: time-weighted average of the initial momenta, to the right: momenta of the measurement images with respect to the base image).

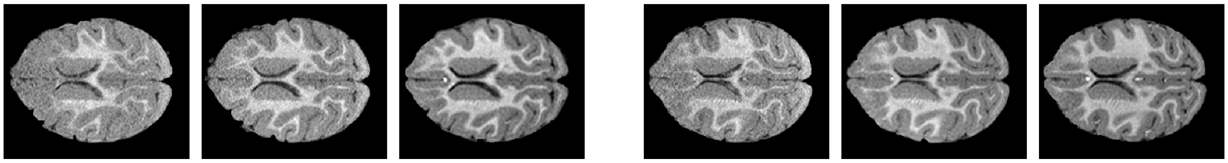


Figure 3.14: Two representative image scans at 3, 6 and 12 months (left to right).

a multi-Gaussian kernel K with $\{K_{1.0}, K_{0.75}, K_{0.5}, K_{0.4}, K_{0.3}, K_{0.2}, K_{0.1}\}$ and set $\rho = 5.0$. Metamorphic regression successfully captures the spatial transformation and the intensity changes of the square even when adding vertical oscillations. As expected, metamorphic regression eliminates the oscillations while capturing the intensity change and the movement to the right. We see from the time-weighted average of the initial momenta that intensity changes are controlled by the values inside the square region (dark: decreasing intensity; bright: increasing intensity). The spatial transformations are mainly controlled by the momenta on the edges of the square.

Real Images. Fig. 3.14 shows two representative longitudinal MRI time-series (300×250 with spacing 0.2734 mm) of nine macaque monkeys at age 3, 6, and 12 months. Some subjects have no visible myelination in the anterior parts of the brain at 3 months, while others show substantial myelination. Here, we use metamorphic geodesic regression not for an individual

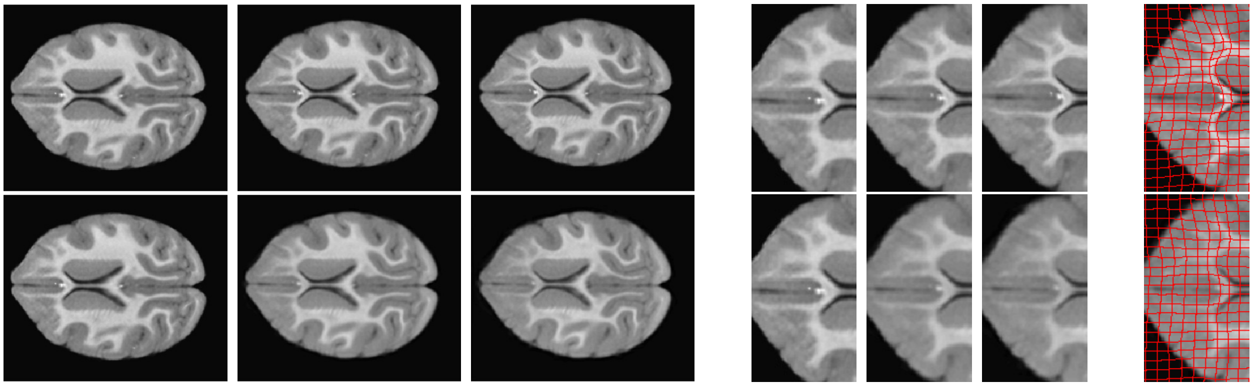


Figure 3.15: Regression results for monkey data: LDDMM (top) metamorphosis (bottom). (a) Images on geodesic at 12, 6, 3 months; (b) Zoom in for images on geodesic at 12, 6, 3 months; (c) Zoom in for images at 3 months to illustrate spatial deformation.

longitudinal image set, but for all nine monkeys and all time-points simultaneously. We use an unbiased atlas for images at 12 months as the base image. Metamorphic geodesic regression is applied over the remaining 18 images at 3 and 6 months. We use a multi-Gaussian kernel, K, with $\{K_{40}, K_{20}, K_{15}, K_{10}, K_{5}, K_{2.5}\}$, and $\rho = 500$.

Fig. 3.15 shows regression results for the simple metamorphic model and for its LDDMM version [Hong et al., 2012b] which cannot capture intensity changes. The metamorphic regression geodesic captures intensity changes of the brain well (increase in white matter intensity with age caused by myelination) while capturing spatial deformations, most notably a subtle expansion of the ventricles.

3.5 Model Criticism

In previous sections, to measure the goodness-of-fit of these regression models, we compute the sum of squared errors (SSE, or energy), R^2, and the mean squared error (MSE) using cross-validation. While these measures assess quality of fit, they do not directly check if the underlying model assumptions hold. In contrast, *model criticism* does exactly that: it checks if a model's assumptions (including the noise model) are consistent with the observed data. It thereby provides valuable additional information beyond the classical measures for model fit. Recently, a statistical model criticism approach [Lloyd and Ghahramani, 2015] using a kernel-based two-sample test [Gretton et al., 2012] has been proposed and its utility to evaluate regression models in Euclidean space was demonstrated.

In this section, we take this approach one step further and use the fact that the strategy of [Lloyd and Ghahramani, 2015] depends on a suitable kernel function for the data and can thus be extended to manifolds given such a kernel function. For instance, given a population of data samples on the Grassmannian, we (1) perform regression, then (2) generate samples from the regression model and (3) assess whether the observed data could have been generated by the fitted model. We demonstrate the approach by criticizing our three regression models on the Grassmannian using both synthetic and real data. We argue that model criticism approach is complementary to traditional measures of model fit, but it has the advantage of directly assessing the suitability of a statistical model and its fit for given observed data.

3.5.1 Model Criticism for Regression in Euclidean Space

The objective of model criticism for regression is to test for discrepancies between observed data and a model estimated from the data [Lloyd and Ghahramani, 2015]. We assume the data observations are independently and identically distributed (i.i.d.) and that we can draw i.i.d. samples from the model respecting the noise assumptions; then, the key ingredient of model criticism is to measure whether these two samples are drawn from the same underlying distribution. To perform this *two-sample test*, a “kernelized” variant of the maximum mean discrepancy (MMD) [Gretton et al., 2012] has been proposed as one choice of the test-statistic.

Review of kernel-based two-sample testing [Gretton et al., 2012]. Assume we have i.i.d. samples $X = \{x_i\}_{i=1}^m$ and $Y = \{y_i\}_{i=1}^n$, drawn randomly from distributions p, q, defined on a domain $\mathcal{X}$. The goal of two-sample testing is to assess if $p = q$. One choice of a test-statistic is the MMD, defined as

$$\mathrm{MMD}[\mathcal{F}, p, q] = \sup_{f \in \mathcal{F}} (\mathbb{E}_{x \sim p}[f(x)] - \mathbb{E}_{y \sim q}[f(y)]) \ , \tag{3.54}$$

where $\mathcal{F}$ is a suitable class of functions $f : \mathcal{X} \to \mathbb{R}$. To uniquely measure whether $p = q$, Gretton et al. [Gretton et al., 2012] let $\mathcal{F}$ be the unit ball in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$, i.e., $\mathcal{F} = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} \leq 1\}$, with associated reproducing kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. The kernel can be written as $k(x, y) = \langle \phi(x), \phi(y) \rangle_{\mathcal{H}}$, where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ is the inner product in $\mathcal{H}$ and $\phi : \mathcal{X} \to \mathcal{H}$ denotes the feature map.

According to [Gretton et al., 2012], Eq. (3.54) can then be expressed as the RKHS distance between the mean embeddings $\mu[p]$ and $\mu[q]$ of the distributions p and q, in particular $\mu[p] := \mathbb{E}_{x \sim p}[\phi(x)]$. Since the mean embedding satisfies $\mathbb{E}_{x \sim p}[f(x)] = \langle \mu[p], f \rangle_{\mathcal{H}}$, Eq. (3.54) can be written as

$$\mathrm{MMD}[\mathcal{F}, p, q] = \sup_{\|f\|_{\mathcal{H}} \leq 1} \langle \mu[p] - \mu[q], f \rangle = \|\mu[p] - \mu[q]\|_{\mathcal{H}} \tag{3.55}$$

with the empirical estimate (using the kernel function k)

$$\widehat{\mathrm{MMD}}[\mathcal{F}, X, Y] = \left[\frac{1}{m^2} \sum_{i,j=1}^{m} k(x_i, x_j) - \frac{2}{mn} \sum_{i,j=1}^{m,n} k(x_i, y_j) + \frac{1}{n^2} \sum_{i,j=1}^{n} k(y_i, y_j) \right]^{\frac{1}{2}} . \tag{3.56}$$

In the following, we omit $\mathcal{F}$ and let $\widehat{\mathrm{MMD}}(X, Y)$ denote the computation of the MMD statistic on two samples X and Y using a suitable kernel function k.

Model criticism using two-sample testing. Assume we have data observations $\{Y_i^{obs}\}_{i=1}^N$, drawn from some distribution p, conditioned on an associated independent value t_i. A regression model M is estimated from the tuples $\{(t_i, Y_i^{obs})\}_{i=1}^N$. If the regression model is based on a Gaussian noise assumption[6], we can generate i.i.d. samples from the model; let this distribution be denoted by q and a sample by $\{Y_i^{est} = M(t_i) + n_i\}_{i=1}^N$, where $n_i \sim N(0, \sigma^2)$ and σ is the standard deviation of the residuals. Criticizing the model M now means to perform a two-sample test between $\{(t_i, Y_i^{obs})\}_{i=1}^N$ and $\{(t_i, Y_i^{est})\}_{i=1}^N$ under the null-hypothesis $H_0 : p = q$. This is done by computing the test-statistic between data observations and samples drawn from the regression model, i.e., $T^* = \widehat{\mathrm{MMD}}(\{(t_i, Y_i^{obs})\}_{i=1}^N, \{(t_i, Y_i^{est})\}_{i=1}^N)$. Then, to obtain the distribution of T under H_0, we repeatedly draw (from q) N i.i.d. samples to form two populations, $\{(t_i, Y_i^a)\}_{i=1}^N$ and $\{(t_i, Y_i^b)\}_{i=1}^N$. For each such draw j, we compute $T_j = \widehat{\mathrm{MMD}}(\{(t_i, Y_i^a)\}_{i=1}^N, \{(t_i, Y_i^b)\}_{i=1}^N)$ and thereby obtain the empirical distribution of T under H_0. Note that in case of observations in $\mathbb{R}^n$, computing the MMD statistic for model criticism is straightforward, since we can simply add t_i as an additional dimension to our data, i.e., we obtain observations in $\mathbb{R}^{n+1}$. The well-known RBF kernel can then be used to compute Eq. (3.56). We will see that this needs to be handled differently on manifolds.

Finally, we count the number of times that the bootstrapped statistics under H_0 are larger than the test statistic T^*, which results in a p-value estimate. Because T^* will be affected by the added random noise, we can also sample it a large number of times, resulting in a distribution of T^* and associated p-values.

[6] Other noise models can also be used as long as one can sample from them.

3.5.2 Model Criticism for Regression on the Grassmannian

We now extend the model criticism approach of the previous section to the Grassmann manifold. The test objects are regression models on the Grassmannian, i.e., generalizations of classical regression in Euclidean space which minimize the sum of squared (geodesic) distances to the regression curves. Noise in these models is assumed to be Gaussian. To generalize the model criticism idea, one key ingredient is to draw random samples on the Grassmannian at each point on the regression curves. The other key ingredient is a suitable kernel k for Eq. (3.56) and a strategy to include the independent value into the kernel.

Drawing random samples on the Grassmannian. Similar to the Euclidean case, we assume we have N data observations on the Grassmannian $\mathcal{G}(r,s)$[7] with associated independent values, i.e., $\{(t_i, \mathbf{Y}_i^{obs})\}_{i=1}^N$. Using a regression model M estimated from this data, we can compute the corresponding data points on the regressed curve for each t_i as $\bar{\mathbf{Y}}_i^{est} = M(t_i)$. To draw sample points at each t_i under a Gaussian noise model, we adhere to the following strategy (although, other approaches such as the one outlined in [Zhang and Fletcher, 2013] are possible). First, we compute the empirical standard deviation of the residuals as

$$\sigma = \sqrt{\frac{\sum_{i=1}^N d^2(\mathbf{Y}_i^{obs}, \bar{\mathbf{Y}}_i^{est})}{N-1}} \;, \tag{3.57}$$

where $d(\cdot,\cdot)$ denotes the geodesic distance on $\mathcal{G}(r,s)$. For each estimated data point $\bar{\mathbf{Y}}_i^{est}$, we then generate a tangent vector $\dot{\mathbf{Y}}_i^{est}$ as the projection of an $s \times r$ random matrix $\mathbf{Z}_i = [z_{uv}], z_{uv} \sim \mathcal{N}(0, \hat{\sigma}^2)$ onto the tangent space at $\bar{\mathbf{Y}}_i^{est}$; this is done via $\dot{\mathbf{Y}}_i^{est} = (\mathbf{I}_s -$

[7] The Grassmannian $\mathcal{G}(r,s)$ is the manifold of r-dimensional subspaces of $\mathbb{R}^s$. A point on $\mathcal{G}(r,s)$ is identified by an $s \times r$ matrix $\mathbf{Y}$ with $\mathbf{Y}^\top \mathbf{Y} = \mathbf{I}_r$.

$\bar{\mathbf{Y}}_i^{est}(\bar{\mathbf{Y}}_i^{est})^\top)\mathbf{Z}_i$ where $\mathbf{I}_s$ is the $s \times s$ identity matrix. The random point $\mathbf{Y}_i^{est}$ (at t_i), is eventually computed via the Riemannian exponential map as

$$\mathbf{Y}_i^{est} = \mathrm{Exp}(\bar{\mathbf{Y}}_i^{est}, \dot{\mathbf{Y}}_i^{est}) \ . \tag{3.58}$$

We note that the standard deviation $\hat{\sigma}$ of the samples in $\mathbf{Z}_i$ is proportional to the standard deviation σ as computed by Eq. (3.57). In fact, it can be shown[8] that the resulting geodesic distance between $\bar{\mathbf{Y}}_i^{est}$ and $\mathbf{Y}_i^{est}$ has standard deviation $\hat{\sigma}\sqrt{rs}$. Consequently, we set $\hat{\sigma} = \sigma/\sqrt{rs}$ when creating the $\mathbf{Z}_i$.

Kernels for model criticism on the Grassmannian. The next step is to adjust the kernel-based two-sample test of [Gretton et al., 2012] for model criticism on the $\mathcal{G}(r,s)$ by selecting a suitable kernel. In [Harandi et al., 2014], several positive definite (and universal) kernels on $\mathcal{G}(r,s)$ have been proposed, e.g., RBF, Laplace and Binomial kernels. These kernels are constructed by using the Binet-Cauchy kernel k_{bc} [Wolf and Shashua, 2003] and the projection kernel k_p [Hamm and Lee, 2008]. In our experiments we selected the kernel $k_{l,p}(\mathbf{X},\mathbf{Y}) = \exp\left(-\beta\sqrt{r - \|\mathbf{X}^\top\mathbf{Y}\|_F^2}\right)$, $\beta > 0$, from [Harandi et al., 2014][9]. However, for the model criticism approach that was proposed in [Lloyd and Ghahramani, 2015] to be used on regression models, we need to be able to compute the MMD test for $(t_i, \mathbf{Y}_i)$ (i.e., including the information about the independent variable). While this is simple in the Euclidean case (cf. Section 3.5.1), the situation on manifolds is more complicated since we cannot

[8] Say we have samples $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ with Fréchet mean $\bar{\mathbf{Y}}$. The Fréchet variance is defined as $\sigma^2 = \min_{\mathbf{Y}\in\mathcal{G}(r,s)} {}^1\!/\!_n \sum_{i=1}^n d^2(\mathbf{Y},\mathbf{Y}_i)$; this can equivalently be written as $\sigma^2 = \frac{1}{n}\sum_{i=1}^n \mathrm{tr}(\dot{\mathbf{Y}}_i^\top\dot{\mathbf{Y}}_i)$ with $\dot{\mathbf{Y}}_i = \mathrm{Exp}^{-1}(\bar{\mathbf{Y}},\mathbf{Y}_i)$. Expanding the trace yields $\sigma^2 = \sum_{j=1}^s\sum_{k=1}^r[{}^1\!/\!_n \sum_{i=1}^n y_{i,j,k}^2]$ where the term in brackets is $\hat{\sigma}^2$ since $y_{i,j,k}$ is i.i.d. as $\mathcal{N}(0,\hat{\sigma}^2)$. We finally obtain $\sigma^2 = rs\hat{\sigma}^2$.

[9] Here, r denotes the subspace dimension in $\mathcal{G}(r,s)$, as defined before.

simply add t_i to our observations. Our strategy is to use an RBF kernel for the t_i, i.e., $k_{rbf}(t, t') = \exp(-(t-t')^2/(2\gamma^2))$ and then leverage the closure properties of positive definite kernels, which allow multiplication of positive definite kernels. This yields our *combined kernel* as

$$k((t_i, \mathbf{X}), (t_j, \mathbf{Y})) = \exp\left(-\frac{(t_i - t_j)^2}{2\gamma^2}\right) \cdot \exp\left(-\beta\sqrt{r - \|\mathbf{X}^\top \mathbf{Y}\|_F^2}\right) \quad . \tag{3.59}$$

In all the experiments, we set both β and γ to 1 for simplicity.

Model criticism on the Grassmannian. The computational steps for model criticism on the Grassmannian are listed below:

(1) Compute the points $\{\bar{\mathbf{Y}}_i^{est} = M(t_i)\}_{i=1}^N$ for each data observation $\mathbf{Y}_i^{obs}$ on $\mathcal{G}(r, s)$ from the estimated regression model M.

(2) Estimate the standard deviation of residuals, σ, using Eq. (3.57).

(3) Generate noisy samples, $\mathbf{Y}_i^{est}$ at each t_i using Eq. (3.58) and $\hat{\sigma}^2 = \sigma^2/(rs)$.

(4) Compute $T^* = \widehat{\text{MMD}}(\{(t_i, \mathbf{Y}_i^{est})\}_{i=1}^N, \{(t_i, \mathbf{Y}_i^{obs})\}_{i=1}^N)$ using Eqs. (3.56) & (3.59).

(5) Repeat (3) and (4) many times to obtain a population of T^*.

(6) Generate two groups of samples using (3), $\{(t_i, \mathbf{Y}_i^a)\}_{i=1}^N$, and $\{(t_i, \mathbf{Y}_i^b)\}_{i=1}^N$, and compute $T = \widehat{\text{MMD}}(\{(t_i, \mathbf{Y}_i^a)\}_{i=1}^N, \{(t_i, \mathbf{Y}_i^b)\}_{i=1}^N)$.

(7) Repeat (6) many times to obtain a distribution of T under H_0.

(8) Compute a p-value for each T^* in (5) with respect to the distribution of T from (7), resulting in a population of p-values. This allows rejecting the null-hypothesis that the observed data distribution is the same as the sampled data distribution of the model at a chosen significance level α (e.g., 0.05).

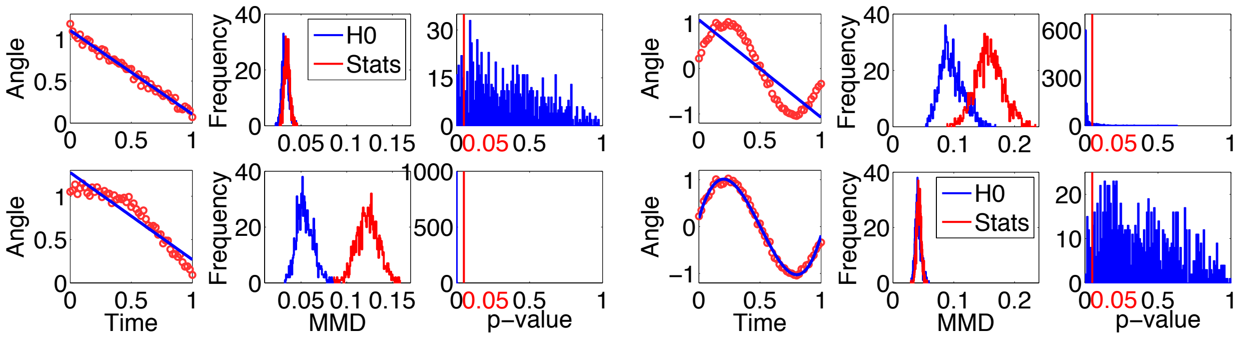
3.5.3 Experimental Results

We criticize three different regression models on the Grassmannian discussed in the previous section: Std-GGR, TW-GGR, and CS-GGR, which are generalizations of linear least squares, time-warped, and cubic-spline regression, respectively.

Synthetic data. For easy visualization, we synthesize data on the Grassmannian $\mathcal{G}(1, 2)$, i.e., the space of lines through the origin in $\mathbb{R}^2$. Each point uniquely corresponds to an angle in $[-\pi/2, \pi/2]$ with respect to the horizontal axis.

(1) Different data distributions. The first experiment is to perform model criticism on *one* regression model, but for different data distributions. To generate this data, we select two points on $\mathcal{G}(1, 2)$ to define a geodesic curve and then uniformly sample 51 points along this geodesic from one point at time 0 to the other point at time 1. Using the strategy described in Section 3.5.2, Gaussian noise with $\sigma = 0.05$ is then added to the sampled points along the geodesic, resulting in the 1st group (Group 1) of synthetic data. The 2nd group (Group 2) of synthetic data is simulated by concatenating two different geodesics. Again, 51 points are uniformly sampled along the curve and Gaussian noise is added. The left column in Fig. 3.16(a) shows the two synthetic data sets using their corresponding angles.

Both groups are fitted using a standard geodesic regression model (Std-GGR); the qualitative and quantitative results of model criticism are reported in Fig. 3.16(a) and Table 3.4, respectively. Among 1000 trials, H_0 is rejected in only 8.1% of all cases for Group 1 (generated from one geodesic), but H_0 is always rejected for Group 2 (i.e., generated from two geodesics). As expected, Std-GGR is not an appropriate model to capture the distribution of the data belonging to multiple geodesics. Model criticism correctly identifies this, while the R^2 values are hard to interpret with respect to model suitability.



(a) Different data distributions

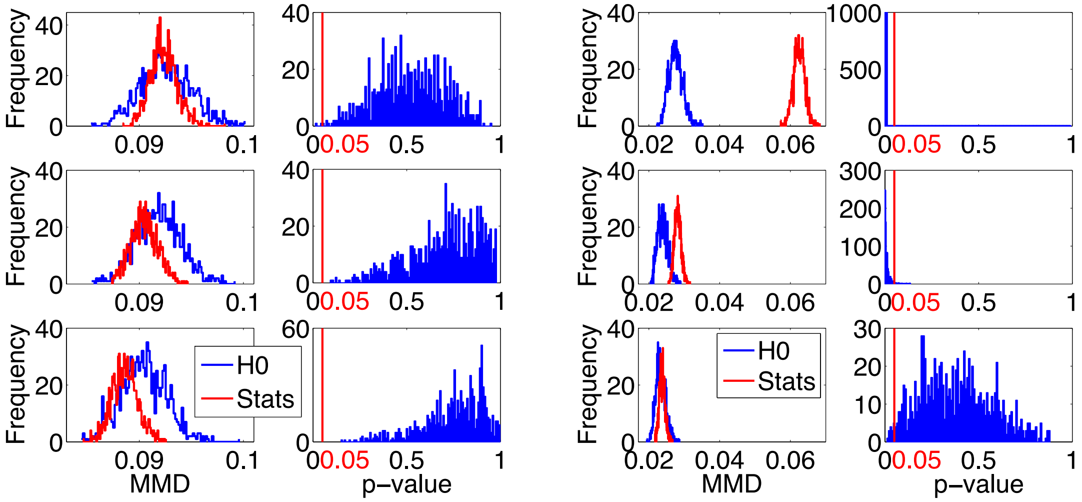
(b) Different regression models

Figure 3.16: Model criticism for synthetic data on the Grassmannian. (a) Different data distributions are fitted by one regression model (Std-GGR); (b) One data distribution is fitted by different regression models (*top*: Std-GGR, *bottom*: CS-GGR).

	Different data distributions		Different regression models	
	Group 1	Group 2	Std-GGR	CS-GGR
R^2	0.99	0.89	0.73	0.99
%(p-values < 0.05)	8.1%*	100%	88.0%	4.5%*

Table 3.4: Comparison of R^2 measure and model criticism for synthetic data. *In theory, this number should approximate 5% with enough trials, e.g., 10000.

(2) Different regression models. The second experiment is to generate one data distribution but to estimate different regression models. We first generate a section of a sine curve with x-axis as time and y-axis as the angle. Each angle θ corresponds to a point on the Grassmannian, i.e., $[\cos\theta; \sin\theta] \in \mathcal{G}(1, 2)$. In this way, we can generate polynomial data on $\mathcal{G}(1, 2)$ with associated time $t \in [0, 1]$. A visualization of the generated data with added Gaussian noise ($\sigma = 0.05$) is shown in the left column of Fig. 3.16(b). The data points are fitted by the standard regression model (Std-GGR) and its cubic spline variant (CS-GGR), respectively. The results of model criticism are shown in Fig. 3.16(b) and Table 3.4; in 88.0% of 1000 trials, H_0 is rejected for the Std-GGR model, while we only reject H_0 in 4.5% of all trials for CS-GGR. As designed, CS-GGR has better performance than Std-GGR and can appropriately capture the distribution of the generated data, as confirmed by model



(a) Corpus callosum

(b) Rat calvarium

Figure 3.17: Model criticism for real data. From *top* to *bottom*: the regression model corresponds to Std-GGR, TW-GGR, and CS-GGR respectively.

	Corpus callosum			**Rat calvarium**		
	Std-GGR	TW-GGR	CS-GGR	Std-GGR	TW-GGR	CS-GGR
R^2	0.12	0.15	0.21	0.61	0.79	0.81
%(p-values < 0.05)	0.2%	0%	0%	100%	98.0%	1.3%

Table 3.5: Comparison of R^2 measure and model criticism for real data.

criticism.

Real data. Two real applications, shape changes of the corpus callosum during aging and landmark changes of the rat calvarium with age, are used to evaluate model criticism for three regression models on the Grassmannian.

(1) Corpus callosum shapes. The population of corpus callosum shapes is the same one used in Section 3.3.4. In Fig. 3.17(a) and Table 3.5, although the R^2 values of the three regression models are relatively low, our model criticism results with 1000 trials suggest that all three models may be appropriate for the observed data.

(2) Rat calvarium landmarks. We use the same Vilmann rat dat in Section 3.3.4. From the model criticism results shown in Fig. 3.17(b) and Table 3.5 we can see that the equivalent linear (Std-GGR) and time-warped (TW-GGR) models cannot faithfully capture the distribution of the rat calvarium landmarks. However, the cubic-spline model is not rejected by model criticism and therefore appears to be the best among the three models. This result is consistent with the R^2 values and also provides more information about the regression models. As we can see, the R^2 values of TW-GGR and CS-GGR are very close, but the model criticism suggests CS-GGR is the one that should be chosen for regression on this dataset.

3.6 Conclusion

In this chapter, a general theory was developed for parametric regression on manifolds from an optimal-control perspective. By introducing the basic principles for fitting models of increasing order for the special case of $\mathcal{M} = \mathbb{R}^n$, we established the framework that was then used for a generalization to Riemannian manifolds and, in particular, the Grassmann manifold, as well as the manifold of diffeomorphisms.

From an application point of view, we have seen that quite different vision problems can be solved within the same framework under minimal data preprocessing. We compared our regression approaches to two alternative approaches in the literature. In comparison, we achieved similar or better performance, while providing a unified formulation and straightforward implementation. Our approaches also scale better to larger problems, thereby allowing for experiments on the traffic and the pedestrian data sets. While the presented applications are limited to shape analysis and surveillance video processing, our method should be widely applicable to other problems on the Grassmann manifold, *e.g.*, domain adaptation [Gong et al., 2012], facial pose regression, or the recently proposed domain evolution problems [Rematas et al., 2013].

Regarding the limitations of the proposed approach, we note that the issue of model selection is critical. In fact, whether we should use Std-GGR, TW-GGR or CS-GGR highly depends on our prior knowledge of the data. In shape regression, for instance such prior knowledge is frequently available, since the medical / biological literature already provides evidence for different growth and saturation effects as a function of age. For applications where the prediction of the independent variable is of importance, *e.g.*, traffic or or crowd surveillance, we additionally have computational constraints in many cases. Interestingly, a simple geodesic curve as a model for regression can often provide sufficiently good performance, as we observed in the crowd counting experiment. We hypothesize that this can be explained, to some extent, by the fact that geodesic regression respects the geometry of the underlying space. In this space, the relationship between the dependent and the independent variable might be relatively simple to model. In contrast, approaches where video content is represented by feature vectors and conventional regression approaches with standard kernels are used, more flexible models might be needed. TW-GGR can serve as a hybrid solution when we have prior knowledge about the data; however, samples throughout the range of the independent variable are needed to avoid degenerate cases of the warping function, which could be avoided via regularization.

Furthermore, the model criticism approach provides another alternative for model selection. This approach provides complementary information to the existing measure(s) for checking the goodness-of-fit for regression models, such as the customary R^2 statistic. While we developed the approach for the Grassmannian, the general principle can be extended to other smooth manifolds, as long as one knows how to add the modeled noise into samples and a positive definite kernel is available for the underlying manifold. Two important ingredients of model criticism are the estimation of the parameters of the noise model on the manifold (in our case Gaussian noise and therefore we estimate the standard deviation) and the selection of the kernel. Exploring, in detail, the influence of the estimator and the kernel and its parameters is left as future work.

CHAPTER 4 : ESTIMATION OF A SPATIOTEMPORAL STATISTICAL ATLAS

This chapter[1] presents a method to build a spatiotemporal atlas, which not only includes a summary representer but also retains variation information of the data distribution. This is achieved by a weighted variant of the functional boxplot, which enables the use of kernel regression to build spatiotemporal atlases. The method is applied to functions, shapes, and images. To demonstrate how an atlas can robustly be augmented with statistical data, the method has been tested in two applications: capturing changes in the pediatric airway development and changes of the corpus callosum over time. Furthermore, an age-adapted atlas can be used to score the severity of a child suffering from airway obstruction before and after surgery. This quantitative assessment shows significant differences among normal controls, pre- and post-surgery subjects.

In particular, Section 4.1 presents the strategy of using weighted functional boxplots to construct a spatiotemporal atlas. Section 4.2 shows the effectiveness of the method in comparison to pointwise analysis and its difference to the point distribution model. In Section 4.3, it demonstrates the applicability of the method to functions, shapes, and images, as well as the use of the presented method for quantitative assessment in pediatric airways.

[1] The work presented in this chapter is based on previous papers [Hong et al., 2013a, Hong et al., 2013b, Hong et al., 2014a].

4.1 Statistical Atlas Building

This section introduces a weighted variant of the functional boxplot and extends it for use with kernel regression of functional data. We first cover the preliminaries on kernel regression and later present the concept of weighted band-depth essential to defining the weighted functional boxplot. This method is applicable to the analysis of function, shape, and image populations to create non-parametric regression models with associated subject characteristics. For example, we consider subject age and demonstrate the effectiveness of weighted functional boxplots and kernel smoothing [Wand and Jones, 1994] to build a spatiotemporal atlas.

4.1.1 Atlas Building with Kernel Regression

Given spatially aligned data objects we want to capture population changes, e.g., with respect to age. Spatial alignment refers to a preprocessing step that transforms all data objects to common coordinates for further analysis. The type of alignment depends on the objectives of a particular study. For instance, this alignment may be a rigid transformation when the statistical analysis needs to be performed modulo translations and rotations only. An atlas with population changes can be built through kernel regression which assigns weights to data-objects with respect to the regressor (say a desired age $\bar{a}$). For example, we can use a Gaussian weighting function $w_i = f(a_i; \sigma, \bar{a}) = ce^{-(a_i - \bar{a})^2/2\sigma^2}$, where a_i is the age of the observation i, σ is the Gaussian standard deviation and c a normalization constant to assure that the weights sum up to one.

Boundary bias. Kernel-based methods exhibit a bias near the boundary of the available data. This is usually attributed to the asymmetric averaging of limited information at the boundaries. Many solutions have been proposed to address this issue [Schuster, 1985, Jones, 1993, Marron and Ruppert, 1994]. If the target age for the atlas is located within the interior part of the observed population, boundary effects can be ignored. However, for studies involving models for growth, aging or memory decline, we often build atlases for very young or very old subjects. This usually requires averaging kernel weights with respect to an age near the boundary. In such models, to mitigate the boundary bias, we adjust the weights around the boundaries based on the approach proposed in [Schuster, 1985], which relies on adjustment through boundary reflection.

We assume observations are given in the age range $[b_l, b_h]$. We adjust weights for observations at the boundaries in kernel regression by folding using reflection. In particular, given the kernel bandwidth, σ, and the location for each observation, a_i, with respect to the regression location, $\bar{a}$, the adjusted weights over the complete range are given as

$$w_i = f(a_i; \sigma, \bar{a}) = \begin{cases} c(g(a_i) + g(2b_l - a_i)), & a_i \in [b_l, b_l + \sigma) \\ cg(a_i), & a_i \in [b_l + \sigma, b_h - \sigma] \\ c(g(a_i) + g(2b_h - a_i)), & a_i \in (b_h - \sigma, b_h]. \end{cases} \tag{4.1}$$

Here $g(\cdot)$ denotes the Gaussian function, $g(\cdot; \sigma, \bar{a})$, with the mean, $\bar{a}$, and the variance, σ^2.

Bandwidth for kernel. An appropriate choice of the bandwidth, σ, for kernel regression depends upon the application. In general, the bandwidth should be chosen based on the expected variation in the data. A small bandwidth is able to express fast changes at the potential cost of becoming noise-sensitive, whereas a bandwidth that is too large gives overly smooth kernel regression results. A compromise can be achieved by selecting the bandwidth through cross-validation based model selection procedures [Hardle and Marron, 1985]. We will cover more details on how to choose σ in the experimental sections.

Note that for scalar-valued data, the weights for regression can be used to define a weighted mean. For kernel regression on deformations, these weights can be incorporated during the atlas-building procedure, e.g., in atlas construction using images [Davis et al., 2010]. Our goal is to augment the atlas with statistical information about observations and hence develop a weighted functional boxplot. The weighted functional boxplot will allow obtaining a regressed median. The median will be one of the data-objects from the population, which represents the center at the target age. We will further define the α central region, compute the interquartile-range and the maximum non-outlying envelope, and detect outliers.

4.1.2 Weighted Functional Boxplots

The challenge in defining a *functional* boxplot is to develop a notion of ordering for the space of functions. Once this ordering has been defined order statistics can be computed. Hence, the equivalent to a scalar-valued boxplot which makes use of the median and percentiles of the data can be defined. [Sun and Genton, 2011] proposed an ordering of functions based on the concept of band-depth. Essentially, band-depth measures how deeply a particular function is buried within all the other functions of the data population. The deepest one is then declared the median curve. The band-depth itself is used to define the ordering among functions. To define a *weighted* functional boxplot consistent with the functional boxplot introduced by [Sun and Genton, 2011], it requires the definition of a consistent *weighted band depth* for functional data.

Weighted band-depth. The functional boxplot is defined through the concept of band-depth [López-Pintado and Romo, 2009, Sun and Genton, 2011]. Since in our case, each observation has a different weight, we first need to define a weighted band-depth. Such a definition would naturally generalize the functional boxplot to its weighted variant.

To motivate our choice, assume we want to compute a standard weighted median of scalar values. This is given by

$$\mu^* = \underset{\mu}{\operatorname{argmin}} \sum_{i=1}^{n} w_i |x_i - \mu|, \tag{4.2}$$

where μ is the sought-for median, n is the number of measurements, $\{x_i\}$ are the measurements, and $w_i > 0$ are weights for the individual measurements. Assume that all weights are natural numbers, i.e., $w_i \in \mathbb{N}^+$. This can be realized exactly for arbitrary rational, w_i, and in general by multiplying the energy with a suitable constant, which does not change the minimizer. Hence, we replace the weighted problem with the equivalent unweighted minimization problem

$$\mu^* = \underset{\mu}{\operatorname{argmin}} \sum_{i=1}^{n} \sum_{j=1}^{m_i} |x_i - \mu|, \tag{4.3}$$

where the individual measurements are simply repeated based on their multiplicities, i.e., $m_i = w_i$.

Using a similar strategy, we can derive the weighted band-depth. The band-depth introduced in [López-Pintado and Romo, 2009, Sun and Genton, 2011] is defined for a population of n functions, y_i (for $i = 1 \ldots n$), defined on a domain $\mathbb{I}$, where $\mathbb{I}$ is an interval in $\mathbb{R}$. It is a graph-based approach that computes the fraction of bands delimited by the subset of the population containing the curve, $\mathrm{y}(\mathbf{x})$. In particular, it is defined as

$$BD_n^{(j)}(\mathrm{y}) = \frac{1}{C} \sum_{1 \leq i_1 < i_2 < \cdots < i_j \leq n} I\{G(\mathrm{y}) \subseteq B(\mathrm{y}_{i_1}, \cdots, \mathrm{y}_{i_j})\}. \tag{4.4}$$

Here j is the number of observations used for defining the band, C is a normalization constant equal to the number of admissible permutations. $G(\mathrm{y})$ is the graph of the function, $G(\mathrm{y}) = \{(\mathbf{x}, \mathrm{y}(\mathbf{x})) : \mathbf{x} \in \mathbb{I}\}$. B is the band delimited by the observations given as its arguments. That is, $B(\mathrm{y}_{i_1}, \cdots, \mathrm{y}_{i_j}) = \{(\mathbf{x}, \mathrm{y}(\mathbf{x})) : \mathbf{x} \in \mathbb{I}, min_{r=i_1,\cdots,i_j} \mathrm{y}_r(\mathbf{x}) \leq \mathrm{y}(\mathbf{x}) \leq max_{r=i_1,\cdots,i_j} \mathrm{y}_r(\mathbf{x})\}$. $I\{\cdot\}$ denotes the indicator function, which evaluates to 1 if the graph of the function is within the band or to 0 otherwise.

Now we want to define a weighted variant of the above definition of band-depth. For the weighted variant, say, we are now given a population of functions, y_i, for $i = 1 \ldots n$, each with its associated weight, w_i. Before we present the actual expression for weighted band-depth, we first write its repeated version. We notice that, similar to the scalar case, we could write the band-depth for this population of functions by repeating each observation as per its given weight. The band-depth with repeats is then given as

$$BD_{\overline{n}}^{(j)}(\mathrm{y}) = \frac{1}{C'} \sum_{1 \leq i_1 < i_2 < \cdots < i_j \leq \overline{n}} I\{G(\mathrm{y}) \subseteq B(\overline{\mathrm{y}}_{i_1}, \cdots, \overline{\mathrm{y}}_{i_j})\},$$

$$\text{s.t. } \{\overline{\mathrm{y}}_{i_1}, \cdots, \overline{\mathrm{y}}_{i_j}\} \text{ contains unique observations,} \tag{4.5}$$

where C' is the normalization constant representing admissible permutations adjusted for repeats and $\overline{n}$ is the number of observations including the repeats. The $\{\overline{\mathrm{y}}_i\}$ contain the original observations $\{\mathrm{y}_i\}$ but with repeats, according to their respective multiplicity given by their weights. The band with repeated observations is given as $B(\overline{\mathrm{y}}_{i_1}, \cdots, \overline{\mathrm{y}}_{i_j}) = \{(\mathbf{x}, \mathrm{y}(\mathbf{x})) : \mathbf{x} \in \mathbb{I}, min_{r=i_1,\cdots,i_j} \overline{\mathrm{y}}_r(\mathbf{x}) \leq \mathrm{y}(\mathbf{x}) \leq max_{r=i_1,\cdots,i_j} \overline{\mathrm{y}}_r(\mathbf{x})\}$. We made use of the fact that, according to our definition, we only consider unique observations for the depth measure.

Finally, we define the weighted band-depth by rewriting the sampled band-depth as

$$\begin{aligned} WBD_n^{(j)}(\mathrm{y}) = & \frac{1}{\sum_{1 \leq i_1 < i_2 < \cdots < i_j \leq n} w_{i_1} w_{i_2} \cdots w_{i_j}} \cdot \\ & \sum_{1 \leq i_1 < i_2 < \cdots < i_j \leq n} w_{i_1} w_{i_2} \cdots w_{i_j} I\{G(\mathrm{y}) \subseteq B(\mathrm{y}_{i_1}, \cdots, \mathrm{y}_{i_j})\}, \end{aligned} \tag{4.6}$$

which generalizes to non-natural-numbered weights $w_i \in \mathbb{R}^+$. This is a natural way to define a weighted band-depth and, in further consequence, a weighted functional boxplot. Computing the weighted band-depth in this way is intuitive, as only bands with large weights for all its individual observations have a large impact. Furthermore, this weighted version can also be adapted to the modified band-depth proposed in [Sun and Genton, 2011], i.e.,

$$\begin{aligned} WMBD_n^{(j)}(\mathrm{y}) = & \frac{1}{\sum_{1 \le i_1 < i_2 < \ldots < i_j \le n} w_{i_1} w_{i_2} \cdots w_{i_j}} \cdot \\ & \sum_{1 \le i_1 < i_2 < \ldots < i_j \le n} w_{i_1} w_{i_2} \cdots w_{i_j} \lambda_m \{ A(\mathrm{y}; \mathrm{y}_{i_1}, ..., \mathrm{y}_{i_j}) \} \end{aligned} \tag{4.7}$$

where $A_j(\mathrm{y}) \equiv A(\mathrm{y}; \mathrm{y}_{i_1}, ..., \mathrm{y}_{i_j}) \equiv \{\mathbf{x} \in \mathbb{I} : min_{r=i_1,...,i_j} \mathrm{y}_r(\mathbf{x}) \le \mathrm{y}(\mathbf{x}) \le max_{r=i_1,...,i_j} \mathrm{y}_r(\mathbf{x})\}$, m is the observation's dimension, $\lambda_m(\mathrm{y}) = \lambda(A_j(\mathrm{y}))/\lambda(\mathbb{I})$ and $\lambda \in \mathbb{R}^m$ is the Lebesgue measure.

With the above definitions, the band depths of all the sampled observations can be calculated and ranked in descending order, $\mathrm{y}_{[1]}(\mathbf{x}) \ge ... \ge \mathrm{y}_{[n]}(\mathbf{x})$. Here, $\mathrm{y}_{[1]}(\mathbf{x})$ is the deepest observation and regarded as a notion of the median of the population, whereas $\mathrm{y}_{[n]}(\mathbf{x})$ is the "most outlying" observation which is a potential outlier.

α **central region.** The concept of the central region was introduced in [Liu et al., 1999]. We define the α central region for the weighted functional boxplot based on the weights of the observations. The band of the α central region is delimited by the α proportion of all weights, i.e., the accumulated weights of the first $\hat{p}$ deepest observations.

We first compute the value of $\hat{p}$ based on the weights by

$$\hat{p} = \{p \in \mathbb{N}^+ : \sum_{r=1,...,p-1} w_{[r]} < \alpha, \sum_{r=1,...,p} w_{[r]} \ge \alpha, and\ p \le n\}, \tag{4.8}$$

where $w_{[r]}$ corresponds to the weight for the r-th deepest observation and $0 \le \alpha \le 1$. Here, we assume that the weights are normalized so they sum up to one. Then the α central region can be generated using these first $\hat{p}$ observations through

$$WCR_{\alpha} = \{(\mathbf{x}, \mathrm{y}(\mathbf{x})) : \underset{r=1,\ldots,\hat{p}}{min} \mathrm{y}_{[r]}(\mathbf{x}) \leq \mathrm{y}(\mathbf{x}) \leq \underset{r=1,\ldots,\hat{p}}{max} \mathrm{y}_{[r]}(\mathbf{x})\}. \tag{4.9}$$

When $\alpha = 0.5$, Eq. (4.9) corresponds to the 50% central region $WCR_{0.5}$. In practice, the 50% central region is commonly chosen as the confidence region for analysis because it 1) is a robust range for interpretation and 2) enables visualization of the data spread which is less affected by outliers or extreme-values.

Outlier detection. In classical boxplots, the outliers can be detected by the 1.5 IQR (interquartile range) [Frigge et al., 1989]. This is comparable to 1.5 times the corresponding range of the 50% central region for the weighted functional boxplot. The weights of the observations also need to be taken into consideration during the outlier detection. For a Gaussian distribution, the IQR encompasses the most central 50% of the distribution and the fence defined by 1.5 IQR covers 99.3% of the distribution. Therefore, we use this threshold, 0.993, to find the first $\hat{q}$ deepest observations that would be within the fences by

$$\hat{q} = \{q \in \mathbb{N}^{+} : \sum_{r=1,\ldots,q-1} w_{[r]} < \beta, \sum_{r=1,\ldots,q} w_{[r]} \geq \beta, and\ q \leq n\}, \tag{4.10}$$

where $\beta = 0.993$. The next step is to narrow the fences with the 1.5 IQR, so that the fences defined in weighted functional boxplots are the combination of the fence defined by the 1.5 IQR with the accumulated weights consistent with the 1.5 IQR of the normal distribution:

$$\begin{aligned} C_{fences} = \{(\mathbf{x}, \mathrm{y}(\mathbf{x})) : max(min_{r=1,\ldots,\hat{q}}\mathrm{y}_{[r]}(\mathbf{x}), min(WCR_{0.5}) - 1.5 * IQR) \cup \\ min(max_{r=1,\ldots,\hat{q}}\mathrm{y}_{[r]}(\mathbf{x}), max(WCR_{0.5}) + 1.5 * IQR)\}. \end{aligned} \tag{4.11}$$

Any objects outside the fences defined by C_{fences} are flagged as outliers.

4.1.3 Implementation and Algorithm Complexity

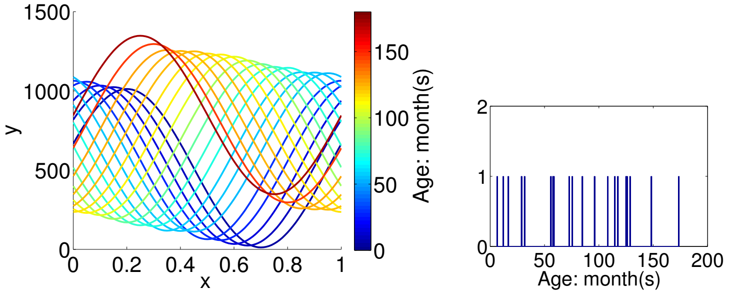
In this section, we discuss how to implement our statistical atlas-building method based on the weighted functional boxplots (see Algorithm 4), as well as the time complexity of the algorithm. From the observations and their associated independent values, e.g., ages, the algorithm generates a statistical atlas at a target age, which consists of the median, the confidence region, the maximal non-outlying region, and the outliers. Details about converting between the functional representation of the boxplot and shapes and images are covered in Section 4.3.1.

Algorithm 4 Statistical atlas-building based on weighted functional boxplots

```
**Input**: $\{a_i, \mathrm{y}_i\}_{i=1}^N$ ($N$ observations with ages), $\bar{a}$ (the target age), $J$ (the number of observations for a band)
**Output**: $\mathrm{y}_{[1]}$ (the median), $WCR_\alpha$ (the $\alpha$ center region), $C_{fences}$ (the fences of the atlas)
Choose the bandwidth $\sigma$ and compute the weight $w_i$ for each $\mathrm{y}_i$ with Gaussian function centered at $\bar{a}$ (Section 4.1.1).
**for** i := 1:N **do**
  **for** j := 2:J **do**
    Loop through all combinations, choosing j from N observations, and compute the weighted band depth for $\mathrm{y}_i$ using Eq. (4.6) or (4.7).
  **end for**
**end for**
Sort $\{\mathrm{y}_i\}_{i=1}^N$ based on the weighted band-depth in the decreasing order, $\mathrm{y}_{[1]} \geq \cdots \geq \mathrm{y}_{[N]}$.
Compute $WCR_\alpha$ and $C_{fences}$ according to Section 4.1.2.
```

Since in practice $J \ll n$, the complexity of this algorithm is $O(MN^{J+1})$ where M is the dimension of each observation and N is the number of the observations. We usually choose $J = 2$ resulting in a time complexity of $O(MN^3)$.

For our experiments we use the weighted version of the modified band-depth, Eq. (4.7), because it results in fewer depth ties compared to the unmodified band-depth. Note that as we are dealing with a generalization of the median, continuity with respect to the regression



(a) Observations (b) Age histogram

Figure 4.1: (a) 20 observations generated based on Eq. (4.12) and colored by age. (b) The age histogram of the observations.

variable (here, age) *cannot be guaranteed.* Assuming that the underlying data is continuous, a "more continuous" behavior may be achieved using more and sufficiently dense sampled data. In particular, we would expect that additional samples in sparsely sampled regions of a dataset would result in solutions with less severe discontinuities.

4.2 Comparisons of Boxplots for Analysis

Synthetic data. We compare the atlases built by weighted functional boxplots and those built by 1) weighted pointwise boxplots and 2) functional boxplots, using synthetic observations defined by

$$y_i(x) = 500 * (1 + sin(2\pi x + 0.1\pi i)) + 2 * age_i, \tag{4.12}$$

where $x \in [0, 1]$, i ranges from 1 to 20 (i.e., we generate 20 observations for analysis, shown in Fig. 4.1(a)), and age_i is the simulated age corresponding to the ith spike in Fig. 4.1(b). The age varies from 0 to 200 months, that is, $b_l = 0$ and $b_h = 200$.

4.2.1 Comparison with Weighted Pointwise Boxplots

The left image of Fig. 4.2 shows an atlas built with the weighted pointwise boxplot including four typical percentiles and the pointwise median. The weights are computed

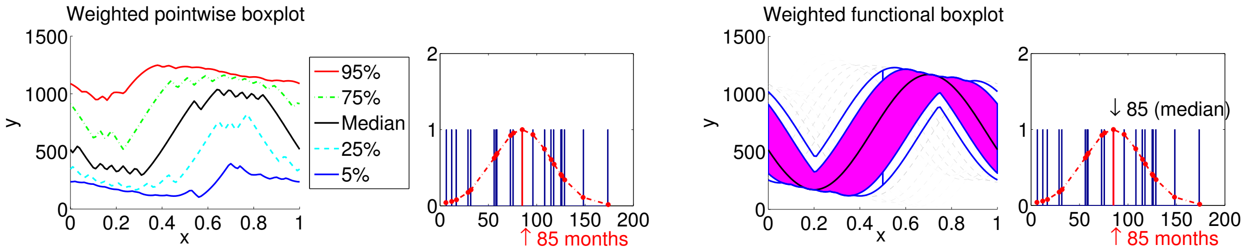


Figure 4.2: Comparisons of the atlases built by the weighted pointwise boxplot (left) and the weighted functional boxplot (right) on the synthetic data. The atlases are adapted to the age of 85 months. The median computed by the weighted pointwise boxplot is a pointwise median, and the median computed by the weighted functional boxplots corresponds to an existing observation at 85 months.

based on the Gaussian function in Section 4.1.1 with $\sigma = 30$ months. While the median curve follows the overall population trend, it is not “close” to any of the observations because weighted boxplots, applied in a pointwise manner to a population of functions, disregard the spatial aspect of the functional data. In contrast, our method shown in the right image of Fig. 4.2 1) provides a median curve which corresponds to a curve in the data set, i.e., the one at the age of 85 months, and 2) allows for the detection of *functional* outliers (gray dashed lines) which results in a more robust statistical description for the atlas.

4.2.2 Comparison with Functional Boxplots

To construct an atlas at a particular age using standard functional boxplots, we use a uniform window to pick curves centered around the age of interest. As shown in Fig. 4.3, only two curves are available in the uniform window for atlas-building with functional boxplots, and one of them is flagged as an outlier. This atlas includes little information about the population. However, the atlas built using the weighted functional boxplot (with a Gaussian window size comparable to that of the uniform window used in the standard functional boxplots according to [Marron and Nolan, 1988]) captures the population data much better as it suffers less from the local data sparsity.

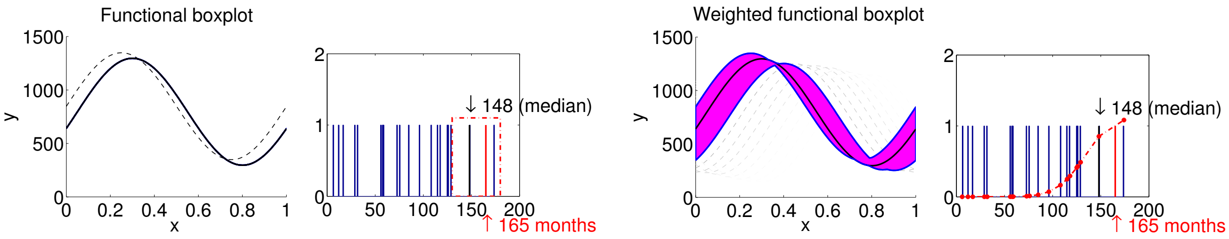


Figure 4.3: Comparisons of atlases built by the functional boxplot (left) and the weighted functional boxplot (right) on the synthetic data. The atlases are built at age 165 months and for both methods the observation at 148 months is selected as the median curve.

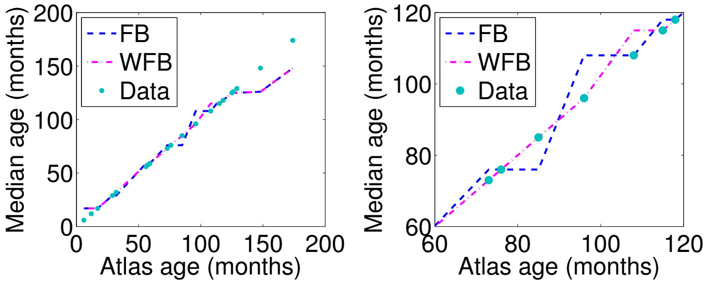


Figure 4.4: Comparison of the atlas age and the median age between the functional boxplot (FB, the blue dashed line) and the weighted functional boxplot (WFB, the magenta dashed line). The cyan dots show the ideal case, that is, a method has a better performance if it passes through more cyan dots. The right image is a close-up view of the left one.

For further illustration, we build a set of atlases from the synthetic curves at the age associated to each curve using functional boxplots and weighted functional boxplots. Each age-matched atlas has a median curve, and ideally the age of the atlas matches with the age of the median when the age of the population is evenly distributed, indicated by the cyan dots in Fig. 4.4. For this synthetic dataset, we want to determine which one provides a better approximation of the median age to the atlas age, the functional boxplot or the weighted functional boxplot. As shown in Fig. 4.4, the magenta line estimated by the weighted functional boxplot (WFB) is much smoother than the blue line estimated by the functional boxplot (FB), and the magenta line is closest to most of the cyan dots, indicating that the weighted functional boxplot has a better performance than the functional boxplot for spatiotemporal atlas construction.

Table 4.1 provides quantitative measurements on the median ages computed by the func-

	FB	**WFB**
Mean of relative errors	15.25%	**13.99%**
Equal to atlas ages	35%	**50%**
Closer to atlas ages*	5%	**20%**

Table 4.1: Comparison of the median ages estimated by functional boxplots (FB) and weighted functional boxplots (WFB) on synthetic data. *This measure counts the frequency with which the estimated median ages are closer to the true age for FB and WFB respectively. In 75% of the cases the median ages from these two methods are identical.

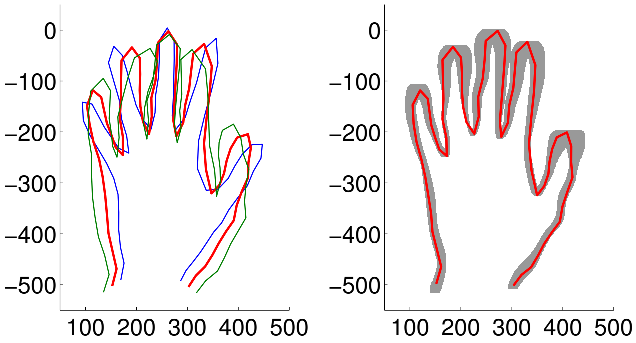


Figure 4.5: Comparison between the point distribution model (left) and the functional boxplot (right) applied to 18 2D hand contours. Left: mean shape in red and shape variation along the first mode for −3 standard deviations (blue) and for +3 standard deviations (green). Right: median shape in red and 50% confidence region in gray.

tional boxplot and weighted functional boxplots with respect to the atlas ages. In particular, we evaluate the relative age error, how frequently the methods return a median curve of exactly the correct age and with what frequency the estimated median curve's age is closer to the real age for the functional versus the weighted functional boxplot. The weighted functional boxplot outperforms the standard functional boxplot for all these measures.

4.2.3 Comparison with the Point Distribution Model

The point distribution model (PDM) [Cootes et al., 2004] is a powerful method to statistically describe shape variations. Shape variation is captured by computing a mean shape and the principal shape variations around this mean through principal component analysis. It is important to note that the objective of a PDM is different from that of the functional

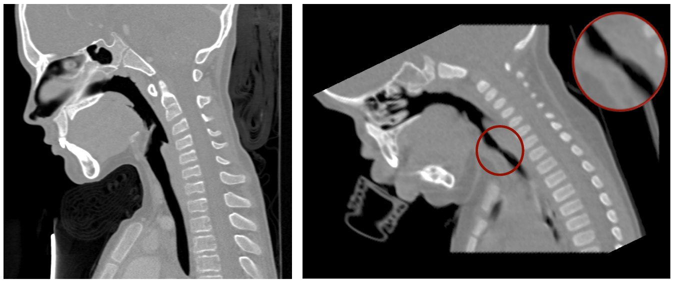


Figure 4.6: CT scans for a control subject (left, CRL04) and a subglottic stenosis patient (right, SGS03). The zoomed-in part in the red circle shows the location of subglottic stenosis, the narrowing of the airway.

boxplot. Whereas PDM is used to capture the major modes of shape variation through a multi-variate Gaussian distribution, the functional boxplot is free of distributional assumptions as it is a form of order statistics. The functional boxplot is therefore robust to outliers and readily allows for the computation of α-central regions, such as the interquartile range, to quantify data variation.

To demonstrate the difference in behavior between the PDM and the functional boxplot, Fig. 4.5 shows the shape variation of 18 2D hand outlines from [Cootes et al., 1995] as captured through a PDM and the functional boxplot. The PDM readily allows for the visualization of the principal modes of shape variation, whereas the functional boxplot provides an intuitive way of looking at the spatial differences observable within for example the 50% confidence region (see Section 4.3.1 for details on how to compute the confidence region). Hence, both methods provide useful but complementary information.

4.3 Experimental Results

In this section we show example applications using the weighted functional boxplot. The examples involve functions, shapes, and images.

Functions. Our first application is the construction of a pediatric airway atlas from

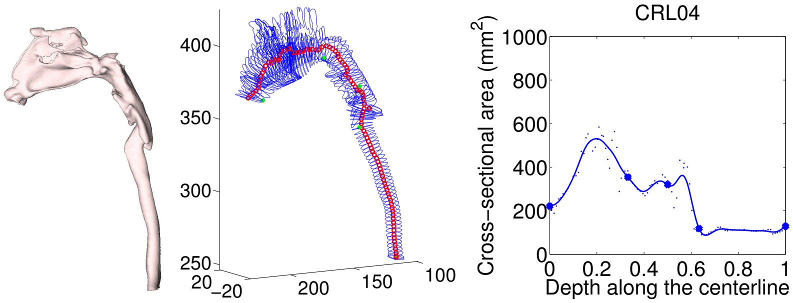


Figure 4.7: The simplified airway model for converting a 3D airway geometry to a 1D curve. Left: the geometry segmented from a CT image, CRL04; middle: the centerline of the airway with cross sections along the centerline; right: the curve of the cross-sectional area with the depth along the centerline.

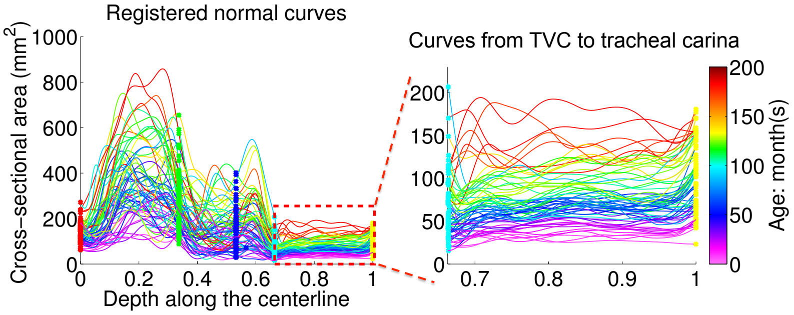


Figure 4.8: Normal curves for pediatric airway atlas construction, which are registered based on the following five landmarks: nasal spine, choana, epiglottis tip, true vocal cord (TVC) and tracheal carina (from left to right). Zoomed-in: the sub-region from TVC to tracheal carina where the subglottis is located.

normal subjects to assess airway obstruction (i.e. subglottic stenosis, SGS) [Daniel, 2006], as shown in the computed tomography (CT) images in Fig. 4.6. The observations are a population of 1D functions describing airway cross-sectional areas parameterized along the centerline of the airway as shown in Fig. 4.8. These functions are generated from 3D CT data for 68 normal subjects using a simplified airway model [Hong et al., 2013b], shown in Fig. 4.7. The functions are registered using a landmark-based spatial alignment [Ramsay and Silverman, 2005]. The alignment is based on five key anatomic landmarks: nasal spine, choana, epiglottis tip, true vocal cord and tracheal carina. For each landmark, there is a physical position on the centerline and a mean position of that landmark for all subjects. We estimate a warping function parameterized as a spline smoothly

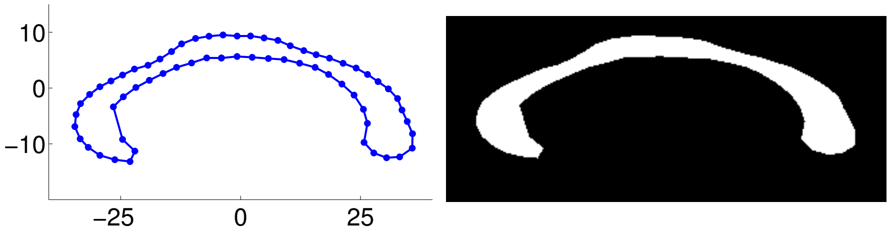


Figure 4.9: Examples of the corpus callosum shape (left) and the binary image of the corresponding segmentation (right).

passing through pairs of physical and mean position for all landmarks for the registration of the functions. We focus the analysis on the region between the true vocal cord (TVC) and the tracheal carina where the subglottis is located, as shown in the right image of Fig. 4.8. The 68 normal functions are used to build a normal control pediatric airway atlas to assess 19 SGS subjects pre- or/and post-surgery.

Shapes. The second application is to build a corpus callosum atlas and to explore its shape changes with age. We use the same dataset in Section 3.3.4.

Images. The third application is to understand age-related changes of the corpus callosum using binary images. The images are converted from the aligned corpus callosum shapes, and one example is shown in the right image of Fig. 4.9.

4.3.1 Functional Representation of Shapes and Images

In our experiments, we treat shapes and images as functions by vectorizing the data. After analysis, we convert the functional form back to the original representation of the data objects. It is instructive to look at the effect of this vectorizing step in the context of binary images which we use as an image-based method to represent shapes (contours). In this case images represent shapes through indicator functions. Note that we discuss curves in 2D in this work, but the principle extends to the representation of any closed data objects, e.g., closed surfaces in 3D. Assume the shapes are represented by images through indicator functions $\mathbb{1}_{I_i}$, where I_i is the set which indicates the interior of the shape S_i, i.e., $I_i = \{x : \ x \text{ inside } S_i\}$ and $\mathbb{1}_{I_i}(x) := 1$ if $x \in I_i$ and 0 otherwise. We can then write intersections and unions of sets through the indicator functions as

$$\mathbb{1}_{S_i \cap S_j} = min\{\mathbb{1}_{S_i}, \mathbb{1}_{S_j}\} \quad \text{and} \quad \mathbb{1}_{S_i \cup S_j} = max\{\mathbb{1}_{S_i}, \mathbb{1}_{S_j}\}. \tag{4.13}$$

The band-depth defined in Section 4.1.2 is based on evaluating $I\{G(\mathrm{y}) \subseteq B(\mathrm{y}_1, \cdots, \mathrm{y}_i)\}$. Applied to indicator functions this expression is equivalent to

$$I\left\{\bigcap_i S_{\mathrm{y}_i} \subseteq S_{\mathrm{y}} \subseteq \bigcup_i S_{\mathrm{y}_i}\right\} \tag{4.14}$$

as the band B is constructed by taking the minima and maxima over all functions. For the indicator functions the minimum and maximum operators correspond to the set intersections and unions respectively (due to the associativity of set union and intersection). Applying the functional boxplot to vectorized indicator functions of image-representing shapes is therefore equivalent to the definition of contour boxplots proposed independently by [Whitaker et al., 2013], unifying the two methods. [Whitaker et al., 2013] introduce contour boxplots to quantify uncertainty in feature sets from simulation ensembles for example obtained from fluid simulations. Our weighted functional boxplot can be interpreted as an extension of the method [Whitaker et al., 2013] to weighted contour boxplots. That is, the vectorization approach is quite natural in the context of indicator-function-based shape representation. The relation for the contour-based representations is theoretically less clear. However, our experiments indicate that in practice this method achieves similar results to the indicator-function-based shape representation. Specifically, bands for shapes and images are computed as follows:

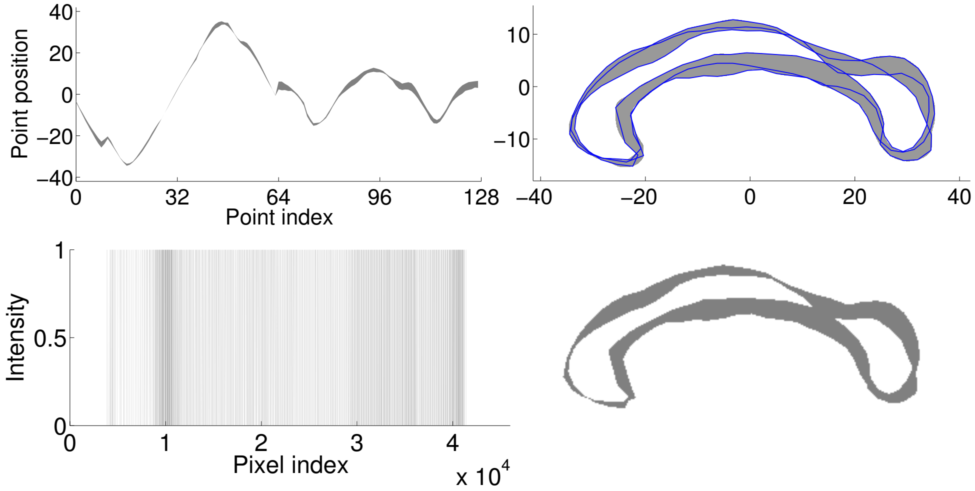


Figure 4.10: The functional bands (left), delimited by three corpus callosum shapes (the blue contours), and their corresponding shape band (top right) and image band (bottom right).

Shape band. To compute the band for aligned shapes, taking the three blue curves in the top right of Fig. 4.10 as an example, we first vectorize them to compute the functional band, shown in the top left of Fig. 4.10. Then, for a 2D point on the shape, (p, q), its variation is within the rectangular region with the diagonal given by the two points on the band's boundary, $(min(p), min(q))$ and $(max(p), max(q))$. For a 3D point, its variation is within a rectangular solid. The union of these rectangular regions then forms the shape band. With a sufficiently dense sampling of the functions, we obtain a smooth shape band as illustrated in the top right of Fig. 4.10. This shape band contains all three curves.

Image band. Compared with the shape band, the image band is much easier to construct. As discussed above for binary images, the standard functional boxplot theory can be directly applied. Converting the obtained bands back to the image domain immediately results in the desired image band. The image band can be constructed for non-binary images in the same way. Fig. 4.10 (bottom) shows the functional band for three binary images of corpora callosa on the left and the corresponding image band on the right.

Fig. 4.10 also shows that both shape and image bands are similar and correctly capture the range of the observations.

4.3.2 Comparison with Pointwise Boxplots

We compare the functional boxplot to the pointwise approach on real datasets to demonstrate the advantages of our method. Fig. 4.11 shows the median (the black curve) and the confidence region (the 50% central region, magenta) for both pointwise and functional boxplots. We count the number of data objects inside the confidence region shown in Table 4.2: for the pointwise boxplots only 12 (of 68) functions and none of the shapes or images are fully within the confidence region. However, the functional boxplot, by construction, provides a confidence region containing 50% of the data objects. We consider this a more intuitive representation of true data-object variation. To construct the pointwise confidence regions for shapes we locally compute distances with respect to the median point which establishes an (unsigned) ordering. The confidence region is then the convex hull of the closest half of the points. This strategy would extend to constructing approximate confidence regions with respect to manifold embedding coordinates. Specifically, in the coordinate system after manifold embedding, each observation is represented as a point and the median is defined as the point with the minimal sum of the squared geodesic distances to other points on the manifold. The confidence region can then be defined as the convex hull on the manifold formed by half of the points with the closest geodesic distances to the median point. This is conceptually similar to the way we construct shape bands.

4.3.3 Atlas Construction with Weighted Functional Boxplots

The weighted functional boxplot is used to build a pediatric airway atlas with a standard deviation of $\sigma = 24$ months for the weighting function, and to build the corpus callosum

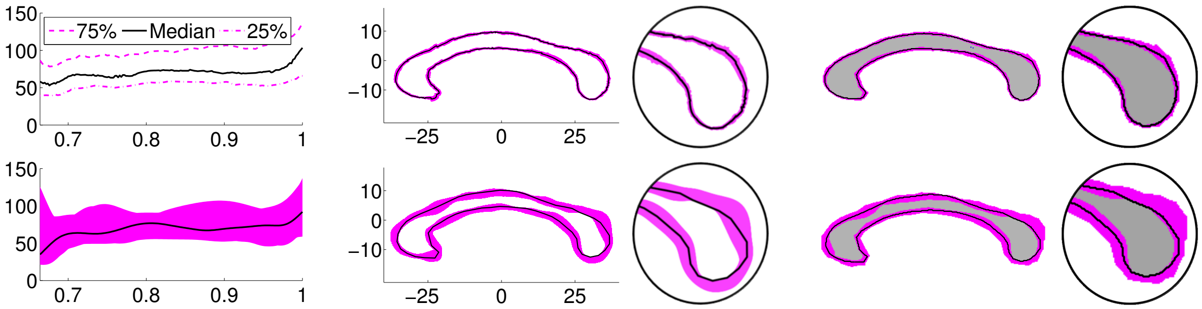


Figure 4.11: Comparison between pointwise (top) and functional (bottom) boxplots on functions, shapes and images. The black curve is the median and for the pointwise boxplots it is the pointwise median. The magenta region is the 50% confidence region.

	Functions	Shapes	Images
Pointwise boxplots	12/68	0/32	0/32
Functional boxplots	**34/68**	**16/32**	**16/32**

Table 4.2: The number of data objects inside the 50% central region for functions, shapes and images in Fig. 4.11. The first number is the sum of the data objects inside the central region, and the second number is the total number of observations.

shape/image atlases with $\sigma = 10$ years. This allows capturing large changes for both cases, while keeping the changes to be smooth. For the pediatric airway application, the age range varies from 0 to 200 months, that is, $b_l = 0$ and $b_h = 200$. For the corpus callosum application, we set the age within $(0, 100 \text{ years})$, that is, $b_l = 0$ and $b_h = 100$.

Pediatric airway atlas. Fig. 4.12 shows two pediatric airway atlases at different ages, 20 months and 180 months respectively. The pediatric airway atlases capture increases in cross-sectional airway area with age which is consistent with the growth pattern for pediatric airways and indicates the necessity of building an age-adapted atlas as a reference. Furthermore, in Table 4.3 we measure the difference of median ages estimated by functional boxplots and weighted functional boxplots. Similar to Section 4.2.2, we build an age-matched atlas for each control subject and use the age of the selected median subject for comparison. The weighted functional boxplot leads to a smaller mean relative error and more median ages

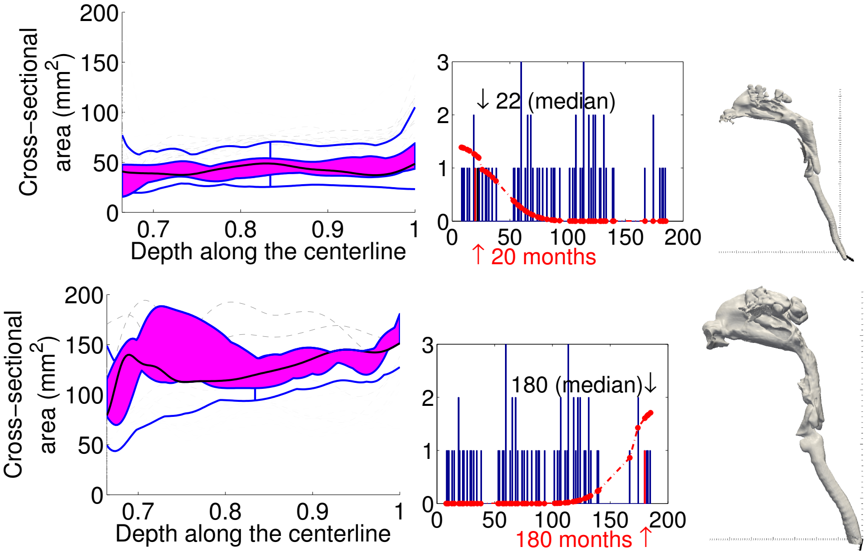


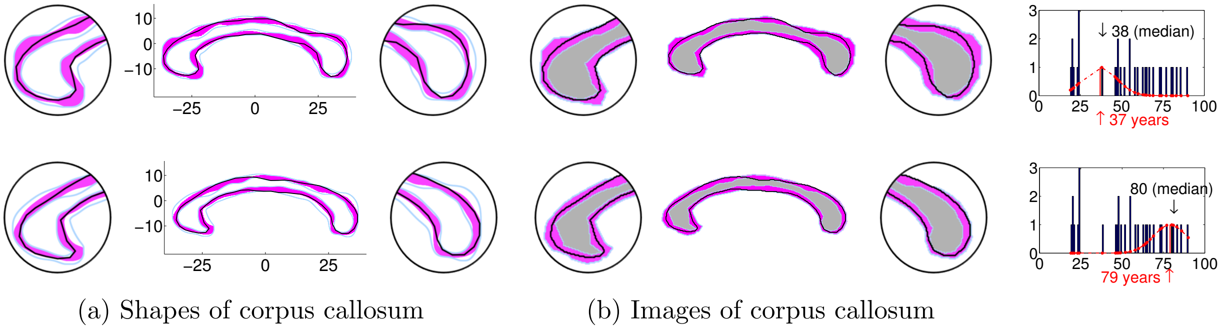
Figure 4.12: Age-adapted atlases for functions: pediatric airway atlases at 20 and 180 months respectively. The two airway geometries correspond to the median subjects selected by the age-matched atlases. The older atlas has a larger airway size compared to the younger atlas, indicating the importance of building age-matched atlases.

	FB	**WFB**
Mean of relative errors	19.75%	**15.05%**
Equal to atlas ages	**11.76%**	7.35%
Closer to atlas ages*	20.59%	**26.47%**

Table 4.3: Comparison of the median ages estimated from the functional boxplot (FB) and the weighted functional boxplot (WFB) on the pediatric airway dataset. *This counts the number of the median ages that are closer to the atlas ages between functional boxplots and weighted functional boxplots; 52.94% of the median ages from these two methods are equal.

are closer to the atlas ages. However, fewer median ages agree exactly with the atlas ages. This is acceptable because the cross-sectional area of pediatric airways increases with age in general while small variations may be caused, e.g., by measurement errors and difference in true developmental age. Overall, the weighted functional boxplot performs well at building the age-adapted pediatric airway atlas.

Corpus callosum atlas. In Fig. 4.13, we select atlases at age 37 and 79 years for both the shape and the segmentation of corpus callosum to demonstrate atlas changes with respect to age. The two corpus callosum atlases reveal the thinning in the shape and the decreasing



(a) Shapes of corpus callosum

(b) Images of corpus callosum

Figure 4.13: Age-adapted atlases for shapes and images: corpus callosum atlases at 37 (top) and 79 (bottom) years respectively. Zoomed-in: the anterior (the splenium, on the right of the atlas) and posterior (the genu, on the left of the atlas) portions of corpus callosum atlases. The atlases at different ages, especially the zoomed-in parts, clearly show the thinning of the corpus callosum with age.

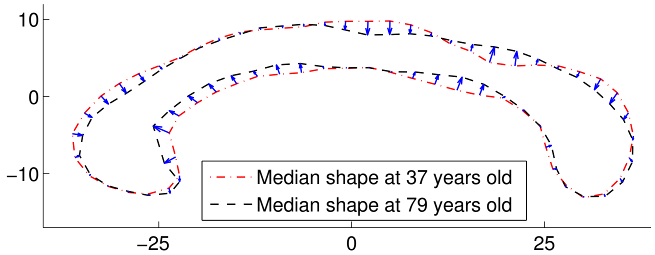


Figure 4.14: The median shapes of two corpus callosum atlases at different ages and the direction of change of the corresponding points on the boundaries.

volume in the image with age, especially at the anterior (the splenium) and posterior (the genu) portions consistent with [Driesen and Raz, 1995, Fletcher, 2011]. To further visualize these changes, we overlap the median shapes of the corpus callosum atlases in Fig. 4.13, and we display the directions of change for all corresponding points on the boundary in Fig. 4.14. Most parts of the median shape, especially the anterior and posterior regions, show the thinning of the corpus callosum with age.

4.3.4 Computational Cost for Building a Statistical Atlas

The algorithm is implemented in Matlab. All the experiments were run on an Intel® Xeon(R) CPU E5645 system with 2.4GHz. Table 4.4 shows the computation times for building atlases from populations of observations with different numbers of datasets and

	Observations		Computational
	Size	Number	Cost (s)
Synthetic functions	101 1D points	20	0.024
Pediatric airway	169 1D points	68	0.50
Corpus callosum (shape)	64 2D points	32	0.34
Corpus callosum (image)	157 × 456 pixels	32	29.56

Table 4.4: Computational cost of building an atlas based on the weighted functional boxplot.

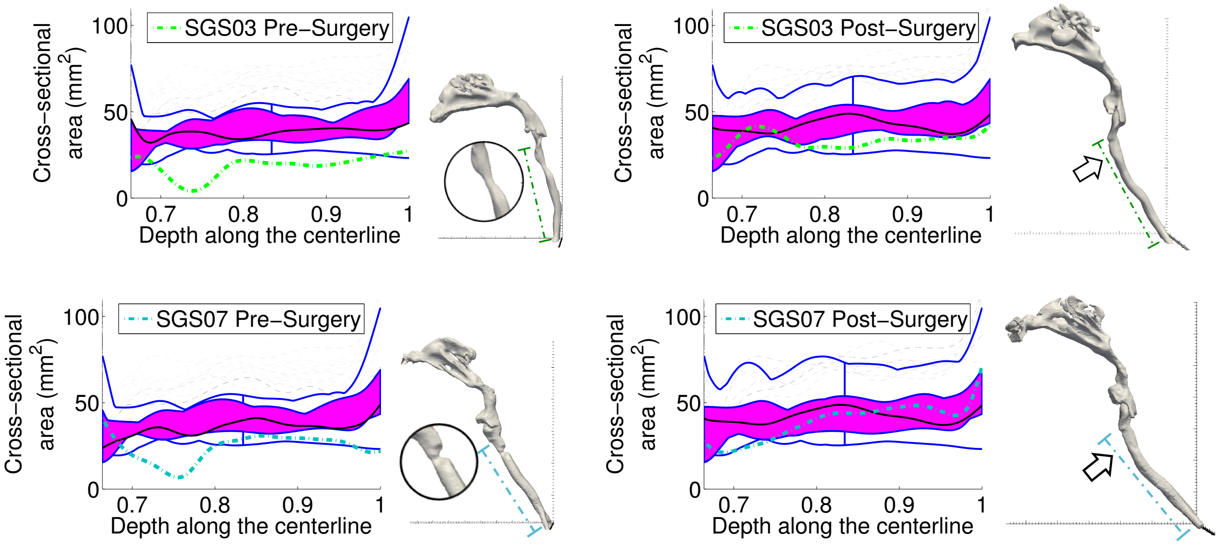


Figure 4.15: Airway changes for two subjects, SGS03 and SGS07, pre- and post-surgery (cyan dashed lines) compared to their age-matched atlases respectively. For each subject, the stenosis of the airway is marked by the zoomed-in circle on the pre-surgery geometry and no visible stenosis exists in the post-surgery geometry (the arrow in the right image indicates the subglottic area).

different numbers of 1D/2D points or pixels. Most experiments required less than one second of runtime. The image case, while still reasonably fast, is as expected slowest because the dataset size is largest. This computational time can be greatly reduced if we use the fast algorithm presented in Chapter 5 to compute the band-depth for binary images.

4.3.5 Assessment with Statistical Atlas

Comparison Pre- and Post-Surgery. To test the utility of the statistical atlas built by weighted functional boxplots we show the airway changes of two SGS subjects before and

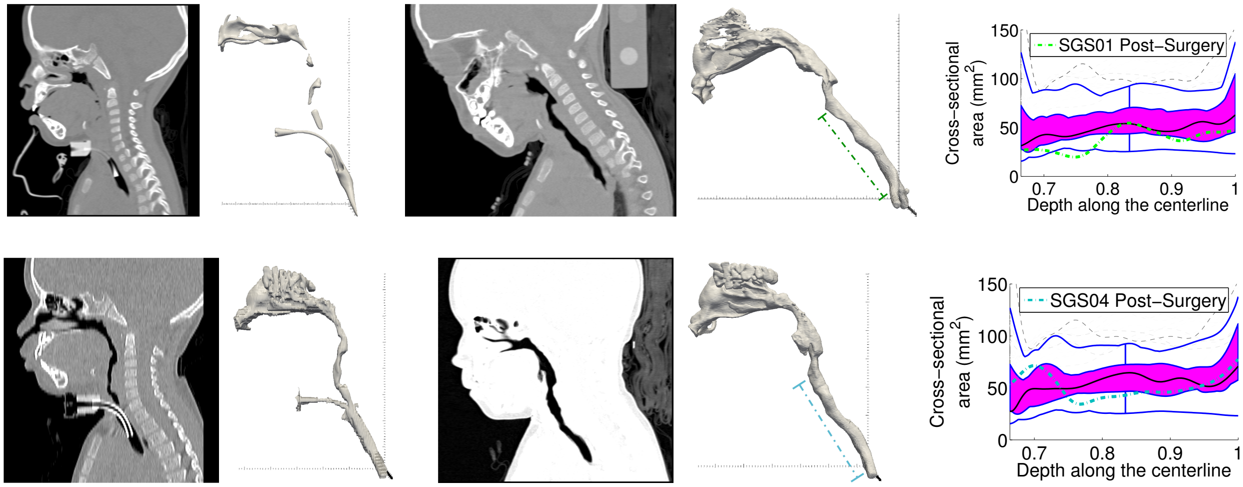

150
SGS01 Post–Surgery
100
50
0
Cross–sectional area (mm^2)
0.7
0.8
0.9
1
Depth along the centerline
150
SGS04 Post–Surgery
100
50
0
Cross–sectional area (mm^2)
0.7
0.8
0.9
1
Depth along the centerline

Figure 4.16: Airway changes for SGS01 and SGS04 pre- and post-surgery. For both subjects, before surgery there is a tracheostomy tube in the airway. After surgery the subglottic stenosis is resolved. Compared with the age-matched atlas almost all of the corresponding curve is within the maximal non-outlying envelope, indicating a successful surgery.

after surgery compared to the age-matched normal control airway atlases. The subject shown on the top of Fig. 4.15 is SGS03, a male who had two CT scans, one before surgery at 9 months and the other after surgery at 20 months. Fig. 4.15 shows another male (SGS07) who had a CT scan before surgery at 6 months and another one after surgery at 15 months. Before treatment, there was a constricted region outside the atlas for both children, corresponding to the dip in the cross-sectional area curve and the zoomed-in circle of the geometry in both cases. After treatment, the airway size increased and the corresponding curve is almost entirely within the maximal non-outlying envelope of the atlas. Also there is no visible stenosis in the geometry, indicating that the surgeries for these children were successful.

Fig. 4.16 show two additional children with subglottic stenosis. We can see the tracheostomy tubes in the CT scans and the airway geometries. We do not compute the cross-sectional areas for such cases, because their airways appear disconnected before surgery and breathing is accomplished through the tracheostomy tubes. Minimal cross-sectional areas are set to zero. After surgery, the airway cross-sectional areas greatly increase. For SGS01

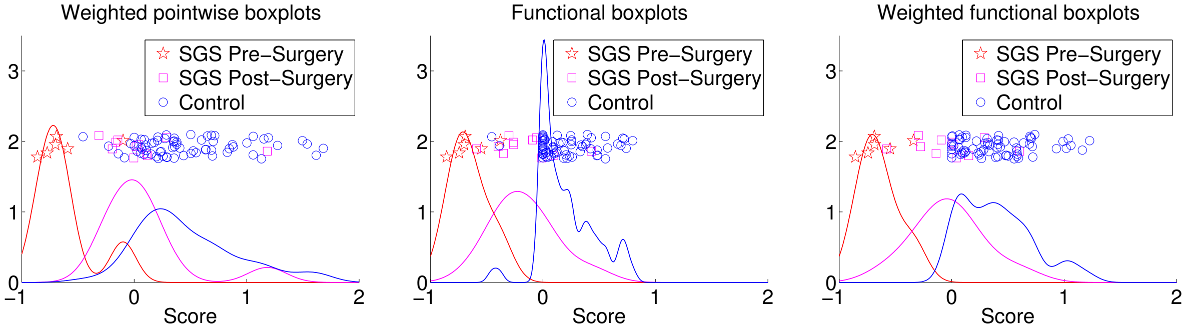


Figure 4.17: The scores for all subjects, including three groups, SGS pre-surgery, SGS post-surgery and control subjects, based on the atlases built by weighted pointwise boxplots, functional boxplots, and weighted functional boxplots (from left to right). The curves in different colors represent the kernel density estimations for different groups. Note: the y-axis in the plots is a random height to visualize the scores clearly.

only a small part of the corresponding curve is slightly outside the maximal non-outlying envelope of the atlas, and for SGS04 its whole corresponding curve is totally inside the maximal non-outlying envelope, indicating successful surgeries also for these two cases.

Quantitative Measurements

Definition of the scoring system. The Myer-Cotton grading system [Myer et al., 1994] is commonly used in clinical diagnosis for estimating the severity of subglottic stenosis in the pediatric upper airway. It describes the stenosis by the relative percentage reduction of the cross-sectional area at the subglottis. In practice, this is determined by using different sizes of endotracheal tubes. Similar to the Myer-Cotton system, we define a scoring system based on the age-matched atlas to quantitatively measure the severity of subglottic stenosis for the pediatric upper airway. Compared with the Myer-Cotton system, our measurement is non-invasive and not limited by the size of the endotracheal tubes.

For each individual curve y, from TVC to tracheal carina, we build an atlas that is adapted to the age of the corresponding subject, and we compare the curve y with the minimal curve of the atlas, C_{lower_fence}, because this minimal curve can be considered the

	Two sample t-test		
	Weighted pointwise boxplots	Functional boxplots	Weighted functional boxplots
Pre v.s. CRL	2.1e-07	**1.4e-11**	2.6e-11
Pre v.s. Post	1.7e-03	1.4e-03	**5.2e-04**
Pre v.s. Post&CRL	7.7e-07	1.7e-09	**1.4e-09**
Post v.s. CRL	9.6e-03	**4.4e-05**	9.7e-05
	Wilcoxon rank sum test		
	Weighted pointwise boxplots	Functional boxplots	Weighted functional boxplots
Pre v.s. CRL	7.2e-05	6.0e-05	**5.6e-05**
Pre v.s. Post	1.9e-03	1.1e-03	**6.5e-04**
Pre v.s. Post&CRL	8.2e-05	6.7e-05	**5.7e-05**
Post v.s. CRL	1.6e-03	**9.9e-05**	3.9e-04

Table 4.5: P-values of two types of tests on the scores for pediatric upper airways estimated based on weighted pointwise boxplots, functional boxplots and weighted functional boxplots. Notes: Pre represents the SGS pre-surgery group, Post represents the SGS post-surgery group, CRL represents the normal control group, and Post&CRL represents the union of the SGS post-surgery and normal control groups.

minimal cross-sectional area of a *normal* airway at that age. With the minimal curve as the reference of the airway's cross-sectional area, our scoring system is defined as:

$$Score(\mathrm{y}) = \min_{x}((\mathrm{y}(x) - C_{lower\_fence}(x))/C_{lower\_fence}(x)). \tag{4.15}$$

If the whole curve y is above the minimal curve, the score will be non-negative; otherwise it will be negative. Since all $\mathrm{y}(x) \geq 0$, the lower bound for the score is -1. While there is no theoretical upper bound to this definition, the score will be upper-bounded in practice by the largest observable cross-sectional areas for a given age. That is, the score of the cross-sectional area curve for a pediatric upper airway is within $[-1, \infty)$, where -1 indicates a fully closed airway with a zero cross-sectional area somewhere. A negative score indicates a potential stenosis and a normal control subject usually has a non-negative score. Overall, the higher the score, the more normal the corresponding subject will be. Note that our measurement is not directly comparable to the Myer-Cotton system, as the Myer-Cotton system computes within-subject scores by estimating the cross-sectional area of what should be considered a non-constricted airway. Our score on the other hand makes use of population data contained in the normal control atlas to define what a minimum normal cross-sectional area should be. Nevertheless, the two scoring systems can be made "roughly comparable" by setting all positive atlas-derived scores to zero (indicating a healthy airway) and negating all negative scores.

Scores for all subjects. Based on our scoring system, we score the pediatric upper airways not only for SGS patients but also for the normal controls. The scores shown in Fig. 4.17 are estimated based on the atlases built by weighted pointwise boxplots, functional boxplots, and weighted functional boxplots. The subjects shown in the plots include 68 normal controls and 17 SGS patients (6 pre-surgery, 11 post-surgery). Among the total 19 SGS patients two subjects that have completely obstructed airways are directly scored as -1 and are not shown in Fig. 4.17. Within the 11 post-surgery subjects, some of them have no stenoses after surgery, others show improvement in the airway but still exhibit slight stenoses.

To verify whether there is a statistically significant score difference among groups, we use two types of hypothesis tests, the two sample t-test [Snedecor and Cochran, 1989] with the normal distribution assumption for samples and the Wilcoxon rank sum test [Siegel, 1956], a non-parametric statistical hypothesis test for populations that cannot be assumed to be normally distributed. Table 4.5 shows the testing results among the following three groups: SGS pre-surgery, SGS post-surgery, and control subjects. We use three different analysis approaches: weighted pointwise boxplots, functional boxplots, and weighted functional boxplots. In each test between two groups the smallest p-value is in highlighted boldface in Table 4.5. Overall, the results suggest that the weighted functional boxplot is superior to the standard functional boxplot and the weighted pointwise boxplot in separating the SGS pre-surgery subjects from the SGS post-surgery subjects or/and the normal controls, though all results are highly statistically significant. Note that it is not obvious that post-surgery and normal control subjects should be well distinguishable as a successful surgery should result in a post-surgery airway which should be close to normal.

A closer look at the scores resulting from the three different analysis methods reveals that the scores for the weighted pointwise boxplot (shown in Fig. 4.17(left)) mix the SGS post-surgery subjects with the normal controls. While this could be desired, as a successful surgery should result in a more “normal-looking” airway, more importantly the weighted pointwise boxplot assigns negative weights to some of the normal controls. This suggests potential stenoses in the control airways and conflicts with the definition of our scoring system. This negative score effect for normal controls is not present for the weighted functional analysis approach, but it also appears when using the un-weighted functional analysis (see details below). This suggests that the weighted functional analysis is more appropriate for this application.

Considering the scores of the functional boxplot (Fig. 4.17 (middle)), there are two normal control subjects scored with negative values: CRL32 and CRL102, whose curves and age-matched atlases are shown in Fig. 4.18. For these two subjects, parts of their curves are below the atlases built by the functional boxplot. The age of CRL32 is 8 months which is near the lower age boundary with limited information for atlas-building, and CRL102 is at the age of 182 months and therefore suffering from local data sparsity in our current

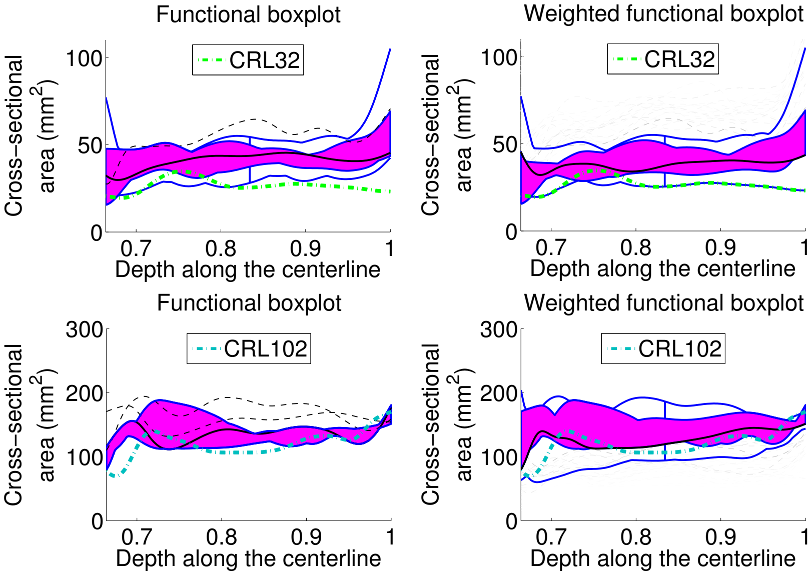


Figure 4.18: Two control subjects, represented by colored dashed curves and their age-matched atlases. The curves obtain negative scores when using the functional boxplot and non-negative scores when using the weighted functional boxplot.

dataset. Compared with the functional boxplot, the weighted functional boxplot shows better performance given the limited data information and the local data sparsity as also shown in Section 4.2. The curves for these two subjects are fully within the atlases built by weighted functional boxplots and scored with non-negative values, as shown in the right column of Fig. 4.18.

Score comparison of pre- and/or post-surgery. Table 4.6 shows the scores for SGS subjects using weighted pointwise boxplots, functional boxplots, and weighted functional boxplots, and it compares them with the clinical diagnosis based on the Myer-Cotton grading system. The scores computed using our scoring system are converted to be roughly comparable to the corresponding Myer-Cotton values as described in Section 4.3.5. Scores that are outside of the ±20% deviation of the clinical diagnosis and that are zero for subjects with stenoses or non-zero for subjects without stenoses are shown in boldface. Table 4.6 shows that the weighted pointwise boxplot frequently gives results which are not what would

Patient Id	**Surgery**	**Myer-Cotton**	**WPB**	**FB**	**WFB**
SGS03	Pre	80-90%	86.0%	85.6%	85.6%
SGS07	Pre	85%	77.4%	74.5%	74.5%
SGS11	Pre	50%	59.2%	54.8%	54.8%
SGS12	Pre	70%	70%	70%	68.7%
SGS13	Pre	60-70%	**9.9%**	**37.8%**	**34.0%**
SGS18	Pre	60-70%	69.2%	69.2%	68.8%
SGS03_V3	Post	0%	**0.3%**	**1.1%**	0%
SGS05	Post	0%	0%	0%	0%
SGS06	Post	40-50%	**0%**	35.1%	**13.9%**
SGS08	Post	0%	0%	0%	0%
SGS09	Post	50%	**19.6%**	59.1%	57.8%
SGS10	Post	grade I	**0%**	39.9%	27.6%
SGS14	Post	50%	**0%**	**26.1%**	**0%**
SGS17	Post	0%	**16.3%**	0%	0%
SGS07_V3	Post	30%	**14.5%**	**9.3%**	**9.3%**
SGS04_V3	Post	grade I: 10%	**0%**	5.6%	**0%**
SGS01_V3	Post	15-20%	31.4%	30.2%	29.6%

Table 4.6: Comparison of the scores for SGS subjects using three different methods with the clinical diagnosis based on the Myer-Cotton grading system. Notes: Weighted pointwise boxplots (WPB), functional boxplots (FB), and weighted functional boxplots (WFB). The scores are converted based on the correspondence between our scoring system and the Myer-Cotton system in Section 4.3.5. Grade I represents an obstruction within (0% - 50%].

be expected from the Myer-Cotton scores. The scores based on the functional boxplot and the weighted functional boxplot both give results which are more comparable to the Myer-Cotton scoring.

To further reveal the differences between functional boxplots and weighted functional boxplots, we quantitatively compare the four SGS subjects pre- and post-surgery shown before in Section 4.3.5. In general, as shown in Fig. 4.19 the scores of these four subjects increase after surgery for both methods, which indicates that all subjects' airways improved from the surgeries.

Classification of Control and SGS subjects. To demonstrate the classification accuracy

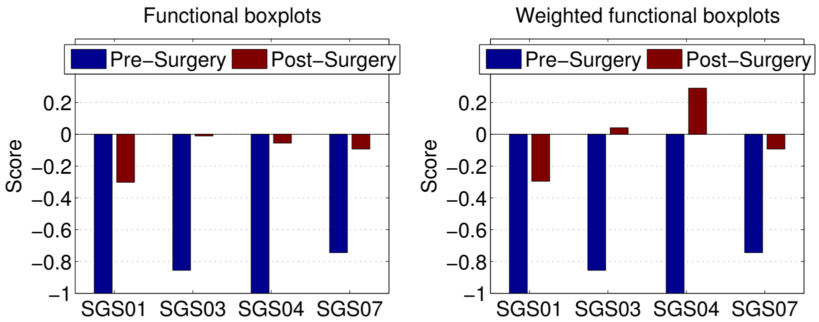


Figure 4.19: Quantitative comparison of the scores for four SGS subjects before and after surgery using functional boxplots and weighted functional boxplots for atlas-building.

	Pre (P)	**Pre (P)**	**Pre (P)**
	vs. CRL (N)	**vs. Post (N)**	**vs. Post&CRL (N)**
	TP = 5 FP = 1	TP = 5 FP = 0	TP = 5 FP = 2
Weighted	FN = 1 TN = 67	FN = 1 TN = 11	FN = 1 TN = 77
pointwise	TPR = 0.83, FPR = 0.01	TPR = 0.83, FPR = 0.0	TPR = 0.83, FPR = 0.03
boxplots	PPV= 0.83, ACC = 0.97	PPV = 1.0, ACC = 0.94	PPV = 0.71, ACC = 0.96
	TP = 6 FP = 2	TP = 5 FP = 1	TP = 5 FP = 8
Functional	FN = 0 TN = 66	FN = 1 TN = 10	FN = 1 TN = 71
boxplots	TPR = 1.0, FPR = 0.03	TPR = 0.83, FPR = 0.09	TPR = 0.83, FPR = 0.10
	PPV= 0.75, ACC = 0.97	PPV = 0.83, ACC = 0.88	PPV = 0.38, ACC = 0.89
	TP = 6 FP = 0	TP = 5 FP = 1	TP = 6 FP = 3
Weighted	FN = 0 TN = 68	FN = 1 TN = 10	FN = 0 TN = 76
functional	TPR = 1.0, FPR = 0.0	TPR = 0.83, FPR = 0.09	TPR = 1.0, FPR = 0.04
boxplots	PPV=1.0, ACC = 1.0	PPV = 0.83, ACC = 0.88	PPV = 0.67, ACC = 0.96

Table 4.7: The confusion matrices among groups: SGS pre-surgery (Pre), SGS post-surgery (Post), and control (CRL). Notes: P (positive), N (negative), TP (true positive), FP (false positive), FN (false negative), TN (true negative), TPR (true positive rate), FPR (false positive rate), PPV (positive predictive value), and ACC (accuracy).

for separating SGS pre-surgery subjects from SGS post-surgery and/or control subjects we compute the confusion matrices [Provost and Kohavi, 1998] based on the scores estimated from weighted pointwise boxplots, functional boxplots and weighted functional boxplots, as shown in Table 4.7. We label the SGS pre-surgery subjects as positive (P), and we label both SGS post-surgery and control subjects as negative (N). We repeatedly take one subject out for testing (i.e., leave-one-patient-out) and trained a Support Vector Machine (SVM) [Cortes and Vapnik, 1995] on the data from the remaining subjects. We use a linear

SVM for our experiments. For the confusion matrices, we calculate the numbers of true positive (TP), true negative (TN), false positive (FP), false negative (FN) instances. Besides, we use the true positive rate (TPR = TP/(TP+FN)), the false positive rate (FPR = FP/(FP+TN)), the positive predictive value (PPV = TP/(TP+FP)), and the accuracy (ACC = (TP + TN)/(P+N)) to further assess the performance of the classification between SGS pre-surgery subjects and others.

In the classification between SGS pre-surgery and control subjects, for weighted pointwise boxplots one SGS pre-surgery subject is regarded as a normal control and one control subject is regarded as pre-surgery; for functional boxplots there are two false positive subjects, which means two children test positive but actually do not have subglottic stenoses. In contrast, the weighted functional boxplot result shows no false positives or false negatives and yields 100% classification accuracy. In the classification between SGS pre- and post-surgery subjects, the weighted pointwise boxplot has one misclassified subject, and both functional boxplots and weighted functional boxplots have one false positive and one false negative. The misclassified cases will be discussed in detail in the next section. The accuracy of classifying the pre- and post-surgery subjects using the weighted functional boxplot is about 88%. If we combine the SGS post-surgery subjects with the normal controls, the accuracy of the weighted functional boxplot increases to about 96%, which is higher than that of the functional boxplot.

It is important to note that no fully conclusive statements can be made based on the presented classification results. While Table 4.7 indicates better prediction performance when using WFBs, further tests with larger sample sizes are needed to substantiate our claims.

Discussion of SGS outliers. Based on the above discussion, the scores computed from

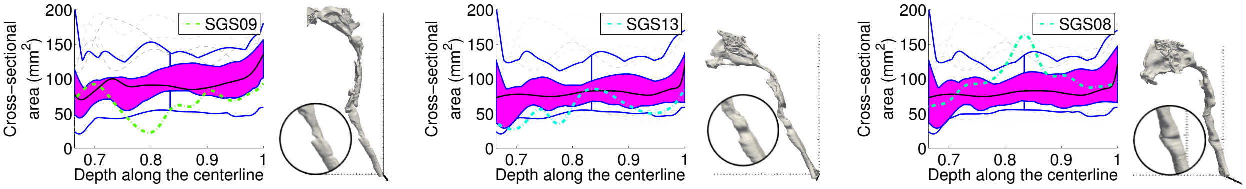


Figure 4.20: Three outliers in Fig. 4.17 for both functional boxplots and weighted functional boxplots. (a) SGS09 is post-surgery while having a low score more consistent with a pre-surgery subject; (b) SGS13 is pre-surgery while mixed into the post-surgery group; (c) SGS08 is post-surgery appearing as a normal control subject consistent with near normal post-operative airway.

weighted functional boxplots can be used to roughly divide the pediatric upper airways into three different groups: the SGS pre-surgery group (score in [-1, -0.5)), the SGS post-surgery group (score in [-0.5, 0)), and the normal control group with a score larger or equal to zero. In this classification, three representative subjects should be discussed. Namely, SGS09, SGS13 and SGS08, which are shown in Fig. 4.20 together with their cross-sectional area curves in the age-matched atlases and their airway geometries.

SGS09 shown in Fig. 4.20(a) corresponds to the false positive subject in the confusion matrix of the weighted functional boxplot in Table 4.7. This subject is post-surgery with a 50% airway obstruction based on the clinical diagnosis. From the cross-sectional area curve and the zoomed-in part of the geometry we can clearly see the subglottic stenosis with a score of -57.8%, thus resulting in being classified as a SGS pre-surgery case, which is sensible.

SGS13 is a pre-surgery subject and according to the clinical diagnosis has a $60-70\%$ obstruction in the airway. From Fig. 4.20(b), we see two stenoses in the airway, as confirmed by the surgeon. However, because of its score, -34.0%, it is the false negative subject in the confusion matrix of the weighted functional boxplot in Table 4.7 and it is classified as belonging to the SGS post-surgery group.

The last case, SGS08, is post-surgery and has a very high score of 60.5%, indicating a normal subject. As shown in Fig. 4.20(c), it has a comparable airway size to the atlas and its airway geometry also indicates no stenosis existing in the airway. This subject is confirmed by the surgeon as near normal caliber airway and hence could also be sensibly classified as normal. This case suggests that our scoring system can reliably be used to assess abnormalities in pediatric upper airways.

4.4 Conclusion

In this chapter a method was proposed to compute weighted functional boxplots and used for building spatiotemporal atlas. It was applied to construct a pediatric airway atlas to assess children with subglottic stenosis and to construct a corpus callosum atlas capturing the impact of aging. Also, a scoring system was defined for pediatric airways based on the statistical atlas to quantitatively measure the severity of subglottic stenosis in children. The proposed method is general and easy to compute, and it allows robust statistical description of functional, shape, and image data.

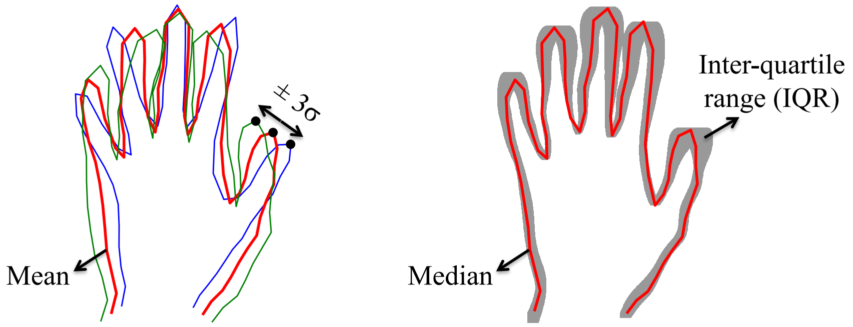
CHAPTER 5 : STATISTICS OF GROUP DIFFERENCES

This chapter[1] presents two statistical analysis approaches to identify group-level shape differences for both cross-sectional and longitudinal data. In particular, a new method is proposed for cross-sectional data, which is based on the ordering of shapes using band-depth, as discussed in Section 5.1. This band-depth is used to non-parametrically define a global depth for a shape, which allows globally quantifying differences. Using the depth-ordering of shapes also allows the detection of localized shape differences. For longitudinal data studied in Section 5.2, the presented method further considers the within-subject correlation that is ignored in the cross-sectional data, and it introduces shape trajectories that summarize shape variations for testing group differences. A generalized Bhattacharyya distance is presented to study statistical differences in two distributions of trajectories. This not only allows considering second-order statistics, but also it serves as the test-statistic during hypothesis testing. Both methods are tested on synthetic and real data.

5.1 Shape Analysis for Cross-Sectional Data

Analyzing and comparing three-dimensional brain structures or objects in general can be as simple as comparing volumes. While informative, such a global measure cannot fully describe the changes between objects. Shape analysis approaches have been proposed to assess object properties beyond global volume and to characterize shape variations across subjects

[1] The work presented in this chapter is based on previous papers [Hong et al., 2014b, Hong et al., 2015a, Hong et al., 2015c].



(a) Point distribution model (PDM) (b) Depth-ordering model (DOM)

Figure 5.1: Comparison between two types of models for capturing shape variations. The three shapes in PDM (a) correspond to the mean and varying shapes along the first mode at ±3 standard deviations. In DOM (b), the red shape is the median of the shape population and the grey area is the region covered by 50% of the shapes at the top ranking list of the populations, similar to the inter-quartile range (IQR) for scalar values visualized as part of a box-plot.

and between subject populations (*cf*. [Nitzken et al., 2014]). For instance, many shape analysis methods are based on the classical point distribution model (PDM) [Cootes et al., 2004], which captures shape variations by computing a mean shape and the major modes around the mean of corresponding points in a set of shapes as illustrated in Fig. 5.1(a). The PDM assumes a Gaussian distribution of the points around the mean shape.

In contrast, shape characterizations built on concepts of order statistics have been explored recently [Whitaker et al., 2013, Hong et al., 2013a, Hong et al., 2014a], as discussed in Chapter 4. These methods utilize depth-ordering of shapes to generalize order statistics, for example, the median and the inter-quartile range (IQR), to shapes, effectively obtaining the equivalent of a box-plot for shapes. Fig. 5.1(b) illustrates the depth-ordering model (DOM). Using shape-descriptions based on depth-ordering makes it possible to perform shape analysis without making strong distributional assumptions. While these approaches focus on using DOM to analyze shape variations within one population, in this chapter we address the challenge of differentiating subject populations, for example subjects with a disorder versus normal controls, based on the depth-ordering model.

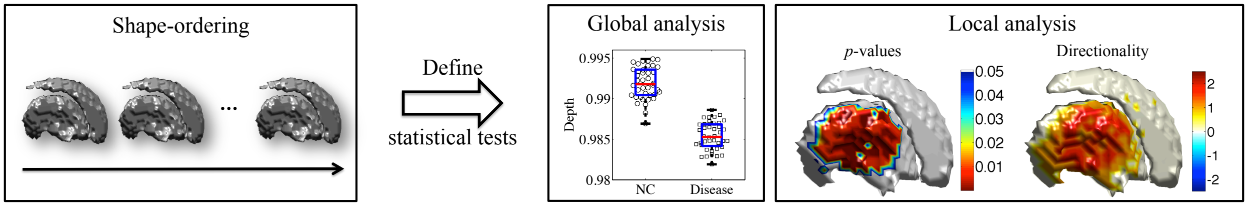


Figure 5.2: Overview of the depth-ordering-based shape analysis. Based on the depth and the ordering of a shape population, statistical tests are defined to globally separate control and disease groups using global analysis with a scalar value (depth) for each shape. Statistical difference for global analysis results can be established through permutation testing. Equivalently, local shape differences can be detected using local analysis with a corresponding local permutation test, resulting in *p*-values on the surface of a shape to establish local shape differences between populations. The directionality of the shape differences (inflation versus deflation) can also be determined.

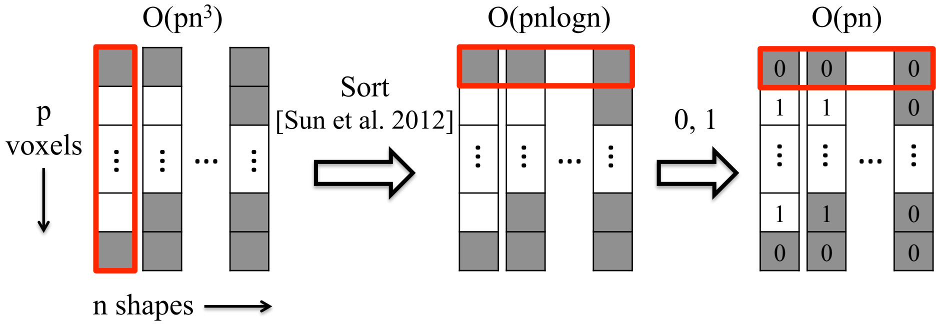
Existing methods for population-based shape analysis can be roughly subdivided into two categories: methods for global analysis and methods for local analysis. Global analysis methods are designed to detect whether population shape differences exist [Loncaric, 1998, Reuter et al., 2009, Wachinger et al., 2014], but they generally cannot locate or characterize these shape differences. This limits interpretability of results and consequentially insights into the underlying biological processes. The main attraction of such methods is that they often avoid establishing dense correspondences between shapes through registration. In contrast, local analysis methods require some form of point-to-point correspondence between shapes to allow precise local shape analysis. Establishing these correspondences is highly non-trivial and arguably one of the main sources of inaccuracy, because any misregistration may create artifacts in the final shape analysis results. Nevertheless, a variety of methods for local shape analysis have been proposed and successfully used [Miller, 2004, Davies et al., 2008, Cates et al., 2008, Chung et al., 2008, Hosseinbor et al., 2014]. Shape representations that fit an a-priori model to the data have also been used successfully, although they still need to establish some form of one-to-one correspondence via model fitting [Styner et al., 2004, Yushkevich and Zhang, 2013].

The work presented in this section aims to explore a new method that allows for both global and local shape analyses, and the method would only need limited (e.g., rigid or affine) spatial alignment of shapes. This is achieved by using the notion of band-depth in statistics [López-Pintado and Romo, 2009, Hong et al., 2013a], which provides a notion of centrality of a shape with respect to a reference dataset. The deeper the shape the more similar it is to all other shapes in the dataset. We use this notion to detect small morphometric differences between two populations of neuroanatomical structures. Fig. 5.2 shows the general principle of the approach. First, shapes are depth-ordered with respect to a reference dataset. We then define a global statistical test directly on the shape depth, as well as a local statistical test to assess local shape differences. This method is free from strong distributional assumptions by using principles from non-parametric order-statistics.

5.1.1 Depth-Ordering of Shapes

Generalizing concepts from order-statistics to shape analysis faces the challenge that there is no canonical ordering of shapes. To define such an ordering we can use the concept of band-depth and ordering of functions and extend it to shapes, as discussed in Section 4.3.1. The intuition of ordering functions based on band-depth is that the deeper a function is buried within a dataset the more central it is. The deepest function corresponds to the within-sample median function. Once defined, this ordering can be used to generalize traditional order statistics, such as the median or the inter-quartile range, to functions. What makes band-depth attractive for *shape*-ordering is that shapes can be represented by *indicator functions*, i.e., by binary functions that are 1 inside and 0 outside of a shape.

Despite its conceptual simplicity, one of the main limitations of band-depth is its computational complexity. As we discussed in Section 4.1.3, the time complexity of functional



(a) Original algorithm (b) Sun's algorithm (c) Our algorithm

Figure 5.3: Illustration of the computational complexity of band-depth calculation for all observations. The computational complexity of the original algorithm is cubic, and our algorithm reduces it to linear, with respect to the number of observations.

boxplots is $O(pn^{J+1})$, where p is the dimension of each observation, n is the number of the observations, and J is the number of observations used to define the band. Take $J = 2$ for example, that is, the band is defined by two observations[2]. The original algorithm [Sun and Genton, 2011] has $O(pn^3)$ complexity for computing band-depth for all observations, as illustrated in Fig. 5.3(a). This makes this algorithm unsuitable for large numbers of functions. To allow for faster computations recently a fast method to compute band-depth has been proposed [Sun et al., 2012]. This method is based on computing local curve ranks and requires sorting of values from all curves at the same position. The algorithm uses a rank list to compute the frequency with which one point of a curve is contained by other curves at the same location. However, the proposed algorithm is ill-suited for binary shape representations as it does not consider ranking ties during sorting, which frequently appear when sorting binary values. Fig. 5.3(b) illustrates this ranking approach.

We observed that for binary representations sorting can be avoided, as at any location only two values are possible, that is, no sorting but simple addition and subtraction is

[2] Due to computational complexity this is typically how band-depth is evaluated, because choosing $J > 2$ results in even higher computational costs.

enough to measure whether a point of a shape is contained by the band formed by shapes of a population at that location. In this work we use this binary function representation to compare shape populations. To improve the computational efficiency of our model, we propose a novel fast algorithm to compute the band-depth of shapes represented by binary maps. Most importantly we demonstrate how band-depth can be used to provide both global and local statistical tests to differentiate between shape populations.

Given a set of n shapes represented as 3D binary volumes, $\{\mathrm{Y}_1, \mathrm{Y}_2, ..., \mathrm{Y}_n\}$, with dimension of (s_x, s_y, s_z), we vectorize them into binary vectors $\mathrm{y}_i \in \{0,1\}^p$ $(1 \leq i \leq n)$, where $p = s_x \times s_y \times s_z$. The computation of modified band-depth (MBD, see Section 4.1.2) can then be accomplished efficiently for $J = 2$ using our algorithm as follows:

S0) Given n binary volumes, $\{\mathrm{Y}_i\}_{i=1}^n$, vectorize them: $\mathrm{y}_i \in \{0,1\}^p$, p is the number of voxels in a binary volume.

S1) At each location of a volume, k $(1 \leq k \leq p)$, for a given value $v(k) \in \{0,1\}$, we count the number of functions that have a value larger (n_a), smaller (n_b) or equal (n_t) to $v(k)$:

a) if $v(k) = 0$, then $n_a = \sum_{i=1}^n \mathrm{y}_i(k)$, $n_b = 0$, and $n_t = n - n_a - 1$;

b) if $v(k) = 1$, then $n_a = 0$, $n_b = \sum_{i=1}^n (1 - \mathrm{y}_i(k))$, and $n_t = n - n_b - 1$.

This procedure allows evaluating how often a voxel of a volume is larger (or smaller) than voxels of other volumes at the same location.

S2) Based on the numbers n_a and n_b defined in step S1, we can calculate the number of bands containing $v(k)$, for example, from y_i. First, let B be a band such that y_i is not one of its observations, and its bounds at k are denoted as b_{min} and b_{max}. For

$J = 2$, we have the following possible scenarios, i.e., the combinations of possible B that contain $v(k)$:

a) $b_{min} < v(k) < b_{max}$ ($n_a n_b$ such combinations),

b) $b_{min} = v(k)$ and $v(k) < b_{max}$ ($n_a n_t$ such combinations),

c) $b_{min} < v(k)$ and $v(k) = b_{max}$ ($n_b n_t$ such combinations),

d) $b_{min} = v(k)$ and $v(k) = b_{max}$ ($n_t(n_t - 1)/2$ such combinations).

In addition, we have to account for bands that have y_i as one of its observations. There are $n-1$ such bands. By collecting all combinations the total number of bands ($J = 2$) containing the value $v(k)$ is:

$$C_k(v(k)) = n_a n_b + n_a n_t + n_b n_t + n_t(n_t - 1)/2 + (n - 1) \ .$$

For binary functions n_a and n_b cannot simultaneously be different from zero, which simplifies the expression to

$$C_k(v(k)) = (n_a + n_b)n_t + n_t(n_t - 1)/2 + n - 1 \ .$$

The above equation provides us with the number of times a voxel of a volume is contained by voxels of any two volumes at the same location.

S3) The modified band-depth for a shape y_i is then

$$MBD(\mathrm{y}_i) = \frac{1}{p}\binom{n}{2}^{-1} \sum_{k=1}^{p} C_k(\mathrm{y}_i(k)) \ ,$$

where the notation $C_k(\mathrm{y}_i(k))$ denotes computing C_k in step S2 based on the coefficients n_a, n_b, n_t given by the value of y_i at location k.

We can connect this special case scenario to the original MBD definition. In particular, the MBD definition in [Sun and Genton, 2011] is

$$MBD_{n,J}(\mathrm{y}) = \sum_{j=2}^{J} MBD_n^{(j)}(\mathrm{y}), \quad \text{and}$$
$$MBD_n^{(j)}(\mathrm{y}) = \frac{1}{C} \sum_{1 \leq i_1 < i_2 < \ldots < i_j \leq n} \lambda_p\{A(\mathrm{y}; \mathrm{y}_{i_1}, \ldots, \mathrm{y}_{i_j})\} \ ,$$

where $A_j(\mathrm{y}) \equiv A(\mathrm{y}; \mathrm{y}_{i_1}, ..., \mathrm{y}_{i_j})$ and $A_j(\mathrm{y}) \equiv \{\mathbf{x} \in \mathbb{I} : min_{r=i_1,...,i_j}\mathrm{y}_r(\mathbf{x}) \leq \mathrm{y}(\mathbf{x}) \leq max_{r=i_1,...,i_j}\mathrm{y}_r(\mathbf{x})\}$, $\lambda_p(\mathrm{y}) = \lambda(A_j(\mathrm{y}))/\lambda(\mathbb{I})$, λ is the Lebesgue measure on $\mathbb{R}^p$ and p is the observation's dimension, i.e., the number of voxels in a binary represented shape. First, for bands made of two observations we have $J = 2$ and thus $C = \binom{n}{2}$. Second, for volumes of size p, we have $\lambda(\mathbb{I}) = p$. Finally, for binary volumes $\sum_{1 \leq i_1 < i_2 < \cdots < i_j \leq n} \lambda(A_j(\mathrm{y}_i))$ is equivalent to $\sum_{k=1}^{p} C_k(\mathrm{y}_i(k))$.

This algorithm computes MBD of all shapes in a population at the same time, not sequentially. Compared to the original band-depth algorithm our computational complexity is reduced from $O(pn^3)$ to $O(pn)$, as shown in Fig. 5.3. This makes it possible to compute band-depth for large populations and large multi-dimensional shapes.

Note that, when choosing two shapes from a population to define a band, we do not exclude the one that is being evaluated from the admissible permutations of the population. Instead, all the shapes in the population will be considered to define the bands. So, in the algorithm the number of observations is n, i.e., the number of shapes in a population. One can also exclude the current shape and compute its depth value with respect to a population of the remaining shapes. This will not change the ranking because the current shape is always contained within the band formed by shapes including itself, resulting in the same constant difference in the depth value of each shape, compared to using n shapes as a population.

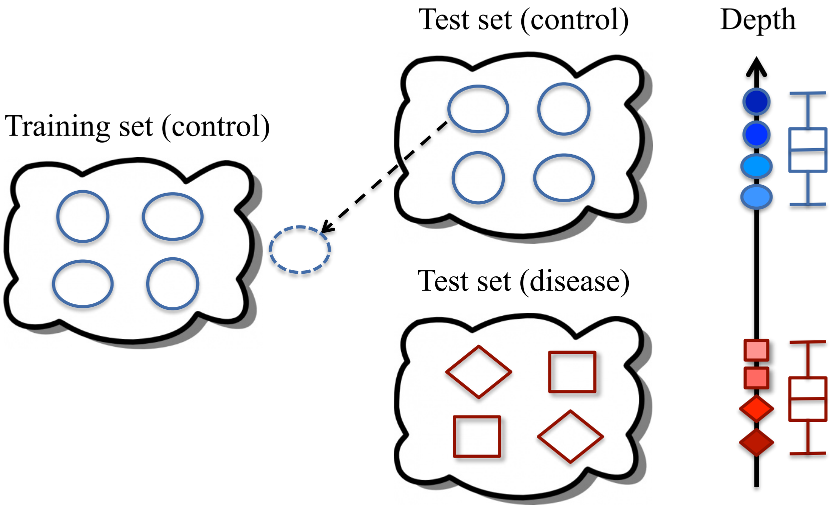
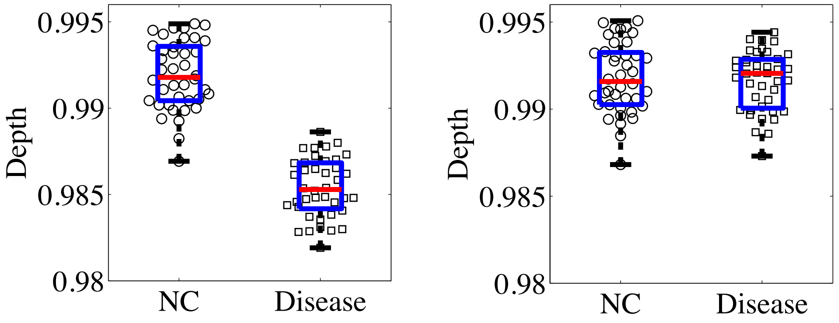


Figure 5.4: Illustration of global shape analysis. A group of control subjects is chosen as a training set, and the band-depth of each test shape from control or disease groups, is computed with respect to the training set. The boxplots on the right show that in general the control group would have larger depths than the disease group if control subjects are selected as the reference/training population.

5.1.2 Statistics Using Depth-Ordering

Band-depth computed in Section 5.1.1 can measure the relationship between a shape and a reference population. A higher value indicates the shape is closer to the median of the reference population, and a lower one indicates the shape is a potential outlier with respect to the reference population. Based on this property of band-depth we can perform global shape analysis as described in Section 5.1.2 as well as local shape analysis as described in Section 5.1.2. For all these analyses we assume that shapes have been pre-aligned as appropriate. Typically, this will either involve a rigid, similarity or affine alignment of shapes to a given template or to a template obtained by some form of unbiased atlas-building method. The choice of transform will depend on the objective of a given study. For example, if size differences should be included, rigid alignment would be appropriate.

The key ingredient to performing statistics using depth-ordering is to compute depth-ordering with respect to a reference population of shapes that are used as a non-parametric



(a) w.r.t a training set (b) without a training set

Figure 5.5: Global shape analysis on synthetic data using band-depth computed with (a) and without (b) a training population. The training population allows detection of global shape differences.

model of shapes. In particular, while band-depth would typically be computed for each shape of a population with respect to all the shapes, in our definition of statistical tests we will make use of a reference population to which other shapes are compared, one at a time. Intuitively, one establishes a reference population (for example a population of normal control subjects) and then for another given shape not part of this population, one tests how deep this shape would be buried with respect to the reference. Note that if a new subject is added into the reference population, the band-depth of each shape should be recomputed with respect to the new reference population.

Global Shape Analysis

The goal of global shape analysis is to establish if there are shape differences between two different groups. To this end we assign a scalar description of shape, here a measure of depth, to each shape. Given a dataset $\{R_i\}$ containing a training population of shapes, we compute the band-depth for a given datum D, from a set of input test shapes $\{D_j\}$. We first compute the band-depth for all the data in a population of $\{R_i\} \bigcup D$ using the algorithm described in Section 5.1.1, and then we assign the resulting band-depth of D to the datum D, denoted as $BD(D; \{R_i\})$, as illustrated in Fig. 5.4. A larger depth value indicates the test shape being closer to the training set. Typically, we choose a subgroup of control subjects as the training population, so in general, the control test set would have larger depths than the disease one.

In the proposed method the training population forms a “yard-stick” by which to judge data-depth. This is substantially different from directly computing the band-depths for the full dataset $\{D_j\}$, which would be problematic. Most easily this effect is illustrated by considering scalar-valued data of two groups with the same variance but strongly different means. The most central element (the median) in such a case would be half-way between the means. This median element will have the largest band-depth. Band-depth will decrease moving away from this median element towards smaller *and* larger elements. Hence, the groups would become indistinguishable even though they are clearly different. Equally problematic would be to compute band-depth *separately* for both groups in which case they cannot be meaningfully compared directly as the band-depth values become relative to each group. To illustrate this effect Fig.5.5 shows the results of the proposed approach using a reference population and the approach using separate band-depth computations for the groups of the synthetic striatum data described in detail in Section 5.1.4. Our proposed approach using a reference group can clearly differentiate the populations whereas a computation of the band-depth without a reference group is not successful.

Permutation test. To measure if global shape differences are statistically significant between control and disease groups, we use a permutation test on the mean depth of the control group versus that of the disease group. The null hypothesis of the permutation test is that both the control and the disease groups have the same mean depth with respect to the training group. Specifically, since the depth of each test shape has been computed with respect to

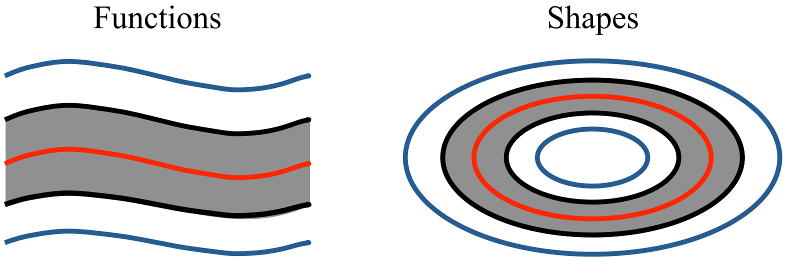


Figure 5.6: Illustration of the median and α-central region. The median function / shape (red) is the most central one of a population, and the α-central region (grey) is the band delimited by the α proportion of the deepest functions/shapes. For example, the grey region is the 0.6-central region because it is built by three out of five functions/shapes.

the training group, we first compute the mean depth difference between the control test group and the disease test group with no permutation. Then, in each permutation we exchange the subjects in the two test groups and compute the mean depth difference between the two permuted test groups[3]. We count the number of times that the mean depth difference is larger than the one computed without permutation, and the proportion of larger values to the total number of values is the associated p-value. We test for statistical significance at a level of 0.05, i.e., we declare the statistical significance for a p-value ≤ 0.05.

Local shape analysis

For the local shape analysis the goal is not only to establish whether there are shape differences, but also where these shape differences occur. To better understand the local analysis, we first introduce its key ingredient, the α-central region as shown in Fig. 5.6, which allows us to define a local shape analysis approach using depth-ordering. Specifically, based on the ordering of shapes, we can not only compute the median shape, the most central shape of a population, but also α-central regions ($0 < \alpha \leq 1$), similar to for example

[3] Since the permutation is performed between test groups and there is no change in the training group, the depth value of each test subject will keep unchanged during this process. Hence, we do not need to recompute their band-depth values.

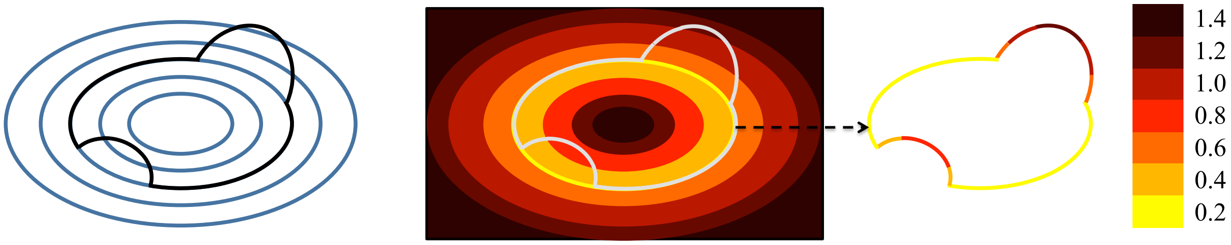
the inter-quartile-range (IQR) for scalar-valued cases. The band of the α-central region is delimited by the α proportion of the deepest shapes in the ranking list, which is defined as

$$CR_\alpha = \{B(\mathrm{y}_{[1]}, \mathrm{y}_{[2]}, ..., \mathrm{y}_{[q]}),\ q = \lceil \alpha n \rceil\}\ , \tag{5.1}$$

where $\mathrm{y}_{[1]}, ..., \mathrm{y}_{[q]}$ $(q \geq 1)$ denote the ordered shapes, from the deepest one to the qth-deepest one, $B(\cdot)$ is the band as defined in Section 5.1.1. The grey band shown in Fig. 5.6 illustrates a central region with α equal to 0.6, because the three deepest shapes, out of a total of five shapes in a population, are used to build this central region.

The local analysis is based on the α-central regions of a reference population to detect the location of shape differences. For the shapes in the reference population, starting from the deepest shape (its median shape) we gradually add shapes in order of decreasing band-depth. Each set of shapes defines a specific α-central region and a particular location is assigned the α value of the first α-central region that covers it. The resulting map describes the "centrality" of a shape population at each point in the domain. The test shape can be overlaid on this centrality map and the corresponding α-values will be recorded on its surface, thus providing a local measure of shape abnormality.

Fig. 5.7 illustrates this concept for a population of two-dimensional shapes. Given the reference shapes, the blue shapes shown in Fig. 5.7(a), we build a centrality map based on the band-depth of each shape. As shown in Fig. 5.7(b), the deepest shape has the lowest α-value (light yellow) and the most outlying shape has the highest α-value (dark red). A local measure of "belonging" to the population can then be computed for a test shape by tracing the α-central region it traverses, see Fig. 5.7(c). Note that some regions of the shape



(a) Reference population and a test shape (b) Centrality map (c) α-values on the test shape

Figure 5.7: Illustration of local shape analysis. Reference shape population (a) (blue contours) defines a centrality map (b) (light yellow to dark red corresponds to the most to the least central region of the reference population), which provides a local measure (c) (α-values) of how deeply a test shape (the non-ellipse shape) is buried with respect to the reference population. The dilated region is colored by the darkest red with α-values greater than 1.

may not be covered by the reference shape population, for example the top right part of the test shape. To assign an α value in such cases, we use a dilation procedure starting from the boundary of the $\alpha = 1$ central region, evolving at a constant speed until all locations of the volume are covered. This results in α values greater than 1, shown as the regions colored with the darkest red in Fig. 5.7.

Overall, $\alpha > 0$, and a larger α-value indicates a local region with a higher potential to be different with respect to the reference group. Note that the centrality map has no notion of the directionality of shape differences. Instead, it captures both the inflated and deflated shape differences with respect to the reference group without distinguishing them, as the bump and the indented region shown in Fig. 5.7.

Permutation test. To measure whether a local region of a test group is significantly different from another test group, we design a permutation test using the α-values. Specifically, similar to the global analysis, we have three groups of subjects: a group of control subjects as the training set, another group of control subjects as the control test set, and a group of disease subjects as the disease test set. We first compute the band-depth for each shape from the two test groups with respect to the training set. That is, each test shape will be assigned a band-depth value, and its band-depth will be used for ordering in the permutation test. Then, the median of one test group is chosen as the template, and the other test group is left as the reference population. Our goal is to measure the shape differences of the template compared to the reference population, which is ordered based on the associated band-depths of its shapes when computing the α-values for the template.

We perform a permutation test with the null hypothesis being that the α-values recorded on the surface of the template shape are the same for the two test populations. In each permutation, we exchange the subjects in the two test groups to reconstruct a new reference group and reorder its shapes according to the band-depth associated with them. Here, we do not recompute these band-depths but always keep the band-depth values that are associated to a particular shape. Then, new α-values for the template with respect to the new reference population are computed. At each position we count the number of the α-values that are larger than the one computed without permutation, resulting in p-values on the surface of the template. When simultaneously testing numerous hypotheses it increases the risk of false positive (type I errors). This is the problem of multiple comparisons. The false discovery rate (FDR) is one way to control the expected proportion of those incorrectly-rejected null hypotheses during multiple comparisons [Benjamini and Hochberg, 1995]. After FDR-correction, a p-value smaller than 0.05 (e.g., with 10000 permutations) indicates that the corresponding local region of the template is significantly different from that region of the reference population at a significance level of 0.05.

5.1.3 Directionality of Shape Differences

While the concept of band-depth allows us to measure how central a shape is with respect to a reference population, this measure is not signed and thus cannot represent whether a

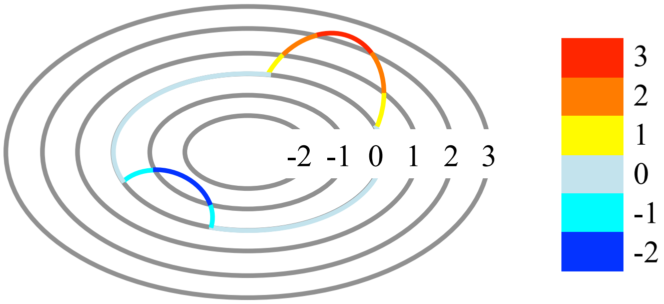


Figure 5.8: Illustration of directionality for a template shape (the ellipse with a bump and an indented region) with respect to the median shape (the ellipse at the zero level-set) of a reference population.

particular region of a shape is inflated or deflated (i.e., atrophied). In other words, band-depth ordering lacks *directionality*. For example, in Fig. 5.7 two different regions can be flagged as abnormal using the centrality map, but it is not possible to tell that one is ”thinner” and the other ”thicker” purely based on band-depth or the assigned α-values.

To measure the directionality of shape differences, we propose to use the signed distance transform. The signed distance map of the median shape of the reference population is computed. The test shape is then overlaid onto this signed distance map, so that the corresponding values can be recorded on the surface of the test shape, as illustrated in Fig. 5.8. As a result, a positive value indicates an enlarged region with respect to the reference median, and a negative value indicates an atrophied region.

5.1.4 Experimental Results

We applied our method on both synthetic and real data. All shapes are pre-aligned using a rigid transformation and represented as binary functions. If the size of a shape is of interest in some applications, one should use a similarity transformation or an affine transformation for the alignment.

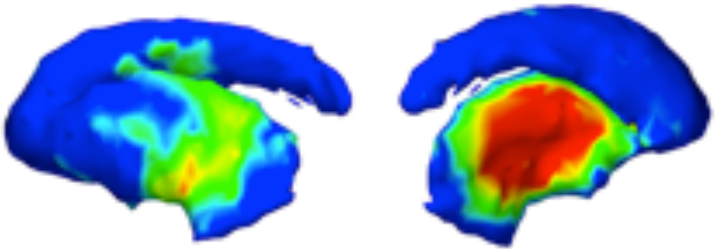
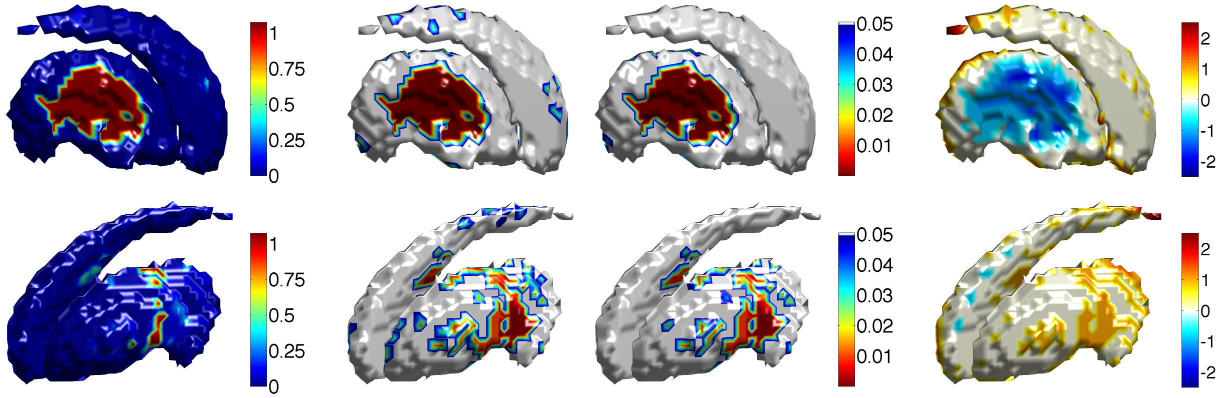


Figure 5.9: Ground-truth of a sample shape (two views from the lateral and medial sides). Colormap on the shape indicates the location and magnitude of the artificial deformations compared to the undeformed shape. Red color corresponds to a large deformation distance.

Synthetic data

In synthetic data we introduce a predefined shape change which we wish to recover using our proposed approach. We use the technique described in [Gao and Bouix, 2012] to generate large data sets of realistic shapes with known deformations. In short, a manifold learning technique is used to generate arbitrarily many shapes from a small training sample. A joint clustering algorithm is then applied to parcellate each shape's surface into small regions which are consistently located across all shapes. Finally, a Log-Euclidean framework is used to introduce smooth, invertible and anatomically realistic deformations to one or multiple regions as defined by the clustering.

For this application, we generated 160 shapes based on 27 manually traced striata. We then modified 80 of them by thickening the putamen. A sample shape with deformation distance is shown in Fig. 5.9. Here, the medial side of the putamen was "pulled out", and because of this large deformation the lateral side of the putamen was slightly deflated due to the diffeomorphic nature of the deformation. We evenly divided 80 normal controls into two groups. One group is used as the training set (NC-Train), and the other for testing (NC-Test). From the 80 abnormal subjects, we randomly picked 40 of them for testing.

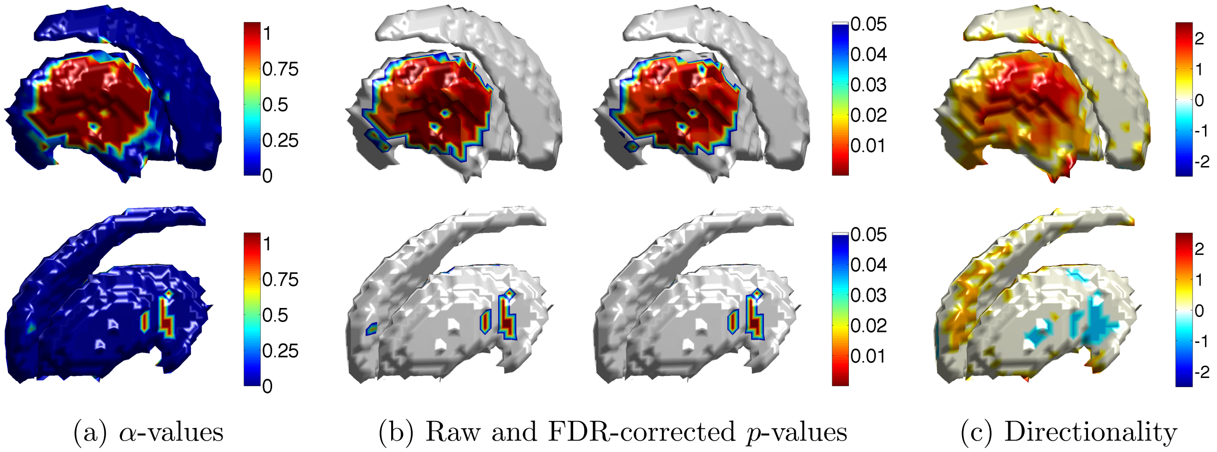


(a) α-values (b) Raw and FDR-corrected p-values (c) Directionality

Figure 5.10: Local shape analysis on synthetic striatum shown from two views, the medial (top) and the lateral (bottom) views, with the median of NC-Test as the template and the disease test group as the reference. (a) The α-values on the template. (b) The corresponding raw p-values with 10000 permutations and FDR corrected p-values. (c) The directionality of shape differences on the template with respect to the median of the reference group.

Global analysis. To test for global group separability, we followed the strategy described in Section 5.1.2 to perform a permutation test (10000 permutations). When using the NC-Train as the training set to compute band-depth for both control and disease test sets, the resulting p-value is less than 1e-4, indicating the normal controls and the disease subjects are significantly different. On the other hand, as shown in Fig. 5.5, when pooling all shapes together to compute their band depths, no significant difference is detected. Since we know the control and disease groups of the synthetic data are significantly different according to the ground truth, this experimental result indicates that the training population is essential for the discrimination of subject populations.

Local analysis. For the local analysis, we first need to choose the template and the reference population for the permutation test. Two possible choices are: (i) taking the median of the NC-Test as the template and the disease test group as the reference population; (ii) using the median of the disease test group as the template and the NC-Test group as the reference population. We demonstrate our method using both strategies with experimental results



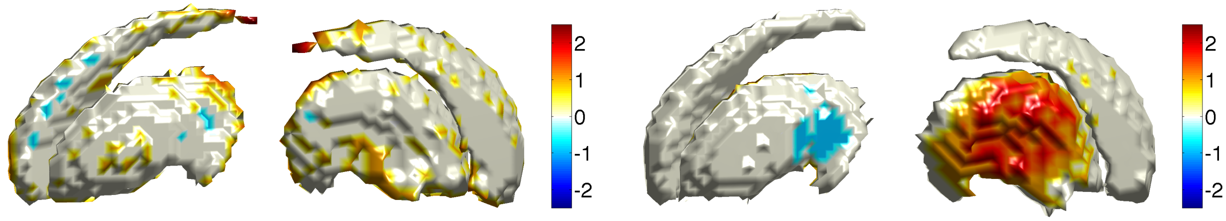
(a) α-values (b) Raw and FDR-corrected p-values (c) Directionality

Figure 5.11: Local shape analysis on synthetic striatum shown from two views, the medial (top) and the lateral (bottom) views, with the median of the disease test group as the template and the NC-Test group as the reference population. (a) The α-values on the template. (b) The corresponding raw p-values with 10000 permutations and FDR corrected p-values. (c) The directionality of shape differences on the template with respect to the median of the reference group.

shown in Fig. 5.10 and Fig. 5.11.

We first use the NC-Train group to compute the band-depth for all test shapes, and then we estimate α-values of the template with respect to the reference population which is ordered using the computed band-depth. The α-values on the NC-Test median are shown in Fig. 5.10(a) and those of the disease median are shown in Fig. 5.11(a). Both results demonstrate that the introduced deformed (or different) regions on both medial and lateral sides of the template are detected as expected.

Fig. 5.10(b) and Fig. 5.11(b) show the raw p-values and the false-discovery-rate (FDR) corrected p-values, with 10000 permutations. Based on the ground truth shown in Fig. 5.9, the "pulled-out" region on the medial side of the putamen is significantly deformed, while the deflated region on the lateral side is slightly deformed, and some significantly deformed region may exist. We detect, regardless of which median shape is chosen as the template, the significantly deformed region. The main difference is that Fig. 5.10(b) shows more significantly different regions on the lateral side of the putamen than Fig. 5.11(b). Both



(a) The median of the NC-Test group (b) The median of the disease group

Figure 5.12: The directionality of shape differences on the median test shapes with respect to the median shape of the NC-Train group shown from two views, lateral (left) and medial (right) sides.

strategies provide reasonable results compared to the ground truth (Fig. 5.9).

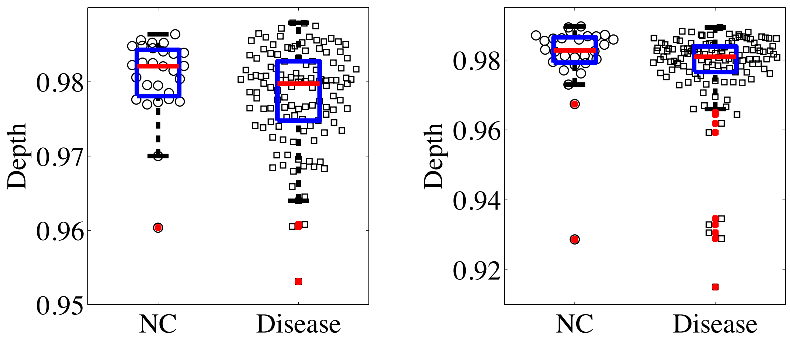
Directionality of shape differences. To test our strategy of measuring the directionality of shape differences, we first applied it on the median shapes of NC-Test and disease groups with respect to the median of NC-Train, as shown in Fig. 5.12. As expected, the median shapes from NC-Train and NC-Test have relatively small differences, c.f. Fig. 5.12(a). But, in Fig. 5.12(b) (the disease median with respect to the NC-Train median), we can see positive values on the medial side of the putamen, indicating that the disease median has an inflated region compared to the NC-Train median, and negative values on the lateral side of putamen, indicating that the disease median has a deflated region there. This is consistent with the introduced deformation in the synthetic data.

Since the directionality test among the control and the disease median shapes correctly reveals the directions of their shape differences, according to the introduced deformations, we further applied the directionality computation on the template with respect to the median shape of the reference population to augment the local analysis results based on depth-ordering. Fig. 5.10(c) shows the NC-Test median with respect to the disease median. The negative values on the medial side of the putamen indicate that this region of the NC-Test median (template) is deflated compared to the corresponding part of the disease median (from the reference population). And the positive values on the lateral side of the putamen indicate that this region of the NC-Test median is inflated compared to that part of the disease median. This is also consistent with the ground truth. By also considering the p-values in Fig. 5.10(b), we can determine the directionality for those significantly different regions. Consistent results are shown in Fig. 5.11(c) for the disease median with respect to the NC-Test median.

Real data

Magnetic Resonance Images (MRIs) of the brains of 123 subjects (including 102 males) diagnosed with first-episode schizophrenia and of 56 normal control subjects (including 37 males) were acquired on a 1.5-T scanner. Multi-site SPGR T1-weighted images (voxel dimensions $0.9375 \times 0.9375 \times 1.5$ mm) were obtained. These MRI scans were rigidly aligned to a standard coordinate space, from which hippocampus structures were segmented. Each binary segmentation was fit with a mesh model. This dataset was used in a previous shape analysis study [McClure et al., 2013]. All the hippocampus shapes were pre-aligned using a rigid transformation. To get the binary representation of the hippocampus shapes, we used voxelizations with equal spacing in each direction. To measure the sensitivity with respect to spacing we tested using spacings of 0.3 and 0.4 mm. Results were similar. Hence, only results for a spacing of 0.4 mm are presented in what follows.

Global analysis. Similar to the synthetic data experiment, we divided the 56 normal control subjects into two groups, NC-Train and NC-Test, and we considered all 123 subjects with first-episode schizophrenia as the disease group. Fig. 5.13 shows the global differences between NC-Test and the disease groups, for both left and right hippocampi. We also used a permutation test with 10000 permutations to determine whether these groups' mean depth values differed significantly. This resulted in p-values of 0.03 for the left hippocampus and

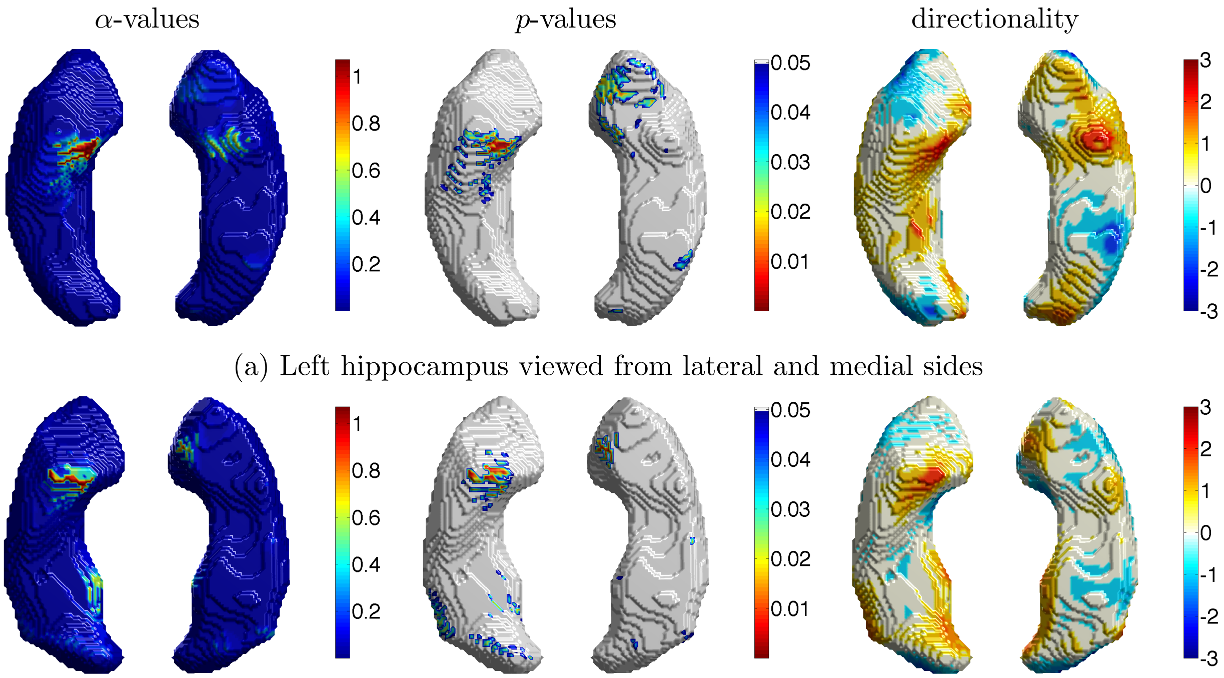


(a) Left hippocampus (b) Right hippocampus

Figure 5.13: Global analysis on both left and right hippocampi. The disease group indicates subjects with first-episode schizophrenia.

0.15 for the right hippocampus. This indicates based on the global depth-based analysis, that the difference in the left hippocampi from disease and normal control populations is marginally significant at a significance level of 0.05.

Local analysis. For the local shape analysis, we used the NC-Train group as the training set to compute the depth for test subjects, and we took the median of the disease group as the template and the NC-Test group as the reference population. Fig. 5.14 shows the local analysis results on both left and right hippocampi. We can see that our method captures some abnormal regions. However, based on the local p-values only relatively small regions of the median disease shape are significantly different from the normal controls. The p-values also show that the response is stronger on the left hippocampus, which is consistent with a previous study by [McClure et al., 2013]. The shown p-values for this experiment were not corrected using FDR, because no significant regions remained after FDR-correction. Note that in our case we work directly on the voxel level of shapes thereby generating many local comparisons. Considering the analysis at a coarser spatial scale, e.g., averaged responses over regions as adopted in [McClure et al., 2013], could potentially reveal shape differences which persist under FDR correction.



(a) Left hippocampus viewed from lateral and medial sides

(b) Right hippocampus viewed from medial and lateral sides

Figure 5.14: Local analysis on the disease median of both left and right hippocampi with respect to the normal control test group.

Directionality of shape differences. The right column in Fig. 5.14 shows the directionality of shape differences on the template (the disease median) with respect to the median of the reference group, NC-Test. We observe for both the left and right hippocampi of the disease median that few detected regions are locally deflated with respect to the normal median, but instead most of the other detected regions are found to be locally inflated compared to the normal median. Since the detected regions are shown before FDR correction, and no regions are significant after FDR correction, we cannot suggest if the disease shape is deflated or inflated compared to the normal one.

Comparison with other methods

In [Gao et al., 2014], three standard methods, SPHARM-PDM, Shapeworks, and Tensor Based Morphometry (TBM), were tested on the same synthetic dataset in our experiments.

To quantitatively measure the significantly deformed region, they extracted the "region of deformation" (ROD) from the ground truth data and computed the ratio between the area of deformation and the area of the entire shape. The average of these ratios is referred to as the ground truth. For the striatum dataset used in our work the average ratio is 0.33. When measuring the detected significant region in the local analysis results, the ROD is defined as the region whose FDR-corrected p-value is ≤ 0.05. The area ratio reported in [Gao et al., 2014] is 0.61 for SPHARM-PDM, 0 for ShapeWorks, and 0.17 for TBM.

The ROD cannot measure how well a shape analysis method localizes the differences, but it reveals whether the detected significant region has a reasonable size compared to the ground truth. Using the combination of ROD with the visual results allows one to assess both the location and extent of the detected deformation with respect to the ground truth. We follow this measurement and compute the area ratio of our FDR-corrected results on the synthetic data, resulting in 0.2 when using the median of NC-Test as the template (see Fig. 5.10) and 0.15 when using the median of the disease test group as the template (see Fig. 5.11). According to the ground truth of the area ratio for the synthetic data, our method has better performance (closer to 0.33) than SPHARM-PDM and Shapeworks, and it provides comparable, even slightly better, quantitative measures than TBM.

Compared to the qualitative results of the three methods evaluated in [Gao et al., 2014], our method provides more accurate locations of the detected deformations, according to the introduced deformations. Furthermore, in contrast to these methods, our results reveal that most of the significantly different regions are located on the medial side of the putamen, instead of the lateral side, which is also consistent with the synthetic data.

In addition, our real data experiment is consistent with previous results on shape differences in the hippocampus for first-episode schizophrenia [McClure et al., 2013], specifically with strong differences for the left hippocampus.

5.2 Hypothesis Testing for Longitudinal Data

The shape analysis method developed for cross-sectional data can effectively detect both global and local shape differences, but it is time-independent. Recently, longitudinal data designs frequently arise in medical research that involve repeated measurements during follow-up studies. Analysis of such longitudinal data often involves constructing statistical models to summarize growth, aging and disease progression over time. For example, longitudinal studies in new-borns and young children use imaging at multiple follow-up visits to understand the process of early brain development [Gilmore et al., 2012]. Similarly, recent collective efforts have enabled longitudinal data collection to facilitate the study of neurodegeneration due to aging and age-related neurological disorders, such as Alzheimer's disease [Marcus et al., 2010]. Conventional cross-sectional models of regression that do not take into account the temporal dependencies of measurements are inappropriate for modeling such longitudinal data designs.

Recent methods for analyzing longitudinal, manifold-valued data have enabled modeling and even detection of changes over time [Fletcher, 2013, Niethammer et al., 2011]. These methods allow for the estimation of trajectories, *i.e.*, smooth paths estimated from the longitudinal data of subjects. Based on them, Riemannian approaches for computing averages of trajectories, e.g., mean trajectories, have been proposed [Muralidharan and Fletcher, 2012, Singh et al., 2013a]. The registration and comparison of trajectories has been studied in [Durrleman et al., 2013, Su et al., 2014a, Su et al., 2014b]. In general, statistical methods for longitudinal manifold-valued data focus on first-order statistics, such as computing the mean, which only captures limited information of the data distribution. Capturing higher-order statistics on the trajectories themselves would be useful for a more comprehensive description of the underlying distributions and for designing test-statistics that go beyond a simple comparison of means; an example would be testing differences in variances.

Motivated by this, we develop an approach that leverages second-order statistics of shape trajectories for group testing. In particular, we propose a generalization of principal component analysis (PCA) and principal geodesic analysis (PGA) [Fletcher et al., 2004] to the tangent bundle [Lee, 2012] of a shape space. Similar to PCA/PGA, the first principal direction characterizes the dominant variability in a *population of trajectories*, and each point along this principal direction is a trajectory. This differs from previous studies which have focused on computing averages on the tangent bundle. Incorporating second-order statistics additionally allows identifying differences between groups of trajectories in situations where the average longitudinal trend over time is similar (or equal) between two groups. We refer to this approach as *principal geodesic analysis on the tangent bundle.* Having estimated both variance and principal directions of shape trajectories, we then introduce a generalization of the Bhattacharyya distance to manifold-valued data. This enables the assessment of statistical differences between groups of trajectories.

5.2.1 Distribution of Trajectories in Euclidean Space

We first illustrate the concept of analyzing populations of trajectories in Euclidean space, which is a trivial case of a Riemannian manifold.

Consider the case of two groups of subjects such that each subject is measured at multiple points in time, see Fig. 5.15(b). If we ignore the within-subject correlations and model the data with a cross-sectional design, illustrated in Fig. 5.15(a), the two groups cannot be

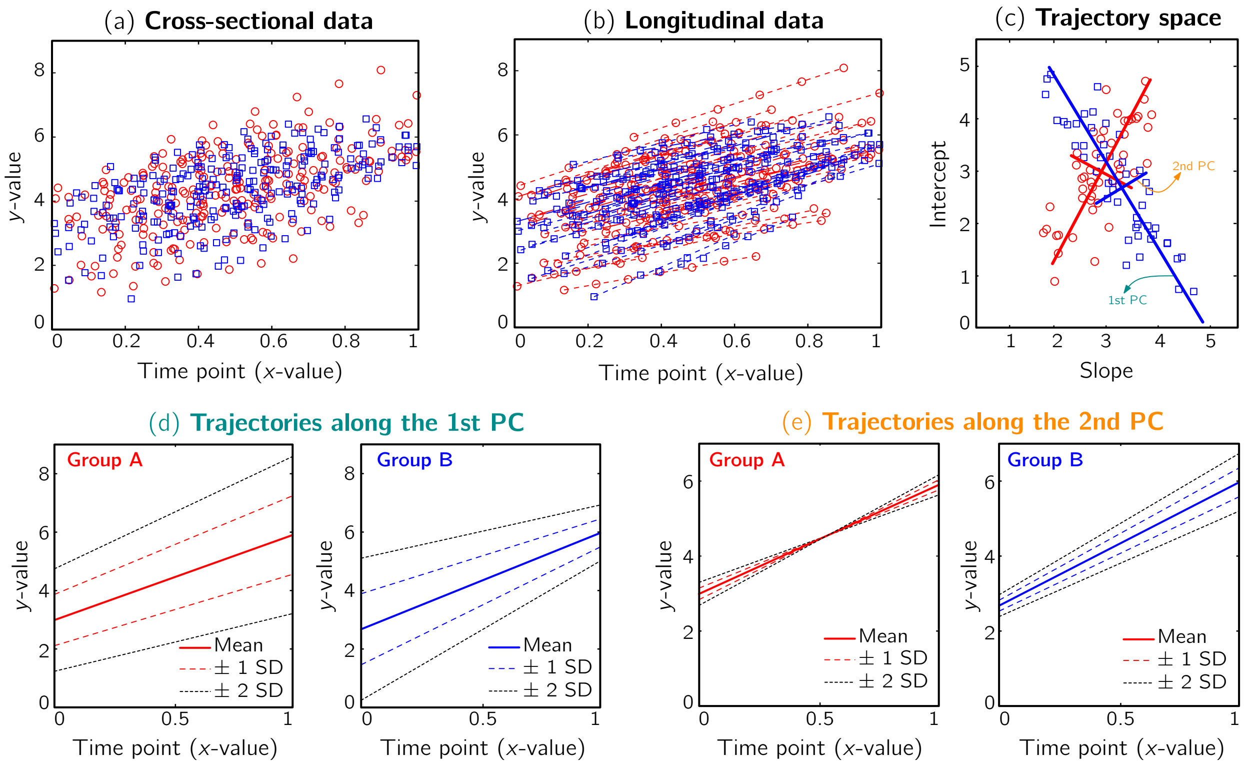


Figure 5.15: A toy example in Euclidean space. *Top*: (a) Cross-sectional data of two groups, illustrated as red circles and blue squares; (b) the same data *with* longitudinal information where points on the same line are observations from one subject; (c) the trajectory space, represented by a slope and an intercept. Every point in this space corresponds to a straight line in (b). *Bottom*: (d) Trajectories generated by points along the 1st principal component (PC) of standard PCA in trajectory space with {0, ±1, ±2} standard deviations (SD); (e) trajectories generated along the 2nd PC (best-viewed in color).

separated using statistical tests that rely on a comparison of means only (*cf*. Table 5.1). Hence, to leverage longitudinal information, we first estimate linear regression models on each subject to summarize its trend. The regression line, a smooth trajectory approximating a subject's data points, is parameterized by the tuple of *slope* and *intercept*, which can be represented as a point in the space of trajectories. As shown in Fig. 5.15(c), representing the data in this trajectory space separates the populations (at least visually) in this example. In fact, Table 5.1 indicates that including longitudinal information allows us to identify differences between the two groups statistically.

To further analyze the group differences, we explore the distribution of trajectories within the (slope, intercept) space, *i.e.*, the trajectory space. Under a Gaussian assumption, principal component analysis (PCA) is a standard tool to estimate the variance and principal directions of a sample. By applying PCA to (slope, intercept) data, we obtain a representation of the population of trajectories, namely their variances and their principal components. For example, the solid lines with different colors in Fig. 5.15(c) show the principal components of the two groups, respectively. By moving along these two principal components, we generate new points in the trajectory space such that each point represents a straight line in the original space of the data points. Figures 5.15(d) and (e) visualize the trajectories along the principal components for different standard deviations. The five trajectories in Figure 5.15(d), for instance, show the five points along the first principal component in the trajectory space for each group. This Euclidean case illustrates that the proposed approach is a potentially useful tool in the analysis of longitudinal time-varying data.

Bhattacharyya distance. Visualization of trajectories along principal directions can qualitatively demonstrate differences between groups. However, to quantitatively assess the differences, we need a suitable distance measure that serves as a test-statistic. One possible candidate for this is the Bhattacharyya distance [Bhattacharyya, 1946], which measures the similarity of two probability distributions. Given two multivariate Gaussians, with means (μ_1, μ_2) and covariance matrices (Σ_1, Σ_2), the Bhattacharyya distance D_B has the closed-form expression

$$D_B((\mu_1, \Sigma_1), (\mu_2, \Sigma_2)) = \frac{1}{8}(\mu_1 - \mu_2)\Sigma^{-1}(\mu_1 - \mu_2)^\top + \frac{1}{2}\ln\left(\frac{|\Sigma|}{\sqrt{|\Sigma_1| \cdot |\Sigma_2|}}\right), \tag{5.2}$$

where $\Sigma = (\Sigma_1 + \Sigma_2)/2$, and $|\cdot|$ denotes the matrix determinant. The first term in Eq. (5.2)

	Cross-sectional data			**Longitudinal data**		
	$\bar{D}_E$	$\bar{D}_M$	D_B	$\bar{D}_E$	$\bar{D}_M$	D_B
Distance	0.0003	0.0047	0.0077	0.2438	0.3332	0.6722
p-value	0.9232	0.7487	0.1249	0.0347	0.0186	**1e-4**

Table 5.1: Distances and estimated p-values (10000 random permutations) on toy data using (1) the mean difference in Euclidean space ($\bar{D}_E$), (2) the Mahalanobis distance ($\bar{D}_M$), and (3) the Bhattacharyya distance (D_B) as a test-statistic.

measures the separability of the distributions w.r.t. their means. It is related to the squared Mahalanobis distance [Mahalanobis, 1936], which can be considered a special case of Eq. (5.2) when the difference between the covariances (as measured by the second term in the summation) is not considered. This additional term makes D_B more suitable, compared to the Mahalanobis distance, in cases where the distributions differ in variances. In particular, the Mahalanobis distance is zero when two distributions have equal means. However, as D_B only satisfies three conditions of a distance metric (non-negativity, identity of indiscernibles, and symmetry), but lacks the triangle inequality, it is only a semi-metric.

In fact, Eq. (5.2) allows us to compute a distance between the two distributions (assuming Gaussianity) in Fig. 5.15(c), and thereby to define a test-statistic to test for group differences in a permutation testing setup. The null-hypothesis H_0 of the permutation test is that the two distributions (say P, Q) to be tested are the same, *i.e.*, $H_0 : P = Q$. We estimate the empirical distribution of the test-statistic under H_0 by repeatedly permuting the group labels of the points in Fig. 5.15(c) and re-computing D_B between the two groups that result from the permuted labels. The p-value under H_0 then is the proportion of the area under the empirical distribution of samples for which the distance is less than the one estimated for the original (unpermuted) label assignments. In Table 5.1, D_B, tested on the longitudinal data, exhibits the best performance in separating the groups with an estimated p-value of

<1e-4 under 10000 permutations.

5.2.2 Distribution of Trajectories on Manifolds

To explore the distribution of trajectories for manifold-valued data, *e.g.*, images or shapes, we need to generalize the statistical test of the previous section from Euclidean space to manifolds. Specifically, let $\{P_{i,j,k}\}$ be a population of longitudinal data on the same manifold, where i is the group identifier, j is the subject identifier, and k identifies the time point. Further assume we have N groups: group i has S_i subjects ($i = 1, \ldots, N$), and each subject has multiple time points, $\{t_{i,j,k}\}, k = 1, \ldots, T_{i,j}$. Our objective is to characterize the distribution of trajectories for each group, $\{D_i\}$, *i.e.*, to estimate its variance and principal directions, and to assess whether two groups are significantly different.

Individual trajectories for longitudinal data. To perform statistical tests on subjects with associated longitudinal data, our first step is to summarize the variations within a subject as a smooth trajectory. The parametric geodesic regression approaches for data in Kendall's shape space [Fletcher, 2013] or images [Niethammer et al., 2011, Hong et al., 2012a], which generalize linear regression in Euclidean space, provide a compact representation of the continuous trajectory for each subject. The trajectory of subject j from group i is parametrized by the initial point $\hat{p}_{i,j}$ and the initial velocity $\hat{u}_{i,j}$. This trajectory minimizes the sum-of-squared geodesic distances between the observations and their corresponding points on the trajectory, *i.e.*,

$$(\hat{p}_{i,j}, \hat{u}_{i,j}) = \operatorname{argmin}_{(p_{i,j}, u_{i,j})} \sum_{k=1}^{T_{i,j}} d_g^2(\operatorname{Exp}(p_{i,j}, t_{i,j,k} \cdot u_{i,j}), P_{i,j,k}) \ , \tag{5.3}$$

where $d_g(\cdot, \cdot)$ is the geodesic distance and $\operatorname{Exp}(\cdot, \cdot)$ denotes the exponential map on some manifold $\mathcal{M}$ [Fletcher, 2013]. This compact representation, $(\hat{p}_{i,j}, \hat{u}_{i,j})$, is a point in the tangent bundle $\mathcal{TM}$ of $\mathcal{M}$. $\mathcal{TM}$ is also a smooth manifold, which can be equipped with a Riemannian metric, such as the *Sasaki metric* [Sasaki, 1958]. Since each subject's longitudinal data is represented as a point on $\mathcal{TM}$, we work in this space, instead of the space of the data points, to perform appropriate hypothesis testing.

Principal geodesic analysis (PGA) for trajectories. We adopt principal geodesic analysis to estimate the variance and the principal directions of trajectories on the tangent bundle for each group. We follow the definitions of the exponential- and the log-map on $\mathcal{TM}$ in [Muralidharan and Fletcher, 2012] and use the Sasaki metric. Specifically, given two points $(p_1, u_1), (p_2, u_2) \in \mathcal{TM}$, the log-map outputs the tangent vector such that $(v, w) = \mathrm{Log}_{(p_1,u_1)}(p_2, u_2)$. The exponential map enables us to shoot forward with a given base point and a tangent vector, *i.e.*, $(p_2, u_2) = \mathrm{Exp}_{\mathcal{TM}}((p_1, u_1), (v, w))$. Furthermore, using the log-map, the geodesic distance on $\mathcal{TM}$ can be computed as $d_{\mathcal{TM}}((p_1, u_1), (p_2, u_2)) = \| \mathrm{Log}_{(p_1,u_1)}(p_2, u_2)\|$.

Before computing the variance and the principal directions, we first need to estimate the mean of the trajectories for each group. This is done by minimizing the sum-of-squared geodesic distances, for each group, on $\mathcal{TM}$ as

$$\forall i : (\bar{p}_i, \bar{u}_i) = \mathrm{argmin}_{(p_i,u_i)} \sum_{j=1}^{S_i} d^2_{\mathcal{TM}}((p_i, u_i), (\hat{p}_{i,j}, \hat{u}_{i,j})) \ . \tag{5.4}$$

Then, following the PGA algorithm of [Fletcher et al., 2004], we compute the variance and principal directions w.r.t. the estimated mean of the trajectories. Specifically, we first compute the tangent vector from the mean of group i to the trajectory of its subject j, $(v_{i,j}, w_{i,j}) = \mathrm{Log}_{(\bar{p}_i,\bar{u}_i)}(\hat{p}_{i,j}, \hat{u}_{i,j})$ and then calculate its corresponding covariance matrix

$\Sigma_i = \frac{1}{S_i - 1}\sum_{j=1}^{S_i}(v_{i,j}, w_{i,j})(v_{i,j}, w_{i,j})^\top$. The principal decomposition of Σ_i results in the eigenvalues $\lambda_{i,q} \in \mathbb{R}_0^+$ and eigenvectors $(v_{i,q}, w_{i,q}) \in \mathcal{T}_{(\bar{p}_i, \bar{u}_i)}\mathcal{M}$ with $q = 1, \ldots, Q_i$ for group i. As a result, we can identify the distribution of trajectories for each group by $D_i = \{(\bar{p}_i, \bar{u}_i), \Sigma_i\}$ with $i = 1, \ldots, N$. By moving along a principal direction, we can generate points on $\mathcal{TM}$, which correspond to trajectories on the manifold of the data points.

Generalized Bhattacharyya distance. Since we can characterize the distribution of trajectories on $\mathcal{TM}$ for each group, to measure the distance between them, we generalize the Bhattacharyya distance from Euclidean space to $\mathcal{TM}$. Again, the distribution D_i on $\mathcal{TM}$, is identified by a mean $\mu_i = (\bar{p}_i, \bar{u}_i) \in \mathcal{TM}$ and a covariance matrix Σ_i associated to the mean.

Generalizing the first term of the Bhattacharyya distance in Eq. (5.2), *i.e.*, the pooling of covariance matrices $\Sigma = (\Sigma_1 + \Sigma_2)/2$, is not as straightforward on $\mathcal{TM}$ as it is in Euclidean space because the covariance matrices Σ_1 and Σ_2 of the two groups reside in tangent spaces at different points on $\mathcal{TM}$. Hence, we follow the strategy in [Muralidharan and Fletcher, 2012] and replace the first term with the average of two squared-Mahalanobis distances, *i.e.*, $(\mathrm{Log}_{\mu_1}\mu_2{}^\top\Sigma_1^{-1}\mathrm{Log}_{\mu_1}\mu_2 + \mathrm{Log}_{\mu_2}\mu_1{}^\top\Sigma_2^{-1}\mathrm{Log}_{\mu_2}\mu_1)/2$. Furthermore, because most manifold-valued data in medical applications is high dimensional and low sample size, the resulting covariance matrix is usually semi-positive-definite (SPD) with zero eigenvalues. This means that in many applications Σ_1 and Σ_2 are not invertible[4]. To address this issue, we approximate the covariance matrix via eigen-decomposition by dropping the eigenval-

[4] A better estimate of the covariance matrix may be obtained, *e.g.*, by using [Ledoit and Wolf, 2004] or [Bickel and Levina, 2008].

ues that are smaller than a cutoff value, ϵ[5]. In this way, the covariance matrix can be decomposed approximately as $\Sigma_i \approx U_{i,Q_i}\Lambda_{i,Q_i}U_{i,Q_i}^\top$, where $\lambda_{i,q} < \epsilon$ if $q > Q_i$, resulting in $\Sigma_i^{-1} \approx U_{i,Q_i}\Lambda_{i,Q_i}^{-1}U_{i,Q_i}^\top$ [Oliver, 1998].

To generalize the second term of the Bhattacharyya distance, which involves the computation of the determinant of a covariance matrix, we use the pseudo-determinant, *i.e.*, the product of all non-zero eigenvalues of a square matrix. For consistency, the same number of eigenvalues as for the first term is used, *i.e.*, $|\Sigma_i| = \prod_{q=1}^{Q_i} \lambda_{i,q}$. Since it is non-trivial to compute the pooled covariance matrix Σ, we replace its determinant in Eq. (5.2) with the averaged determinants of Σ_1 and Σ_2. While this changes the original definition of the Bhattacharyya distance, most of its properties are kept, see the proof at the end of this subsection. Also, it can be shown that the value of the second term increases as the difference in the determinants gets larger. Hence, the generalized second term can serve as a distance measure of generalized variances of covariance matrices on $\mathcal{TM}$. In summary, we define the *generalized Bhattacharyya distance* between two Gaussians D_1, D_2 on $\mathcal{TM}$ as

$$D_B^{\mathcal{TM}}(D_1, D_2) = \frac{1}{16}(D_M^{\mathcal{TM}}(\mu_1, D_2) + D_M^{\mathcal{TM}}(\mu_2, D_1)) + \frac{1}{2}\ln\left(\frac{(|\Sigma_1| + |\Sigma_2|)}{2\sqrt{|\Sigma_1| \cdot |\Sigma_2|}}\right) . \tag{5.5}$$

Here, $D_M^{\mathcal{TM}}$ is a generalized version of the squared Mahalanobis distance, i.e., $D_M^{\mathcal{TM}}(\mu_i, D_j) = \langle \mathrm{Log}_{\mu_j} \mu_i, U_{j,Q_j}\rangle^\top \Lambda_{j,Q_j}^{-1} \langle \mathrm{Log}_{\mu_j} \mu_i, U_{j,Q_j}\rangle$, and $\langle \cdot, \cdot \rangle$ is the inner product on the tangent bundle. $D_B^{\mathcal{TM}}$ is a *pseudo-semimetric*, *i.e.*, it satisfies (1) non-negativity, (2) symmetry, and (3) $D_B^{\mathcal{TM}}(D_i, D_i) = 0$ for all D_i (required for the identity of indiscernibles). As shown in the proof at the end of this subsection, although Eq. (5.5) does not satisfy the positivity

[5] The threshold ϵ varies with the application. In our experiments, we set it to 1e-6. Usually, the eigenvalues larger than ϵ cover almost 99% of the variances.

property, *i.e.*, for all $D_1 \neq D_2$, $D_B^{\mathcal{TM}}(D_1, D_2) > 0$, only the distance between two distributions with equal mean *and* generalized variance is zero. This may bring a problem when two distributions have equal mean and generalized variance (i.e., equal determinate of the covariance matrix), but their covariance matrices actually differ. In such a case, our generalized Bhattacharyya distance cannot differentiate them. Consequently, we can distinguish two distributions of trajectories that have different means and/or different determinants of the covariance matrices.

We use Eq. (5.5) as our test-statistic in the same permutation testing setup as described in Section 5.2.1. The null-hypothesis H_0 is that the samples of trajectories from the two groups were drawn from the same underlying distribution. The distribution of test-statistics under H_0 is estimated by randomly permuting the group label assignments. We then count the number of times that the distance is larger than the one computed without permutation to obtain a p-value estimate. Compared to the Hotelling T^2 statistic used in [Muralidharan and Fletcher, 2012], which tests for the difference in sample means (based on the squared Mahalanobis distance), our permutation test is based on Eq. (5.5), which is more appropriate in situations where two distributions have similar means but different variances.

Properties of the generalized Bhattacharyya distance

Non-negativity. In the first term of Eq. (5.5), $D_M^{\mathcal{TM}}$ is the generalized squared-Mahalanobis distance which is non-negative; consequently, the first term in Eq. (5.5) is non-negative. Furthermore, the determinant of a covariance matrix in the second term is also non-negative, since it is the product of all non-negative eigenvalues. Besides, it is easy to demonstrate that $(|\Sigma_1| + |\Sigma_2|)/(2\sqrt{|\Sigma_1||\Sigma_2|}) \geq 1$, indicating the second term is non-negative. Hence,

$D_B^{\mathcal{TM}}(D_1, D_2) \geq 0$.

Identity of indiscernibles. If $D_1 = D_2$, *i.e.*, $\mu_1 = \mu_2$ and $\Sigma_1 = \Sigma_2$, we see that (1) $\mathrm{Log}_{\mu_1} \mu_2$ and $\mathrm{Log}_{\mu_2} \mu_1$ are zero tangent vectors, and (2) $|\Sigma_1| = |\Sigma_2|$. Hence, $D_M^{\mathcal{TM}}(\mu_1, D_2) = D_M^{\mathcal{TM}}(\mu_2, D_1) = 0$, *i.e.*, the first term of Eq. (5.5) is 0; also, the second term is 0. Now, if $D_1 = D_2$ then $D_B^{\mathcal{TM}}(D_1, D_2) = 0$. On the other hand, assuming $D_B^{\mathcal{TM}}(D_1, D_2) = 0$, we can only obtain $\mu_1 = \mu_2$ and $|\Sigma_1| = |\Sigma_2|$, because of the non-negativity properties of the two terms in Eq. (5.5). But, we *cannot* draw the conclusion that the two covariance matrices are equal. Therefore, if $D_1 = D_2$ then $D_B^{\mathcal{TM}}(D_1, D_2) = 0$, but it is possible that $D_B^{\mathcal{TM}}(D_1, D_2) = 0$ for some $D_1 \neq D_2$, if $\mu_1 = \mu_2$ and $|\Sigma_1| = |\Sigma_2|$.

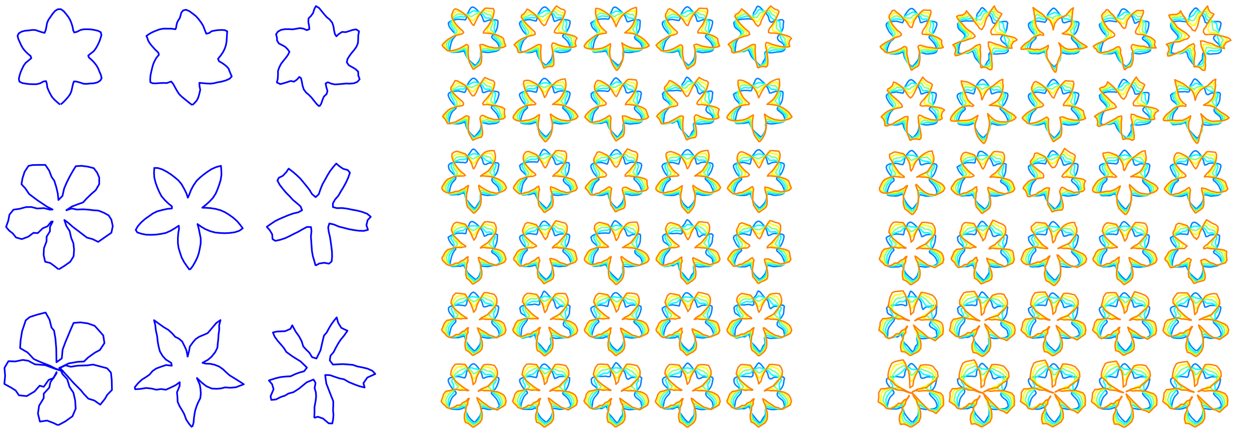
Symmetry. Because both terms of Eq. (5.5) are symmetric, the sum of them is also symmetric, *i.e.*, $D_B^{\mathcal{TM}}(D_1, D_2) = D_B^{\mathcal{TM}}(D_2, D_1)$.

Triangle inequality. Since, Eq. (5.2) in $\mathbb{R}^n$ does not satisfy the triangle inequality, our generalized variant will not satisfy it either.

5.2.3 Experimental Results

We demonstrate our method on (1) a toy example in Euclidean space, (2) a 2D example with synthetic shapes, and (3) real corpus callosum shapes. All shapes are represented in (2D) Kendall's shape space.

Toy example in Euclidean space. Fig. 5.15 shows the generated toy data and the qualitative comparison between two groups using PCA in the trajectory space. Both groups have 50 subjects each, measured at 3 to 7 time points. Table 5.1 reports the quantitative comparison, *i.e.*, permutation testing with 10000 permutations and three different distances: the Euclidean distance $\bar{D}_E$ (*i.e.*, the squared mean differences), the Mahalanobis distance $\bar{D}_M$ (*i.e.*, the squared mean difference based on the pooled covariance matrix), and the Bhat-



(a) Basic shapes (b) Group A (c) Group B

Figure 5.16: Synthetic shapes: (a) Basic shapes used to generate the population on the right; (b) and (c) show the two groups of trajectories (best-viewed in color).

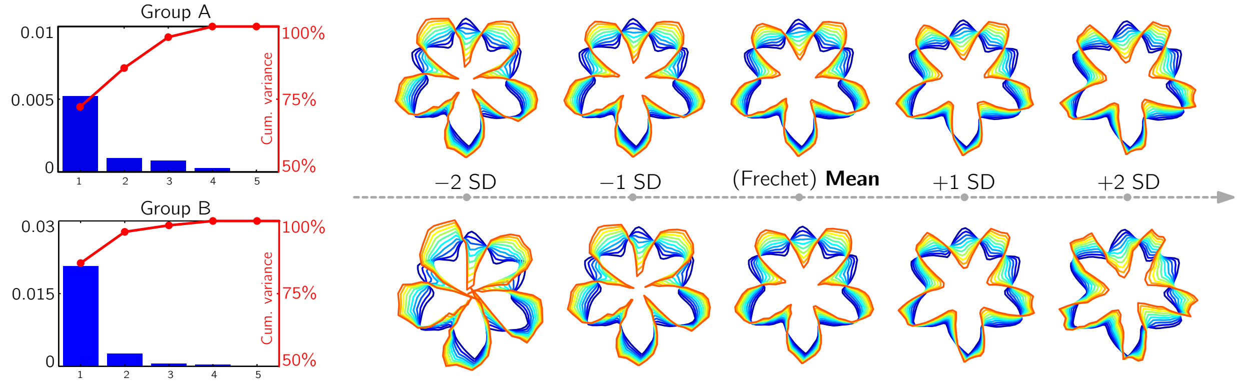


Figure 5.17: Visualization of the variances (left) and principal directions (right) of trajectory distributions for the synthetic data (best-viewed in color).

tacharyya distance D_B. The results of the cross-sectional *vs*. longitudinal tests indicate that leveraging the longitudinal information greatly improves our ability to identify differences, as indicated by low p-values. Besides, among the three evaluated distance measures, the Bhattacharyya distance most clearly highlights this difference with a p-value of <1e-4 (given the number of permutations).

Synthetic shapes in Kendall's shape space. To verify the advantage of the generalized Bhattacharyya distance over the generalized Mahalanobis distance, we generate two groups of 2D shapes with similar mean trajectories but different variances, see Figs. 5.16b and 5.16c.

	$(\hat{p}, \hat{u})$		$(\hat{p}, 0)$		$(0, \hat{u})$	
	$\bar{D}_M^{\mathcal{TM}}$	$D_B^{\mathcal{TM}}$	$\bar{D}_M^{\mathcal{TM}}$	$D_B^{\mathcal{TM}}$	$\bar{D}_M^{\mathcal{TM}}$	$D_B^{\mathcal{TM}}$
Distance on $\mathcal{TM}$	0.7212	2.2833	0.0232	0.0152	0.7439	2.3057
p-value	0.1817	**0.0234**	0.8486	0.6801	0.1650	**0.0297**

Table 5.2: Distances and estimated p-values (10000 random permutations) on synthetic shapes using the averaged Mahalanobis distance ($\bar{D}_M^{\mathcal{TM}}$) and the generalized Bhattacharyya distance ($D_B^{\mathcal{TM}}$). The last two columns report the test results when dropping one of the initial conditions.

Hence, the distributions are different by design. In particular, we use the three shapes in the first row of Fig. 5.16a to uniformly sample 60 shapes within the triangle region in Kendall's shape space, spanned by the three shapes[6]. We call them the *base shapes*. In the same way, the shapes in the second and third row are used to sample 30 shapes each; we refer to these shapes as the *target shapes*. In summary, we have 60 base shapes from the same distribution and two groups of target shapes from two different distributions. By splitting the 60 base shapes into two subsets of equal size and connecting each base shape with one target shape (via a geodesic), we obtain 30 trajectories per group. Assuming every base shape is at time 0 and every target shape is at time 1, we sample 5 shapes along each trajectory to represent one subject. To make sure these two groups of trajectories have similar means, the shapes in the third row of Fig. 5.16a are not picked randomly but generated using the shapes in the second row. This is done by computing the mean of the shapes in the second row, then shooting a geodesic from the mean to each of the three shapes and continuing to move beyond time 1 (for two times) to generate the shapes in the third row. Essentially, this has the effect that the means of the trajectories of both groups are similar, but the variances differ.

[6] We use two geodesics to connect three given shapes and uniformly sample points on these two geodesics. Then, by connecting opposing points, we obtain new geodesics which are located within the triangle region to sample a population of shapes.

Fig. 5.17 shows the results of PGA in trajectory space for the synthetic shapes. The largest eigenvalue of the trajectories in *Group A* is 0.005 at 72% cumulative variance, compared to the largest eigenvalue of 0.02 at 85% cumulative variance in *Group B*. Also, as expected, the trajectories visualized by 10 shapes in Fig. 5.17 show that the shapes of *Group B* change faster than in *Group A*. Table 5.2 reports the quantitative measures of the difference between the two groups. Since, by design, the mean trajectories are similar, it is difficult to identify significant deviations from the null-hypothesis H_0 using the generalized Mahalanobis distance; this is indicated by the relatively high p-values of $\bar{D}_M^{\mathcal{T}\mathcal{M}}$ in Table 5.2[7]. As desired, $D_B^{\mathcal{T}\mathcal{M}}$ is sensitive w.r.t. differences in variance, indicated by the relatively low p-value. This would allow to reject H_0 at the customary significance level of 0.05.

Furthermore, since all base shapes are uniformly sampled from within the shape triangle spanned by the first row of Fig. 5.16a, *i.e.*, the initial points of the two groups have similar means, it is *not* possible to only use the initial points to separate the two groups; this is confirmed by the high p-values for both distance measures in the $(\hat{p}, 0)$ column of Table 5.2. In fact, even when specifically testing for differences in the initial velocity, the generalized Bhattacharyya distance exhibits better behavior than the generalized Mahalanobis distance in terms of lower p-values (*cf*. column $(0, \hat{u})$ of Table 5.2).

Corpora callosa in Kendall's shape space. The longitudinal corpus callosum dataset used in [Muralidharan and Fletcher, 2012], contains 23 subjects, 11 of which are males with dementia, and the rest are normal controls. Every subject has been measured at three time points within the age range of 60 to 92 years old, and each corpus callosum shape is

[7] The average of two generalized squared-Mahalanobis distances is related to the first term of the generalized Bhattacharyya distance in Eq. (5.5).

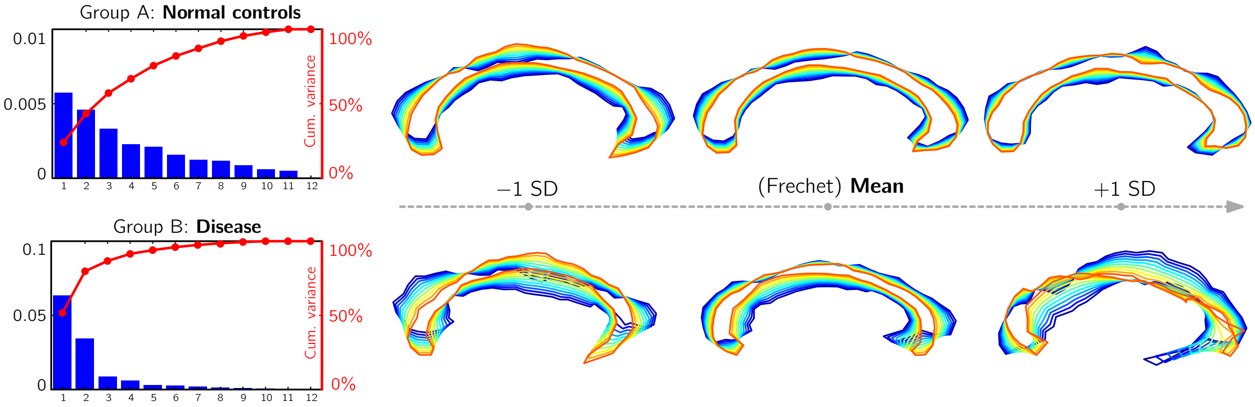


Figure 5.18: Visualization of the variances (left) and principal directions (right) of trajectory distributions for the *normal control* (top) and *disease* group (bottom) of corpus callosum shapes (best-viewed in color, blue to red: young to old).

	$(\hat{p}, \hat{u})$		$(\hat{p}, 0)$		$(0, \hat{u})$	
	$\bar{D}_M^{\mathcal{TM}}$	$D_B^{\mathcal{TM}}$	$\bar{D}_M^{\mathcal{TM}}$	$D_B^{\mathcal{TM}}$	$\bar{D}_M^{\mathcal{TM}}$	$D_B^{\mathcal{TM}}$
Distance on $\mathcal{TM}$	3.1817	4.0029	3.7377	3.6863	4.1537	4.3765
p-value	0.0241	**0.0054**	0.2014	0.0654	0.0319	**0.0046**

Table 5.3: Distances and estimated p-values (10000 random permutations) on corpora callosa using the averaged Mahalanobis distance ($\bar{D}_M^{\mathcal{TM}}$) and the generalized Bhattacharyya distance ($D_B^{\mathcal{TM}}$). The last two columns report the test results of dropping one of the initial conditions during the distance computation.

represented by 64 2D boundary landmarks.

Fig. 5.18 demonstrates the variances and the principal directions of the trajectories from the normal controls and the disease group. As shown in Fig. 5.18, the largest eigenvalue of the normal control group only accounts for 24% variability with a numeric value of 0.006, while the largest eigenvalue of the disease group accounts for 52% variability with a numeric value of 0.06. Fig. 5.18 (right) further shows the trajectories of each group along the first principal direction with standard deviations changing from -1 to 1. The plots indicate that the corpora callosa with dementia degenerate faster than the normal controls.

Table 5.3 reports the quantitative measures of the group tests on the corpus callosum shapes with 10000 permutations. Compared to the generalized squared-Mahalanobis distance, the generalized Bhattacharyya distance consistently exhibits better behavior in identifying the group differences. Similar to the experiments on the synthetic shapes, during the distance computation we drop one term of the initial conditions to measure which one plays a more important role in the group tests. As shown in Table 5.3, regardless of the distance measure, the initial velocity is most relevant in identifying group differences; this is consistent with [Muralidharan and Fletcher, 2012]. If we declare the statistical significance at the level of 0.01, the p-value of the generalized Bhattacharyya distance, either using both initial conditions or only the initial velocity, indicates that the disease group of corpus callosum shapes is significantly different from the normal control group.

5.3 Conclusions

This chapter presented a shape analysis framework for cross-sectional data that can provide both global and local information, yet does not require complex processes to establish point-to-point correspondences. It used the notion of band-depth of functions to order shapes according to how well they "fit in" a shape ensemble. This method allowed for the definition of a median and α-central regions of a population, which can then be used to compare different populations of shapes without strong distributional assumptions.

Different from the work presented in Chapter 4, which focuses on augmenting a population atlas with statistical information using weighted band-depth, the work in this chapter proposed a fast algorithm to compute the band-depth of shapes represented by binary maps, and most importantly it showed how band-depth can be used to provide both global and local statistical tests to differentiate between populations. In contrast to other deformation based tools for shape analysis, the presented approach is non-parametric and naturally captures how likely a shape belongs to a population. Although it does not provide physical measurements of displacement, these can be computed by deformation or a distance transform to the population median.

One limitation of our method is the reliance on a representative training set. Also, the size of the training set will affect the computed band-depth. Determining an appropriate size of the training set to obtain sufficient statistical power is left for future work. In addition, as discussed our current method works on the pixel/voxel level. In future work we will explore a multi-scale approach to better adapt the analysis results to the spatial scale of the expected differences. To further increase the statistical power of our method, we could include more factors, e.g., age, gender, into the shape analysis, and we explore their relationship with the estimated band-depths of shapes. This extension is not straightforward and therefore left to future work.

On the other hand, we have proposed an approach for studying group differences in the distributions of shape *trajectories*, estimated from longitudinal data. By means of a generalized version of the Bhattacharyya distance, we demonstrated, on both real and toy data, that taking second-order statistics into account can be beneficial in assessing group differences. However, the proposed approach also has limitations. For instance, although the compact representation of a trajectory is an efficient way to summarize longitudinal data, its accuracy inevitably influences the test-statistics. Currently, the adopted regression approach for estimating a trajectory is a generalization of linear regression in Euclidean space. Hence, we expect poor fitting performance on data that cannot be represented by a geodesic. For that reason, our test-statistic may not be appropriate under such a model. Furthermore, our real dataset only contains a limited number of subjects, which does not allow strong conclusions and requires to interpret results in the context of the low sample size.

A potential direction for future work is to apply our method to other types of longitudinal data, *e.g.*, images, which is straightforward but slightly more involved due to the complexity of the tangent bundle.

CHAPTER 6 : DISCUSSION AND FUTURE WORK

This chapter reviews and discusses the contributions of this dissertation in Section 6.1, followed by the discussion of future work in Section 6.2.

6.1 Summary of Contributions

This dissertation studied computational methods for spatiotemporal data in three parts: 1) regression models to capture time-varying changes in spatiotemporal data, 2) statistical atlas construction to summarize a representative data object and population variation information, and 3) statistical testing methods to identify shape differences between populations. In general, six approaches had been proposed to address the corresponding problems discussed in Chapter 1. Each contribution is restated with a discussion of how it was accomplished in this dissertation.

1. *To develop a uniform framework for fitting curves of increasing order for different types of spatiotemporal data, a model of parametric regression on the Grassmannian was proposed, by generalizing linear, time-warped, and cubic-spline regression in Euclidean space to the Grassmannian via optimal-control.*

   In Chapter 3, linear regression was revisited as an optimization problem with a line represented by second-order dynamics, which provides the intuition to generalize linear regression to geodesic regression on the Grassmann manifold. Time-warped geodesic regression was achieved by transforming the domain of the independent variable with a warping function. In the warped domain a simple geodesic regression can be used to achieve a good fitting result. For cubic spline regression, an extra force was added into the geodesic equation to allow controlled deviations from a geodesic, resulting in a spline curve. By further gluing spline pieces together, a formulation was obtained for cubic spline regression.

These different orders of regression models on the Grassmann manifold provided the flexibility to capture changes in spatiotemporal data, e.g., shape sequences and videos. They were applied to capture rat calvarium development and corpus callosum degeneration, as well as to predict traffic speed from videos and to count people in crowds. In general, these regression approaches can be applied to any data object that has a subspace representation. Apart from the applications studied in this dissertation, one could also use these methods to explore face recognition [Lui and Beveridge, 2008] and activity recognition [Slama et al., 2015].

2. *To simultaneously capture spatial deformations and intensity changes in image time-series, a model of metamorphic geodesic regression was proposed to jointly account for the appearance and shape changes and was efficiently solved using a simple, approximate algorithm through pairwise shooting metamorphosis.*

   The metamorphic geodesic regression approach was presented in Chapter 3. It was based on a shooting strategy under the framework of LDDMM registration. Shooting metamorphosis was parametrized by the initial image and the initial momentum, similar to the intercept and the slope of a straight line. This made it possible to regress over an image time-series through the generalization of linear least squares. To overcome the high computational cost of image regression, the initial image was assumed to be fixed, simplifying the regression solution to a weighted average of the initial momenta of pairwise shooting metamorphoses.

This method was able to capture white matter intensity changes caused by myelination in longitudinal MR images of macaque monkey brains. Generally speaking, this approach is designed for applications that focus on capturing diffeomorphic spatial deformation and image intensity changes simultaneously. For example, it could be applied to image morphing [Wolberg, 1998] and furthermore multi-image morphing [Georgiev and Wainer, 2001].

3. *To check if the model assumptions hold, a model criticism for regression models on the Grassmannian was proposed using kernel two-sample tests.*

   Model criticism was discussed in Chapter 3. It was achieved by performing a two-sample test with a kernelized variant of maximum mean discrepancy as the test statistic. The sampled data was generated on the Grassmannian under a Gaussian assumption. Two real applications were studied: degeneration of the corpus callosum during aging and developmental shape changes of the rat calvarium. The three regression models on the Grassmann manifold, i.e., geodesic regression, time-warped geodesic regression, and cubic spline regression, were tested to check model validity. While this model criticism approach was developed for the Grassmannian, the principles are applicable to smooth manifolds in general.

4. *To augment a single regressed curve with the local distribution of spatiotemporal data, e.g., confidence bounds or outliers, a model of statistical atlas construction via the weighted functional boxplot was proposed. This approach was applied to time-varying shapes and images.*

The time-varying statistical atlas was presented in Chapter 4. It was built based on a weighted functional boxplot, which enabled the generalization of concepts such as the median, percentiles, or outliers to spaces where the data objects are functions, shapes, or images. Also, it allowed spatiotemporal atlas-building through kernel regression. The utility of the approach was demonstrated by constructing statistical atlases for pediatric upper airways and corpora callosa, which revealed their growth patterns. In a pediatric airway study this method showed sensitivity in determining which children with subglottic stenoses received surgical intervention. Overall, this method is suitable for data objects that can be vectorized as functional data with application to robustly building a statistical atlas for them. For example, one could construct an atlas with a confidence region for image or shape segmentations obtained from different experts.

5. *To explore a new method in analyzing and comparing populations, e.g., normal controls and subjects with disease, a model of shape analysis based on depth-ordering was proposed to define statistical tests for identifying global and local shape differences without computing explicit dense correspondences or making strong distributional assumptions.*

   Ordering-based shape analysis was discussed in Chapter 5. In general, band-depth was used to non-parametrically define a global depth for a shape with respect to a reference population, typically consisting of normal control subjects. This allowed us to globally quantify differences with respect to “normality”. Using depth-ordering of shapes also allowed the detection of localized shape differences by using α-central values of shapes. Permutation tests were adopted to statistically assess global and local shape differences. The proposed method was evaluated on a synthetically generated striatum dataset, and it was also applied to detect shape differences in the hippocampus between subjects with first-episode schizophrenia and normal controls. Typically, this method can be used for analyzing any shapes that can be approximated using a binary representation and identifying shape differences among different populations.

6. *To facilitate statistical testing for trajectory distributions with similar mean and different variances, a model of hypothesis testing for longitudinal data was proposed by leveraging second-order statistics, i.e., variances of trajectory distributions, to identify group differences of shapes.*

   A hypothesis test for group differences of longitudinal data was presented in Chapter 5. And a generalization of principal geodesic analysis was introduced to the tangent bundle of a shape space. This allowed the estimation of the variance and principal directions of the distribution of trajectories that summarize shape variations within the longitudinal data. Each trajectory was parameterized as a point in the tangent bundle of the manifold. To study statistical differences in two distributions of trajectories, the Bhattacharyya distance in Euclidean space was generalized to the tangent bundle space. This not only allowed taking second-order statistics into account but also served as the test-statistic during the permutation test. This study shed new light on group differences in longitudinal corpus callosum shapes of subjects with dementia versus normal controls.

   In general, the assumptions of this method are that 1) individual changes can be captured using standard geodesic regression and 2) the mean and its covariance matrix can represent most information of interest in each trajectory population. Under these assumptions the method is able to analyze group differences within longitudinal data.

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

Thesis: *Advanced regression models or a time-varying statistical atlas can efficiently capture individual or population changes in spatiotemporal image and shape data. Statistical differences between shape populations can be detected using depth-ordering and statistics on shape trajectories.*

The first claim summarized the dissertation work on parametric regression in Chapter 3 and statistical atlas construction in Chapter 4. The regression models on the Grassmann manifold and the manifold of diffeomorphisms captured time-varying changes within image or shape time-series, e.g., the process of brain development or degeneration. On the other hand, the spatiotemporal statistical atlas captured changes, while further including population variations such as a confidence region and outliers. The second claim summarized the dissertation work on statistical shape analysis in Chapter 5. The hypothesis testing approaches were applied to detect group differences in shape populations. For cross-sectional data the tests were defined on shapes ordered with a depth value, which detected both global and local differences between populations. For longitudinal data the test was defined on shape trajectories that were summarized based on the data of individual subjects. Both methods identified statistical differences between shape populations.

6.2 Future Work

There are several directions that can be explored in the future. Section 6.2.1 will discuss four potential directions for further improving regression models on manifolds. Then, some thoughts on statistical shape analysis will be presented in Section 6.2.2, which will be followed by the discussion of model computing and visualization in Section 6.2.3. Finally, some other application areas will be discussed in Section 6.2.4.

6.2.1 Regression on Manifolds

In this dissertation the problem of regression on manifolds focused on estimating a function that describes the approximate relationship between manifold-valued data and its independent variable. Following this direction, there are several potential ways to further improve the estimated regression model. On one hand, a hierarchical model could be built to account for a more complex relationship among dependent and independent variables. If there are multiple independent variables, multivariate regression could be considered. On the other hand, one may wonder about the uncertain regions or the confidence intervals of the estimated regression parameters. This is the regression uncertainty problem of interest. Furthermore, when dealing with noisy data, typically a robust regression model is required to ensure that the model performance is not affected by the presence of outliers. These four potential directions for regression on manifolds will be discussed in the following.

(1) Hierarchical regression model. The regression models presented in Chapter 3 typically work on time-varying data collected from one subject or a population, without considering differences between subjects. A large data set with longitudinal data collected from each subject often has a hierarchical structure that allows modeling both individual and group trends. Hence, building a scalable regression model under a multilevel mixed-effects (hierarchical) framework is a reasonable strategy. Another advantage is that this approach can be parallelized due to its multilevel nature. In [Singh et al., 2013a] a hierarchical model was proposed for the manifold of diffeomorphisms to perform regression on longitudinal image sequences. It provided an approximate solution for its optimization problem by estimating individual and group trends separately. In future work I will develop a hierarchical regression model on the Grassmann manifold that can simultaneously learn the individual and group trends. This could be achievable because the geometry of the Grassmann manifold is relatively simpler than that of the manifold of diffeomorphisms. The big difference between the hierarchical model and the original one would be the number of regression parameters. The unknown parameters would include not only the initial conditions of the group trend but also the initial conditions of individual trajectories for each subject.

(2) Multivariate regression. The regression models presented in this dissertation are typically simple regressions, because the independent variable is a scalar number. In some scenarios we may need to consider multiple independent variables, e.g., age, weight, and gender, which could be modeled using multivariate regression [Mardia et al., 1980]. In particular, in a simple regression model the position of the underlying dynamic system is a function of one independent variable. While in a multivariate regression model the position could be a function of more than one independent variable. This formulation could extend the manifold-valued regression model presented in this dissertation to a general one that could handle more complex relationship among variables. Furthermore, if one has more independent variables than observations, or the independent variables are highly correlated, principal component regression (PCR) [Jolliffe, 1982] or partial least squares (PLS) [Geladi and Kowalski, 1986] could be adopted and extended to manifolds.

(3) Regression uncertainty. In Chapter 4 a confidence region (i.e., uncertainty) was estimated for the predicted data object at a specific time point. This was obtained under the non-parametric regression setting. Confidence regions of the prediction results could also be estimated for parametric regression. Take least squares for example. Uncertainties can be estimated for the intercept and the slope of the regression line by estimating their covariance matrix (i.e., the inverse of the Hessian matrix of regression parameters)[1]. In particular, these uncertainties result in a confidence interval (CI) for estimated parameters and a prediction interval (PI) for future observations [Rawlings et al., 1998]. Similarly, these concepts could be extended to regression on manifolds. For manifold-valued data we could compute uncertainties for the parameters of a regression model, e.g., the initial conditions, the initial position and the initial velocity. This could be achieved by estimating their Hessian matrix and taking the inverse of the Hessian matrix as the required covariance matrix. Furthermore, by propagating the covariance matrix at the initial position forward along the regression curve on manifolds, we could have a sequence of covariance matrices that describe how the variances of the estimators change with the independent value.

(4) Robust regression. The non-parametric kernel regression in Chapter 4 detected outliers in a data set when building an atlas. However, in Chapter 3 the parametric regression models were estimated under the assumption that the input data is clean without outliers. Since the presence of outliers might greatly affect the accuracy of the regression results, one may want to develop a robust regression which can be used on noisy data. To solve this problem, one could perform regression with random sample consensus (RANSAC) [Fischler and Bolles, 1981] to detect outliers. Another promising strategy is to generalize the least trimmed squares (LTS) or the least median of squares (LMS) regression [Rousseeuw and Leroy, 2005] to manifolds for robust regression.

[1] Each parameter may have different units, resulting in values at different scales. Here, the units are ignored, while it will be interesting to consider them in the model.

6.2.2 Statistical Shape Analysis

(1) Statistics of shape trajectories. Chapter 5 focused on identifying shape differences between two populations. For data from different populations or data from one population while having multiple subgroups, other interesting topics are classification or clustering problems. In future work I would like to discriminate subjects with disease from normal controls, for example, using the shape trajectories estimated for each subject. Furthermore, when analyzing shape differences in both cross-sectional and longitudinal data I observed that the disease group usually has a much larger variance than the normal control group. It may be that subgroups exist in the disease group. It will be interesting to see if we can cluster disease subjects into different subgroups and to study the shared structures or features in each subgroup.

(2) Shape domain adaptation. Domain adaptation [Ben-David et al., 2010] aims at learning models from sufficient data in source domains and deploying them across target domains with less data. In the real world of shape analysis we often encounter the task of domain adaptation. For instance, we may have a large population of shape data collected from subjects at middle ages. With enough data in this source domain we could learn a model for further analysis, e.g., a shape subspace using principal component analysis (PCA). While we may lack data collected from subjects at older ages to learn a reasonable subspace in this target domain. To address this issue, one could transport or predict a new subspace from available ones that are learnt from source domains. Since we can represent a subspace as an element on the Grassmann manifold, the regression models presented in this dissertation could be further applied to predict a new subspace for the target domain.

6.2.3 Model Computing and Visualization

Apart from the above theoretical directions for future work, in practice efficient model computing and visualization would greatly promote the widespread use of the proposed computational models. This is especially important and worth the effort for models that are built on high-dimensional complex data, e.g., the manifold-valued data studied in this dissertation.

(1) Efficient model computing. Computational models for analyzing high-dimensional data in non-Euclidean spaces often require efficient algorithms to estimate the model parameters. For example, an unoptimized program for the original LDDMM could take hours to register two 3D images, making the analysis of a large number of images an intractable task. One direction is to seek a "sweet spot" to trade accuracy for efficiency, which is often achieved by reformulating models under reasonable assumptions. The simple image regression in Chapter 3 followed this approach. The resulting pairwise image registrations were independent tasks which can be easily parallelized. In future work, I will adopt similar strategies to speed up our regression models so that they can deal with large-scale data sets.

In another direction, as commodity computing hardware is increasingly affordable, an ideal system should be able to exploit the computing power of multiple machines and coordinate parallel model computation on them. The key strategy of this approach is to divide a large problem into smaller ones and solve each of them in parallel on a cluster of machines. One could build such a data processing system on top of a parallel computing framework (e.g., MapReduce [Dean and Ghemawat, 2008] and Spark [Zaharia et al., 2012]) with customized coordination among all computing nodes, enabling data processing at scale.

(2) Data and model visualization. The data objects studied in this dissertation are typically high-dimensional. It is crucial to understand the data before solving a problem or formulating a model; however, humans have difficulty grasping high-dimensional data. Visualization provides an intuitive way to interact with such data; techniques such as parallel coordinates and feature extraction can be used to visualize high-dimensional data. For complex high-dimensional data (e.g., manifold-valued data), efficient visualization techniques rarely exist. Furthermore, visualization of model fitness for data sets has not been well explored, despite being of great importance in checking data quality, model performance, and even parameter settings. In future work, I will explore the visualization of high-dimensional manifold-valued data, and the interaction between the data and model estimates. This will allow exploring how high-dimensional data is distributed in a non-Euclidean space and which subset of the data is well represented by the model and which subset is not.

6.2.4 Other Application Areas

(1) Video prediction. In computer vision researchers are interested in predicting a new frame for a video or a new entire video. Since a video can be treated as a dynamic system, this problem could be reduced to predict a new dynamic system. In Chapter 3 we modeled the underlying dynamics of a video with the observability matrix obtained from linear dynamic system (LDS) identification. Using the regression models on the Grassmannian, we can predict a new observability matrix. As a result, we can reconstruct the dynamic and appearance matrices, i.e., $\mathbf{A}$ and $\mathbf{C}$, respectively (see Section 2.1.3). By further estimating the initial state $\mathbf{x}_0$ and associated noise models $\mathbf{w}_k$ and $\mathbf{v}_k$ from the input data, it might be possible to synthesize a new video for a specific independent value. This could also be used to predict a video of a cardiac or respiratory cycle in medical image analysis.

(2) Disease diagnosis. One important task of medical image analysis is to help diagnosis of a disease. Using discrepancies among shapes or shape trajectories of objects of interest, one could design a grading system for a disease based on the shape or trajectory distribution of the training data. For example, given shape populations collected from both normal controls and disease subjects as training sets, one could compute band depth for a test shape with respect to the control training group and the disease training group respectively. This would result in two depth values. Their ratio, e.g., the disease one divided by the normal control one, could be used as the diagnosis score. In particular, a normal control subject would have a relatively lower score, which means this subject has a lower risk for having this disease. In contrast, a subject with disease would be expected to have a relatively higher score. For longitudinal data, according to the study in this dissertation the velocity of a shape trajectory is dominant when separating disease subjects from normal controls. Therefore, we could look at the changing speed of a shape and use it to define a score for diagnosis.

BIBLIOGRAPHY

[Absil et al., 2004] Absil, P.-A., Mahony, R., and Sepulchre, R. (2004). Riemannian geometry of Grassmann manifolds with a view on algorithmic computation. *Acta Appl. Math.*, 80(2):199–220.

[Ahlberg et al., 1967] Ahlberg, J. H., Nilson, E. N., and Walsh, J. L. (1967). *The Theory of Splines and Their Applications.* Academic Press.

[Aljabar et al., 2009] Aljabar, P., Heckemann, R., Hammers, A., Hajnal, J., and Rueckert, D. (2009). Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy. *NeuroImage*, 46:726–739.

[Ashburner and Friston, 2011] Ashburner, J. and Friston, K. (2011). Diffeomorphic registration using geodesic shooting and Gauss-Newton optimization. *Neuroimage*, 55(3):954–967.

[Banyaga, 1997] Banyaga, A. (1997). *The structure of classical diffeomorphism groups.* Springer.

[Batzies et al., 2015] Batzies, E., Hüper, K., Machado, L., and Leite, F. S. (2015). Geometric mean and geodesic regression on Grassmannians. *Linear Algebra Appl.*, 466:83–101.

[Beg et al., 2005] Beg, M., Miller, M., Trouvé, A., and Younes, L. (2005). Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *IJCV*, 61(2):139–157.

[Begelfor and Werman, 2006] Begelfor, E. and Werman, W. (2006). Affine invariance revisited. In *CVPR*.

[Ben-David et al., 2010] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., and Vaughan, J. W. (2010). A theory of learning from different domains. *Machine learning*, 79(1-2):151–175.

[Benjamini and Hochberg, 1995] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300.

[Bhattacharyya, 1946] Bhattacharyya, A. (1946). On a measure of divergence between two multinomial populations. *Sankhyā: The Indian Journal of Statistics*, pages 401–406.

[Bickel and Levina, 2008] Bickel, P. and Levina, E. (2008). Covariance regularization by thresholding. *Ann. Stat.*, pages 2577–2604.

[Bookstein, 1991] Bookstein, F. (1991). Morphometric tools for landmark data: geometry and biology. *Cambridge Univ. Press.*

[Boothby, 1986] Boothby, W. (1986). *An Introduction to Differentiable Manifolds and Riemannian Geometry.* Academic Press.

[Boumal, 2013] Boumal, N. (2013). Interpolation and regression of rotation matrices. In Nielsen, F. and Barbaresco, F., editors, *Geometric Science of Information*, volume 8085 of *Lecture Notes in Computer Science*, pages 345–352. Springer Berlin Heidelberg.

[Boumal and Absil, 2011a] Boumal, N. and Absil, P.-A. (2011a). A discrete regression method on manifolds and its application to data on $\mathcal{SO}(n)$. In *IFAC*.

[Boumal and Absil, 2011b] Boumal, N. and Absil, P.-A. (2011b). Discrete regression methods on the cone of positive-definite matrices. In *ICASSP*.

[Boyd and Vandenberghe, 2004] Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

[Burns and Iliffe, 2009] Burns, A. and Iliffe, S. (2009). Alzheimer's disease. *BMJ*, 338(b158).

[Camarinha et al., 1995] Camarinha, M., Leite, F. S., and Crouch, P. (1995). Splines of class C^k on non-Euclidean spaces. *IMA J. Math. Control Info.*, 12(4):399–410.

[Cates et al., 2008] Cates, J., Fletcher, P. T., Styner, M., Hazlett, H. C., and Whitaker, R. (2008). Particle-based shape analysis of multi-object complexes. *Med Image Comput Comput Assist Interv*, 11(Pt 1):477–485.

[Chan and Vasconcelos, 2005] Chan, A. and Vasconcelos, N. (2005). Classification and retrieval of traffic video using auto-regressive stochastic processes. In *IV*.

[Chan and Vasconcelos, 2012] Chan, A. and Vasconcelos, N. (2012). Counting people with low-level features and Bayesian regression. *IEEE Trans. Image Process.*, 12(4):2160–2177.

[Chung et al., 2008] Chung, M. K., Dalton, K. M., and Davidson, R. J. (2008). Tensor-based cortical surface morphometry via weighted spherical harmonic representation. *IEEE Trans Med Imaging*, 27(8):1143–1151.

[Cootes et al., 1995] Cootes, T. F., Taylor, C. J., Cooper, D. H., and Graham, J. (1995). Active shape models – their training and application. *Comput. Vis. Image Underst.*, 61(1):38–59.

[Cootes et al., 2004] Cootes, T. F., Taylor, C. J., et al. (2004). Statistical models of appearance for computer vision.

[Cortes and Vapnik, 1995] Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20:273–297.

[Crouch and Leite, 1995] Crouch, P. and Leite, F. S. (1995). The dynamic interpolation problem: On Riemannian manifolds, Lie groups, and symmetric spaces. *J. Dyn. Control Syst.*, 1(2):177–202.

[Dale et al., 1999] Dale, A. M., Fischl, B., and Sereno, M. I. (1999). Cortical surface-based analysis: I. segmentation and surface reconstruction. *Neuroimage*, 9(2):179–194.

[Daniel, 2006] Daniel, S. (2006). The upper airway: Congenital malformations. *Pediatric Respiratory Reviews*, 7S:S260–S263.

[Davies et al., 2008] Davies, R., Twining, C., and Taylor, C. (2008). *Statistical Models of Shape: Optimisation and Evaluation*. Springer, London.

[Davis et al., 2007] Davis, B., P. T. Fletcher, Bullit, E., and Joshi, S. (2007). Population shape regression from random design data. In *ICCV*.

[Davis et al., 2010] Davis, B. C., Fletcher, P. T., Bullitt, E., and Joshi, S. (2010). Population shape regression from random design data. *International journal of computer vision*, 90(2):255–266.

[Dean and Ghemawat, 2008] Dean, J. and Ghemawat, S. (2008). Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113.

[Do Carmo, 1992] Do Carmo, M. P. (1992). Riemannian geometry, mathematics: Theory & applications.

[Doretto et al., 2003] Doretto, G., Chiuso, A., Wu, Y., and Soatto, S. (2003). Dynamic textures. *Int. J. Comput. Vision*, 51(2):91–109.

[Driesen and Raz, 1995] Driesen, N. and Raz, N. (1995). The influence of sex, age, and handedness on corpus callosum morphology: a meta-analysis. In *Psychobiology*, volume 23(3), pages 240–247.

[Dryden and Mardia, 1998] Dryden, I. L. and Mardia, K. V. (1998). *Statistical shape analysis*, volume 4. J. Wiley Chichester.

[Durrleman et al., 2013] Durrleman, S., Pennec, X., Trouvé, A., Braga, J., Gerig, G., and Ayache, N. (2013). Toward a comprehensive framework for the spatiotemporal statistical analysis of longitudinal shape data. *Int. J. Comput. Vision*, 103(1):22–59.

[Edelman et al., 1998] Edelman, A., Arias, T., and Smith, S. T. (1998). The geometry of algorithms with orthogonality constraints. *SIAM J. Matrix Anal. Appl.*, 20(2):303–353.

[Ernst et al., 2004] Ernst, M. D. et al. (2004). Permutation methods: a basis for exact inference. *Statistical Science*, 19(4):676–685.

[Evans et al., 2006] Evans, A. C., Group, B. D. C., et al. (2006). The nih mri study of normal brain development. *Neuroimage*, 30(1):184–202.

[Fekedulegn et al., 1999] Fekedulegn, D., Mac Siurtain, M., and Colbert, J. (1999). Parameter estimation of nonlinear growth models in forestry. *Silva Fennica*, 33(4):327–336.

[Fischler and Bolles, 1981] Fischler, M. A. and Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395.

[Fletcher et al., 2004] Fletcher, P., Lu, C., Pizer, S., and Joshi, S. (2004). Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE TMI*, 23(8):995–1005.

[Fletcher et al., 2009] Fletcher, P., Venkatasubramanian, S., and Joshi, S. (2009). The geometric median on Riemannian manifolds with application to robust atlas estimation. *NeuroImage*, 45(Suppl 1):S143–S152.

[Fletcher, 2013] Fletcher, P. T. (2013). Geodesic regression and the theory of least squares on Riemannian manifolds. *Int. J. Comput. Vision*, 105(2):171–185.

[Fletcher, 2011] Fletcher, T. (2011). Geodesic regression on Riemannian manifolds. In *3rd MICCAI workshop on mathematical foundations of computational anatomy*, pages 75–86.

[Freedman, 2009] Freedman, D. A. (2009). *Statistical models: theory and practice*. cambridge university press.

[Frigge et al., 1989] Frigge, M., Hoaglin, D. C., and Iglewicz, B. (1989). Some implementations of the boxplot. *The American Statistician*, 43(1):50–54.

[Gallivan et al., 2003] Gallivan, K., Srivastava, A., Xiuwen, L., and Dooren, P. V. (2003). Efficient algorithms for inferences on Grassmann manifolds. In *Statistical Signal Processing Workshop*, pages 315–318.

[Gao and Bouix, 2012] Gao, Y. and Bouix, S. (2012). Synthesis of realistic subcortical anatomy with known surface deformations. In Levine, J. A., Paulsen, R. R., and Zhang, Y., editors, *Mesh Processing in Medical Image Analysis*, pages 80–88. Springer, Heidelberg.

[Gao et al., 2014] Gao, Y., RiklinRaviv, T., and Bouix, S. (2014). Shape analysis, a field in need of careful validation. *Human brain mapping*.

[Garcin and Younes, 2005] Garcin, L. and Younes, L. (2005). Geodesic image matching: A wavelet based energy minimization scheme. In *EMMCVPR*, volume 3757 of *LNCS*, pages 349–364.

[Geladi and Kowalski, 1986] Geladi, P. and Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Analytica chimica acta*, 185:1–17.

[Georgiev and Wainer, 2001] Georgiev, T. and Wainer, M. (2001). Compression and editing of movies by multi-image morphing. US Patent 6,285,794.

[Gerber et al., 2010] Gerber, S., Tasdizen, T., Fletcher, P. T., Joshi, S., and Whitaker, R. (2010). Manifold modeling for brain population analysis. *Medical image analysis*, 14(5):643–653.

[Gilmore et al., 2012] Gilmore, J., Shi, F., Woolson, S., Knickmeyer, R., Short, S., Lin, W., Zhu, H., Hamer, R., Styner, M., and Shen, D. (2012). Longitudinal development of cortical and subcortical gray matter from birth to 2 years. *Cereb. Cortex*, 22(11):2478–2485.

[Gong et al., 2012] Gong, B., Shi, Y., Sha, F., and Grauman, K. (2012). Geodesic flow kernel for unsupervised domain adaptation. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2066–2073. IEEE.

[Gretton et al., 2012] Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *J. Mach. Learn. Res.*, 13:723–773.

[Hamm and Lee, 2008] Hamm, J. and Lee, D. D. (2008). Grassmann discriminant analysis: a unifying view on subspace-based learning. In *ICML*.

[Harandi et al., 2014] Harandi, M. T., Salzmann, M., Jayasumana, S., Hartley, R., and Li, H. (2014). Expanding the family of Grassmannian kernels: An embedding perspective. In *ECCV*.

[Hardle and Marron, 1985] Hardle, W. and Marron, J. S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *The Annals of Statistics*, pages 1465–1481.

[Hart et al., 2010] Hart, G., Shi, Y., Zhu, H., Sanchez, M., Styner, M., and Niethammer, M. (2010). DTI longitudinal atlas construction as an average of growth models. *MICCAI STIA*.

[Hart et al., 2009] Hart, G., Zach, C., and Niethammer, M. (2009). An optimal control approach for deformable registration. In *MMBIA*, pages 9–16.

[Hinkle et al., 2014] Hinkle, J., P. T. Fletcher, and Joshi, S. (2014). Intrinsic polynomials for regression on Riemannian manifolds. *J. Math. Imaging Vis.*, 50:32–52.

[Holm et al., 2009] Holm, D. D., Trouvé, A., and Younes, L. (2009). The Euler-Poincaré theory of metamorphosis. *Quarterly of Applied Mathematics*, 67:661–685.

[Hong et al., 2013a] Hong, Y., Davis, B., Marron, J. S., Kwitt, R., and Niethammer, M. (2013a). Weighted functional boxplot with application to statistical atlas construction. In Mori, K., Sakuma, I., Sato, Y., Barillot, C., and Navab, N., editors, *MICCAI 2013, Part III. LNCS*, volume 8151, pages 584–591. Springer, Heidelberg.

[Hong et al., 2014a] Hong, Y., Davis, B., Marron, J. S., Kwitt, R., Singh, N., Kimbell, J. S., Pitkin, E., Superfine, R., Davis, S. D., Zdanski, C. J., and Niethammer, M. (2014a). Statistical atlas construction via weighted functional boxplots. *Medical image analysis*, 18(4):684–698.

[Hong et al., 2014b] Hong, Y., Gao, Y., Niethammer, M., , and Bouix, S. (2014b). Depth-based shape-analysis. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2014)*, pages 17–24.

[Hong et al., 2015a] Hong, Y., Gao, Y., Niethammer, M., and Bouix, S. (2015a). Shape analysis based on depth-ordering. *Medical image analysis*, 25(1):2–10.

[Hong et al., 2012a] Hong, Y., Joshi, S., Sanchez, M., Styner, M., and Niethammer, M. (2012a). Metamorphic geodesic regression. In Ayache, N., Delingette, H., Golland, P., and Mori, K., editors, *MICCAI, 2012, Part III. LNCS, vol. 7512*, pages 197–205.

[Hong et al., 2015b] Hong, Y., Kwitt, R., and Niethammer, M. (2015b). Model criticism for regression on the Grassmannian. In *MICCAI*.

[Hong et al., 2014c] Hong, Y., Kwitt, R., Singh, N., Davis, B., Vasconcelos, N., and Niethammer, M. (2014c). Geodesic regression on the Grassmannian. In *ECCV*.

[Hong et al., 2016] Hong, Y., Kwitt, R., Singh, N., Vasconcelos, N., and Niethammer, M. (2016). Parametric regression on the Grassmannian. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP(99).

[Hong et al., 2013b] Hong, Y., Niethammer, M., Andruejol, J., Kimbel, J., Pitkin, E., Superfine, R., Davis, S., Zdanski, C., and Davis, B. (2013b). A pediatric airway atlas and its application in subglottic stenosis. In *International symposium on biomedical imaging: from nano to macro*, pages 1194–1197.

[Hong et al., 2012b] Hong, Y., Shi, Y., Styner, M., Sanchez, M., and Niethammer, M. (2012b). Simple geodesic regression for image time-series. In *WBIR*.

[Hong et al., 2014d] Hong, Y., Singh, N., Kwitt, R., and Niethammer, M. (2014d). Time-warped geodesic regression. In *MICCAI*.

[Hong et al., 2015c] Hong, Y., Singh, N., Kwitt, R., and Niethammer, M. (2015c). Group testing for longitudinal data. In *Information Processing in Medical Imaging*, pages 139–151. Springer.

[Hopper et al., 1994] Hopper, K., Patel, S., Cann, T., Wilcox, T., and Schaeffer, J. (1994). The relationship of age, gender, handedness and sidedness to the size of the corpus callosum. *Acad. Radiol.*, 1:243–248.

[Hosseinbor et al., 2014] Hosseinbor, A. P., Kim, W. H., Adluru, N., Acharya, A., Vorperian, H. K., and Chung, M. K. (2014). The 4D hyperspherical diffusion wavelet: A new method for the detection of localized anatomical variation. *Med Image Comput Comput Assist Interv*, 17(Pt 3):65–72.

[Hughes et al., 1978] Hughes, P., Tanner, J., and Williams, J. (1978). A longitudinal radiographic study of the growth of the rat skull. *J. Anat.*, 127(1):83–91.

[Johnson et al., 1994] Johnson, S., Farnworth, T., Pinkston, J., Bigler, E., and Blatter, D. (1994). Corpus callosum surface area across the human adult life span: Effect of age and gender. *Brain Res. Bull.*, 35(4):373–377.

[Jolliffe, 1982] Jolliffe, I. T. (1982). A note on the use of principal components in regression. *Applied Statistics*, pages 300–303.

[Jones, 1993] Jones, M. C. (1993). Simple boundary correction for kernel density estimation. *Statistics and Computing*, 3:135–146.

[Joshi et al., 2004] Joshi, S., Davis, B., and Jomier, M. (2004). Unbiased diffeomorphic atlas construction for computational anatomy. *Neuroimage*, 23(Suppl 1):S151–S160.

[Kalman, 1959] Kalman, R. (1959). On the general theory of control systems. *IRE Transactions on Automatic Control*, 4(3):110–110.

[Kedem and Fokianos, 2005] Kedem, B. and Fokianos, K. (2005). *Regression models for time series analysis*, volume 488. John Wiley & Sons.

[Kendall, 1984] Kendall, D. G. (1984). Shape manifolds, procrustean metrics, and complex projective spaces. *Bulletin of the London Mathematical Society*, 16(2):81–121.

[Kogure et al., 2000] Kogure, D., Matsuda, H., Ohnishi, T., Asada, T., Uno, M., Kunihiro, T., Nakano, S., and Takasaki, M. (2000). Longitudinal evaluation of early alzheimer's disease using brain perfusion spect. *Journal of nuclear medicine*, 41(7):1155–1162.

[Kuklisova-Murgasova et al., 2011] Kuklisova-Murgasova, M., Aljabar, P., Srinivasan, L., Counsell, S. J., Doria, V., Serag, A., Gousias, I. S., Boardman, J. P., Rutherford, M. A., Edwards, A. D., et al. (2011). A dynamic 4d probabilistic atlas of the developing brain. *NeuroImage*, 54(4):2750–2763.

[Ledoit and Wolf, 2004] Ledoit, O. and Wolf, M. (2004). A well-conditioned estimator for large-dimensional covariance matrices. *J. Multivar. Anal.*, 88(2):365–411.

[Lee, 2012] Lee, J. (2012). *Introduction to smooth manifolds*. Springer.

[Liu et al., 1999] Liu, R., Parelius, J., and Singh, K. (1999). Multivariate analysis by data depth: Descriptive statistics, graphics and inference. *The Annals of Statistics*, 27:783–858.

[Lloyd and Ghahramani, 2015] Lloyd, J. R. and Ghahramani, Z. (2015). Statistical model criticism using kernel two sample tests. Preprint at `http://mlg.eng.cam.ac.uk/Lloyd/papers/`.

[Loncaric, 1998] Loncaric, S. (1998). A survey of shape analysis techniques. *Pattern Recognition*, 31(8):983–1001.

[López-Pintado and Romo, 2009] López-Pintado, S. and Romo, J. (2009). On the concept of depth for functional data. *Journal of the American Statistical Association*, 104:718–734.

[Lui, 2012] Lui, Y. (2012). Human gesture recognition on product manifolds. *JMLR*, 13:3297–3321.

[Lui and Beveridge, 2008] Lui, Y. M. and Beveridge, J. R. (2008). Grassmann registration manifolds for face recognition. In *Computer Vision–ECCV 2008*, pages 44–57. Springer.

[Machado et al., 2010] Machado, L., Silva Leite, F., and Krakowski, K. (2010). Higher-order smoothing splines versus least-squares problems on Riemannian manifolds. *J. Dyn. Control Syst.*, 16(1):121–148.

[Mahalanobis, 1936] Mahalanobis, P. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences (Calcutta)*, 2:49–55.

[Marcus et al., 2010] Marcus, D., Fotenos, A., Csernansky, J., Morris, J., and Buckner, R. (2010). Open access series of imaging studies: longitudinal mri data in nondemented and demented older adults. *J. Cognitive Neurosci.*, 22(12):2677–2684.

[Mardia et al., 1980] Mardia, K. V., Kent, J. T., and Bibby, J. M. (1980). *Multivariate analysis*. Academic press.

[Marron and Nolan, 1988] Marron, J. and Nolan, D. (1988). Canonical kernels for density estimation. *Statistics and probability letters*, 7:195–199.

[Marron and Ruppert, 1994] Marron, J. and Ruppert, D. (1994). Transformations to reduce boundary bias in kernel density estimation. *Journal of the royal statistical society*, 56:653–671.

[Marron and Alonso, 2014] Marron, J. S. and Alonso, A. M. (2014). Overview of object oriented data analysis. *Biometrical Journal*, 56(5):732–753.

[McClure et al., 2013] McClure, R. K., Styner, M., Maltbie, E., Lieberman, J. A., Gouttard, S., Gerig, G., Shi, X., and Zhu, H. (2013). Localized differences in caudate and hippocampal shape are associated with schizophrenia but not antipsychotic type. *Psychiatry Research: Neuroimaging*, 211(1):1–10.

[Miller, 2004] Miller, M. I. (2004). Computational anatomy: Shape, growth, and atrophy comparison via diffeomorphisms. *NeuroImage*, 23:S19–S33.

[Miller and Younes, 2001] Miller, M. I. and Younes, L. (2001). Group actions, homeomorphisms, and matching: A general framework. *International Journal of Computer Vision*, 41:61–84.

[Moussa and Cheema, 1998] Moussa, M. and Cheema, M. (1998). Non-parametric regression in curve fitting. *The Statistician*, 41:209–225.

[Muralidharan and Fletcher, 2012] Muralidharan, P. and Fletcher, P. (2012). Sasaki metrics for analysis of longitudinal data on manifolds. In *CVPR*, pages 1027–1034.

[Myer et al., 1994] Myer, C. r., O'Connor, D., and Cotton, R. (1994). Proposed grading system for subglottic stenosis based on endotracheal tube sizes. *Ann Otol Rhinol Laryngol*, 103(4 Pt 1):319–323.

[Niethammer et al., 2011] Niethammer, M., Huang, Y., and Vialard, F.-X. (2011). Geodesic regression for image time-series. In Fichtinger, G., Martel, A., and Peters, T., editors, *MICCAI 2011, Part II. LNCS*, volume 6892, pages 655–662. Springer, Heidelberg.

[Nitzken et al., 2014] Nitzken, M. J., Casanova, M. F., Gimelfarb, G., Inanc, T., Zurada, J. M., and El-Baz, A. (2014). Shape analysis of the human brain: a brief survey. *IEEE J Biomed Health Inform*, 18(4):1337–1354.

[Noakes et al., 1989] Noakes, L., Heinzinger, G., and Paden, B. (1989). Cubic splines on curved spaces. *IMA J. Math. Control Info.*, 6(4):465–473.

[Nocedal and Wright, 2006] Nocedal, J. and Wright, S. (2006). *Numerical optimization*. Springer Science & Business Media.

[Oliveira and Tavares, 2014] Oliveira, F. P. and Tavares, J. M. R. (2014). Medical image registration: a review. *Computer methods in biomechanics and biomedical engineering*, 17(2):73–93.

[Oliver, 1998] Oliver, D. (1998). Calculation of the inverse of the covariance. *Math. Geol.*, 30(7):911–933.

[Pizer et al., 2003] Pizer, S. M., Fletcher, P. T., Joshi, S., Thall, A., Chen, J. Z., Fridman, Y., Fritsch, D. S., Gash, A. G., Glotzer, J. M., Jiroutek, M. R., et al. (2003). Deformable m-reps for 3d medical image segmentation. *International Journal of Computer Vision*, 55(2-3):85–106.

[Provost and Kohavi, 1998] Provost, F. and Kohavi, R. (1998). On applied research in machine learning. In *Machine learning*, pages 127–132.

[Ramsay and Silverman, 2005] Ramsay, J. and Silverman, B. (2005). *Functional Data Analysis*. Springer.

[Rawlings et al., 1998] Rawlings, J. O., Pantula, S. G., and Dickey, D. A. (1998). *Applied regression analysis: a research tool*. Springer Science & Business Media.

[Rematas et al., 2013] Rematas, K., Fernando, B., Tommasi, T., and Tuytelaars, T. (2013). Does evolution cause a domain shift. In *The First International Workshop on Visual Domain Adaptation and Dataset Bias*.

[Rentmeesters, 2011] Rentmeesters, Q. (2011). A gradient method for geodesic data fitting on some symmetric Riemannian manifolds. In *CDC-ECC*.

[Reuter et al., 2009] Reuter, M., Wolter, F.-E., Shenton, M., and Niethammer, M. (2009). Laplace–beltrami eigenvalues and topological features of eigenfunctions for statistical shape analysis. *Computer-Aided Design*, 41(10):739–755.

[Risser et al., 2011] Risser, L., Vialard, F., Wolz, R., Murgasova, M., Holm, D., and Rueckert, D. (2011). Simultaneous multiscale registration using large deformation diffeomorphic metric mapping. *IEEE Transactions on Medical Imaging*, 30(10):1746–1759.

[Rohlfing et al., 2009] Rohlfing, T., Sullivan, E., and Pfefferbaum, A. (2009). Regression models of atlas appearance. In *Information Processing in Medical Imaging*, pages 151–162. Springer.

[Rousseeuw and Leroy, 2005] Rousseeuw, P. J. and Leroy, A. M. (2005). *Robust regression and outlier detection*, volume 589. John Wiley & Sons.

[Samir et al., 2012] Samir, C., Absil, P., Srivastava, A., and Klassen, E. (2012). A gradient-descent method for curve fitting on Riemannian manifolds. *Found. Comp. Math*, 12(1):49–73.

[Sasaki, 1958] Sasaki, S. (1958). On the differential geometry of tangent bundles of Riemannian manifolds. *TMJ*, 10(3):338–354.

[Scahill et al., 2003] Scahill, R. I., Frost, C., Jenkins, R., Whitwell, J. L., Rossor, M. N., and Fox, N. C. (2003). A longitudinal study of brain volume changes in normal aging using serial registered magnetic resonance imaging. *Archives of neurology*, 60(7):989–994.

[Schuster, 1985] Schuster, E. (1985). Incorporating support constraints into nonparametric estimators of densities. *Communications in Statistics - Theory and Methods*, 14:1123–1136.

[Sepiashvili et al., 2003] Sepiashvili, D., Moura, J., and Ha, V. (2003). Affine-permutation symmetry: Invariance and shape space. In *IEEE Workshop on Statistical Signal Processing*.

[Siegel, 1956] Siegel, S. (1956). *Nonparametric statistics for the behavioral sciences.* New York: McGraw-Hill.

[Singh et al., 2013a] Singh, N., Hinkle, J., Joshi, S., and P. T. Fletcher (2013a). A hierarchical geodesic model for diffeomorphic longitudinal shape analysis. In Gee, J., Joshi, S., Pohl, K., Wells, W., and Zöllei, L., editors, *IPMI 2013, LNCS*, volume 7917, pages 560–571. Springer.

[Singh et al., 2013b] Singh, N., Hinkle, J., Joshi, S., and P. T. Fletcher (2013b). A vector momenta formulation of diffeomorphisms for improved geodesic regression and atlas construction. In *ISBI*.

[Singh and Niethammer, 2014] Singh, N. and Niethammer, M. (2014). Splines for diffeomorphic image regression. In *MICCAI*.

[Slama et al., 2015] Slama, R., Wannous, H., Daoudi, M., and Srivastava, A. (2015). Accurate 3d action recognition using learning on the grassmann manifold. *Pattern Recognition*, 48(2):556–567.

[Snedecor and Cochran, 1989] Snedecor, G. W. and Cochran, W. G. (1989). *Statistical Methods, Eighth Edition.* Iowa State University Press.

[Styner et al., 2004] Styner, M., Lieberman, J. A., Pantazis, D., and Gerig, G. (2004). Boundary and medial shape analysis of the hippocampus in schizophrenia. *Medical Image Analysis*, 8(3):197–203.

[Su et al., 2012] Su, J., Dryden, I., Klassen, E., Le, H., and Srivastava, A. (2012). Fitting smoothing splines to time-indexed, noisy points on non-linear manifolds. *Image Vision Comput.*, 30:428–442.

[Su et al., 2014a] Su, J., Kurtek, S., Klassen, E., and Srivastava, A. (2014a). Statistical analysis of trajectories on Riemannian manifolds: Bird migration, hurricane tracking and video surveillance. *Ann. Appl. Stat.*, 8(1):530–552.

[Su et al., 2014b] Su, J., Srivastrava, A., de Souza, F., and Sarkar, S. (2014b). Rate-invariant analysis of trajectories on Riemannian manifolds with application in visual speech recognition. In *CVPR*.

[Sun and Genton, 2011] Sun, Y. and Genton, M. (2011). Functional boxplots. *Journal of Computational and Graphical Statistics*, 20:316–334.

[Sun et al., 2012] Sun, Y., Genton, M. G., and Nychka, D. W. (2012). Exact fast computation of band depth for large functional datasets: How quickly can one million curves be ranked? *Stat*, 1:68–74.

[Trouvé and Vialard, 2012] Trouvé, A. and Vialard, F.-X. (2012). Shape splines and stochastic shape evolutions: A second order point of view. *Quart. Appl. Math*, 70(2):219–251.

[Turaga et al., 2010] Turaga, P., Biswas, S., and Chellappa, R. (2010). The role of geometry for age estimation. In *ICASSP*.

[Vialard et al., 2012] Vialard, F., Risser, L., Rueckert, D., and Cotter, C. (2012). Diffeomorphic 3D image registration via geodesic shooting using an efficient adjoint calculation. *International Journal of Computer Vision*, 97(2):229–241.

[Wachinger et al., 2014] Wachinger, C., Golland, P., and Reuter, M. (2014). Brainprint: Identifying subjects by their brain. *Med Image Comput Comput Assist Interv*, 17(Pt 3):41–8.

[Wand and Jones, 1994] Wand, M. and Jones, M. (1994). *Kernel Smoothing*. CRC Press.

[Whitaker et al., 2013] Whitaker, R. T., Mirzargar, M., and Kirby, R. M. (2013). Contour boxplots: A method for characterizing uncertainty in feature sets from simulation ensembles. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2713–2722.

[Wolberg, 1998] Wolberg, G. (1998). Image morphing: a survey. *The visual computer*, 14(8):360–372.

[Wolf and Shashua, 2003] Wolf, L. and Shashua, A. (2003). Learning over sets using kernel principal angles. *J. Mach. Learn. Res.*, 4:913–931.

[Yushkevich and Zhang, 2013] Yushkevich, P. A. and Zhang, H. G. (2013). Deformable modeling using a 3D boundary representation with quadratic constraints on the branching structure of the Blum skeleton. *Inf Process Med Imaging*, 23:280–291.

[Zaharia et al., 2012] Zaharia, M., Chowdhury, M., Das, T., Dave, A., Ma, J., McCauley, M., Franklin, M. J., Shenker, S., and Stoica, I. (2012). Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*, pages 2–2. USENIX Association.

[Zhang and Fletcher, 2013] Zhang, M. and Fletcher, P. T. (2013). Probabilistic principal geodesic analysis. In *NIPS*, pages 1178–1186.

[Zitova and Flusser, 2003] Zitova, B. and Flusser, J. (2003). Image registration methods: a survey. *Image and vision computing*, 21(11):977–1000.