UNIVERSITY OF GLASGOW

Statistical models for the evolution of facial curves

by

Irene Mariñas del Collado

A thesis submitted in partial fulfillment for the degree of Doctor of Philosophy

in the

College of Science and Engineering

School of Mathematics and Statistics

December 2017

Declaration of Authorship

I, Irene Mariñas del Collado, declare that this thesis, titled 'Statistical models for the evolution of facial curves', and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.

Part of the work in Chapters 4 and 5 has been presented at the 31^{st} and the 32^{nd} International Workshop on Statistical Modelling (IWSM) in Rennes (2016) and Groningen (2017) with the titles "Modelling the Shape of Emotions" and "Gaussian Process model for evolving 3D lip curves" and is included in the corresponding journals of conference proceedings.

Signed: _____

Date: _____

"If we knew what it was we were doing, it would not be called research, would it?"

Albert Einstein

Abstract

This thesis presents statistical models for the study of the evolution of shape. Particularly, it focuses on the evolution of facial curves. Evolution can be modelled viewing time as a linear, continuous variable, i.e., one curve that is gradually changing in a particular situation. Alternatively, it can play the role of evolutionary time, where branching points in the evolution can occur: ancestors diverging into multiple daughters. Two applications are studied: the evolution of the shape of the lips during the performance of an emotion (linear evolution) and the evolution of nose shape within and between ethnic groups (phylogenetic evolution).

The facial images available are in the form of three-dimensional point clouds which characterize each facial surface. Each face is represented by around 100,000 points. Anatomical curves are studied to provide a rich characterization of the full anatomical surface. The curves define the boundaries of morphological features of interest, using information of the facial surface curvature. Methods for the identification of facial three-dimensional curves are studied, and an algorithm to track four-dimensional curves (three spatial dimensions plus time) proposed.

The physical characterisation of facial expression involves a set of human facial movements. This thesis considers the shape of the lips as a unique facial feature to characterise emotions. Different approaches are proposed to model the lip shape and its change during the performance of an emotion. A first analysis of the evolving curves is performed using techniques of Procrustes analysis and a model based on B-splines. The thesis then moves to Gaussian Process (GP) models as an alternative approach. Models for k-dimensional curves and k-dimensional evolving curves are proposed. One direct application of the GP models is to study the grouping of different expressions of emotions in a space defined in terms of correlation parameters. To model the evolution of facial curves over many generations, the GP model for evolving k-dimensional curves is extended, using the phylogenetic covariance function, to allow for branching points in the evolution. A case study is conducted on data specially collected from different ethnic groups, where the phylogenetic model is applied to points on two curves defining the shape of the nose.

Acknowledgements

First and foremost I offer my profoundest gratitude and admiration to my supervisors. None of this work would have been possible without their help and motivation. Prof. Adrian Bowman took me under his wing as a Masters student, encouraged me to pursue a PhD degree and never stopped being supportive. Dr. Vincent Macaulay shared his valuable knowledge, and provided guidance, patience, continuous support and enthusiasm over the last few years. I feel truly privileged to have been their student.

I am also extremely grateful to the School of Mathematics and Statistics at the University of Glasgow for funding me throughout my research, as well as a period as a Graduate Teaching Assistant. The support and friendship of all the staff and students will be my enduring memory of the School. Special thanks go to my office mates for keeping me sane and having their drawers always ready 'in case of emergency'.

Thanks to all the friends that became my Glasgow family. I have genuinely felt at home here and that was thanks to the amazing people that I met. My time in Glasgow would have not been the same without each and everyone of them. And thanks to the friends from back home who stood by me no matter the distance, always at the other end of the phone when needed.

Last, but not least, thanks to my parents for their support and encouragement throughout my life to be the best I can be. And to my sister, for being the best personal cheerleader one could wish for and a constant in my ever-changing life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

When introduced to a new person, much of our initial, largely unconscious, perception is dependent on their facial appearance. The importance of the face in social interaction and social intelligence is widely recognized [Little et al., 2011] and faces have long been a source of interest to scientists in a wide range of disciplines. The face is usually the first type of visual information available to a perceiver and is visible continually through almost all types of interaction. Consequently, a fundamental question in social perception, and thus in understanding the social world of humans, is exactly what information a human face conveys [Jack and Schyns, 2015]. There are also issues of medical and biological interest. Recent advances in image-capture technologies have made available detailed three-dimensional facial images, which permit a quantitative investigation of these effects.

A research project based in the School of Mathematics and Statistics in the University of Glasgow, and involving both the Institute of Neuroscience and Psychology and the Dental School, is currently capturing data on facial shape in three-dimensions, with different medical and psychological questions in mind. For example, a long-standing clinical application is in the analysis of the facial shape of those who have undergone surgery for conditions such as cleft lip and palate [Bell et al., 2014; Millar et al., 2013]. More recently attention has shifted to quantifying the effects of surgical operations in adult faces.

The facial data are in the form of three-dimensional point clouds which characterize each facial surface. Each face is represented by around 100,000 points (See Figure 1.1). The protocol for collecting this type of data is explained in Chapter 2. Analysis of the data however raises very interesting questions about how to measure shape and shape-change over time. The first aim of this thesis lies in the estimation and analysis of facial curves, more specifically lip curves (the motivation for using lip curves is explained below).

FIGURE 1.1: Illustration of the cloud of three-dimensional points that form a facial surface. Different angles of view can be seen when viewed in digital form (using Adobe reader).

Statistical shape analysis is a geometric analysis of a set of shapes in which statistics are computed to describe geometrical properties and so to compare similar shapes, estimate population average shapes and the population shape variability

and so forth. The word 'shape' is used in everyday language most commonly to refer to the appearance of an object. A more formal definition is based on the intuition that shape is: "what is left when the differences which can be attributed to translations, rotations, and dilatations have been quotiented out"[Kendall, 1984]. For example, the shape of a human skull consists of all the geometrical properties invariant under 'pose'. A traditional starting point for the analysis of shape is to identify a finite number of landmarks. These are points on the anatomical surface which represent well-defined positions on the objects of interest, in this case, a human face. Methods of statistical analysis of landmarks are well developed, with an excellent description given by Dryden and Mardia [1998]. However, the expectation of this approach is that the number of landmarks will be small and they cannot therefore do justice to the very rich surface representations. The approach of identifying anatomical curves, rather than just points, the landmarks, is studied as an alternative with the aim of providing a richer characterization of the full anatomical surface [Bowman et al., 2015]. Curves can define the boundaries of anatomical features of interest, allowing the position of these to be identified and, if appropriate, extracted from the larger object for separate analysis.

In the literature, most of the approaches to identifying lips are based on detection of the lip boundaries in two-dimensional images, where the change in colour from skin to lip is the key information. However, due to the strong influence of the lighting conditions and shading in the camera images, this approach is not foolproof [Kakumanu et al., 2007]. Anatomically defined curves have the advantage of providing a much richer expression of shape. The key features of the face can be viewed as a set of ridges and valleys. For example, the mid-line of the lip is a valley, while the nose profile is a ridge. The key issue is to use information from the surface curvature to characterize the outline of the lips. This theory will be introduced in Chapter 3, together with the identification of the outline of the lips and the lip curves that represent the upper and lower lip boundaries. The data (processed from the raw 3d point cloud) is then a set of three-dimensional points that lie on the lip curves. In this thesis, each lip curve is represented by 24 points on the facial surface. Once these lip points are identified, the substantial literature on curve fitting can be employed. Lancaster and Šalkauskas [1986] summarised the foundations and major features of several basic methods for curve and surface fitting, from polynomial interpolation to splines, which is the first approach adopted in this thesis.

When studying facial curves, one cannot help but consider the changes in these. Facial curves change in many different contexts. They vary between and within people and they have changed over the evolution of modern humans. The statistical modelling of evolution is a topic of sizeable interest with a extended range of possibilities. In everyday language, the word evolution has two definitions attached to it, first, as a process of gradual development in a particular organism over a period of time and second, as a gradual change in the characteristics of a population over successive generations, accounting for the origin of existing species from ancestors unlike them [Collins Dictionary, 2017].

Statistically, the first definition can be interpreted as modelling time as linear and directional. The statistical modelling of an evolving curve, and moreover a three-dimensional curve that changes over time, raises many interesting issues. Many authors have considered the motion of curves in the plane through changes in curvature and direction of the normal [Kimmel et al., 1997]. The study of three-dimensional curves that change over time usually involves the reduction to low-dimensional summaries or discrete characters. In this thesis, Gaussian process models are proposed for tracking the evolution of k-dimensional curves, expressed on a continuous and multivariate scale. An application of this, in this thesis, is to the change in the shape of facial curves during the performance of an emotion.

To model the evolution of facial curves over many generations, one has to account for branching points in the evolution. In this scenario, the curves evolve along a phylogenetic tree, allowing time to play the role of 'evolutionary time', rather than a continuous variable that can be modelled linearly. Time is then not just a coordinate, but also indicates a branch of the phylogeny (an introduction to phylogenies is given below). Both classical and Bayesian approaches to the resulting inference problems are possible in this scenario. Methods of modelling the evolution of k-dimensional curves evolving along the branches of phylogenetic trees are presented as a complement to methods that use genetic information, on which extensive research has been carried out [Page and Holmes, 1998]. These models are also based on Gaussian processes. The application chosen here to illustrate the modelling of the phylogenetic evolution is to study the evolution of nose shape within and between ethnic groups.

1.2 Linear evolution of facial curves: tracking of emotions

Central to all human interaction is the mutual understanding of emotions, achieved primarily through a set of biologically rooted social signals evolved for this purpose: facial expressions of emotion. Facial expressions are a means of communication that is more rapid than language, from which people can quickly infer the state of mind of their companions. Facial expressions allow a group to easily understand the opinions and attitudes of others, and thus constitute a powerful tool in social coordination. Since Darwin's seminal work [Darwin, 1872], the universality of facial expressions of emotion has remained one of the longest standing debates in the biological and social sciences. Specifically, the universality hypothesis (which can be traced back to Darwin's *The Expression of the Emotions in Man and Animals* [Darwin, 1872]) proposes that six basic internal human emotions (i.e., happiness, surprise, fear, disgust, anger and sadness) are expressed using the same facial movements across all cultures, supporting universal recognition. Ekman and Friesen [1971] reported that these six emotions are readily recognized across very different cultures. The discernment of how emotions are expressed has a big impact in maxillofacial surgery where it is important to measure the success of surgical operations in terms of the ability of the patient to produce a 'normal' facial expression [Kau et al., 2007], as well as in the case of people with psychological problems as schizophrenia [Gaebel and Wölwer, 1992]. Emotion recognition is currently gaining importance also for its increasing scope of applications in human-computer interactive systems [Fragopanagos and Taylor, 2005].

The physical characterisation of facial expression involves a set of human facial movements, including raising or lowering of the eyebrows, cheek fluctuation, opening and closing of the eyes, as well as head position. While several modalities of emotion recognition, including facial expression, voice, gesture and posture have been studied, there exists hardly any significant work on emotion recognition by a single facial feature [Halder et al., 2011]. This thesis considers the shape of the lips as a unique facial feature to characterise emotions. The lips are a facial component which shows great variability of changes. While changes in the brow or in the cheek can be observed too, emotions are more easily recognised by lip features: happiness is mostly shown by a smile, surprise by a sudden opening on

the mouth, etc. I.e., by the shape of the lip curves directly. To be able to recognize facial emotions in the eyes, however, one needs to detect whether the eyes are open or closed, the degree of eye opening and the location and radius of the iris [Tian et al., 2001]. When observing the Emotional Facial Action Coding System (See Section 2.2), movements in the eyes involve both the eye lids and the brow (different facial curves), while movement in the lips requires only the lip curves. A particular aim of the thesis focuses on the changes in lip shape associated with the expression of emotions: how to estimate the evolving lip curves. This applies to the evolution over short time scales and is a four-dimensional problem, three spatial dimensions plus time.

1.2.1 Models for lip curves

As mentioned above, to exemplify the evolution of curves along a linear time dimension, this thesis focuses on the communication of emotions through facial expression. For this purpose, different approaches are proposed to model lip shape and its change during the performance of an emotion.

B-Splines

In Chapter 3 models for the outline of the mouth are constructed using B-splines to represent the lip as a curve through a set of coefficients of basis functions instead of individual points.

Spline functions have diverse applications. They have been used to provide solutions to mathematical problems in interpolation, data and function approximation, ordinary and partial differential equations, and integral equations. Splines have also been employed in many scientific and engineering applications [Cox, 1982]. B-splines have been used as a convenient set of basis functions for problems of interpolation and smoothing by means of spline functions with fixed knots since at least the 1960s, with a great deal of work done on their numerical evaluation by Cox [1972].

B-splines are attractive for non-parametric modelling since their computation is relatively inexpensive. They have been used before in the literature for curve fitting and, more specifically, for lip tracking [Hoch and Girod, 1994; Piegl and Tiller,

1987; Sánchez et al., 1997]. B-splines are constructed from polynomial pieces, joined at certain locations, known as knots. Once the knots are given, it is easy to compute the B-splines, recursively, for any desired degree of the polynomial. The theory for B-splines is presented in Section 3.3.1. The advantage of the use of B-splines is that they can be used both to model the data and also to smooth the lip points both spatially and temporally. B-splines are applied to the series of coordinate values for the 24 three-dimensional points chosen to identify a lip path. Furthermore, an approach to smooth the B-splines coefficients at each particular time point over time is proposed.

Gaussian Processes

In many real-world applications the practical problem consists of 'learning' an underlying function that can associate some observed inputs (e.g., characteristics of the three-dimensional points, such as spatial and temporal position in the emotion sequence), with the corresponding outputs (e.g., the coordinates values of the three-dimensional points). The chosen approach is usually to train parametric models, i.e., assume that some (ideally small number of) parameters can adequately describe the underlying function. The main limitation of parametric models is the original assumption regarding a finite set of parameters. A more flexible approach is provided by non-parametric models, which assume that the distribution of the data can only be defined in terms of an infinite dimensional set of parameters. This infinite set can be naturally modelled within the framework of Gaussian processes [Eleftheriadis, 2016]. Gaussian processes are mathematically equivalent to many well-known models, including Bayesian linear models, spline models and large neural networks. They are closely related to other methods, such as vector machines [Rasmussen and Williams, 2006]. Gaussian processes are a generalization of the Gaussian probability distribution. They can be understood as regression models or thought of as defining a distribution over functions, making inference directly in the space of functions.

A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution. The random variables will be chosen to represent the values of the three-dimensional points that define a lip curve, as a function of their location along the curve arc-length, for lip curves, or as a function of the

spatial location plus the time point in the sequence of pictures representing an emotion. One of the advantages in the use of Gaussian Processes to map directly the spatio-temporal characteristics to the output facial features is that Gaussian processes allows one to specify various types of covariance functions that can capture complex data structures. This is important to be able to model the interactions among the different coordinates, as well as the space and time location of the points.

The first, most simple, model to be proposed is a Gaussian process model for the one-dimensional curve created by the values of just one coordinate from a lip curve, in terms of the arc-length. The model is then extended to three-dimensional curves. For the latter model, the outputs (values of the coordinates) are described not only as a function of a spatial component, the arc-length, but also the relationship between the three coordinates. These models represents lip curves in a resting position, without expression of any emotion. The Gaussian process models for k-dimensional curves, together with a larger account of the theory behind Gaussian processes, are presented in Chapter 4.

The challenge then considered is to enrich the data type modelled to capture more of the information in the three-dimensional shape of an evolving lip curve (whether it is one- or three-dimensional). A key issue is to capture adequately the covariance structure in space and time. Models for evolution are presented in Chapter 5. The aim is to learn how a three-dimensional curve evolves over time and, more specifically, how the shape of the lip curves changes during the performance of different emotions. One direct application of this model is to study the grouping of different expressions of emotions in terms of correlation parameters. One could also view this problem as a prototype of the next aim: a phylogenetic setting (introduced below) where branching points in the evolution can occur.

1.3 Phylogenetic evolution of facial curves: changes in nose morphology

The final part of the thesis focuses on the evolution of facial curves not through the performance of an emotion, but through evolutionary time from our ancestors, along the branches of a phylogenetic tree.

Phylogenies, or evolutionary trees, are the basic structures necessary to think clearly about differences between species, and to analyse those differences statistically. The use of both metaphors (figures of speech) and models (diagrams) comparing evolution to branching trees go back at least to Charles Darwin's *On the Origin of Species* [Darwin, 1859], but the major advances from a statistical, computational and algorithmic point of view have been mostly made in the past 50 years. Felsenstein well summarised in [Felsenstein, 2004] the major advances that had been achieved in the course of the previous four decades.

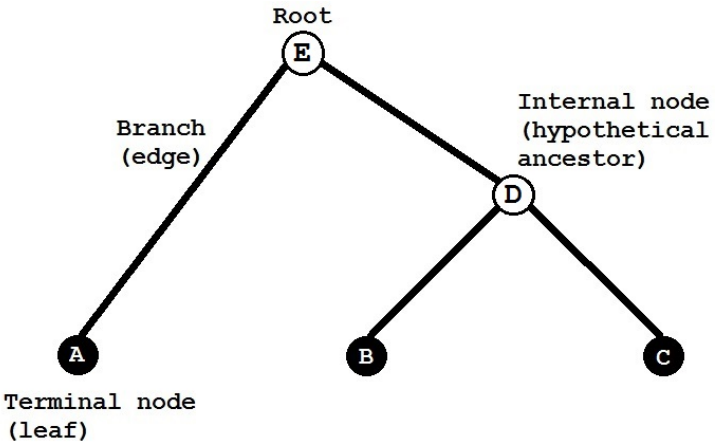


FIGURE 1.2: A simple tree and associated terms.

A tree can be defined as a mathematical structure which is used to model the evolutionary history of a group of organisms, DNA sequences, or some other kind of data. The actual pattern of historical relationships is the phylogeny or evolutionary tree for which estimation is the aim [Page and Holmes, 1998]. A tree consists of nodes connected by branches (edges) (Figure 1.2). In most cases, the terminal nodes or leaves are the sequences or organisms for which data are available, usually at the present time. Internal nodes represent the hypothetical ancestors, and the ancestor of all the sequences that comprise the tree is the root of the tree.

The nodes and branches of a tree may have various kinds of information associated with them and one of the issues of interest is to reconstruct the (missing) data at each hypothetical ancestor (phylogeny reconstruction). Most methods also try to estimate the amount of change that takes place between each pair of nodes, which is represented as the branch/edge length. These types of trees with branch lengths are sometimes referred to as weighted trees.

The number of adjacent branches possessed by an internal node is referred to as the node's degree. If a node has a degree greater than three (one ancestor and

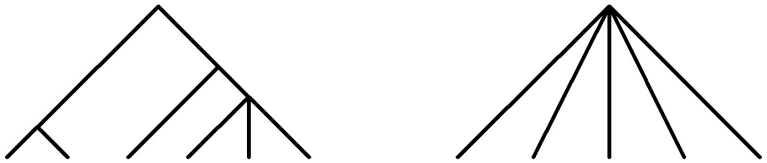


FIGURE 1.3: Examples of: partially resolved tree (left) and star tree (right).

more than two immediate descendants), then the node is said to be a polytomy. A fully resolved tree is one that has no polytomies. For example, the tree in Figure 1.2 is fully resolved, while the ones shown in Figure 1.3 are not.

A widely-used shorthand notation (Newick format [Huson et al., 2010]) for trees is to label the terminal nodes and use nested parentheses: each internal node is represented by a pair of parentheses that enclose all descendants of that node. In this notation, the tree from Figure 1.2 would be written as $(A, (B, C))$.

The trees in Figure 1.2 and Figure 1.3 are cladograms, which simply show relative recency of common ancestry. For example, given the three sequences A, B and C, the cladogram in Figure 1.2 represents that B and C share a common ancestor more recently than either does with A. This kind of tree, in which there is also a node identified as the root from which ultimately all other nodes descend, is called a rooted tree.

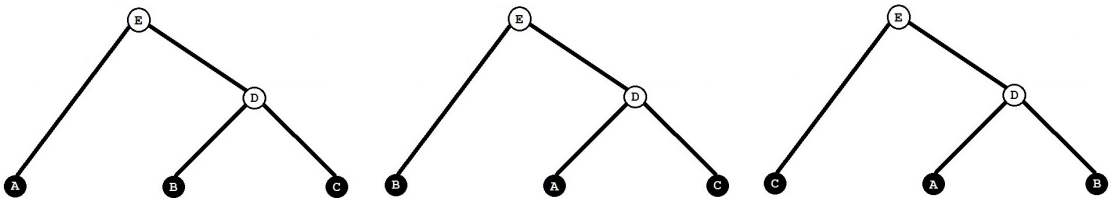


FIGURE 1.4: Possible topologies for a rooted tree with 3 leaves.

The particular branching pattern of a tree is called its topology. It represents all of the evolutionary relationships, but does not represent the distance or time between nodes. The topology is the graph and the leaf labels. The tree in Figure 1.2 has its nodes labelled from A to E (Terminal nodes A, B and C; internal node D and root E), then there are three possible topologies for it, illustrated in Figure 1.4. The first possible topology is the one mentioned above, where B and C share a common ancestor more recently than either does with A. The topology can be written as the series of immediate ancestors for each node (except the root). Therefore, the first topology would be written as (E, D, D, E), corresponding to

the ancestors of (A, B, C, D). The other two possibles topologies, (Figure 1.4, middle and right) are: (D, E, D, E), and (D, D, E, E), respectively.

Where there are models of evolution for the data, standard statistical methods can be used to make estimates of the phylogeny. Perhaps the most standard framework of all is to use maximum likelihood [Felsenstein, 2004]. Depending on the different types of data that can be observed at the leaves of the tree, a likelihood function can be associated with the tree. This likelihood function can be maximised with respect to the tree topology, the branch lengths, and other model parameters. Phylogenetic maximum likelihood algorithms proceed by iterating between two major algorithmic steps:

1. for a given tree topology, find optimal branch lengths (i.e., the branch lengths that make the observed data most likely) and the rest of the model parameters, which characterise the evolution process;
2. find the tree topology that maximizes the likelihood, given the branch lengths and evolution process parameters.

Because there exist finitely many tree topologies, it is possible, in principle, to optimize branch lengths and model parameters for every possible tree topology and choose the tree that has the highest likelihood value as the maximum likelihood tree. However this approach is only viable for trees with a small number of leaves, where the number of topologies is limited. For trees with a larger number of leaves the number of possible topologies increases super-exponentially so naïvely searching over this tree space is computationally infeasible. Various heuristics are used to find the topology that has the highest likelihood. All these methods use local modifications of the previously visited tree topologies to find a new tree with a higher likelihood. For example, common methods traverse the tree topology space greedily by comparing the likelihood values between modified trees and by choosing the topology that increases the likelihood the most; the procedure will end if there are no trees that increase the likelihood [Dhar and Minin, 2015].

Chapter 6 builds on the idea of developing statistical methods through which shape information on organisms can be used to construct a phylogenetic tree and to learn how shape evolves. Genetic information is most commonly used for this purpose, based on DNA sequences observed in the organisms. The approach proposed here is to use shape information, i.e., facial curves, to study relationships

between different ethnic groups and their (our) ancestors. The use of shape information, expressed on a continuous and multivariate scale, raises a number of very interesting issues from a methodological perspective. The main challenge is to model the data without sacrificing information, as traditionally happens, for example, in distance-based methods [Singh, 2015], while maintaining computational feasibility.

The general idea of distance-based methods (also known as distance matrix methods) is: calculate a measure of the distance between each pair of leaves (of DNA sequence data, most commonly) and then find a tree that predicts the observed set of distances as closely as possible. This destroys information from higher-order (more than pairwise) combinations of character states[1], reducing the data matrix to a simple table of pairwise distances [Felsenstein, 2004]. It is important to understand, however, that in distance matrix methods branch lengths are not simply a function of time. They reflect expected amounts of evolution in different branches of the tree. Two branches may reflect the same elapsed time, but they can have different expected amounts of evolution. This allows different branches to have different rates of evolution. Distance methods have been widely studied and are considered the easiest phylogeny methods to program. They produce reasonable estimates of phylogenies. Their drawback comes from their phenetic[2] approach, as they attempt to classify organisms based on overall similarity, and this does not aim to reflect evolutionary descent. Moreover, the relationship between the individual characters and the tree is lost in the process of reducing characters to distances and the strength of the technique is dependent on accuracy of the distance estimate, and thus dependent on the model used to obtain the distance matrix. They are regarded as more sensitive to systematic errors than maximum likelihood methods.

Facial curves have both of the following features:

1. they are functions rather than single numbers or vectors;

2. they can be correlated owing to phylogenetic relationships when considering the evolution of facial morphology within and between families or ethnic groups.

[1] In genetics, a character is a heritable trait possessed by an organism. Characters are usually described in terms of their states, e.g.: "hair present" vs. "hair absent", where "hair" is the character, and "present" and "absent" are its states.

[2] In biology, phenetics, also known as taximetrics, is an attempt to classify organisms based on overall similarity, usually in morphology or other observable traits, regardless of their phylogeny or evolutionary relation.

Jones and Moriarty Jones and Moriarty [2013] presented a flexible statistical model for such data, by combining assumptions from phylogenetics with Gaussian processes. Their approach generalizes the Brownian motion and Ornstein-Uhlenbeck models of continuous-time character evolution from quantitative genetics. A summary of their work is presented in Chapter 6, together with how the theory from Chapters 4 and 5 can be then extended to the phylogenetic setting by adapting the evolution correlation between curves.

During the last few decades several authors have tried to clarify the anthropological aspects of the shape of the human nose [Mladina et al., 2009]. Modern humans possess a unique projecting, external nose whose basic structure is reflected in a series of skeletal features [Franciscus and Trinkaus, 1988]. It seems that the erectile posture of Homo sapiens caused remarkable morphological changes to skull shape. Erectile posture enabled man to better see around, to more easily recognise potential dangers, enemies, sources of food, etc. The olfactory function of the nose, so important for quadrupeds, started therefore slowly to diminish over time. Its respiratory function became the leading role in man's nose physiology.

Nasal configuration in modern humans seems to be associated with the internal nasal cavity broadness and nasal bridge elevation which happened because of newly developed physiological needs. For instance, the large size of the nasal cavity in Neanderthals represents an anatomical accommodation to the cold environmental conditions and low humidity of that time. The increased surface of the internal nose as a specific, natural air-conditioning machine enabled better regulation of the temperature and humidity of the inspired air [Mladina et al., 2009].

Chapter 6 uses as a case study the evolution of nose curves across three ethnic groups. The nose curves used are:

- the mid-line nasal profile: ridge points from the nasal root along the dorsum of the nose and the columella.
- the nasal bridge: which outlines the width of the nose from one alar facial groove to the other.

Figure 1.5 shows points on the two curves chosen to characterise the nose shape. The points are shown over a facial mesh rotated at different angles for a better illustration of the three-dimensional curves.

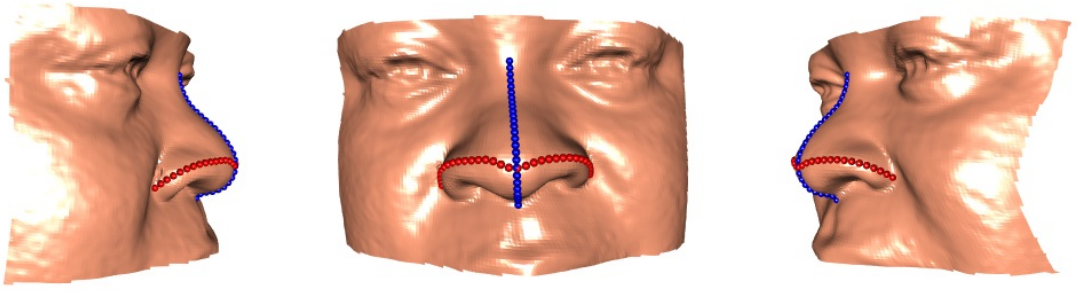


FIGURE 1.5: Nose curves: mid-line nasal profile (blue) and nasal bridge (red).

1.4 Thesis structure

To summarise, this thesis explores and develops a range of statistical models for the study of the evolution of different facial curves. The concept of evolution is expressed in two ways. The chapters are structured as follows:

Chapter 2 presents a brief explanation of the methods for capturing 3D facial surfaces and how the emotions are recorded.

Chapter 3 comprises the study related to shape analysis and the estimation of anatomical curves from a three-dimensional facial surface. A summary of the theory used for the study of surface shape is presented, followed by the statistical tools developed to help identify the outline of the lips. A first analysis of the evolving curves is performed using techniques of Procrustes analysis and a model based on B-splines proposed.

Chapter 4 introduces a Gaussian process model as an alternative method for modelling the data of the lip curves. It addresses the problem of estimating a ridge curve embedded in a three-dimensional surface, with no evolution involved. The theory for Gaussian processes is introduced with the study of one-dimensional curves and extended to the model for three-dimensional curves (lip curves). In Chapter 5, the models proposed in Chapter 4 are extended to capture adequately the covariance structure in space and time.

These chapters study the evolution understood in the first definition, namely, the evolution of the lip curves during the performance of an emotion.

Chapter 6 refers to evolution as a process of gradual change that takes place over many generations, during which species, or sub-populations within species,

change some of their physical characteristics. The notion of shape evolving in time is extended to the phylogenetic setting, where branching points in the evolution can occur: ancestors diverging into multiple daughters. A small case study is conducted, where the phylogenetic model is applied to compare and classify facial characteristics (nose curves) among different ethnic groups, Africans, East Asian and Europeans.

The conclusions are given in Chapter 7, which summarizes and discusses the main findings of the thesis. Suggestions for further lines of investigation are explored.

Chapter 2

Facial shape imaging

2.1 Methods for capturing three-dimensional facial surfaces

The data analysed in this thesis is collected from a stereo-camera system which is able to construct a three-dimensional model of the surface of a face. The ©*Di3D* [Dimensional Imaging Ltd, 2017] 3D facial image capture system is used. This is a passive stereo-photogrammetry system for the creation of high-resolution, accurate, full colour 3D facial surface images. The data that is retrieved from the image capture is a collection of multiple polygonal meshes and accompanying colour maps (See Figure 1.1). The system consists in a set of four cameras (Figure 2.1). The technique of stereo-photogrammetry cleverly uses the mismatch between images from the adjacent cameras to reconstruct the three-dimensional surface shape of the face. Stereo-photogrammetry involves identifying common points on each image and constructing a line of sight (or ray) from the camera location to the point on the object. It is the intersection of these rays (triangulation) that determines the three-dimensional location of the point. This is, however, done by the proprietary software of [Dimensional Imaging Ltd, 2017].

FIGURE 2.1: ©*Di3D* Imaging Camera System [McNeil, 2012]

The data which these images provide offer a very helpful route to quantitative analysis.

Multiple pictures are usually taken from each subject and the best selected. There are potential issues/loss of detail with the data capture. First, if the system is not properly calibrated and in the correct position, i.e. the participant's eyes lined up at the centre of each camera view and the centre of the camera system at a distance of roughly 95 cm from the participant's cheek (See Figure 2.2), the software may fail to build the proper three-dimensional surface from the four component pictures. This results in unrealistic facial surfaces where parts of the face 'collapse'.

FIGURE 2.2: ©*Di3D* Imaging Camera System in use.

The second problem is the so-called 'orange peel effect'. This occurs when an individual's skin is thin. In infants and older individuals, light can easily penetrate the skin and bounce off, causing subsurface artefacts. Young adult individuals tend to have thicker skin, and hence a smoother surface is produced by the system. The

area that is most difficult to capture is the ear region. With the angle of the four cameras, the ears tend to be 'smeared' backwards, not being able to capture the full dimension of the head [Vittert, 2015]. Problems arise too with facial hair such as beards and fringes. All participants are asked to pull their hair away from their face using hair clips or hair bands provided. Prior to their arrival participants are also asked to refrain from using heavy make-up and to shave any facial hair.

The process of capturing the data is based upon the protocols of the Face3D project, which was supported by the Wellcome Trust. In order to address the specific research questions of the Face3D project, as well as adhere to the reasonable concept of comparing facial shape amongst specific ethnic groups, some biological and medical information is gathered through a series of questions. These are:

- Date of Birth.
- Sex.
- Have you ever had any serious facial injury? (Y/N)
- Have you had, or are you seeking, any serious facial surgery, cosmetic or otherwise? (Y/N)
- Is there any history of cleft lip and/or palate in your family? (Y/N)
- From the following, please choose the number associated with the ethnic background (*based on the Scottish Census*) of you and the following family members: mother, maternal grandmother, maternal grandfather, father, paternal grandmother, paternal grandfather. :

    1. White Scottish
    2. Black Scottish
    3. White British
    4. Black British
    5. White Irish
    6. White Gypsy/Traveller
    7. White Polish
    8. White European (Please specify country)
    9. Arab (Please specify country)
    10. Caribbean
    11. African (Please specify country)
    12. Asian Chinese
    13. Asian Bangladeshi
    14. Asian Indian
    15. Asian Pakistani
    16. Other (Please Specify)

These questions are asked in a one-on-one environment to protect each subject's privacy and then the image capture process is conducted. Participants are also

asked to sign the statement that they understand the data collected from them in this study can be used for the purposes of the research involved, and that the facial image, together with information on age, sex and ethnic background may be made available freely to others in an anonymous manner, thereafter.

Each facial image captured from the camera system is loaded into a software package called ©Landmark, developed by the Institute for Data Analysis and Visualisation (IDAV), which allows the facial image to be visualised in three-dimensions and manually landmarked. The virtual three-dimensional image can be rotated in any direction and specific areas of interest magnified so that the facial surface can be viewed from every angle (Figure 2.3). When identifying anatomical landmarks, the ability to display the facial shape in this manner is highly advantageous, especially when aiming to identify points of maximum curvature.

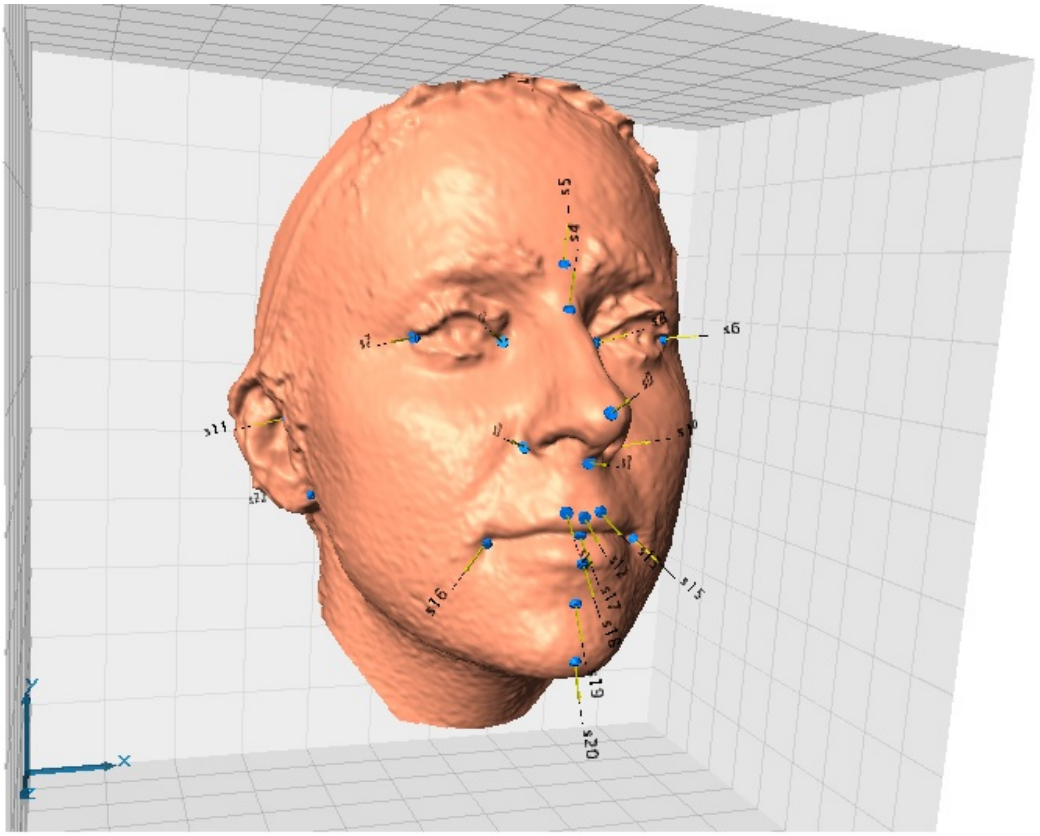


FIGURE 2.3: Facial mesh and anatomical landmarks [McNeil, 2012]

2.2 Recording of emotions

To record the expressions, a large number of pictures of the person producing the emotion are taken with the stereo-photogrammetry camera system. For this research, there are available sequences from 60 images, in the case of the expression of *disgust*, to sequences of about 180 images, in the case of *happiness*. Data on emotions was collected by Oliver Garrod and colleagues from the School of Psychology at the University of Glasgow from a 25 year old actress performing each expression between three and five times. The actress was given no stimulation to represent the emotions; she based the expressions on the Ekman prototypes, explained below.

2.2.1 Facial Action Coding System

The Facial Action Coding System (FACS) is a system to taxonomise human facial movements by their appearance on the face, based on a system originally developed by Swedish anatomist Carl-Herman Hjortsj. Charles Darwin theorized that emotions were biologically determined and universal to human culture in his *The Expression of the Emotions in Man and Animals* published in 1872 [Darwin, 1872]. However, the more popularized belief during the 1950s was that facial expressions and their meanings were culturally determined through behavioural learning processes. Through a series of studies, Ekman found a high agreement across members of diverse Western and Eastern literate cultures on selecting emotional labels that fit facial expressions. Expressions he found to be universal included those indicating anger, disgust, fear, happiness, sadness and surprise. Ekman supplemented these in the 1990s with 11 additional emotions (amusement, contempt, contentment, embarrassment, excitement, guilt, pride in achievement, relief, satisfaction, sensory pleasure, and shame). The present thesis focuses only in the six universal emotions.

The FACS is a comprehensive, anatomically based system for measuring all visually discernible facial movement. FACS describes all visually distinguishable facial activity on the basis of unique actions units (AUs), as well as several categories of head and eye positions and movements [Ekman and Rosenberg, 1997]. Table 2.1 shows the descriptions of the most used AUs, and some examples are shown in Figure 2.4.

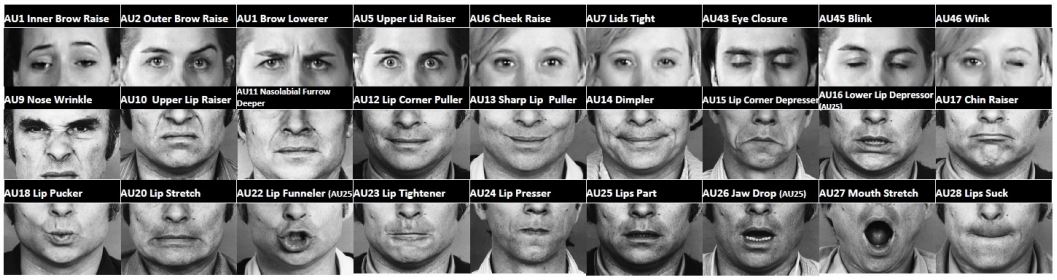


FIGURE 2.4: Illustration for facial Action Units (AUs) [Rudovic, 2013]

AU number	FACS name	AU number	FACS name
1	Inner Brow Raiser	27	Mouth Stretch
2	Outer Brow Raiser	28	Lip Suck
4	Brow Lowerer	19	Tongue out
5	Upper Lid Raiser	21	Neck Tightener
6	Cheek Raiser	29	Jaw Thrust
7	Lid Tightener	30	Jaw Sideways
9	Nose Weinkler	31	Jaw Clencher
10	Upper Lip Raiser	32	Lip Bite
11	Nasolabial Fold Deepener	33	Cheek Blow
12	Lip Corner Puller	34	Cheek Puff
13	Cheek Puffer	35	Cheek Suck
14	Dimpler	36	Tongue Bulge
15	Lip Corner Depressor	37	Lip Wipe
16	Lower Lip Depressor	38	Nostril Dilator
17	Chin Raiser	39	Nostril Compressor
18	Lip Puckerer	41	Lid Droop
20	Lip Stretcher	42	Slit
22	Lip Funnerler	43	Eyes Closed
23	Lip Tightener	44	Squint
24	Lip Pressor	45	Blink
25	Lips Part	46	Wink
26	Jax Drop		

TABLE 2.1: Descriptions of AUs in the FACS.

The EMFACS (Emotional Facial Action Coding System) [Ekman et al., 1983] considers only emotion-related facial actions. This can range from real-life observation of a person interacting in a group to videotaped interactions in which facial expressions are documented under laboratory conditions upon experimental triggering of a specific emotion. In other words, FACS identify anatomically distinct muscle movements (e.g., lip corner puller) labelled action units (AUs), but they are not asked to make inferences about emotional state (e.g., happiness, sadness). EMFACS is an abbreviated version of FACS that assesses only those muscle movements believed to be associated with emotional expressions. Table 2.2 shows the definition of the emotions under study in this thesis.

Emotion	AUs	Emotion	AUs
Anger	4+5+7+23	Happiness	6+12
Disgust	9+15+16	Sadness	1+4+15
Fear	1+2+4+5+7+20+26	Surprise	1+2+5+26

TABLE 2.2: Descriptions of AUs in the EMFACS.

For clarification, FACS/EMFACS creates an index of facial expressions, but does not actually provide any bio-mechanical information about the degree of muscle activation.

Chapter 3

Estimation and analysis of lip curves

3.1 Introduction and background

The study of human lip curves is vital as part of facial analysis as lips form an important part of facial expression. The present chapter focuses on the shape of the lips in a three-dimensional facial image and their variation over the expression of different emotions such as fear, anger, happiness, etc. (see Figure 3.1). Methods for identifying the outline of the mouth are investigated, using ideas of shape index and curvature.

The approach investigated in this thesis is motivated by the three-dimensional setting and it is based on geometric characteristics of surface shape. In the case of a closed mouth, the lip boundaries are characterized by three well-defined curves: upper lip, lower lip and the valley between both lips (mid-line lip). The lower and upper boundaries lie on the ridges formed by the different orientations of lip and skin tissue. The meeting of the lips is characterized by a sharp change in the opposite direction. The aim is to identify first the positions where the upper and lower lip meet, i.e., to get information on the surface curvature to identify first the two corners of the lips, and, from them, to identify a series of points in the curves along the lips. Once the paths in the lips are identified, B-splines are used to describe the curve. The valley between both lips will be identified in the first image of the sequence to help to separate the upper and lower lip; the lips in the

FIGURE 3.1: Sample of pictures from the sequences (across rows) of sadness, happiness, surprise, fear and disgust, respectively.

subsequent images will be estimated with the assistance of the information of their coordinates in the previous image.

Before going into a deeper explanation of the process, it is helpful to present a summary of the theory used: the concept of 'landmarks' and the definition of shape index to study the surface shape.

3.1.1 Shape analysis

3.1.1.1 Landmarks

A **landmark** is a point of correspondence on each object that matches between and within populations. Shape analysis has a wide variety of applications, including archaeology, biology, chemistry, geography, image analysis and medicine. Dryden and Mardia [1998] gave a classification of three types of landmarks that can be commonly identified: biological, mathematical and pseudo-landmarks.

- *Biological landmarks*, also known as 'anatomical': are assigned by an expert to locations that corresponds between organisms in some biological meaningful way, e.g., the corner of an eye. They designate parts (said to be homologous) of an organism that correspond in terms of biological function.
- *Mathematical landmarks*: points located on an object according to some mathematical or geometrical property of the figure, e.g., an extreme point. They are particularly useful in automatic recognition and analysis.
- *Pseudo-landmarks*: constructed points on an organism, located either around the outline or in between anatomical or mathematical landmarks, e.g., many equally spaced points might be placed on the outline of a bone. Continuous curves can be approximated by a large number of pseudo-landmarks along the curve.

Landmarks can be alternatively divided in the three following types [Bookstein, 1997]:

- *Type I landmarks*, which occur at the joints of tissues/bones.
- *Type II landmarks*, which are defined by local properties such as maximum curvatures.
- *Type III landmarks*, which take place at extremal points or constructed landmarks, such as maximal diameters and centroids.

Biological landmarks are usually of type I or II, whilst mathematical ones are of type II or III in general. Pseudo-landmarks are commonly taken as equi-spaced points along outlines between pairs of landmarks of type I or II; this is what makes them type III landmarks. Point configurations can be labelled or unlabelled. Labelled configurations have each landmark assigned a name or number and this corresponds in some meaningful way to the landmark with the same name or number on another specimen.

3.1.1.2 Curvature and shape index

Curvature is a two-dimensional property of a three-dimensional space curve. More commonly curvature is a scalar quantity, but one may also define a curvature vector

that takes into account the direction of the bend as well as its sharpness. Curvature describes how bent a curved line is at a particular point, i.e., how much the curve deviates from a straight line at this point. For a particular point on a curve, its curvature is defined as the rate of change of direction in the curve [Roberts, 2001]. Mathematically, for a particular point P on a curve, the curvature can be defined in terms of the radius of the osculating circle. Given any curve C and a point P on it, there is a unique circle or line which most closely approximates the curve near P. This is the osculating circle. This circle possesses a common tangent with the curve at P. The radius of this circle is defined as the radius of curvature, R. A circle is of course bent by the same amount all the way around its circumference and therefore has a constant curvature, K. Curvature is simply the reciprocal of the radius of curvature: $K = 1/R$. From this it can be seen that the smaller the radius of curvature, the more bent the curve is and the larger the curvature. When considering the limiting case where R is infinite, then locally the circle would approximate a straight line and hence have zero curvature.

The two-dimensional concept of curvature can easily be extended into three dimensions. A curve can be constructed by mathematically cutting the surface with a plane. The intersection the plane makes with the surface describes a curve from which the curvature can be calculated at any point along the curve. Out of the infinite number of curvatures that can be extracted, it is found [Roberts, 2001] that the most useful subset of curvatures are the *normal curvatures*, defined by planes which are orthogonal to the surface.

From the infinite number of *normal curvatures* that pass through a particular point on a surface, there exists one curve that defines the largest absolute curvature, and its perpendicular defines the smallest. These two surface attributes are called the *principal curvatures*, denoted κ_1 (maximum value) and κ_2 (minimum value).

In differential geometry, the two *principal curvatures* at a given point on a surface provide the key characteristics. They measure how the surface bends by different amounts in different directions at that point. Parabolic points are characterised by the vanishing of one principal curvature and occur generically on curves, the so called 'parabolic lines' [Koenderink and van Doorn, 1992]. Such curves are smooth loops that never intersect. Parabolic curves strictly separate regions of convexities ($\kappa_{1,2} < 0$), concavities ($\kappa_{1,2} > 0$), and saddle-like points (κ_1 and κ_2 of different sign). Two classical shape measures combine the *principal curvatures*:

- *Gaussian curvature*, which equals the product of the *principal curvatures*. It is also referred to as the 'total curvature'. If the product is a positive number, the surface is a dome, while a negative result would correspond to a saddle. If the product equals zero, it is a parabolic point. Gaussian curvature is named after Carl Friedrich Gauss, who introduced the measure in his *'Theorema egregium'* [Pressley, 2001].
- *Mean curvature*, which numerically equals the arithmetic average of the two *principal curvatures*.

A shape index was suggested [Koenderink and van Doorn, 1992] as a means of combining the two principal curvatures, as an alternative technique to the classical surface measures, since neither by themselves are considered to capture the intuitive notion of ''local shape'' very well. The shape index combines the principal curvatures into a single shape indicator rather than having to work with a pair of numbers. It is a number in the range $[-1, +1]$ and it is scale invariant. The shape index is defined as:

$$S = \frac{2}{\pi} \tan^{-1} \left(\frac{\kappa_2 + \kappa_1}{\kappa_2 - \kappa_1} \right) \qquad (\kappa_1 \geq \kappa_2). \tag{3.1}$$

This representation has many intuitively natural properties. For instance, the convexities and concavities find their places on opposite sides of the shape scale. These basic shapes are separated by those shapes which are neither convex nor concave, namely saddle-like objects. The range $[-1, +1]$ is interpreted as follows:

Range of S	Interpretation
$S = \pm 1$	Extremes, Umbilical points: the outside and inside of a spherical surface.
$S = \pm 0.5$	Cylindrical shapes: ridge ($S = 0.5$) and rut ($S = -0.5$).
$-1 < S < -0.5$	Concavities: concave ruts or trough-shapes.
$0.5 < S < 1$	Convexities: convex ridges or dome-shapes.
$-0.5 < S < 0.5$	Saddle-like shapes: saddle ruts and ridges, with the symmetrical saddle at $S = 0$.

Table 3.1: Range and interpretation of Shape Index.

The shape index (S) is a function of the ratio of κ_1 and κ_2, such that S describes the type, but not the size of the curvature. The curvedness at point p is defined

as [Vittert, 2015]:

$$K(p) = \sqrt{\frac{\kappa_1^2(p) + \kappa_2^2(p)}{2}}, \tag{3.2}$$

where κ_1 and κ_2 are the principal curvatures at the fixed point p. Gaussian curvature is used to describe the strength of curvature at any point on the surface.

Moreover, the shape index can be mapped onto an intuitively natural colour scale. [Koenderink and van Doorn, 1992] proposed the design summarised in Table 3.2 and illustrated in Figure 3.2.

Mnemonic	Index range	Colour
Spherical cup	$S \in [-1, -\frac{7}{8})$	Green
Trough	$S \in [-\frac{7}{8}, -\frac{5}{8})$	Cyan
Rut	$S \in [-\frac{5}{8}, -\frac{3}{8})$	Blue
Saddle rut	$S \in [-\frac{3}{8}, -\frac{1}{8})$	Pale Blue
Saddle	$S \in [-\frac{1}{8}, +\frac{1}{8})$	White
Saddle ridge	$S \in [+\frac{1}{8}, +\frac{3}{8})$	Pale yellow
Ridge	$S \in [+\frac{3}{8}, +\frac{5}{8})$	Yellow
Dome	$S \in [+\frac{5}{8}, +\frac{7}{8})$	Orange
Spherical cap	$S \in [+\frac{7}{8}, +1]$	Red

Table 3.2: Colour scheme for shape index.

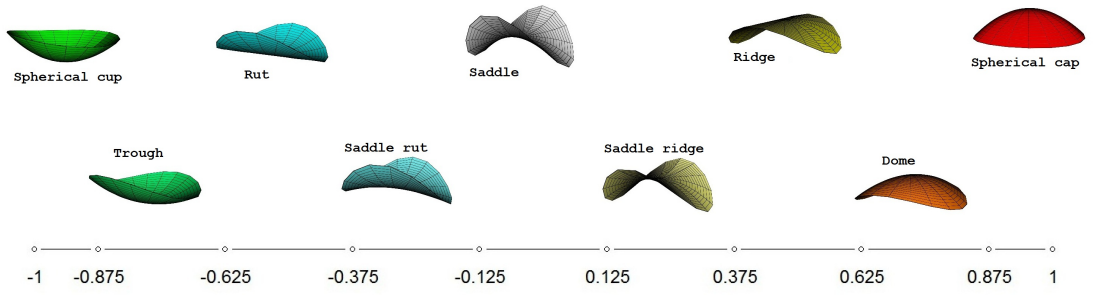


Figure 3.2: Illustration of shape index scale divided into the nine categories of Table 3.2

To summarise, the shape index describes the information contained either in the two principal curvatures, or in the Gaussian and mean curvature taken as a pair, with the advantage that shape index specifies shape independently of size and does not depend on the (arbitrary) assignment of principal direction. It is important to remember that when condensing the principal curvatures into a single number, it

is inevitable that some information will be lost. Here, it is only the relative values of the principal curvatures that matter, not their absolute values.

3.2 Estimation of lip curves in motion

Statistical tools are developed to help identify points along the lip curves. Methods to estimate these curves require the landmarks at the corners of the lips as starting points. Each landmark associated with the corners of the lips, the most lateral points of the mouth fissure, where both lips meet, is labelled a *Cheilion* (abbreviated as *ch*). For differentiation, they will be referred as *chL* (left cheilion) and *chR* (right cheilion).

A brief explanation of the manual estimation of the mouth landmarks in a static image is given below, followed by the two methods used for the estimation of static 3D curves: plane-path and principal curves. The estimation of the landmarks is later enhanced for their estimation in the 4D (three dimensions plus time) lip data, and an algorithm that allows estimation of the curves in the 4D data developed.

3.2.1 Estimation of 3D mouth landmarks

The main problem in identifying the lip curves is to set the landmarks at the corner of the lips. This can be done manually using the ©Landmark software package, developed by the Institute for Data Analysis and Visualisation (IDAV) and introduced in Chapter 2. An alternative method is to use the curvature of the mouth.

As a facial image contains thousands of points, it is convenient to subset just the coordinates corresponding to the mouth, for easier and faster analysis. There are multiple methods for isolating the desired facial feature area. When no landmark has been identified, it can be done manually according to coordinate values. For the mouth, the section between certain values of y can be selected, as well as for values of the x and z coordinates.

Calculation of the shape index for all the points on the mouth will characterise the surface. By displaying the mouth using the shape index scale (see left panel

of Figure 3.3) an initial indication of shapes is obtained. Red and green represent points whose curvatures are similar in strength but different in sign. For a particular image, a logical approach to estimate the *chL* and *chR* landmarks is to identify those parts of the mouth with an appropriate shape index. The corners of the lips are characterized by a spherical cup or trough i.e., a shape index under -0.6. After sub-setting from the mouth image those parts with a shape index of -0.6 or below (see middle panel of Figure 3.3), the two largest parts connected are, in most cases, the corners of the mouth. The next step is to identify a single point in each area. An easy solution is to calculate the mean of the points in the area and then select the closest point in the surface to that. This approach is used for the first mouth of every sequence. The right panel of Figure 3.3 shows the mouth surface with the estimated landmarks.

FIGURE 3.3: On the left, a mouth coloured by shape index. In the middle, those areas with a shape index below -0.625, and the middle point identified. On the right, the mouth is displayed with the estimated *chL* and *chR* landmarks.

3.2.2 Estimation of 3D curves

Once the two landmarks at the corners of the lips are identified in the 3D facial surface, the aim is to estimate the three curves that lie between these two landmarks: the mid-line lip and the upper and lower lip lines.

3.2.2.1 Plane-path

The idea behind a plane-path is to identify a curve in the surface by the points lying on the intersection between the surface and a cutting plane. To determine this plane there are different approaches. In this thesis, the plane is forced to contain two given points. From all the possible planes that contain those two points, the one that makes the shortest path can be selected, or from a series of

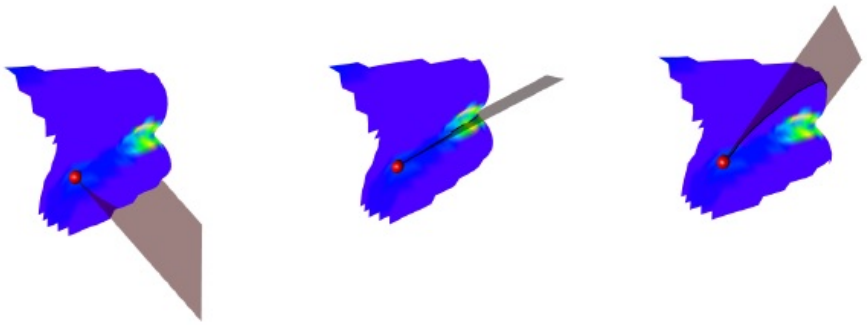


FIGURE 3.4: Three planes from the set of planes containing the landmarks at the corners of the lips, at different rotation angles. [Vittert et al., 2017].

values associated with each point in the surface, identify the path that minimizes the integral value along the path. A rotation angle γ can also be specified to be the angle by which the cutting plane containing the two points (or from the one point in the given direction), is rotated. Three of this possible planes are illustrated in Figure 3.4. The lighter area indicates the targeted shape index for the mid-line lip. The values used to select the cut are usually related to the curvature values, e.g., to maximize the maximum value of the largest principal curvature (in absolute value) for each point in the path. The simple, shortest distance path which lies on the surface does not take into account any curvature information, but simply minimises on the distance.

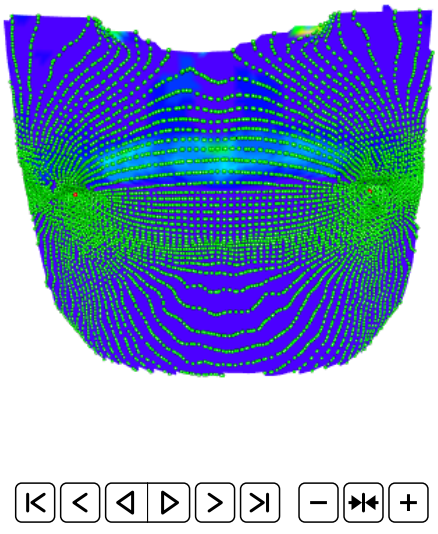


FIGURE 3.5: Plane-path cuts and target area.

Figure 3.5 (when in Adobe reader) shows a visualization of the different cuts the planes make in the surface as the rotation angle γ changes. The lighter area

indicates the targeted shape index for an upper lip. For each γ, the intersection of the plane with the surface describes a path $\mathbf{p}_\gamma$ from *chL* to *chR*.

For the calculation of the path which captures most curvature, a mechanism based on optimising integrated curvature is defined in [Vittert et al., 2017]. Let every three-dimensional point in the surface be referred to as $p(s)$ (i.e., as a function of the arc-length argument s), and $\upsilon = \max(|\kappa_1|, |\kappa_2|)$ a measure of the strength of the curvature. The standardised integral of the maximum curvature along a path $\mathbf{p}_\gamma$ is given by:

$$\left[\int \upsilon(s)ds\right] \Big/ \left[\int 1ds\right] \approx \left[\sum_{j=2}^{n} \omega_j \upsilon(p_j)\right] \Big/ \left[\sum_{j=2}^{n} \omega_j\right], \tag{3.3}$$

where the expression on the right hand is the discrete approximation based on the points that form the path and the weights ω_j are the distance between successive points, $\omega_j = \|p_j - p_{j-1}\|$. Note that (3.3) is not symmetric, i.e., $\sum_{j=2}^{n}$ gives a different (but similar) sum to $\sum_{j=1}^{n-1}$. The curve associated with the path which maximises this expression captures as much curvature as possible. The standardisation by curve length $\int 1ds$ is required to penalise curves which 'pick up' large amounts of curvature by travelling long distances.

3.2.2.2 Principal curve

Principal component analysis (PCA) is perhaps the most commonly used dimensionality reduction method. It is defined using the linear projection that maximizes the variance in the projected space. [Hastie and Stuetzle, 1989] proposal of self-consistent principal curves pointed out a different track for non-linear dimensionality reduction [Ozertem and Erdogmus, 2011] .

Principal curves are defined as self-consistent smooth curves passing through the middle of a k-dimensional data set, providing a non-linear summary of the data. 'Self-consistent' means that if we pick any point on the curve, collect all of the data that project onto this point, and average them, then this average coincides with the point on the curve.

[Hastie and Stuetzle, 1989] defined a projection index $\lambda_r : R^p \to R^1$ as:

$$\lambda_r(x) = \sup_{\lambda}\{\lambda : || \, x - f(\lambda) \, || = \inf_{\mu} || \, x - f(\mu) \, ||\}. \tag{3.4}$$

$\lambda_r(x)$, a function of x, is the value of λ for which $f(\lambda)$ is closest to x. x belongs to the random vector X in R^p with density h and f represents a smooth unit-speed curve. The curve f is then called self-consistent, or a principal curve of h, if

$$E(X|\lambda_r(X) = \lambda) = f(\lambda) \qquad \text{for } a.e.\lambda. \tag{3.5}$$

[Hastie and Stuetzle, 1989] also proved that if a straight line is self-consistent then it is a principal component and that, based on the mean squared error criterion, self-consistent principal curves are saddle points of the distance function. Their original principal curve definition forms a strong basis for many algorithms. The most intuitive algorithm for finding the principal curve starts with any smooth curve (such as the largest principal component). This first curve would be projected and averaged to test whether it is self-consistent or not. If it is not, the procedure is repeated, using the new curve obtained by averaging as a starting guess. This is iterated until convergence.

In this project, principal curves are used as an alternative to a plane-path for those cases where the nature of the surface (for example in an open mouth) does not allow a cut that would satisfactorily represent the lip curves (see Figure 3.6). The algorithm is modified to find the principal curve that contains the path that goes from one given point to another given one (the landmarks), with the plane-path as the starting curve.

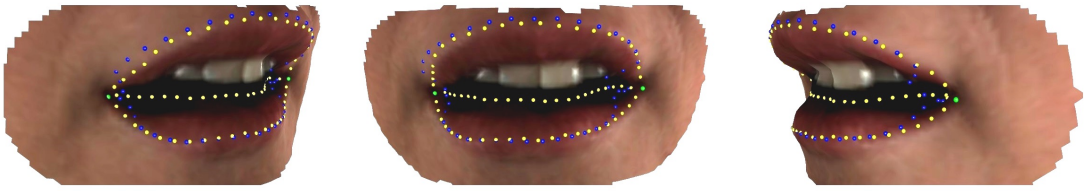


FIGURE 3.6: Comparison of Plane-path (blue) and smooth Principal Curve (yellow) in an open mouth expressing surprise.

3.2.2.3 Algorithm

The general strategy is to identify the local region of interest, use shape index to identify the relevant surface shapes and then identify the path of points in the upper and lower lip curves.

The mesh of points defining the mouth is first cut to work only in the area where the corresponding lip is expected to lie. All the points on the lip curve between chL and chR should lie within a cylinder with axis $(chR - chL)$ and radius r. Empirical investigations with human faces led to the choice of $r = \|chR - chL\|/2$ to identify the local region of interest. For the points in the cylinder, the targeted shape index is identified and set to zero for those points that do not meet the criterion.

For each lip curve, a path is then estimated by either plane-path or principal curves, as explained above. The problem with the principal curve is that the points on it are not guaranteed to lie on the mouth surface. The closest point in the surface for each one is used, and therefore it can only be regarded as a first approximation to an estimate of the curve of interest. Similarly, the plane-path is constrained by the plane. For this, [Vittert, 2015] proposed an algorithm that allows for a more flexible estimate, using the previous paths as reference starting paths. The points in the target area are projected from 3D to 2D. Points in the surface can be represented by two coordinate axes: one representing the signed distance of each point on the surface from its closest point on the reference path (perpendicular distance), and the other describing the arc-length along the reference path of these closest points. Smoothing in 2D is carried out with P-Splines and the resulting smoothed curve is then projected again to the 3-dimensional space. Details of the method are given in [Vittert et al., 2017].

3.2.3 Estimation of 4D mouth landmarks

The landmarks at the corners of the lips are identified manually for the first picture in a sequence as defined in Section 3.2.1. For the rest of the sequence, it was expected to use the closest point in the new surface to the landmarks from the previous image, but, because of head movement and the building of the 3D surface from the photos, pictures from the same sequence do not share the same origin in the 3D space of coordinates, and consequently, the coordinates of the landmarks from the previous image may actually be far from the real ones in the new surface.

Experimentation showed that it was valuable to use information from the previous picture but it was also necessary to find a criterion to determine which point would define better the union of the lips given a target area. To avoid starting landmarks which are too far away, instead of only the closest point to the previous landmarks,

two tubes are calculated around the upper and lower lips paths of the previous image in the sequence. The new landmark would be in the areas where the tubes overlap in the new face surface, more specifically, the mean point will be selected as a starting estimations of the landmarks, denoted by $chL.s$ and $chR.s$, for the left and right corners of the mouth respectively.

An algorithm was created to improve the starting landmarks. First, given the starting landmarks $chL.s$ and $chR.s$, a distance d and a target shape index S_T, the points in the area within the given distance from the starting landmarks and with a shape index equal to or smaller than the one stated are selected as candidate points. That is, $\{\mathbf{p}_{chL} : \|p_{chL} - chL.s\| < d \,\&\, S(p_{chL}) \leq S_T\}$ and $\{\mathbf{p}_{chR} : \|p_{chR} - chR.s\| < d \,\&\, S(p_{chL}) \leq S_T\}$, for the left and right corners respectively, are the sets of possible landmarks.

Then, all the possible paths from the candidate points are evaluated and those points, p_{chL} and p_{chR}, which maximize the sum of the criteria for upper and lower lip are selected as the optimal chL and chR. The criterion selected was based on, as in plane-path, optimising the integral of maximum curvature along the path. This is estimated by the sum of the curvature values for every point in the curve, weighted along the curve. This criterion can be calculated easily for every path. The number of candidate landmarks in each corner of the lip can vary depending on the distance and/or the shape index chosen. If both upper and lower paths had to be calculated for every combination of the two landmarks, the process could become lengthy, i.e., if there are l candidate points for chL and r for chR, there are $l \times r$ possible combinations, and $l \times r \times 2$ paths (upper and lower lips are used) would have to be estimated. To reduce the work involved, in the paths calculated with the starting landmarks ($\mathbf{p} : chL.s \rightarrow chR.s$) the middle point of the upper and lower paths (um and lm) are taken and used to calculate only half of each path. This way, each corner landmark will be optimised with the information of the corresponding half upper and lower lips, which are faster to compute (only $(l \times 2) + (r \times 2)$ paths would have to be calculated), i.e.:

$$\max_{\mathbf{p}_{chL}} \left\{ \left[\int_{\mathbf{p}:p_{chL} \rightarrow um} \upsilon(s)ds \right] + \left[\int_{\mathbf{p}:p_{chL} \rightarrow lm} \upsilon(s)ds \right] \right\} \text{ and}$$

$$\max_{\mathbf{p}_{chR}} \left\{ \left[\int_{\mathbf{p}:p_{chR} \rightarrow um} \upsilon(s)ds \right] + \left[\int_{\mathbf{p}:p_{chR} \rightarrow lm} \upsilon(s)ds \right] \right\}. \quad (3.6)$$

This approach can be done without losing information about the lip's curvatures because the trend in the middle of the path has been proved to change very little between paths with close landmarks and the candidate points should never be too far from the starting landmarks. For every sequence, the selection of the distance and the shape index threshold is a subjective choice in the hands of the researcher. Figure 3.7 shows the starting landmarks, the candidate points and the final optimal landmarks for one 3D image.

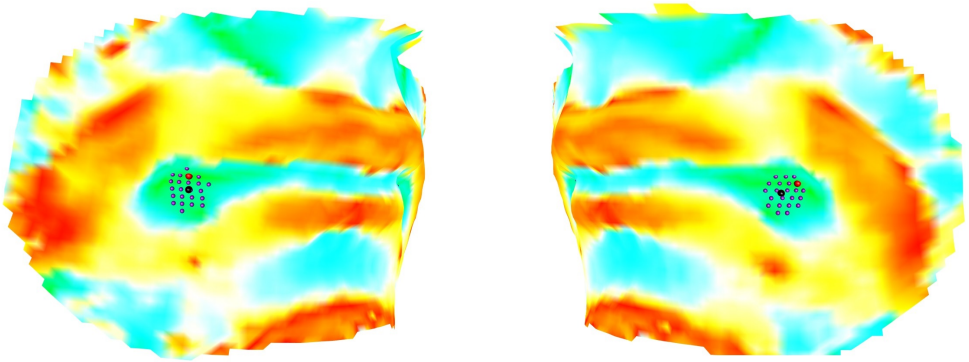


FIGURE 3.7: Starting landmarks for *chR.s* and *chL.s* (red), candidate points within 3 mm and appropriate shape index (purple) and final optimal landmarks (black).

In addition, the shape index scale obtained for a certain picture will depend on the number of nearby points that are used to determinate the shape of a particular point: smaller distances lead to more variability in estimation (Figure 3.8). To identify the starting paths and landmarks, different distances were tried. To avoid misinterpreting information for the ridges in the upper and lower lip, it was decided to split the mouth by the existing valley between the lips, which is easy to identify (shape index below -0.5). Once this mid line was calculated, the shape index was calculated for each part separately and both parts linked again afterwards. After this, the optimal landmarks would be calculated.

3.2.4 Algorithm for the estimation of 4D curves

When analysing the whole set of images for an emotion, for the first picture everything is calculated manually: the subset of the mouth by the coordinates, the starting landmarks by shape index. For the remaining images, the mouth is subsetted by those points within a distance of 25mm from the middle path of the previous image, and the starting landmarks as stated in Section 3.2.1. Once the

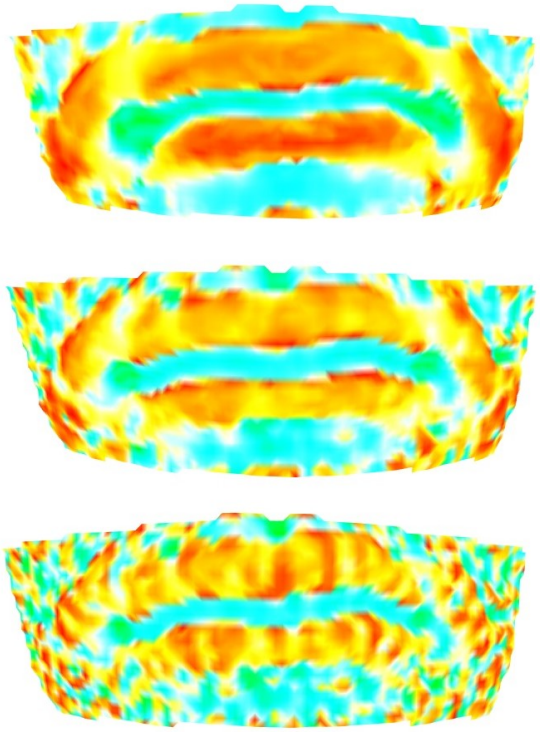


FIGURE 3.8: Illustration of shape index with distances 10, 5 and 3 (top to bottom).

mouth and the initial landmarks are identified, the three paths (mid-line, upper and lower lip) are created using the cutting plane and used to calculate the shape index separately for the upper and lower parts of the mouth. The landmarks are optimised using the created algorithm. With the new optimal landmarks, the paths can be calculated using either plane-path or principal curves. The algorithm to identify the paths was also updated so we can save a set of 'old' curves in the face object, i.e., paths from previous image or first starting paths, to help identify the area where the new path should lie.

This process was carried out recursively, for every image in the sequence: landmarks are optimised from the starting ones and paths are identified with the help of the previous ones to target the area. It should be noted that both with the landmark optimisation and the path identification, there are several parameters for the researcher to control, i.e., the distance and shape index to select the candidate new landmarks and the type of path. The process should be carefully followed and checked to ensure appropriate changes are made when needed. While the algorithm performs well most of the time it does need supervision to account for unusual features of the data, such as the coordinate system changing abruptly due

to the building of the 3D surface, head movement, etc. However, this happens very occasionally and the algorithm is still a huge improvement over manual procedures. In addition, the algorithm is slightly different for each emotion, and the preview of the images will help the researcher to decide the distance over which to search for the optimal landmarks, as well as the method to identify the path (plane-path or principal curve).

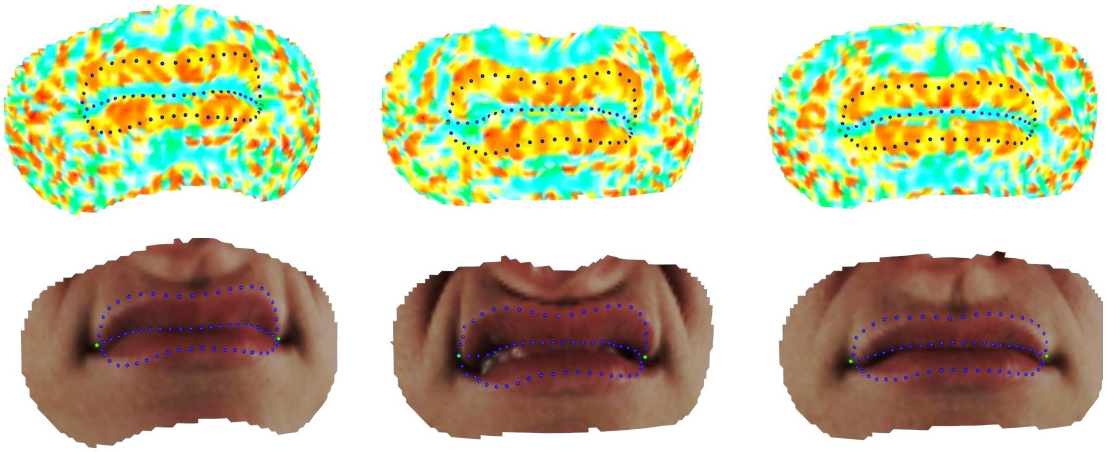


FIGURE 3.9: Snapshots of the emotion *disgust* along the sequence.

Figure 3.9 shows some of the 4D curves in one of the sequences for the emotion *Disgust*. In the upper row the facial surface is printed in the Shape Index colour scheme, where it can be observed that the curves track as much of the ridges as possible. The same curves are plotted in the facial surface on the second row, showing that anatomically defined curves don't always match where the change in skin tissue happens.

3.3 Analysis of 4D lip curves

Before comparing the different replicates of the expression of an emotion, a series of modifications have to be made to make the data comparable. Models for the outline of the mouth are constructed using B-splines to represent the lip as a curve in a set of coefficients instead of individual points. Moreover, B-splines are used to smooth the paths on each image and for every point through the sequence of paths. Procrustes Analysis (Ordinary and General) is performed to match the coordinates of the different paths inside each sequence, as well as among replicates

for better comparison. Below, the theory used for these purposes is described, followed by the full description of the transformations.

3.3.1 B-Splines

Usually in curve fitting, a set of data points is fitted with a curve defined by some mathematical function. In our case, these points will correspond to the coordinates of the path and the function is defined using B-splines. The theory below presented can be found in [De Boor, 1978; Piegl and Tiller, 1987], amongst others. Spline functions consist of polynomial segments which are joined together smoothly at pre-defined subintervals such that a curve estimate can be expressed as: $f(x) = \sum \beta_j b_j(x)$, where the b_j are called the basis functions and the β_j the basis coefficients. The points at which the joins occur are called knots. B-Splines, or basis splines, are a type of spline basis functions, indeed, the most commonly used, as they are particularly flexible and computationally efficient for model fitting. Let x be the variable and $t = (t_0, t_1, \dots)$ the knots vector with $t_0 \leq t_1 \leq \cdots$. Let $j = 0, \dots, p$ be the number of basis functions and d the degree of the polynomial. Then, each of the individual B-Spline functions, $B_j^d(x)$, can be defined recursively from the $d-1$ B-splines as follows (for $d > 0$):

$$B_j^d(x) = \left(\frac{x - t_j}{t_{j+d} - t_j}\right) B_j^{d-1}(x) + \left(\frac{t_{j+d+1} - x}{t_{j+d+1} - t_{j+1}}\right) B_{j+1}^{d-1}(x). \tag{3.7}$$

The B-Spline function of degree zero is defined as:

$$B_j^0 = \begin{cases} 1 & \text{if} \quad t_j \leq x < t_{j+1} \\ \\ 0 & \text{otherwise} \end{cases} \tag{3.8}$$

The curve can be defined as a linear combination of B-Splines:

$$f(x) = \sum_{j=0}^{p} \beta_j B_j^d(x), \tag{3.9}$$

where we shall call the coefficients, β_j, the 'beta-coefficients'.

If the degree of the basis functions is d, and the number basis functions of degree d is $p+1$ then $k = p+d+1$, with $k+1$ ($t = (t_0, t_1, \dots, t_k)$) the number of knots. If

$B_p^d(x)$ is the last basis function of degree d, it is non-zero on $[t_p, t_{p+d+1})$ (given the local support property: $B_j^d(x)$ is a non-zero polynomial on $[t_j, t_{j+d+1})$). Since it is the last basis function, t_{p+d+1} must be the last knot, t_k. Therefore, $t_{p+d+1} = t_k$ and $k = p + d + 1$. In summary, given k and d, let $p = k - d - 1$ and the degree d basis functions are $B_0^d(x), B_1^d(x)...B_p^d(x)$ [Lowther and Shene, 2003].

For n observed values of x: $\mathbf{x} = (x_1, \ldots, x_n)$, the B-Spline basis functions can be stored in a matrix $\mathbf{B} = \mathbf{B}(\mathbf{x})$, such as:

$$\begin{bmatrix} B_1^d(x_1) & \cdots & \cdots & B_p^d(x_1) \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ B_1^d(x_n) & \cdots & \cdots & B_p^d(x_n) \end{bmatrix},$$

and the beta-coefficients estimated then as: $\hat{\beta}_j = (\mathbf{B}^{\mathrm{T}}\mathbf{B})^{-1}\mathbf{B}^{\mathrm{T}}\mathbf{x}$.

In this research, the coordinates from the lip paths are smoothed using B-Splines, first by path, for each of the coordinates against its arc-length, and then each of the beta-coefficients that define the path are also smoothed by time through the sequence of images of one emotion. The Splines used are order 3-cubic B-Splines.

3.3.2 Procrustes Analysis

In statistics, Procrustes analysis is a form of statistical shape analysis used to register a set of shapes into a common coordinate system. The name Procrustes refers to a bandit from Greek mythology who made his victims fit his bed either by stretching their limbs or cutting them off. Procrustes methods are useful for estimating an average shape and for exploring the structure of shape variability in a dataset. To compare the shape of two or more objects, the objects must first be optimally 'superimposed'. Procrustes superimposition (PS) is performed by optimally translating, rotating and uniformly scaling the objects. In other words, both the placement in space and the size of the objects are freely adjusted.

Below, the Ordinary Procrustes Analysis (OPA) and the Generalized Procrustes Analysis (GPA) are explained in detail, following the theory presented in [Dryden

and Mardia, 1998], and the new edition [Dryden and Mardia, 2016]. Before presenting their theory, some definitions need to be clarified: A *configuration* is a set of landmarks on a particular object, and a *configuration matrix, X*, is a $k \times m$ matrix of the coordinates of the k landmarks in m dimensions. Procrustes Analysis involves matching configurations with similarity transformations to be as close as possible in terms of Euclidean distance, using least squares techniques. The Euclidean Similarity Transformations of a configuration matrix X are the rotation matrix, Γ, the translation vector, γ, and the scaling parameter, ξ.

3.3.2.1 Ordinary Procrustes Analysis

Ordinary Procrustes Analysis (OPA) is used for matching two configurations. Consider two configuration matrices X_1 and X_2 (both $k \times m$ matrices of coordinates from k points in m dimensions). The aim is to match the configurations as closely as possible. Estimation of the similarity parameters Γ, γ and ξ is carried out by minimizing the squared Euclidean distance [Dryden and Mardia, 1998]:

$$D^2_{OPA}(X_1, X_2) = || \, X_2 - \xi X_1 \Gamma - 1_k \gamma^{\mathrm{T}} \, ||^2, \tag{3.10}$$

where:

- $||X|| = \text{trace}(X^{\mathrm{T}} X)^{1/2}$: Euclidean norm;
- $\Gamma = (m \times m)$ rotation matrix;
- $\xi > 0$: scale parameter;
- $\gamma = (m \times 1)$ location vector.

The minimum of (3.10) is written as $OSS(X_1, X_2)$, which stands for Ordinary (procrustes) Sum of Squares, and has solution:

- $\hat{\gamma} = 0$,
- $\hat{\Gamma} = UV^t$, where:
  - $X_2^{\mathrm{T}} X_1 = \|X_1\| \, \|X_2\| V \Lambda U^t$.

- $U, V \in SO(m)$, $m \times m$ orthogonal matrices with determinant $+1$.

- Λ, a diagonal $m \times m$ matrix of positive elements.

- $\hat{\xi} = \dfrac{\text{trace}(X_2^{\text{T}} X_1 \hat{\Gamma})}{\text{trace}(X_1^{\text{T}} X_1)}$.

It can be proved that:

$$OSS(X_1, X_2) = \|X_2\|^2 \sin^2 \rho(X_1, X_2), \tag{3.11}$$

where $\rho(X_1, X_2)$ is the Procrustes distance (sum of distances between corresponding landmarks). In broad terms, the sine function can be thought of as a measure of distance. In shape space, when scale is taken out, the configurations lie in a sphere and the $\sin^2$ of the angle between the two projected shapes is a measure of the distance between them. The detailed trigonometry can be found in the book of Dryden and Mardia [1998], where this expression is derived below Equation 5.6.

3.3.2.2 Generalized Procrustes Analysis

Consider the general case where $n \geq 2$ configuration matrices are available $X_1, \cdots, X_n$. Generalized Procrustes Analysis (GPA) involves the superimposition of all configurations placed 'on top of each other' in optimal positions by translating, rotating and rescaling each figure so as to minimize the sum of squared Euclidean distances.

A general idea of the algorithm is:

1. Choose an arbitrary reference shape. This is typically selected among the available instances.

2. Superimpose all instances to current reference shape.

3. Compute the mean shape of the current set of superimposed shapes.

4. If the Procrustes distance between mean and reference shape is above a threshold, set reference to mean shape and repeat from step 2.

Mathematically, the algorithm for GPA is as follows [Dryden and Mardia, 1998]:

1. **Translations.** Centre the configurations to remove location. Initially let

$$X_i^P = X_i,\ i = 1, ..., n.$$

2. **Rotations.** Let, for the ith configuration:

$$\overline{X}_{(i)} = \tfrac{1}{n-1} \textstyle\sum_{j \neq i} X_j^P.$$

The new X_i^P is taken to be the ordinary Procrustes superimposition, involving the rotation, of the old X_i^P on $\overline{X}_{(i)}$. The n figures are rotated in turn, repeatedly until

$$G(X_1, ..., X_n) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=i+1}^{n} \|(\xi_i X_i \Gamma_i + 1_k \gamma_i^{\mathrm{T}}) - (\xi_j X_j \Gamma_j + 1_k \gamma_j^{\mathrm{T}})\|^2, \quad (3.12)$$

the Procrustes sum of Squares, cannot be reduced further.

3. **Scaling.** Let Φ be the $n \times n$ correlation matrix of the $\mathrm{vec}(X_i^P)$, with eigenvector $\phi = (\phi_1, ..., \phi_n)^{\mathrm{T}}$, corresponding to the largest eigenvalue. Then, take for each i:

$$\hat{\xi}_i = \left(\frac{\sum_{k=1}^{n} \|X_k^P\|^2}{\|X_i^P\|^2} \right)^{1/2} \phi_i. \quad (3.13)$$

4. Repeat steps 2 and 3 until $G(X_1, ..., X_n)$, the Procrustes sum of squares (3.12) cannot be reduced any further.

The algorithm generally converges quickly. The concept of GPA was originally proposed by [Kristof and Wingersky, 1971]. Note that in OPA one of the two configurations is regarded as fixed and the other rotated to fit it. When dealing with data for more than two individuals it is natural to consider rotations with respect to the common centroid or consensus configuration [Gower, 1975]. While OPA is not symmetrical in the ordering of the objects, GPA (even when $n = 2$) is invariant under re-ordering of the objects [Dryden and Mardia, 1998].

3.3.3 Pre-analysis transformations

Before being able to compare the replicates of one emotion, it is convenient to perform some transformations so the data are comparable. For each of the two lip curve (upper and lower lips) there are 24 points. For this, each of the sequences

has to first be put together in the same coordinate space and smoothed. Then, all the replicates are matched together and smoothed again. The steps for this are:

For each of the sequences of the same expression:

- GPA to match the lip paths coordinates.
- Smoothing by path: Three B-Splines models for (x, y, z) coordinates, for upper and lower lip.
- Smoothing by time: smoothing of the B-Splines coefficients estimated over time.
- Calculate the new curves from the coefficients and translate them back to the original image coordinates. OPA is used to obtain the transformations between the new curve and the original one.

Once all repetitions of the emotions have been smoothed, one proceeds to:

- Matching of the coordinates of the first image of each repetition with GPA.
- Comparison of the matched coordinates with each of the original images: OPA to obtain the transformations and apply these to the rest of each sequence. All the images in all the repetitions lie then in the same coordinates space.
- Smoothing by path and time (time for each sequence). Coefficients of B-Splines are replaced for the new ones and the resulting curves updated.

To compare the expression, the difference from each beta in the sequence to the corresponding one in the first image is used. When adding the differences for the betas of the x, y and z coordinates an overall difference can be seen, giving a general idea or change. Figure 3.10 shows the absolute beta coefficient differences for each of the sequences of *surprise*. The index in the x-axis represents the picture in the sequence.

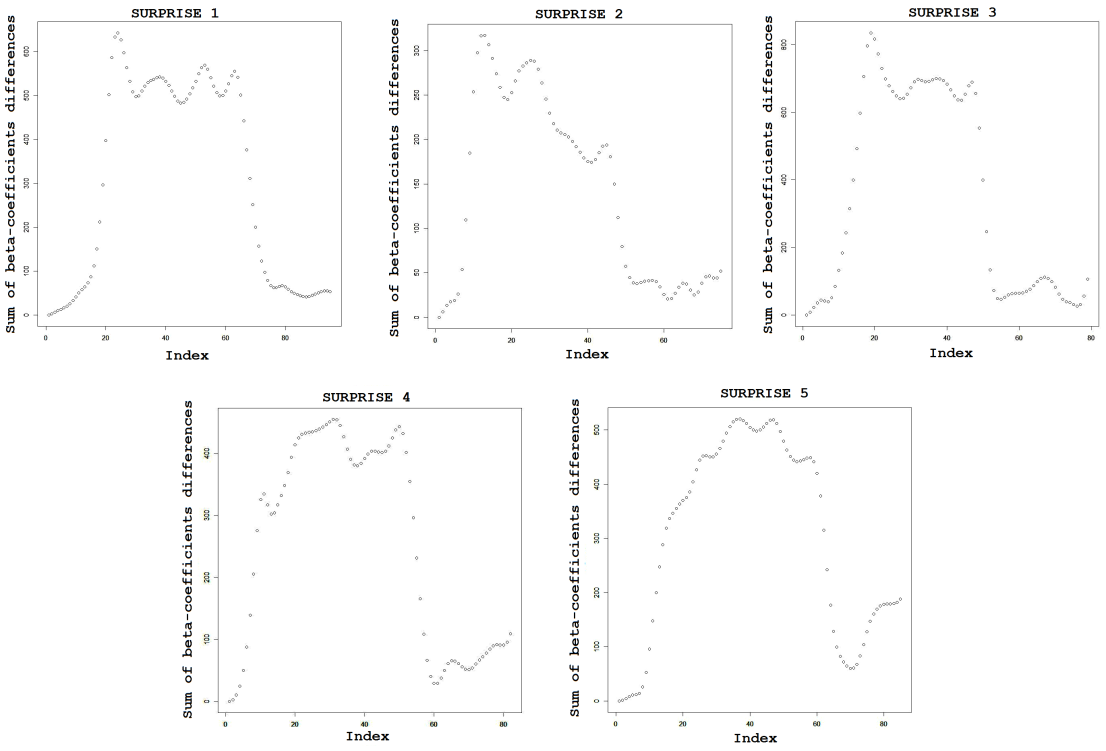


FIGURE 3.10: Sum of absolute differences in the beta coefficients for coordinates x, y and z from the first image of the sequence in the five replicates of the emotion 'surprise'.

Due to the resting position of the lips at the beginning and the end of each of the sequence, finding a principled approach to cut the extremes of the sequences was considered. For this purpose, different procedures were studied:

- For each time point represented by the sum of the differences of the x, y and z betas from the first image, average the following 3, 5 or 10 differences between points. The first big change could be considered to be at the point with the maximum associated 'mean change'. As the differences are decreasing as the lips go back to the rested position, the second big change could be considered at the minimum 'mean change'. Moreover, as the changes should have occurred before the second cutting point, for the second half of the sequence, instead of taking the average of the following difference, the previous 3, 5 or 10 differences were used.

  The problem with this procedure was that if the number of differences was too small, the points obtained were in the middle of the change, where the bigger differences are found, but the starting of the movement of the lips was

missing. If the number of differences examined was too high, the very first and last points were often identified as the changing points.

- Assuming that the start and the end of the emotion representation should look similar, the next approach aimed to compare the vector of mean changes calculated previously and, instead of choosing the maximum and minimum values, compare the value associated with each of the first point with the values associated to the points in the end section. The first pair of points identified with mean changes less than a certain threshold apart were chosen as the starting and ending points.

  This approach needed to have both the number of differences used to calculate the 'mean change' and the threshold chosen by the researcher. Furthermore, this procedure would work assuming the changes in both directions (from rested lip to the expression of the emotion and vice-versa) were to happen at the same 'speed', which was not always the case.

Subsequently, finding no consistent way to state when the difference between two set of betas (β) were significantly large to identify that point as the starting of the emotion, it was decided to use the full sequence of betas. However, as not all the replicates of the same expressions are the same length, the betas coefficients were interpolated between the first and last image to obtain the same number of equally-spaced time points in each sequence. Specifically, it was decided to use 20 time points.

3.3.4 Mean shape of the emotions

Once the different replicates of an emotion are comparable in terms of the beta-coefficients, Principal Components Analysis (PCA) is used to study how the mean shape varies over space and time.

The principal components represent the underlying structure in the data. They are the directions where there is the most variance, where the data is most spread out. Principal components are calculated from eigenvectors and eigenvalues of the data. Eigenvectors and eigenvalues exist in pairs: every eigenvector has a corresponding eigenvalue. An eigenvector is a direction, whereas an eigenvalue is a number representing how much variance there is in the data in that direction. The eigenvector with the highest eigenvalue is taken as the first principal component.

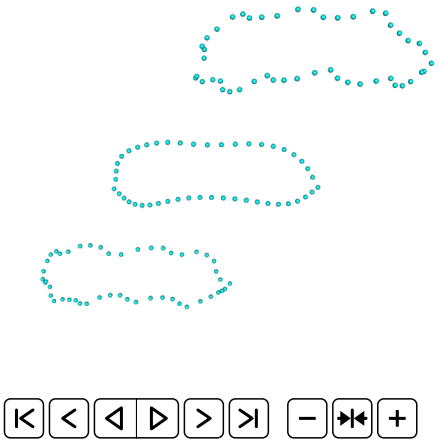
To analyse the variation across the replicates of each of the emotions, PCA is applied to the whole set of beta coefficients for each of the time points in each of the coordinates, over the replicates. For this, the data are stored in a matrix where the columns represent each of the repetitions of the emotion and the rows the coefficients. The forty-six x-coordinate coefficients for each of the twenty time points are stored one after another (nine-hundred twenty rows), and after them, the same for the y and z coordinates. The result is a matrix of $2760 \times n$, n being the number of replicates: four in the emotions *Anger, Disgust* and *Sadness*, five for *Fear* and *Surprise* and three in the case of *Happiness*.

To explore the mean shape of an emotion over space and time, and the variation around it, the eigenvalues and eigenvectors returned from the PCA are used as follows:

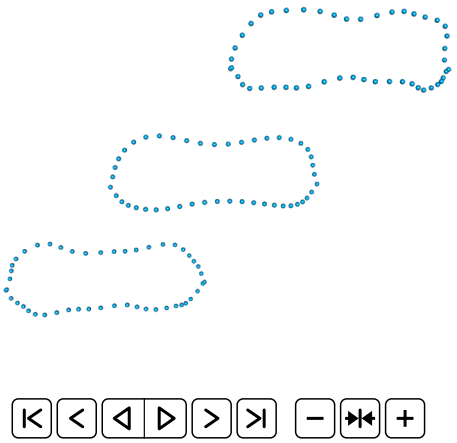
$$\text{mean} \pm 2\sqrt{\text{eigenvalue}} \times \text{eigenvector}, \tag{3.14}$$

for each of the time points. The curves can be recreated from these new coefficients using the design matrix created on the smoothing step when computing the B-Spline representation.

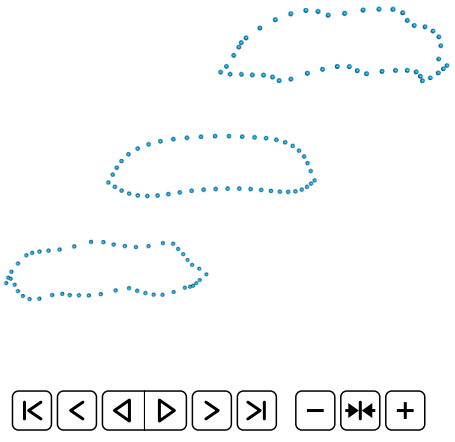
The plots in Figure 3.11 show, for each emotion, the average of the of the lip curves through the replicates (in the middle) with the variation around it. In the printed version, only the middle time-point is shown, every time-point can be viewed when reproduced in Adobe reader. It can be appreciated that the smaller the change along the time points in the mean, the more noisy is the variation around it. The change along the mean emotion is smooth in all of the emotions regardless of how wiggly the variation around it looks. For the mean average of *Happiness* one should take into account that for some replicates the actress performed a smile with an open mouth whereas in two of them it was with closed mouth. The difference in the number of replicated does not seem to be significant to extrapolate any result on how it affects the mean representation. *Anger*, *Happiness* and *Sadness* are the emotions with the most wiggly variations, but this could be both for having less replicates or, most likely, due to the general amount of change in the emotion being small.



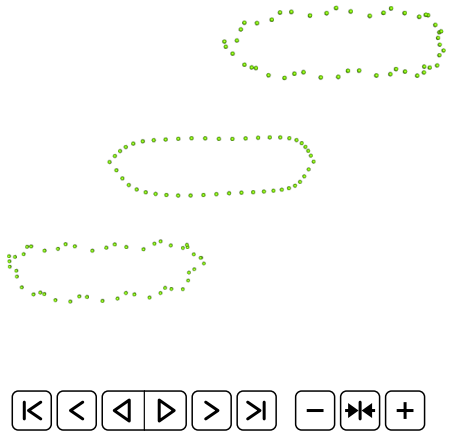
(a) Anger



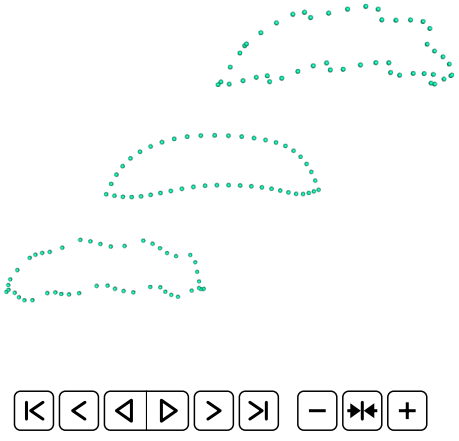
(b) Disgust



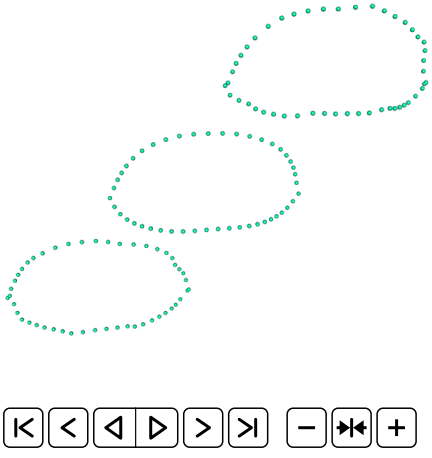
(c) Fear



(d) Happiness



(e) Sadness



(f) Surprise

FIGURE 3.11: Average emotions displayed with their variation, calculated from the first Principal Component.

3.4 Discussion

The matter of estimating 4D face curves, specifically lip curves, can still be improved. Although the use of shape index instead of skin tissue colour change gives a better representation, the allocation of the landmarks and the track of the ridges in open mouths still has room for improvement. A candidate can be using curves from the software that builds the 3D image, on which the Institute of Psychology of the University of Glasgow is working. However, one of the motivations for doing the research was to investigate methods which use shape information to track the curves, the ©*Di3D* methods don't follow that procedure. The allocation of landmarks by using an estimation from the previous image is better than manually allocating them in terms of time of computation and effectiveness.

The use of B-Splines to model the curves and smooth the paths has shown good results and made the comparison of the betas feasible. Since as many knots as data points are being used in smoothing the path, however, the B-Splines are not actually smoothing here (this has been done as part of the path identification in the algorithm by Vittert et al. [2017], see Section 3.2.2.3), but rather providing basis functions that allow the spatial points to be interpolated. Further research might include some model selection to select fewer knots, which would provide additional smoothing. Choosing the number and locations of the knots is an equivalent problem to choosing the number and values of the support vectors [Murphy, 2012]. Different approaches are used for this, e.g. cross-validation. Alternatively, penalized regression splines can be used. Moreover, it would be natural to compare the differences between the results of applying a B-Spline model in time to the registered points directly.

The process of the pre-analysis transformations and the PCA has led to a good representation of the average emotion shape. The problems with the variation over time and space observed, especially in those emotions where the lips do not exhibit big changes might most probably be related to the tracking of the original paths and the estimation of the landmarks in images that do not change much from one another but whose coordinates system might be too distant for an optimal allocation of the landmarks and paths. The fact that in emotions such as *Disgust* or *Surprise*, where the changes are much larger, the PCA results in a good representation supports the claim that the methods discussed in this chapter are promising means of averaging the expression of an emotion.

Chapter 4

Gaussian Process model for k-dimensional curves

4.1 Introduction and background

The aim is to model the data of the lip curves in terms of Gaussian Processes (GPs), thinking of them as defining a distribution over functions. They are extensions to, and, indeed, defined in terms of, the multivariate normal distribution, which is reviewed in the next section.

Gaussian processes are mathematically equivalent to many well-known models, including spline models [Murphy, 2012]. In Section 3.3.1, it was introduced that a curve can be defined as a linear combination of B-Splines:

$$f(x) = \sum_{j=0}^{p} \beta_j B_j^d(x). \tag{4.1}$$

Consider the case where the vector of beta-coefficients ($\boldsymbol{\beta}$) follows a normal distribution:

$$\boldsymbol{\beta} \sim N(0, \mathbf{C}_\beta). \tag{4.2}$$

This is equivalent to a Gaussian process prior model with a prior covariance matrix $\mathbf{C}_f = \mathbf{B}\mathbf{C}_\beta\mathbf{B}^{\mathrm{T}}$, where $\mathbf{B}$ is the basis matrix composed of the $\mathbf{B}(\mathbf{x})$ basis functions (theory on the definition of the Gaussian process model and its covariance function will be presented in Section 4.2). In other words, the basis chosen and the prior

over the coefficients implies a Gaussian process for the function with a particular covariance function [Paciorek, 2003].

4.1.1 The multivariate normal distribution

The joint Gaussian distribution, or multivariate normal distribution, is a generalization of the one-dimensional normal distribution. Most of the theory presented below can be found in [McColl, 2004]. If every linear combination of the k components of a random vector has a univariate normal distribution, the vector is said to be k-variate normally distributed. A more formal definition starts with the definition of the multivariate standard normal. Let $Z_1, ..., Z_k$ be independent $N(0,1)$ random variables. Then $Z = [Z_1, ..., Z_k]^{\mathrm{T}}$ is said to have a k-variate standard normal distribution and is written $Z \sim N_k(0, \mathbf{I})$. The general multivariate normal can be seen as a linear transformation of the standard multivariate normal distribution. Suppose that $Z \sim N_k(0, \mathbf{I})$ and $X = \mathbf{A}Z + b$, where $b \in \mathbb{R}^p$ and $\mathbf{A}$ is a $p \times k$ matrix of constants with $\mathrm{rank}(\mathbf{A}) = k$. Then X is said to have a p-variate normal distribution of rank p, with mean vector

$$\mu = \mathbb{E}(X) = \mathbf{A}\mathbb{E}(Z) + b = b, \tag{4.3}$$

and variance-covariance matrix

$$\Sigma = \mathrm{Cov}(X) = \mathbf{A}\mathrm{Cov}(Z)\mathbf{A}^{\mathrm{T}} = \mathbf{A}\mathbf{A}^{\mathrm{T}}. \tag{4.4}$$

Written $X \sim N_p(\mu, \Sigma)$, the p.d.f. of X is

$$p(x) = (2\pi)^{-\frac{p}{2}}|\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu)\right\}, \tag{4.5}$$

where $|\cdot|$ indicates the determinant.

The log-likelihood of a single observation from $N_p(\mu, \Sigma)$ is therefore given by:

$$\log p(x) = -\frac{p}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu). \tag{4.6}$$

Moreover, for a random variable with a new mean μ_a and a covariance multiplied by a scalar $b > 0$, i.e., $R \sim N_p(\mu_a, b\Sigma)$, the corresponding log-likelihood is:

$$\begin{aligned} \log p(r) &= -\frac{p}{2}\log(2\pi) - \frac{1}{2}\log|b\Sigma| - \frac{1}{2}(r-\mu_a)^{\mathrm{T}}[b\Sigma]^{-1}(r-\mu_a) \\ &= -\frac{p}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{p}{2}\log(b) - \frac{1}{2b}(r-\mu_a)^{\mathrm{T}}\Sigma^{-1}(r-\mu_a). \end{aligned} \tag{4.7}$$

One key property of the multivariate normal distribution is the 'marginalization property': if the distribution specifies, for example, $(X_1, X_2)^{\mathrm{T}} \sim N(\mu, \Sigma)$, then it must also specify $X_1 \sim N(\mu_1, \Sigma_{11})$, with Σ_{11} being the relevant sub-matrix of Σ:

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \sim N_p\left(\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}\right). \tag{4.8}$$

Furthermore, the distribution of X_1 conditional on $X_2 = x_2$ is a multivariate normal: $X_1 \mid X_2 = x_2 \sim N_{p^*}(\bar{\mu}, \bar{\Sigma})$, with:

$$\begin{aligned} \bar{\mu} &= \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \\ \bar{\Sigma} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \end{aligned} \tag{4.9}$$

4.2 Gaussian Processes for 1D curves

Gaussian Processes provide a flexible model for continuous functions. A GP is a collection of random variables, any finite number of which have a joint Gaussian distribution (multivariate normal distribution). A GP $x(s)$ is defined by its mean function $m(s)$ and the covariance function $k(s, s')$ [Rasmussen and Williams, 2006]:

$$\begin{aligned} m(s) &= E[x(s)], \\ k(s, s') &= E[(x(s) - m(s))(x(s') - m(s'))]. \end{aligned} \tag{4.10}$$

Consider the case of a single 3D lip curve, for example the shape of one upper lip (Figure 4.1). It is defined by the values of the three coordinates, x, y, z, of all the points along the upper lip. In this scenario, the variable s is the continuous index labelling points along the curve, e.g., the arc-length of the curve, rescaled to be from 0 to 1, $s \in [0, 1]$. The GP for the values of one coordinate (e.g., the x

coordinate) in terms of the arc-length, can be written as:

$$x(s) \sim GP\big(m(s), k_s(s, s')\big). \tag{4.11}$$

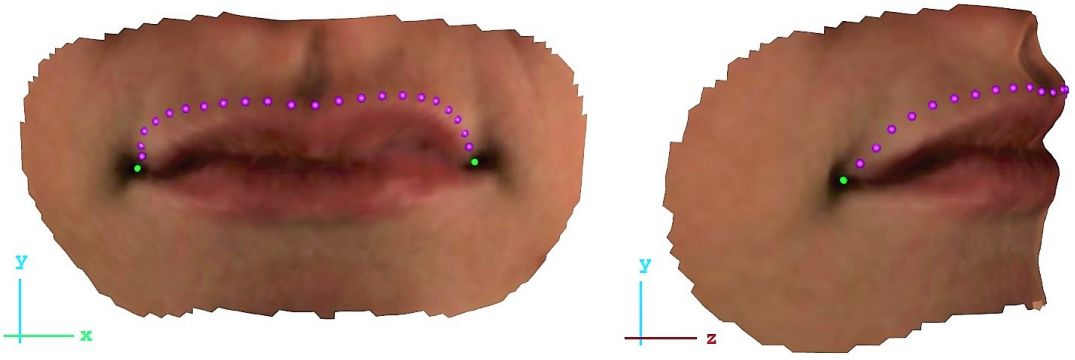


FIGURE 4.1: One 3D mouth and its upper lip curve, defined by 24 points. Picture on the left shows the x and y coordinates and the picture on the right, y and z.

To fit the GP, a set of observations (called the training points) at locations $\mathbf{s} = (s_1, ..., s_n)^{\mathrm{T}}$, with n the number of observed points along the curve (in our application, the observed points in the arc-length are equally spaced but nothing of what follows requires that), is used to infer the relationship between inputs s and targets $x(s)$. The observed curve can be then defined as:

$$\mathbf{x} \equiv x(\mathbf{s}) \equiv \begin{pmatrix} x(s_1) \\ \vdots \\ x(s_n) \end{pmatrix}. \tag{4.12}$$

Then:

$$\mathbf{x} \sim N_n\left(\mathbf{m}, \mathbf{K}\right), \tag{4.13}$$

where $\mathbf{m}$, the mean, is assumed to be zero, and $\mathbf{K}$ is the covariance matrix (see Section 4.2.1). Similarly, GPs $y(s)$ and $z(s)$ can be specified for the remaining coordinates. Figure 4.2 shows the values for the observed points from Figure 4.1 on the curves $x(s)$, $y(s)$ and $z(s)$ curves.

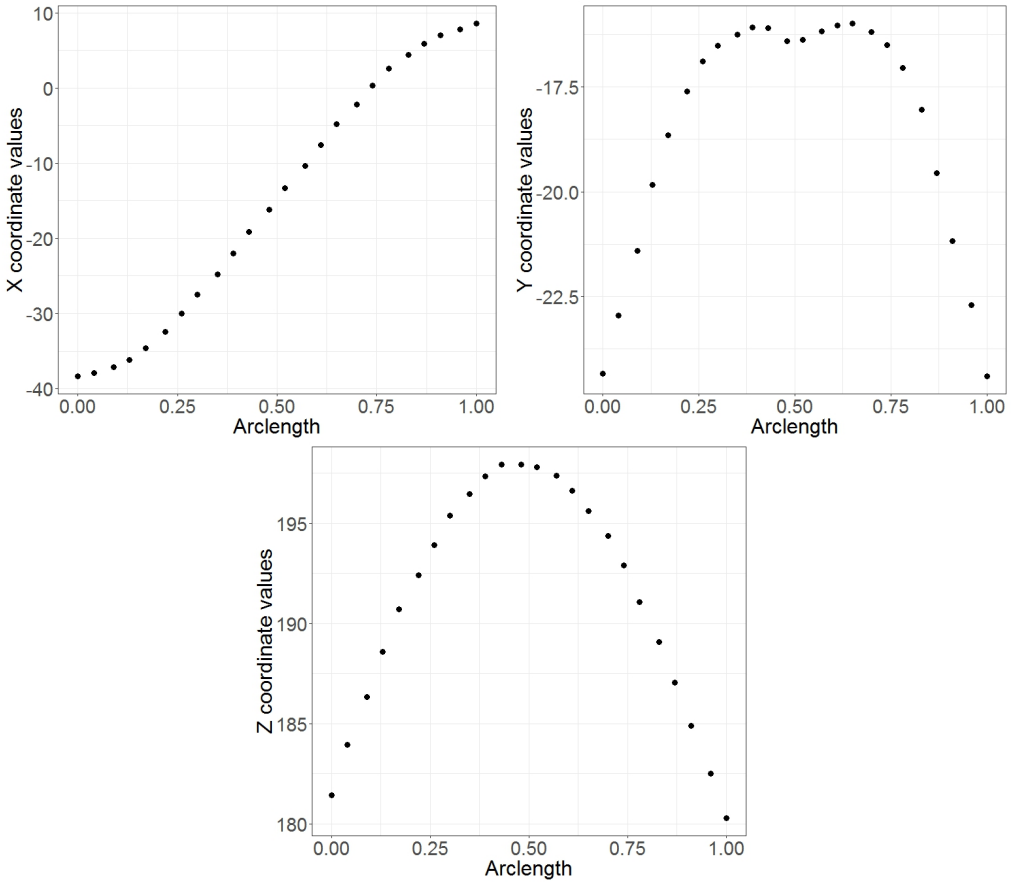


FIGURE 4.2: Data points for x, y and z coordinates plotted against the arc-length of the curve, rescaled to [0,1].

4.2.1 Covariance function

Whilst the mean of the process is widely assumed to be zero [Rasmussen and Williams, 2006], the covariance function used to define the GP is a crucial ingredient, as it encodes dependency assumptions about the function, for example, related to smoothness. Note that the covariance between the outputs, $x(s)$ and $x(s')$, is written as a bi-function of the inputs, s and s' (4.10). It is safe to assume that points with close inputs are likely to have similar target values, and thus, observed points should be informative about predictions near them. It is the covariance function that defines nearness or similarity. Basic aspects that can be defined through the covariance function are the process' stationarity, isotropy, smoothness and periodicity.

Stationarity refers to the process' behaviour regarding the separation of any two points s and s'. If the process is stationary the joint probability distribution does

not change when shifted in s. It depends on the separation, $s - s'$. Consequently, parameters such as the mean and variance, also do not change over s. If the process is non-stationary it depends on the actual position of the points s and s' (not just their difference). If, further, the covariance is a function only of the Euclidean distance (not the direction) between s and s', then the process is called 'isotropic'. Isotropic means it is invariant to all rigid motions (in 1D, just inversion $s \to -s$). An isotropic covariance function is invariant to shifts and rotations of the input space (when s is of higher dimension).

A general name for a function k of two arguments mapping a pair of inputs s and s' into $\mathbb{R}$ is a 'kernel'. A real kernel is said to be symmetric if $k(s, s') = k(s, s')$; covariance functions must be symmetric by the definition of covariance. If k is a covariance function, the covariance matrix $\mathbf{K}$ can be computed with entries $\mathbf{K}_{ij} = k(s_i, s_j)$. This covariance matrix should be positive semi-definite (PSD). A symmetric matrix is PSD if and only if all its eigenvalues are non-negative.

This work uses the squared exponential (SE) covariance function to describe the curves indexed by the arc-length, s. This covariance function is infinitely differentiable, which means that the GP with this covariance function has mean square derivatives of all orders, and is therefore very smooth (as the lip curves are). The SE is probably the most widely-used kernel [Rasmussen and Williams, 2006].

It is defined by:

$$k_s(s, s') = \sigma_f^2 \exp\left(-\frac{1}{2\lambda^2}(s - s')^2\right), \tag{4.14}$$

with free parameters: σ_f^2, the signal variance or magnitude, and λ, the length-scale. The signal variance determines the variation of function values from their mean. Small values of σ_f^2 characterize functions that stay close to their mean value, larger values allow more variation. The length-scale describes how smooth the function is. Small length-scale values means function values change quickly, large values characterize functions that change slowly. If it is considered that the measured values of the GP $x(s)$, $\tilde{x}(s)$, are contaminated with i.i.d. Gaussian noise of variance σ_n^2, i.e., $\tilde{x}(s) = x(s) + \epsilon(s)$, with $\epsilon(s) \sim N(0, \sigma_n^2)$, then $\text{cov}(\tilde{x}(s)) = \text{cov}(x(s)) + \text{cov}(\epsilon(s)) = \mathbf{K} + \sigma_n^2 \mathrm{I}_n$ (by independence of $x(s)$ and $\epsilon(s)$). Therefore, the kernel is modified as follows:

$$k_n(s_i, s_j) = \sigma_f^2 \exp\left(-\frac{1}{2\lambda^2}(s_i - s_j)^2\right) + \sigma_n^2 \delta_{ij}, \tag{4.15}$$

where δ_{ij} is the Kronecker delta, which is one iff $i = j$ and zero otherwise. Note the difference between k_s, the covariance function for the noise-free latent $x(s)$, and k_n, for the noisy targets.

In general, the free parameters are called *hyperparameters* and, in this thesis, will be denoted by $\boldsymbol{\theta}$. The hyperparameters are on a higher level of the model, sitting above the parameters which would be the basis coefficients in a basis-function view of a Gaussian Process, a viewpoint which was already alluded to at the start of the Chapter, but would not be used again in this thesis.

4.2.2 Likelihood

To select the best combination of the hyperparameters, the likelihood is maximised for the n observations, in the noise-less case. This is achieved in practice by maximising with respect to $\boldsymbol{\theta}$ the log-likelihood:

$$\log p(\mathbf{x} \mid \mathbf{s}, \boldsymbol{\theta}) = -\frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{K}_s^{-1}\mathbf{x} - \frac{1}{2}\log \mid \mathbf{K}_s \mid -\frac{n}{2}\log(2\pi), \tag{4.16}$$

where $\mathbf{K}_s$ is the covariance matrix for the n arc-length inputs, with $(i,j)^{th}$ element equal to $k_s(s_i, s_j)$ and depends on the hyperparameters σ_f and λ.

In the case of noisy data, one needs to maximise the marginal likelihood (where the marginalisation is over the unknown noise-free version of the data). In practice this is achieved by replaicing $\mathbf{K}_s$ in (4.16) with $\mathbf{K}_n$, whose elements are given in (4.15).

4.2.3 Hessian matrix

Some measure of the uncertainty of the maximum likelihood estimates is necessary, and, furthermore, it is useful to get a general idea of the relationships between those estimates. The Hessian matrix can provide both, via, respectively, standard errors and correlations.

In general, the Hessian matrix is the square matrix of second-order partial derivatives of a scalar-valued function. In this case, the function is the log-likelihood, evaluated at the maximum likelihood estimates ($\hat{\boldsymbol{\theta}}$) of the hyperparameters. The negative Hessian evaluated at $\hat{\boldsymbol{\theta}}$ is the same as the observed Fisher information

matrix evaluated at $\hat{\boldsymbol{\theta}}$. The Fisher information matrix essentially describes the amount of information data provide about an unknown parameter. If $l(\boldsymbol{\theta})$ is the log-likelihood function for a set of k hyperparameters, $\boldsymbol{\theta}$, then the Fisher information matrix $\mathbf{I}(\boldsymbol{\theta})$ is given by the $k \times k$ symmetric matrix whose $(i,j)^{th}$ element is given by the covariance between first partial derivatives of the log-likelihood,

$$\mathbf{I}(\boldsymbol{\theta})_{ij} = \mathrm{Cov}\left[\frac{\partial l(\boldsymbol{\theta})}{\partial \theta_i}, \frac{\partial l(\boldsymbol{\theta})}{\partial \theta_j}\right]. \tag{4.17}$$

Under certain regularity conditions (exchangeability of the order of differentiation and integration), the Fisher information matrix can also be expressed in terms of the expected values of the second partial derivatives:

$$\mathbf{I}(\boldsymbol{\theta})_{ij} = -\mathrm{E}\left[\frac{\partial^2}{\partial \theta_i \partial \theta_j} l(\boldsymbol{\theta})\right]. \tag{4.18}$$

This holds when the expected value of the partial derivative is 0 and $\log p(X \mid \theta)$ is twice differentiable with respect to θ. This can be shown in the scalar case, let $l(\theta) = \log p(X \mid \theta)$ for a data set X and a scalar θ, then:

$$\begin{aligned}
\mathrm{E}\left[\frac{\partial}{\partial \theta} \log p(X \mid \theta)\right] &= \int \left(\frac{\partial}{\partial \theta} \log p(x \mid \theta)\right) p(x \mid \theta) dx \\
&= \int \frac{\frac{\partial}{\partial \theta} p(x \mid \theta)}{p(x \mid \theta)} p(x \mid \theta) dx \\
&= \int \frac{\partial}{\partial \theta} p(x \mid \theta) dx \\
&= \frac{\partial}{\partial \theta} \int p(x \mid \theta) dx = \frac{\partial}{\partial \theta} 1 = 0.
\end{aligned} \tag{4.19}$$

The second derivative is:

$$\begin{aligned}
\frac{\partial^2}{\partial \theta^2} \log p(x \mid \theta) &= \frac{\frac{\partial^2}{\partial \theta^2} p(x \mid \theta)}{p(x \mid \theta)} - \left(\frac{\frac{\partial}{\partial \theta} p(x \mid \theta)}{p(x \mid \theta)}\right)^2 \\
&= \frac{\frac{\partial^2}{\partial \theta^2} p(x \mid \theta)}{p(x \mid \theta)} - \left(\frac{\partial}{\partial \theta} \log p(x \mid \theta)\right)^2,
\end{aligned} \tag{4.20}$$

with

$$\mathrm{E}\left[\frac{\frac{\partial^2}{\partial \theta^2} p(x \mid \theta)}{p(x \mid \theta)}\right] = \frac{\partial^2}{\partial \theta^2} \int p(x \mid \theta) dx = 0. \tag{4.21}$$

Therefore (still subject to regularity conditions):

$$\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\log p(x \mid \theta)\right] = -\mathrm{E}\left[\left(\frac{\partial}{\partial\theta}p(x \mid \theta)\right)^2\right] \tag{4.22}$$

The Fisher information, described above as the covariance between first partial derivatives, is

$$I(\theta) = \mathrm{Var}\left[\frac{\partial}{\partial\theta}\log p(X \mid \theta)\right] = \mathrm{E}\left[\left(\frac{\partial}{\partial\theta}\log p(X \mid \theta)\right)^2\right], \tag{4.23}$$

and so:

$$\mathrm{I}(\theta) = -\mathrm{E}\left[\frac{\partial^2}{\partial\theta^2}\log p(x \mid \theta)\right]. \tag{4.24}$$

Strictly, this definition corresponds to the expected Fisher information. If no expectation is taken, the data-dependent quantity, called the observed Fisher information, is obtained [Myung and Navarro, 2004]. The observed Fisher information matrix, $\mathbf{I}(\hat{\boldsymbol{\theta}})$, the information matrix evaluated at the MLE, can be written as:

$$\mathbf{I}(\hat{\boldsymbol{\theta}}) = -\left.\frac{\partial^2}{\partial\theta_i\partial\theta_j}l(\boldsymbol{\theta})\right|_{\hat{\theta}_i,\hat{\theta}_j}. \tag{4.25}$$

The Hessian is defined as:

$$\mathbf{H}(\boldsymbol{\theta}) = \frac{\partial^2}{\partial\theta_i\partial\theta_j}l(\boldsymbol{\theta}). \tag{4.26}$$

So $\mathbf{I}(\hat{\boldsymbol{\theta}}) = -\mathbf{H}(\hat{\boldsymbol{\theta}})$.

Further, the inverse of the Fisher information matrix is an estimator of the asymptotic (large sample size) covariance matrix:

$$\mathrm{Var}(\hat{\boldsymbol{\theta}}) = \left[\mathbf{I}(\hat{\boldsymbol{\theta}})\right]^{-1}, \tag{4.27}$$

under certain assumptions [Pawitan, 2001]. The asymptotic distribution of a maximum likelihood estimate is normal:

$$\hat{\boldsymbol{\theta}} \sim N\left(\boldsymbol{\theta}_0, \left[\mathbf{I}(\hat{\boldsymbol{\theta}})\right]^{-1}\right), \tag{4.28}$$

where $\boldsymbol{\theta}_0$ denotes the true hyperparameters values.

The estimated standard error of the maximum likelihood estimate of the i^{th} hyperparameter is:

$$\text{SE}(\hat{\theta}_i) = \sqrt{\left[-\mathbf{H}(\hat{\boldsymbol{\theta}})^{-1}\right]_{ii}}. \tag{4.29}$$

Note this is an asymptotic result. Moreover, from the covariance matrix, relationships between the different hyperparameters can be observed. The correlation between $\hat{\theta}_i$ and $\hat{\theta}_j$ can be estimated from:

$$\text{Corr}(\hat{\theta}_i, \hat{\theta}_j) = \frac{\left[-H(\hat{\boldsymbol{\theta}})^{-1}\right]_{ij}}{\sqrt{\left[-H(\hat{\boldsymbol{\theta}})^{-1}\right]_{ii}\left[-H(\hat{\boldsymbol{\theta}})^{-1}\right]_{jj}}}. \tag{4.30}$$

4.2.4 Predictive distributions

Consider the simple problem of mapping from the input s to the output $x(s)$, essentially a regression problem. The set of values of s where data are collected are called training points. Often, one would like to make inferences (predictions) of the function values at other values of s, generally called test points.

If the observations are considered noise free, the joint distribution of the training outputs, $\mathbf{x}$, and the test outputs, $\mathbf{x}^*$, for a set of n training points $\mathbf{s} = (s_1, \ldots, s_n)^{\text{T}}$ and a set of n^* test points $\mathbf{s}^* = (s_1^*, \ldots, s_{n^*}^*)^{\text{T}}$, according to the GP model, is:

$$\begin{bmatrix} \mathbf{x} \\ \mathbf{x}^* \end{bmatrix} = N_{n+n^*}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K}(\mathbf{s},\mathbf{s}) & \mathbf{K}(\mathbf{s},\mathbf{s}^*) \\ \mathbf{K}(\mathbf{s}^*,\mathbf{s}) & \mathbf{K}(\mathbf{s}^*,\mathbf{s}^*) \end{bmatrix}\right), \tag{4.31}$$

where $\mathbf{K}(\mathbf{s},\mathbf{s}^*)$ denotes the $n \times n^*$ matrix of the covariances evaluated at all pairs of training and test points, with $(i,j)^{th}$ element equal to $k_s(s_i, s_j^*)$. $\mathbf{K}(\mathbf{s},\mathbf{s}), \mathbf{K}(\mathbf{s}^*,\mathbf{s}^*)$ and $\mathbf{K}(\mathbf{s}^*,\mathbf{s})$ are defined analogously. For brevity, let $\mathbf{K}(\mathbf{s},\mathbf{s}) = \mathbf{K}_s$, $\mathbf{K}(\mathbf{s}^*,\mathbf{s}^*) = \mathbf{K}_{s^*}$, $\mathbf{K}(\mathbf{s}^*,\mathbf{s}) = \mathbf{K}_{s^*s}$ and $\mathbf{K}(\mathbf{s},\mathbf{s}^*) = \mathbf{K}_{ss^*}$.

To obtain the posterior, it is necessary to restrict this joint prior distribution (4.31) to contain only those functions which agree with the observed data points. This corresponds to conditioning on the observations, using (4.9):

$$\mathbf{x}^* \mid \mathbf{x}, \mathbf{s}, \mathbf{s}^* \sim N_{n^*}\big(\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{x}, \quad \mathbf{K}_{s^*} - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}\big). \tag{4.32}$$

Values of $\mathbf{x}^*$, which correspond to test inputs $\mathbf{s}^*$, can be sampled from this joint posterior by generating samples of a multivariate normal with the mean and covariance matrix from (4.32).

If the prediction is to be performed from noisy observations, assuming additive independent identically distributed noise with variance σ_n^2, the conditional distribution corresponding to (4.32) becomes:

$$\mathbf{x}^* \mid \mathbf{x}, \mathbf{s}, \mathbf{s}^* \sim N_{n^*}\big(\mathbf{K}_{s^*s}[\mathbf{K}_s + \sigma_n^2\mathbf{I}]^{-1}\mathbf{x}, \quad \mathbf{K}_{s^*} - \mathbf{K}_{s^*s}[\mathbf{K}_s + \sigma_n^2\mathbf{I}]^{-1}\mathbf{K}_{ss^*}\big). \tag{4.33}$$

4.2.5 Optimization of hyperparameters

The log (marginal) likelihood allows us to make inferences about all the hyperparameters in the light of the data as it represents the probability of the data given the hyperparameters. The log-likelihood (4.16) consists of three terms: The first term, $-\frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{K}_s^{-1}\mathbf{x}$, is the only term which depends on the training set output values and plays the role of a data-fit measure; it is a Mahalanobis distance between the model predictions and the data [Shahriari et al., 2016]. The second term, $-\frac{1}{2}\log \mid \mathbf{K}_s \mid$, is a complexity penalty term, which measures and penalizes the complexity of the model [Rasmussen, 2004]. Smoother covariance matrices will have smaller determinants and therefore lower complexity penalties [Shahriari et al., 2016]. The third and last term is simply a log normalization term, independent of data and hyperparameters. Note that the trade-off between penalty and data-fit in the GP model is automatic. Choosing the hyperparameters via optimisation of the (marginal) likelihood becomes then a widely used approach in the literature, partly because of the intuitive motivation of maximizing the probability of occurrence and partly because of its strong asymptotic properties (consistency and efficiency) [Robert, 2001].

Nonetheless, the topic of whether the hyperparameters should be optimised or drawn from their posterior distribution has been the long-standing debate as to whether Bayesian or frequentist methods are more desirable. Frequentists are often unhappy about the setting of priors, claiming them to be 'arbitrary' and hence the Bayesian framework of questionable worth for any form of comparison. Even if the Bayesian theory is accepted, it may be considered computationally impractical [Rasmussen, 1996]. On the other hand, Bayesians claim their models are superior since the incorporation of cogent prior knowledge can have a significant effect.

Any prior placed on an object is determined by the prior knowledge of that object and, hence, objective not arbitrary.

Applying Bayesian methods, however, is not always straightforward. The problem with priors lies not with subjectivity but with the challenging process of assigning prior probabilities when prior beliefs may be very hard to express in terms of the models [Gibbs, 1998]. On the other hand, maximum likelihood estimation (MLE) has frequently been suggested as a way to optimise the hyperparameters. MLE was recommended, analysed and widely popularized by Ronald Fisher between 1912 and 1922 [Aldrich, 1997]. Since then [Mardia and Marshall, 1984], it has been widely used for the estimation of hyperparameters in Gaussian Processes. In this research hyperparameters are chosen by MLE.

The lip curve is represented by 24 highly correlated points (due to their smoothness). This causes the covariance matrix $\mathbf{K}_s$ (in the noise-free case) to be potentially ill-conditioned, making numerical calculation of its inverse (required to calculate the log-likelihood) unstable. Spectral decomposition was used to get around this problem.

4.2.5.1 Spectral decomposition

Spectral decomposition, also known as eigen-decomposition, permits any symmetric, positive semi-definite matrix $\mathbf{A}$ to be expressed in terms of its eigenvalues and eigenvectors as follows:

$$\mathbf{A} = \sum_{i=1}^{n} \lambda_i \mathrm{u}_i \mathrm{u}_i^{\mathrm{T}} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^{\mathrm{T}}, \tag{4.34}$$

where the matrix $\mathbf{U} = [\mathrm{u}_1, \ldots, \mathrm{u}_n]$ is orthogonal (i.e., $\mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{U}\mathbf{U}^{\mathrm{T}} = \mathbf{I}_n$) and contains the eigenvectors of $\mathbf{A}$, and the diagonal matrix $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_{\mathrm{n}})$ contains the eigenvalues of $\mathbf{A}$. The inverse and determinant of $\mathbf{A}$ can be expressed as:

$$\mathbf{A}^{-1} = \sum_{i=1}^{n} \frac{1}{\lambda_i} \mathrm{u}_i \mathrm{u}_i^{\mathrm{T}} \qquad \det \mathbf{A} = \prod_{i=1}^{n} \lambda_i. \tag{4.35}$$

When working with an ill-conditioned matrix, the inverse and determinant could be approximated by discarding those eigenvalues (and corresponding eigenvectors) that are very small or even negative (note that the true eigenvalues must be non-negative, and so the negative values reflect the lost of accuracy of the numerical analysis). Small eigenvalues have a large contribution to the matrix $\mathbf{A}^{-1}$. However,

paradoxically, as explained by Press et al. [1992], setting these eigenvalues to zero is advantageous since it is finding an approximate solution of $\mathbf{Ax} = \mathbf{b}$ for an ill-conditioned $\mathbf{A}$, which has small residuals $|\mathbf{Ax} - \mathbf{b}|$. The problem now consists in determining how small the eigenvalue has to be in order to be disregarded, and hence, to select the most appropriate number of eigenvalues to use.

4.2.5.2 Choosing the number of eigenvalues

If all 24 eigenvalues could be used, the targets would be fitted exactly (in the noise free case). By investigating the surface of the log-likelihood, it was found that it was multimodal for a large number of eigenvalues, making the optimisation unreproducible when starting from different initial values of the hyperparameters. That is another reason to use spectral decomposition, as a smaller number of eigenvalues results in a smoother log-likelihood function. Nonetheless, having too few eigenvalues makes the surface overly flat, which also complicates the search for optimal hyperparameters. The study of the eigenvalues for different sets of values for the hyperparameters showed that usually only the first three or four eigenvalues are large-positive, dropping rapidly to close to zero or becoming negative within the first 10 eigenvalues (Figure 4.3). As the data are assumed noise free, the only two

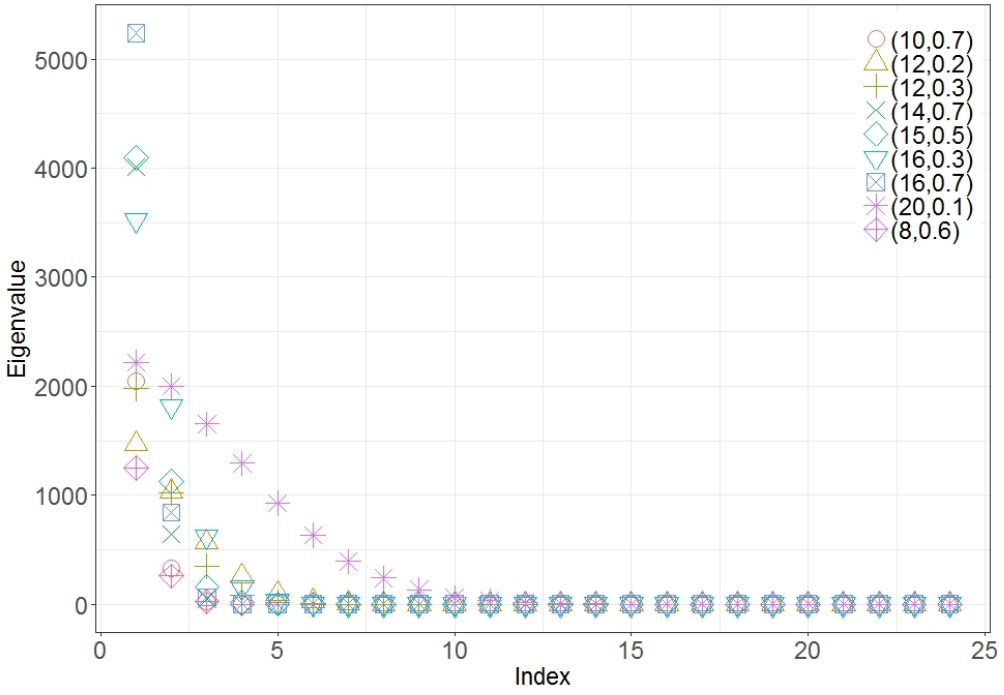


FIGURE 4.3: Values of the 24 eigenvalues (indexed from largest to smallest) for 9 pairs of hyperparameters (σ_f, λ)

parameters to estimate are σ_f and λ. To decide exactly how many eigenvalues to use, a number of simulations were performed with different values for σ_f and λ. For each pair of chosen values, 1000 curves were simulated. As the problem arose from the lip curves, where the arc-length s consists of 24 equally spaced points between zero and one, it was decided to use this same input. For each curve, starting points for the optimisation were randomly selected from the real values plus an error. The error for σ_f was drawn from a normal distribution with mean zero and standard deviation two ($N(0, 2^2)$), whereas for λ it was from a $N(0, 0.15^2)$. The absolute value of λ is taken at the end (since the models with λ and $-\lambda$ are equivalent). From these starting points, optimisation was done with the Conjugate Gradients method coded in the R-function *optim*. Details of the Conjugate Gradients method are given by [Press et al., 1992]. The optimisation was done six times, with the spectral approximation done with five to ten eigenvalues (recall, from Figure 4.3 only the first 5 to 10 eigenvalues are clearly positive).

Nine sets of simulations were performed with values of σ_f between 8 and 20 and of λ between 0.1 to 0.9. For each set of curves, the mean squared error (MSE) of all the estimators was calculated, for each number of retained eigenvalues. The bias of the estimators were also calculated. The nine sets were: $(\sigma_f, \lambda) = (8, 0.6), (10, 0.7), (12, 0.2), (12, 0.3), (14, 0.7), (15, 0.5), (16, 0.3), (16, 0.7), (20, 0.1)$.

One of the first observations was that the smaller the scale-length parameter, λ, is, the more the R-function *optim* tends to fail. This is because smaller values of λ indicate more independent points, and hence, the log-likelihood spikes closer to zero. Furthermore, when using fewer eigenvalues, the log-likelihood surface becomes flatter and therefore global maxima are harder to locate. Both problems occurring together result in extremely large 'optimal' values of the hyperparameters. For this reason, some simulated data sets had to be removed from a few of the simulations, particularly in those with smaller λ. All the information from that particular simulated data set was disregarded, not only the problematic estimation results, so as not to make the comparison between numbers of eigenvalues unbalanced. Figure 4.4 shows the values of the MSEs for each pair of (σ_f, λ) for the 5 to 10 eigenvalues used, ignoring just those cases were the MSE was over 1000 (hence some lines being incomplete). Figure 4.5 shows corresponding bias. The number of problematic datasets was reduced and didn't cause concern.

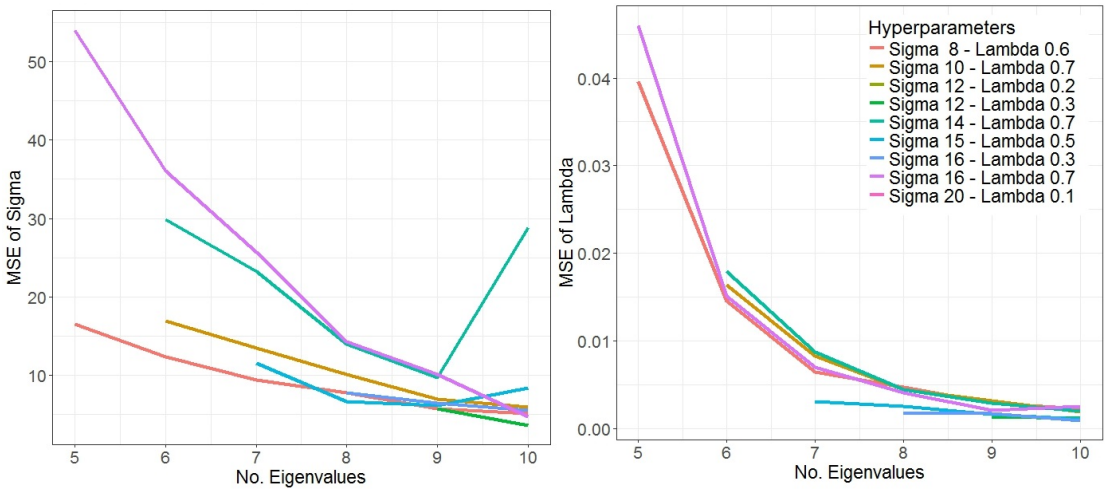


FIGURE 4.4: MSEs for σ_f and λ estimators for varying numbers of retained eigenvalues.

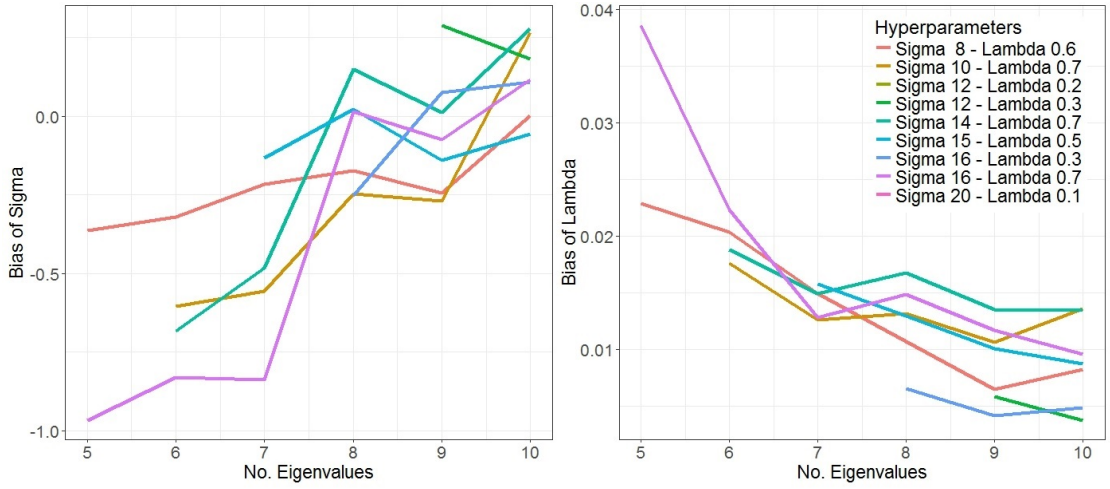


FIGURE 4.5: Bias for σ_f and λ estimators for varying numbers of retained eigenvalues.

The mean MSEs and bias across simulations (after removing failed cases) are shown in Table 4.2.5.2.

No. Eigenvalues	5	6	7	8	9	10
MSE for σ_f	32.9650	23.7577	15.2405	12.0447	9.5264	10.1868
MSE for λ	0.0343	0.0151	0.0067	0.0044	0.0027	0.0021
Bias for σ_f	−0.9899	−0.9386	−0.5257	−0.4118	−0.2810	−0.1451
Bias for λ	0.0060	−0.0038	0.0023	0.0002	0.0010	0.0016

TABLE 4.1: Mean MSE and bias by number of retained eigenvalues.

Approximate 95% Wald confidence intervals for MSE and bias were calculated too, in the case of the MSE (with n^* cases removed), using:

$$95\%\text{ CI for MSE} = \text{MSE} \pm 1.96 \times \frac{\text{SD(se)}}{\sqrt{n - n^*}}, \tag{4.36}$$

where SD(se) is the standard deviation of the squared errors and n is the total number of simulations (1000). The 95% CI for the Bias is done equivalently. Figure 4.6 shows the MSEs and bias together with the 95% CIs.

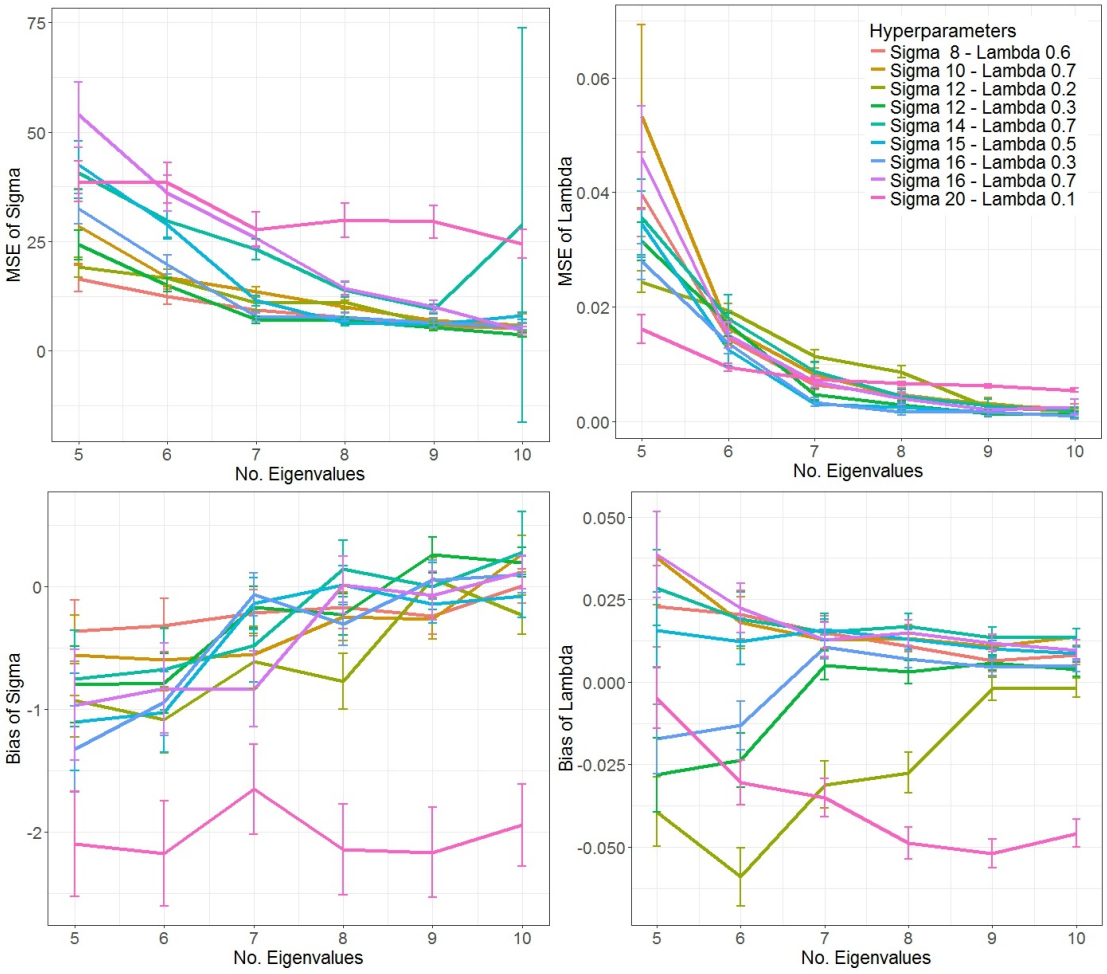


FIGURE 4.6: MSEs and Bias for σ_f and λ estimators with error bands.

Although the differences between nine or ten eigenvalues are not large, when using ten eigenvalues there are cases in which the last one is already negative (or close to zero) and is in fact disregarded[1]. This motivates the final choice of nine eigenvalues.

[1]The log-likelihood function is programmed to exclude negative eigenvalues regardless of the number asked to use and gives a notification of how many have been used in the end.

4.2.6 Fitting the model for one coordinate

As previously mentioned, the hyperparameters are optimised by maximum likelihood (MLE). Given the assumption of zero expectation, the curves have their mean subtracted. For the x coordinate, a series of images were studied, giving values for the hyperparameter σ_f around 15 with standard error (SE), estimated from the Hessian matrix, of about four or five units. The values for λ vary between 0.26 and 0.3, with SE of approximately 0.02 units.

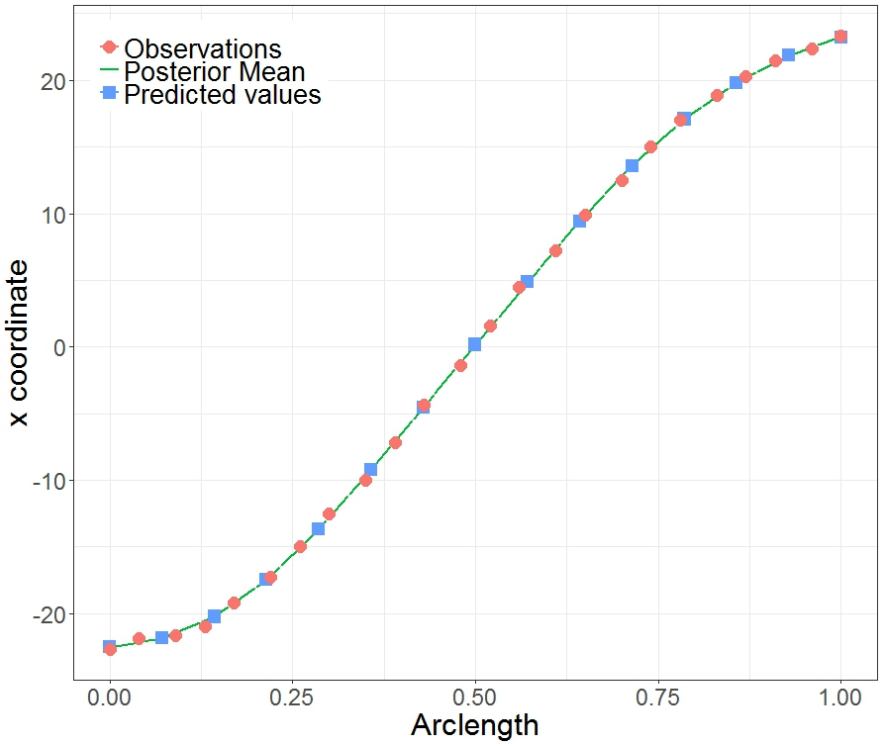


FIGURE 4.7: Fitted GP to x coordinate of an upper lip curve. Observed values (orange), fitted GP posterior mean (green), one set of sampled predictions (blue) and two SE confidence bands (shaded area).

Once the optimal hyperparameters were found, the fitted GP permits interpolation/prediction[2]. With a test input of 15 points between 0 and 1, and the joint posterior distribution (4.32), predicted values can be generated. Figure 4.7 shows the fit of x for a resting upper lip. The optimal hyperparameters found are: $\hat{\theta} = (\hat{\sigma}_f, \hat{\lambda}) = (16.51, 0.31)$, with SE 5.42 and 0.02 respectively. Given the optimal hyperparameters and the test points, the posterior mean and covariance of the process can be calculated, and so, a sample from the process can be drawn. The

[2]Predictions/interpolations are based on the estimated $\hat{\sigma}_f$ and $\hat{\lambda}$, and so, they ignore uncertainty in their values. A Bayesian approach would marginalize over this uncertainty.

fact that the 2 standard errors confidence bands (shaded area) can be barely seen is due to the fact that noise-free data has been assumed.

The y and z coordinates data can each be treated separately. Figure 4.8 shows the results for y and z coordinate for a resting upper lip curve. For y, $\hat{\sigma}_f = 6.33$ and $\hat{\lambda} = 0.29$ with SE 2.25 and 0.03 units, respectively. For z, $\hat{\sigma}_f = 19$ and $\hat{\lambda} = 0.35$ with SE 6.98 and 0.03 units, respectively. Samples were drawn from the posterior distribution of each process, given the optimal hyperparameters and using the same test points as for the x coordinate.

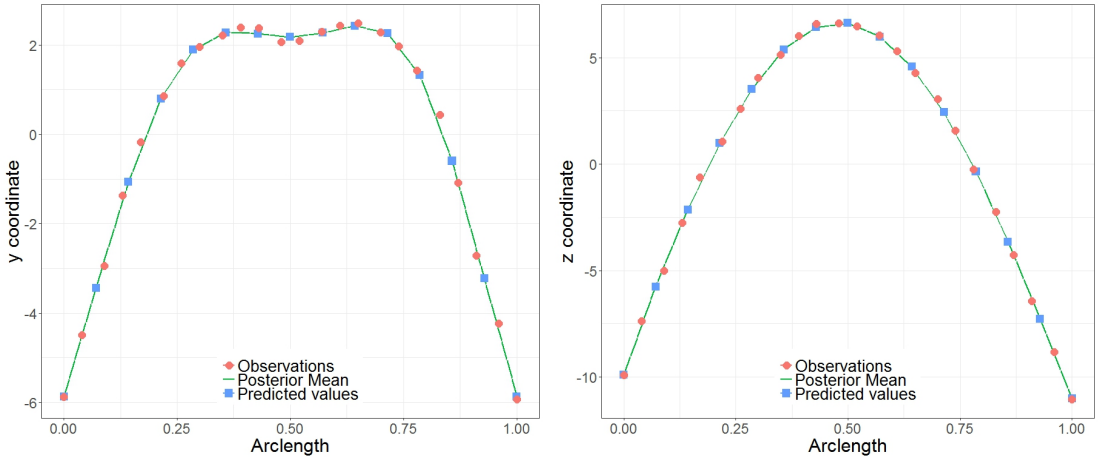


FIGURE 4.8: Fitted GP to y (left) and z (right) coordinates of an upper lip curve. For each: observed values (orange), fitted GP posterior mean (green), one set of sampled predictions (blue) and two SE confidence bands (shaded area).

4.3 Gaussian Process model for 3D curves (lip curves)

In the previous section, separate GP models were fitted independently for each of the three coordinates of an upper lip curve. However, given that the three coordinates are naturally related, it might be helpful to jointly model them using a mixed GP model. This can be achieved using a mixed GP model for the continuous spatial index (the arc-length of the curve, rescaled to be from 0 to 1), $s \in [0, 1]$, and the discrete label (the coordinate), $c \in \{x, y, z\}$. The GP, r, is defined as:

$$r(s, c) \sim GP\big(m(s, c), k(s, s', c, c')\big). \tag{4.37}$$

This represents each coordinate as a function of the arc-length: $r(s,x) = x(s)$, $r(s,y) = y(s)$ and $r(s,z) = z(s)$. Figure 4.9 shows each of the 24 points (equally spaced in s) from the upper lip in Figure 4.1, for each coordinate (x, y, z), plotted against the arc-length, s, of the curve, rescaled from 0 to 1.

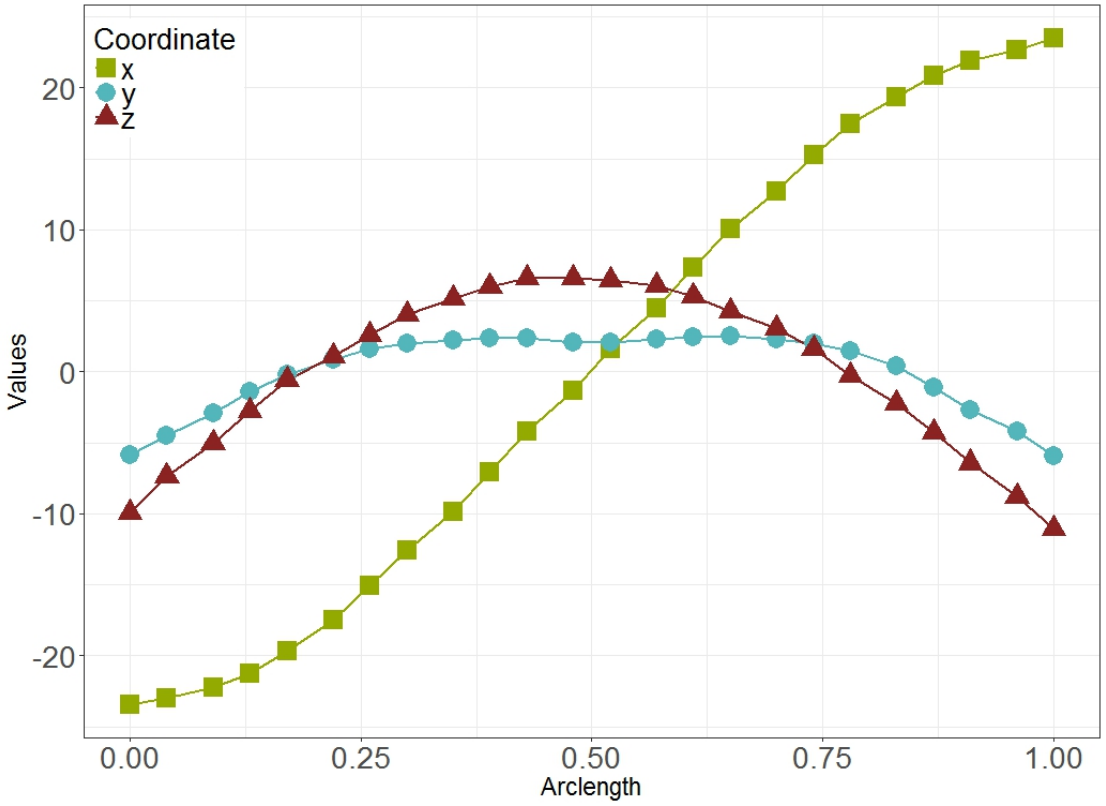


FIGURE 4.9: Upper lip 3 coordinates (x, y, z) plotted against arc-length s.

Let: $\mathbf{s} = (s_1 \cdots s_n)^{\mathrm{T}}$ for a choice of n values of s and $\mathbf{x} = \big(x(s_1) \cdots x(s_n)\big)^{\mathrm{T}}$, $\mathbf{y} = \big(y(s_1) \cdots y(s_n)\big)^{\mathrm{T}}$ and $\mathbf{z} = \big(z(s_1) \cdots z(s_n)\big)^{\mathrm{T}}$. Then:

$$\mathbf{r} = \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \\ \mathbf{z} \end{bmatrix} \sim N_{3n}\left(\mathbf{m}, \mathbf{K}\right), \tag{4.38}$$

where $\mathbf{m} = m(\mathbf{s}, c) = \big(m(s_1,x), \cdots, m(s_n,x), m(s_1,y), \cdots, m(s_n,y), m(s_1,z), \cdots, m(s_n,z)\big)^{\mathrm{T}}$ is the mean and $\mathbf{K}$ is the covariance matrix. Separability is assumed such that:

$$k(s,s',c,c') = k_s(s,s')k_c(c,c'). \tag{4.39}$$

The space-covariance function used is, as in Section 4.2, the Squared-Exponential (SE), i.e., $k_s(s, s') = \sigma_f^2 \exp\left(-\frac{1}{2\lambda^2}(s - s')^2\right)$, with hyperparameters: σ_f^2, the signal variance and λ, the length-scale. The full covariance matrix is then the Kronecker product of the 'coordinate' covariance matrix $\mathbf{K}_c$ and the 'spatial' covariance matrix $\mathbf{K}_s$: $\mathbf{K} = \mathbf{K}_c \otimes \mathbf{K}_s$, where the $(i, j)^{th}$ element of $\mathbf{K}_s$ is equal to $k_s(s_i, s_j)$. For the 3×3 matrix $\mathbf{K}_c$, given the nature of the lip curves, where the y and z coordinates seem highly correlated in space, two hyperparameters were specified: κ_1, the correlation between x and y or z, and κ_2, between y and z:

$$\mathbf{K}_c = \begin{pmatrix} 1 & \kappa_1 & \kappa_1 \\ \kappa_1 & 1 & \kappa_2 \\ \kappa_1 & \kappa_2 & 1 \end{pmatrix}. \tag{4.40}$$

One important fact to take into consideration is that for certain combinations of κ_1 and κ_2, $\mathbf{K}_c$ is not positive definite. The determinant of $\mathbf{K}_c$ needs to be positive and so $(\kappa_1^2 + \kappa_2^2 - 2\kappa_1^2\kappa_2)/(1 - \kappa_1^2)$ needs to be smaller than one. This means that $-1 < \kappa_1 < 1$ and κ_2 has to be contained in the interval $(2\kappa_1^2 - 1, 1)$.

The mean is assumed to be zero as before and therefore:

$$\mathbf{r} \sim N_{3n}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K}_s & \kappa_1\mathbf{K}_s & \kappa_1\mathbf{K}_s \\ \kappa_1\mathbf{K}_s & \mathbf{K}_s & \kappa_2\mathbf{K}_s \\ \kappa_1\mathbf{K}_s & \kappa_2\mathbf{K}_s & \mathbf{K}_s \end{bmatrix}\right). \tag{4.41}$$

4.3.1 Conditional dependences

The joint distribution of $\mathbf{y}$ and $\mathbf{x}$ is:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{x} \end{bmatrix} \sim N_{2n}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K}_s & \kappa_1\mathbf{K}_s \\ \kappa_1\mathbf{K}_s & \mathbf{K}_s \end{bmatrix}\right), \tag{4.42}$$

by the marginalization property. Hence, the conditional distribution of $\mathbf{y}$ given $\mathbf{x}$ is multivariate normal with mean $\kappa_1\mathbf{x}$ and covariance $(1 - \kappa_1^2)\mathbf{K}_s$.

Similarly, to calculate the distribution of $\mathbf{z}$ given the other two coordinates, the joint distribution is as in (4.41), rearranged so that the dependent variable is first:

$$\begin{bmatrix} \mathbf{z} \\ \mathbf{x} \\ \mathbf{y} \end{bmatrix} \sim N_{3n} \left(\mathbf{0}, \begin{bmatrix} \mathbf{K}_s & \kappa_1 \mathbf{K}_s & \kappa_2 \mathbf{K}_s \\ \kappa_1 \mathbf{K}_s & \mathbf{K}_s & \kappa_1 \mathbf{K}_s \\ \kappa_2 \mathbf{K}_s & \kappa_1 \mathbf{K}_s & \mathbf{K}_s \end{bmatrix} \right). \tag{4.43}$$

Now, $\mathbf{z} \mid \mathbf{x}, \mathbf{y}$ is multivariate normal with mean

$$\begin{bmatrix} \kappa_1 \mathbf{K}_s & \kappa_2 \mathbf{K}_s \end{bmatrix} \begin{bmatrix} \mathbf{K}_s & \kappa_1 \mathbf{K}_s \\ \kappa_1 \mathbf{K}_s & \mathbf{K}_s \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}, \tag{4.44}$$

and covariance

$$\mathbf{K}_s - \begin{bmatrix} \kappa_1 \mathbf{K}_s & \kappa_2 \mathbf{K}_s \end{bmatrix} \begin{bmatrix} \mathbf{K}_s & \kappa_1 \mathbf{K}_s \\ \kappa_1 \mathbf{K}_s & \mathbf{K}_s \end{bmatrix}^{-1} \begin{bmatrix} \kappa_1 \mathbf{K}_s \\ \kappa_2 \mathbf{K}_s \end{bmatrix}. \tag{4.45}$$

The key for both calculations is the inverse of the covariance matrix between $\mathbf{x}$ and $\mathbf{y}$. The matrix can be written as the Kronecker product of the coordinate covariance of x and y and $\mathbf{K}_s$. Given that $(\mathbf{A} \otimes \mathbf{B})^{-1} = \mathbf{A}^{-1} \otimes \mathbf{B}^{-1}$,

$$\begin{bmatrix} \mathbf{K}_s & \kappa_1 \mathbf{K}_s \\ \kappa_1 \mathbf{K}_s & \mathbf{K}_s \end{bmatrix}^{-1} = \left[\begin{pmatrix} 1 & \kappa_1 \\ \kappa_1 & 1 \end{pmatrix} \otimes \mathbf{K}_s \right]^{-1} = \frac{1}{1-\kappa_1^2} \begin{bmatrix} 1 & -\kappa_1 \\ -\kappa_1 & 1 \end{bmatrix} \mathbf{K}_s^{-1} \tag{4.46}$$

After some algebra (Appendix A.1), it is concluded that:

$$\begin{aligned} \mathbf{x} &\sim N_n(\mathbf{0}, \mathbf{K}_s), \\ \mathbf{y} \mid \mathbf{x} &\sim N_n\big(\kappa_1 \mathbf{x}, (1-\kappa_1^2)\mathbf{K}_s\big), \\ \mathbf{z} \mid \mathbf{x}, \mathbf{y} &\sim N_n \left(\left[\{\kappa_1 - \kappa_1\kappa_2\} \mathbf{x} + \{\kappa_2 - \kappa_1^2\} \mathbf{y} \right] / \left[1 - \kappa_1^2\right], \right. \\ &\qquad \left. \left[1 - \{\kappa_1^2 + \kappa_2^2 - 2\kappa_1^2\kappa_2\} / \{1 - \kappa_1^2\}\right] \mathbf{K}_s \right). \end{aligned} \tag{4.47}$$

4.3.2 Likelihood

The log-likelihood of the process for a single observation of $\mathbf{r}$ is: $l(\boldsymbol{\theta}) = \log p(\mathbf{x}) + \log p(\mathbf{y} \mid \mathbf{x}) + \log p(\mathbf{z} \mid \mathbf{x}, \mathbf{y})$, where $\boldsymbol{\theta} = (\sigma_f, \lambda, \kappa_1, \kappa_2)$ is the set of hyperparameters,

and from (4.7):

$$\begin{aligned}
\log p(\mathbf{x}) &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{K}_s| - \frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{K}_s^{-1}\mathbf{x}, \\
\log p(\mathbf{y} \mid \mathbf{x}) &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{K}_s| - \frac{n}{2}\log(1-\kappa_1^2) - \\
&\quad \frac{1}{2(1-\kappa_1^2)}(\mathbf{y}-\kappa_1\mathbf{x})^{\mathrm{T}}\mathbf{K}_s^{-1}(\mathbf{y}-\kappa_1\mathbf{x}), \\
\log p(\mathbf{z} \mid \mathbf{x}, \mathbf{y}) &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{K}_s| - \frac{n}{2}\log\left(1 - \frac{\kappa_1^2+\kappa_2^2-2\kappa_1^2\kappa_2}{1-\kappa_1^2}\right) \\
&\quad - \frac{(\mathbf{z}-\bar{\mathbf{z}})^{\mathrm{T}}\mathbf{K}_s^{-1}(\mathbf{z}-\bar{\mathbf{z}})}{2\left(1-\left[\kappa_1^2+\kappa_2^2-2\kappa_1^2\kappa_2\right]/\left[1-\kappa_1^2\right]\right)},
\end{aligned} \tag{4.48}$$

where $\bar{\mathbf{z}}$ denotes the mean of $\mathbf{z} \mid \mathbf{x}, \mathbf{y}$: $\bar{\mathbf{z}} = [(\kappa_1 - \kappa_1\kappa_2)\mathbf{x} + (\kappa_2 - \kappa_1^2)\mathbf{y}]/(1-\kappa_1^2)$.

4.3.3 Predictive distributions

To predict the values of the coordinates at a set of test points $\mathbf{s}^* = (s_1^*, \ldots, s_{n^*}^*)$, the joint distribution of the training outputs $\mathbf{r}$ and the test outputs $\mathbf{r}^*$, i.e., the joint distribution of $(\mathbf{x}, \mathbf{y}, \mathbf{z})$ and $(\mathbf{x}^*, \mathbf{y}^*, \mathbf{z}^*)$, can be written as:

$$\begin{bmatrix} \mathbf{x}^* \\ \mathbf{y}^* \\ \mathbf{z}^* \\ \mathbf{x} \\ \mathbf{y} \\ \mathbf{z} \end{bmatrix} \sim N_{3n^*+3n}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K}_c \otimes \mathbf{K}_{s^*} & \mathbf{K}_c \otimes \mathbf{K}_{s^*s} \\ \mathbf{K}_c \otimes \mathbf{K}_{ss^*} & \mathbf{K}_c \otimes \mathbf{K}_s \end{bmatrix}\right), \tag{4.49}$$

where $\mathbf{K}_{s^*}$, $\mathbf{K}_{s^*s}$, $\mathbf{K}_{ss^*}$ are as before (Section 4.2.4). Then, (see Appendix A.2.1):

$$\begin{bmatrix} \mathbf{x}^* \\ \mathbf{y}^* \\ \mathbf{z}^* \end{bmatrix} \Bigg| \begin{bmatrix} \mathbf{x} \\ \mathbf{y} \\ \mathbf{z} \end{bmatrix} \sim N_{3n^*}\left(\begin{bmatrix} \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{x} \\ \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{y} \\ \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{z} \end{bmatrix}, \begin{bmatrix} 1 & \kappa_1 & \kappa_1 \\ \kappa_1 & 1 & \kappa_2 \\ \kappa_1 & \kappa_2 & 1 \end{bmatrix} \otimes \left[\mathbf{K}_{s^*} - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}\right]\right). \tag{4.50}$$

To make predictions one by one, the distributions of each predicted coordinate can be calculated with the same dependences assumed in Section 4.3.1. The conditional prediction distributions are: (see calculations in Appendix A.2.2)

$$\begin{aligned}
\mathbf{x}^* \mid \mathbf{x} &\sim N_{n^*}\left(\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{x}, \mathbf{K}_{s^*} - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}\right), \\
\mathbf{y}^* \mid \mathbf{x}^*, \mathbf{x}, \mathbf{y} &\sim N_{n^*}\left(\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{y} + \kappa_1[\mathbf{x}^* - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{x}],\right. \\
&\qquad \left.[1-\kappa_1^2][\mathbf{K}_{s^*} - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}]\right), \\
\mathbf{z}^* \mid \mathbf{x}^*, \mathbf{x}, \mathbf{y}^*, \mathbf{y}, \mathbf{z} &\sim N_{n^*}\Bigg(\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{z} + \frac{1}{1-\kappa_1^2}\left[\{\kappa_1 - \kappa_1\kappa_2\}\left\{\mathbf{x}^* - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{x}\right\} + \right. \\
&\qquad \left.\{\kappa_2 - \kappa_1^2\}\left\{\mathbf{y}^* - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{y}\right\}\right], \qquad (4.51) \\
&\qquad \left[1 - \frac{\kappa_1^2 + \kappa_2^2 - 2\kappa_1^2\kappa_2}{1-\kappa_1^2}\right]\left[\mathbf{K}_{s^*} - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}\right]\Bigg).
\end{aligned}$$

4.3.4 Optimization of hyperparameters

To study the model, the upper lip of one resting face was taken. Each coordinate data set of 24 points has its mean subtracted so that each is centred around zero. Figure 4.9 shows the values of the three coordinates against the arc-length (rescaled to be from zero to one) of the upper lip curve. Note that since $\mathbf{y}$ and $\mathbf{z}$ are very similar, they have similar correlation with $\mathbf{x}$. κ_1 is expected to be close to zero and κ_2 close to one.

In this scenario, maximisation of the likelihood is still rather unstable even with the use of spectral decomposition to approximate the inverse of $\mathbf{K}_s$. This is likely due to the smoothness of the data. Two different approaches were employed to optimise the parameters: 1) to thin the data so it is less correlated or 2) to include a noise term in the model. Thinning was motivated by the fact that sometimes high correlation of the data is counter productive in the sense that more points do not give any new information and create flat and smooth likelihood surfaces, where the maxima are harder to locate. It was found that the maximum number of spatial points that could be used with no need to approximate $\mathbf{K}_s$ was 12, half of the total number. The other approach was to add a noise term to the model of the observations (see 4.15), as this causes the ratio between the largest and the smallest eigenvalue to decrease and hence the correlation matrix is no longer ill-conditioned. This is not unreasonable since the data acquisition process is not

noise-free in practice (despite having assumed noise-free data in Section 4.2). With both approaches optimal values for the hyperparameters were found, using a grid search first, followed by a conjugate gradients search starting from the grid point with the highest likelihood.

4.3.5 Fitting the model for a lip curve

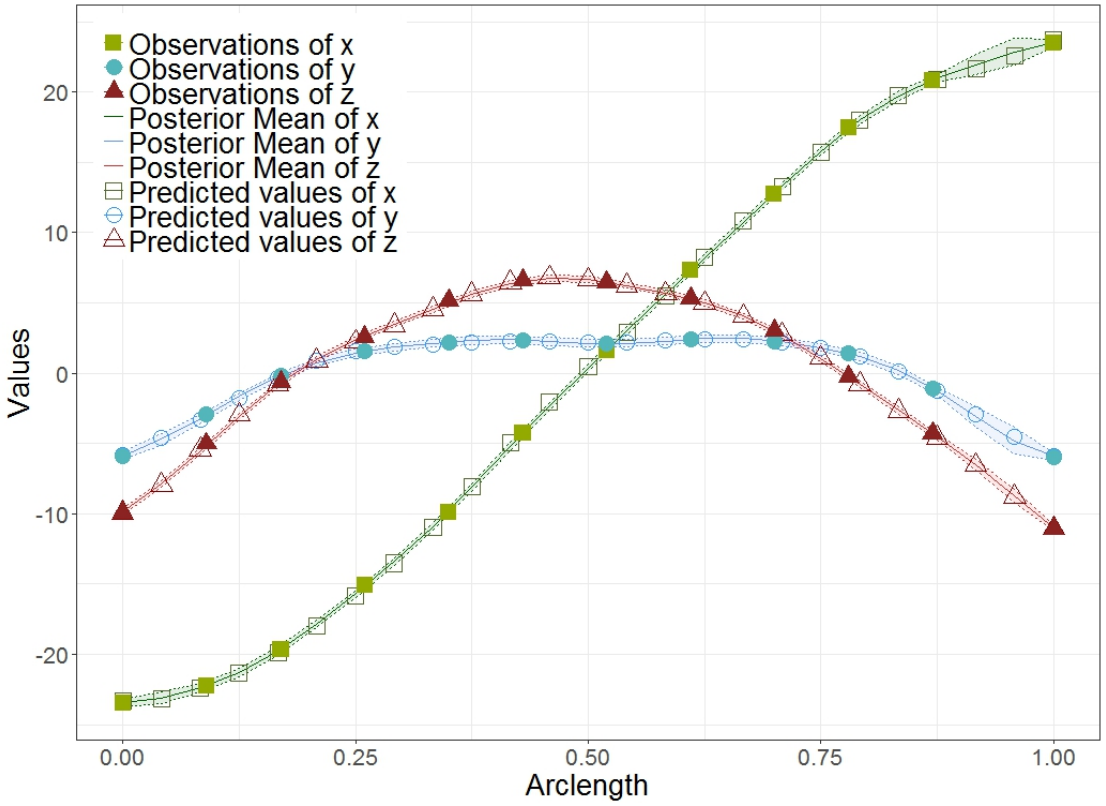


FIGURE 4.10: Observations of 12 upper lip points and predicted values for each coordinate together with the posterior prediction mean and two s.d. error bands.

For the thinning approach, every other point was omitted and both end points included. The indices of points selected are then $(1, 3, 5, 7, 9, 11, 13, 15, 17, 19, 21, 24)$. MLE was carried with no noise and no Spectral Decomposition approximation. The optimal hyperparameters found were $\hat{\boldsymbol{\theta}} = (\hat{\sigma}_f, \hat{\lambda}, \hat{\kappa}_1, \hat{\kappa}_2) = (9.0144, 0.1381, 0.0314, 0.8893)$, with respective SE: $(1.3333, 0.0058, 0.2619, 0.0656)$. It can be seen that the values of the correlation between coordinates correspond to previous intuition: $\hat{\kappa}_1$ is small and $\hat{\kappa}_2$ is large. Figure 4.10 shows the original data points and 25 predicted points using the posterior predictive distributions (4.51). The posterior predictive means are displayed with two standard errors bands. Note how above

$s = 0.8$, as the data points are more spaced out, the uncertainty between them increases, whereas for the rest of the sequence the bands can be barely discerned.

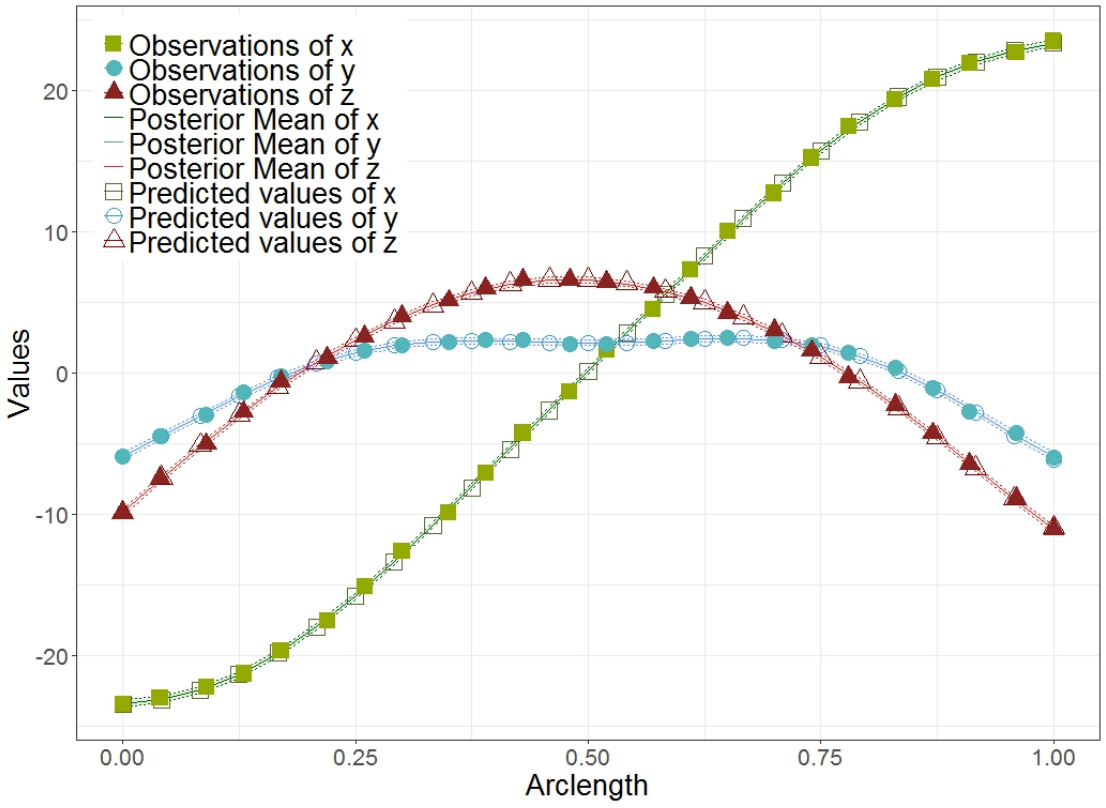


FIGURE 4.11: Observations and predicted values for each coordinate together with the prediction mean and two s.d. error bands for a full upper lip.

For the second (model noise) approach, it was key to establish what amount of noise would be reasonable. The noise variance could have been inferred by maximum likelihood, together with the rest of hyperparameters. However, given the computational time for the grid search and how unstable maximisation of the likelihood was, it was preferred to fix it to a value. It was decided that an error of 0.1 mm had little effect on the lip representation while making optimisation viable and was plausible for the apparatus. Again, MLE was carried with no Spectral Decomposition approximation. The optimal hyperparameters found were: $\hat{\boldsymbol{\theta}} = (\hat{\sigma}_f, \hat{\lambda}, \hat{\kappa}_1, \hat{\kappa}_2) = (9.2895, 0.3054, -0.0223, 0.4334)$, with respective SE: $(1.2867, 0.0168, 0.0939, 0.1215)$. It can be noted that the optimal values are consistent with those found by using fewer data points, except for λ. There is larger variability and length-scale and the correlation between y and z decreases. The posterior predictive mean (Figure 4.11) is very little changed. Since the noise is

small but all 24 of the points are being used, the error bands are too narrow to be properly appreciated.

4.4 Discussion

The use of shape information, expressed as curves via three functions of arc-length raises a number of very interesting issues from a methodological perspective. The model to express the three coordinates (each of them independently or all together) as functions of the arc-length interpolates the data well. Due to the smoothness of lip curves, the covariance matrix $\mathbf{K}_s$ is ill-conditioned, which made the optimisation of the hyperparameters a challenge. Different approaches to this issue were employed. For the one-dimensional curves, truncation by spectral decomposition was an effective approach that made optimisation viable. However, for the 3D joint analysis, the approximation was not enough. With both approaches, i.e., using fewer data points or adding noise to the model of the observations, it was found necessary to perform an initial grid search to locate the approximate volume where the hyperparameters lie. Nonetheless, the size of the grid grows exponentially with the dimension of the distribution. Moreover, each evaluation of the likelihood requires inversion of $\mathbf{K}_s$. Therefore, it is important to have restricted the number of possible values of $\boldsymbol{\theta}$ at which the log-likelihood is computed, while assuring the coverage of the relevant parts of the posterior distribution [Pietilainen, 2010].

It is clear from Figure 4.9 that the coordinates y and z are highly correlated, whereas the relationship between x and y or z is much weaker. The results from both approaches are consistent with these beliefs. However, adding noise to the model seems more reasonable, since it is likely there is actually noise in the data. The standard deviation of the noise should be chosen with care.

One line for further investigation that remains open would be to carry out a full Bayesian modelling, properly propagating posterior uncertainty about the hyperparameters into the predictive distributions, rather than conditioning those distributions on point estimates (in this case maximum likelihood estimates). An MCMC approach would permit sampling from the posterior distribution of $\boldsymbol{\theta}$.

Chapter 5

Gaussian Process models for k-dimensional curves evolving over time

5.1 Introduction and background

This chapter develops the scenario of Chapter 4, to the case when the curve is embedded in a three dimensional surface that now changes and is measured over time. A GP is proposed for evolving one-dimensional curves (representing one coordinate in terms of the arc-length) and then extended to three-dimensional curves. The aim is to learn how a three-dimensional curve evolves over time, and more specifically, how the shape of the lip changes during the performance of different emotions as expressed by the various correlation parameters of the GP.

5.1.1 Ornstein-Uhlenbeck processes

In Section 4.2.1, the concept of stationary process was introduced when presenting the squared exponential covariance function for the spatial variation (indexed by arc-length). For the evolution of the curves, the time component $\mathbf{t} = (t_1, ..., t_T)$ is defined. A process $\{Y_t : t \geq 0\}$ is said to be Markovian if, for all $t_1 < t_2 < \cdots < t_T$, $P(Y_{t_n} \leq y \mid Y_{t_2}, Y_{t_2}, \ldots, Y_{t_{n-1}}) = P(Y_{t_n} \leq y \mid Y_{t_{n-1}})$; the conditional probability distribution of future states of the process (conditional on both past and present

values) depends only upon the present state, that is, given the present, the future does not depend on the past. The most used Markovian processes are the class of Ornstein-Uhlenbeck (OU), which are the unique stationary first-order Gaussian Markov process [Rasmussen and Williams, 2006] and have the covariance function:

$$k_t(t, t') = \exp\left[-\frac{|\, t - t' \,|}{\mu}\right], \tag{5.1}$$

where the hyperparameter μ specifies the characteristic time scale for the evolutionary dynamics. Crudely, it explains how wiggly the function is in time. A large μ corresponds to a process where the change is slow. Given that $k_t(t, t')$ is a function of $|t - t'|$, it is isotropic.

Assuming equidistant points in time, and defining $\kappa = \exp(-1/\mu)$, it can be proved that the process is indeed Markov. The distribution of three time points is:

$$\begin{bmatrix} t_0 \\ t_1 \\ t_2 \end{bmatrix} \sim N_3 \left(\mathbf{0}, \begin{bmatrix} 1 & \kappa & \kappa^2 \\ \kappa & 1 & \kappa \\ \kappa^2 & \kappa & 1 \end{bmatrix} \right). \tag{5.2}$$

Following similar calculations to those in Section 4.3.1, it can be seen that:

$$\begin{aligned} t_0 &\sim N(0, 1), \\ t_1 \mid t_0 &\sim N\big(\kappa t_0, (1 - \kappa^2)\big), \\ t_2 \mid t_0, t_1 &\sim N\left(\kappa t_1, (1 - \kappa^2)\right). \end{aligned} \tag{5.3}$$

5.1.2 Principal Component Analysis

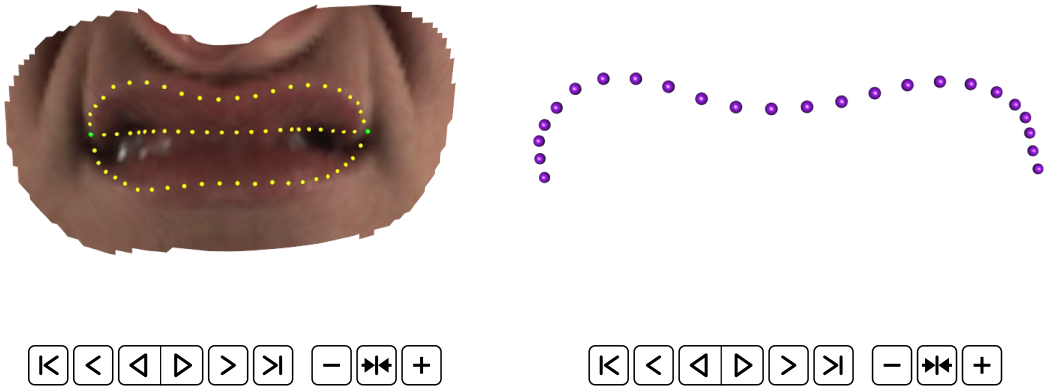
The idea of principal components analysis (PCA) was lightly introduced in Section 3.3.4. PCA is mathematically defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance of the data by some projection comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on. The desired goal is to reduce the dimensions of a d-dimensional dataset by projecting it into a k-dimensional subspace (where $k < d$) while retaining most of the information. The steps to perform PCA can be summarised as:

- Obtain the eigenvectors and eigenvalues of the covariance matrix or correlation matrix.

- Sort eigenvalues in descending order and choose the k eigenvectors that correspond to the k largest eigenvalues (where k is the number of dimensions of the new subspace).
- Construct the projection matrix $\mathbf{W}$ corresponding to the selected k eigenvectors.
- Transform the original dataset $\mathbf{X}$ via $\mathbf{W}$ to obtain $\mathbf{Y}$, the data approximated in a k-dimensional subspace.

5.2 Gaussian Process model for the evolution of 1D curves

Consider the case where the lip shape varies over the performance of an emotion (for example, *disgust*). Figure 5.1, in digital format, shows the 61 images for one of the sequences of *Disgust*. In the printed version, a time-point of the middle of the sequence is shown. The lip curves are shown lying on the facial manifold in Figure 5.1(a), and on their own in Figure 5.1(b). As in Chapter 4, each of these three-dimensional curves can be expressed as the value of each coordinate in terms of the arc-length.



(a) Facial surface with upper, lower and mid-line lip curves.

(b) Points on upper lip curve.

FIGURE 5.1: Sequence of upper lips for the emotion *disgust*.

Figure 5.2 shows some of the curves for the y coordinate, plotted against the arc-length (rescaled to be from 0 to 1). The y coordinate is probably the coordinate that best characterises an emotion, since it is the direction in which most change occurs. In the first set of curves of Figure 5.2, the change from the resting lip to the characteristic shape of *disgust* can be seen. The curves stay in that shape until around time point 55 out of 61, when the lip is in the resting position again.

A GP can be specified for each coordinate evolving through time. In this scenario, the observed values of, say, y depend on two variables: arc-length s (the spatial index, defined in the previous section) and time t.

The GP can be then defined as

$$y(s,t) \sim GP\big(m(s,t), k(s,s',t,t')\big). \tag{5.4}$$

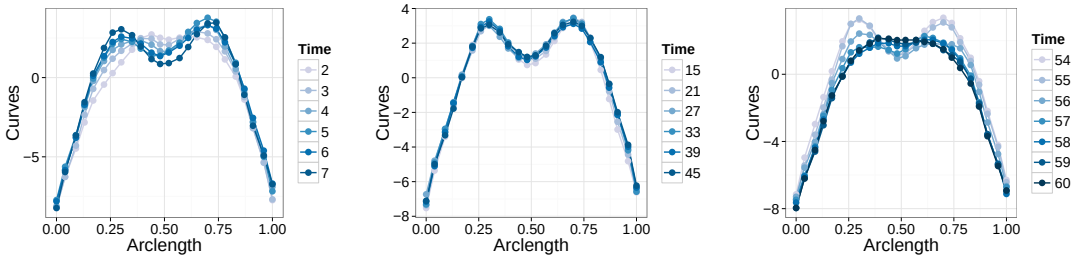


FIGURE 5.2: Samples of the y coordinate evolving during the performance of *disgust*.

The assumption of separability is used heavily in spatiotemporal statistics. Tests of separability for spatio-temporal covariances are reviewed in Mitchell et al. [2005] and Fuentes [2006]. If the spatio-temporal covariance has a specific parametric form, a likelihood ratio test can be derived. Other tests of separability include functional extensions of the Monte Carlo likelihood method or are based on quadratic forms [Constantinou et al., 2017]. This tests are, however, not straightforward. If separability is assumed, i.e., the covariance function is space-time separable, there exists a purely spatial covariance function k_s, and a purely temporal covariance function k_t, such that:

$$k(s,s',t,t') = k_s(s,s')k_t(t,t'). \tag{5.5}$$

It is also assumed that the process is Markovian in time, that is, the position of one lip curve at one particular time point in the sequence is dependent only on the immediately preceding time. Hence, the Ornstein-Uhlenbeck (OU) stationary covariance function is used:

$$k_t(t, t') = \exp\left[-\frac{\mid t - t' \mid}{\mu}\right]. \tag{5.6}$$

The time difference between adjacent time points is considered to be constant along the sequence. Since, specifically, it is assumed to be 1, $\kappa = \exp(-1/\mu)$ is defined.

The spatial covariance function is the Squared Exponential (SE), as in Chapter 4 (recall: $k_s(s, s') = \sigma_f^2 \exp\left(-\frac{1}{2\lambda^2}(s - s')^2\right)$). Note that the overall variance σ_f^2 is included in k_s, rather than k_t or both.

Let $\mathbf{s} = (s_1 \cdots s_n)^{\mathrm{T}}$ (as in Chapter 4) and $\mathbf{t} = (t_1 \cdots t_T)^{\mathrm{T}}$, for a choice of T values of t. Let also:

$$\mathbf{y} = \big(\mathbf{y}(t_1) \cdots \mathbf{y}(t_T)\big)^{\mathrm{T}}, \tag{5.7}$$

where

$$\mathbf{y}(t_i) = \big(y(s_1, t_i) \cdots y(s_n, t_i)\big)^{\mathrm{T}}, \tag{5.8}$$

represents the points on the curve at time t_i.

5.2.1 Conditional dependences

The joint distribution of $\mathbf{y}$ can be factorised into the marginal distribution of $\mathbf{y}(1)$, and all the conditional distributions of the subsequent $\mathbf{y}(i)$ given the immediately preceding $\mathbf{y}(i-1)$ (by the Markov property):

$$p(\mathbf{y}) = p(\mathbf{y}(1)) \prod_{i=2}^{T} p(\mathbf{y}(i) \mid \mathbf{y}(i-1)), \tag{5.9}$$

where,

$$\begin{aligned} \mathbf{y}(1) &\sim N_n(\mathbf{0}, \mathbf{K}_s), \\ \mathbf{y}(t) \mid \mathbf{y}(t-1) &\sim N_n\big(\kappa \mathbf{y}(t-1), (1-\kappa^2)\mathbf{K}_s\big), \quad (t \geq 2). \end{aligned} \tag{5.10}$$

Recall: $\mathbf{K}_s$ is the SE covariance matrix and $\kappa = \exp(-1/\mu)$ from the OU covariance matrix. The second result follows from the properties of conditional Multivariate Normal distributions (see (4.9)), writing the distribution of the tth curve and its predecessor as:

$$\begin{bmatrix} \mathbf{y}(t) \\ \mathbf{y}(t-1) \end{bmatrix} \sim N_{2n} \left(\mathbf{0}, \begin{bmatrix} \mathbf{K}_s & \kappa \mathbf{K}_s \\ \kappa \mathbf{K}_s & \mathbf{K}_s \end{bmatrix} \right), \tag{5.11}$$

Then, the distribution $\mathbf{y}(t)$ given $\mathbf{y}(t-1)$ is multivariate normal with mean $\kappa \mathbf{y}(t-1)$ and covariance $(1-\kappa^2)\mathbf{K}_s$.

5.2.2 Likelihood

Given (5.9), the total log-likelihood of the hyperparameters, $\boldsymbol{\theta} = (\sigma_f, \lambda, \mu)$, is:

$$\log p(\mathbf{y} \mid \boldsymbol{\theta}) = \log p(\mathbf{y}(1) \mid \boldsymbol{\theta}) + \sum_{i=2}^{T} \log p(\mathbf{y}(i) \mid \mathbf{y}(i-1), \boldsymbol{\theta}). \tag{5.12}$$

The marginal log-likelihood for the first curve is:

$$\log p\big(\mathbf{y}(1) \mid \boldsymbol{\theta}\big) = -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log \mid \mathbf{K}_s \mid -\frac{1}{2} \mathbf{y}(1)^{\mathrm{T}} \mathbf{K}_s^{-1} \mathbf{y}(1), \tag{5.13}$$

and the conditional log-likelihood for the ith curve given the $(i-1)$th is:

$$\begin{aligned} \log p\big(\mathbf{y}(i) \mid \mathbf{y}(i-1), \boldsymbol{\theta}\big) \quad &= -\frac{n}{2} \log(2\pi) - \frac{1}{2} \log \mid \mathbf{K}_s \mid -\frac{n}{2} \log(1-\kappa^2) \\ &\quad - \frac{1}{2(1-\kappa^2)} \big(\mathbf{y}(i) - \kappa \mathbf{y}(i-1)\big)^{\mathrm{T}} \mathbf{K}_s^{-1} \big(\mathbf{y}(i) - \kappa \mathbf{y}(i-1)\big). \end{aligned} \tag{5.14}$$

5.2.3 Predictive distributions

Along the sequence of curves that capture the emotion, marginal interpolations (5.15a) at time $q \in \{1, \ldots, T\}$ can be made at a set of test points $\mathbf{s}^* = (s_1^*, \ldots, s_{n^*}^*)$. Furthermore, predictions can also be made at different time points from those in the sequence: extrapolation can be done backwards (5.15b) for $q < 1$ or forward (5.15c) when $q > T$ or interpolation (5.15d) can be done for $t < q < t+1$, where times t and $t+1$ belong to the sample of time points, using, respectively:

$$\mathbf{y}^*(q) \mid \mathbf{y} \sim N_n \left(\mathbf{K}_{s^* s} \mathbf{K}_s^{-1} \mathbf{y}(q), \mathbf{K}_{s^*} - \mathbf{K}_{s^* s} \mathbf{K}_s^{-1} \mathbf{K}_{s s^*} \right), \tag{5.15a}$$

$$\mathbf{y}^*(q) \mid \mathbf{y} \sim N_n\left(\kappa^{1-q}\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{y}(1), \kappa^{2(1-q)}\mathbf{K}_{s^*} - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}\right), \tag{5.15b}$$

$$\mathbf{y}^*(q) \mid \mathbf{y} \sim N_n\left(\kappa^{q-T}\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{y}(T), \kappa^{2(q-T)}\mathbf{K}_{s^*} - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}\right), \tag{5.15c}$$

$$\mathbf{y}^*(q) \mid \mathbf{y} \sim N_n\left(\frac{\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}}{1-\kappa^2}\left[\{\kappa^{q-t} - \kappa^{t-q+2}\}\mathbf{y}(t) + \{\kappa(\kappa^{t-q} - \kappa^{q-t})\mathbf{y}(t+1)\}\right], \mathbf{K}_{s^*} - \frac{\kappa^{2(q-t)} + \kappa^{2(t-q)+2} - 2\kappa^2}{1-\kappa^2}\left[\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}\right]\right), \tag{5.15d}$$

where point estimates of the hyperparameters are plugged into k and the $\mathbf{K}$ matrices. The optimal hyperparameters are found for the sequence including all time points. However, it can be seen that prediction for the curve at time $q \in \{1, \ldots, T\}$ only depend on the data at time point q (5.15a). Equivalently, retrodiction (5.15b) and prediction (5.15c) only depend on the data from the first and last curves, respectively. Interpolation involves the data from the curves at time point t and $t+1$ (5.15d). Details are shown in Appendix B.1.

5.2.4 Inference of the hyperparameter μ

Study of this model started through optimisation of the hyperparameters for the y-coordinate in one of the sequences of the emotion *disgust* (Figure 5.2), where the points on each time-point curve have their mean subtracted. One of the first problems faced when starting optimisation of the hyperparameters was the uncertainty even about the order of magnitude of μ. Since the sequences evolve slowly, it is expected to be large. However, how large it should be is unknown. This problem was approached by focusing solely on the temporal OU process. To understand the behaviour of the OU process on its own, without the spatial component, a 'slice' of the evolving curves, at one particular space point can be modelled. The data for a single space point of the x coordinate of *disgust* can be seen in Figure 5.3.

A Markovian GP for one of the coordinates, say, x, at a particular value of s, characterised by the OU covariance function, can be written as:

$$x(t) \sim GP\big(m(t), k_t(t, t')\big), \tag{5.16}$$

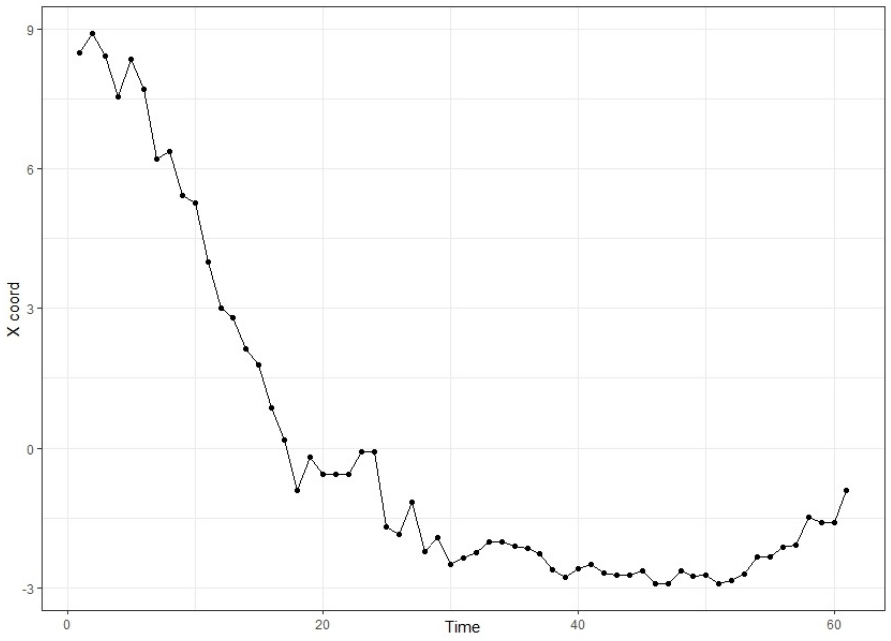


FIGURE 5.3: Evolution of the x coordinate values in the 17th space point.

where the mean is assumed to be zero and $k_t(t, t')$ is as defined in (5.6). The log-likelihood can then be calculated using:

$$\log p(x \mid t, \boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log \mid \sigma_f^2 \mathbf{K}_t \mid -\frac{1}{2}x^{\mathrm{T}}[\sigma_f^2 \mathbf{K}_t]^{-1}x, \tag{5.17}$$

where $\mathbf{K}_t$ is the covariance matrix for the T temporal points, with $(i, j)^{th}$ element equal to $k_t(t_i, t_j)$.

The log-likelihood surface was studied for different space points for each coordinate. Moreover, Wilks confidence regions for the true values of the parameters [in this case $\boldsymbol{\theta} = (\sigma_f, \mu)$] were calculated.

Figure 5.4 shows the log-likelihood surface for the model (5.16) at the 17th space point of the x-coordinate. The red contour surrounds the 95% Wilks confidence region. Note how large the region is and also that $\hat{\sigma}_f$ and $\hat{\mu}$ are strongly correlated. This shows that once μ is 'large', its effect on the log-likelihood values decreases, i.e., the log-likelihood becomes very flat. For each unit increase in the time-correlation parameter μ, the log-likelihood is modified only in its first decimal place. Similar behaviour was found at different space-points in different coordinates. Therefore, it must be assumed that μ will be large and its standard error also large. For this reason, it was decided to look for the maximum of μ on a

logarithmic scale, to ease the grid search, assuring a large enough coverage for the range of μ while maintaining computational efficiency.

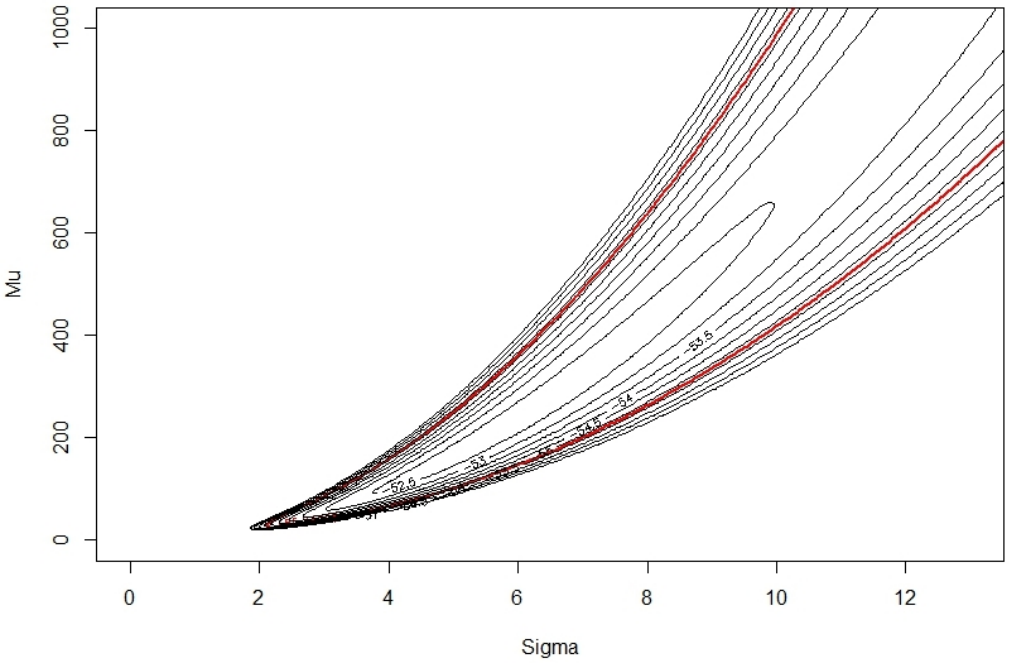
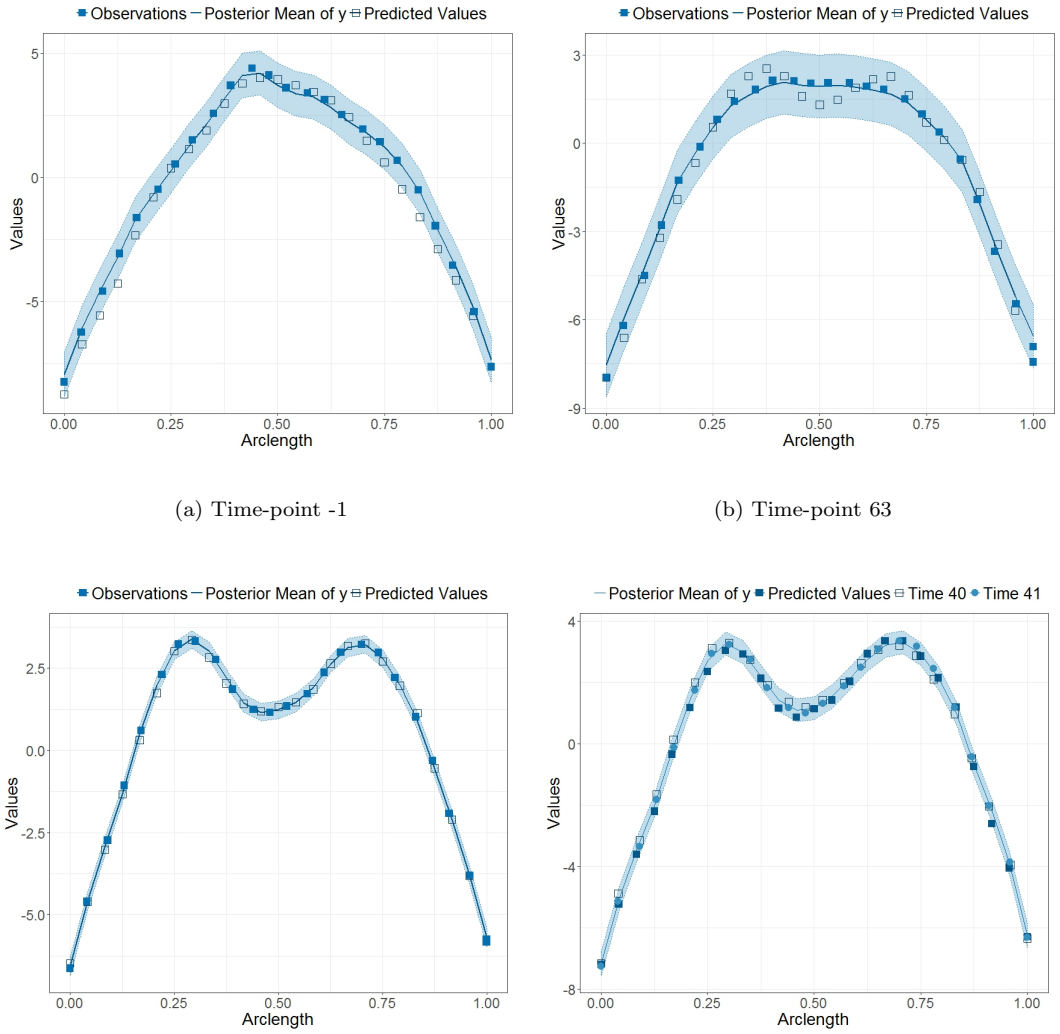


FIGURE 5.4: Contour plot of the log-likelihood surface for the x coordinate at the 17th space point.

5.2.5 Fitting the evolution model

A sequence of data from the upper lip for the emotion *disgust* consisting of 61 pictures, i.e., $\mathbf{t} = (1 \cdots 61)^T$, was used as a case test. We focussed on the y-coordinate, since it tends to be where the most striking changes occur. The data at each time-point had their mean subtracted. Optimal hyperparameters, $\hat{\boldsymbol{\theta}} = (\hat{\sigma}_f, \hat{\lambda}, \hat{\mu})$, were found by maximum likelihood: $\hat{\boldsymbol{\theta}} = (1.74, 677 \times 10^{-4}, 6.44 \times 10)$, with respective SE: $(0.0751, 9 \times 10^{-4}, 0.54 \times 10)$. A grid search was done first. Refinement of the values was done using the Conjugate Gradients method in the R function *Optim*. As in Section 4.3.4, a noise term with standard deviation 0.1 mm was added to the model to avoid an ill-conditioned spatial covariance matrix and descibe noise in the raw data.

Figure 5.5 shows the original data points and 25 predicted values for the test points at time points: -1 (retrodiction), 63 (prediction), 30 (at a time where data are observed) and 40.75 (interpolation between time points 40 and 41). The posterior means are displayed with 2 standard deviations bands (shown dotted).



(a) Time-point -1

(b) Time-point 63

(c) Time-point 30

(d) Time-point 40.75

FIGURE 5.5: Observations, posterior means and predicted values for the y coordinate curves of *disgust* at four time-points.

For the retrodiction (5.5(a)), since the sequence is considered to start at time 1, a 2 time-point jump was made backward, using the data available at time point 1. Equivalently, for prediction (5.5(b)), the jump was done 2 time-points ahead, using data from the last curve. The confidence bands are considerably wider than in the rest of predictions, given that the time interval from the observed values is larger. Prediction at time point 30 (5.5(c)), uses the data only from that time point. Interpolation (5.5(d)) uses the data observed at time points 40 and 41. In the last two cases, the confidence bands are much narrower and can be barely appreciated.

The theory can be applied to the x and z coordinates equivalently. Figure 5.6 shows the data for the x and z coordinates. The changes in the first are minimal, while the z coordinate shows more movement, especially at the middle values of the arc-length. The optimisation of the hyperparameters was carried out in the same way as for the y coordinate. A grid search was done for each coordinate, using a noise s.d. of 0.1 mm.

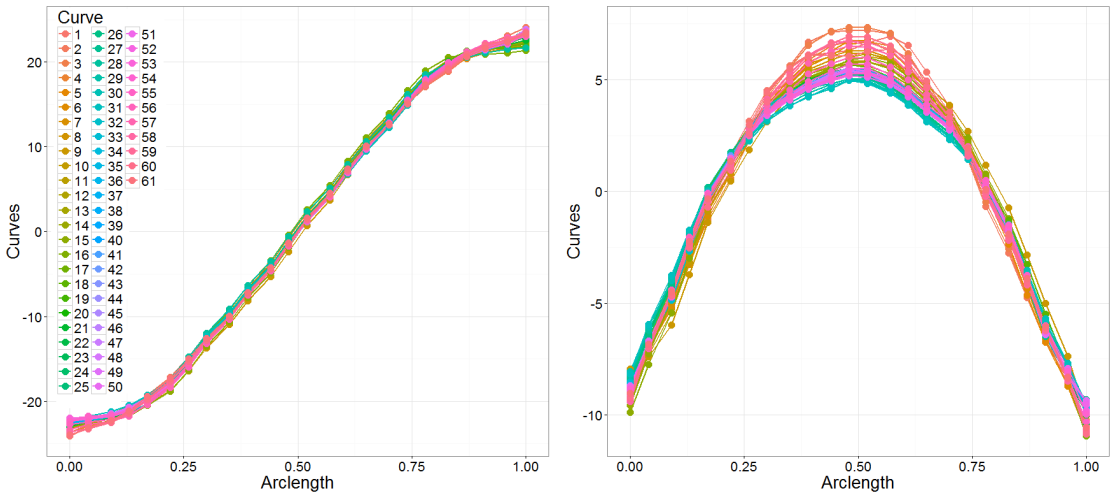
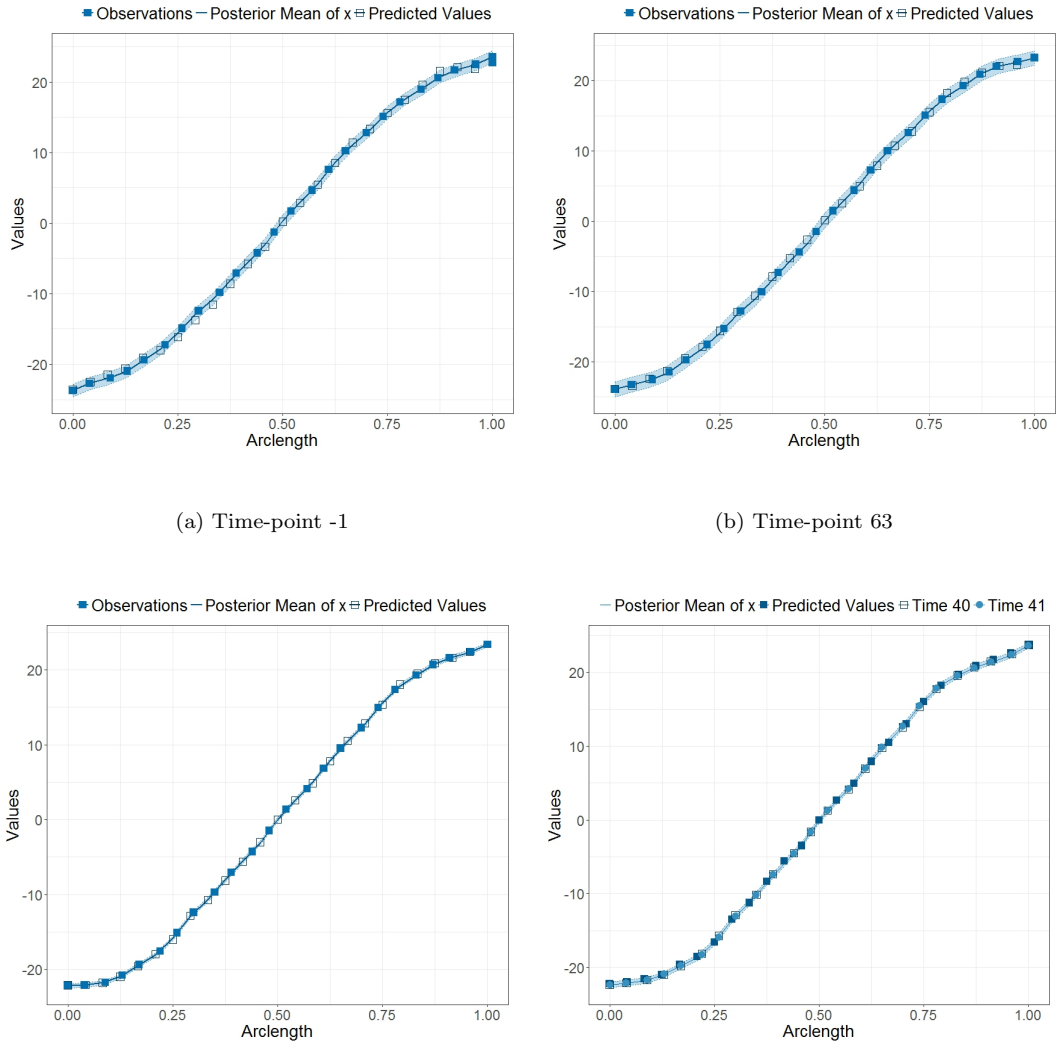


FIGURE 5.6: On the left, values of the x coordinate evolving over time. On the right, the z coordinate.

For the evolution of the x-coordinate curves in the upper lip of the emotion *Disgust*, the maximum likelihood estimates found were $\hat{\boldsymbol{\theta}} = (\hat{\sigma_f}, \hat{\lambda}, \hat{\mu}) = (10.45, 627 \times 10^{-4}, 2.60 \times 20^3)$, with respective SE: $(1.9, 5 \times 10^{-4}, 0.93 \times 10^3)$. The same time-points as for y were considered to make predictions, at 25 spatial-points. Figure 5.7 shows the original data points, together with the predicted values, the posterior means and the 2 standard deviations bands (shown dotted). In this case, the confidence bands can barely be appreciated due to the very small variability in the curves. Even when we move further away from the observed data in time, the bands are still narrow around the posterior mean, in contrast to the predictions for those time points in the y-coordinate.



(a) Time-point -1

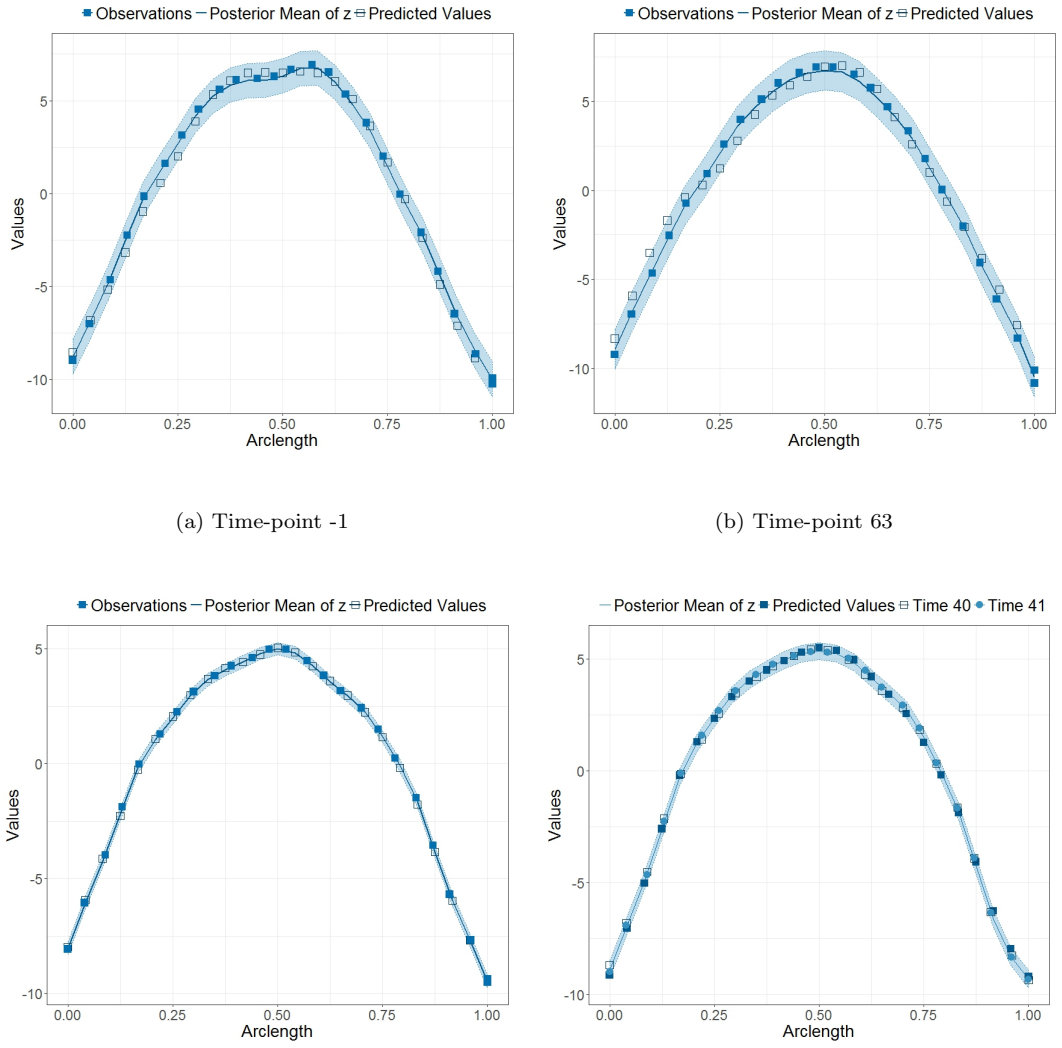
(b) Time-point 63

(c) Time-point 30

(d) Time-point 40.75

FIGURE 5.7: Observations, posterior means and predicted values for the x coordinate curves of *disgust* at four time-points.

The optimal hyperparameters found for the curves corresponding to the z coordinate were $\boldsymbol{\theta} = (\sigma_f, \lambda, \mu) = (2.29, 73 \times 10^{-3}, 10.5 \times 10)$, with respective SE: $(0.1, 1.1 \times 10^{-3}, 0.86 \times 10)$. Figure 5.8 shows the observed data points of the z-coordinate, together with the predicted values (at the same 25 equally spaced spatial points as for the x and y coordinates), the posterior means and the 2 standard deviations bands (shown dotted) which, as for the y-coordinate, expand as predictions are further away from the observations.



(a) Time-point -1

(b) Time-point 63

(c) Time-point 30

(d) Time-point 40.75

FIGURE 5.8: Observations, posterior means and predicted values for the z coordinate curve of *disgust* at four time-points.

5.2.6 Discussion

The sequences of pictures are taken at such short intervals that the lip curves expressing the emotion change rather slowly over time. As a result the hyperparameter μ may be large. Through experimentation it was observed that the larger the value of μ, the smaller its effect on the likelihood function. While for small μ values, an increase of one unit on it can alter the log-likelihood by large amounts, for larger values, an increase (or decrease) of 50 units may only modify the log-likelihood value by decimals. Thus, the grid search could be made more efficient

by choosing values for the parameter evenly on the log scale. This allows for the grid to cover a large enough range of values for μ, while maintaining computational feasibility having the grid more dense at small values and more sparse for larger values. Recall the faces have not been registered in advance, before extracting the curves. Registration may have an effect in this problem and should therefore be considered.

Each coordinate changes differently over time, with the y coordinate having the most characteristic change. The amount of change in each coordinate can be assessed by the hyperparameter μ, which is at its largest for the x coordinate (2597.3984), meaning there is much less variability in this set of curves, and therefore a higher correlation between different time points. Moreover, its associated SE is also much larger (933.6749 vs 5.4307 (y) and 8.64 (z)). This result shows once again the problems associated when the curves are highly correlated in time. The maximum value of the log-likelihood for the x coordinate curves is 1057.532; if μ is reduced by 100 units (2497.3984), the log-likelihood is reduced only by 1 unit (1056.296). If μ is increased by 100 units, the log-likelihood increases only by 0.2 unit (1057.7), reflecting the results from Section 5.2.4 where it was shown that the log-likelihood function becomes flat for large values of the time-scale parameter μ (Figure 5.4).

For the y and z coordinate, μ takes the values 64.4236 and 104.99, respectively. While both of these values are much smaller than for the x coordinate, the value for the z coordinate is still around 1.6 times the value of μ for the y coordinate, consistent with the observation that most of the characteristic movements in the lip curves during the emotion *disgust* occur on the y axis.

The process has been assumed Markovian for mathematical convenience in this section (and the following). Other choices may describe the data better. Despite innumerable studies rooted in Markov processes, there are few existing tests for the Markov property in the literature, summarised, together with the presentation of a new approach, in [Chen and Hong, 2012].

5.3 Gaussian Process model for the evolution of 3D curves

In the previous section, separate GP models have been fitted for each of the three coordinates of an upper lip curve evolving during the emotion *disgust*. The theory presented in Section 5.2 can be combined with the theory shown in Section 4.3 to model the evolution of the three evolving coordinates jointly. Figure 5.9 shows a three-dimensional upper lip curve changing along the emotion *disgust* in a sample of time point (the sequence consists of 61 pictures). Each coordinate value is plotted against the arc-length, rescaled from 0 to 1.

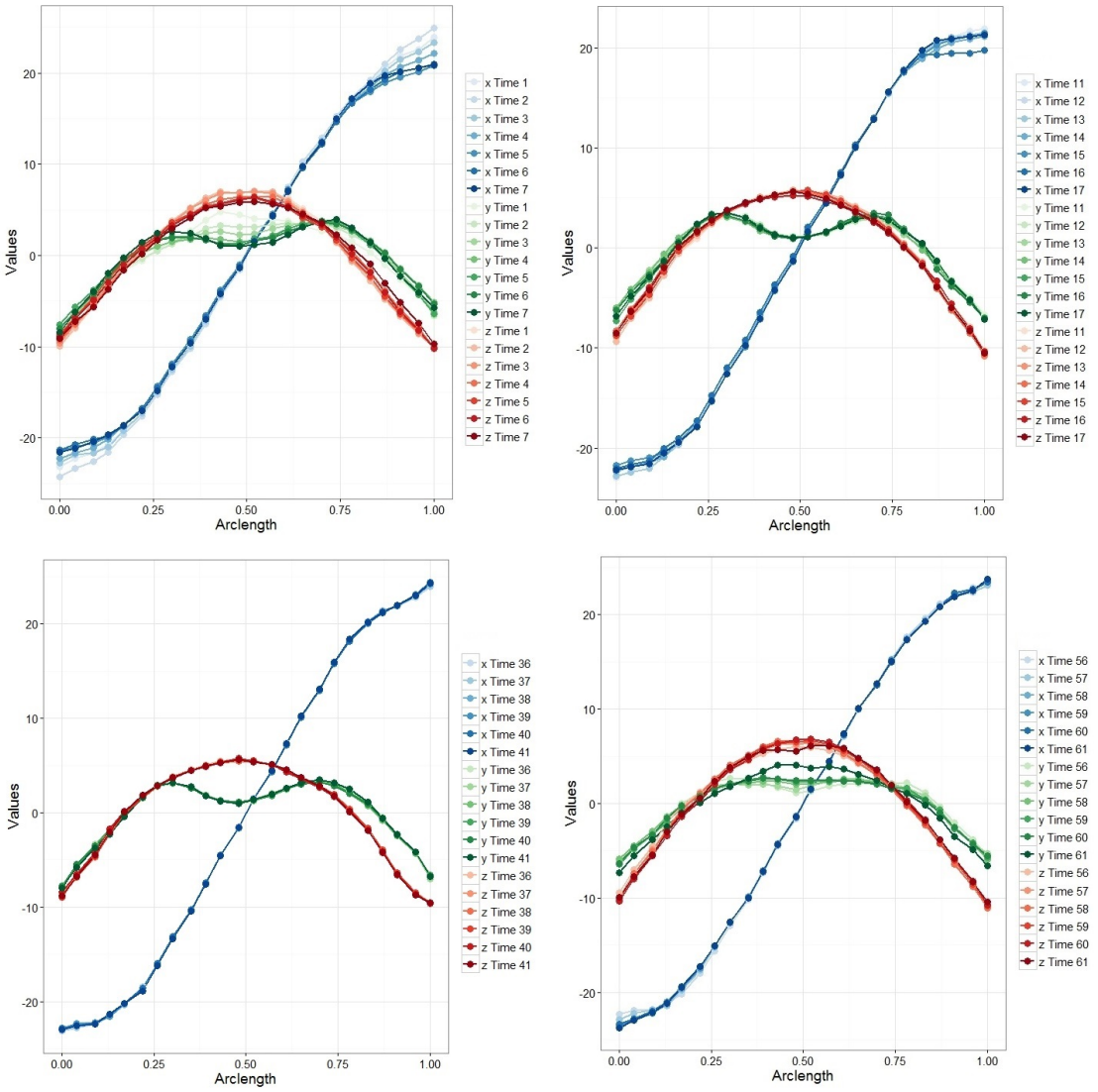


FIGURE 5.9: Samples of the x, y and z coordinate evolving during the performance of *disgust*.

A GP model for the evolving three coordinates along the lip curve can be specified as:

$$w(s,c,t) \sim GP\big(m(s,c,t), k(s,s',c,c',t,t')\big), \tag{5.18}$$

a mixed GP for the spatial component (indexed by the arc-length) $s \in [0,1]$, the discrete coordinate $c = \{x, y, z\}$ and the temporal component (indexed by the time t). The sampled three-dimensional curve at time t is notated as:

$$\mathbf{W}(t) = \begin{bmatrix} \mathbf{x}(t) \\ \mathbf{y}(t) \\ \mathbf{z}(t) \end{bmatrix}, \tag{5.19}$$

where, as defined in Section 5.2, $\mathbf{y}(t) = \big(y(s_1,t) \cdots y(s_n,t)\big)^{\mathrm{T}}$, and similarly for $\mathbf{x}(t)$ and $\mathbf{z}(t)$. The entire sequence of $\mathbf{W}$s stacked for all time points $(t_1, \ldots, t_T)$ where data are observed, is denoted by $\mathbf{W}$. Separability is assumed such that:

$$k(s,s',c,c',t,t') = k_t(t,t')k_c(c,c')k_s(s,s'). \tag{5.20}$$

The space-covariance function k_s used is, as before, the Squared-Exponential (SE), i.e. $k_s(s,s') = \sigma_f^2 \exp\left(-\frac{1}{2\lambda^2}(s-s')^2\right)$, with hyperparameters σ_f^2, the signal variance and λ, the length-scale. The 3×3 matrix $\mathbf{K}_c$ remains as in Section 4.3, with hyperparameters κ_1, the correlation between x and y or z, and κ_2, between y and z:

$$\mathbf{K}_c = \begin{pmatrix} 1 & \kappa_1 & \kappa_1 \\ \kappa_1 & 1 & \kappa_2 \\ \kappa_1 & \kappa_2 & 1 \end{pmatrix}. \tag{5.21}$$

As the process is assumed Markovian in time, the Ornstein-Uhlenbeck (OU) covariance function is used, i.e., $k_t(t,t') = \exp(-\mid t - t' \mid /\mu)$, with hyperparameter μ.

The mean is assumed to be zero and hence:

$$\mathbf{W} \sim N_{3nT}\left(\mathbf{0}, \mathbf{K}_t \otimes \mathbf{K}_c \otimes \mathbf{K}_s\right), \tag{5.22}$$

where $\mathbf{K}_s$ is the covariance matrix for the n arc-length inputs, with $(i,j)^{th}$ element equal to $k_s(s_i, s_j)$ and $\mathbf{K}_t$ is the covariance matrix for the T time points with $(i,j)^{th}$ element equal to $k_t(t_i, t_j)$. Since the time differences are considered constant along

the sequence, and assumed equal to one, using $\kappa = \exp(-1/\mu)$:

$$\mathbf{K_t} = \begin{bmatrix} 1 & \kappa & \kappa^2 & \kappa^3 & \cdots \\ \kappa & 1 & \kappa & \kappa^2 & \cdots \\ \kappa^2 & \kappa & 1 & \kappa & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \tag{5.23}$$

5.3.1 Conditional dependences

At the first time point marginally, the three coordinates follow the same model as in Section 4.3:

$$\mathbf{W}(1) = \begin{bmatrix} \mathbf{x}(1) \\ \mathbf{y}(1) \\ \mathbf{z}(1) \end{bmatrix} \sim N_{3n} \left(\mathbf{0}, \begin{bmatrix} \mathbf{K}_s & \kappa_1 \mathbf{K}_s & \kappa_1 \mathbf{K}_s \\ \kappa_1 \mathbf{K}_s & \mathbf{K}_s & \kappa_2 \mathbf{K}_s \\ \kappa_1 \mathbf{K}_s & \kappa_2 \mathbf{K}_s & \mathbf{K}_s \end{bmatrix} \right). \tag{5.24}$$

The joint distribution of all the $\mathbf{W}(t)$ for $t = (t_1, \ldots, t_T)$, can be factorised, using the Markov assumption in time, into the product of this marginal and all of the conditional distributions of subsequent times given the previous time. These conditional distributions an be obtained from:

$$\begin{bmatrix} \mathbf{W}(t) \\ \mathbf{W}(t-1) \end{bmatrix} \sim N_{6n} \left(\mathbf{0}, \begin{bmatrix} \mathbf{K}_c \otimes \mathbf{K}_s & \kappa \mathbf{K}_c \otimes \mathbf{K}_s \\ \kappa \mathbf{K}_c \otimes \mathbf{K}_s & \mathbf{K}_c \otimes \mathbf{K}_s \end{bmatrix} \right), \tag{5.25}$$

where $\kappa = \exp(-1/\mu)$, as the time difference between adjacent sampled curves is assumed to be one.

From (5.25):

$$\mathbf{W}(t) \mid \mathbf{W}(t-1) \sim N_{3n} \left(\kappa \begin{bmatrix} \mathbf{x}(t-1) \\ \mathbf{y}(t-1) \\ \mathbf{z}(t-1) \end{bmatrix}, (1-\kappa^2) \begin{bmatrix} \mathbf{K}_s & \kappa_1 \mathbf{K}_s & \kappa_1 \mathbf{K}_s \\ \kappa_1 \mathbf{K}_s & \mathbf{K}_s & \kappa_2 \mathbf{K}_s \\ \kappa_1 \mathbf{K}_s & \kappa_2 \mathbf{K}_s & \mathbf{K}_s \end{bmatrix} \right). \tag{5.26}$$

Moreover, the same conditional dependences as in Section 4.3 can be used ti further factorise the full joint distribution. From (5.26):

$$\begin{aligned}
\mathbf{x}(t) \mid \mathbf{W}(t-1) &\sim N_n\left(\kappa\mathbf{x}(t-1), (1-\kappa^2)\mathbf{K}_s\right), \\
\mathbf{y}(t) \mid \mathbf{x}(t), \mathbf{W}(t-1) &\sim N_n\big(\kappa\mathbf{y}(t-1) + \kappa_1(\mathbf{x}(t) - \kappa\mathbf{x}(t-1)), \\
&\quad (1-\kappa^2)\mathbf{K}_s(1-\kappa_1^2)\big). \qquad (5.27) \\
\mathbf{z}(t) \mid \mathbf{x}(t), \mathbf{y}(t), \mathbf{W}(t-1) &\sim N_n\Big(\kappa\mathbf{z}(t-1) + \frac{1}{1-\kappa_1^2}\Big[\big(\kappa_1 - \kappa_1\kappa_2\big)\big(\mathbf{x}(t) - \\
&\quad \kappa\mathbf{x}(t-1)\big) + \big(\kappa_2 - \kappa_1^2\big)\big(\mathbf{y}(t) - \kappa\mathbf{y}(t-1)\big)\Big], \\
&\quad (1-\kappa^2)\left(1 - \frac{\kappa_1^2 + \kappa_2^2 - 2\kappa_1^2\kappa_2}{1-\kappa_1^2}\right)\mathbf{K}_s\Big).
\end{aligned}$$

The details can be seen in Appendix B.2.

5.3.2 Likelihood

The full likelihood of the hyperparameters $\boldsymbol{\theta} = (\sigma_f, \lambda, \mu, \kappa_1, \kappa_2)$ is then:

$$p(\mathbf{W} \mid \boldsymbol{\theta}) = p\big(\mathbf{W}(1) \mid \boldsymbol{\theta}\big) \prod_{i=2}^{T} p\big(\mathbf{W}(t) \mid \mathbf{W}(t-1), \boldsymbol{\theta}\big). \qquad (5.28)$$

Hence, the full log-likelihood for the sequence is:

$$\log p\big(\mathbf{W} \mid \boldsymbol{\theta}\big) = \log p\big(\mathbf{W}(1) \mid \boldsymbol{\theta}\big) + \sum_{i=2}^{T} \log p\big(\mathbf{W}(t) \mid \mathbf{W}(t-1), \boldsymbol{\theta}\big). \qquad (5.29)$$

The term in the log-likelihood coming from time-point one is exactly (4.48), with $\mathbf{x} = \mathbf{x}(1)$, $\mathbf{y} = \mathbf{y}(1)$ and $\mathbf{z} = \mathbf{z}(1)$. Each of the remaining terms in the log-likelihood can be calculated from:

$$\begin{aligned}
&\log p(\mathbf{W}(t) \mid \mathbf{W}(t-1), \boldsymbol{\theta}) = \log p(\mathbf{x}(t) \mid \mathbf{W}(t-1), \boldsymbol{\theta}) + \\
&\quad \log p(\mathbf{y}(t) \mid \mathbf{x}(t), \mathbf{W}(t-1), \boldsymbol{\theta}) + \log p(\mathbf{z}(t) \mid \mathbf{x}(t), \mathbf{y}(t)\mathbf{W}(t-1), \boldsymbol{\theta}). \qquad (5.30)
\end{aligned}$$

Let:

$$\begin{aligned}
\text{m1} &= \kappa \mathbf{x}(t-1), \\
\text{cov1} &= (1-\kappa^2), \\
\text{m2} &= \kappa \mathbf{y}(t-1) + \kappa_1 \left[\mathbf{x}(t) - \kappa \mathbf{x}(t-1)\right], \\
\text{cov2} &= (1-\kappa^2)(1-\kappa_1^2), \\
\text{m3} &= \kappa \mathbf{z}(t-1) + \frac{1}{1-\kappa_1^2}\left[\kappa_1(1-\kappa_2)(\mathbf{x}(t) - \kappa \mathbf{x}(t-1)) + \right. \\
&\qquad \left. (\kappa_2 - \kappa_1^2)(\mathbf{y}(t) - \kappa \mathbf{y}(t-1))\right], \\
\text{cov3} &= (1-\kappa^2)\left(\frac{\kappa_1^2 + \kappa_2^2 - 2\kappa_1^2\kappa_2}{1-\kappa_1^2}\right).
\end{aligned}$$

Then:

$$\begin{aligned}
\log p(\mathbf{x}(t) \mid \mathbf{W}(t-1), \boldsymbol{\theta}) &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{K}_s| - \frac{n}{2}\log \text{cov1} \\
&\quad -\frac{1}{2\text{cov1}}(\mathbf{x}(t) - \text{m1})^{\mathrm{T}}\mathbf{K}_s^{-1}(\mathbf{x}(t) - \text{m1}), \\
\log p(\mathbf{y}(t) \mid \mathbf{x}(t), \mathbf{W}(t-1), \boldsymbol{\theta}) &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{K}_s| - \frac{n}{2}\log \text{cov2} \\
&\quad -\frac{1}{2\text{cov2}}(\mathbf{y}(t) - \text{m2})^{\mathrm{T}}\mathbf{K}_s^{-1}(\mathbf{y}(t) - \text{m2}), \\
\log p(\mathbf{z}(t) \mid \mathbf{x}(t), \mathbf{y}(t), \mathbf{W}(t-1), \boldsymbol{\theta}) &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{K}_s| - \frac{n}{2}\log \text{cov3} \\
&\quad -\frac{1}{2\text{cov3}}(\mathbf{z}(t) - \text{m3})^{\mathrm{T}}\mathbf{K}_s^{-1}(\mathbf{z}(t) - \text{m3}).
\end{aligned} \tag{5.31}$$

5.3.3 Predictive distributions

Marginal predictions at time $q \in \mathbb{R}$ can be done at a set of test points $\mathbf{s}^* = (s_1^*, \ldots, s_{n^*}^*)$ for each coordinate x, y and z. The joint distribution for a three-dimensional curve at time q, $\mathbf{W}^*(q)$, and the whole sequence, $\mathbf{W}$, can be written as:

$$\begin{bmatrix} \mathbf{W}^*(q) \\ \mathbf{W} \end{bmatrix} \sim N_{3n^*+3nT}\Big(\mathbf{0}, \begin{bmatrix} \mathbf{K}_c \otimes \mathbf{K}_{s^*} & \mathbf{L} \otimes \mathbf{K}_c \otimes \mathbf{K}_{s^*s} \\ (\mathbf{L} \otimes \mathbf{K}_c \otimes \mathbf{K}_{s^*s})^{\mathrm{T}} & \mathbf{K}_t \otimes \mathbf{K}_c \otimes \mathbf{K}_s \end{bmatrix}\Big), \tag{5.32}$$

where

$$\mathbf{L} = \begin{bmatrix} \exp\left(-\frac{|\, q-1\,|}{\mu}\right) & \exp\left(-\frac{|\, q-2\,|}{\mu}\right) & \cdots & \exp\left(-\frac{|\, q-T\,|}{\mu}\right) \end{bmatrix} \tag{5.33}$$

Then

$$\mathbf{W}^*(q) \mid \mathbf{W} \sim N_{3n^*}\left([\mathbf{L}\otimes\mathbf{K}_c\otimes\mathbf{K}_{s^*s}][\mathbf{K}_t\otimes\mathbf{K}_c\otimes\mathbf{K}_s]^{-1}\mathbf{W},\right.$$
$$\left.\mathbf{K}_c\otimes\mathbf{K}_{s^*} - [\mathbf{L}\otimes\mathbf{K}_c\otimes\mathbf{K}_{s^*s}][\mathbf{K}_t\otimes\mathbf{K}_c\otimes\mathbf{K}_s]^{-1}[\mathbf{L}\otimes\mathbf{K}_c\otimes\mathbf{K}_{s^*s}]^{\mathrm{T}}\right).$$

The matrix $\mathbf{L}$ will change depending on the value of q. After some algebra (see Appendix B.3), it emerges that the distributions of $\mathbf{W}^*(q) \mid \mathbf{W}$ for different values of q is a combination of the distribution for the prediction of a three-dimensional curve, scaled by the time correlation parameter as in the predictive distributions for a one-dimensional curve evolving over time. Analogous conditional dependences between coordinates as in (4.47) are:

When $q \in \{1, ..., T\}$,

$$\begin{aligned}
\mathbf{x}^*(q) \mid \mathbf{W} &\sim N_{n^*}\left(\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{x}(q), \mathbf{K}_{s^*} - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}\right),\\
\mathbf{y}^*(q) \mid \mathbf{x}^*(q), \mathbf{W} &\sim N_{n^*}\left(\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{y}(q) + \kappa_1[\mathbf{x}^*(q) - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{x}(q)],\right.\\
&\qquad\left.[1-\kappa_1^2][\mathbf{K}_{s^*} - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}]\right),\\
\mathbf{z}^*(q) \mid \mathbf{x}^*(q), \mathbf{y}^*(q), \mathbf{W} &\sim N_{n^*}\Bigg(\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{z}(q) + \frac{1}{1-\kappa_1^2}\left[\{\kappa_1 - \kappa_1\kappa_2\}\right.\\
&\qquad\left.\left\{\mathbf{x}^*(q) - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{x}(q)\right\} + \left\{\kappa_2 - \kappa_1^2\right\}\left\{\mathbf{y}^* - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{y}\right\}\right],\\
&\qquad\left[1 - \frac{\kappa_1^2 + \kappa_2^2 - 2\kappa_1^2\kappa_2}{1-\kappa_1^2}\right]\left[\mathbf{K}_{s^*} - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}\right]\Bigg).
\end{aligned}$$

When $q > T$,

$$\begin{aligned}
\mathbf{x}^*(q) \mid \mathbf{W} &\sim N_{n^*}\left(\kappa^{q-T}\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{x}(T), \mathbf{K}_{s^*} - (\kappa^{q-T})^2\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}\right),\\
\mathbf{y}^*(q) \mid \mathbf{x}^*(q), \mathbf{W} &\sim N_{n^*}\left(\kappa^{q-T}\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{y}(T) + \kappa_1[\mathbf{x}^*(q) -\right.\\
&\qquad\left.\kappa^{q-T}\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{x}(T)],\, [1-\kappa_1^2][\mathbf{K}_{s^*} - (\kappa^{q-T})^2\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}]\right),\\
\mathbf{z}^*(q) \mid \mathbf{x}^*(q), \mathbf{y}^*(q), \mathbf{W} &\sim N_{n^*}\Bigg(\kappa^{q-T}\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{z}(T) + \frac{1}{1-\kappa_1^2}\\
&\qquad\left[\{\kappa_1 - \kappa_1\kappa_2\}\left\{\mathbf{x}^*(q) - \kappa^{q-T}\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{x}(T)\right\} +\right.\\
&\qquad\left.\left\{\kappa_2 - \kappa_1^2\right\}\left\{\mathbf{y}^*(q) - \kappa^{q-T}\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{y}(T)\right\}\right],\\
&\qquad\left[1 - \frac{\kappa_1^2 + \kappa_2^2 - 2\kappa_1^2\kappa_2}{1-\kappa_1^2}\right]\left[\mathbf{K}_{s^*} - (\kappa^{q-T})^2\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}\right]\Bigg).
\end{aligned}$$

When $q < 1$,

$$\begin{aligned}
\mathbf{x}^*(q) \mid \mathbf{W} &\sim N_{n^*}\left(\kappa^{1-q}\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{x}(1), \mathbf{K}_{s^*} - (\kappa^{1-q})^2\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}\right), \\
\mathbf{y}^*(q) \mid \mathbf{x}^*(q), \mathbf{W} &\sim N_{n^*}\left(\kappa^{1-q}\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{y}(1) + \kappa_1[\mathbf{x}^*(q) - \kappa^{1-q}\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{x}(1)],\right. \\
&\qquad \left.[1-\kappa_1^2][\mathbf{K}_{s^*} - (\kappa^{1-q})^2\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}]\right), \\
\mathbf{z}^*(q) \mid \mathbf{x}^*(q), \mathbf{y}^*(q), \mathbf{W} &\sim N_{n^*}\Bigg(\kappa^{1-q}\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{z}(1) + \frac{1}{1-\kappa_1^2} \\
&\qquad \left[\{\kappa_1 - \kappa_1\kappa_2\}\left\{\mathbf{x}^*(q) - \kappa^{1-q}\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{x}(1)\right\} + \right. \\
&\qquad \left.\left\{\kappa_2 - \kappa_1^2\right\}\left\{\mathbf{y}^*(q) - \kappa^{1-q}\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{y}(1)\right\}\right], \\
&\qquad \left[1 - \frac{\kappa_1^2 + \kappa_2^2 - 2\kappa_1^2\kappa_2}{1-\kappa_1^2}\right]\left[\mathbf{K}_{s^*} - (\kappa^{1-q})^2\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}\right]\Bigg).
\end{aligned}$$

When $t < q < t+1$, if:

$$\begin{aligned}
\text{mx} &= \frac{1}{1-\kappa^2}\Big(\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\left[\left(\kappa^{q-t} - \kappa^{t-q+2}\right)\mathbf{x}(t) + \left(\kappa(\kappa^{t-q} - \kappa^{q-t})\right)\mathbf{x}(t+1)\right]\Big), \\
\text{my} &= \frac{1}{1-\kappa^2}\Big(\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\left[\left(\kappa^{q-t} - \kappa^{t-q+2}\right)\mathbf{y}(t) + \left(\kappa(\kappa^{t-q} - \kappa^{q-t})\right)\mathbf{y}(t+1)\right]\Big), \\
\text{mz} &= \frac{1}{1-\kappa^2}\Big(\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\left[\left(\kappa^{q-t} - \kappa^{t-q+2}\right)\mathbf{z}(t) + \left(\kappa(\kappa^{t-q} - \kappa^{q-t})\right)\mathbf{z}(t+1)\right]\Big), \\
\text{cov} &= \frac{\kappa^{2(q-t)} + \kappa^{2(t-q)+2} - 2\kappa^2}{1-\kappa^2},
\end{aligned}$$

then:

$$\begin{aligned}
\mathbf{x}^*(q) \mid \mathbf{W} &\sim N_{n^*}\big(\text{mx}, \mathbf{K}_c \otimes (\mathbf{K}_{s^*} - \text{cov}\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*})\big), \\
\mathbf{y}^*(q) \mid \mathbf{x}^*(q), \mathbf{W} &\sim N_{n^*}\big(\text{my} + \kappa_1[\mathbf{x}(q) - \text{mx}], \\
&\qquad [1-\kappa_1^2][\mathbf{K}_c \otimes (\mathbf{K}_{s^*} - \text{cov}\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*})]\big), \\
\mathbf{z}^*(q) \mid \mathbf{x}^*(q), \mathbf{y}^*(q), \mathbf{W} &\sim N_{n^*}\Big(\text{mz} + \frac{1}{1-\kappa_1^2}\left[\{\kappa_1 - \kappa_1\kappa_2\}\{\mathbf{x}^*(q) - \text{mx}\} + \right. \\
&\qquad \left.\left\{\kappa_2 - \kappa_1^2\right\}\{\mathbf{y}^*(q) - \text{my}\}\right], \\
&\qquad \left[1 - \frac{\kappa_1^2 + \kappa_2^2 - 2\kappa_1^2\kappa_2}{1-\kappa_1^2}\right][\mathbf{K}_c \otimes (\mathbf{K}_{s^*} - \text{cov}\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*})]\Big).
\end{aligned}$$

5.3.4 Optimization of hyperparameters

For the model of three coordinates evolving over time, $\boldsymbol{\theta} = (\sigma_f, \lambda, \mu, \kappa_1, \kappa_2)$, i.e., there are 5 hyperparameters to be optimised. This makes the optimisation more

challenging and increases the computational time for the grid search. To reduce the number of hyperparameters, it was decided to exploit the fact that the optimal value of the signal variance, σ_f^2, that maximises the log-likelihood function, as a function of the other hyperparameters, can be determined exactly.

5.3.4.1 Estimation of the signal variance, σ_f^2

If

$$\mathrm{L}(\boldsymbol{\theta}) \propto \frac{1}{\sigma_f^m} \exp\left(-\frac{1}{2\sigma_f^2}\mathrm{Z}\right), \tag{5.34}$$

so that

$$\log \mathrm{L}(\boldsymbol{\theta}) \propto m \log \sigma_f - \frac{1}{2\sigma_f^2}\mathrm{Z}, \tag{5.35}$$

the log-likelihood can be optimised w.r.t. σ_f. Differentiating and setting to zero, the maximum can be found:

$$\left.\frac{\partial \log \mathrm{L}}{\partial \sigma_f}\right|_{\hat{\sigma}_f} = \frac{-m}{\hat{\sigma}_f} + \frac{1}{\hat{\sigma}_f^3}\mathrm{Z} = 0. \tag{5.36}$$

Hence:

$$\hat{\sigma}_f^2 = \frac{\mathrm{Z}}{m}. \tag{5.37}$$

This approach was carried out first for the simplest model: the GP of a single 1D curves (Section 4.2). In this scenario: $\mathbf{x} \sim N_n(\mathbf{0}, \mathbf{K}_s)$, where the $(i,j)^{th}$ element of $\mathbf{K}_s$ is equal to $\sigma_f^2 \exp\left(-\frac{1}{2\lambda^2}(s_i - s_j)^2\right)$. The signal variance can be taken out of the definition of $\mathbf{K}_s$ as a common factor so that the covariance matrix can be written as $\sigma_f^2\mathbf{K}_s$, with $(i,j)^{th}$ element of $\mathbf{K}_s$ equal $\exp\left(-\frac{1}{2\lambda^2}(s_i - s_j)^2\right)$. With $\boldsymbol{\theta} = (\sigma_f, \lambda)$, then:

$$\begin{aligned}
\log p(\mathbf{x} \mid \boldsymbol{\theta}) &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log \mid \sigma_f^2\mathbf{K}_s \mid -\frac{1}{2}\mathbf{x}^{\mathrm{T}}\left[\sigma_f^2\mathbf{K}_s\right]^{-1}\mathbf{x} \\
&= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma_f^2) - \frac{1}{2}\log \mid \mathbf{K}_s \mid -\frac{1}{2\sigma_f^2}\mathbf{x}^{\mathrm{T}}\mathbf{K}_s^{-1}\mathbf{x} \\
&= \text{const} - n\log\sigma_f - \frac{1}{2\sigma_f^2}\mathbf{x}^{\mathrm{T}}\mathbf{K}_s^{-1}\mathbf{x}.
\end{aligned} \tag{5.38}$$

Using (5.37),

$$\hat{\sigma}_f^2 = \frac{\mathbf{x}^{\mathrm{T}}\mathbf{K}_s^{-1}\mathbf{x}}{n}, \tag{5.39}$$

which, via $\mathbf{K}_s$, is a function of λ.

Substituting (5.39) for σ_f^2 in (5.38) gives the profile log-likelihood of λ:

$$\log p(\mathbf{x} \mid \lambda) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\left(\frac{\mathbf{x}^{\mathrm{T}}\mathbf{K}_s^{-1}\mathbf{x}}{n}\right) - \frac{1}{2}\log \mid \mathbf{K}_s \mid -\frac{n}{2} \tag{5.40}$$

Optimization is then performed over the length-scale, λ, giving the same results (not shown) as in Section 4.2, thus validating the approach.

In the model for the evolution of three-dimensional curves, calculations are not as straightforward (see Appendix B.4). The signal variance is estimated as:

$$\hat{\sigma}_f^2 = \frac{\mathrm{Z1} + \sum_{t=2}^{T} \mathrm{Z2}_t}{3n(T+1)}, \tag{5.41}$$

where:

$$\begin{aligned}\mathrm{Z1} = &\big(\mathbf{x}(1)^{\mathrm{T}}\mathbf{K}_s^{-1}\mathbf{x}(1)\big) + \\ &\frac{1}{a}\big(\mathbf{y}(1) - \kappa_1\mathbf{x}(1)\big)^{\mathrm{T}}\mathbf{K}_s^{-1}\big(\mathbf{y}(1) - \kappa_1\mathbf{x}(1)\big) + \frac{1}{c}\big(\mathbf{z}(1) - b\big)^{\mathrm{T}}\mathbf{K}_s^{-1}\big(\mathbf{z}(1) - b\big),\end{aligned}$$

$$\begin{aligned}\mathrm{Z2}_t = &\frac{\big(\mathbf{x}(t) - \mathrm{m1}\big)^{\mathrm{T}}\mathbf{K}_s^{-1}\big(\mathbf{x}(t) - \mathrm{m1}\big)}{\mathrm{cov1}} + \\ &\frac{\big(\mathbf{y}(t) - \mathrm{m2}\big)^{\mathrm{T}}\mathbf{K}_s^{-1}\big(\mathbf{y}(t) - \mathrm{m2}\big)}{\mathrm{cov2}} + \frac{\big(\mathbf{z}(t) - \mathrm{m3}\big)^{\mathrm{T}}\mathbf{K}_s^{-1}\big(\mathbf{z}(t) - \mathrm{m3}\big)}{\mathrm{cov3}},\end{aligned}$$

and

$$\begin{aligned}
a &= (1 - \kappa_1^2), \\
b &= \frac{(\kappa_1 - \kappa_1\kappa_2)\mathbf{x}(1) + (\kappa_2 - \kappa_1^2)\mathbf{y}(1)}{a}, \\
c &= 1 - \frac{\kappa_1^2 + \kappa_2^2 - 2\kappa_1^2\kappa_2}{a}, \\
\mathrm{m1} &= \kappa\mathbf{x}(t-1), \\
\mathrm{cov1} &= 1 - \kappa^2, \\
\mathrm{m2} &= \kappa\mathbf{y}(t-1) + \kappa_2\left[\mathbf{x}(t) - \kappa\mathbf{x}(t-1)\right], \\
\mathrm{cov2} &= (1 - \kappa_1^2)(1 - \kappa^2), \\
\mathrm{m3} &= \kappa\mathbf{z}(t-1) + \frac{1}{1-\kappa_1^2}\big[\kappa_1(1-\kappa_2)\big(\mathbf{x}(t) - \kappa\mathbf{x}(t-1)\big) + \\
&\quad (\kappa_2 - \kappa_1^2)\big(\mathbf{y}(t) - \kappa\mathbf{y}(t-1)\big)\big], \\
\mathrm{cov3} &= (1 - \kappa^2)\left(\frac{\kappa_1^2 + \kappa_2^2 - 2\kappa_1^2\kappa_2}{1 - \kappa_1^2}\right).
\end{aligned}$$

Then the full (profile) log-likelihood for the model of three-dimensional curves evolving over time is:

$$\begin{aligned}\log p(\mathbf{W} \mid \boldsymbol{\theta}) = &-\frac{3}{2}\left[n\log(2\pi) + n\log\left(\frac{\mathrm{Z1} + \sum_{t=2}^{T}\mathrm{Z2}_t}{3n(T+1)}\right) + \log \mid \mathbf{K}_s \mid\right] \\ &- \frac{n}{2}\left[\log(a) + \log(c)\right] - \frac{\mathrm{Z13}n(T+1)}{2(\mathrm{Z1} + \sum_{t=2}^{T}\mathrm{Z2}_t)} \\ &+ \sum_{t=2}^{T}\left(-\frac{3}{2}\left[n\log(2\pi) + n\log\left(\frac{\mathrm{Z1} + \sum_{t=2}^{T}\mathrm{Z2}_t}{3n(T+1)}\right) + \log \mid \mathbf{K}_s \mid\right]\right. \\ &\left. - \frac{n}{2}\left[\log(\mathrm{cov1}) + \log(\mathrm{cov2}) + \log(\mathrm{cov3})\right] - \frac{\mathrm{Z2}_t 3n(T+1)}{2(\mathrm{Z1} + \sum_{t=2}^{T}\mathrm{Z2}_t)}\right). \quad (5.42)\end{aligned}$$

Since the signal variance is taken out of the SE covariance function, where measurement noise was previously added as an additional variance σ_n^2 on the diagonal, it is now accounted for by specifying a noise-to-signal ratio σ_n^2/σ_f^2 as an additive term on the diagonal of the new $\mathbf{K}_s$.

The conditional SE of $\hat{\sigma}_f$ can be estimated taking the square root of the inverse of the negative second derivative of the full log-likelihood function, at $\hat{\boldsymbol{\theta}}$.

5.3.5 Simulated data

To test the implementation of the model, points on a set of three-dimensional curves were simulated with similar structures to the lip curves. Starting from an actual 3D lip curve and using hyperparameters $\boldsymbol{\theta} = (\sigma_f, \lambda, \mu, \kappa_1, \kappa_2) = (11.8985, 0.33, 2000, 0.0218, 0.7)$, based on estimates from the starting lip, the points shown in Figure 5.10 were simulated.

Starting with a grid search and then using the conjugate gradients method, with an additive normal error of standard deviation 0.5 mm added to the model, optimal hyperparameters were found by maximizing the log-likelihood function (5.42). The optimal values found are $\hat{\boldsymbol{\theta}} = (\hat{\sigma}_f, \hat{\lambda}, \hat{\mu}, \hat{\kappa}_1, \hat{\kappa}_2) = (5.56, 27.23 \times 10^{-2}, 2.6 \times 10^3, -0.07, 0.67)$, with respective SE $(0.06, 0.28 \times 10^{-2}, 0.45 \times 10^3, 0.03, 0.02)$. Note the large SE for the hyperparameter μ. This is consistent with the findings in Section 5.2.4. The optimal value found by maximum likelihood is larger than the true hyperparameter. However, the true hyperparameter is contained within the

errors, except for the signal variance, which is underestimated. A range of further simulations were performed (not shown), which generated no cause for concern.

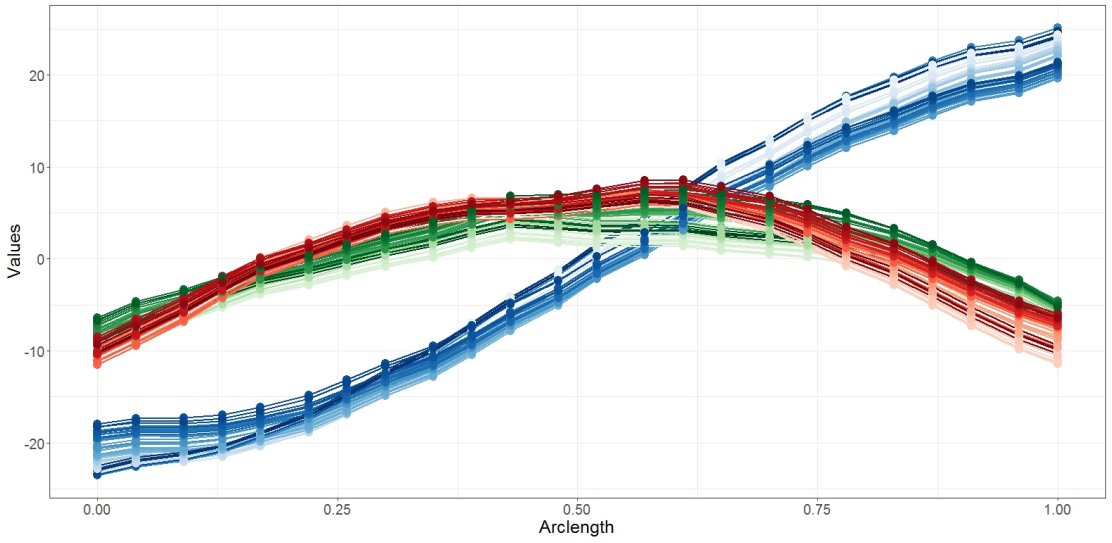


FIGURE 5.10: Points on simulated three-dimensional evolving curves. The x coordinate points are shown in blue, those from the y coordinate in green and, finally, the z coordinate in red. The first curves simulated are shown in lighter colours and the last simulations in darker shades.

5.3.6 Evolution of three-dimensional lip curves

Starting with the sequence of upper lips for the emotion *disgust* (Figure 5.9), of 61 pictures $\mathbf{t} = (1 \cdots 61)^T$, as in Section 5.2.5, the main problem faced was the very high correlation between points at adjacent time points, since it makes the hyperparameter μ very large. It was observed that the larger the value of μ, the smaller its effect on the likelihood function. Moreover, it was found, from the Hessian matrix at the maximum likelihood estimates, that μ and κ_2, the correlation between coordinates y and z, were negatively correlated. Two different approaches were investigated:

1. Assume there is no relationship between coordinates, i.e., fix $\kappa_1 = \kappa_2 = 0$.

2. Optimise the values for hyperparameters κ_1 and κ_2 for a series of individual time points across the sequence, using the model for three-dimensional curves (Section 4.3), and then fix them to the mean of these values.

For the second approach, it was considered to carry out a free maximum likelihood optimisation of all the hyperparameter, using these points estimates as starting points in the optimization, however, the strong negative correlation between μ and κ_2 still made optimisation of both simultaneously unviable. Therefore, pragmatically, κ_1 and κ_2 had to be fixed to carry out optimisation over the rest of hyperparameters. Recall that, to ease the optimisation process, the number of hyperparameters is reduced by finding the signal variance, σ_f, that maximises the log-likelihood function analytically. Optimization is then carried out, by maximum likelihood, for the remaining hyperparameters. Optimal hyperparameters found when setting κ_1 and κ_2 both to zero are: $\hat{\boldsymbol{\theta}}_1 = (\hat{\sigma}_f, \hat{\lambda}, \hat{\mu}) = (4.93, 17.7 \times 10^{-2}, 50.8)$, with respective SE: $(0.05, 0.33 \times 10^{-2}, 5.2)$. For the second approach, κ_1 and κ_2 were optimised every 5 time-points. The mean values of these estimates were fixed, and the resulting optimal values are: $\hat{\boldsymbol{\theta}}_2 = (\hat{\sigma}_f, \hat{\lambda}, \hat{\mu}, \hat{\kappa}_1, \hat{\kappa}_2) = (5.34, 15.5 \times 10^{-2}, 25.5, -1.64 \times 10^{-}2, 78.5 \times 10^{-2})$, with respective SE: $(0.06, 0.27 \times 10^{-2}, 2.6, 0.59 \times 10^{-}2, 2.86 \times 10^{-2})$. The SE for κ_1 and κ_2 can be approximated from the series of estimates, dividing their standard deviation by the square root of the number of them. While the two different approaches give similar values for the signal variance and the spatial length-scale, the value of μ is halved by the second approach. This illustrates the inverse relationship between μ and κ_2: when the latter increases from 0 to 0.7850, μ is reduced from 50.79 to 25.50. For both approaches, an error ratio $\eta = \sigma_n^2/\sigma_f^2$ was added to the covariance matrix to accommodate errors in the observed values. For optimisation, η was fixed at 0.01. Much smaller values failed to remove the problems with the covariance matrix being ill-conditioned. Given the resulting optimal values for $\hat{\sigma}_f$, the final additive normal error has standard deviation (σ_n) of 0.5 mm approximately (0.4911 and 0.5321 for each approach, respectively).

The same time-points ($\mathbf{t} = (1, \ldots, 61)^{\mathrm{T}}$) were considered to make predictions at 25 spatial-points, conditioning on both sets of optimal hyperparameters, $\hat{\boldsymbol{\theta}}_1$ (Figure 5.11) and $\hat{\boldsymbol{\theta}}_2$ (Figure 5.12). Retrodiction was done at time $q = -1$, using the data at time point 1. Prediction at time $q = 63$, conditioned on the data from the last curve available, i.e. at time 61, was also made. Marginal prediction was done at time $q = 30$, using the observed values at that time, and interpolation at $q = 40.75$, using the observed values at times 40 and 41. Both approaches produce plausible predictions. Again, note how the error bands expand as predictions are made further away from the observations, both in time and space. Both figures show the observed values are shown in filed symbols, while the predicted values

are shown in unfilled symbols. The lines represent the posterior means, displayed with 2 standard deviations bands (shown dotted).

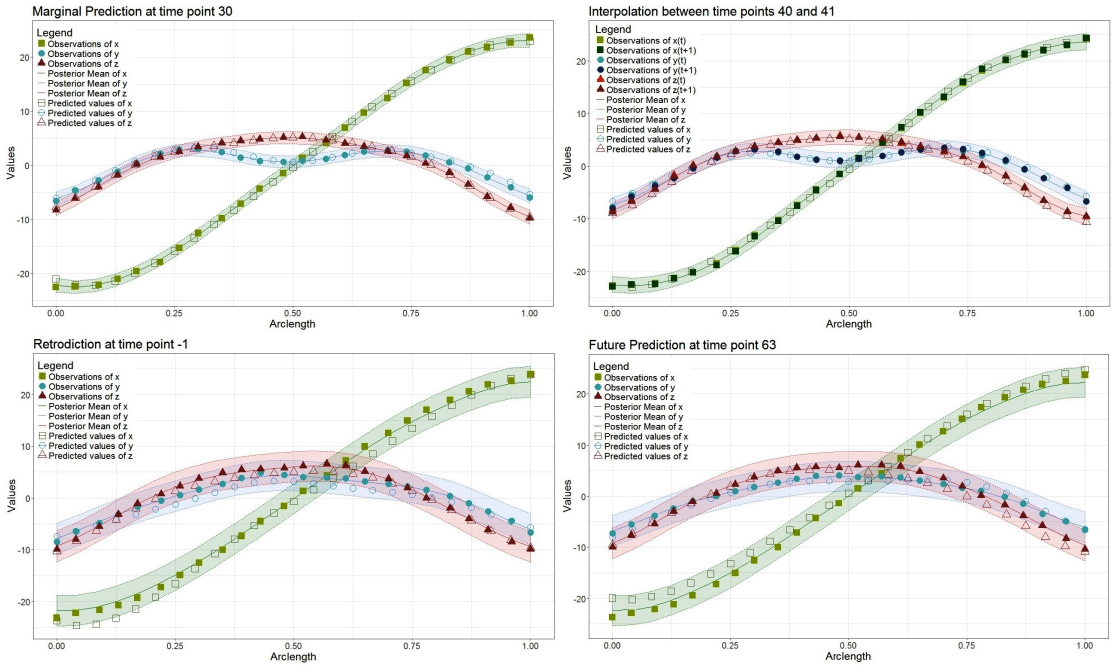


FIGURE 5.11: Observations, posterior means and one draw from the predictive distributions, for 3-dimensional lip curves of the emotion *Disgust*, using $\hat{\boldsymbol{\theta}}_1$.

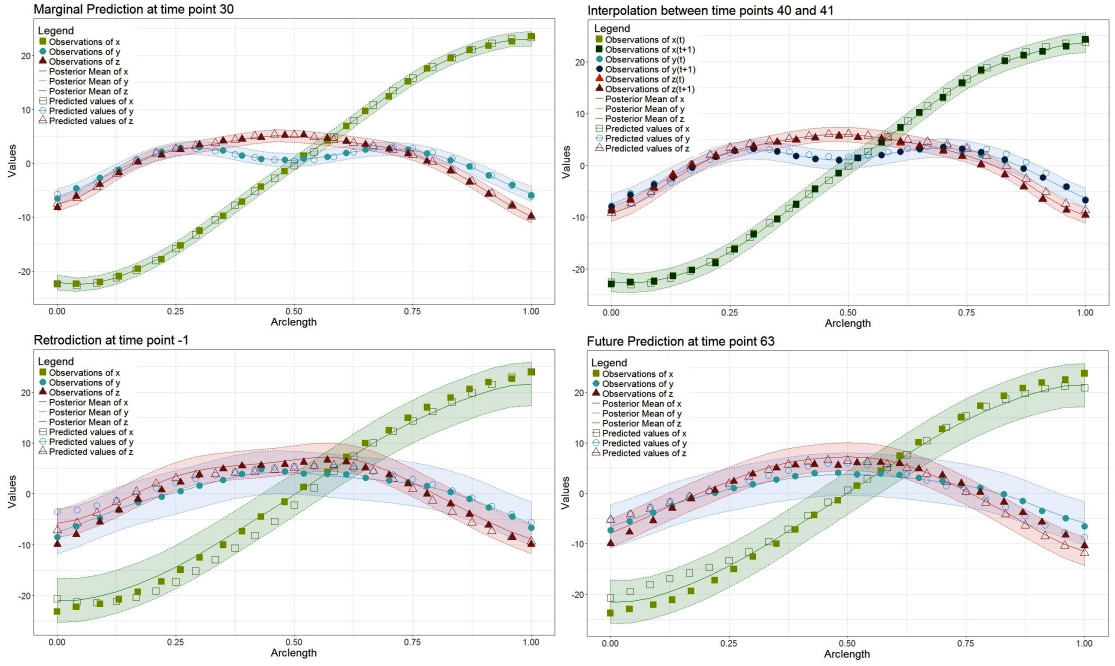


FIGURE 5.12: Observations, posterior means and one draw from the predictive distributions, for 3-dimensional lip curves of the emotion *Disgust*, using $\hat{\boldsymbol{\theta}}_2$.

5.4 Grouping of emotions

The way that the emotions of people who have undergone maxillofacial surgery are perceived from their facial expression relates directly to the success of the surgical operation, in respect of the ability of the patient to produce 'typical' facial expressions. The complexity of the lip shape, as an example, evolving through the expression of the emotion can be simplified by studying the differences between the emotions in terms of the covariance parameters.

Optimisation by maximum likelihood was performed for all the replicates in each emotion. Data were available for six different emotions: *anger, disgust, fear, happiness, sadness* and *surprise*. These consist of sequences 60 images long, in the case of the expression of *disgust*, to sequences of about 180 images, in the case of *happiness*.

For the second approach, where the hyperparameters for the correlation between coordinates, κ_1 and κ_2, are estimated at a series of time-points in the sequence, Figure 5.13 shows the different estimates. It can be seen how the correlation between x and y or z is always close to 0, while the estimates of κ_2 (correlation between y and z coordinates), is always above 0.7. The estimates are shown with confidence intervals of ± 2 SE.

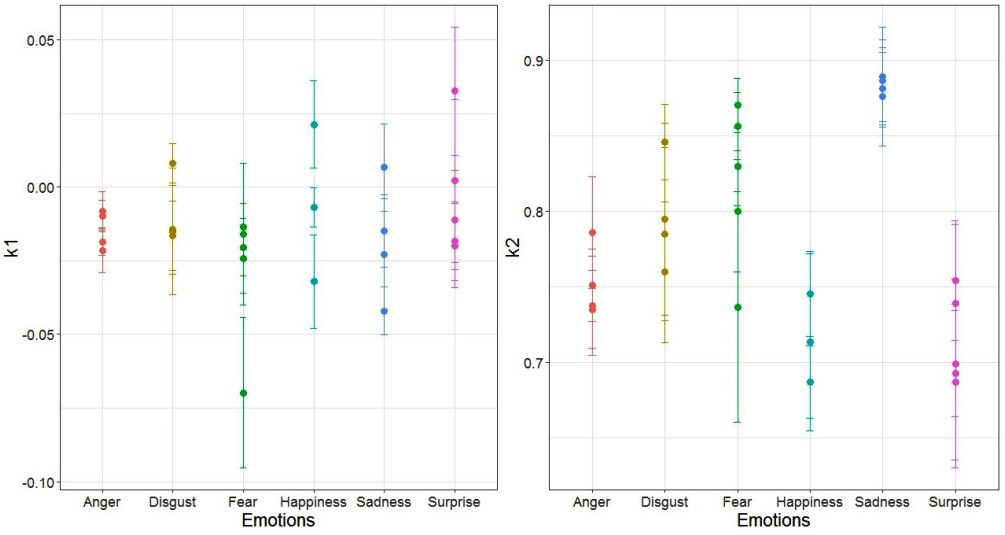


FIGURE 5.13: Summaries of $\hat{\kappa}_1$ (left) and $\hat{\kappa}_2$ (right) with error bands, across emotions.

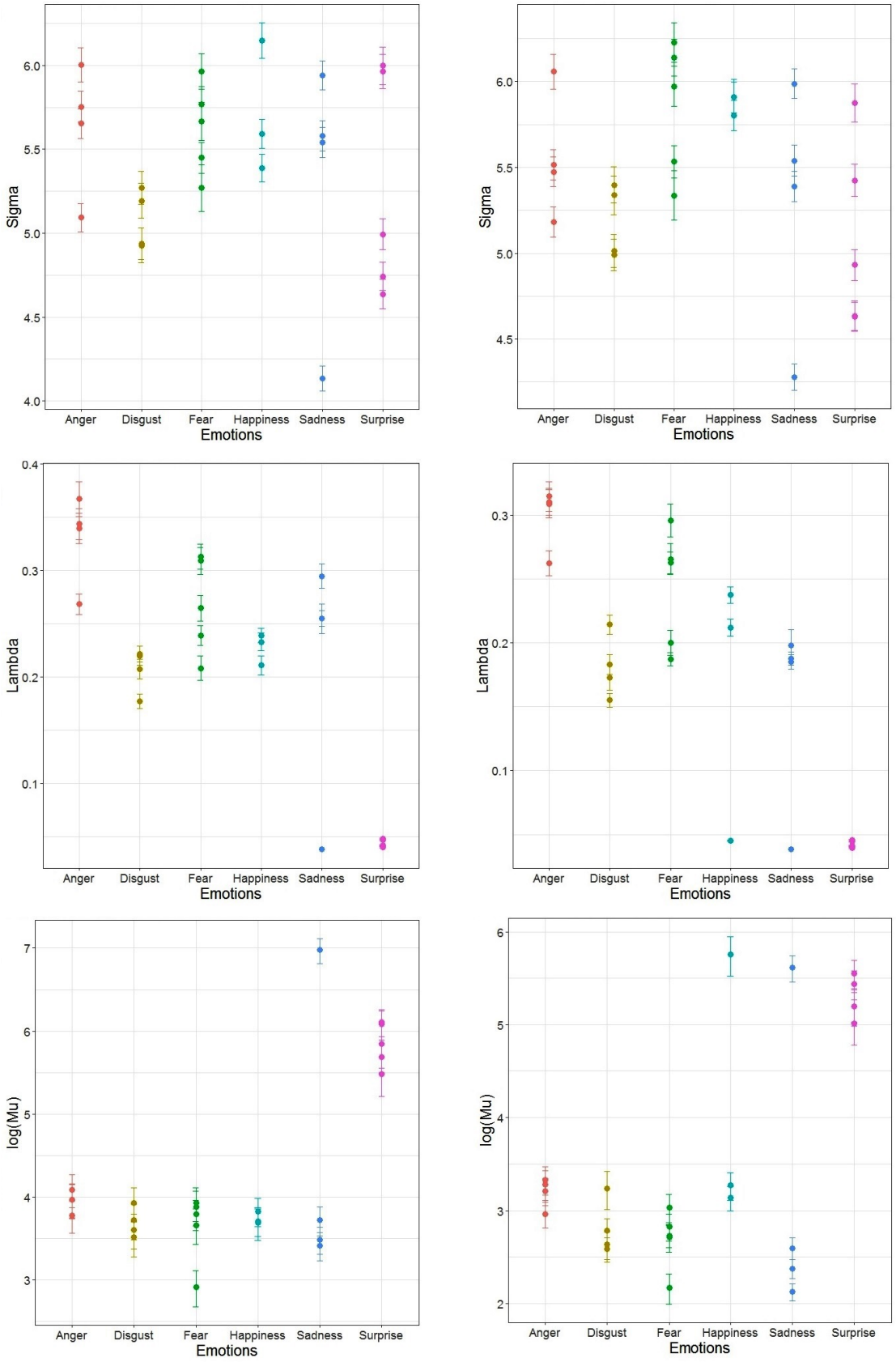


FIGURE 5.14: Summaries of $\hat{\sigma}_f$ (top), $\hat{\lambda}$ (middle) and $\log(\hat{\mu})$ (bottom), using approach 1 (left) and approach 2 (right), across emotions.

Figure 5.14 shows how the remaining hyperparameters change within the different replicates of an emotion and across the different emotions. The logarithm of $\hat{\mu}$ is taken for a better appreciation of the differences. The values of the optimal hyperparameters for all the replicates of the six emotions can be found in Appendix B.5.

To study how the emotions could be grouped in terms of their covariance parameters (the hyperparameters), principal component analysis (PCA) was performed. For Approach 1, PCA was applied over hyperparameters σ_f, λ and μ, although other possibilities might include $\log \sigma_f$, $\log \lambda$ and $\log \mu$. The first principal component (PC) explains 74.8% of the variation in the data, and together with the second component, they explain 91.3% of the variability, which makes Figure 5.15 a good representation of the data. Most of the emotions lie together with no clear distinction between them, except for *Surprise*, which is further apart from the rest. The emotion *Disgust* seems to be the most tightly clustered.

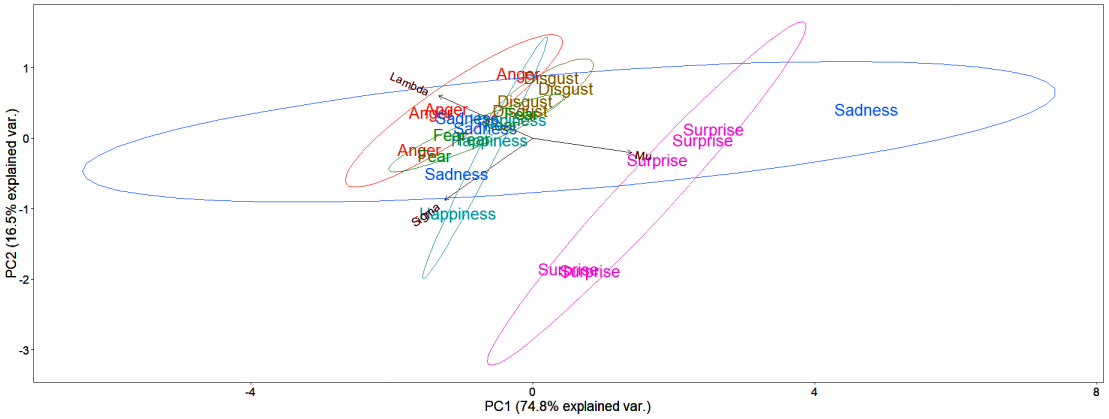


FIGURE 5.15: Biplot of the (scaled) first two principal components for Approach 1, with 95% normal probability ellipsoids.

From the rotated component matrix, shown in Table 5.1, it can be seen that the highest component loading for the first PC is the hyperparameters μ, i.e. the temporal correlation, but fairly close to the other two parameters. For the second PC, the signal variance, σ_f, is the highest component loading.

	PC1	PC2	PC3
λ	-0.5799	0.5549	-0.5963
μ	0.6069	-0.1938	-0.7707
σ_f	-0.5433	-0.8089	-0.2244

TABLE 5.1: Rotated Components Matrix for Approach 1.

Figure 5.16 shows the representation of the emotions based on the first two PCs resulting from applying PCA to the results of Approach 2. Given there is now a higher number of hyperparameters, there is also more variability in the data, and therefore, the first two PC are able to capture only 72.4% of the total variability, which is less than the variability explained by just the first PC with Approach 1.

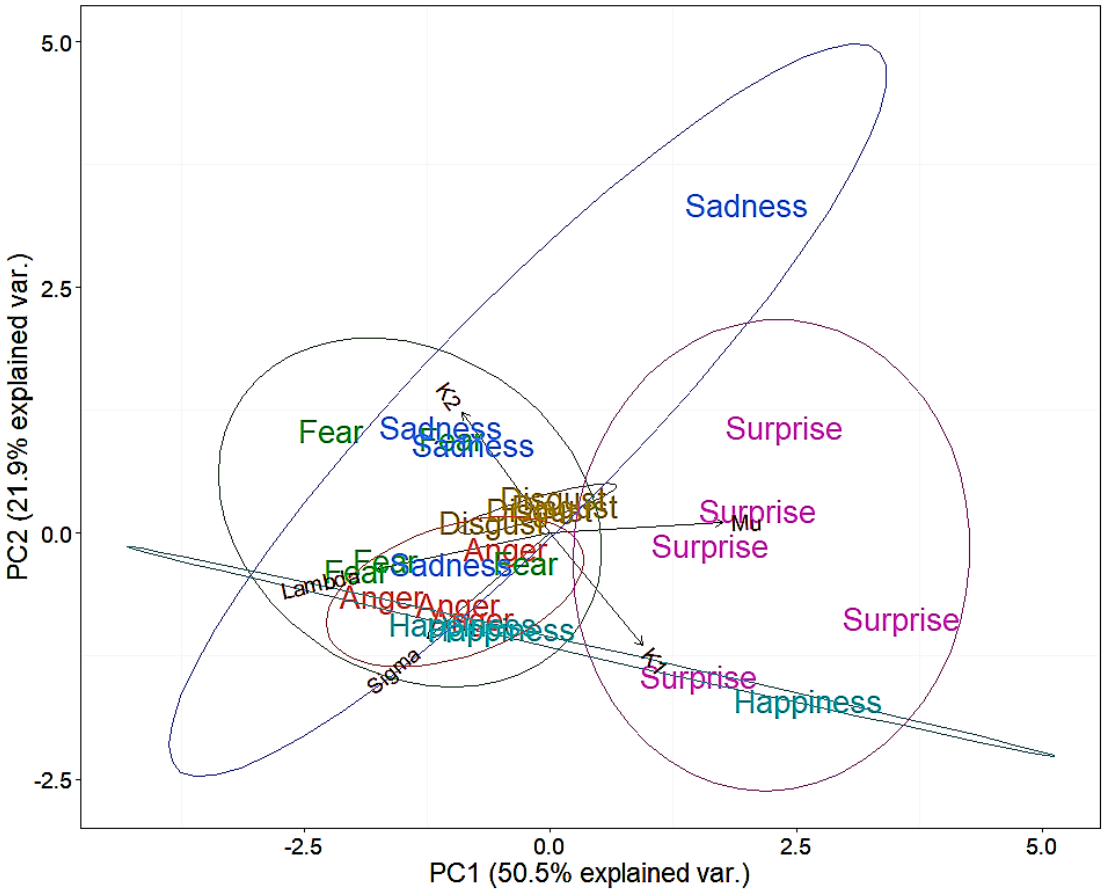


FIGURE 5.16: Biplot of the (scaled) first two principal components for Approach 2, with 95% normal probability ellipsoids.

It is clear from the biplot that *Surprise* is the most different from the rest of the emotions, and *Disgust* the one with less variability between replicates, which can be interpreted as having a series of strong unique characteristics that do not allow much change.

The highest component loading for the first PC is now λ. However, μ has a value almost as large. The σ_f component loading is now lower than before. For the second PC, in this case, is the correlation between the y and z coordinates. Given the shape of the lips, these two coordinates are strongly correlated, but for certain emotions where there is a lot of variation in the coordinate y, this correlation will decrease, and hence, it is expected to assume this hyperparameter can help differentiate the emotions. All the loadings are included in Table 5.2.

	PC1	PC2	PC3	PC4	PC5
κ_1	0.3059	-0.5637	-0.7212	0.2600	0.0299
κ_2	-0.2918	0.6075	-0.6874	-0.2559	0.0882
λ	-0.5726	-0.1784	0.0502	0.3432	0.7210
μ	0.5710	0.0559	0.0631	-0.4537	0.6789
σ_f	-0.4088	-0.5274	-0.0312	-0.7370	-0.1023

TABLE 5.2: Rotated Components Matrix for Approach 2.

5.4.1 Discussion

The model for three-dimensional curves evolving over time has raised a number of very interesting issues from a methodological perspective. The number of hyperparameters and the problems found when the change is subtle have resulted in the proposal of different approaches to cope with it. The lip curve data represent a peculiar scenario due to their high smoothness both spatially and temporally, and therefore special measures had to be taken. It has been seen however that these are not necessarily needed in other sequences of less correlated three-dimensional curves, using simulation. Both approaches proposed for the three-dimensional lip curves appear to lead to the same conclusions. Given the nature of the lips, however, it is clear that the y and z coordinates are strongly correlated. Setting the correlation parameters κ_1 and κ_2 to zero does not take full advantage of all the information contained in the data. Moreover, it can be seen from Table 5.2 that κ_2 is the highest component loading for the second PC, which explains 21.9% of the variance in the data, making it an important hyperparameters to help cluster the different emotions.

For a better understanding of the differences between the emotions, in terms of their correlation parameters, it would be necessary to increase the number of replicates, as well as adding more subjects to the study, to account for the variability across people.

5.5 Gaussian Process model for the evolution of 2D curves

The main goal was a model for three-dimensional curves evolving over time, however, between models for one- and three-dimensional curves, it is natural to consider the corresponding model for two-dimensional curves. This model can be regarded as a particular case of the three-dimensional model. Appendix B.6 presents a summary of the key elements of a Gaussian Process model for two-dimensional curves evolving over time.

Chapter 6

Phylogenetic Gaussian Process models for k-dimensional curves

6.1 Introduction and background

Whilst the concept of evolution has been regarded in previous chapters as a gradual process of change and development, the notion of shape evolving in time is extended in this chapter to the phylogenetic setting. The term evolution is referred to now as a process of gradual change that takes place over many generations, where branching points can occur: ancestors diverging into daughter species. A framework for modelling many-to-one functions that evolve and branch (''functional phylogenetics'') has recently been developed, permitting inference, e.g., of the unknown branching pattern of unobserved ancestral data, as well as characteristics of the evolutionary process itself. Gaussian processes (GPs) can be used to model continuous trait evolution in statistical phylogenetics. Under such processes, observations at the tips of a phylogenetic tree have a multivariate Gaussian distribution.

6.1.1 Phylogenetic covariance function

Jones and Moriarty [2013] proposed a Gaussian process model for the evolution of a function-valued trait along a phylogenetic tree $\mathbf{T}$. $\mathbf{T}$ specifies both the time of a point in the tree and also which branch it sits on, generalises the linear time

variable t used in Chapter 5. The term 'function-valued' is meant in the sense that the data object is a continuous function $f(x)$ indexed by a variable x.

Let $\mathbf{p}$ be a set of observations such that each corresponds to a point $(s, \mathbf{t})$, where $s \in S$ is the continuous-spatial index, the arc-length of the previous chapters, and $\mathbf{t} \in \mathbf{T}$ is the point in the tree (at a specific time on a specific branch) under consideration. To construct a covariance function for this type of data, two assumptions natural to the context of evolution are made:

- **Assumption 1**: Conditional on their common ancestors in the phylogenetic tree $\mathbf{T}$, any two function-valued traits are statistically independent.
- **Assumption 2**: The statistical relationship between a function-valued trait and any of its descendants in $\mathbf{T}$ is independent of the topology of the tree.

In many cases, the covariance function may be assumed to be separable in both space and time, so that we need to specify just a space-only covariance function and a time-only covariance function. The latter was called the *phylogenetic covariance function* by Jones and Moriarty [2013]. They also assumed that, conditional on any given trait value, its ancestor and progenitor trait values are statistically independent. This corresponds to choosing a temporal component which is Markovian. As seen in previous chapters, the most common Markovian GPs are the class of Ornstein-Uhlenbeck (OU) processes, which have covariance function: $k_t(t, t') = \exp(-|t - t'|/\mu)$, with hyperparameter μ, the time-scale. The phylogenetic covariance function can be defined as:

$$k_{\mathbf{T}}(\mathbf{t}_i, \mathbf{t}_j) = \exp\left(\frac{-d_{\mathbf{T}}(\mathbf{t}_i, \mathbf{t}_j)}{\mu_{\mathbf{T}}}\right), \tag{6.1}$$

where $d_{\mathbf{T}}(\mathbf{t}_i, \mathbf{t}_j)$ denotes the 'patristic' distance between $\mathbf{t}_i$ and $\mathbf{t}_j$, i.e., the distance along the tree branches from node $\mathbf{t}_i$ to node $\mathbf{t}_j$. For two terminal nodes, the patristic distance is equal to twice the time of the most recent common ancestor. The hyperparameter $\mu_{\mathbf{T}}$ specifies the characteristic time scale for the evolutionary dynamics. As previously, the overall process variance is included in the spatial covariance function.

Consider the tree in Figure 6.1, and let t_0 be the time at present, i.e. $t_0 = 0$ (this will follow throughout the Chapter). The patristic distance between (terminal) nodes B and C is twice the age of their most common recent ancestor, i.e., the

time back to node D: $d_{\mathbf{T}}(B, C) = 2t_1$. The time between A and D is the sum of the times from each of them to their most common recent ancestor, E: $d_{\mathbf{T}}(A, D) = t_2 + (t_2 - t_1) = 2t_2 - t_1$.

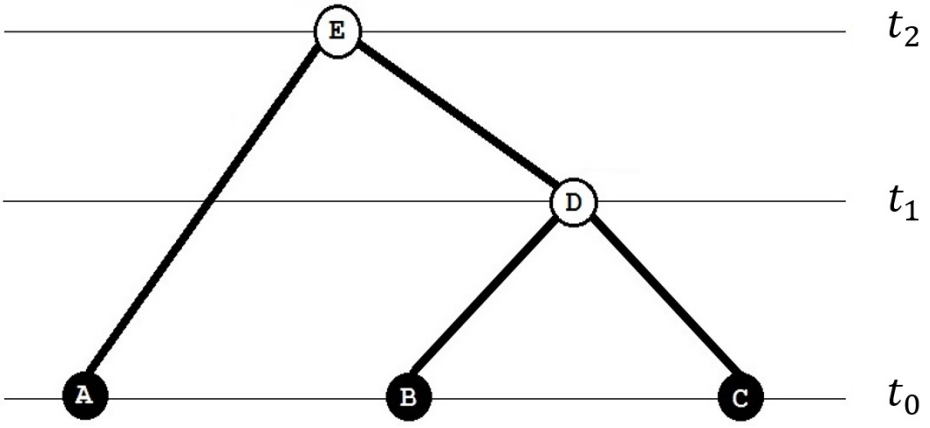


FIGURE 6.1: A simple tree with associated times.

The matrix of patristic distances for the tree is:

$$\begin{array}{c} \\ A \\ B \\ C \\ D \\ E \end{array} \begin{array}{c} \begin{array}{ccccc} A & B & C & D & E \end{array} \\ \begin{bmatrix} 0 & 2t_2 & 2t_2 & 2t_2 - t_1 & t_2 \\ 2t_2 & 0 & 2t_1 & t_1 & t_2 \\ 2t_2 & 2t_1 & 0 & t_1 & t_2 \\ 2t_2 - t_1 & t_1 & t_1 & 0 & t_2 - t_1 \\ t_2 & t_2 & t_2 & t_2 - t_1 & 0 \end{bmatrix} \end{array}. \tag{6.2}$$

The phylogenetic covariance function offers a straightforward approach to the prediction of unobserved function-valued traits using GPs. The associated likelihood can be used for inference over the phylogenetic tree or the evolutionary process. The approach is suitable for dense observations of function-valued traits and for sparsely and even irregularly sampled traits, with missing observations.

6.2 Phylogenetic Gaussian Process model for 3D curves

It has been illustrated in previous chapters how the covariance structure of a three-dimensional curve can be expressed in terms of the arc-length and the relationship

between coordinates, and also how a series of evolving three-dimensional curves requires additional specification of the temporal covariance structure. This model is now extended to the phylogenetic setting, where the linear time index is replaced by a tree.

Imagine a tree, $\mathbf{T}$, where the data object at each location in the tree is a three-dimensional curve. Each of this curves can be represented as a mixed GP using the space component, indexed by the arc-length $s \in [0, 1]$, the discrete label $c = \{x, y, z\}$ and a phylogenetic component associated with the position in the tree where the curve occurs, $\mathbf{t} \in \mathbf{T}$. The GP can then be defined as:

$$p(\mathbf{t}, c, s) \sim GP\big(m(\mathbf{t}, c, s), k(\mathbf{t}, \mathbf{t}', c, c', s, s')\big), \tag{6.3}$$

For a series of spatial points $\mathbf{s} = (s_1 \cdots s_n)^{\mathrm{T}}$ and using the notation from Section 5.3, points on a three-dimensional curve at node $\mathbf{t}$ can be notated as:

$$\mathbf{W}(\mathbf{t}) = \begin{bmatrix} \mathbf{x}(\mathbf{t}) \\ \mathbf{y}(\mathbf{t}) \\ \mathbf{z}(\mathbf{t}) \end{bmatrix}, \tag{6.4}$$

where $\mathbf{x}(\mathbf{t}) = \big(x(s_1, \mathbf{t}) \cdots x(s_n, \mathbf{t})\big)^{\mathrm{T}}$ represents the values of the x coordinate, and similarly for $\mathbf{y}(\mathbf{t})$ and $\mathbf{z}(\mathbf{t})$. $\mathbf{W}(\mathbf{t})$ will be observed at a collection of points in the tree $\{\mathbf{t}_i\}$. These are collected together in the vector $\mathbf{W_T}$.

Let l represent the number of terminal nodes (leaves) in the tree, then the total number of nodes, m, is $m = 2l - 1$ (in a fully resolved rooted bifurcating tree). If data are available at all nodes, the distribution for the set of points of all the three-dimensional curves in the tree can be written as:

$$\mathbf{W_T} = \begin{bmatrix} \mathbf{W}(\mathbf{t}_1) \\ \vdots \\ \mathbf{W}(\mathbf{t}_m) \end{bmatrix} \sim N_{3mn}\left(\mathbf{m}, \mathbf{K}\right), \tag{6.5}$$

where $\mathbf{m}$ represents the mean, which is assumed to be zero (as before), and $\mathbf{K}$ is the covariance matrix. Once again, separability is assumed, such that:

$$k(\mathbf{t}, \mathbf{t}', c, c', s, s') = k_{\mathbf{T}}(\mathbf{t}, \mathbf{t}')k_c(c, c')k_s(s, s'). \tag{6.6}$$

Therefore, $\mathbf{K} = \mathbf{K_T} \otimes \mathbf{K}_c \otimes \mathbf{K}_s$, such that:

- $\mathbf{K_T}$ represents the covariance of points on curves at different nodes. The phylogenetic covariance function is used, i.e., $k_{\mathbf{T}}(\mathbf{t}, \mathbf{t}') = \exp\left(-d_{\mathbf{T}}(\mathbf{t}, \mathbf{t}')/\mu_{\mathbf{T}}\right)$, and hence, the matrix $\mathbf{K_T}$ has $(i,j)^{th}$ element equal to $k_{\mathbf{T}}(\mathbf{t}_i, \mathbf{t}_j)$, where $\mathbf{t}_i$ is the position of the i^{th} node in the tree.

- For the 3×3 matrix $\mathbf{K}_c$, (up to) three hyperparameters are specified: κ_1, the correlation between x and y, κ_2, between y and z and κ_3, between x and z:

$$\mathbf{K}_c = \begin{pmatrix} 1 & \kappa_1 & \kappa_3 \\ \kappa_1 & 1 & \kappa_2 \\ \kappa_3 & \kappa_2 & 1 \end{pmatrix}. \tag{6.7}$$

- The space-covariance function used is the Squared-Exponential (SE), i.e., $k_s(s, s') = \sigma_f^2 \exp\left(-\frac{1}{2}(s - s')^2/\lambda^2\right)$, with hyperparameters: σ_f^2, the signal variance and λ, the length-scale. Therefore, $\mathbf{K}_s$ represents the covariance matrix for the n arc-length inputs, with $(i,j)^{th}$ element equal to $k_s(s_i, s_j)$.

If only data at the l terminal nodes are available, the distribution has its dimensions reduced, such that the set of curves at the leaves $\mathbf{W_L} \sim N_{3ln}(\mathbf{m}, \mathbf{K})$. The mean is also assumed to be zero, and the covariance function separable. The covariance matrix $\mathbf{K}$ will have the same structure. It is important to note that the phylogenetic covariance matrix, although calculated only for the leaves now, reflects the relationship among all nodes, since it takes into account internal nodes to calculate the patristic distances between the leaves.

6.2.1 Likelihood

The total log-likelihood of the tree can be calculated, given the hyperparameters $\boldsymbol{\theta} = (\sigma_f, \lambda, \mu_{\mathbf{T}}, \kappa_1, \kappa_2, \kappa_3)$:

$$\log\left(L(\hat{\boldsymbol{\theta}})\right) = \log p(\mathbf{W_T} \mid \boldsymbol{\theta}) = \frac{-3mn}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{K}| - \frac{1}{2}\mathbf{W}_{\mathbf{T}}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{W_T}. \tag{6.8}$$

As introduced in Section 5.3.4.1, to ease the optimisation process, the number of hyperparameters can be reduced by finding the signal variance, σ_f, that maximises the log-likelihood function analytically:

$$\hat{\sigma}_f^2 = \frac{\mathbf{W}_{\mathbf{T}}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{W_T}}{3mn}. \tag{6.9}$$

Let $\tilde{\boldsymbol{\theta}} = (\lambda, \mu_{\mathbf{T}}, \kappa_1, \kappa_2, \kappa_3)$, the profile log-likelihood can then be rewritten as:

$$\log p(\mathbf{W_T} \mid \tilde{\boldsymbol{\theta}}) = -\frac{3mn}{2}\log(2\pi) - \frac{3mn}{2}\log\left(\frac{\mathbf{W_T^T K^{-1} W_T}}{3mn}\right) - \frac{1}{2}\log|\mathbf{K}| - \frac{3mn}{2} \tag{6.10}$$

The (conditional) standard error of the estimated variance signal can be calculated from the square root of the negative of the inverse of the second derivative:

$$\text{SE}(\hat{\sigma_f}) = \sqrt{-\left[\frac{3mn}{\sigma_f^2} - \frac{3(\mathbf{W_T^T K^{-1} W_T})}{\sigma_f^4}\right]^{-1}}. \tag{6.11}$$

6.2.2 Identifiability

A basic question concerning any statistical model is whether it is identifiable, that is, given a distribution of observations which the model predicts, is it theoretically possible to recover the parameters of the model? In other words, it is possible to learn the true values of the model's parameters after obtaining an infinite number of observations (for consistency as well as for identifiability). Mathematically, this is equivalent to saying that different values of the parameters must generate different probability distributions of the observable variables. Understanding what parameters are identifiable for a model is crucial in order to know what can be inferred from the data. Identifiability of the tree topology and its branch lengths is essential for any model that is to be used to infer evolutionary information from data. If a tree is not uniquely determined by an expected joint distribution, then one has no hope of using the model to infer trees well from data. Indeed, proofs of the statistical consistency of an inference method such as maximum likelihood begin by establishing identifiability of parameters [Allman and Rhodes, 2006].

The model proposed above assumes the times of the nodes are known, and the evolution rate $\mu_{\mathbf{T}}$ is to be estimated. In practice, both the times and $\mu_{\mathbf{T}}$ are unknown. There is then a parameter identification problem, since the times and $\mu_{\mathbf{T}}$ cannot be optimised at the same time. Ho and An [2013] proved that if the tree is ultra-metric (a tree where all the path-lengths from the root to the leaves are equal) and the height of the tree is unbounded, $\mu_{\mathbf{T}}$ cannot be estimated consistently as the tree grows indefinitely, regardless of the estimation method. This is implicit from the definition of the phylogenetic covariance matrix, where the ratio between the patristic distance and $\mu_{\mathbf{T}}$ is needed. If the times are unknown, we are not only

trying to optimise the hyperparameter $\mu_{\mathbf{T}}$ but also the patristic distances. In this scenario, larger distances with larger values for $\mu_{\mathbf{T}}$ result in the same likelihood value as shorter distances with smaller $\mu_{\mathbf{T}}$, i.e., two or more parametrizations are observationally equivalent. A model like this, that fails to be identifiable, is said to be non-identifiable or unidentifiable. In this case, the asymptotic properties of maximum likelihood estimation cannot be used. Such scenarios are manifest in the likelihood surface containing ridges of equal (maximal) likelihood.

Therefore, there is a need to choose whether to optimise the change rate $\mu_{\mathbf{T}}$ or the times of the nodes. If $\mu_{\mathbf{T}}$ is fixed to a constant, then the optimised times are effectively being measured in some arbitrary units (not years or generations, as would be ideal). It is conventional to choose these units so that the average rate of change is 1 per time unit [Yang, 2006]. In this case the 'times' become a measure of the proportion of time that has passed from divergence in the data. They can be used to study for how long a trait has evolved before changing, but only in comparison to the time spent in other branch. The longer the branch, the larger the amount of change between the nodes it links. One needs to also pay attention to the fact that assuming that $\mu_{\mathbf{T}}$ is constant implies it does not vary among groups. This is analogous to the 'molecular clock hypothesis' of molecular evolution [Thorpe, 1982]. In this thesis $\mu_{\mathbf{T}}$ is fixed to the value of 1 for simplicity, and the branch lengths (times) optimised.

6.2.3 Predictive distributions

In practice, the most common scenario is to have data available only at the leaves of the tree. In which case, it is natural to try to predict the data at internal nodes (ancestors). Predictions can be done both spatially and temporally. If the set of points on curves at the l leaves $\mathbf{W_L} \sim N_{3ln}(\mathbf{m}, \mathbf{K})$, with $\mathbf{K} = \mathbf{K_T} \otimes \mathbf{K}_c \otimes \mathbf{K}_s$, marginal predictions at node $\mathbf{q} \in \mathbf{T}$ can be done at a set of test points $\mathbf{s}^* = (s_1^*, \ldots, s_{n^*}^*)$ using:

$$\begin{pmatrix} \mathbf{W}^*(\mathbf{q}) \\ \mathbf{W_L} \end{pmatrix} \sim N_{3n^*+3ln} \left(\mathbf{0}, \begin{bmatrix} \mathbf{K}_c \otimes \mathbf{K}_{s^*} & \mathbf{L} \otimes \mathbf{K}_c \otimes \mathbf{K}_{s^*s} \\ \mathbf{L} \otimes \mathbf{K}_c \otimes \mathbf{K}_{ss^*} & \mathbf{K_T} \otimes \mathbf{K}_c \otimes \mathbf{K}_s \end{bmatrix} \right), \tag{6.12}$$

where $\mathbf{L}$ is the covariance matrix between the terminal nodes and node q:

$$\mathbf{L} = \left[\exp\left(-\frac{d_{\mathbf{T}}(\mathbf{t}_q, \mathbf{t}_1)}{\mu_{\mathbf{T}}} \right) \quad \exp\left(-\frac{d_{\mathbf{T}}(\mathbf{t}_q, \mathbf{t}_2)}{\mu_{\mathbf{T}}} \right) \quad \cdots \quad \exp\left(-\frac{d_{\mathbf{T}}(\mathbf{t}_q, \mathbf{t}_l)}{\mu_{\mathbf{T}}} \right) \right]. \tag{6.13}$$

Then the posterior predictive distribution of $\mathbf{W}^*(q) \mid \mathbf{W_L}$ (conditioned on the maximum likelihood vaues of the hyperparameters) is:

$$\mathbf{W}^*(q) \mid \mathbf{W_L} \sim N_{3n^*}\Big(\left(\left[\mathbf{L}\mathbf{K}_{\mathbf{T}}^{-1}\right] \otimes \mathbf{I}_3 \otimes \left[\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\right]\right) \mathbf{W_L}, \\ \mathbf{K}_c \otimes \mathbf{K}_{s^*} - \left(\left[\mathbf{L}\mathbf{K}_{\mathbf{T}}^{-1}\mathbf{L}\right] \otimes \left[\mathbf{I}_3\mathbf{K}_c\right] \otimes \left[\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}\right] \right)\Big). \quad (6.14)$$

This distribution depends on the number of leaves in the tree, as well as the topology and the position of the internal node for which prediction is made. $\mathbf{I}_3$ denotes a 3×3 identity matrix. As in previous chapters, as there are n training points and n^* test points, $\mathbf{K}_{ss^*}$) denotes the $n \times n^*$ matrix of spatial covariances evaluated at all pairs of training and test points, with $(i, j)^{th}$ element equal to $k_s(s_i, s_j^*)$. $\mathbf{K}_s, \mathbf{K}_{s^*}$ and $\mathbf{K}_{s^*s}$ are as before.

6.2.4 An application to simulated data

To study the model, a number of simulations were done. A simple example is presented here, with $l = 3$ terminal nodes and total number of nodes $m = 2l - 1 = 5$. A set of points of three-dimensional curves was simulated using hyperparameters $\boldsymbol{\theta} = (\sigma_f, \lambda, \mu_{\mathbf{T}}, \kappa_1, \kappa_2, \kappa_3) = (1, 0.3, 1, -0.5, 0.17, 0.3)$, with a set of 15 training points along the arc-length, s, equally spaced from 0 to 1. The curves and tree structure are shown in Figure 6.2, where each coordinate is plotted as a function of the arc-length. The nodes are labelled from 1 to 5, and therefore, the tree can be written as $((1, 2), 3)$ in the Newick format [Huson et al., 2010], or as a vector of ancestors, the i^{th} element of which is the ancestor node of node i $(i = 1, \ldots, m)$: $(4, 4, 5, 5)$. The curves at the leaves are considered to be at time 0, the curve at node 4, the most recent common ancestor of 1 and 2, was simulated at time $t_4 = 0.1$ and the root at time $t_5 = 0.3$. This can be written as a m-vector of times $(0, 0, 0, 0.1, 0.3)$, the i^{th} element of which is the time of node i.

The hyperparameters are estimated by maximum likelihood. Recall the hyperparameter $\mu_{\mathbf{T}}$ was chosen to be one, to make the model identifiable. The model can be re-parametrised so that the differences in node times are optimised. Then two time hyperparameters are defined:

- t_1, representing the time between node 4 and the leaves, and
- t_2, representing the time from the root to node 4.

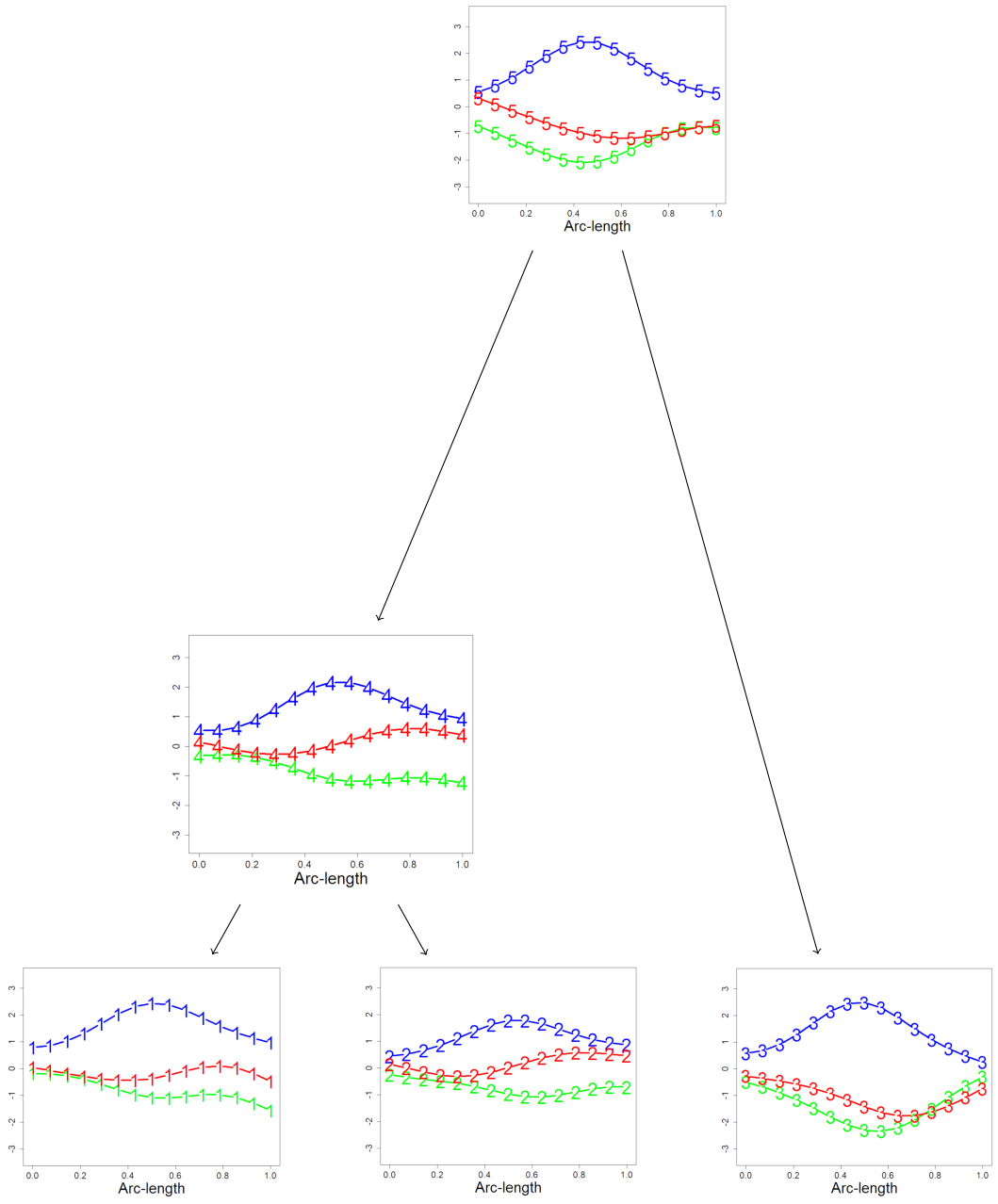


FIGURE 6.2: Simulated 3D curves: the x coordinate is indicated in blue, y in green and z in red.

Then, the set of hyperparameters to be estimated is $\boldsymbol{\theta} = (\sigma_f, \lambda, \kappa_1, \kappa_2, \kappa_3, t_1, t_2)$, with true values $(1, 0.3, -0.5, 0.17, 0.3, 0.1, 0.2)$. Different optimisations were done to check the implementation of the model. First, optimal hyperparameters for the tree are found, using all data available, at the leaves, internal node and root ($\mathbf{W_T}$) and the true topology from which the curves were simulated, represented by the ancestors vector $(4, 4, 5, 5)$. If optimisation is carried out using a different topology, for example $(5, 4, 4, 5)$, the maximal log-likelihood value is expected to be smaller.

When the topology is unknown, the key is to find the maximum likelihood for each possible topology, and then estimate the tree as the one that produces the largest value. This approach is only practical in small trees, where the number of topologies is small. This number rises very rapidly with the number of leaves l, since there are $(2l-3)!! = (2l-3)!/\left[2^{l-2}(l-2)!\right]$ rooted bifurcating trees [Felsenstein, 2004]. Optimal hyperparameters for the three possible topologies for this tree are summarised in Table 6.1.

All data available							
$\hat{\boldsymbol{\theta}}$	$\hat{\sigma}_f$	$\hat{\lambda}$	$\hat{\kappa}_1$	$\hat{\kappa}_2$	$\hat{\kappa}_3$	$\hat{t}_1$	$\hat{t}_2$
Topology	(4, 4, 5, 5)						(Right)
Estimates	0.4662	0.2498	−0.6392	0.4303	−0.0920	0.1043	0.2863
SE	0.0219	0.0085	0.0643	0.0933	0.1252	0.0241	0.0754
$\log\left(L(\hat{\boldsymbol{\theta}})\right)$	322.4511						
Topology	(5, 4, 4, 5)						(Wrong)
Estimates	0.4984	0.2643	−0.6884	0.3495	0.0320	0.3227	0.2726
SE	0.0235	0.0091	0.0543	0.1017	0.1246	0.0744	0.0756
$\log\left((L(\hat{\boldsymbol{\theta}})\right)$	273.8693						
Topology	(4, 5, 4, 5)						(Wrong)
Estimates	0.4963	0.2630	−0.7268	0.3595	−0.0054	0.3609	0.2920
SE	0.0234	0.0091	0.0494	0.1019	0.1203	0.0805	0.0851
$\log\left((L(\hat{\boldsymbol{\theta}})\right)$	272.8213						

Table 6.1: Maximum likelihood estimates for data at all nodes under all three possible topologies.

As expected, the correct tree does have the maximal log-likelihood. Most of the hyperparameter estimates remain similar across the optimisations with the three different topologies, except the first time difference t_1, which increases for the wrong topologies. This is expected since these topologies have nodes 1-3 and 2-3 evolving from the same common ancestor, 4, but they are in fact more different to 4 than 1 and 2, and therefore a longer branch (and hence more variation) is inferred by the model.

The model for which data are available only at the leaves ($\mathbf{W_L}$) was also studied, by optimising the hyperparameters using only the points on the curves at nodes 1, 2 and 3. In this case, the true topology is also expected to result in a higher value of the log-likelihood. Table 6.2 shows these results. Again, the maximal log-likelihood corresponds with the true topology.

For the right topology, note that the estimates of the hyperparameters are almost the same as when optimising for $\mathbf{W_T}$, and therefore, also close to the true values. In the case of the wrong topologies, t_1 is again over-estimated, resulting in longer branch lengths. Moreover, the second time difference, t_2, is now pushed towards the lower boundary, where it is almost equal to zero. This means that the model inferred for these two wrong topologies involves all three leaves descending from just one common ancestor, in a trifurcation.

Data at leaves available							
$\hat{\boldsymbol{\theta}}$	$\hat{\sigma}_f$	$\hat{\lambda}$	$\hat{\kappa}_1$	$\hat{\kappa}_2$	$\hat{\kappa}_3$	$\hat{t}_1$	$\hat{t}_2$
Topology	(4, 4, 5, 5)						(Right)
Estimates	0.4781	0.2504	−0.6059	0.6712	−0.1693	0.1613	0.2584
se	0.0291	0.0108	0.0863	0.0760	0.1442	0.0496	0.1307
$\log\left(L(\hat{\boldsymbol{\theta}})\right)$	157.6314						
Topology	(5, 4, 4, 5)						(Wrong)
Estimates	0.4836	0.2557	−0.6232	0.6196	−0.1016	0.3083	0.0001
se	0.0294	0.0111	0.0883	0.0815	0.1428	0.0957	0.1126
$\log\left(L(\hat{\boldsymbol{\theta}})\right)$	153.6477						
Topology	(4, 5, 4, 5)						(Wrong)
Estimates	0.4836	0.2557	−0.6232	0.6196	−0.1016	0.3084	0.0001
se	0.0294	0.0111	0.0858	0.0824	0.1424	0.0810	0.1482
$\log\left(L(\hat{\boldsymbol{\theta}})\right)$	153.6467						

Table 6.2: Maximum likelihood estimates for data just at the leaves under all three possible topologies.

For both models, a noise ratio $\eta = \sigma_n^2/\sigma_f^2$ was added to the diagonal of the covariance matrix, as a means of dealing with ill-conditioning. It was fixed at $\eta = 0.085$. Given the resulting optimal values for $\hat{\sigma}_f$, the final additive normal noise has standard deviation (σ_n): for $\mathbf{W_T}$, $\hat{\sigma}_n = 0.1361$ and for $\mathbf{W_L}$, $\hat{\sigma}_n = 0.1398$. Note the data is different from that in Chapter 5 and therefore will have a different σ_f, hence the different η.

6.3 Phylogenetic Gaussian Process model for 2D curves

In some particular scenarios, it could be that one of the three coordinates does not provide any meaningful information about the shape, or maybe the data available is directly in the form of two-dimensional curves. A simpler model for just two

coordinates can be fitted to the data, which is just a particular case of the model presented in Section 6.2. A GP can be specified as:

$$p(\mathbf{t}, c, s) \sim GP\big(m(\mathbf{t}, c, s), k(\mathbf{t}, \mathbf{t}', c, c', s, s')\big), \tag{6.15}$$

a mixed GP for the spatial component (indexed by the arc-length) $s \in [0, 1]$, the discrete label c, which consists now of two coordinates, i.e., $c = \{x, y\}$ and the phylogenetic component $\mathbf{t} \in \mathbf{T}$ specifying the location within the phylogeny. Points on a two-dimensional curve at node $\mathbf{t}$ can be written as:

$$\mathbf{W}(\mathbf{t}) = \begin{bmatrix} \mathbf{x}(\mathbf{t}) \\ \mathbf{y}(\mathbf{t}) \end{bmatrix}. \tag{6.16}$$

Given there are only two coordinates now, only one correlation hyperparameter, κ_1, say, is needed. Then the distribution of the set of two-dimensional curves in a tree is:

$$\mathbf{W_T} \sim N_{2mn}\left(\mathbf{m}, \mathbf{K}\right), \tag{6.17}$$

where $\mathbf{m}$ is the mean, assumed to be zero, and $\mathbf{K}$ is the covariance matrix, $\mathbf{K} = \mathbf{K_T} \otimes \mathbf{K}_c \otimes \mathbf{K}_s$, where both $\mathbf{K_T}$ and $\mathbf{K}_s$ remain the same as in Section 6.2 and $\mathbf{K}_c$ is now defined as:

$$\mathbf{K}_c = \begin{pmatrix} 1 & \kappa_1 \\ \kappa_1 & 1 \end{pmatrix}. \tag{6.18}$$

The same theory applies to the data only at the leaves, $\mathbf{W_L} \sim N_{3ln}\left(\mathbf{m}, \mathbf{K}\right)$, with $\mathbf{K_T}$ having entries corresponding to the correlation between terminal nodes.

6.3.1 Likelihood

The profile likelihood can be calculated as before, using the estimation of σ_f that maximises the full likelihood analytically, which is now:

$$\hat{\sigma}_f^2 = \frac{\mathbf{W_T^T K^{-1} W_T}}{2mn}. \tag{6.19}$$

Let $\tilde{\boldsymbol{\theta}} = (\lambda, \mu_{\mathbf{T}}, \kappa_1, \kappa_2, \kappa_3)$, the profile log-likelihood can then be rewritten as:

$$\log p(\mathbf{W_T} \mid \tilde{\boldsymbol{\theta}}) = -mn\log(2\pi) - mn\log\left(\frac{\mathbf{W_T^T K^{-1} W_T}}{2mn}\right) - \frac{1}{2}\log|\mathbf{K}| - mn. \tag{6.20}$$

6.3.2 An example based on simulation

A set of two-dimensional curves was simulated using hyperparameters $\boldsymbol{\theta} = (\sigma_f, \lambda, \mu_{\mathbf{T}}, \kappa_1) = (0.23, 0.7, 1, -0.5)$, times $(0, 0, 0, 0.2, 0.5)$ and ancestors vector $(5, 4, 4, 5)$. Points on the simulated curves are plotted as a function of the arc-length in Figure 6.3. The training points chosen this time for the arc-length, s, were 21 equally spaced points between 0 and 1. Using the same parametrisation as previously, the internal node time differences are $t_1 = 0.2$ and $t_2 = 0.3$, and therefore the set of hyperparameters to be optimised $\boldsymbol{\theta} = (\sigma_f, \lambda, \kappa_1, t_1, t_2) = (0.23, 0.7, -0.5, 0.2, 0.3)$.

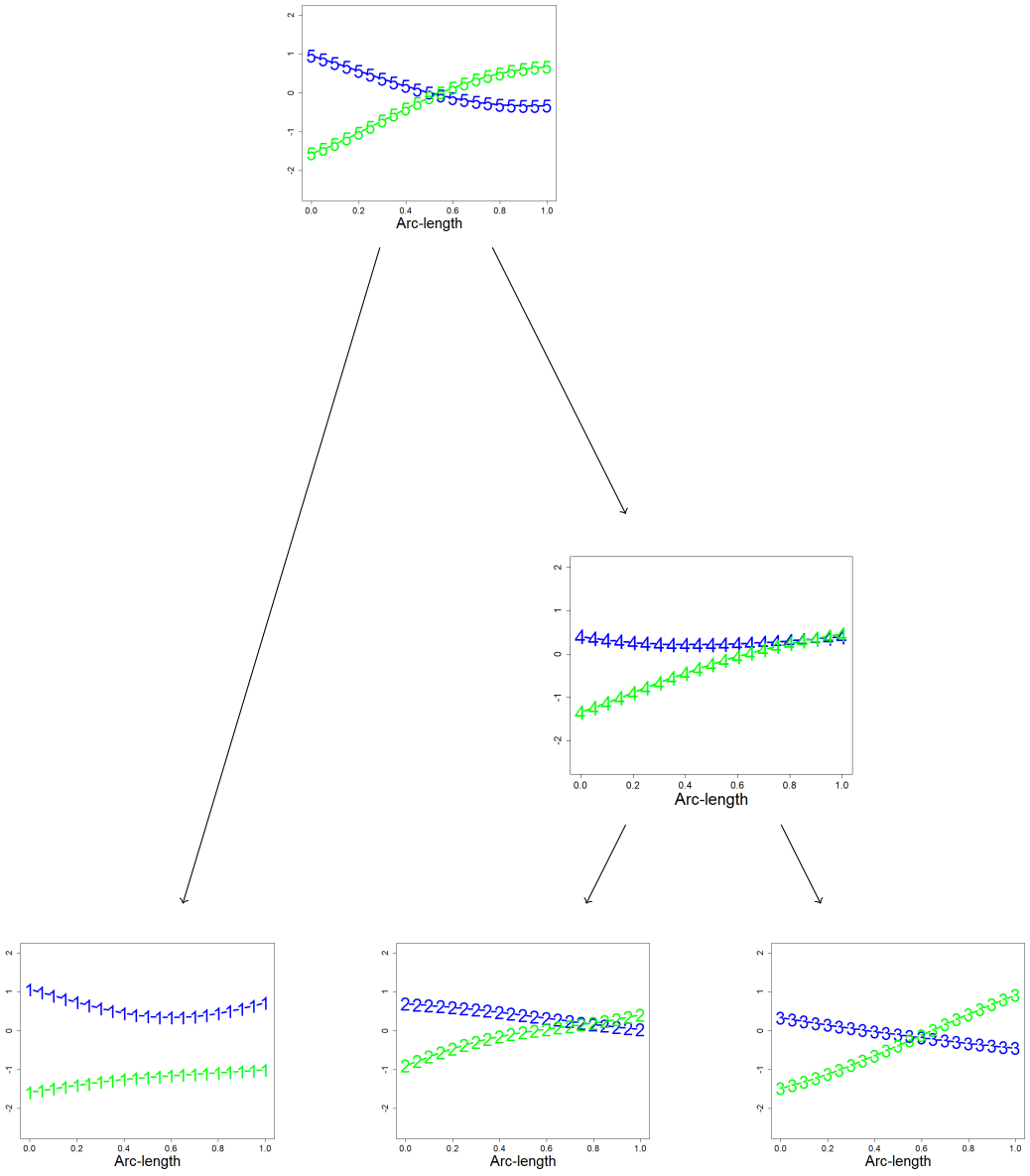


FIGURE 6.3: Simulated curves: the x coordinate is indicated in blue and y in green.

As before, different assumptions were tried, using each of the three possible topologies, and either $\mathbf{W_T}$ or $\mathbf{W_L}$ as data. These results are shown in Table 6.3. Pleasingly, for both models, the topology from which the curves were simulated produces the largest maximised log-likelihood values and hyperparameter estimates closest to the true values, whilst the wrong topologies tend to decrease the value of the second time difference. Particularly, when we only have data at the terminal nodes available, the model pulls t_2 towards 0, meaning it infers that the three two-dimensional curves branch straight from the root, with no internal nodes. Estimates for the other hyperparameters remain remarkably constant across optimisations. It is surprising how robust these correlation parameters are to incorrect specification of the topology.

All data available

$\hat{\boldsymbol{\theta}}$	$\hat{\sigma}_f$	$\hat{\lambda}$	$\hat{\kappa}_1$	$\hat{t}_1$	$\hat{t}_2$
Topology	(5, 4, 4, 5)				(Right)
Estimates	0.2932	0.4235	−0.6603	0.2038	0.2289
SE	0.0143	0.0250	0.0566	0.0483	0.0657
$\log\left(L(\hat{\boldsymbol{\theta}})\right)$	372.3341				
Topology	(4, 4, 5, 5)				(Wrong)
Estimates	0.2859	0.4230	−0.5255	0.3122	0.1400
SE	0.0139	0.0261	0.0756	0.0705	0.0353
$\log\left(L(\hat{\boldsymbol{\theta}})\right)$	358.703				
Topology	(4, 5, 4, 5)				(Wrong)
Estimates	0.2876	0.4204	−0.6120	0.3410	0.1550
SE	0.0140	0.0256	0.0629	0.0726	0.0399
$\log\left(L(\hat{\boldsymbol{\theta}})\right)$	359.1795				

Data at leaves available

$\hat{\boldsymbol{\theta}}$	$\hat{\sigma}_f$	$\hat{\lambda}$	$\hat{\kappa}_1$	$\hat{t}_1$	$\hat{t}_2$
Topology	(5, 4, 4, 5)				(Right)
Estimates	0.2957	0.4342	−0.6046	0.1647	0.2960
SE	0.0186	0.0308	0.0888	0.0563	0.1422
$\log\left(L(\hat{\boldsymbol{\theta}})\right)$	198.8515				
Topology	(4, 4, 5, 5)				(Wrong)
Estimates	0.3001	0.4284	−0.6763	0.3691	0.0000
SE	0.0189	0.0320	0.0700	0.1083	0.1422
$\log\left(L(\hat{\boldsymbol{\theta}})\right)$	194.9868				
Topology	(4, 5, 4, 5)				(Wrong)
Estimates	0.3005	0.4284	−0.6763	0.3691	0.0000
SE	0.0189	0.0331	0.0827	0.0963	0.1713
$\log\left(L(\hat{\boldsymbol{\theta}})\right)$	194.9868				

Table 6.3: Maximum likelihood estimates for data at all nodes and just at the leaves under all three possible topologies.

6.4 A case study: the evolution of nose shape within and between ethnic groups

To study the model on real data, a small case study was conducted. Using the ©*Di3D* 3-dimensional surface-imaging device (see Chapter 2), facial images were collected from volunteer subjects recruited in the local community. Ethical permissions were obtained from the Ethics Committee of the College of Science and Engineering, University of Glasgow.

Three ethnic groups were selected for the study: African, Asian and White European. Due to the variability in the facial characteristics within these ethnic groups, we focused on three subgroups, namely: Sub-Saharan African, Chinese and White British. The number of male and female volunteers was unbalanced so it was decided to conduct the study using only male subjects. The final database consists of 12 Sub-Saharan African, 20 White British and 12 Chinese males. Table 6.4 shows the mean age and the age range for each group.

Group	Age range	Mean age
Sub-Saharan African	20-50	34.58
White British	19-50	28.55
Chinese	22-28	24.75

TABLE 6.4: Age range and mean for each ethnic group.

For each subject, the two curves chosen to identify the nose are extracted. These curves are:

- mid-line nasal profile: ridge points from the nasal root along the dorsum of the nose and the columella, defined by 28 equally-spaced points;
- nasal bridge: which outlines the width of the nose from one alar facial groove to the other, defined by 13 equally-spaced points.

To be able to properly study the similarities and dissimilarities of the curves, these have to be first made comparable. This was done using General Procrustes Analysis, introduced in Section 3.3.2. An example of the nose curves from one participant from each ethnic group is shown in Figure 6.4. The curves are shown over facial surfaces rotated at different angles for a better appreciation of the three-dimensional curves.

FIGURE 6.4: Nose curves: mid-line nasal profile (blue) and nasal bridge (green). From top to bottom: African, British and Chinese subjects.

6.4.1 Evolution of mean nose shape

To study the evolution of nose shape, it was decided to start by modelling the evolution of the mean shape of the nose through a phylogenetic tree. For this, the corresponding three-dimensional points of each curve were averaged for each group, resulting in sets of points from two three-dimensional mean curves per group (mid-line nasal profile and nasal bridge).

These means can be plotted for each coordinate as a function of the arc-length re-scaled from 0 to 1, as before. Moreover, the points on each coordinate curve have the mean over that curve subtracted to match the assumption of zero mean in the GP model. The profile means, consisting of 28 equally-spaced points, are

shown in Figure 6.5, with each group represented by its first letter, conveniently, *A*-African, *B*-British and *C*-Chinese. It should be appreciated how, by the very nature of the nose profile, the x coordinate remains constant (no subjects with broken noses were included in the sample) and therefore it adds no information regarding the differences between groups. For this reason, it was decided to model the nose profiles as two-dimensional curves, omitting coordinate x.

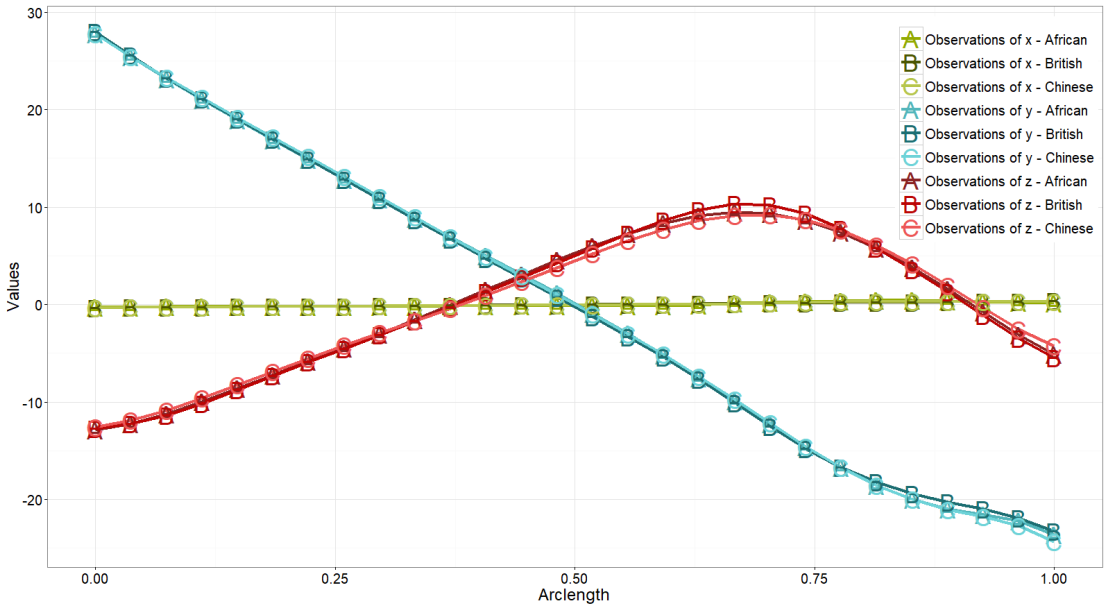


FIGURE 6.5: Mean mid-line nasal profile for the three ethnic groups, each coordinate plotted against the arc-length.

The means of the nasal bridges for each group are shown in Figure 6.6. The bridges are characterised by 13 equally-spaced points. More differences between groups can be appreciated in these curves. Moreover, there is enough variation in all the coordinates to justify applying the model for three-dimensional curves.

In both cases, for the two- and three-dimensional models, data are only available at the terminal nodes. Given there are three leaves, the tree structure is like that used in the simulations presented above, where two of the leaves share one common ancestor (say D) and the three terminal nodes have the root (say E) of the tree as most recent common ancestor. If the three groups had evolved simultaneously from just one common ancestor, without any internal ancestors, this would be shown by analyses based on all the three bifurcating topologies estimating the time difference between the root and the internal node (i.e., between D and E) to be zero.

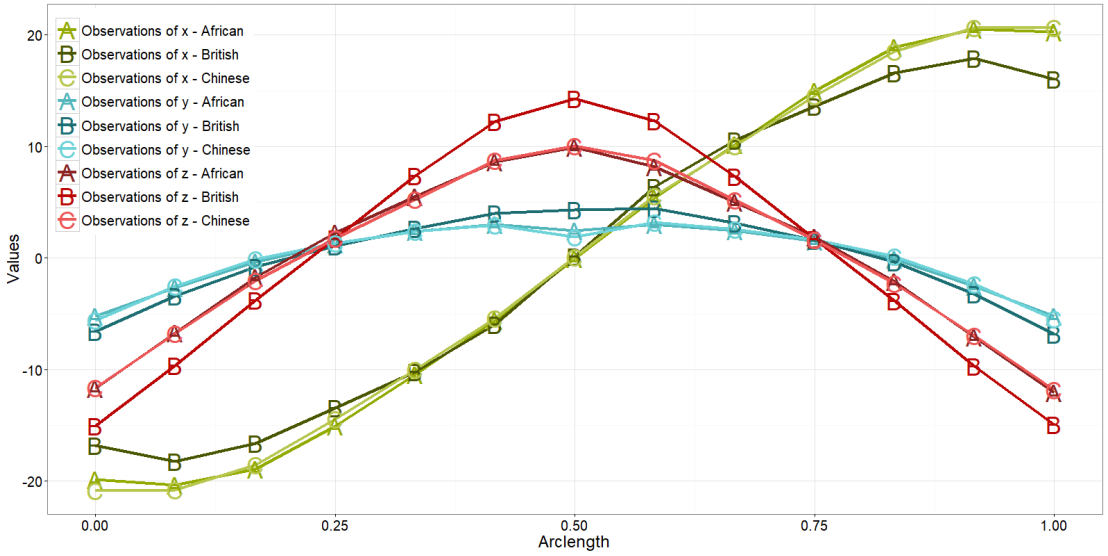


FIGURE 6.6: Mean nasal bridge for the three ethnic groups, each coordinate plotted against the arc-length.

The three possible topologies are $(A,(B,C))$, $(B,(A,C))$ and $(C,(A,B))$ (in Newick format, as defined in Section 1.3). Figure 6.7 illustrates one of the possible tree structures. Moreover, ordering the nodes from A to E, and assigning the data at the leaves time zero, the vectors of times could be written $(0,0,0,t_D,t_E)$, and hence the same time differences as before, t_1 and t_2 can be optimised.

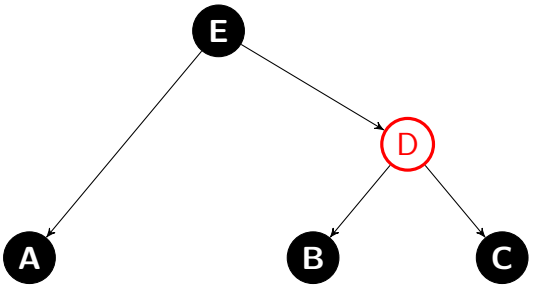


FIGURE 6.7: Illustration of one possible topology.

Let $\boldsymbol{\theta}_{2D}$ denote the set of hyperparameters for the model of two-dimensional nose profile curves. Since the model is for coordinates y and z, let the correlation between these be κ_2 and hence $\boldsymbol{\theta}_{2D} = (\sigma_{f2D}, \lambda_{2D}, \kappa_{22D}, t_{12D}, t_{22D})$. Similarly, for the model for the nasal bridges, let $\boldsymbol{\theta}_{3D} = (\sigma_{f3D}, \lambda_{3D}, \kappa_{13D}, \kappa_{23D}, \kappa_{33D}, t_{13D}, t_{23D})$ be its set of hyperparameters. Note $\sigma_{f2D} \neq \sigma_{f3D}$, $\lambda_{2D} \neq \lambda_{3D}$, and so on. Both models can be optimised by maximum likelihood, either separately or by adding the sum of the log-likelihood values of the two- and three-dimensional datasets and optimising

the hyperparameters simultaneously. If both models are optimised simultaneously, there are three possible scenarios:

1. Force both sets of curves to have the same time differences: $t_{12\mathtt{D}} = t_{13\mathtt{D}}$ and $t_{22\mathtt{D}} = t_{23\mathtt{D}}$, i.e., assume the nose profile curves and the nasal bridges have evolved at the same rate, and have diverged at the same time points in history. Therefore, the full set of hyperparameters: $\boldsymbol{\theta} = (\sigma_{f2\mathtt{D}}, \lambda_{2\mathtt{D}}, \kappa_{22\mathtt{D}}, \sigma_{f3\mathtt{D}}, \kappa_{13\mathtt{D}}, \kappa_{23\mathtt{D}}, \kappa_{33\mathtt{D}}, t_1, t_2)$.

2. Optimise the same time difference for both sets, but allow for a scaling parameter multiplying the rate of change ($\mu_{\mathbf{T}}$). In this scenario, $\mu_{\mathbf{T}}$ is fixed to 1 for the three-dimensional data, and a hyperparameter $\mu_{\mathbf{T}2\mathtt{D}}$ is optimised, representing the relative rate of change of profile to bridge. Therefore $\boldsymbol{\theta} = (\sigma_{f2\mathtt{D}}, \lambda_{2\mathtt{D}}, \mu_{\mathbf{T}2\mathtt{D}}, \kappa_{22\mathtt{D}}, \sigma_{f3\mathtt{D}}, \kappa_{13\mathtt{D}}, \kappa_{23\mathtt{D}}, \kappa_{33\mathtt{D}}, t_1, t_2)$.

3. Allow for each set of curves to have different times t_1 and t_2. This is equivalently to model each set of curves independently and the set of hyperparameters is $\boldsymbol{\theta} = (\boldsymbol{\theta}_{2\mathtt{D}}, \boldsymbol{\theta}_{3\mathtt{D}})$.

Model selection can be performed to select the best scenario. When fitting models, it is possible to increase the likelihood by adding parameters, but doing so may result in over-fitting. In this study, the first option has two time hyperparameters, the second three and the last one four (10, 11 and 12 hyperparameters in total, respectively). Whilst the model with more free hyperparameters may result in a larger log-likelihood since the other options are merely special cases of it, it may not be completely necessary to have all the time differences to represent the data well.

Results

The first task is to find the right topology for both the nose profiles and the nasal bridges, for each possible model. Hyperparameters were optimised by maximum likelihood for the three possible topologies and the topology with the largest log-likelihood value chosen as the real topology. For every possible model and for modelling both sets of curves independently, the topology producing the highest values was $(B, (A, C)))$. That is, the African and Chinese mean curves have a most recent ancestor more recently than each with the British.

Same times

$\hat{\boldsymbol{\theta}}$	$\hat{\sigma}_{f2\mathtt{D}}$	$\hat{\lambda}_{2\mathtt{D}}$	$\hat{\kappa}_{22\mathtt{D}}$	$\hat{\sigma}_{f3\mathtt{D}}$	$\hat{\lambda}_{3\mathtt{D}}$	$\hat{\kappa}_{13\mathtt{D}}$	$\hat{\kappa}_{23\mathtt{D}}$	$\hat{\kappa}_{33\mathtt{D}}$	$\hat{t}_1$	$\hat{t}_2$
Estimates	8.2505	0.1519	0.1351	8.589	0.1116	0.0052	0.7464	−0.0115	0.0073	0.0107
SE	0.5197	0.0039	0.1164	0.5614	0.0030	0.1588	0.0670	0.1581	0.0017	0.0032
$\log\left(L(\hat{\boldsymbol{\theta}})\right)$	−56.59313									

Same times + Scaling parameter for $\mu_{\mathbf{T}}$

$\hat{\boldsymbol{\theta}}$	$\hat{\sigma}_{f2\mathtt{D}}$	$\hat{\lambda}_{2\mathtt{D}}$	$\hat{\mu}_{\mathbf{T}2\mathtt{D}}$	$\hat{\kappa}_{22\mathtt{D}}$	$\hat{\sigma}_{f3\mathtt{D}}$	$\hat{\lambda}_{3\mathtt{D}}$	$\hat{\kappa}_{13\mathtt{D}}$	$\hat{\kappa}_{23\mathtt{D}}$	$\hat{\kappa}_{33\mathtt{D}}$	$\hat{t}_1$	$\hat{t}_2$
Estimates	7.5779	0.1520	0.9463	0.1408	6.9561	0.1113	0.0085	0.7572	−0.0183	0.0069	0.0104
SE	0.4773	0.0041	0.3783	0.1168	0.4547	0.0033	0.1614	0.0628	0.1601	0.0024	0.0036
$\log\left(L(\hat{\boldsymbol{\theta}})\right)$	−56.58461										

Different times

$\hat{\boldsymbol{\theta}}$	$\hat{\sigma}_{f2\mathtt{D}}$	$\hat{\lambda}_{2\mathtt{D}}$	$\hat{\kappa}_{22\mathtt{D}}$	$\hat{t}_{12\mathtt{D}}$	$\hat{t}_{22\mathtt{D}}$	$\hat{\sigma}_{f3\mathtt{D}}$	$\hat{\lambda}_{3\mathtt{D}}$	$\hat{\kappa}_{13\mathtt{D}}$	$\hat{\kappa}_{23\mathtt{D}}$	$\hat{\kappa}_{33\mathtt{D}}$	$\hat{t}_{13\mathtt{D}}$	$\hat{t}_{23\mathtt{D}}$
Estimates	7.3722	0.1496	0.0716	0.0091	0.0062	6.8717	0.1068	0.0306	0.7363	−0.0023	0.0036	0.0149
SE	0.4644	0.0041	0.1204	0.0026	0.0035	0.4492	0.0035	0.1577	0.0698	0.1592	0.0012	0.0052
$\log\left(L(\hat{\boldsymbol{\theta}})\right)$	−53.59132											

Table 6.5: Optimal hyperparameters for the three scenarios.

The optimal hyperparameters values are shown in Table 6.5. Note the small differences in the values of the log-likelihood, particularly between the model for the same times and the model that allows for a scaling parameter for $\mu_{\mathbf{T2D}}$. This is due to the fact that the scaling parameter is estimated to be close to one, and in fact, its approximate 95% Wald confidence interval contains one. This implies that there is indeed not a significant difference between the rate of change of the nose profiles and the nasal bridges. This can also be illustrated by calculating the 95% confidence intervals for all the optimised time difference values (Table 6.6).

	$\hat{t}_1$	$\hat{t}_2$
3D	$[0.0012, 0.0060]$	$[0.0045, 0.0253]$
2D	$[0.0039, 0.0143]$	$[0.0000, 0.0132]$
Same times	$[0.0039, 0.0107]$	$[0.0043, 0.0171]$
Same times - diff $\mu_{\mathbf{T}}$	$[0.0021, 0.0070]$	$[0.0032, 0.0176]$

TABLE 6.6: 95 % Confidence intervals for time differences.

All intervals for both t_1 and t_2 overlap, and therefore there is no significant difference between the optimal values from the different scenarios. Regarding the other hyperparameters, they remain more or less stable, when allowing for the large SE of some of them, particularly the signal standard deviations σ_{f2D} and σ_{f3D}. An error ratio η was added to the diagonal of the covariance matrices to accommodate errors in the observed values, defined as $\eta = \sigma_n^2/\sigma_f^2$. It was fixed to $\eta = 0.01$. Given the resulting optimal values for $\hat{\sigma}_{f2D}$ and $\hat{\sigma}_{f3D}$, the final additive normal error have standard deviations (σ_n) of approximately 0.74 mm and 0.86 mm, respectively.

	AIC	AICc	BIC
1. Same times	133.19	133.99	169.71
2. Same times, different rates	135.17	136.14	175.35
3. Different times	131.18	132.33	175.01

TABLE 6.7: AIC, AICc and BIC for the three possible models.

It would appear that the simplest model would be best. Nevertheless, the Akaike information criterion (AIC), its version corrected for finite sample sizes (AICc) and Bayesian information criterion (BIC) [Posada and Buckley, 2004] are calculated for each model (Table 6.7). The model with the lowest criterion is preferred. Both, AIC and BIC, balance improvements in fit (coming from the likelihood) against a penalty for introducing additional parameters. Despite various subtle theoretical differences, their only difference in practice is the size of the penalty; BIC penalizes model complexity more heavily. Model selection has to be based on a well-justified

criterion of what is the 'best' model. The criterion must be estimable from the data for each fitted model and must fit into a general statistical inference framework, i.e., either a likelihood or Bayesian framework. In broad terms, AIC represents the information-theoretic selection based on Kullback-Leibler information loss and BIC is an approximation to the Bayesian model selection [Burnham and Anderson, 2004]. AIC can also be viewed as a measure of the relative quality of a statistical model for a given set of data [Burnham and Anderson, 2003] while BIC is a criterion for model selection among a finite set of models [Raftery, 1995]. The AIC or BIC for a model is usually written in the form $[-2\log\left(L(\hat{\boldsymbol{\theta}})\right) + kp]$, where $L(\hat{\boldsymbol{\theta}})$ is the maximised likelihood, p is the number of parameters in the model, and k is 2 for AIC and $\log(n)$ for BIC, with n being the number of observations. The corrected version of AIC is defined as AICc $=$ AIC $+ (2p(p+1))/(n-p-1)$. AICc is essentially AIC with a greater penalty for extra parameters. Using AIC, instead of AICc, when n is not many times larger than p^2, increases the probability of selecting models that have too many parameters, i.e., overfitting [Claeskens et al., 2008].

Since BIC penalises more complex models more severely, and given that the time differences are not significantly different between models, it was decided to keep the model with the same times for the profiles and the bridges, i.e., the model with fewest parameters (scenario 1).

Prediction at nodes D and E

The topology $(B,(A,C))$, which had the maximal log-likelihood value, is considered to make predictions at the internal node D and at the root, E, with the aim of visualising the estimated mean nose shape of the ancestors. Recall for the two-dimensional nose profile curves $\mathbf{s}_{\texttt{2D}} = (s_1 \cdots s_{28})^{\mathrm{T}}$ and for the three-dimensional nasal bridge curves, $\mathbf{s}_{\texttt{3D}} = (s_1 \cdots s_{13})^{\mathrm{T}}$. The set of optimal hyperparameters, $\hat{\boldsymbol{\theta}} = (\hat{\sigma}_{f\texttt{2D}}, \hat{\lambda}_{\texttt{2D}}, \hat{\kappa}_{\texttt{22D}}, \hat{\sigma}_{f\texttt{3D}}, \hat{\lambda}_{\texttt{3D}}, \hat{\kappa}_{\texttt{13D}}, \hat{\kappa}_{\texttt{23D}}, \hat{\kappa}_{\texttt{33D}}, \hat{t}_1, \hat{t}_2) = (8.2505, 0.1519, 0.1351, 8.589, 0.1116, 0.0052, 0.7464, -0.0115, 0.0073, 0.0107)$, is used to make predictions at 30 equally spaced spatial-points, $\mathbf{s}^*_{\texttt{2D}}$, for the profile curves, and 20 spatial-points, $\mathbf{s}^*_{\texttt{3D}}$, also equally spaced, for the nasal bridges. The predicted curves at node D, the common ancestor of A (African) and C (Chinese), are shown in Figure 6.8, for the profile and Figure 6.9, for the nasal bridge. It can be seen how, especially in the nasal bridges, the posterior predictive mean is much closer to A and C than

to B (British). The posterior means are displayed with two standard deviations confidence bands. As there is little variation among the nose profile curves, it can be seen all the observed values from the three ethnic groups are contained within them. For the bridges, the observed values of B differ from A, C and the predictions at D remarkably, with the middle spatial-points not even contained in the confidence bands of D.

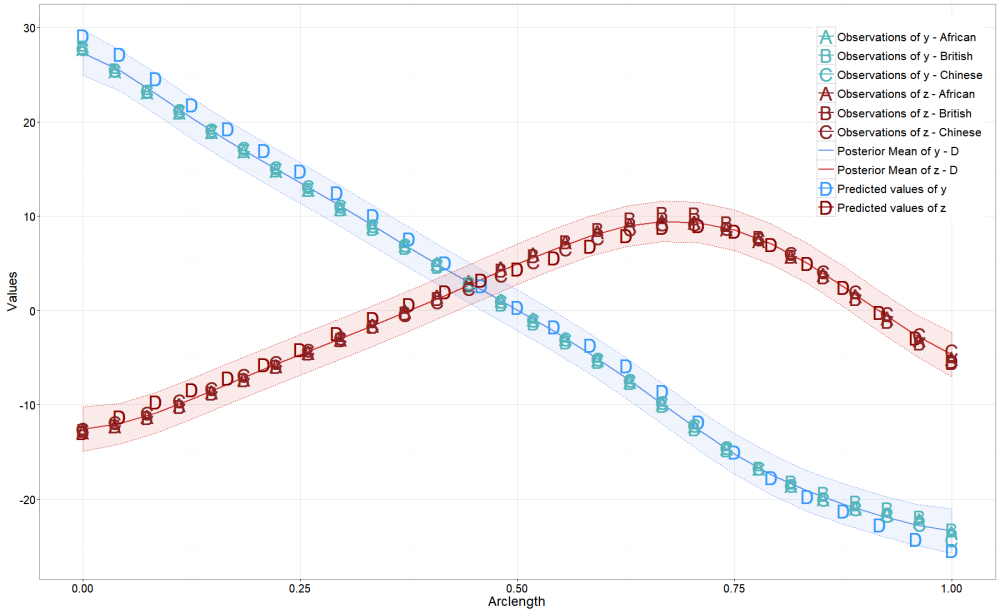


FIGURE 6.8: Observations, posterior means and one draw from the predictive distribution for two-dimensional nose profile curves at node D.

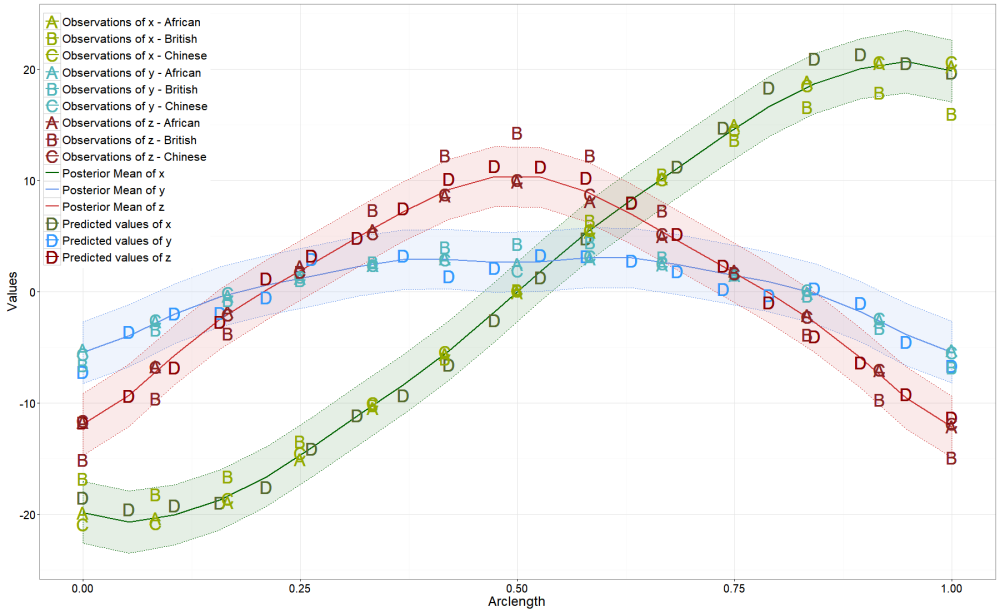


FIGURE 6.9: Observations, posterior means and one draw from the predictive distributions for three-dimensional nasal bridge curves at node D.

The predictions for the root of the tree, node E, are shown in Figure 6.10, for the profile and Figure 6.11, for the nasal bridge. Note how much wider the confidence bands are in this case, given that predictions are being made even further back in time from the observed values, and therefore there is more uncertainty. It can also be seen how in this case, the bands contain all the observed values, since E is the most recent common ancestor for the three ethnic groups together. Note also how the posterior mean of D is closer to the observed values of A and C, than the posterior mean of E is to all the observed values. This is due to the time difference between A and C and D being 0.0073, whilst between the root E and the leaves, it is 0.018 $(0.0073 + 0.0107)$. Nonetheless, it is important to recall the data are measured in millimetres, and therefore, even the widest bands represent uncertainty of no more than 4 mm. The predicted values for D and E in both the two- and three-dimensional curves, are a random draw using the predictive distributions from (6.14). Viewing all the observed and predicted values, it is clear too that most changes have occurred in the nasal bridge, i.e., in the broadness of the nose, and the size of the nostrils. The subtle changes in the profile curves are associated with the position of the tip of the nose.

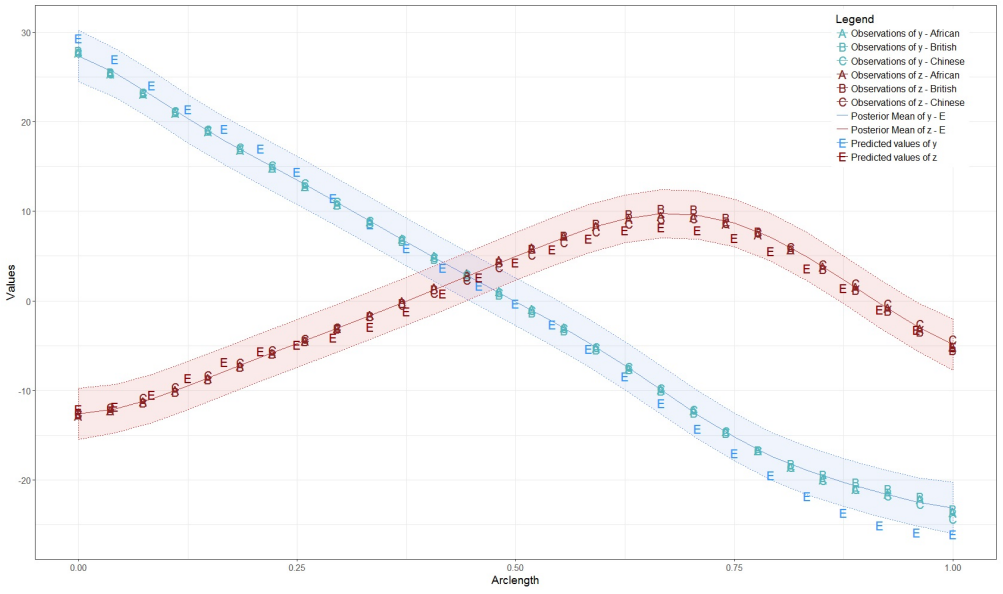


FIGURE 6.10: Observations, posterior means and one draw from the predictive distribution for two-dimensional nose profile curves at node E.

Using Procrustes analysis, the posterior means can be translated back to the original three-dimensional space of the nose curves, allowing a more comprehensive exploration of the ancestors' nose shape. These shapes are illustrated in Figure

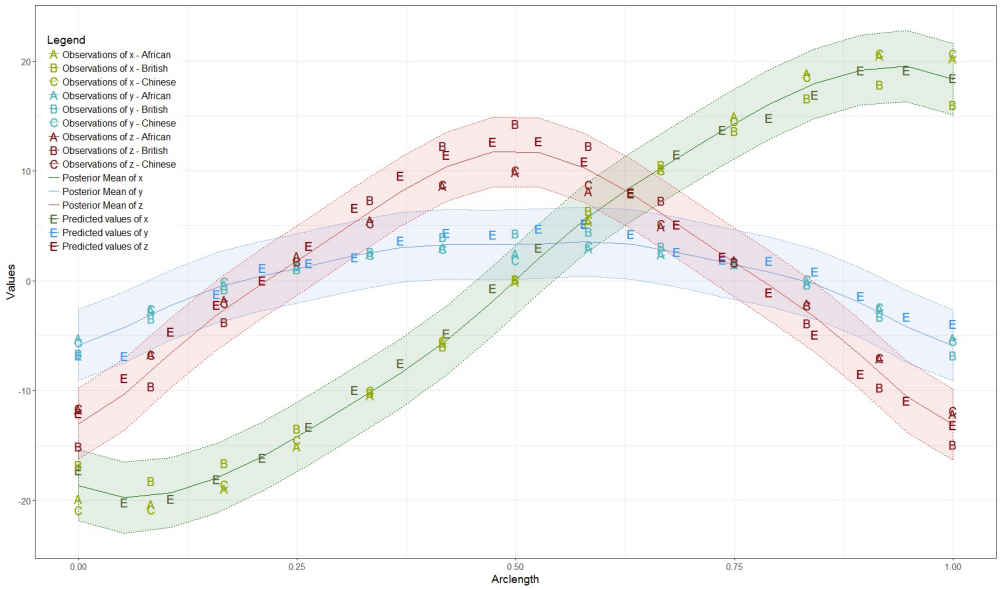
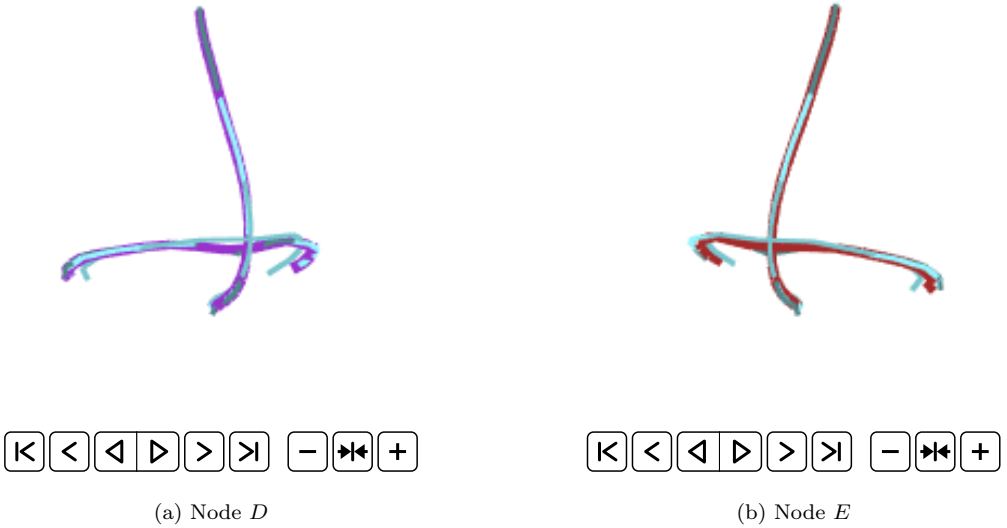


FIGURE 6.11: Observations, posterior means and one draw from the predictive distribution for three-dimensional nasal bridge curves at node E.

6.12. Different angles of view can be seen when viewed in digital form (using Adobe reader). The posterior mean for node D, the common ancestor for A and C, is shown on the left. The posterior mean for the root is displayed on the right. The means of the three ethnic groups are displayed in blue, from the African in the lightest blue, from the British in an intermediate shade and from the Chinese in the darkest shade of blue.



(a) Node D (b) Node E

FIGURE 6.12: Posterior mean for prediction at nodes D and E, displayed with the means of the original data.

6.4.2 Inter- and intra-group variation in nose shape

In the study above, the mean nose curves for each ethnic group are modelled. This approach leaves out the variability within groups. Although the sample size in each group is small for a comprehensive analysis, an attempt to fit a model of all data is illustrated here. The database consists of 12 African, 20 British and 12 Chinese, so that there are 44 terminal nodes. All the curves in each group are assumed to have simultaneously diverged from one common ancestor, different between groups. This is to reduce the number of possible topologies. Moreover, if this assumption was not made, it could be that curves from different ethnic groups would be grouped under one common ancestor, rather than each group having its own common ancestor, which should be more recent than the ancestors common to different ethnic groups. It is expected that the topology with the highest likelihood value will match the one of the mean curves, i.e., $(B,(A,C))$ (shown in Figure 6.13). Simulation studies for this kind of multifurcating tree can be found in Appendix C.

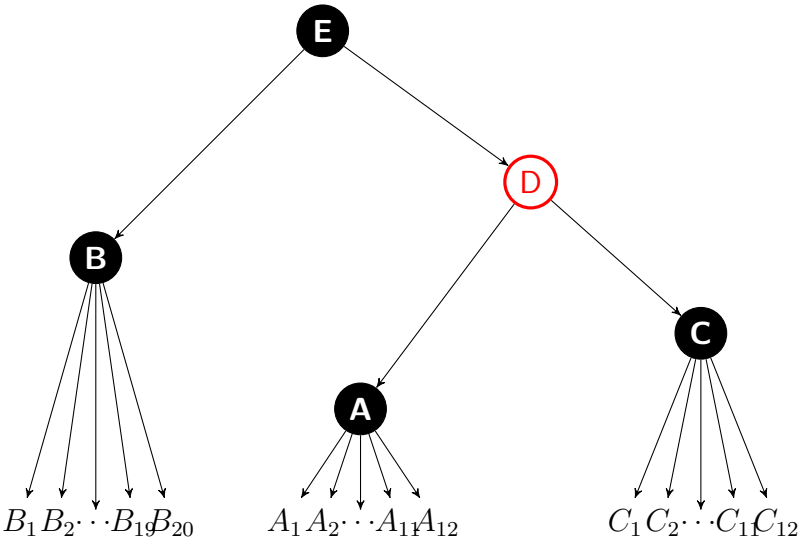


FIGURE 6.13: Illustration of tree with all curves at leaves.

As before, if there were no internal node D, it would be shown by the time difference between D and E having an optimal value of zero. The times between the curves at the leaves and their common ancestor can differ from one ethnic group to another, accounting for each group's distinct variability. The larger the time difference between curves and their most recent common ancestor, the larger the variability. The times have been randomly chosen in Figure 6.13 for illustration.

If the time of the leaves is assumed to be zero (present time), the vector of node times is $(0, \ldots, 0, t_A, t_B, t_C, t_D, t_E)$.

All the data are shown in Figure 6.14. Each nasal curve is plotted with each coordinate as a function of the arc-length.

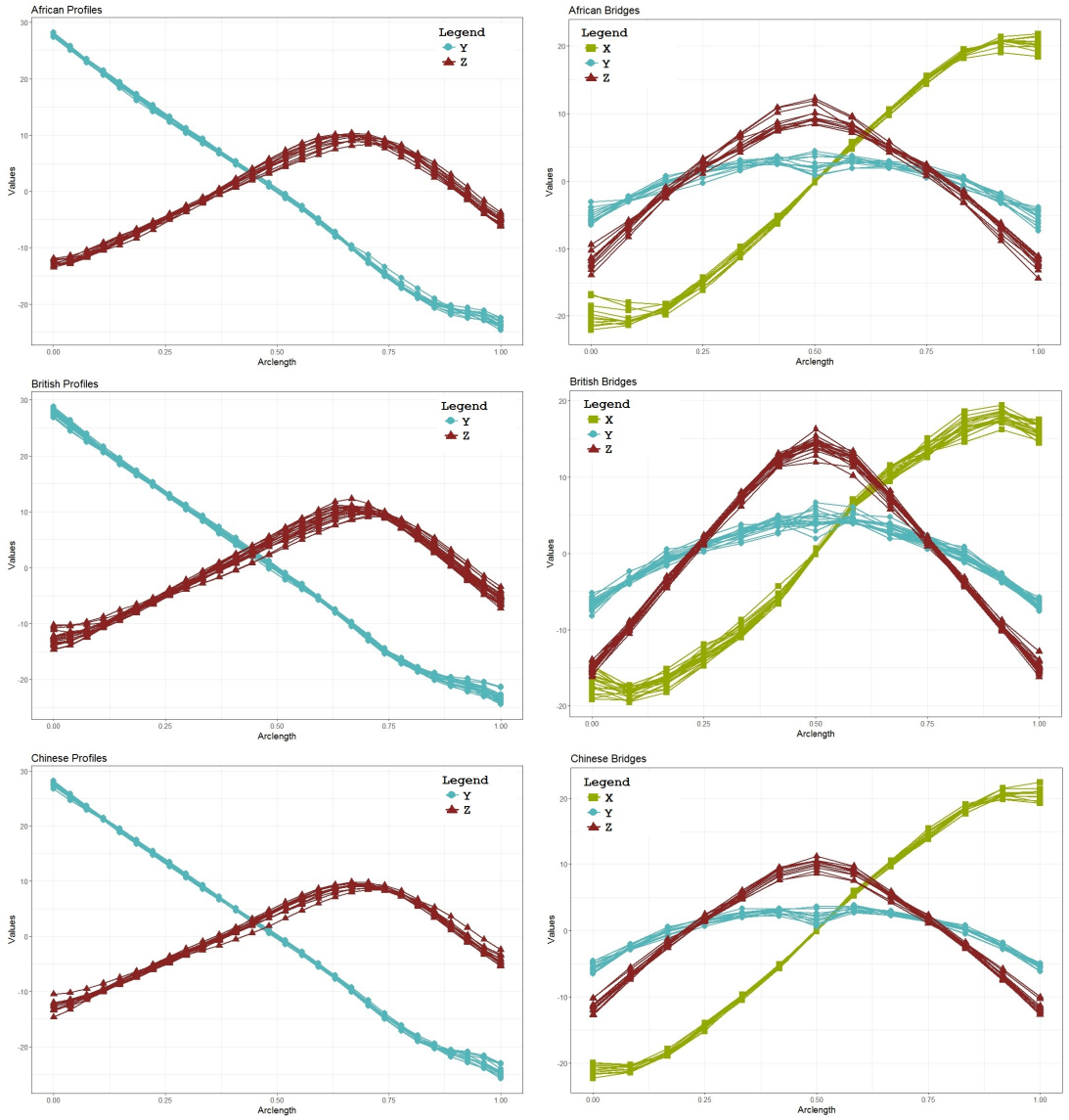


FIGURE 6.14: Nose profile (2D) and nasal bridge (3D) curves available from each ethnic groups. All curves plotted with each coordinate as a function of the arc-length.

As before, to select the best topology, optimal hyperparameters by maximum likelihood were found for each of the three possible topologies. From these results, the topology with the highest log-likelihood value is chosen. Two trees were studied,

one for the evolution of three-dimensional nasal bridge curves, and one for the two-dimensional nose profile curves.

Optimisation was done, as previously, for the time differences rather than the node times. In this scenario, there are five time differences, and one needs to be careful to satisfy constraints on those parameters. Depending on the distance between each group's leaves and their ancestor, some time differences are allowed to be negative. However, there can be no negative node times and the resulting times for the internal nodes need to be smaller than the time of the root. The differences are ordered as: $t_1 = t_B$, $t_2 = t_A - t_B$, $t_3 = t_C - t_A$, $t_4 = t_D - t_C$ and $t_5 = t_E - t_D$.

Nasal Bridges (3D) results

Estimates for the hyperparameters for the set of three-dimensional curves are shown in Table 6.8, together with their standard errors SE.

Nasal bridges (3D)

$\hat{\boldsymbol{\theta}}$	$\hat{\sigma}_{f3D}$	$\hat{\lambda}_{3D}$	$\hat{\kappa}_{13D}$	$\hat{\kappa}_{23D}$	$\hat{\kappa}_{33D}$		
Ests	7.1679	0.0830	0.1108	0.0162	−0.0181		
SE	0.1224	0.0010	0.0382	0.0423	0.0466		
			$\hat{t}_1$	$\hat{t}_2$	$\hat{t}_3$	$\hat{t}_4$	$\hat{t}_5$
		Ests	0.0055	−0.0006	−0.0021	0.0027	0.0059
		SE	0.0010	0.0007	0.0003	0.0003	0.0011
$\log\left(L(\hat{\boldsymbol{\theta}})\right)$	−1511.343						

Table 6.8: Optimal hyperparameters for the multifurcating tree for the nasal bridges.

Nasal bridges (3D)

	$\hat{t}_A$	$\hat{t}_B$	$\hat{t}_C$	$\hat{t}_D$	$\hat{t}_E$
Ests	0.0049	0.0055	0.0028	0.0055	0.0114

Table 6.9: Estimated node times for the nasal bridges.

Note how small these differences are in general, and their rather large SE. To better understand the times, the differences are translated back to the node times in Table 6.9 and displayed in Figure 6.15, which provides a clear picture of the relationships between and within ethnic groups. As when studied the mean nasal bridge curves, the topology with the maximum likelihood is the one that links Africans and Chinese under a more recent common ancestor (D) than the common ancestor of all groups, E.

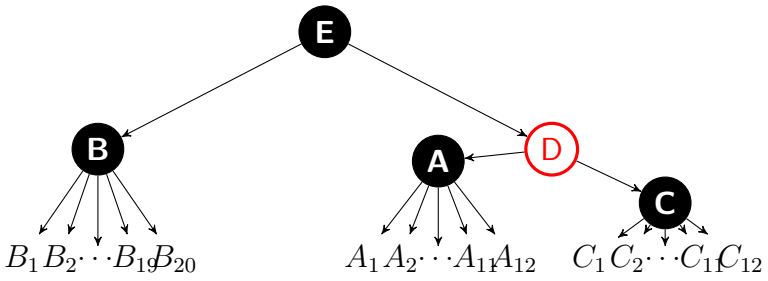


FIGURE 6.15: Tree for nasal bridge curves, with branch lengths corresponding to the fitted model.

Nose Profiles (2D) results

When studying the multifurcating tree for the nose profiles (two-dimensional curves), problems arose due to the very little variability between the curves. The differences between node times are even smaller than those found for the nasal bridges. This resembles the problems found when modelling the sequence of lip curves for an emotion (Section 5.3.6), where the change from one picture to the next was very small and the estimation of a large μ was problematic. Small distances between the nodes are equivalent to a large $\mu_{\mathbf{T}}$. However, since the hyperparameter $\mu_{\mathbf{T}}$ is now fixed to 1, a different approach to the problem had to be investigated, in relationship with the branch lengths. A heuristic search to find the topology that could optimise this data is shown below, starting from the tree in Figure 6.16. The optimal hyperparameters are shown in Table 6.10. Recall, the time differences are ordered as: $t_1 = t_B$, $t_2 = t_A - t_B$, $t_3 = t_C - t_A$, $t_4 = t_D - t_C$ and $t_5 = t_E - t_D$. The optimal node times are shown in Table 6.11.

NOSE PROFILES (2D)

$\hat{\boldsymbol{\theta}}$	$\hat{\sigma}_{f2D}$	$\hat{\lambda}_{2D}$	$\hat{\kappa}_{22D}$				
Ests	7.9106	−0.0532	−0.0888				
			$\hat{t}_1$	$\hat{t}_2$	$\hat{t}_3$	$\hat{t}_4$	$\hat{t}_5$
			0.0021	−0.0005	−0.0003	0.0005	0.0003
$\log\left(L(\hat{\boldsymbol{\theta}})\right)$	415.34						

TABLE 6.10: Optimal hyperparameters for the multifurcating tree for the nose profiles.

NOSE PROFILES (2D)

	$\hat{t}_A$	$\hat{t}_B$	$\hat{t}_C$	$\hat{t}_D$	$\hat{t}_E$
Ests	0.0016	0.0021	0.0013	0.0018	0.0021

TABLE 6.11: Node times for the two-dimensional nose profiles curves.

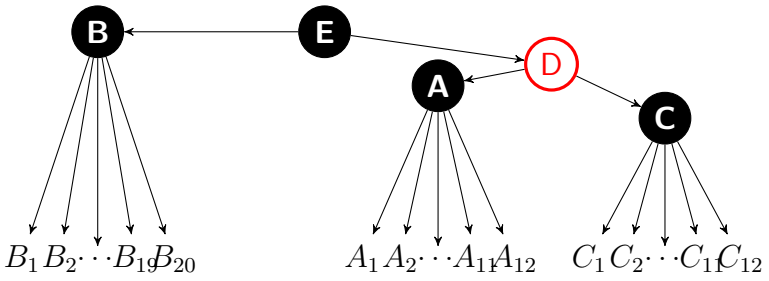


FIGURE 6.16: Tree for nose profile curves.

It should be noted that the root of the tree has the same time as that of the common ancestor for the British curves, i.e., $\hat{t}_B = \hat{t}_E$. This is illustrated in Figure 6.16. This implies that t_5 lies on the boundary. If t_5 was smaller than 0.0003, it would result in $\hat{t}_E$ being smaller than 0.0021, and therefore node B would have occurred before the root of the tree, which is not allowed. The use of the Hessian matrix to calculate the standard errors cannot be trusted when the maximum is at a boundary. Recall the Hessian matrix is defined as the matrix of second derivatives of the log-likelihood function with respect to the hyperparameters. The inverse of minus the Hessian is the variance-covariance matrix of the maximum likelihood estimates. The standard errors of the estimates can be calculated as the square roots of the diagonal terms in the variance-covariance matrix. Intuitively, the precision of the estimates depends on the curvature of the log-likelihood function near these. The more peaked the log-likelihood function, the less the uncertainty in the hyperparameter estimate. The information delivered by the Hessian is the local curvature of a function, based on the assumption that the gradient (vector of first derivatives of the log-likelihood) is equal to zero. If an estimate lies on a boundary, these properties of the curvature are no longer applicable. Hence, the SEs calculated from the Hessian matrix are not reliable, and the estimates are unpredictable.

The results shown above indicate that the branch between E and D is as short as it can be, which could imply it may not exist, i.e. there is no internal ancestor D. For this reason, a multifurcating tree with no internal node was studied. The differences are ordered as: $t_1 = t_B$, $t_2 = t_A - t_B$, $t_3 = t_C - t_A$ and $t_4 = t_E - t_C$. Optimal hyperparameters are shown in Table 6.12, and the final node times are shown in Table 6.13.

Nose profiles (2D)

$\hat{\boldsymbol{\theta}}$	$\hat{\sigma}_{f2D}$	$\hat{\lambda}_{2D}$	$\hat{\kappa}_{22D}$			
Ests	7.3011	-0.0532	-0.0888			
			$\hat{t}_1$	$\hat{t}_2$	$\hat{t}_3$	$\hat{t}_4$
			0.002311	-0.0003	-0.000301	0.000601
$\log\left(L(\hat{\boldsymbol{\theta}})\right)$	412.5987					

Table 6.12: Optimal hyperparameters for the second multifurcating tree for the nose profiles.

Nose profiles (2D)

	$\hat{t}_A$	$\hat{t}_B$	$\hat{t}_C$	$\hat{t}_E$
Ests	0.002011	0.002311	0.001710	0.002311

Table 6.13: Node times for the second multifurcating tree for nose profiles curves.

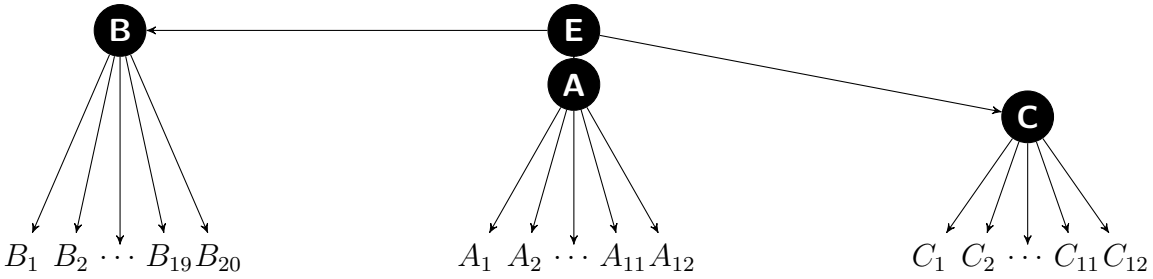


Figure 6.17: Second tree for nose profile curves.

In this scenario, nodes B and E are also estimated at the same time, i.e., $\hat{t}_B = \hat{t}_E$. The last time difference $\hat{t}_4$ lies therefore on a boundary and, again, the Hessian matrix can not be used to calculate the SEs.

Since both of the above trees have estimated $\hat{t}_B = \hat{t}_E$, it was decided to set node B to be the root of the tree. This multifurcating tree is shown in Figure 6.18, with the branch lengths, as usual, proportional to the optimal time differences. In this tree, none of the time differences hits a boundary (see Table 6.14 for all the optimal hyperparameters and Table 6.15 for the corresponding node times). In this case, the time differences are defined as: $t_1 = t_C, t_2 = t_A - t_C$ and $t_3 = t_B - t_A$.

NOSE PROFILES (2D)

$\hat{\boldsymbol{\theta}}$	$\hat{\sigma}_{f2D}$	$\hat{\lambda}_{2D}$	$\hat{\kappa}_{22D}$		
Ests	7.1535	-0.053200	-0.0888		
SE	0.1019	2.5×10^{-4}	2.9×10^{-2}		
			$\hat{t}_1$	$\hat{t}_2$	$\hat{t}_3$
		Ests	0.0018	0.000301	0.0003
		SE	1.9×10^{-6}	1.9×10^{-6}	1.9×10^{-6}
$\log\left(L(\hat{\boldsymbol{\theta}})\right)$	412.4838				

TABLE 6.14: Optimal hyperparameters for the last multifurcating tree for the nose profiles.

NOSE PROFILES (2D)

	$\hat{t}_A$	$\hat{t}_B$	$\hat{t}_C$
Ests	0.002101	0.002401	0.0018

TABLE 6.15: Node times for the last multifurcating tree two-dimensional nose profiles curves.

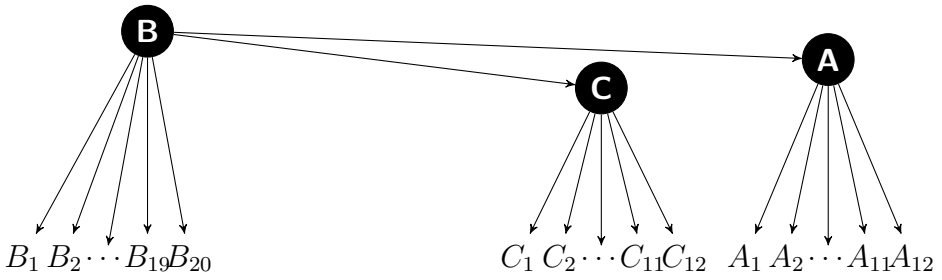


FIGURE 6.18: Last tree for nose profile curves.

Note that the estimates for the hyperparameters not related to time are stable across the different trees. Moreover, the distance between node B and all the B_i curves is stable at approximately 0.002, and the distance between the curves A_i and their ancestor A is always larger than between node C and the curves C_i. This is interpreted, in all the trees studied, as there being more variability amongst the British curves than within the other ethnic groups, which agrees with the topologies previously found that had the British evolving more separately. The Chinese group is the one with more similar nose profile curves. This can be seen in Figure 6.14. The final log-likelihood value is around 3 units smaller than in the original tree, but there are two hyperparameters less in the later tree. The BIC for the first model is -773.74 and -782.27, hence the simpler model with less hyperparameters is preferred. Furthermore, there are no estimates lying on the boundary so SEs should be reliable.

6.5 Discussion

The use of GPs models combined with the phylogenetic covariance function provide a new, powerful, tool for the study of shape combined with genealogical analysis. From simulations, it was shown the model performs well, capturing adequately the covariance structure in space and time, for both two- and three-dimensional curves. Particularly, the model performs well when only data at the leaves are available, which will be the case in most studies. This is shown with the study of the evolution of the mean nose shape.

When studying the structure of the tree for the mean nose curves for the three ethnic groups A - African, B - British and C - Chinese, the first thing observed is that the optimal topology links A and C with one common ancestor more recent than the common ancestor of the three groups. From the genetic study of ethnic groups, there is general agreement that the human lineage evolved in Africa, and then spread to southern Eurasia as *Homo erectus*. After the evolution of modern humans in Africa, a second expansion occurred out of Africa between 60000-80000 years ago that resulted in a global replacement [Macaulay et al., 2005]. Therefore, it might have been expected to find a topology that links B and C under one more recent ancestor, having A evolving on their own. However, these results are based on DNA analysis, and it could be that the morphology of the nose has evolved differently to adapt to different environmental conditions. Noses adapted to cold weather may function differently from those that evolved in hot and humid climates. People of African descent typically have shorter noses, with wider nostrils, whilst people of northern European descent typically have longer, thinner, noses. Researchers in Germany showed that individuals from cold, dry climates had higher and narrower nasal cavities than those from hot, humid climates [Noback et al., 2011]. This topology was found to be optimal for both sets of nose curves for the mean of each group, as well as when using all the curves available. The estimation of the optimal topology by the study of all possible topologies is a reliable method but only practical for trees with a reduced number of nodes. For larger trees the number of possible topologies increases exponentially, and different methods to optimise the topology should be studied, since the optimisation for each of them can become expensive.

The predictive distributions provide a powerful tool to estimate ancestral shapes. Spatial marginal predictions to interpolate the data at one known node can also

be done, but most interest lies in being able to reconstruct data which one could never directly obtain. In the second study, where all the curves at the leaves are studied, prediction could also be done for the common ancestor of each ethnic group. Given the small amount of data available, the predictions in this case are not expected to differ largely from the mean of the curves in each .

The results presented here are meant just as an illustration of what these models could accomplish. A bigger study with more subjects, that takes into account more ethnic and sub-ethnic groups would permit one to test the idea that nose shape is correlated with climate condition. The models could clearly also be applied to other facial curves. If data could be collected from various members of a family, it would be interesting to model the facial morphology within its members, apart from anything else, to study quantitatively the often-quoted observation that 'she has her grandfather's nose'.

Chapter 7

Discussion and further lines of investigation

This thesis presents and discusses a number of approaches for modelling the evolution of k-dimensional curves. The use of k-dimensional curves is motivated by its application to facial curves, so special attention is given to three-dimensional curves. Two scales of evolution were considered. First, time is modelled as a linear continuous variable, i.e., one curve that is gradually changing in a particular situation. The notion of a shape evolving in time is then extended to a phylogenetic setting, where branching points in the evolution can occur.

7.1 Facial curve estimation: limitations and further directions

Methods for the estimation of 4D facial curves were investigated. Many of the difficulties of curve estimation are due to the use of manual landmarks. Manually allocating the landmarks means there is always human error possible. The beginning and end points of the curves rely solely on the landmarks. The allocation of landmarks by using an estimate from the coordinates of the previous image in the temporal sequence is better than manually and independently allocating them at each time point, both in terms of the time to compute the landmarks and also effectiveness. However, there are a number of steps in the process where room for error is still present, such as the choice of neighbourhoods that are used to

determine the shape of a particular point: smaller neighbourhoods lead to more variability in estimation. Moreover, there are still some issues arising from the differences in the origin of the coordinate system of each image. A better approach might be to centre all facial point clouds to a common origin first, and then estimate the 4D curves.

When applied to the evolution of lip curves during the performance of an emotion, the algorithm for the estimation of 4D curves, proposed in Chapter 3, has to be slightly tuned depending on the emotion and a presight of the images is helpful to decide the distance out to which to look for the optimal landmarks, as well as the method to identify the path (plane-path versus principal curve). The use of 'plane-path' methods is well established for curve identification. The inclusion of principal curves as an option has provided a solution to the problem where the nature of the surface (i.e., open mouth) creates some difficulties in representing the lip curves.

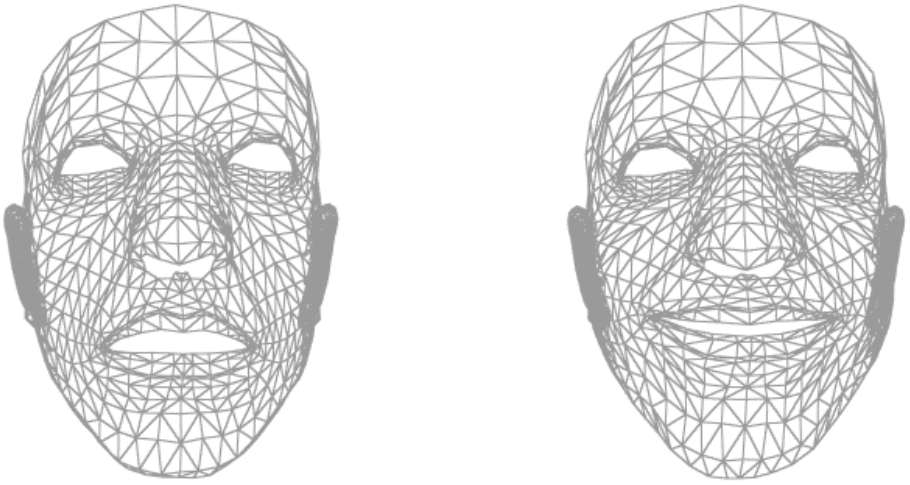


FIGURE 7.1: Conformed meshes from the ©*Di4D* system for a sequence of *happiness*: a resting position of the face (left) and from the middle of the sequence (right).

There are also limitations in the capture of the data itself. The 'orange peel effect' and the issues identified around the ear can sometimes distort the facial images and possibly the curves and landmarks on the surface as well. An enormously high curvature value in some location due to an issue in image capture can at times lead a curve to go very much off target. A new ©*Di4D* [Dimensional Imaging Ltd, 2017] system for high-resolution 4D facial motion capture has been released, using digital video cameras. The captured video sequence is post-processed using the

©*Di4D* dense passive stereo-photogrammetry software and then a fixed mesh can be tracked through the sequence using the ©*Di4D* optical flow-tracking software. Figure 7.1 shows these conformed meshes for two images in a sequence of *happiness*. These data arrived too late for this thesis, but the analysis of the lip curves extracted from these meshes is promising.

7.2 Discussion on models for lip curves

The use of B-Splines to model the curves and smooth the paths has shown good results and moved the analysis of variation of the curves to differences in the spline coefficients. These models led to a good representation of the average emotion shape, given the small number of replicates per emotion. PCA was used to study how the mean shape varies over space and time, identifying the directions in three-dimensional space where most variation is happening. The first PC was used to produce confidence intervals around the mean. Problems were faced especially for emotions with a small amplitude of change, when very noisy variation around the mean was observed. Improvements in the methods of curve estimation might help to fix these problems. Moreover, the number of replicates is small, and the length of image sequence varies. Due to the prolonged resting period of the lips at the beginning and the end of each of the sequences, a principled approach to cutting the extremes of the sequences was sought, however, unsuccessfully. It would be interesting to perform the analysis over more standardised (in terms of length) replicates, as well as incorporating data from different people. Approaches to comparing the 4D curves accross subject groups, for example, between females and males, would be of interest. PCA and the mean shape should also be studied using data derived from the ©*Di4D* system.

The extension to Gaussian process models opened a new line for the study of the emotions in terms of their correlation parameters. Several models with GPs were proposed for k-dimensional evolving (and non-evolving) curves. All the models interpolated the data well. Understanding three-dimensional curves as a set of three functions of the arc-length provides a powerful solution to capture the relationship between the coordinates. Separable covariance functions were chosen as a simple means of erecting a full covariance function in space, time and coordinate (future work may explore non-separable alternatives). These models offer a powerful tool for the analysis of evolving curves. Problems faced with the models were due to

the particularities of the lip curves. The lip curves are very smooth and therefore highly correlated spatially. Moreover, the sequences of pictures are taken in such short intervals that the curves are also highly correlated temporally. Different approaches were investigated to deal with the high correlation: spectral decomposition, reducing the number of space-points and adding noise to the model of the observations. For noise-free one-dimensional curves, spectral decomposition may be the best approach; it was shown to make optimisation viable. Unfortunately, for the 3D lip data, this approximation was not enough. The thinning of the sample points is practical in cases with high number of observations. The lip curves consist of only 24 space-points, and moreover, they have been identified with methods the limitations of which have been discussed in Section 7.1. Therefore, it seems more reasonable to opt for including an additive normal error to the model of observations. This not only solves problems with the ill-conditioned covariance matrix but, on top of that, it accommodates actual noise in the observed facial surface. The special adjustments needed to model the lip curves can be set as a reference for other applications which involve very smooth curves.

7.3 The shape of emotions

Differences between emotions were studied based on the set of estimated hyperparameters: the spatial and temporal length scales, the signal variance and the two correlation parameters between coordinates. Each emotion is drastically reduced to just a few numbers, all related to the covariance structure of the GP, so the question was whether or not these were enough to group the emotions. The main limitation, however, is not so much in the reduced number of hyperparameters but in the small number of replicates available. Two approaches were proposed for the three-dimensional lip curves: setting to zero the correlation parameters between coordinates or estimating them from a series of time-points and averaging them. Both approaches led to the same conclusions, as well as produced plausible predictions. However, given the nature of the lips, it is clear that the y and z coordinates are strongly correlated, and indeed important to characterise an emotion. Leaving out the correlation parameters κ_1 and κ_2 does not take full advantage of all the information contained in the data. PCA was used to understand which hyperparameters vary the most from one emotion to another, these being the temporal length-scale and the correlation between the y and z coordinates,

another indication that the correlations between coordinates should be kept in the model. Despite the reduced number of hyperparameters and replicates, the first two principal components explain roughly three quarters of the variability, which is perhaps more than one might have expected. For a better understanding of the differences between emotions, in terms of their correlation parameters, it would be necessary to increase the number of replicates, as well as to add more subjects to the study, to account for the variability across people. Only then could one robustly conclude whether a reduced number of hyperparameters can be used to group different sets of, in this case, lip curves. Models for classification could then be applied, so that, given a new sequence (and its estimated hyperparameters), the emotion it represents can be inferred.

7.4 Phylogenetic GP models and the evolution of nose shape: limitations and possibilities

The use of the phylogenetic covariance function [Jones and Moriarty, 2013] allows the fields of shape analysis and phylogenetics to be combined. The phylogenetic GP models are able to capture adequately the covariance structure in space and time. Particularly, the model performs well when only the data at the leaves of the tree are available, which will be the case in most studies. One of the main limitations with the approach shown in this thesis is the strategy for the choice of topology by an exhaustive search, i.e., by finding the maximum likelihood of the hyperparameters for each possible topology and choosing the one with the highest maximised-likelihood value. This approach is possible when there is a small number of possible topologies, but becomes impractical when the number of leaves increases. When one cannot find the best tree by examining all possible trees, heuristic search techniques are commonly used [Felsenstein, 2004]. The main idea is to take an initial estimate of the tree and make small rearrangements of its branches to reach 'neighbouring' trees. If any of these neighbours are better, they are accepted and one continues attempting more rearrangements, with the hope that there will be a point at which no small rearrangement can improve the tree structure. Such a tree is at a local optimum in the tree space, but there is no guarantee that it is a global optimum [Felsenstein, 1989]. A small illustration of this heuristic search was performed when studying the multifurcating tree of nose

profiles. When the optimal hyperparameters were on the boundary, a neighbouring tree was studied, with the arrangements motivated by the hyperparameter estimates for the previous tree. A natural continuation for this work would be to use these heuristic search techniques to optimise the tree topology for the models presented in this thesis, when more leaves are present.

Even though genetic drift has played a predominant role in human evolution, external physical traits such as facial shape and skin pigmentation have also probably been influenced by natural selection. How selection may have affected facial shape, a trait that is also quite variable between populations, has received less attention than other biological traits. Given the complexity of the human face, it was decided to study one particularly interesting and variable part of it: the nose.

The diversity of facial features across human populations and the evolutionary reasons for variation in nose shape across human populations have been subject to debate in recent years. An important function of the nose and nasal cavity is to condition inspired air before it reaches the lower respiratory tract. For this reason, it is thought the observed differences in nose shape among populations are not simply the result of genetic drift, but may be adaptations to climate. Zaidi et al. [2017] have recently studied the variation in nose shape using three-dimensional images to support the claim that local adaptation to climate may have had a role in the evolution of nose shape differences between human populations. They used linear distances and areas to characterize the shape of the nose, rather than anatomical curves. One of their findings was a positive selection for lighter skin pigmentation and narrower nostrils in higher latitude populations (i.e., Europeans). This is consistent with the results reported in Section 6.4. Climate may not have been the only factor in contributing to nose shape differences across populations. Zaidi et al. [2017] also show that temperature is only weakly correlated with nose width, especially when compared with the correlation between skin pigmentation and UVB. Models proposed in Chapter 6 pave the way for even more powerful methods of analysis through possible combinations of genetic and shape information. The small sample size in the case studies was a hindrance in trying to determine anything conclusively, especially in the case of the nose profile curves, where there is little variation. However, in future work, more data could easily be collected, leading to a greater level of conclusive material.

One should note that the parameter which governs the rate of change of shape, μ, has been assumed constant between groups. When modelling the mean curves for

each ethnic groups, a scaling parameter for the evolution rates was incorporated. That scaling parameter, however, was meant to capture the different evolution rates of the nose profile and the nasal bridge curves, not a difference between ethnic groups. A natural development of the study of the phylogenetic GP models would be to include different rates within the same tree, whilst, however, keeping in mind the identifiability issues that affect μ and the times of internal nodes of the tree (see Section 6.2.2).

7.5 Further lines of investigation

In all the GP models, for both time as a linear variable and branching along a tree, the estimation of the hyperparameters was performed by maximum likelihood, which is a widely-used approach in the literature of GPs. However, it would be very interesting to carry out a fully Bayesian analysis, when integration over the hyperparameters would be called for, in one way or another.

In this thesis, shape has been represented by a small number of three-dimensional curves. The next logical step would be to ask oneself what are other forms the data could take to capture more information about shape, while maintaining computational feasibility. In the case of the nose, for example, its shape was defined by only two curves, but there are more curves that could be calculated. Can models for a cloud of three-dimensional points rather than points on curves evolving in time be constructed? The last part of this thesis (Chapter 6) aimed to develop statistical methods by which shape information on organisms can be used to construct phylogenetic trees. Even more powerful methods of analysis could be developed through fusions of genetic and shape information. The results presented here leave an open door for this interdisciplinary field.

Appendix A

A.1 Conditional distribution of z given x and y

Inverse matrix:

$$\left[\begin{pmatrix} 1 & \kappa_1 \\ \kappa_1 & 1 \end{pmatrix} \otimes \mathbf{K}_s\right]^{-1} = \frac{1}{1-\kappa_1^2}\begin{pmatrix} 1 & -\kappa_1 \\ -\kappa_1 & 1 \end{pmatrix} \otimes \mathbf{K}_s = \frac{1}{1-\kappa_1^2}\begin{pmatrix} \mathbf{K}_s^{-1} & -\kappa_1\mathbf{K}_s^{-1} \\ -\kappa_1\mathbf{K}_s^{-1} & \mathbf{K}_s^{-1} \end{pmatrix}.$$

Mean:

$$0 + \begin{pmatrix} \kappa_1\mathbf{K}_s & \kappa_2\mathbf{K}_s \end{pmatrix}\begin{bmatrix} \mathbf{K}_s & \kappa_1\mathbf{K}_s \\ \kappa_1\mathbf{K}_s & \mathbf{K}_s \end{bmatrix}^{-1}\left(\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix} - \mathbf{0}\right)$$

$$= \begin{pmatrix} \kappa_1\mathbf{K}_s & \kappa_2\mathbf{K}_s \end{pmatrix}\frac{1}{1-\kappa_1^2}\begin{pmatrix} \mathbf{K}_s^{-1} & -\kappa_1\mathbf{K}_s^{-1} \\ -\kappa_1\mathbf{K}_s^{-1} & \mathbf{K}_s^{-1} \end{pmatrix}\begin{bmatrix} \mathbf{x} \\ \mathbf{y} \end{bmatrix}$$

$$= \frac{1}{1-\kappa_1^2}\left[(\kappa_1 - \kappa_1\kappa_2)\mathbf{x} + (\kappa_2 - \kappa_1^2)\mathbf{y}\right].$$

Covariance:

$$\mathbf{K}_s - \begin{pmatrix} \kappa_1\mathbf{K}_s & \kappa_2\mathbf{K}_s \end{pmatrix}\begin{bmatrix} \mathbf{K}_s & \kappa_1\mathbf{K}_s \\ \kappa_1\mathbf{K}_s & \mathbf{K}_s \end{bmatrix}^{-1}\begin{pmatrix} \kappa_1\mathbf{K}_s \\ \kappa_2\mathbf{K}_s \end{pmatrix}$$

$$= \mathbf{K}_s - \frac{1}{1-\kappa_1^2}\begin{pmatrix} \kappa_1\mathbf{K}_s & \kappa_2\mathbf{K}_s \end{pmatrix}\begin{pmatrix} \mathbf{K}_s^{-1} & -\kappa_1\mathbf{K}_s^{-1} \\ -\kappa_1\mathbf{K}_s^{-1} & \mathbf{K}_s^{-1} \end{pmatrix}\begin{pmatrix} \kappa_1\mathbf{K}_s \\ \kappa_2\mathbf{K}_s \end{pmatrix}$$

$$= \left(1 - \frac{\kappa_1^2 + \kappa_2^2 - 2\kappa_1^2\kappa_2}{1-\kappa_1^2}\right)\mathbf{K}_s.$$

A.2 Prediction for a 3D curve

A.2.1 Joint predictive distribution

$$\textbf{Mean: } \mathbf{0} + (\mathbf{K}_c \otimes \mathbf{K}_{s^*s})(\mathbf{K}_c \otimes \mathbf{K}_s)^{-1} \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \\ \mathbf{z} \end{pmatrix}$$

Here, $(\mathbf{K}_c \otimes \mathbf{K}_s)^{-1}$

$$= \mathbf{K}_c^{-1} \otimes \mathbf{K}_s^{-1}$$

$$= \frac{1}{1 + 2\kappa_1^2\kappa_2 - 2\kappa_1^2 - \kappa_2^2} \begin{pmatrix} 1-\kappa_2^2 & \kappa_1\kappa_2 - \kappa_1 & \kappa_1\kappa_2 - \kappa_1 \\ \kappa_1\kappa_2 - \kappa_1 & 1-\kappa_2^2 & \kappa_1\kappa_2 - \kappa_1 \\ \kappa_1\kappa_2 - \kappa_1 & \kappa_1\kappa_2 - \kappa_1 & 1-\kappa_2^2 \end{pmatrix} \otimes \mathbf{K}_s^{-1}$$

$$= \frac{1}{1 + 2\kappa_1^2\kappa_2 - 2\kappa_1^2 - \kappa_2^2} \begin{pmatrix} (1-\kappa_2^2)\mathbf{K}_s^{-1} & (\kappa_1\kappa_2 - \kappa_1)\mathbf{K}_s^{-1} & (\kappa_1\kappa_2 - \kappa_1)\mathbf{K}_s^{-1} \\ (\kappa_1\kappa_2 - \kappa_1)\mathbf{K}_s^{-1} & (1-\kappa_2^2)\mathbf{K}_s^{-1} & (\kappa_1\kappa_2 - \kappa_1)\mathbf{K}_s^{-1} \\ (\kappa_1\kappa_2 - \kappa_1)\mathbf{K}_s^{-1} & (\kappa_1\kappa_2 - \kappa_1)\mathbf{K}_s^{-1} & (1-\kappa_2^2)\mathbf{K}_s^{-1} \end{pmatrix},$$

and so, $(\mathbf{K}_c \otimes \mathbf{K}_{s^*s})(\mathbf{K}_c \otimes \mathbf{K}_s)^{-1}$

$$= \begin{pmatrix} \mathbf{K}_{s^*s} & \kappa_1\mathbf{K}_{s^*s} & \kappa_1\mathbf{K}_{s^*s} \\ \kappa_1\mathbf{K}_{s^*s} & \mathbf{K}_{s^*s} & \kappa_2\mathbf{K}_{s^*s} \\ \kappa_1\mathbf{K}_{s^*s} & \kappa_2\mathbf{K}_{s^*s} & \mathbf{K}_{s^*s} \end{pmatrix} \frac{1}{C}$$

$$\times \begin{pmatrix} (1-\kappa_2^2)\mathbf{K}_s^{-1} & (\kappa_1\kappa_2 - \kappa_1)\mathbf{K}_s^{-1} & (\kappa_1\kappa_2 - \kappa_1)\mathbf{K}_s^{-1} \\ (\kappa_1\kappa_2 - \kappa_1)\mathbf{K}_s^{-1} & (1-\kappa_2^2)\mathbf{K}_s^{-1} & (\kappa_1\kappa_2 - \kappa_1)\mathbf{K}_s^{-1} \\ (\kappa_1\kappa_2 - \kappa_1)\mathbf{K}_s^{-1} & (\kappa_1\kappa_2 - \kappa_1)\mathbf{K}_s^{-1} & (1-\kappa_2^2)\mathbf{K}_s^{-1} \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{K}_{s^*s}\mathbf{K}_s^{-1} & 0 & 0 \\ 0 & \mathbf{K}_{s^*s}\mathbf{K}_s^{-1} & 0 \\ 0 & 0 & \mathbf{K}_{s^*s}\mathbf{K}_s^{-1} \end{pmatrix},$$

with $C = 1 + 2\kappa_1^2\kappa_2 - 2\kappa_1^2 - \kappa_2^2$, and so

$$(\mathbf{K}_c \otimes \mathbf{K}_{s^*s})(\mathbf{K}_c \otimes \mathbf{K}_s)^{-1} \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \\ \mathbf{z} \end{pmatrix} = \begin{pmatrix} \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{x} \\ \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{y} \\ \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{z} \end{pmatrix}.$$

Covariance: $(\mathbf{K}_c \otimes \mathbf{K}_{s^*s}) - (\mathbf{K}_c \otimes \mathbf{K}_{s^*})(\mathbf{K}_c \otimes \mathbf{K}_s)^{-1}(\mathbf{K}_c \otimes \mathbf{K}_{ss^*})$.
Here, $(\mathbf{K}_c \otimes \mathbf{K}_{s^*})(\mathbf{K}_c \otimes \mathbf{K}_s)^{-1}(\mathbf{K}_c \otimes \mathbf{K}_{ss^*})$

$$= \begin{pmatrix} \mathbf{K}_{s^*s}\mathbf{K}_s^{-1} & 0 & 0 \\ 0 & \mathbf{K}_{s^*s}\mathbf{K}_s^{-1} & 0 \\ 0 & 0 & \mathbf{K}_{s^*s}\mathbf{K}_s^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{K}_{ss^*} & \kappa_1\mathbf{K}_{ss^*} & \kappa_1\mathbf{K}_{ss^*} \\ \kappa_1\mathbf{K}_{ss^*} & \mathbf{K}_{ss^*} & \kappa_2\mathbf{K}_{ss^*} \\ \kappa_1\mathbf{K}_{ss^*} & \kappa_2\mathbf{K}_{ss^*} & \mathbf{K}_{ss^*} \end{pmatrix}$$

$$= \begin{pmatrix} \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*} & \kappa_1\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*} & \kappa_1\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*} \\ \kappa_1\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*} & \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*} & \kappa_2\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*} \\ \kappa_1\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*} & \kappa_2\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*} & \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*} \end{pmatrix}.$$

and so, $(\mathbf{K}_c \otimes \mathbf{K}_{s^*s}) - (\mathbf{K}_c \otimes \mathbf{K}_{s^*})(\mathbf{K}_c \otimes \mathbf{K}_s)^{-1}(\mathbf{K}_c \otimes \mathbf{K}_{ss^*})$

$$= \begin{pmatrix} \mathbf{K}_{s^*} - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*} & \kappa_1(\mathbf{K}_{s^*} - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}) & \kappa_1(\mathbf{K}_{s^*} - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}) \\ \kappa_1(\mathbf{K}_{s^*} - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}) & \mathbf{K}_{s^*} - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*} & \kappa_2(\mathbf{K}_{s^*} - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}) \\ \kappa_1(\mathbf{K}_{s^*} - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}) & \kappa_2(\mathbf{K}_{s^*} - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}) & \mathbf{K}_{s^*} - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*} \end{pmatrix}$$

$$= \begin{pmatrix} 1 & \kappa_1 & \kappa_1 \\ \kappa_1 & 1 & \kappa_2 \\ \kappa_1 & \kappa_2 & 1 \end{pmatrix} \otimes \left(\mathbf{K}_{s^*} - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}\right).$$

A.2.2 Conditional predictive distributions

From:

$$\begin{matrix} \mathbf{x}^* \\ \mathbf{y}^* \\ \mathbf{z}^* \end{matrix} \;\middle|\; \begin{matrix} \mathbf{x} \\ \mathbf{y} \\ \mathbf{z} \end{matrix} \sim N\left(\begin{bmatrix} \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{x} \\ \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{y} \\ \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{z} \end{bmatrix}, \begin{bmatrix} \begin{pmatrix} 1 & \kappa_1 & \kappa_1 \\ \kappa_1 & 1 & \kappa_2 \\ \kappa_1 & \kappa_2 & 1 \end{pmatrix} \otimes \left(\mathbf{K}_{s^*} - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}\right) \end{bmatrix} \right).$$

- For $\mathbf{x}^* \mid \mathbf{x}$ take the element of the mean vector corresponding to $\mathbf{x}^*$ and the first diagonal element of the covariance.
- For $\mathbf{y}^* \mid \mathbf{x}^*$, the joint distribution of $\mathbf{y}^*$ and $\mathbf{x}^*$ given the three coordinates is needed.

$$\begin{matrix}\mathbf{y}^* \\ \mathbf{x}^*\end{matrix}\left|\begin{matrix}\mathbf{x} \\ \mathbf{y} \\ \mathbf{z}\end{matrix}\right. \sim N\left(\begin{pmatrix}\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{y} \\ \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{x}\end{pmatrix}, \left[\begin{pmatrix}1 & \kappa_1 \\ \kappa_1 & 1\end{pmatrix} \otimes \left(\mathbf{K}_{s^*} - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}\right)\right]\right).$$

By the conditional distribution properties:

Mean:

$$\begin{aligned}&\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{y}+ \\ &\kappa_1(\mathbf{K}_{s^*} - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*})\left[\mathbf{K}_{s^*} - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}\right]^{-1}(\mathbf{x}^* - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{x}) \\ &= (\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{y}) + \kappa_1(\mathbf{x}^* - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{x}).\end{aligned}$$

Covariance:

$$\begin{aligned}&(\mathbf{K}_{s^*} - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*})- \\ &\kappa_1(\mathbf{K}_{s^*} - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*})\left[\mathbf{K}_{s^*} - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}\right]^{-1}\kappa_1(\mathbf{K}_{s^*} - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*})^{\mathrm{T}} \\ &= (\mathbf{K}_{s^*} - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}) - \kappa_1^2(\mathbf{K}_{s^*} - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*})^{\mathrm{T}} \\ &= (1-\kappa_1^2)(\mathbf{K}_{s^*} - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*})^{\mathrm{T}}.\end{aligned}$$

Given that $(\mathrm{ABC})^{\mathrm{T}} = \mathrm{C}^{\mathrm{T}}\mathrm{B}^{\mathrm{T}}\mathrm{A}^{\mathrm{T}}$ and $(\mathrm{D}-\mathrm{E})^{\mathrm{T}} = \mathrm{D}^{\mathrm{T}} - \mathrm{E}^{\mathrm{T}}$, then:

$$(\mathbf{K}_{s^*} - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*})^{\mathrm{T}} = \mathbf{K}_{s^*}^{\mathrm{T}} - \mathbf{K}_{ss^*}^{\mathrm{T}}(\mathbf{K}_s^{-1})^{\mathrm{T}}\mathbf{K}_{s^*s}^{\mathrm{T}} = \mathbf{K}_{s^*} - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}.$$

This result shows that $\mathbf{y}^* \mid \mathbf{x}^*$ implies $\mathbf{y}^* \mid \mathbf{x}^*, \mathbf{x}, \mathbf{y}$.

-For $\mathbf{z}^* \mid \mathbf{x}^*, \mathbf{y}^*$, the joint distribution of $\mathbf{z}^*$, $\mathbf{y}^*$ and $\mathbf{x}^*$ given the three coordinates is needed.

$$\begin{matrix}\mathbf{z}^* \\ \mathbf{x}^* \\ \mathbf{y}^*\end{matrix}\left|\begin{matrix}\mathbf{x} \\ \mathbf{y} \\ \mathbf{z}\end{matrix}\right. \sim N\left(\begin{bmatrix}\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{z} \\ \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{x} \\ \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{y}\end{bmatrix}, \left[\begin{pmatrix}1 & \kappa_1 & \kappa_2 \\ \kappa_1 & 1 & \kappa_1 \\ \kappa_2 & \kappa_1 & 1\end{pmatrix} \otimes \left(\mathbf{K}_{s^*} - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}\right)\right]\right).$$

Denote $\mathbf{B} = \mathbf{K}_{s^*} - \mathbf{K}_{s^*s}Ks^{-1}\mathbf{K}_{ss^*}$.

Mean:

$$\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{z} + \left((\kappa_1\ \kappa_2) \otimes \mathbf{B}\right)\left[\begin{pmatrix}1 & \kappa_1 \\ \kappa_1 & 1\end{pmatrix} \otimes \mathbf{B}\right]^{-1}\left(\begin{bmatrix}\mathbf{x}^* \\ \mathbf{y}^*\end{bmatrix} - \begin{pmatrix}\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{x} \\ \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{y}\end{pmatrix}\right).$$

Here, $\left[\begin{pmatrix}1 & \kappa_1 \\ \kappa_1 & 1\end{pmatrix} \otimes \mathbf{B}\right]^{-1} = \frac{1}{1-\kappa_1^2}\begin{pmatrix}\mathbf{B}^{-1} & -\kappa_1\mathbf{B}^{-1} \\ -\kappa_1\mathbf{B}^{-1} & \mathbf{B}^{-1}\end{pmatrix},$

so that $\big((\kappa_1\ \kappa_2)\otimes\mathbf{B}\big)\left[\begin{pmatrix}1 & \kappa_1\\ \kappa_1 & 1\end{pmatrix}\mathbf{B}\right]^{-1}$

$$= \begin{pmatrix}\kappa_1\mathbf{B} & \kappa_2\mathbf{B}\end{pmatrix}\frac{1}{1-\kappa_1^2}\begin{pmatrix}\mathbf{B}^{-1} & -\kappa_1\mathbf{B}^{-1}\\ -\kappa_1\mathbf{B}^{-1} & \mathbf{B}^{-1}\end{pmatrix},$$

$$= \frac{1}{1-\kappa_1^2}\begin{pmatrix}(\kappa_1-\kappa_1\kappa_2)\mathbf{I} & (\kappa_2-\kappa_1^2)\mathbf{I}\end{pmatrix}.$$

so that $\big((\kappa_1\ \kappa_2)\otimes\mathbf{B}\big)\left[\begin{pmatrix}1 & \kappa_1\\ \kappa_1 & 1\end{pmatrix}\mathbf{B}\right]^{-1}\left(\begin{bmatrix}\mathbf{x}^*\\ \mathbf{y}^*\end{bmatrix}-\begin{pmatrix}\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{x}\\ \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{y}\end{pmatrix}\right)$

$$= \frac{1}{1-\kappa_1^2}\begin{pmatrix}(\kappa_1-\kappa_1\kappa_2)\mathbf{I} & (\kappa_2-\kappa_1^2)\mathbf{I}\end{pmatrix}\begin{pmatrix}\mathbf{x}^*-\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{x}\\ \mathbf{y}^*-\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{y}\end{pmatrix}$$

$$= \frac{1}{1-\kappa_1^2}\left[(1-\kappa_1^2)(\mathbf{x}^*-\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{x})+(\kappa_2-\kappa_1^2)(\mathbf{y}^*-\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{y})\right].$$

Therefore the mean is :

$$\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{z}+\frac{1}{1-\kappa_1^2}\left[(\kappa_1-\kappa_1\kappa_2)(\mathbf{x}^*-\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{x})\right.$$
$$\left.+(\kappa_2-\kappa_1^2)\left(\mathbf{y}^*-\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{y}\right)\right].$$

Covariance:

$$\mathbf{B}-\big((\kappa_1\ \kappa_2)\otimes\mathbf{B}\big)\left[\begin{pmatrix}1 & \kappa_1\\ \kappa_1 & 1\end{pmatrix}\otimes\mathbf{B}\right]^{-1}\big((\kappa_1\ \kappa_2)\otimes\mathbf{B}\big)^{\mathrm{T}}$$

$$= \mathbf{B}-\frac{1}{1-\kappa_1^2}\begin{pmatrix}(\kappa_1-\kappa_1\kappa_2)\mathbf{I} & (\kappa_2-\kappa_1^2)\mathbf{I}\end{pmatrix}\begin{pmatrix}\kappa_1\mathbf{B}\\ \kappa_2\mathbf{B}\end{pmatrix}$$

$$= \mathbf{B}-\frac{1}{1-\kappa_1^2}(\kappa_1^2+\kappa_2^2-2\kappa_1^2\kappa_2)\mathbf{B}$$

$$= \left(1-\frac{\kappa_1^2+\kappa_2^2-2\kappa_1^2\kappa_2}{1-\kappa_1^2}\right)\left(\mathbf{K}_{s^*}-\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}\right).$$

Thus $\mathbf{z}^* \mid \mathbf{x}^*, \mathbf{y}^*$ implies $\mathbf{z}^* \mid \mathbf{x}^*, \mathbf{x}, \mathbf{y}^*, \mathbf{y}, \mathbf{z}$.

Appendix B

B.1 Predictive distributions for the evolution of 1D curves

Data:

- $\mathbf{s} = (s_1, ..., s_n)$
- $\mathbf{y} = \big(y(s_1) \cdots y(s_n)\big)^{\mathrm{T}}$
- At time i, $\mathbf{y}(t_i) = \big(y(s_1, t_i) \cdots y(s_n, t_i)\big)^{\mathrm{T}}$, in this section denoted as $\mathbf{y}_i$ for abbreviation.
- All times $1, ..., T$: $\mathbf{y} = (\mathbf{y}_1 \cdots \mathbf{y}_T)^{\mathrm{T}}$
- $\mathbf{y}_i \in \mathbb{R}^n$

Predictions:

- $\mathbf{s}^* = (s_1^*, ..., s_m^*)$
- $\mathbf{y}_q \in \mathbb{R}^m$
- If:
  - $q \in [1, T]$ Marginal prediction for existing time points or interpolation between existing timepoints.
  - $q < 1$ Retrodiction
  - $q > 1$ Prediction

Then:

$$\begin{bmatrix} \mathbf{y}_q^* \\ \mathbf{y} \end{bmatrix} \sim N_{m+nT}(\mathbf{0}, \mathbf{\Sigma}),$$

where the covariance matrix has the form:

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{K}_{s^*} & \mathbf{L} \otimes \mathbf{K}_{s^*s} \\ (\mathbf{L} \otimes \mathbf{K}_{s^*s})^{\mathrm{T}} & \mathbf{K}_t \otimes \mathbf{K}_s \end{bmatrix},$$

where

$$\mathbf{L} = \begin{bmatrix} \exp\left(\frac{-\mid q-1 \mid}{\mu}\right) & \exp\left(\frac{-\mid q-2 \mid}{\mu}\right) & \cdots & \exp\left(\frac{-\mid q-T \mid}{\mu}\right) \end{bmatrix}$$

and

$$\mathbf{K}_t = \begin{bmatrix} 1 & \kappa & \kappa^2 & \kappa^3 & \cdots \\ \kappa & 1 & \kappa & \kappa^2 & \cdots \\ \kappa^2 & \kappa & 1 & \kappa & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

with $\kappa = \exp(-1/\mu)$. Then

$$\mathbf{y}_q^* \mid \mathbf{y} \sim \left([\mathbf{L} \otimes \mathbf{K}_{s^*s}][\mathbf{K}_t \otimes \mathbf{K}_s]^{-1}\mathbf{y}, \mathbf{K}_{s^*} - [\mathbf{L} \otimes \mathbf{K}_{s^*s}][\mathbf{K}_t \otimes \mathbf{K}_s]^{-1}[\mathbf{L} \otimes \mathbf{K}_{s^*s}]^{\mathrm{T}}\right).$$

It can be shown that:

$$[\mathbf{K}_t \otimes \mathbf{K}_s]^{-1} = \mathbf{K}_t^{-1} \otimes \mathbf{K}_s^{-1} = \frac{1}{1-\kappa^2} \begin{bmatrix} 1 & -\kappa & 0 & \cdots & \cdots & 0 \\ -\kappa & 1+\kappa^2 & -\kappa & 0 & \cdots & \vdots \\ 0 & -\kappa & 1+\kappa^2 & -\kappa & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & -\kappa \\ 0 & \vdots & \vdots & \vdots & -\kappa & 1 \end{bmatrix} \otimes \mathbf{K}_s^{-1}.$$

B.1.1 Marginal prediction

Let $q \in \{1, ..., T\}$, then:

$$[\mathbf{L} \otimes \mathbf{K}_{s^*s}] = \begin{pmatrix} \kappa^{q-1} & \kappa^{q-2} & \cdots & \kappa & 1 & \kappa & \cdots & \kappa^{T-q} \end{pmatrix} \otimes \mathbf{K}_{s^*s},$$

with the '1' being at position q, in the row vector.

Using properties of the Kronecker product:

$$[\mathbf{L} \otimes \mathbf{K}_{s^*s}][\mathbf{K}_t \otimes \mathbf{K}_s]^{-1} = [\mathbf{L}\mathbf{K}_t^{-1}] \otimes [\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}].$$

$$\mathbf{L}\mathbf{K}_t^{-1} = \frac{1}{1-\kappa^2}\begin{pmatrix}\kappa^{q-1} & \kappa^{q-2}\cdots & \kappa & 1 & \kappa & \cdots\kappa^{T-q}\end{pmatrix} \times \begin{bmatrix} 1 & -\kappa & 0 & \cdots & \cdots & 0 \\ -\kappa & 1+\kappa^2 & -\kappa & 0 & \cdots & \vdots \\ 0 & -\kappa & 1+\kappa^2 & -\kappa & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & -\kappa \\ 0 & \vdots & \vdots & \vdots & -\kappa & 1 \end{bmatrix}$$

$$= \frac{1}{1-\kappa^2}\cdot\begin{pmatrix}0 & \cdots & 0 & (1-\kappa^2) & 0 & \cdots & 0\end{pmatrix} = \begin{pmatrix}0 & \cdots & 0 & 1 & 0 & \cdots & 0\end{pmatrix},$$

again, at position q.

Therefore:

$$([\mathbf{L}\mathbf{K}_t^{-1}] \otimes [\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}])\mathbf{y} = \begin{pmatrix}0 & \cdots & 0 & (\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}) & 0 & \cdots & 0\end{pmatrix}\begin{pmatrix}\mathbf{y}_1 \\ \vdots \\ \mathbf{y}_q \\ \vdots \\ \mathbf{y}_T\end{pmatrix}$$

$$= \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{y}_q.$$

Moreover:

$$\begin{pmatrix}0 & \cdots & 0 & (\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}) & 0 & \cdots & 0\end{pmatrix}(\mathbf{L} \otimes \mathbf{K}_{ss^*}) = \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}.$$

Hence:

$$\mathbf{y}_q^* \mid \mathbf{y} \sim N_n\left(\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{y}_q, \mathbf{K}_{s^*} - \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}\right).$$

Thus, prediction for space points at time $q \in \{1, ..., T\}$ only depends on the data at time q.

B.1.2 Prediction for future time points

Let $q > T$, then:

$$\mathbf{L} \otimes \mathbf{K}_{s^*s} = \begin{pmatrix} \kappa^{q-1} & \kappa^{q-2} & \cdots & \kappa^{q-T} \end{pmatrix} \otimes \mathbf{K}_{s^*s}.$$

This time:

$$\mathbf{L}\mathbf{K}_t^{-1} = \frac{1}{1-\kappa^2} \begin{pmatrix} \kappa^{q-1} & \kappa^{q-2} & \cdots & \kappa^{q-T} \end{pmatrix} \times \begin{bmatrix} 1 & -\kappa & 0 & \cdots & \cdots & 0 \\ -\kappa & 1+\kappa^2 & -\kappa & 0 & \cdots & \vdots \\ 0 & -\kappa & 1+\kappa^2 & -\kappa & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & -\kappa \\ 0 & \vdots & \vdots & \vdots & -\kappa & 1 \end{bmatrix}$$

$$= \frac{1}{1-\kappa^2} \cdot \begin{pmatrix} 0 & \cdots & 0 & (1-\kappa^2)\kappa^{q-T} \end{pmatrix} = \begin{pmatrix} 0 & \cdots & 0 & \kappa^{q-T} \end{pmatrix}.$$

Also:

$$\begin{aligned} ([\mathbf{L}\mathbf{K}_t^{-1}] \otimes [\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}])\mathbf{y} &= \begin{pmatrix} 0 & \cdots & 0 & \kappa^{q-T}(\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}) \end{pmatrix} \begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_T \end{pmatrix} \\ &= \kappa^{q-T}(\mathbf{K}_{s^*s}\mathbf{K}_s^{-1})\mathbf{y}_T, \end{aligned}$$

and

$$\begin{pmatrix} 0 & \cdots & 0 & \kappa^{q-T}(\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}) \end{pmatrix} \begin{pmatrix} (\kappa^{q-1} & \kappa^{q-2} & \cdots & \kappa^{q-T}) \otimes \mathbf{K}_{ss^*} \end{pmatrix} = (\kappa^{q-T})^2 \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}.$$

Hence:

$$\mathbf{y}_q^* \mid \mathbf{y} \sim N_n \left(\kappa^{q-T}(\mathbf{K}_{s^*s}\mathbf{K}_s^{-1})\mathbf{y}_T, (\kappa^{q-T})^2 \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*} \right).$$

Thus, extrapolation forward in time only involves using the data from the last time point.

B.1.3 Retrodiction for previous time points

Let $q < 1$, then:

$$\mathbf{L} \otimes \mathbf{K}_{s^*s} = \begin{pmatrix} \kappa^{1-q} & \kappa^{2-q} & \cdots & \kappa^{T-q} \end{pmatrix} \otimes \mathbf{K}_{s^*s}.$$

This time:

$$\mathbf{L}\mathbf{K}_t^{-1} = \frac{1}{1-\kappa^2} \begin{pmatrix} \kappa^{1-q} & \kappa^{2-q} & \cdots & \kappa^{T-q} \end{pmatrix} \begin{bmatrix} 1 & -\kappa & 0 & \cdots & \cdots & 0 \\ -\kappa & 1+\kappa^2 & -\kappa & 0 & \cdots & \vdots \\ 0 & -\kappa & 1+\kappa^2 & -\kappa & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & -\kappa \\ 0 & \vdots & \vdots & \vdots & -\kappa & 1 \end{bmatrix}$$

$$= \frac{1}{1-\kappa^2} \cdot \begin{pmatrix} (1-\kappa^2)\kappa^{1-q} & 0 & \cdots & 0 \end{pmatrix} = \begin{pmatrix} \kappa^{1-q} & 0 & \cdots & 0 \end{pmatrix}.$$

Then:

$$\begin{aligned} ([\mathbf{L}\mathbf{K}_t^{-1}] \otimes [\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}])\mathbf{y} &= \begin{pmatrix} \kappa^{1-q}(\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}) & 0 & \cdots \end{pmatrix} \begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_T \end{pmatrix} \\ &= \kappa^{1-q}(\mathbf{K}_{s^*s}\mathbf{K}_s^{-1})\mathbf{y}_1, \end{aligned}$$

and

$$\begin{pmatrix} \kappa^{1-q}(\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}) & 0 & \cdots \end{pmatrix} \begin{pmatrix} (\kappa^{1-q} & \kappa^{2-q} & \cdots & \kappa^{T-q}) \otimes \mathbf{K}_{ss^*} \end{pmatrix}$$
$$= (\kappa^{1-q})^2 \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}.$$

Hence:

$$\mathbf{y}_q^* \mid \mathbf{y} \sim N_n \left(\kappa^{1-q}(\mathbf{K}_{s^*s}\mathbf{K}_s^{-1})\mathbf{y}_1, \kappa^{1-q})^2 \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*} \right).$$

Thus, extrapolation backward in time only involves using the data from the first time point.

B.1.4 Interpolation between time points

Let $t < q < t+1$, where $t \in [1, T-1]$, then:

$$\mathbf{L} \otimes \mathbf{K}_{s^*s} = \begin{pmatrix} \kappa^{q-1} & \kappa^{q-2} & \cdots & \kappa^{q-T} & \kappa^{t+1-q} & \kappa^{t+2-q} & \cdots & \kappa^{t-q} \end{pmatrix} \otimes \mathbf{K}_{s^*s}.$$

In this case:

$$\bullet\ \mathbf{L}\mathbf{K}_t^{-1} = \frac{1}{1-\kappa^2} \begin{pmatrix} \kappa^{q-1} & \cdots & \kappa^{q-T} & \kappa^{t+1-q} & \kappa^{t+2-q} & \cdots & \kappa^{t-q} \end{pmatrix}$$

$$\times \begin{bmatrix} 1 & -\kappa & 0 & \cdots & \cdots & 0 \\ -\kappa & 1+\kappa^2 & -\kappa & 0 & \cdots & \vdots \\ 0 & -\kappa & 1+\kappa^2 & -\kappa & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & -\kappa \\ 0 & \vdots & \vdots & \vdots & -\kappa & 1 \end{bmatrix}$$

$$= \frac{1}{1-\kappa^2} \cdot \begin{pmatrix} 0 & \cdots & 0 & (\kappa^{q-t} - \kappa^{t-q+2}) & (\kappa(\kappa^{t-q} - \kappa^{q-t})) & 0 & \cdots & 0 \end{pmatrix},$$

where the non-zero entries are at positions t and $t+1$.

Therefore:

$$([\mathbf{L}\mathbf{K}_t^{-1}] \otimes [\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}])\mathbf{y}$$

$$= \left[\frac{1}{1-\kappa^2} \begin{pmatrix} 0 \cdots & (\kappa^{q-t} - \kappa^{t-q+2}) & (\kappa(\kappa^{t-q} - \kappa^{q-t})) & 0 \cdots \end{pmatrix} \otimes (\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}) \right] \begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_t \\ \mathbf{y}_{t+1} \\ \vdots \\ \mathbf{y}_T \end{pmatrix}$$

$$= \frac{1}{1-\kappa^2} (\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}) \left[(\kappa^{q-t} - \kappa^{t-q+2})\mathbf{y}_t + (\kappa(\kappa^{t-q} - \kappa^{q-t}))\mathbf{y}_{t+1} \right].$$

Moreover:

$$\left[\begin{pmatrix}0 & \cdots & 0 & (\kappa^{q-t}-\kappa^{t-q+2}) & (\kappa(\kappa^{t-q}-\kappa^{q-t})) & 0 & \cdots & 0\end{pmatrix}\frac{1}{1-\kappa^2}\otimes(\mathbf{K}_{s^*s}\mathbf{K}_s^{-1})\right]$$
$$\times\left[\begin{pmatrix}(\kappa^{q-1} & \cdots & \kappa^{q-T} & \kappa^{t+1-q} & \cdots & \kappa^{t-q})\otimes\mathbf{K}_{ss^*}\end{pmatrix}\right]$$
$$=\mathbf{K}_{s^*}-\frac{\kappa^{2(q-t)}+\kappa^{2(t-q)+2}-2\kappa^2}{1-\kappa^2}(\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*})$$

Hence:

$$\mathbf{y}_q^* \mid \mathbf{y}\sim N_n\left(\frac{1}{1-\kappa^2}(\mathbf{K}_{s^*s}\mathbf{K}_s^{-1})\left[(\kappa^{q-t}-\kappa^{t-q+2})\mathbf{y}_t+(\kappa(\kappa^{t-q}-\kappa^{q-t}))\mathbf{y}_{t+1}\right],\right.$$
$$\left.\mathbf{K}_{s^*}-\frac{\kappa^{2(q-t)}+\kappa^{2(t-q)+2}-2\kappa^2}{1-\kappa^2}(\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*})\right).$$

As expected, interpolation between two time points depends on the data from those two times.

B.2 Conditional distributions of evolving 3D curves

B.2.1 Evolution step

For $\mathbf{W}(t)\mid\mathbf{W}(t-1)$, using:

$$\begin{bmatrix}\mathbf{W}(t)\\ \mathbf{W}(t-1)\end{bmatrix}\sim N_{6n}\left(\mathbf{0},\begin{bmatrix}\mathbf{K}_c\otimes\mathbf{K}_s & \kappa(\mathbf{K}_c\otimes\mathbf{K}_s)\\ \kappa(\mathbf{K}_c\otimes\mathbf{K}_s) & \mathbf{K}_c\otimes\mathbf{K}_s\end{bmatrix}\right),$$

then the mean equals

$$\mathbf{0}+\kappa(\mathbf{K}_c\otimes\mathbf{K}_s)\left[\mathbf{K}_c\otimes\mathbf{K}_s\right]^{-1}(\mathbf{W}(t)-0)=\kappa\mathbf{W}(t-1),$$

and the covariance equals

$$(\mathbf{K}_c\otimes\mathbf{K}_s)-\kappa(\mathbf{K}_c\otimes\mathbf{K}_s)\left[\mathbf{K}_c\otimes\mathbf{K}_s\right]^{-1}\kappa(\mathbf{K}_c\otimes\mathbf{K}_s)=(1-\kappa^2)(\mathbf{K}_c\otimes\mathbf{K}_s).$$

Therefore:

$$\mathbf{W}(t) \mid \mathbf{W}(t-1) \sim N_{3n}\left(\kappa \begin{bmatrix}\mathbf{x}(t-1)\\ \mathbf{y}(t-1)\\ \mathbf{z}(t-1)\end{bmatrix}, (1-\kappa^2)\begin{bmatrix}\mathbf{K}_s & \kappa_1\mathbf{K}_s & \kappa_1\mathbf{K}_s\\ \kappa_1\mathbf{K}_s & \mathbf{K}_s & \kappa_2\mathbf{K}_s\\ \kappa_1\mathbf{K}_s & \kappa_2\mathbf{K}_s & \mathbf{K}_s\end{bmatrix}\right).$$

B.2.2 Conditional distributions between coordinates

$\mathbf{x}(t) \mid \mathbf{W}(t)$ can be extracted from the previous results.

For $\mathbf{y}(t) \mid \mathbf{x}(t), \mathbf{W}(t-1)$, we use:

$$\begin{bmatrix}\mathbf{y}(t)\\ \mathbf{x}(t)\end{bmatrix}\Bigg| \mathbf{W}(t-1) \sim N_{2n}\left(\kappa \begin{bmatrix}\mathbf{y}(t-1)\\ \mathbf{x}(t-1)\end{bmatrix}, (1-\kappa^2)\begin{bmatrix}\mathbf{K}_s & \kappa_1\mathbf{K}_s\\ \kappa_1\mathbf{K}_s & \mathbf{K}_s\end{bmatrix}\right).$$

The mean equals

$$\kappa\mathbf{y}(t-1) + (1-\kappa^2)\kappa_1\mathbf{K}_s\left[(1-\kappa^2)\mathbf{K}_s\right]^{-1}(\mathbf{x}(t) - \kappa\mathbf{x}(t-1)) \\ = \kappa\mathbf{y}(t-1) + \kappa_1(\mathbf{x}(t) - \kappa(t-1)),$$

and the covariance equals

$$(1-\kappa^2)\mathbf{K}_s - (1-\kappa^2)\kappa_1\mathbf{K}_s\left[(1-\kappa^2)\mathbf{K}_s\right]^{-1}(1-\kappa^2)\kappa_1\mathbf{K}_s = (1-\kappa^2)\mathbf{K}_s(1-\kappa_1^2).$$

For $\mathbf{z}(t) \mid \mathbf{x}(t), \mathbf{y}(t), \mathbf{W}(t-1)$:

$$\begin{bmatrix}\mathbf{z}(t)\\ \mathbf{x}(t)\\ \mathbf{y}(t)\end{bmatrix}\Bigg| \mathbf{W}(t-1) \sim N_{3n}\left(\kappa \begin{bmatrix}\mathbf{z}(t-1)\\ \mathbf{x}(t-1)\\ \mathbf{y}(t-1)\end{bmatrix}, (1-\kappa^2)\begin{bmatrix}\mathbf{K}_s & \kappa_1\mathbf{K}_s & \kappa_2\mathbf{K}_s\\ \kappa_1\mathbf{K}_s & \mathbf{K}_s & \kappa_1\mathbf{K}_s\\ \kappa_2\mathbf{K}_s & \kappa_1\mathbf{K}_s & \mathbf{K}_s\end{bmatrix}\right).$$

The mean equals

$$\kappa\mathbf{z}(t-1) + (1-\kappa^2)\begin{pmatrix}\kappa_1\mathbf{K}_s & \kappa_2\mathbf{K}_s\end{pmatrix} \\ \left[(1-\kappa^2)\begin{pmatrix}\mathbf{K}_s & \kappa_1\mathbf{K}_s\\ \kappa_1\mathbf{K}_s & \mathbf{K}_s\end{pmatrix}\right]^{-1}\left(\begin{bmatrix}\mathbf{x}(t)\\ \mathbf{y}(t)\end{bmatrix} - \kappa\begin{bmatrix}\mathbf{x}(t-1)\\ \mathbf{y}(t-1)\end{bmatrix}\right),$$

but

$$\begin{pmatrix} \mathbf{K}_s & \kappa_1\mathbf{K}_s \\ \kappa_1\mathbf{K}_s & \mathbf{K}_s \end{pmatrix}^{-1} = \frac{1}{1-\kappa_1^2}\begin{pmatrix} 1 & \kappa_1 \\ -\kappa_1 & 1 \end{pmatrix} \otimes \mathbf{K}_s^{-1},$$

so that

$$\begin{aligned}(\kappa_1\mathbf{K}_s \quad \kappa_2\mathbf{K}_s)\frac{1}{1-\kappa_1^2}\begin{pmatrix} 1 & -\kappa_1 \\ -\kappa_1 & 1 \end{pmatrix} \otimes \mathbf{K}_s^{-1} \\ = \frac{1}{1-\kappa_1^2}\left[(\kappa_1-\kappa_1\kappa_2)\mathbf{I} \quad (\kappa_2-\kappa_1^2)\mathbf{I}\right].\end{aligned}$$

Then the mean is:

$$\begin{aligned}&\kappa\mathbf{z}(t-1) + \frac{1}{1-\kappa_1^2}\left[(\kappa_1-\kappa_1\kappa_2)\mathbf{I} \quad (\kappa_2-\kappa_1^2)\mathbf{I}\right]\left(\begin{bmatrix}\mathbf{x}(t)\\ \mathbf{y}(t)\end{bmatrix} - \kappa\begin{bmatrix}\mathbf{x}(t-1)\\ \mathbf{y}(t-1)\end{bmatrix}\right) \\ &= \kappa\mathbf{z}(t-1) + \frac{1}{1-\kappa_1^2}\left[(\kappa_1-\kappa_1\kappa_2)(\mathbf{x}(t)-\kappa\mathbf{x}(t-1)) + (\kappa_2-\kappa_1^2)(\mathbf{y}(t)-\kappa\mathbf{y}(t-1))\right].\end{aligned}$$

The covariance is:

$$\begin{aligned}(1-\kappa^2)\mathbf{K}_s - (1-\kappa^2)(\kappa_1\mathbf{K}_s \quad \kappa_2\mathbf{K}_s)\left[(1-\kappa^2)\begin{pmatrix}\mathbf{K}_s & \kappa_1\mathbf{K}_s \\ \kappa_1\mathbf{K}_s & \mathbf{K}_s\end{pmatrix}\right]^{-1}(1-\kappa^2)\begin{pmatrix}\kappa_1\mathbf{K}_s \\ \kappa_2\mathbf{K}_s\end{pmatrix} \\ = (1-\kappa^2)\left(1 - \frac{\kappa_1^2+\kappa_2^2-2\kappa_1^2\kappa_2}{1-\kappa_1^2}\right)\mathbf{K}_s\end{aligned}$$

B.3 Predictive distributions for the evolution of 3D curves

As stated in Chapter 5, marginal predictions at time $q \in \{1,\ldots,T\}$ can be done at a set of test points $\mathbf{s}^* = (s_{1^*},\ldots,s_{n^*})$ for each coordinate x, y and z. The distribution for a three-dimensional curve at time q and the whole sequence can be written as:

$$\begin{pmatrix}\mathbf{W}^*(q)\\ \mathbf{W}\end{pmatrix} \sim N_{n^*+n*3*T}\left(\mathbf{0}, \begin{pmatrix}\mathbf{K}_c\otimes\mathbf{K}_{s^*} & \mathbf{L}\otimes\mathbf{K}_c\otimes\mathbf{K}_{s^*s} \\ (\mathbf{L}\otimes\mathbf{K}_c\otimes\mathbf{K}_{s^*s})^{\mathrm{T}} & \mathbf{K}_t\otimes\mathbf{K}_c\otimes\mathbf{K}_s\end{pmatrix}\right),$$

where

$$\mathbf{L} = \left[\exp\left(\frac{-\mid q-1\mid}{\mu}\right) \quad \exp\left(\frac{-\mid q-2\mid}{\mu}\right) \quad \cdots \quad \exp\left(\frac{-\mid q-T\mid}{\mu}\right)\right]$$

and

$$\mathbf{K}_t = \begin{bmatrix} 1 & \kappa & \kappa^2 & \kappa^3 & \cdots \\ \kappa & 1 & \kappa & \kappa^2 & \cdots \\ \kappa^2 & \kappa & 1 & \kappa & \cdots \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix},$$

with $\kappa = \exp(-1/\mu)$. Then

$$\mathbf{W}^*(q) \mid \mathbf{W} \sim \big([\mathbf{L}\otimes\mathbf{K}_c\otimes\mathbf{K}_{s^*s}][\mathbf{K}_t\otimes\mathbf{K}_c\otimes\mathbf{K}_s]^{-1}\mathbf{W}, \\ \mathbf{K}_c\otimes\mathbf{K}_{s^*} - [\mathbf{L}\otimes\mathbf{K}_c\otimes\mathbf{K}_{s^*s}][\mathbf{M}\otimes\mathbf{K}_c\otimes\mathbf{K}_s]^{-1}[\mathbf{L}\otimes\mathbf{K}_c\otimes\mathbf{K}_{s^*s}]^{\mathrm{T}}\big).$$

By the properties of the Kronecker product,

$$[\mathbf{M}\otimes\mathbf{K}_c\otimes\mathbf{K}_s]^{-1} = \mathbf{K}_t^{-1}\otimes\mathbf{K}_c^{-1}\otimes\mathbf{K}_s^{-1}.$$

$\mathbf{K}_t^{-1}$ was calculated in B.1, and

$$\mathbf{K}_c^{-1} = \frac{1}{1-2\kappa_1^2+\kappa_2}\begin{pmatrix} \kappa_2+1 & -\kappa_1 & -\kappa_1 \\ -\kappa_1 & \frac{1-\kappa_1^2}{1-\kappa_2} & \frac{\kappa_1^2-\kappa_2}{1-\kappa_2} \\ -\kappa_1 & \frac{\kappa_1^2-\kappa_2}{1-\kappa_2} & \frac{1-\kappa_1^2}{1-\kappa_2} \end{pmatrix}.$$

Then:

$$\mathbf{K}_t^{-1}\otimes\mathbf{K}_c^{-1}\otimes\mathbf{K}_s^{-1} = \frac{1}{(1-\kappa^2)(1-2\kappa_1^2+\kappa_2)}\cdot \\ \left[\begin{pmatrix} 1 & -\kappa & 0 & \cdots & \cdots & 0 \\ -\kappa & 1+\kappa^2 & -\kappa & 0 & \cdots & \vdots \\ 0 & -\kappa & 1+\kappa^2 & -\kappa & \cdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \ddots & -\kappa \\ 0 & \vdots & \vdots & \vdots & -\kappa & 1 \end{pmatrix} \otimes \begin{pmatrix} \kappa_2+1 & -\kappa_1 & -\kappa_1 \\ -\kappa_1 & \frac{1-\kappa_1^2}{1-\kappa_2} & \frac{\kappa_1^2-\kappa_2}{1-\kappa_2} \\ -\kappa_1 & \frac{\kappa_1^2-\kappa_2}{1-\kappa_2} & \frac{1-\kappa_1^2}{1-\kappa_2} \end{pmatrix} \otimes \mathbf{K}_s^{-1}\right].$$

B.3.1 Marginal prediction

Let $q \in \{1, ..., T\}$, then:

$$\mathbf{L} = \begin{pmatrix} \kappa^{q-1} & \kappa^{q-2} & \cdots & \kappa & 1 & \kappa & \cdots & \kappa^{T-q} \end{pmatrix},$$

with the '1' being at position q.

Using properties of the Kronecker product and the result for $\mathbf{L}\mathbf{K}_t^{-1}$ for the Marginal Prediction in B.1:

$$(\mathbf{L} \otimes \mathbf{K}_c \otimes \mathbf{K}_{s^*s})(\mathbf{K}_t^{-1} \otimes \mathbf{K}_c^{-1} \otimes \mathbf{K}_s^{-1}) =$$
$$(\mathbf{L}\mathbf{K}_t^{-1}) \otimes (\mathbf{K}_c\mathbf{K}_c^{-1}) \otimes (\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}) = \begin{pmatrix} 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \end{pmatrix} \otimes \mathbf{I}_{3\times 3} \otimes (\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}),$$

with the 1 being at position q.

Then the mean $[\mathbf{L} \otimes \mathbf{K}_c \otimes \mathbf{K}_{s^*s}][\mathbf{K}_t \otimes \mathbf{K}_c \otimes \mathbf{K}_s]^{-1}\mathbf{W}$

$$= \begin{pmatrix} 0 & \cdots & 0 \end{pmatrix} \begin{pmatrix} \mathbf{K}_{s^*s}\mathbf{K}_s^{-1} & 0 & 0 \\ 0 & \mathbf{K}_{s^*s}\mathbf{K}_s^{-1} & 0 \\ 0 & 0 & \mathbf{K}_{s^*s}\mathbf{K}_s^{-1} \end{pmatrix} \begin{pmatrix} 0 & \cdots & 0 \end{pmatrix} \begin{bmatrix} \mathbf{W}(1) \\ \vdots \\ \mathbf{W}^*(q) \\ \vdots \\ \mathbf{W}(T) \end{bmatrix}$$

$$= \begin{pmatrix} \mathbf{K}_{s^*s}\mathbf{K}_s^{-1} & 0 & 0 \\ 0 & \mathbf{K}_{s^*s}\mathbf{K}_s^{-1} & 0 \\ 0 & 0 & \mathbf{K}_{s^*s}\mathbf{K}_s^{-1} \end{pmatrix} \mathbf{W}^*(q) = \begin{pmatrix} \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{x}^*(q) \\ \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{y}^*(q) \\ \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{z}^*(q) \end{pmatrix}.$$

The covariance is

$$\mathbf{K}_c \otimes \mathbf{K}_{s^*} - [\mathbf{L} \otimes \mathbf{K}_c \otimes \mathbf{K}_{s^*s}][\mathbf{M} \otimes \mathbf{K}_c \otimes \mathbf{K}_s]^{-1}[\mathbf{L} \otimes \mathbf{K}_c \otimes \mathbf{K}_{s^*s}]^{\mathrm{T}} =$$
$$(\mathbf{K}_c \otimes \mathbf{K}_{s^*}) - (\mathbf{L} \otimes \mathbf{K}_c \otimes \mathbf{K}_{s^*s})(\mathbf{K}_t^{-1} \otimes \mathbf{K}_c^{-1} \otimes \mathbf{K}_s^{-1})(\mathbf{L}^{\mathrm{T}} \otimes \mathbf{K}_c \otimes \mathbf{K}_{ss^*}).$$

But,

$$
\begin{aligned}
&(\mathbf{L}\otimes\mathbf{K}_c\otimes\mathbf{K}_{s^*s})(\mathbf{K}_t^{-1}\otimes\mathbf{K}_c^{-1}\otimes\mathbf{K}_s^{-1})(\mathbf{L}^{\mathrm{T}}\otimes\mathbf{K}_c\otimes\mathbf{K}_{ss^*})\\
&=\begin{pmatrix} 0 & \cdots & 0 \end{pmatrix}\begin{pmatrix} \mathbf{K}_{s^*s}\mathbf{K}_s^{-1} & 0 & 0\\ 0 & \mathbf{K}_{s^*s}\mathbf{K}_s^{-1} & 0\\ 0 & 0 & \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\end{pmatrix}\begin{pmatrix} 0 & \cdots & 0\end{pmatrix}\begin{pmatrix} 0\\ \vdots\\ \mathbf{K}_c\otimes\mathbf{K}_{ss^*}\\ \vdots\\ 0\end{pmatrix}\\
&=\begin{pmatrix} \mathbf{K}_{s^*s}\mathbf{K}_s^{-1} & 0 & 0\\ 0 & \mathbf{K}_{s^*s}\mathbf{K}_s^{-1} & 0\\ 0 & 0 & \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\end{pmatrix}\begin{pmatrix} \mathbf{K}_{ss^*} & \kappa_1\mathbf{K}_{ss^*} & \kappa_1\mathbf{K}_{ss^*}\\ \kappa_1\mathbf{K}_{ss^*} & \mathbf{K}_{ss^*} & \kappa_2\mathbf{K}_{ss^*}\\ \kappa_1\mathbf{K}_{ss^*} & \kappa_2\mathbf{K}_{ss^*} & \mathbf{K}_{ss^*}\end{pmatrix}\\
&=\begin{pmatrix} \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*} & \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\kappa_1\mathbf{K}_{ss^*} & \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\kappa_1\mathbf{K}_{ss^*}\\ \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\kappa_1\mathbf{K}_{ss^*} & \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*} & \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\kappa_2\mathbf{K}_{ss^*}\\ \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\kappa_1\mathbf{K}_{ss^*} & \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\kappa_2\mathbf{K}_{ss^*} & \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}\end{pmatrix}=\mathbf{K}_c\otimes(\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}).
\end{aligned}
$$

Therefore:

$$
\mathbf{K}_c\otimes\mathbf{K}_c-\mathbf{K}_c\otimes(\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*})=\mathbf{K}_c\otimes(\mathbf{K}_{s^*}-\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}).
$$

And hence

$$
\begin{bmatrix}\mathbf{x}^*(q)\\ \mathbf{y}^*(q)\\ \mathbf{z}^*(q)\end{bmatrix}\Bigg|\,\mathbf{W}\sim N_{3m}\left(\begin{bmatrix}\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{x}(q)\\ \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{y}(q)\\ \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{z}(q)\end{bmatrix},\begin{pmatrix}1 & \kappa_1 & \kappa_1\\ \kappa_1 & 1 & \kappa_2\\ \kappa_1 & \kappa_1 & 1\end{pmatrix}\otimes(\mathbf{K}_{s^*}-\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*})\right),
$$

which is the same as the prediction for a three-dimensional curve (A.2), and the conditional distributions between coordinates can be written as:

$$
\begin{aligned}
\mathbf{x}^*(q)\mid\mathbf{W} &\sim N\left(\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{x}(q),\mathbf{K}_{s^*}-\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}\right)\\
\mathbf{y}^*(q)\mid\mathbf{x}^*(q),\mathbf{W} &\sim N\left(\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{y}(q)+\kappa_1[\mathbf{x}^*(q)-\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{x}(q)],\right.\\
&\qquad\left.[1-\kappa_1^2][\mathbf{K}_{s^*}-\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}]\right)\\
\mathbf{z}^*(q)\mid\mathbf{x}^*(q),\mathbf{y}^*(q),\mathbf{W} &\sim N\Bigg(\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{z}(q)+\frac{1}{1-\kappa_1^2}\left[\{\kappa_1-\kappa_1\kappa_2\}\right.\\
&\qquad\left\{\mathbf{x}^*(q)-\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{x}(q)\right\}+\left\{\kappa_2-\kappa_1^2\right\}\left\{\mathbf{y}^*-\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{y}\right\}\Big],\\
&\qquad\left[1-\frac{\kappa_1^2+\kappa_2^2-2\kappa_1^2\kappa_2}{1-\kappa_1^2}\right]\left[\mathbf{K}_{s^*}-\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}\right]\Bigg).
\end{aligned}
$$

B.3.2 Prediction for future time points

Let $q > T$, then:

$$\mathbf{L} = \begin{pmatrix} \kappa^{q-1} & \kappa^{q-2} & \cdots & \kappa^{q-T} \end{pmatrix}.$$

Following similar calculations as in the previous section, it can be seen that in this case, the predictive distribution follow the same distributions as the three-dimensional curve, with the time-scaling parameter from 'Prediction for future time points' in B.1. Therefore, when $q > T$:

$$\begin{bmatrix} \mathbf{x}^*(q) \\ \mathbf{y}^*(q) \\ \mathbf{z}^*(q) \end{bmatrix} \Bigg| \mathbf{W} \sim N_{3m} \left(\kappa^{q-T} \begin{bmatrix} \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{x}(T) \\ \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{y}(T) \\ \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{z}(T) \end{bmatrix}, \mathbf{K}_c \otimes \left(\mathbf{K}_{s^*} - (\kappa^{q-T})^2 \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*} \right) \right),$$

and the conditional predictive distributions are:

$$\begin{aligned}
\mathbf{x}^*(q) \mid \mathbf{W} &\sim N\left(\kappa^{q-T}\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{x}(T), \mathbf{K}_{s^*} - (\kappa^{q-T})^2\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}\right) \\
\mathbf{y}^*(q) \mid \mathbf{x}^*(q), \mathbf{W} &\sim N\left(\kappa^{q-T}\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{y}(T) + \kappa_1[\mathbf{x}^*(q) - \kappa^{q-T}\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{x}(T)],\right. \\
&\quad \left.[1-\kappa_1^2][\mathbf{K}_{s^*} - (\kappa^{q-T})^2\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}]\right) \\
\mathbf{z}^*(q) \mid \mathbf{x}^*(q), \mathbf{y}^*(q), \mathbf{W} &\sim N\Bigg(\kappa^{q-T}\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{z}(T) + \frac{1}{1-\kappa_1^2} \\
&\quad \left[\{\kappa_1 - \kappa_1\kappa_2\}\left\{\mathbf{x}^*(q) - \kappa^{q-T}\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{x}(T)\right\} + \right. \\
&\quad \left.\left\{\kappa_2 - \kappa_1^2\right\}\left\{\mathbf{y}^*(q) - \kappa^{q-T}\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{y}(T)\right\}\right], \\
&\quad \left[1 - \frac{\kappa_1^2 + \kappa_2^2 - 2\kappa_1^2\kappa_2}{1-\kappa_1^2}\right]\left[\mathbf{K}_{s^*} - (\kappa^{q-T})^2\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}\right]\Bigg).
\end{aligned}$$

B.3.3 Retrodiction for previous time points

Let $q < 1$, then:

$$\mathbf{L} = \begin{pmatrix} \kappa^{1-q} & \kappa^{2-q} & \cdots & \kappa^{T-q} \end{pmatrix}.$$

Now:

$$\begin{bmatrix} \mathbf{x}^*(q) \\ \mathbf{y}^*(q) \\ \mathbf{z}^*(q) \end{bmatrix} \Bigg| \mathbf{W} \sim N_{3m} \left(\kappa^{1-q} \begin{bmatrix} \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{x}(1) \\ \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{y}(1) \\ \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{z}(1) \end{bmatrix}, \mathbf{K}_c \otimes \left(\mathbf{K}_{s^*} - (\kappa^{1-q})^2 \mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*} \right) \right),$$

and the conditional predictive distributions are:

$$\begin{aligned}
\mathbf{x}^*(q) \mid \mathbf{W} &\sim N\left(\kappa^{1-q}\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{x}(1), \mathbf{K}_{s^*} - (\kappa^{1-q})^2\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}\right) \\
\mathbf{y}^*(q) \mid \mathbf{x}^*(q), \mathbf{W} &\sim N\left(\kappa^{1-q}\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{y}(1) + \kappa_1[\mathbf{x}^*(q) - \kappa^{1-q}\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{x}(1)],\right. \\
&\quad \left.[1-\kappa_1^2][\mathbf{K}_{s^*} - (\kappa^{1-q})^2\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}]\right) \\
\mathbf{z}^*(q) \mid \mathbf{x}^*(q), \mathbf{y}^*(q), \mathbf{W} &\sim N\Bigg(\kappa^{1-q}\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{z}(1) + \frac{1}{1-\kappa_1^2} \\
&\quad \left[\{\kappa_1 - \kappa_1\kappa_2\}\left\{\mathbf{x}^*(q) - \kappa^{1-q}\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{x}(1)\right\} + \right. \\
&\quad \left.\left\{\kappa_2 - \kappa_1^2\right\}\left\{\mathbf{y}^*(q) - \kappa^{1-q}\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{y}(1)\right\}\right], \\
&\quad \left[1 - \frac{\kappa_1^2 + \kappa_2^2 - 2\kappa_1^2\kappa_2}{1-\kappa_1^2}\right]\left[\mathbf{K}_{s^*} - (\kappa^{1-q})^2\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*}\right]\Bigg).
\end{aligned}$$

B.3.4 Interpolation between time points

Let $t < q < t+1$, where $t \in [1, T-1]$, then:

$$\mathbf{L} = \begin{pmatrix} \kappa^{q-1} & \kappa^{q-2} & \dots & \kappa^{q-T} & \kappa^{t+1-q} & \kappa^{t+2-q} & \dots & \kappa^{t-q} \end{pmatrix}.$$

In this scenario:

$$\begin{aligned}
&\begin{bmatrix} \mathbf{x}^*(q) \\ \mathbf{y}^*(q) \\ \mathbf{z}^*(q) \end{bmatrix} \Bigg| \mathbf{W} \sim \\
&N_{3m}\left(\begin{bmatrix} \frac{1}{1-\kappa^2}\left[(\kappa^{q-t} - \kappa^{t-q+2})\mathbf{x}(t) + (\kappa(\kappa^{t-q} - \kappa^{q-t}))\mathbf{x}(t+1)\right]\mathbf{K}_{s^*s}\mathbf{K}_s^{-1} \\ \frac{1}{1-\kappa^2}\left[(\kappa^{q-t} - \kappa^{t-q+2})\mathbf{y}(t) + (\kappa(\kappa^{t-q} - \kappa^{q-t}))\mathbf{y}(t+1)\right]\mathbf{K}_{s^*s}\mathbf{K}_s^{-1} \\ \frac{1}{1-\kappa^2}\left[(\kappa^{q-t} - \kappa^{t-q+2})\mathbf{z}(t) + (\kappa(\kappa^{t-q} - \kappa^{q-t}))\mathbf{z}(t+1)\right]\mathbf{K}_{s^*s}\mathbf{K}_s^{-1} \end{bmatrix}, \right. \\
&\qquad \left. \mathbf{K}_c \otimes (\mathbf{K}_{s^*} - \frac{\kappa^{2(q-t)} + \kappa^{2(t-q)+2} - 2\kappa^2}{1-\kappa^2}\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*})\right).
\end{aligned}$$

Let:

$$\begin{aligned}
\text{mx} &= \frac{1}{1-\kappa^2}\left[(\kappa^{q-t}-\kappa^{t-q+2})\mathbf{x}(t)+(\kappa(\kappa^{t-q}-\kappa^{q-t}))\mathbf{x}(t+1)\right]\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\\
\text{my} &= \frac{1}{1-\kappa^2}\left[(\kappa^{q-t}-\kappa^{t-q+2})\mathbf{y}(t)+(\kappa(\kappa^{t-q}-\kappa^{q-t}))\mathbf{y}(t+1)\right]\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\\
\text{mz} &= \frac{1}{1-\kappa^2}\left[(\kappa^{q-t}-\kappa^{t-q+2})\mathbf{z}(t)+(\kappa(\kappa^{t-q}-\kappa^{q-t}))\mathbf{z}(t+1)\right]\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\\
\text{cov} &= \frac{\kappa^{2(q-t)}+\kappa^{2(t-q)+2}-2\kappa^2}{1-\kappa^2},
\end{aligned}$$

so that the conditional predictive distributions are:

$$\begin{aligned}
\mathbf{x}^*(q)\mid\mathbf{W} &\sim N\left(\text{mx},\mathbf{K}_c\otimes(\mathbf{K}_{s^*}-\text{cov}\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*})\right)\\
\mathbf{y}^*(q)\mid\mathbf{x}^*(q),\mathbf{W} &\sim N\left(\text{my}+\kappa_1[\mathbf{x}(q)-\text{mx}],\right.\\
&\quad\left.[1-\kappa_1^2][\mathbf{K}_c\otimes(\mathbf{K}_{s^*}-\text{cov}\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*})]\right)\\
\mathbf{z}^*(q)\mid\mathbf{x}^*(q),\mathbf{y}^*(q),\mathbf{W} &\sim N\left(\text{mz}+\frac{1}{1-\kappa_1^2}\left[\{\kappa_1-\kappa_1\kappa_2\}\{\mathbf{x}^*(q)-\text{mx}\}+\right.\right.\\
&\quad\left\{\kappa_2-\kappa_1^2\right\}\{\mathbf{y}^*(q)-\text{my}\}\Big],\\
&\quad\left.\left[1-\frac{\kappa_1^2+\kappa_2^2-2\kappa_1^2\kappa_2}{1-\kappa_1^2}\right][\mathbf{K}_c\otimes(\mathbf{K}_{s^*}-\text{cov}\mathbf{K}_{s^*s}\mathbf{K}_s^{-1}\mathbf{K}_{ss^*})]\right).
\end{aligned}$$

B.4 Estimation of the signal variance, σ_f^2, for the evolution of 3D curves

Recall the log-likelihood:

$$\log p(\mathbf{W}\mid\boldsymbol{\theta})=\log p(\mathbf{W}(1)\mid\boldsymbol{\theta})+\sum_{i=2}^{T}\log p(\mathbf{W}(t)\mid\mathbf{W}(t-1),\boldsymbol{\theta}).$$

Writing the covariance matrix as $\sigma_f^2\mathbf{K}_s$, the log-likelihood at time point one is

$$\begin{aligned}
&\log p(\mathbf{W}(1) \mid \boldsymbol{\theta}) = \log p\big(\mathbf{x}(1) \mid \boldsymbol{\theta}\big) + \log p\big(\mathbf{y}(1) \mid \mathbf{x}(1), \boldsymbol{\theta}\big) + \log p\big(\mathbf{z}(1) \mid \mathbf{x}(1), \mathbf{y}(1), \boldsymbol{\theta}\big) \\
&= 3\left[\frac{n}{2}\log(2\pi) - \frac{n}{2}\log \mid \sigma_f^2\mathbf{K}_s \mid\right] - \frac{1}{2}\mathbf{x}(1)^{\mathrm{T}}\left[\sigma_f^2\mathbf{K}_s\right]^{-1}\mathbf{x}(1) - \frac{n}{2}\log(a) - \frac{n}{2}\log(c) \\
&-\frac{1}{2a}\big(\mathbf{y}(1)-\kappa_1\mathbf{x}(1)\big)^{\mathrm{T}}\left[\sigma_f^2\mathbf{K}_s\right]^{-1}\big(\mathbf{y}(1)-\kappa_1\mathbf{x}(1)\big) - \frac{1}{2c}\big(\mathbf{z}(1)-b\big)^{\mathrm{T}}\left[\sigma_f^2\mathbf{K}_s\right]^{-1}\big(\mathbf{z}(1)-b\big) \\
&\qquad \propto -3n\log(\sigma_f) - \frac{1}{2\sigma_f^2}\mathrm{Z1}
\end{aligned}$$

with:

$$\begin{aligned}
a &= 1-\kappa_1^2, \\
b &= \frac{(\kappa_1-\kappa_1\kappa_2)\mathbf{x}(1) + (\kappa_2-\kappa_1^2)\mathbf{y}(1)}{a}, \\
c &= 1 - \frac{\kappa_1^2+\kappa_2^2-2\kappa_1^2\kappa_2}{a}, \\
\mathrm{Z1} &= \big(\mathbf{x}(1)^{\mathrm{T}}\mathbf{K}_s^{-1}\mathbf{x}(1)\big) + \frac{1}{a}\big(\mathbf{y}(1)-\kappa_1\mathbf{x}(1)\big)^{\mathrm{T}}\mathbf{K}_s^{-1}\big(\mathbf{y}(1)-\kappa_1\mathbf{x}(1)\big) + \\
&\quad \frac{1}{c}\big(\mathbf{z}(1)-b\big)^{\mathrm{T}}\mathbf{K}_s^{-1}\big(\mathbf{z}(1)-b\big).
\end{aligned}$$

For the conditional log-likelihood:

$$\begin{aligned}
\log p(\mathbf{W}(t) \mid \mathbf{W}(t-1), \boldsymbol{\theta}) &= \log p(\mathbf{x}(t) \mid \mathbf{W}(t-1), \boldsymbol{\theta}) \\
&\quad + \log p(\mathbf{y}(t) \mid \mathbf{x}(t), \mathbf{W}(t-1), \boldsymbol{\theta}) + \log p(\mathbf{z}(t) \mid \mathbf{x}(t), \mathbf{y}(t)\mathbf{W}(t-1), \boldsymbol{\theta}).
\end{aligned}$$

For $\mathbf{x}(t) \mid \mathbf{W}(t-1), \boldsymbol{\theta}$, let:

$$\begin{aligned}
\mathrm{m1} &= \kappa\mathbf{x}(t-1), \\
\mathrm{cov1} &= 1-\kappa^2.
\end{aligned}$$

Then:

$$\begin{aligned}
&\log p\big(\mathbf{x}(t) \mid \mathbf{W}(t-1), \boldsymbol{\theta}\big) = \\
&-\frac{n}{2}\log(2\pi) - \frac{1}{2}\log \mid \sigma_f^2\mathbf{K}_s \mid - \frac{n}{2}\log(\mathrm{cov1}) - \frac{1}{2\mathrm{cov1}}\big(\mathbf{x}(t)-\mathrm{m1}\big)^{\mathrm{T}}\left[\sigma_f^2\mathbf{K}_s\right]^{-1}\big(\mathbf{x}(t)-\mathrm{m1}\big) \\
&\qquad \propto -n\log(\sigma_f) - \frac{1}{2\sigma_f^2}\frac{\big(\mathbf{x}(t)-\mathrm{m1}\big)^{\mathrm{T}}\mathbf{K}_s^{-1}\big(\mathbf{x}(t)-\mathrm{m1}\big)}{\mathrm{cov1}}.
\end{aligned}$$

For $\mathbf{y}(t) \mid \mathbf{x}(t), \mathbf{W}(t-1), \boldsymbol{\theta}$, let:

$$\begin{aligned} \mathrm{m2} &= \kappa \mathbf{y}(t-1) + \kappa_2 \left[\mathbf{x}(t) - \kappa \mathbf{x}(t-1)\right] \\ \mathrm{cov2} &= (1-\kappa_1^2)(1-\kappa^2). \end{aligned}$$

Then:

$$\log p\big(\mathbf{y}(t) \mid x(t), \mathbf{W}(t-1), \boldsymbol{\theta}\big) \propto -n \log(\sigma_f) - \frac{1}{2\sigma_f^2} \frac{\big(\mathbf{y}(t) - \mathrm{m2}\big)^{\mathrm{T}} \mathbf{K}_s^{-1} \big(\mathbf{y}(t) - \mathrm{m2}\big)}{\mathrm{cov2}}.$$

For $\mathbf{z}(t) \mid \mathbf{x}(t), \mathbf{y}(t), \mathbf{W}(t-1), \boldsymbol{\theta}$, let:

$$\begin{aligned} \mathrm{m3} &= \kappa \mathbf{z}(t-1) + \frac{1}{1-\kappa_1^2} \big[\kappa_1(1-\kappa_2)\big(\mathbf{x}(t) - \kappa \mathbf{x}(t-1)\big) \\ &\quad + (\kappa_2 - \kappa_1^2)\big(\mathbf{y}(t) - \kappa \mathbf{y}(t-1)\big)\big] \\ \mathrm{cov3} &= (1-\kappa^2)\left(\frac{\kappa_1^2 + \kappa_2^2 - 2\kappa_1^2\kappa_2}{1-\kappa_1^2}\right). \end{aligned}$$

Then:

$$\log p\big(\mathbf{z}(t) \mid \mathbf{x}(t), \mathbf{y}(t), \mathbf{W}(t-1), \boldsymbol{\theta}\big) \propto -n \log(\sigma_f) - \frac{1}{2\sigma_f^2} \frac{\big(\mathbf{z}(t) - \mathrm{m3}\big)^{\mathrm{T}} \mathbf{K}_s^{-1} \big(\mathbf{z}(t) - \mathrm{m3}\big)}{\mathrm{cov3}}.$$

Then the total log-likelihood:

$$\log p(\mathbf{W}(t) \mid \mathbf{W}(t-1), \boldsymbol{\theta}) \propto -3n \log(\sigma_f) - \frac{1}{2\sigma_f^2} \mathrm{Z2}_t,$$

with

$$\mathrm{Z2}_t = \frac{\big(\mathbf{x}(t) - \mathrm{m1}\big)^{\mathrm{T}} \mathbf{K}_s^{-1} \big(\mathbf{x}(t) - \mathrm{m1}\big)}{\mathrm{cov1}} + \frac{\big(\mathbf{y}(t) - \mathrm{m2}\big)^{\mathrm{T}} \mathbf{K}_s^{-1} \big(\mathbf{y}(t) - \mathrm{m2}\big)}{\mathrm{cov2}} + \frac{\big(\mathbf{z}(t) - \mathrm{m3}\big)^{\mathrm{T}} \mathbf{K}_s^{-1} \big(\mathbf{z}(t) - \mathrm{m3}\big)}{\mathrm{cov3}}.$$

Overall:

$$\log p(\mathbf{W} \mid \boldsymbol{\theta}) \propto -3n \log(\sigma_f) - \frac{1}{2\sigma_f^2} \mathrm{Z1} + \sum_{t=2}^{T} \left(-3n \log(\sigma_f) - \frac{1}{2\sigma_f^2} \mathrm{Z2}_t\right).$$

Therefore:

$$\hat{\sigma}_f^2 = \frac{\mathrm{Z1} + \sum_{t=2}^{T} \mathrm{Z2}_t}{3n(T+1)}.$$

The total log-likelihood for the model of three-dimensional curves evolving over time can therefore be written as:

$$\begin{aligned}
\log p(\mathbf{W} \mid \boldsymbol{\theta}) = &-\frac{3}{2}\left[n\log(2\pi) + n\log\left(\frac{\mathrm{Z1} + \sum_{t=2}^{T}\mathrm{Z2}_t}{3n(T+1)}\right) + \log \mid \mathbf{K}_s \mid\right] \\
&- \frac{n}{2}\left[\log(a) + \log(c)\right] - \frac{\mathrm{Z1}3n(T+1)}{2(\mathrm{Z1} + \sum_{t=2}^{T}\mathrm{Z2}_t)} \\
&+ \sum_{t=2}^{T}\left(-\frac{3}{2}\left[n\log(2\pi) + n\log\left(\frac{\mathrm{Z1} + \sum_{t=2}^{T}\mathrm{Z2}_t}{3n(T+1)}\right) + \log \mid \mathbf{K}_s \mid\right]\right. \\
&\left. - \frac{n}{2}\left[\log(\mathrm{cov1}) + \log(\mathrm{cov2}) + \log(\mathrm{cov3})\right] - \frac{\mathrm{Z2}_t 3n(T+1)}{2(\mathrm{Z1} + \sum_{t=2}^{T}\mathrm{Z2}_t)}\right).
\end{aligned}$$

B.5 Optimal hyperparameters across replicates of the six emotions

Anger:

$\hat{\boldsymbol{\theta}}_1$	Replicate 1	Replicate 2	Replicate 3	Replicate 4	Mean
$\hat{\sigma_f}$	6.0038	5.6528	5.7536	5.0925	5.6257
$\hat{\sigma_f}$ SE	0.0503	0.0441	0.0459	0.0427	0.0457
$\hat{\lambda}$	0.3397	0.3437	0.3671	0.2683	0.3297
$\hat{\lambda}$ SE	0.0072	0.0072	0.0083	0.0048	0.0069
$\hat{\mu}$	43.8534	53.0714	59.6807	53.0745	52.4200
$\hat{\mu}$ SE	4.1950	5.2685	5.7820	5.5430	5.1971
$\log\left(L(\hat{\boldsymbol{\theta}}_1)\right)$	837.6156	1908.3593	2281.3865	1815.6018	1710.7408
$\hat{\boldsymbol{\theta}}_2$					
$\hat{\sigma_f}$	6.0569	5.4746	5.5162	5.1826	5.5576
$\hat{\sigma_f}$ SE	0.0507	0.0427	0.0440	0.0434	0.0452
$\hat{\lambda}$	0.3091	0.3106	0.3150	0.2626	0.2993
$\hat{\lambda}$ SE	0.0055	0.0053	0.0058	0.0048	0.0053
$\hat{\mu}$	19.3102	24.7835	27.8526	26.5300	24.6191
$\hat{\mu}$ SE	1.3184	1.7946	2.1015	2.0831	1.8244
$\hat{\kappa_1}$	−0.0215	−0.0081	−0.0098	-0.0187	-0.0145
$\hat{\kappa_1}$ SE	0.0038	0.0033	0.0027	0.0023	0.0030
$\hat{\kappa_2}$	0.7861	0.7353	0.7512	0.7377	0.7526
$\hat{\kappa_2}$ SE	0.0184	0.0129	0.0119	0.0164	0.0149
$\log\left(L(\hat{\boldsymbol{\theta}}_2)\right)$	−1066.1383	−10.8338	473.3732	204.8549	−99.6860

Disgust:

$\hat{\boldsymbol{\theta}}_1$	Replicate 1	Replicate 2	Replicate 3	Replicate 4	Mean
$\hat{\sigma_f}$	4.9265	5.1932	4.9377	5.2712	5.0822
$\hat{\sigma_f}$ SE	0.0521	0.0514	0.0469	0.0485	0.0497
$\hat{\lambda}$	0.1770	0.2215	0.2071	0.2198	0.2064
$\hat{\lambda}$ SE	0.0033	0.0038	0.0046	0.0044	0.0040
$\hat{\mu}$	50.7923	33.6945	41.3702	36.7914	40.6621
$\hat{\mu}$ SE	5.2325	3.5543	4.3866	3.8063	4.2449
$\log\left(L(\hat{\boldsymbol{\theta}}_1)\right)$	440.5479	−212.9106	370.3420	−107.5121	122.6168
$\hat{\boldsymbol{\theta}}_2$					
$\hat{\sigma_f}$	5.3383	5.3984	5.0139	4.9924	5.1858
$\hat{\sigma_f}$ SE	0.0565	0.0534	0.0476	0.0459	0.0508
$\hat{\lambda}$	0.1553	0.2145	0.1832	0.1729	0.1815
$\hat{\lambda}$ SE	0.0027	0.0038	0.0039	0.0048	0.0038
$\hat{\mu}$	25.5058	13.9328	13.2789	16.1759	17.2233
$\hat{\mu}$ SE	2.5851	1.0161	0.8688	1.0941	1.3910
$\hat{\kappa_1}$	−0.0164	−0.0150	0.0080	−0.0144	−0.0095
$\hat{\kappa_1}$ SE	0.0059	0.0107	0.0033	0.0076	0.0069
$\hat{\kappa_2}$	0.7850	0.7952	0.8460	0.7599	0.7966
$\hat{\kappa_2}$ SE	0.0286	0.0317	0.0126	0.0234	0.0241
$\log\left(L(\hat{\boldsymbol{\theta}}_2)\right)$	−965.4963	−1728.9216	−1812.8688	−1881.9284	−1597.3038

Fear:

$\hat{\boldsymbol{\theta}}_1$	Replicate 1	Replicate 2	Replicate 3	Replicate 4	Replicate 5	Mean
$\hat{\sigma_f}$	5.9636	5.7687	5.2686	5.6660	5.4497	5.6233
$\hat{\sigma_f}$ SE	0.0524	0.0534	0.0703	0.0560	0.0461	0.0557
$\hat{\lambda}$	0.3128	0.3090	0.2081	0.2645	0.2389	0.2667
$\hat{\lambda}$ SE	0.0059	0.0062	0.0057	0.0060	0.0046	0.0057
$\hat{\mu}$	50.7921	48.6073	18.4677	39.0036	44.5156	40.2773
$\hat{\mu}$ SE	5.0931	4.9845	1.9764	4.0975	4.0859	4.0475
$\log\big(L(\hat{\boldsymbol{\theta}}_1)\big)$	1211.7346	1077.8808	−1057.1032	176.6569	757.2048	433.2748
$\hat{\boldsymbol{\theta}}_2$						
$\hat{\sigma_f}$	6.1394	6.2272	5.3368	5.9717	5.5337	5.8418
$\hat{\sigma_f}$ SE	0.0539	0.0577	0.0712	0.0591	0.0468	0.0577
$\hat{\lambda}$	0.2630	0.2960	0.2003	0.2658	0.1872	0.2425
$\hat{\lambda}$ SE	0.0043	0.0065	0.0049	0.0061	0.0026	0.0049
$\hat{\mu}$	20.7547	16.9314	8.7386	15.0471	15.2742	15.3492
$\hat{\mu}$ SE	1.5487	1.2160	0.7030	1.1160	0.9183	1.1004
$\hat{\kappa_1}$	−0.0205	−0.0135	−0.0159	−0.0698	−0.0244	−0.0288
$\hat{\kappa_1}$ SE	0.0049	0.0039	0.0120	0.0128	0.0058	0.0079
$\hat{\kappa_2}$	0.8297	0.8567	0.7367	0.8000	0.8704	0.8187
$\hat{\kappa_2}$ SE	0.0130	0.0111	0.0382	0.0201	0.0089	0.0182
$\log\big(L(\hat{\boldsymbol{\theta}}_2)\big)$	−866.9068	−1065.5393	−1701.8763	−1482.4651	−1988.9865	−1421.1548

Happiness:

$\hat{\boldsymbol{\theta}}_1$	Replicate 1	Replicate 2	Replicate 3	Mean
$\hat{\sigma_f}$	6.1484	5.3884	5.5910	5.7093
$\hat{\sigma_f}$ SE	0.0537	0.0408	0.0427	0.0457
$\hat{\lambda}$	0.2109	0.2329	0.2387	0.2275
$\hat{\lambda}$ SE	0.0045	0.0041	0.0034	0.0040
$\hat{\mu}$	40.1740	45.8766	40.7691	42.2732
$\hat{\mu}$ SE	3.8945	3.8671	3.3601	3.7072
$\log\big(L(\hat{\boldsymbol{\theta}}_1)\big)$	−508.7242	1117.4325	464.0200	357.5761
$\hat{\boldsymbol{\theta}}_2$				
$\hat{\sigma_f}$	5.9089	5.8037	5.9078	5.8735
$\hat{\sigma_f}$ SE	0.0516	0.0440	0.0451	0.0469
$\hat{\lambda}$	0.0451	0.2120	0.2377	0.1649
$\hat{\lambda}$ SE	0.0003	0.0034	0.0032	0.0023
$\hat{\mu}$	317.2565	26.3194	23.0199	122.1986
$\hat{\mu}$ SE	33.3193	1.9482	1.5325	12.2667
$\hat{\kappa_1}$	0.0212	−0.0068	−0.0321	−0.0059
$\hat{\kappa_1}$ SE	0.0074	0.0033	0.0079	0.0062
$\hat{\kappa_2}$	0.7135	0.7457	0.6872	0.7155
$\hat{\kappa_2}$ SE	0.0292	0.0140	0.0119	0.0184
$\log\big(L(\hat{\boldsymbol{\theta}}_2)\big)$	−1497.0443	−1017.4359	−1440.8328	−1318.4377

Sadness:

$\hat{\boldsymbol{\theta}}_1$	Replicate 1	Replicate 2	Replicate 3	Replicate 4	Mean
$\hat{\sigma_f}$	4.1324	5.5799	5.5404	5.9400	5.2982
$\hat{\sigma_f}$ SE	0.0371	0.0450	0.0455	0.0432	0.0427
$\hat{\lambda}$	0.0378	0.2947	0.2546	0.2548	0.2105
$\hat{\lambda}$ SE	0.0003	0.0058	0.0069	0.0037	0.0042
$\hat{\mu}$	1070.7751	32.7227	30.4087	41.3702	293.8192
$\hat{\mu}$ SE	81.3627	2.6516	2.5378	3.5121	22.5161
$\log\big(L(\hat{\boldsymbol{\theta}}_1)\big)$	2743.3723	83.4536	−502.0028	316.1499	660.2433
$\hat{\boldsymbol{\theta}}_2$					
$\hat{\sigma_f}$	4.2756	5.3885	5.5399	5.9873	5.2978
$\hat{\sigma_f}$ SE	0.0384	0.0434	0.0455	0.0436	0.0427
$\hat{\lambda}$	0.0383	0.1853	0.1881	0.1981	0.1524
$\hat{\lambda}$ SE	0.0003	0.0028	0.0026	0.0063	0.0030
$\hat{\mu}$	274.0069	10.7485	8.3789	13.3917	76.6315
$\hat{\mu}$ SE	19.2391	0.5432	0.3969	0.7774	5.2391
$\hat{\kappa_1}$	−0.0420	−0.0150	−0.0228	0.0066	−0.0183
$\hat{\kappa_1}$ SE	0.0041	0.0061	0.0095	0.0074	0.0068
$\hat{\kappa_2}$	0.8869	0.8762	0.8891	0.8813	0.8834
$\hat{\kappa_2}$ SE	0.0135	0.0163	0.0166	0.0120	0.0146
$\log\big(L(\hat{\boldsymbol{\theta}}_2)\big)$	63.5302	−3243.0298	−3942.5536	−3448.8873	−2642.7351

Surprise:

$\hat{\boldsymbol{\theta}}_1$	Replicate 1	Replicate 2	Replicate 3	Replicate 4	Replicate 5	Mean
$\hat{\sigma_f}$	4.7427	5.9638	4.6369	5.9985	4.9934	5.2670
$\hat{\sigma_f}$ SE	0.0426	0.0507	0.0443	0.0559	0.0457	0.0478
$\hat{\lambda}$	0.0411	0.0478	0.0398	0.0466	0.0415	0.0434
$\hat{\lambda}$ SE	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003
$\hat{\mu}$	437.9642	346.4177	450.9916	240.1470	294.8397	354.0720
$\hat{\mu}$ SE	38.0562	44.4479	36.7390	28.0531	25.4490	34.5490
$\log\big(L(\hat{\boldsymbol{\theta}}_1)\big)$	−392.1987	−1743.2259	−309.4153	−2696.3497	−1788.9301	−1386.0239
$\hat{\boldsymbol{\theta}}_2$						
$\hat{\sigma_f}$	4.6338	5.4249	4.6364	5.8742	4.9324	5.1003
$\hat{\sigma_f}$ SE	0.0416	0.0461	0.0443	0.0547	0.0451	0.0464
$\hat{\lambda}$	0.0405	0.0447	0.0396	0.0457	0.0411	0.0423
$\hat{\lambda}$ SE	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003
$\hat{\mu}$	229.8172	181.7791	258.4055	150.2450	181.7791	200.4052
$\hat{\mu}$ SE	17.6672	17.3570	19.5921	15.3961	14.3135	16.8652
$\hat{\kappa_1}$	-0.0184	-0.0112	0.0326	0.0021	-0.0199	-0.0030
$\hat{\kappa_1}$ SE	0.0067	0.0084	0.0109	0.0138	0.0071	0.0094
$\hat{\kappa_2}$	0.7546	0.7391	0.6994	0.6871	0.6928	0.7146
$\hat{\kappa_2}$ SE	0.0198	0.0262	0.0176	0.0258	0.0313	0.0241
$\log\big(L(\hat{\boldsymbol{\theta}}_2)\big)$	−1403.6009	−2858.4654	−1206.2260	−3399.5303	−2526.7138	−2278.9073

B.6 Gaussian Process model for the evolution of 2D curves

The machinery is similar to that in Section 4.3, a GP can be specified as:

$$w(t,c,s) \sim GP\big(m(t,c,s), k(t,t',c,c',s,s')\big), \tag{B.1}$$

a mixed GP for the space component, indexed by the arc-length, $s \in [0,1]$, the discrete label includes c, which consists of two coordinates, i.e., $c = \{x, y\}$ and the time component $\mathbf{t} = (t_1, \ldots, t_T)$. Points on a two-dimensional curve at time t can be written as:

$$\mathbf{W}(t) = \begin{bmatrix} \mathbf{x}(t) \\ \mathbf{y}(t) \end{bmatrix}. \tag{B.2}$$

The sequence for a choice of T values of t, $\big(\mathbf{W}(t_1) \cdots \mathbf{W}(t_T)\big)^{\mathrm{T}}$ is denoted as $\mathbf{W}$. Separability is also assumed such that:

$$k(s,s',c,c',t,t') = k_t(t,t')k_c(c,c')k_s(s,s'). \tag{B.3}$$

The space-covariance function used is, as before, the Squared-Exponential (SE), the time-covariance function is the Ornstein-Uhlenbeck (OU) covariance function (Markov assumption) and the matrix $\mathbf{K}_c$ is now a 2×2 matrix, with hyperparameter κ_1, the correlation between x and y:

$$\mathbf{K}_c = \begin{pmatrix} 1 & \kappa_1 \\ \kappa_1 & 1 \end{pmatrix}. \tag{B.4}$$

B.6.1 Conditional dependencies

At $t = 1$, the distribution can be written as:

$$\mathbf{W}(1) = \begin{bmatrix} \mathbf{x}(1) \\ \mathbf{y}(1) \end{bmatrix} \sim N_{2n}\left(\mathbf{0}, \mathbf{K}_c \otimes \mathbf{K}_s\right). \tag{B.5}$$

The joint distribution for $\mathbf{W}(t)$ and $\mathbf{W}(t-1)$ is:

$$\begin{bmatrix} \mathbf{W}(t) \\ \mathbf{W}(t-1) \end{bmatrix} \sim N_{4n}\left(\mathbf{0}, \begin{bmatrix} \mathbf{K}_c \otimes \mathbf{K}_s & \kappa\mathbf{K}_c \otimes \mathbf{K}_s \\ \kappa\mathbf{K}_c \otimes \mathbf{K}_s & \mathbf{K}_c \otimes \mathbf{K}_s \end{bmatrix}\right), \tag{B.6}$$

so that

$$\mathbf{W}(t) \mid \mathbf{W}(t-1) \sim N_{2n}\left(\kappa \mathbf{W}(t-1), (1-\kappa^2)\mathbf{K}_c \otimes \mathbf{K}_s\right), \tag{B.7}$$

and therefore, the distributions for each curve separately, with the y-coordinate conditioned on the x-coordinate curve are:

$$\begin{aligned} \mathbf{x}(t) \mid \mathbf{W}(t-1) &\sim N_n\left(\kappa \mathbf{x}(t-1), (1-\kappa^2)\mathbf{K}_s\right), \\ \mathbf{y}(t) \mid \mathbf{x}(t), \mathbf{W}(t-1) &\sim N_n\Big(\kappa \mathbf{y}(t-1) + \kappa_1\big(\mathbf{x}(t) - \kappa \mathbf{x}(t-1)\big), \\ &\qquad (1-\kappa^2)\mathbf{K}_s(1-\kappa_1^2)\Big). \end{aligned} \tag{B.8}$$

The same as for the three dimensional model (5.27).

B.6.2 Likelihood

The total log-likelihood for the hyperparameters $\boldsymbol{\theta} = (\sigma_f, \lambda, \mu, \kappa_1)$ is:

$$\log p(\mathbf{W} \mid \boldsymbol{\theta}) = \log p(\mathbf{W}(1) \mid \boldsymbol{\theta}) + \sum_{i=2}^{T} \log p(\mathbf{W}(t) \mid \mathbf{W}(t-1), \boldsymbol{\theta}). \tag{B.9}$$

The log-likelihood at time-point 1 is:

$$\log p(\mathbf{W}(1) \mid \boldsymbol{\theta}) = \log p(\mathbf{x}(1) \mid \boldsymbol{\theta}) + \log p(\mathbf{y}(1) \mid \mathbf{x}(1), \boldsymbol{\theta}), \tag{B.10}$$

where:

$$\begin{aligned} \log p(\mathbf{x}(1) \mid \boldsymbol{\theta}) &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{K}_s| - \frac{1}{2}\mathbf{x}(1)^{\mathrm{T}}\mathbf{K}_s^{-1}\mathbf{x}(1), \\ \log p(\mathbf{y}(1) \mid \mathbf{x}(1), \boldsymbol{\theta}) &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{K}_s| - \frac{n}{2}\log(1-\kappa_1^2) \\ &\quad -\frac{1}{2(1-\kappa_1^2)}\big(y(1) - \kappa_1\mathbf{x}(1)\big)^{\mathrm{T}}\mathbf{K}_s^{-1}\big(\mathbf{y}(1) - \kappa_1\mathbf{x}(1)\big). \end{aligned} \tag{B.11}$$

For the remaining curves, the log-likelihood given the data at the previous time point

$$\begin{aligned} &\log p(\mathbf{W}(t) \mid \mathbf{W}(t-1), \boldsymbol{\theta}) = \\ &\qquad \log p(\mathbf{x}(t) \mid \mathbf{W}(t-1), \boldsymbol{\theta}) + \log p(\mathbf{y}(t) \mid \mathbf{x}(t), \mathbf{W}(t-1), \boldsymbol{\theta}). \end{aligned} \tag{B.12}$$

Let:

$$\begin{aligned} \text{m1} &= \kappa \mathbf{x}(t-1), \\ \text{cov1} &= (1-\kappa^2), \\ \text{m2} &= \kappa \mathbf{y}(t-1) + \kappa_1 \left[\mathbf{x}(t) - \kappa \mathbf{x}(t-1)\right], \\ \text{cov2} &= (1-\kappa^2)(1-\kappa_1^2). \end{aligned}$$

Then:

$$\begin{aligned} \log p\big(\mathbf{x}(t) \mid \mathbf{W}(t-1), \boldsymbol{\theta}\big) &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{K}_s| - \frac{n}{2}\log \text{cov1} \\ &\quad -\frac{1}{2\text{cov1}}\big(\mathbf{x}(t) - \text{m1}\big)^{\mathrm{T}}\mathbf{K}_s^{-1}\big(\mathbf{x}(t) - \text{m1}\big). \\ \log p\big(\mathbf{y}(t) \mid \mathbf{x}(t), \mathbf{W}(t-1), \boldsymbol{\theta}\big) &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{K}_s| - \frac{n}{2}\log \text{cov2} \\ &\quad -\frac{1}{2\text{cov2}}\big(\mathbf{y}(t) - \text{m2}\big)^{\mathrm{T}}\mathbf{K}_s^{-1}\big(\mathbf{y}(t) - \text{m2}\big). \end{aligned} \tag{B.13}$$

B.6.3 Discussion

Predictive distributions can be derived in a similar manner as for the three-dimensional curves. Given the equivalences between the models, the results found for the x and y coordinates are equivalent to those found adding the z coordinate. Optimal hyperparameters can be found by maximum likelihood. Since these results were presented in Section 5.3.6, they are not included here.

Appendix C

C.1 GP Phylogenetic model: multifurcating tree simulations

C.1.1 Two-dimensional curves simulation

A set of two-dimensional curves of 15 equally spaced points (arc-length from 0 to 1) was simulated using hyper-parameters $\boldsymbol{\theta} = (\sigma_f, \lambda, \mu, \kappa_1) = (0.5, 0.3, 0.7)$, shown in Figure C.1. The times for nodes A to E are (0.04, 0.05, 0.02, 0.06, 0.07), the leaves are at time 0 (same proportions are kept in the illustration).

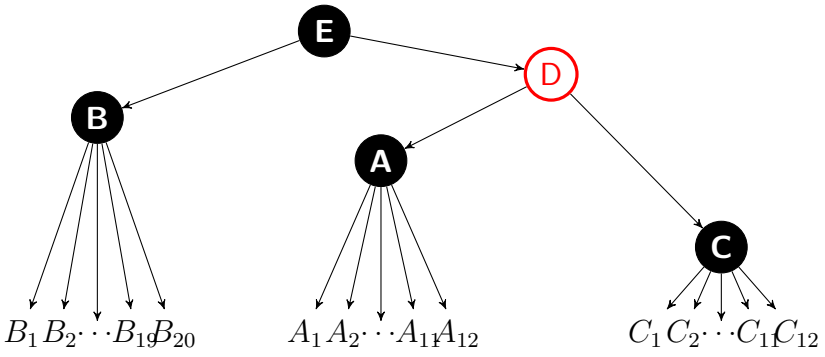


FIGURE C.1: Illustration of the simulated tree.

The curves are shown in Figure C.2. Tne x-coordinate values in blue and the y-coordinate values in green. The branches lengths are, again, proportional to the chosen distances.

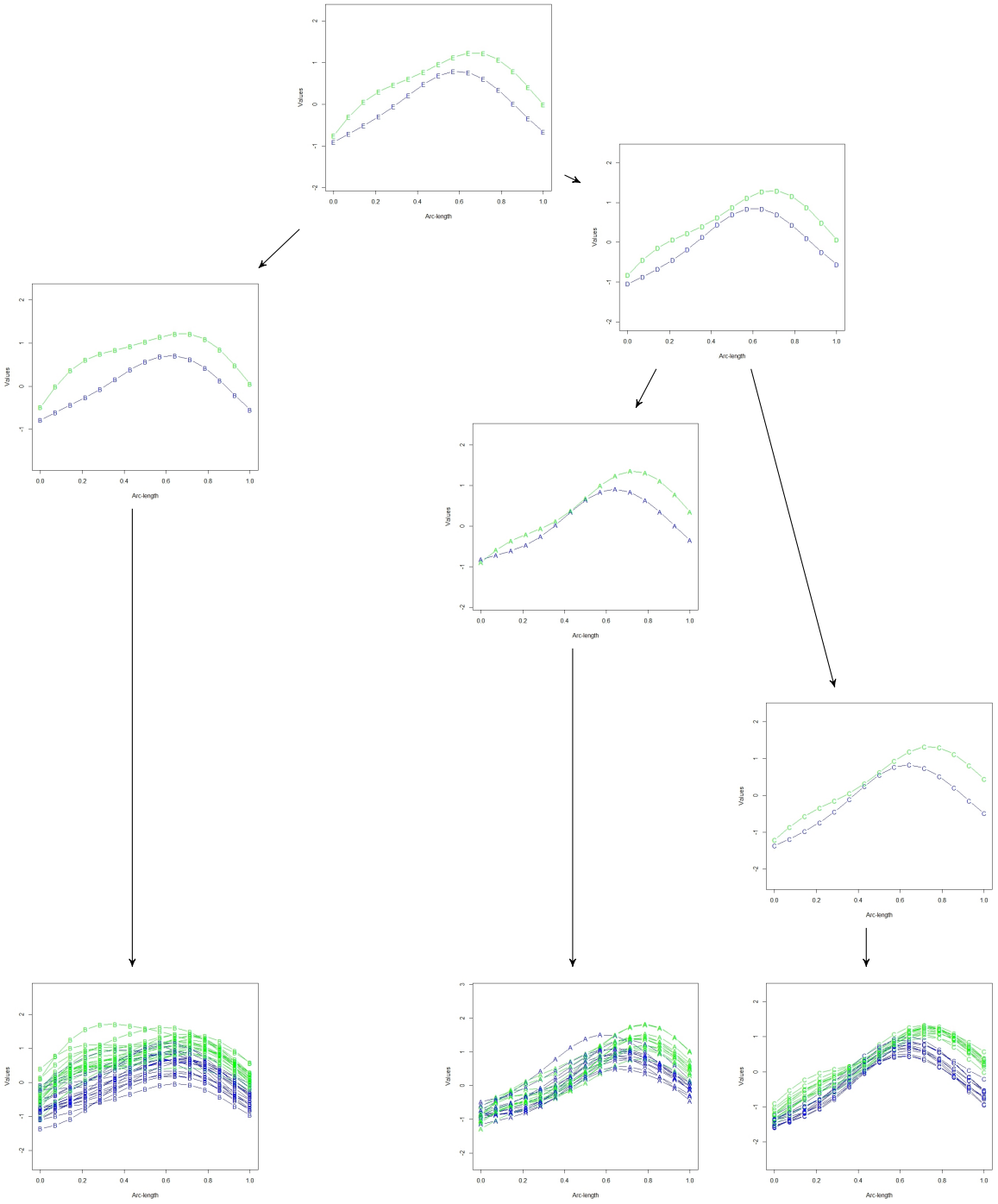


FIGURE C.2: Simulated 2D curves.

Given the small differences in times, there is not much variability between the curves. This is done to test the model in an scenario where finding differences may be hard. Defining $t_1 = t_B$, $t_2 = t_A - t_B$, $t_3 = t_C - t_A$, $t_4 = t_D - t_C$ and $t_5 = t_E - t_D$, the hyper-parameters to optimise are therefore: $\boldsymbol{\theta} = (\sigma_f, \lambda, \kappa_1, t_1, t_2, t_3, t_4, t_5) =$

$(0.5, 0.3, 0.7, 0.05, -0.01, -0.02, 0.04, 0.01)$. The likelihood value at these hyper-parameters is $\log(L(\boldsymbol{\theta})) = 2347.447$. Optimising by maximum likelihood, the optimal values for the hyper-parameters are:

$\hat{\boldsymbol{\theta}}$	$\hat{\sigma}_{f2\text{D}}$	$\hat{\lambda}_{2\text{D}}$	$\hat{\kappa}_{22\text{D}}$	$\hat{t}_1$	$\hat{t}_2$	$\hat{t}_3$	$\hat{t}_4$	$\hat{t}_5$
Ests	0.4562	0.2553	0.7034	0.0642	-0.0199	-0.0215	0.0445	0.0531
SE	0.0089	0.0042	0.0198	0.0155	0.0073	0.0065	0.0143	0.0236
$\log\left(L(\hat{\boldsymbol{\theta}})\right)$	2405.973							

It can be seen that the optimal values are really close to the true parameters, particularly for σ_f, λ and κ_1. The times differences, despite small, are also well estimated by the model, with all the true values contained within the confidence intervals calculated with the standard errors:

- $t_1 = 0.05 \in (0.0332, 0.0952)$
- $t_2 = -0.01 \in (-0.0345, -0.0053)$
- $t_3 = -0.02 \in (-0.0345, -0.0085)$
- $t_4 = 0.04 \in (0.0159, 0.0731)$
- $t_5 = 0.01 \in (0.0059, 0.1003)$

Different topologies were also tried, resulting in lower maximal log-likelihood values.

C.1.2 Three-dimensional curves simulation

A set of three-dimensional curves of 15 equally spaced points (arc-length from 0 to 1) was simulated using hyper-parameters $\boldsymbol{\theta} = (\sigma_f, \lambda, \mu, \kappa_1, \kappa_2, \kappa_3) = (1, 0.3, 1, -0.5, 0.17, 0.3)$, and the same times as before, for nodes A to E: (0.04, 0.05, 0.02, 0.06, 0.07). The curves are shown in Figure C.3. The values for the x coordinate are shown in blue, for y in green and for z in red.

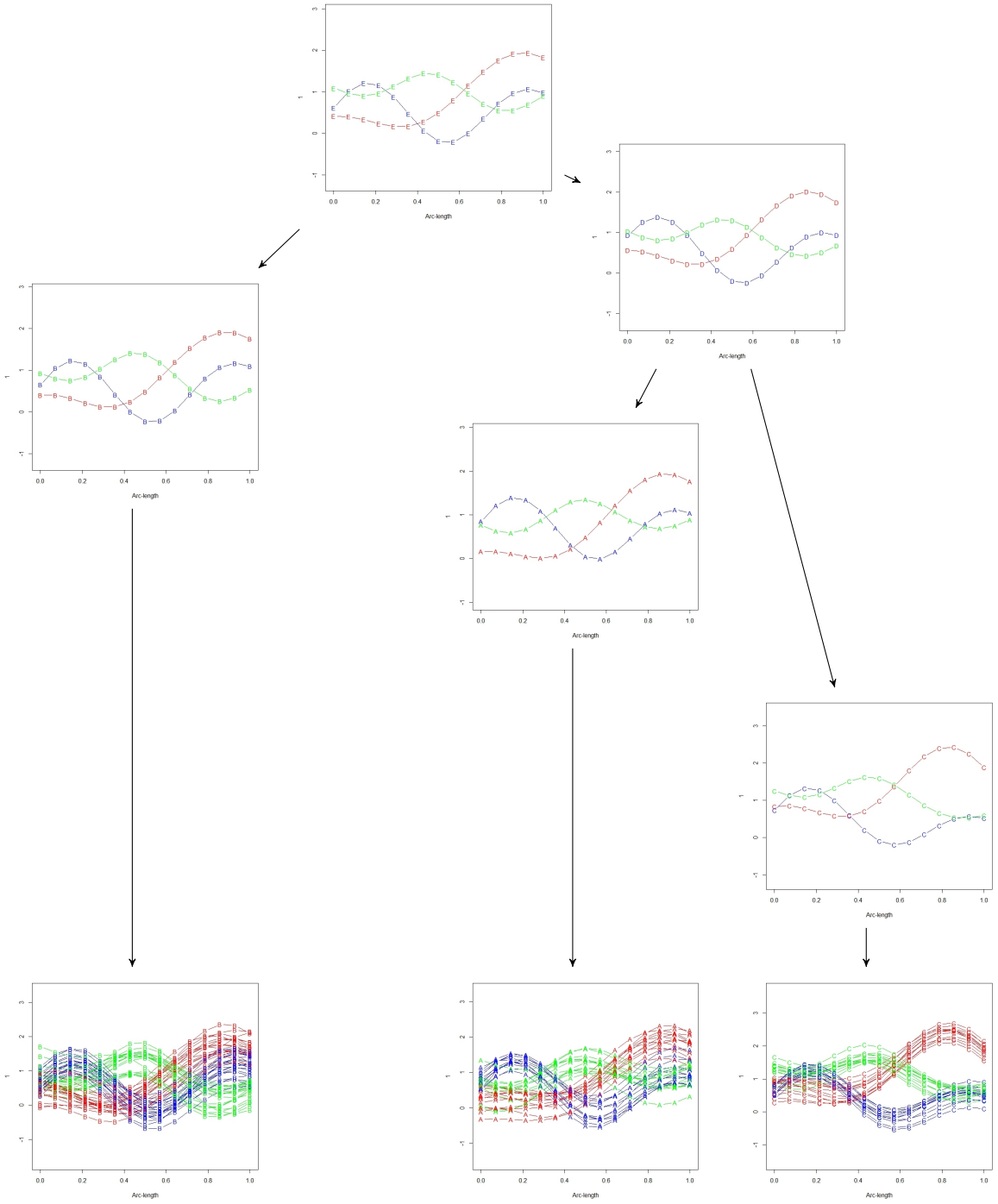


FIGURE C.3: Simulated 3D curves.

Specifying also the same time differences as before, the set of hyper-parameters is $\boldsymbol{\theta} = (\sigma_f, \lambda, \kappa_1, \kappa_2, \kappa_3, t_1, t_2, t_3, t_4, t_5) = (1, 0.3, -0.5, 0.17, 0.3, 0.05, -0.01, -0.02, 0.04, 0.01)$, with $\log\big(L(\boldsymbol{\theta})\big) = 3163.536$.

$\hat{\boldsymbol{\theta}}$	$\hat{\sigma}_{f3D}$	$\hat{\lambda}_{3D}$	$\hat{\kappa}_{13D}$	$\hat{\kappa}_{23D}$	$\hat{\kappa}_{33D}$			
Ests	0.4951	0.2557	-0.4024	0.2245	0.2929			
SE	0.0079	0.0036	0.0325	0.0333	0.0341			
				$\hat{t}_1$	$\hat{t}_2$	$\hat{t}_3$	$\hat{t}_4$	$\hat{t}_5$
			Ests	0.0521	-0.0066	-0.0263	0.0455	0.0003
			SE	0.0107	0.0042	0.0063	0.0097	1.99e-06
$\log\left(L(\hat{\boldsymbol{\theta}})\right)$	3236.701							

In this case, σ_f is underestimated, however λ and the correlations between the coordinates are close to the true values. The confidence intervals for the time differences can also be calculated:

- $t_1 = 0.05 \in (0.0307, 0.0735)$
- $t_2 = -0.01 \in (-0.015, 0.0018)$
- $t_3 = -0.02 \in (-0.0389, -0.0137)$
- $t_4 = 0.04 \in (0.0261, 0.0649)$
- $t_5 = 0.01 \notin (0.00029, 0.00030)$

All the true values are contained in the intervals, except for the last difference, t_5. This is because the curves at node E and D are so similar the model interprets there is no difference between them, and hence, the estimated branch length is really small.

Bibliography

Aldrich, J. (1997). R. A. Fisher and the making of maximum likelihood 1912-1922. *Statist. Sci.*, 12(3):162–176.

Allman, E. S. and Rhodes, J. A. (2006). The identifiability of tree topology for phylogenetic models, including covarion and mixture models. *Journal of Computational Biology*, 13(5):1101–1113.

Bell, A., Lo, T.-W. R., Brown, D., Bowman, A., Siebert, J., Simmons, D., Millett, D., and Ayoub, A. (2014). Three-dimensional assessment of facial appearance following surgical repair of unilateral cleft lip and palate. *Cleft Palate-Craniofacial Journal*, 51(4):462–471.

Bookstein, F. L. (1997). *Morphometric Tools for Landmark Data: Geometry and Biology*. Cambridge University Press.

Bowman, A. W., Katina, S., Smith, J., and Brown, D. (2015). Anatomical curve identification. *Computational Statistics and Data Analysis*, 86:52–64.

Burnham, K. P. and Anderson, D. R. (2003). *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media.

Burnham, K. P. and Anderson, D. R. (2004). Multimodel inference: Understanding aic and bic in model selection. *Sociological Methods & Research*, 33(2):261–304.

Chen, B. and Hong, Y. (2012). Testing for the markov property in time series. *Econometric Theory*, 28(1):130–178.

Claeskens, G., Hjort, N. L., et al. (2008). *Model selection and model averaging*, volume 330. Cambridge University Press Cambridge.

Collins Dictionary (2017). 'Evolution'. Available at https://www.collinsdictionary.com/dictionary/english/evolution Accessed: 1st September 2017.

Constantinou, P., Kokoszka, P., and Reimherr, M. (2017). Testing separability of space-time functional processes. *Biometrika*, 104(2):425–437.

Cox, M. G. (1972). The numerical evaluation of b-splines. *IMA Journal of Applied Mathematics*, 10(2):134.

Cox, M. G. (1982). *Practical Spline Approximation*, pages 79–112. Springer Berlin Heidelberg, Berlin, Heidelberg.

Darwin, C. (1859). *On the Origin of Species by Means of Natural Selection.* London: J. Murray.

Darwin, C. (1872). The expression of the emotions in man and animals. *London, UK: John Murray.*

De Boor, C. (1978). *A practical guide to splines*, volume 27. Springer-Verlag New York.

Dhar, A. and Minin, V. N. (2015). Maximum likelihood methods for phylogenetic inference.

Dimensional Imaging Ltd (2017). ©*Di4D* and ©*Di3D*. http://www.di4d.com/.

Dryden, I. and Mardia, K. (1998). *Statistical Shape Analysis.* John Wiley.

Dryden, I. and Mardia, K. (2016). *Statistical Shape Analysis: With Applications in R.* John Wiley & Sons.

Ekman, P. and Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2):124.

Ekman, P., Levenson, R., and Friesen, W. (1983). Autonomic nervous system activity distinguishes among emotions. *Science*, 221(4616):1208–1210.

Ekman, P. and Rosenberg, E. (1997). *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression using the Facial Action Coding System (FACS).* Oxford University Press, USA.

Eleftheriadis, S. (2016). *Gaussian Processes for Modeling of Facial Expressions.* PhD thesis.

Felsenstein, J. (1989). PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics*, 5:164–166.

Felsenstein, J. (2004). *Inferring Phylogenies*. Sinauer Associates, Inc.

Fragopanagos, N. and Taylor, J. (2005). Emotion recognition in human-computer interaction. *Neural Networks*, 18(4):389 – 405.

Franciscus, R. G. and Trinkaus, E. (1988). Nasal morphology and the emergence of Homo erectus. *American Journal of Physical Anthropology*, 75(4):517–527.

Fuentes, M. (2006). Testing for separability of spatial–temporal covariance functions. *Journal of statistical planning and inference*, 136(2):447–466.

Gaebel, W. and Wölwer, W. (1992). Facial expression and emotional face recognition in schizophrenia and depression. *European Archives of Psychiatry and Clinical Neuroscience*, 242(1):46–52.

Gibbs, M. (1998). *Bayesian Gaussian Processes for Regression and Classification.* PhD thesis, University of Cambridge, England.

Gower, J. C. (1975). Generalized procrustes analysis. *Psychometrika*, 40(1):33–51.

Halder, A., Rakshit, P., Chakraborty, A., Konar, A., and Janarthanan, R. (2011). *Emotion Recognition from the Lip-Contour of a Subject Using Artificial Bee Colony Optimization Algorithm*, pages 610–617. Springer Berlin Heidelberg, Berlin, Heidelberg.

Hastie, T. and Stuetzle, W. (1989). Principal curves. *Journal of the American Statistical Association*, 84(406):502–516.

Ho, L. S. T. and An, C. (2013). Asymptotic theory with hierarchical autocorrelation: Ornstein-uhlenbeck tree models. *Annals of Statistics*, 41(2):957–981.

Hoch, M., F. G. and Girod, B. (1994). Modeling and animation of facial expressions based on b-splines. *Visual Computer*, 11(2):87–95.

Huson, D. H., Rupp, R., and Scornavacca, C. (2010). *Phylogenetic Networks: Concepts, Algorithms and Applications.* Cambridge University Press.

Jack, R. E. and Schyns, P. G. (2015). The human face as a dynamic tool for social communication. *Current Biology*, 25(14):R621–R634.

Jones, N. S. and Moriarty, J. (2013). Evolutionary inference for function-valued traits: Gaussian process regression on phylogenies. *Journal of the Royal Society Interface*, 10(78):20120616.

Kakumanu, P., Makrogiannis, S., and Bourbakis, N. (2007). A survey of skin-color modeling and detection methods. *Pattern Recognition*, 40(3):1106–1122.

Kau, C. H., Richmond, S., Incrapera, A., English, J., and Xia, J. J. (2007). Three-dimensional surface acquisition systems for the study of facial morphology and their application to maxillofacial surgery. *International Journal of Medical Robotics and Computer Assisted Surgery*, 3(2):97–110.

Kendall, D. (1984). Shape manifolds, procrustean metrics, and complex projective spaces. *Bulletin of the London Mathematical Society*, 16(2):81–121.

Kimmel, R., Kiryati, N., and Bruckstein, A. M. (1997). Analyzing and Synthesizing Images by Evolving Curves with the Osher-Sethian Method. *International Journal of Computer Vision*, 24(1):37–55.

Koenderink, J. and van Doorn, A. (1992). Surface shape and curvature scales. *Image and Vision Computing*, 10(8):557–565.

Kristof, W. and Wingersky, B. (1971). Generalization of the orthogonal procrustes rotation procedure for more than two matrices. *Proceedings of the Annual Convention of the American Psychological Association*, 6(1):89–90.

Lancaster, P. and Šalkauskas, K. (1986). *Curve and Surface Fitting: an Introduction*. Computational Mathematics and Applications. Academic Press.

Little, A. C., Jones, B. C., and DeBruine, L. M. (2011). The many faces of research on face perception. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 366(1571):1634–1637.

Lowther, J. and Shene, C.-K. (2003). Teaching b-splines is not difficult! *ACM SIGCSE Bulletin*, 35(1):381–385.

Macaulay, V., Hill, C., Achilli, A., Rengo, C., Clarke, D., Meehan, W., Blackburn, J., Semino, O., Scozzari, R., Cruciani, F., et al. (2005). Single, rapid coastal

settlement of Asia revealed by analysis of complete mitochondrial genomes. *Science*, 308(5724):1034–1036.

Mardia, K. V. and Marshall, R. J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, 71(1):135–146.

McColl, J. H. (2004). *Multivariate Probability*. John Wiley and Sons.

McNeil, K. (2012). Analysis of Three-Dimensional Facial Shape. Master's thesis, University of Glasgow.

Millar, K., Bell, A., Bowman, A., Brown, D., Lo, T.-W., Siebert, P., Simmons, D., and Ayoub, A. (2013). Psychological status as a function of residual scarring and facial asymmetry after surgical repair of cleft lip and palate. *Cleft Palate-Craniofacial Journal*, 50(2):150–157.

Mitchell, M. W., Genton, M. G., and Gumpertz, M. L. (2005). Testing for separability of space–time covariances. *Environmetrics*, 16(8):819–831.

Mladina, R., Skitareli, N., and Vukovi, K. (2009). Why do humans have such a prominent nose? The final result of phylogenesis: A significant reduction of the splanchocranium on account of the neurocranium. *Medical Hypotheses*, 73(3):280–283.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Myung, J. I. and Navarro, D. J. (2004). Information matrix. *Encyclopedia of Statistics in Behavioral Science*.

Noback, M. L., Harvati, K., and Spoor, F. (2011). Climate-related variation of the human nasal cavity. *American Journal of Physical Anthropology*, 145(4):599–614.

Ozertem, U. and Erdogmus, D. (2011). Locally defined principal curves and surfaces. *Journal of Machine Learning Research*, 12:1249–1286.

Paciorek, C. J. (2003). *Nonstationary Gaussian processes for regression and spatial modelling*. PhD thesis, Carnegie Mellon University, Pittsburgh, Pennsylvania.

Page, R. and Holmes, E. (1998). Molecular evolution: a phylogenetic approach. *Blackwell Science Oxford*.

Pawitan, Y. (2001). *In All Likelihood: Statistical Modelling and Inference using Likelihood*. Oxford University Press.

Piegl, L. and Tiller, W. (1987). Curve and surface constructions using rational b-splines. *Computer-Aided Design*, 19(9):485–498.

Pietilainen, V. (2010). Approximations for integration over the hyperparameters in gaussian processes.

Posada, D. and Buckley, T. R. (2004). Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. *Systematic Biology*, 53(5):793–808.

Press, W. H., Teukolsky, S., Vetterling, W., and Flannery, B. (1992). *Numerical Recipes in C (2nd edn.): The Art of Scientific Computing*. Cambridge University Press, New York, NY, USA.

Pressley, A. (2001). *Gauss's Theorema Egregium*, pages 229–246. Springer London, London.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological methodology*, pages 111–163.

Rasmussen, C. (1996). *Evaluation of Gaussian processes and other methods for non-linear regression*. PhD thesis, University of Toronto.

Rasmussen, C. (2004). Gaussian processes in machine learning. In Bousquet, O., von Luxburg, U., and Ratsch, G., editors, *Advanced Lectures on Machine Learning*, volume 3176 of *Lecture Notes in Computer Science*, pages 63–71. Springer Berlin Heidelberg.

Rasmussen, C. and Williams, C. (2006). Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning).

Robert, C. (2001). *The Bayesian Choice: from Decision-theoretic Foundations to Computational Implementation*. Springer Texts in Statistics. Springer, New York.

Roberts, A. (2001). Curvature attributes and their application to 3d interpreted horizons. *First Break*, 19(2):85–100.

Rudovic, O. (2013). *Machine Learning Techniques for Automated Analysis of Facial Expressions*. PhD thesis, Imperial College London.

Sánchez, M. U. R., Matas, J., and Kittler, J. (1997). *Statistical chromaticity models for lip tracking with B-splines*, pages 69–76. Springer Berlin Heidelberg, Berlin, Heidelberg.

Shahriari, B., Swersky, K., Wang, Z., Adams, R. P., and de Freitas, N. (2016). Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104(1):148–175.

Singh, G. B. (2015). *Distance Based Methods*, pages 253–260. Springer International Publishing, Cham.

Thorpe, J. P. (1982). The molecular clock hypothesis: biochemical evolution, genetic differentiation and systematics. *Annual Review of Ecology and Systematics*, 13(1):139–168.

Tian, Y.-I., Kanade, T., and Cohn, J. F. (2001). Recognizing action units for facial expression analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 23(2):97–115.

Vittert, L. (2015). *Facial Shape Analysis*. PhD thesis, University of Glasgow.

Vittert, L., Bowman, A., and Katina, S. (2017). Statistical models for manifold data with applications to the human face. *Preprint arXiv:1701.07328*.

Yang, Z. (2006). *Computational Molecular Evolution*. Oxford University Press.

Zaidi, A. A., Mattern, B. C., Claes, P., McEcoy, B., Hughes, C., and Shriver, M. D. (2017). Investigating the case of human nose shape and climate adaptation. *PLOS Genetics*, 13(3):1–31.