Gaussian Process Arc Lengths, Functional Regression and Applications

Justin Dragon Bewsher

St Peter's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2017

Acknowledgements

First, I would like to thank my supervisors Michael Osborne and Stephen Roberts. They kindly took me on board and patiently allowed me to explore my research topic, letting it evolve naturally throughout my studies. Your guidance, insight and patience has been invaluable.

To all those who have been in the lab with me, from the start and in the years in-between, thanks for keeping me company, bouncing ideas and being there for lunch; Ali Rizvi, Tom Nickson, Tom Gunter, Ahsan Alvi, Rory Beard, Chris Lloyd, Logan Graham, Jack Fitzsimons, Bernardo Perez Orozco, and Nafisa Sharif.

A special thank you to my collaborator Alessandra Tosi: it was a pleasure and delight working through the science of the arc length paper with you.

Thanks to my first Oxford family at St Hughs, who helped me settle in and find my place in Oxford, especially Josh Cowls, Liam Mulroy and Alex Salam. To those at St Peters, Rose de Geus, Genevieve Martin and Elise Maes; thank you for the lunches, brunches and wonderful support.

A big part of what kept me sane during my Oxford experience was fencing, so I would be remiss not to mention the fencer. Thank you to Harriet Dixon, Trevelyan Wing, Will Halliwell, Alex Robinson, Antoine Grey, Doga Basaran, Sean Jamshidi and Samuel Bradley. Big thanks to my coach Tomek Walicki who honed my skills, and my friend-coach, who patiently waited for me to finish, Steven Williams.

Thank you to Diane Bouchacourt, for the adventures, always being hungry and the great friendship. A heartfelt thanks to Jonathan Downing, who supported me academically in my research, through the challenging last months and has the kindest of hearts. Thank you to Elmarie van Heerden for keeping me company, enjoying great tunes and spurring me on with unrelenting stubbornness and kindness: the lab would not have been the same without you.

Thank you to Victoria Cox for her unwavering belief in me and showing me anything is possible. Those final months would have been all the much harder without your support.

Finally, thank you to my parents. You have let me find my own path and have been proud of me at every step. Thanks to my dad, Ian Bewsher, for nourishing my love of maths and science from a young age. Lastly a huge thanks to my amazing and wonderful mum, Jacqueline Swift, who loved and cherished me, encouraged and supported me throughout my DPhil; and who proof read this thesis despite being unable to understand the mathematics.

Abstract

This thesis presents novel functional regression methods for non-linear data and develops core Gaussian Process theory for arc lengths. The need for richer functional regression methods and theory around Gaussian Process arc lengths is established in the introduction.

Chapter 1 commences with the requisite background material and Chapter 2 works through the fundamentals of probability, highlighting key messages with a worked example. Gaussian Processes are introduced in Chapter 3, developing the core theory needed to extend the functional methods outlined in Chapter 4 and to prepare for considering the arc length distribution of Gaussian Processes.

Three novel Gaussian Process functional regression models are introduced in Chapter 5, building upon existing models and furthering the probabilistic approach to functional regression, namely the Gaussian Process Functional Index Model, the Gaussian Process Functional Additive Model and the Gaussian Process Functional Generalised Additive Model. In each model a kernel function is derived whilst outlining inference and prediction. It is clearly demonstrated that the functional regression models out perform their competitors on synthetic examples.

In Chapter 6, the arc length distribution of a Gaussian Process is derived. By tackling the arc length integrand, a novel approach to the one dimensional arc length mean is presented. This equips us with the tools to approximate the arc length of a vector Gaussian Process. We derive the arc length distribution for a prior and posterior Gaussian Process in terms of its kernel functions and hyperparamaters.

Chapter 7 signals a return to functional regression where two challenging real world functional data sets are tackled. The novel Gaussian Process methods provide competitive results in a number of cases, supporting the potential for the new methods.

The thesis concludes with a vision for future work.

Contents

List of Figures

Notation

Symbol	Description
x	a single number.
$\mathbf{x}$	a vector.
$\mathbf{X}$	a matrix.
x_k	the kth element of $\mathbf{x}$.
s	the arc length of a curve.
$f(\cdot)$	a function.
$X(t)$	a functional input on the domain $\mathcal{I}$.
k	a kernel function.
K	a kernel matrix.
$p()$	a probability density function.

Chapter 1

Introduction

1.1 Preamble

Probabilistic approaches to quantitative data analysis are at the core of modern machine learning. They enable us to move from mere point estimates to expressing predictions as distributions. These distributions capture our uncertainty, providing us with confidence bands that permit robust predictions, whilst simultaneously reducing significant error due to unfounded overconfidence. Quantification of uncertainty also clarifies where and how our assumptions fail, thereby enhancing our understanding of the world we are modelling. The need to move from point estimates to full distributions drives the need for probabilistic modelling and as we believe in the power of data to drive our modelling, a non-parametric approach is favoured.

The Gaussian Process (GP) is the probabilistic model that is the focus of this thesis (Rasmussen and Williams, 2006). Gaussian Processes are a state-of-the-art, ubiquitous tool in machine learning, that provides a flexible non-parametric approach to non-linear data modelling. Gaussian Processes have been used in a number of machine learning problems, including latent variable modelling (Lawrence, 2004), dynamical time-series modelling (Wang et al., 2005) and Bayesian optimisation (Snoek

et al., 2012). Though much attention has been given to the application of GPs, there are still theoretical areas that have yet to be addressed, and, two aspects that have received minimal attention are: functional regression with GPs and the arc length of a GP. This thesis is concerned with modelling and exploring the statistical features of curves using GPs: we explore functional data models using GPs and the arc length of a GP. Importantly, these GP tools and insights can be leveraged to aide the advancement, progress and development of machine learning.

1.2 Problem Statement

Typically, modelling with a Gaussian Process focusses on mapping real-valued inputs to outputs (Rasmussen and Williams, 2006). Instead, consider the problem of functional regression, where the inputs are partially-observed functions. That is, we are interested in situations in which we map observations of some underlying, possibly stochastic, process, to a real valued output.

Functional regression emerges from many fields of science and engineering and is becoming increasingly prevalent. For example, the problem of state of charge measurements of lithium-ion batteries can be framed as functional regression, where the response is the state of charge and the input is the impedance spectra obtained for a range of excitation frequencies (Andre et al., 2011, Xu et al., 2013). In chemometrics, the functional response is the prediction of a chemical variable on the basis of a digitized signal such as near infrared reflectance spectroscopic information (Aguilera et al., 2013), whilst diffusion tensor imaging profiles can be considered functional predictors for performance on an auditory memory test (Goldsmith et al., 2011a). Other examples include the prediction of temperature across Canadian weather stations (Ramsay and Silverman, 2005a), classification of the lifetime of medflies (Müller and Stadtmüller, 2005), and analysis of the relationship between log-spectra of sequences

of spoken syllables and phoneme classification (Cao and Fan, 2009).

Extensive research has been done in the field of functional regression (Acharyya and Ghosh, 2015, Cardot and Sarda, 2005, Fan et al., 2014, James, 2002, Müller and Stadtmüller, 2005, Shi et al., 2007). Functional data analysis includes, but is not limited to, modelling of curve dependence on covariates, deviation of curves from an underlying mean function, and prediction in regression and classification problems (Ramsay and Silverman, 2005a). Convergence rates of models is a common focus of research (Li and Hsing, 2007, Müller and Stadtmüller, 2005), and the resulting estimation methods needed to determine the functional coefficients (Cardot and Sarda, 2005).

The range of potential application domains is limited by a lack of non-linear models functional regression methods (Morris, 2015, Wang et al., 2016), and underutilisation of non-parametric methods (Ferraty and Vieu, 2006). Bayesian treatments of functional regression have received only limited attention, and the models proposed predominantly focus on the linear case (Crainiceanu and Goldsmith, 2010, James, 2002). Whilst Shi et al. (2007) use GPs, they focus only on the functional concurrent model (Maity, 2017), and do not address the function to scalar problem. In Wang et al. (2017) the authors use a functional metric as a GP distance measure to directly map the functional predictors to a scalar. There is significant scope for enriching the literature on functional regression by the use of flexible, probabilistic models, namely Gaussian Processes.

A measure of distance, under represented in functional regression problems, is the length of the curve. For a parametrised curve, the arc length of a curve between two points is an intrinsic and fundamental property. Furthermore, when considering properties of functions, number of turning points, infimum or supremum, to measure their distance, arc length is under represented. Arc length is an important mathematical feature of the world, allowing us to measure space, distance, quantity used

and energy required: in many physical systems, or optimisation problems, arc length is the critical issue. The cutting edge of science research is driven by our ability to precisely measure the world (Abbott and et Al, 2016), supporting the idea that the ability to better determine lengths, and understand our uncertainty around them, has great significance in many fields, including functional regression.

The question of uncertainty drives a consideration of the length of a random curve and it is important to consider: how long is the curve drawn at random from a probability density function over functions? The answer is simple in that it varies: an unsatisfying answer and one that has only been minimally explored for GPs (Barakat and Baumann, 1970, Corrsin and Phillips, 1961, Miller and Freund, 1956). It is not immediately apparent how to quantify the distribution of the arc length, though the kernel hyperparameters seem to play an intimate part in the resulting curves; a fact that is well known to the GP community. Observing sample paths of GPs, drawn from the same covariance, it is clear that there is consistency in the lengths of the curves. Dependent on its kernel, a GP has a well-defined structure and we imagine that we can exploit this structure and leverage other statistical properties to determine the moments of the arc length distribution.

At present, a gap exists in the literature: no one has derived the distribution of the arc length of a vector-valued GP. Previous work tackles only the univariate case (Barakat and Baumann, 1970), making it inapplicable to important applications, for example in spatial path planning (Marchant and Ramos, 2014). An understanding of the arc length properties of a GP will open up promising avenues of research. Presently arc length statistics have been used to analyse multivariate time series modelling (Wickramarachchi et al., 2015), and minimise the arc length to obtain a geodesic in a high dimensional space (Tosi et al., 2014). Minimal arc lengths are known as geodesics (Hauberg et al., 2012) and there exists a connection between GP solutions to ordinary differential equations and shortest paths (Hennig and Hauberg,

2013, Schober et al., 2014).

1.3 Objectives

This thesis attempts to address two problems. Firstly, we attempt to develop novel non-linear functional regression methods using GPs. We investigate whether or not it is possible to develop a GP version of the Functional Index Model (Chen et al., 2011), the Functional Additive Model (Müller and Yao, 2008), and the Functional Generalized Additive Model (McLean et al., 2012). Gaussian Process versions of state-of-the-art functional regression models will be compared against baseline synthetic models to measure their predictive capability in terms of root mean square prediction error. Thereafter, we will test the proposed models on several real world functional data sets to determine their improvement. We investigate whether the models provide interpretable by-products that strengthen our understanding of the natural phenomena, giving us meaningful uncertainty on our response.

Secondly, we aim to quantify the distribution of the arc length of a GP and answer the question: how long is a GP? Given the well defined structure of a GP specified by its kernel and choice of hyperparameters, we attempt to develop a relationship between covariance functions, hyperparameters and arc length distributions. We attempt to re-derive the one dimensional results, with new methods and using these we strive to derive the first description of the arc length of a vector valued GP, validating derivations numerically. We envision the arc length as a cost function in Bayesian optimisation, as a tool in path planning problems (Marchant and Ramos, 2014) and a way to construct meaningful features from functional data. Given the importance of length and the need for non parametric probabilistic methods, a greater understanding of the fundamental properties of GPs will lead to exciting research opportunities.

1.4 Scope and limitations

This thesis is concerned with functional regression problems using GPs, and the fundamental properties of the arc length of a GP. We explore function to scalar problems under the assumption that functional trajectories are fully observed. Aiding inference, we make the simplifying assumption that our functional predictors are completely observed: that our inputs are in essence free of noise. In effect, we are assuming perfect observation of the underlying process. Though this may be somewhat unrealistic in practice, it allows the development of models that are simple and straightforward to implement and interpret. The possibility for further research with uncertain and sparse trajectory observations is highlighted in the conclusion of the thesis. The study is limited to the Functional Index Model (Chen et al., 2011), the Functional Additive Model (Müller and Yao, 2008), and the Functional Generalized Additive Model (McLean et al., 2012). Function to function models are not explored.

The arc length distributions are derived for the prior and posterior of a GP. We derive an exact formula in the one dimensional case and an approximated distribution for the vector case. Our approximation is performed using a moment matched integrand distribution, and we do not compare against another approximation. The theory is verified via numerical simulations using a variety of GP kernels and hyperparameters values. It is an open question as to how to apply these formula and properties in real world applications. The potential scope for real world applications is discussed and a novel application re-interpreting arc lengths as functional features is presented. Arc lengths are not considered for any other random curves.

1.5 Outline

The first part of the thesis focuses on introducing the background necessary to understand the novel contributions. In Chapter 2, the basics of probability theory are

detailed and emphasise the importance of the Bayesian approach and the use of non parametric statistics. Chapter 3 introduces GPs and the corresponding theory; outlining their construction, basic properties and how to perform inference. The state-of-the art in functional regression is presented in Chapter 4, where the methodology, techniques and limitations are discussed. The basics of arc lengths of functions are also discussed in preparation for the statistical properties of GP arc lengths that are later visited in Chapter 6.

Chapter 5 presents three novel approaches to functional regression with GPs: the Gaussian Process Functional Index Model (GP-IND), Gaussian Process Functional Additive Model (GP-FAM) and Gaussian Process Functional Generalized Additive Model (GP-FGAM). For each model we develop the core theory using GPs and describe their technical details. Under the assumption of fully observed trajectories we develop methods to perform inference and make prediction, in a simple, straightforward implementation. A number of synthetic examples are considered for each model and we demonstrate the performance capability of the GP models against their basic counterparts in improved predictive capability coupled with meaningful uncertainty estimates.

Gaussian Process Arc Lengths are the subject of Chapter 6. This chapter considers one dimensional and vector-valued GPs, presenting the first treatment of the vector case. We quantify the relationship between the distribution of the arc length and the parameters of the GP. The mean of the arc length of a one dimensional GP is derived using a new method and the first consideration of the vector valued arc length is presented. Approximations for the arc length moments are derived in the vector valued case using an approximation to the arc length integrand. Numerical simulations verify the high fidelity of the approximation.

In Chapter 7 we consider a number of real functional data sets. We compare the GP functional regression models developed in Chapter 5 against a suite of func-

tional regression benchmark models. We quantify the new GP models on a number of real data sets compared to benchmark models, showing empirical improvement in a number of experiments and consider the use of arc lengths as functional features.

Discussion and remarks are presented in Chapter 8. We discuss the findings and summarise the novel contributions of this thesis. Directions for future research are detailed and we speculate on the future of functional regression with machine learning and the importance of fundamental GP properties in Chapter 9.

1.6 Contribution

Chapter 5 contributes three new GP models. I developed the theory for the GP-IND and the GP-FAM, implementing both in GPflow (Matthews et al., 2016) and running the synthetic examples. The GP-FGAM was developed jointly by myself and Jonathan Downing, with Jonathan Downing implementing it, again in GPflow (Matthews et al., 2016). I identified the real world data in Chapter 7 and experiments were run jointly with Jonathan Downing. Michael Osborne provided advice on general theory.

In Chapter 6 I further expand on details around arc length distributions the core of which is based on the publication:

> J D Bewsher, A Tosi, M A Osborne and S J Roberts. Distribution of Gaussian Process Arc Lengths. In *Artificial Intelligence and Statistics (AISTATS)* 2017.

I developed the core theory and validated the distributions numerically. Alessandra Tosi helped developed intuition and provided geometric insight into the problem. Michael Osborne and Steve Roberts provided comments and advice on the theory.

Chapter 2

Probabilistic Thinking

2.1 Introduction

Understanding the mathematical language of probability is pertinent to making concrete statements regarding uncertainty. This chapter introduces the concept of probability theory and includes important distributions that arise in machine learning, demonstrating how they can be manipulated, and working through a Bayesian linear regression problem. This will lead to the introduction of Gaussian Processes (GPs) in Chapter 3.

2.2 Why Probability?

There are two main philosophical thoughts on probability frequently discussed. First of all, there is the use of probability to analyse the frequency at which events occur; for example the number of times a fair coin lands heads, the occurrence of 3's in successive die rolls or how often one would expect a card drawn from a full deck to be a heart. Importantly, in each of these cases the event is repeatable, and thus a probability value, p, is understood as the frequency of an event occurring in the limit of infinite repetitions. This interpretation of probability is coined 'frequentist'.

Though this definition appears clear, immediately one is unstuck when one reads a *50% chance of rain tomorrow*, or that a person's favourite sporting team has a *1:10 chance of making the finals*. These are discrete events and conditions associated with each are not able to be repeated. How can we make sense of these statements probabilistically? The frequentist approach offers no solution to this quandary.

The second thought on probability refers to the Bayesian paradigm, where we define probabilities as our degree of belief about a statement or event, given all background knowledge. In the Bayesian view, two people that are given the same background context should arrive at the same probability: two weather forecasters, given the same seasonal trend data and meteorological information, should arrive at the same probability of rain fall for the next day. Probability has been proposed as an extension of logic (Jaynes, 2003), whereby it is a formal language used to determine the likelihood of one proposition being true given another proposition on which it depends. Fortunately, common sense reasoning properties imply Bayesian probability has the same rules of frequentist probability.

Bayesian probability has a surprisingly long history; as early as the 1800's Laplace was using the Bayesian framework to compute the mass of planets (Laplace, 1815). Laplace used his understanding of the celestial motion of planets as his 'prior' and the orbits of neighbouring planets as his observation to estimate the corresponding mass of Saturn. It is a testament to his mathematical prowess and the power of the Bayesian framework that his calculation differs from that of today by about 0.5%.

2.3 The Rules of Probability

There is significant literature which expands on the theory and justification of our probabilistic approach (Papoulis, 1965, Pitman, 1993). The concept of a random variable is introduced to deal with uncertainty. A random variable maps outcomes

of physical phenomena to a numerical value. We will define our random variables in upper case letters, X, and the values they can take by lower case, x: they may take on discrete values, such as the side of a die, or continuous values, such as the height of a person. Furthermore, the possible values of the random variable are exhaustive and mutually exclusive; exactly one is true and the random variable can take on only one value at a time.

A random variable needs to be coupled with a probability distribution that indicates how likely each event is. The statement $P(X = x)$ is read as the probability that the random variable X takes on the value x, or under our Bayesian view, how likely is the event $X = x$. For the discrete case, we define a Probability Mass Function (PMF): P a mapping from the set of discrete states to a real number. A probability value of 1 corresponds to an event that is certain, whilst the probability of 0 corresponds to an event that is impossible. A PMF has the following characteristics:

1. P must be defined for the entire set of $\{x\}$.

2. $0 \leq P(x) \leq 1$; formally dictating that no state has less than an impossible chance and nothing is more than certain.

3. $\sum_{x \in X} P(x) = 1$; probabilities normalise, thus ensuring we don't invalidate the second characteristic.

Probabilities may be defined over multiple variables; a joint probability for random variables X and Y taking on values x and y is written as $P(X = x, Y = y) = P(x, y)$; we drop dependence on the random variable if it is clear from the context. For a joint probability distribution $P(x, y)$, we can obtain the marginal distribution of one variable by summing over the values of the other:

$$P(X = x) = \sum_{y \in Y} P(X = x, Y = y). \tag{2.1}$$

Equation (2.1) is known as the sum rule of probability.

For a continuous random variable we define a Probability Density Function (PDF). A PDF does not give the probability of an event, instead it prescribes the probability of an event occurring in an infinitesimal volume δx, which is given by $p(x)\delta x$, and can be considered the limit of:

$$p(X = x) \doteq \lim_{\delta x \to 0} \frac{P(x \leq X < x + \delta x)}{\delta x}, \tag{2.2}$$

This limit is generally non-trivial and it has been noted that a lack of rigour may lead to errors or paradoxes (Jaynes, 2003). However, if we limit ourselves to normalisable PDFs, we can proceed straightforwardly (Bretthorst, 1999); whereby sums transform into integrals. Echoing the characteristics of a PMF, for a PDF we have:

1. The domain of p must be all of x.
2. $0 \leq p(x), \forall\ x \in X$.
3. $\int_{x \in X} p(x)\mathrm{d}x = 1$.

Marginal probabilities are likewise computed by integrating out the 'nuisance' variables, as they may be called:

$$p(x) = \int_{y \in Y} p(x, y)\mathrm{d}y. \tag{2.3}$$

For a pdf we define the Cumluative Distribution Function (CDF), which is the probability of a random variable taking a values less than, or equal to, x:

$$F_X(X \leq x) = \int_{-\infty}^{x} p(x')\mathrm{d}x'. \tag{2.4}$$

Any joint distribution can be decomposed into products of one variable in terms of

the rest:

$$p(x_1, x_2, \ldots, x_n) = p(x_1) \prod_{i=2}^{n} p(x_i | x_1, \ldots, x_{i-1}). \tag{2.5}$$

This is the 'chain' or 'product' rule of probability and allows us to describe one variable in terms of another. Conditional probability is defined as:

$$p(y|x) = \frac{p(x, y)}{p(x)}. \tag{2.6}$$

Combining the product and chain rule, we can write our conditional distribution in the following form:

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} = \frac{p(x|y)p(y)}{\sum p(x|y)p(y)}. \tag{2.7}$$

This is Bayes' rule and in the context of Bayesian theory this statement lays out the mechanism by which we perform Bayesian inference. We revisit this statement in further detail later in this chapter.

Two variables are independent if knowledge of one does not provide knowledge of the other, meaning that the joint probability of the two decomposes, or that the conditional of one does not affect the other, and we write: $p(x, y) = p(x)p(y)$. Another form of independence, conditional independence, is where knowledge of one variable factorises the probability of the others $p(x, y|z) = p(x|z)p(y|z)$.

We are interested in informative quantities related to our probability distributions: these may include the maximum value, the most common value, or the halfway point. A common quantity of interest is the mean or expected value; the probability-weighted average of all values. More generally, we can consider the expected value of a function

$f()$ with respect to our probability distribution:

$$\mathbb{E}_{x\sim P(x)}[f(x)] = \sum_{x\in X} P(x)f(x), \tag{2.8}$$

$$\mathbb{E}_{x\sim p(x)}[f(x)] = \int_{x\in X} p(x)f(x)\mathrm{d}x. \tag{2.9}$$

Again, we will drop the explicit dependence on the probability distribution for the expectation if there is no ambiguity in doing so. The variance is the squared deviation from the mean value:

$$\mathbb{V}[f(x)] = \mathbb{E}[f(x) - \mathbb{E}[f(x)]^2]. \tag{2.10}$$

The square root of the variance is known as the standard deviation, and quantifies the dispersion of our values around the mean. Covariance measures variability between two random variables:

$$\mathrm{Cov}(f(x), g(y)) = \mathbb{E}[(f(x) - \mathbb{E}[f(x)])(f(y) - \mathbb{E}[g(y)])]. \tag{2.11}$$

High values of covariance indicate a strong (linear) relationship between two variables, whereas the sign indicates the direction of the relationship. A value close to zero indicates the variables do not affect each other linearly. Independent variables have zero covariance, as can be observed by the decomposition of their joint distribution. However, the converse is not true; zero covariance does not imply that the variables are independent. Correlation is the normalised covariance and ranges between -1 and 1. Higher order moments of variables can be computed in a similar fashion to Equations (2.8) and (2.9).

Formal analysis of probability theory requires measure theory. To ensure the rules of probability are not violated, one needs to ensure appropriate sets are chosen to integrate over. Measure theory sets out the rules for specifying sets and for assigning

a measure to sets that are negligibly small: thus probability is laid on a firm theoretical footing (Mikosch and Kallenberg, 1998).

2.4 Transformations

Consider $y = h(x)$, where $x \in X$ has a PDF, $p_X(x)$, and h is a differentiable function. Our interest is in determining the distribution of $y \in Y$. In order to ensure conservation of probability we need to be careful about how we represent our new distribution in terms of the original. The transformation h can be viewed as a warping from one space to another; as such, areas in one do not necessarily correspond to areas in the other. In order to ensure conservation of probability we require volumes of probability to be the same:

$$p_Y(y)\mathrm{d}y = p_X(x)\mathrm{d}x, \tag{2.12}$$

$$\implies p_Y(y)\mathrm{d}y = p_X(x)\left|\frac{\partial x}{\partial y}\right|, \tag{2.13}$$

An alternative derivation involves the cumulative distribution of y. Consider:

$$F_Y(y) = F_Y(Y < y), \tag{2.14}$$

$$= F_Y(h(X) < y), \tag{2.15}$$

$$= F_X(X < h^{-1}(y)), \tag{2.16}$$

$$= \int_{-\infty}^{h^{-1}(y)} p_X(x')\mathrm{d}x'. \tag{2.17}$$

Upon taking the derivative of the cumulative distribution function with respect to y we arrive at the distribution of y:

$$p_Y(y) = p_X\left(h^{-1}(y)\right)\left|\frac{\mathrm{d}}{\mathrm{d}y}h^{-1}(y)\right|. \tag{2.18}$$

For higher dimensions we use the Jacobian, $\mathbf{J}$:

$$p_Y(\mathbf{y}) = p_X(\mathbf{x})\mathbf{J}. \tag{2.19}$$

Many common distributions are related to each other through a simple transformations: the Chi-squared is the sum of squared normal variables. We examine a less well know transformation; the square root of a gamma random variable (Papoulis, 1965), whose PDF is given by:

$$f(y : k, \theta) = \frac{y^{k-1}}{\Gamma(k)\theta^k} \exp(-y/\theta). \tag{2.20}$$

The transformation is $h(y) = \sqrt{y}$, therefore $\frac{\mathrm{d}}{\mathrm{d}y}h^{-1}(x) = 2x$. The transformed distribution is therefore:

$$\begin{aligned} g(x) &= f\big(h^{-1}(x)\big)\frac{\mathrm{d}}{\mathrm{d}y}h^{-1}(x), & (2.21)\\ &= \frac{x^{2(k-1)}}{\Gamma(k)\theta^k} \exp\left(-\frac{x^2}{\theta}\right) 2x, & (2.22)\\ &= \frac{2x^{2k-1}}{\Gamma(k)\theta^k} \exp\left(-\frac{x^2}{\theta}\right). & (2.23) \end{aligned}$$

Defining $k = m$ and $\theta = \frac{\Omega}{m}$, we recognise $x \geq 0$, as begin Nakagami distributed (Hoffman, 1958) with parameters m and Ω:

$$g(x : m, \Omega) = \frac{2m^m}{\Omega^m\Gamma(m)} x^{2m-1} \exp\left(-\frac{m}{\Omega}x^2\right). \tag{2.24}$$

Figure 2.1 shows the change of variable alongside the resulting transformation of probability mass. This technique, and in particular this relation, will prove useful when considering distributions over arc lengths in Chapter 6.

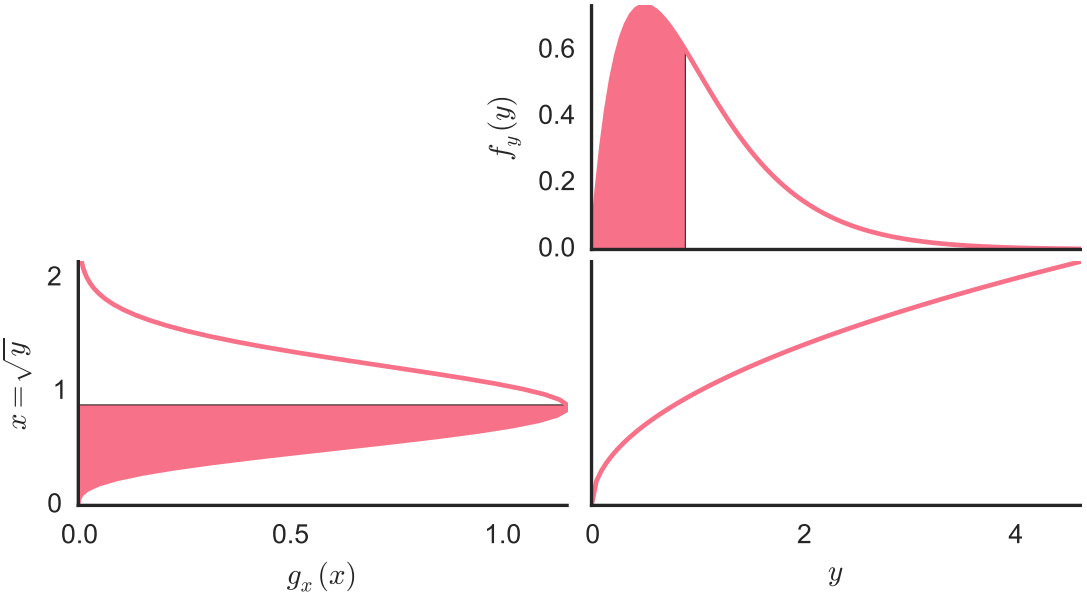

Figure 2.1: A gamma distribution (f) and the resulting transformed Nakagami distribution (g). The values $k = 2$ and $\theta = 0.5$ were used. The squeezing effect of the square root is visible. The shaded area is conserved through the transformation.

2.5 The Gaussian Distribution

The multivariate Gaussian or normal distribution, used interchangeably, is one of the most widely used probability distributions. A variable $\mathbf{x} \in \mathbb{R}^d$ that is normally distributed with mean $\boldsymbol{\mu}$ and variance Σ is defined as:

$$p(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right). \tag{2.25}$$

The Gaussian distribution is a continuous distribution defined over vectors in $\mathbb{R}^d$. It is a sensible choice for many problems, especially in the absence of stronger prior knowledge. Of all distributions with the same mean and variance, the Gaussian encodes maximum prior uncertainty (entropy). The Central Limit theorem also tells us that in the limit of large samples, many distributions, under certain assumptions, are approximately normal. We write a normally distributed $\mathbf{x}$ variable in terms of its two parameters, the mean μ and covariance Σ as:

$$\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}), \tag{2.26}$$

where $\sim$ means 'distributed as'. The properties of the Gaussian distribution make it easy to manipulate and are outlined in the Appendix.

2.6 Bayesian Inference for Linear Regression

Returning to Bayes' rule, it is the fundamental relation on which Bayesian inference is built. It is helpful to elucidate some of the principles of the Bayesian paradigm with a concrete example, whilst simultaneously developing the corresponding theory. This allows us to move from intuition to theory to understanding. A natural example is Bayesian Linear Regression (Box and Tiao, 1992), in which we are interested in

determining the output variable from a linear combination of predictor variables:

$$f(\mathbf{x}) = \mathbf{x}^T\boldsymbol{\beta}, \quad y = f(\mathbf{x}) + \epsilon, \tag{2.27}$$

Here $\mathbf{x} = [x_1, \ldots, x_d]^T$ is the observed input, $\boldsymbol{\beta}$ is a coefficient weight vector of length d, and y is a scalar output. In this setting we make the assumption that our observations of the true linear function, f, are corrupted with additive Gaussian noise;

$$\epsilon \sim \mathcal{N}(0, \sigma_y^2). \tag{2.28}$$

Consider being presented with a set of independent observations of predictors and outputs, which we collectively denote as $\mathcal{D} = \{\mathbf{x}_i, y_i\}_i^N$. Under our Gaussian noise model, we can write down the probability of having observed each point given our model parameters:

$$\begin{aligned} p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}) &= \prod_{i=1}^{N} p(y_i|\mathbf{x}_i, \boldsymbol{\beta}), && (2.29)\\ &= \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}\sigma_y} \exp\left(-\frac{(y_i - \mathbf{x}_i^T\boldsymbol{\beta})^2}{2\sigma_y^2}\right), \\ &\qquad \text{(Likelihood factorises over independent observations)} \\ &= \frac{1}{(2\pi\sigma_y^2)^{N/2}} \exp\left(-\frac{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^2}{2\sigma_y^2}\right), && (2.30)\\ &= \mathcal{N}(\mathbf{y}; \mathbf{X}\boldsymbol{\beta}, \sigma_y^2\mathrm{I}). && (2.31) \end{aligned}$$

We have $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^T$ and I is an identity matrix of size $N \times N$. Equation (2.31) is the likelihood of the data – how likely is the observed data for this model and set of parameters. Working in the Bayesian paradigm we need to choose a prior, $p(\boldsymbol{\beta})$, for the model weights, in order to specify a well-defined model. Typically any knowledge of our model is encoded into our likelihood distributions and our choice

of prior. We could potentially fix $\boldsymbol{\beta}$, which corresponds to a delta function prior, or we could specify a narrow range of permissible values. Given no other information, a natural choice, is to assign a Gaussian prior over the weights:

$$\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\beta}; 0, \Sigma_\beta). \tag{2.32}$$

Now that the likelihood and the prior have been defined, we are able to generate data from our model, make predictions and assess the validity of our model assumptions. Inference over the posterior distribution of $\boldsymbol{\beta}$'s given the data observation is now performed using Bayes' rule. Recall that the posterior is equal to the prior times the likelihood divided by the marginal likelihood:

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{marginal likelihood}}. \tag{2.33}$$

For our Bayesian linear regression case we have:

$$p(\boldsymbol{\beta}|\mathbf{y}, \mathbf{X}) = \frac{p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta})p(\boldsymbol{\beta})}{p(\mathbf{y}|\mathbf{X})}. \tag{2.34}$$

The marginal likelihood, or model evidence, $p(\mathbf{y}|\mathbf{X}) = \int p(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta})p(\boldsymbol{\beta})\mathrm{d}\boldsymbol{\beta}$ is independent of the model parameters, and acts as a normalising factor; it is often written as $p(\mathcal{D})$. The greater this value, the higher the probability that the observed data is drawn from our model. The posterior combines our prior belief over $\boldsymbol{\beta}$ and the observations of the data:

$$p(\boldsymbol{\beta}|\mathbf{X}, \mathbf{y}) \propto \exp\left(-\frac{1}{2\sigma_y^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right) \exp\left(-\frac{1}{2}\boldsymbol{\beta}^T\Sigma_\beta^{-1}\boldsymbol{\beta}\right). \tag{2.35}$$

To compute the posterior over $\boldsymbol{\beta}$ we need to compute the product of multivariate Gaussians. By completing the square in the exponential in Equation (2.35), lead us

to:

$$p(\boldsymbol{\beta}|\mathbf{X},\mathbf{y}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_{\text{post}})^T\left(\frac{\mathbf{X}^T\mathbf{X}}{\sigma_y^2}+\Sigma_\beta^{-1}\right)(\boldsymbol{\beta}-\boldsymbol{\beta}_{\text{post}})\right), \tag{2.36}$$

where:

$$\boldsymbol{\beta}_{\text{post}} = \sigma_y^{-2}\left(\frac{\mathbf{X}^T\mathbf{X}}{\sigma_N^2}+\Sigma_\beta^{-1}\right)\mathbf{X}^T\mathbf{y}. \tag{2.37}$$

The full posterior over the weights is thus:

$$p(\boldsymbol{\beta}|\mathbf{X},\mathbf{y}) = \mathcal{N}\left(\boldsymbol{\beta};\boldsymbol{\beta}_{\text{post}},\Sigma_{\text{post}}\right), \tag{2.38}$$

with the posterior covariance over the weights given by $\Sigma_{\text{post}}^{-1} = \left(\frac{\mathbf{X}^T\mathbf{X}}{\sigma_N^2}+\Sigma_\beta^{-1}\right)$. Inspecting the form of $\boldsymbol{\beta}_{\text{post}}$, we see that this is exactly the weight vector given by standard linear regression with a ridge regression penalty: a quadratic penalty in $\boldsymbol{\beta}$, expressed as $\boldsymbol{\beta}^T\boldsymbol{\beta}$. Thus we see that Bayesian linear regression with a weight prior is equivalent to a linear regression problem with an appropriate penalty.

Given new observations, $\mathcal{D}_{\text{new}}$, the Bayesian framework gives us a straightforward method to update our beliefs:

$$p(\boldsymbol{\beta}|\mathcal{D}_{\text{new}},\mathcal{D}_{\text{old}}) \propto p(\mathcal{D}_{\text{new}}|\mathcal{D}_{\text{old}},\boldsymbol{\beta})p(\boldsymbol{\beta}|\mathcal{D}_{\text{old}}). \tag{2.39}$$

The old posterior becomes our new prior and we update our beliefs from the combination of the prior and likelihood. Difficulty may emerge when our prior is poorly specified and/or far from the posterior. Samples from the likelihood may take a long time to move away from the prior and towards the correct posterior distribution. Specifying an appropriate prior becomes essential to correct Bayesian modelling.

An important difference arises from standard linear regression when making pre-

dictions for a new test point $\mathbf{x}_*$. Standard linear regression would lead to a single point estimate given by $f_* = \mathbf{x}_*^T \boldsymbol{\beta}_{\text{post}}$. A probabilistic framework tells us that to make predictions from data about future observations we must evaluate:

$$p(y_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}) = \int p(y_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}, \boldsymbol{\beta}) p(\boldsymbol{\beta}|\mathbf{X}, \mathbf{y}) \mathrm{d}\boldsymbol{\beta}. \tag{2.40}$$

In the Bayesian formulation we average our test point over the posterior distribution of the weights $\boldsymbol{\beta}$. This integral can be computed exactly, as we are dealing only with Gaussian distributions, with the distribution given by:

$$p(y_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}) = \mathcal{N}\left(y_* : \mathbf{x}_*^T \boldsymbol{\beta}_{\text{post}}, \mathbf{x}_*^T \Sigma_{\text{post}} \mathbf{x}_* + \sigma_y^2\right). \tag{2.41}$$

The predictive mean is equivalent to the standard linear regression, but we now have a predictive uncertainty for our test point. Figure 2.2 depicts a visualisation of Bayesian linear regression as we observe data, update the posterior for $\boldsymbol{\beta}_{\text{post}}$ and predict new values. It is possible to simultaneously place a prior over σ_y^2, with an appropriate conjugate prior being the inverse gamma distribution, and perform joint inference for $\boldsymbol{\beta}$ and σ_y^2.

A Bayesian framework also provides an answer to the challenging question of model selection. Suppose we were not certain that the choice of a linear model was correct, we could specify models for increasing orders of polynomials, indexed by m, and then consider the probability of a model given the observed data:

$$p(m|\mathcal{D}) = \frac{p(\mathcal{D}|m)p(m)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|m)p(m)}{\sum_{m \in \mathcal{M}} p(\mathcal{D}|m)p(m)}. \tag{2.42}$$

The marginal likelihood would now capture a preference for simpler models that explain the data. Thus, Bayesian methods provide a natural way to ensure we do not over fit a model: Occam's razor for Bayesian models (MacKay, 1992, 2003).

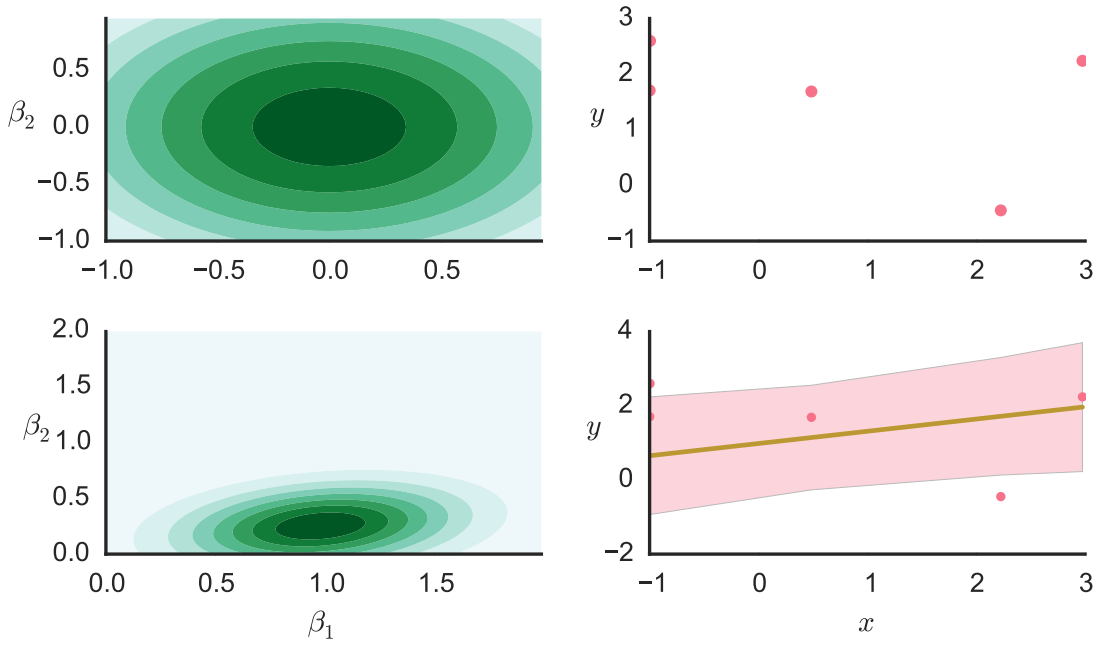

Figure 2.2: Bayesian linear regression in action. Top left panel shows a prior distribution over $\boldsymbol{\beta}$. Top right shows pairs of data observations. Bottom left is the posterior of $\boldsymbol{\beta}$: $p(\boldsymbol{\beta}|\mathcal{D})$. Bottom right is the posterior mean with one standard deviation, colour shading, either side of y_*.

2.7 Bayesian Nonparametrics

Having examined the Bayesian linear model we ask a simple question - where is the information stored in our model? You might say that it is the data, $\mathcal{D}$, and in a sense that is correct, as the information in the model is learnt from the data during fitting. However, when performing predictions, only the model parameters are required; thus the data is immediately discarded when performing predictions. By *a priori* fixing the number of parameters, we limit the flexibility of our models, reducing capacity and complexity of the model; furthermore, we become highly dependent on correct model selection. If instead we allow the number of parameters of our model to grow as the volume of data increases, we are able to obtain richer and more accurate *non-parametric* models.

A non-parametric model uses the data to perform future calculations, where we no longer 'parametrise' our model by a finite set of parameters: the data now forms part of the model parameters. Non-parametric models consider a potentially infinite set of parameters that could describe our model: thus have the flexibility to let the number of model parameters grow as we increase the number of observations available. In the language of probability, we say non-parametric models are memory-based and depend on the data $p(x|\mathcal{D})$: all the data is required to make predictions. The information bottle neck encountered in the parametric model has been removed with a now potentially infinite dimensional set of parameters.

Generally, in any practical situation or inference problem we will only be required to explain and or predict a finite set of observables. By marginalising out all but finitely many of the parameters, we able to model our data and perform inference. In order to capture the nature of our infinite dimensional objects we must consider distributions over our parameters. As we are routinely dealing with exchangeable data, then by the consequences of de Finetti's theorem (Kallenberg, 2005), we are justified in using probability distributions to describe our models. In practice, many

non-parametric models can be derived as natural extensions of finite-dimensional models. Non-parametric models have been developed to deal with clustering and classification problems: these include the Dirichlet Process Mixture Model, the Indian Buffet Process and various extensions thereof; on overview of which can be found in Ghahramani (2011).

2.8 Conclusion

In this chapter we have introduced probability theory: its foundational rules and the background that will be required for further chapters. As proponents of the Bayesian framework, we have outlined its motivation and highlighted it practically through Bayesian linear regression. Finally we have noted the limit of parametric methods therefore highlighting the need for non-parametric methods. Our interest lies in the development of Bayesian non-parametric regression models and their properties, which leads us naturally to the construction of the GP, that is introduced in the next chapter, and the focus of this thesis.

Chapter 3

Gaussian Processes

3.1 Introduction

In the previous chapter we explored a Bayesian treatment of the linear regression model. Despite our Bayesian framework, we were fundamentally limited by the linear assumption – no amount of data would overcome that limitation. In this chapter we overcome that limitation by introducing the Gaussian Process (GP), the non-linear probabilistic model that is at the core of this thesis. We detail the properties of a GP, and outline a number of ways that GPs are used in non-linear data modelling.

3.2 Gaussian Processes

As our interest lies in modelling uncertain functions, which are likely non-linear, we need a highly expressive function: though choosing that function is non-trivial. Whilst we could restrict the types of functions, we face the task of choosing which functions to limit ourselves to, with no clear method of how we would achieve this.

Gaussian Processes provide a natural and powerful way to perform inference around such functions (Rasmussen and Williams, 2006). Under the GP framework we instead place a probability on the types of functions we expect to encounter and

observe. They give us a flexible way to choose our function and possess a number of desirable properties making them easy to manipulate. The GP defines a distribution over functions and can be seen as an extension of the multivariate Gaussian distribution; an often helpful guide for intuition.

By definition, a stochastic process is a set of random variables $\{f(x) : x \in \mathcal{X}\}$, indexed by a set, $\mathcal{X}$. A GP is a stochastic process such that for any finite set of function evaluations, $\{f(x_1), \ldots, f(x_n)\}$, f is multivariate Gaussian distributed. Formally, we say that for any finite set of elements drawn from $\mathcal{X}$, f is a GP described by a mean, $m()$, and covariance function, $k()$, which we write as:

$$f(x) \sim \mathcal{GP}(m(x), k(x, x')), \tag{3.1}$$

where,

$$m(x) = \mathbb{E}[f(x)], \tag{3.2}$$

$$k(x, x') = \mathbb{E}[(f(x) - m(x))(f(x') - m(x'))]. \tag{3.3}$$

The types of functions admitted by $m()$ and $k()$ are precisely those that allow for a marginal multivariate Gaussian; thus m can be any parametric function and k must admit a semi-positive definite matrix when evaluated at points $x \in \mathcal{X}$. They are functions that map from the index set to the reals:

$$\mu : \mathcal{X} \to \mathbb{R}, \tag{3.4}$$

$$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}. \tag{3.5}$$

How do we deal with an infinite-dimensional function space in practice? The number of possible values is infinite, however, in practice our interest will lie in a subset of values.

In practice we will consider our functions as very long vectors where each entry corresponds to a possible function evaluation. We will consider inputs $\mathbf{x}$, that are generally considered known, and corresponding function evaluations $f(\mathbf{x})$. From the definition of a GP we immediately not that there is a marginalisation property; therefore working with a finite number of variables is the same as having considered the whole distribution but only looking at a subset of it. This is an artefact of working with the Gaussian: we have inherited the convenient mathematical properties of the multivariate Gaussian distribution. This allows us to marginalise over unseen variables and thus permits us to perform inference.

We observe that the model constraint encountered in parametric regression is overcome: as a GP is a non-parametric model it will fit to the observed data. We now have a mechanism to inform our choice of a non-linear function f; having placed a prior over our types of f without parametrising it. Though, we caution, there is still the need to specify our beliefs about the types of f's that we expect to observe, however, now the data informs our selection of f.

Both the mean and covariance are specified by a set of hyperparameters, the parameters of GP. Marginalising and performing inference over these is a key GP challenge.

3.3 Mean Functions

Our prior mean expresses our expectation about the values of f prior to observation and is captured in the prior mean function, $m(x)$, which may depend on a set of hyperparameters, θ. A commonplace and simple choice is a zero mean prior $m(x) = 0$, indicating that our best guess of the prior function value is zero. In the absence of prior knowledge it is generally good practice to set the prior mean to the empirical mean of our observations, this is equivalent to zero-meaning the data as a preprocessing step.

We note, however, that such an approach is non-Bayesian by explicitly incorporating data into the prior. Likewise, other simple choices include specifying a constant mean or a polynomial, e.g. a linear mean function, $m(x) = x$. It is possible to specify any arbitrary function as a mean corresponding to our knowledge of the observed quantity f. When one has specific domain knowledge relevant to the problem at hand, it is better to include this in the GP model. For example, in modelling a complex physical system, we could and probably should include knowledge of the underlying physics in our model.

3.4 Covariance Functions

The covariance function, $k(x, x')$, also depends on parameters θ, and indicates the correlation between the outputs f. $k(\cdot, \cdot)$ is also referred to as the kernel function and we will use the terms interchangeably. A GP prior is specified by a choice of kernel, encoding our belief about the nature of our function behaviour. Many choices exist for the covariance function, with some references including Abrahamsen (1997) and Gibbs (1997).

As noted, any covariance function that admits a semi-positive definite covariance matrix, is permissible. We will generally be interested in kernels that are a function of the distance between inputs only:

$$k(x, x') = k(\tau), \tag{3.6}$$

where we use the L^2 norm:

$$\tau = |x - x'|. \tag{3.7}$$

Under this assumption, values of our function at each point are highly correlated with

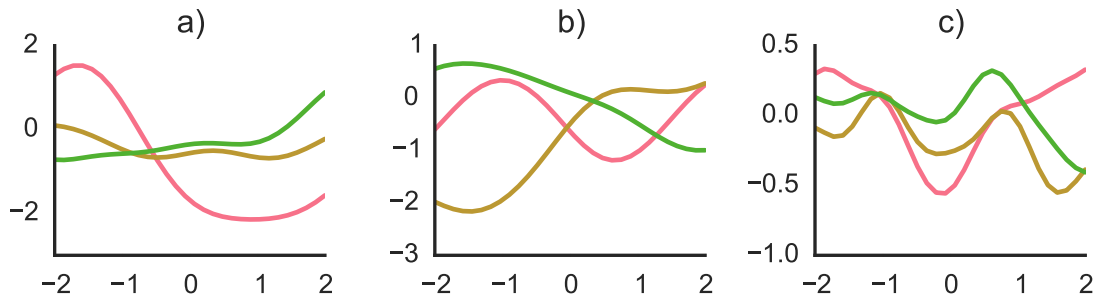

Figure 3.1: The effect of lengthscale and variance hyperparameters on draws from the SE kernel. The parameters are: a) $\lambda^2 = 1.0$, $l = 1.0$, b) $\lambda^2 = 0.5$, $l = 1.0$, c) $\lambda^2 = 0.1$, $l = 0.5$. We observe the range of interaction change as the length scale and output variance vary.

nearby points and that the correlation decreases as we move away further. These are known as stationary kernels: the general shape of samples functions is independent of the location at which it is observed.

A common choice of kernel for smooth functions is the Squared Exponential (SE):

$$k_{\text{SE}}(x, x') = \lambda^2 \exp\left(-\frac{1}{2}\frac{\tau^2}{l^2}\right). \tag{3.8}$$

The rate at which that correlation decreases, l, is known as the lengthscale, and can be considered the distance at which points will significantly interact with other. λ^2 is the output variance and determines the dynamic range of the function. Figure 3.1 shows the effect of varying hyperparameters on the shape of sample functions from the SE kernel. In Equation (3.8) we have implicitly assumed the function varies at the same rate in either direction of x – an isotropic kernel.

Though often the first choice of kernel, the SE can be too smooth for many applications (Stein, 1999). A more flexible choice is the Matérn class of kernels:

$$k_{\text{Matérn}}(x, x') = \lambda^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu}\frac{\tau}{l}\right)^{\nu} \kappa_\nu \left(\sqrt{2\nu}\frac{\tau}{l}\right). \tag{3.9}$$

Here, $\nu > 0$ is a smoothness parameter (larger ν leads to smoother functions), κ_ν is the modified Bessel function and $\Gamma()$ is the gamma function. This simplifies for half integer ν, and can be derived as (Abramowitz et al., 1965):

$$k_{\text{Matérn32}}(x, x') = \lambda^2 \left(1 + \frac{\sqrt{3}\tau}{l}\right) \exp\left(-\sqrt{3}\frac{\tau}{l}\right), \tag{3.10}$$

$$k_{\text{Matérn52}}(x, x') = \lambda^2 \left(1 + \frac{\sqrt{5}\tau}{l} + \frac{5}{3}\frac{\tau^2}{l^2}\right) \exp\left(-\sqrt{5}\frac{\tau}{l}\right). \tag{3.11}$$

Another example is the Rational Quadratic (RQ):

$$k_{\text{RQ}}(x, x') = \lambda^2 \left(1 + \frac{1}{2\alpha}\frac{\tau^2}{l^2}\right)^{-\alpha}. \tag{3.12}$$

The RQ can be derived as an infinite sum of squared exponential kernels each with a different input scale. Figure 3.2 show samples for the SE, Matérn32 (MAT32) and Matérn52 (MAT52) kernels: draws from each kernel result in distinct structure.

It is also possible to construct non stationary kernels, with one of the most simple being the Polynomial (POLY) kernel:

$$k_{\text{Poly}}(x, x') = (1 + xx')^p, \tag{3.13}$$

where p is the order of the polynomial. Other non stationary kernels have been developed to incorporate varying lengthscales (Paciorek and Schervish, 2004) or to help model changepoints (Garnett et al., 2010).

We have introduced a number of kernels for dimensional inputs. It is straightforward to consider multidimensional inputs, $\mathbf{x} = [x_1, \ldots, x_D]$, and corresponding kernels as described. For high dimensional inputs the function might vary across inputs at different scales, therefore, we can put a lengthscale on each dimension and arrive at the Automatic Relevance Determination (ARD) kernel (MacKay, 1992). More generally,

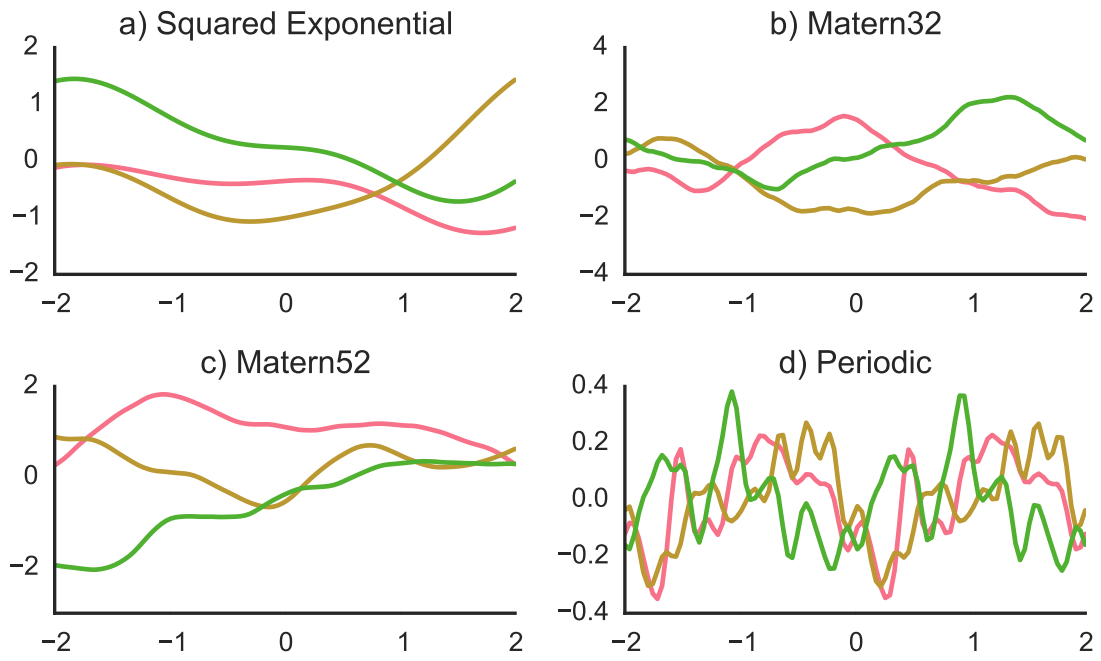

Figure 3.2: Examples of draws from a range of different covariance functions. We note the varying degree of smoothness: MAT32 gives rougher paths than MAT52 which is rougher than SE. The periodic nature of PER is clearly evident.

a Mahalanobis distance may be used instead of the Euclidean distance to compute distance between inputs (Titsias and Lázaro-Gredilla, 2013).

Kernel choice and construction is an active area of research. We are able to construct new kernels from old kernels through simple arithmetic operations, such as adding, multiplying, or performing a differentiable transformation on a function (Scholkopf and Smola, 2002). For example, a PER kernel (see Figure 3.2) can be constructed by transforming a function, with any of the above kernels, using the transformation $h(x) = (\cos(f(x), \sin(f(x))$. Gaussian Processes are part of kernel learning methods (Hofmann et al., 2008) and therefore have theoretical similarities with Reproducing Kernel Hilbert Spaces (RKHSs) (Rasmussen and Williams, 2006) and Support Vector Machines. Some recent work on kernel choice and development include spectral kernels (Wilson and Adams, 2013), and string kernels (Kom Samo and Roberts, 2016), and the use of Fourier features (Rahimi and Recht, 2007).

3.5 Posterior Distributions

Given a set of hyperparameters, the selection of which we discuss in the next section, we can evaluate our mean and covariance functions at a set of observations, and likewise we are able to generate draws from the prior GP. Now given a set of data observations, $\mathcal{D} = \{\mathbf{x}_i, f_i\}_{i=1}^n$, we are interested in making predictions around a previously unseen point, $\mathbf{x}_*$. We represent our observed inputs as $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T$ with $\mathbf{f}$ a $N \times 1$ vector of corresponding outputs. Consider the joint distribution $p(\mathbf{f}, \mathbf{f}_*)$:

$$\begin{bmatrix} \mathbf{f} \\ \mathbf{f}_* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(\mathbf{X}) \\ m(\mathbf{x}_*) \end{bmatrix}, \begin{bmatrix} K(\mathbf{X}, \mathbf{X}) & K(\mathbf{X}, \mathbf{X}_*) \\ K(\mathbf{X}_*, \mathbf{X}) & K(\mathbf{X}_*, \mathbf{X}_*) \end{bmatrix} \right), \tag{3.14}$$

with $K(\mathbf{X}, \mathbf{X})$, the kernel k evaluated at $\mathbf{X}$. The posterior GP can be determined by the conditional rules of joint Gaussian variables, detailed in the Appendix:

$$\mathbf{f}_*|\mathbf{f}, \mathbf{X}, \mathbf{y} \sim \mathcal{N}(\mathbf{f}_*; m(\mathbf{x}_*), \mathbb{V}(f(\mathbf{x}_*)), \tag{3.15}$$

$$m(\mathbf{x}_*) = \mu(\mathbf{x}_*) + K(\mathbf{x}_*, \mathbf{X})K(\mathbf{X}, \mathbf{X})^{-1}(\mathbf{f} - m(\mathbf{X})), \tag{3.16}$$

$$\mathbb{V}(f(\mathbf{x}_*)) = K(\mathbf{x}_*, \mathbf{x}_*) - K(\mathbf{x}_*, \mathbf{X})K(\mathbf{X}, \mathbf{X})^{-1}K(\mathbf{X}, \mathbf{x}_*). \tag{3.17}$$

These posterior equations are key to the GP. These equations provide a straightforward mechanism to update our posterior GP in light of new data; the use of a GP means that our obtained posterior is also Gaussian. We are able to to use previous observations to predict new data points or to interpolate between observations. In fact, the GP can be seen as a data smoother: the posterior mean can be re-written as a linear combination, in f, of kernel smoothers at each data point. Thus we see a limitation of the GP: it can only represent new data as a linear combination of previous observations.

We note that the variance of an observed point depends only on the location of the observations, not the values, and is made up of two parts. The first term captures the prior variance whilst the second represents the information given to us from the observations. We note, that it is possible to obtain a posterior despite the infinite values of $\mathbf{f}$ that we do not observe: a result of the marginalisation property of the GP. A more general covariance between outputs can be obtained for f_* and $f_\#$, given by:

$$\mathbb{C}(f(\mathbf{x}_*), f(\mathbf{x}_\#)) = K(\mathbf{x}_*, \mathbf{x}_\#) - K(\mathbf{x}_*, \mathbf{X})K(\mathbf{X}, \mathbf{X})^{-1}K(\mathbf{X}, \mathbf{x}_\#). \tag{3.18}$$

Typically in application settings we do not observe actual values of f, but instead

our observations will be corrupted by additive noise:

$$y(\mathbf{x}) = f(\mathbf{x}) + \sigma_y^2, \tag{3.19}$$

$$p(\mathbf{y}|\mathbf{f}) = \mathcal{N}(\mathbf{y}; \mathbf{f}, \sigma^2 \mathrm{I}), \tag{3.20}$$

in which I is an identity matrix of correct size, and σ_y^2 is Gaussian noise. Generally we assume independent Gaussian noise with fixed variance, though one can assume heteroscedastic noise models (Kersting et al., 2007, Le et al., 2005). Thus our likelihood model is Gaussian. This noise can be included in our covariance function, becoming another hyperparameter, by writing $k(f(\mathbf{x}_i), f(\mathbf{x}_j)) = k(\mathbf{x}_i, \mathbf{x}_j) + \delta_{ij}\sigma_y^2$, where δ_{ij} is the Kronecker[1] delta function. Noise is now present on our observations and the joint distribution over test and training points is augmented in the covariance of Equation (3.14):

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{y}_* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} m(\mathbf{X}) \\ m(\mathbf{x}_*) \end{bmatrix}, \begin{bmatrix} K(\mathbf{X}, \mathbf{X}) + \sigma_y^2 \mathrm{I} & K(\mathbf{X}, \mathbf{x}_*) \\ K(\mathbf{x}_*, \mathbf{X}) & K(\mathbf{x}_*, \mathbf{x}_*) + \sigma_y^2 \end{bmatrix} \right). \tag{3.21}$$

The resulting posterior equations are then:

$$\mathbf{y}_*|\mathbf{f}, \mathbf{X}, \mathbf{y} \sim \mathcal{N}(m(\mathbf{x}_*), \mathbb{V}(\mathbf{x}_*)), \tag{3.22}$$

$$m(\mathbf{x}_*) = m(\mathbf{x}_*) + K(\mathbf{x}_*, \mathbf{X})\big(K(\mathbf{X}, \mathbf{X}) + \sigma_y^2 \mathrm{I}\big)^{-1}(\mathbf{y} - m(\mathbf{X})), \tag{3.23}$$

$$\mathbb{V}(\mathbf{x}_*) = K(\mathbf{x}_*, \mathbf{x}_*) - K(\mathbf{x}_*, \mathbf{X})\big(K(\mathbf{X}, \mathbf{X}) + \sigma_y^2 \mathrm{I}\big)^{-1} K(\mathbf{X}, \mathbf{x}_*) + \sigma_y^2. \tag{3.24}$$

Figure 3.3 shows draws from a GP with a SE kernel and the resulting posterior conditioned on observations ($\star$). Near observed data points our posterior uncertainty decreases, and as we move away form the data the posterior mean reverts to the prior mean.

[1]The Kronecker delta function is defined: $\delta_{ij} = 1$, if $i = j$, or zero otherwise.

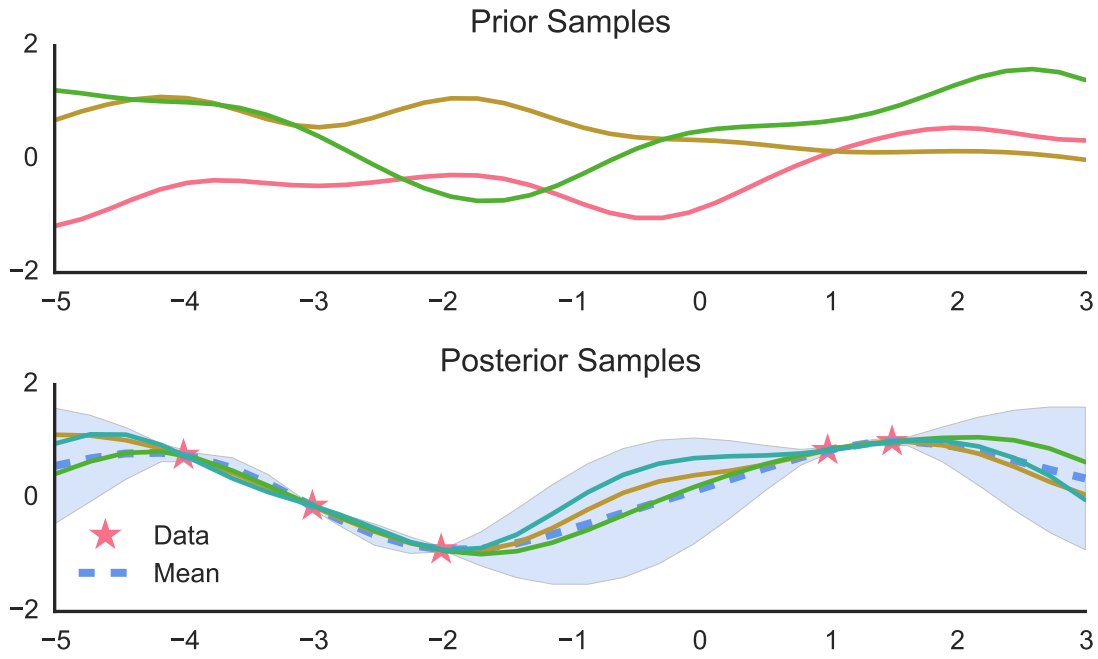

Figure 3.3: Prior and posterior draws of a zero-mean GP. Top: draws from a zero mean prior with a SE kernel. Bottom: the posterior conditioned on observations (markers), the dotted line is the posterior mean and the shaded area is equal to two standard deviations either side of the mean. Solid lines correspond to draws from the prior and posterior.

3.6 Inference & Learning

Now we have the machinery in place to compute posterior distributions, we return to the question of hyperparameter selection. This is the quintessential GP learning problem - how do we fit our GP model to data? Under a fully Bayesian framework we would specify a prior for the hyperparameters and then marginalise out the hyperparameters. In a general setting, with a judicious choice of priors, Bayesian marginalisation can be performed in closed-form. However, in the GP framework the likelihood and posterior have non-trivial dependence on hyperparameters, rendering

our integrals intractable. Resorting to approximate methods is inevitable.

The most straightforward of these is numerical integration: we could employ an Markov Chain Monte Carlo (MCMC) approach, generating samples of the hyperparameters and performing an MCMC estimate; as described in Richey (2010). This can be highly inefficient and it can be difficult to construct samplers that converge.

We are dealing with a probabilistic model, so a natural inference method would be to maximise the probability that the data comes from the underlying model. That is, given our observations, what parameters maximise the likelihood of the observed data? Our assumption when using a GP is that the data has been drawn from some underlying function space which the GP is able to model. The marginal likelihood, or evidence, of the model is:

$$p(\mathbf{y}|\mathbf{X}) = \int p(\mathbf{y}|\mathbf{f}, \mathbf{X}) p(\mathbf{f}|\mathbf{X}) \mathrm{d}\mathbf{f}. \tag{3.25}$$

Under our GP prior and assuming a zero mean prior, we can write this explicitly as:

$$\log p(\mathbf{y}|\mathbf{X}, \theta) = -\frac{1}{2}\mathbf{y}^T \left(K + \sigma_y^2 \mathrm{I}\right)^{-1} \mathbf{y} - \frac{1}{2}\log|K + \sigma_y^2 I| - \frac{D}{2}\log(2\pi). \tag{3.26}$$

Let's take a moment to inspect those terms, the first represents the model fit to our data, whilst the second represents the complexity of our model. Maximising the first term corresponds to fitting the model to the observations and is achieved by making K more complicated. However as the complexity of K increases the second term decreases and thus we penalise overly complicated models. Thus we see that the marginal likelihood will favour the simplest model that fits the data: Occam's Razor once again. Learning our GP now involves optimisation over the

model hyperparameters, for which any gradient based method would work using:

$$\frac{\partial}{\partial\theta}\log p(\mathbf{y}|\mathbf{X},\theta) = -\frac{1}{2}\mathbf{y}^T\left(K+\sigma_y^2\mathrm{I}\right)^{-1}\frac{\partial K}{\partial\theta}\left(K+\sigma_y^2\mathrm{I}\right)^{-1}\mathbf{y} - \frac{1}{2}\mathrm{tr}\left(\left(K+\sigma_y^2\mathrm{I}\right)^{-1}\frac{\partial K}{\partial\theta}\right). \tag{3.27}$$

Thus maximising the log likelihood or equivalently minimising the negative log likelihood will train our GP model: a Maximum Likelihood Estimate (MLE). The likelihood surface of the GP can be highly multimodal with no guarantee that a local maximum fits the data correctly. In fact every local maximum corresponds to one potential interpretation of the data, we could have a long length scale, large noise or alternatively small length with small noise. We can improve the chance of locating the global maximum of the likelihood by performing multiple restarts of a local optimiser. Closely related to maximum likelihood is Maximum a Posteriori (MAP) optimisation, which involves specifying a prior over the hyperparameters and then minimising:

$$\log p(\mathbf{y}|\mathbf{X},\theta)p(\theta) = -\frac{1}{2}\mathbf{y}^T(K+\sigma_y^2\mathrm{I})^{-1}\mathbf{y} - \frac{1}{2}\log|K+\sigma_n^2 I| - \frac{D}{2}\log(2\pi) + \log p(\theta). \tag{3.28}$$

The $\log p(\theta)$ term can be thought of as a regulariser; a term that penalises models for which θ deviates far from the hyperprior $p(\theta)$. The MLE and MAP are useful approximations in many practical cases. If we were considering a full Bayesian model, then in terms of marginalisation we can view MLE as specifying a delta prior at the MLE of θ, $p(\theta) = \delta(\theta - \theta_{\mathrm{MLE}})$ and the MAP approximation as specifying a delta prior at the MAP estimate $p(\theta) = \delta(\theta - \theta_{\mathrm{MAP}})$. Ideally, a full Bayesian approach would also involve marginalisation over models: a difficult and complex task.

3.7 Classification

Discussion on GPs has thus far focused on the regression problem, where the values for $\mathbf{y}$ are continous real-valued numbers. Consider instead the supervised learning problem of associating each input $\mathbf{x}$ with a class label $y \in \{-1, 1\}$. Gaussian Process classification models the likelihood of observing the class label conditional on the input as a Bernoulli random variable. The latent function, f, is modelled as a GP with the probability of the output $p(y = 1|x) = \Phi(f(x))$, where $\Phi()$ represents the cumulative normal distribution; an alternative choice would be to use the logistic function. We assume independent observations y_i, therefore the likelihood of the data factorises as:

$$p(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^{n} p(y_i|f_i), \tag{3.29}$$

$$= \prod_{i=1}^{n} \Phi(y_i f_i), \tag{3.30}$$

where the latent function $f()$ is pushed through the probit, or logit, function, $\Phi()$. The aim is to perform posterior inference over the class labels:

$$p(\mathbf{f}|\mathcal{D}, \theta) = \frac{\mathcal{N}(\mathbf{f}; 0, K)}{p(\mathcal{D}|\theta)} \prod_{i=1}^{n} \Phi(y_i f_i), \tag{3.31}$$

where,

$$p(\mathcal{D}|\theta) = \int p(\mathbf{y}|\mathbf{f}) p(\mathbf{f}|X, \theta) \mathrm{d}\mathbf{f}. \tag{3.32}$$

Unfortunately these quantities cannot be computed analytically due to the non-conjugate likelihood, so an approximation must be performed. The two prevailing approximations are the Laplace approximation (Riihimäki and Vehtari, 2014) and the Expectation Propagation algorithm (Minka, 2001); a review of the competing

methods is presented in Kuss and Rasmussen (2006). It would also be possible to design an efficient MCMC sampler to compute the posterior.

Once a posterior is obtained, prediction for an unobserved class label is performed via the quantities:

$$q(f_*|\mathcal{D}, , \mathbf{x}_*) = \mathcal{N}(f_*; m_*, \sigma_*^2), \tag{3.33}$$

$$q(y_* = 1|\mathcal{D}, \mathbf{x}_*) = \Phi(m_*/\sqrt{1+\sigma_*^2}), \tag{3.34}$$

where are m_* and σ_*^2 are the posterior mean and variance of the latent function. Gaussian Process classification is known to work well in practice.

3.8 Approximate Inference

Throughout this chapter we have neglected to confront the computational difficulty of working with GPs. Difficulty arises in computing the inverse of the covariance matrix, K^{-1}, which fast becomes infeasible for problems exceeding even thousands of points; the computational cost of the GP in inference is $\mathcal{O}(n^3)$ and for storage $\mathcal{O}(n^2)$. A plethora of approaches have been developed to tackle this problem, including: learning using pseudo-inputs (Snelson and Ghahramani, 2006), exploiting Kronecker structure (Flaxman et al., 2015, Wilson and Nickisch, 2015), stochastic variational inference (Hensman et al., 2013) and general sparse approximations, a summary of which is provided in Quinonero Candela and Rasmussen (2005). Most of these approaches attempt to approximate the kernel matrix and thus the inverse in a sensible way. Approximate inference is also necessary in non-conjugate likelihoods, which arise in Generalized GP models (Chan and Dong, 2011, Sheth et al., 2015) which require intractable integrals; as in Equation (3.32) for GP classification.

3.9 Derivative Gaussian Processes

At times it may be useful to have access to derivative observations of a function, for example in a dynamic system (Solak et al., 2002). Differentiation is a linear operator, therefore the derivative of a GP is also a GP (Rasmussen and Williams, 2006). Therefore we can use a GP to perform predictions about derivatives and conversely use derivative observations in predictions. The mean of the derivative is equal to the derivative of the mean:

$$\mathbb{E}\left[\frac{\partial f(\mathbf{x})}{\partial x_d}\right] = \frac{\partial \mathbb{E}\left[f(\mathbf{x})\right]}{\partial x_d}. \tag{3.35}$$

The covariance over the function values also implies mixed covariances between derivatives (Papoulis, 1965):

$$\mathbb{C}\left(f(\mathbf{x}_i), \frac{\partial f(\mathbf{x}_j)}{\partial x_{j,d}}\right) = \frac{\partial k(\mathbf{x}_i, \mathbf{x}_j)}{\partial x_{j,d}}, \tag{3.36}$$

$$\mathbb{C}\left(\frac{\partial f(\mathbf{x}_i)}{\partial x_{i,e}}, \frac{\partial f(\mathbf{x}_j)}{\partial x_{j,d}}\right) = \frac{\partial^2 k(\mathbf{x}_i, \mathbf{x}_j)}{\partial x_{i,e} \partial x_{j,d}}, \tag{3.37}$$

where $x_{i,e}$ is the eth element of $\mathbf{x}_i$. This informs us that we could define a prior derivative GP in terms of the prior GP:

$$f \sim \mathcal{GP}\left(m, k\right), \quad f' \sim \mathcal{GP}\left(m', k'\right), \tag{3.38}$$

with m' defined as in Equation (3.35) and k' defined in Equation (3.37). In turn, given observed values, we can determine the derivative of the GP posterior. The mean is

straightforwardly calculated, for zero mean prior:

$$\mathbb{E}\left[\frac{\partial f_*}{\partial x_{*,d}}\right] = \frac{\partial \mathbb{E}[f_*]}{\partial x_{*,d}}, \tag{3.39}$$

$$= \frac{\partial m(\mathbf{x}_*)}{\partial x_{*,d}}, \tag{3.40}$$

$$= \frac{\partial K(\mathbf{x}_*, \mathbf{X})}{\partial x_{*,d}}[K(\mathbf{X}, \mathbf{X}) + \sigma_y^2 I]^{-1}\mathbf{y}. \tag{3.41}$$

Computation of the variance is somewhat more involved with the final result (Riihimäki and Vehtari, 2010):

$$\mathbb{V}\left[\frac{\partial f_*}{\partial x_{*,d}}\right] = \frac{\partial^2 K(\mathbf{x}_*, \mathbf{x}_*)}{\partial x_{*,d}\partial x_{*,d}} - \frac{\partial K(\mathbf{x}_*, \mathbf{X})}{\partial x_{*,d}}[K(\mathbf{X}, \mathbf{X}) + \sigma_y^2]^{-1}\frac{K(\mathbf{X}, \mathbf{x}_*)}{\partial x_{*,d}}. \tag{3.42}$$

A full distribution can now be defined for the posterior derivative process:

$$\frac{\partial f_*}{\partial x_{*,d}} \sim \mathcal{N}\left(\frac{\partial f_*}{\partial x_{*,d}}; \mathbb{E}\left[\frac{\partial f_*}{\partial x_{*,d}}\right], \mathbb{V}\left[\frac{\partial f_*}{\partial x_{*,d}}\right]\right). \tag{3.43}$$

Figure 3.4 shows samples from a GP posterior and the corresponding derivative posterior. You can see the points at which the function has been observed lead to high uncertainty in the derivative function. As we constrain the function location we make less stipulation about which direction it must be either side.

We can use derivative observations alongside our function observations to make predictions. As we did earlier we consider the joint distribution over the functions value and the $n(D+1)$ derivative observations. We augment the covariance matrix to include the derivative observations and the corresponding cross covariances induced by the derivatives, and then we condition on those derivative observations to perform predictions. The effect of observing the derivative lowers the predictive uncertainty in those regions (Solak et al., 2002).

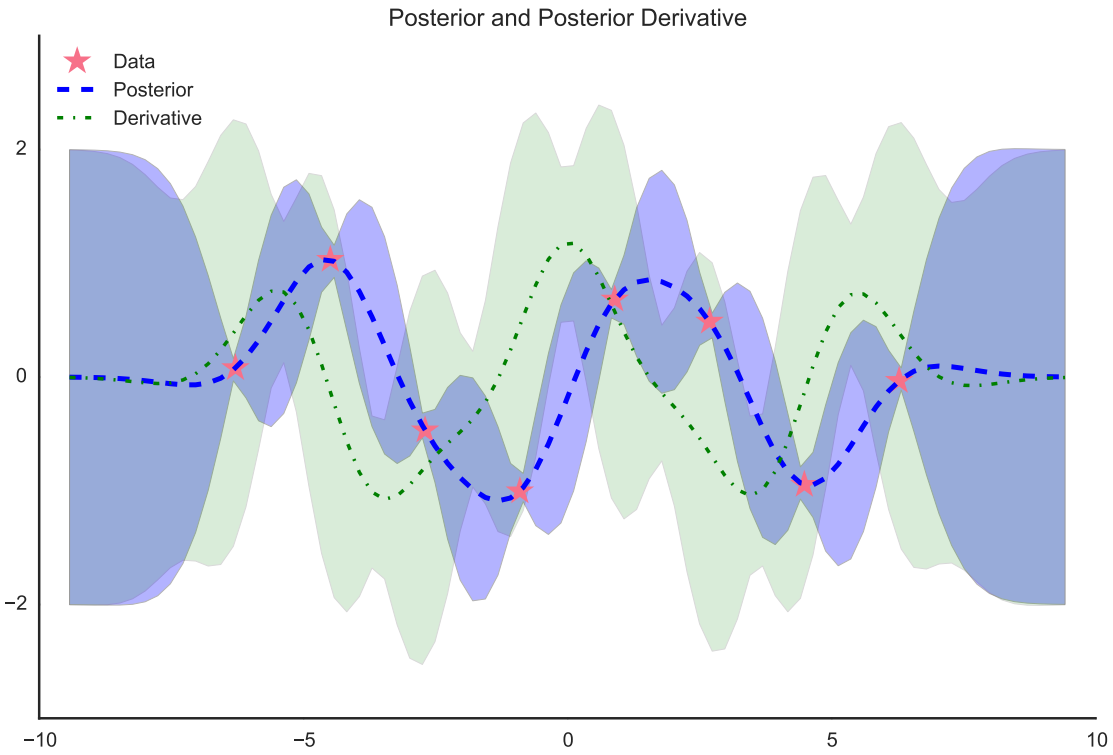

Figure 3.4: A posterior GP and its corresponding derivative process. Markers correspond to observations, dotted lines to the mean and shaded area are two standard deviations either side of the mean. We note the increase in uncertainty of the posterior derivative around observed points.

3.10 Bayesian Quadrature

Continuing the theme of treating quantities as uncertain variables we consider the evaluation of an integral under a Bayesian framework using a GP. Given a deterministic function, f, we wish to evaluate integrals of the form:

$$Z = \int f(x)p(x)\mathrm{d}x. \tag{3.44}$$

These types of integrals turn up across machine learning, for example in MCMC estimates, and as the marginal likelihood. To estimate Z we evaluate the function f at a number of locations and use a GP as a prior over the function, which then induces a posterior over the integral. Specifically, if f is modelled by a GP with mean $m(x)$ and covariance $K(x, x')$, then Z is a univariate Gaussian with mean and variance given by:

$$\mathbb{E}[Z] = \int m(x)p(x)\mathrm{d}x, \quad \mathbb{C}[Z] = \int\int k(x, x')p(x)p(x')\mathrm{d}x\mathrm{d}x'. \tag{3.45}$$

We condition on the observations of f to inform an estimate of Z. This idea was first introduced by O'Hagan (1991) where it was called Bayes-Hermite Quadrature and later re-introduced as Bayesian Monte Carlo (Ghahramani and Rasmussen, 2002). Such an approach is helpful when the function in question is expensive to evaluate. It is possible to analytically compute the integrals of the posterior mean and covariance function for specific choices of the covariance function, for examples the SE. It may appear unusual to consider an integral as random,, but under the Bayesian approach we encode uncertainty about all values using probability. More broadly this is an example of the probabilistic numerics approach to machine learning (Hennig et al., 2015), which ascribes uncertainty to calculation of numerical quantities and has been used in differential equations, and linear algebra.

3.11 Vector Gaussian Processes

Treatment of GPs until this point has focused on the single-output case: each entry of $\mathbf{y}$ has been one dimensional. Naturally we do not inhabit a one dimensional world, and many modelling situations require multiple outputs. In this section we develop the framework for multi output, vector valued GPs. Development proceeds in a manner similar to the single output case (Alvarez et al., 2012). The main idea is to share knowledge across outputs, with examples including multi-task learning (Caruana, 1997), sensor networks (Osborne et al., 2008) and geostatistics modelling (Goovaerts, 1997).

The outputs are random variables associated with different processes evaluated at potentially different values of $\mathbf{x}$. We consider a vector valued GP:

$$\mathbf{f} \sim \mathcal{GP}(\mathbf{m}, \mathbf{K}), \tag{3.46}$$

where $\mathbf{m} \in \mathbb{R}^D$ is the mean vector with $\{m_d(x)\}_{d=1}^D$ the mean functions associated with each output and $\mathbf{K}$ is now a positive definite matrix valued function. $(\mathbf{K}(\mathbf{x}, \mathbf{x}'))_{d,d'}$ is the covariance between $f_d(x)$ and $f_{d'}(x')$. Given input $\mathbf{X}$, our prior over $\mathbf{f}(\mathbf{X})$ is now

$$\mathbf{f}(\mathbf{X}) \sim \mathcal{N}(\mathbf{m}_*(\mathbf{X}), \mathbf{K}(\mathbf{X}, \mathbf{X})). \tag{3.47}$$

Here, $\mathbf{m}_*(\mathbf{X})$ is a DN-length vector that concatenates the mean vectors for each output and $\mathbf{K}(\mathbf{X}, \mathbf{X})$ is a $ND \times ND$ block partitioned matrix. In the vector valued case the predictive equations for an unseen datum, $\mathbf{x}_*$, become:

$$\mathbf{m}(\mathbf{x}_*) = \mathbf{K}_{\mathbf{x}_*}^T (\mathbf{K}(\mathbf{X}, \mathbf{X}) + \mathbf{\Sigma})^{-1} \mathbf{y}, \tag{3.48}$$

$$\mathbf{C}_*(\mathbf{x}_*, \mathbf{x}_*) = \mathbf{K}(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{K}_{\mathbf{x}_*}^T (\mathbf{K}(\mathbf{X}, \mathbf{X}) + \mathbf{\Sigma})^{-1} \mathbf{K}_{\mathbf{x}_*}, \tag{3.49}$$

where $\mathbf{K}_{\mathbf{x}_*} = \mathbf{K}(\mathbf{X}, \mathbf{x}_*)$ and $\mathbf{\Sigma}$ is a block diagonal matrix with the prior noise of each

output along the diagonal.

3.11.1 Kernels

The problem now focuses on specifying the form of the covariance matrix $\mathbf{K}$. Work was pioneered in kriging and geostatistics (Goovaerts, 1997).

We are interested in separable kernels of the form:

$$\mathbf{K}(\mathbf{x}, \mathbf{x}')_{d,d'} = k(\mathbf{x}, \mathbf{x}')k_T(d, d'), \tag{3.50}$$

where k and k_T are themselves valid kernels. In the simplest case we have $k_T(d, d') = \delta_{d,d'}$ implying the outputs are independent.

The kernel can then be specified in the form:

$$\mathbf{K}(\mathbf{x}, \mathbf{x}') = \mathbf{k}(\mathbf{x}, \mathbf{x}')\mathrm{B}, \tag{3.51}$$

where B is a $D \times D$ positive semi-definite matrix. For a data set $\mathbf{X}$:

$$\mathbf{K}(\mathbf{X}, \mathbf{X}) = \mathrm{B} \otimes k(\mathbf{X}, \mathbf{X}), \tag{3.52}$$

with $\otimes$ representing the Kronecker product. B specifies the degree of correlation between the outputs. Various choices of B result in what is known as the Intrinsic Model of Coregionalisation (IMC) and Linear Model of Coregionalisation (LMC) (Goovaerts, 1997, Journel and Huijbregts, 1978).

Parameter estimation follows by considering the objective function, the marginal

likelihood:

$$\log p(\mathbf{y}|\mathbf{X},\theta) = \log \mathcal{N}(\mathbf{y}|0,\mathbf{K}(\mathbf{X},\mathbf{X})+\mathbf{\Sigma}), \tag{3.53}$$

$$= -\frac{1}{2}\mathbf{y}^T\big(\mathbf{K}(\mathbf{X},\mathbf{X})+\mathbf{\Sigma}\big)^{-1}\mathbf{y} - \frac{1}{2}\log|(\mathbf{K}(\mathbf{X},\mathbf{X})+\mathbf{\Sigma})| - \frac{ND}{2}\log 2\pi. \tag{3.54}$$

Optimization follows by employing for example a gradient based approach:

$$\frac{\partial}{\partial\theta}\log p(\mathbf{y}|\mathbf{X},\theta) = -\frac{1}{2}\mathbf{y}^T\big(\mathbf{K}+\mathbf{\Sigma}\big)^{-1}\frac{\partial\mathbf{K}}{\partial\theta}\big(\mathbf{K}+\mathbf{\Sigma}\big)^{-1}\mathbf{y} - \frac{1}{2}\mathrm{tr}\left(\big(\mathbf{K}+\mathbf{\Sigma}\big)^{-1}\frac{\partial\mathbf{K}}{\partial\theta}\right), \tag{3.55}$$

to compute an MLE or MAP estimate of θ.

It is also possible to construct convolution based kernels, effectively linking the outputs in a non-linear fashion (Majumdar and Gelfand, 2007). As we did in the single output GP, we can compute the derivatives of the vector valued output. Chapter 6 will make use of vector valued GPs and their derivatives.

3.12 Mercer & Karhunen Loéve Theorems

We recall that K was by construction a positive semi-definite matrix, which implies that it can be decomposed in terms of its eigenvalues and eigenvectors. As K was derived from evaluation of our kernel function at locations, we are able to generalise the idea of eigen-analysis to our kernel function. We define the eigenfunction expansion of $k(\mathbf{x},\mathbf{x}')$ as

$$k(\mathbf{x},\mathbf{x}') = \sum_i^\infty \lambda_i\phi_i(\mathbf{x})\phi_i(\mathbf{x}'), \tag{3.56}$$

with

$$\int k(\mathbf{x}, \mathbf{x}')\phi_i(x)p(\mathbf{x})\mathrm{d}x = \lambda_i\phi_i(\mathbf{x}'). \tag{3.57}$$

Mercer's Theorem (Mercer, 1909) describes how a kernel can be expanded in terms of orthogonal basis functions. The rate of decay of the eigenvalues gives us information regarding the smoothness of our function: rougher functions are known to have more power within their higher frequencies, thus their eigenvalue spectrum decays more slowly.

The Karhunen-Loéve theorem then enables us to expand our GP functions in combinations of eigenfunctions of the kernel function. Closed form solutions are available for certain kernels; the SE for example can be represented in terms of products of exponentials and Hermite polynomials (Zhu et al., 1998). Approximate eigenfunctions for a kernel can be computed from the covariance matrix, which equates to a discrete evaluation of principal components of the covariance matrix. Many approximation methods described earlier require an approximation to the covariance matrix and we see that we are in essence choosing to represent our kernel function by a subset of its eigenfunctions.

3.13 Concluding Remarks

Gaussian Process provide a natural, flexible tool and a principled, probabilistic approach to modelling non-linear functions. The choice of a GP is a reflection of our understanding of the world we are seeking to model. If after training, our GP model fails to capture our expectations, then we must update our priors and modelling accordingly; this is the Bayesian approach. In this chapter we have introduced and detailed numerous properties of a GP. GPs have seen widespread use and been adopted to a number of applications. In this thesis we develop GPs in two major ways related

to functions and we outline the background of those areas, functional regression and arc lengths, in the next chapter.

Chapter 4

Functional Regression and Arc Lengths

4.1 Introduction

Functional regression is a paradigm in which units of observations are curves. We review the basics underpinning function regression, examine the common models and their limitations. The approaches and current methodologies are described from both the statistical literature and the Bayesian paradigm. Finally we review arc lengths of curves, which we can consider as a functional map of a curve to its length.

4.2 Functional Data Analysis

We do not attempt to cover all developments and results in Functional Data Analysis (FDA). Instead we focus on the core theory, developments and areas that we later expand upon with novel Gaussian Process (GP) extensions. For more comprehensive reviews of the literature please refer to the excellent reviews found in Morris (2015) and Wang et al. (2016).

Functional data is the paradigm where we assume a relationship between a func-

tional predictor, $X(t)$, and a corresponding response y; the units of observations, our inputs, are now the $X(t)$. Predictors are hence infinite dimensional objects and as such any method that is employed must necessarily perform dimensionality reduction. The introductory text by Ramsay and Silverman (2005a) developed the field of FDA and research has continued steadily ever since, as functional data becomes more commonplace. Examples include the prediction of temperature across Canadian weather stations (Ramsay and Silverman, 2005a), the classification of the lifetime of medflies (Müller and Stadtmüller, 2005) or prediction of the state-of-charge of a lithium-ion battery (Andre et al., 2011, Xu et al., 2013).

Functions are generally specified on a domain with observations consisting of groups or samples of $X(t)$ (whose arguments are often univariate, such as time) sampled on a grid. We face a unique problem where we must simultaneously consider the relationship within functions and between functions. Thus we must deal with two problems – replication, where we derive statistical relationships between repeated pairs of inputs and response, and regularisation, where we borrow strength across the functions points. The grid upon which functions are observed can be sparse, regular or irregular; we may be fortunate and have common observation points or face differing locations between replications of subjects.

A key idea in functional data analysis is to consider each $X(t)$ as a single unit, a structured object, rather than the traditional viewpoint of multivariate data analysis where we have a collection of data points. This view equips us with a simplicity of thought that empowers us to build models that contain complex structure both inter- and intra-function. The functional linear model relies on the assumption that the output, y, depends on the whole functional trajectory, $X(t)$, which is modelled through a weighted integral between the predictor and a coefficient function, $\beta(t)$:

$$y = \int X(t)\beta(t)\mathrm{d}t. \tag{4.1}$$

This can be seen as the extension of linear regression where we move from an inner product between independent vectors to the corresponding inner product in an infinite dimensional space. A large portion of the literature has focused on point estimates of $\beta(t)$ (Cardot and Sarda, 2005), though there exist some inferential capabilities and models. Non-parametric forms of functional regression have been proposed with the philosophy to let the data speak for themselves. Although these approaches may suffer due to sparse and irregularly spaced data, and the large gaps between datapoints. In contrast, parametric forms are able to overcome these issues by assuming a parametric form for the underlying function.

A salient feature of functional data problems is that the $X(t)$ are assumed smooth and continuous, however, crucially, only discrete data can be collected; we observe function values $\mathbf{x} = [x_1, \ldots, x_n]$, at corresponding points $\mathbf{t} = [t_1, \ldots, t_n]$. It becomes necessary to convert this data to functions for modelling purposes: representing as a function helps to reduce measurement error and leads to less bias. Often one would employ a two step process: 1) smoothing the trajectories and 2) modelling the response with the smoothed predictors. The main approaches for smoothing include using a basis function representation of the predictors typically with splines, Fourier series or wavelets. In the case where observations are dense, we can treat our functions as fully observed with low noise.

4.2.1 Illustrative Data

We describe two examples of functional data: Diffusion Tensor Imaging (DTI) scans and battery impedance measurements. From the John Hopkins University, diffusion tensor imaging was used to collect Fractional Anisotropy (FA) tract profiles for a number of patients. There are 381 scans from 142 patients, 100 of which have multiple sclerosis with 42 healthy patients. FA profiles are obtained along the Corpus Callosum (CCA) and Right Corticospinal (RCST) tracts; providing two functional predictors.

There are two scalar outputs: the MS status of the subject and their Paced Auditory Serial Additional Test (PASAT) score – a measurement of cognitive function related to speed and flexibility in processing auditory information. Either of these values could be regressed from the functional series (Goldsmith et al., 2011a).

A motivating use of GPs is their flexibility and ease of use for multiple forms of data. Another data set, Battery Impedance (BAT) data, presents a data analysis problem requiring the flexibility of GPs. The BAT data[1] contains real and imaginary impedance measurements of a Li-Ion battery cell over a range of frequencies (41 frequencies from 1Hz to 10 000 Hz evenly distributed in log-space). The corresponding scalar outputs are the cell's State of Charge (SOC) and Temperature (TEMP), which are correlated.

Figure 4.1 shows data samples from the FA-CCA and FA-RCST tracts and corresponding impedance curves for the BAT: we see highly structured curves, not simply disjoint points, supporting a functional approach.

4.2.2 Basis functions

The building blocks of FDA are the basis functions, $\{\phi(t)\}_{i=1}^{N}$ used to parametrise the functions $X(t)$. They also provide the mechanism by which we can incorporate regularisation, enforcing smoothness. We represent $X(t)$ in a basis expansion, with coefficients ζ_i, and write:

$$X(t) = \sum_{i=1}^{N} \zeta_i \phi_i(t). \tag{4.2}$$

Basis functions allow us to describe a function in terms of a linear combination of observation locations and provide a way to capture correlations in the function and create a framework for us to borrow strength across the function: without parametris-

[1]Thank you to David Howey, Energy and Power Group, University of Oxford, for providing the data set.

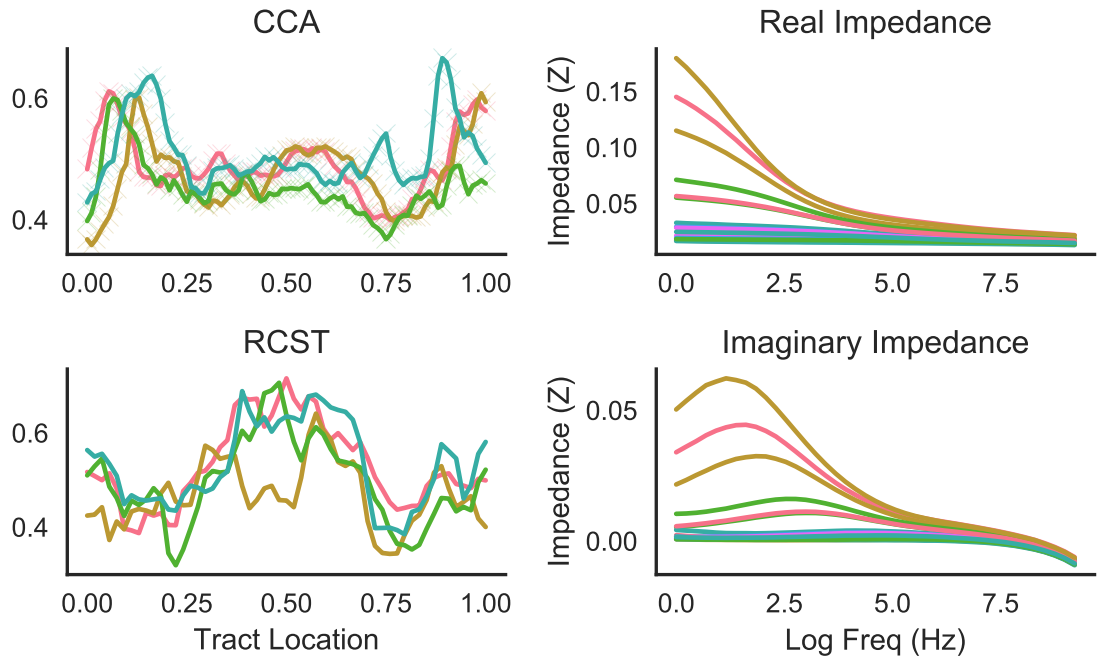

Figure 4.1: Left: DTI data; top image CCA tract, bottom image RCST tract. Right: Impedance curves for the battery cell. The x-axis is log(freq) and the y-axis the impedance value. Top image: real impedances, bottom image: imaginary impedances. The functional nature of the inputs is clearly visible.

ing we are left with only a collection of data points. In practice, basis representations enable us to compactly represent our infinite-dimensional objects in terms of finite-dimensional objects in terms of basis coefficients, thus allowing estimation and inference to be performed.

A number of common basis representations exist: such as splines, Fourier, wavelets, and principal components; each suited for different applications and data sets (Ramsay and Silverman, 2005a). For low-dimensional smooth data splines are preferable. For periodic data, Fourier basis are recommended, whilst wavelets have good local support making them a good choice for spiky data with discontinuities and non-stationarity. Principal components must be estimated from the data and are a viable choice for simple, structured data that have rapidly decaying eigenvalues; capturing local, high-frequency smooth data (Wang et al., 2016).

It is possible to regularise the basis coefficients in three main ways. First via truncation by eliminating all basis functions past a certain point: with Functional Principal Component Analysis (FPCA) we keep the first eigenfunctions that explain most of the functional variability, with wavelet and Fourier basis we keep only the first frequencies (ie a low-pass filter), for orthogonal polynomials we neglect above a certain order and with splines we truncate by considering knots at limited locations. Roughness penalties are included by penalising the second squared derivative, which involves a form of L^2 penalty on the basis coefficients, with fitting done via a penalised least squares. Finally we can enforce sparsity in the number of coefficients using L^1 penalisation.

4.2.3 Functional Principal Components

FPCA is a fundamental part of functional analysis. Our interest lies in determining the dominant modes of functional data; just as Principal Component Analysis (PCA) (Hotelling, 1933) determines the dominant directions in multivariate data, FPCA de-

termines the corresponding functional ones (Hotelling, 1933). We consider realisations of a square integrable stochastic process in L^2 defined on an interval $\mathcal{I}$ with mean and covariance:

$$\mu(t) = \mathbb{E}[X(t)], \tag{4.3}$$

$$\Sigma(s,t) = \mathrm{Cov}[X(s), X(t)] = \sum_{k=1}^{\infty} \lambda_k \psi_k(s)\psi_k(t), \tag{4.4}$$

where we have represented our covariance in an orthonormal basis expansion as a result of Mercer's Theorem. By the Karhunen-Loéve theorem, Chapter 3, we are then able to express the functions in terms of the eigenfunctions:

$$X(t) - \mu(t) = \sum_{k=1}^{\infty} \varepsilon_k \psi_k(t). \tag{4.5}$$

The coefficients of the expansion are given by:

$$\varepsilon_k = \int (X(t) - \mu(t))\psi_k(t)\mathrm{d}t. \tag{4.6}$$

Furthermore the following properties of the coefficients are assumed:

$$\mathbb{E}(\varepsilon_k) = 0, \mathrm{Var}(\varepsilon_k) = \lambda_k \text{ and } \mathbb{E}(\varepsilon_k\varepsilon_l) = 0 \text{ for } k \neq l. \tag{4.7}$$

Thus indicating that the coefficients are independently normally distributed with variance λ_k.

The ith realisation of the process is $X_i = X_i()$, and a data set would consist of n independent realisations of the data. In general we would assume the sampling to vary across realisations and denote the locations for the ith as $t_{i1}, \ldots, t_{in_i}$, with corresponding observations, $X_i = \{X_{i1}, \ldots, X_{in_i}\}$, where $X_{ij} = X_i(t_{ij})$. Additionally we assume measurement error of X_{ij} with random noise $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_{ij}^2)$, meaning we

have observations $Y_{ij} = X_{ij} + \epsilon_{ij}$. Generally errors are assumed to be homoscedastic and from an underlying smooth variance function $\sigma^2(t)$.

Estimation of the mean, for densely sampled points, can be the obtained with the average:

$$\hat{\mu}(t_{ij}) = \frac{1}{n}\sum_{i=1}^{n} Y_{ij}. \tag{4.8}$$

For sparse observations, one needs to smooth the data from across all observations to obtain the mean estimate (Yao et al., 2005), via any smoothing technique. The raw covariances are needed to estimate the covariance surface:

$$\Sigma_i(t_{ij}, t_{il}) = (Y_{ij} - \hat{\mu}(t_{ij}))(Y_{il} - \hat{\mu}(t_{il})), \quad j \neq l, i = 1, \ldots, n. \tag{4.9}$$

An estimate for the covariance $\hat{\Sigma}(s,t)$ is then obtained by averaging or smoothing the raw covariances from Equation (4.9) for the dense and sparse cases respectively. Practically we would discretise the covariance function and numerically estimate the eigenvalues $\hat{\lambda}_k$ and eigenvectors, with the resulting eigenfunctions, $\hat{\psi}_k$, obtained by interpolation of the eigenvectors.

The fitted covariance should by construction be positive definite and symmetric:

$$\tilde{G}(s,t) = \sum_{\lambda_k > 0} \hat{\lambda}_k \hat{\varphi}_k(s) \hat{\varphi}_k(t). \tag{4.10}$$

Finally the signal variance estimate $\hat{\sigma}^2$ is obtained by smoothing the diagonal estimates of the raw covariances to obtain $\hat{V}(t)$, and then computing:

$$\hat{\sigma}^2 = \frac{2}{|\mathcal{I}|}\int_{\mathcal{I}} (\hat{V}(t) - \tilde{G}(t,t))\mathrm{d}t, \text{ if } \hat{\sigma}^2 > 0; \text{ otherwise } \hat{\sigma}^2 = 0. \tag{4.11}$$

If observation of the functions is dense, then the k-th Functional Principal Com-

ponent (FPC) ξ_k can be estimated by numerical integration, implementing Equation (4.6) using the estimated eigenfunctions. Instead, if the observations are spare, we will obtain poor estimates, thus we will use the best linear unbiased predictors (Yao et al., 2005):

$$\hat{\xi}_k = \hat{\lambda}_k \hat{\varphi}_k^T \hat{\Sigma}_{Y_i}^{-1} (Y_i - \hat{\mu}), \tag{4.12}$$

where:

$$\hat{\Sigma}_{Y_i} = \tilde{G} + \hat{\sigma}^2 \mathbf{I}_{m_i}. \tag{4.13}$$

FPCA is a major tool in the analysis of longitudinal functional data: it is used as tool for simple and smooth functional data which can be explained with only the first few principal components. For all basis expansions that use p components, FPCA will explain the most variation in $X_i()$ in the L^2 sense, thus representing the function in the most parsimonious form. Care, however, is needed for complex, high-dimensional data with slowly decaying eigenvalues, as the FPCs may fail to adequately capture the structure of the functions.

4.2.4 Functional Linear Regression

We look at the most prevalent model in functional regression, the functional linear model. In classical linear regression, features are assumed to be i.i.d and the matrix of coefficients $\boldsymbol{\beta}$ is inferred using a pre-chosen loss function. In the situation where the inputs are a function of a latent variable, say t, we can utilise the functional equivalent. The equation that maps the functional inputs to the outputs is then:

$$y_i = \beta_0 + \int X_i(t)\beta(t)\mathrm{d}t + \epsilon_i, \tag{4.14}$$

where β_0 is a bias, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ is Gaussian noise and $\beta(t)$ is a weighting function; the functional extension of the classical regression coefficient β. Equation (4.14) is the Functional Linear Model (FLM); most work in functional predictor regression is based upon variants thereof.

The functional regression problem is now framed as solving for the $\beta(t)$ that minimises some loss function. Difficulties arise in dealing with the infinite-dimensional object on the right-hand side of Equation (4.14) and solving for an arbitrary function. The problem is simplified by expressing $X(t)$ and $\beta(t)$ as a sum of orthogonal basis functions, for example a Fourier or orthogonal polynomial basis:

$$X_i(t) = \sum_{j=1}^{\infty} \zeta_j^{(i)} \phi_j(t), \quad \beta(t) = \sum_{j=1}^{\infty} \beta_j \phi_j(t). \tag{4.15}$$

Each $X_i(t)$ is a single functional input, and could for example be a mixture of sinusoidal basis functions, with each $\phi_j(t)$ being the basis function evaluated at time t, determined by the coefficients $\zeta_j^{(i)}$.

Exploiting the orthogonality of our basis choice, $\int \phi_i(t)\phi_j(t)\mathrm{d}t = \delta_{ij}$, we can now write:

$$\int X_i(t)\beta(t)\mathrm{d}t = \sum_{j=1}^{\infty} \beta_j \zeta^{(i)}. \tag{4.16}$$

It would appear that not much has been gained: we have traded an integral for an infinite sum, another difficult problem. In practice, however, we can deal with a truncated basis, using p basis functions to represent $X_i(t)$ and $\beta(t)$ (Müller and Stadtmüller, 2005). The orthogonality condition is not a necessary condition, nor do they need to have the same basis expansion; thus, it can be relaxed. This would lead to Equation (4.16), but with β_j replaced by $\int \beta(t)\psi_j(t)\mathrm{d}t$, for the new (non-orthogonal) basis functions $\psi_i(t)$. Consequently, this integral never needs to be evaluated: our interest lies in solving for $\{\beta_j\}_{j=1}^{p}$. The functional regression problem can now be

simplified to:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j \zeta_j^{(i)} + \epsilon_i \tag{4.17}$$

$$y_i = \beta_0 + \boldsymbol{\beta}^T \boldsymbol{\zeta}^{(i)} + \epsilon_i, \tag{4.18}$$

where we have combined our parameters into vectors to simplify representation. It is now possible to infer the $\{\beta_j\}_{j=1}^p$. Many statistical approaches involve minimising a loss function with respect to the $\{\beta_j\}_{j=1}^p$ (Cardot and Sarda, 2005, Müller and Stadtmüller, 2005).

For a non-Gaussian response we can introduce a link function $g()$:

$$y_i = g\left(\beta_0 + \int X_i(t)\beta(t)\mathrm{d}t\right) + \epsilon_i, \tag{4.19}$$

The link function $g(\cdot)$ is typically assumed to be from the exponential family to simplify inference, however, it is also possible to estimate parameters without assuming a particular form of the link function (Acharyya and Ghosh, 2015, Müller and Stadtmüller, 2005).

Solving Equation (4.18) involves regularisation via truncation, penalisation or sparsity applied to either $X_i(t)$ or $\beta(t)$. Regularisation of the coefficient function, $\beta(t)$: 1) reduces collinearity in the regression fit: 2) increases interpretability of the coefficient estimates and: 3) potentially increases estimation and prediction efficiency by making use of the functional nature of the data to borrow strength across t. Regularization of the predictor functions, $X_i(t)$, serves the two fold purpose of reducing measurement error in predictors and accommodating functional predictors with sparse and/or irregular grids varying across the sampled functions. The best choice of basis and regularization will often be problem dependent.

Most methods in functional predictor regression follow this methodology, with

varying choices for the basis functions and differing regularization strategies. It is possible to introduce other model components such as non-Gaussian or correlated responses, inclusion of other non-functional fixed or random effects, extension to multiple functional predictors, and specific approaches for variable selection or inference (Morris, 2015).

The Bayesian Generalized Linear Model (James, 2002) utilises the classical functional regression model within a Bayesian framework. The response is assumed to be distributed according to an exponential family and a prior is placed over the natural cubic spline coefficients, ζ_i; regularization is implicit via knot selection. The ζ_i are treated as unobserved allowing inference to be performed using an Expectation Maximisation algorithm that infers the model parameters and the distribution over the coefficients.

In Cardot et al. (1999) the FLM is fit with FPCA through truncation over the FPC components and some relevant theory is presented. Generalised Functional Linear Models (GFLMs) are introduced by Müller and Stadtmüller (2005) who use a smoothed FPC basis with regularisation by truncation of the FPC scores. For functions that are sparse or irregular, the FPC scores are computed using PACE: Principal Analysis by Conditional Expectation (PACE) (Yao et al., 2005).

Group lasso is used by Fan et al. (2015) to perform variable selection across multiple functional predictors whilst James et al. (2009) introduce a method that encourages sparsity for interpretability, via an L^1 shrinkage penalty in the derivative.

In Goldsmith et al. (2011b) the authors develop a Variational Bayes method approximation for a Bayesian GFLM with FPCs for $X_i(t)$, B-Splines for $\beta(t)$, and regularization using a random walk prior on the spline coefficients (Crainiceanu and Goldsmith, 2010).

4.2.5 Non Linear Functional Regression

The FLM is a linear model and hence limited for non-linear data. A number of methods have been developed to incorporate non-linear structure into the functional regression model; this is achieved by introducing non-linear interactions in Equation (4.14).

Penalised spline regression is extended by Li and Marx (2008) to include general additive polynomial terms in Equation (4.14), with terms such as $\int\{X_i(t)\}^2\beta_2(t)\mathrm{d}t$. Full quadratic terms in Equation (4.14) are considered by Yao and Müller (2010):

$$y_i = \beta_0 + \int X_i(t)\beta_1(t) + \iint X_i(t)X_i(r)\beta_2(t,r)\mathrm{d}t\mathrm{d}s + \epsilon_i. \tag{4.20}$$

A principal component decomposition was used to estimate an empirical basis for $X_i(t)$ which was then used as a basis for the $\beta(t)$ with regularisation.

One of the first non-parametric extension of the FLM, was presented by James and Silverman (2005) who introduced Functional Adaptive Model Estimation, which extended projection pursuit regression (Ferraty et al., 2013) to functional regression. The model is given by:

$$y_i = \beta_0 + \sum_{k=1}^{N_k} f_k\left\{\int X_i(t)\beta_k(t)\mathrm{d}t\right\}, \tag{4.21}$$

with the f_k smooth unspecified functions to be learnt from the data. Natural cubic splines are used to represent the predictors $X_i(t)$, the coefficient functions $\beta_k(t)$ and $f_k()$, regularized using roughness penalties.

Non-linearities have been incorporated using index models,

$$y = f\left(\int X_i(t)\beta_1(t)\mathrm{d}t, \ldots, \int X_i(t)\beta_q(t)\mathrm{d}t\right) + \epsilon, \tag{4.22}$$

and additive index models,

$$y = f_1\left(\int X_i(t)\beta_1(t)\mathrm{d}t\right) + \cdots + f_q\left(\int X_i(t)\beta_q(t)\mathrm{d}t\right) + \epsilon, \tag{4.23}$$

for smooth f and f_i. A single index model (Jiang and Wang, 2011) is recovered for $q = 1$, with models for which $q > 1$ being difficult to fit due to the curse of dimensionality (Chen et al., 2011). Fan et al. (2015) allowed separate additive index models to account for multiple functional inputs.

Functional Additive Models (FAMs) (Müller and Yao, 2008) extends functional principal component regression to non-linear models involving additive non-parametric functions of the FPC scores;

$$y_i = \beta_0 + \sum_{k=1}^{N_k} f_k(\xi_{ik}) + \epsilon_i, \tag{4.24}$$

where the ξ_{ik} are the FPC scores truncated to N_k terms and the f_k are smoothed functions, using local kernels. The authors provide a time continuous version in Muller et al. (2013).

A Functional Generalized Additive Model (FGAM) (McLean et al., 2012) is developed extending generalised additive models (Hastie and Tibshirani, 1986) to the functional case for noise-free observations:

$$y_i = \beta_0 + \int f\{X_i(t), t\}\mathrm{d}t, \tag{4.25}$$

with $f\{\cdot,\cdot\}$ a smooth bivariate surface, parametrized using tensor B-splines, regularised by P-spline-type L^2 penalization. A Bayesian version was introduced (McLean et al., 2014) for sparsely observed functions, on irregular grids. A Markov Chain Monte Carlo (MCMC) and variational scheme were used to updated the FPC scores conditioned on eigenfunctions and mean curves, alongside the model coefficients. In

Wang and Ruppert (2013) the authors consider optimal prediction in an additive functional model in the framework of reproducing kernel Hilbert Space.

A functional extension for Reproducing Kernel Hilbert Spaces (RKHS) was developed in Kadri et al. (2010) to tackle non-linear functional regression. The authors developed an extension to RKHS to handle functional inputs and outputs, they demonstrated basic properties of these functional RKHS and provide the theoretical framework to develop non-linear regression models. Their model still requires modification to properly formulate methods to learn model parameters and have multiple data attributes. Other work (Yuan and Cai, 2010) has explored a regularisation approach under an RKHS framework.

4.2.6 Function to Function Models

The focus of this thesis is on function to scalar methods. However, the development of function to function methods is an important research activity and we present the main models; again we refer to Morris (2015) for a historical review. In Chapter 8 we highlight how our methodologies may be extended to function to function models.

Ramsay and Dalzell (1991) introduced the functional response regression model:

$$y_i(t) = \beta_0(t) + \int X_i(r)\beta(r,t)\mathrm{d}r + \epsilon_i(t), \tag{4.26}$$

where $\beta(t,r)$ is now a coefficient surface in t and r. There are several issues in solving 4.26 including: de-noising the predictor functions, reducing collinearity between functions, accommodating variable grid sizes in both functions, regularisation of the coefficient surface to construct efficient estimators and ensuring interpretability of models and regularisation within function modelling (Morris, 2015).

In Ramsay and Dalzell (1991), Equation (4.26) is fit using piecewise Fourier basis for $\beta(r,t)$, Besse (1991) develop a spline approach and in Ramsay and Silverman

(2005a) a separate basis is prescribed for both $X_i(r)$ and $Y_i(t)$,$\mathbf{\Psi}(r)$ & $\mathbf{\Phi}(t)$, leading to the expansion, $\beta(r,t) = \mathbf{\Psi}(r)'\mathbf{B}\mathbf{\Phi}(t)$, where $\mathbf{B}$ is a $N_x \times N_y$ matrix containing the coefficient surface, truncated to N_x and N_y basis functions.

The Functional Additive Model (FAM) (Müller and Yao, 2008) extends to functional responses, with both $Y(t)$ and $X(r)$ modelled using their FPC decompositions, computed using PACE. The estimation of the functional coefficient reduces to a series of independent linear models between the FPCs.

An alternative function to function model is the concurrent functional model where the observation of $y(t)$ depend only on the currently observed predictor function $X(t)$, thus simplifying the coefficient surface, $\beta(t,r) = \beta(t)$; which can be considered a special case of the varying coefficient model (Hastie and Tibshirani, 1993). This is the model given by:

$$y_i(t) = \alpha(t) + \beta(t)X_i(t) + \epsilon(t). \tag{4.27}$$

This model has been the focus of GPs approaches to functional regression (Shi et al., 2007, 2012, Wang and Shi, 2014). For a survey of non parametric concurrent modelling see Maity (2017).

Finally there has been work done on directly mapping function to functions without assuming a particular relationship as in Equation (4.26) – non-parametric functional regression. In this case one assumes a general form:

$$Y_i = \mathcal{F}(X_i) + \epsilon_i, \quad i = 1, \ldots, n \tag{4.28}$$

for a given i.i.d. sequence, where both X_i and Y_i are functions and $\mathcal{F}()$ the non-linear mapping between them. Ferraty and Vieu (2006), Ferraty et al. (2006) present a general purpose non-parametric approach to functional regression, however, their models are not competitive (McLean et al., 2012) whilst providing no uncertainty

around estimates, and Ferraty and Vieu (2009) introduce a non-parametric additive model using boosting.

4.3 Arc Lengths

Our attention now focuses on a slightly different problem concerning curves and associated characteristics. Consider the problem of determining the length of a curve. In the simplest case, a straight line, we would measure the straight line distance between the start and end points. How about a general curve? If we break down the curve into segments we can imagine an approximation to the length of the curve can be found by summing the straight line distances for each segment. In the limit of infinite segments we recover the true length of the curve. Let us make this precise.

Consider a Euclidean space $X = \mathbb{R}^n$ and a differentiable injective function $\gamma : [a, b] \to \mathbb{R}^n$. Then the image of the curve, $\gamma(t)$, is a curve with length given by the sum of all line segments as we take the length of the lines to zero:

$$\text{length}(\gamma) = \lim_{N\to\infty} \sum_{i=1}^{N} \big|\gamma(t_i) - \gamma(t_{i-1})\big|, \tag{4.29}$$

with $t_i = a + \frac{i(b-a)}{N} = a + i\Delta t$. Rewriting our sum:

$$\text{length}(\gamma) = \lim_{N\to\infty} \sum_{i=1}^{N} \left|\frac{\gamma(t_i) - \gamma(t_{i-1})}{\Delta t_i}\right| \Delta t_i, \tag{4.30}$$

$$= \int_a^b |\gamma'(t)| \mathrm{d}t. \tag{4.31}$$

Importantly the length of a curve is independent of the choice of curve parametrization. If we have a bijection defined as $\zeta : [a, b] \to [c, d]$, then we have $g = f(\zeta()) :$

$[c, d] \to \mathbb{R}^n$ is also a differentiable function.

$$\begin{aligned}
\text{length}(\gamma) &= \int_a^b |\gamma'(t)| \mathrm{d}t, && && (4.32)\\
&= \int_a^b |g'(\zeta(t))\zeta'(t)| \mathrm{d}t, && && (4.33)\\
&= \int_a^b |g'(\zeta(t))| \zeta'(t) \mathrm{d}t, && \zeta \text{ is non decreasing} && (4.34)\\
&= \int_c^d |g'(u)| \mathrm{d}u. && \text{change of variables} && (4.35)\\
&= \text{length}(g). && && (4.36)
\end{aligned}$$

For the specific case where $X = \mathbb{R}$, with the parametrization in terms of t, $\gamma = (y(t), x(t))$, we have:

$$\text{length}(\gamma) = \int_a^b |\gamma'(t)| \mathrm{d}t = \int_a^b \sqrt{y'(t)^2 + x'(t)^2} \mathrm{d}t. \tag{4.37}$$

If we can write $y = f(x)$, $x = t$, then our expression reduces to the commonly known expression for the arc length of a function:

$$s = \text{length}(\gamma) = \int_a^b |\gamma'(t)| \mathrm{d}t = \int_a^b \sqrt{1 + \left(\frac{df}{dx}\right)^2} \mathrm{d}t. \tag{4.38}$$

A curve with shortest length is known as a geodesic. Under a Euclidean metric this is the straight line. This can be derived by minimizing Equation (4.38) using the Euler-Lagrange equations. The concept of curve length can be defined on a Reimannian manifold with a given metric:

$$\text{length}(\gamma) = \int_a^b \mathrm{g}(\gamma'(t), \gamma'(t)) \mathrm{d}t \tag{4.39}$$

For an illustrative example, consider the curve defined by $\gamma(t) = (t, \cos t, \sin t)$. The

length of the curve is given by:

$$s = \int_a^b |\gamma'(t)| \mathrm{d}t, \tag{4.40}$$

$$= \int_a^b \sqrt{1 + (\sin t)^2 + (-\cos t)^2} \mathrm{d}t, \tag{4.41}$$

$$= \int_a^b \sqrt{2} \mathrm{d}t, \tag{4.42}$$

$$= (b - a)\sqrt{2}. \tag{4.43}$$

A quantity closely related to the arc length, is the energy of a curve:

$$\text{Energy}(\gamma) = \int_a^b |\gamma'(t)|^2 \mathrm{d}t. \tag{4.44}$$

The quadratic function, $f(x) = x^2$, is monotonic, hence, curves of minimal energy correspond to curves of minimal arc length (Hauberg et al., 2012). A similar form is obtained for the energy of a curve on a Riemannian manifold.

We can interpret the length of a curve as functional mapping from a parametrised curve to its length and can consider lengths as a functional problem. The length of a one dimensional GP was considered by Barakat and Baumann (1970), Miller and Freund (1956) who provide a closed form expression for the mean of the length by direct calculation.

4.4 Conclusion

In this chapter we have outlined the theory around functional regression, highlighting a lack of probabilistic non-linear models. In the next chapter we use GPs to extend a number of functional regression models. Then in Chapter 6 we compute a distribution for the arc length of a GPs.

Chapter 5

Functional Regression with Gaussian Processes

5.1 Overview

In this chapter, functional regression models using Gaussian Processes (GPs) are introduced and developed. As discussed in Chapter 4, a number of non-linear functional regression models exist in the literature; however, they lack flexibility and the probabilistic benefit afforded by a GP. As such, three functional regression models introduced in Chapter 4 are extended; developing a GP version of the Functional Index Model (Chen et al., 2011), the Functional Additive Model (Müller and Yao, 2008), and the Functional Generalized Additive Model (McLean et al., 2012). In each model, fully observed functional predictors are assumed, allowing inference and prediction to be performed in a straightforward manner. For each model, the GP method's predictive efficacy is tested against its counterparts on synthetic examples, showing improvement across the board. Experiments on real world data will be later discussed in Chapter 7.

5.2 Functional Index Models

5.2.1 Model

The first functional regression model investigated is the Functional Index Model (INDEX). A INDEX maps the weighted integral of the functional predictor and coefficient function into a one dimensional response using a non-linear function g:

$$y_i = g\left(\int_{\mathcal{I}} X_i(t)\beta(t)\mathrm{d}t\right) + \epsilon_i, \tag{5.1}$$

with g a non-linear function that is to be estimated alongside the coefficient function $\beta(t)$. The functional predictor is projected into a single index, $\int_{\mathcal{I}} X_i(t)\beta(t)\mathrm{d}t$, from which the nomenclature stems.

Previous approaches considered non parametric linear estimation (Müller and Stadtmüller, 2005) and local smoothing techniques (Chen et al., 2011) in order to simultaneously estimate $g()$ and $\beta(t)$. A natural non-parametric approach is to model g with a GP; where we now map the inner product of the functional predictors and the coefficient function to the response via a GP. This is a Gaussian Process Functional Index Model (GP-IND) which extends GP index models (Choi et al., 2011, Gramacy and Lian, 2010) to functional data and functional index models to incorporate GPs.

In order to perform inference, the integral in Equation (5.1) must be transformed. We proceed by representing $X(t)$ and $\beta(t)$ in a basis expansion:

$$X(t) = \sum_{i=1}^{N_x} \zeta_i \phi_i(t), \quad \beta(t) = \sum_{i=1}^{N_x} \beta_i \phi_i(t), \tag{5.2}$$

for an orthogonal basis $\{\phi_i(t)\}_{i=1}^{N_x}$ on the domain $\mathcal{I}$, where ζ_i and the β_i are the basis coefficients of the predictor and coefficient function. As discussed in Chapter 4, noting that the basis functions are orthogonal, we express the inner product of the functional

predictor and coefficient function as:

$$\int_{\mathcal{I}} X(t)\beta(t)\mathrm{d}t = \sum_{i=1}^{N_x} \zeta_i \beta_i = \boldsymbol{\zeta}^T \boldsymbol{\beta}, \tag{5.3}$$

with $\boldsymbol{\zeta} = [\zeta_1, \ldots, \zeta_{N_x}]^T$ and $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_{N_x}]^T$.

We are now in a position to describe the full generative model. For a functional input, each scalar response, y, is determined by a generative model, which can be written compactly as:

$$y = g(\boldsymbol{\zeta}^T \boldsymbol{\beta}) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma_y^2), \tag{5.4}$$

$$g \sim \mathcal{GP}(\mu(\cdot), K(\cdot)), \tag{5.5}$$

$$X(t) = \Phi(t)\boldsymbol{\zeta} + \varepsilon, \qquad \boldsymbol{\zeta} \sim \mathcal{N}(\boldsymbol{\mu}_\zeta, \Sigma_\zeta), \quad \varepsilon \sim \mathcal{N}(0, \sigma_x^2 \mathrm{I}), \tag{5.6}$$

where $\Phi(t)$ is the basis matrix with each entry given by the evaluation of the basis function at the point t_j: $\Phi(t)_{jk} = \phi_k(t_j)$, with $\phi_k(\cdot)$ the k^{th} basis function, $\boldsymbol{\zeta} \in \mathbb{R}^{N_x}$ are the basis coefficients, ε is Gaussian noise on the inputs and ϵ is Gaussian noise on the outputs. For a full generative model we have also prescribed a normal distribution for the coefficients $\boldsymbol{\zeta}$. The GP mean and covariance are given by $\mu(\boldsymbol{\zeta}^T\boldsymbol{\beta})$ and $K(\boldsymbol{\zeta}^T\boldsymbol{\beta}, \boldsymbol{\zeta}'^T\boldsymbol{\beta})$.

Performing inference now results in jointly learning $\boldsymbol{\beta}$ and the GP hyper-parameters, Θ, of our covariance function.

Consider a matrix of observed functions $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T$, their corresponding matrix of basis expansion $\mathbf{Z} = [\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_n]^T$ and a set of responses $\mathbf{y} = [y_1, \ldots, y_n]^T$.

The log-likelihood is then:

$$\log p(\mathcal{D}|\Theta) = \ln[p(\mathbf{y}|\mathbf{Z}\boldsymbol{\beta})p(\mathbf{X}|\boldsymbol{Z})p(\boldsymbol{\zeta})], \tag{5.7}$$

$$\begin{aligned} &= -\frac{1}{2}\mathbf{y}^T[K(\mathbf{Z}\boldsymbol{\beta},\mathbf{Z}\boldsymbol{\beta}) + \sigma_y^2]^{-1}\mathbf{y} - \frac{1}{2}\ln|K(\mathbf{Z}\boldsymbol{\beta},\mathbf{Z}\boldsymbol{\beta}) + \sigma_y^2| - \frac{N}{2}\ln(2\pi) \\ &- \sum_{i=1}^{N}\left[\frac{n_i}{2}\log\sigma_x^2 + \frac{1}{2\sigma_x^2}(\mathbf{x}_i - \Phi(t)\boldsymbol{\zeta}_i)^T(\mathbf{x}_i - \Phi(t)\boldsymbol{\zeta}_i)\right] \\ &- \sum_{i=1}^{N}\left[\frac{1}{2}\log|\Sigma_\zeta| + \frac{1}{2}(\boldsymbol{\zeta_i} - \boldsymbol{\mu}_\zeta)^T\Sigma_\zeta^{-1}(\boldsymbol{\zeta_i} - \boldsymbol{\mu}_\zeta)\right]. \end{aligned} \tag{5.8}$$

Taking a fully Bayesian approach would require us to integrate out the prior distributions:

$$p(\mathbf{y}|\mathbf{X}) = \int_\Omega p(\mathbf{y}|g)p(g|\mathbf{Z}\boldsymbol{\beta})p(\mathbf{X}|\mathbf{Z})p(\boldsymbol{\zeta})\mathrm{d}g\mathrm{d}\Theta\mathrm{d}\zeta. \tag{5.9}$$

We note that the $\boldsymbol{\zeta}$ appear non-linearly in the integral, making such an integral intractable. Thus, in order to perform inference we would, for example, employ a Laplace approximation or a variational inference approach (Titsias and Lawrence, 2010).

However, we can simplify things by assuming, as is common in the literature, that our trajectories are fully observed (Chen et al., 2011, McLean et al., 2012, Morris, 2015). Therefore, we assume we can represent our functional inputs exactly in terms of the fitted basis coefficients, and that we have access to them. Thus we do not need to consider the distributions over the parameters or the noise on the functional predictors.

Under this assumption our log likelihood is:

$$\log p(\mathbf{y}|\Theta) = -\frac{1}{2}\mathbf{y}^T[K(\mathbf{Z}\boldsymbol{\beta},\mathbf{Z}\boldsymbol{\beta}) + \sigma_y^2\mathrm{I}]^{-1}\mathbf{y} - \frac{1}{2}\ln|K(\mathbf{Z}\boldsymbol{\beta},\mathbf{Z}\boldsymbol{\beta}) + \sigma_y^2| - \frac{N}{2}\ln(2\pi). \tag{5.10}$$

5.2.2 Functional Index Kernel

By inspecting the distance term in our GP kernel, we can reinterpret our index kernel:

$$|\boldsymbol{\zeta}^T\boldsymbol{\beta} - \boldsymbol{\zeta}'^T\boldsymbol{\beta}|^2 = (\boldsymbol{\beta}^T\boldsymbol{\zeta} - \boldsymbol{\beta}^T\boldsymbol{\zeta}')^T(\boldsymbol{\beta}^T\boldsymbol{\zeta} - \boldsymbol{\beta}^T\boldsymbol{\zeta}'), \tag{5.11}$$

$$= (\boldsymbol{\zeta}^T\boldsymbol{\beta} - \boldsymbol{\zeta}'^T\boldsymbol{\beta})(\boldsymbol{\beta}^T\boldsymbol{\zeta} - \boldsymbol{\beta}^T\boldsymbol{\zeta}'), \tag{5.12}$$

$$= (\boldsymbol{\zeta} - \boldsymbol{\zeta}')^T\boldsymbol{\beta}\boldsymbol{\beta}^T(\boldsymbol{\zeta} - \boldsymbol{\zeta}'), \tag{5.13}$$

$$= (\boldsymbol{\zeta} - \boldsymbol{\zeta}')^T\Lambda(\boldsymbol{\zeta} - \boldsymbol{\zeta}'), \tag{5.14}$$

where $\Lambda = \boldsymbol{\beta}\boldsymbol{\beta}^T$. Henceforth, we can interpret our kernel with inputs $\boldsymbol{\zeta}^T\boldsymbol{\beta}$ as a kernel with the coefficients, $\boldsymbol{\zeta}$, as inputs where Λ is the Mahalanobis-like distance which allows us to write:

$$K(\boldsymbol{\zeta}^T\boldsymbol{\beta}, \boldsymbol{\zeta}'^T\boldsymbol{\beta}) = K(\boldsymbol{\zeta}, \boldsymbol{\zeta}'), \tag{5.15}$$

$$= \lambda^2 \exp\left(-\frac{(\boldsymbol{\zeta} - \boldsymbol{\zeta}')^T\Lambda(\boldsymbol{\zeta} - \boldsymbol{\zeta}')}{2\theta^2}\right), \tag{5.16}$$

where in the last line we have used an Squared Exponential (SE) kernel. This representation holds for any kernel with an isotropic distance measure. We now have a kernel with the coefficients as inputs and consider the $\{\beta_j\}_{j=1}^p$ as hyperparameters that need to be learned.

5.2.3 Identifiability

It is noted that to ensure identifiability of the index model, we require $\int_{\mathcal{I}} \beta(t)^2 \mathrm{d}t = 1$, or equivalently $\sum_{i=1}^{N_x} \beta_i^2 = 1$. Constrained optimisation is employed in the INDEX (Chen et al., 2011); however, for the GP-IND an appropriate prior needs to be specified. For GP index models Choi et al. (2011) ensure that the constraint $|\boldsymbol{\beta}| = 1$ by specifying either a von Mises distribution or uniform distribution on the N_x-unit sphere; such a prior is overly complicated for our needs. Due to the presence of the θ in Equation

(5.16) we can reparametrise our prior distribution over $p(\boldsymbol{\beta})$ as do the authors in Gramacy and Lian (2010); the θ acts as a scaling on $\boldsymbol{\beta}$ that allows us to relax the constraint. As such we pick an appropriate $p(\boldsymbol{\beta})$, such as a Laplace prior, and perform Maximum a Posteriori (MAP) inference jointly over the kernel and index parameters. Furthermore, our main objective is prediction and inference, thus we are satisfied with finding a $\boldsymbol{\beta}$ that performs well and not recovering the 'true' $\boldsymbol{\beta}$.

5.2.4 Inference & Prediction in the Index Model

A maximum a posteriori estimate of our model parameters $\Theta = \{\lambda, \{\beta_j\}_{j=1}^{p}, \theta\}$ requires the derivatives of our log-likelihood with respect to those parameters, namely:

$$\frac{\partial}{\partial \Theta_i} \log p(\mathbf{y}|\Theta)p(\boldsymbol{\beta}) = -\frac{1}{2}\text{Tr}\left(K^{-1}\frac{\partial K}{\partial \Theta_i}\right) + \frac{1}{2}\mathbf{y}^T K^{-1}\frac{\partial K}{\partial \Theta_i}K^{-1}\mathbf{y} + \frac{\partial}{\partial \Theta_i}p(\boldsymbol{\beta}), \quad (5.17)$$

where θ is the vector of hyperparameters of K. Setting the derivatives with respect to each parameter to zero and performing gradient descent optimises our objective.

Having learnt the model parameters we can compute the predictive means, $\mathbb{E}[y^*]$, and variances, $\mathbb{V}[y^*]$ using the equations derived in Chapter 3. Given training coefficients $\mathbf{Z} = [\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_n]^T$ with corresponding scalar responses $\mathbf{y} = [y_1, \ldots, y_n]^T$ we find that the prediction of a new functional predictor $X^*(t)$, with corresponding basis coefficients, $\boldsymbol{\zeta}^*$, is given by:

$$\mathbb{E}[y^*] = K(\boldsymbol{\zeta}^{*T}\boldsymbol{\beta}, \mathbf{Z}\boldsymbol{\beta})\big(K(\mathbf{Z}\boldsymbol{\beta}, \mathbf{Z}\boldsymbol{\beta}) + \sigma_y^2\mathrm{I}\big)^{-1}\mathbf{y}, \quad (5.18)$$

$$\mathbb{V}[y^*] = K(\boldsymbol{\zeta}^{*T}\boldsymbol{\beta}, \boldsymbol{\zeta}^{*T}\boldsymbol{\beta}) - K(\boldsymbol{\zeta}^{*T}\boldsymbol{\beta}, \mathbf{Z}\boldsymbol{\beta})\big(K(\mathbf{Z}\boldsymbol{\beta}, \mathbf{Z}\boldsymbol{\beta}) + \sigma_y^2\mathrm{I}\big)^{-1}K(\mathbf{Z}\boldsymbol{\beta}, \boldsymbol{\zeta}^{*T}\boldsymbol{\beta}). \quad (5.19)$$

Furthermore, visualisation of $\beta(t)$ can be performed by using the learned coefficients, $\{\beta_i\}_{i=1}^{N_x}$, and plotting $\beta(t) = \sum_{i=1}^{N_x}\beta_i\phi_i(t)$. Some authors suggest that the coefficient function provides interpretable insight into the importance of the functional

trajectories (James et al., 2009). We imagine that with more careful specification of the prior over $\boldsymbol{\beta}$ that we can ensure a more interpretable model; however, we do not address this.

5.2.5 Multiple Input Index Models

At times we may have multiple functional inputs. In the case of multiple inputs the index model can be readily extended in one of two ways:

$$y = g\left(\int_{\mathcal{I}} X_1(t)\beta_1(t)\mathrm{d}t, \ldots, \int_{\mathcal{I}} X_n(t)\beta_n(t)\mathrm{d}t\right), \tag{5.20}$$

or

$$y = g_1\left(\int_{\mathcal{I}} X_1(t)\beta_n(t)\mathrm{d}t\right) + \cdots + g_n\left(\int_{\mathcal{I}} X_n(t)\beta_n(t)\mathrm{d}t\right). \tag{5.21}$$

The functional predictors are again represented as a basis expansion resulting in a GP-IND model over multiple predictors. In the first form we have an Automatic Relevance Determination (ARD) like kernel and in the second we have separate additive kernels. Inference and prediction would be performed using maximum a posteriori.

5.2.6 Synthetic Validation

In this section we test GP-IND on synthetic non-linear functional to scalar data. It is shown that the functional kernel can capture the non-linear functional behaviour and performance is competitive when compared to the standard INDEX model. We use an experiment similar to a synthetic one previously described in (Chen et al., 2011). The functional predictors are:

$$X_i(t) = \sum_{k=1}^{4} \xi_{ik}\phi_k(t), \quad i = 1, \ldots, N, \tag{5.22}$$

where:

$$\phi_i(t) = \frac{1}{\sqrt{2}} \cos(i\pi t), \tag{5.23}$$

with each ξ_{ik} distributed as $\mathcal{N}(0, \lambda_k)$, $\lambda_1 = 1, \lambda_2 = \frac{1}{2}, \lambda_3 = \frac{1}{4}$ and $\lambda_4 = \frac{1}{8}$. We pick a small number of Fourier basis functions as in regular place in the literature (Chen et al., 2011). Scalar outputs are generated from the functional predictors via four different index functions:

$$\text{Linear}: \quad y = \int X(t)\beta(t)\mathrm{d}t + \epsilon, \tag{5.24}$$

$$\text{Qudratic}: \quad y = \left(\int X(t)\beta(t)\mathrm{d}t \right)^2 + \epsilon, \tag{5.25}$$

$$\text{Cosine}: \quad y = \cos\left(0.25 \int X(t)\beta(t)\mathrm{d}t \right) + \epsilon, \tag{5.26}$$

$$\text{Sine Hill}: \quad y = 0.1 \left(\int X(t)\beta(t)\mathrm{d}t \right) + \sin\left(0.5 \int X(t)\beta(t)\mathrm{d}t \right)^3 + \epsilon. \tag{5.27}$$

A set of 100 trajectories for $X_i(t)$ are generated over 50 equally spaced points in the interval $t \in [0, 1]$: samples of functional trajectories are shown in Figure 5.1. In each case we fix $\beta = \frac{1}{\sqrt{3}}\phi_1 + \frac{1}{\sqrt{3}}\phi_2 + \frac{1}{\sqrt{6}}\phi_3 + \frac{1}{\sqrt{6}}\phi_4$.

To ensure a fair comparison, and to reflect the real-world situation where we do not know the true functional form of $X_j(t)$ or $\beta(t)$, we implement our models without knowing the true number of Fourier basis. Coefficients of the Fourier basis are estimated from the observed $X_i(t)$, using least squares: we use 8 basis functions to fit the curves.

The GP-IND model is compared to two other non-parametric functional index models. The first is the Generalised Functional Linear Model (GFLM) (Müller and Stadtmüller, 2005), which fits a linear non-parametric link function for g. Secondly we compare against an index model with the Fourier coefficients as the inputs using the R package SIMEST (Kuchibhotla and Patra, 2017) based on (Kuchibhotla et al.,

2016), which we call INDEX. It was impractical to compare to Chen et al. (2011) for two reasons: firstly we pre-specify the number of coefficients ahead of time, whilst they cross-validate over the number of basis functions, and secondly, we were unable to develop a model that could reproduce their results, thus unable to provide a fair comparison. We use the squared exponential as the kernel function for the GP-IND.

Each model is trained on 67 curves, and we compare predictive performance on the remaining 33 held-out curves. For each model we compute the Root Mean Square Error (RMSE) of the predicted test points, y^*, compared to the test points, y:

$$\text{RMSE} = \sqrt{\frac{1}{33}\sum_{i=1}^{33}(y_i^* - y_i)^2}. \tag{5.28}$$

We consider a Signal to Noise (SNR) of 2 and 10, and generate 50 runs of the experiment.

Table 5.1 shows the average RMSE values for each model across the different index functions. In each case the GP-IND produces lower RMSE than the competitor models, empirically demonstrating the improved predictability of the GP-IND. For the Sine Hill situation, performance is slightly worse than the GFLM for SNR=2, as the noise signal makes it difficult to disentangle the true signal from a linear signal.

Functional to scalar data is difficult to visualise, however, under the assumption of a functional index model, we are able to visualise the function g by plotting the functional index against the response. In Figure 5.2 we plot the predicted values against the functional index, $\int X(t)\beta(t)\mathrm{d}t$, overlaid with the true values. We observe that we are able to both predict the correct values and recover a structurally correct link function: g has the right functional form, but sightly scaled and stretched.

Link Function	GP-IND		INDEX	
	SNR = 2	SNR = 10	SNR = 2	SNR = 10
Linear	2.229 (0.073)	**0.475**(**0.043**)	10.145 (4.75)	2.641 (0.516)
Quadratic	**10.562**(**0.52**)	**4.089**(**0.389**)	58.035 (17.268)	6.843 (1.096)
Cosine	**0.175**(**0.006**)	**0.056**(**0.007**)	0.855 (0.165)	0.153 (0.031)
Sine Hill	0.568 (0.018)	**0.120**(**0.009**)	3.411 (0.816)	0.639 (0.173)

Link Function	GFLM	
	SNR = 2	SNR = 10
Linear	**1.812**(**0.050**)	1.179 (0.070)
Quadratic	11.983 (0.450)	10.056 (0.440)
Cosine	0.226 (0.010)	0.165 (0.010)
Sine Hill	**0.514**(**0.010**)	0.346 (0.010)

Table 5.1: Experimental results comparing INDEX, GFLM and GP-IND for a range of link functions and two SNR values (best results in bold). For each combination of link function and SNR we generate 100 functional trajectories, 67 are used for training and 33 to test the models. We use the RMSE as the measurement of predictive performance and report the average RMSE (and standard error) over 50 experimental runs. The GP-IND gives lower values in almost every case and performs extremely well for low SNR values. The GFLM does well in the low SNR Sine Hill, likely due to the fact that the noise makes the response effectively linear. The INDEX model does poorly across the board, particularly for low SNR.

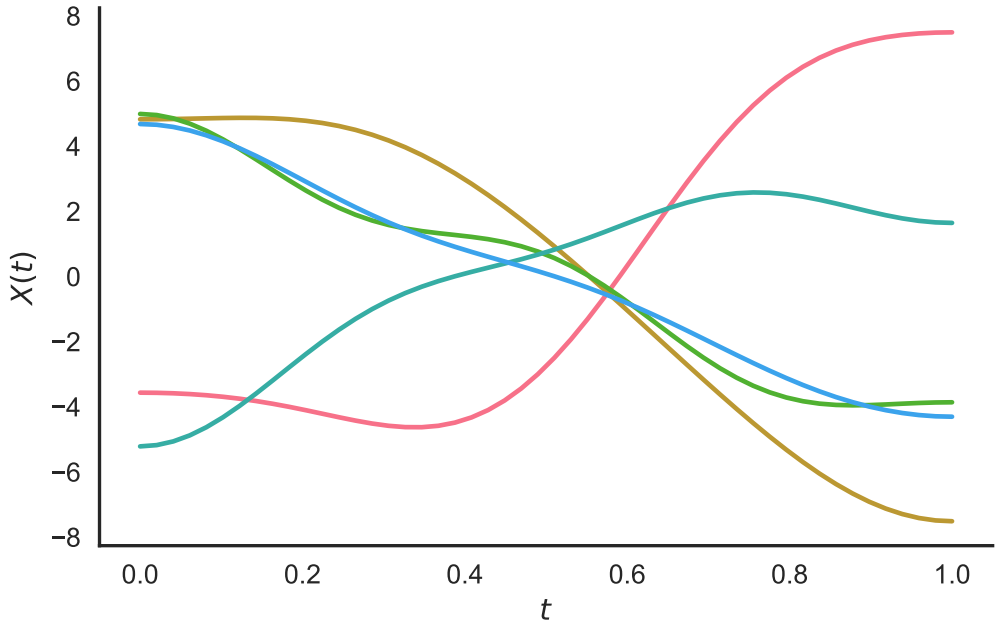

Figure 5.1: Five sample trajectories from the Fourier basis. These are the functional inputs used in the GP-IND synthetic experiment.

5.3 Functional Additive Models

5.3.1 Model

The GP-IND assumed that we could represent the relationship between the functional predictors and the response with an index model. However, such a representation might not always be true or we may not be able to represent our functions parsimoniously in a given time basis. As such, we now consider a functional model that is additive in the principal components, the Functional Additive Model (FAM) (Müller and Yao, 2008).

Consider the decomposition of $X(t)$ and $\beta(t)$ into their Functional Principal Component Analysis (FPCA) basis representation as in Yao et al. (2005), then we can write

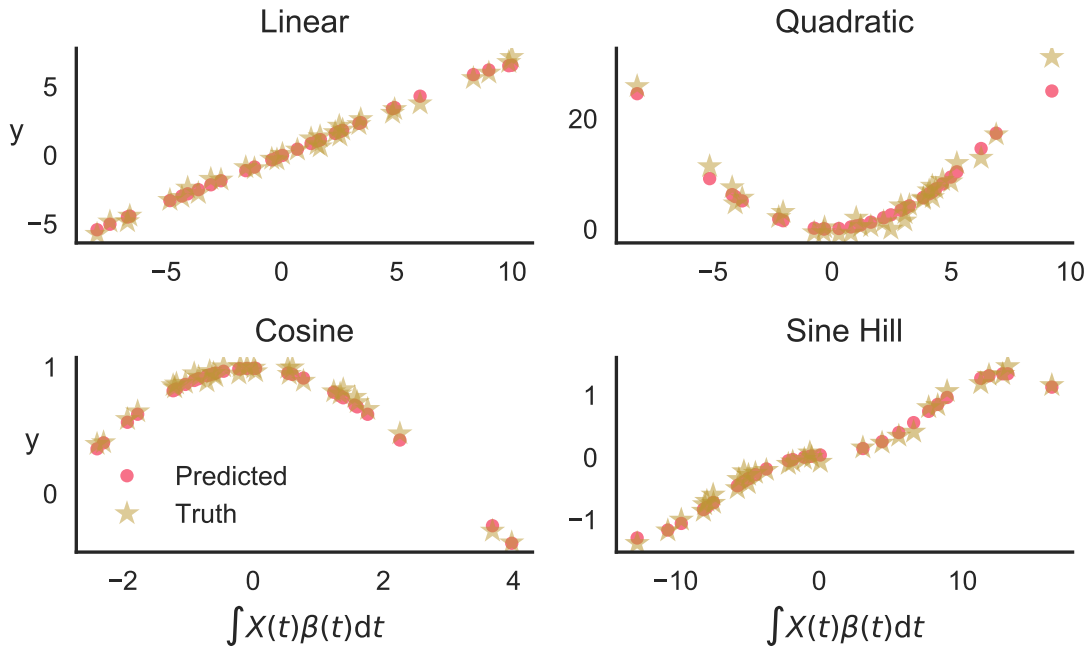

Figure 5.2: Plots of GP-IND responses, y, against the index, $\int X(t)\beta(t)\mathrm{d}t$, for each link function, with SNR $= 10$: predicted and true values are plotted. Clockwise from top left: Linear, Quadratic, Sine Hill and Cosine link functions. The GP-IND gives both the correct prediction and the right functional form.

the linear functional regression problem as:

$$y = \sum_{k=1}^{\infty} \beta_k \xi_k, \tag{5.29}$$

where the ξ_k are the Functional Principal Components of $X(t)$. By truncating to the first N_k principal components, linear regression is performed by solving Equation (5.29) for the β_k. Complexity can be naturally introduced by assuming a more general, non-linear, relationship between the $\{\xi_k\}$ and y, leading to the functional additive model (Müller and Yao, 2008):

$$y = \sum_{k=1}^{N_k} f_k(\xi_k), \tag{5.30}$$

where we have truncated to the first N_k Functional Principal Components (FPCs), and the f_k are non-linear functions to be learned from the data. In Müller and Yao (2008) the authors suggest any smoothing technique, though they use local linear smoothing to learn the $f_k()$. Such an approach requires a multitude of design choices and unnecessary fitting complications: which type of functions f_k to pick, determining an appropriate smoothing kernel and corresponding bandwidth, all whilst fitting parameters by undertaking a cross validation exercise. More generally the specification of an appropriate f_k is difficult.

We circumvent this difficulty by utilising a GP in Equation (5.30), thus equipping us firstly with a single design choice, the kernel, and allowing us to determine uncertainty bounds around predictions. Using a GP to fit the f_k, means we are fitting an additive model in the FPCs components. As such an additive kernel (Duvenaud et al., 2011) is required as each component is assumed independently distributed. The kernel is therefore:

$$k(\boldsymbol{\xi}, \boldsymbol{\xi}') = \sigma^2 \sum_{i=1}^{N_k} k_i(\xi_i, \xi_i'), \tag{5.31}$$

with $\boldsymbol{\xi} = [\xi_1, \ldots, \xi_D]^T$ the vector of FPCs and k_i a kernel for the ith FPC. This is a GP additive model over the FPCs, which we call the Gaussian Process Functional Additive Model (GP-FAM). If we were interested in higher order terms they could be included, for example, as:

$$k_{\text{add}_n}(\boldsymbol{\xi}, \boldsymbol{\xi}') = \sigma_n^2 \sum_{1 \leq i_1 \leq \cdots \leq i_n \leq N_k} \prod_{d=1}^{N_k} k_{id}(\xi_{id}, \xi_{id}'). \tag{5.32}$$

For interactions of order N_k, using the squared exponential we recover the standard ARD square exponential kernel. Using an FPCA decomposition implies independence between the inputs that feed into a GP, supporting the use of an additive model with only first order interactions. If we were to use a basis representation of another kind,

for example a Fourier decomposition, or lessen our assumption that the coefficients are independent, we could consider higher order interactions between the coefficients as in Equation (5.32).

Specifying a GP kernel enables us to learn our model probabilistically, thus equipping us with the benefits of GP regression. Learning the model now amounts to performing a maximum likelihood, maximum a posteriori or Markov Chain Monte Carlo (MCMC) estimate using the kernel in Equation (5.31).

The GP-FAM is simple to train: first we decompose our functions into their FPCs, and then train a Gaussian Process Generalized Additive Model (GP-GAM) model using the principal components as inputs. Let $\Xi = [\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_n]^T$, be the matrix of FPCs for the observed functions, with corresponding outputs, $\mathbf{y} = [y_1, \ldots, y_n]^T$; the model is then learned by optimising the log likelihood, with respect to the kernel hyperparamters, Θ:

$$\log p(\mathbf{y}|\Theta) = -\frac{1}{2}\mathbf{y}^T[K(\Xi,\Xi) + \sigma_y^2\mathrm{I}]^{-1}\mathbf{y} - \frac{1}{2}\ln|K(\Xi,\Xi) + \sigma_y^2\mathrm{I}| - \frac{N}{2}\ln(2\pi). \quad (5.33)$$

Prediction at a new functional predictor $X(t)^*$ follows by decomposing it into its FPC, $\boldsymbol{\xi}^*$, and using the GP predictive equations:

$$\mathbb{E}[y^*] = k(\boldsymbol{\xi}^*,\Xi)\big(K(\Xi,\Xi) + \sigma_y^2\mathrm{I}\big)^{-1}\mathbf{y}, \quad (5.34)$$

$$\mathbb{V}[y^*] = k(\boldsymbol{\xi}^*,\boldsymbol{\xi}^*) - k(\boldsymbol{\xi}^*,\Xi)\big(K(\Xi,\Xi) + \sigma_y^2\mathrm{I}\big)^{-1}k(\Xi,\boldsymbol{\xi}^*). \quad (5.35)$$

We demonstrate the power of the GP-FAM compared to FAM on synthetic data in the next section and real data in Chapter 7.

5.3.2 Synthetic Validation

We consider the experimental set up from Müller and Yao (2008) and show that the GP-FAM model provides better prediction quality whilst simultaneously providing us with uncertainty estimates. The functional predictors are:

$$X_i(t) = t + \sin(t) + \sum_{k=1}^{K} \xi_{ik}\phi_k, \tag{5.36}$$

The first part is a common mean function and the component basis functions are:

$$\phi_1(t) = -\cos\left(\frac{\pi t}{10}\right), \quad \phi_2(t) = \sin\left(\frac{\pi t}{10}\right), \quad 0 \leq t \leq 10, \tag{5.37}$$

The principal component scores are generated from two possible distributions:

1. $\xi_{ik} \sim \mathcal{N}(0, \lambda_k)$; and
2. $\xi_{ik} = \sqrt{\lambda_k}(G_{ik} - 4)/2, \quad G_{ik} \sim \text{gamma}(4, 1)$.

The first is a standard normal distribution, whilst the second is a right-skewed distribution.

The responses, y_i, are generated as $y_i = \sum_{k=1}^{2} f_k(\xi_{ik}) + \epsilon_i$. For the linear case we use $f_1(x) = 2x$, and $f_2(x) = 0.5x$, whilst in the non-linear case we choose $f_k(x) = x^2 - \lambda_k$ for $k = 1, 2$, to relate the principal components to the response. The noise on the response is specified as $\sigma_y^2 \sim \mathcal{N}(0, 0.1^2)$.

We generate 200 functional trajectories: 100 each for training and test. Prediction quality is measured by the RMSE. We compare the the GP-FAM against, the FAM and Functional Linear Functional Principal Component Model (FLM-PCA), using PACE: Principal Analysis by Conditional Expectation (PACE)[1]. We repeat our experiments 50 times for each combination of FPC generation, linear/ non-linear responses and for dense and sparse functional inputs. The results for all experiments are presented in

[1]Matlab software package available from `http://www.stat.ucdavis.edu/PACE/`.

Table 5.2. For all non-linear experiments, and all but one linear case, the GP-FAM, provides better predictive values as shown by a lower RMSE, demonstrating that the GP-FAM has superior predictive performance.

			FLM-PCA	FAM	GP-FAM
Normal	Linear	Dense	**0.695**(**0.008**)	0.789 (0.029)	0.704 (0.007)
		Sparse	4.352 (0.130)	11.168 (3.641)	**1.859**(**0.033**)
	Non Linear	Dense	6.000 (0.137)	2.494 (0.084)	**1.664**(**0.073**)
		Sparse	6.971 (0.194)	14.143 (1.066)	**4.078**(**0.137**)
Gamma	Linear	Dense	0.710 (0.011)	0.882 (0.049)	**0.707**(**0.008**)
		Sparse	4.616 (0.183)	9.806 (1.829)	**1.900**(**0.030**)
	Non Linear	Dense	6.471 (0.244)	2.882 (0.302)	**2.292**(**0.320**)
		Sparse	8.130 (0.323)	22.807 (4.668)	**4.956**(**0.253**)

Table 5.2: Synthetic data results for FLM-PCA, FAM and GP-FAM on linear and non-linear functions of the FPC, (best results in bold). Two cases are considered when generating the FPC to test the models. Dense corresponds to functional trajectories are densely observed in the unit interval, and sparse corresponds to sparsely observed trajectories. The FPC are generated either from a Normal: $\xi_{ik} \sim \mathcal{N}(0, \lambda_k)$, or Gamma distribution: $\xi_{ik} = \sqrt{\lambda_k}(G_{ik} - 4)/2, \quad G_{ik} \sim \text{gamma}(4, 1)$. For each combination we generate 200 functional trajectories, 100 are used for training and 100 to test the models. We use the RMSE as the measurement of predictive performance and report the average RMSE (and standard error) over 50 experimental runs. GP-FAM outperforms the FAM in all cases and provides good estimates for all linear experiments. Interestingly the FAM does much worse on the sparse case. One possible explanation is that the FPCs are very similar in the sparse case, making it difficult to learn accurate additive functions, further supporting the case for the GP-FAM.

5.4 Functional Generalised Additive Models

Finally we introduce the Gaussian Process Functional Generalized Additive Model (GP-FGAM), a new continuous functional additive model, building upon the Functional Generalized Additive Model (FGAM) introduced previously by McLean et al. (2012). The GP-FAM provides an additive non-linear model in the principal components of the functional predictors. Ultimately this is an assumption that our responses are fully explained by the frequency components of our inputs. For high frequency

predictors, it becomes difficult to capture all the functional dependency in the principal components. As such, we consider the continuous version of the functional additive mode in the time domain. The continuous version of the functional additive model is:

$$y = \int_{\mathcal{I}} f\{X(t), t\}\mathrm{d}t + \epsilon, \tag{5.38}$$

where $f\{\cdot, \cdot\}$ is a bivariate non-linear surface. This model was previously considered in McLean et al. (2012), where the term FGAM was introduced. The same authors later extended it to a Bayesian version with sparse and uncertain functional inputs in McLean et al. (2014). As already highlighted, we focus on the noise free functional predictor case and fill a gap by providing a probabilistic non-linear, time-additive model for functional regression. McLean et al. (2012) assume a separable form for $f\{\cdot, \cdot\}$ using a tensor spline basis, which they fit using penalized least squares. Splines are difficult to fit and their approach requires a number of model decisions a priori, including the number of basis functions (in both x and t), knot locations, the level of smoothing; all which make the model cumbersome to use. Furthermore, the FGAM provide no uncertainty guarantees about predictions.

We use a GP to model the function $f\{\cdot, \cdot\}$, providing a highly flexibly, probabilistic form for the latent function. Additionally, we retain the interpretability of the FGAM model, as we are able to infer the latent surface $f\{\cdot, \cdot\}$ as part of the inference process. In order to specify the GP we need to decide upon an appropriate kernel function for $f\{\cdot, \cdot\}$, which creates a large challenge in the GP-FGAM. From Equation (5.38) we see that $f\{\cdot, \cdot\}$ is a a two dimensional surface in t, and $X(t)$, which is also a function of t, and it is not immediately clear how we should define our kernel.

Following the approach of McLean et al. (2012) we assume that our function is

separable in x and t and that therefore the kernel for $f\{\cdot,\cdot\}$ inputs factorises as:

$$k_f(X(t), X(t'), t, t') = k_x(X, X')k_t(t, t'), \tag{5.39}$$

where $k_x()$ and $k_t()$ are appropriate kernels defined over X and t respectively. An open question is how to define the kernel for y from Equation (5.39)? If we were to ignore the $X(t)$ momentarily and consider the kernel as only a function of t, then we would write:

$$y = \int_{\mathcal{I}} f(t)\mathrm{d}t. \tag{5.40}$$

This tells us that for a GP f with kernel k_f, the variance for y would be (O'Hagan, 1992):

$$\mathbb{V}(y) = \int_{\mathcal{I}}\int_{\mathcal{I}} k_f(t, t')\mathrm{d}t\mathrm{d}t'. \tag{5.41}$$

Substituting in our specific k_f:

$$\mathbb{V}(y) = \int_{\mathcal{I}}\int_{\mathcal{I}} k_f(X(t), X(t'), t, t')\mathrm{d}t\mathrm{d}t', \tag{5.42}$$

$$= \int_{\mathcal{I}}\int_{\mathcal{I}} k_x(X(t), X(t'))k_t(t, t')\mathrm{d}t\mathrm{d}t'. \tag{5.43}$$

The covariance between two input functions $X_i(t)$ and $X_j(t')$ can now be readily observed as:

$$\mathbb{C}(y_i, y_j) = \int_{\mathcal{I}}\int_{\mathcal{I}} k_x(X_i(t), X_j(t'))k_t(t, t')\mathrm{d}t\mathrm{d}t'. \tag{5.44}$$

Specification of the GP model is now a question of choosing appropriate kernels for $X()$ and t. Ideally we would like to compute the integral in Equation (5.44) analytically. Difficulty arises due to the dependence of $k_x()$ on $X()$ which itself depends on t,

making such integrations intractable for a large choice of kernels.

Alternatively if we consider our observations of $X_i()$ to be sufficiently dense, then without loss of model accuracy we may move away from parametrising the $X_i()$ and using the values as direct inputs into $k_x()$. We then compute the covariance in Equation (5.44) numerically using a quadrature rule, for example Simpson's rule. Should we be presented with sparse samples then we could pre-smooth the functional predictors: this approach is commonly used in the literature and suggested in McLean et al. (2012).

Training the GP-FGAM model now proceeds by estimating the kernel parameters via a MCMC method or Maximum Likelihood Estimate (MLE). We note that learning in the model can be slow due to the construction of the covariance function. For each i, j entry a double integral needs to be computed. Furthermore, the covariance matrix must be reconstructed for each step taken in the optimiser or when a new sample is drawn. The upside of using the GP-FGAM is a principled learning approach, coupled with uncertainties on the outputs alongside the potential to include arbitrarily complex kernels for $k_x()$ and $k_t()$.

Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]^T$ be a matrix of the functional predictors, with $K(\mathbf{X}, \mathbf{X})_{ij} = \int_{\mathcal{I}} \int_{\mathcal{I}} k_x(X_i(t), X_j(t'))k_t(t, t')\mathrm{d}t\mathrm{d}t'$ the full covariance matrix between each of the functional predictors, and $\mathbf{y} = [y_1, \ldots, y_n]^T$ the corresponding responses. Prediction of the output of a new functional predictor $X(t)^*$ is then given by the equations:

$$\mathbb{E}[y^*] = k(X(t)^*, \mathbf{X})(K(\mathbf{X}, \mathbf{X}) + \sigma_y^2 \mathrm{I})^{-1}\mathbf{y}, \tag{5.45}$$

$$\mathbb{V}[y^*] = k(X(t)^*, X(t)^*) - k(X(t)^*, \mathbf{X})(K(\mathbf{X}, \mathbf{X}) + \sigma_y^2 \mathrm{I})^{-1}k(\mathbf{X}, X(t)^*), \tag{5.46}$$

where $k(X(t)^*, \mathbf{X})$ is a $1 \times n$ matrix with the jth column given by:

$$k(X(t)^*, \mathbf{X})_j = \int_{\mathcal{I}} \int_{\mathcal{I}} k_x(X(t)^*, X_j(t'))k_t(t, t')\mathrm{d}t\mathrm{d}t'. \tag{5.47}$$

5.4.1 Generating the surface

A by-product of the GP-FGAM is the ability to construct the underlying surface $f\{\cdot,\cdot\}$ conditioned on observations of y. This is a slightly unusual situation, where we are interested in unobserved function values conditioned on observed integral quantities. Fortunately, by using a GP to represent the surface, we are able to compute the expected values and obtain an uncertainty estimate for the surface. Writing down the joint distribution of y and $f\{\cdot,\cdot\}$ in matrix notation:

$$\begin{bmatrix}\mathbf{y}\\ \mathbf{f}\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}0\\ 0\end{bmatrix}, \begin{bmatrix}K(\mathbf{X},\mathbf{X})+\sigma_y^2\mathrm{I} & k_{fy}(f,\mathbf{X})\\ k_{yf}(\mathbf{X},f) & k_f(X(t^*),X(t^*),t^*,t^*)\end{bmatrix}\right) \tag{5.48}$$

with k_f as defined in Equation (5.39) and the cross covariance is given as:

$$[k_{fy}(f,\mathbf{X})]_j = \int_{\mathcal{I}} k_x(X(t^*),X_j(t'))k_t(t^*,t')\mathrm{d}t'. \tag{5.49}$$

Therefore, the posterior over the surface at a point (x^*,t^*) is given by:

$$\mathbb{E}(f(x^*,t^*)) = k_{fy}(f,\mathbf{y})(K(\mathbf{X},\mathbf{X})+\sigma_y^2\mathrm{I})^{-1}\mathbf{y}, \tag{5.50}$$

$$\mathbb{V}(f(x^*,t^*)) = k_f(x^*,x^*,t^*,t^*) - k_{fy}(f,\mathbf{y})(K(\mathbf{X},\mathbf{X})+\sigma_y^2\mathrm{I})^{-1}k_{fy}(\mathbf{y},f). \tag{5.51}$$

5.4.2 Synthetic Data Validation

We investigate the power of the GP-FGAM in synthetic experiments, replicating those in McLean et al. (2012), but only comparing the GP-FGAM against FGAM; using the implementation of McLean et al. (2012) provided in R. One hundred replicates of data are generated, each consisting of 100 curves sampled at 200 equally-spaced points in

the interval $[0, 1]$, with functional predictors defined as:

$$X_i(t) = \sum_{j=1}^{J} \frac{2}{j}[\mathcal{Z}_{1ij}\phi_{1j}(t) + \mathcal{Z}_{2ij}\phi_{2j}(t)], \tag{5.52}$$

with basis functions:

$$\phi_{1j}(t) = \sqrt{2}\cos(\pi jt), \quad \phi_{2j}(t) = \sqrt{2}\sin(\pi jt). \tag{5.53}$$

The coefficients are drawn from a normal distribution $\mathcal{Z}_{hij} \sim \mathcal{N}(0, \frac{4}{j^2})$, $h = 1, 2$ and $\gamma_j = \frac{2}{j}$. We consider $J = 5$ and $J = 500$, with the former resulting in smoother predictor trajectories. Two true surfaces are considered, one where the linear assumption holds, Equation (5.54), and one where the non-linear one is true, Equation (5.55):

$$f(x, t) = xt, \tag{5.54}$$

$$f(x, t) = -0.5 + \exp\left(-\frac{x^2}{5^2} - \frac{(t - 0.5)^2}{0.3^2}\right). \tag{5.55}$$

The error variance in the signal for each samples is given as:

$$\sigma_y^2 = \frac{1}{N-1}\sum_{i}^{N}\left(\int_{\mathcal{I}} f(X_i(t), t)\mathrm{d}t - \frac{1}{N}\sum_{i=1}^{N}\int_{\mathcal{I}} f(X_i(t), t)\mathrm{d}t\right), \tag{5.56}$$

with the resulting signal to noise ratio (SNR) defined by $\text{SNR} = \frac{\sigma_y}{\sigma}$. Four values of signal to noise are investigated SNR$= 1, 2, 4, 8$.

For each data replicate, 67 curves are used to fit the models and 33 are used for prediction. Performance of each model is again measured using the RMSE. Table 5.3 shows the results for all SNR. The GP-FGAM provides lower RMSE for every case, showing improvement over the FGAM.

In Figure 5.3 we plot the estimated mean surface for the hill for each SNR. As the SNR increases we obtain a more accurate estimate to the true surface, demonstrating

Surface	Fourier	FGAM			
		1	2	4	8
hill	J=5	0.107 (0.004)	0.077 (0.003)	0.058 (0.002)	0.049 (0.004)
	J=500	0.097 (0.003)	0.067 (0.002)	0.049 (0.001)	0.041 (0.003)
linear	J=5	1.170 (0.024)	0.883 (0.024)	0.609 (0.020)	0.428 (0.011)
	J=500	1.203 (0.032)	0.849 (0.019)	0.598 (0.015)	0.430 (0.014)
Surface	Fourier	GP-FGAM			
		1	2	4	8
hill	J=5	**0.095(0.017)**	**0.067(0.012)**	**0.048(0.009)**	**0.034(0.007)**
	J=500	**0.089(0.017)**	**0.062(0.012)**	**0.043(0.008)**	**0.032(0.006)**
linear	J=5	**1.136(0.024)**	**0.820(0.019)**	**0.564(0.012)**	**0.391(0.008)**
	J=500	**1.115(0.023)**	**0.799(0.017)**	**0.555(0.011)**	**0.392(0.009)**

Table 5.3: Experimental results comparing FGAM and GP-FGAM, (best results in bold). We compare for the hill and linear surfaces, across a range of SNR values (1,2,4,8) and two values of J (5, 500), corresponding to the number of Fourier components in the functional inputs. For each surface function and SNR we generate 100 functional trajectories, 67 are used for training and 33 to test the models. We use the RMSE as the measurement of predictive performance and report the average RMSE (and standard error) over 100 experimental runs. The GP-FGAM provides lower RMSE values in all cases, outperforming the FGAM in predictive capability.

the models ability to learn the true surface, indicating the potential for interpretable by-products of the GP-FGAM.

5.4.3 Multiple Predictors

For cases which involve multiple predictors we can easily incorporate these into our model:

$$y = \int_{\mathcal{I}} f_1\{X_1(t), t\}\mathrm{d}t + \int_{\mathcal{I}} f_2\{X_2(t), t\}\mathrm{d}t. \tag{5.57}$$

Specification of the GP-FGAM follows by including an additional additive term in the covariance function to account for $f_2\{\cdot,\cdot\}$. In theory, we could arbitrarily introduce

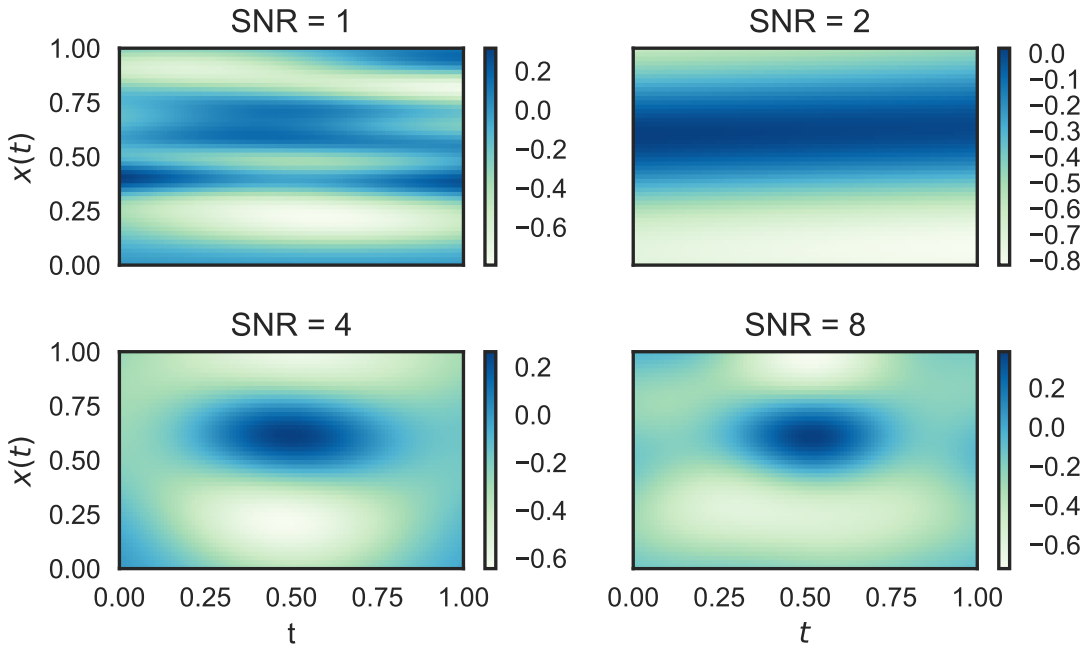

Figure 5.3: The fit hill surface using GP-FGAM. Each plot shows the mean surface for a different value of SNR. The GP-FGAM is able to better determine the true surface as the SNR increases.

Feature	GP-IND	GP-FAM	GP-FGAM
Inputs:	$\boldsymbol{\zeta}$	$\boldsymbol{\xi}$	$X(t)$
Inference:	MLE or MAP	MLE or MAP	MLE or MAP
Additive:	No	In frequency	In time
Interpretable:	Yes – $\beta(t)$	No	Yes – $f\{\cdot,\cdot\}$
Multiple Inputs:	Yes	Yes	Yes
Strength:	Smooth functions	Low frequency signals	Dense trajectories
Weakness:	Rough functions	High frequency signals	Computational Cost

Table 5.4: Summary of the GP functional models, GP-IND, GP-FAM, GP-FGAM. We outline the core components of each, their strengths and weaknesses.

extra functional predictors into Equation (5.57), though with high dimensional functional inputs it may be computational prohibitive to learn an accurate model without careful prior specification.

5.5 Conclusion and Future Directions

In this chapter, we have extended functional regression methods in three ways using GPs by introducing the GP-IND, GP-FAM and the GP-FGAM. For each model that was developed we made the assumption that the functional predictors were fully observed. Table 5.4 summarises the core features of each model, their strengths and weaknesses. An inference method is developed to implement these models and superior performance, measured by RMSE, is demonstrated in comparison with synthetic experiments. For a full Bayesian approach we would incorporate uncertain or sparse functional predictors, and this is an aim for future work. In this chapter each model has been considered in isolation. However, it would be of interest to see to what extent each model performs under the synthetic example of the other and compare to a GP baseline. We defer application of these functional methods to real world data until Chapter 7. The next chapter considers a different functional GP problem – the arc length of a GP.

Chapter 6

Gaussian Process Arc Lengths

6.1 Overview

In this chapter we examine the distribution of GP arc lengths. We begin discussion by considering the one dimensional case before moving onto the vector valued situation. The mean and variance for the arc length of both the prior and posterior of the GP are computed. A novel approach, whereby we use the distribution of the integrand to derive the arc length distribution, is presented. Numerical results show the fidelity of our results and we examine the resulting distributions and their properties.

6.2 One Dimensional Gaussian Processes

6.2.1 Samples and Lengths

First we consider the one dimensional case, where we develop a new method to derive the expected length: an approach that we will later use to derive results in the vector case. Consider a GP, f and a corresponding derivative process f':

$$f \sim \mathcal{GP}(\mu, K), \quad f' \sim \mathcal{GP}\left(\mu_{f'}, \sigma^2_{f'}\right), \tag{6.1}$$

where $\mu_{f'}$ and $\sigma^2_{f'}$ are the mean and variance of the derivative GP which can be found in terms of μ and K. The posterior of a newly observed point t^* is again normally distributed and defined in Chapter 3:

$$\bar{f}_* = \mathbb{E}[y_*|t_*, \mathbf{t}, \mathbf{y}] = K(t_*, \mathbf{t})(K(\mathbf{t}, \mathbf{t}) + \sigma_n^2 \mathrm{I})^{-1}\mathbf{y}, \tag{6.2}$$

$$\mathbb{V}[y_*|t_*, \mathbf{t}, \mathbf{y}] = K(t_*, t_*) - K(t_*, \mathbf{t})(K(\mathbf{t}, \mathbf{t}) + \sigma_n^2 \mathrm{I})^{-1}K(\mathbf{t}, t_*), \tag{6.3}$$

where $\mathbf{t} = [t_1, \ldots, t_n]$ are the observation locations and σ_n^2 is the observation noise. The derivative of the posterior mean can be calculated:

$$\frac{\partial \bar{f}_*}{\partial t_*} = \frac{\partial K(t_*, \mathbf{t})}{\partial t_*}(K(\mathbf{t}, \mathbf{t}) + \sigma_n^2 \mathrm{I})^{-1}\mathbf{y}, \tag{6.4}$$

whenever the derivative of the kernel function can be calculated. The full distribution over the derivative process is:

$$p\left(\frac{\partial f_*}{\partial t_*}\right) = \mathcal{N}\left(\frac{\partial \bar{f}_*}{\partial t_*}; \frac{\partial^2 K(t_1, t_2)}{\partial t_1^* \partial t_2^*} - \frac{\partial K(t_*, \mathbf{t})}{\partial t_*}(K + \sigma_n^2 \mathrm{I})^{-1}\frac{\partial K(\mathbf{t}, t_*)}{\partial t_*}\right). \tag{6.5}$$

Then the arc length is:

$$s = \int_0^T \sqrt{1 + (f')^2}\mathrm{d}t. \tag{6.6}$$

We are interested in computing $\mathbb{E}[s]$ and $\mathbb{V}[s]$, which require integrating s and s^2 against the distribution over f'. Instead of attempting to compute these quantities directly, we first determine the probability distribution over the arc length integrand $(1 + (f')^2)^{1/2}$, and then compute the arc length distribution.

Before we dive into any computations let us first inspect the shape of the distributions of the arc length. We look at two common kernels: the SE and the Matérn32 (MAT32) and a range of kernel parameters. Arc lengths are a positive variable, thus

we expect the distribution over arc lengths to have positive support.

In Figure 6.1 we plot a number of sample draws from our GP and the corresponding distribution of lengths. To compute the histogram distribution we generate draws from a GP prior and compute the length numerically. We can see similarity within the distributions of lengths and an apparent dependence on both the kernel and its hyperparameters. In this chapter we quantify this relationship.

6.2.2 Direct Computation of Arc Length Statistics

We are interested in calculating the distribution of the arc length. A direct approach would involve directly computing the statistics of the arc length – the mean and the variance. To highlight the difficulty of direct computation, consider $\mathbb{E}[s]$; integrating the expression for the arc length against the distribution of f':

$$\mathbb{E}[s]_{p_{f'}} = \int_{-\infty}^{\infty}\int_{0}^{T} \sqrt{1+(f'(t))^2}p(f'(t))\mathrm{d}f'\mathrm{d}t, \tag{6.7}$$

$$= \int_{-\infty}^{\infty}\int_{0}^{T} \sqrt{1+(f'(t))^2}\mathcal{N}\left(f'(t);\mu_{f'},\sigma^2_{f'}\right)\mathrm{d}f'\mathrm{d}t, \tag{6.8}$$

$$= \int_{-\infty}^{\infty}\int_{0}^{T} \sqrt{1+(f'(t))^2}\frac{1}{\sqrt{2\pi}(\sigma^2_{f'})^{1/2}}\exp\left(-\frac{(f'(t)-\mu_{f'})^2}{2\sigma^2_{f'}}\right)\mathrm{d}f'\mathrm{d}t, \tag{6.9}$$

where $\mu_{f'}$ and $\sigma^2_{f'}$ could also be functions of t; for example in the case of $p(f'(t))$ representing a posterior distribution, we would use the expressions in Equation (6.5). The difficulty lies in evaluating the nested integrals in Equation (6.9). It is possible to rearrange the order of the integrals to get:

$$\mathbb{E}[s]_{p_{f'}} = \int_{0}^{T}\mathrm{d}t\int_{-\infty}^{\infty}\mathrm{d}f'\sqrt{1+(f'(t))^2}\frac{1}{\sqrt{2\pi}(\sigma^2_{f'})^{1/2}}\exp\left(-\frac{(f'(t)-\mu_{f'})^2}{2\sigma^2_{f'}}\right). \tag{6.10}$$

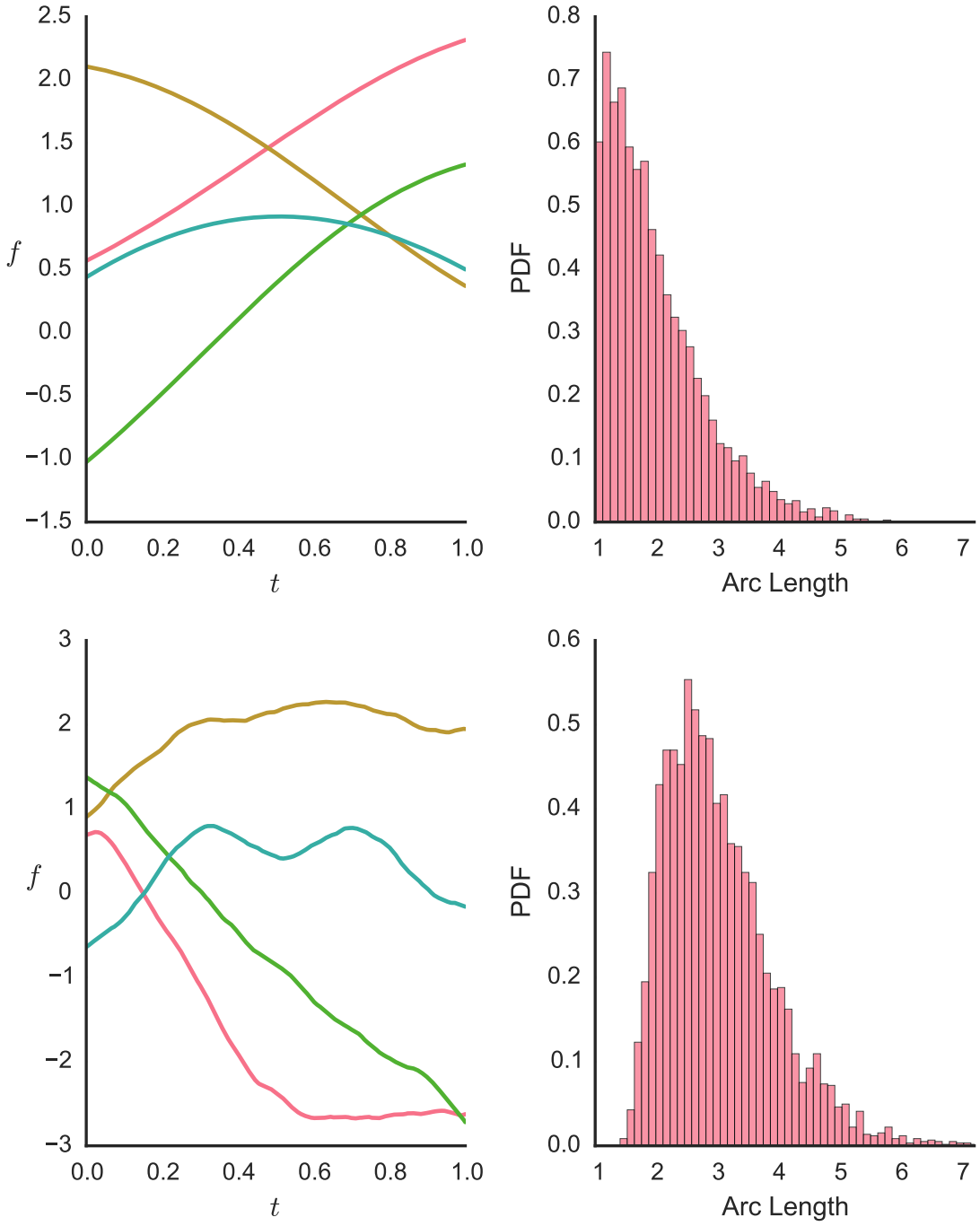

Figure 6.1: Example draws from the SE (top) and Matérn32 (bottom) kernels and the corresponding distribution of lengths. Changing kernels and parameters affects the distribution of length.

However, the innermost Gaussian integral is intractable due to the non-linear arc length function. Even if we are able to solve that, we would still have an intractable integral, as $\mu_{f'}$ and $\sigma^2_{f'}$ are both non-linear functions of t.

Linearisation around the mode

A possible solution to our non-linear quandary would be to linearise our function around the mode of the GP, $\mu_{f'}$. To this end we consider the Taylor expansion of $g(f) = (1 + (f')^2)^{1/2}$, around $\mu_{f'}$:

$$g(f) = g(\mu_{f'}) + g'(\mu_{f'})(f' - \mu_{f'}) + \text{higher order terms}, \tag{6.11}$$

$$\approx (1 + \mu^2_{f'})^{1/2} + \mu_{f'}(1 + \mu^2_{f'})^{-1/2}(f' - \mu_{f'}). \tag{6.12}$$

Now we can compute:

$$\mathbb{E}[s]_{p_{f'}} = \int_0^T \mathrm{d}t \int_{-\infty}^{\infty} \sqrt{1 + (f')^2}\mathcal{N}\left(f'(t)|\mu_{f'}, \sigma_{f'}\right) \mathrm{d}f', \tag{6.13}$$

$$\approx \int_0^T \mathrm{d}t \int_{-\infty}^{\infty} [(1 + \mu^2_{f'})^{1/2} + \mu_{f'}(1 + \mu^2_{f'})^{-1/2}(f' - \mu_{f'})]\mathcal{N}\left(f'(t)|\mu_{f'}, \sigma_{f'}\right) \mathrm{d}f', \tag{6.14}$$

$$= \int_0^T \mathrm{d}t[(1 + \mu^2_{f'})^{1/2} + \mu_{f'}(1 + \mu^2_{f'})^{-1/2}(\mu_{f'} - \mu_{f'})], \tag{6.15}$$

$$= \int_0^T \mathrm{d}t\sqrt{1 + \mu^2_{f'}}. \tag{6.16}$$

From Equation (6.16) we see that the expected value of the arc length is the arc length of the posterior computed using the derivative mean function; this is intuitively what we would expect.

Previous work

Previously Barakat and Baumann (1970) consider the derivative process directly. They consider a GP $f(t)$ with zero mean and autocovariance function $R_f(\tau)$, $\tau = |t_1 - t_2|$. Its derivative process $f'(t)$ is also a zero mean GP with density:

$$p(f') = \frac{1}{\sqrt{2\pi}\sigma_{f'}} \exp\left(-\frac{f'^2}{2\sigma_{f'}^2}\right), \tag{6.17}$$

where, $\sigma_f^2 = R_{f'}(0)$. The covariance function of the derivative process is related to the derivative of the original process via the relation (Middleton, 1960):

$$R_{f'}(\tau) = -\frac{d^2}{d\tau^2} R_f(\tau). \tag{6.18}$$

Given this, the expected length of $f(t)$ can be computed:

$$\mathbb{E}[s] = \int_{-\infty}^{\infty} \left[\int_0^T \mathrm{d}s\right] p(f')\mathrm{d}f', \tag{6.19}$$

$$= \frac{T}{\sigma_{f'}\sqrt{2\pi}} \int_{-\infty}^{\infty} \sqrt{1 + f'^2} \exp\left(-\frac{f'^2}{2\sigma_{f'}^2}\right) \mathrm{d}f'. \tag{6.20}$$

Using a change of variables and properties of Whittaker functions (Slater, 2010) we obtain:

$$\mathbb{E}[s] = \frac{\beta T \exp(\beta^2)}{\sqrt{2\pi}} [\mathrm{BF}_0(\beta^2) + \mathrm{BF}_1(\beta^2)], \tag{6.21}$$

where $\beta = (2\sigma_{f'})^{-1}$, and BF_0 and BF_1 are modified Bessel functions of the second kind. This approach suggests we should also be able to directly compute the arc length statistics. However, we are interested in the more general non-zero-mean case (such as would arise from most Gaussian process posterior distributions) and thus their method proves intractable, as we highlighted earlier.

6.2.3 Integrand Distribution

As mentioned earlier, we now present a new method for deriving the mean of the arc length of a one-dimensional GP by first considering the transformation of a normally distributed variable under the non-linear transformation $g(x) = (1 + x)^{1/2}$. Specifically, we consider the distribution of a normally distributed random variable under the transformation g:

$$Y = g(X) = \sqrt{1 + X^2}, \quad X \sim \mathcal{N}(\mu, \sigma^2). \tag{6.22}$$

We consider the more general case where $\mu \neq 0$. Intuitively, we expect our distribution for Y to be a skewed Chi distribution, from the form of the integrand function and inspection of Figure 6.1. We are able to directly compute the probability density function for Y by considering the cumulative distribution and using standard rules for the transformation of probability functions:

$$\begin{aligned} P(Y < y) &= P(|X - \mu| < \sqrt{y^2 - 1}), && (6.23)\\ &= P(-\sqrt{y^2 - 1} < X - \mu < \sqrt{y^2 - 1}), && (6.24)\\ &= P(-\sqrt{y^2 - 1} + \mu < X < \sqrt{y^2 - 1} + \mu), && (6.25)\\ &= F_X(\sqrt{y^2 - 1} + \mu) - (1 - F_X(\sqrt{y^2 - 1} - \mu)), && (6.26) \end{aligned}$$

where F_X is the cumulative probability distribution of X. The probability density function (pdf) of Y is obtained by taking the derivative of $P(Y < y)$ with respect to

y:

$$p_Y(y) = \frac{\mathrm{d}}{\mathrm{d}y} P(Y < y), \tag{6.27}$$

$$= \frac{\mathrm{d}}{\mathrm{d}y}\left[F_X(\sqrt{y^2-1}+\mu) - (1 - F_X(\sqrt{y^2-1}-\mu))\right], \tag{6.28}$$

$$= \frac{\mathrm{d}}{\mathrm{d}y}\left[\int_{-\infty}^{\sqrt{y^2-1}+\mu} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right) \mathrm{d}x + \int_{-\infty}^{\sqrt{y^2-1}-\mu} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2}{2\sigma^2}\right) \mathrm{d}x\right], \tag{6.29}$$

$$= \frac{1}{\sqrt{2\pi}\sigma}\left[\exp\left(-\frac{(\sqrt{y^2-1}+\mu)^2}{2\sigma^2}\right) + \exp\left(-\frac{(\sqrt{y^2-1}-\mu)^2}{2\sigma^2}\right)\right] \frac{\mathrm{d}}{\mathrm{d}y}\sqrt{y^2-1}, \tag{6.30}$$

$$= \frac{1}{\sqrt{2\pi}\sigma}\left[\exp\left(-\frac{(\sqrt{y^2-1}+\mu)^2}{2\sigma^2}\right) + \exp\left(-\frac{(\sqrt{y^2-1}-\mu)^2}{2\sigma^2}\right)\right] \frac{y}{\sqrt{y^2-1}}. \tag{6.31}$$

This probability distribution is valid for $y > 1$. We confirm this is a probability distribution by integrating Equation (6.31) over $y \in (0, \infty)$ to compute $\int_\Omega p_Y(y)\mathrm{d}y$.

We make the following substitutions:

$$y^2 - 1 = x^2, \quad y = (x^2+1)^{1/2}, \quad y\mathrm{d}y = x\mathrm{d}x, \quad y \in (1,\infty), \quad x \in (0,\infty). \tag{6.32}$$

The integral can now be computed:

$$\int_{\Omega} p_Y(y)\mathrm{d}y \tag{6.33}$$

$$= \frac{1}{\sqrt{2\pi}\sigma}\int_0^{\infty}\left[\exp\left(-\frac{(x+\mu)^2}{2\sigma^2}\right)+\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\right]\mathrm{d}x, \tag{6.34}$$

$$= \frac{1}{\sqrt{2\pi}\sigma}\left[\int_{\mu}^{\infty}\exp\left(-\frac{z_1^2}{2\sigma^2}\right)\mathrm{d}z_1+\int_{-\mu}^{\infty}\exp\left(-\frac{z_2^2}{2\sigma^2}\right)\mathrm{d}z_2\right], \quad z_1=x+\mu, z_2=x-\mu \tag{6.35}$$

$$= \frac{1}{\sqrt{2\pi}\sigma}\left[\int_{\mu}^{\infty}\exp\left(-\frac{z^2}{2\sigma^2}\right)\mathrm{d}z+\int_{-\mu}^{\infty}\exp\left(-\frac{z^2}{2\sigma^2}\right)\mathrm{d}z\right], \quad \text{Let } z_1=z, z_2=z \tag{6.36}$$

$$= \frac{1}{\sqrt{2\pi}\sigma}\left[\int_{\mu}^{\infty}\exp\left(-\frac{z^2}{2\sigma^2}\right)\mathrm{d}z+\int_{-\mu}^{0}\exp\left(-\frac{z^2}{2\sigma^2}\right)\mathrm{d}z+\int_{0}^{\infty}\exp\left(-\frac{z^2}{2\sigma^2}\right)\mathrm{d}z\right], \tag{6.37}$$

$$= \frac{1}{\sqrt{2\pi}\sigma}\left[\int_{\mu}^{\infty}\exp\left(-\frac{z^2}{2\sigma^2}\right)\mathrm{d}z+(-1)\int_{-0}^{-(-\mu)}\exp\left(-\frac{(-z)^2}{2\sigma^2}\right)\mathrm{d}(-z)\,, \right. \tag{6.38}$$

$$\left. +\int_{0}^{\infty}\exp\left(-\frac{z^2}{2\sigma^2}\right)\mathrm{d}z\right], \tag{6.39}$$

$$= \frac{1}{\sqrt{2\pi}\sigma}\left[\int_{\mu}^{\infty}\exp\left(-\frac{z^2}{2\sigma^2}\right)\mathrm{d}z+\int_{0}^{\mu}\exp\left(-\frac{z^2}{2\sigma^2}\right)\mathrm{d}z+\int_{0}^{\infty}\exp\left(-\frac{z^2}{2\sigma^2}\right)\mathrm{d}z\right], \tag{6.40}$$

$$= \frac{1}{\sqrt{2\pi}\sigma}\left[\int_{0}^{\infty}\exp\left(-\frac{z^2}{2\sigma^2}\right)\mathrm{d}z+\int_{0}^{\infty}\exp\left(-\frac{z^2}{2\sigma^2}\right)\mathrm{d}z\right], \tag{6.41}$$

$$= \frac{1}{\sqrt{2\pi}\sigma}\left[2\frac{1}{2}\sqrt{2\pi}\sigma\right], \tag{6.42}$$

$$= 1 \tag{6.43}$$

We find that $\int_{y\in Y} p_y(y)\mathrm{d}y = 1$, confirming that we have a probability distribution, with support $y > 1$.

Expectation of the integrand

Computation of the expectation of the integrand can now be done in closed form. We need to compute:

$$\mathbb{E}_{p_Y(y)}[y] \tag{6.44}$$

$$= \int_1^\infty y \; p_Y(y) \mathrm{d}y, \tag{6.45}$$

$$= \int_1^\infty y \frac{1}{\sqrt{2\pi}\sigma} \left[\exp\left(-\frac{(\sqrt{y^2-1}+\mu)^2}{2\sigma^2} \right) + \exp\left(-\frac{(\sqrt{y^2-1}-\mu)^2}{2\sigma^2} \right) \right] \frac{y}{\sqrt{y^2-1}} dy, \tag{6.46}$$

$$= \frac{1}{\sqrt{2\pi}\sigma} \int_1^\infty \left[\exp\left(-\frac{(\sqrt{y^2-1}+\mu)^2}{2\sigma^2} \right) + \exp\left(-\frac{(\sqrt{y^2-1}-\mu)^2}{2\sigma^2} \right) \right] y^2 (y^2-1)^{-\frac{1}{2}} \mathrm{d}y. \tag{6.47}$$

At first glance this looks to be an intractable integral, however, with an appropriate change of variables we will be able to derive a closed form for this expression. Consider the change of variables:

$$y^2 - 1 = x^2, \quad y = (x^2+1)^{1/2}, \quad y\mathrm{d}y = x\mathrm{d}x, \quad y \in (1,\infty), \quad x \in (0,\infty). \tag{6.48}$$

Our integral can now be written:

$$\mathbb{E}_{p_Y(y)}[y] = \frac{1}{\sqrt{2\pi}\sigma} \int_0^\infty \left[\exp\left(-\frac{(x+\mu)^2}{2\sigma^2} \right) + \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right). \right] (1+x^2)^{1/2} x^{-1} \mathrm{d}x \tag{6.49}$$

We expand the square of the exponentials and rearrange:

$$\mathbb{E}_{p_Y(y)}[y] \tag{6.50}$$

$$= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\mu^2}{2\sigma^2}\right) \int_0^\infty \exp\left(-\frac{x^2}{2\sigma^2}\right) \left[\exp\left(-\frac{x\mu}{\sigma^2}\right) + \exp\left(\frac{x\mu}{\sigma^2}\right)\right] (1+x^2)^{1/2} x^{-1} \mathrm{d}x. \tag{6.51}$$

The cross exponential terms can be expanded in their Maclaurin Series:

$$\exp\left(-\frac{x\mu}{\sigma^2}\right) + \exp\left(+\frac{x\mu}{\sigma^2}\right) = \sum_{k=0}^{\infty} \frac{(-1)^k}{k!} \left(\frac{\mu}{\sigma^2}\right)^k x^k + \sum_{k=0}^{\infty} \frac{1}{k!} \left(\frac{\mu}{\sigma^2}\right)^k x^k, \tag{6.52}$$

$$= \sum_{k=0}^{\infty} \frac{1}{(2k)!} \left(\frac{\mu}{\sigma^2}\right)^{2k} x^{2k}. \tag{6.53}$$

Substituting in this expression:

$$\mathbb{E}[y] = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\mu^2}{2\sigma^2}\right) \int_0^\infty \exp\left(-\frac{x^2}{2\sigma^2}\right) \left[\sum_{k=0}^{\infty} \frac{1}{(2k)!} \left(\frac{\mu}{\sigma^2}\right)^{2k} x^{2k}\right] (1+x^2)^{1/2} x^{-1} \mathrm{d}x, \tag{6.54}$$

$$= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\mu^2}{2\sigma^2}\right) \sum_{k=0}^{\infty} \frac{1}{(2k)!} \left(\frac{\mu}{\sigma^2}\right)^{2k} \int_0^\infty \exp\left(-\frac{x^2}{2\sigma^2}\right) x^{2k} (1+x^2)^{1/2} x^{-1} \mathrm{d}x. \tag{6.55}$$

Now we have integrals of the form:

$$\mathrm{I}_k = \int_0^\infty \exp\left(-\frac{x^2}{2\sigma^2}\right) x^{2k} (1+x^2)^{1/2} x^{-1} \mathrm{d}x. \tag{6.56}$$

Letting $z = x^2$, $\mathrm{d}z = 2x\mathrm{d}x$:

$$\mathrm{I}_k = \int_0^\infty \exp\left(-\frac{z}{2\sigma^2}\right) z^k (1+z)^{1/2} z^{-1/2} \mathrm{d}z, \tag{6.57}$$

$$= \int_0^\infty \exp\left(-\frac{z}{2\sigma^2}\right) z^{k-1/2} (1+z)^{1/2} \mathrm{d}z, \tag{6.58}$$

$$= \Gamma(k+1/2)\mathrm{U}(k+1/2, k+2, 1/2\sigma^2). \tag{6.59}$$

Here $\Gamma(n)$ is the gamma function and $\mathrm{U}(a,b,z)$ is the confluent hypergeometric function of the second kind (Slater, 2010), defined by the integral expression:

$$\mathrm{U}(a,b,z) = \frac{1}{\Gamma(a)} \int_0^\infty \exp\left(-zt\right) t^{a-1} (1+t)^{b-a-1} \mathrm{d}t. \tag{6.60}$$

The expression for I_k, Equation (6.59) holds for $k \geq 0$. Therefore the expected value is:

$$\mathbb{E}_{p_Y(y)}[y] = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\mu^2}{2\sigma^2}\right) \sum_{l=0}^{\infty} \frac{\Gamma\left(l+\frac{1}{2}\right)}{(2l)!} \left(\frac{\mu}{\sigma^2}\right)^{2l} \mathrm{U}\left(l+\frac{1}{2}, l+2, \frac{1}{2\sigma^2}\right). \tag{6.61}$$

Second moment of the integrand

A similar process will allows us to derive an exact expression for $\mathbb{E}_{p_Y(y)}[y^2]$ and hence $\mathbb{V}_{p_Y(y)}[y]$. By definition the second moment is:

$$\mathbb{E}_{p_Y(y)}[y^2] \tag{6.62}$$

$$= \int_1^\infty y^2 \frac{1}{\sqrt{2\pi}\sigma} \left[\exp\left(-\frac{(\sqrt{y^2-1}+\mu)^2}{2\sigma^2}\right) + \exp\left(-\frac{(\sqrt{y^2-1}-\mu)^2}{2\sigma^2}\right)\right] \frac{y}{\sqrt{y^2-1}} \mathrm{d}y, \tag{6.63}$$

$$= \frac{1}{\sqrt{2\pi}\sigma} \int_1^\infty \left[\exp\left(-\frac{(\sqrt{y^2-1}+\mu)^2}{2\sigma^2}\right) + \exp\left(-\frac{(\sqrt{y^2-1}-\mu)^2}{2\sigma^2}\right)\right] y^3 (y^2-1)^{-1/2} \mathrm{d}y. \tag{6.64}$$

Changing variables:

$$y^2 - 1 = x^2, \quad y = (x^2+1)^{1/2}, \quad y\mathrm{d}y = x\mathrm{d}x, \quad y \in (1,\infty), \quad x \in (0,\infty), \tag{6.65}$$

so the integral can now be written:

$$\mathbb{E}_{p_Y(y)}[y^2] = \frac{1}{\sqrt{2\pi}\sigma}\int_0^\infty \left[\exp\left(-\frac{(x+\mu)^2}{2\sigma^2}\right) + \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\right](1+x^2)x^{-1}\mathrm{d}x. \tag{6.66}$$

Re-arranging and expanding the exponential terms:

$$\mathbb{E}_{p_Y(y)}[y^2] = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\mu^2}{2\sigma^2}\right)\int_0^\infty \exp\left(-\frac{x^2}{2\sigma^2}\right)\left[\sum_{k=0}^\infty \frac{1}{(2k)!}\left(\frac{\mu}{\sigma^2}\right)^{2k} x^{2k}\right](1+x^2)x^{-1}\mathrm{d}x, \tag{6.67}$$

$$= \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\mu^2}{2\sigma^2}\right)\sum_{k=0}^\infty \frac{1}{(2k)!}\left(\frac{\mu}{\sigma^2}\right)^{2k}\int_0^\infty \exp\left(-\frac{x^2}{2\sigma^2}\right)x^{2k-1}(1+x^2)\mathrm{d}x. \tag{6.68}$$

Now we have integrals of the form:

$$\mathrm{I}_k = \int_0^\infty \exp\left(-\frac{x^2}{2\sigma^2}\right)x^{2k-1}(1+x^2)\mathrm{d}x, \tag{6.69}$$

$$= \int_0^\infty \exp\left(-\frac{z}{2\sigma^2}\right)z^{k-1/2}(1+z)\mathrm{d}z, \qquad z = x^2, \mathrm{d}z = 2x\mathrm{d}x \tag{6.70}$$

$$= \Gamma(k+1/2)\mathrm{U}(k+1/2, k+5/2, 1/2\sigma^2), \tag{6.71}$$

where we have followed the same logic as we did for the first moment. The final result is:

$$\mathbb{E}_{p_Y(y)}[y^2] = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\mu^2}{2\sigma^2}\right)\sum_{k=0}^\infty \frac{1}{(2k)!}\left(\frac{\mu}{\sigma^2}\right)^{2k}\Gamma\left(k+\frac{1}{2}\right)\mathrm{U}(k+1/2, k+5/2, 1/2\sigma^2). \tag{6.72}$$

Variance of the integrand

The variance of the integrand can be calculated as:

$$\mathrm{Var}[y] = \mathbb{E}[y^2] - \mathbb{E}[y]^2, \tag{6.73}$$

$$= \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\mu^2}{2\sigma^2}\right) \sum_{k=0}^{\infty} \frac{1}{(2k)!} \left(\frac{\mu}{\sigma^2}\right)^{2k} \Gamma(k+1/2)\mathrm{U}(k+1/2, k+5/2, 1/2\sigma^2) \tag{6.74}$$

$$- \left[\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\mu^2}{2\sigma^2}\right) \sum_{k=0}^{\infty} \frac{1}{(2k)!} \left(\frac{\mu}{\sigma^2}\right)^{2k} \Gamma(k+1/2)\mathrm{U}(k+1/2, k+2, 1/2\sigma^2)\right]^2. \tag{6.75}$$

For the zero mean case, $\mu = 0$, the expression reduces to:

$$\mathrm{Var}[y] = \mathbb{E}[y^2] - \mathbb{E}[y]^2 \tag{6.76}$$

$$= \frac{1}{\sqrt{2\pi}\sigma}\Gamma(1/2)\mathrm{U}(1/2, 5/2, 1/2\sigma^2) - \left[\frac{1}{\sqrt{2\pi}\sigma}\Gamma(1/2)\mathrm{U}(1/2, 2, 1/2\sigma^2)\right]^2. \tag{6.77}$$

Figure 6.2 shows draws of $g(X)$, overlaid with $p_Y(y)$ for a range of μ and σ^2. We observe the squashing effect of the square root and the strong dependence on the relationship between μ and σ^2 in determining the exact shape of the distribution.

6.2.4 Arc Length Statistics

Having derived expressions for the arc length integrand distribution we are now able to evaluate the moments of the arc length. We will consider a zero mean GP with kernel K and its corresponding derivative process:

$$f \sim \mathcal{GP}(0, K), \quad f' \sim \mathcal{GP}(0, \partial^2 K). \tag{6.78}$$

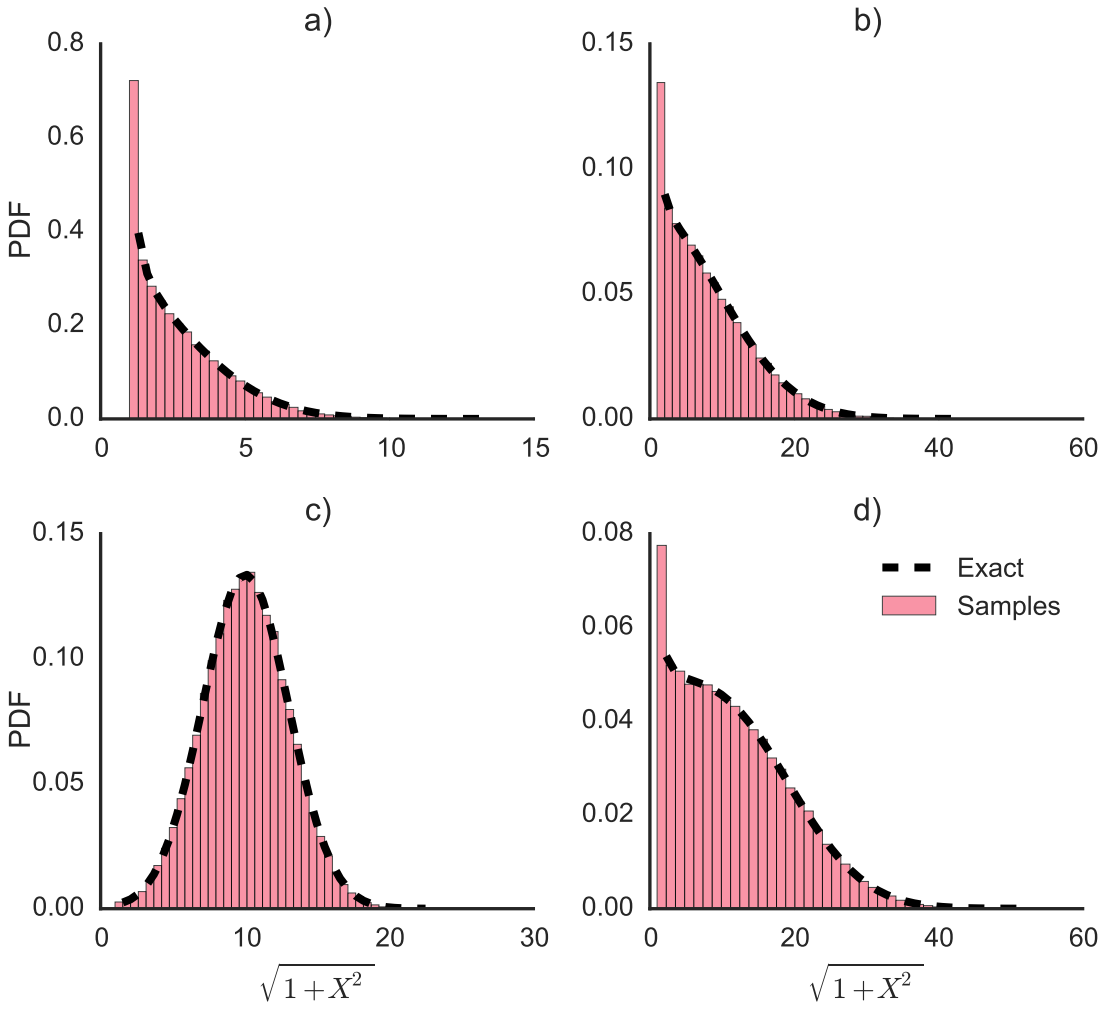

Figure 6.2: Histogram of samples from $\sqrt{1+X^2}$, where $X \sim \mathcal{N}(\mu, \sigma^2)$, overlaid with the corresponding distribution. We display the effects of varying μ and σ^2: a) $\mu = 0, \sigma = 3$, b) $\mu = 0, \sigma = 10$, c) $\mu = 10, \sigma = 3$, and d) $\mu = 10, \sigma = 10$. The fit is computed directly from Equation (6.31) and not a density estimator applied to the samples.

Taking the expectation of the arc length, noting that the integrand is non-negative, therefore by Fubini's Theorem (Fubini, 1907) we can interchange the expectation and integral:

$$\mathbb{E}[s] = \mathbb{E}\left[\int_0^T \sqrt{1 + (f')^2}\mathrm{d}t\right], \tag{6.79}$$

$$= \int_0^T \mathbb{E}\left[\sqrt{1 + (f')^2}\right]\mathrm{d}t. \tag{6.80}$$

The variance of f' is given by $\sigma^2_{f'} = R_{f'}(0)$. At each point along the integral the expectation of the integrand is the same, therefore we arrive at:

$$\mathbb{E}[s] = \int_0^T \left[\frac{1}{\sqrt{2\pi}\sigma_{f'}}\Gamma\left(\frac{1}{2}\right)\mathrm{U}\left(\frac{1}{2}, 2, \frac{1}{2\sigma^2_{f'}}\right)\right]\mathrm{d}t, \tag{6.81}$$

$$= T\frac{1}{\sqrt{2\pi}\sigma}\Gamma\left(\frac{1}{2}\right)\mathrm{U}\left(\frac{1}{2}, 2, \frac{1}{2\sigma^2_{f'}}\right), \tag{6.82}$$

where we have used the expectation of the integrand for the zero-mean case. Using identities related to the Confluent Hypergeometric function we can rewrite the mean as:

$$\mathbb{E}[s] = \frac{T\exp(1/4\sigma^2_{f'})}{2\sqrt{2\pi}\sigma_{f'}}\left[\mathrm{BF}_0\left(\frac{1}{4\sigma^2_{f'}}\right) + \mathrm{BF}_1\left(\frac{1}{4\sigma^2_{f'}}\right)\right], \tag{6.83}$$

where BF_i is the modified Bessel function of the second kind of order i. For a posterior distribution of the arc length, given data observations, we would use Equation (6.61) along with the the posterior derivative mean, $\mu_{f'} = \frac{\partial m_*}{\partial x_*}$ and variance function of the posterior GP, $\sigma^2_{f'}$ to compute the expected length:

$$\mathbb{E}[s] = \sum_{l=0}^{\infty}\frac{\Gamma\left(l + \frac{1}{2}\right)}{(2l)!}\int_0^T \frac{1}{\sqrt{2\pi}\sigma_{f'}}\exp\left(-\frac{\mu^2_{f'}}{2\sigma^2_{f'}}\right)\left(\frac{\mu_{f'}}{\sigma^2_{f'}}\right)^{2l}\mathrm{U}\left(l + \frac{1}{2}, l + 2, \frac{1}{2\sigma^2_{f'}}\right)\mathrm{d}t, \tag{6.84}$$

where $\mu_{f'}$ and $\sigma_{f'}$ depend on t. We have derived a closed form expression for the mean of the arc length of a one dimensional zero mean GP, reproducing the original result from Barakat and Baumann (1970) whilst providing a way to compute the arc length mean of a GP posterior distribution. The variance involves the computation of the second moment, a calculation involving the bi-variate form of the integrand distribution: we do not derive that in this thesis, instead we attempt to approximate the variance using a Taylor expansion. An alternate derivation is also reported in Barakat and Baumann (1970).

Kernel Derivatives

The value of the prior mean arc length, Equation (6.82), is determined solely by the derivative variance, $\sigma^2_{f'}$. For stationary kernels, $k(t, t') = k(t - t')$, this equates to:

$$\sigma^2_{f'} = \frac{\partial^2}{\partial t \partial t'} k(t - t') \bigg|_{t=t'}. \tag{6.85}$$

Table 6.1 summarises common kernels (Rasmussen and Williams, 2006) and the variance of the effective length scale in terms of their hyperparameters: derivations in the Appendix. The effect of the choice of hyperparameters on the expected length is shown in Figure 6.4.

Table 6.1: Derivative process variance, $\sigma^2_{f'}$, in terms of kernel hyperparameters for a range of common kernels. In each case λ^2 is the output (signal) variance hyperparameter and σ is the input dimension length scale hyperparameter.

Square Exponential	Matérn, $\nu = \frac{3}{2}$	Matérn, $\nu = \frac{3}{2}$	Rational Quadratic
λ^2/σ^2	$3\lambda^2/\sigma^2$	$5\lambda^2/3\sigma^2$	λ^2/σ^2

6.2.5 Variance of arc length

In order to derive the variance we need to compute the second moment:

$$\mathbb{E}[s^2] = \mathbb{E}\left[\int_0^T \int_0^T (1 + f_1'^2)^{\frac{1}{2}}(1 + f_2'^2)^{\frac{1}{2}} \mathrm{d}t_1 \mathrm{d}t_2\right] \tag{6.86}$$

$$= \int\int \mathbb{E}\left[(1 + f_1'^2)^{\frac{1}{2}}(1 + f_2'^2)^{\frac{1}{2}}\right] \mathrm{d}t_1 \mathrm{d}t_2, \tag{6.87}$$

Which presents us with some difficulties. The second moment of the arc length of a zero mean GP is derived in Barakat and Baumann (1970) in an infinite sum expansion that is difficult to work with.

We attempt to approximate the variance of the arc length. using a Taylor series expansion of the function $h(uv) = u^{\frac{1}{2}}v^{\frac{1}{2}}$ and then compute the expectation of the Taylor approximation around $h(\mu_u, \mu_v)$, where

$$\mu_v = \mathbb{E}[v] = \mathbb{E}[1 + f_1'^2] = 1 + \sigma_{f'}^2 = \mathbb{E}[1 + f_2'^2] = \mathbb{E}[u] = \mu_u, \tag{6.88}$$

using the derivatives:

$$h_u = \frac{1}{2}u^{-\frac{1}{2}}v^{\frac{1}{2}}, \quad h_{uu} = -\frac{1}{4}u^{-\frac{3}{2}}v^{\frac{1}{2}}, \quad h_{uv} = \frac{1}{4}u^{-\frac{1}{2}}v^{-\frac{1}{2}}, \tag{6.89}$$

$$h_v = \frac{1}{2}u^{\frac{1}{2}}v^{-\frac{1}{2}}, \quad h_{vv} = -\frac{1}{4}u^{\frac{1}{2}}v^{-\frac{3}{2}}, \quad h_{vu} = \frac{1}{4}u^{-\frac{1}{2}}v^{-\frac{1}{2}}. \tag{6.90}$$

$$\tag{6.91}$$

We hence obtain:

$$\mathbb{E}[h(u,v)] \tag{6.92}$$

$$\approx \mathbb{E}\left[h(\mu_u,\mu_v) + h_u(\mu_u,\mu_v)(u-\mu_u) + h_v(\mu_u,\mu_v)(v-\mu_v) + \frac{1}{2}h_{uu}(\mu_u,\mu_v)(u-\mu_u)^2\right. \tag{6.93}$$

$$\left. + \frac{1}{2}h_{vv}(\mu_u,\mu_v)(v-\mu_v)^2 + 2\frac{1}{2}h_{uv}(\mu_u,\mu_v)(u-\mu_u)(v-\mu_v)\right] \tag{6.94}$$

$$= h(\mu_u,\mu_v) + h_u(\mu_u,\mu_v)\mathbb{E}[(u-\mu_u)] + h_v(\mu_u,\mu_v)\mathbb{E}[(v-\mu_v)] + \frac{1}{2}h_{uu}(\mu_u,\mu_v)\mathbb{E}[(u-\mu_u)^2] \tag{6.95}$$

$$+ \frac{1}{2}h_{vv}(\mu_u,\mu_v)\mathbb{E}[(v-\mu_v)^2] + h_{uv}(\mu_u,\mu_v)\mathbb{E}[(u-\mu_u)(v-\mu_v)] \tag{6.96}$$

$$= h(\mu_u,\mu_v) + \frac{1}{2}h_{uu}(\mu_u,\mu_v)\mathbb{V}[u] + \frac{1}{2}h_{vv}(\mu_u,\mu_v)\mathbb{V}[v] \; + h_{uv}(\mu_u,\mu_v)\mathbb{C}[(u,v)]. \tag{6.97}$$

We can compute the variance of u and v:

$$\mathbb{V}[u] = \mathbb{V}[1 + f_1'^2] = \mathbb{V}[f_1'^2] = 2\sigma_{f'}^4, \tag{6.98}$$

$$\mathbb{V}[v] = \mathbb{V}[1 + f_2'^2] = \mathbb{V}[f_2'^2] = 2\sigma_{f'}^4. \tag{6.99}$$

The only remaining quantity we require is the covariance term:

$$\mathbb{C}[u,v] = \mathbb{C}[1 + f_1'^2, 1 + f_2'^2] = \mathbb{C}[f_1'^2, f_2'^2] = 2(\mathbb{C}[f_1', f_2'])^2, \tag{6.100}$$

where we have used the identity that for two normally distributed variables f_1' and f_2', with given covariance ρ, the covariance of the squares is twice the covariance of the variables squared. Combining the previous components gives our approximation

for the second moment as:

$$\mathbb{E}[s^2] \tag{6.101}$$

$$\approx \int_0^T \int_0^T \left(1 + \sigma_{f'}^2 - \frac{2}{4}\sigma_{f'}^4(1+\sigma_{f'}^2)^{-\frac{3}{2}}(1+\sigma_{f'}^2)^{\frac{1}{2}} + \frac{2}{4}\rho^2(1+\sigma_{f'}^2)^{-\frac{1}{2}}(1+\sigma_{f'}^2)^{-\frac{1}{2}}\right) \mathrm{d}t\mathrm{d}t', \tag{6.102}$$

$$= T^2\left(1 + \sigma_{f'}^2 - \frac{1}{2}\frac{\sigma_{f'}^4}{1+\sigma_{f'}^2}\right) + \int_0^T \int_0^T \frac{1}{2}\frac{\rho^2}{(1+\sigma_{f'}^2)}\mathrm{d}t\mathrm{d}t'. \tag{6.103}$$

6.2.6 Numerical Simulations

Here we examine the samples from the GP process and the derivative process. Consider our GP, with zero mean and squared exponential kernel. We know that:

$$\mathbb{E}[s] = \frac{T\exp(1/4\sigma_{f'}^2)}{2\sqrt{2\pi}\sigma_{f'}}\left[\mathrm{BF}_0\left(\frac{1}{4\sigma_{f'}^2}\right) + \mathrm{BF}_1\left(\frac{1}{4\sigma_{f'}^2}\right)\right], \tag{6.104}$$

where, BF_i is the modified Bessel function of the second kind. From the above we can relate $\mathbb{E}[s]$ to the parameters of the SE kernel via the relation:

$$\mathbb{E}[s] = \frac{T\exp(\sigma^2/4\lambda^2)}{2\sqrt{2\pi}\lambda/\sigma}\left[\mathrm{BF}_0\left(\frac{\sigma^2}{4\lambda^2}\right) + \mathrm{BF}_1\left(\frac{\sigma^2}{4\lambda^2}\right)\right]. \tag{6.105}$$

To confirm this theoretical result we sample from the 1D derivative kernel:

$$k_{f'}(t,t') = \frac{\lambda^2}{\sigma^2}\left(1 - (t-t')^2\sigma^{-2}\right)\exp\left(-\frac{(t-t')^2}{2\sigma^2}\right). \tag{6.106}$$

Noting the arc length formula is $s = \int_0^T \sqrt{1+f'(t)^2}\mathrm{d}t$, we use the samples f' and numerical quadrature (Simpson's rule) to compute s. We also validate the length of the sampled curves f. From a GP draw the length of f can be computed via the

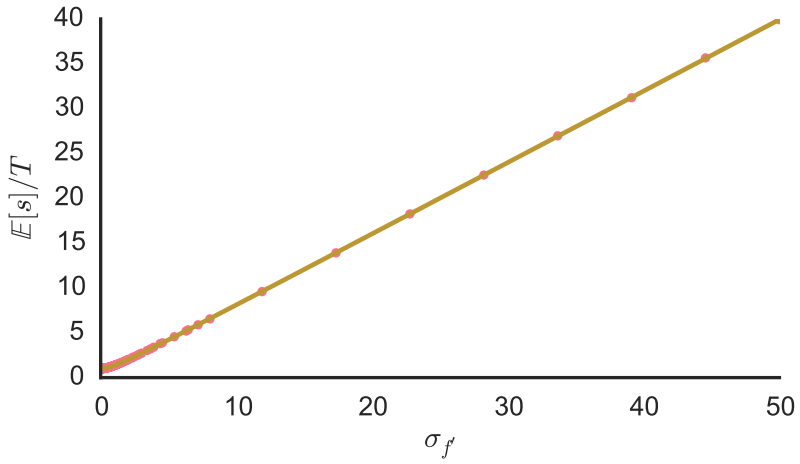

Figure 6.3: Theoretical $\mathbb{E}[s]/T$ (straight line) against lengths calculated from derivative samples (dots). x-axis is $\sigma_{f'}$, y-axis is $\mathbb{E}[s]/T$. Empirical aligns with theoretical for the SE kernel. As $\sigma_{f'}$ increases the arc length goes to infinity.

formula:

$$s_f = \sum_{i=1}^{n} |f_{i+1} - f_i|_2 = \sum_{i=1}^{n} \sqrt{(t_{i+1} - t_i)^2 + (f_{i+1} - f_i)^2}. \tag{6.107}$$

Figure 6.3 shows the theoretical and empirical results match. We see that the expected length increases with length scale; likewise the expected length increases with the interval, as is naturally expected.

Figure 6.4 shows the heat map of expected length values for a grid of kernel parameters. We are able to see the effect of kernel parameters on the expected length of the curves. The expected length is dominated by the input length scale.

Numerical experiments indicate that Equation (6.103) is a poor approximation to the variance, we are unable to achieve accurate results. This tell us that a second order Taylor approximation is insufficient, likely due to the highly non-linear terms in Equation (6.87). As such we would need to transform the joint integrand distribution to determine the arc length variance.

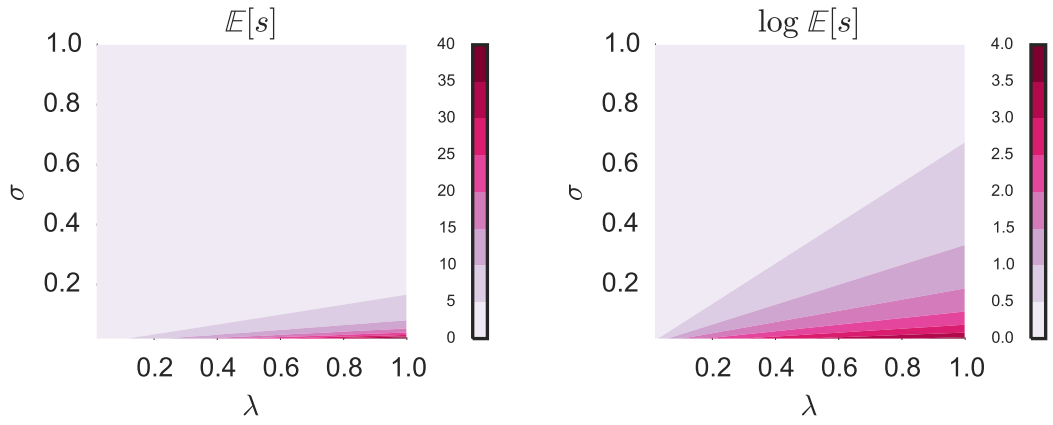

Figure 6.4: Values of the expected arc length (colour shading) for various values of the SE kernel parameters. Each plot shows the heat map of the arc length to different scale to show sufficient detail. The length is more sensitive to changes in the input scale parameter.

6.3 Vector Valued Gaussian Processes

6.3.1 Prior Vector Lengths

Until now we have been dealing with one dimensional problems. A lot of interesting problems, namely tractography, tracking, and non-stationarity problems, occur in three dimensions, $\mathbb{R}^3$: thus we pursue theoretical treatment of the three dimensional case. A vector valued function, parametrised by t, in $\mathbb{R}^3$ is written:

$$\gamma(t) = \langle x(t), y(t), z(t) \rangle \tag{6.108}$$

This traces out a curve in $\mathbb{R}^3$. A reparameterisation will give a different functional representation of γ whilst tracing the same curve. The arc length is given by the

expression:

$$s = \int_0^T |\gamma'(t)| \mathrm{d}t \tag{6.109}$$

$$= \int_0^T \sqrt{x'(t)^2 + y'(t)^2 + z'(t)^2} \mathrm{d}t \tag{6.110}$$

In this section we present the first treatment of the arc length of a GP in more than one output dimension. We present an approximation to the arc length integrand distribution and use this to compute the moments of the arc length. For the vector case, we now consider a vector GP and its corresponding derivative process:

$$\mathbf{f} \sim \mathcal{GP}(0, \mathbf{K}), \quad \mathbf{f}' \sim \mathcal{GP}(0, \partial^2 \mathbf{K}), \tag{6.111}$$

where $\mathbf{K} = \mathrm{B} \otimes \mathbf{k}$, with a coregionalised matrix B and a stationary kernel $\mathbf{k}$ (Alvarez et al., 2012). The arc length for the vector case is given by:

$$s = \int_0^T |\mathbf{f}'| \mathrm{d}t. \tag{6.112}$$

As we did in the one-dimensional case we first consider the distribution of the arc length integrand $|\mathbf{f}'|$ and then use this to derive the moments of the arc length itself.

6.3.2 Mode Linearisation

Before we proceed with the integrand distribution we again highlight the difficulty of direct computation, this time for the vector case. If we consider a vector valued function $\mathbf{f}$ that is a GP, with posterior mean and covariance for example:

$$\mathbf{f} \sim \mathcal{GP}(\boldsymbol{\mu}, \mathbf{K}) \tag{6.113}$$

with $\boldsymbol{\mu} \in \mathbb{R}^D$ in the mean vector with mean component functions $\{\mu_d(\mathbf{x})\}_{d=1}^D$ of each output and $\mathbf{K}$ is a positive-definite matrix function defined as in Chapter 3. The expected value over the posterior mean can be computed:

$$\mathbb{E}[s] = \int_\Omega \int_0^T \mathrm{d}t |\mathbf{f}'(t)| p(\mathbf{f}'|\boldsymbol{\mu}_{f'}, \boldsymbol{\Sigma}_{f'}) \mathrm{d}\mathbf{f}', \tag{6.114}$$

$$= \int_0^T \mathrm{d}t \int_\Omega \sqrt{\mathbf{f}'^{\mathbf{T}}\mathbf{f}'} \mathcal{N}(\mathbf{f}'|\boldsymbol{\mu}_{f'}, \boldsymbol{\Sigma}_{f'}) \mathrm{d}\mathbf{f}' \tag{6.115}$$

where the integral over $\mathbf{f}'$ is taken over its support Ω. This is a formidable looking integral but we can approximate it around the mode of our distribution. Expanding the non-linear term $\mathbf{g}(f') = \sqrt{\mathbf{f}'^{\mathbf{T}}\mathbf{f}'}$ around the mode, $\mu_{f'}$ and neglecting higher order terms:

$$\mathbf{g}(f') \approx \mathbf{g}(\boldsymbol{\mu}_{f'}) + D_{\mathbf{f}'}\mathbf{g}|_{\mathbf{f}'=\mu_{f'}}(\mathbf{f}' - \boldsymbol{\mu}_{f'}), \tag{6.116}$$

$$= \sqrt{\boldsymbol{\mu}_{f'}^T \boldsymbol{\mu}_{f'}} + D_f(\mathbf{f}' - \boldsymbol{\mu}_{f'}), \tag{6.117}$$

where $D_f = D_{\mathbf{f}'}\mathbf{g}|_{\mathbf{f}'=\boldsymbol{\mu}_{f'}}$ is constant with respect to f'. Substituting in our linearised term gives:

$$\mathbb{E}[s] = \int_0^T \mathrm{d}t \int_\Omega \left[\sqrt{\boldsymbol{\mu}_{f'}^T \boldsymbol{\mu}_{f'}} + D_f(\mathbf{f}' - \mu_{f'})\right] p(\mathbf{f}'|\boldsymbol{\mu}_{f'}, \boldsymbol{\Sigma}_{f'}) \mathrm{d}\mathbf{f}', \tag{6.118}$$

$$= \int_0^T \mathrm{d}t \left[\sqrt{\boldsymbol{\mu}_{f'}^T \boldsymbol{\mu}_{f'}} + D_f(\boldsymbol{\mu}_{f'} - \boldsymbol{\mu}_{f'})\right], \tag{6.119}$$

$$= \int_0^T \mathrm{d}t |\boldsymbol{\mu}_{f'}(t)|. \tag{6.120}$$

Thus we see that the approximate expected length of the posterior is the length of the posterior mean, as we would intuitively expect and as was shown in the one-dimensional case. As in the one dimensional case, this approach yields no additional information about the nature of the arc length distribution, supporting the case for considering the integrand distribution.

6.3.3 Integrand Distribution

We use the approach outlined for the one dimensional case and thus we are interested in the distribution over the arc length integrand. Ultimately we are interested in $\mathbb{R}^3$, however, the theory we present is valid for any $\mathbb{R}^n$. We consider the random variable W, defined by:

$$\mathrm{W} = |\mathbf{x}| = (\mathbf{x}^T\mathbf{x})^{1/2} = \sqrt{\sum_i^n \mathbf{x}_i^2}, \tag{6.121}$$

$$\mathbf{x} \sim \mathcal{N}(\mu, \Sigma), \tag{6.122}$$

with $\mathbf{x}, \mu \in \mathbb{R}^n$ and $\Sigma \in \mathbb{R}^{n\times n}$ is a full-rank covariance matrix. W is the square root of the sum of squares of correlated normal variables. It is well known that the sum of squares of independent identically distributed normal variables is Chi-squared distributed and that the corresponding square root is Chi distributed (Johnson et al., 1994). At first glance it seems that we should easily be able to identify this transformed distribution, however, the full-covariance between the elements of $\mathbf{x}$ hinder the derivation of a straightforward distribution.

Substantial work has been done on the distribution of quadratic forms, $Q(\mathbf{x}) = \mathbf{x}^T A\mathbf{x}$ (Mathai and Provost, 1992), where $\mathbf{x}$ is an $n \times 1$ normal vector defined previously and A is a symmetric $n \times n$ matrix. It is possible to write:

$$Q(\mathbf{x}) = \mathbf{x}^T A\mathbf{x} = \sum_i^n \lambda_i (U_i + b_i)^2, \tag{6.123}$$

where the U_i are i.i.d. normal variables with zero mean and unit variance, the λ_i are the eigenvalues of Σ and b_i is the ith component of $b = P^T\Sigma^{\frac{1}{2}}\mu$, with P a matrix that diagonalises $\Sigma^{\frac{1}{2}}A\Sigma^{\frac{1}{2}}$.

Observing the summation of the quadratic form in Equation (6.123), we see that our distribution is a weighted sum of Chi-squared variables. Unfortunately, there

exists no simple closed-form solution for this distribution. However, it is possible to express this distribution via a power-series of Laguerre polynomials and some approximations have been used (Mathai and Provost, 1992).

We note that a Chi-squared distribution is a gamma distributed variable for the case where the shape parameter is $v/2$ and the scale factor is 2. Therefore we will approximate $Q(\mathbf{x}) = \mathbf{x}^T\mathbf{x}$ with a single gamma random variable by moment matching the first two moments. The mean and variance of $Q(\mathbf{x})$ are given by:

$$\mathbb{E}[Q(\mathbf{x})] = \mathrm{tr}(\Sigma) + \mu^T\mu, \tag{6.124}$$

$$\mathbb{V}[Q(\mathbf{x})] = 2\mathrm{tr}(\Sigma\Sigma) + 4\mu^T\Sigma\mu, \tag{6.125}$$

where tr() denotes that trace of a matrix. Derivation of the mean follows:

$$\begin{aligned}
\mathbb{E}[Q(\mathbf{x})] &= \mathbb{E}[\mathbf{x}^T\mathbf{x}], && (6.126)\\
&= \mathbb{E}[\mathrm{tr}(\mathbf{x}^T\mathbf{x})], && (6.127)\\
&= \mathbb{E}[\mathrm{tr}(\mathbf{x}\mathbf{x}^T)], && (6.128)\\
&= \mathrm{tr}(\mathbb{E}[\mathbf{x}\mathbf{x}^T]), && (6.129)\\
&= \mathrm{tr}(\Sigma + \mu\mu^T) && (6.130)\\
&= \mathrm{tr}(\Sigma) + \mu^T\mu. && (6.131)
\end{aligned}$$

The variance is likewise computed. The pdf of a gamma distribution with shape k_G and scale θ_G is given by:

$$p_G(x : k_G, \theta_G) = \frac{x^{k_G-1}\exp\left(-\frac{x}{\theta_G}\right)}{\theta_G^{k_G}\Gamma(k_G)}. \tag{6.132}$$

The first two moments are:

$$\mu_G = k_G\theta_G, \quad \sigma_G^2 = k_G\theta_G^2. \tag{6.133}$$

Solving for k_G and θ_G:

$$k_G = \frac{\mu_G^2}{\sigma_G^2}, \quad \theta_G = \frac{\sigma_G^2}{\mu_G}. \tag{6.134}$$

Equating moments, we set $\mu_G = \mathbb{E}[Q(\mathbf{x})]$ and $\sigma_G^2 = \mathbb{V}[Q(\mathbf{x})]$. Thus, Q is approximated as a gamma random variable and we write, $Q(\mathbf{x}) \sim \text{Gamma}(k_G, \theta_G)$.

Now we are in a position to consider the quantity $\sqrt{Q}$. Here we use that fact that if a random variable $Q \sim \text{Gamma}(k_G, \theta_G)$, then the random variable $\text{W} = \sqrt{Q}$ is a Nakagami random variable $\text{W} \sim \text{Nakagami}(m, \Omega)$, with parameters given by $m = k_G$ and $\Omega = k_G\theta_G$. The Nakagami distribution (Hoffman, 1958) is:

$$p_{\text{Nak}}(x; m, \theta) = \frac{2m^m}{\Gamma(m)\Omega^m} x^{2m-1} \exp\left(-\frac{m}{\Omega}x^2\right). \tag{6.135}$$

Using the value for k and θ obtained via our moment matched approximation and transforming to the Nakagami distribution we say $\sqrt{Q}$ is approximated as a Nakagami distribution with parameters:

$$m = \frac{\mu_G^2}{\sigma_G^2}, \quad \Omega = \mu_G. \tag{6.136}$$

In terms of our original distribution $\mathbf{x} \sim \mathcal{N}(\mu, \Sigma)$, we therefore have $\text{W} = \sqrt{\mathbf{x}^T\mathbf{x}} \sim \text{Nakagami}(m, \Omega)$, with:

$$m = \frac{[\text{tr}(\Sigma) + \mu^T\mu]^2}{2\text{tr}(\Sigma\Sigma) + 4\mu^T\Sigma\mu}, \quad \Omega = \text{tr}(\Sigma) + \mu^T\mu. \tag{6.137}$$

The mean and variance of the Nakagami distribution, provided in the Appendix, are:

$$\mathbb{E}[\mathrm{W}] = \frac{\Gamma(m+\frac{1}{2})}{\Gamma(m)}\left(\frac{\Omega}{m}\right)^{\frac{1}{2}}, \tag{6.138}$$

$$\mathbb{V}[\mathrm{W}] = \Omega\left(1 - \frac{1}{m}\left(\frac{\Gamma(m+\frac{1}{2})}{\Gamma(m)}\right)^2\right). \tag{6.139}$$

The method we have used to derive the distribution of the arc length integrand is summarised as:

$$\mathcal{N}(\mu,\Sigma) \underset{\text{Approximate}}{\overset{Q}{\rightarrow}} \text{Gamma}(k_G,\theta_G) \overset{\sqrt{Q}}{\underset{\text{Exact}}{\rightarrow}} \text{Nakagami}(m,\Omega). \tag{6.140}$$

Numerical samples of $Q(\mathbf{x})$ and $\sqrt{Q(\mathbf{x})}$ and the pdf of the corresponding gamma and Nakagami distributions are show in Figure 6.5 for $d = 3$. The approximated distributions show a reasonable approximation for a range of μ and Σ.

The quadratic form approximated to the gamma distribution is exact when all the eigenvalues of the covariance are identical, in that case we have only a single gamma random variable. This approximation is known to work well in the literature (Covo and Elalouf, 2014). We were unable to generate a synthetic covariance with given eigenvalues that performed poorly under the given approximation: this provides empirical support for the accuracy of the approximation.

6.3.4 Vector Arc Length Statistics

We are now in a position to consider the arc length directly. Taking the expectation of the arc length, recalling that expectation is a linear operator and using Fubini's

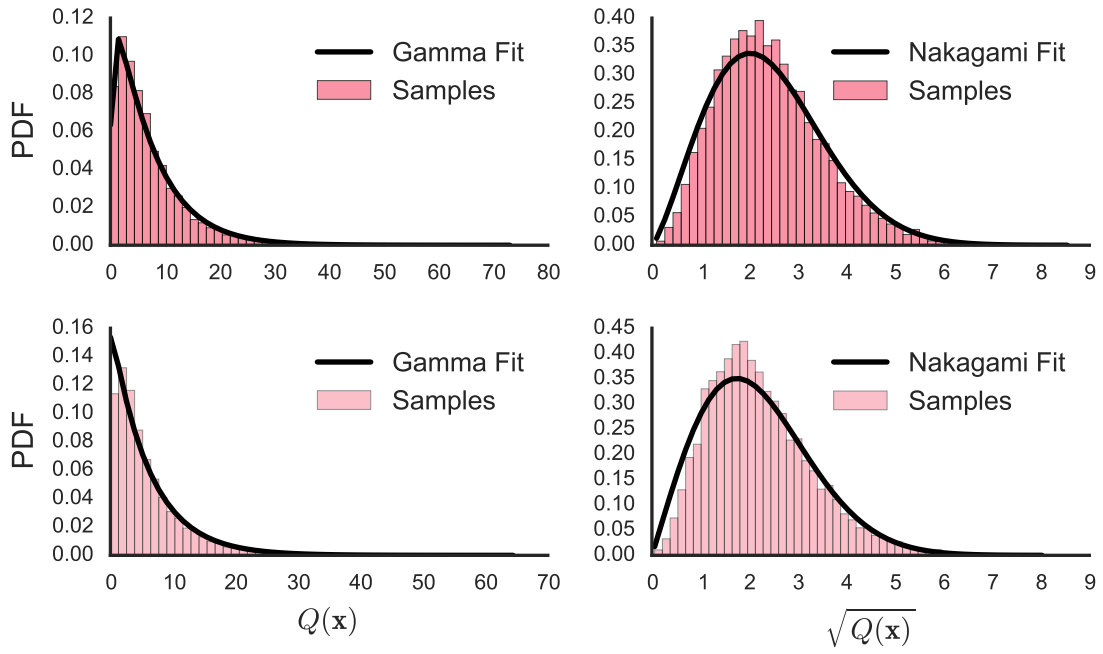

Figure 6.5: Samples from $Q(\mathbf{x})$ and $\sqrt{Q(\mathbf{x})}$ overlaid with the approximated gamma and Nakagami distributions. $\mathbf{x} \sim \mathcal{N}(0, \Sigma)$ in the top row, and $\mathbf{x} \sim \mathcal{N}(\mu, \Sigma)$ in the bottom row with μ and Σ randomly generated. Similar plots are obtained for different values of μ and Σ. The gamma and Nakagami distributions provide a reasonable approximation to the shape of the distribution, whilst capturing the true mean and variance. The fit is computed directly from the Nakagami and gamma distributions given their parameters in Equations 6.134 and 6.137, and is not a density estimator applied to the samples.

theroem:

$$\mathbb{E}[s] = \mathbb{E}\left[\int_0^T |\mathbf{f}'| \mathrm{d}t\right], \tag{6.141}$$

$$= \int_0^T \mathbb{E}\left[(\mathbf{f}'^T \mathbf{f}')^{\frac{1}{2}}\right] \mathrm{d}t. \tag{6.142}$$

Recalling the form of our kernel as $\mathbf{K}(t,t') = \mathrm{B} \otimes k(t,t')$, the integrand distribution of $\mathbf{f}'$ is constant with respect to t with covariance given by:

$$\Sigma_{f'} = \mathrm{B} \otimes \frac{\partial^2}{\partial t \partial t'} k(t,t')\bigg|_{t=t'} = \mathrm{B}\ \sigma_{f'}^2. \tag{6.143}$$

The coregionalised matrix B is constant in space: as we recall it captures the correlation between outputs, not the input dependence. Thus we note the instantaneous distribution of the integrand has $\Sigma_{f'} = \mathrm{B} \times \sigma_{f'}^2$. Therefore $\mathrm{tr}(\Sigma_{f'}) = \mathrm{tr}[\mathrm{B} \times \sigma_{f'}^2]$. The approximated expected arc length is thus:

$$\mathbb{E}[s] \approx T \frac{\Gamma(m_{f'} + \frac{1}{2})}{\Gamma(m_{f'})} \left(\frac{\Omega_{f'}}{m_{f'}}\right)^{\frac{1}{2}}, \tag{6.144}$$

with:

$$m_{f'} = \frac{[\mathrm{tr}(\Sigma_{f'})]^2}{2\mathrm{tr}(\Sigma_{f'}\Sigma_{f'})}, \quad \Omega_{f'} = \mathrm{tr}(\Sigma_{f'}), \tag{6.145}$$

where we have used the Nakagami approximation to the arc length integrand to evaluate the mean. As in the one-dimensional case, the expected length of the GP is determined solely by the choice of kernel and the length of the interval. The

calculation of the variance is somewhat more involved:

$$\mathbb{V}[s] = \mathbb{V}\left[\int_0^T |\mathbf{f}'_t| \mathrm{d}t\right], \tag{6.146}$$

$$= \int_0^T \mathbb{V}\left[|\mathbf{f}'_t|\right] \mathrm{d}t + 2\int_0^T \int_0^T \mathrm{Cov}[|\mathbf{f}'_{t_1}|, |\mathbf{f}'_{t_2}|] \mathrm{d}t_1 \mathrm{d}t_2. \tag{6.147}$$

The first term can be evaluated following similar reasoning to the mean:

$$\int_0^T \mathbb{V}\left[|\mathbf{f}'_t|\right] \mathrm{d}t = T\Omega_{f'}\left(1 - \frac{1}{m_{f'}}\left(\frac{\Gamma(m_{f'} + \frac{1}{2})}{\Gamma(m_{f'})}\right)^2\right), \tag{6.148}$$

with $m_{f'}$ and $\Omega_{f'}$ as defined previously. Though we could attempt to proceed with this approach, we note that it is as straightforward to directly compute the variance using the second moment of s:

$$\mathbb{E}[s^2] = \mathbb{E}\left[\int_0^T \int_0^T |\mathbf{f}'_{t_1}| \ |\mathbf{f}'_{t_2}| \mathrm{d}t_1 \mathrm{d}t_2\right], \tag{6.149}$$

$$= \int_0^T \int_0^T \mathbb{E}\left[|\mathbf{f}'_{t_1}| \ |\mathbf{f}'_{t_2}|\right] \mathrm{d}t_1 \mathrm{d}t_2. \tag{6.150}$$

Making use of the Nakagami approximation to our integrand we need the mixed moment of two correlated Nakagami variables. Let us write $|\mathbf{f}'_{t_1}| \approx \mathrm{W}_1$, $|\mathbf{f}'_{t_2}| \approx \mathrm{W}_2$, with $\mathrm{W}_1 \sim \text{Nakagami}(m_{f'}, \Omega_{f'})$ and $\mathrm{W}_2 \sim \text{Nakagami}(m_{f'}, \Omega_{f'})$. The mixed moments of two correlated Nakagami variables with the same parameters is given by Reig et al. (2002):

$$\mathbb{E}[\mathrm{W}_1^n \mathrm{W}_2^l] = \left(\frac{\Omega_{f'}}{m_{f'}}\right)^{\frac{n}{2}} \left(\frac{\Omega_{f'}}{m_{f'}}\right)^{\frac{l}{2}} \frac{\Gamma(m_{f'} + n/2)\Gamma(m_{f'} + l/2)}{[\Gamma(m_{f'})]^2} {}_2F_1\left(-\frac{n}{2}, -\frac{l}{2}, m_{f'} : \rho(\tau)\right), \tag{6.151}$$

where $\rho(\tau)$ is the correlation between the gamma variables that the Nakagami

distribution was derived from and ${}_2F_1(a, b, c : z)$ is the hypergeometric function:

$$
\begin{aligned}
{}_2F_1(a, b, c : z) &= \sum_{n=0}^{\infty} \frac{(a)_n (b)_n}{(c)_n} \frac{z^n}{n!}, && (6.152)\\
&= 1 + \frac{ab}{1.c} z + \frac{a(a+1)b(b+1)}{1 \cdot 2 \cdot c(c+1)} z^2 + \ldots, && (6.153)
\end{aligned}
$$

with the Pochhammer $(q)_n$ symbol defined as $(q)_n = q(q+1)\ldots(q+n-1)$ and $(q)_0 = 1$. The second moment can now be expressed as a power series in ρ:

$$
\begin{aligned}
\mathbb{E}[s^2] &= \int_0^T \int_0^T \mathbb{E}[\mathrm{W}_1 \mathrm{W}_2] \mathrm{d}t_1 \mathrm{d}t_2, && (6.154)\\
&= \int_0^T \int_0^T \frac{\Omega_{f'}}{m_{f'}} \frac{[\Gamma(m_{f'} + 1/2)]^2}{[\Gamma(m_{f'})]^2} {}_2F_1\left(-\frac{1}{2}, -\frac{1}{2}, m_{f'} : \rho(|t_1 - t_2|)\right) \mathrm{d}t_1 \mathrm{d}t_2, && (6.155)\\
&= \frac{\Omega_{f'}}{m_{f'}} \frac{[\Gamma(m_{f'} + 1/2)]^2}{[\Gamma(m_{f'})]^2} \sum_{n=0}^{\infty} \frac{\left(-\frac{1}{2}\right)_n \left(-\frac{1}{2}\right)_n}{(m_{f'})_n} \frac{1}{n!} \int_0^T \int_0^T \rho(t_1 - t_2)^n \mathrm{d}t_1 \mathrm{d}t_2. && (6.156)
\end{aligned}
$$

We now derive the correlation function. Consider two gamma variables $\mathcal{G}_x = X^T X$ and $\mathcal{G}_y = Y^T Y$, with $X, Y \sim \mathcal{N}(0, \Sigma)$, both drawn from the same multivariate normal. Now we compute the covariance directly using the rules of expectation and in

particular Isserlis' Theorem (Isserlis, 1918):

$$\mathbb{C}[\mathcal{G}_x, \mathcal{G}_y] = \mathbb{C}[X^T X, Y^T Y], \tag{6.157}$$

$$= \mathbb{E}[X^T X Y^T Y] - \mathbb{E}[X^T X]\mathbb{E}[Y^T Y], \tag{6.158}$$

$$= \mathbb{E}\left[\sum_i^n \sum_j^n x_i^2 y_j^2\right] - \mathrm{tr}(\Sigma)\mathrm{tr}(\Sigma), \tag{6.159}$$

$$= \sum_i^n \sum_j^n \mathbb{E}\left[x_i^2 y_j^2\right] - \mathrm{tr}(\Sigma)^2, \tag{6.160}$$

$$= \sum_i^n \sum_j^n \left(\mathbb{E}\left[x_i x_i\right]\mathbb{E}\left[y_j y_j\right] + \mathbb{E}\left[x_i y_j\right]\mathbb{E}\left[x_i y_j\right] + \mathbb{E}\left[y_j x_i\right]\mathbb{E}\left[x_i y_j\right]\right) - \mathrm{tr}(\Sigma)^2, \tag{6.161}$$

$$= \sum_i^n \mathbb{E}\left[x_i x_i\right] \sum_j^n \mathbb{E}\left[y_j y_j\right] + 2\sum_i^n \sum_j^n \mathbb{E}\left[x_i y_j\right]\mathbb{E}\left[x_i y_j\right] - \mathrm{tr}(\Sigma)^2, \tag{6.162}$$

$$= \mathrm{tr}(\Sigma)\mathrm{tr}(\Sigma) + 2\mathrm{tr}(\Sigma_{xy}\Sigma_{xy}) - \mathrm{tr}(\Sigma)^2, \tag{6.163}$$

$$= 2\mathrm{tr}(\Sigma_{xy}\Sigma_{xy}), \tag{6.164}$$

where:

$$\Sigma_{xy} = \mathrm{B} \otimes k_{f'}(x, y) = \mathrm{B} \times k_{f'}(x, y). \tag{6.165}$$

The correlation function is thus:

$$\rho(|x - y|) = \frac{2\mathrm{tr}(\Sigma_{xy}\Sigma_{xy})}{2\mathrm{tr}(\Sigma_{xx}\Sigma_{xx})}, \tag{6.166}$$

$$= \frac{\mathrm{tr}(\mathrm{B} \times k_{f'}(x, y) \times \mathrm{B} \times k_{f'}(x, y))}{\mathrm{tr}(\mathrm{B} \times k_{f'}(x, x) \times \mathrm{B} \times k_{f'}(x, x))}, \tag{6.167}$$

$$= \frac{\mathrm{tr}(\mathrm{BB})k_{f'}(x, y)^2}{\mathrm{tr}(\mathrm{BB})k_{f'}(x, x)^2}, \tag{6.168}$$

$$= \frac{k_{f'}(x, y)^2}{k_{f'}(0)^2}, \tag{6.169}$$

where we write $k_{f'}(0) = k_{f'}(|x - x|) = k_{f'}(x, x)$. Interestingly we see that the corre-

lation between outputs depends only on the choice of spatial covariance function:

$$\rho(t-t') = \left[\frac{\partial^2}{\partial t \partial t'} k(t,t')\right]^2 \frac{1}{\sigma_{f'}^4}. \tag{6.170}$$

Equation (6.156) can be solved numerically (noting that the two-dimensional integral is readily tackled using traditional methods of quadrature) and the variance is then computed by $\mathbb{V}[s] = \mathbb{E}[s^2] - \mathbb{E}[s]^2$.

6.3.5 Numerical Simulations

In this section we generate samples from our GP prior and compute the arc length, for the vector case. We show the effect of the kernel choice and show the fidelity of our theoretical results. To generate our curves we specify a zero mean GP kernel, $\mathrm{K} = \mathrm{B} \otimes k(t,t')$, with fixed B and we use the Matérn Kernel with $\nu = 3/2$:

$$k_{\text{MAT32}}(t,t') = \lambda^2 \left(1 + \frac{\sqrt{3}|t-t'|}{\sigma}\right) \exp\left(-\frac{\sqrt{3}|t-t'|}{\sigma}\right). \tag{6.171}$$

We draw a sample $f_i = (x_i, y_i, z_i)$ evaluated at evenly spaced t. The arc length of the GP draw is then computed numerically:

$$s_i = \sum_{j=1}^{n} \sqrt{(x_i - x_{i-1})^2 + (y_i - y_{i-1})^2 + (z_i - z_{i-1})^2}. \tag{6.172}$$

Unit variance and length scale parameters are chosen and the arc length is computed over the interval $t = [0,1]$. Figure 6.6 shows the sample lengths, the theoretical mean and variance, and the Nakagami distribution of a single arc length integrand. Figure 6.8 gives a heat map of the expected length. Our theoretical results are close to the numerically generated values. Figure 6.7 shows the arc length distribution for range of kernel functions and hyperparameter values. No estimation methods are required to calculate the arc length statistics. Our approximated equations are closed

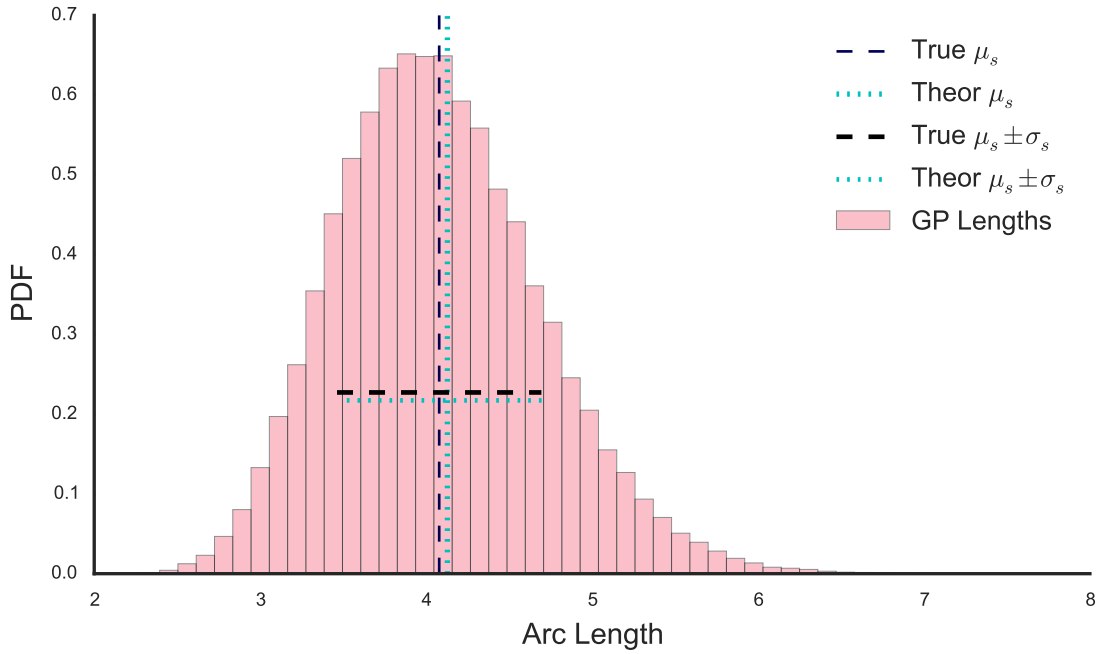

Figure 6.6: Histogram of GP lengths. The theoretical and empirical mean ($\mathbb{E}[s] = \mu_s$) are shown and one standard deviation $\mathbb{V}[s] = \sigma_s^2$. The true mean and variance are well approximated as evidenced in the figure.

form (for the mean) and a quadrature problem (for the variance). A Monte Carlo method is not required to compute these values.

To compute a Monte Carlo estimate of the arc length statistics we would need to specify the interval we wish to compute our arc length over and then draw sufficient GP samples (1000s) to get a good estimate. For an interval of another length we would need to repeat this entire process. This work enables one to compute these statistics with a single computation (for the mean) and a numerical double integration for the variance.

Importantly, if we wish to compute the statistics for a different length interval we are able to either reuse our computations (for a longer interval) or quickly compute it for a shorter interval. If we were to use a Monte Carlo method, we would need to generate new samples for a new interval; a process that would be slower than using the equations derived.

6.3.6 Effects of Kernel Parameters

Let's analyse the effects on the arc length statistics in terms of the kernel and kernel parameters. We can write:

$$\Sigma_{f'} = \sigma_{f'}^2 \mathrm{B}, \tag{6.173}$$

where we are using the notation $\sigma_{f'}^2 = \partial^2 k(\tau)|_{\tau=0}$. We can now write our Nakagami parameters:

$$\Omega_f = \mathrm{tr}(\Sigma_f) = \sigma_{f'}^2 \mathrm{tr}(\mathrm{B}), \tag{6.174}$$

$$m_f = \frac{\mathrm{tr}(\Sigma_f)^2}{2\mathrm{tr}(\Sigma_f \Sigma_f)} = \frac{\sigma_{f'}^4 \mathrm{tr}(\mathrm{B})^2}{2\sigma_{f'}^2 \sigma_{f'}^2 \mathrm{tr}(\mathrm{BB})} = \frac{\mathrm{tr}(\mathrm{B})^2}{2\mathrm{tr}(\mathrm{BB})}. \tag{6.175}$$

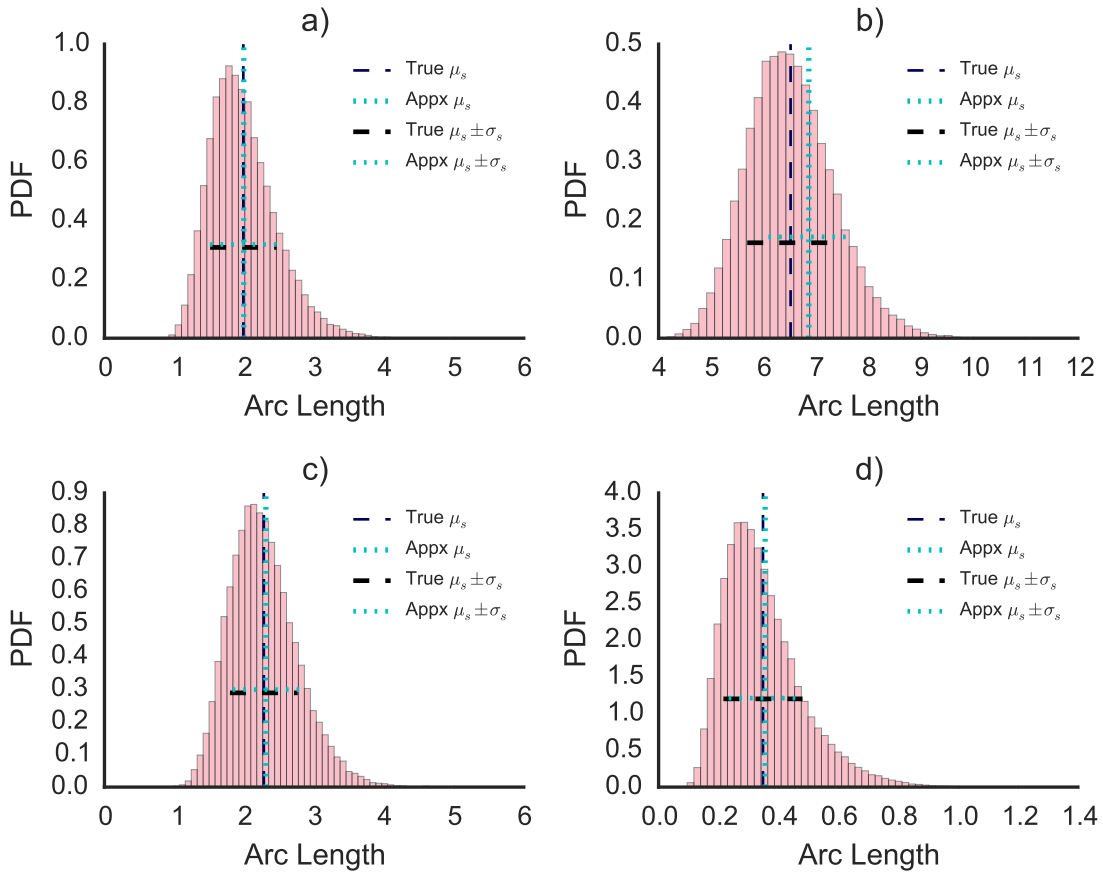

Figure 6.7: Histogram of GP lengths as in Figure6.6. For the Matérn32 kernel over a range of hyperparameters: a) $\lambda^2 = 0.5, \sigma = 1.0$: b) $\lambda^2 = 0.5, \sigma = 0.25$ c) $\lambda^2 = 1.0, \sigma = 1.0$ d) $\lambda^2 = 0.5, \sigma = 4.0$. The lengths are all computed over the unit interval except for c) where $t \in [0, 2]$

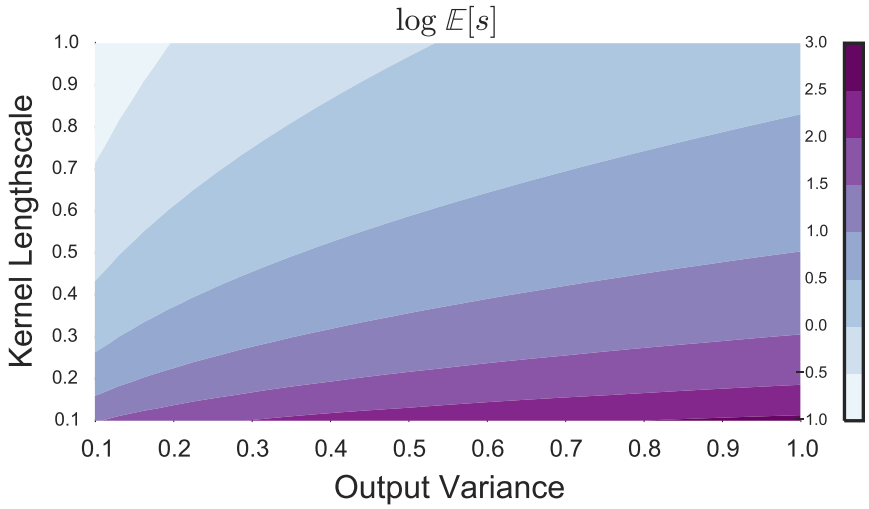

Figure 6.8: Heat map for the log expected length of a vector valued GP for the SE kernel. Again we see clear dependence on the kernel hyperparameters.

We observe that only Ω_f depends on the variance of the derivative kernel. Therefore we can write:

$$\mathbb{E}[s] \approx T\frac{\Gamma(m_{f'}+\frac{1}{2})}{\Gamma(m_{f'})}\left(\frac{\Omega_{f'}}{m_{f'}}\right)^{\frac{1}{2}}, \tag{6.176}$$

$$= T\frac{\Gamma(\frac{\mathrm{tr(B)}^2}{2\mathrm{tr(BB)}}+\frac{1}{2})}{\Gamma(\frac{\mathrm{tr(B)}^2}{2\mathrm{tr(BB)}})}\left(\frac{\sigma_{f'}^2\mathrm{tr(B)}}{\frac{\mathrm{tr(B)}^2}{2\mathrm{tr(BB)}}}\right)^{\frac{1}{2}}, \tag{6.177}$$

$$= \sigma_{f'}T\frac{\Gamma(\mathrm{tr(B)}^2/(2\mathrm{tr(BB)})+\frac{1}{2})}{\Gamma(\mathrm{tr(B)}^2/(2\mathrm{tr(BB)}))}\left(\frac{2\mathrm{tr(BB)}}{\mathrm{tr(B)}}\right)^{\frac{1}{2}}. \tag{6.178}$$

We examine some cases for B.

B = I: For the simplest case the trace simplifies $\mathrm{tr(I)} = 3$. Our parameters are thus:

$$\Omega = 3\sigma_{f'}^2, \quad m = \frac{3^2}{2.3} = \frac{3}{2}. \tag{6.179}$$

The expected length now becomes:

$$\mathbb{E}[s] \approx T\frac{\Gamma(3/2+1/2)}{\Gamma(3/2)}\left(\frac{3\sigma_{f'}^2}{3/2}\right)^{\frac{1}{2}}, \tag{6.180}$$

$$= \sigma_{f'}T\frac{1}{\sqrt{\pi}/2}\sqrt{2}, \tag{6.181}$$

$$= \frac{2\sqrt{2}\sigma_{f'}T}{\sqrt{\pi}}. \tag{6.182}$$

General B**:** A general B will have three positive eigenvalues (in the 3D case). This means we can write:

$$\text{tr(B)} = \lambda_1 + \lambda_2 + \lambda_3 = \sum_i^3 \lambda_i, \quad \text{tr(BB)} = \sum_i^3 \lambda_i^2, \tag{6.183}$$

$$\Omega = \sigma_{f'}^2 \sum \lambda_i, \quad m = \frac{1}{2}\frac{(\sum \lambda_i)^2}{\sum \lambda_i^2}. \tag{6.184}$$

Entering these formulas into the expected length provides no straightforward simplifications:

$$\mathbb{E}[s] \approx \sigma_{f'}T\frac{\Gamma((\sum_i^3 \lambda_i)^2/(2\sum_i^3 \lambda_i^2)+\frac{1}{2})}{\Gamma((\sum_i^3 \lambda_i)^2/(2\sum_i^3 \lambda_i^2))}\left(\frac{2\sum_i^3 \lambda_i^2}{\sum_i^3 \lambda_i}\right)^{\frac{1}{2}}. \tag{6.185}$$

We now have an explicit formula for the expected length in terms of the hyperparameters of the chosen kernel, $\sigma_{f'}$ and the eigenvalues of our co-regionalising matrix.

We see that the arc length is directly a function of the hyperparameters via, $\sigma_{f'}$ and the co-regionalising matrix. With careful inspection of the eigenvalues of B we should be able to bound the values of Ω and m and subsequently the expected length. This suggests that it would be possible to construct our GP kernel such that we generate curves of specific lengths.

6.4 Arc Length of the Posterior

Our discussion till this point has focussed on the arc length for a GP prior, with focus on a zero mean prior. Given a GP with a specified kernel and hyperparameters we know how lengths of GP draws from the prior will be distributed. We now turn our attention to the arc length of a GP posterior, which we shall call the arc length posterior. In many applications we are often interested in conditioning our GP on observations. Our GP will now be conditioned on observations and we expect this to affect the statistics of our arc length.

As opposed to having the samples dependent only on the covariance structure we now have the added information from our data points. In this case we would expect our function to be 'anchored' around these observations. In this way the statistics of the arc length now become a function of both the kernel and data. Naturally the kernel choice and resulting parameters are a result of the data as well, but the data points themselves are involved in arc length computations. The moments of the arc length of a GP posterior follow a similar derivation to the moments of the GP prior.

6.4.1 One Dimensional Posterior

The posterior distribution was sketched earlier, here we make it explicit. We have established that the mean of the integrand for a non-zero mean, can be written as:

$$\begin{aligned}&\mathbb{E}[(1+f'^2)^{1/2}]\\&= \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\mu_{f'}^2}{2\sigma_{f'}^2}\right)\sum_{k=0}^{\infty}\frac{1}{(2k)!}\left(\frac{\mu_{f'}}{\sigma_{f'}^2}\right)^{2k}\Gamma(k+1/2)\mathrm{U}(k+1/2,k+2,1/2\sigma_{f'}^2).\end{aligned} \tag{6.186}$$

Therefore the mean of the posterior arc length can be computed as:

$$\mathbb{E}[s] = \sum_{k=0}^{\infty} \frac{1}{(2k)!} \int_0^T \frac{1}{\sqrt{2\pi}\sigma_{f'}} \exp\left(-\frac{\mu_{f'}^2}{2\sigma_{f'}^2}\right) \left(\frac{\mu_{f'}}{\sigma_{f'}^2}\right)^{2k} \Gamma(k+1/2)\mathrm{U}(k+1/2, k+2, 1/2\sigma_{f'}^2)\mathrm{d}t. \tag{6.187}$$

where $\mu_{f'}$ and $\sigma_{f'}^2$ are the derivative mean and variance of the posterior which now depend on t.

The result is expressed in an infinite series of terms, in practice, computations will only require a finite number of terms for sufficient accuracy. This would be dependent on the exact relationship between $\mu_{f'}$ and $\sigma_{f'}^2$; such an analysis would provide for interesting further research. The one dimensional arc length variance of the posterior is not derived in this thesis.

6.4.2 Vector Posterior

In the vector case, the expectation for the approximated integrand for a non-zero GP is given by:

$$\mathbb{E}\left[(\mathbf{f}'^T\mathbf{f}')^{\frac{1}{2}}\right] \approx \frac{\Gamma(m_{f'}+\frac{1}{2})}{\Gamma(m_{f'})}\left(\frac{\Omega_{f'}}{m_{f'}}\right)^{\frac{1}{2}}, \tag{6.188}$$

where $m_{f'}$ and $\Omega_{f'}$ are the Nakagami parameters which now depend on the mean and covariance functions of the GP posterior:

$$m_{f'} = \frac{[\mathrm{tr}(\Sigma_{f'}) + \mu_{f'}^T\mu_{f'}]^2}{2\mathrm{tr}(\Sigma_{f'}\Sigma_{f'}) + 4\mu_{f'}^T\Sigma_{f'}\mu_{f'}}, \quad \Omega_{f'} = \mathrm{tr}(\Sigma_{f'}) + \mu_{f'}^T\mu_{f'}, \tag{6.189}$$

where $\Sigma_{f'}$ and $\mu_{f'}$ are now the posterior mean and variance of the vector GP and are a function of t. This means we could approximate the mean arc length of the posterior

with the following:

$$\mathbb{E}[s] = \int_0^T \mathbb{E}\left[(\mathbf{f}'^T\mathbf{f}')^{\frac{1}{2}}\right] \mathrm{d}t, \tag{6.190}$$

$$\approx \int_0^T \frac{\Gamma(m_{f'} + \frac{1}{2})}{\Gamma(m_{f'})} \left(\frac{\Omega_{f'}}{m_{f'}}\right)^{\frac{1}{2}} \mathrm{d}t. \tag{6.191}$$

This non tractable expression now requires an integration (which, again, can be efficiently approximated with quadrature). The second moment of the posterior can similarly be expressed as (Reig et al., 2002):

$$\mathbb{E}[\mathrm{W}_1\mathrm{W}_2] = \left(\frac{\Omega_1}{m_1}\right)^{\frac{1}{2}} \left(\frac{\Omega_2}{m_2}\right)^{\frac{1}{2}} \frac{\Gamma(m_1 + \frac{1}{2})\Gamma(m_2 + \frac{1}{2})}{\Gamma(m_1)\Gamma(m_2)} \, {}_2F_1\left(-\frac{1}{2}, -\frac{1}{2}; m_2; \rho\right), \tag{6.192}$$

where m_i and Ω_i again depend on the mean and covariance functions of the GP posterior and are evaluated at t_i. The second moment of the posterior arc length is therefore:

$$\mathbb{E}[s^2]$$

$$\approx \int_0^T \int_0^T \mathbb{E}[\mathrm{W}_1\mathrm{W}_2]\mathrm{d}t_1\mathrm{d}t_2, \tag{6.193}$$

$$= \int_0^T \int_0^T \left(\frac{\Omega_1}{m_1}\right)^{\frac{1}{2}} \left(\frac{\Omega_2}{m_2}\right)^{\frac{1}{2}} \frac{\Gamma(m_1 + \frac{1}{2})\Gamma(m_2 + \frac{1}{2})}{\Gamma(m_1)\Gamma(m_2)} \, {}_2F_1\left(-\frac{1}{2}, -\frac{1}{2}; m_2; \rho\right) \mathrm{d}t_1\mathrm{d}t_2, \tag{6.194}$$

$$= \sum_{n=0}^{\infty} \frac{\left[\left(-\frac{1}{2}\right)_n\right]^2}{n!} \int_0^T \int_0^T \left(\frac{\Omega_1}{m_1}\right)^{\frac{1}{2}} \left(\frac{\Omega_2}{m_2}\right)^{\frac{1}{2}} \frac{\Gamma(m_1 + \frac{1}{2})\Gamma(m_2 + \frac{1}{2})}{\Gamma(m_1)\Gamma(m_2)(m_2)_n} \rho(|t_1 - t_2|)^n \mathrm{d}t_1\mathrm{d}t_2. \tag{6.195}$$

The correlation function for the posterior is of the same form as previously:

$$\rho(t - t') = \left[\frac{\partial^2}{\partial t \partial t'} k(t, t')\right]^2 \frac{1}{\sigma_{f'}^4}. \tag{6.196}$$

Numerical Simulations

As we did for the prior distribution, we empirically validate our theoretical results. We also compare posterior and prior length distribution side by side. Figure 6.9 show the prior and posterior lengths alongside the theoretically calculated mean and variance overlaid with the empirical values

6.4.3 Kernel Combinations

It is possible to effortlessly extend the results here to more complicated form of kernel than we have demonstrated, such as sums or products. The results of this chapter may be extended to piecewise kernels as well, where we would now consider the length for each domain separately.

6.5 Concluding Thoughts

In this chapter we have derived the distribution of the arc length of a GP. By considering the integrand distribution we were able to derive the moments for the arc length distribution for a single output GP, re-deriving previously computed results and providing the posterior distribution.

This novel approach equipped us with the necessary tools to tackle the vector valued case: the first known work to do so. An approximation to the integrand was specified in terms of Nakagami distribution whose parameters depended on the mean and covariance of the GP. Using this integrand distribution we could then derive expressions for the mean and variance of the arc length distribution for both a prior and posterior GP.

Further specifying the limits of the approximation by considering the limits of the integral expressions and assessing exactly when the integrand approximation breaks down would provide greater insight into the distribution. Finally, it remains to apply

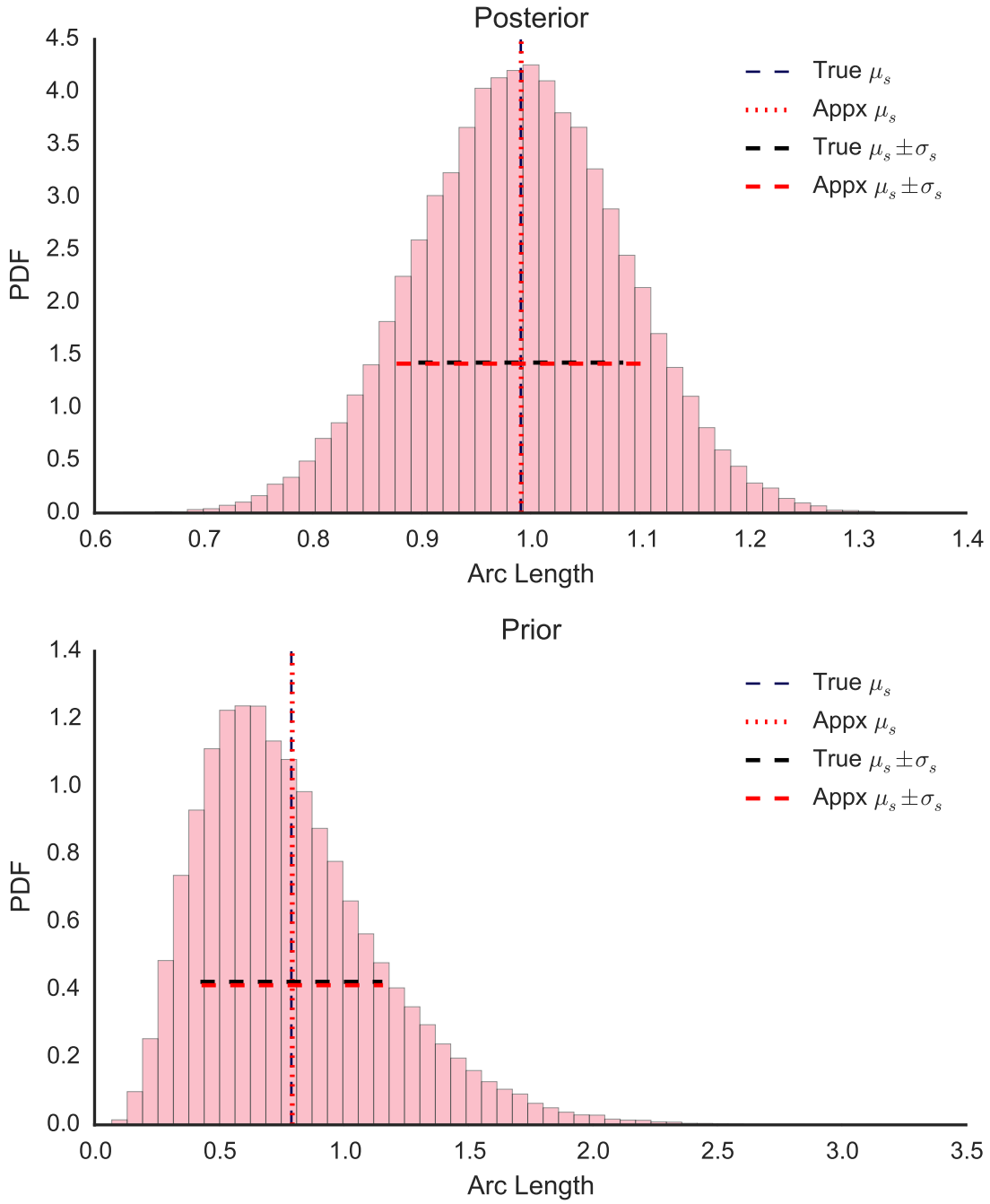

Figure 6.9: GP prior lengths against GP posterior lengths for the SE kernel with unit output variance and a lengthscale of 2.0. The theoretical and empirical mean are shown and the corresponding variance. The posterior values appear to fall symmetrically around the mean value, supporting the notion that draws are centered around the posterior mean function, 'anchored" by the data.

these equations to further theoretical problems and real world applications. We give some potential insight into that problem in Chapter 8.

In the next chapter we investigate a number of real world functional regression problems using the methods of Chapter 5, and present a novel use of arc lengths as functional features.

Chapter 7

Real World Experiments & Applications

7.1 Overview

In Chapter 5 we introduced three novel Gaussian Process (GP) functional regression models, the Gaussian Process Functional Index Model (GP-IND), Gaussian Process Functional Additive Model (GP-FAM) and Gaussian Process Functional Generalized Additive Model (GP-FGAM). Each GP model showed improvement in predicting values for synthetic examples. In this chapter we consider a number of challenging real world functional data sets and apply the GP methods to them, which are compared against the original functional models as well as a number of extra non-linear functional regression models. Finally we conclude, by considering the arc length of the functional trajectories as inputs to a GP model.

7.2 Functional Data

7.2.1 Spectrometric Data

Several Spectrometric (SPECTRO) data sets are explored; focusing our methods on real-world examples of function to scalar data. Spectrometric data provide rich examples of challenging functional data to analyse and they are regular repeated measurements at close wavelength intervals: the functional predictors are dense and with minimal noise. $X(\cdot)$ is the Near Infra-Red Reflectance (NIR) value at the wavelength ω – these are the functional inputs as seen in Figure 7.1. These functional inputs then map to a scalar output: $X(\omega) \rightarrow y$. We investigate four NIR data sets, each with different scalar responses.

The Moisture (MOST) (Kalivas, 1997) data set consists of NIR spectra of 100 wheat samples, measured in 2 nm intervals from 1100 to 2500nm. The associated scalar response variables are the samples' moisture content.

In the Octane (OCT) (Kalivas, 1997) data set we have at our disposal spectra from 60 gasoline samples, measured in 2 nm intervals from 900 to 1700 nm. The response variable is the octane numbers of the samples.

The Biscuit (BISC) (Brown et al., 2001, Osborne et al., 1984) contains the results of an experiment involving the variation of the compositions of biscuit dough pieces. Two sets of dough pieces were measured: a calibration set and a prediction set. They were created and measured as two distinct sets, on separate occasions, and do not result from a random (or any other) split of a larger set. There are 40 examples in the calibration sample and 32 samples in the prediction set. For each piece we have the NIR spectra over the range 1100-2498 nm in steps of 2nm. The four variables being measured are Fat (FAT), Sucrose (SUC), Flour (FLR) and Water (WAT) levels. Therefore we can consider this data set as containing four functional $\rightarrow$ scalar prediction problems. We combine the calibration and prediction set into one corpus of

observations for our purpose; resulting in a data set totalling 72 observations.

Tecator (TEC)[1] (Febrero-Bande and Oviedo de la Fuente, 2012) is data characterising a set of 215 pieces of finely chopped meat. For each piece we observe a single spectrometric curve which corresponds to the absorbance measured at 100 wavelengths (from 852 to 1050 in steps of 2nm). Our output in each case is the measured FAT, WAT and Protein (PRO) content of the piece, obtained by analytic chemical processing.

The spectral data provide a challenging regression problem. They are high dimensional data with only a comparatively small number of scalar observations. In this case fitting a standard GP or multivariate model becomes difficult due to the large number of variables required; for the BISC data set this would require joint optimisation of over 700 parameters using an Automatic Relevance Determination (ARD) kernel. The functional inputs are plotted in Figure 7.1; visually confirming that the predictors can be considered as functions over the NIR frequencies. The data is prepared as described in Section 7.3 and described in Table 7.1.

7.2.2 Diffusion Tract Imaging Data

As described in Chapter 4 we consider the Diffusion Tensor Imaging (DTI) functional data (Goldsmith et al., 2011a). From the John Hopkins University, diffusion tensor imaging was used to collect Fractional Anisotropy (FA) and Mean Diffusivity (MD) tract profiles for a number of patients. We have access to the Right Corticospinal (RCST) and Corpus Callosum (CCA) tracts for FA and MD, totalling four data sets, where we aim to regress the Paced Auditory Serial Additional Test (PASAT) score from the functional predictor. Table 7.2, outlines the number of response observations, functional predictor points, basis functions fit and Functional Principal Components (FPCs), as described in Section 7.3.

[1]This data set can be found at http://lib.stat.cmu.edu/datasets/tecator

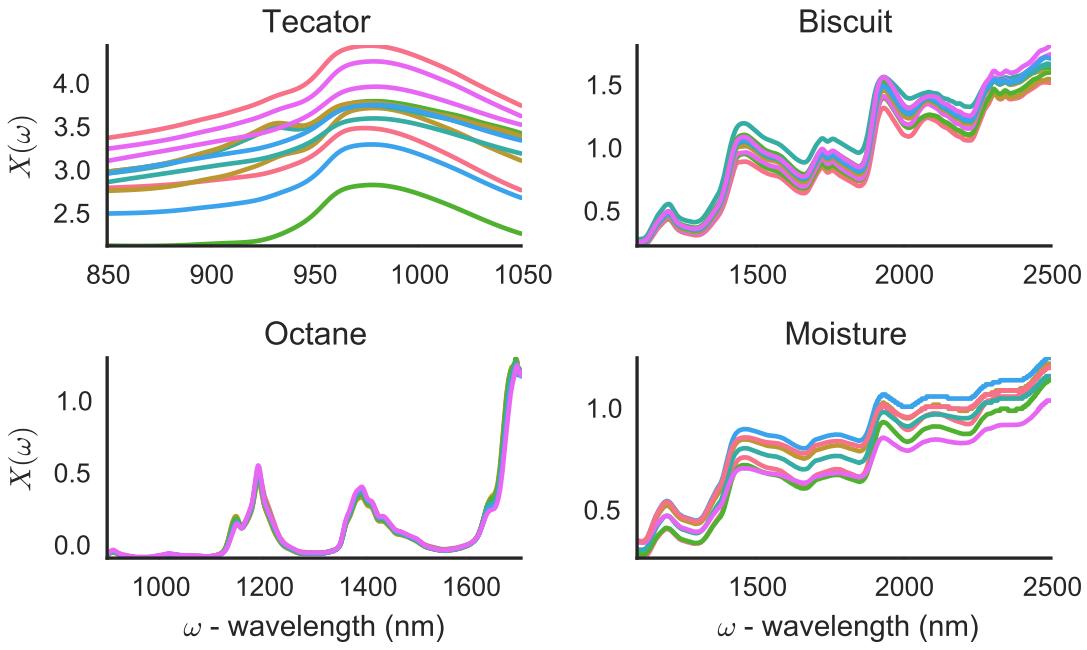

Figure 7.1: SPECTRO functional inputs. Clockwise from top left: TEC, BISC, MOST and OCT. The x-axis is wavelength (nm), the y-axis the spectrometric reading. Clear functional form is observed.

The scalar output we are interested in is the PASAT score for each patient. This is a real positive number which we pre-transform through a log transform and then standardise by removing the mean and dividing by the standard deviation. Figure 7.2 shows data samples from the MD and FA profiles along the CCA and RCST tracts; again the functional form of the inputs is clearly visible.

7.3 Experimental Setup

It is important to prepare the data such that the GP models are regressing on appropriately transformed responses: the GP values must be distributed on the real number line. As such, percentages values, which are between 0% and 100% (MOST, TEC and BISC), and purely positive numerical values (OCT) are inappropriate for a GP model – therefore the data needs to be transformed.

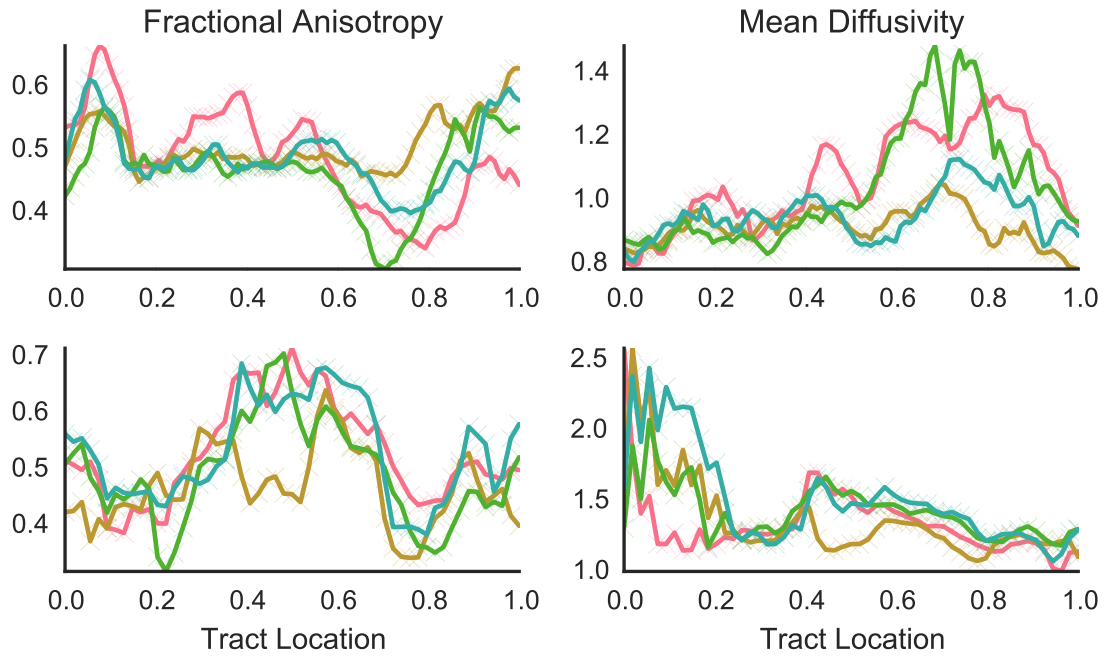

Figure 7.2: DTI functional input data. Top row: CCA, bottom row: RCST, left image FA, right image MD. We observe distinct functional behaviour along each tract.

For positive numbers, the log function is used to map the positive numbers to real valued numbers, whilst for percentage values the logit transformation is utilised:

$$y_{\text{trans}} = \log\left(\frac{y_{\text{orig}}/100 + \varepsilon}{1 - y_{\text{orig}}/100 + \varepsilon}\right) \tag{7.1}$$

A small noise value of $\varepsilon = 1 \times 10^{-6}$ is added to ensure that we can still capture values close to 0 or 100 percent. Once transformed, we whiten our responses: the mean is removed and we normalise the values to a standard deviation of one.

Each predictor is shifted to the domain $\omega \to t \in [0, 1]$: this provides consistency across all models and helps to ensure we avoid numerical difficulties associated by using values outside the unit interval. Similarly, we whiten the values of the functional predictors, removing the mean and normalising the values to a standard deviation of

one.

As the functional predictors depict frequency response behaviour, we choose the Fourier cosine basis to represent the functions. Each functional predictor is fit to the Fourier basis with n_b basis functions such that the error between the function values and the fitted values, divided by the norm of the function, is below a pre-specified threshold:

$$\text{error} = \frac{|X(t) - \Phi(t)\boldsymbol{\zeta}|^2}{|X(t)|^2}, \tag{7.2}$$

with $\Phi(t)$ the Fourier basis matrix, and $\boldsymbol{\zeta}$ the fit coefficients. We choose error $= 5 \times 10^{-2}$, as it provides a good balance between ensuring a representative fit and selecting a parsimonious number of basis coefficients; it is a value common in the literature and referred to as the Fraction of Explained Variance (FVE). For methods built on Functional Principal Component Analysis (FPCA), the number of principal components are selected by choosing the first k that explain 95% of the FVE. PACE: Principal Analysis by Conditional Expectation (PACE) is used to estimate the FPC on a grid of time points (100 for the spectrometric data and 50 for the DTI).

We compare against the three functional models of Chapter 5, Functional Additive Model (FAM) (Müller and Yao, 2008), Functional Index Model (INDEX) and Functional Generalized Additive Model (FGAM) (McLean et al., 2012). Two linear models are included as baselines, one using the FPC, Functional Linear Functional Principal Component Model (FLM-PCA) (Yao et al., 2004) and one that uses penalised spline regression for the functional basis, Functional Linear Basis Model (FLM-BASIS) (Ramsay and Silverman, 2005b). In addition, we compare against a number of other non-linear and non-parametric models: Functional Quadratic Model (FQM) (Yao and Müller, 2010), Generalised Functional Linear Model (GFLM) (Müller and Stadtmüller,

2005) and Functional Kernel Model (FV) (Ferraty and Vieu, 2006).

For FV we pick $\min(40, t)$ basis knots for the kernel function: default settings for the available code[2]. Similarly we compute FLM-BASIS using R, picking a smoothing parameter by Generalised Cross Validation (GCV) and admitting up to 25 basis functions.

For the FGAM we choose default setting as in McLean et al. (2012), using 6 basis functions for t and 7 for $X(\cdot)$. The spectrometric inputs are high-dimensional input and difficult to compute with GP-FGAM, requiring extensive computational resources to optimise the hyperparameters. In order to overcome this, we use subsamples of the functional trajectories at t_l points instead of t. This allows us to optimise in a reasonable timeframe, though will necessarily introduce error into our predictions.

Finally we compare against two baseline GP models. The first is an Squared Exponential (SE) ARD kernel treating the functional predictor as a high dimensional input vector with independent features, whist the second, Coefficients (COFS), uses the Fourier coefficients $\boldsymbol{\zeta}$ as inputs into an ARD SE kernel. All the GP methods are implemented in GPflow (Matthews et al., 2016).

We are interested in comparing the predictive capability of each model. Our underlying assumption is that the GP models will provide better predictive performance. We compare each model's predictive capability using a leave one out cross validation exercise on all data sets: we leave one sample out in training each model on the remaining samples, using the held out sample to test how well we've learnt the model. We iterate over the data and compute the Root Mean Square Error (RMSE) for each model, and the Log Predictive Probability (LPP) for the GP models:

[2]Code available from https://www.math.univ-toulouse.fr/staph/npfda/

	BISC	MOST	OCT	TEC
n	72	100	60	215
t	700	701	401	100
t_l	234	176	401	50
n_b	25	24	43	13
FPC	3	3	2	2

Table 7.1: Data information for each spectrometric data set

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i}^{N}(y_{i*} - y_i)^2}, \tag{7.3}$$

$$\text{LPP} = \sum_{i}^{n} \log p(y_*|x_*, \mathcal{D}), \tag{7.4}$$

with y_i the true value, x_* the test input and y_* the prediction at the test input. We then compute the average RMSE and standard error across all the leave one out experiments.

Tables 7.1 and 7.2 summarise the number of observations (n), time points of each functional input (t), the subsampled size (t_l) (for SPECTRO), the number of Fourier basis function (n_b) and the number of FPCs for each data set.

	FA		MD	
	CCA	RCST	CCA	RCST
n	99	66	99	66
t	47	55	47	55
n_b	47	55	47	55
FPC	10	11	11	8

Table 7.2: Data information for the DTI data.

7.4 Results

7.4.1 Spectrometric Data

The RMSE (smaller numbers are better) and LPP (bigger numbers are better) values from our experiments are shown in Table 7.3 and Table 7.4.

We observe competitive prediction using the novel GP methods, obtaining low RMSE values, with good prediction for the GP-FGAM and GP-IND models. Methods based on the FPC inputs do not perform as well. We note that the GP-FAM performs the worst out of the new GP models: this is likely a result of the FPC being unable to adequately capture the high-frequency behaviour of the spectrometric data. The models do not comprehensively outperform the baseline results in every instance. Interestingly the GP baselines perform well, indicating that it is possible to accurately train an ARD kernel. Further, representation of the functional predictors in their basis is clearly an effective way to compress the information in our predictors as evidence by the experimental results.

Figure 7.3 plots the true values against the predicted values, demonstrating that we are capturing the relationship between the functional inputs and the response.

In Chapter 5 we derived a posterior distribution for GP-FGAM surface, f. Figure 7.4 shows the surfaces for the GP-FGAM models, with the functional overlaid on the surface mean. From these we observe that the GP-FGAM is working as expected, where the surface mean returns to zero and the surface variance increases as we move away from observed points. Higher surface mean values should indicate areas that are strongly related to the output, however, this is not straightforward to disentangle – in part due to having whitened both the inputs and responses.

Figure 7.5 plots the $y = g(\int X(t)\beta(t)\mathrm{d}t)$ and the corresponding coefficient function for the TEC data. We observe a clear non-linear relationship between the index and the response; the values of $\beta(t)$ could relate to important wavelengths of the

functional predictor, however, as indicated in Chapter 5 this would require more careful prescription of the prior to ensure a fully interpretable model. Plots for the other data sets do not reveal a clear relationship between the index and the response.

The GP-FGAM may be limited by kernel choice: we investigated only the SE kernel and it is possible that for the SPECTRO data we require a more complex kernel. Furthermore, for each of the GP models we could try complex combinations of kernel, sums or products, which would improve performance: the baseline models provide no such mechanism. Additionally we were unable to use the full functional predictors, thus introducing error in our integral, likely leading to an under-performance by the model.

7.4.2 Diffusion Tract Imaging Data

The DTI is a difficult set to model: the functional models perform poorly, overall, failing to adequately capture the relationship between the tract profiles and the PASAT response. Only GP-IND performs well, capturing a relationship as evidence in Figure 7.6, for all but MD RCST.

Figure 7.7 plots the GP-FGAM surfaces, revealing potential tract locations that reveal cognitive capability. We resist making strong conclusions from this plot as, though we fare better than the baselines, we still do not fully capture the relationship between the tract profiles and the PASAT score.

Plots of $y = g(\int X(t)\beta(t)\mathrm{d}t)$ and the corresponding coefficient function for the GP-IND do not provide further insight into the DTI data. In this case the GP baselines do not perform well, thus the need for functional regression models is apparent and we cannot naively use the ARD kernel for every problem.

	BISC FAT	BISC WAT	BISC SUC	BISC FLR	MOST
GP-FGAM	0.159 (0.020)	0.166 (0.029)	1.182 (0.085)	0.184 (0.033)	0.121 (0.010)
GP-IND	**0.093**(**0.008**)	0.192 (0.035)	**0.408**(**0.074**)	**0.132**(**0.016**)	0.125 (0.014)
GP-FAM	0.501 (0.043)	0.242 (0.033)	1.272 (0.097)	0.358 (0.032)	0.164 (0.014)
ARD	0.137 (0.019)	0.137 (0.025)	0.509 (0.093)	0.198 (0.037)	0.130 (0.010)
COFS	0.098 (0.010)	**0.118**(**0.025**)	0.544 (0.103)	0.199 (0.047)	**0.105**(**0.008**)
FGAM	0.097 (0.010)	0.155 (0.031)	0.633 (0.125)	0.217 (0.046)	0.115 (0.009)
FV	0.498 (0.037)	0.270 (0.032)	1.038 (0.092)	0.326 (0.031)	0.264 (0.026)
INDEX	0.136 (0.015)	0.148 (0.024)	0.542 (0.077)	0.165 (0.028)	0.125 (0.010)
FLM-BASIS	0.115 (0.011)	0.146 (0.023)	0.485 (0.074)	0.167 (0.027)	0.127 (0.011)
FAM	0.645 (0.045)	0.551 (0.050)	1.412 (0.099)	0.510 (0.041)	0.942 (0.035)
FQM	0.472 (0.046)	0.338 (0.040)	1.396 (0.142)	0.431 (0.049)	0.904 (0.084)
GFLM	0.501 (0.039)	0.437 (0.045)	1.450 (0.111)	0.511 (0.038)	0.597 (0.048)
FLM-PCA	0.480 (0.047)	0.288 (0.039)	1.261 (0.103)	0.403 (0.036)	0.372 (0.033)

	OCT	TEC FAT	TEC pro	TEC WAT
GP-FGAM	0.650 (0.059)	0.236 (0.016)	0.086 (0.005)	0.059 (0.004)
GP-IND	0.154 (0.018)	**0.123**(**0.009**)	0.074 (0.005)	0.069 (0.004)
GP-FAM	0.784 (0.066)	0.631 (0.043)	0.201 (0.012)	0.207 (0.012)
ARD	0.143 (0.018)	0.123 (0.012)	0.102 (0.009)	0.065 (0.008)
COFS	0.133 (0.014)	0.086 (0.014)	0.077 (0.007)	**0.045**(**0.004**)
FGAM	0.150 (0.020)	0.229 (0.018)	0.103 (0.011)	0.070 (0.006)
FV	0.347 (0.045)	0.318 (0.022)	0.118 (0.009)	0.112 (0.007)
INDEX	0.156 (0.023)	0.495 (0.036)	0.104 (0.007)	0.101 (0.006)
FLM-BASIS	**0.115**(**0.011**)	0.464 (0.03)	**0.073**(**0.005**)	0.123 (0.006)
FAM	0.871 (0.066)	103.868 (102.638)	0.288 (0.078)	0.990 (0.365)
FQM	0.797 (0.059)	0.674 (0.044)	0.204 (0.012)	0.219 (0.013)
GFLM	0.872 (0.078)	0.632 (0.048)	0.209 (0.012)	0.218 (0.016)
FLM-PCA	0.878 (0.064)	0.675 (0.044)	0.207 (0.012)	0.216 (0.014)

Table 7.3: Spectrometric data RMSE (standard error) experiment results. We perform a leave one out cross validation on each data set; training on all but one sample, which we use as the test point. We repeat this for the entire data set, retraining each time. We compare the RMSE (and standard error) as a measure of the predictive performance, with best values in bold. The GP-FGAM and GP-IND give competitive results. The BISC Sucrose appears as an outlier, demonstrating that there may be a tenuous relationship between the sucrose level and functional predictor. It is interesting that the GP baseline models perform extremely well, indicating that the coefficients in the COFS model provide a strong signal as an input and that it is possible to optimise an ARD model well.

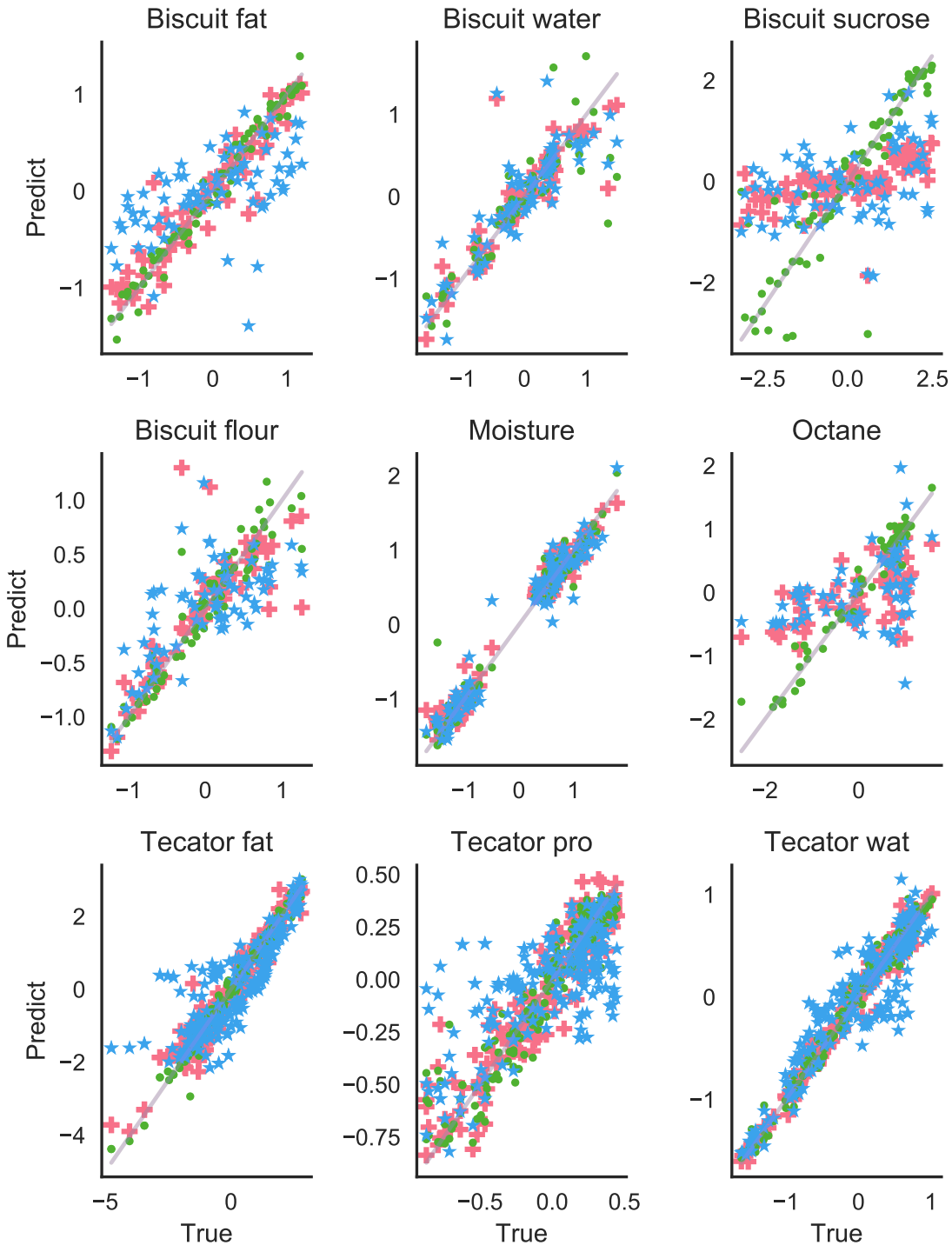

Figure 7.3: True (x-axis) vs predicted (y-axis) values for GP-IND ($\cdot$), GP-FAM ($\star$) and GP-FGAM ($+$). Low RMSE values are obtained and these plots demonstrate that we are capturing the relationship between the functional inputs and the response. The BISC Sucrose appears as an outlier, demonstrating that there may be only be a weak relationship between the sucrose level and functional predictor.

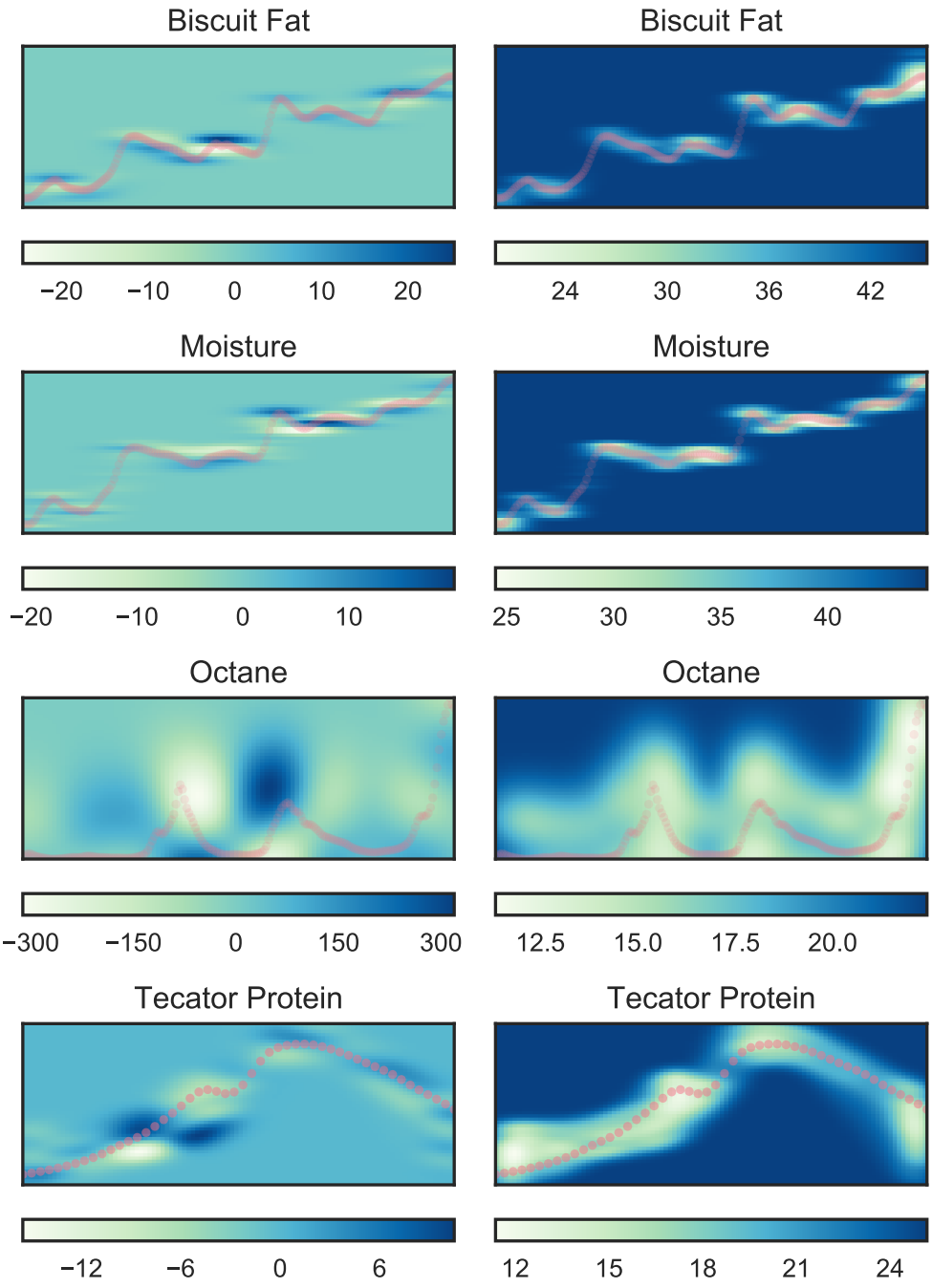

Figure 7.4: GP-FGAM surface plots for each spectrometric data set: on the left the expected mean of the surface $\mathbb{E}[f(x^*, t^*)]$ and on the right, the surface variance $\mathbb{V}[f(x^*, t^*)]$. The x-axis is the wavelength and the y-axis is the normalised NIR value; (refer to original Figure 7.1). As we move away from the functional trajectories, the surface mean decreases to zero whilst the variance increases – confirming that we are observing expected GP posterior behaviour. The surfaces have correctly captured high values around the functional predictors: with more careful learning and preparation of the data these plots could be used to determine the most informative wavelengths of the NIR values.

	BISC WAT	BISC PRO	BISC SUC	BISC FLR	MOST
GP-FGAM	7.9	-22.4	-127.9	-31.6	45.3
GP-IND	31.5	-443.2	-493.5	-78.3	41.3
GP-FAM	-68.4	-33.6	-135.7	-43.6	10.0
ARD	33.6	**29.1**	**−53.8**	**13.3**	31.9
COFS	**47.5**	13.9	-81.9	-24.1	**54.4**

	OCT	TEC fat	TEC pro	TEC wat
GP-FGAM	-72.0	-62.7	153.8	237.7
GP-IND	-277.0	50.8	175.2	201.9
GP-FAM	-81.9	-282.1	-23.0	-30.2
ARD	**131.8**	**220.4**	107.1	209.3
COFS	26.0	135.6	**189.4**	**325.5**

Table 7.4: SPECTRO data LPP experiment results for the GP models (best results in bold). We perform a leave one out cross validation on each data set; training on all but one sample, which we use as the test point. We repeat this for the entire data set, retraining each time. The LPP is reported as a measure of predictive performance. We see varying performance between the GP models and baselines. Again we note the GP baseline models perform extremely well. The GP-FAM performs poorly across the data sets.

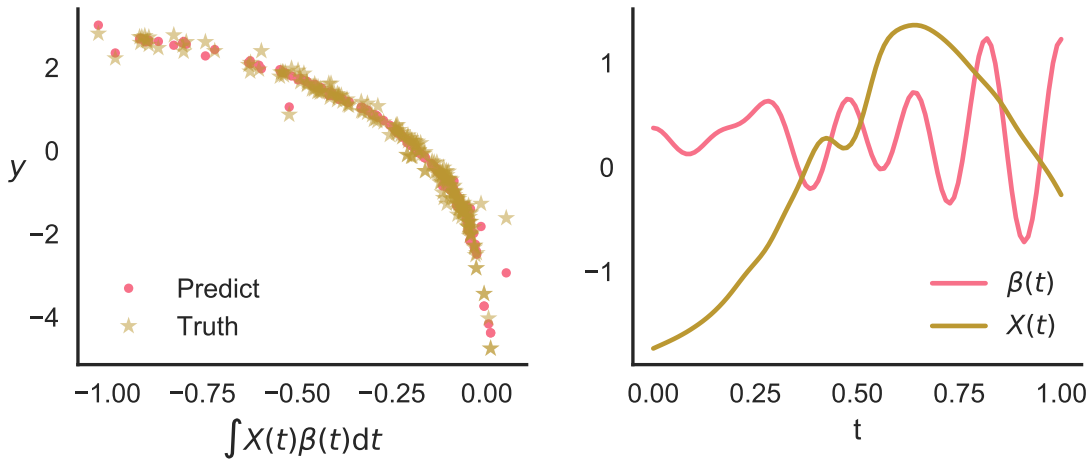

Figure 7.5: Left column: plot of y against the functional index, $\int X(t)\beta(t)\mathrm{d}t$ for TEC TEC. We observe that a non-linear relationship is captured by the GP-IND. Right column: $\beta(t)$ with a corresponding functional input: with more careful prior specification, this function could indicate the most salient frequencies for the prediction problem.

	FA		MD	
	CCA	RCST	CCA	RCST
GP-FGAM	0.801 (0.061)	0.850 (0.076)	0.801 (0.061)	0.792 (0.082)
GP-IND	**0.295**(**0.048**)	**0.638**(**0.093**)	**0.188**(**0.049**)	**0.800**(**0.077**)
GP-FAM	0.892 (0.064)	0.884 (0.083)	0.923 (0.068)	0.883 (0.082)
ARD	0.822 (0.057)	0.784 (0.074)	0.828 (0.057)	0.860 (0.078)
COFS	0.822 (0.057)	0.786 (0.074)	0.827 (0.057)	0.844 (0.078)
FGAM	0.815 (0.067)	0.913 (0.080)	0.918 (0.065)	0.914 (0.082)
FV	0.854 (0.064)	0.862 (0.084)	0.856 (0.064)	0.842 (0.082)
INDEX	1.003 (0.072)	1.077 (0.106)	0.915 (0.075)	0.977 (0.094)
FLM-BASIS	0.834 (0.059)	0.841 (0.073)	0.840 (0.060)	0.828 (0.074)
FAM	0.831 (0.057)	0.819 (0.073)	0.831 (0.057)	0.819 (0.073)
FQM	1.308 (0.114)	5.676 (0.663)	1.622 (0.168)	1.464 (0.173)
GFLM	0.883 (0.066)	0.850 (0.083)	0.883 (0.066)	0.850 (0.083)
FLM-PCA	0.868 (0.063)	0.842 (0.087)	0.872 (0.064)	0.868 (0.081)

Table 7.5: DTI RMSE (standard error) data experiment results. We perform a leave one out cross validation on each data set; training on all but one sample, which we use as the test point. We repeat this for the entire data set, retraining each time. We compare the RMSE (and standard error) as a measure of the predictive performance, with best values in bold. Almost none of the models are able to provide good RMSE results. We see that GP-IND gives the best RMSE. However, it is clearly overconfident as evidenced by the low LPP in Table 7.6.

	FA		MD	
	CCA	RCST	CCA	RCST
GP-FGAM	**83.9**	**56.1**	**83.2**	**52.3**
GP-IND	-56342.1	-59077.1	-54151.5	-72056.6
GP-FAM	-151.5	-104.0	-155.3	-104.8
ARD	-145.6	-101.2	-146.0	-106.1
COFS	-145.6	-101.1	-145.9	-105.1

Table 7.6: DTI data LPP experiment results for the GP models (best results in bold). We perform a leave one out cross validation on each data set; training on all but one sample, which we use as the test point. We repeat this for the entire data set, retraining each time. The LPP is reported as a measure of predictive performance. We see that though GP-IND gave the lowest best RMSE, it is clearly overconfident as evidenced by the low LPP. The GP-FGAM emerges as the best predictive model as measured by the LPP.

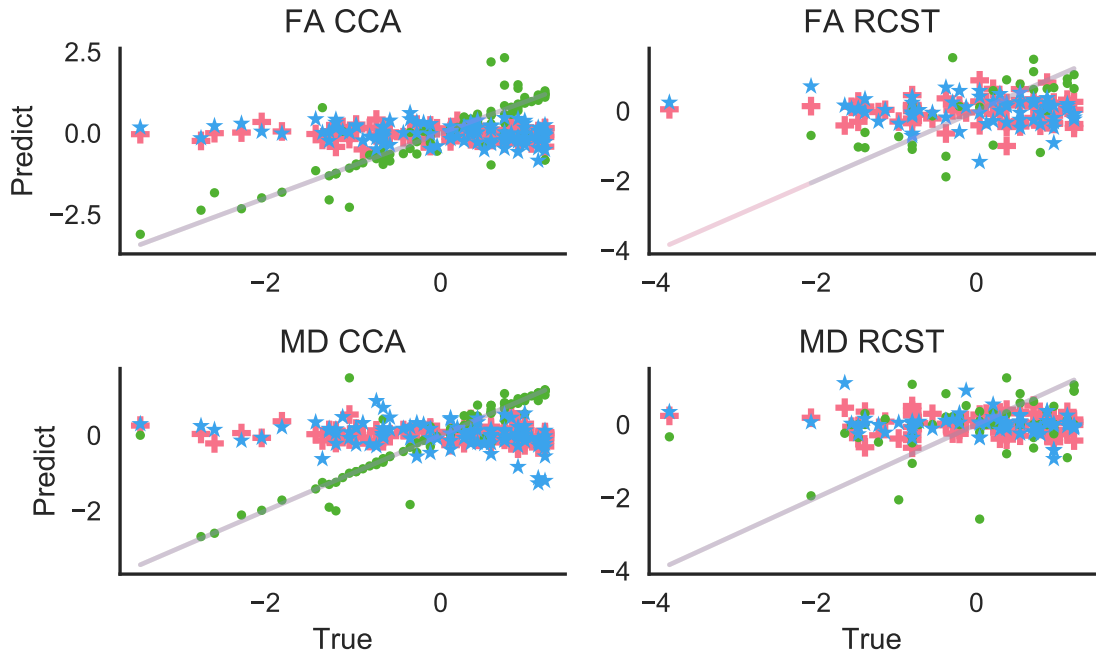

Figure 7.6: A comparison of true and predicted values for the DTI data for each of the GP functional models proposed: GP-IND ($\cdot$), GP-FAM ($\star$) and GP-FGAM ($+$). It is difficult to accurately capture the relationship between the functional predictors and the PASAT score, except for in the GP-IND model.

7.5 Arc Lengths as Functional Features

Previously the possibility of using arc lengths as functional features was highlighted and we now investigate the veracity of such an approach. We compute the arc length for each functional predictor numerically. These are now used as our inputs to the functional responses. We perform the same experiments as detailed previously, using a leave one out cross validation and the SE kernel.

Results are reported in Table 7.7. We see that using the arc length as input enables us to map the predictors to the response. Though we are not competitive against the functional models, these results indicate that we may be able to improve our functional regression models by augmenting our feature space to include arc lengths.

	RMSE	LPP
BISC WAT	0.43 (0.04)	-61.45
BISC PRO	0.36 (0.03)	-47.01
BISC SUC	1.34 (0.10)	-139.84
BISC FLR	0.44 (0.03)	-56.34
MOST	0.65 (0.05)	-123.65
OCT	0.85 (0.07)	-86.3
TEC FAT	1.17 (0.06)	-391.14
TEC PRO	0.27 (0.01)	-73.6
TEC WAT	0.52 (0.03)	-208.74

DTI		RMSE	LPP
FA	CCA	0.83 (0.06)	-143.7
	RCST	0.86 (0.09)	-112.66
MD	CCA	0.83 (0.06)	-143.7
	RCST	0.81 (0.07)	-95.66

Table 7.7: Experimental results for regression using arc length as the input. We perform a leave one out cross validation on each data set; training on all but one sample, which we use as the test point. We repeat this for the entire data set, retraining each time. RMSE and LPP are reported as a measure of predictive performance. These values, whilst not competitive against the functional regression models introduced, suggest that there is signal present in the length of the predictor.

7.6 Concluding Remarks

Throughout this chapter a number of real world functional data sets have been investigated. We compared the three GP functional methods against a range of baseline models. The GP functional methods provide competitive predictive capability across the board. Furthermore we are able to generate the latent surface for the GP-FGAM that we believe may aid interpretability and understanding of the functional data sets. With further attention to kernel choice we believe we can improve the predictive performance of the GP functional methods. Results using arc lengths as functional features indicate that there is potential for inclusion as features for predictive models.

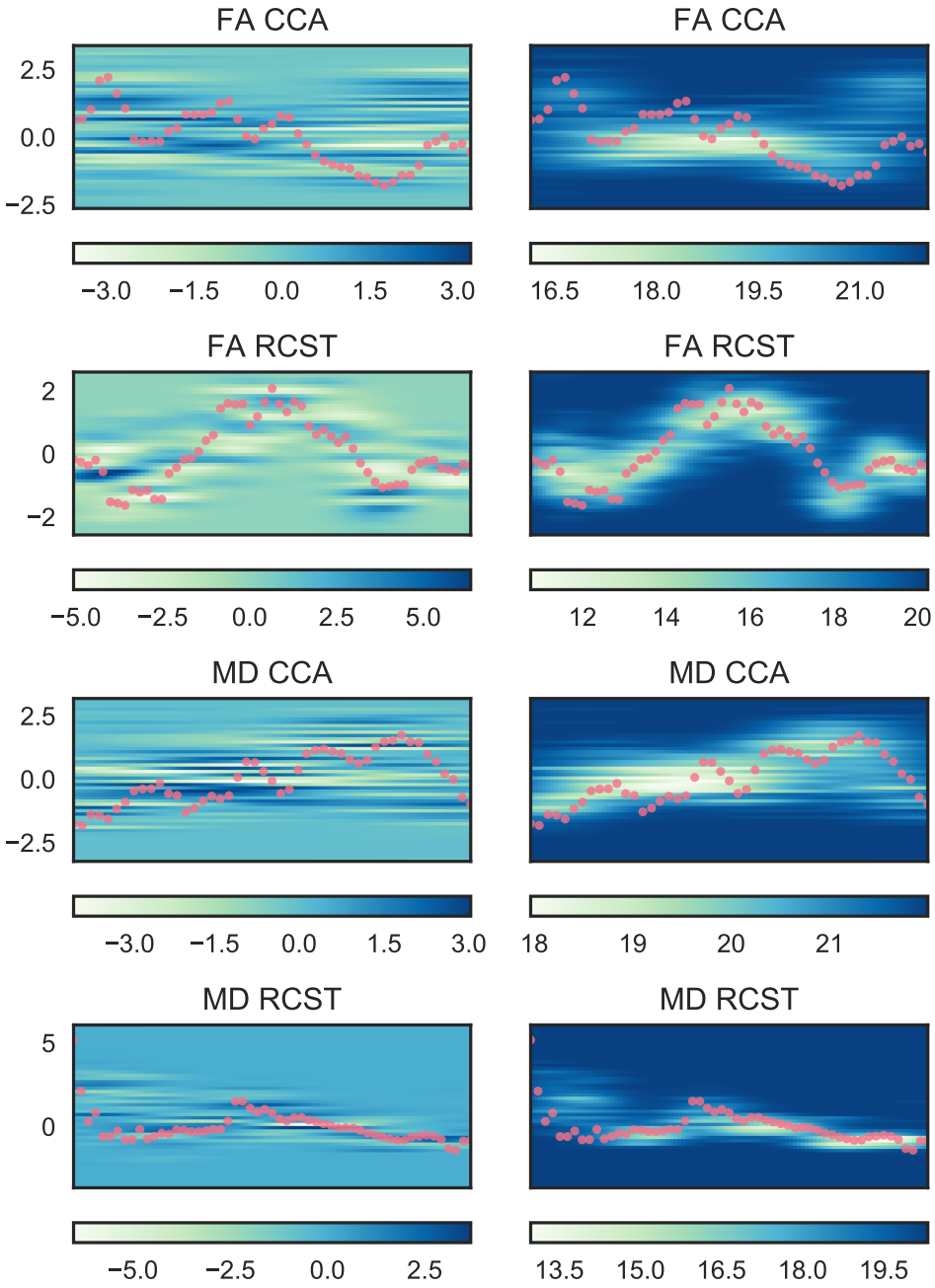

Figure 7.7: GP-FGAM surface plots for each DTI data set: on the left the expected mean of the surface $\mathbb{E}[f(x^*, t^*)]$ and on the right, the surface variance $\mathbb{V}[f(x^*, t^*)]$, with an example trajectory overlaid. From top to bottom, FA-CCA, FA-RCST, MD-CCA and MD-RCST. The x-axis is the tract location and the y-axis is the normalised DTI value; for more information refer to Figure 7.2. As we move away from the functional trajectories, the surface mean decreases to zero whilst the variance increases, as expected. The surfaces have correctly captured high values around the functional predictors: with more careful learning and preparation of the data these plots could be used to aid in determining the most informative tract locations related to cognitive performance.

Chapter 8

Discussion & Future Work

8.1 Overview

This chapter details a summary and discussion of the contributions of this thesis. We recap novel work and experiments, and highlight how machine learning has been advanced. Numerous avenues for future research naturally emerge from this thesis, which are clearly articulated with a forward trajectory.

8.2 Discussion

Three new functional regression models were introduced in Chapter 5: Gaussian Process (GP) extensions of the Functional Index Model (INDEX), Functional Additive Model (FAM) and Functional Generalized Additive Model (FGAM). Each model developed was built without considering noise on the functional inputs, allowing us to develop: the Gaussian Process Functional Index Model (GP-IND), the Gaussian Process Functional Additive Model (GP-FAM) and the Gaussian Process Functional Generalized Additive Model (GP-FGAM).

We demonstrated how we would perform Maximum Likelihood Estimate (MLE) or Maximum a Posteriori (MAP) inference in each case, and then showed how to

compute the posterior distribution for latent surface in the GP-FGAM model. In synthetic examples the GP models were compared against their counterparts, giving lower Root Mean Square Error (RMSE) in experiments, demonstrating their predictive strength.

We retain the interpretable aspects of the FGAM in the GP-FGAM where the use of the GP allows quantification of the surface uncertainty. We did not compare the models against each other on joint synthetic examples and it would prove interesting to see in what circumstance each fails and performs well.

In Chapter 7 we applied the GP functional regression methods to a number of real world data examples. The Spectrometric (SPECTRO) data provided high-dimensional, challenging functional regression problems. The GP-FGAM and GP-IND provided competitive results across the experiments. For the SPECTRO data the GP-FAM did not perform as well: likely due to the inability of the Functional Principal Components (FPCs) to capture sufficient detail in the functional inputs.

The GP baseline competitor models were competitive; an indication that treating the inputs as a long vector with a GP is a reasonable approach to functional data modelling – further supporting the case for using GPs to model complex functions. It may be that the Squared Exponential (SE)-Automatic Relevance Determination (ARD) kernel is learning a functional distance metric, and that we have approximated:

$$\sum_{d=1}^{N} \frac{(x_{id} - x_{jd})^2}{\theta_d^2} \approx \lambda \int_{\mathcal{I}} |X_i(t) - X_j(t)|^2 \mathrm{d}t, \tag{8.1}$$

for ARD values θ_d and a value λ. Comparing the ARD kernel against a functional distance metric, may further elucidate the situation. However, we note that in that case we do not gain any interpretable insights into the functional data, and only a predictive model.

The Diffusion Tensor Imaging (DTI) data proved to be more difficult to predict

than the SPECTRO data; none of the baselines achieved good predictive results. Our newly introduced model GP-IND gave the best predictive results, emphasising the strength of our GP model and further reinforcing that a range of functional regression models are needed for different data types.

Generated surfaces from the GP-FGAM provided insight into the trajectories: we were able to get visualisations that may provide insights into the underlying functional data, with a more careful prescription of hyperparameters. The GP-FGAM is time consuming due to the need to numerically compute the integrals. Thus, though it is a compelling model, it is difficult to implement for high dimensional data and further improvements in implementation and numerical libraries for GPs would allow it be more readily used.

In Chapter 6 we investigated the distribution of the arc length of a GP. Initially we considered the one dimensional case, approaching the problem by considering the arc length integrand. We were able to derive an exact expression for the arc length integrand distribution, confirming it was a probability distribution and calculating its mean and variance.

Using the expectation of the integrand we were able to compute the expectation of the arc length using Fubini's theorem. We re-derived the result of Barakat and Baumann (1970), whilst providing insight into the shape of the integrand distribution. A novel approximation of the variance was proposed which proved inaccurate, suggesting a careful transformation of the integrand variance needs to be further investigated. This approach equipped us to compute the arc length distribution for non-zero GPs, a limitation of Barakat and Baumann (1970), and thus we were able to compute the arc length distribution of a posterior GP represented as a sum of hypergeometric confluent functions.

Numerical experiments confirmed the theoretical results, and by investigating the form of the expected length, we could see a relationship between the choice of kernel,

kernel hyperparameters and the length: given in terms of the derivative variance $\sigma_{f'}^2$ of the GP.

Tackling the integrand distribution proved to be a key insight, allowing us to address the vector output case. Investigating the integrand, we see that is the square root of the sum of correlated normal variables; with the correlation defined by the GP kernel. Unfortunately no simple closed-form expression could be found and thus we resorted to an approximation of the integrand.

We approximated the sum of squares, a quadratic form, of the integrand with a moment matched gamma distribution. Then, recognising the square root of a gamma function as a Nakagami variable (an exact relation) we were able to approximate the entire arc integrand as a Nakagami distributed variable. Samples of the approximations and the true integrand distributions showed that our approximation holds well.

By approximating the integrand with a Nakagami distribution we were then able compute the expected length. For a zero mean prior, we have a compact expression in terms of the Nakagami parameters. Using the mixed moments of a Nakagami variable we can also derive the second moment, and hence the variance of the arc length. This approach grants the ability to compute simply an arc length distribution for a posterior vector GP.

We showed that we could obtain bounds on the expected length in terms of kernel hyperparameters and the eigenvalues of the coregionalistion matrix.

The arc length distributions were validated by considering samples drawn from GP kernels and computing the lengths numerically. High fidelity between our theoretical values and the empirically generated values provide evidence for the strength of the approximation.

Due to the additive nature of integration, our methods are easily extendable for a range of kernels: additive kernels, or even piecewise kernels over disjoint domains. We

used the computed arc length as a functional feature in Chapter 7. It provided some accuracy above pure noise, however, it was not competitive compared to the pure functional methods. It may be worth considering arc lengths of derivatives, which may be more widely distributed in the input space.

8.3 Avenues for Future Work

8.3.1 Functional Regression

In Chapter 5 it was demonstrated that each GP model provided better empirical performance than the competitors. However, we did not compare them in alternate situations and it would provide further insight into the models to see under what circumstance each fails. Unfortunately it does not appear that a detailed and thorough comparison has been considered in the literature.

Considering the experiments of Chapter 7, there are a number of themes that could be expanded. First, getting the GP-FGAM to run on the full data – this could be achieved by splitting the functional predictors in time or space and distributing the calculations on a Graphics Processing Unit.

We tested using the SE kernel; our interest lay in testing the models, not optimising over kernels. It would be fruitful to see what performance improvements may be gained by changing kernels across the models. The DTI data are particularly rough, suggesting that a Matérn32 (MAT32) or Matérn52 (MAT52) may be more appropriate, whilst for the SPECTRO data a spectral kernel may be more appropriate, especially for the GP-FGAM model. Alternatively, for the DTI data it may be worth smoothing the curves, with a GP, and using denser trajectories to model.

Though we have introduced all the models under the auspice of regression, they are easily amenable to the problem of classification. We would augment each of our models by mapping our latent GP function through a probit or logit function

as described in Chapter 3, performing inference over the combined model. It has previously been suggested to use the derivatives of curves to further separate them (Ferraty and Vieu, 2006, Ferraty et al., 2013); looking at the derivatives of the data functions may enrich the discriminative power of the GP functional models.

Another class of models that were not addressed in this thesis were function-to-function models. The GP-FAM is a natural candidate for function-to-function in modelling, mapping the FPC of the input trajectories to the output trajectories, as demonstrated by Müller and Yao (2008). As with the function to scalar case, we could map the principle scores using a GP. However, our GP now outputs a distribution over the $y(t)$ FPCs, not the trajectories themselves: what this uncertainty would mean and how it would relate to observations of $y^*(t)$ is not immediately clear.

It is commonplace in the literature to consider the convergence rate of the functional model, an issue not considered in this thesis. Assessing the convergence and consistency of the GP models would provide valuable insight into the type of applications we could use the model for and under what conditions we should expect good performance.

For each functional regression model considered, we made the assumption that our input trajectories were completed observed. This enabled simple and straightforward inference models to be developed. Furthermore it helped provide us with interpretable models to understand the relationship between functional inputs and scalar outputs. For our experiments we found that with sufficiently dense sampled trajectories we were able to obtain competitive results. However, in reality we rarely face fully observed inputs. In fact, our observations of the real world are by construction discrete, thus we necessarily never truly observe a continuous variable.

Dealing with uncertain inputs is an existing concept within the GP community (Deisenroth and Rasmussen, 2011, Girard and Murray-Smith, 2003, Mchutchon and Rasmussen, 2011). Naturally research has focussed on multivariate inputs corrupted

with noise.

For these situations we place a noise model over the inputs and augment our GP inference, resulting in increased uncertainty on our targets. This is a naturally satisfying result: when we are uncertain in our input this should propagate forward to our ultimate subject of interest.

Consequentially we would like to transfer such ideas to the functional regression case, incorporating uncertainty around noisy and potentially sparse functional predictors. Such problems have been tackled previously in McLean et al. (2014) in the FGAM model, where the trajectories are modelled using their Functional Principal Component Analysis (FPCA) decompositions, whilst in Radchenko et al. (2015) the predictors were drawn from sparsely observed GPs.

Parametrising the functions allows us to borrow strength across trajectories; this is particularly useful for missing or spare data situations. We could for example take the approach of McLean et al. (2014) and represent our functions using their FPCA decompositions with a noise model or in a given basis with a noise model (James, 2002, Radchenko et al., 2015). Gaussian Processes are our preferred probabilistic tool to model non-linear functions, as such we would use them to model our underlying trajectories, for example:

$$x(t) \sim \mathcal{GP}(\mu(t), \Sigma(t, t')). \tag{8.2}$$

Each trajectory would then be modelled jointly as a GP. We would then couple the trajectories to the response using the appropriate kernel: GP-IND, GP-FAM or GP-FGAM kernels. Inference now becomes more complicated and we envision the need to appeal to approximation methods such as variational inference.

Placing a distribution over the functional predictors would allow us to model

sparse and noisy functional inputs. We envision modelling situations, such as the Battery Impedance (BAT) described in Chapter 4, where it is expensive to observe full inputs: thus we would use sparse samples coupled with a robust model to both accurately predict our response, coupled with meaningful uncertainty.

8.3.2 Arc Lengths

In this thesis substantial theory is presented around GP arc length distributions. A number of interesting theoretical questions remain open along with the important question of practical applications.

The first question we ask is: can we better characterise the variance of the arc length for a single output GP? We presented an approximate value for the variance, that was inaccurate. It may be possible to derive an un-approximated form following Barakat and Baumann (1970) and obtain a variance for the posterior GP. The expression for a non-zero mean is given as a sum of hypergeometric functions; clearly it is impractical to evaluate such expressions. Observing the functional form it is clear that for values of $\mu_{f'}$ and $\sigma^2_{f'}$, we can truncate the expression to a small number of terms: how exactly that relationships manifests warrants further exploration.

Similarly, the variance term for the second moment of the vector arc length is expressed in an infinite sum; in experiments we found that five terms gave reasonable results. How exactly this variance decays would be crucial to understanding how to use these expressions in practice. Further, placing bounds on the approximation distribution would help understand either how to improve the approximation or the limitations it presents. Given that the approximation depends crucially on how well a single gamma variable approximates the sum of squared normal variables, we believe that a bound can be derived in terms of that relationship.

Additionally we may consider alternative ways to approximate the integrand distribution. We may use a variational approach, and attempt to bound the distribution

of the arc length. For our moment matched approach, we used the first two moments, as such we could naturally extend to three to improve the fidelity of our approximation. Whatever approach is taken, it is unlikely that we will avoid the need to integrate over the correlation function in the variance, as this is the crucial element that dictates the behaviour of the GP.

It is important to consider how to use the distribution of the arc length in practice. We note that a length corresponds to a measurable quantity in a physical problem. Thus, if we need to constrain the length in a practical application then we can model our problem using a GP and constrain the expected length. We envision using the arc length to help define a prior over types of functions. Generally a prior can be thought of as a regulariser or a penalty term. Using the expected length as an example we could penalise functions that are too long or too short. For example, if we wanted to specify a path of given length, passing through five waypoints, we could penalise with respect to the length of the GP and optimise for a curve with given smoothness. This allows us to be more flexible and removes the need to use splines or parametrise our curves. Furthermore, in Chapter 6 we demonstrated that we could place bounds on the expected length of our GP samples given the hyperparameters. If we were interested in generating lengths of a given length we could engineer our kernel using those distributions to meet our requirements.

Another quantity, closely related to arc length is the energy of the GP:

$$\mathcal{E}(f) = \int_{\mathcal{T}} |f'|^2 \mathrm{d}t. \tag{8.3}$$

For a GP modelling a physical system, this value could correspond to important characteristics, such as power used or fuel remaining. Due to the squared term in the integrand, we envision computation of the moments being less involved then the arc length. Minimising arc length is the same as minimising the energy of a curve, as

previously discussed in Chapter 4. Generally this is done with respect to the curve and results in a differential equation, however, we could also minimize with respect to the GP hyperparameters, which may lead to interesting solutions.

This thesis demonstrates an intimate relationship between the arc length of GP and the kernel parameters; notions of stationarity are often closely related to a varying length scale in time. We suggest that it may be possible that we may be able to learn a paramaterisation of a curve given empirical computation of its length. Alternatively, we could move away from parametrising entirely, representing a curve by its length over an interval.

8.4 Concluding Remarks

This chapter has revisited the main contributions of the thesis. We summarised the core additions to the literature, articulating strengths and weaknesses of our research. Finally, we described a multitude of research opportunities for future development.

Chapter 9

Conclusions

This thesis has clearly addressed the two Gaussian Process (GP) issues and achieved the aims identified at the start of this thesis: developing functional regression models with GPs and quantifying the arc length distribution of a GP.

The thesis highlights a lack of probabilistic non-linear methods for functional problems in Chapter 1. To address this gap, three models were introduced: the Gaussian Process Functional Index Model (GP-IND), Gaussian Process Functional Additive Model (GP-FAM) and Gaussian Process Functional Generalized Additive Model (GP-FGAM) where the inference and prediction schemes were specified in each. The GP models clearly outperformed established benchmarks on synthetic data sets and we also achieved competitive performance on a number of real world data applications. Furthermore, we retained the interpretability of the Functional Generalized Additive Model (FGAM) model.

The ability to extend functional regression models and obtain competitive results was clearly demonstrated, emphasising the under-utilised potential of GPs for functional regression problems.

We sought to quantify the arc length distribution of a GP. Investigating the one dimensional case led us to consider the arc length integrand. This insight proved

crucial to approximating the vector-valued case; with the obtained approximation aligning with high fidelity to the ground truth. We were able to quantify the arc length distribution for a prior and posterior GP. As foreshadowed the relationship depends on the choice of kernel function and hyperparameters; achieving the second aim of the thesis, quantifying the arc length distribution of a GP.

Future work lies in further understanding this approximation, and leveraging application opportunities.

Appendix A

Mathematical Identities

A.1 Gaussian Identities

We outline the two main properties of the Gaussian distributions (Rasmussen, 2006), which we write as usual as:

$$\mathcal{N}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{\det 2\pi\boldsymbol{\Sigma}}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right) \tag{A.1}$$

Arguably the two most important properties are that its marginal and conditional are Gaussian. That is, if we have:

$$p(\mathbf{x},\mathbf{y}) = \mathcal{N}\left(\begin{bmatrix}\mathbf{x}\\ \mathbf{y}\end{bmatrix};\begin{bmatrix}\boldsymbol{\mu}_1\\ \boldsymbol{\mu}_2\end{bmatrix},\begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{bmatrix}\right) \tag{A.2}$$

then the marginal and conditional distributions are given as:

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x};\boldsymbol{\mu}_1,\boldsymbol{\Sigma}_{11}) \tag{A.3}$$

$$p(\mathbf{x}|\mathbf{y}) = \mathcal{N}\big(\mathbf{x};\boldsymbol{\mu}_1+\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mu_2-\mathbf{x}_2),\boldsymbol{\Sigma}_{11}-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\big) \tag{A.4}$$

$$\tag{A.5}$$

A.2 The Nakagami Distribution

The Nakagami Distribution is

$$p(x;m,\Omega) = \frac{2m^m}{\Gamma(m)\Omega^m}x^{2m-1}\exp\left(-\frac{m}{\Omega}x^2\right), \tag{A.6}$$

$m \geq 1/2$ and $\Omega > 0$.

A.2.1 Moments

The first moment is:

$$\mathbb{E}[x] = \int_0^\infty x \frac{2m^m}{\Gamma(m)\Omega^m} x^{2m-1} \exp\left(-\frac{m}{\Omega}x^2\right) \mathrm{d}x \tag{A.7}$$

$$= \frac{2m^m}{\Gamma(m)\Omega^m} \int_0^\infty x^{2m} \exp\left(-\frac{m}{\Omega}\right) \mathrm{d}x \tag{A.8}$$

Changing variables, let $x = \sqrt{\Omega/m}\sqrt{x'}$, then $\mathrm{d}x = \sqrt{\Omega/m}‘\frac{1}{2}x'^{-1/2}\mathrm{d}x'$:

$$\mathbb{E}[x] = \frac{2m^m}{\Gamma(m)\Omega^m} \int_0^\infty \left(\frac{\Omega}{m}\right)^2 x'^m \exp\left(-x'\right) \sqrt{\frac{\Omega}{m}} \frac{1}{2} x'^{-1/2} \mathrm{d}x' \tag{A.9}$$

$$= \frac{1}{\Gamma(m)} \int_0^\infty x'^{m+\frac{1}{2}-1} \exp\left(-x'\right) \mathrm{d}x' \tag{A.10}$$

$$= \frac{\Gamma\left(m+\frac{1}{2}\right)}{\Gamma(m)} \tag{A.11}$$

The second moment is:

$$\mathbb{E}[x^2] = \int_0^\infty x^2 \frac{2m^m}{\Gamma(m)\Omega^m} x^{2m-1} \exp\left(-\frac{m}{\Omega}x^2\right) \mathrm{d}x \tag{A.12}$$

$$= \frac{2m^m}{\Gamma(m)\Omega^m} \int_0^\infty x^{2m+1} \exp\left(-\frac{m}{\Omega}\right) \mathrm{d}x \tag{A.13}$$

Changing variables, let $x = \sqrt{\Omega/m}\sqrt{x'}$, then $\mathrm{d}x = \sqrt{\Omega/m}‘\frac{1}{2}x'^{-1/2}\mathrm{d}x'$:

$$\mathbb{E}[x^2] = \frac{2m^m}{\Gamma(m)\Omega^m} \int_0^\infty \left(\frac{\Omega}{m}\right)^{\frac{1}{2}(2m+1)} x'^{m+\frac{1}{2}} \exp\left(-x'\right) \sqrt{\frac{\Omega}{m}} \frac{1}{2} x'^{-1/2} \mathrm{d}x' \tag{A.14}$$

$$= \frac{m^m}{\Gamma(m)\Omega^m} \left(\frac{\Omega}{m}\right)^{m+1} \int_0^\infty x'^{m+1-1} \exp\left(-x'\right) \mathrm{d}x' \tag{A.15}$$

$$= \frac{\Omega}{m} \frac{1}{\Gamma(m)} \Gamma(m+1) \tag{A.16}$$

$$= \frac{\Omega}{m} \frac{1}{\Gamma(m)} \Gamma(m)m \tag{A.17}$$

$$= \Omega \tag{A.18}$$

Therefore the variance is:

$$\mathbb{V}[x] = \Omega\left(1 - \frac{1}{m}\left(\frac{\Gamma(m+\frac{1}{2})}{\Gamma(m)}\right)\right) \tag{A.19}$$

A.2.2 Mixed Distribution

The distribution for a bi-variate Nakagami random variable is (Reig et al., 2002):

$$\begin{aligned} p_{W_1,W_2}(w_1,w_2) = 4(1-\rho)^{m_2} \sum_{k=0}^{\infty} \frac{(m_1)_k}{k!} \rho^k \left(\frac{m_1}{\Omega_1(1-\rho)}\right)^{m_1+k} \\ \times w_1^{2(m_1+k)-1} \frac{\exp(-m_1 w_1^2/(\Omega_1(1-\rho)))}{\Gamma(m_1+k)} \left(\frac{m_2}{\Omega_2(1-\rho)}\right)^{m_2+k} \\ \times w_2^{2(m_2+k)-1} \frac{\exp(-m_2 w_2^2/(\Omega_2(1-\rho)))}{\Gamma(m_2+k)} \\ \times {}_2F_1\left(m_2-m_1, m_2+k; \frac{m_2\rho}{\Omega_2(1-\rho)} w_2^2\right), \quad w_1, w_2 \geq 0. \end{aligned} \tag{A.20}$$

Where ${}_2F_1(\cdot,\cdot;\cdot)$ is the hypergeometric function. The moments are given as (Reig et al., 2002):

$$\mathbb{E}[W_1^n W_2^l] = \left(\frac{\Omega_1}{m_1}\right)^{n/2} \left(\frac{\Omega_2}{m_2}\right)^{l/2} \frac{\Gamma(m_1+n/2)\Gamma(m_2+l/2)}{\Gamma(m_1)\Gamma(m_2)} {}_2F_1\left(-\frac{n}{2}, -\frac{l}{2}, m_2 : \rho\right) \tag{A.21}$$

A.3 Kernel Derivatives

The one-dimensional derivatives of the commonly used covariance functions are derived and effective length scale $\sigma_{f'}^2$. We use the relation:

$$\sigma_{f'}^2 = -\frac{d^2}{\mathrm{d}\tau^2} k(\tau)\bigg|_{\tau=0}, \tag{A.22}$$

where $k(\tau)$ is the covariance function.

A.3.1 Squared Exponential

The Squared Exponential (SE) kernel is:

$$k(\tau) = \lambda^2 \exp\left(-\frac{\tau^2}{2\sigma^2}\right) \tag{A.23}$$

Taking derivatives:

$$\rho'(\tau) = -\frac{d^2}{\mathrm{d}\tau^2}\rho(\tau) \tag{A.24}$$
$$= -\frac{d^2}{\mathrm{d}t^2}\lambda^2 \exp\left(-\frac{\tau^2}{2\sigma^2}\right) \tag{A.25}$$
$$= \frac{d}{\mathrm{d}t}\lambda^2\frac{\tau}{\sigma^2}\exp\left(-\frac{\tau^2}{2\sigma^2}\right) \tag{A.26}$$
$$= \frac{\lambda^2}{\sigma^2}\left(1-\frac{\tau}{\sigma^2}\right)\exp\left(-\frac{\tau^2}{2\sigma^2}\right) \tag{A.27}$$
$$\tag{A.28}$$

Therefore;

$$\sigma^2_{f'} = -\frac{d^2}{d\tau^2}k(\tau)\bigg|_{\tau=0} = \frac{\lambda^2}{\sigma^2} \tag{A.29}$$

A.3.2 Matérn $\nu = \frac{3}{2}$

The Matérn32 (MAT32) kernel is:

$$k(\tau) = \lambda^2\left(1+\frac{\sqrt{3}}{\sigma}\tau\right)\exp\left(-\frac{\sqrt{3}}{\sigma}\tau\right) \tag{A.30}$$

Taking derivatives:

$$\frac{\mathrm{d}}{\mathrm{d}\tau}k(\tau) = \frac{\mathrm{d}}{\mathrm{d}\tau}\lambda^2\left(1+\frac{\sqrt{3}}{\sigma}\tau\right)\exp\left(-\frac{\sqrt{3}}{\sigma}\tau\right) \tag{A.31}$$
$$= \lambda^2\frac{\sqrt{3}}{\sigma}\exp\left(-\frac{\sqrt{3}}{\sigma}\tau\right) + \lambda^2\left(1+\frac{\sqrt{3}}{\sigma}\tau\right)\left(-\frac{\sqrt{3}}{\sigma}\right)\exp\left(-\frac{\sqrt{3}}{\sigma}\tau\right) \tag{A.32}$$
$$= -\tau\lambda^2\frac{3}{\sigma^2}\exp\left(-\frac{\sqrt{3}}{\sigma}\tau\right) \tag{A.33}$$
$$\frac{\mathrm{d}^2}{\mathrm{d}\tau^2}k(\tau) = \frac{\mathrm{d}}{\mathrm{d}\tau}\left(-\tau\lambda^2\frac{3}{\sigma^2}\exp\left(-\frac{\sqrt{3}}{\sigma}\tau\right)\right) \tag{A.34}$$
$$= -\frac{\sqrt{3}\lambda^2}{\sigma^3}\exp\left(-\frac{\sqrt{3}}{\sigma}\tau\right) + \lambda^2\frac{3\sqrt{3}}{\sigma^3}\tau\exp\left(-\frac{\sqrt{3}}{\sigma}\tau\right) \tag{A.35}$$
$$= \lambda^2\left(\frac{3\sqrt{3}\tau}{\sigma^3}-\frac{3}{\sigma^2}\right)\exp\left(-\frac{\sqrt{3}}{\sigma}\tau\right) \tag{A.36}$$

Thus:

$$\sigma^2_{f'} = -\frac{d^2}{d\tau^2}k(\tau)\bigg|_{\tau=0} = 3\frac{\lambda^2}{\sigma^2} \tag{A.37}$$

A.3.3 Matérn $\nu = \frac{5}{2}$

The Matérn52 (MAT52) kernel is:

$$k(x, x') = \lambda^2 \left(1 + \frac{\sqrt{5}}{\sigma}|x - x'| + \frac{5}{3\sigma^2}|x - x'|^2\right) \exp\left(-\frac{\sqrt{5}|x - x'|^2}{\sigma}\right) \tag{A.38}$$

Writing in terms of $\tau = |x - x'|$

$$k(\tau) = \lambda^2 \left(1 + \frac{\sqrt{5}}{\sigma}\tau + \frac{5}{3\sigma^2}\tau^2\right) \exp\left(-\frac{\sqrt{5}\tau}{\sigma}\right) \tag{A.39}$$

Taking derivatives:

$$\frac{\mathrm{d}}{\mathrm{d}\tau}k(\tau) = \frac{\mathrm{d}}{\mathrm{d}\tau}\lambda^2 \left(1 + \frac{\sqrt{5}}{\sigma}\tau + \frac{5}{3\sigma^2}\tau^2\right) \exp\left(-\frac{\sqrt{5}\tau}{\sigma}\right) \tag{A.40}$$

$$= \lambda^2 \left(\frac{\sqrt{5}}{\sigma} + \frac{10}{3\sigma^2}\tau\right) \exp\left(-\frac{\sqrt{5}\tau}{\sigma}\right) \tag{A.41}$$

$$+ \lambda^2 \left(1 + \frac{\sqrt{5}}{\sigma}\tau + \frac{5}{3\sigma^2}\tau^2\right) \left(-\frac{\sqrt{5}}{\sigma}\right) \exp\left(-\frac{\sqrt{5}\tau}{\sigma}\right) \tag{A.42}$$

$$= \lambda^2 \left(\frac{10}{3\sigma^2}\tau - \frac{5\tau}{\sigma^2} - \frac{5\sqrt{5}}{3\sigma^2}\tau^2\right) \exp\left(-\frac{\sqrt{5}\tau}{\sigma}\right) \tag{A.43}$$

$$\frac{\mathrm{d}}{\mathrm{d}\tau}\frac{\mathrm{d}}{\mathrm{d}\tau}k(\tau) = \frac{\mathrm{d}^2}{d\tau^2}\lambda^2 \left(\frac{10}{3\sigma^2}\tau - \frac{5\tau}{\sigma^2} - \frac{5\sqrt{5}}{3\sigma^2}\tau^2\right) \exp\left(-\frac{\sqrt{5}\tau}{\sigma}\right) \tag{A.44}$$

$$= \lambda^2 \left(\frac{10}{3\sigma^2} - \frac{5}{\sigma^2} - \frac{10\sqrt{5}}{3\sigma^2}\tau\right) \exp\left(-\frac{\sqrt{5}\tau}{\sigma}\right) \tag{A.45}$$

$$+ \lambda^2 \left(\frac{10}{3\sigma^2}\tau - \frac{5\tau}{\sigma^2} - \frac{5\sqrt{5}}{3\sigma^2}\tau^2\right) \left(-\frac{\sqrt{5}}{\sigma}\right) \exp\left(-\frac{\sqrt{5}\tau}{\sigma}\right) \tag{A.46}$$

$$= \lambda^2 \left(-\frac{5}{3\sigma^2} - \frac{5\sqrt{5}\tau}{3\sigma^2} + \frac{25}{3\sigma^3}\tau^2\right) \exp\left(-\frac{\sqrt{5}\tau}{\sigma}\right) \tag{A.47}$$

Thus:

$$\sigma_{f'}^2 = -\frac{d^2}{d\tau^2}k(\tau)\bigg|_{\tau=0} = \frac{5}{3}\frac{\lambda^2}{\sigma^2} \tag{A.48}$$

A.3.4 Rational Quadratic

The Rational Quadratic (RQ) kernel is:

$$k(\tau) = \sigma^2 \left(1 + \frac{\tau^2}{2\alpha l^2}\right)^{-\alpha} \tag{A.49}$$

Taking derivatives:

$$\frac{\mathrm{d}}{\mathrm{d}\tau}k(\tau) = \frac{\mathrm{d}}{\mathrm{d}\tau}\sigma^2\left(1+\frac{\tau^2}{2\alpha l^2}\right)^{-\alpha} \tag{A.50}$$

$$= -\sigma^2\alpha\left(1+\frac{\tau^2}{2\alpha l^2}\right)^{-\alpha-1}\frac{2\tau}{2\alpha l^2} \tag{A.51}$$

$$= -\frac{\tau\sigma^2}{l^2}\left(1+\frac{\tau^2}{2\alpha l^2}\right)^{-(\alpha+1)} \tag{A.52}$$

$$\frac{\mathrm{d}}{\mathrm{d}\tau}k(\tau) = -\frac{\sigma^2}{l^2}\left(1+\frac{\tau^2}{2\alpha l^2}\right)^{-(\alpha+1)} + \frac{\tau\sigma^2}{l^2}\frac{2\tau}{2\alpha l^2}(\alpha+1)\left(1+\frac{\tau^2}{2\alpha l^2}\right)^{-(\alpha+2)} \tag{A.53}$$

$$= -\frac{\sigma^2}{l^2}\left(1+\frac{\tau^2}{2\alpha l^2}\right)^{-(\alpha+1)} + \frac{\sigma^2\tau^2(\alpha+1)}{\alpha l^4}(\alpha+1)\left(1+\frac{\tau^2}{2\alpha l^2}\right)^{-(\alpha+2)} \tag{A.54}$$

Thus:

$$\sigma^2_{f'} = -\frac{d^2}{d\tau^2}k(\tau)\bigg|_{\tau=0} = \frac{\sigma^2}{l^2} \tag{A.55}$$

Bibliography

B. P. Abbott and et Al. Observation of Gravitational Waves from a Binary Black Hole Merger. *Physical Review Letters*, 116(6):061102, feb 2016. ISSN 0031-9007. doi: 10.1103/PhysRevLett.116.061102. URL `https://link.aps.org/doi/10.1103/PhysRevLett.116.061102`.

P. Abrahamsen. A review of Gaussian random fields and correlation functions. Technical report, Oslo, Norway, 1997.

M. Abramowitz, I. A. Stegun, and D. Miller. Handbook of Mathematical Functions With Formulas, Graphs and Mathematical Tables (National Bureau of Standards Applied Mathematics Series No. 55), 1965. ISSN 00218936.

S. Acharyya and J. Ghosh. Parameter Estimation of Generalized Linear Models without Assuming their Link Function. *Proceedings of the 18th International Conference on Artifical Intelligence and Statistics (AISTATS)*, 38, 2015.

A. M. Aguilera, M. Escabias, M. J. Valderrama, and M. C. Aguilera-Morillo. Functional Analysis of Chemometric Data. *Open Journal of Statistics*, 3:334–343, 2013. ISSN 2161-718X. doi: 10.4236/ojs.2013.35039. URL `http://dx.doi.org/10.4236/ojs.2013.35039http://www.scirp.org/journal/ojs`.

M. A. Alvarez, L. Rosasco, and N. D. Lawrence. Kernels for Vector-Valued Functions : A Review. pages 1–37, 2012.

D. Andre, M. Meiler, K. Steiner, C. Wimmer, T. Soczka-Guth, D. U. Sauer, H. Walz, T. Soczka-Guth, and D. U. Sauer. Characterization of high-power lithium-ion batteries by electrochemical impedance spectroscopy. I. Experimental investigation. *Journal of Power Sources*, 196(12):5334–5341, 2011. ISSN 03787753. doi: 10.1016/j.jpowsour.2010.12.102.

R. Barakat and E. Baumann. Mean and variance of the arc length of a Gaussian process on a finite interval. *International Journal of Control*, 12(3):377–383, 1970. ISSN 0020-7179. doi: 10.1080/00207177008931855. URL `http://www.tandfonline.com/doi/abs/10.1080/00207177008931855`.

P. C. Besse. Approximation spline de la prevision d ' un processus fonctionnel autoregressif d ' ordre 1 1 Introduction. *Source The Canadian Journal of Statistics / La Revue Canadienne de Statistique*, 24(4):1–25, 1991. URL `http://www.jstor.org/stable/3315328`.

G. E. P. Box and G. C. Tiao. *Bayesian Inference in Statistical Analysis*. 1992. ISBN 0471574287. doi: 10.1002/9781118033197.

G. L. Bretthorst. The near-irrelevance of sampling frequency distributions. pages 21–46, 1999.

P. J. Brown, T. Fearn, and M. Vannucci. Bayesian Wavelet Regression on Curves With Application to a Spectroscopic Calibration Problem. *Journal of the American Statistical Association*, 96(454):398–408, 2001. ISSN 0162-1459. doi: 10.1198/016214501753168118.

J. Cao and G. Fan. Functional Data Classification with Kernel-Induced Random Forests. *Biometrics*, pages 1–15, 2009. URL `http://people.stat.sfu.ca/{~}cao/Research/FunctionalDataClassification`.

H. Cardot and P. Sarda. Estimation in generalized linear models for functional data via penalized likelihood. *Journal of Multivariate Analysis*, 92(1):24–41, 2005. ISSN 0047259X. doi: 10.1016/j.jmva.2003.08.008.

H. Cardot, F. Ferraty, and P. Sarda. Functional linear model. *Statistics & Probability Letters*, 45(1):11–22, oct 1999. ISSN 01677152. doi: 10.1016/S0167-7152(99)00036-X. URL `http://linkinghub.elsevier.com/retrieve/pii/S016771529900036X`.

R. Caruana. Multitask Learning. *Machine Learning*, 28(1):41–75, 1997. ISSN 1573-0565. doi: 10.1023/A:1007379606734.

A. B. Chan and D. Dong. Generalized Gaussian process models. *Cvpr 2011*, pages 2681–2688, 2011. ISSN 1063-6919. doi: 10.1109/CVPR.2011.5995688. URL `http://visal.cs.cityu.edu.hk/static/pubs/conf/cvpr11-ggpm.pdf`.

D. Chen, P. Hall, and H. G. Müller. Single and multiple index functional regression models with nonparametric link. *Annals of Statistics*, 39(3):1720–1747, 2011. ISSN 00905364. doi: 10.1214/11-AOS882.

T. Choi, J. Q. Shi, and B. Wang. A Gaussian process regression approach to a single-index model. *Journal of Nonparametric Statistics*, 23(1):21–36, 2011. ISSN 1048-5252. doi: 10.1080/10485251003768019. URL `http://www.tandfonline.com/doi/abs/10.1080/10485251003768019`.

S. Corrsin and O. M. Phillips. Contour Length and Surface Area of Multiple-Valued Random Variables. *Journal of the Society for Industrial and Applied Mathematics*, 9(3):395–404, 1961.

S. Covo and A. Elalouf. A novel single-gamma approximation to the sum of independent gamma variables, and a generalization to infinitely divisible distributions. *Electronic Journal of Statistics*, 8(1):894–926, 2014. ISSN 19357524. doi: 10.1214/14-EJS914. URL `https://projecteuclid.org/download/pdfview{_}1/euclid.ejs/1403812157`.

C. M. Crainiceanu and A. J. Goldsmith. Bayesian Functional Data Analysis Using WinBUGS. *Journal of statistical software*, 32(11), jan 2010. ISSN 1548-7660.

M. P. Deisenroth and C. E. Rasmussen. PILCO: A Model-Based and Data-Efficient Approach to Policy Search. *Proceedings of the International Conference on Machine Learning*, pages 465–472, 2011. URL `http://mlg.eng.cam.ac.uk/pub/pdf/DeiRas11.pdf`.

D. Duvenaud, H. Nickisch, and C. E. Rasmussen. Additive Gaussian Processes. *Advances in Neural Information Processing Systems 24*, pages 1–9, 2011. URL `http://eprints.pascal-network.org/archive/00008445/`.

Y. Fan, N. Foutz, G. M. James, and W. Jank. Functional Response Additive Model Estimation with Online Virtual Stock Markets. pages 1–31, 2014.

Y. Fan, G. M. James, and P. Radchenko. Functional additive regression. *Annals of Statistics*, 43(5):2296–2325, 2015. ISSN 00905364. doi: 10.1214/15-A0S1346.

M. Febrero-Bande and M. Oviedo de la Fuente. Statistical computing in functional data analysis: the R package fda. usc. *Journal of Statistical Software*, 51(4):1–28, 2012. ISSN 1548-7660. doi: 10.18637/jss.v051.i04. URL `http://www.jstatsoft.org/v51/i04/paper`.

F. Ferraty and P. Vieu. *Nonparametric Functional Data Analysis - Theory and Practice*. 2006. ISBN 978-0387-30369-7. doi: 10.1007/0-387-36620-2.

F. Ferraty and P. Vieu. Additive prediction and boosting for functional data. *Computational Statistics and Data Analysis*, 53(4):1400–1413, feb 2009. ISSN 01679473. doi: 10.1016/j.csda.2008.11.023. URL `http://linkinghub.elsevier.com/retrieve/pii/S0167947308005628`.

F. Ferraty, A. Mas, and P. Vieu. Advances on nonparametric regression for functional variables. 2006. URL `http://arxiv.org/abs/math/0603084`.

F. Ferraty, A. Goia, E. Salinelli, and P. Vieu. Functional projection pursuit regression. *Test*, 22(2):293–320, 2013. ISSN 11330686. doi: 10.1007/s11749-012-0306-2.

S. Flaxman, A. G. Wilson, D. B. Neill, H. Nickisch, H. Org, A. J. Smola, and A. Org. Fast Kronecker Inference in Gaussian Processes with non-Gaussian Likelihoods. 2015. URL `http://proceedings.mlr.press/v37/flaxman15.pdf`.

G. Fubini. *Sugli Integrali Multipli*. 1907.

R. Garnett, M. A. Osborne, S. Reece, A. Rogers, and S. J. Roberts. Sequential Bayesian Prediction in the Presence of Changepoints and Faults. *The Computer Journal*, 53(9):1430–1446, feb 2010. ISSN 0010-4620. doi: 10.1093/comjnl/bxq003. URL `http://comjnl.oxfordjournals.org/cgi/doi/10.1093/comjnl/bxq003`.

Z. Ghahramani. Bayesian nonparametrics and the probabilistic approach to modelling. *Phil. Trans. R. Soc. A 1-27*, pages 1–27, 2011. doi: 10.1098/rspa.00000000.

Z. Ghahramani and C. E. Rasmussen. Bayesian monte carlo. *Advances in neural information processing systems*, (1), 2002. URL `http://machinelearning.wustl.edu/mlpapers/paper{_}files/AA01.pdf`.

M. N. Gibbs. *Bayesian Gaussian processes for regression and classification*. PhD thesis, University of Cambridge, 1997. URL `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.147.1130{&}rep=rep1{&}type=pdf`.

A. Girard and R. Murray-Smith. Learning a Gaussian Process Model with Uncertain Inputs. Technical Report c, 2003. URL `http://www.dcs.gla.ac.uk/{~}rod/publications/GirMur03-tr-144.pdf`.

J. Goldsmith, J. Bobb, C. M. Crainiceanu, B. Caffo, and D. Reich. Penalized Functional Regression. *Journal of Computational and Graphical Statistics*, 20(4):830–851, jan 2011a. ISSN 1061-8600. doi: 10.1198/jcgs.2010.10007.

J. Goldsmith, M. P. Wand, and C. Crainiceanu. Functional regression via variational bayes. *Electronic Journal of Statistics*, 5:572–602, 2011b. ISSN 19357524. doi: 10.1214/11-EJS619.

P. Goovaerts. *Geostatistics for natural resources evaluation*. Oxford University Press, 1997. ISBN 0195115384. URL `https://books.google.co.uk/books/about/Geostatistics{_}for{_}Natural{_}Resources{_}Eval.html?id=CW-7tHAaVR0C`.

R. B. Gramacy and H. Lian. Gaussian process single-index models as emulators for computer experiments. (Cv):1–23, 2010. URL `http://arxiv.org/abs/1009.4241`.

T. Hastie and R. Tibshirani. Generalized Additive Models. *Statistical Science*, 1(3):297–310, aug 1986. ISSN 0883-4237. doi: 10.1214/ss/1177013604. URL `http://projecteuclid.org/euclid.ss/1177013604http://www.jstor.org/discover/10.2307/2245459?uid=3738776{&}uid=2{&}uid=4{&}sid=21101370099757`.

T. J. Hastie and R. Tibshirani. Varying-coefficient Models. *Journal of the Royal Statistical Society*, 55(4):757–796, 1993. ISSN 0035-9246. doi: 10.2307/2345993. URL `https://www.jstor.org/stable/pdf/2345993.pdf?refreqid=excelsior:fdfbd89dc665261b12bf1f20dee4bfbf`.

S. Hauberg, O. Freifeld, and M. Black. A Geometric take on Metric Learning. *Neural Information Processing Systems*, (1):1–9, 2012. ISSN 10495258. URL `https://papers.nips.cc/paper/4539-a-geometric-take-on-metric-learning.pdf`.

P. Hennig and S. Hauberg. Probabilistic Solutions to Differential Equations and their Application to Riemannian Statistics. *Artificial Intelligence and Statistics (AISTATS)*, 33(1):11, 2013. ISSN 15337928. URL `http://www2.compute.dtu.dk/{~}sohau/papers/aistats2014/AISTATS2014{_}probODEs.pdfhttp://arxiv.org/abs/1306.0308`.

P. Hennig, M. A. Osborne, and M. Girolami. Probabilistic numerics and uncertainty in computations. *Proceedings. Mathematical, physical, and engineering sciences / the Royal Society*, 471(2179):20150142, 2015. ISSN 1364-5021.

doi: 10.1098/rspa.2015.0142. URL http://rspa.royalsocietypublishing.org/content/471/2179/20150142.

J. Hensman, N. O. Fusi, and N. D. Lawrence. Gaussian Processes for Big Data. *Uncertainty in Artificial Intelligence (UAI)*, 2013. URL http://www.auai.org/uai2013/prints/papers/244.pdf.

W. Hoffman. Statistical Methods in Radio Wave Propagation: Proceedings of a Symposium Held at the University of California, Los Angeles, June 1820, 1958. 1958.

T. Hofmann, B. Schölkopf, and A. J. Smola. Kernel methods in machine learning. *Annals of Statistics*, 36(3):1171–1220, 2008. ISSN 00905364. doi: 10.1214/009053607000000677.

H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417–441, 1933. ISSN 0022-0663. doi: 10.1037/h0071325. URL http://content.apa.org/journals/edu/24/6/417.

L. Isserlis. On a Formula for the Product-Moment Coefficient of any Order of a Normal Frequency Distribution in any Number of Variables. *Biometrika*, 12(1/2):134–139, 1918. ISSN 00063444. doi: 10.2307/2331932. URL http://www.jstor.org/stable/pdf/2331932.pdf?refreqid=excelsior{%}3A3eb6f8c650450b62f6e8f24180ad60cchttp://biomet.oxfordjournals.org/content/12/1-2/134.short{%}5Cnhttp://www.jstor.org/stable/2331932?origin=crossref.

G. James. Generalized Linear Models with Functional Predictor Variables. *Journal of the Royal Statistical Society, Series B*, 64(2):411–432, 2002.

G. M. James and B. W. Silverman. Functional Adaptive Model Estimation. *Journal of the American Statistical Association*, (1993):1–28, 2005. URL http://amstat.tandfonline.com/doi/abs/10.1198/016214504000001556.

G. M. James, J. Wang, and J. Zhu. Functional linear regression that's interpretable 1. *Annals of Statistics*, 37(5 A):2083–2108, 2009. ISSN 00905364. doi: 10.1214/08-AOS641.

E. T. Jaynes. *Probability Theory: The Logic of Science.* Cambridge University Press, 2003. ISBN 0521592712. doi: 10.1007/BF02985800. URL http://medcontent.metapress.com/index/A65RM03P4874243N.pdf.

C.-R. Jiang and J.-L. Wang. Functional single index models for longitudinal data. 39(1):362–388, 2011. ISSN 0090-5364. doi: 10.1214/10-AOS845. URL http://arxiv.org/abs/1103.1726.

N. L. Johnson, S. Kotz, and N. Balakrishnan. *Continuous univariate distributions*. Wiley, 1994. ISBN 0471584940.

A. G. Journel and C. J. Huijbregts. *Mining geostatistics*. Academic Press, 1978. ISBN 0123910501.

H. Kadri, E. Duflos, P. Preux, S. Canu, and M. Davy. Nonlinear functional regression: a functional RKHS approach. pages 374–380, 2010. ISSN 15324435. URL http://eprints.pascal-network.org/archive/00007035/.

J. H. Kalivas. Two data sets of near infrared spectra. *Chemometrics and Intelligent Laboratory Systems*, 37(2):255–259, jun 1997. ISSN 01697439. doi: 10.1016/S0169-7439(97)00038-5. URL http://www.sciencedirect.com/science/article/pii/S0169743997000385.

O. Kallenberg. *Probabilistic Symmetries and Invariance Principles*. Springer, 2005.

K. Kersting, C. Plagemann, P. Pfaff, and W. Burgard. Most Likely Heteroscedastic Gaussian Process Regression. In *24th International Conference on Machine Learning (ICML 2007)*, pages 393–400, 2007. ISBN 978-1-59593-793-3. doi: 10.1145/1273496.1273546. URL http://people.csail.mit.edu/kersting/papers/kersting07icml{_}mlHetGP.pdf.

Y.-L. Kom Samo and S. J. Roberts. String and Membrane Gaussian Processes. *Journal of Machine Learning Research*, 17:1–87, 2016. ISSN 15337928.

A. K. Kuchibhotla and R. K. Patra. {simest}: Single Index Model Estimation with Constraints on Link Function, 2017. URL https://cran.r-project.org/package=simest.

A. K. Kuchibhotla, R. K. Patra, and B. Sen. Efficient Estimation in Convex Single Index Models. *arXiv*, 2016. URL http://stat.ufl.edu/{~}rohitpatra/PapersandDraft/Cvx{_}SIMpaper.pdf.

M. Kuss and C. E. Rasmussen. Assessing Approximations for Gaussian Process Classification. *Advances in Neural Information Processing Systems 18: Proceedings of the 2005 Conference*, pages 699–706, 2006. ISSN

10495258. URL http://www.is.tuebingen.mpg.de/fileadmin/user{_}upload/files/publications/NIPS2005{_}0163{_}3530[1].pdf.

P. S. Laplace. Sur L ' Application Du Calcul Des Probabilites A La Philosophie Naturelle. *Connaissance des Temps*, pages 361–377, 1815.

N. D. Lawrence. Gaussian Process Latent Variable Models for Visualisation of High Dimensional Data. *Advances in Neural Information Processing Systems 16*, 16 (329-336):3, 2004.

Q. V. Le, A. J. Smola, and S. Canu. Heteroscedastic Gaussian process regression. *Proceedings of the 22nd International Conference on Machine Learning (ICML)*, pages 489–496, 2005. ISSN 01697439. doi: 10.1145/1102351.1102413. URL https://cs.stanford.edu/{~}quocle/LeSmoCan05.pdfhttp://dl.acm.org/citation.cfm?id=1102413.

B. Li and B. D. Marx. Sharpening P-spline signal regression. *Statistical Modelling*, 8 (4):367–383, 2008. ISSN 1471-082X. doi: 10.1177/1471082X0800800403.

Y. Li and T. Hsing. On rates of convergence in functional linear regression. *Journal of Multivariate Analysis*, 98(9):1782–1804, 2007. ISSN 0047259X. doi: 10.1016/j.jmva.2006.10.004.

D. J. C. MacKay. Bayesian interpolation. *Neural computation*, 4(3):415–447, 1992.

D. J. C. MacKay. *Information theory, inference and learning algorithms*. Cambridge University Press, 2003. ISBN 0521642981.

A. Maity. Nonparametric functional concurrent regression models. *Wiley Interdisciplinary Reviews: Computational Statistics*, 9(2):1–11, 2017. ISSN 19390068. doi: 10.1002/wics.1394.

A. Majumdar and A. E. Gelfand. Multivariate spatial modeling for geostatistical data using convolved covariance functions. *Mathematical Geology*, 39(2):225–245, may 2007. ISSN 08828121. doi: 10.1007/s11004-006-9072-6. URL http://link.springer.com/10.1007/s11004-006-9072-6.

R. Marchant and F. Ramos. Bayesian Optimisation for informative continuous path planning. In *Proceedings - IEEE International Conference on Robotics and Automation*, pages 6136–6143. IEEE, may 2014. ISBN 9781479936847. doi: 10.1109/ICRA.2014.6907763. URL http://ieeexplore.ieee.org/document/6907763/.

A. M. Mathai and S. B. Provost. *Quadratic forms in random variables: theory and applications*. Marcel Dekker Inc., 1992. ISBN 0824786912.

A. G. d. G. Matthews, M. van der Wilk, T. Nickson, K. Fujii, A. Boukouvalas, P. León-Villagrá, Z. Ghahramani, and J. Hensman. GPflow: A Gaussian process library using TensorFlow. *Journal of Machine Learning Research*, 18:1–6, 2016. ISSN 15337928. URL `http://www.jmlr.org/papers/volume18/16-537/16-537.pdfhttp://arxiv.org/abs/1610.08733`.

A. Mchutchon and C. E. Rasmussen. Gaussian Process Training with Input Noise. *Advances in Neural Information Processing Systems*, pages 1341–1349, 2011.

M. W. McLean, G. Hooker, A.-M. Staicu, F. Scheipl, and D. Ruppert. Functional Generalized Additive Models. *Journal of computational and graphical statistics : a joint publication of American Statistical Association, Institute of Mathematical Statistics, Interface Foundation of North America*, 23(1):249–269, 2012. ISSN 1061-8600. doi: 10.1080/10618600.2012.729985. URL `http://www.ncbi.nlm.nih.gov/pubmed/24729671`.

M. W. McLean, F. Scheipl, G. Hooker, S. Greven, and D. Ruppert. Bayesian Functional Generalized Additive Models with Sparsely Observed Covariates. *arXiv*, page 36, 2014. URL `https://arxiv.org/pdf/1305.3585.pdfhttp://arxiv.org/abs/1305.3585`.

J. Mercer. Functions of positive and negative type and their conection with the theory of integral equations. *Philosophical Transactions of the royal society, London*, 209(441-458), 1909. URL `http://rsta.royalsocietypublishing.org/content/209/441-458/415`.

D. Middleton. *An Introduction to Statistical Communication Theory*. IEEE Press, 1960. ISBN 0780311787.

T. Mikosch and O. Kallenberg. Foundations of Modern Probability. *Journal of the American Statistical Association*, 93(443):1243, sep 1998. ISSN 01621459. doi: 10.2307/2669881. URL `http://www.jstor.org/stable/2669881?origin=crossref`.

I. Miller and J. E. Freund. Expected Arc Length of a Gaussian Process on a Finite Interval. *Journal of the Royal Statistical Society. Series B (Methodology)*, 18(2): 257–258, 1956.

T. P. Minka. Expectation Propagation for Approximate Bayesian Inference. *Uncertainty in Artificial Intelligence (UAI)*, 17(2):362–369, 2001. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.86.1319{&}rep=rep1{&}type=pdf.

J. S. Morris. Functional Regression. *Annual Review of Statistics and Its Application*, 2(1):321–359, 2015. ISSN 2326-8298. doi: 10.1146/annurev-statistics-010814-020413. URL http://www.annualreviews.org/doi/abs/10.1146/annurev-statistics-010814-020413.

H. G. Müller and U. Stadtmüller. Generalized functional linear models. *Annals of Statistics*, 33(2):774–805, 2005. ISSN 00905364. doi: 10.1214/009053604000001156.

H.-G. Müller and F. Yao. Functional Additive Models. *Journal of the American Statistical Association*, 103(484):1534–1544, 2008. ISSN 0162-1459. doi: 10.1198/016214508000000751.

H.-G. Muller, Y. Wu, and F. Yao. Continuously additive models for nonlinear functional regression. *Biometrika*, 100(3):607–622, 2013. ISSN 0006-3444. doi: 10.1093/biomet/ast004.

A. O'Hagan. Bayeshermite quadrature. *Journal of Statistical Planning and Inference*, 1991. URL http://www.sciencedirect.com/science/article/pii/037837589190002V.

A. O'Hagan. Some Bayesian Numerical Analysis, 1992.

B. G. Osborne, T. Fearn, A. R. Miller, and S. Douglas. Application of near infrared reflectance spectroscopy to the compositional analysis of biscuits and biscuit doughs. *Journal of the Science of Food and Agriculture*, 35(1):99–105, jan 1984. ISSN 10970010. doi: 10.1002/jsfa.2740350116. URL http://doi.wiley.com/10.1002/jsfa.2740350116.

M. A. Osborne, S. J. Roberts, A. Rogers, S. D. Ramchurn, and N. R. Jennings. Towards Real-Time Information Processing of Sensor Network Data Using Computationally Efficient Multi-output Gaussian Processes. *2008 International Conference on Information Processing in Sensor Networks (ipsn 2008)*, pages 109–120, apr 2008. doi: 10.1109/IPSN.2008.25. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4505467.

C. J. Paciorek and M. J. Schervish. Nonstationary covariance functions for Gaussian process regression. *Advances in Neural Information Processing Systems 16: Proceedings of the 2003 Conference*, page 273, 2004. URL papers2://publication/uuid/8FEC5270-3B33-414F-9ED9-FCFAA79A56F4.

A. Papoulis. Probability, random variables, and stochastic processes, 1965. ISSN 0036-1445.

J. Pitman. *Probability*. Springer-Verlag, 1993.

J. Quinonero Candela and C. E. Rasmussen. A Unifying View of Sparse Approximate Gaussian Process Regression. 6:1939–1959, 2005. ISSN 1533-7928. URL http://eprints.pascal-network.org/archive/00002632/.

P. Radchenko, X. Qiao, and G. M. James. Index Models for Sparsely Sampled Functional Data. *Journal of the American Statistical Association*, 110(510):824–836, 2015. ISSN 1537274X. doi: 10.1080/01621459.2014.931859. URL http://www-bcf.usc.edu/{~}gareth/research/SIMFE.pdf.

A. Rahimi and B. Recht. Random features for large-scale kernel machines. *Advances in neural information ...*, (1):1–8, 2007. ISSN 0033-6599. doi: 10.1.1.145.8736.

J. Ramsay and C. Dalzell. Some tools for functional data analysis. *Journal of the Royal Statistical Society*, 53(3):539–572, 1991.

J. O. Ramsay and B. W. Silverman. *Functional Data Analysis (2nd Ed)*. New York: Springer, 2005a.

J. O. Ramsay and B. W. Silverman. *Functional data analysis*. 2005b. ISBN 9780387400808. doi: 10.1007/978-0-387-98135-2. URL http://www.knovel.com/knovel2/Toc.jsp?BookID=1804.

C. Rasmussen and C. K. Williams. *Gaussian Processes for Machine Learning*. the MIT Press, apr 2006. ISBN ISBN-10 0-262-18253-X. URL http://www.gaussianprocess.org/gpml/chapters/.

C. E. Rasmussen. Gaussian processes in machine learning. *International journal of neural systems*, 14(2):69–106, 2006. ISSN 0129-0657. doi: 10.1142/S0129065704001899.

J. Reig, L. Rubio, and N. Cardona. Bivariate Nakagami-m distribution with arbitrary fading parameters. *Electronics Letters*, 38(25):1715–1717, 2002.

M. Richey. The evolution of Markov chain Monte Carlo methods. *American Mathematical Monthly*, 117(5):383–413, 2010. ISSN 00029890. doi: 10.4169/000298910X485923.

J. Riihimäki and A. Vehtari. Gaussian processes with monotonicity information. *Journal of Machine Learning Research*, 9:645–652, 2010. ISSN 15324435. URL `http://machinelearning.wustl.edu/mlpapers/paper{_}files/AISTATS2010{_}RiihimakiV10.pdf`.

J. Riihimäki and A. Vehtari. Laplace approximation for logistic gaussian process density estimation and regression. *Bayesian Analysis*, 9(2):425–448, 2014. ISSN 19316690. doi: 10.1214/14-BA872. URL `https://projecteuclid.org/download/pdfview{_}1/euclid.ba/1401148315`.

M. Schober, N. Kasenburg, A. Feragen, P. Hennig, and S. Hauberg. Probabilistic shortest path tractography in DTI using Gaussian process ODE solvers. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 8675 LNCS, pages 265–272, 2014. ISBN 9783319104423. doi: 10.1007/978-3-319-10443-0_34. URL `http://www2.compute.dtu.dk/{~}sohau/papers/miccai2014/MICCAI2014.pdf`.

B. Scholkopf and A. J. Smola. *Learning with kernels : support vector machines, regularization, optimization, and beyond*. MIT Press, 2002. ISBN 9780262194754. URL `https://mitpress.mit.edu/books/learning-kernels`.

R. Sheth, Y. Wang, and R. Khardon. Sparse Variational Inference for Generalized Gaussian Process Models. *International Conference on Machine Learning*, 37, 2015. URL `http://proceedings.mlr.press/v37/sheth15.pdf`.

J. Q. Shi, B. Wang, R. Murray-Smith, and D. M. Titterington. Gaussian process functional regression modeling for batch data. *Biometrics*, 63(3):714–723, 2007. ISSN 0006341X. doi: 10.1111/j.1541-0420.2007.00758.x.

J. Q. Shi, B. Wang, E. J. Will, and R. M. West. Mixed-effects Gaussian process functional regression models with application to dose-response curve prediction. *Statistics in Medicine*, 31(26):3165–3177, 2012. ISSN 02776715. doi: 10.1002/sim.4502.

L. J. Slater. *Confluent hypergeometric functions.* Cambridge University Pres, 2010. ISBN 1108013244. URL https://books.google.co.uk/books/about/Confluent{_}Hypergeometric{_}Functions.html?id=IA7ekQEACAAJ{&}redir{_}esc=y.

E. Snelson and Z. Ghahramani. Sparse Gaussian Processes using Pseudo-inputs. *Advances in Neural Information Processing Systems 18*, pages 1257–1264, 2006. ISSN 1049-5258. doi: 10.1.1.60.2209. URL http://papers.nips.cc/paper/2857-sparse-gaussian-processes-using-pseudo-inputs.pdf.

J. Snoek, H. Larochelle, and R. P. Adams. Practical Bayesian Optimization of Machine Learning Algorithms. *NIPS*, 2012. URL http://papers.nips.cc/paper/4522-practical-bayesian-optimization-of-machine-learning-algorithms.pdf.

E. Solak, R. Murray-Smith, W. Leithead, D. Leith, and C. Rasmussen. Derivative observations in Gaussian process models of dynamic systems. *Nips 15*, page 8, 2002. ISSN 1049-5258. URL https://papers.nips.cc/paper/2287-derivative-observations-in-gaussian-process-models-of-dynamic-systems.pdf.

M. L. Stein. *Integration of Random Fields*. 1999. doi: 10.1007/978-1-4612-1494-6_5. URL http://link.springer.com/10.1007/978-1-4612-1494-6{_}5.

M. Titsias and N. Lawrence. Bayesian Gaussian Process Latent Variable Model. *Artificial Intelligence*, 9:844–851, 2010. URL http://eprints.pascal-network.org/archive/00006343/.

M. K. Titsias and M. Lázaro-Gredilla. Variational Inference for Mahalanobis Distance Metrics in Gaussian Process Regression. *Advances in Neural Information Processing Systems*, pages 279–287, 2013. ISSN 10495258.

A. Tosi, S. Hauberg, A. Vellido, and N. D. Lawrence. Metrics for Probabilistic Geometries. *Uncertainty in Artificial Intelligence*, page 800, 2014. URL http://arxiv.org/abs/1411.7432{%}5Cnhttp://auai.org/uai2014/proceedings/individuals/171.pdf.

B. Wang and J. Q. Shi. Generalized Gaussian Process Regression Model for Non-Gaussian Functional Data. *Journal of the American Statistical Association*,

(March):00–00, 2014. ISSN 0162-1459. doi: 10.1080/01621459.2014.889021. URL http://www.tandfonline.com/doi/abs/10.1080/01621459.2014.889021.

B. Wang, T. Chen, and A. Xu. Gaussian process regression with functional covariates and multivariate response. *Chemometrics and Intelligent Laboratory Systems*, 163 (February):1–6, 2017. ISSN 18733239. doi: 10.1016/j.chemolab.2017.02.001. URL http://dx.doi.org/10.1016/j.chemolab.2017.02.001.

J. Wang, D. Fleet, and A. Hertzmann. Gaussian process dynamical models. *Advances in Neural Information Processing Systems*, pages 1441–1448, 2005. ISSN 0162-8828. doi: 10.1109/TPAMI.2007.1167.

J.-L. Wang, J.-M. Chiou, and H.-G. Müller. Functional Data Analysis. *Annual Review of Statistics and Its Application*, 3(1):257–295, 2016. ISSN 2326-8298. doi: 10.1146/annurev-statistics-041715-033624. URL http://www.annualreviews.org/doi/pdf/10.1146/annurev-statistics-041715-033624http://arxiv.org/abs/1507.05135{%}5Cnhttp://www.annualreviews.org/doi/10.1146/annurev-statistics-041715-033624.

X. Wang and D. Ruppert. Optimal Prediction in an Additive Functional Model. *arXiv preprint arXiv:1301.4954*, pages 1–30, 2013. ISSN 10170405. doi: 10.5705/ss.2013.074.

T. D. Wickramarachchi, C. Gallagher, and R. Lund. Arc length asymptotics for multivariate time series. *Applied Stochastic Models in Business and Industry*, 31 (2):264–281, 2015. ISSN 15264025. doi: 10.1002/asmb.2030.

A. Wilson and R. Adams. Gaussian process kernels for pattern discovery and extrapolation. *Proceedings of the 30th ...*, 28(3):1067–1075, 2013.

A. G. Wilson and H. Nickisch. Kernel Interpolation for Scalable Structured Gaussian Processes (KISS-GP). *International Conference on Machine Learning*, 37:1–19, 2015. doi: 10.1007/978-0-8176-8172-2_11.

J. Xu, C. C. Mi, B. Cao, and J. Cao. A new method to estimate the state of charge of lithium-ion batteries based on the battery impedance model. *Journal of Power Sources*, 233(September 2017):277–284, 2013. ISSN 03787753. doi: 10.1016/j.jpowsour.2013.01.094.

F. Yao and H. G. Müller. Functional quadratic regression. *Biometrika*, 97(1):49–64, 2010. ISSN 00063444. doi: 10.1093/biomet/asp069.

F. Yao, H.-G. Muller, and J.-l. Wang. Functional Linear Regression Analysis for Longitudinal Data. *The Annals of Statictics*, 33(6):2873–2903, 2004. ISSN 0090-5364. doi: 10.1214/009053605000000660.

F. Yao, H. G. Müller, and J. L. Wang. Functional Data Analysis for Sparse Longitudinal Data. *Journal of American Statistical Association*, 100(470):577–590, 2005. ISSN 0162-1459. doi: 10.1198/016214504000001745.

M. Yuan and T. T. Cai. A reproducing kernel Hilbert space approach to functional linear regression. *The Annals of Statistics*, 38(6):3412–3444, 2010. ISSN 0090-5364. doi: 10.1214/09-AOS772.

H. Zhu, C. K. I. Williams, R. Rohwer, and M. Morciniec. Gaussian Regression and Optimal Finite Dimensional Linear Models. *Neural Networks and Machine Learning*, pages 167–184, 1998. URL `http://www.ncrg.aston.ac.uk/`.