Evolutionary Inference for Functional Data: Using Gaussian Processes on Phylogenies of Functional Data Objects

Mhairi Kerr

A Dissertation Submitted to the
University of Glasgow
for the degree of
Master of Science

School of Mathematics & Statistics

December 2011

Abstract

This thesis explores the use of phylogenetics and functional data analysis for the analysis of continuous ancestral data such as continuous curves. Gaussian processes (GPs) are placed on phylogenies in order to perform evolutionary inferences on the functional data objects. The mean and covariance functions of the GP model the relationships between different states on the phylogeny. The functional data objects are completely described by the spatial and temporal parameters within the covariance functions, allowing inferences to be made, for example, by the method of maximum likelihood estimation. Inferences are successfully made on known phylogenies, phylogenies with missing ancestral data and on phylogenies of unknown topology. This work is potentially useful for those wanting to compute evolutionary inferences on continuous ancestral data, for which phylogenetic GPs are shown to be an efficient and promising tool.

Acknowledgements

I would like to begin by thanking my supervisor Dr Vincent Macaulay for his guidance and support throughout the last year. I would also like to acknowledge the Engineering and Physical Sciences Research Council (EPSRC) which provided the financial support for this research.

Most importantly I would like to thank my family for the support they have given me throughout my education. Lastly I would like to thank my fellow Masters students who have helped make it a fun year and not all about work.

Contents

List of Tables

List of Figures

Chapter 1

Background

1.1 Background

Functional data analysis (FDA) and phylogenetics will be used in the analysis of continuous ancestral data. The ancestral data to be analysed will take the form of the 'simplest' continuous shape that can be easily created: continuous curves, i.e., functional data. Phylogenetics involves the reconstruction of the evolutionary past in the form of a tree structure (phylogeny), while FDA involves the analysis of information on curves and contours. The aim of this thesis is to explore the effect of placing Gaussian processes (GPs) on phylogenies of functional data (Figure 1.1), in order to perform evolutionary inference on the functional data curves. We wish to learn about the spatial and temporal aspects of these curves, in particular about the covariance function for time-space GPs.

Currently most phylogenetic inference is from discrete data ('characters') [1], [21], [25], [42], as opposed to continuous data, where only a small number of investigations have been carried out [9], [10], [29]. There are many discussions as to whether continuous characters are the most suitable way of making inferences about unobserved shapes [18], [24]. The motivation for this thesis originates from a paper by Felsenstein [9], that presents a method for comparative studies of continuous real-valued phenotypes. Felsenstein

tackles a common mistake that is regularly made in comparative studies of phenotypes, namely to assume that species which are part of a hierarchically structured phylogeny can be regarded as being drawn independently from the same distribution. This faulty assumption of independence results, for example, in the overstatement of the significance of hypothesis tests. Felsenstein's method to overcome this problem is to correct for branching phylogenies. His method requires that the phylogeny is known, and the phenotypes are modelled by Brownian motion. The method Felsenstein proposes is possibly the simplest GP model, which allows the random continuous evolution of the continuous phenotypes to be modelled. This paper provides an excellent starting point for the discussion of key issues for inference on phylogenetic GPs. Martins and Hansen [26] built on Felsenstein's GP model by introducing generalised least squares (GLS), which are special cases of GPs. Their paper demonstrates a general approach for the analysis of comparative evolutionary data by the use of GLS. The authors demonstrate how this approach can be used to deal with many single-character evolution questions, e.g., the rate of phenotypic evolution or the ancestral state of a character. They also show how GLS can be designed to include different covariance functions depending on the evolutionary assumptions made, and how parameter estimates can be easily made by GLS estimates. Most importantly, the authors demonstrate that for each situation a new phylogenetic comparative method can be created by the use of a different covariance matrix.

The use of phylogenetics and GPs for evolutionary inference has very recently been explored by Jones and Moriarty [20]. Their work is also related and built on that of Felsenstein [9], which gives a method for comparative studies of real-valued traits corrected for a phylogeny. In their paper Jones and Moriarty look at combining concepts from phylogenetics with GPs, in order to allow evolutionary inference for function-valued traits that are correlated through a phylogeny. By using GPs a non-parametric Bayesian model for such data is created. From this model ancestral function-valued traits

can be inferred, rates of evolution across a phylogeny can be compared, or the most likely phylogenies consistent with observed data can be identified. The ideas and concepts raised in this paper are very similar to those that will be seen throughout this thesis.

1.2 Statistical Methods

Functional Data Analysis and Gaussian Processes: Curves are examples of functions. For this reason data on curves are commonly referred to as 'functional data' and the analysis of such data is referred to as 'functional data analysis' (FDA). FDA is a relatively new field, where data from the continuous curves are not considered as single discrete or continuous points in a finite or Euclidean space, but rather each datum is treated as a complete continuous function embedded in a function space. In the words of Ramsay and Silverman [45], "Though the individual methods and techniques are important, functional data analysis is also a general way of thinking, where the basic unit of information is the entire observed function rather than a string of numbers". Although FDA is used in many fields, from growth curve estimation to weather analysis, methods to treat functional data as a realisation of a process defined simultaneously over a function space and a phylogeny have only begun to be explored. Papers recently published by The Functional Phylogenies Group [14], and Jones and Moriarty [20] describe methods in which functional data can be defined simultaneously over a function space and a phylogeny. Also discussed in these papers is how evolutionary inference can be computed on function-valued data that are correlated through a phylogeny.

Currently, most phylogenetic inference for shape data are via morphology [25], [37]. Morphology involves a small number of discrete or continuous variables being used to summarise the shape of interest. However this method reduces the dimension of the data and hence loses information, while FDA

allows one to make inferences directly on the curves themselves. Methods for the analysis of functional data can bear a strong resemblance to those for conventional multivariate data [30], including non-linear and linear regression, principal components analysis and many other techniques. In this thesis, a non-parametric approach to FDA will be adopted [12]. It will be assumed that the functional data points are a sample from a GP. One property of these processes is that they are completely specified by their mean and covariance functions; it therefore follows that the functional data curves will be uniquely specified by these functions also. By assuming the functional data to be Gaussian, the time evolution of any function on the spatial domain can, by the specification of the mean and covariance functions of the GP, be represented as a function on the space-time domain. This assumption of the data being Gaussian allows the FDA to become more tractable [12], [30]. The FDA will involve the estimation of the mean and covariance functions that encode the curves. As these functions contain space and time parameters, that define characteristics of a curve in space and time, then the functions themselves can be learned about by the estimation of these parameters, and this can be carried out using the inference method of maximum likelihood or by a corresponding Bayesian approach. To make inferences on phylogenies on which GPs have been placed, the existing literature on machine learning for inference with GPs [32] will be exploited. This method of using GPs to compute inferences on phylogenies of functional data is extremely similar to the methods used by Jones and Moriarty [20], where the authors also used the covariance function of the GP to encode the structure of the functional data.

Gaussian Processes (GPs): A GP model is a stochastic process (or random process) whose output consists of random values associated with every point $\mathbf{x}$ in a range of time and/or space, such that each random variable $f(\mathbf{x})$ has a normal distribution, and any finite number of values of the output for dif-

ferent $\mathbf{x}$s has a multivariate normal distribution.

The study of GPs is a building block in the theory of stochastic processes and they have recently began to appear in statistical machine learning [32]. Prior to this, GPs under the guise of 'kriging' have also been applied extensively in geostatistics [4].

As already stated, one property of GPs is that they are completely specified by their mean and covariance functions. The mean function $m(\mathbf{x})$ and the covariance function $k(\mathbf{x}, \mathbf{x}')$ are defined by

$$m(\mathbf{x}) = \mathbb{E}\big[f(\mathbf{x})\big], \tag{1.1}$$

$$k(\mathbf{x}, \mathbf{x}') = \mathbb{E}\Big[\big(f(\mathbf{x}) - m(\mathbf{x})\big)\big(f(\mathbf{x}') - m(\mathbf{x}')\big)\Big] \tag{1.2}$$

and the GP is written as:

$$f(\mathbf{x}) \sim GP\big(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\big). \tag{1.3}$$

As a GP is defined as a collection of normal random variables, this implies that GPs possess the marginalisation property. This property means, e.g., that if a GP specifies $(y_1, y_2) \sim N(\boldsymbol{\mu}, \sigma)$, where $\boldsymbol{\mu}$ is the 2-component mean vector and σ is the 2×2 covariance matrix, then $y_1 \sim N(\mu_1, \sigma_{11})$, i.e., all marginal distributions of a GP are normally distributed.

GPs also have many features which make them relatively convenient to anaylse by Bayesian methods [15], [32], although in this thesis, a likelihood approach will be taken.

1.3 Understanding Phylogenies

Phylogenetics involves the reconstruction of the evolutionary past in the form of tree structures, where these structures are commonly known as phylogenies [11]. Phylogenies are similar to family trees, but unlike these trees, the branch lengths on the phylogenies are proportional to the predicted evolutionary time between organisms (extant or extinct in the original biological context),

or more abstractly, between functional data objects, in the application here. As the time to evolve from one curve to another increases, the branch length between these curves in the phylogeny also increases.

Phylogenies are either rooted or unrooted. Rooted trees have a specified common ancestor, whilst an unrooted tree does not. In the rooted tree, the curve at the top of the tree is the common ancestor (ancestral node) of all descendants and the curves at the tips (leaves or terminal nodes) of the tree are therefore the descendants of that common ancestor. Note that time is moving forward as you move from the root of the tree to the tips.

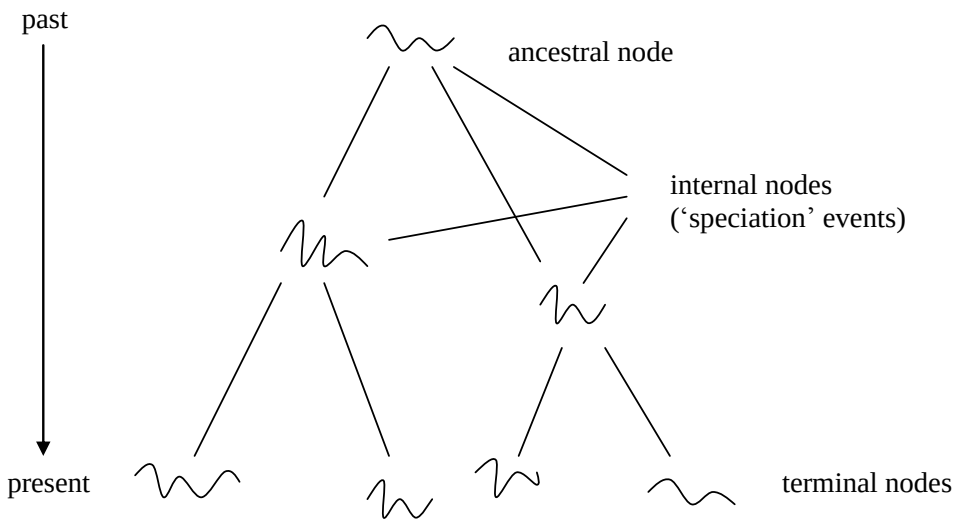

Figure 1.1: Example of a Phylogeny of Functional Data Curves

A curve which is evolving over time can split to create two closely related curves which is represented on the phylogeny by the branch splitting (internal nodes). Such an event is known as speciation after the most common application of phylogenies to the speciation of organisms, generating the tree of life. Importantly when a curve splits like this, the two derived curves continue to evolve over time, but independently of each other.

1.4 Literature Review

Phylogenetic Inference: The maximum likelihood approach to the estimation of evolutionary trees is well established in the phylogenetic field and the literature is extensive in this area. One of the best known papers on the maximum likelihood approach is that by Felsenstein [8] which details the application of maximum likelihood techniques to the estimation of evolutionary trees. The likelihood approach has several advantages over the traditional parsimony methods [3], [5], which aims to find the shortest tree consistent with the observed data at its tips. As Felsenstein explains the parsimony method is a valid statistical method if the amount of change in a specific evolutionary timeframe is small. However most data involves moderate to large amounts of change, and in these cases the parsimony method runs into problems since parallel evolution in nearby branches of the tree will not be allowed for. In a previous paper by Felsenstein [7], he showed that when the amount of "evolutionary change in different lineages are sufficiently unequal", then the parsimony method makes estimates on evolutionary trees that are inconsistent.

To estimate evolutionary trees by maximum likelihood estimation, Felsenstein first constructs an algorithm for computing the likelihood of a given tree, based on a previous algorithm constructed by Felsenstein [6]. Once the likelihood can be computed for a given evolutionary tree, Felsenstein is then left with the exercise of finding the maximum likelihood tree. As Felsenstein explains, one problem is finding values of the lengths of the branches between nodes which maximises the likelihood of the tree given a particular topology. Potentially this could be done by direct search but this would require the likelihood being calculated for each set of branch lengths, which is not feasible due to the possible number of computations. Instead Felsenstein constructs an algorithm which changes one branch length at a time, which is altered to that value which produces the largest likelihood. A further search

over different tree topologies is also performed. Overall this paper provides excellent groundwork for using the method of maximum likelihood estimation to estimate an evolutionary tree.

More recently the Bayesian inference method has become popular in phylogenetics [23], [27], [31]. As described by Hulsenbeck and Ronquist [19], "Bayesian inference has several advantages over other methods of phylogenetic inference, including easy interpretation of results, the ability to incorporate prior information (if such information is available), and some computational advantages" [22]. Bayesian inferences of phylogenies should be based on the posterior probabilities of phylogenetic trees. The method Hulsenbeck and Ronquist suggest to be used to approximate these posterior probabilities is Markov-Chain Monte Carlo (MCMC) [16], [17], [28]. The MCMC method is explained in some detail in their paper.

Yang and Rannala [43] present an improved Bayesian method for estimating phylogenetic trees specifically from DNA sequence data. The method used in their paper is an improvement on a method the authors themselves previously devised [31]. There, Rannala and Yang used a birth-death process to "specify the prior distribution of phylogenetic trees and ancestral speciation times, and a Markov process was used to model nucleotide substitution". Using maximum likelihood, the parameters of the birth-death process and the substitution model were estimated. The posterior probabilities of trees were then evaluated by replacing the true parameters with the estimated parameters, a so-called empirical Bayes approach. Two sets of DNA sequences were analysed in this paper where the Bayesian method generated the same best tree as the already established maximum likelihood method. However the authors found that their method is adequate for analysing data sets which include a small number of species only, as the calculations involve a sum over all tree topologies, and, as the number of species increases, the number of topologies increase rapidly.

In the paper of Yang and Rannala [43], the authors describe a method practical for analysing a larger number of species. They begin by using Monte Carlo integration "to evaluate more efficiently the integral over the ancestral speciation times for a given tree", and, to avoid having to sum over all possible topologies, the posterior probabilities of the trees are evaluated using the MCMC method. Yang and Rannala also make two improvements to the model for the prior distribution of trees and speciation times. The first is to consider species sampling by biologists. This is done as the original birth-death process tends to create trees with longer internal branches than external branches, and so by taking species sampling into account the internal branch lengths can be reduced which results in a more realistic prior distribution of trees. The second change made is that the "birth and death rates of the prior distribution are treated as random variables and eliminated by integration". The authors state that this is "expected to make the posterior probabilities more robust to violations of assumptions about the birth-death prior".

The authors trial this new method on the phylogeny of primates, where the topology that is considered correct is found with relative ease, despite the increase in the number of species. For many data sets, most trees are unlikely and therefore have very small posterior probabilities. The efficiency of this method could therefore be improved by preferentially electing trees with larger posterior probabilities rather than each tree being given an equal probability.

These papers touch on the three well-established methods for estimating phylogenies: parsimony, maximum likelihood and Bayesian inference. The methods that will be applied in this thesis are maximum likelihood and Bayesian inference.

Gaussian Processes: GPs have been extensively applied in many fields to model spatial and temporal data which can be illustrated with a couple of

examples. Ferris, Hähnel and Fox [13] implemented GPs in their research exactly for this purpose. This paper is concerned with using wireless signal-strength information to estimate the location of mobile devices or robots. As well as their ability to model spatial data, GPs are being applied to this problem due to their ability to produce prediction models for locations at which no data are available. Schwaighofer, Grigoras, Tresp and Hoffmann [35] have tackled the use of GPs for signal-strength-based location estimation previously.

The authors detail how GPs can be used to model signal-strength measurements. The mean of the GP model is set to zero and so it is the covariance function only that models the signal-strength data. The covariance function used is the squared exponential, which contains two parameters, one is the signal variance and the other is the spatial correlation between measurements. This GP model is taken to be the prior, and from that the authors show how easily the GP posterior predictive distribution can be obtained for locations at which no data are available. The authors are also interested in obtaining estimates for the parameters within the model. By computing these predictions and parameter estimates, the authors can begin to identify the location of the mobile device or robot. The authors propose that spatial GP models can be successfully applied to various robotic problems and minor sensor network problems.

GPs also have the ability to model spatial and temporal information together. Zhao, Fu and Liu [44] present a problem which incorporates both spatial and temporal data, where the objective is the estimation of human pose via motion tracking. The paper begins by explaining some of the previous approaches used to tackle this problem, such as the discriminative approach [36] and nearest neighbour approach [38]. GPs within the discriminative [39] and generative [40] frameworks have also been applied to human pose/motion estimation in the past.

Regardless of the approach used, the objective is to model the direct map-

ping from visual observations to well-defined human pose configurations. It is explained by the authors that in order to "build the mapping, existing techniques usually involve a large set of training samples in the learning process", but these techniques are known to be limited in their ability to model this data which follows a multimodal distribution (a continuous probability distribution with two or more modes). This is why the authors introduced GPs to this problem; rather than modelling the data with a full GP they instead choose a mixture of local GP experts, which incorporate spatial and temporal information, to model different local neighbourhoods as this helps to handle the multimodality. To obtain these local GP experts, the input space has to be "divided into different regions by a gating network, each of which is dominated by a specific GP expert". It is the covariance function within the GPs that are localised to adapt to the different regions.

To make human pose estimations, the authors first build the temporal GP experts on the input space, and then the temporal and spatial experts are integrated into a hybrid system to make predictions about human pose (details are reported within the paper). Zhao, Fu and Liu tested this method on two databases, where the results have "validated the efficiency of the proposed model by achieving accurate human motion-tracking results". The authors also claim that the model's adaption to other scenarios is straightforward.

These papers illustrate how GPs can be easily adapted, by a change in the covariance function, to model spatial and/or temporal data. It is this property that has led to GPs becoming increasing popular, and why they are appearing in many new research fields.

1.5 Data Collection

To establish the feasibility of using phylogenetic inference on functional data, two types of data will be used, namely, simulated and controlled experimental.

- Simulated data. For each covariance function considered, samples will be simulated from corresponding phylogenetic GPs and the regression, ancestor and phylogeny problems will be attempted.

- Controlled experimental data. A spatial Chinese whispers method will be used to create curves with the aim being to produce the simplest experimental evolutionary system for continuous data. The creation of these curves begins with a single-valued ancestral curve being drawn and removed from sight, and then either one or two copies of that curve are then made from memory. The evolutionary system is created by regarding these copied curves to have evolved from the ancestral curve. These copies can then themselves be removed from sight and redrawn, and this process can be repeated as many times as required. In this way, a phylogeny of evolving single-valued curves is generated, where a branch splitting (speciation event) in the phylogeny is created by making two copies of the splitting curve. The data was generated on an iPad by Dr Nick Jones (U. Oxford, pers.comm:) and the curves extracted as a discrete set of points, using a script written in Matlab by Dr Jones. Note that we are only interested in modelling curves that can be written as $(\mathbf{x}, \mathrm{f}(\mathbf{x}))$ in this thesis, and so Dr Jones was instructed to create curves of this kind only, i.e., curves void of loops and curves that cannot come back to the same "x" position.

  Five single lineages were created using this method with each lineage containing six curves in total (Appendix A). These curves are for the regression problem in Chapter 2. Simulated data were used to validate the ancestor and phylogeny problems of Chapters 3, 4 and 5.

1.6 Overview of Thesis

The remainder of this thesis will be laid out as follows; Chapters 2 and 3 present the regression problem. Chapter 2 details the simulation of evolv-

ing curves along a single lineage and how inferences can be made on such curves. Also included in this chapter are the inferences computed on the Chinese whispers experimental curves. Chapter 3 is an extension of the work in Chapter 2 and involves the simulation of a GP phylogeny on which inferences will be computed. Chapter 4 presents the ancestor problem which is to obtain posterior probability distributions for one or more unobserved ancestral states, when other data for the GP phylogeny is available. Chapter 5 details the phylogeny problem, namely to infer the tree topology, when only the descendant curves at the tips of the tree are known. Lastly Chapter 6 contains a discussion and the conclusions.

Chapter 2

Inference on Functional Data from a Single Lineage

The regression problem will be attempted in this chapter with the main objective being to fit a GP model to evolving Chinese whispers curves on a single lineage. Inferences will be computed over the single lineages to learn about the curves in space and time, i.e., the evolution of the curves. In order to do this an assumption will be made that the discrete-time evolution of a d-dimensional GP spatial function can be viewed as a sample from a $d+1$-dimensional GP in space and time. In our example, $d = 1$.

Each set of curves in Appendix A can be thought as representing a different single lineage (i.e., path through a phylogeny) of curves that have been extracted from some phylogeny. Figure 2.1 illustrates an example of a single lineage that has been extracted, where the lineage chosen is the right-hand branch; the curve at the top of the lineage is the ancestral curve of the phylogeny and the curve at the bottom is the descendant curve in the present time from this right-hand branch. In between the ancestor and descendant are the intermediate curves produced as the ancestral curve evolves into the descendant curve.

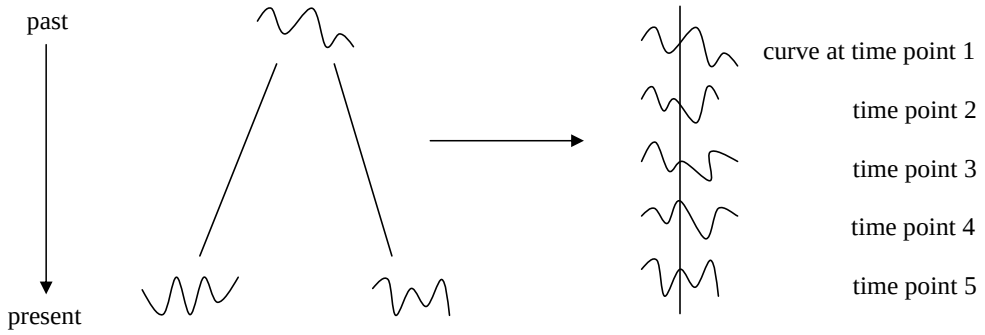

Figure 2.1: Example of a Single Lineage Taken From a Phylogeny

In order to compute evolutionary inferences over these lineages, GPs can be placed on them; of course, it is possible that the Chinese whispers curves are not well described by a GP. So it is natural first to explore what happens with data that have been simulated from a GP.

This chapter will be split into three sections; the first section will involve the simulation and inference of curves in space only, i.e., curves that have not evolved and so have no time factor. The second section will focus on the simulation and inference of evolving curves, in space and time, along a single lineage, and lastly the third section will involve the computing of inferences on the experimental Chinese whispers curves.

2.1 Simulation and Inference of Curves in Space

This section details how curves can be simulated in space from a GP model, and how inferences can then be computed from such curves. In order to do this the mean and covariance functions for the GP first have to be assigned. Once curves are simulated, inferences can be computed from them with the hope being to learn about the spatial structure of the curves (i.e., the parameters in the GP).

2.1.1 Spatial Covariance Matrix

As stated in section 1.2, GPs are completely specified by their mean and covariance functions (1.1) and (1.2). The curves created by the Chinese whispers method are smooth, and so it is important that the GP model to which they are fitted favours smooth functions. Therefore the mean and covariance functions must be chosen with care. It will be assumed throughout that the mean of the GP model is zero (i.e., there is no trend in the curves) and so the problem is somewhat simplified as it is then the covariance function only that encodes the structure of the curves. The spatial covariance function that will be used throughout is the squared exponential (SE):

$$K_x = cov\big[f(x_1), f(x_2)\big] = k(x_1, x_2) = \sigma^2 \exp\left[-\frac{1}{2\lambda^2}(x_1 - x_2)^2\right] \qquad (2.1)$$

Thus $f(x_1)$ and $f(x_2)$ are jointly Gaussian and the covariance between the outputs is written as a function of the inputs, x_1 and x_2.

If the GP is to be evaluated on a grid of x values, collected in the vector $\mathbf{x}$, then the covariance matrix K_x contains $k(x_i, x_j)$ in its (i, j)th position. The SE covariance matrix has σ^2 on the diagonal and is both symmetric and positive-definite, which are required properties for a covariance function. The GP model that the Chinese whispers curves will be fitted with in space therefore has the form

$$\mathbf{y} = f(\mathbf{x}) \sim GP(\mathbf{0}, K_x). \qquad (2.2)$$

It is the parameters σ and λ in the SE covariance function (2.1) that control how a function will appear spatially. The standard deviation σ can be described as the variability parameter (controlling the variability on the y-axis) and λ is the spatial smoothing parameter or length-scale parameter (controlling the smoothness on the x-axis). To explore how these parameters control the curve spatially, we first set the spatial smoothing parameter $\lambda = 0.3$ and the variability parameter $\sigma = 2$. A function on a fine grid of x values from a

GP model with these parameters is shown in Figure 2.2(a): the function produced can be described as being wiggly due to the length-scale being short. In contrast, if we set the spatial smoothing parameter $\lambda = 2$ with σ being kept the same, we expect to see functions like that in Figure 2.2(b), where the sample function is one that can be described as being very smooth due to the length-scale being considerably lengthened. In general, the smoother the curve, the larger λ.

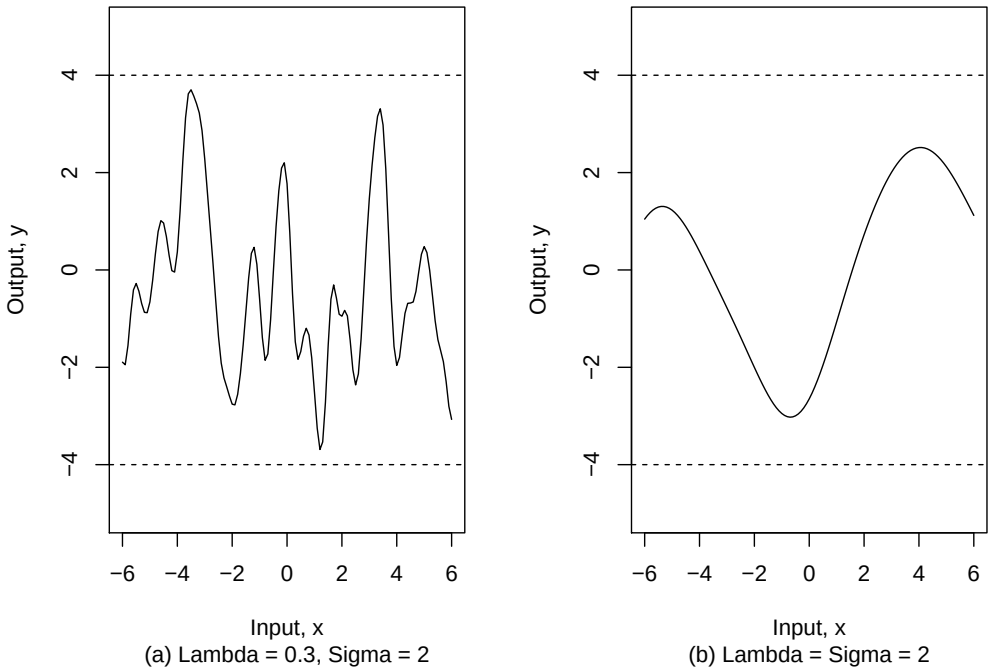

Figure 2.2: The Effect of Varying λ and σ

σ, on the other hand, has a relatively straightforward effect on the sample functions; an increase in σ results in an increase in the variability of the output and similarly a decrease in σ results in a decrease in the variability of the output. Those functions observed in Figure 2.2 were produced with $\sigma = 2$: the curves lie within 2σ of the mean zero, i.e., within $(-4, 4)$.

2.1.2 Simulation of Curves in Space

Having identified the GP model that will be fitted to the experimental Chinese curves in space as

$$f(\mathbf{x}) \sim GP(\mathbf{0}, K_x), \tag{2.3}$$

then we have the prior Gaussian distribution from which curves will be simulated from in space. The distribution over all curves (in space) is the joint multivariate normal distribution (as GPs follow the normal distribution as described in section 1.2) with mean $\boldsymbol{\mu}$ and covariance matrix K_x, denoted by $N_n(\boldsymbol{\mu}, K_x)$. It then follows that in order to simulate a curve in space, the multivariate normal distribution has to be simulated from.

The multivariate normal probability density function is

$$f_x(\mathbf{x}) = \frac{1}{(2\pi)^{\frac{n}{2}}(\det K_x)^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T K_x^{-1}(\mathbf{x}-\boldsymbol{\mu})\right], \tag{2.4}$$

where $\boldsymbol{\mu}$ is the mean vector (of length n) and K_x is the (symmetric, positive-definite) SE spatial covariance matrix (of size $n \times n$). Draws from this distribution can be made with the R [33] function *mvrnorm* within the MASS package [41].

Before the simulating from this distribution, values for the parameters σ and λ first have to be specified in the SE covariance function (2.1), and also a systematic grid of values over the x-axis must be chosen, $\mathbf{x}$ (input). Using these inputs, $\mathbf{x}$, the corresponding covariance matrix can be written out using (2.1). One can then simulate random Gaussian vectors $f(\mathbf{x})$ with this covariance function and plot the generated values as a function of the inputs. Figure 2.3 displays some examples of these simulated sample functions.

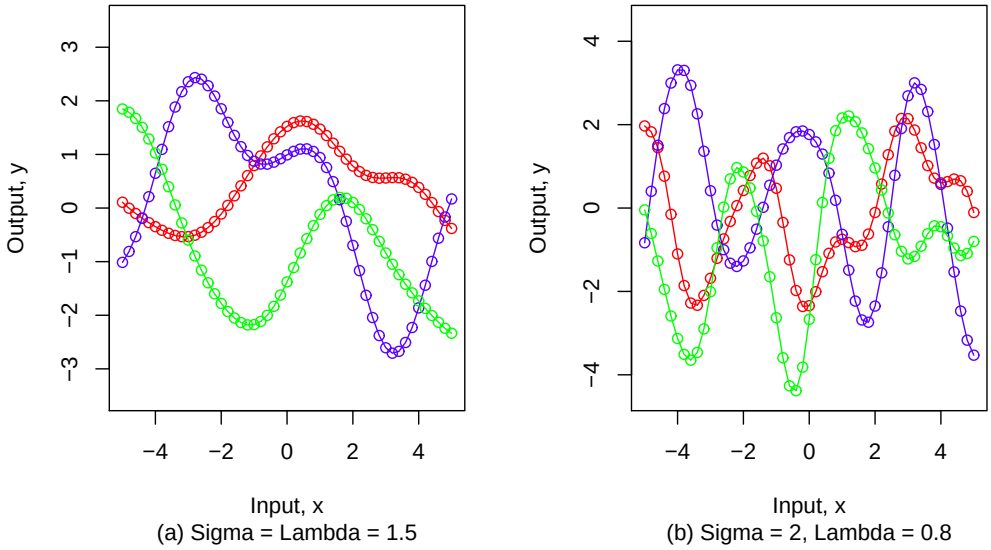

Figure 2.3: Spatial Functions Drawn at Random from a GP

Figure 2.3(a) illustrates three sample spatial functions drawn at random from a GP with $\sigma = \lambda = 1.5$, while Figure 2.3(b) similarly shows three sample spatial functions drawn at random from a GP with $\sigma = 2$ and $\lambda = 0.8$. For all functions the input vector, $\mathbf{x}$, consisted of 51 points ranging from -5 to 5 with an interval of 0.2. As expected of functions simulated from GP models with a SE covariance function, the outputs are randomly fluctuating, and, as required, the generated functions are relatively smooth.

2.1.3 Inferences on Curves Simulated in Space

To learn about these simulated curves in space then the spatial covariance function that encodes them has to be learned about. The SE covariance function contains two parameters, σ and λ, and by using the inference method of maximum likelihood estimation to provide estimates for these parameters the spatial structure of the curves can be determined.

The method of maximum likelihood estimation works by selecting values of the parameters that produce a distribution that gives the observed

data the greatest likelihood (i.e., the parameters that maximise the likelihood function). To obtain maximum likelihood estimates for σ and λ, the likelihood function needs to be determined, which is simply the probability of obtaining the observed data with the parameters set to particular values of σ and λ. As seen in section 2.1.2 these curves are simulated from a mulitvariate Gaussian, and so the likelihood function is simply the multivariate normal probability density function

$$L(\sigma, \lambda; \mathbf{y}) = \frac{1}{(2\pi)^{\frac{n}{2}} (\det K_x)^{\frac{1}{2}}} \exp\left[-\frac{1}{2}\mathbf{y}^T K_x^{-1}\mathbf{y}\right] \tag{2.5}$$

where n is the number of spatial points and y_i is the curve output corresponding to input x_i. Note that K_x is implicitly a function of σ and λ. For simplicity the logarithm of the likelihood function will be taken as it makes the problem mathematically simpler. The log-likelihood function therefore has the form:

$$l(\sigma, \lambda, ; \mathbf{y}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log(\det K_x) - \frac{1}{2}\mathbf{y}^T K_x^{-1}\mathbf{y} \tag{2.6}$$

The function (2.6) can then be maximised over σ and λ using the *optim* command, within the stats package [34] in the statistical program R [33], to obtain estimates for σ and λ, where these estimates will be denoted by $\hat{\sigma}$ and $\hat{\lambda}$. Note that the same values of $\hat{\sigma}$ and $\hat{\lambda}$ that maximises the log-likelihood function will also maximise the likelihood function. For each parameter estimate approximate 95% confidence intervals (CIs) will be obtained, where the CIs provide a range of plausible values for the parameters of interest. These CIs are calculated by the Wald method from the Hessian matrix, which contains the second derivatives of the log-likelihood function with respect to the parameters evaluated at $\hat{\sigma}$ and $\hat{\lambda}$:

$$H = \begin{bmatrix} \frac{\partial^2 l}{\partial \sigma^2} & \frac{\partial^2 l}{\partial \sigma \partial \lambda} \\ \frac{\partial^2 l}{\partial \sigma \partial \lambda} & \frac{\partial^2 l}{\partial \lambda^2} \end{bmatrix}.$$

To calculate the standard errors for the parameters the inverse of the negative Hessian matrix has to be obtained and the diagonal elements taken, and so the 95% CIs are calculated from

$$\hat{\sigma} \pm 1.96\sqrt{-H^{-1}[1,1]}, \tag{2.7}$$

$$\hat{\lambda} \pm 1.96\sqrt{-H^{-1}[2,2]}. \tag{2.8}$$

In general, once the maximum likelihood estimates and CI's have been obtained, it needs to be checked that these estimated maxima are feasible by ensuring that the parameter estimates do produce a distribution that gives the observed data the greatest probability. These estimated maxima can be checked by producing profile plots for the log-likelihood function and/or the likelihood function. To produce such plots a sequence of values for the estimated parameters to take has to be chosen. For each of these parameter values in the sequence, the log-likelihood function (2.6) and the exponential of the log-likelihood function will be calculated and plotted against the corresponding parameter values. Note that only one parameter can be varied in this way at the one time, and so all other parameters in the log-likelihood function (2.6) are set to their estimated value. If the maximum likelihood estimates are feasible then the log-likelihood and likelihood curves should peak at $\hat{\sigma}$ and $\hat{\lambda}$. Note however that it is not hard to show analytically that the profile likelihood is unimodal in sigma and hence so long as the optimisation has found a turning point, it will be the global maximum.

Putting this into practice, the curve in Figure 2.4(a) was created with $\sigma = \lambda = 1$ where the input consisted of 51 spatial points ranging from -25 to 25 with an interval of 1, and the curve in Figure 2.4(b) was created with

$\sigma = 2$ and $\lambda = 0.6$ where the input consisted of 101 spatial points ranging from -50 to 50 with the same interval.

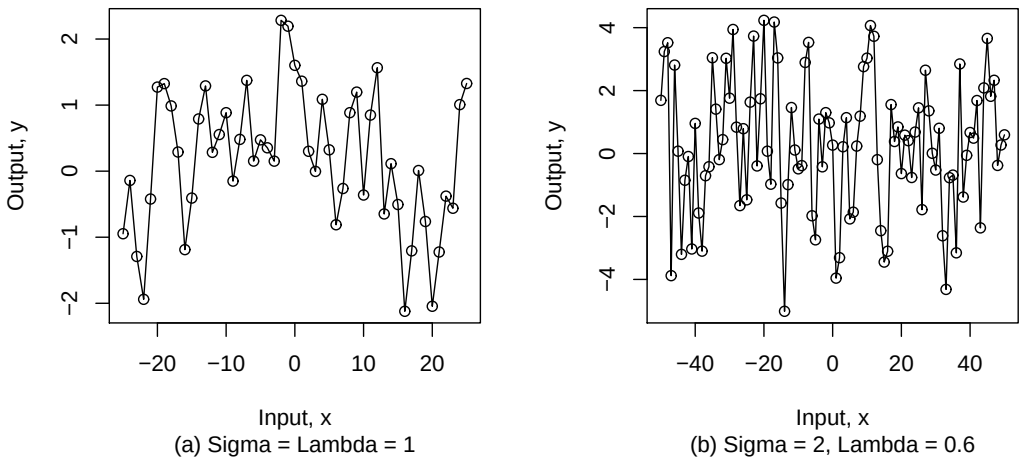

Figure 2.4: Functions Drawn at Random from a GP for Inferences to be Computed From

Note that these curves in Figure 2.4 don't appear to be very smooth despite the use of a smoothing covariance function. This is due to the chosen interval between the input values. If this interval were smaller, e.g., 0.2, then the simulated curves would be smoother. For the function in Figure 2.4(a), the following parameter estimates and CIs were obtained:

$\hat{\sigma} = 1.115$ (0.823, 1.407)

$\hat{\lambda} = 1.004$ (0.824, 1.184)

with the corresponding profile plots in Figure 2.5. The plots on the left-hand side are the calculated log-likelihood functions plotted against the corresponding sequence of parameter values and those on the right hand side are the calculated likelihood functions.

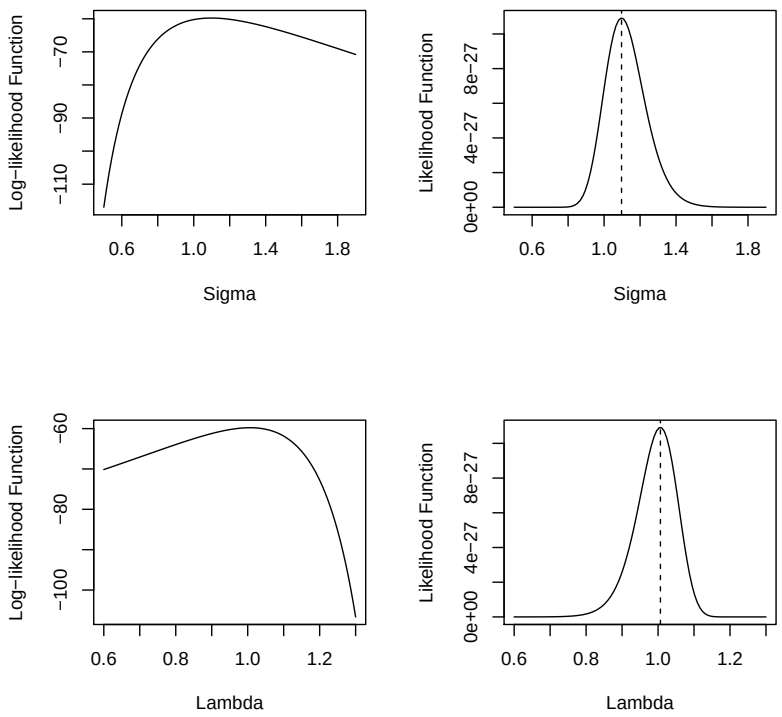

Figure 2.5: Profile Plots for the Function in Figure 2.4(a)

Similarly for the function in Figure 2.4(b),

$\hat{\sigma} = 1.982$ (1.897, 2.067)

$\hat{\lambda} = 0.618$ (0.542, 0.694)

with the corresponding profile plots in Figure 2.6.

Here we have the advantage of knowing the real parameter values and so can easily see that the estimates of $\hat{\sigma}$ and $\hat{\lambda}$ are very good. All CIs contain the real parameter values indicating that it is plausible that these functions were produced with these chosen parameter values. The profile plots also help to confirm that all estimates of $\hat{\sigma}$ and $\hat{\lambda}$ are feasible as they do indeed peak at the estimated maxima. The CIs for the function in Figure 2.4(b) are slightly narrower than those for the function in Figure 2.4(a). This is because there are more spatial points for the function in Figure 2.4(b) so there is more information about the curve available, and therefore the standard errors for

the estimates of $\hat{\sigma}$ and $\hat{\lambda}$ are smaller.

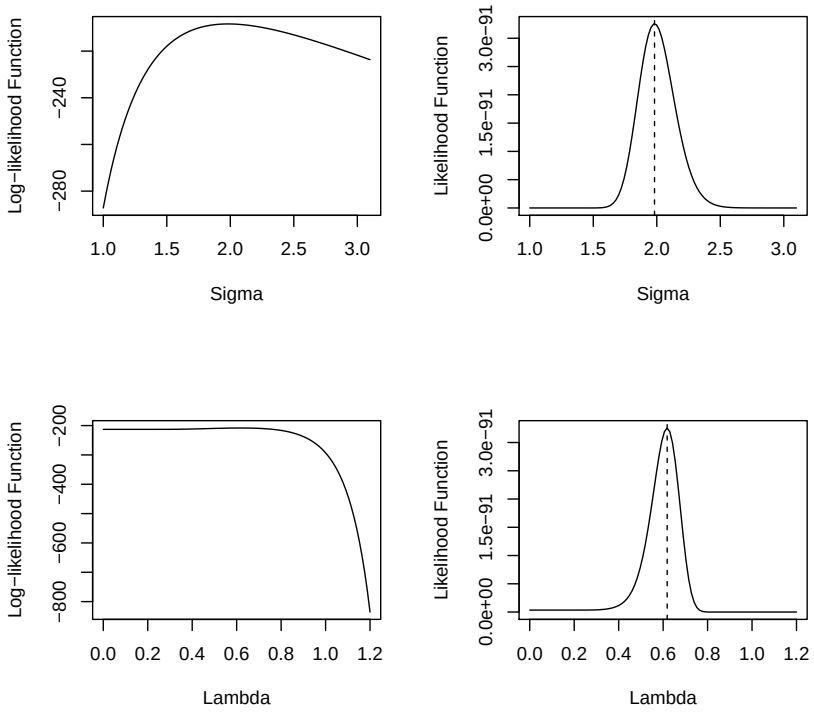

Figure 2.6: Profile Plots for the Function in Figure 2.4(b)

2.2 Simulation and Inference of Curves in Space and Time

Having established how to simulate curves in space from a GP model and identified how to compute inferences on such curves, then now this problem is to be extended to the simulation and inference of evolving curves similar to those in Appendix A and those in Figure 2.7.

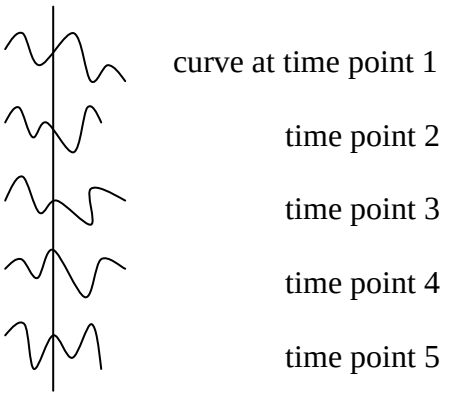

Figure 2.7: Curves Evolving over a Known Single Lineage

The simulation of these curves must be completed in such a way that ensures it is feasible that the curve at time points $t = 2, ..., m$ evolved from the corresponding previous curve(s). In section 2.1.3 it has been shown how easy it is to learn about the curves spatially by using the inference method of maximum likelihood to estimate the spatial parameters in the spatial covariance functions. The aim of this section is to try and estimate the spatial and temporal parameters within the space-time covariance functions in order to learn about the evolutionary aspects of the evolving simulated curves. The inference method of maximum likelihood, similar to that seen in section 2.1.3, will be used in order to attempt to estimate these parameters.

2.2.1 Space-Time Covariance Matrix

So far we have a GP model for the curve $f(\mathbf{x})$ sampled at the spatial points $\mathbf{x} = \{x_1, \ldots, x_n\}$ with zero mean and a SE spatial covariance function (2.1). Now suppose that this curve changes in time, i.e., evolves, so instead of $f(\mathbf{x})$ we have $f(\mathbf{x}, \mathbf{t})$, where $\mathbf{t} = \{t_1, \ldots, t_m\}$ is a vector containing m-time points: at each of these time points is an evolved curve. All that has now changed is that f is defined on a 2-dimensional space (of x and t) rather than just a 1-dimensional space (of x). In order to model curves that are evolving by a GP, a temporal covariance function is required as well as the spatial

covariance function (2.1). The new covariance function of the GP model will have the form

$$cov\big(f(x_1,t_1),f(x_2,t_2)\big)=k_{xt}(x_1,x_2,t_1,t_2), \tag{2.9}$$

where the subscript xt denotes that this is a space-time covariance function. It will be assumed that these functions are separable, i.e., a product of two functions k_x and k_t, one in space and one in time:

$$cov\big(f(x_1,t_1),f(x_2,t_2)\big)=k_{xt}(x_1,x_2,t_1,t_2)=k_x(x_1,x_2)\times k_t(t_1,t_2) \tag{2.10}$$

The temporal covariance function could also have the form of the SE covariance function:

$$k_x(x_1,x_2)=\sigma^2\exp\left[-\frac{1}{2\lambda^2}(x_1-x_2)^2\right], \tag{2.11}$$

$$k_t(t_1,t_2)=\exp\left[-\frac{1}{2\mu^2}(t_1-t_2)^2\right]. \tag{2.12}$$

k_t is a SE covariance function. It could have a separate σ_t^2 but as $\sigma_x^2\sigma_t^2$ would multiply given (2.10) to give some constant, there is no point in including it here. From (2.12), μ is a correlation-in-time parameter, where the correlation between curves will decay with an increasing time difference between the curves. K_t is the matrix that contains all the pairwise covariances between all time points of the curves.

To simulate curves at n-spatial points and m-time points a covariance matrix of size $nm\times nm$ with the following block structure has to be created:

$$
K_{xt} = \begin{array}{cc} & \begin{array}{cccc} t_1 & t_2 & & t_m \\ x_1 \dots x_n & x_1 \dots x_n & \dots \ \dots & x_1 \dots x_n \end{array} \\ \begin{array}{cc} & x_1 \\ t_1 & \vdots \\ & x_n \\ & x_1 \\ t_2 & \vdots \\ & x_n \\ & \vdots \\ & \vdots \\ & x_1 \\ t_m & \vdots \\ & x_n \end{array} & \left(\begin{array}{c|c|c|c} & & & \\ & & & \\ & & & \\ \hline & & & \\ & & & \\ & & & \\ & & & \\ & & & \\ \hline & & & \\ & & & \\ & & & \end{array}\right) \end{array}
$$

This matrix has the obvious potential to be very large. For example, letting $n = 100$ and $m = 10$ would result in a 1000×1000 matrix having to be stored which is problematic computationally. This matrix does have a simple block structure though:

$$
K_{xt} = \begin{bmatrix} X & Xk_t(t_1, t_2) & \dots & Xk_t(t_1, t_m) \\ Xk_t(t_2, t_1) & X & & \vdots \\ \vdots & & \ddots & \vdots \\ Xk_t(t_m, t_1) & \dots & \dots & X \end{bmatrix}.
$$

where X is the spatial covariance matrix, and $k_t(t_i, t_j)$ is the correlation between the curves i and j, and t_i and t_j are the time points of curves i and j, respectively, and so $k_t(t_i, t_j)$ is simply a constant. Therefore the space-time covariance matrix is simply made up of blocks of X multiplied by a corresponding $k_t(t_i, t_j)$ value. Note that the diagonal blocks are not multiplied by a k_t constant as $k_t(t_i, t_i)$ is the temporal correlation between

curve i and itself which is obviously one. K_{xt} is infact a Kronecker product of the spatial and temporal matrices. This allows for very efficient inference, as K_{xt} is never required to be evaluated or stored. There also exists neat formulae for determinants and inverses of the Kronecker products. K_{xt} is also symmetric and positive-definite.

For example, let's say $n = 10$ and $m = 3$, so X is a 10×10 spatial matrix which is multiplied by appropriate $k_t(t_i, t_j)$ constants. Hence K_{xt} is a block matrix of size 30×30 and has the following structure

$$K_{xt} = \begin{array}{c} \\ 10 \\ 10 \\ 10 \end{array} \begin{array}{c} \begin{array}{ccc} 10 & 10 & 10 \end{array} \\ \begin{pmatrix} X & Xk_t(t_1,t_2) & Xk_t(t_1,t_3) \\ Xk_t(t_2,t_1) & X & Xk_t(t_2,t_3) \\ Xk_t(t_3,t_1) & Xk_t(t_3,t_2) & X \end{pmatrix} \end{array}$$

This matrix has the form

$$K_{xt} = \left[\begin{array}{c|c|c} \begin{array}{c} covariance\ between \\ curve\ at\ t_1\ \&\ itself \end{array} & \begin{array}{c} covariance\ between \\ curves\ at\ t_1\ \&\ t_2 \end{array} & \begin{array}{c} covariance\ between \\ curves\ at\ t_1\ \&\ t_3 \end{array} \\ \hline \begin{array}{c} covariance\ between \\ curves\ at\ t_2\ \&\ t_1 \end{array} & \begin{array}{c} covariance\ between \\ curve\ at\ t_2\ \&\ itself \end{array} & \begin{array}{c} covariance\ between \\ curves\ at\ t_2\ \&\ t_3 \end{array} \\ \hline \begin{array}{c} covariance\ between \\ curves\ at\ t_3\ \&\ t_1 \end{array} & \begin{array}{c} covariance\ between \\ curves\ at\ t_3\ \&\ t_2 \end{array} & \begin{array}{c} covariance\ between \\ curve\ at\ t_3\ \&\ itself \end{array} \end{array}\right].$$

As stated earlier there is one drawback to this approach and that is the potential size of K_{xt}. Obviously the more time points there are then the larger K_t becomes, and as a result the larger K_{xt} becomes. Luckily there is a temporal covariance function that allows us to overcome this problem. This covariance function is not the SE but the Ornstein-Uhlenbeck (OU) process

covariance function

$$k_t(t_i, t_j) = \exp\left[-\frac{1}{\mu}\left|t_i - t_j\right|\right], \tag{2.13}$$

where μ is still the correlation-in-time parameter and $|t_i - t_j|$ is the time between curve i and j. The advantage of this covariance function over the SE temporal covariance function (2.12) will be explained further in the next section.

2.2.2 Simulation of Curves in Space and Time

Using the SE temporal covariance function (2.12) it will be demonstrated how curves in space and time can be simulated along a single lineage, like those in Figure 2.7. The first curve to be simulated has to be the ancestral curve which is simulated in space only (as the curve is yet to evolve and so has no time factor) from the following GP model:

$$f(\mathbf{x}, t_1) \sim N(\mathbf{0}, K_x). \tag{2.14}$$

From Figure 2.7, this ancestral curve evolves into the curve at t_2, and so $f(\mathbf{x}, t_2)$ must be conditional on $f(\mathbf{x}, t_1)$, i.e., $f(\mathbf{x}, t_2)|f(\mathbf{x}, t_1)$. All marginals of the multivariate normal distribution are themselves multivariate normal, and so the mean and the covariance functions of the GP model from which $f(\mathbf{x}, t_2)|f(\mathbf{x}, t_1)$ will be simulated are easily obtained by using the conditional multivariate probability formula which has the basic form

$$f(\mathbf{x}, t_i)|f(\mathbf{x}, t_j) \sim N(\boldsymbol{\mu}_i + \Sigma_{ij}\Sigma_{jj}^{-1}(\mathbf{y}_j - \boldsymbol{\mu}_j), \Sigma_{ii} - \Sigma_{ij}\Sigma_{jj}^{-1}\Sigma_{ij}^T), \tag{2.15}$$

where $\boldsymbol{\mu}_i$ and $\boldsymbol{\mu}_j$ denote the mean of the curves $f(\mathbf{x}, t_i)$ and $f(\mathbf{x}, t_j)$ respectively, and $\mathbf{y}_j$ is a vector containing the y-coordinates of the curve $f(\mathbf{x}, t_j)$, with the Σ's being the relevant block matrices from K_{xt}. From (2.15) it follows that the conditional distribution of $f(\mathbf{x}, t_2)|f(\mathbf{x}, t_1)$ is

$$f(\mathbf{x}, t_2)|f(\mathbf{x}, t_1) \sim N(\boldsymbol{\mu}_2 + K_{21}K_{11}^{-1}(\mathbf{y}_1 - \boldsymbol{\mu}_1), K_{22} - K_{21}K_{11}^{-1}K_{21}^T) \quad (2.16)$$

where $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$ denote the mean of the curves $f(\mathbf{x}, t_1)$ and $f(\mathbf{x}, t_2)$ respectively and $\mathbf{y}_1$ is a vector containing the y-coordinates of the simulated curve $f(\mathbf{x}, t_1)$. The space-time covariance matrix has the form

$$K_{xt} = \begin{bmatrix} X & Xk_t(t_1, t_2) \\ Xk_t(t_2, t_1) & X \end{bmatrix} = \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix}$$

and by plugging the relevant information into (2.16) the mean and covariance functions can be calculated and $f(\mathbf{x}, t_2)$ can be simulated from the GP model.

Suppose that a curve $f(\mathbf{x}, t_3)$ then evolves from $f(\mathbf{x}, t_2)$. To simulate $f(\mathbf{x}, t_3)$ then $f(\mathbf{x}, t_2)$ and $f(\mathbf{x}, t_1)$ have to be conditioned on respectively, i.e., $f(\mathbf{x}, t_3)|f(\mathbf{x}, t_2), f(\mathbf{x}, t_1)$ determined. The conditional distribution of $f(\mathbf{x}, t_3)|f(\mathbf{x}, t_2), f(\mathbf{x}, t_1)$ is still multivariate normal but with a complicated mean and covariance that depends on both time points 1 and 2.

As stated earlier the obvious drawback to having more curves needing to be simulated down a single lineage is that K_t increases in size, and hence K_{xt} also increases in size. This leads not only to computational issues of K_{xt} but conditional issues also. The conditional issues become more troublesome as the number of time points at which curves have to be simulated down a lineage increases, as then the number of previous curves that have to be conditioned on to simulate a curve at time point m also increases, i.e., we need to evaluate $f(\mathbf{x}, t_m)|f(\mathbf{x}, t_{m-1}), \ldots, f(\mathbf{x}, t_1)$, and therefore the more messy the conditional formula becomes. This is where the OU process temporal covariance function (2.13) comes into its own. The OU process possesses the Markov property [2]. A Markov process is a stochastic process which has a common description of being memoryless; this description arises from the fact that when conditional on the present state of a system its future and past are independent. This means that a curve simulated at any time point

except t_1 or t_2 (which can only be dependent on t_1 at all times) is conditional only on the previous curve, i.e., $f(\mathbf{x}, t_m)|f(\mathbf{x}, t_{m-1})$. The OU process is the only non-trivial process that is stationary, Gaussian and Markov. The problem then becomes a lot simpler as all that is required to be simulated is:

$$\begin{aligned}
&\text{first } f(\mathbf{x}, t_1) \\
&\text{then } f(\mathbf{x}, t_2)|f(\mathbf{x}, t_1) \\
&\text{then } f(\mathbf{x}, t_3)|f(\mathbf{x}, t_2) \\
&\qquad\vdots \\
&\text{then } f(\mathbf{x}, t_m)|f(\mathbf{x}, t_{m-1}),
\end{aligned}$$

where the simulation of the curves at time points $t = 2, \ldots, m$ only requires the relevant 2×2 block time-space covariance matrix to be constructed. Hence the larger K_{xt} matrix is avoided and the conditioning made simpler. The simulation of these curves is still carried out using the conditional multivariate formula (2.15). From this moment onwards the temporal covariance function that will be used throughout this thesis is the OU process. It's important to note that although the Markov property of the OU process is important it's role in reducing the workload is made less crucial by the fact that the covariance matrix is a a Kronecker product of a spatial and temporal matrix. As explained previously the Kronecker product allows for efficient inference, as K_{xt} is never required to be stored or evaluated.

Figure 2.8 is an example of evolving curves simulated along a single lineage from a GP model with the parameters set as follows: $\sigma = 1.5$, $\lambda = 1$, $\mu = 1$.

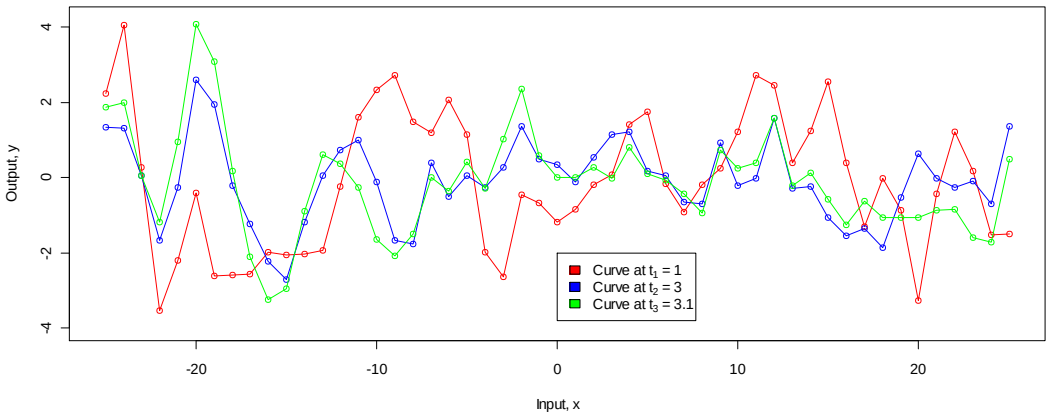

Figure 2.8: (1) Functions Simulated using the OU Temporal Covariance Function

From the time points of these curves, their positions on a single lineage would be as shown in Figure 2.9.

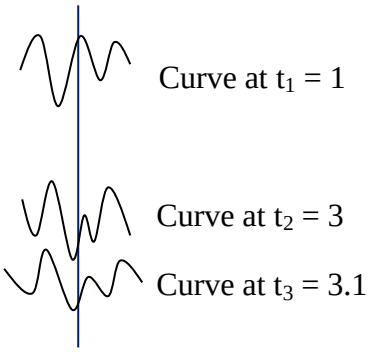

Figure 2.9: Positioning of the Simulated Curves in Figure 2.8

As correlation decreases between curves as the time difference between curves increases, then from Figure 2.9 one would expect the curves at t_2 and t_3 to be highly correlated. In contrast, it would be expected that the curves at t_2 and t_3 are not highly correlated with the curve at t_1. From Figure 2.8, it is clear that this is indeed the case, and so this simulation method does generate feasible evolving curves. Similarly Figure 2.10 shows three sample curves simulated along a lineage at the same time points as those in Figure 2.8, but with the parameters $\sigma = 1.5$, $\lambda = 1$, $\mu = 3$, where this increase in the correlation-in-time parameter does indeed produce functions that are

more correlated with one another, as expected.

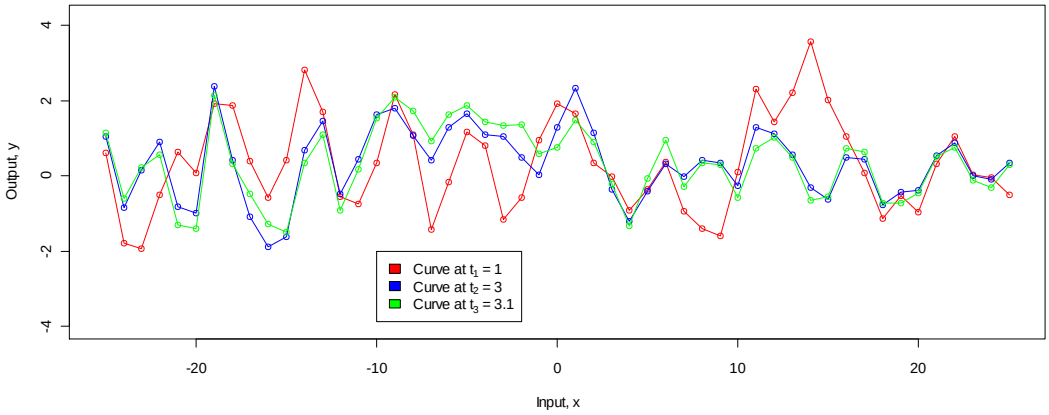

Figure 2.10: (2) Functions Simulated using the OU Temporal Covariance Function

2.2.3 Inferences on Curves Simulated in Space and Time

In order to learn about the spatial and temporal aspect of these curves, the space-time parameters within the space-time covariance functions have to be estimated. Again, the inference method of maximum likelihood will be used to obtain estimates for these parameters. The log-likelihood function (2.6) seen in section 2.1.3 must be adapted to allow for the fact that the curves are evolving and so now has the following form:

$$l(\boldsymbol{\theta}) = -\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\mathbf{f}_i^T K_x^{-1}\mathbf{f}_j k_t^{-1}(i,j) - \frac{1}{2}\log(\det K_{xt}) - \frac{mn}{2}\log(2\pi), \quad (2.17)$$

where the covariance matrices are implicitly functions of $\boldsymbol{\theta} = [\sigma, \lambda, \mu]^T$, $(\det K_{xt}) = (\det K_t)^n \times (\det K_x)^m$, $k_t^{-1}(i,j)$ is the (i,j)th element from the inverse temporal covariance matrix, and $\mathbf{f}_i$ and $\mathbf{f}_j$ are vectors containing the y-coordinates of the ith and jth curves. This function will also be maximised using the *optim* command in order to obtain estimates for σ, λ and μ. For each estimate of $\hat{\sigma}$, $\hat{\lambda}$ and $\hat{\mu}$, approximate 95% CIs will be calculated in the

same way as seen previously and the corresponding profile plots will also be constructed. The parameter estimates and CIs obtained on the evolving curves in Figure 2.8 are

$\hat{\sigma} = 1.549$ (1.293, 1.804),

$\hat{\lambda} = 1.051$ (0.978, 1.123),

$\hat{\mu} = 1.057$ (0.587, 1.528).

Again, we have the advantage of knowing the real parameter values and so can see that the estimates of $\hat{\sigma}$, $\hat{\lambda}$ and $\hat{\mu}$ are very good. Also all CIs contain the real parameter values and all profile plots peak at the estimated maxima, confirming that these estimates are all feasible. The CI for μ is considerably wider than the CIs for the two spatial parameters, which is partly due to the time differences between the curve at t_1 and those at t_2 and t_3; as a direct result of this time difference the curve at t_1 is not highly correlated with the curves at t_2 and t_3, and so the amount of information available between the bottom two curves and that at the top is reduced and therefore the standard error for the temporal parameter increases. Despite this lack of information though the CI is still relatively narrow which is down to how highly correlated the curves at t_2 and t_3 are. Although there is little information available between the curve at t_1 and the curves at t_2 and t_3 there is however a lot of information available between the curves at t_2 and t_3 which helps keep the standard error relatively low. Effectively we have only two time points to learn about μ from. The profile plots obtained for these parameter estimates is shown in Figure 2.11.

As well as computing inferences over all curves on a lineage it is also possible to compute inferences over a subset of curves on a lineage. For example suppose from Figure 2.8 one wanted to estimate the evolutionary parameters for the curve at t_1 evolving into that at t_2 only. In order to do this the log-likelihood function (2.17) will be calculated over these two curves only, i.e., the curve at t_3 will simply be ignored. The parameter estimates

and CIs obtained from these two curves are:

$\hat{\sigma} = 1.479$ (1.224, 1.733),

$\hat{\lambda} = 1.008$ (0.918, 1.099),

$\hat{\mu} = 1.036$ (0.035, 2.036).

Again the spatial and temporal parameter estimates returned are good, and all are feasible estimates. Not surprisingly the CI for μ is considerably wider than previously, where without the two highly correlated curves at t_2 and t_3 the standard error increases substantially as parameter estimates are being made on only two curves that are highly uncorrelated with each other. Despite this lack of information though, the parameter estimates returned are good which is the main issue here. The profile plots for these parameter estimates can be found in Figure 2.12, where the likelihood profile plot for the temporal parameter in Figure 2.12 helps highlight further the issue of limited information, with its shoulder at $\mu = 0$.

Having established that the inference method of maximum likelihood can be used to obtain estimates of the spatial and temporal parameters within the covariance functions that encode the simulated Gaussian curves, then the next step is to try and compute inferences on the experimental Chinese curves.

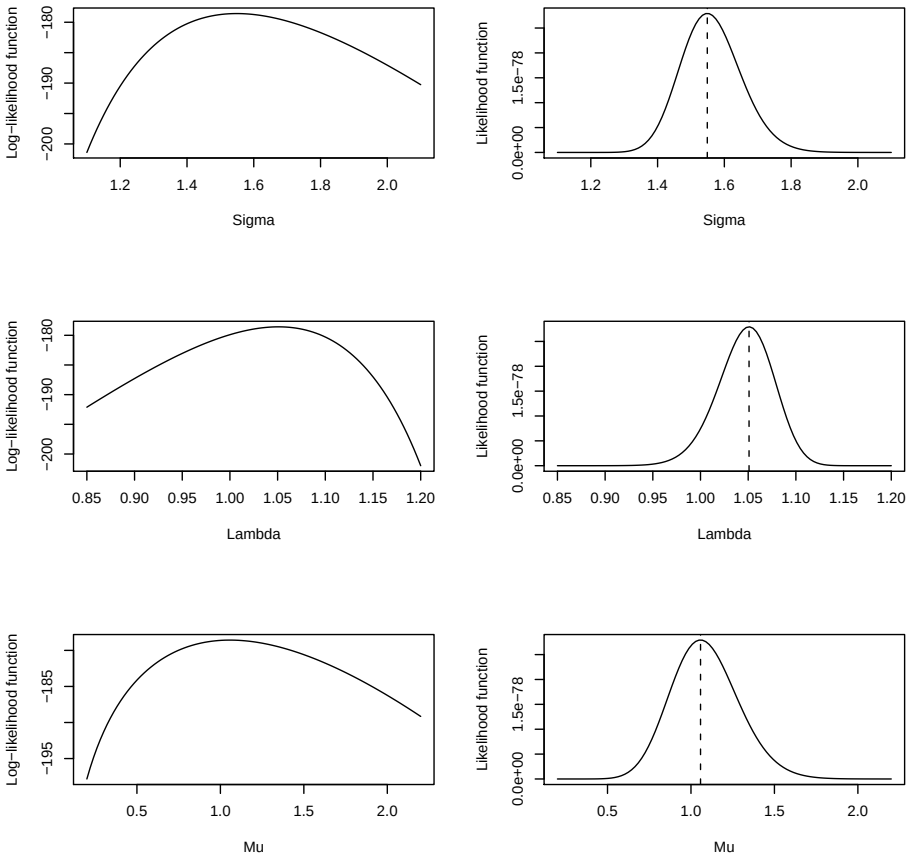

Figure 2.11: Profile Plots for the Functions in Figure 2.8

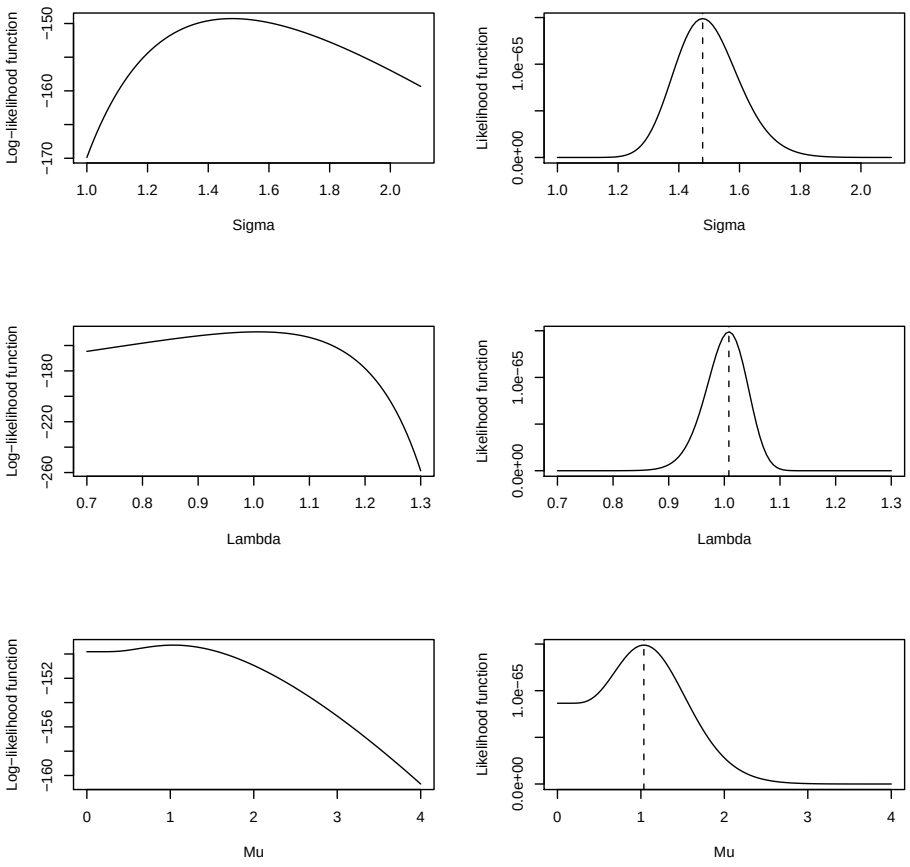

Figure 2.12: Profile Plots for the First Two Functions in Figure 2.8

2.3 Inference on the experimental Chinese whispers Curves

It will be assumed that the curves in Appendix A can be fitted with GP models that have the same form as the GPs from which the evolving curves were simulated in section 2.2.2, i.e., the ancestral curves at the top of each of the lineages will be assumed to be fitted with a model of the form

$$f(\mathbf{x}, t_1) \sim N(\mathbf{0}, K_x), \tag{2.18}$$

and the remaining curves at the discrete time points t_i and t_j, where $i =$

$2, ..., 6$ and $j = i - 1$, will be assumed to be fitted with an appropriate conditional model of the form

$$f(\mathbf{x}, t_i)|f(\mathbf{x}, t_j) \sim N(\boldsymbol{\mu}_i + \Sigma_{ij}\Sigma_{jj}^{-1}(\mathbf{y}_j - \boldsymbol{\mu}_j), \Sigma_{ii} - \Sigma_{ij}\Sigma_{jj}^{-1}\Sigma_{ij}^{T}) \qquad (2.19)$$

Estimates for the parameters will be obtained in the same way as those in section 2.2.3. Concentrating on the first set of curves only in Appendix A the following parameter estimates and CIs were obtained by maximising the log-likelihood function (2.17) over all curves on this lineage:

$\hat{\sigma} = 21.446$ (19.235, 23.656),

$\hat{\lambda} = 15.058$ (14.633, 15.482),

$\hat{\mu} = 1.864$ (1.312, 2.415).

Profile plots can be found in Figure 2.13.

It is also possible to compute inferences over a subset of these curves. For example, inferences can be made over the first three curves on this single lineage where the estimates and CIs obtained are:

$\hat{\sigma} = 20.486$ (17.635, 23.336),

$\hat{\lambda} = 15.655$ (15.099, 16.212),

$\hat{\mu} = 1.537$ (0.915, 2.159).

The profile plots for these parameter estimates can be found in Figure 2.14.

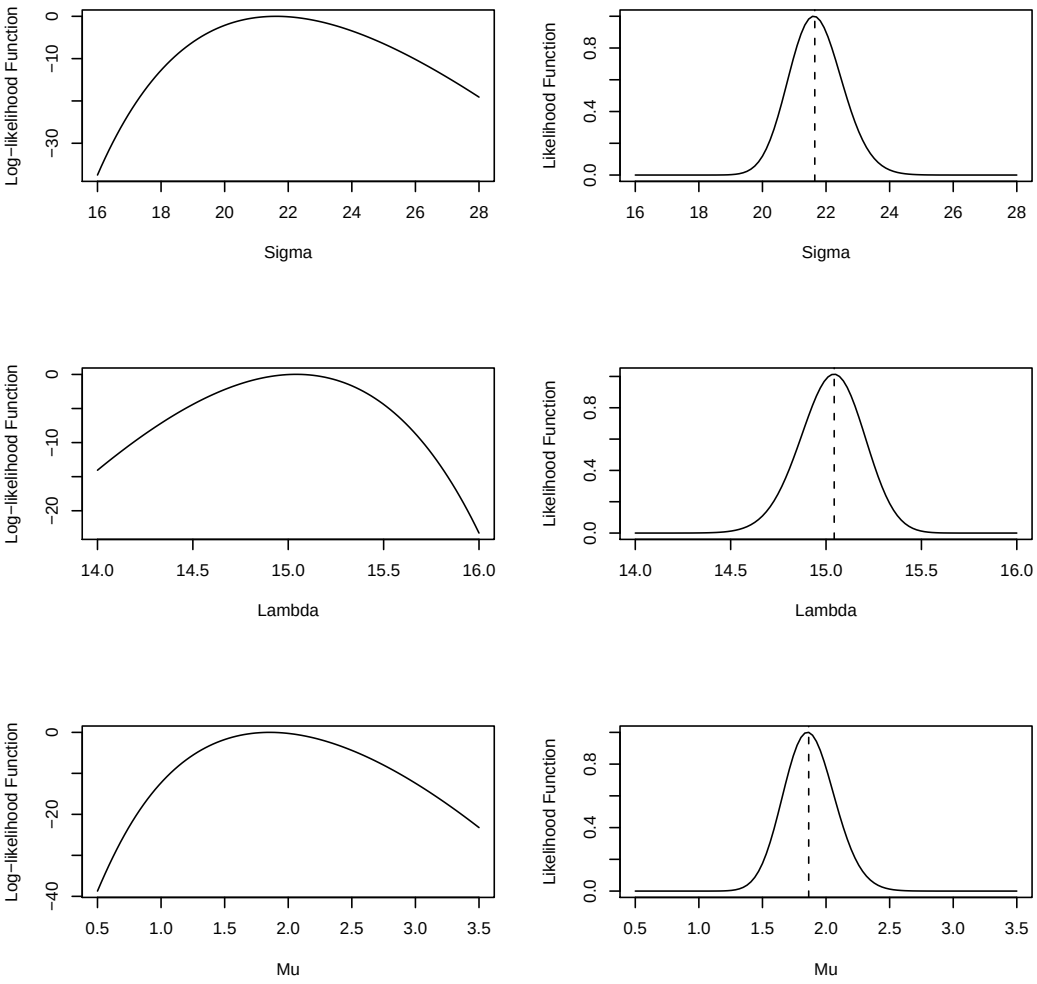

Figure 2.13: Profile Plots for the First Lineage in Appendix A

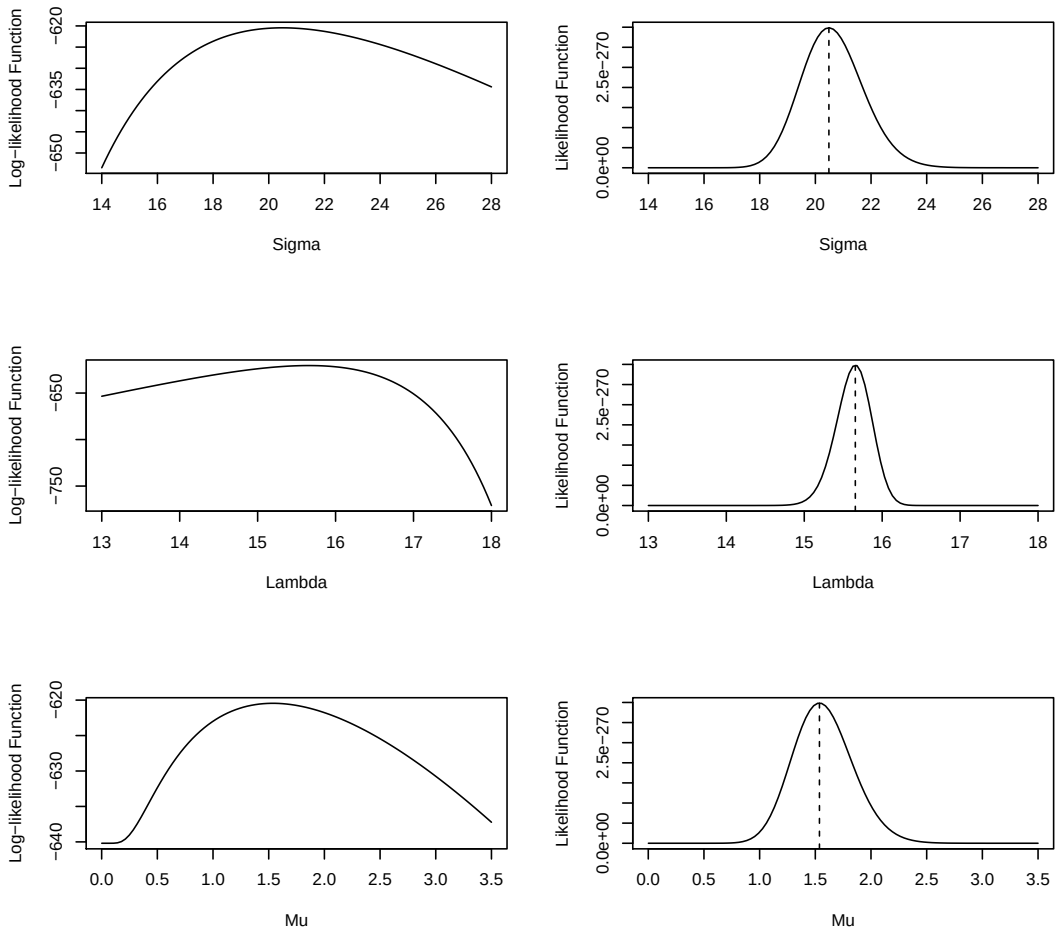

Figure 2.14: Profile Plots for the First Three Curves on the First Lineage in Appendix A

As a check that it is reasonable to fit these curves with the chosen GP model, the estimated parameters will be used to simulate curves from that GP. If these simulated curves are similar in space and time to the experimental Chinese whispers curves then the assumption that the Chinese whispers curves can be fitted with a GP model of the form in section 2.2.2 can be deemed, informally at least to be correct. The parameter estimates obtained over the first three curves from the first lineage in Appendix A were $\hat{\sigma} = 20.486$, $\hat{\lambda} = 15.655$ and $\hat{\mu} = 1.537$, where these estimates were obtained when the x-grid of input values consisted of 58 spatial points ranging from 58 to 742 with an interval of 12. The same input grid was used to simulate curves with the above parameter estimates.

Figure 2.15 displays the first three curves from the first single lineage in Appendix A.

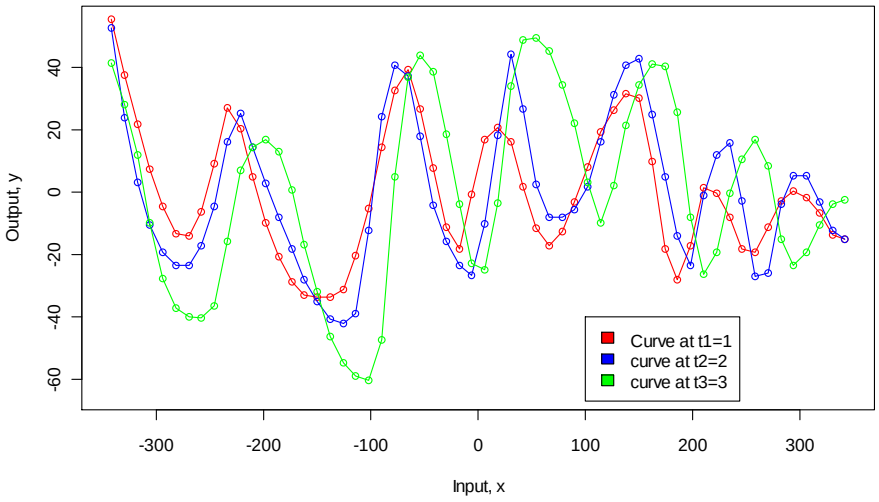

Figure 2.15: First Three curves from the First Lineage in Appendix A

Figure 2.16 in contrast displays three evolving curves that were simulated from a GP using the parameter estimates that were obtained from those curves in Figure 2.15. Details of simulation can be found in section 2.2.2.

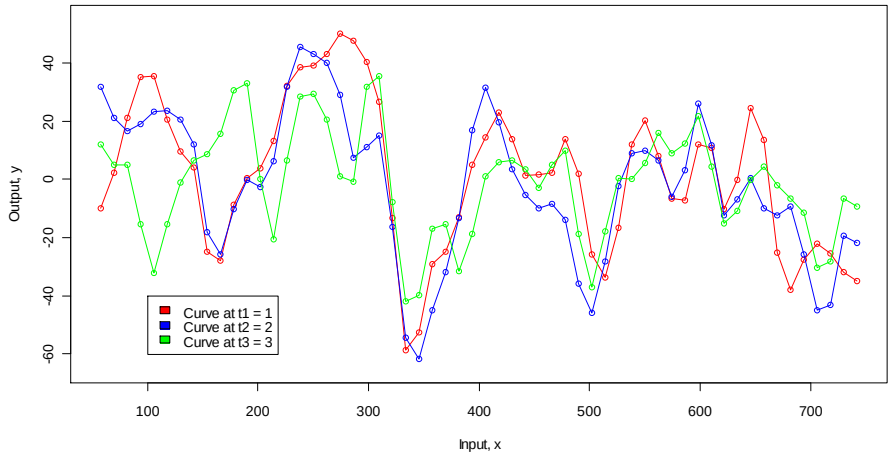

Figure 2.16: Simulated Evolving Curves using the Parameter Estimates obtained from the Curves in Figure 2.15

Clearly the curves in Figure 2.16 look different in character to those in Figure 2.15. The problem appears to be with the parameter λ, where the estimates obtained on the experimental Chinese whispers curves are too small, and hence the simulated curves are too wiggly. To check for any underlying problems the covariance functions which encode the curves might be estimated. As this is deemed to be a spatial problem then it is best to exclude time and concentrate on estimating the spatial covariance matrix only. Figure 2.17 displays a curve simulated in space from a GP using the parameter estimates obtained on the ancestral curve of the lineage in Figure 2.15, i.e., the curve at t_1.

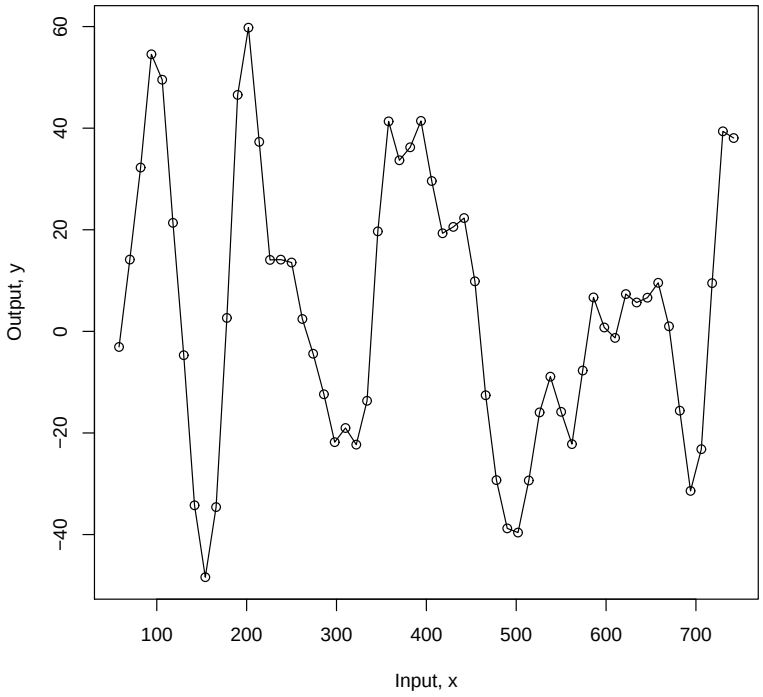

Figure 2.17: Simulation of the Ancestral Curve in Figure 2.15

To estimate the covariance function of the simulated curve the correlation between the curve and itself at lag $= 1, ..., \tau$ is measured. The correlation measurements are then plotted against the lags. As this curve is simulated from the SE GP in space then the estimated covariance plot should be bell-shaped. The same is also done for the ancestral Chinese whisper curve, and if it is correct to model this curve with a SE GP the estimated covariance plot should also be bell-shaped. Figure 2.18(a) is the estimated covariance function for the simulated curve and Figure 2.18(b) is the estimated covariance for the ancestral Chinese whisper curve.

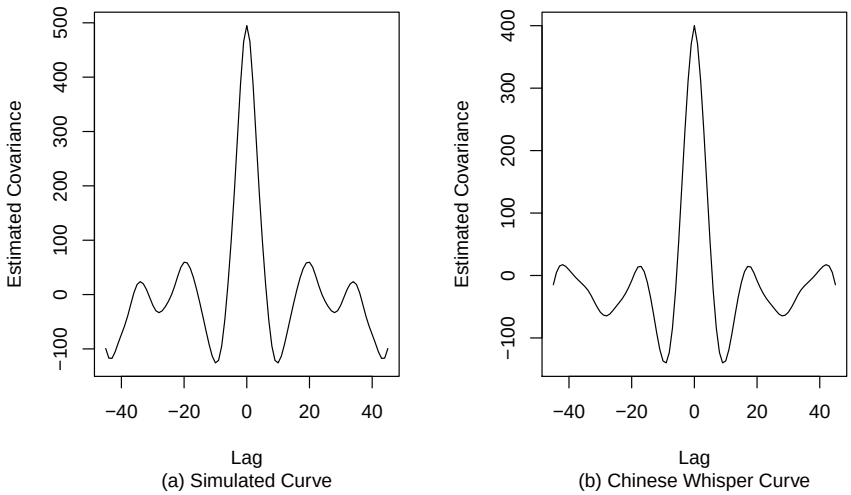

Figure 2.18: Estimated Covariance Functions

Neither estimate of covariance looks particularly like a SE. Figure 2.15 and 2.16 suggest that a much less smooth covaraince function is needed, where a covariance function is required that will capture the "kinks" in the curves, e.g., the curves at t_1 and t_2 in Figure 2.15 around t$\sim$200. As the SE covariance function does not allow for such "kinks" in the curves this leads to λ being under-estimated, and hence the wiggly curves in Figure 2.16 are produced. The same is also true for the four other lineages in Appendix A, where Appendix B contains the inferences computed on these lineages. Despite this being a rather unsatisfactory GP model for the experimental Chinese whispers curves this method is promising. Luckily there are many covariance functions for GP models so one would hope that there is a form of covariance that could describe these curves better, e.g., the Matérn covariance function, which may have yielded a larger estimate of λ, and therefore producing less wiggly simulated curves in Figure 2.16. The SE spatial covariance function (2.1) will continue to be used in the next three chapters of this thesis as no experimental Chinese whispers curves will be treated there.

2.4 Summary

In this chapter it has been shown how the time evolution of a function on the spatial domain can be easily represented as a function on the space-time domain by the specification of the mean and space-time covariance functions of a GP model. Also shown is how the inference method of maximum likelihood can be successfully applied to obtain estimates of the spatial and temporal parameters, within the space-time covariance functions, that encode the structure of the continuous curves evolving along a single lineage.

Unfortunately it was also discovered that a GP model with the SE spatial covariance function (2.1) was not a good fit for the experimental Chinese whispers curves. As explained earlier these curves require to be fitted with a GP model which favours smoother curves.

Chapter 3

Inference on GP Phylogenies

Having identified and illustrated how inferences can be computed on curves simulated down a single lineage, then the next step is to compute inferences over a simulated GP phylogeny (like that in Figure 3.1). Note that inferences in this chapter will only be computed on known tree topologies, although branch lengths will be unknown parameters.

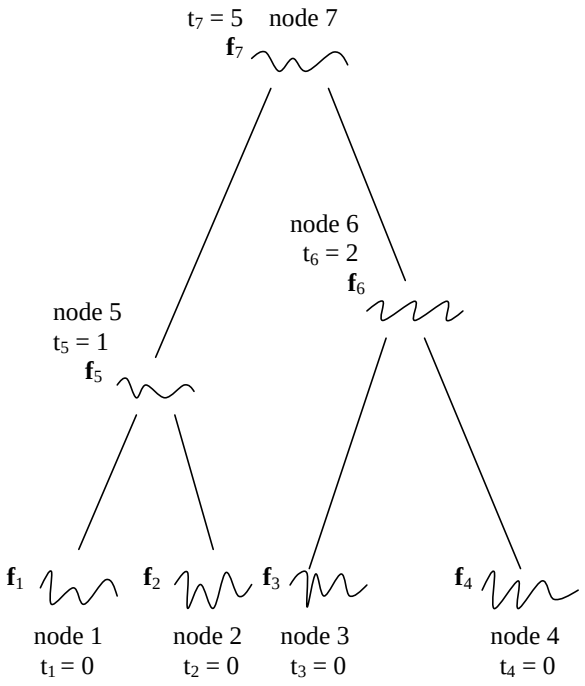

Figure 3.1: Example of a GP Phylogeny

The nodes in a phylogeny are numbered $1, ..., i$ and are labelled in order from the present to the past. In order to simulate a phylogeny, curves must be simulated at each of the nodes, i.e., $\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_i$. Before curves can be simulated though, time points must be chosen for each node in the phylogeny. The nodes, and hence the curves, at the s terminal nodes will be allocated time 0, since they are assumed to be in the present. The curves at the remaining $s-1$ nodes are allocated times in increasing order (note the reversed arrow of time) as one moves from the present to the past. For example, if one wanted to simulate a phylogeny like that in Figure 3.1, all curves at the terminal nodes will be allocated time 0, leaving 3 times having to be chosen for the 3 remaining nodes. One possible example for this phylogeny with $s = 4$ is: $\mathbf{f}_5$ (node $s+1$) has time 1, $\mathbf{f}_6$ (node $s+2$) has time 2 and $\mathbf{f}_7$ (node $s+3 = 2s-1$) has time 5. With these choices the branch above node 5 has duration $t_7 - t_5 = 5 - 1 = 4$, the branch above node 6 duration 3, those above tips 1 and 2 duration 1 and those above tips 3 and 4 duration 2. The phylogeny with these time points has a structure like that in Figure 3.1, with longer branch lengths between $\mathbf{f}_7$ and $\mathbf{f}_5$, and $\mathbf{f}_6$ and $\mathbf{f}_3$ or $\mathbf{f}_4$.

Once a phylogenetic GP has been simulated by generating curves at all nodes, then the knowledge of ancestors along any pathway is sufficient to estimate the covariance functions. As described in the previous chapter to estimate a covariance function, it is actually the parameters within the function that have to be estimated. The inference method of maximum likelihood will again be used to obtain estimates for the spatial and temporal parameters within these covariance functions.

3.1 Inference on a Simulated Simple Phylogeny

The first objective is to simulate a simple phylogenetic GP which has one ancestral curve and two terminal curves (Figure 3.2).

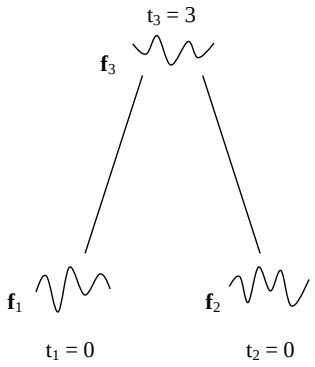

Figure 3.2: Simple Phylogeny

Similar to the simulation of lineages in Chapter 2, the first curve to be generated is at the top of the phylogeny which is simulated in space only (as this curve is yet to evolve). As in Chapter 2, this is done by simulating from the multivariate normal distribution with mean zero and covariance matrix derived from the SE spatial covariance function (2.1). The generation of the curves at the terminal nodes is facilitated by the fact that when a branch splits the evolution that occurs down the right-hand branch is independent of the evolution occurring down the left-hand branch, and so the curves $\mathbf{f}_1$ and $\mathbf{f}_2$ are completely independent of each other given $\mathbf{f}_3$. Due to this independence, both $\mathbf{f}_1$ and $\mathbf{f}_2$ are simulated by conditioning on the ancestral curve $\mathbf{f}_3$, i.e., $\mathbf{f}_1$ from $\mathbf{f}_1|\mathbf{f}_3$ and $\mathbf{f}_2$ from $\mathbf{f}_2|\mathbf{f}_3$. Focus on the simulation of $\mathbf{f}_1$ from $\mathbf{f}_1|\mathbf{f}_3$. First, the space-time covariance matrix has to be constructed, $K_{xt} = K_x \otimes K_t$, where K_x is derived from the SE covariance function (2.1) and K_t from the temporal covariance function which will be that of the OU process (2.13). The symbol $\otimes$ indicates the kronecker product. In this case, K_t has the form:

$$K_t = \begin{bmatrix} e^{-\frac{|t_3-t_3|}{\mu}} & e^{-\frac{|t_3-t_1|}{\mu}} \\ e^{-\frac{|t_1-t_3|}{\mu}} & e^{-\frac{|t_1-t_1|}{\mu}} \end{bmatrix} = \begin{bmatrix} 1 & e^{-\frac{t_3}{\mu}} \\ e^{-\frac{t_3}{\mu}} & 1 \end{bmatrix}$$

and so

$$K_{xt} = \begin{bmatrix} XK_t(1,1) & XK_t(1,2) \\ XK_t(2,1) & XK_t(2,2) \end{bmatrix} = \begin{bmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{bmatrix},$$

where X is the spatial covariance matrix, and K_{12} and K_{21} measure the covariance between the curves $\mathbf{f}_1$ and $\mathbf{f}_3$. The relevant terms can then be plugged into (2.15) to obtain the mean and covariance functions of the multivariate normal conditional distribution $\mathbf{f}_1|\mathbf{f}_3$ from which $\mathbf{f}_1$ will be simulated (full details of the simulation can be found in section 2.2.2). The branch length between $\mathbf{f}_2$ and $\mathbf{f}_3$ is the same as that between $\mathbf{f}_1$ and $\mathbf{f}_3$, and so K_t and hence K_{xt} keep the same form for the simulation of $\mathbf{f}_2$ from $\mathbf{f}_2|\mathbf{f}_3$ (again plugging the relevant information into (2.15)).

Once curves at the nodes of a phylogeny have been simulated, the covariance functions that encode the curves can be estimated in order to learn about the phylogenetic GP. As described previously it is the parameters within the covariance functions that have to be estimated. The parameters σ and λ determine how a curve appears in space and the parameter μ controls the temporal aspects of the curves. In order to compute inferences to obtain parameter estimates, the likelihood function first has to be determined. Because of the independence of different branches, the likelihood is a product of the distribution at the root and a set of conditional probabilities for each branch:

$$L(\boldsymbol{\theta}) = P(\textit{curve at node } 2s-1)\times \prod_{i=1}^{2s-2} P(\textit{curve at node } i \,|\, \textit{curve at ancestor of node } i), \tag{3.1}$$

where $\boldsymbol{\theta} = [\sigma, \lambda, \mu]^T$. From (3.1) the likelihood function for the simple phylogeny in Figure 3.2 is

$$L(\boldsymbol{\theta}) = P(\mathbf{f}_1|\mathbf{f}_3) \times P(\mathbf{f}_2|\mathbf{f}_3) \times P(\mathbf{f}_3). \tag{3.2}$$

This likelihood function can be split into two parts. The first part will concern the probability of the curve at the ancestral node and the second part will concern the conditional probabilities. The likelihood function for

the first part is

$$L_1(\boldsymbol{\theta}) = \frac{1}{(2\pi)^{\frac{n}{2}}(\det K_x)^{\frac{1}{2}}} \exp\left[-\frac{1}{2}\mathbf{f}_3^T K_x^{-1}\mathbf{f}_3\right]. \tag{3.3}$$

As described earlier it is easier to work with the logarithm of the likelihood function

$$l_1(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log(\det K_x) - \frac{1}{2}\mathbf{f}_3^T K_x^{-1}\mathbf{f}_3. \tag{3.4}$$

The conditional multivariate distribution for $P(\mathbf{f}_i|\mathbf{f}_{a(i)})$ has to be determined from part 2 of the likelihood function, where $i = 1, \ldots, 2s-2$ and $a(i) =$ ancestor of node i. The conditional distribution has the form

$$P(\mathbf{f}_i|\mathbf{f}_{a(i)}) \sim N(\boldsymbol{\mu}_i', \Sigma_i'), \tag{3.5}$$

where the mean and covariance matrix are, respectively,

$$\begin{aligned}\boldsymbol{\mu}_i' &= \mathbf{0} + K_x k_t(t_i, t_{a(i)}) K_x^{-1}(\mathbf{f}_{a(i)} - \mathbf{0}) \\ &= k_t(t_i, t_{a(i)})\mathbf{f}_{a(i)}\end{aligned} \tag{3.6}$$

and

$$\begin{aligned}\Sigma_i' &= K_x - K_x k_t^2(t_i, t_{a(i)}) K_x^{-1} K_x \\ &= K_x(1 - k_t^2(t_i, t_{a(i)})),\end{aligned} \tag{3.7}$$

from the general expression for the conditional distribution of a multivariate normal. Inserting $\boldsymbol{\mu}_i'$ and Σ_i' into the density of the standard multivariate normal distribution (2.4) provides the following likelihood function

$$\begin{aligned}L_2(\boldsymbol{\theta}) = \prod_{i=1}^{2s-2} & \frac{1}{(2\pi)^{\frac{n}{2}}\Big[\det K_x\big(1 - k_t^2(t_i, t_{a(i)})\big)\Big]^{\frac{1}{2}}} \\ & \times \exp\Big[-\frac{1}{2}\big(\mathbf{f}_i - k_t(t_i, t_{a(i)})\mathbf{f}_{a(i)}\big)^T \\ & \times K_x^{-1}\big(1 - k_t^2(t_i, t_{a(i)})\big)^{-1} \\ & \times \big(\mathbf{f}_i - k_t(t_i, t_{a(i)})\mathbf{f}_{a(i)}\big)\Big]\end{aligned} \tag{3.8}$$

and the logarithm of this function is

$$
\begin{aligned}
l_2(\boldsymbol{\theta}) = \sum_{i=1}^{2s-2} \Bigg[& -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\Big[\det K_x\big(1-k_t^2(t_i,t_{a(i)})\big)\Big] \\
& -\frac{1}{2}\big(\mathbf{f}_i - k_t(t_i,t_{a(i)})\mathbf{f}_{a(i)}\big)^T \\
& \times K_x^{-1}\big(1-k_t^2(t_i,t_{a(i)})\big)^{-1} \\
& \times \big(\mathbf{f}_i - k_t(t_i,t_{(a(i)})\mathbf{f}_{a(i)}\big)\Bigg]
\end{aligned} \tag{3.9}
$$

Bringing together $l_1(\boldsymbol{\theta})$ and $l_2(\boldsymbol{\theta})$, the complete log-likelihood function to be maximised over $\boldsymbol{\theta}$ and node times is

$$
\begin{aligned}
l(\boldsymbol{\theta}) = & -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log(\det K_x) - \frac{1}{2}\mathbf{f}_3^T K_x^{-1}\mathbf{f}_3 \\
& + \sum_{i=1}^{2s-2} \Bigg[-\frac{n}{2}\log(2\pi) - \frac{1}{2}\log\Big[\det K_x\big(1-k_t^2(t_i,t_{a(i)})\big)\Big] \\
& -\frac{1}{2}\big(\mathbf{f}_i - k_t(t_i,t_{a(i)})\mathbf{f}_{a(i)}\big)^T \\
& \times K_x^{-1}\big(1-k_t^2(t_i,t_{a(i)})\big)^{-1} \\
& \times \big(\mathbf{f}_i - k_t(t_i,t_{(a(i)})\mathbf{f}_{a(i)}\big)\Bigg]
\end{aligned} \tag{3.10}
$$

Having identified the log-likelihood function to be maximised then curves now need to be simulated for inferences to be computed on. Putting this into practice, three curves that make up a simple phylogenetic GP were simulated with the parameter values set as follows: $\sigma = 1$, $\lambda = 2$ and $\mu = 3$. The time of the ancestral node was set to 3. Note that as $k_t(t_1,t_3) = \exp\{-\frac{1}{\mu}|t_1 - t_3|\} = \exp(-t_3/\mu) = k_t(t_2,t_3)$ then μ and t_3 cannot be estimated separately, but only the ratio t_3/μ. Parameter estimates can be obtained on these simulated curves by maximising the log-likelihood function (3.10), and this was done using the *optim* command in R [33]. If it is feasible that we can learn about the spatial and temporal properties of the curves in a phylogeny, then the parameter estimates returned should be similar to the parameters used to simulate the curves. Listed below are the parameter estimates and 95% CIs

obtained on these simulated curves:

$\hat{\sigma} = 0.996$ (0.881, 1.110),

$\hat{\lambda} = 2.010$ (1.984, 2.037),

$\widehat{t_3/\mu} = 1.096$ (0.719, 1.473), (Note: $t_3/\mu = 3/3 = 1$).

The corresponding profile plots can be found in Figure 3.3.

Another set of three curves that make up a simple phylogenetic GP was simulated, but with the following parameters: $\sigma = 1$, $\lambda = 2$ and $\mu = 6$. The time of the ancestral node remained set to 3. The parameter estimates and CIs obtained from these curves were:

$\hat{\sigma} = 1.053$ (0.909, 1.196),

$\hat{\lambda} = 2.003$ (1.973, 2.033),

$\widehat{t_3/\mu} = 0.467$ (0.322, 0.612), (Note: $t_3/\mu = 3/6 = 0.5$).

The profile plots for these parameter estimates can be found in Figure 3.4.

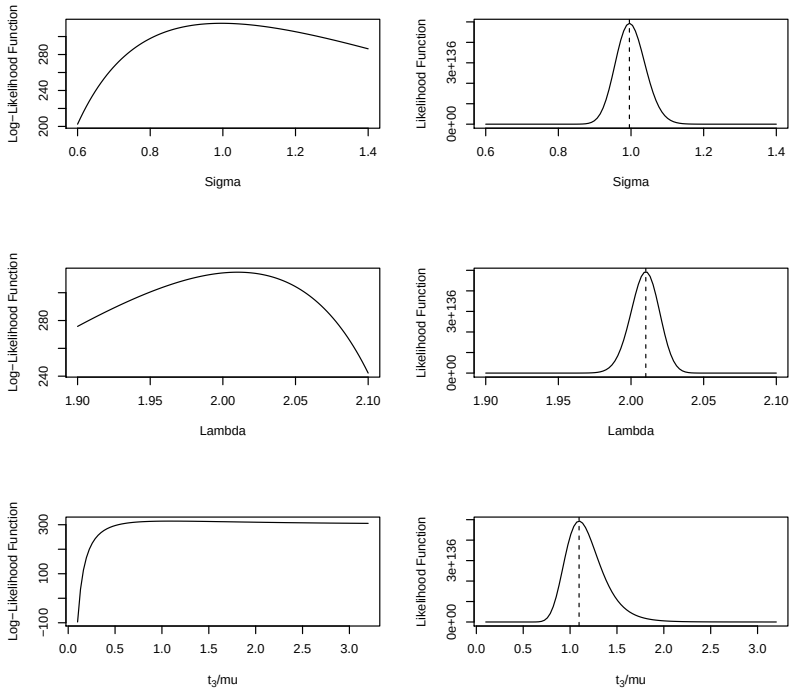

Figure 3.3: Profile Plots for the First Simulation of Figure 3.2

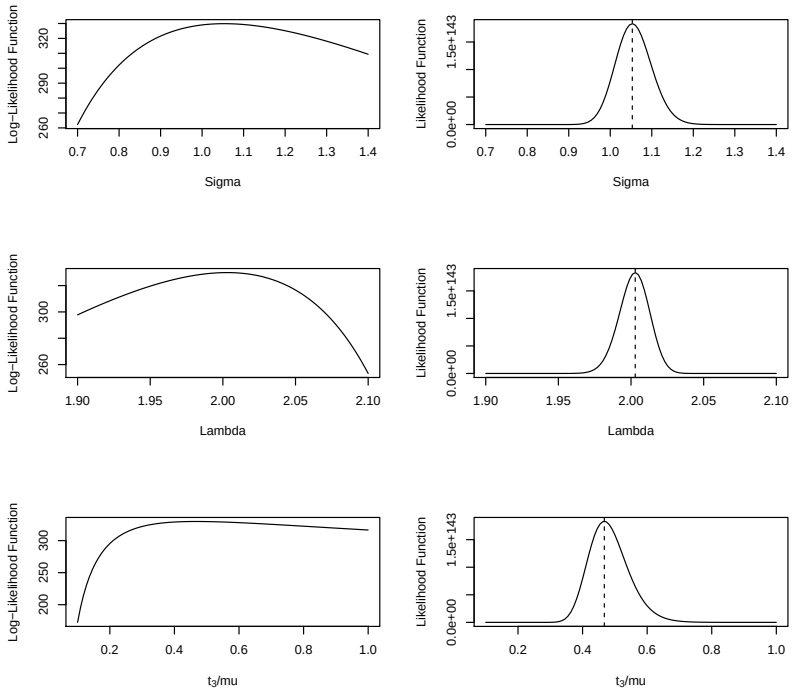

Figure 3.4: Profile Plots for the Second Simulation of Figure 3.2

The spatial parameter estimates obtained are all very good, with all CIs containing the real parameter values and all profile plots peaking at the estimated maxima, confirming that these estimates are feasible. The temporal parameter has also been well estimated in both cases. As expected the CI for t_3/μ is considerably narrower the second time around when the temporal parameter μ was set to 6. As explained previously, this is simply due to the fact that as μ increases the correlation between the curves increases. This in turn leads to an increase in the amount of information available to estimate μ, and so the CI for t_3/μ narrows.

The parameter estimates obtained on these phylogenies indicate that one can learn about the spatial and temporal aspects of all curves in a known phylogeny by the use of maximum likelihood estimation. In particular, being able to obtain good estimates for t_3/μ is important, as the time taken for one curve to evolve into another is of considerable interest and is one of the key factors in determining the strength of relationships between the curves.

In particular, if we had an independent estimate of μ, being able to estimate t_3/μ would permit estimation of t_3, the depth of the tree. However this is only a simple phylogeny with two tips, and so this theory needs to be tested further.

3.2 Inference on Simulated More Complex Phylogenies

Now we want to consider more complex phylogenies which include more terminal and interior curves. As a phylogeny grows in size the time differences between the curves, within the phylogeny, increase and so parameter estimation can potentially become more difficult as the correlation between the curves decreases. To test whether this phylogenetic GP estimation process is limited to simple phylogenies only, inferences will be attempted on the phylogeny in Figure 3.5.

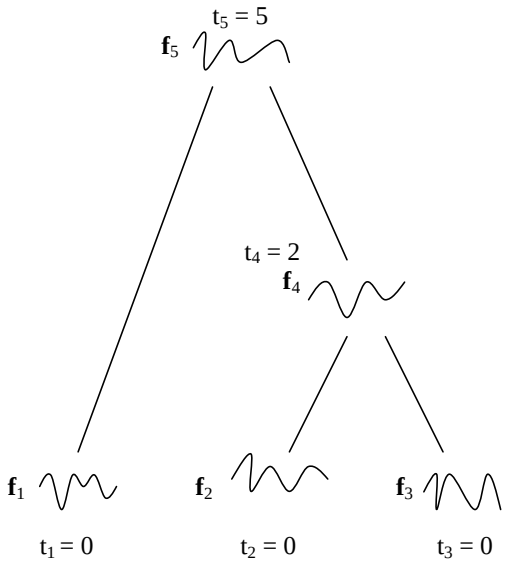

Figure 3.5: More Complex Phylogeny (1)

To simulate data from a phylogeny of this form, the first curve to be generated is the ancestral curve. Again this curve is generated in space only with mean zero and the SE covariance matrix (2.1). The curves $\mathbf{f}_1$ and $\mathbf{f}_4$ can be simulated by conditioning on the ancestral curve $\mathbf{f}_5$, where the correct K_t matrix must be constructed in each case. For $\mathbf{f}_1|\mathbf{f}_5$, K_t has the form

$$K_t = \begin{bmatrix} e^{-\frac{|t_5-t_5|}{\mu}} & e^{-\frac{|t_5-t_1|}{\mu}} \\ e^{-\frac{|t_1-t_5|}{\mu}} & e^{-\frac{|t_1-t_1|}{\mu}} \end{bmatrix} = \begin{bmatrix} 1 & e^{-\frac{5}{\mu}} \\ e^{-\frac{5}{\mu}} & 1 \end{bmatrix}$$

and for $\mathbf{f}_4|\mathbf{f}_5$

$$K_t = \begin{bmatrix} e^{-\frac{|t_5-t_5|}{\mu}} & e^{-\frac{|t_5-t_4|}{\mu}} \\ e^{-\frac{|t_4-t_5|}{\mu}} & e^{-\frac{|t_4-t_4|}{\mu}} \end{bmatrix} = \begin{bmatrix} 1 & e^{-\frac{3}{\mu}} \\ e^{-\frac{3}{\mu}} & 1 \end{bmatrix}.$$

The relevant terms can then be inserted into (2.15) to simulate $\mathbf{f}_1$ from $\mathbf{f}_1|\mathbf{f}_5$ and $\mathbf{f}_4$ from $\mathbf{f}_4|\mathbf{f}_5$. As the temporal covariance function being used is that of the OU process (which possesses the Markov property), then $\mathbf{f}_2$ and $\mathbf{f}_3$ are generated by conditioning on $\mathbf{f}_4$ only. The simulation of these curves is analogous to the simulation of the simple phylogeny seen in section 3.1 with the only difference being that the time of node 4 in this case is 2 (as opposed to 3 at the ancestral node previously).

Once the phylogeny is simulated, the parameters within the covariance functions that encode the curves must be estimated. The likelihood function for this phylogeny is very similar to that for the simple phylogeny in section 3.1:

$$L(\boldsymbol{\theta}) = P(\mathbf{f}_1|\mathbf{f}_5) \times P(\mathbf{f}_4|\mathbf{f}_5) \times P(\mathbf{f}_2|\mathbf{f}_4) \times P(\mathbf{f}_3|\mathbf{f}_4) \times P(\mathbf{f}_5). \tag{3.11}$$

(3.10) can be used to calculate this likelihood, where now $s = 3$ and $a(1) = a(4) = 5$, $a(2) = a(3) = 4$.

To determine whether it is possible to obtain accurate parameter estimates on a more complex phylogeny like that in Figure 3.5 data needs to be simulated. Five curves that make up a phylogenetic GP like that in Figure

3.5 were simulated with the parameters set to $\sigma = 1$, $\lambda = 2$ and $\mu = 4$. The time points chosen were $t_4 = 2$ and $t_5 = 5$. The parameter estimates obtained from maximising the log-likelihood with respect to σ, λ, t_4/μ and t_5/μ are:

$\hat{\sigma} = 0.970$ (0.862, 1.079),

$\hat{\lambda} = 1.985$ (1.960, 2.010),

$\widehat{t_4/\mu} = 0.451$ (0.318, 0.583), (Note: $t_4/\mu = 2/4 = 0.5$),

$\widehat{t_5/\mu} = 1.285$ (0.860, 1.709), (Note: $t_5/\mu = 5/4 = 1.25$).

The corresponding profile plots can be found in Figure 3.6.

A second set of five curves was simulated, with the following parameters: $\sigma = 1$, $\lambda = 2$ and $\mu = 8$. The time points were kept the same. The estimates and CIs obtained from these curves were:

$\hat{\sigma} = 1.068$ (0.939, 1.197),

$\hat{\lambda} = 2.017$ (1.993, 2.041),

$\widehat{t_4/\mu} = 0.249$ (0.176, 0.321), (Note: $t_4/\mu = 2/8 = 0.25$),

$\widehat{t_5/\mu} = 0.667$ (0.469, 0.864), (Note: $t_5/\mu = 5/8 = 0.625$).

The profile plots can be found in Figure 3.7.

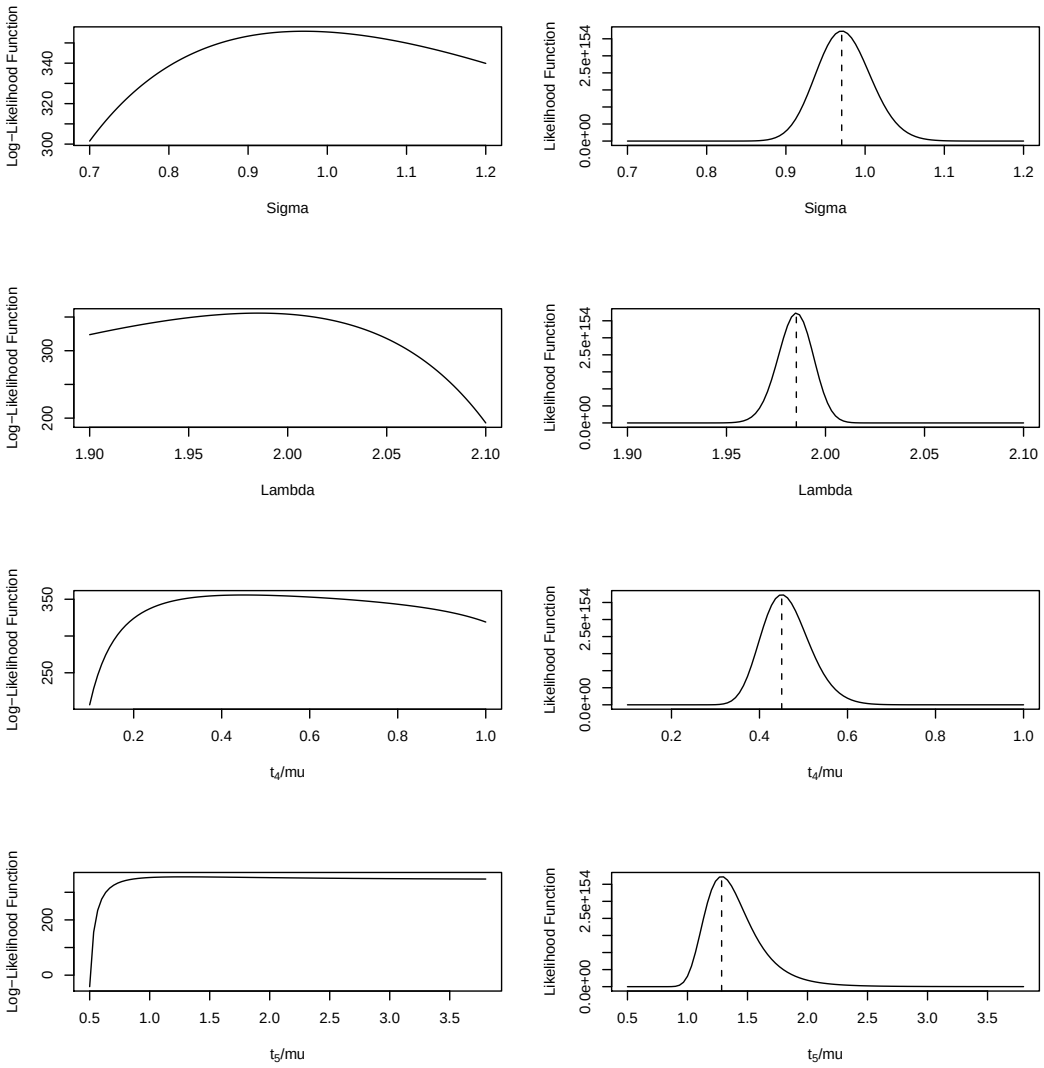

Figure 3.6: Profile Plots for the First Simulation of Figure 3.5

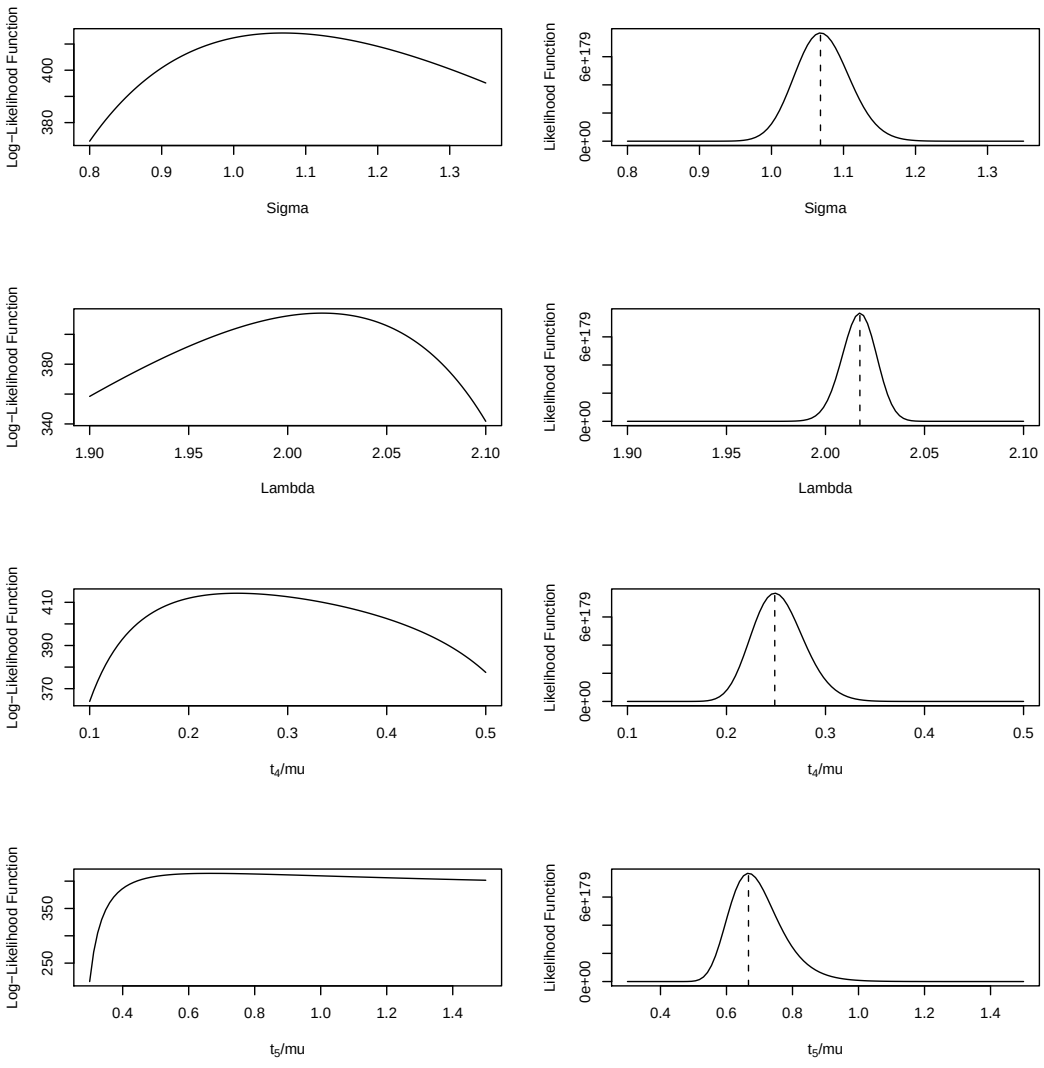

Figure 3.7: Profile Plots for the Second Simulation of Figure 3.5

Despite this being a more complex phylogeny with larger time differences between the curves, the spatial and temporal parameter estimates returned are as good as those obtained on the simple phylogeny (Figure 3.2). Also as one would expect, the CIs for t_5/μ are considerably wider than those for t_4/μ simply because of the larger time difference that exists between the curve at node 5 and those in the present time.

As parameter estimates have been successfully obtained on the phylogeny in Figure 3.5 then inferences will be attempted on a more complex phylogeny with more curves, and with larger time differences between the curves (Figure 3.8).

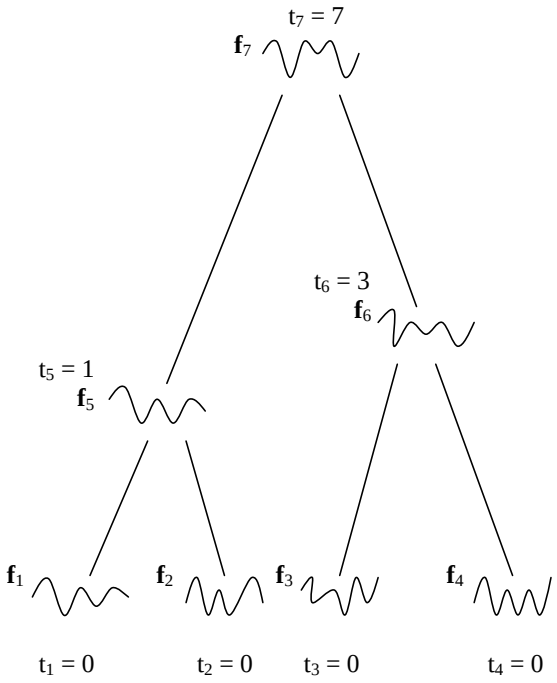

Figure 3.8: More Complex Phylogeny (2)

The simulation of this phylogeny is very similar to the simulation of the others. Seven curves that make up a phylogeny like that in Figure 3.8 were simulated with the parameters $\sigma = \lambda = 1$ and $\mu = 3$. The time points chosen were $t_5 = 1$, $t_6 = 3$ and $t_7 = 7$. The parameter estimates obtained from

maximising the log-likelihood (3.10) with respect to σ, λ, t_5/μ, t_6/μ and t_7/μ are:

$\hat{\sigma} = 1.085$ (0.975, 1.166),

$\hat{\lambda} = 1.021$ (0.984, 1.059),

$\widehat{t_5/\mu} = 0.304$ (0.205, 0.404), (Note: $t_5/\mu = 1/3 = 0.33$),

$\widehat{t_6/\mu} = 1.165$ (0.609, 1.721), (Note: $t_6/\mu = 3/3 = 1.00$),

$\widehat{t_7/\mu} = 2.127$ (1.366, 2.893), (Note: $t_7/\mu = 7/3 = 2.33$).

The corresponding profile plots can be found in Figure 3.9.

A second set of curves was also simulated but with the following parameters: $\sigma = \lambda = 1$ and $\mu = 12$. The estimates and CIs obtained from these curves were:

$\hat{\sigma} = 1.004$ (0.920, 1.088),

$\hat{\lambda} = 0.981$ (0.942, 1.020),

$\widehat{t_5/\mu} = 0.085$ (0.058, 0.112), (Note: $t_5/\mu = 1/12 = 0.08$),

$\widehat{t_6/\mu} = 0.259$ (0.174, 0.344), (Note: $t_6/\mu = 3/12 = 0.25$),

$\widehat{t_7/\mu} = 0.626$ (0.447, 0.804), (Note: $t_7/\mu = 7/12 = 0.58$).

The corresponding profile plots can be found in Figure 3.10.

Good parameter estimates have also been successfully obtained on these phylogenies. This is perhaps surprising as the correlation between the curves is reduced due to the larger time differences. By increasing μ to 12 for the second set of simulated curves there is a higher level of correlation which results in an increase in the amount of information available, and hence smaller standard errors for the estimated temporal parameters are returned.

3.3 Summary

From the analysis carried out in this chapter, it would seem that the method of maximum likelihood estimation has successfully reconstructed spatial and

temporal parameters on known GP phylogenies. As detailed earlier we were wary as to whether accurate parameter estimates could be obtained on complex phylogenies with large time differences between the curves, but it has been found that parameter estimates can successfully be obtained on such phylogenies. As expected, the CIs for the temporal parameters widen as the branch lengths increase, due to the information loss, from reduced correlation between the curves.

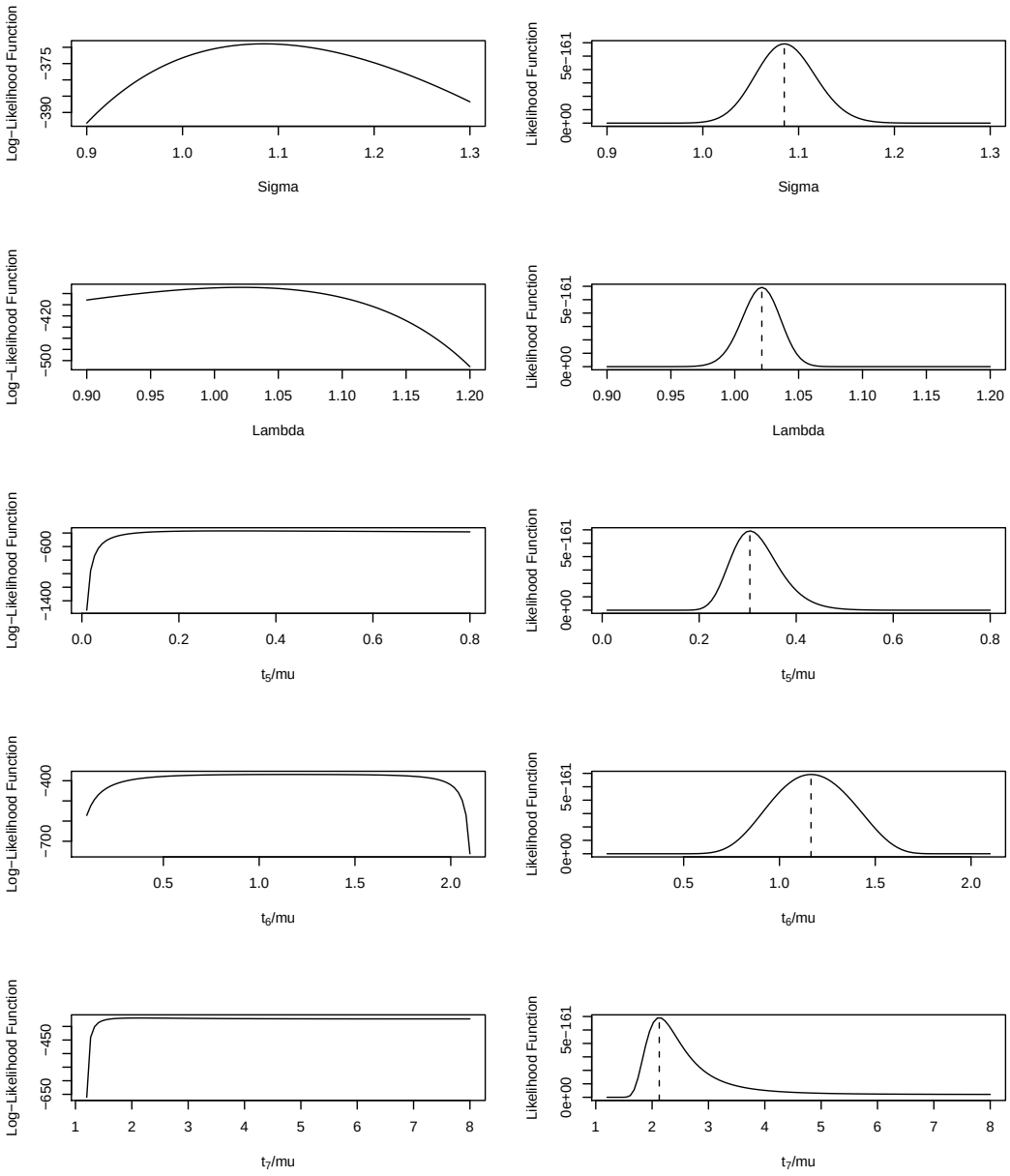

Figure 3.9: Profile Plots for the First Simulation of Figure 3.8

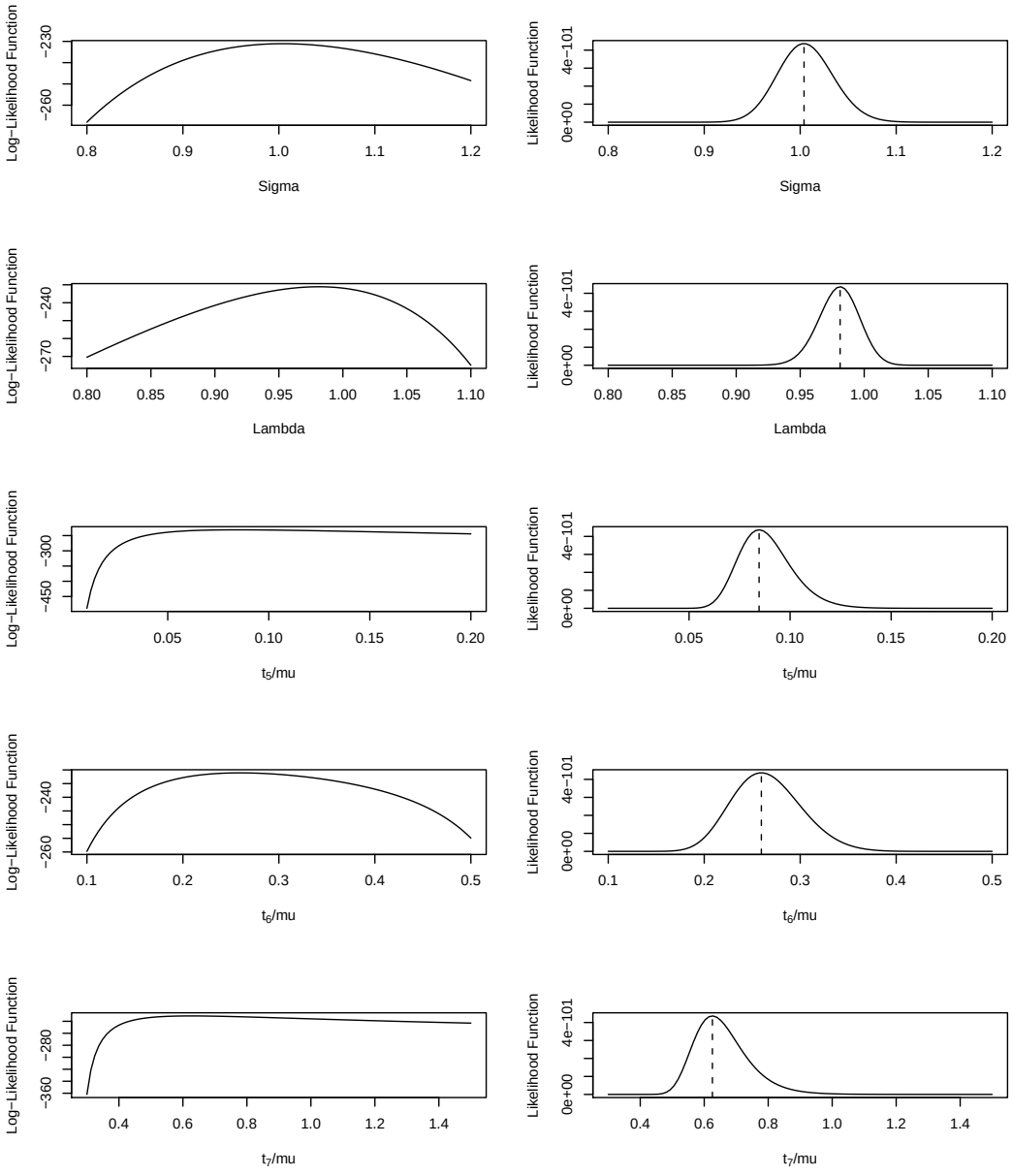

Figure 3.10: Profile Plots for the Second Simulation of Figure 3.8

Chapter 4

Unknown Ancestral Curves

In reality, we almost never have the advantage of knowing the curves in the interior of the phylogeny and in particular at the ancestral nodes, as we presumed we did in Chapter 3. Instead a typical phylogenetics problem is one in which one has a set of curves in the present time only (i.e., those curves at the tips or leaves of the phylogeny), where these curves are assumed to be related in some way. The idea is to use the information available from these curves to try and determine their ancestors, which in turn will allow the relationship between all the curves to be established.

The ancestor problem might involve in a Baysian context obtaining posterior probability distributions for one or more unobserved ancestral states, given all other data for the GP phylogeny, as well as estimating the parameters in the phylogenetic GP that were the forms of Chapter 3 (the branch lengths and the parameters in the covariance functions). Importantly, phylogenetic GPs do not distinguish between leaf (terminal) and ancestral states but only between observed and unobserved states. This has the vital computational advantage that only the unobserved ancestral states of interest need to be estimated. This property can greatly reduce the search space compared to equivalent genetic sequence inference problems.

Note that we assume in this chapter that the topology of the tree is known. This assumption is relaxed in Chapter 5.

4.1 Phylogeny with one Unknown Ancestor

The ancestor problem will first be treated on the simplest phylogeny, which has one unknown ancestor (Figure 4.1).

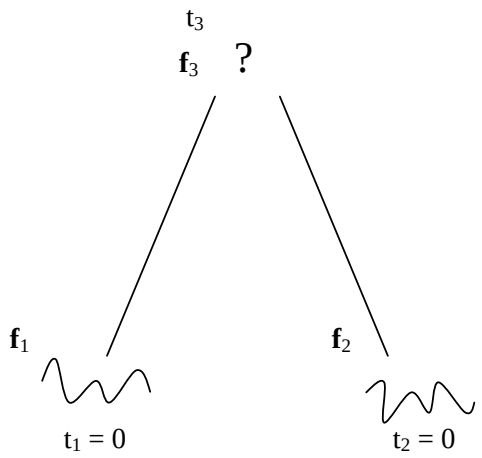

Figure 4.1: Simplest Phylogeny with One Unknown Ancestor

In order to attempt this ancestor problem, known related curves in the present time must be generated. As explained previously the idea of the ancestor problem is that the curve at the ancestral node in Figure 4.1 is unknown. However, in order to generate related curves at the tips of the phylogeny, an ancestral curve has to be simulated. The simulation of this curve follows the procedure outlined in Chapter 2, where the curve at the ancestor node is simulated first with mean zero and the SE covariance function (2.1). The curves at nodes 1 and 2 can then be generated by conditioning on that curve at node 3 (see sections 2.2.2 and 3.1 for details of the simulation). However, to proceed with this problem, the simulated ancestral curve will be discarded, and so the only data that will be regarded as known are the curves at the bottom of the phylogeny in the present time, as well as the tree topology, which in this example is trivial. This information will be used to obtain a likelihood function for the parameters in the model.

To obtain the likelihood, the joint distribution of $\mathbf{f}_1$ and $\mathbf{f}_2$ must be deter-

mined. From the rules of probability, this joint distribution can be obtained from

$$P(\mathbf{f}_1, \mathbf{f}_2) = \int P(\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3)\ d\mathbf{f}_3. \tag{4.1}$$

The direct integration over the joint distribution of $\mathbf{f}_1$, $\mathbf{f}_2$ and $\mathbf{f}_3$ with respect to $\mathbf{f}_3$ can be avoided by using one nice property of the multivariate normal distribution. This property is that all marginals of the multivariate normal distribution are themselves multivariate normal. $P(\mathbf{f}_1, \mathbf{f}_2)$ can be obtained by first computing $P(\mathbf{f}_1,\ \mathbf{f}_2,\ \mathbf{f}_3)$, i.e.,

$$\left[\begin{array}{c} \mathbf{f}_1 \\ \mathbf{f}_2 \\ \hline \mathbf{f}_3 \end{array}\right] \sim N_{3n}\left(\left[\begin{array}{c} \mathbf{0} \\ \mathbf{0} \\ \hline \mathbf{0} \end{array}\right], \left[\begin{array}{cc|c} K_{11} & K_{12} & K_{13} \\ K_{21} & K_{22} & K_{23} \\ \hline K_{31} & K_{32} & K_{33} \end{array}\right]\right),$$

and then, by pulling out the relevant information, $P(\mathbf{f}_1, \mathbf{f}_2)$ can be obtained:

$$\left[\begin{array}{c} \mathbf{f}_1 \\ \mathbf{f}_2 \end{array}\right] \sim N_{2n}\left(\left[\begin{array}{c} \mathbf{0} \\ \mathbf{0} \end{array}\right], \left[\begin{array}{cc} K_{11} & K_{12} \\ K_{21} & K_{22} \end{array}\right]\right). \tag{4.2}$$

Looking more closely at the space-time covariance matrix of $P(\mathbf{f}_1, \mathbf{f}_2)$, $K_{xt} = K_x \otimes K_t$, where K_x is still the SE covariance matrix (2.1) and K_t is still the OU temporal covariance matrix (2.13). In the previous chapter in section 3.1, it was seen that the temporal covariance matrix for both $\mathbf{f}_1|\mathbf{f}_3$ and $\mathbf{f}_2|\mathbf{f}_3$ had the form

$$K_t = \left[\begin{array}{cc} 1 & e^{-\frac{t_3}{\mu}} \\ e^{-\frac{t_3}{\mu}} & 1 \end{array}\right],$$

where t_3 is the time of the ancestral node. In this chapter the distribution is the joint distribution of $\mathbf{f}_1$ and $\mathbf{f}_2$ (irrespective of $\mathbf{f}_3$), and now the temporal

covariance matrix has the form:

$$K_t = \begin{array}{c} \\ \mathbf{f}_1 \\ \mathbf{f}_2 \end{array} \begin{array}{c} \begin{array}{cc} \mathbf{f}_1 & \mathbf{f}_2 \end{array} \\ \begin{bmatrix} 1 & e^{-\frac{2t_3}{\mu}} \\ e^{-\frac{2t_3}{\mu}} & 1 \end{bmatrix} \end{array},$$

where $K_t[1,2] = K_t[2,1] = e^{-2t_3/\mu}$ is obtained by tracing the length of the pathways between tips 1 and 2 in the phylogeny. From Figure 4.1, the pathway from $\mathbf{f}_1$ to $\mathbf{f}_3$ has length t_3 and the pathway from $\mathbf{f}_3$ down to $\mathbf{f}_2$ also has length t_3, and so the total time taken to pass from $\mathbf{f}_1$ to $\mathbf{f}_2$ is $2t_3$, yielding a covariance $e^{-2t_3/\mu}$.

The method of maximum likelihood will again be used to obtain estimates for the spatial and temporal parameters and branch lengths. The likelihood function is, from (4.2),

$$L(\boldsymbol{\theta}) = \frac{1}{(2\pi)^{\frac{n}{2}}(\det K_{xt})^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\mathbf{f}_1^T, \mathbf{f}_2^T)K_{xt}^{-1}\begin{pmatrix}\mathbf{f}_1 \\ \mathbf{f}_2\end{pmatrix}\right], \tag{4.3}$$

and so the log-likelihood function is:

$$l(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log(\det K_{xt}) - \frac{1}{2}(\mathbf{f}_1^T, \mathbf{f}_2^T)K_{xt}^{-1}\begin{pmatrix}\mathbf{f}_1 \\ \mathbf{f}_2\end{pmatrix}. \tag{4.4}$$

To determine whether it is possible to learn about the model parameters given that the data at the tips of the phylogeny are known, then two related curves in the present time have to be simulated. Two curves were simulated with the following parameters: $\sigma = \lambda = 1$ and $\mu = 6$. The time chosen for the ancestral node was 3. Again μ and t_3 are not separately identifiable, but only t_3/μ, since these parameters only occur in this combination in the likelihood function. By maximising function (4.4), the following estimates and CIs were obtained:

$\hat{\sigma} = 1.018$ (0.893, 1.144),

$\hat{\lambda} = 0.987$ (0.925, 1.049),

$\widehat{t_3/\mu} = 0.546$ (0.287, 0.805), (Note: $t_3/\mu = 3/6 = 0.5$).

The corresponding profile plots are shown in Figure 4.2.

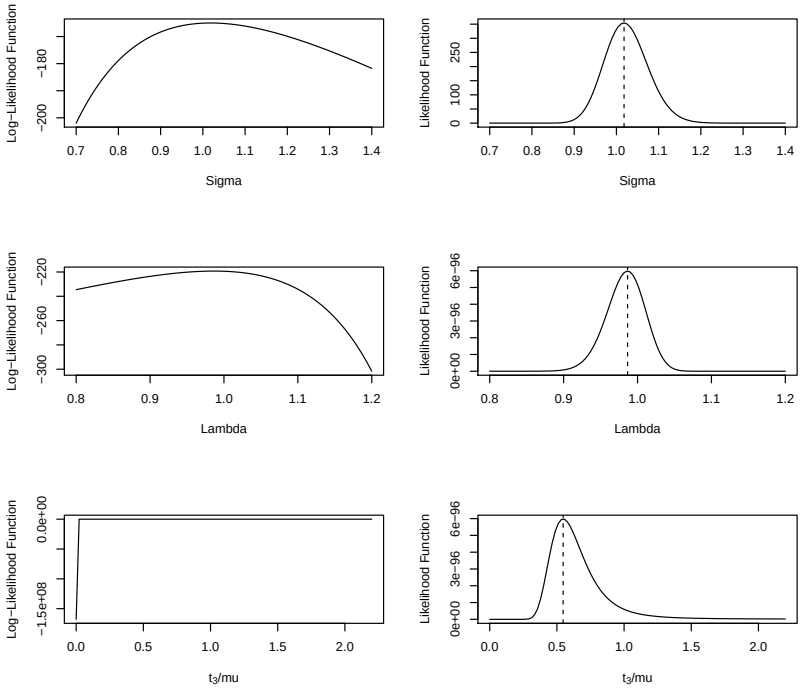

Figure 4.2: Profile Plots for the Simulation of Figure 4.1

Both spatial parameter estimates are good, as indicated by both the CIs and the profile plots. The parameter t_3/μ has also been well estimated, indicating that information about a phylogenetic GP can be obtained when only the curves in the present are known. However, this is the simplest phylogeny with one unknown ancestor, and so, to challenge this theory further, we will try to obtain parameter estimates on more complex phylogenies which have more than one unknown ancestor.

4.2 Phylogeny with Two Unknown Ancestors

Figure 4.3 illustrates a phylogeny with two unknown ancestors.

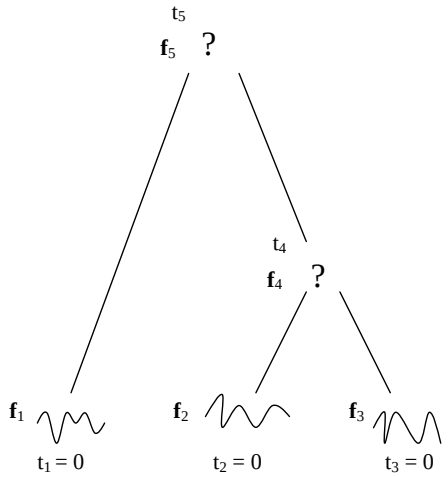

Figure 4.3: Phylogeny with Two Unknown Ancestors

The simulation of the curves $\mathbf{f}_1$, $\mathbf{f}_2$ and $\mathbf{f}_3$ is explained in section 3.2. To obtain the likelihood of the covariance parameters and the branch lengths, we need $P(\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3)$. By the marginalisation property

$$\begin{bmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \\ \mathbf{f}_3 \end{bmatrix} \sim N_{3n} \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} X & Xe^{-\frac{2t_5}{\mu}} & Xe^{-\frac{2t_5}{\mu}} \\ Xe^{-\frac{2t_5}{\mu}} & X & Xe^{-\frac{2t_4}{\mu}} \\ Xe^{-\frac{2t_5}{\mu}} & Xe^{-\frac{2t_4}{\mu}} & X \end{bmatrix} \right)$$

where X is the spatial covariance matrix. The covariance matrix above is $K_t \otimes X$, where K_t depends on the lengths of the paths between each pair of tips (as in section 4.1):

$$K_t = \begin{bmatrix} 1 & e^{-\frac{2t_5}{\mu}} & e^{-\frac{2t_5}{\mu}} \\ e^{-\frac{2t_5}{\mu}} & 1 & e^{-\frac{2t_4}{\mu}} \\ e^{-\frac{2t_5}{\mu}} & e^{-\frac{2t_4}{\mu}} & 1 \end{bmatrix}.$$

The log-likelihood function to be maximised over the model parameters has the form

$$l(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log(\det K_{xt}) - \frac{1}{2}(\mathbf{f}_1^T, \mathbf{f}_2^T, \mathbf{f}_3^T)K_{xt}^{-1}\begin{pmatrix}\mathbf{f}_1\\ \mathbf{f}_2\\ \mathbf{f}_3\end{pmatrix}. \qquad (4.5)$$

Three related curves were simulated in the present time with the parameter values set as follows: $\sigma = 2$, $\lambda = 1.5$ and $\mu = 6$. The time points chosen were $t_4 = 1$ and $t_5 = 3$. The parameter estimates and CIs obtained from maximising function (4.5) were:

$\hat{\sigma} = 1.943$ (1.687, 2.198),

$\hat{\lambda} = 1.470$ (1.432, 1.507),

$\widehat{t_4/\mu} = 0.140$ (0.085, 0.193), (Note: $t_4/\mu = 1/6 = 0.167$),

$\widehat{t_5/\mu} = 0.489$ (0.283, 0.696), (Note: $t_5/\mu = 3/6 = 0.5$).

The corresponding profile plots can be found in Figure 4.4.

As in the previous section, all estimates obtained are very good. The CI for t_5/μ is considerably wider than that for t_4/μ, due to the fact that there is less temporal correlation between $\mathbf{f}_5$ and both $\mathbf{f}_1$ and $\mathbf{f}_4$, than there is between $\mathbf{f}_4$ and both $\mathbf{f}_2$ and $\mathbf{f}_3$. As explained in section 2.2.1, the correlation between curves decreases as the time difference between them increases, leading to less information being available to obtain good parameter estimates.

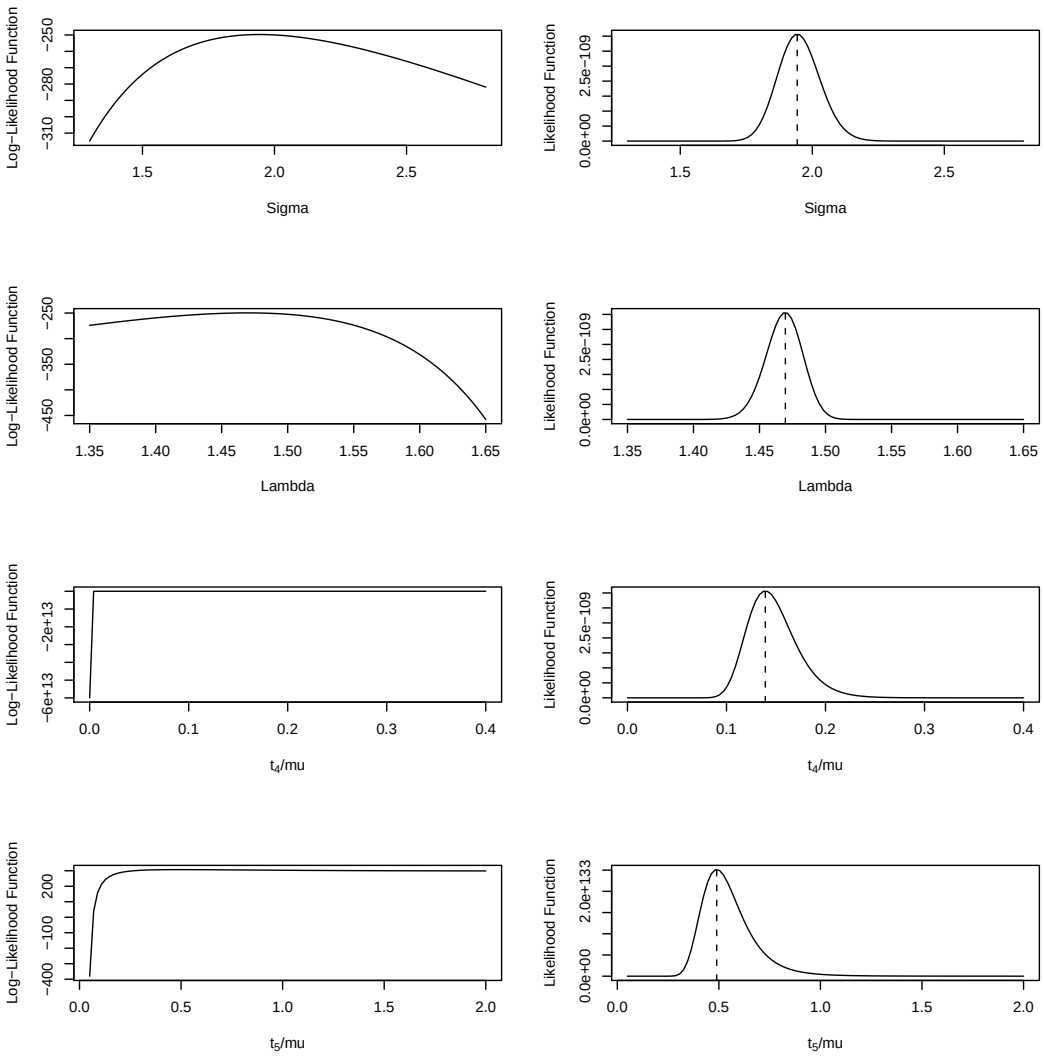

Figure 4.4: Profile Plots for the Simulation of Figure 4.3

4.3 Phylogeny with Three Unknown Ancestors

This problem will be extended even further to a phylogeny with three unknown ancestors, (Figure 4.5).

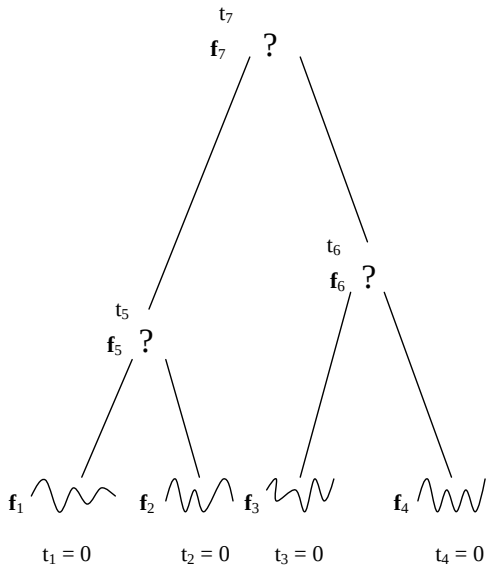

Figure 4.5: Phylogeny with Three Unknown Ancestors

The simulation of the terminal curves is explained in section 3.2. $P(\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3, \mathbf{f}_4)$ is obtained from:

$$\begin{bmatrix} \mathbf{f}_1 \\ \mathbf{f}_2 \\ \mathbf{f}_3 \\ \mathbf{f}_4 \end{bmatrix} \sim N_{4n} \left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} X & Xe^{-\frac{2t_5}{\mu}} & Xe^{-\frac{2t_7}{\mu}} & Xe^{-\frac{2t_7}{\mu}} \\ Xe^{-\frac{2t_5}{\mu}} & X & Xe^{-\frac{2t_7}{\mu}} & Xe^{-\frac{2t_7}{\mu}} \\ Xe^{-\frac{2t_7}{\mu}} & Xe^{-\frac{2t_7}{\mu}} & X & Xe^{-\frac{2t_6}{\mu}} \\ Xe^{-\frac{2t_7}{\mu}} & Xe^{-\frac{2t_7}{\mu}} & Xe^{-\frac{2t_6}{\mu}} & X \end{bmatrix} \right),$$

where the covariance matrix is $K_t \otimes X$ and K_t is:

$$K_t = \begin{bmatrix} 1 & e^{-\frac{2t_5}{\mu}} & e^{-\frac{2t_7}{\mu}} & e^{-\frac{2t_7}{\mu}} \\ e^{-\frac{2t_5}{\mu}} & 1 & e^{-\frac{2t_7}{\mu}} & e^{-\frac{2t_7}{\mu}} \\ e^{-\frac{2t_7}{\mu}} & e^{-\frac{2t_7}{\mu}} & 1 & e^{-\frac{2t_6}{\mu}} \\ e^{-\frac{2t_7}{\mu}} & e^{-\frac{2t_7}{\mu}} & e^{-\frac{2t_6}{\mu}} & 1 \end{bmatrix}.$$

The log-likelihood function to be maximised is analogous to (4.5). Four curves in the present time were created with the following parameter values: $\sigma = 2$, $\lambda = 1.5$ and $\mu = 6$. The time points chosen were $t_5 = 1$, $t_6 = 2$ and $t_7 = 5$. The estimates and CIs obtained from maximising the log-likelihood were:

$\hat{\sigma} = 2.097$ (1.874, 2.321),

$\hat{\lambda} = 1.517$ (1.485, 1.549),

$\widehat{t_5/\mu} = 0.168$ (0.108, 0.228), (Note: $t_5/\mu = 1/6 = 0.167$),

$\widehat{t_6/\mu} = 0.359$ (0.210, 0.509), (Note: $t_6/\mu = 2/6 = 0.333$),

$\widehat{t_7/\mu} = 0.875$ (0.443, 1.307), (Note: $t_7/\mu = 5/6 = 0.833$).

The corresponding profile plots can be found in Figure 4.6.

Another set of four curves was created, but with $\sigma = 2$, $\lambda = 1.5$ and $\mu = 16$. The time points remained the same. The estimates and CIs obtained are:

$\hat{\sigma} = 2.021$ (1.776, 2.267),

$\hat{\lambda} = 1.514$ (1.482, 1.546),

$\widehat{t_5/\mu} = 0.062$ (0.039, 0.085), (Note: $t_5/\mu = 1/16 = 0.063$),

$\widehat{t_6/\mu} = 0.108$ (0.068, 0.148), (Note: $t_6/\mu = 2/16 = 0.125$),

$\widehat{t_7/\mu} = 0.355$ (0.219, 0.490), (Note: $t_7/\mu = 5/16 = 0.313$).

The corresponding profile plots can be found in Figure 4.7.

The spatial and temporal parameter estimates returned on both sets of curves are very good. Looking at Figure 4.6, the likelihood profile plot for t_7/μ has a very long right tail, which indicates a lack of information about this parameter. The depth of the tree t_7 is sufficiently long (compared to μ)

that temporal correlation between $\mathbf{f}_7$ and both $\mathbf{f}_5$ and $\mathbf{f}_6$ has decreased to the point where the method knows that t_7/μ is large but there is little information to determine quite how large. By increasing μ to 16 for the second set of simulated curves and hence decreasing t_7/μ there is no longer such a drastic loss of information.

4.4 Summary

In this chapter it has been shown how to perform estimation for a phylogenetic GP when the tree topology is known but data exists only at the tips, by using the marginalisation property of the multivariate normal distribution. It was a relatively simple task to use the method of maximum likelihood to obtain estimates of the spatial and temporal parameters of the GP. Good parameter estimates were obtained on each of the three phylogenies in this chapter indicating that when only the curves in the present time and the phylogeny structure are known, there is still enough information to make useful estimates. A further object of interest might be to estimate the ancestral curves themselves. This could be straightforwardly accomplished either in a likelihood or Bayesian framework.

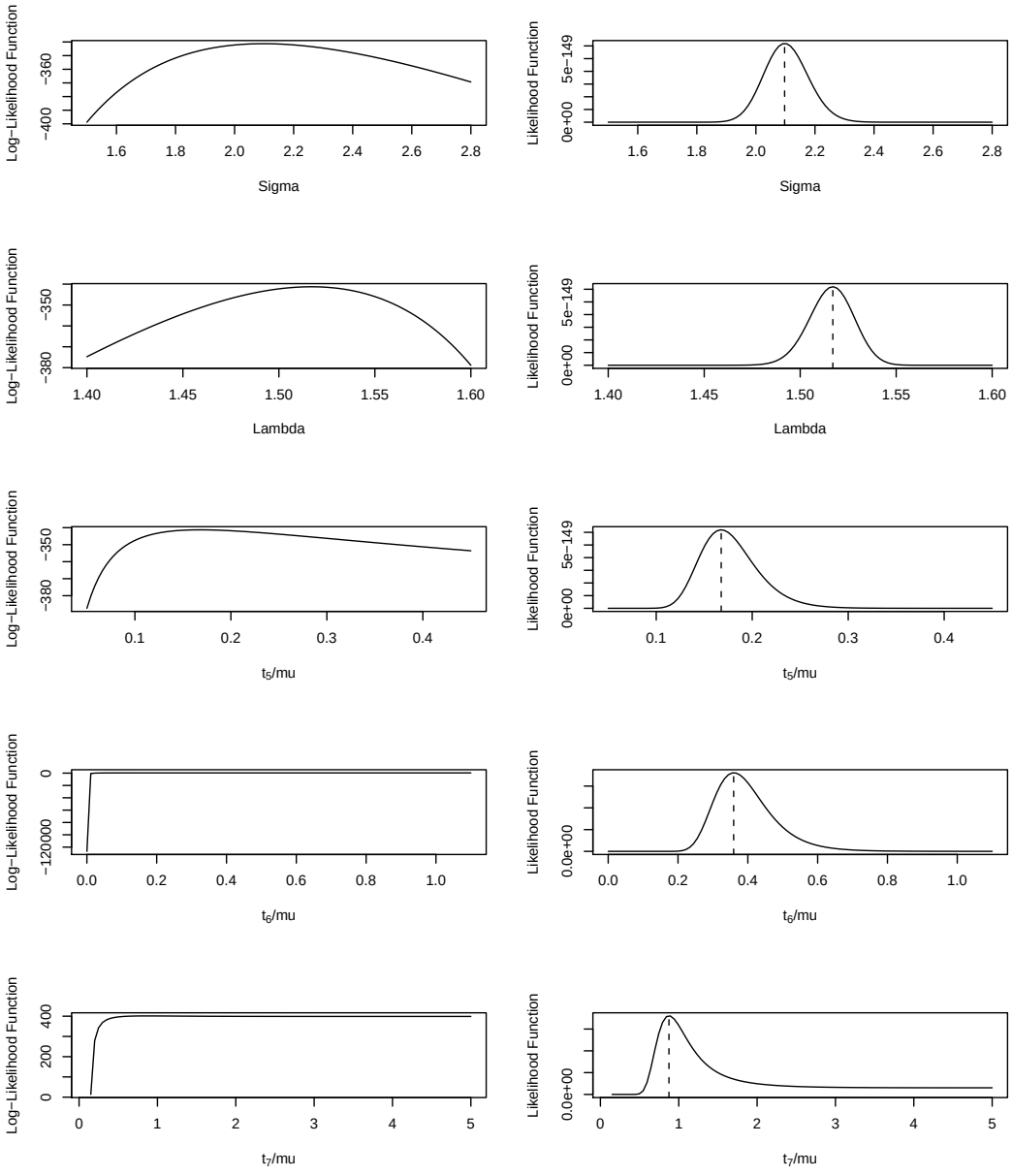

Figure 4.6: Profile Plots for the First Simulation of Figure 4.5

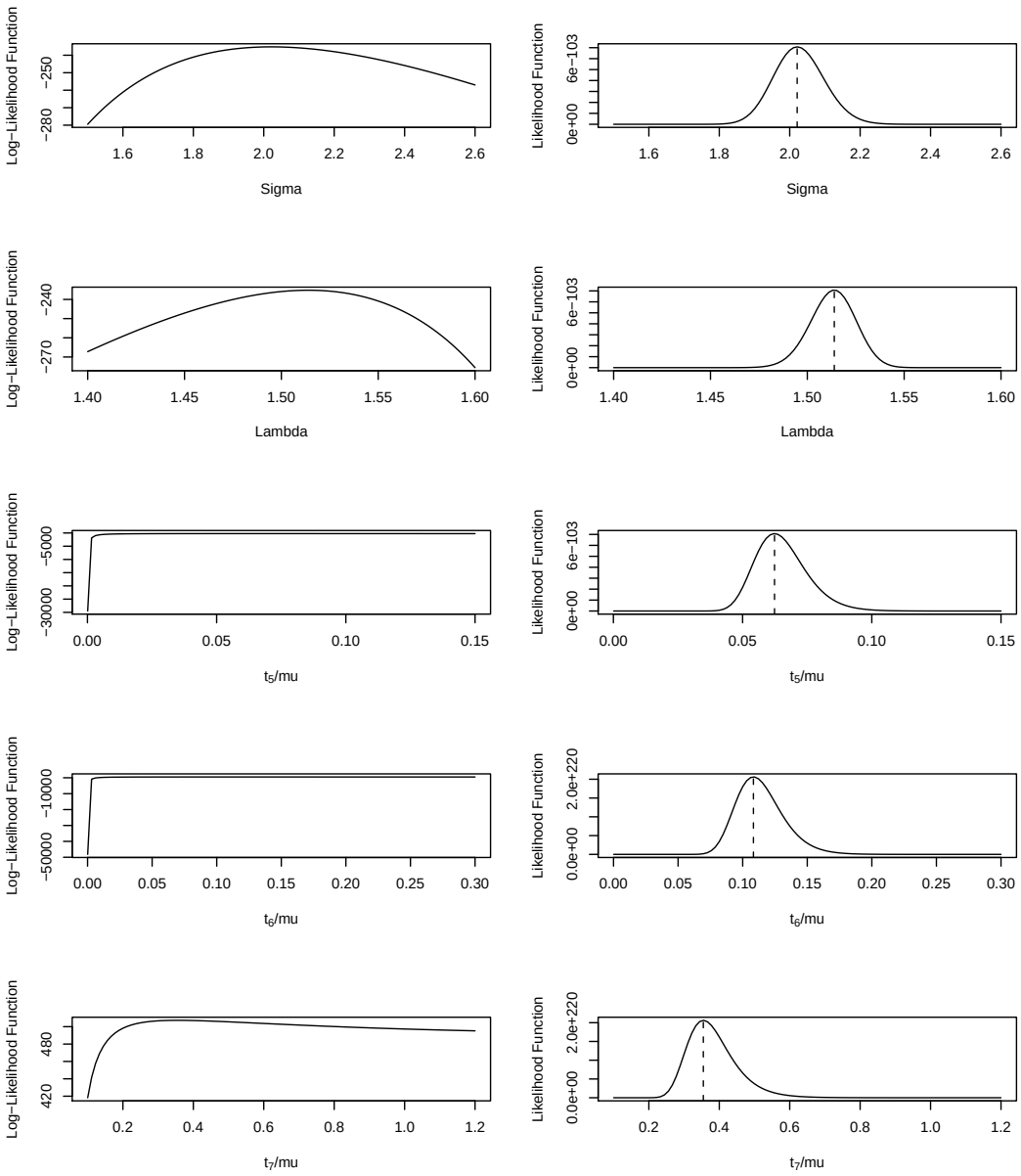

Figure 4.7: Profile Plots for the Second Simulation of Figure 4.5

Chapter 5

Unknown Phylogeny

In a real-life phylogenetics problem, the phylogeny topology and the curves at the ancestral nodes are usually unknown. So the only information available are the curves in the present time, which are hypothesised to be related in some way, but the exact relationship is unknown. The focus of this chapter is to establish the correct phylogeny topology that connects these curves so this relationship can be understood.

In a Bayesian context, the phylogeny problem would be to obtain a posterior distribution on the space of phylogenies. Each feasible phylogeny maps to a phylogenetic GP, and therefore to a likelihood given the observed states. Given a prior for phylogenies, Bayes' rule therefore gives the posterior distribution. We shall approach the problem however from a likelihood perspective, seeking to find the phylogeny (topology and branch lengths) and covariance parameters that maximise the likelihood function. Here again, the GP framework reduces the computational load greatly since unobserved states need not be fitted.

5.1 Choosing the Correct Phylogeny with Three Terminal Nodes

Starting simply we will consider the problem of having only three curves in the present time. The method of maximum likelihood estimation will be used to determine the correct phylogeny topology for these curves. In order to establish the correct topology one has to consider all possible ways these curves can be related. Luckily a phylogeny which has three terminal nodes has only three possible rooted structures, which are those in Figure 5.1.

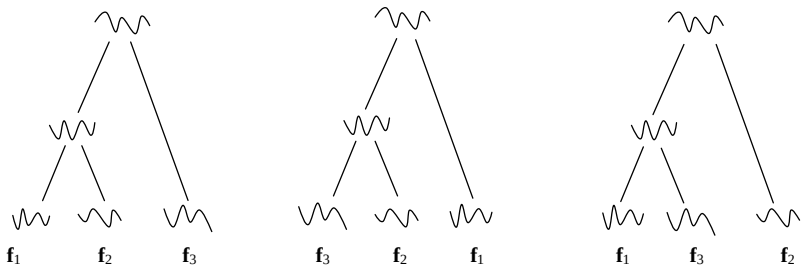

Figure 5.1: Possible Phylogenies with Three Terminal Nodes

As throughout this thesis, the log-likelihood function will be obtained for each of these phylogenies separately, and will be maximised to obtain estimates of the spatial and temporal parameters. Also for each tree the maximum log-likelihood value will be noted. These values are obtained by inserting the estimated parameters into the relevant log-likelihood functions. As the maximum likelihood technique works by selecting values of the parameters that produce a distribution that gives the observed data the greatest probability, then the phylogeny that returns the highest maximum log-likelihood value will provide the estimate of the true topology.

Putting this theory to the test, let us simulate data from a phylogeny, of the form in Figure 5.2, with the following parameters: $\sigma = 1$, $\lambda = 1$ and $\mu = 100$. The temporal parameter has been set to a large number here to try to make sure that, in this test case, information about the true topology has

not been lost by too much decay in temporal correlation. The time points chosen are $t_4 = 1$ and $t_5 = 5$, i.e., $t_4/\mu = 0.01$ and $t_5/\mu = 0.05$.

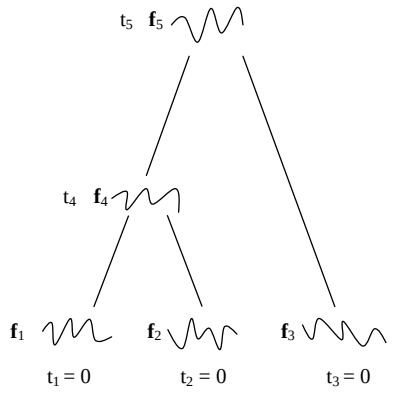

Figure 5.2: Correct Phylogeny

The simulation of this phylogeny requires the following curves to be generated; $\mathbf{f}_5$ (in space only), $\mathbf{f}_4$ from $\mathbf{f}_4|\mathbf{f}_5$, $\mathbf{f}_3$ from $\mathbf{f}_3|\mathbf{f}_5$, $\mathbf{f}_1$ from $\mathbf{f}_1|\mathbf{f}_4$ and $\mathbf{f}_2$ from $\mathbf{f}_2|\mathbf{f}_4$. The process of simulating a phylogeny has been explained in detail in chapter 3, and will not be repeated here. Obviously having simulated this data, we know how these curves relate to one another. However let us presume that all that is known about this phylogeny are those curves in the present time $\mathbf{f}_1$, $\mathbf{f}_2$ and $\mathbf{f}_3$, i.e., the ancestral curves *and* the phylogeny topology are unknown.

As explained above the log-likelihood function will be obtained for each possible topology (Figure 5.1), and will be maximised to obtain estimates of the spatial and temporal parameters, and the maximum log-likelihood value will also be calculated for each topology. Concentrating on the first phylogeny in Figure 5.1, as the ancestral curves are deemed to be unknown then the likelihood function can be obtained as in Chapter 4:

$$L(\boldsymbol{\theta}) = \frac{1}{(2\pi)^{\frac{n}{2}}(\det K_{xt})^{\frac{1}{2}}} \exp\left[-\frac{1}{2}(\mathbf{f}_1^T, \mathbf{f}_2^T, \mathbf{f}_3^T)K_{xt}^{-1}\begin{pmatrix}\mathbf{f}_1\\ \mathbf{f}_2\\ \mathbf{f}_3\end{pmatrix}\right], \qquad (5.1)$$

where $\boldsymbol{\theta} = [\sigma, \lambda, \mu]^T$ and $K_{xt} = K_x \otimes K_t$, where K_x is still the covariance matrix derived from the SE covariance function (2.1) and K_t is the OU process temporal covariance matrix (2.13) (see chapter 4 for details). The log-likelihood function for this phylogeny therefore has the form

$$l_1(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log(\det K_{xt}) - \frac{1}{2}(\mathbf{f}_1^T, \mathbf{f}_2^T, \mathbf{f}_3^T)K_{xt}^{-1}\begin{pmatrix}\mathbf{f}_1\\ \mathbf{f}_2\\ \mathbf{f}_3\end{pmatrix}. \qquad (5.2)$$

By symmetry, the log-likelihood functions for the remaining two phylogenies in Figure 5.1 are

$$l_2(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log(\det K_{xt}) - \frac{1}{2}(\mathbf{f}_3^T, \mathbf{f}_2^T, \mathbf{f}_1^T)K_{xt}^{-1}\begin{pmatrix}\mathbf{f}_3\\ \mathbf{f}_2\\ \mathbf{f}_1\end{pmatrix}, \qquad (5.3)$$

$$l_3(\boldsymbol{\theta}) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log(\det K_{xt}) - \frac{1}{2}(\mathbf{f}_1^T, \mathbf{f}_3^T, \mathbf{f}_2^T)K_{xt}^{-1}\begin{pmatrix}\mathbf{f}_1\\ \mathbf{f}_3\\ \mathbf{f}_2\end{pmatrix}. \qquad (5.4)$$

These log-likelihood functions will not be maximised using the *optim* command in R [33], which has been used up to now, as when the wrong tree structure is being maximised the *optim* command allows the parameter estimate of t_4/μ to be greater than that of t_5/μ, i.e., $t_4 > t_5$. This cannot be allowed to happen as this would imply that some branches within the phylogeny have negative length. Instead a similar R [33] command is used called *constrOptim*, also within the stats package [34], which allows a function to be minimised subject to linear constraints, in this case to minimise the negative log-likelihood subject to $t_4 \leq t_5$. Table 5.1 contains the parameter estimates and maximum log-likelihood values for each of the phylogenies in Figure 5.1.

Table 5.1: Estimates Obtained on the Phylogenies in Figure 5.1

Topology	$\hat{\sigma}$	$\hat{\lambda}$	$\widehat{t_4/\mu}$	$\widehat{t_5/\mu}$	max log.lik.
1. ((1, 2), 3)	**1.095**	**1.048**	**0.011**	**0.057**	**−49.997**
2. ((3, 2), 1)	1.046	1.058	0.049	0.049	−68.646
3. ((1, 3), 2)	1.047	1.058	0.050	0.050	−68.637

From Table 5.1, it would appear that the first phylogeny has the correct topology for these curves as the largest maximum log-likelihood value has been obtained for this one. As this data was simulated by us we have the advanatge of knowing that this is indeed the correct structure. As well as being able to identify the correct structure the spatial and temporal parameters have also been accurately estimated.

Perhaps surprisingly, the spatial parameters for the two wrong topologies have not deviated far from the true parameter values despite the model being wrong. This robustness is good, and probably reflects the fact that most of the information about the spatial structure comes directly from the three observed curves. The temporal parameters returned on the wrong trees do behave badly as expected though, where t_4/μ has been estimated as being larger than the true parameter value, which is due to the curves $\mathbf{f}_3$ and $\mathbf{f}_2$, and $\mathbf{f}_1$ and $\mathbf{f}_3$ being wrongly regarded as being closely related in the second and third phylogeny respectively. As t_4/μ increases, the branch length between $\mathbf{f}_4$ and $\mathbf{f}_5$ shortens, and in fact from the estimates made on these phylogenies the branch is estimated to disappear altogether as $\hat{t_4} = \hat{t_5}$. The similarity between the temporal parameters made on the two wrong phylogenies is due to the symmetry of the two incorrect topologies.

The profile plots for these parameter estimates can be found Figures 5.3 - 5.4. The plots for phylogeny 3 are indistinguishable to those for phylogeny 2. Note that the profile plots for t_4/μ and t_5/μ on the wrong phylogenies appear slightly different. This is due to the constraint put on the temporal

parameters so that t_4 cannot be greater than t_5. The profile plot for t_4/μ is only evaluated up to $t_4/\mu = \widehat{t_5/\mu}$ since $t_4 \leq t_5$ and in fact the maximum is found at $t_4/\mu = \widehat{t_5/\mu}$. Above $\widehat{t_5/\mu}$ higher values of log-likelihood can be found but they correspond to infeasible values since $t_4 > t_5$ and so are irrelevant.

It has been shown how the correct phylogeny structure can be identified when only three curves in the present time are known. However to test this method further one might consider the problem of having more than three curves in the present time, which will increase the search significantly.

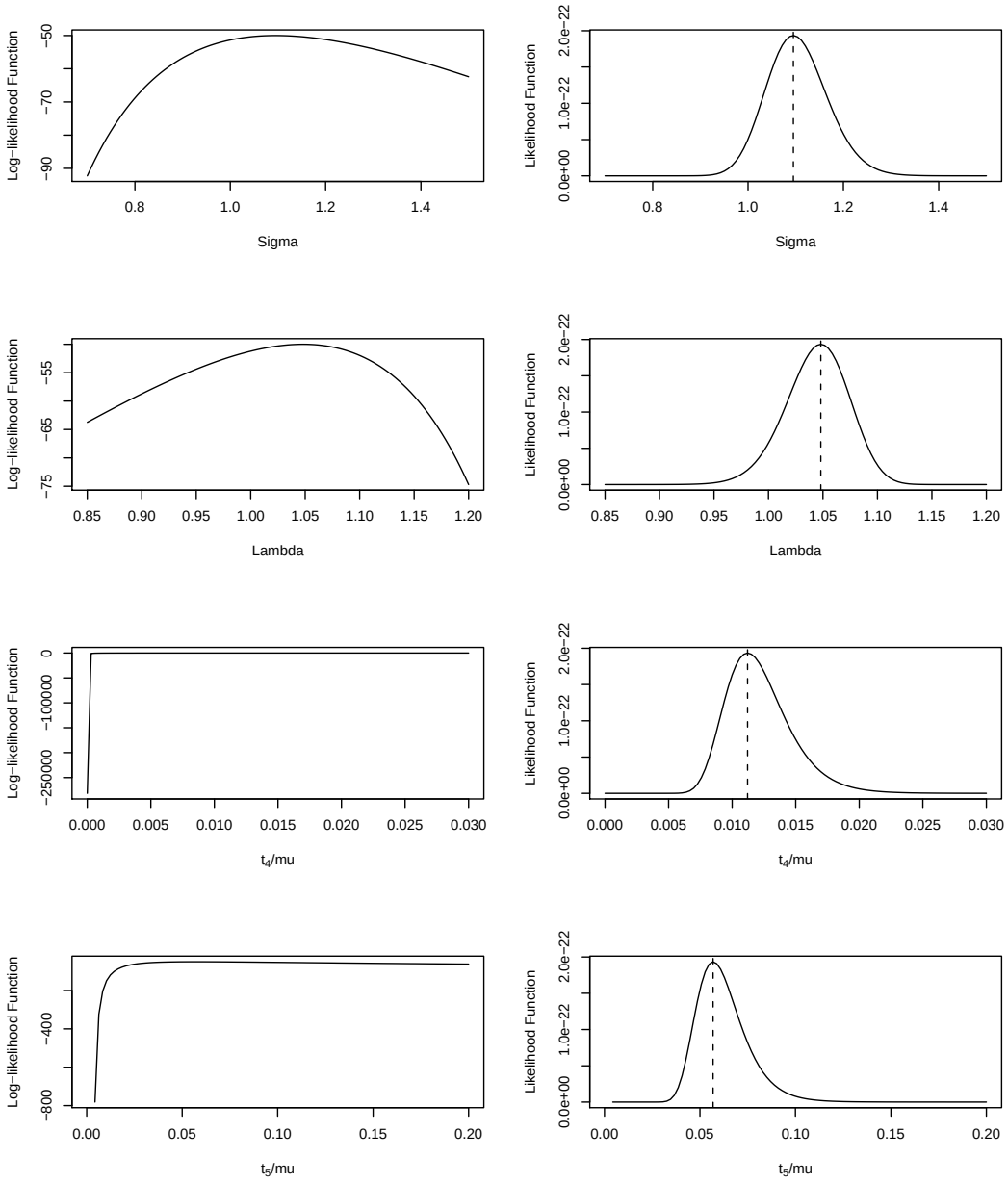

Figure 5.3: Profile Plots for Phylogeny 1 in Table 5.1

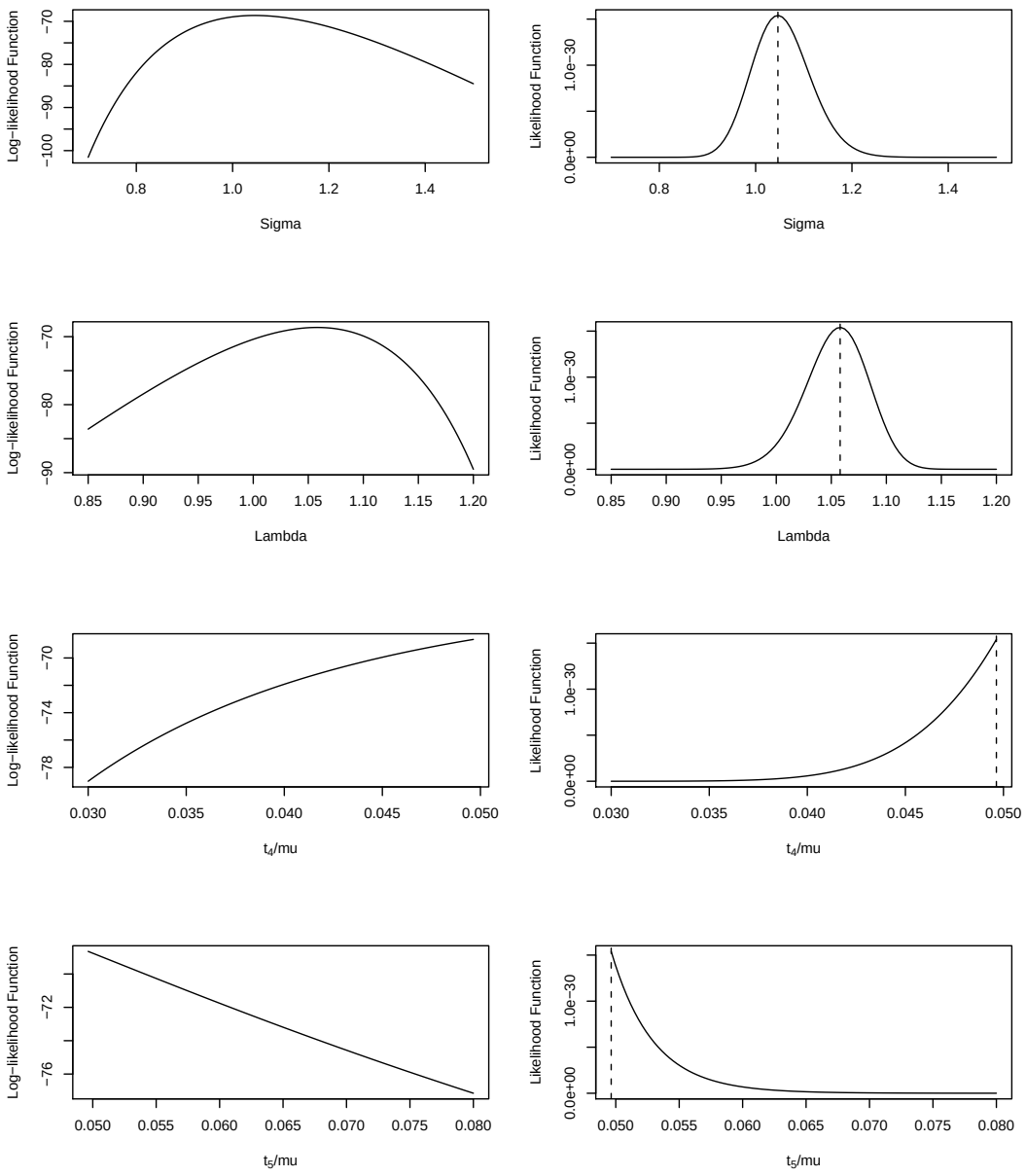

Figure 5.4: Profile Plots for Phylogeny 2 in Table 5.1

5.2 Choosing the Correct Phylogeny with Four Terminal Nodes

Moving onto a more complicated example, we consider having four known curves in the present. By increasing the number of curves in the present time by just one the number of possible topologies has increased to eighteen. Figure 5.5 displays the two basic structures that a phylogeny with four terminal nodes can take.

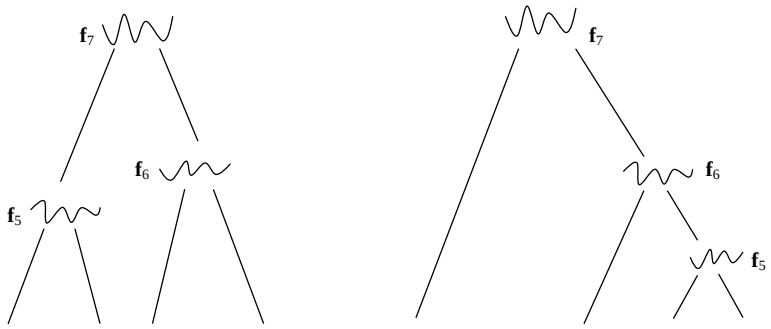

Figure 5.5: Basic Phylogeny Structures with Four Terminal Nodes

For the first phylogeny in Figure 5.5, there are six permutations of how the curves in the present time can be related (note that we are requiring $\mathbf{f}_5$, $\mathbf{f}_6$ and $\mathbf{f}_7$ to be time-ordered, so for example, $\mathbf{f}_1$ and $\mathbf{f}_2$ joining in $\mathbf{f}_5$ is considered distinct from joining them in $\mathbf{f}_6$) and for the second phylogeny there are twelve permutations. As in section 5.1, the log-likelihood function has to be obtained for each of these possible topologies, and these functions maximised to obtain estimates for the spatial and temporal parameters. The maximum log-likelihood value will also be noted for each tree. The log-likelihood functions will not be defined here, but are obtained in the same way as those in section 5.1. Each function will be maximised using the *constrOptim* command, where the constraints that have to be made are $t_5 \leq t_6 \leq t_7$.

We simulate a phylogeny, of the form in Figure 5.6, with the following

parameters: $\sigma = 1$, $\lambda = 1$ and $\mu = 100$. The time points chosen are $t_5 = 1$, $t_6 = 5$ and $t_7 = 9$.

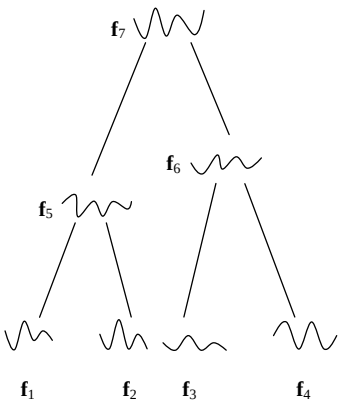

Figure 5.6: Correct Phylogeny with Four Terminal Nodes

Again only those four curves in the present time will be regarded as known. Table 5.2 contains the parameter estimates and maximum log-likelihood values obtained on each possible topology.

Again the phylogeny with the highest maximum log-likelihood value will be regarded as the correct phylogeny structure. This turns out to be the first phylogeny in Table 5.2, which is indeed the correct phylogeny structure. Also the spatial and temporal parameter estimates returned are very good. The spatial parameter estimates returned on the wrong phylogenies again never deviate far from the true parameter values, and the temporal parameter estimates returned on the wrong phylogenies behave poorly as expected.

Table 5.2: Estimates Obtained on Those Phylogenies in Figure 5.5

Topology	$\hat{\sigma}$	$\hat{\lambda}$	$\widehat{t_5/\mu}$	$\widehat{t_6/\mu}$	$\widehat{t_7/\mu}$	max log.lik.
1. ((1, 2), (3, 4))	**1.030**	**0.982**	**0.008**	**0.051**	**0.092**	**−67.773**
2. ((1, 3), (2, 4))	1.053	0.997	0.068	0.068	0.068	−109.025
3. ((1, 4), (3, 2))	1.035	0.998	0.073	0.073	0.073	−108.891
4. ((2, 3), (1, 4))	1.022	1.003	0.078	0.078	0.078	−109.032
5. ((2, 4), (1, 3))	1.021	0.996	0.076	0.076	0.076	−108.943
6. ((3, 4), (1, 2))	1.020	0.976	0.029	0.029	0.093	−109.025
7. (1, (2, (3, 4)))	1.034	0.989	0.043	0.073	0.073	−94.287
8. (1, (3, (2, 4)))	1.050	0.997	0.069	0.069	0.069	−108.866
9. (1, (4, (2, 3)))	1.066	1.005	0.068	0.068	0.068	−105.103
10. (2, (1, (3, 4)))	1.023	0.998	0.045	0.077	0.077	−94.194
11. (2, (3, (1, 4)))	1.000	0.996	0.080	0.080	0.080	−108.866
12. (2, (4, (1, 3)))	1.020	0.995	0.075	0.075	0.075	−108.922
13. (3, (1, (2, 4)))	1.066	1.004	0.051	0.051	0.086	−92.864
14. (3, (2, (1, 4)))	1.021	1.002	0.058	0.058	0.095	−94.289
15. (3, (4, (1, 2)))	1.039	0.989	0.008	0.077	0.079	−70.556
16. (4, (1, (2, 3)))	1.002	0.999	0.082	0.082	0.082	−109.166
17. (4, (2, (1, 3)))	1.004	0.994	0.079	0.079	0.079	−109.024
18. (4, (3, (1, 2)))	1.049	0.989	0.008	0.076	0.076	−70.781

5.3 Summary

The method of maximum likelihood was used to obtain spatial and temporal parameter estimates for the interior/ancestral curves for every possible tree topology, and the maximum log-likelihood value was calculated. The phylogeny with the highest maximum log-likelihood value was proposed as the estimate of the true topology. This worked well in a couple of simple

examples.

The obvious limitation to this approach is the number of possible topologies over which the likelihood has to be maximised. As shown in this chapter, by increasing the number of curves in the present time from three to four the number of possible phylogeny structures jumps from three to eighteen. In general, the number of topologies increases very rapidly with the number of tips. In practice, exhaustive optimisation over all topologies would not be feasible.

Chapter 6

Conclusion & Discussion

The aim of this thesis has been to explore the placing of GPs on phylogenies of functional data in order to perform evolutionary inferences. This began with single lineages in Chapter 2. There, it was shown how evolving curves could be easily simulated from a GP prior when an appropriate space-time covariance function was specified. The spatial and temporal aspects of the simulated GP curves were completely controlled by the parameters within these space-time covariance functions. To learn about these, it then became a simple case of estimating them by the method of maximum likelihood. Also covered in this chapter were the inferences computed on the experimental Chinese whispers curves. There were signs of a lack of fit of the experimental curves to a GP model with a squared exponential spatial covariance function. The OU process temporal covariance function was also introduced in this chapter, the Markov property of which helped greatly in lightening the load involved in simulating and fitting the temporal process. The inference carried out on the known phylogenies in Chapter 3 was a natural extension of the work completed on the single lineages. Phylogenies were generated by simulating curves at each of the nodes in the tree topology. These curves were simulated from a GP prior with an appropriate space-time covariance function. The likelihood of the parameters in the covariance function and the branch lengths were maximised numerically. Importantly, the branch length

parameter estimates returned were good. As these parameters determine the strength of the relationships between the curves in the phylogeny, it is important that these are well estimated.

The unknown ancestor problem (Chapter 4) treated the case where data at the interior nodes were discarded. Due to the marginalisation property of GPs, the likelihood corresponding to data only at the tips of the tree was relatively straightforward to derive. This was optimised to obtain estimates for the spatial and temporal parameters. An unknown tree topology for data curves in the present time is the most general and useful real-life phylogenetics problem. To identify the correct topology (Chapter 5), all possible relationships between the curves were considered. By maximising the log-likelihood for each of these possible phylogenies, then the correct phylogeny was able to be identified as that with the highest maximum log-likelihood value. Spatial and temporal parameter estimates were also computed.

Despite the GP framework reducing the computational load greatly for evolutionary inferences there are still some limitations to this method. The first limitation occurs when there is very little correlation between curves. This can happen as a direct result of there being large time differences between curves and/or the temporal-correlation parameter μ being small. The method of maximum likelihood estimation relies on there being a sufficient level of correlation between the curves in order to make good parameter estimates. Therefore, if the correlation between curves in a phylogeny is low then potentially the maximum likelihood method and indeed any method may struggle to return parameter estimates that are close to the correct parameter values. The second limitation is in the unknown phylogeny method. As the number of curves in the present time increases the number of possible phylogeny structures increases rapidly, so much so that to maximise the log-likelihoods for each possible phylogeny structure individually becomes an almost impossible task. To get round this problem, more heuristic searches, that are not exhaustive, through the space of different tree topologies, would

need to be employed, as is routine in other areas of phylogenetics. Much can probably be learned from these existing approaches.

Future work would include developing a test similar to the likelihood ratio test for the phylogeny problem. This would allow one to assess formally whether the phylogeny with the highest maximum log-likelihood value is a significantly better fit to the data than other phylogenies. Unfortunately the likelihood ratio test itself cannot be used for this as this test requires the models to be nested. But p-values might be estimated by simulation. Alternatively, in a Bayesian approach, one would need to calculate Bayes factors for the different tree topologies. Other future work would involve applying the regression, ancestor and phylogeny problems to experimental data, perhaps of the Chinese whisper variety. It would also be worthwhile to explore other covariance functions that could better model the single lineage experimental curves in Appendix A. The covariance function of a GP can take many forms, so one would hope that these curves could be modelled by some GP, and then evolutionary inferences could be made. Also all curves analysed in this thesis are rather straightforward looking curves but in reality this might not be the case. For example, real curves might have loops or not be single-valued. It would therefore be of interest to develop techniques to analyse curves with a more complicated structure in order to see how the methods proposed in this thesis would perform.

Throughout this thesis it has been shown how the GP framework significantly reduces the workload required to make evolutionary inferences on phylogenies of functional data. The fact that GPs are completely defined by their mean and covariance functions has been their most advantageous property, as then the simulated Gaussian curves were encoded by the parameters within these functions which could be rather easily estimated. Also, by the use of GPs, curves in space and time could be easily modelled by a simple change in the covariance function. Overall GPs have proved to be a promising way forward in tackling a very long-standing problem: how to best make

use of continuous data in an evolutionary inference context.

Appendix A

Chinese Whispers Curves

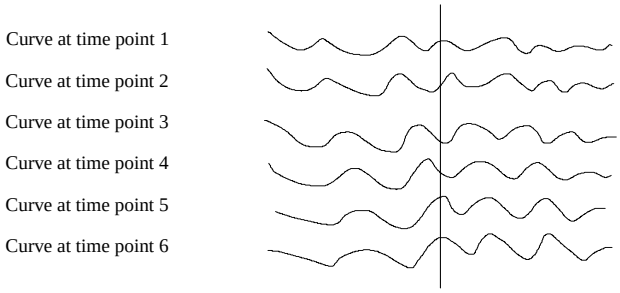

Figure A.1: First Single Lineage of Curves

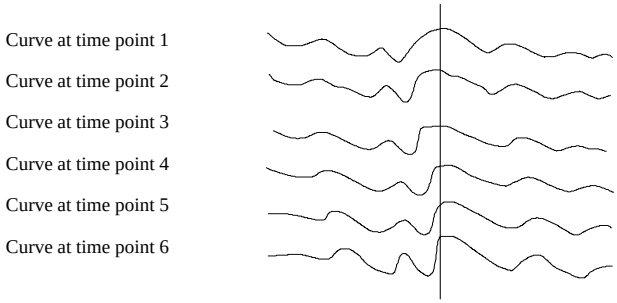

Figure A.2: Second Single Lineage of Curves

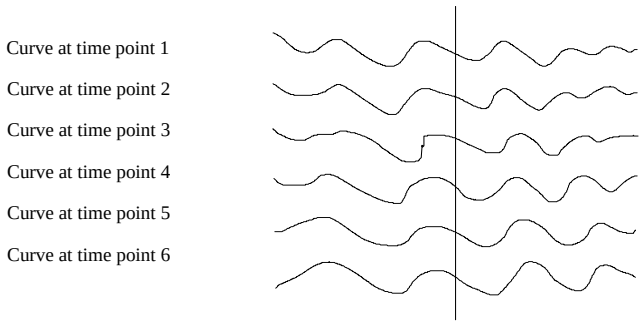

Figure A.3: Third Single Lineage of Curves

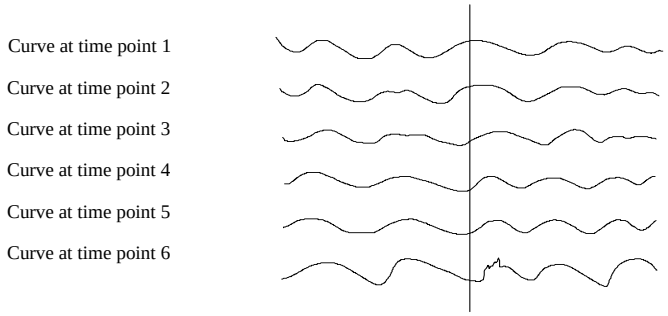

Figure A.4: Fourth Single Lineage of Curves

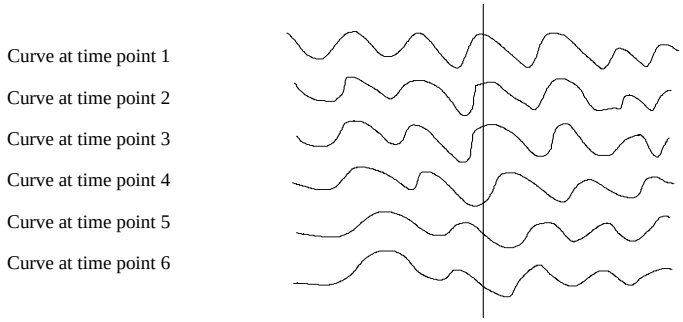

Figure A.5: Fifth Single Lineage of Curves

Appendix B

Inference on Chinese Whispers Curves

Parameter estimates and approximate 95% CIs obtained on the first three curves on the lineage in Figure A.2 are:

$\hat{\sigma} = 21.114$ (17.965, 24.263)

$\hat{\lambda} = 12.704$ (11.882, 13.527)

$\hat{\mu} = 3.613$ (1.793, 5.433)

Figure B.1 displays the first three curves from the single lineage in Figure A.2

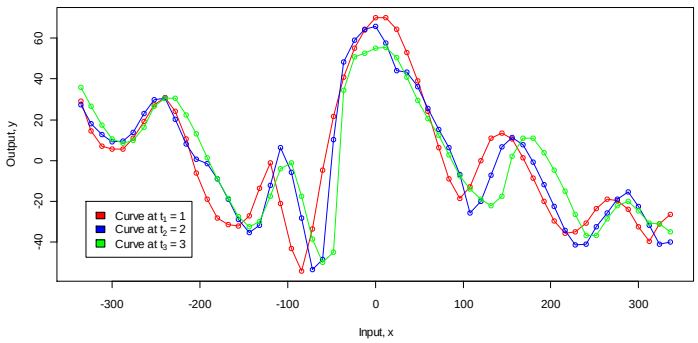

Figure B.1: First Three Curves from Figure A.2

and Figure B.2 displays three curves simulated from a GP prior with the above parameter estimates.

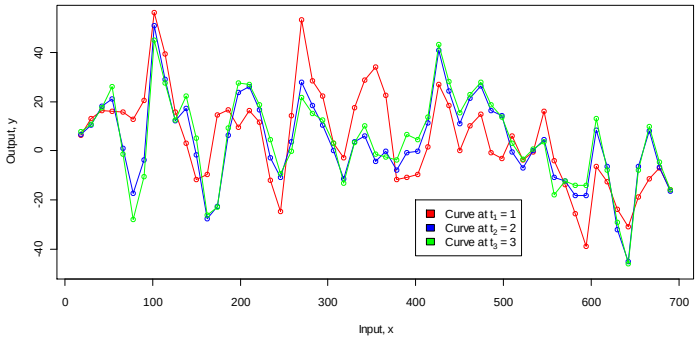

Figure B.2: Simulation of the curves in Figure B.1

Figure B.3 contains the estimated covariance plots for the simulated ancestral curve and the Chinese whisper ancestral curve.

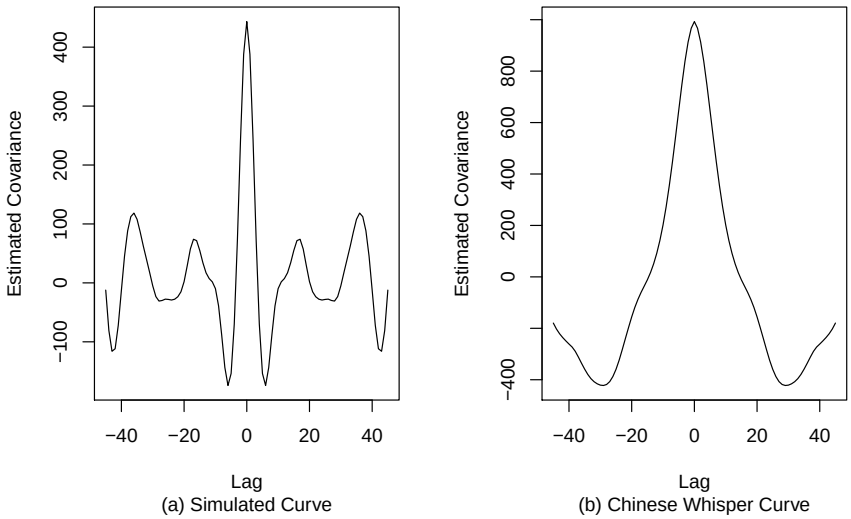

Figure B.3: Estimated Covariances for the Ancestral Curves in Figures B.1 and B.2

Parameter estimates and approximate 95% CIs obtained on the first three curves on the lineage in Figure A.3 are:

$\hat{\sigma} = 21.328$ (18.443, 24.214)

$\hat{\lambda} = 13.689$ (13.097, 14.281)

$\hat{\mu} = 1.215$ (0.647, 1.783)

Figure B.4 displays the first three curves from the single lineage in Figure A.3

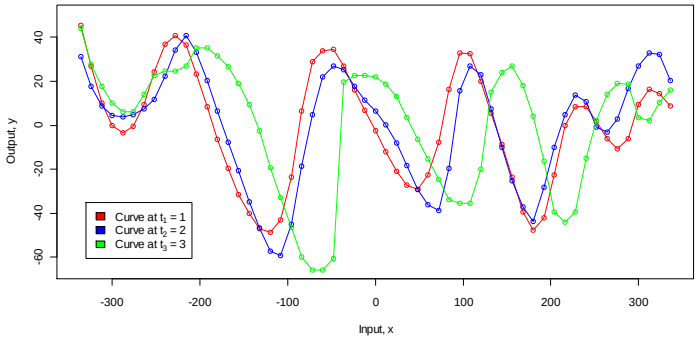

Figure B.4: First Three Curves from Figure A.3

and Figure B.5 displays three curves simulated from a GP prior with the above parameter estimates.

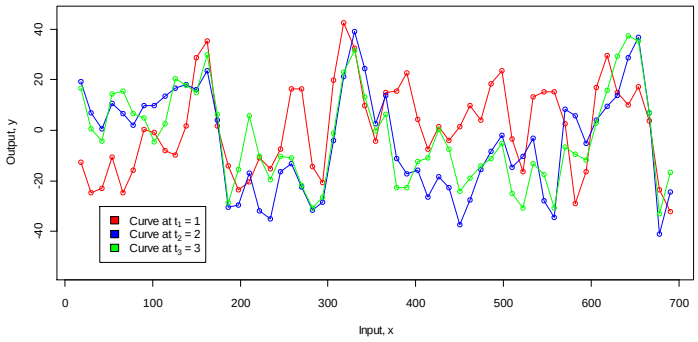

Figure B.5: Simulation of the curves in Figure B.4

Figure B.6 contains the estimated covariance plots for the simulated ancestral curve and the Chinese whisper ancestral curve.

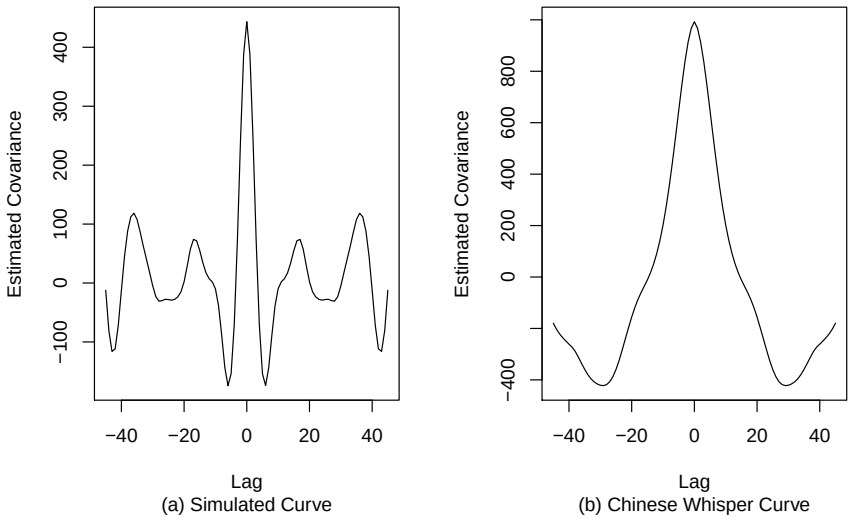

Figure B.6: Estimated Covariances for the Ancestral Curves in Figures B.4 and B.5

Parameter estimates and approximate 95% CIs returned on the curves at the first three time points on the lineage in Figure A.4 are:

$\hat{\sigma} = 12.176$ (10.577, 13.776)

$\hat{\lambda} = 15.462$ (14.890, 16.035)

$\hat{\mu} = 1.334$ (0.711, 1.957)

Figure B.7 displays the first three curves from the single lineage in Figure A.4

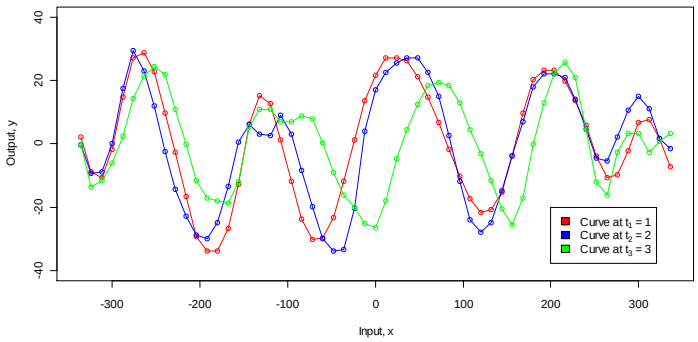

Figure B.7: First Three Curves from Figure A.4

and Figure B.8 displays three curves simulated from a GP prior with the above parameter estimates.

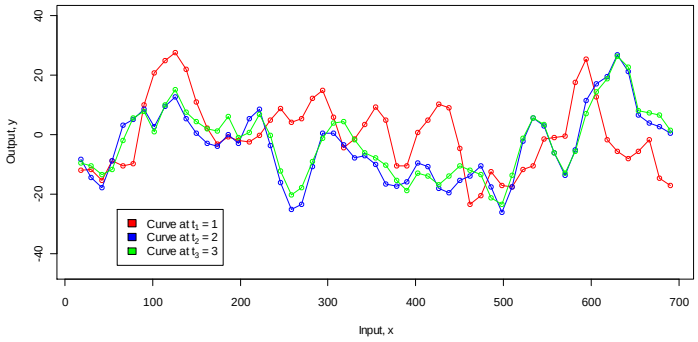

Figure B.8: Simulation of the curves in Figure B.7

Figure B.9 contains the estimated covariance plots for the simulated ancestral curve and the Chinese whisper ancestral curve.

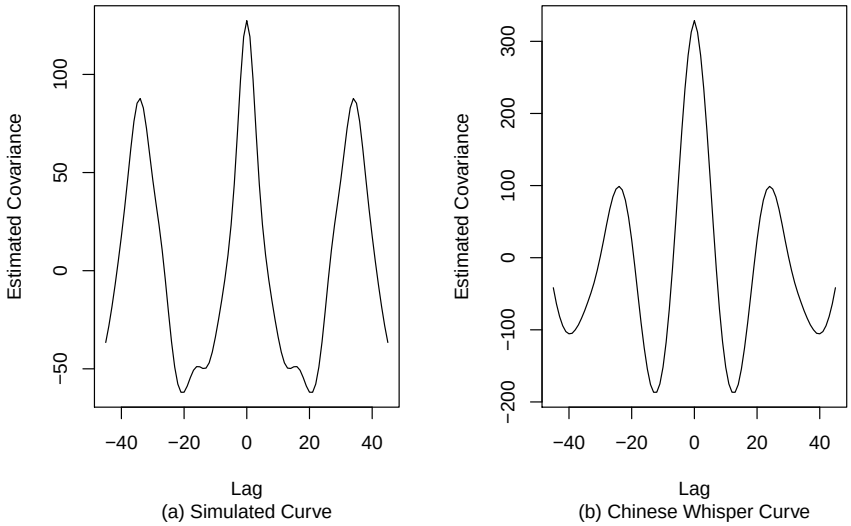

Figure B.9: Estimated Covariances for the Ancestral Curves in Figures B.7 and B.8

Parameter estimates and approximate 95% CIs returned on the curves at the first three time points on the lineage in Figure A.5 are:

$\hat{\sigma} = 27.632$ (23.811, 31.649)

$\hat{\lambda} = 12.730$ (11.553, 13.192)

$\hat{\mu} = 2.581$ (1.374, 3.858)

Figure B.10 displays the first three curves from the single lineage in Figure A.5

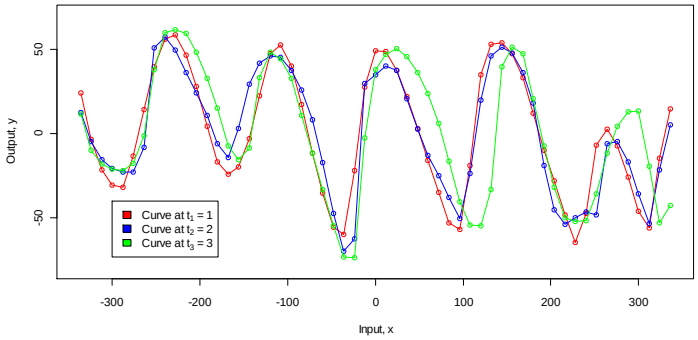

Figure B.10: First Three Curves from Figure A.5

and Figure B.11 displays three curves simulated from a GP prior with the above parameter estimates.

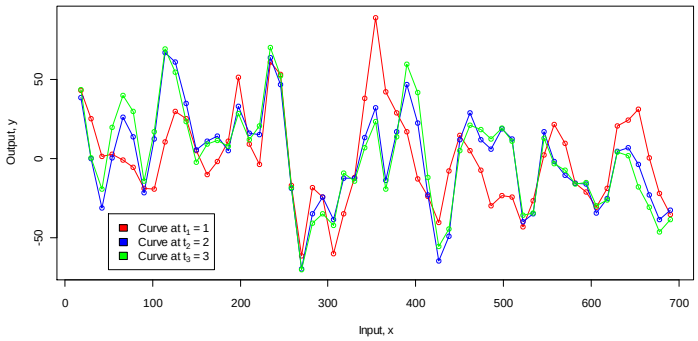

Figure B.11: Simulation of the curves in Figure B.10

Figure B.12 contains the estimated covariance plots for the simulated ancestral curve and the Chinese whisper ancestral curve.

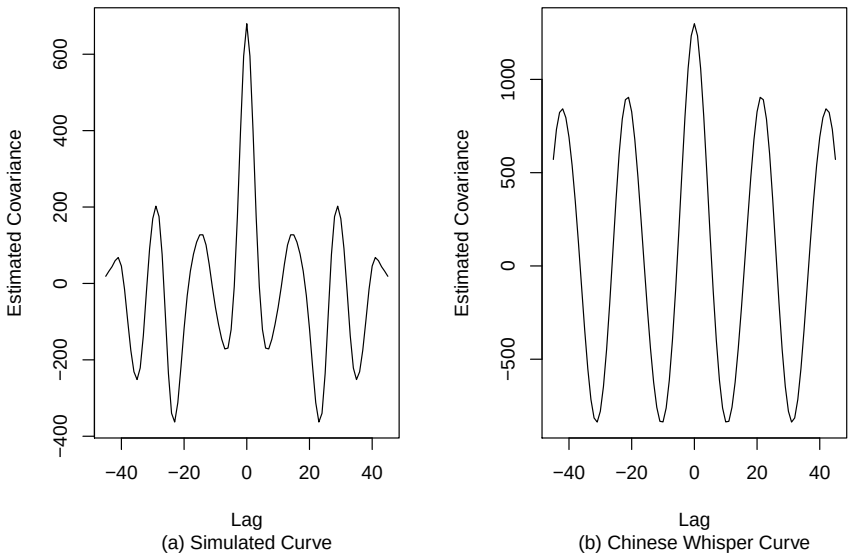

Figure B.12: Estimated Covariances for the Ancestral Curves in Figures B.10 and B.11

Bibliography

[1] Berger, S.A., (2010) Accuracy of Morphology-Based Phylogenetic Fossil Placement under Maximum Likelihood. *ACS/IEEE International Conference on Computer Systems and Application - AICCSA*, pp. 1-9.

[2] Blumenthal, B.M., and Getoor, R.K., (2007) Markov Processes and Potential Theory. *Dover Publications: NY.*

[3] Camin, J.H., and Sokal, R.R., (1965) A Method for Deducing Branching Sequences in Phylogeny. *J. Evolution*, 19(3): pp. 311-326.

[4] Clifton, P.M., and Neuman, S.P., (1982) Effects of Kriging and Inverse Modelling on Conditional Simulation of the Aura Valley Aquifer in Southern Arizona. *J. Water Resources Research*, 18(4): pp. 1215-1234.

[5] Edwards, A.W.F., and Cavalli-Sforza, L.L., (1964) Phenetic and Phylogenetic Classifiction. *J. Systematics Association*, 6: pp. 67-76.

[6] Felsenstein, J., (1973) Maximum Likelihood and Minimum-Steps Methods for Estimating Evolutionary Trees from Data on Discrete Characters. *J. Systematic Zoology*, 22: pp. 240-249.

[7] Felsenstein, J., (1978) Cases in which Parsimony or Compatibility Methods will be Positively Misleading. *J. Systematic Zoology*, 27: pp. 401-410.

[8] Felsenstein, J., (1981) Evolutionary Trees from DNA Sequences: A Maximum Likelihood Approach. *J. Molecular Evolution*, 17: pp. 368-376.

[9] Felsenstein, J., (1985) Phylogenies and the Comparative Method. *J. The American Naturalist*, 125(1): pp. 1-15.

[10] Felsenstein, J., (1988) Phylogenies and Quantitative Characters. *Annual Review of Ecology and Systematics*, 19: pp. 445-471.

[11] Felsenstein, J., (2003) Inferring Phylogenies (2nd ed.). *Sinauer Associates: MA*.

[12] Ferraty, F., and Vieu, P., (2006) Nonparametric Functional Data Analysis: Theory and Practice. *Springer: New York*.

[13] Ferris, B., Hähnel, D., and Fox, D., (2006) Gaussian Processes for Signal Strength-Based Location Estimation. *In Proc. of Robotics Science and Systems*.

[14] The Functional Phylogenies Group (2012) Phylogenetic Inference for Function-valued Traits: Speech Sound Evolution. *J. Trends in Ecology and Evolution*, 27(3): pp. 160-166.

[15] Gestel, T.V., Suykens, J.A.K., Lanckriet, G., Lambrechts, A., De Moor, B., and Vandewalle, J., (2006) Bayesian Framework for Least-Squares Support Vector Machine Classifiers, Gaussian Processes, and Kernal Fisher Discriminant Analysis. *J. Neural Computation*, 14(5): pp. 1115-1147.

[16] Green, P.J., (1995) Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination. *J. Biometrika*, 82: pp. 711-732.

[17] Hastings, W.K., (1970) Monte Carlo Sampling Methods using Markov Chains and their Applications. *J. Biometrika*, 57: pp. 97-109.

[18] Hendrixson, B.E., and Bond, J.E., (2009) Evaluating the Efficiency of Continuous Quantitatative Characters for Reconstructing the Phylogeny

of a Morphologically Homogeneous Spider Taxon (Araneae, Mygalomorphae, Antrodiaetidae, Antrodiaetus). *J. Molecular Phylogenetics and Evolution*, 53(1): pp. 300-313.

[19] Huelsenbeck, J.P., and Ronquist, F., (2001) MRBAYES: Bayesian Inference of Phylogenetic Trees. *Bioinformatics Applications Note*, 17(8): pp. 754-755.

[20] Jones, N., and Moriarty, J., (2011) Evolutionary Inference for Function-valued Traits: Gaussian Process Regression on Phyloenies. *arXiv:1004.4668v2*.

[21] Kishino, H., and Hasegawa, M., (1989) Evaluation of the Maximum Likelihood Estimate of the Evolutionary Tree Topologies from DNA Sequence Data, and the Branching Order in Hominoidea. *J. Molecular Evolution*, 29(2): pp. 170-179.

[22] Larget, B., and Simon, D., (1999) Markov Chain Monte Carlo Algorithms for the Bayesian Analysis of Phylogenetic Trees. *J. Molecular Biology and Evolution*, 16: pp. 750-759.

[23] Li, S., (1996) Phylogenetic Tree Construction using Markov-Chain-Monte-Carlo. *PhD Dissertation, Ohio State University, Columbus*.

[24] Macholan, M., (2008) The Mouse Skull as a Source of Morphometric Data for Phylogeny Inference. *J. Zoologischer Anzeiger*, 247(4): pp. 315-327.

[25] MacLeod, N., and Forey, P.L., (eds) (2002) Morphology, Shape and Phylogeny. *Taylor and Francis: London*.

[26] Martins, E.P., and Hansen, T.F., (1997) Phylogenetics and the Comparative Method: A General Approach to Incorporating Phylogenetic Information into the Analysis of Interspecific Data. *J. The American Naturalist*, 149(4): pp. 646-667.

[27] Mau, B., and Newton, M., (1997) Phylogenetic Inference for Binary Data on Dendrograms using Markov Chain Monte Carlo. *J. Computational and Graphical Statistics*, 6: pp. 122-131.

[28] Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A.H., and Teller, E., (1953) Equations of State Calculations by Fast Computing Machines. *J. Chemical Physics*, 21(6): pp. 1087-1092.

[29] O'Meara, B.C., Ané, M.J., Sanderson, M.J., and Wainwright, P.C., (2006) Testing for Different Rates of Continuous Trait Evolution using Likelihood. *J. Evolution*, 60(5): pp. 922-933.

[30] Ramsay, J.O., and Silverman, B.W., (2005) Functional Data Analysis (2nd ed.). *Springer: Berlin*.

[31] Rannala, B., and Yang, Z., (1996) Probability Distribution of Molecular Evolutionary Trees: A New Method of Phylogenentic Inference. *J. Molecular Evolution*, 43: pp. 304-311.

[32] Rasmussen, C., and Williams, C., (2006) Gaussian Processes for Machine Learning. *MIT: MA*.

[33] R Development Core Team (2005) R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing: Vienna, Austria*. ISBN 3-900051-07-0. http://www.R-project.org.

[34] Ripley, B.D., R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing*. ISBN 3-900051-07-0. http://www.R-project.org/.

[35] Schwaighofer, A., Grigoras, M., Tresp, V., and Hoffman, C., (2004) GPPS: A Gaussian Process Positioning System for Cellular Networks. *In Advances in Neural Information Processing Systems (NIPS), MIT*.

[36] Sminchisescu, C., Kanaujia, A., Li, Z., and Metaxas, D., (2005) Discriminataive Density Propagation for 3-D Human Motion Estimation.

In Proc. IEEE Conf. Conputing Vision Pattern Recognition, 1: pp. 390-398.

[37] Thomas, M.G., Hagelberg, E., Jone, H.B., Yang, Z., and Lister, A.M., (2000) Molecular and Morphological Evidence on the Phylogeny of the Elephantidae. *J. Biological Science*, 267(1461): pp. 2493-2500.

[38] Tomasi, C., Petrov, S., and Sastry, A., (2003) 3D Tracking = Classification + Interpolation. *In Proc. IEEE Int. Conf. Computing Vision*, 2: pp. 1441-1448.

[39] Urtasun, R., and Darrell, T., (2008) Local Probabilistic Regression for Activity-Independent Human Pose Inference. *In. Proc. IEEE Conf. Computing Vision Pattern Recognition*, 2: pp. 1-8.

[40] Urtasun, R., Fleet, D.J., Hertzmann, A., and Fua, P., (2005) Priors for People Tracking from Small Training Sets. *In Proc. IEEE Int. Conf. Computing Vision*, 1: pp. 403-410.

[41] Venables, W.N., and Ripley, B.D., (2002) Modern Applied Statistics with S (4 ed.). *Springer: NY* ISBN 0-387-95457-0. http://www.stats.ox.ac.uk/pub/MASS4

[42] Wiens, J.J., (2001) Character Analysis in Morphological Phylogenetics: Problems and Solutions. *J. Systematic Biology*, 50(5): pp. 689-699.

[43] Yang, Z., and Rannala, B., (1997) Bayesian Phylogenetic Inference using DNA Sequences: A Markov Chain Monte Carlo Method. *J. Molecular Evolution*, 14: pp. 717-724.

[44] Zhao, X., Fu, Y., and Liu, Y., (2011) Human Motion Tracking by Temporal-Spatial Local Gaussian Process Experts. *IEEE Transactions on Image Processing*, 20(4): pp. 1141-1151.

[45] http://genetics.agrsci.dk/sorenh/public/FDA2008/iesbs.pdf