Non-parametric Bayesian Time Series With Gaussian Processes

Iiro Tiihonen
University of Helsinki, Faculty of Science
Supervisor: Petteri Piiroinen

November 27, 2020

Tiedekunta/Osasto – Fakultet/Sektion – Faculty Matemaattis-luonnontieteellinen tiedekunta		
Tekijä – Författare – Author Iiro Tiihonen		
Työn nimi – Arbetets titel – Title Non-Parametric Bayesian Time Series With Gaussian Processes		
Oppiaine – Läroämne – Subject Soveltava matematiikka		
Työn laji – Arbetets art – Level Pro gradu -tutkielma	Aika – Datum – Month and year 11/2020	Sivumäärä– Sidoantal – Number of pages 82

Tiivistelmä – Referat – Abstract

Työni aihe on Gaussisten prosessien (Gp) soveltaminen aikasarjojen analysointiin. Erityisesti lähestyn aikasarjojen analysointia verrattain harvinaisen sovellusalan, historiallisten aikasarja-aineistojen analysoinnin näkökulmasta. Bayesilaisuus on tärkeä osa työtä: parametreja itsessään kohdellaan satunnaismuuttujina, mikä vaikuttaa sekä mallinnusongelmien muotoiluun että uusien ennusteiden tekemiseen työssä esitellyillä malleilla.

Työni rakentuu paloittain. Ensin esittelen Gp:t yleisellä tasolla, tilastollisen mallinnuksen työkaluna. Gp:eiden keskeinen idea on, että Gp-prosessin äärelliset osajoukot noudattavat multinormaalijakaumaa, ja havaintojen välisiä yhteyksiä mallinnetaan ydinfunktiolla (kernel), joka samaistaa havaintoja niihin liittyvien selittäjien ja parametriensa funktiona. Oikeanlaisen ydinfunktion valinta ja datan suhteen optimoidut parametrit mahdollistavat hyvinkin monimutkaisten ja heikosti ymmärrettyjen ilmiöiden mallintamisen Gp:llä. Esittelen keskeiset tulokset, jotka mahdollistavat sekä GP:n sovittamisen aineistoon että sen käytön ennusteiden tekemiseen ja mallinnetun ilmiön alatrendien erittelyyn.

Näiden perusteiden jälkeen siirryn käsittelemään sitä, miten GP-malli formalisoidaan ja sovitetaan, kun lähestymistapa on Bayesilainen. Käsittelen sekä eri sovittamistapojen vahvuuksia ja heikkouksia, että mahdollisuutta liittää Gp osaksi laajempaa tilastollista mallia. Bayesilainen lähestymistapa mahdollistaa mallinnettua ilmiötä koskevan ennakkotiedon syöttämisen osaksi mallin formalismia parametrien priorijakaumien muodossa. Lisäksi se tarjoaa systemaattisen, todennäköisyyksiin perustuvan tavan puhua sekä ennakko-oletuksista että datan jälkeisistä parametreihin ja mallinnetun ilmiön tuleviin arvoihin liittyvistä uskomuksista.

Seuraava luku käsittelee aikasarjoihin erityisesti liittyviä Gp-mallintamisen tekniikoita. Erityisesti käsittelen kolmea erilaista mallinnustilannetta: ajassa tapahtuvan Gp:n muutoksen, useammasta eri alaprosessista koostuvan Gp:n ja useamman keskenään korreloivan Gp:n mallintamista. Tämän käsittelyn jälkeen työn teoreettinen osuus on valmis: aikasarjojen konkreettinen analysointi työssä esitellyillä työkaluilla on mahdollista.

Viimeinen luku käsittelee historiallisten ilmiöiden mallintamista aiemmissa luvuissa esitellyillä tekniikoilla. Luvun tarkoitus on ensisijaisesti esitellä lyhyesti useampi potentiaalinen sovelluskohde, joita on yhteensä kolme. Ensimmäinen luvussa käsitelty mahdollisuus on usein vain repalaisesti havaintoja sisältävien historiallisten aikasarja-aineistojen täydentäminen GP-malleista saatavilla ennusteilla. Käytännön tulokset korostivat tarvetta vahvoille prioreille, sillä historialliset aikasarjat ovat usein niin harvoja, että mallit ovat valmiita hylkäämän havaintojen merkityksen ennustamisessa. Toinen esimerkki käsittelee historiallisia muutoskohtia, esimerkkitapaus on Englannin sisällissodan aikana äkillisesti räjähtävä painotuotteiden määrä 1640-luvun alussa. Sovitettu malli onnistuu päättelemään sisällissodan alkamisen ajankohdan. Viimeisessä esimerkissä mallinnan painotuotteiden määrää per henkilö varhaismodernissa Englannissa, käyttäen ajan sijaan selittäjinä muita ajassa kehittyviä muuttujia (esim. urbanisaation aste), jotka tulkitaan alaprosesseiksi. Tämänkin esimerkin tekninen toteutus onnistui, mikä kannustaa sekä tilastollisesti että historiallisesti kattavampaan analyysiin.

Kokonaisuutena työni sekä esittelee että demonstroi Gp-lähestymistavan mahdollisuuksia aikasarjojen analysoinnissa. Erityisesti viimeinen luku kannustaa jatkokehitykseen historiallisten ilmiöiden mallintamisen uudella sovellusalalla.

Avainsanat – Nyckelord – Keywords Gaussian Process, time series, aikasarjat, Gaussiset prosessit, historia	
Säilytyspaikka – Förvaringställe – Where deposited Kumpulan kampuksen kirjasto	
Muita tietoja – Övriga uppgifter – Additional information	

Contents

Chapter 1

Introduction

1.1 Motivation

Modeling of time series is a significant part of statistical research literature. The number of possible models one can build for a given temporal data set is almost endless. However, this does not solve the question of ”which model to fit?”. Often we only have a limited or even completely lacking understanding of the process that generates our observations, and the reasonable option would be a data-driven approach that does not make strong assumptions about the nature of the phenomenon behind the time series. A significant motivation for this thesis was our need for a modeling framework that could handle time series that are related to non-modern historical processes.

The construction introduced in this thesis is one of the co-called non-parametric methods to time series modeling. This does not mean that the model has no parameters, but the parameters function in a different way than in parametric approach. Whereas in parametric approach we fit the model in a narrow domain of family of models by optimising the parameters, our non-parametric approach

makes it possible to fit the model over a wider function space described only by loose properties like ”observations close to each other have similar values”. The tool of our choice that makes it possible to articulate such vague properties and infer them from the data is called Gaussian Process (GP, and for Gaussian Processes GPs). It builds on the properties of the multivariate Gaussian distribution and its conditional distribution. The GP is introduced in the next chapter of this thesis.

The Bayesian part comes from the probabilistic treatment of the (hyper)parameters of GPs. GPs are usually defined so that the covariance (and sometimes also the mean) is defined by a parametrised function, and these (hyper)parameters of the functions(s) have prior distributions that reflect our knowledge about the phenomenon being modeled. It may sound odd to make assumptions about the parameter values as the motivation for the non-parametric approach is to model relatively unknown processes, but as we'll see, the combination is fruitful. Bayesian approach also allows the construction and fitting of more complicated hierarchical models, in which GP is only one component among others.

The structure of the thesis is to first introduce the GP as a tool of non-parametric modeling, integrate it to the Bayesian framework and then to consider the task of prediction with and fitting of the full Bayesian model. After the general basics of GP modeling have been established, we focus on extensions and modelling approaches from the point of view of time series modeling. The last chapter before conclusions demonstrates the usage of the Bayesian GP framework in the modeling of historical processes. The aim of this endeavour is to bring together research literature about GPs from the point of view of one of the application domains (time series modeling), and especially to consider their potential in my own application field, computational history.

Chapter 2

Gaussian Process

2.1 Prerequisites: Linear Algebra, Miscellaneous Mathematics and Notations

Before moving to the main elements of GP modeling, some preceding terminology and results need to be established. We still assume that the reader is familiar with the very basics of linear algebra, probability calculus and ideas of statistical inference.

2.1.1 Linear Algebra

First thing to note is that in the entire thesis our matrices are real matrices. Then we establish how we write the determinant, as this varies somewhat in the literature.

Notation 2.1.1. *We mark the determinant of a matrix by* $|\cdot|$. *For example, the determinant of matrix* X *is* $|X|$

A similar small clarification is the definition and notation of trace.

Notation 2.1.2. *Let X be $N \times N$ matrix. The trace of the matrix is defined as the sum of its diagonal values, formally:* $\mathrm{tr}(X) = \sum_{i=1}^{N} x_{ii}$.

Terminology and basic results regarding definite matrices are fundamental for this thesis. Many of the most important matrices needed later in this thesis require some version of definiteness. We will return to these terms many times.

Definition 2.1.1. *Symmetric matrix (and only symmetric matrix) $A \in N \times N$ is positive definite it holds that $x^T Ax > 0 \; \forall x \in \mathrm{R}^N \setminus \{0\}$.*

Definition 2.1.2. *Symmetric matrix (and only symmetric matrix) $A \in N \times N$ is positive semi-definite it holds that $x^T Ax \geq 0 \; \forall x \in \mathrm{R}^N \setminus \{0\}$.*

There are two important reasons we focus on positive (semi)definite matrices. The first reason for the emphasis is explained by the following result.

Theorem 2.1.1. *If $N \times N$ matrix A is positive definite, then A is non-singular.*

Proof. Lets prove by assuming a positive definite matrix that is singular. Symmetric matrix is singular if and only if 0 is one of its eigenvalues[1] Now it must hold for some $x \in R^N \setminus \{0\}$ that

$$Ax = 0 \Leftrightarrow x^T Ax = 0$$

Yet we know that this breaks the definition of positive definite matrix, A must be non-singular. □

As non-singular matrices are invertible, we can rely on the inverse of positive definite matrices existing. The inverse of a matrix is required in many occasions, especially we need it for the density function of a multivariate Gaussian distribution that we discuss in the next chapter. Positive

[1] Very concise proof: A singular $\Leftrightarrow det(A) = 0 \Leftrightarrow det(A - 0I_{N\times N}) = 0 \Leftrightarrow 0$ is eigenvalue of A

semi-definiteness is even more fundamental for Gaussian random variables, which is based on the following theorem.[2]

Theorem 2.1.2. *Let A be a symmetric $N \times N$ matrix. Now A is positive semi-definite if and only if there exists $M \times N$ matrix B such that $A = B^T B$*

We will see that this property is needed for our definition of a multivariate Gaussian distribution.

Another property some matrices, for example all positive semi-definite matrices, have is the eigen-decomposition, which is very useful in matrix-related proofs. The definition is presented alongside with the following theorem:

Theorem 2.1.3. *Let A be a symmetric $N \times N$ matrix, which has eigenvalues $\lambda_1, \ldots, \lambda_N$ and corresponding eigenvectors $q_1, \ldots, q_N$. Now it holds that $A = Q\Lambda Q^T$, where $Q = (q_1, \ldots, q_N)$ and $\Lambda = diag(\lambda_1, \ldots, \lambda_N)$[3]. $Q\Lambda Q^T$ is called the eigendecomposition of A.*

Sometimes its handy to write the eigendecomposition in the following manner (with the notation of the previous theorem)

$$A = Q\Lambda Q^T = \sum_{i=1}^{N} \lambda_i q_i q_i^T .$$

In addition to the normal matrix multiplication, we sometimes use Halamard and Kronecker products as well. These are defined below.

Definition 2.1.3. *Let A and B be $N \times K$ matrices. The Halamard product*

$$A \circ B = C = \begin{bmatrix} a_{1,1}b_{1,1} & \cdots & a_{1,K}b_{1,K} \\ \vdots & \ddots & \vdots \\ a_{N,1}b_{N,1} & \cdots & a_{N,K}b_{N,K} \end{bmatrix}$$

[2]Proof: for more general proof that also considers complex valued matrices but from which the real case easily follows, see Horn and Charles 2012, p. 440

[3]For proof, see the Review of Linear Algebra section of R. Wang 2020

is taken element-wise so that $c_{i,j} = a_{i,j}b_{i,j}$

Definition 2.1.4. *Let* A *be* $N \times K$ *and* B *be* $J \times D$ *matrix. The Kronecker product*

$$A \otimes B = C = \begin{bmatrix} a_{1,1}B & \cdots & a_{1,K}B \\ \vdots & \ddots & \vdots \\ a_{N,1}B & \cdots & a_{N,K}B \end{bmatrix}$$

is taken by forming a block matrix where the later matrix is multiplied by one element of the first in each block, the locations of the blocks being defined by the indexes of the first matrix's elements.

2.1.2 Miscellaneous Mathematics

There is a collection of notations, definitions and theorems that are also relevant background for the thesis, but do not fall under the same field of mathematics, if not under a very broad umbrella on Analysis.

Notation 2.1.3. *We mark composition of two function with* $\bullet$*, e.g* $f(g(x))$ *becomes* $f \bullet g$

Definition 2.1.5. *We define the vectorisation of a* $N \times K$ *matrix* X *to be the following operation:*

$$vec(X) = (x_{1,1}, \ldots, x_{1,K}, \ldots, x_{N,1}, \ldots, x_{N,K})$$

Last but not least we characterize the conditional distribution of a continuous random vector (with certain reservations).

Definition 2.1.6. *Let* $Z = (X,Y)$ *be a random vector where* $X = (X_1, \ldots, X_n)$ *and* $Y = (Y_1, \ldots, Y_p)$ *with jointly continuous distribution with a corresponding density function* $f(x,y)$*. Now the density*

of X given $Y = y$ (observed values of Y) is

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_y(y)}$$

when $f(y) > 0$. We define $f_{Y|X}(y|x)$ in a similar manner.

From a technical point of view, this density is not wholly satisfactory. For example, it is not necessarily distinct. However, these problems are not severe for practical applications, as the problem is related to variants of the density function that do not agree on zero measured sets. It is possible to show that this density can be used to compute the conditional distribution of a continuous random vector[4], that is: $P(X \in A|Y = y) = \int_A f_{X|Y}(x|y)dx$.

Now we can move to the part of our vocabulary that we need for discussing about statistical inference.

2.1.3 Terms and Notations of Statistical Inference

If not mentioned otherwise, we follow relatively consistent way of notating different components of statistical inference.

Our main interest is to explain and predict the behaviour of some random vector $Y \in \mathrm{R}^N$ with probabilistic modeling. As time series are our main focus application-wise, Y's values most often relate to different points of time. If we know the values this random vector has (for example, we aim to explain air humidity as a function of time and we have some measurements of air humidity), we mark these "realised" values by $y \in \mathrm{R}^N$. If the random process to be predicted is of different form,

[4]Claim at Chapter 4.5 of Lebanon 2020 and Chapter 2 contains the proof for scalar case, which it claims to be similar as that for random vector. As the focus of this thesis is somewhere else and these are not disputed claims, I move on.

it will be told explicitly.

If, one the other hand, we want to highlight that we want to predict the values the random vector will get, we use the asterisk and write $Y^* \in M$ for the unknown values of the random vector to be predicted. This distinction is especially helpful if we know some values of the random vector and want to use them to predict others.

Although this varies and is established in the specific context within the thesis, g and G are often used to notate a random variable that is a function of Y and g is the related observed value.

Often our aim is to associate the behaviour of the random vector with measurements and indicators of other things. If we do not consider these measurements and indicators to be random variables themselves, we call them **features**. For example, time could be a feature used to predict or at least be associated with air humidity. By default, we mark features with $t \in \mathrm{R}^N$ and consider them vectors were each element relates to an singe element of the explained vector $Y \in \mathrm{R}^N$ (Pairing being index-wise:(Y_i, t_i)).

In cases were we speak of a matrix of features, we mean a situation were each row contains all features related to a single observation (or to be predicted value) from Y and columns contain different features, if not defined otherwise in the specific context. When we refer to the space of possible values the features may obtain, we speak of the feature space.

2.2 Properties of Multivariate Gaussian Distribution

The usefulness of the GP approach in statistical modeling is based on the properties of the multivariate Gaussian distribution. We assume that some of the basic properties of Gaussian random

variables are familiar to the reader, especially the definition of one dimensional Gaussian that we do not cover here, but we will shortly revise some concepts and results that are heavily used in this thesis, as they also set certain requirements to our modeling.[5]

We start by introducing one of the basic blocks for constructing the multivariate Gaussian random variable, the standard normal random vector.

Definition 2.2.1. *A random vector $Y = (Y_1, \ldots, Y_N)^T$ is called standard normal random vector if all of its components Y_n are independent and each is a zero-mean unit-variance normally distributed random variable*

Using the preceding term, we can define the multivariate Gaussian random variable (or normal random vector) as a linear transformation of the standard normal random vector.

Definition 2.2.2. *A real random vector $Y = (Y_1, \ldots, Y_N)^T$ is called a normal random vector if there exists a random -vector Z , which is a standard normal random vector, a vector $\mu \in R^N$, and a $N \times K$ matrix A, such that*

$$Y = AZ + \mu$$

Where $\mathrm{cov}(Y,Y) = \Sigma = AA^T$ *and* $\mathrm{E}(Y) = \mu$. *We mark this distribution by writing:*

$$Y \sim N(\mu, \Sigma).$$

The requirement for a covariance matrix that can be written as a product of some matrix and its transpose also implies (see theorem 2.1.2.) that we require the covariance matrix to be positive

[5]This chapter is, among other sources, based on my B.Sc thesis that was about GPs. I have no intention of self-plagiarism, but it already was an aggregate of previous research, there was no need to cite it for any of the shared content.

semi-definite, which is actually the case with all covariance matrices. This is a significant motivation for our later focus on the positive semi-definiteness of certain matrices. We refer to the normal random vector as multivariate normal or multivariate Gaussian random variable, or we say that a random vector follows multivariate normal or multivariate Gaussian distribution if it is normal random vector.

The multivariate normal distribution has many properties that make it easy to use in practical modeling. We list some of these, the first one being that the affine transformation of a multivariate normal distribution is a multivariate normal distribution.

Theorem 2.2.1. *Affine transformation of a multivariate normal distribution is a multivariate normal distribution*[6]

An easy mistake in probability calculus is to think that the covariance between two random variables and their independence are equivalent in general. With the multivariate normal distribution this equivalence holds.[7]

Theorem 2.2.2. *Components of multivariate Gaussian random variables are uncorrelated if and only if they are independent. Formally:* $X \sim N(\mu_X, \Sigma_X), Y \sim N(\mu_Y, \Sigma_Y)$ *and* (X,Y) *is a multivariate Gaussian random vector, then*
$\operatorname{cov}(X,Y) = 0 \Leftrightarrow X \perp\!\!\!\perp Y$.

Another theorem allows us to conclude that the joint distribution of independent multivariate Gaussian variables is multivariate Gaussian as well.[8]

Theorem 2.2.3. *If* $X = (X_1, \ldots, X_N)$ *is a multivariate Gaussian variable and* $Y = X = (X_1, \ldots, X_M)$ *is multivariate Gaussian variable and X and Y are independent, then* (X,Y) *is a multivariate*

[6] For proof, see chapter 5.7 of Siegrist 2020
[7] For proof, see chapter 5.7 of Siegrist 2020
[8] For proof, see chapter 5.7 of Siegrist 2020

Gaussian variable and

$$(X,Y) \sim N(\begin{bmatrix}\mu_X \\ \mu_Y\end{bmatrix}, \begin{bmatrix}\Sigma_X & 0 \\ 0 & \Sigma_Y\end{bmatrix}).$$

The subvectors of a multivariate Gaussian distribution also follow multivariate Gaussian distribution.

Theorem 2.2.4. *Let Z=(X,Y) follow a multivariate Gaussian distribution. Then X and Y follow multinormal Gaussian distributions*[9] *: formally* $X \sim N(\mu_X, \Sigma_{XX}), Y \sim N(\mu_Y, \Sigma_{YY})$ *and* $\mathrm{Cov}(X,Y) = \Sigma_{X,Y}$*) with the following composition of Z:*

$$Z \sim N(\begin{bmatrix}\mu_X \\ \mu_Y\end{bmatrix}, \begin{bmatrix}\Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY}\end{bmatrix}).$$

Where $\Sigma_{X,Y} = \mathrm{cov}(X,Y)$.

The multivariate normal has a density function under known conditions. This property is fundamental for practical modeling, as it gives us a basis for a likelihood function for a model based on multivariate normal distribution. Here we can see why positive definite correlation matrix is so desirable, as the property guarantees the existence of the needed inverse matrix.

Theorem 2.2.5. *If the covariance matrix of a multivariate Gaussian random variable Y is positive definite, the multinormal distribution has a density function*[10]*:*

$$f(y) = \frac{1}{|\Sigma_Y|^{\frac{1}{2}}(2\pi)^{\frac{n}{2}}} \exp\{-\frac{1}{2}(y-\mu_Y)'\Sigma_Y^{-1}(y-\mu_Y)\}.$$

[9]For proof, see chapter 5.7 of Siegrist 2020
[10]For proof, see chapter 5.7 of Siegrist 2020

For modeling based on GPs, the following formula is equally important. It provides an analytical formula for predicting rest of a Gaussian vector based on some of its observed elements.

Theorem 2.2.6. *Let $Z = (X,Y)$ follow a multivariate Gaussian distribution. If Σ_{XX} is invertible, then the conditional distribution of Y given observed elements of the multivariate Gaussian $X = x$ is:*[11]

$$(Y|X = x) \sim N(\mu_Y + \Sigma_{YX}{\Sigma_{XX}}^{-1}(x - \mu_X), \Sigma_{YY} - \Sigma_{YX}{\Sigma_{XX}}^{-1}\Sigma_{XY}).$$

Finally, we define a generalisation of the multivariate normal distribution, the matrix Gaussian distribution.[12]

Definition 2.2.3. *The random $N \times P$ matrix X is said to follow matrix Gaussian distribution*[13] *with mean $N \times P$ matrix M and covariance matrix $V \otimes U$, where U is positive definite $N \times N$ and V is positive definite $P \times P$ matrix, if*

$$\text{vec}(X) \sim N_{NP}(vec(M), V \otimes U)$$

where $\otimes$ is the Kronecker product[14]

Each element of the covariance matrix is defined by multiplication of an element of U and element of V. Formally: $\text{cov}(X_{ij}, X_{kl}) = u_{ik}v_{jl}$. We can interpret that U describes the row-vise and V the column-wise covariances between elements of the matrix. The matrix Gaussian distribution defined in this manner also has a density.

[11] For proof, see chapter 5.7 of Siegrist 2020

[12] For matrix Gaussian distribution, see the supplement of Ding and Cook 2014.

[13] Also known as matrix variate normal distribution

[14] If you have forgotten the definition, see the Prerequisites Section at the start of this chapter

Theorem 2.2.7. *Let X follow a matrix Gaussian distribution as defined above. Its density function is*[15] *(x now being a fixed value of the matrix):*

$$p(x|M,U,V) = \frac{\exp(-\frac{1}{2}\mathrm{tr}(V^{-1}(x-M)^T U^{-1}(x-M))}{(2\pi)^{\frac{np}{2}}|V|^{\frac{n}{2}}|U|^{\frac{p}{2}}}.$$

With the multivariate Gaussian distribution discussed, we can move to another building block of a GP model, the kernel function.

2.3 The Kernel Function

We start by defining the kernel function itself.

Definition 2.3.1. *If $k_\theta : T \times T \to R$ is a function parametrised with a parameter vector $\theta = (\theta_1 ..., \theta_s)$. We call it a kernel function.*

Our practical motivation is to use kernel functions to describe how the similarity of values obtained by the values of a random variable is connected to the similarity of values of the related features. We do this by using the feature values $(t_1 ..., t_n)$[16] as inputs for the kernel function and then constructing a covariance matrix of the following form:

$$K = K_\theta = \begin{bmatrix} k_\theta(t_1,t_1) & \cdots & k_\theta(t_1,t_n) \\ \vdots & \ddots & \vdots \\ k_\theta(t_n,t_1) & \cdots & k_\theta(t_n,t_n) \end{bmatrix}$$

[15] For proof of the density, see chapter 2 of Gupta and D.Nagar 1999

[16] In this thesis t_i is always either a real value or vector of real values in concrete cases.

Careful reader might have noticed that our definition of a kernel does not guarantee that the resulting matrix K is a proper covariance matrix, that is to say symmetric and positive semi-definite. For this reason we introduce the term valid kernel. We do not mention the symmetricity requirement separately, as it is in our definition a precondition for positive semi-definiteness. However, we also define the looser term of symmetric kernel.

Definition 2.3.2. *We call kernel valid*[17] *if it always (with the allowed parametrisation and allowed inputs) defines a positive semi-definite matrix. We call kernel symmetric if* $k(t_i,t_j) = k(t_j,t_i)$

The valid input $T \times T$ depends on the kernel, but most often it is either a pair of real numbers or real vectors. An important thing to note is that the positive semi-definiteness is only assumed when we compute the matrix row and column wise over the same feature vector/matrix in the same manner as with the matrix K. Clearly a valid kernel needs to be symmetric as well.

The idea of the kernel function is that in modeling we are interested about the connections between different feature values $(t_1,\ldots,t_n)$ and the kernel function k_θ reflects assumptions about the possible nature of such connections. The higher the absolute values obtained by the kernel function, the higher the (positive) connection between the observations related to the features the kernel takes as input.

There is a wide collection of different kind of kernels. We provide some examples here. A very simple but useful kernel is the Exponentiated Quadratic (EQ) kernel

$$k_\theta(t_1,t_2) = \theta_1^2 \exp\left\{-\frac{(t_1-t_2)^2}{2\theta_2^2}\right\}.$$

Here the similarity between points of time is a function of their distance. The choice of θ determines the strength of the connection (θ_1 is clearly the upper limit of the covariance) and the

[17] Definition from S. Roberts et al. 2012, p. 7

smoothness (θ_2 determines how fast the connection between different feature values diminishes) of the function. Another widely used kernel type is suitable for detecting periodic components, an example of which is the following Periodic Kernel (PK)

$$k_\theta(t_1,t_2) = \theta_1^2 \exp\left\{-\frac{2}{\theta_2^2}\sin^2(\frac{(t_1-t_2)}{\theta_3})\right\}.$$

In fact, the above kernel is already quite sophisticated, as it can track the wawelength of the periodic component as well as its smoothness and the overall strength of this periodic connection. We will return to the validity of these kernels in later section of this chapter.

Although we are (mostly) focusing on 1-dimensional time series and mostly introduce kernels that take real value (pair) inputs, many kernels have different kind of inputs. For example, EQ kernel can be generalized to higher dimensional space, and higher dimensions also have applications. GPs are used extensively in geospatial modeling[18], where 2-dimensional kernels are the default. There is no theoretical limitation to dimensionality, although the computational cost will naturally increase as a function of dimensionality.

2.4 Gaussian Process

We return to the situation we talked about in the introduction. We have some observations $y = (y_1,\ldots,y_n)$ from a temporal (or perhaps other kind of) process available, and we would like to fit a model to it. However, we only vaguely understand how these observed values are related, perhaps we assume that similar values of related features tend to cause similarity of observed values of the process, or perhaps it is a periodic process. It is time to define the Gaussian Process (or GP which will remain as the shorthand version), which will in time give us tools to tackle a situation like this.

[18] Gelfland and Schliep 2016

It is a stochastic process, so we need to define the later first.

Definition 2.4.1. *Let T be an arbitrary set. Now a set of random variables $Y(t)$, $t \in T$ defined in the same probability space Ω, is a stochastic process.*[19]

Definition 2.4.2. *Let T be the set of all indices of a stochastic process. Gaussian Process is a stochastic process, for which it holds that any finite subset of it follows multivariate Gaussian distribution. Formally:* $(t_1,\ldots,t_n) \subset T \Rightarrow (Y(t_1),\ldots,Y(t_n)) \sim N(\mu,\Sigma)$.[20]

Using properties of the multivariate Gaussian distribution and kernel functions, we turn the GP to a tool of non-parametric modeling. The outline of the construction is the following: Let $y = (y_1 ..., y_n)$ be our set of observations and $(t_1 ..., t_n)$ the set of related features (in our case temporal measurements). Now we assume that Y is a finite subset of a GP, that is $Y \sim N(m,K)$. Note that we changed the notation for mean vector and covariance matrix. We also abuse notation, since m refers to $(m(t_1),...,m(t_n))$ and K to $(k(t_i,t_j))i,j$. This notation stresses the fact than the essence of modeling with GPs is about the choice of covariance and mean functions. The mean is often set to be 0 or it is defined by a weakly informative prior in Bayesian settings. The covariance is defined by a valid kernel function, as defined in this chapter.

As the kernel describes ways the features associated to the values of the modeled process are related, the idea of GP modeling becomes one of relating process values by the similarity of the values of related features. As there are many kinds of kernels and they can express many kind of and often very general level similarities, GPs are very flexible ways to model different kind of phenomenon. Next we will see that kernels are also flexible w.r.t each other, as new valid kernels can often be consructed from old ones.

[19] Lifshits 2014, p. 47

[20] C. Rasmussen and Williams 2006[13]

2.5 Properties of Kernels

In practice, we want to limit our modeling to valid kernels. Fortunately a relatively small set of valid kernels can be expanded to a larger one by addition, scalar and kernel multiplication and transformation of existing valid kernels, as well as with less obvious methods. Here we mostly drop the subindex θ from the kernel function and related matrices.

Theorem 2.5.1. *Let k_1, k_2 be two valid kernels. Then the kernel $k_1 + k_2$ is valid.*

Proof. Let K_1, K_2 be $N \times N$ matrices corresponding to the two kernels applied over a set of features $t = (t_1, ..., t_N)$ and $x = (x_1, ..., x_N)^T \in R^N \setminus 0$. Now it holds that $x^T K_1 x \geq 0$ and $x^T K_2 x \geq 0$. It follows that

$$x^T (K_1 + K_2) x = x^T K_1 x + x^T K_2 x \geq 0.$$

Since both kernels are valid, it also follows that

$$k_1(t_i, t_j) + k_2(t_i, t_j) = k_1(t_j, t_i) + k_2(t_j, t_i).$$

□

Theorem 2.5.2. *Valid kernel multiplied with a positive scalar is a valid kernel.*

Proof. k be the kernel, K the matrix defined by it and $a > 0$ the scalar. Now k is valid $\Leftrightarrow x^T K x \geq 0 \Leftrightarrow a(x^T K x) \geq 0$.

That scalar multiplication preserves the symmetricity follows by multiplying $K_{i,j} = K_{j,i}$ with scalar to obtain $aK_{i,j} = aK_{j,i}$. □

Theorem 2.5.3. *Let k_1, k_2 be two valid kernels. Then the kernel $k_1 k_2$ is valid.*[21]

[21] McKay 1998, p. 17

Proof. Let K_1, K_2 be $N \times N$ matrices corresponding to the two kernels applied over a set of features and $x = (x_1, ..., x_N)^T \in \mathrm{R}^N \setminus \{0\}$. Since both kernels are valid, it trivially holds that $k_1(t_i, t_j)k_2(t_i, t_j) = k_1(t_j, t_i)k_2(t_j, t_i)$, from which symmetricity follows.

We make the following observation: the matrix that results from multiplying the kernels

$$K_{k_1 \times k_2} = \begin{bmatrix} k_1(t_1,t_1)k_2(t_1,t_1) & \cdots & k_1(t_1,t_n)k_2(t_1,t_n) \\ \vdots & \ddots & \vdots \\ k_1(t_n,t_1)k_2(t_n,t_1) & \cdots & k_1(t_n,t_n)k_2(t_n,t_n) \end{bmatrix} = K_1 \circ K_2$$

Is a Hadamard Product[22] of the matrices K_1, K_2. These matrices are positive semi-definite. Now, we can use eigendecomposition to write $K_1 = \sum_{i=1}^{N} \lambda_i q_i q_i^T, K_2 = \sum_{i=1}^{N} \nu_i m_i m_i^T$, where $q_i = (q_{i,1}, ..., q_{i,N}), m_i = (m_{i,1}, ..., m_{i,N})$ are the eigenvectors and λ_i, ν_i the eigenvalues. Now the Hadamard product can be written as

$$K_1 \circ K_2 = \sum_{i=1}^{N} \lambda_i q_i q_i^T \circ \sum_{i=1}^{N} \nu_i m_i m_i^T = \sum_{i=1}^{N} \sum_{j=1}^{j} \lambda_i \nu_j (q_i q_i^T) \circ (m_j m_j^T) = \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \nu_j (q_i \circ m_j)(q_i \circ m_j)^T$$

Now we note that all $\lambda_i \nu_j \geq 0$ (all eigenvalues of positive semi-definite matrices are greater than or equal to 0) and all $(q_i \circ m_j)(q_i \circ m_j)^T$ are positive semi-definite. The later claim follows from the following (the form of the matrix already proves symmetricity):

$$x^T (q_i \circ m_j)(q_i \circ m_j)^T x = (\sum_{k=1}^{N} q_{i,k} m_{j,k} x_k)^2 \geq 0$$

From these notions it follows that each component in the above sum composition of the Hadamard product is positive semi-definite, since positive semi-definite matrix multiplied with a positive

[22] For the definition, see the Prerequisites Section at the start of this chapter

scalar remains positive semi-definite. And since the sum of positive semi-definite matrixes is positive semi-definite, the above Hadamard product is positive semi-definite. □

Theorem 2.5.4. *Let $k : T \times T \to \mathrm{R}$ be a valid kernel function and let $g : T_g \times T_g \to T \times T$ be a function so that $g(t_{gi}, t_{gj}) = (u(t_{gi}), u(t_{gj}))$ when $t_{gi}, t_{gj} \in T_g$ and $u : T_g \to T$. Then $k \bullet g : T_g \times T_g \to \mathrm{R}$ defines a valid kernel.*

Proof. Let $t_g = \{t_{g1}..., t_{gN}\}$ be our set of features, not necessarily acceptable as inputs of the kernel k but satisfying $t_{gj} \in T_g$. Now, by applying the $k \bullet g$ to the elements of g, we obtain

$$K = \begin{bmatrix} k(g(t_{g1}, t_{g1})) & \cdots & k(g(t_{g1}, t_{gN})) \\ \vdots & \ddots & \vdots \\ k(g(t_{gN}, t_{g1})) & \cdots & k(g(t_{gN}, t_{gN})) \end{bmatrix} = \begin{bmatrix} k(u(t_{g1}), u(t_{g1})) & \cdots & k(u(t_{g1}), u(t_{gN})) \\ \vdots & \ddots & \vdots \\ k(u(t_{gN}), u(t_{g1})) & \cdots & k(u(t_{gN}), u(t_{gN})) \end{bmatrix}$$

Now it is easy to see that this is the same as the matrix K computed over input $(u(t_{g1}), ..., u(t_{gN}))$ where each $u(t_{gj}) \in T$. This is an allowed input of the kernel, hence the valid kernel defines a positive semi-definite matrix when computed over it. The transformed kernel is valid. □

It is also guaranteed that the limit of a sequence of valid kernels that converge pointwise is valid

Theorem 2.5.5. *Let (k_i) be a sequence of valid kernels and that $lim_{i\to\infty} k(t_j, t_k)_i \in \mathrm{R}$. Then $lim_{i\to\infty} k(t_j, t_k)$ is a valid kernel.*

Proof. Let $k(t_i, t_i) = lim_{n\to\infty} k_n(t_i, t_i)$ be a kernel which exists by the assumption and let K be $(k(t_i, t_j))_{i,j}$. Let $t = (t_1, ..., t_N)$ be a set of features and K_n be $(k_n(t_i, t_j))_{i,j}$ for every $n \in \mathrm{N}_+$. Now

$$x^T K x = lim_{n\to\infty} x^T K_n x \geq 0$$

Since the limit of non-negative sequence of real numbers is non negative.[23]

[23] For proof: see Harjulehto, Klén, and Koskenoja 2017, p. 48

We also observe that since $k_n(t_i,t_j) = k_n(t_j,t_i)$ for all n and t_i,t_j, they must converge to the same value. The limit kernel also retains symmetricity, hence it is valid.

□

The following result is very useful when combined with other theorems.

Theorem 2.5.6. *Let $f : T \to \mathrm{R}$ be an arbitrary function. Then $k(t_i,t_j) = f(t_i)f(t_j)$ is valid kernel.*

Proof. Let k be like in the theorem statement,$(t_1,\ldots,t_N)$ be the features and $x = (x_1,...,x_N)^T \in \mathrm{R}^N \setminus 0$. Now we get:

$$x^T Kx = \sum_{i=1}^{N}\sum_{j=1}^{N} x_i x_j f(t_i)f(t_j) = (\sum_{i=1}^{N} x_i f(t_i))^2 \geq 0$$

We also observe that this kind of kernel is always symmetric:

$$k(t_i,t_j) = f(t_i)f(t_j) = k(t_j,t_i).$$

□

These theorems have very favourable implications. They tell us that we get well behaving complicated kernels from simple kernels that are valid. This allows e.g. addition of both linear and periodic components to a kernel, description of processes that undergo change on different time scales (e.g. sum of two exponential kernels with different parameters) or addition of covariance effects between processes.

Next we give a simple chain of demonstrations for how new kernels can be constructed from old ones, the end result being the EQ kernel introduced earlier. We start from a very simple kernel, and build more complicated ones from it. If not mentioned otherwise, the parameters are real values except 0.

Theorem 2.5.7. *Let $k : \mathrm{R} \times \mathrm{R} \to \mathrm{R}, k(t_i, t_j) = t_i t_j$. Then k is called linear kernel, and it is valid.*

Proof. A special case of the previous theorem with f being the identity function. □

We can use the linear kernel to construct a more complicated valid kernel.

Theorem 2.5.8. *Let $k : \mathrm{R} \times \mathrm{R} \to \mathrm{R}, k(t_i, t_j) = \exp(t_i t_j \beta)$ and $\beta > 0$. Then the kernel is called exponential kernel, and it is valid.*

Proof. By definition of exponential function we can write this kernel in the following manner:

$$k(t_i, t_j) = \exp(t_i t_j \beta) = \sum_{k=0}^{\infty} \frac{(\beta t_i t_j)^k}{k!}.$$

If we mark $k_n(t_i t_j) = \sum_{k=0}^{n} \frac{(\beta t_i t_j)^k}{k!}$, we see that the kernel k_n is obtained by multiplication, scaling and addition of linear kernels that are valid, hence the kernel is valid for each n. The kernel k is the limit of the kernel k_n (converges to it pointwise), so k is valid as well. □

Going further, we can prove that the Squared Exponential (SE) kernel is valid.

Theorem 2.5.9. *Let k be kernel so that $k : \mathrm{R} \times \mathrm{R} \to \mathrm{R}, k(t_i, t_j) = \exp(-\frac{1}{2}\frac{(t_i - t_j)^2}{\theta^2})$. Then k is called Squared Exponential kernel, and it is valid.*

Proof. We observe that the kernel can be written in the following manner:

$$k(t_i, t_j) = \exp(-\frac{1}{2}\frac{(t_i - t_j)^2}{\theta^2}) = \exp\left\{-\frac{t_i^2}{2\theta^2}\right\} \exp\left\{-\frac{t_j^2}{2\theta^2}\right\} \exp\left\{\frac{t_i t_j}{\theta^2}\right\}.$$

Now we can define

$$f(x) = \exp\left\{-\frac{x^2}{2\theta^2}\right\}.$$

So we can write

$$k(t_i, t_j) = f(t_i) f(t_j) \exp\left\{\frac{t_i t_j}{\theta^2}\right\}.$$

Therefore, Squared Exponential kernel can be written as a product of an exponential kernel which is valid by theorem 2.58, and another kernel, which is valid by theorem 2.56. We conclude that it is valid. □

Theorem 2.5.10. *Exponentiated Quadratic (EQ) kernel k:*$\mathrm{R} \times \mathrm{R} \to \mathrm{R}$ *is valid.*

Proof. EQ kernel can be obtained by scaling SE kernel with the variance parameter, so it is valid. □

The Periodic Kernel (PK) introduced earlier and other similar periodic kernels can be obtained by transforming the inputs of a kernel that takes real values $x \in \mathrm{R}$ as inputs to a two dimensional value that expresses periodic properties of the value and then using them as inputs in a kernel that takes inputs of the form $\mathrm{R}^2 \times \mathrm{R}^2$. For example, the transformation can be $g(x) = (cos(ax), sin(ax))$ where $a > 0$ (perhaps a positive parameter or its square root). The kernel can be for example the multi-dimensional version of the EQ kernel with the squared distance between real values replaced with the squared distance of two real vectors of R^2. The PK kernel can be acquired by using the described transformation to the inputs of the multi-dimensional EQ kernel. We leave rest as an exercise for the reader.

The kernels covered here are only a very small subset of all kernels that exist. There is a vast number of kernels with different kind of properties, especially the term of stationarity is important for classifying kernels used in time series modeling.

Definition 2.5.1. *Kernel is stationary if its values only depend on the difference of the feature vectors of which it is a function of. That is, it can be thought as a function of* $h = t_1 - t_j$

Stationary kernels only depend on the difference in time when defining the similarity of two time points, more generally they are translation invariant in index domain. Isotropic kernels extend on this idea, and only care about the absolute value of this difference.

Definition 2.5.2. *Kernel is isotropic if its values only depend on the distance of the feature vectors of which it is a function of. That is, it can be thought as a function of* $|h| = |t_1 - t_j|$

The EQ kernel introduced earlier is an example of a kernel that is both stationary and isotropic. Kernels can also be classified by their output range. One important special case is the following:

Definition 2.5.3. *Kernel is called correlation kernel if it holds that its output range is (-1,1)*

My impression is that the selection of kernel for a GP model is currently mostly based on domain knowledge and the flexibility of certain kernels. However, if the method of model fitting and the computational resources allow complex kernels, it is possible to systematically explore different combinations of kernels. Relatively recently attempts have been made to develop tools for automated kernel selection that bears some resemblance to feature selection used with linear models.[24]

The idea of this approach builds on the properties of kernels as defined in this chapter, namely that we can create valid new ones by multiplication and addition. By defining a set of valid base kernels, it is possible to build a selection tree, where new components are added via addition or multiplication from the group of base kernels to the existing kernel combination. The procedure cited here uses similar approach as is often used with automated feature selection with Generalised Linear Models, where new elements are added to the kernel one at a time, always choosing the one that bring the biggest increase of score that balances predictive capability with model complexity, e.g. Bayesian Information Criterion. Tools of systematic kernel selection significantly extend the already remarkable flexibility of GP modeling.

Having covered the basic building blocks of GP modeling, we move to the basics of model fitting and prediction.

[24] For the systematic search over GP models described here, see Duvenaud, Lloyd, et al. 2013 and Loyd, Duvenaud, and Grosse 2014

2.6 The Marginal Likelihood of a Gaussian Process Model

Whether the setting is Bayesian or not, a fundamental formula for model fitting is the marginal likelihood of finite subset of a GP.

Definition 2.6.1. *Let $y = (y_1...,y_n)$ be a vector of observed values from a GP and $t = (t_1...t_n)$ a set of related features. The kernel and mean of the process in this finite subset is marked as in the chapter 2.4. If K is positive definite, the marginal likelihood is*

$$p(y|m,K_\theta) = (2\pi)^{\frac{-n}{2}}|K_\theta|^{-\frac{1}{2}}\exp(-\frac{1}{2}(y-m)^T K_\theta^{-1}(y-m)) \propto |K_\theta|^{-\frac{1}{2}}\exp(-\frac{1}{2}(y-m)^T K_\theta^{-1}(y-m)).$$

The marginal likelihood can be used as a function of the parameters of the kernel and the mean, thus it allows us to optimize these parameters. It is possible and not altogether unreasonable to fit a GP model based on the marginal likelihood alone, without priors for the hyperparameters. This is not the approach that will be used in this thesis, but as the marginal likelihood is - along with the prior - a component in the posterior density of parameters in Bayesian setting, it remains fundamental for our approach as well.

It is noteworthy that all GPs do not have marginal likelihood, positive semi-definite matrix is not always invertible. A practical solution is to add diagonal noise to the matrix.[25]

2.7 Prediction With Gaussian Process Model

There is one very useful analytical result that can be used to make predictions with a GP Model. Let us assume we have a set of features $t = (t_1,\ldots,t_n)$ and related observations $y = (y_1,\ldots,y_n)$ and let us further assume that this data comes from a GP with known parametrisation and kernel function.

[25] For example, the Stan manual uses this trick in the example code for GPs.

Our task is to predict vector y_* related to a feature vector t_* for which we have no observations. This situation can be formalised as:

$$(Y, Y_*) \sim N(\begin{bmatrix} \mu_y \\ \mu_{y^*} \end{bmatrix}, \begin{bmatrix} K_{YY} & K_{YY_*} \\ K_{Y_*Y} & K_{Y_*Y_*} \end{bmatrix}).$$

Setting Y=y and using this partition and the result for the conditional distribution of a multivariate Gaussian introduced earlier we obtain (assuming that K_{YY} is invertible)

$$Y_*|(Y = y) \sim N(\mu_{y^*} + K_{Y_*Y}{K_{YY}}^{-1}(y - \mu_y), K_{Y_*Y_*} - K_{Y_*Y}{K_{YY}}^{-1}K_{YY_*}).$$

As the density of a multivariate normal Gaussian is centered around the mean (element-wise), the mean can be used as a point estimate for our prediction and we can interpret the variance parameter of the conditional distribution as a confidence interval around this mean. Figure 2.1 illustrates this. This conditional distribution regression formula, also called Gaussian Process Regression, is at the heart of GP modeling, and it has been used in this manner from the 1970's onward.[26]

In fact, we can extend this result to be even more helpful. Let $Y_1, ..., Y_n$ be independent GPs and $G = Y_1 + ... + Y_n$ be their sum. Now, since the sum of independent multivariate Gaussian's is a multivariate Gaussian, G follows multivariate Gaussian distribution in all finite sets of features, and therefore it is a GP. Now, we can naturally predict the values of the entire process in unknown locations with the previous formula, but we can also predict the values of the subprocesses Y_k. By

[26] O'Hagan 1978

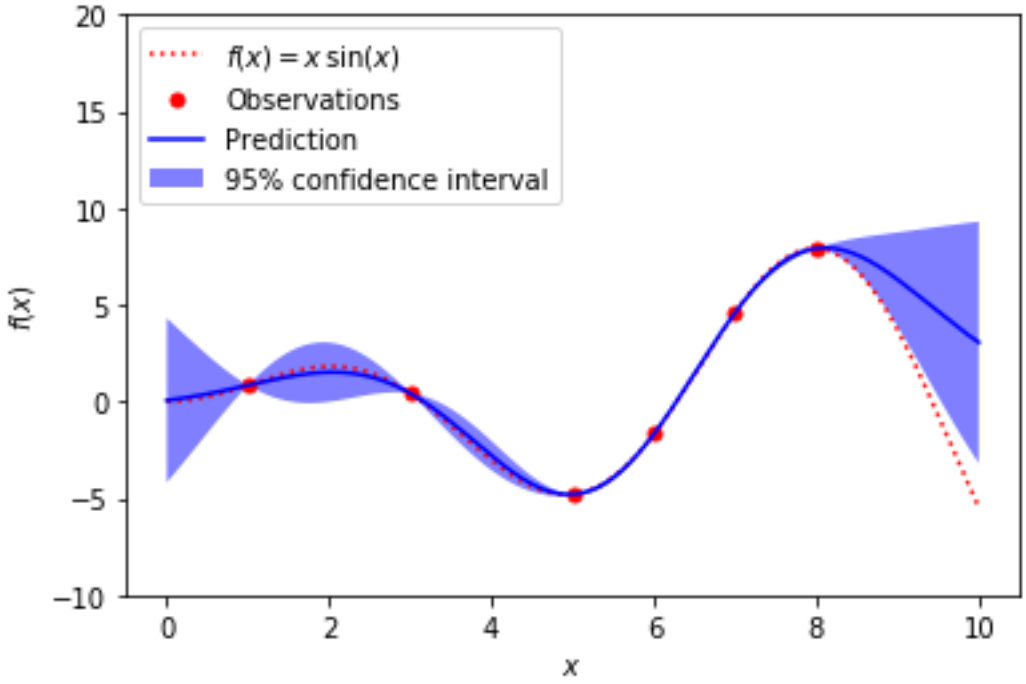

Figure 2.1: Visual illustration of GP prediction with the dark blue line being the mean prediction and the lighter blue area around it the confidence interval. The picture is an example illustration of the skikitlearn-package. For more about the package, see Pedregosa et al. 2011.

making the following partition[27]

$$(G, Y_k^*) \sim N(\begin{bmatrix} \mu_G \\ \mu_{Y_k^*} \end{bmatrix}, \begin{bmatrix} K_{GG} & K_{GY_k^*} \\ K_{Y_k^*G} & K_{Y_k^*Y_k^*} \end{bmatrix})$$

we can once again apply the analytical form of the conditional distibution of the multivariate normal distribution, and using it we get the following predictive formula[28] for the subprocess Y_k:

[27] assuming that K_{YY} is invertible and the block matrix as a whole positive semi-definite, symmetricity clearly holds
[28] Idea presented for example in Duvenaud 2014, p. 18

$$Y_k^*|(G=g) \sim N(\mu_{Y_k}^* + K_{Y_k^*G}{K_{GG}}^{-1}(g-\mu_g), K_{Y_k^*Y_k^*} - K_{Y_k^*G}{K_{GG}}^{-1}K_{GY_k^*}).$$

This formula has several applications, two of which are very relevant for this thesis. Based on joint observations from the entire process, we can still separate our predictions for subprocesses. This allows as to estimate relevancy of e.g. periodic, slow term and long term components of the process separately.[29] We can also predict subprocesses that are themselves sums of individual subprocesses. Another application (and one that is often separately derived, even though it is only a special case) is the prediction of the real value of the process when the model includes a Gaussian white noise term.

Such a term can be seen as a GP with a zero mean and diagonal identity matrix that is scaled with a noise parameter. The corresponding kernel is $k(x_i,x_j,\theta) = \delta(x_i,x_j)\theta^2$, where θ controls the amount of the noise and δ is Dirac's delta function[30]. In this special case, the predictive distribution of the noise-free process given observations skewed by the noise is of the form

$$Y_*|(G=g) \sim N(\mu_{y^*} + K_{Y_*Y}(K_{YY} + I_n\sigma^2)^{-1}(g-\mu_g), K_{Y_*Y_*} - K_{Y_*Y}(K_{YY} + I_n\sigma^2)^{-1}K_{YY_*}).$$

[29]For examples of subtrend detection, see Duvenaud 2014, p. 17 and Cheng et al. 2020, p. 7

[30]Instead of the proper definition we mean a function for which it holds that $\delta(x_i,x_i) = 1$ and $\delta(x_i,x_j) = 0$ if $i \neq j$. Rigorously defined Dirac's delta is a generalized kernel as defined by Laurent Schwartz. For more, see Hörmander 1983, p. 256

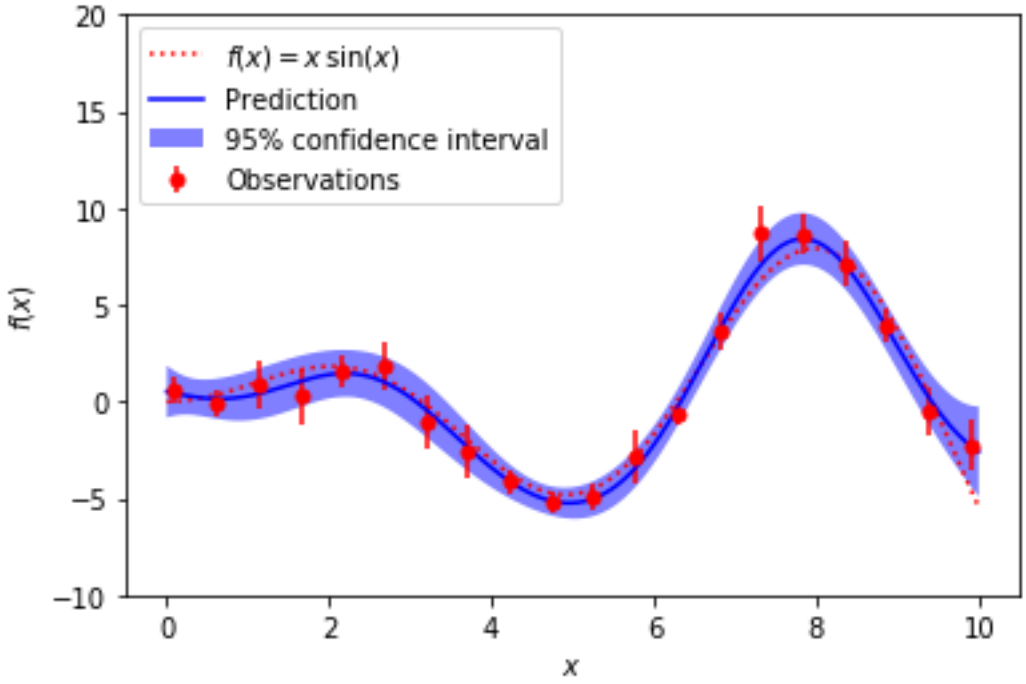

Figure 2.2: GP prediction with white noise. Note that now the regression line does not go through all of the observations. The picture is an example illustration of the skikitlearn-package. For more about the package, see Pedregosa et al. 2011.

2.8 Theoretical Properties and Practical Performance of Gaussian Process Regression

The intuition behind the GP regression is clear enough, but is there any reason beyond ”it just works” to assume that it truly does? Here by ”working” I mean theoretical properties that would guarantee that predictions done with the GP regression would (likely) keep increasing in accuracy with enough data, or demonstrated better performance than with other models. We will leave most of the details behind references that themselves are only a small part of related discussions, as these topics could form theses of their own.

There are multiple results that guarantee that, by choosing a specific kernel, the GP Regression will lead to the same prediction as some other modeling approach, some of which have theoretical guarantees. Such results exist at least for linear models[31] and neural networks[32]. These results are insightful in the sense that they expose connections between different approaches. There are also many empirical studies in which GP models have outperformed other models and they have been suggested as replacements for neural networks in tasks like non-linear regression and classification[33]. There are also convergence results that guarantee that GP models start to approximate the observation generating function in some sense as the amount of data tends to infinity[34], but we will not go to these often technical results in any more detail.

Both theoretical and practical performance of GP models are research fields that are either so deep or so extensively discussed that we can mostly nod to the existence of these topics. However, nothing guarantees sensible modeling if assumptions about the kernels and parameters are badly chosen, and this is our main focus in rest of this thesis.

2.9 Conclusions

To summarize, we have a very flexible tool at our disposal. GPs define distributions over functions, and so they allow modeling based on relatively vague understanding of the data generating process. Assuming that our kernels define positive semi-definite matrices, it is in theory easy to combine elementary processes to more complicated ones. The analytical result for prediction means that with a fixed estimate for parameters, predictive inference is easy as well.

[31] Williams 1998, pp. 602–605
[32] Duvenaud 2014, pp. 60–61
[33] Chen and B. Wang 2018
[34] For example C. Rasmussen and Williams 2006, pp. 155–157 or Teckentrup 2019

However, in practice modeling with GPs is much more difficult for several reasons. The predictive capability of a GP model might be weak in areas of the feature space where we have no observations, and the flexibility of the approach also means that fixed point estimate of parameters might favour a function not corresponding with our preferences or knowledge w.r.t for example the smoothness of the process we are trying to model.

To mitigate these problems, the next chapter adds a layer of depth to GP modeling. By defining probability distributions over the model parameters, we can both regularize the model and encode our prior knowledge about the process we are modeling to it.

Chapter 3

Gaussian Process Modeling in a Bayesian Framework

3.1 Hierarchical Modeling

Bayesian hierarchical model consists of several layers of related random variables. One can, for example, define a random variable with a mean so that the mean itself is random variable with its own parameters that can also have their own distribution etc. GPs are often used in this kind of framework. The simplest cases merely add priors to a similar model described in the previous chapter, but GPs can be and are used in more complex hierarchical models as well.[1] As more complex models are non-trivial to optimize, we will also shortly discuss this topic, as well as predictive inference and model evaluation in a more general Bayesian context.

[1] For Hierarhical models in general, see chapter 5 of Gelman et al. 2014

3.2 Gaussian Process Model With Priors and a Mean Function

The first extension of a GP model to a Bayesian setting simply treats the parameters of the mean and kernel functions as random variables. Here we define that for finite subset Y of a GP (with related features jointly marked here with X, they are not considered random) it holds that $Y \sim N(m(X,\mu),K(X,\theta))$. The parameters for the mean and covariance function follow their own prior distributions, that reflect our assumptions about the properties of the GP. We use the notation $\phi = (\theta,\mu)$. The following set of equations formalises this approach:

$$\begin{aligned}\phi &\sim \pi(a)\\ Y|\phi,X &\sim N(m(X,\mu),K(X,\theta)).\end{aligned}$$

This simple probabilistic extension of a GP model has multiple benefits over a model that is fitted only with the likelihood of the multivariate Gaussian distribution. As our framework is Bayesian, we can evaluate the joint evidence our priors and the data together give to different parametrisations with the posterior distribution of the model parameters

$$p(\phi|y) = \frac{p(y|\phi)p(\phi)}{\int p(y|\phi)p(\phi)d\phi} \propto p(y|\phi)p(\phi).$$

The prior distribution allows us to regularize the model fitting process. We might not prefer all values of θ and μ equally either due to prior knowledge about the process the GP models or we might just want to be conservative in our estimates and assume high amounts of noise (that can be expressed with the right kind of kernel and related parametrisation) and low level of actual dependencies between observations. The exact information and preferences encoded to the choice of the hyperparameter distributions and their parametrisation depends on context, but what is important

to understand is that the Bayesian framework allows such encoding. As GP itself is a very flexible tool, we can at the same time fit our model over a very large set of functions and get the benefits of the Bayesian approach.[2]

The mean function is especially useful for encoding knowledge about the behavior of the process when data is not available. For example, we might only have time series observations of a physical process from a limited interval, but fortunately some understanding about the underlying physics behind the phenomenon. One example given in the research literature is the use of exponential decay function to model certain physical processes.[3] This can lead to a better predictive inference in situations where data is not available from all parts of the feature space.

Before moving to more complicated constructions it is convenient to discuss fitting and predictive inference with GP models.

3.3 Optimisation and Predictive Inference

There are roughly two major approaches to fitting a Bayesian GP model in a Bayesian setting. In the first one the idea is to maximize the expression $p(y|\phi)p(\phi)$ as a function of the parameter $\phi = (\theta, \mu)$. This is a joint expression of the marginal likelihood that we introduced earlier (that alone could be the basis for fitting in a non-Bayesian setting) and the prior distribution of the hyperparameters. The parameter $\hat{\phi} = \text{argmax}_{\phi}(p(y|\phi)p(\phi))$ can be used as a plug-in estimate for the hyperparameters. There are various algorithms for finding the maximizing point estimate. Good point estimate is often easier and faster to find than the full posterior distribution that we will discuss next. On the other hand, point estimate does not treat the inherent uncertainty about the 'real'

[2]For a discussion about the benefits of Bayesian GP modeling, espesically from the point of view of time series. see S. Roberts et al. 2012

[3]S. Roberts et al. 2012

hyperparameter values as rigorously as full probabilistic treatment.[4]

The other major approach aims to obtain the whole posterior distribution $p(\phi|y)$ or its approximation. In this case, we can evaluate the following predictive formula:

$$p(y^*|y) = \frac{\int p(y^*|\phi,y)p(y|\phi)p(\phi)d\phi}{\int p(y|\phi)p(\phi)d\phi} = \int p(y^*|\phi,y)p(\phi|y)d\phi. \tag{3.1}$$

Compared to a plug-in estimate that "breaks the rules" and forces all of the posterior mass to a single point estimate, here we integrate over the posterior distribution of the parameters. This solution explicitly addresses our uncertainty and prior information, and is less prone to overfitting. It is also theoretically satisfying in the sense that we stick to the probabilistic framework from the beginning (prior to using the data) to the end (predictive inference).

It is possible to simulate draws from the posterior distribution explicitly, and use these samples for the evaluation of the above integrand. These sampling techniques are based on Monte Carlo Markow Chain -algorithm (MCMC) that in theory converge to the posterior distribution in some point. Although there are are no ways to be absolutely sure of convergence after any finite set of samples, the existing diagnostic tools are often good enough for practical evaluations. Such tools include summary statistics like $\hat{R}$ that measures how well markov chains have mixed, diagnostic plots and convergence result that tells us that, under certain conditions, as the amount of observations increases, the posterior distribution of parameters will converge to a multivariate Gaussian.[5]

In Bayesian modeling in general, this approach has become vastly more popular in recent years because of increased and cheaper computational power and softwares like Stan that combine fast

[4]These notions about the empirical Bayes approach are based on the lecture notes of the Spatial Modeling and Bayesian Inference -course lectured in Spring 2019 at the University of Helsinki, Vanhatalo 2019, p. 53.

[5]For these and other tools of convergence checking with MCMC, see chapter 11 of Gelman et al. 2014.

description of models with automated and efficient sampling all within one language and one chunk of code. Unfortunately, GPs are not the best suited models for this approach. Fitting of GP models with MCMC is a much discussed topic, and in practice the approach is often slow and inefficient. Methods like the Laplace approximation provide something that falls between full posterior sampling with MCMC and point estimate for the parameters by first finding a good point estimate and then making inference about how the posterior mass is concentrated around that point.[6]

Nevertheless, Stan, the most popular posterior sampling software currently available, supports GP models and its official guide has an introduction for implementing them. As Stan also provides tools for model diagnostic and checks on predictive inference, there is no clear answer to the question of ”how should one fit GP model”. Deterministic approaches might be faster, but Stan allows quick and flexible implementation of the model itself, and provides the possibility to integrate GP to a larger hierarhical model as well. The question is further complicated by the fact that proper choice of prior parametrisation and distributions can drastically change the convergence of the posterior simulation, so domain knowledge about the modeled process and even good statistical intuition can be factors in determining whether posterior sampling is feasible.

In the end, I decided to use this thesis as an opportunity to develop my skills with posterior sampling software, and the practical examples seen in this thesis have been implemented jointly with Stan and R. Some of the example models of the chapter 5 would have most likely converged faster with some other approach, but Stan saved time in implementation, and as a general and widely used framework mastering it is a benefit in its own right.

In addition to practical considerations and future prospects, direct sampling from the posterior

[6]For a discussion about the pros and cons of MCMC and Laplace approximation in Bayesian Context, see Hartmann 2019, pp. 17–19

distribution also makes predictive inference very convenient, as we can use our draws from the posterior distribution of the parameters to sample from the posterior predictive distribution of new observations. We do not even have to evaluate the above integral! The algorithm of this approach is the following:

1) For each sample of parameters $\phi_i|y$ from the posterior distribution:
2) Draw a sample from $Y_*|\phi_i, t_*, y$, where t_*. emphasizes that we are drawing our new samples from (possibly) unobserved locations of the feature space.[7]

As the distribution of GP in an unknown location given the hyperparameters and observed data has the nice closed from presentation discussed in the previous chapter and sampling from a multivariate Gaussian is relatively easy, simulated draws based on the posterior are very handy if sampling of the posterior itself is possible.

Fit of a GP model can be evaluated in many ways. There is a great variety of tools for model assessment, selection and comparison in Bayesian Statistics in general.[8] Here we only recognize these questions very important for any rigorous statistical analysis done with GPs (or with any other tools for that matter), but we will not delve deeper to this direction.

[7] This is the same as drawing from the joint posterior of parameters and predictions discussed e.g in Gelman et al. 2014, pp. 23–24

[8] A good survey is given byVehtari and Ojanen 2012

3.4 General Extensions of Hierarchical Gaussian Process Models

The time series models described can be used as the final layer of a GP, the process itself being the source of our observations. However, it is very common in many application fields that GPs play a different part. The can be used to model e.g. correlated noise[9], they can parametrisize models where the final layer is discrete or binary[10] (the later case is common with classification problems) or otherwise extend more complicated structures. One way to extend the model introduced at the start of this chapter to model time-series (using the same notation) is the following[11]:

$$\begin{aligned}
&\theta \sim \pi(a) \\
&f|\theta, X \sim N(0, K(X, \theta)) \\
&Y|\theta, f, X \sim N(f, I\sigma^2).
\end{aligned}$$

Here, the finite subset of a latent GP process $f = (f(X_{i,.})..., f(X_{N,.})$ determines the mean of a process skewed by Gaussian noise, the size of which is determined by σ. This can be further marginalised to obtain:

$$Y|X, \theta \sim N(0, K(X, \theta) + I\sigma^2).$$

We provide the necessary theorem and the proof here.

[9] Kurtz and Lin 2019
[10] See chapter 3 of C. Rasmussen and Williams 2006
[11] This model is used e.g. in Cheng et al. 2020

Theorem 3.4.1. *If*

$$Y|Z \sim N(Z+\mu_1,\Sigma_1))$$
$$Z \sim N(\mu_2,\Sigma_2)$$

Then

$$(Y,Z) \sim N((\left[\begin{bmatrix}\mu_1+\mu_2\\ \mu_2\end{bmatrix}\right], \begin{bmatrix}\Sigma_1+\Sigma_2 & \Sigma_2\\ \Sigma_2 & \Sigma_2\end{bmatrix})).$$

Proof. Let $Y|Z$ and Z be defined like in above. Now $(Y-Z)|Z \sim N(\mu_1,\Sigma_1)$. Since the distribution does not depend on Z, we can deduce that there is W such that $W \sim N(\mu_1,\Sigma_1)$ so that $W \perp\!\!\!\perp Z$ and $Y-Z=W$. From these it follows that[12] $(Y-Z)|Z=W|Z=W \sim N(\mu_1,\Sigma_1)$

Now we can write (Y,Z) as a affine transformation of (Z,W), which is a multivariate Gaussian random variable by the theorem 2.2.3., in the following manner:

$$\begin{bmatrix}Y\\ Z\end{bmatrix} = \begin{bmatrix}I & I\\ I & 0\end{bmatrix}\begin{bmatrix}Z\\ W\end{bmatrix}$$

This proves that, as an affine transformation of a multivariate Gaussian random variable, it is a multivariate Gaussian random variable. Now $\mathrm{E}(Y,Z) = (\mu_1+\mu_2,\mu_2)^T$ and

$$\mathrm{cov}((Y,Z)) = \begin{bmatrix}I & I\\ I & 0\end{bmatrix}\mathrm{cov(Z,W)}\begin{bmatrix}I & I\\ I & 0\end{bmatrix} = \begin{bmatrix}I & I\\ I & 0\end{bmatrix}\begin{bmatrix}\Sigma_2 & 0\\ 0 & \Sigma_1\end{bmatrix}\begin{bmatrix}I & I\\ I & 0\end{bmatrix} = \begin{bmatrix}\Sigma_1+\Sigma_2 & \Sigma_2\\ \Sigma_2 & \Sigma_2\end{bmatrix}$$

□

The main implication of this theorem combined with the theorem 2.2.4. (that tells us that the

[12] We apply the independence of W from Z

components of a multivariate Gaussian random vector follow multivariate Gaussian distribution with the parametrisation described in the partition) is that the marginal distribution of Y is what we claimed it to be before the theorem. The hierarchical model described above is often replaced with model where Y has been marginalised, which gives the model:

$$\theta \sim \pi(a)$$
$$Y|\theta, X \sim N(0, I\sigma^2 + K(X,\theta)).$$

Now that we have shortly discussed model fitting, prediction, evaluation and the most relevant simple hierarchical extensions of Bayesian GP models, we can move to the time-series specific aspects of this thesis.

Chapter 4

Time Series With Gaussian Processes

4.1 Theoretical Connections and practical tools of Gaussian Process time series

This chapter introduces ways to enhance the GP approach as a tool for time series modeling. We consider three topics: processes that undergo change in nature in their lifetime, additive processes and multiple outputs that are (possibly) correlated. None of these topics are exclusively related to time series applications of GPs, but all of them have been used in modeling time series with GPs and have shown practical utility.

On the other hand, some topics that are relevant for some temporal applications like GP models that are interactively updated as data becomes available[1] and convolutions of kernels[2] have been left out. Our primary interest are tools that are useful for trend detection and are easy to interpret instead of (especially real time) going for the best but possibly intractable predictive inference.

[1] Osborne, Garnett, and Stephen Roberts 2010
[2] Thobar, Bui, and Turner 2015

This focus is motivated by the application domain discussed in the next chapter, modeling of historical phenomenon.

As in the chapter 2, we only nod to the research literature that covers the more theoretical aspects of the topic. GPs are related to many stochastic processes which have both theoretical and practical value. Ornstein-Uhlenbeck Process, the significant extension of Brownian motion, can be intrerpreted as a Gaussian Markov Process (GMP).[3] The relationship between Kalman filters, a standard tool of time series prediction, and GPs has been extensively studied.[4]

It should be emphasized that often these tools are used when GP is a part of a larger model, often more complicated than the simple extensions mentioned in the last chapter. For example, GPs can be used to model the effect of environmental covariates in species distribution modelds that also include other terms (that can also be GPs) like noise and offset[5] or to describe one (group of) parameter(s) in a more complicated equation that describes population replacement and growth[6]. But as more complicated models tend to be problem specific, we focus to the basic components that make the building blocks of such grand constructs.

4.2 Modeling Suddenly Changing Phenomenon With Gaussian Processes

One major problem in time series modeling is the fact that many temporal processes that are for most of the time quite staple can still can go over sudden changes that change the behavior of

[3] C. Rasmussen and Williams 2006, p. 212

[4] Hartikainen and Särkkä 2010,Reece and Stephen Roberts 2013,Reece and Stephen Roberts 2010,Douglas Leith and Ringwood 2004

[5] Vanhatalo, Hartmann, and Veneranta 2020

[6] Vanhatalo, Li, and Sillanpää 2017

the phenomenon significantly. For example, the development of Gross Domestic Product can be predicted to some extent for long periods of time, but wars, institutional crises and pandemics can result in very fast drops or changes of dynamics quickly.

The concept of changepoint was developed to model such changes. It is noteworthy that changepoint as a technique is not limited to GP modeling – in fact it first emerged in other frameworks – but here we focus on Bayesian implementation with GPs.

As the name implies, changepoints are points of change. In temporal case, changepoint is a point of time that divides the GP to two domains (if there are more than one changepoint, the number of domains increase as well) with different kernels. It is possible that the parametrisation before and after the changepoint is completely different, but it can also be limited only to the smoothness, periodicity, output scale or other specific property of the process.

In the Bayesian approach, the changepoint is a parameter of the kernel. In fact it is a relatively nice hyperparameter to interpret and specify, as the the timescope of the process we are investigating gives us a way to narrow its prior distribution to a specific domain, and it has a clear interpretation[7]. Below we present some kernels that apply changepoints. The examples are cases with only one changepoint x_c.

The most general case is to have two completely separate kernels for the process before and after the changepoint.[8] As the idea of the approach is to model a situation where the changepoint divides the process to two completely separate domains, the covariance between before and after

[7] These properties do not guarantee the ease of model fitting though.

[8] The treatment of the changepoint -based GP kernels is mostly based on Garnett, Osborne, and Stephen Roberts 2009, pp. 347–349

the changepoint is assumed to be 0. In this case, the formalism for the kernel is as follows:

$$k(x_1,x_2,x_c,\theta) = \begin{cases} k_1(x_1,x_2,\theta_1), \text{if} x_1,x_2 < x_c \\ k_2(x_1,x_2,\theta_2), \text{if} x_1,x_2 \geq x_c \\ 0, \text{otherwise} \end{cases}$$

If both kernel functions are valid and the inputs are given in an ascending order (e.g. x_1 is the smallest), it is possible to prove that the kernel defined in this manner is valid.

Proof. For any set of observations, we can write the complete covariance matrix as a similar composition as seen in the chapter 2. We can then write the different submatrices of this composition as eigendecompositions of symmetric matrices and develop further to obtain:

$$K(x,\theta) = \begin{bmatrix} K_1 & 0 \\ 0 & K_2 \end{bmatrix} = \begin{bmatrix} Q_1\Lambda_1 Q_1^T & 0 \\ 0 & Q_2\Lambda_2 Q_2^T \end{bmatrix} = \begin{bmatrix} Q_1 & 0 \\ 0 & Q_2 \end{bmatrix} \begin{bmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{bmatrix} \begin{bmatrix} Q_1^T & 0 \\ 0 & Q_2^T \end{bmatrix}.$$

As the last form is an eigenvalue decomposition with all of the diagonal values of the matrix in the centre non-negative (since they are the eigenvalues of two positive semi-definite matrixes) all the eigenvalues of the entire matrix are non-negative, and we can conclude that it is positive semi-definite.

The matrix $K(x,\theta)$ is also symmetric. We can see this by going through three different possibilities and relying on the kernels k_1 and k_2 being valid.
If $x_i < x_c$ and $x_j < x_c$, then $k(x_i,x_j) = k_1(x_i,x_j) = k_1(x_j,x_i) = k(x_j,x_i)$.
Also if $x_i \geq x_c$ and $x_j \geq x_c$, then $k(x_i,x_j) = k_2(x_i,x_j) = k_2(x_j,x_i) = k(x_j,x_i)$.
Otherwise $k(x_i,x_j) = 0 = k(x_j,x_i)$. □

The practical implication of this result is that as long as the kernel of the process is valid before and after the changepoint, the joint kernel will be as well.

We can also change only some part of the kernel. Here we consider two such possibilities for isotropic covariance functions that have parameters that control the smoothness and output scale of the process. The changepoint can change the smoothness of the process or the scale of its values. An example of such kernel is the Exponentiated Quadratic kernel introduced in the chapter 2. Here we use somewhat more abstract notation to include other kernels of similar kind. Our general isotropic kernel that we assume to be valid ($\theta = (\alpha, \sigma)$) and which can be thought as the function of $\frac{|x_c - x_j|}{\sigma}$ is

$$k_I(x_i, x_j, \theta) = \alpha^2 k(\frac{|x_i - x_j|}{\sigma}).$$

Let us now extend this kernel to handle a change in the smoothness of the process, so that we have $\theta = (\alpha, \sigma_1, \sigma_2)$ Let us mark this kernel $k_{I_c}(x_i, x_j, \theta, x_c)$. We can define the kernel in the following manner

$$k_{I_c}(x_1, x_2, x_c, \theta) = \begin{cases} k_I(x_1, x_2, \alpha, \sigma_1), \text{if} x_1, x_2 < x_c \\ k_I(x_1, x_2, \alpha, \sigma_2), \text{if} x_1, x_2 \geq x_c \\ \alpha^2 k(\frac{|x_c - x_i|}{\sigma_1} + \frac{|x_c - x_j|}{\sigma_2}), \text{otherwise} \end{cases}$$

We once again assume that the inputs are given in an ascending order. This kind of kernel has two pleasing properties. First, it is possible to prove that it is positive semi-definite[9] and the fact that it is symmetric can be checked with the exactly same method that was used in the previous proof, hence it is valid. Second, it allows us to describe such a change in the time series where the absolute covariance will often be limited (for example with the QE kernel) by the same upper limit (a^2 with the notation used here) but the time span at which great variation in the values of the process

[9] Garnett, Osborne, and Stephen Roberts 2009, pp. 347–349

is likely to occur changes. For example, the stockmarket during Covid-pandemic certainly became less stable. A more nuanced change than a complete independence between different periods of time.

The change of the output scale can be formalised in a very similar manner. We yet again assume ascending order for the inputs. As we limit ourselves to the same group of kernels as with the change of smoothness, we use the same notation k_{I_c}. However, now the parameter is $\theta = (\alpha_1, \alpha_2, \sigma)$ and the kernel is defined as:

$$k_{I_c}(x_1, x_2, x_c, \theta) = \begin{cases} \alpha_1^2 k(\frac{|x_i - x_j|}{\sigma}), \text{if} x_1, x_2 < x_c \\ \alpha_2^2 k(\frac{|x_i - x_j|}{\sigma}), \text{if} x_1, x_2 \geq x_c \\ \alpha_1 \alpha_2 k(\frac{|x_i - x_j|}{\sigma}), \text{otherwise} \end{cases}$$

Likewise, this kernel allows us to be more specific about the change (now a change in the scale of variance and covariance).

It is also possible to use smooth transitions from a kernel to another instead of fixed changepoints. To do this, we consider the following composition of two valid kernels.[10]

$$k_g = k_1(t_1, t_2, \theta_1) g(t_1, \phi) g(t_2, \phi) + k_2(t_1, t_2, \theta_2)(1 - g(t_1, \phi))(1 - g(t_2, \phi))$$

where g is some kind of Sigmoid[11] function. This kind of kernel transforms from mostly being defined by k_1 to mostly being defined by k_2, ϕ and the specific form of the Sigmoid function

[10] For this approach, see Duvenaud 2014, p. 20

[11] For the definition of a Sigmoid function, see for example Murphy 2012, p. 21.

controls how fast and where this transition occurs. To proof that it is valid is straightforward.

Proof. We have proven that kernel of the form $f(t_1)f(t_2)$ that maps to the real space is valid, $g(t_1,\phi)g(t_2,\phi)$ and $(1-g(t_1,\phi))(1-g(t_2,\phi))$ are both kernels of this form, so they are valid. The proof reduces to the fact that multiplication and addition preserve validity. □

Third way to model change presented in the research literature is to - instead of using multiple (possibly) stationary kernels - use non-stationary (ns) kernels. One way to construct suitable ns kernels is to take a suitable stationary kernel, in our example the EQ, as the basis, and transform its inputs so that the end result will be ns. Let $k_{se}(t_i,t_j,\theta)$ be the EQ kernel and $\omega(t,a,b) = 2c(-\frac{1}{2} + \frac{1}{1+e^{-a(t-b)}})$. By somewhat truncating the notation, we will obtain the ns kernel ($\phi = (\theta,a,b)$)

$$k_{ns}(t_i,t_j,\phi) = \theta_1^2 \exp\left(-\frac{(\omega(t_i,a,b)-\omega(t_j,a,b))^2}{2\theta_2^2}\right).$$

In this approach, there are three predefined parameters a,b and c. The parameter b controls the location where the process will undergo rapid change, the parameter a controls the size of the time window in which the process "activates" and the parameter c controls the maximum range.[12] The parameters that are fitted based on the data are the same as with regular EQ kernel. The kernel is clearly valid as a transformation (as defined in chapter 2) of valid kernel.

These tools extend the already very flexible modeling capabilities of GP by a significant degree, as they allow vaguely described GPs to undergo changes in their development. This allows the modeling of unstable phenomenon, like EEG measurements of epileptic patients[13] and weather data

[12] This approach, based on earlier notions about non-stationary kernels, was introduced in Cheng et al. 2020, p. 8

[13] Garnett, Osborne, and Stephen Roberts 2009, pp. 350–351

provided by sensor networks.[14] A novel application domain we discuss in more detail in chapter 5 are historical processes like print production, that can quickly change during war, revolutions and other large-scale human-induced events.

4.3 Additive Gaussian Processes

Time series modeling can be done with the flow of time being the only explanatory factor of the process, but very often the real interest is to understand the relation between temporal measurements of some other variables and the process values. There are various ways to do such multi-feature -analysis with GPs, here we consider Additive GPs.[15]

The first key idea of Additive GPs is, that it can be seen as a sum of several GPs that are independent. In the GP framework, the formalism would be that with any finite set of observations, with $f_i, i \in \{1,...,n\}$ being finite subsets of independent GPs, the Additive GP follows the multivariate Gaussian distribution

$$f_{add} = f_1 + ... + f_n \sim N(\sum_i \mu_i, \sum_i K_i).$$

Each of the subkernels (the kernels of the individual processes that are part of the sum), only take some subset of features as input. Here we consider the situation in which each of them takes strictly 1 dimension of the feature space as input. The immediate benefit is that we can interpret the process as a sum of very simple processes, and after the model fitting, we can regress not only with the entire model, but with individual GPs that together constitute the model. In this manner,

[14] Osborne, Garnett, and Stephen Roberts 2010

[15] For the article that, given the articles surveyed in this thesis, used this term for the first time, see Duvenaud, Nickisch, and C. E. Rasmussen 2011

Additive GPs can be used to detect subtrends of the investigated process.

One way to use additive processes in time series modeling is to model multiple time series of the same process (e.g. the development of some health indicator of multiple test subjects) at once, and to use different kernels to reflect shared and separating properties of the time series. Let us assume J time series, each with $y_{j1}...y_{jn_j}$ observations and related features x_{ji}. By fitting a GP over the vector $y = (y_{11},...,y_{1n_1},...,y_{J1},...,y_{Jn_J})$ and using matrix of features X ordered the same way as the input for kernels, we can use GPs in the modeling of several time series. One subkernel can capture the shared effect that only depends on time, another can capture group effects like ethnicity or medical history[16], third can capture the specifics of each individual time series (e.g. of a single test subject with a binary variable) etc. This approach was used in a recent article that introduced a GP framework for longitudinal analysis.[17]

Analogous to linear regression, we can construct interaction terms between features. Here $t_i = (t_{i1},\ldots,t_{id})$, where t_{ij} is the feature relating to observation i and subkernel d. One approach is to start with a collection of subkernels, and to build covariate terms as kernel products in the following manner: Let d be the number of dimensions in the feature space. We wish to model this data with an Additive GP, so its kernel is of the form

$$k(t_i,t_j,\theta) = \sum_{l=1}^{d} k_l(t_{il},t_{jl},\theta_l).$$

[16]Even though we have not discussed them, there are kernels for categorical and binary features as well
[17]Cheng et al. 2020

Given this kind of kernel and its part, nth order additive kernel is defined as[18]

$$k_{add_n}(t_i, t_j, \theta, \sigma_n) = \sigma_n^2 \sum_{1 \leq v_1 < v_2 \ldots, < v_n \leq d} \prod_{l=1}^{n} k_{v_l}(t_{iv_l}, t_{jv_l}, \theta_{v_l}).$$

Some special cases of this are the first order interaction $k_{add_1} = \sigma_1 \sum_{l=1}^{d} k_l(t_{il}, t_{jl}, \theta_l)$ and the dth order (highest) interaction term $k_{add_d} = \sigma_d^2 \prod_{l=1}^{d} k_l(t_{il}, t_{jl}, \theta_l)$. This formulation has its own limitations: all terms of the same interaction level share the same variance parameter and the parameters θ are the same of those of the base kernels.

There is some empirical proof that additive GP models enhanced with additive kernels are able to mitigate one of the major disadvantages of GP modeling done in multi-dimensional feature space: the curse of dimensionality.[19] This curse refers to the phenomenon of GP models not generalizing well outside those areas of the feature space where training data is available. Especially the curse refers to such situations where prediction is done to points that lie outside the training data domain w.r.t. several dimensions of the feature space.

For some reason, the possibility of forming interaction terms of continuous features by multiplying features themselves and using these interaction features as inputs of a subkernel is not discussed in the literature surveyed in this thesis. It is possible that this possibility is too obvious to be discussed – a very recent algorithm for the formation of interaction terms did not in any way discuss the formation of interaction terms between continuous features[20] – but it makes it possible to model different interaction terms separately from each other, unlike the additive kernels construction, that forces them the same variance parameter. The additive example at chapter 5 of this thesis had separate kernels with their own parametrisations for continuous interaction terms.

[18] This conctruction is from Duvenaud, Nickisch, and C. E. Rasmussen 2011

[19] Duvenaud, Nickisch, and C. E. Rasmussen 2011, pp. 7–8, Duvenaud 2014, p. 84

[20] Cheng et al. 2020, supplement

4.4 Multiple-output Gaussian Processes

Often our interest is to model not only one but multiple time series at once. The approach introduced in the previous subchapter has been used in the context of modeling multiple time series of the same phenomenon, but often there is the need to consider multiple somehow related outputs from different processes. Some extensions are based on generalisations of multivariate Gaussian distribution, others on convolutions of kernels. The approach chosen here is easy to interpret and has seen recent use in applications.

Our starting point is the same as in the last section. Let $f_1 ..., f_d$ be independent GPs with respective mean and covariance functions k_j, m_j with kernel function parameters θ and mean function parameters β and let Σ be a $d \times d$ symmetric positive-definite matrix. We also set that the kernels are correlation kernels. The unique (since the matrix is positive-definite) Cholesky decomposition of the previous matrix is $L = chol(\Sigma)$. We now define Multivariate Gaussian Process (MGP) as

$$g(t) = \begin{bmatrix} g_1(t_1) \\ \vdots \\ g_d(t_d) \end{bmatrix} = Lf(t) = \begin{bmatrix} L_{1,1} & \cdots & 0 \\ \vdots & \ddots & 0 \\ L_{J,1} & \cdots & L_{J,J} \end{bmatrix} \begin{bmatrix} f_1(t_1) \\ \vdots \\ f_d(t_d) \end{bmatrix}.$$

Note that MGP allows these processes to have their own inputs by setting $t = (t_1 ..., t_d)$. This formalism is somewhat analogous to the construction of a Multivariate Gaussian Distribution when it is defined as a linear combination of independent standard normal distributions, where the linear transformation is the Cholesky decomposition of the covariance matrix of a multivariate Gaussian distribution. The difference is, that here the linear transformation is made for a group of independent GPs that are themselves multivariate Gaussians in any finite set of indexes.

It is relatively easy to derive the distributional form of the MGP. First we note that (using the independence of GPs)

$$\text{Cov}(g(t), g(t')) = L\text{Cov}(f(t), f(t'))L^T = \sum_{i=1}^{d} k_i(t,t',\theta_i)L_i L_i^T$$

where L_i denotes the ith column of the Cholesky decomposition. Similarly, we can write the covariance between different times and between different processes

$$\text{Cov}(g_i(t_i), g_j(t_j)) = L_{i,\cdot}\text{Cov}(f(t_i), f(t_j))L_{j,\cdot}^T = \sum_{l=1}^{d} k_l(t_i,t_j,\theta_l)u_l(i,j)$$

where $u_l(i,j) = L_l L_l^T{}_{[i,j]}$.

Now, given n_j observed values from jth component of g, we can write $g_j = [g_j(t_{j,1})..., g_j(t_{j,n_j})]$, where the first index of t now highlights the component to which the indice is related to. Jointly, we can write $g = [g_1^T ..., g_d^T]^T$. Now, g, with a given parametrisation θ follows multivariate Gaussian distribution of dimensionality $\sum_{j=1}^{d} n_j$. The formal presentation uses the components we derived just before:

$$\begin{bmatrix} g_1 \\ \vdots \\ g_d \end{bmatrix} |\phi \sim N(\begin{bmatrix} m_1 \\ \vdots \\ m_d \end{bmatrix}, \sum_{l=1}^{d} \begin{bmatrix} u_l(1,1)K_{l1,1} & \cdots & u_l(1,d)K_{l1,d} \\ \vdots & \ddots & \vdots \\ u_l(d,1)K_{ld,1} & \cdots & u_l(d,d)K_{ld,d} \end{bmatrix}).$$

With the following definitions:

$$K_{li,j} = \begin{bmatrix} k_l(t_{i,1},t_{j,1},\theta_l) & \cdots & k_l(t_{i,1},t_{j,n_j},\theta_l) \\ \vdots & \ddots & \vdots \\ k_l(t_{i,n_i},t_{j,1},\theta_l) & \cdots & k_l(t_{i,n_i},t_{j,n_j},\theta_l) \end{bmatrix}$$

and $\phi = \{\theta_1..., \theta_d, \beta_1..., \beta_d, \Sigma\}$.

This form of MGP Distribution has its own name, Linear Model of Coregionalisation (LMC), and the variant presented here is fairly new formalism for it.[21] As noted in the definitions, we can have a differing number of outputs for the different components of the process, and the end result will still be a Multivariate Gaussian Distribution. This formulation also has the advantage that, since L is unique for each Σ, we do not face same kind of optimisation problems as with some other approaches where the parameter corresponding to L might not be unique. Σ also offers an understandable interpretation for interrelations between the different processes.

Model fitting becomes quite a bit more complicated due to the need to estimate the covariance matrix Σ, one of the reasons why the demo projects of the next chapter that were implemented with limited computational resources will not include an example using this construction, but the analytical results related to Gaussian Distribution can be used the same way as with one-dimensional GPs. Namely, we can predict values of the process in unknown locations of the feature space with the conditional distribution of a Multivariate Gaussian Distribution, given that we have chosen a parametrisation for the model first.

We also introduce another approach to modeling multiple-output GPs. This approach is more restrictive , but it is easy to implement with the full probabilistic MCMC approach used in this thesis.[22] This makes sense, given that the source for this model (although it is clearly at least inspired by earlier derivations) is the official Stan User's Guide.[23]

[21] For LMC and good discussion about the matters discussed here, see the chapter 2.2 and 2.3 of Hartmann 2019

[22] Although this approach was not finally used in the next chapter either, the technical implementation would have been easy.

[23] See the chapter 10 of Stan n.d.

This approach applies separate kernels for temporal (or in general, feature) and output-channel covariance, and presents the full model in terms of the Gaussian Matrix distribution. Lets assume J separate time series that share the same feature inputs $t_1 ..., t_n$.

Following the notation established for the multivariate Gaussian distribution in the chapter two let K_U be a matrix of feature-wise (or in general column-wise) covariances defined by the kernel $k_U(t_i, t_j, \theta), i, j \in (1..., n)$. This kernel associates observations w.r.t to their temporal/feature-wise similarity.

Another Kernel $k_V(\phi)$ is defined by its parameters, not using features as inputs. It defines $J \times J$ positive definite covariance Matrix K_V that sets the covariances between different (in our case temporal) processes (or in general the row-wise covariances). Jointly, in the multivariate Gaussian form the covariance matrix is defined by the kernel $k(t_i, t_j, \theta, \phi, m, m') = k_U(t_i, t_j, \theta) K_V(\phi)_[m, m']$, where m and m' stand for different output channels.

The Stan manual recommended restrictions for parameters in order to ensure that the model will work. With $k_U = \theta_1^2 \exp(\frac{-(t_i - t_j)^2}{2\theta_2^2})$ being the EQ kernel, the following fixed parametrisation is needed: $\theta_1 = 1$. This is to ensure that the model is identifiable. Identifiable model is such for which each parametrisation defines a distinct probability distribution in a non-zero measured set (so that distributions defined by two different parametrisations are genuinely different), which implies that this condition is for ensuring that there will be an unique optimal parametrisation for the model. This is unfortunate limitation. On the other hand, the presentation of the model with matrix normal distribution

$$X_f \sim MN(m(x), K_U, K_V) \tag{4.1}$$

is very easy to interpret, as the covariance consists of two parts that relate to different axis of

difference (between processes and between feature values).

4.5 Conclusion

The techniques treated in this chapter significantly extend the range of temporal modeling that can be done with GPs. The basic properties that makes Bayesian GP regression – either as a final or middle layer of a Bayesian Hierarchical model – so useful are the possibility of modeling only vaguely understood phenomenon on the one hand and the possibility of encoding prior information about the process in the form of hyperparameters and the choice of kernel on the other. Additionally, the analytic results for prediction and the marginal likelihood ease model fitting and predictive inference.

Techniques like Changepoints and non-stationary kernels make it possible to model processes that undergo either sudden or transitionary change during their lifetime. Additive processes allow a natural way to consider multiple features that can be used in addition or instead of the time itself as the feature inputs of the kernel(s). LMC and other multiple-process extensions allow us to model interrelated time series within the same framework that builds on multivariate Gaussian distribution as in the univariate case.

As the coverage of cited sources proves, GPs and the tools introduced here are used in a range of temporal applications. Ecological modeling, sensor networks and signal processing (understood broadly) are well established fields of application, and there is emerging interest to apply GP framework in longitudinal studies. However, our main motivation is to explore an entirely new potential application domain.

We have now set up a reasonable machinery for the modeling of temporal phenomenon with GPs.

Next chapter demonstrates their usage in practice in a novel application domain, Bibliographical Data Science.

Chapter 5

Gaussian Processes in Bibliographical Data Science

5.1 Gaussian Processes in the Research of History

This chapter examines the possibility of using Bayesian GP time series in a novel application domain: human history. Natural history is already an application field[1], but GP approach is lucrative for the examination of human-related activity of the past as well. On the one hand, GP approach allows us to be data driven, which is desirable, as most historical phenomenon can not be modeled by explicitly describing the dynamics. On the other hand, the kernels of a GP model can be interpreted in a meaningful manner, which allows us to determine which kind of developments we are trying to model. This also means that the posterior distributions often have a meaningful interpretation, as we will see.

Quantitative analysis of long-term historical processes is an age-old idea, but the developments that

[1] As already cited, the behavior of the river Nile has been modeled with the changepoint-approach

have brought the idea very close to reality are quite recent. The approach to which the quantitative examination in thesis builds on is called Bibliographical Data Science (BDS). The BDS-approach is about using bibliographical metadata – collected and organized by national cultural institutions like the British and Finnish national libraries – of print products to examine long-term development of intellectual, cultural, social and economic history. It has been demonstrated that even simple measures like the annual count of distinct print products can offer valuable insight into early modern history[2], and it has been speculated that books and the development of reading might have been the primary contributor to the decrease of violence in the West during the early modern period[3]. This chapter aims to examine how GP models could possibly be used to enhance (primarily) bibliography-based quantitative analysis of history.

In addition to bibliographical metadata, data sets on economic development, urbanisation, literacy rate(s) and other socio-political indicators are increasingly produced and available for historical research. However, they amount of temporal measurements is often small, and our first example considers the possibility of extending scarce time series with GPs. Our approach in the following is exploratory: instead of going in depth to any demonstration, we tested different ways in which GP time series could be used in historical research. Some of findings were interesting enough that they should developed and scrutinised more thoroughly.

5.2 Data Extrapolation With Gaussian Processes

Historical quantitative data is often very hard and laborious to obtain before the era of modern administration, and they are often reliant on interpretation made by the researcher. For example, the rate of literacy or writing capability varies drastically based on the definition used. This makes

[2]For BDS and related research, see Lahti and ”et.al.” 2019 and Lahti, Ilomäki, and Tolonen 2015
[3]Pinker 2011, pp. 171–179

comparison of different estimates hard. On the other hand, temporal and geographical span of individual projects to produce these statistics is often limited or otherwise compromised. For example, perhaps the most comprehensive data set of early modern urbanisation in Europe includes an estimate for most European countries from 1500 to 1800, but the number of temporal measurements for each is only 6.[4]

In cases where there is a reason to assume some kind of continuity between measured values, GP Regression might be useful for extending such data sets. In Geostatistical literature, GP Regression is called kriging and it is used to predict the existence of ores in an area based on a limited amount of test spots.[5] Similar approach could have practical value in the extrapolation of historical data to produce data-driven educated guesses about the missing values. In this thesis we extended estimates of total population, proportion of urban population, average male wage and literacy rate of British Isles for the period 1500-1750.[6]

Extrapolation from a small number of data points also highlights the importance of hyperparameter priors. Small initual values of the multiplier parameter of the Exponentiated Quadratic kernel led to predictions that completely discarded the existing data, and the model had to be given priors that enforced it to seek a connection between the training data points.[7] The Figures from 5.1 to 5.4 shows the extrapolations, based on sampling from the posterior predictive distribution. In this and all the cases considered strong priors seem justified, as population and literacy rates do not change overnight. Still, this kind of ”biased” data extrapolation should be more thoroughly evaluated before endorsed, but it could offer better than nothing extensions of scarce historical data in cases where we somehow understand how the object of measurement behaves temporally.

[4]The tables of Vries 1984
[5]C. Rasmussen and Williams 2006, p. 30
[6]For the data sources, see Appendix.
[7]See the Appendix for priors.

In all of the cases, the mean gives a curve that fits the data well, and it was used as the basis for the data used in the additive GP example. The confidence intervals give us some insight to the uncertainty related to each extrapolation, and they wary drastically, being almost negligible in the average-wage example, and very wide in the case of literacy. A more thorough investigation with test data sets would be needed for proper conclusions, but the usage of similar methods in Geostatistics and practical success in the technical implementation of this demo would encourage to investigate historical data extrapolation with GPs further.

The first step would be to incorporate a white noise kernel to the model. It would allows us to express uncertainty about reliability of the data itself. The current model assumes the estimate to be noiseless, which of course is far from the truth. Choose of a proper prior for the white noise kernels parameter would be an interesting and non-trivial case of expert elicitation in itself. The same applies to the next example.

5.3 Changepoints and Major Historical Turning Points

One major problem of quantitative analysis of long-term historical processes is the fact that their dynamics often undergo change during their lifetime. For example, the increase in the number of annual titles in the English Short Title catalogue (ESTC) - roughly speaking the catalogue of early modern print production of the British Isles - develops in a roughly linear fashion for the first half of the seventeenth century, but the English revolution and related events change the entire landscape of print production at the start of the 1640's, and both the volatility and magnitude of print production suddenly change, and the rest of the century looks very different compared to the relatively peaceful start. Moments of change are as such of great interest to historians, and the suddenly increasing number print products at the eve of the English Civil War has obtained numerous

interpretations in the research literature.[8]

However, we can model sudden (or not so sudden) change and detect it from the data with the tools introduced in the previous chapter. There are many takeoff moments of time series in historical literature, but here we focus on the simplest possible sanity check for this kind of approach. The previously described sudden change in the ESTC is perhaps the best known and most extreme example of this kind of behavior I am aware of, and as such it offers a good sanity check for the viability of this kind modeling. It should be quite easy for a model to detect, at least if the peak of 1660's (that resulted from the return of the king from exile), is not part of the same time series of 17th century print production, which is why our final model was tested only with the first half of the century.

Our model was based on two uncorrelated EQ kernels[9] with a single changepoint parameter that had an uniform distribution within the first half of the 17th century. We were successful in the detection of the change at the start of the fourth decade of the Century. This is illustrated by the Figure 5.5 and the diagnostics of Stan[10] that tell us that roughly 75 percent of the posterior mass of the changepoint parameter belongs to the 1640-1642 range, the first of these years being the starting year for the process that lead to the Civil War (the Long Parliament started in London) and the last being the official starting year. The conclusion of this example is that at least the changepoint approach passes the first test of detecting real historical change.

Historical developments can be read from the other posterior parameters as well. The scale parameter of the process increases and the length parameter (controlling how fast the similarity of two

[8] Parker 2013, pp. 367–368. Zaret 2000, p. 174. Zwicker 2003, p. 189. Hill 1987, p. 40. Peltonen 2011, p. 261. McKenzie 2002b, p. 130. McKenzie 2002a, pp. 559–560. Raymond 2003, p. 168

[9] Details and hyperparameter priors in the Appendix.

[10] See the Appendix for the diagnostics.

feature points diminishes) decreases after the changepoint, which could be interpreted as increase in the total number of print and decreasing similarity between less close observations. As print became both more numerous and more volatile during the early 40's, it is good to see that the model captured this property. The lower panel of the Figure 5.5 gives a similar visual intuition, as both the absolute value and the variance of the predictions increase at the start of the Civil War.

5.4 Longue Durée of Early Modern Knowledge Production With GPs

The early modern period of Europe (c.a. 1500-1800 CE) was as time of economic growth, increase of urbanisation and rising levels of knowledge production in Northwestern Europe. A lot has been written about how many of these developments are interrelated – sometimes under the umbrella on concepts like the Enlightenment – but the statistical analysis of how these different processes were related is often quite vague, only referring to the simultaneous increase by several measures of progress like literacy rate, economic growth and increase in book production.[11]

In a very exploratory manner, I tried to model the normalised amount of books produced per capita (bpcn) between 1500 and 1750 in the British Isles as a function of an additive GP with the following components: proportion of urban population (f_1), the proportion of literate population (f_2), the average male wage (f_3), and the covariates of these temporal measurements $urban \times literacy$, $urban \times wage$ and $literacy \times wage$ (f_4, f_5, f_6) in a given time point. In our case, we used every 10th year starting from 1500, ending up with 26 observations. The joint model was $g(t) = f_1 + f_2 + f_3 + f_4 + f_5 + f_5 + f_6 + \varepsilon$, where $\varepsilon \sim N(0, I_{n_y}\sigma^2)$ and all of the f_j had a Exponentiated Quadratic kernel. Jointly, the model was still a GP with the independence assumption.

[11] For a recent example of such barrage of figures, see Scheidel 2019, pp. 372–375.

The goal of this analysis was to make a relatively simple test to see whether GP models could be used in trend-detection of long-term historical processes. Of course, a single data set and fitting of one particular model is not enough for any conclusive proof, but the results would encourage to do a more robust and thorough investigation both in terms of quantitative analysis and from a historical view as well.

Figure 5.6 illustrates the behavior of the individual processes $f_1, ..., f_6$, plotted against time (although the real "x"-value of all the processes is the input related to the point of time) and Figure 5.7 shows how these subtrends accumulate into the full noiseless process. The visualisations lend themselves to some interpretation. The joint effect of wage and urbanisation varies the most, while the proportion of literate proportion remains relatively insignificant for the entire period. Other trends fall between these extremes, showing moderate upwards development. The last panel of the Figure 5.7 shows the final noiseless process that regresses over the observations.

A more thorough investigation would include model evaluation, feature selection and a more comprehensive data set as well. From historical point of view it would be interesting to expand this kind of approach over multiple observed units (cities, countries etc.) with shared and unit-wise features like was done in the already cited article where GP was used for longitudinal modeling. For example, we could test whether features like urbanisation and literacy rate have universal effect on the amount of books per capita, or is the flow of time or a feature representing some distinct quality of the unit (e.g feature that describes the type of publishing present in the city or country) the strongest subtrend explaining bpcn. But this is something to be tested in the future. The argument we try to motivate with this chapter is that the possibilities - of which the examples seen here are only very modest demonstrations - of GP modeling in the research of only vaguely understood historical developments are exciting.

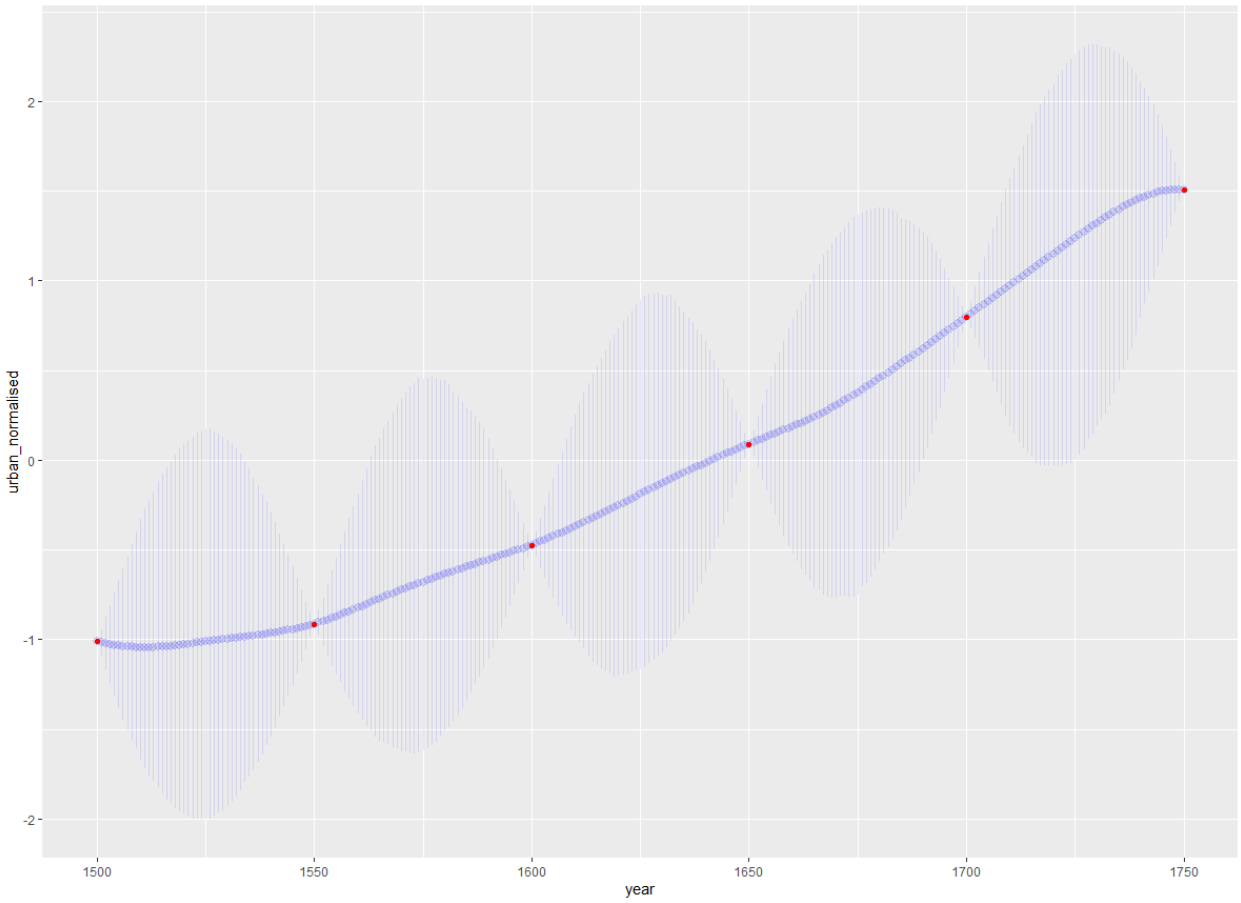

Figure 5.1: Normalised proportion of urban population in the British Isles between 1500-1750 extrapolated. The darker blue line is the mean of the samples from the posterior predictive distribution, the blue interval around it is the 95-percent empirical confidence interval and the red points are the original (normalised) data.

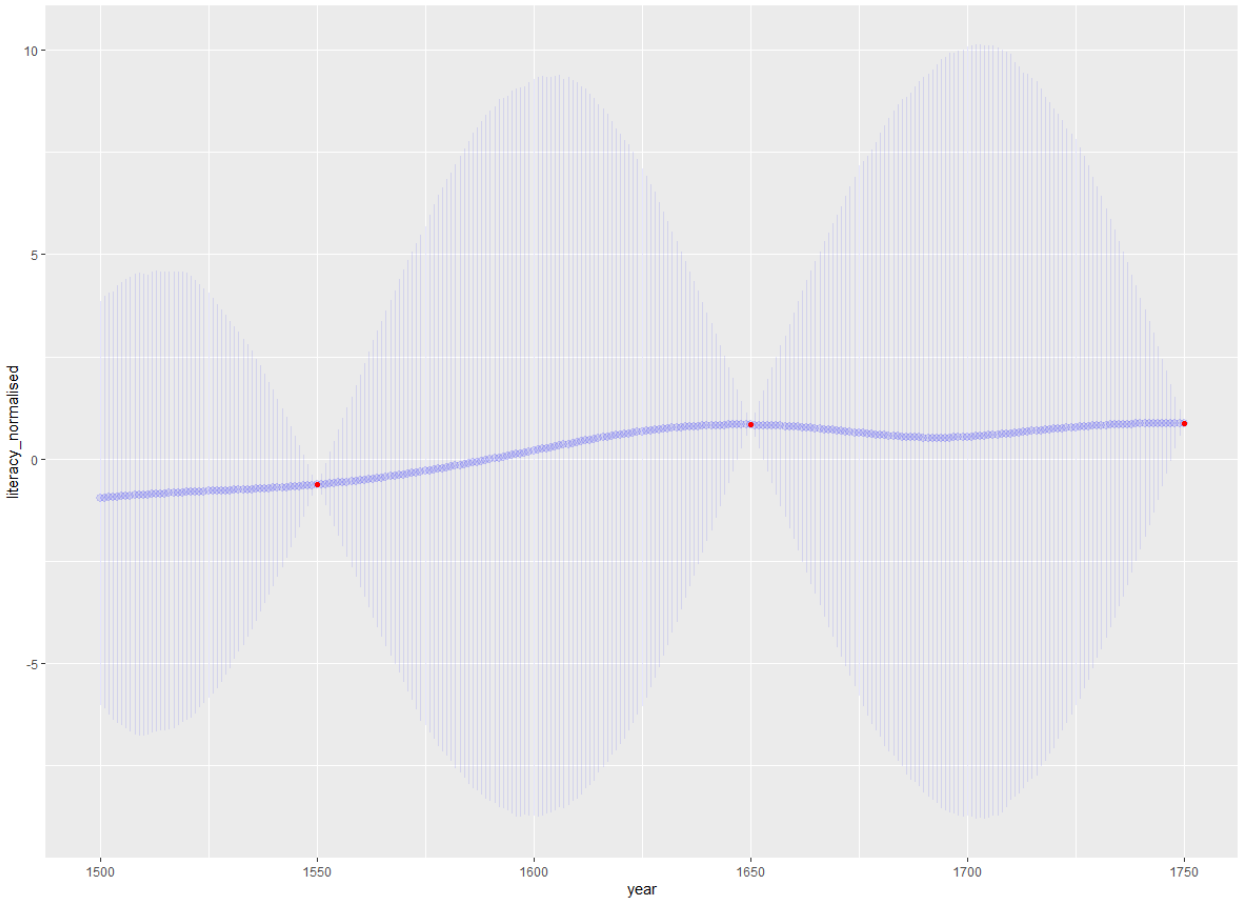

Figure 5.2: Normalised level of literacy in the United Kingdom between 1500-1750 extrapolated. The darker blue line is the mean of the samples from the posterior predictive distribution, the blue interval around it is the 50-percent empirical confidence interval and the red points are the original (normalised) data.

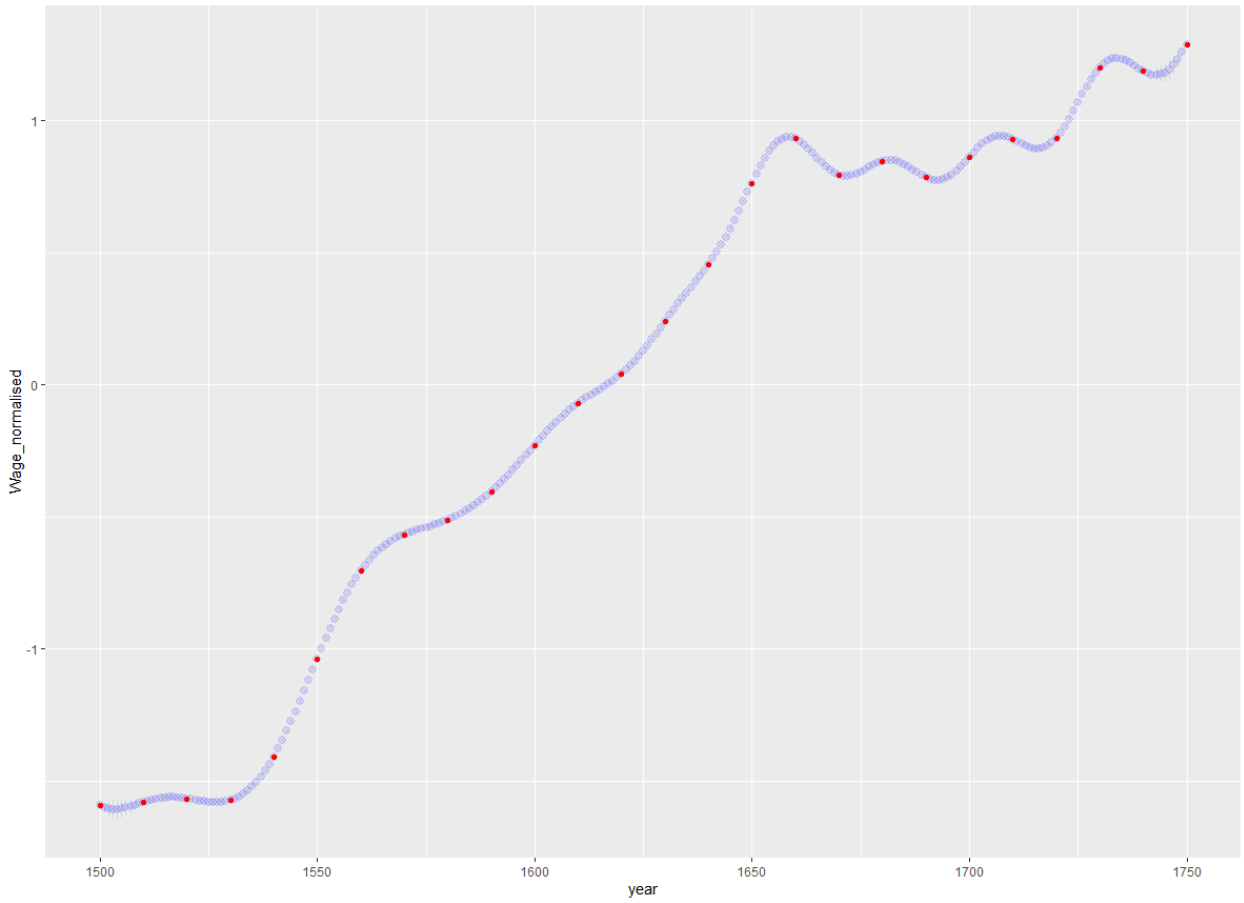

Figure 5.3: Normalised average male wage in England between 1500-1750 extrapolated. The darker blue line is the mean of the samples from the posterior predictive distribution, the blue interval around it is the 75-percent empirical confidence interval and the red points are the original (normalised) data.

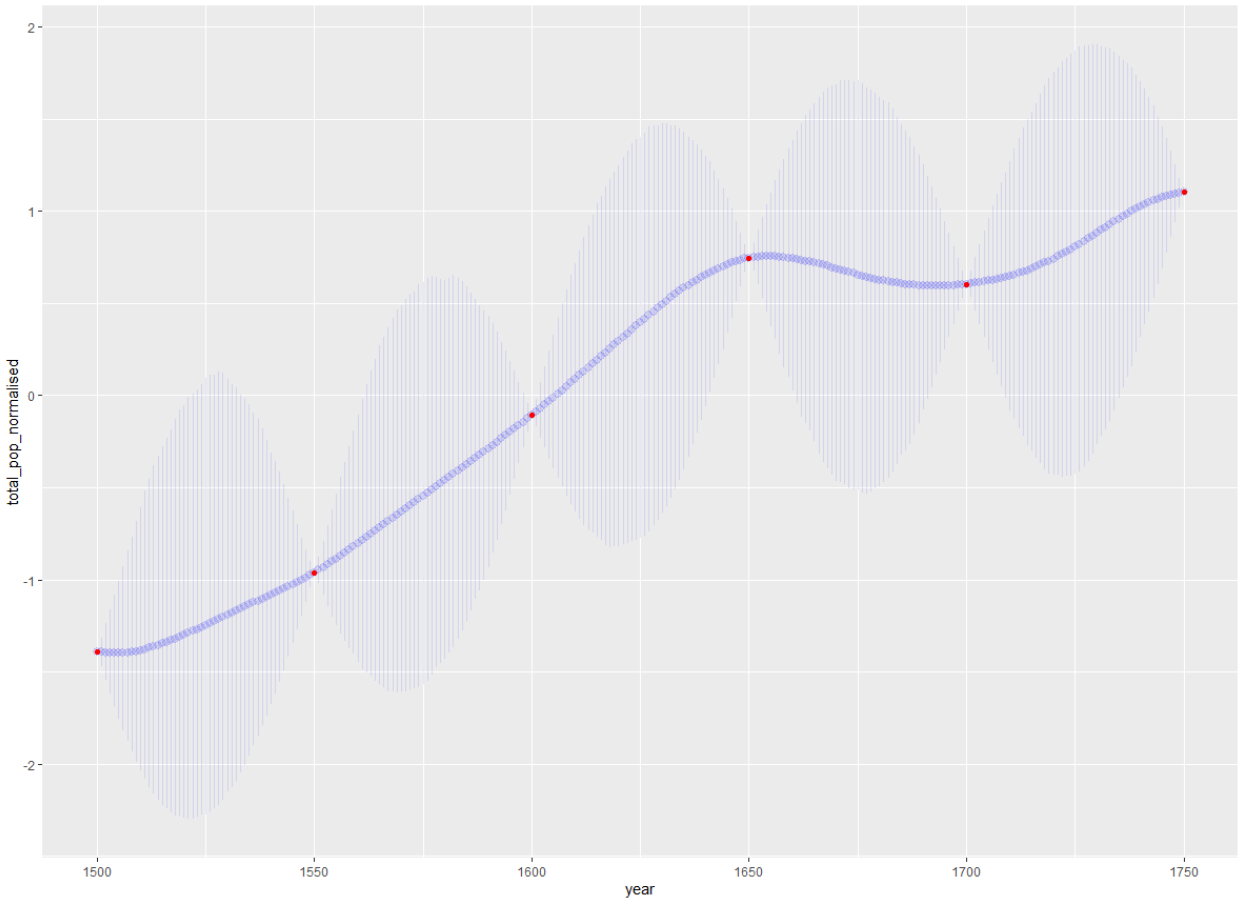

Figure 5.4: Total population in England and Wales 1500-1750 extrapolated. The darker blue line is the mean of the samples from the posterior predictive distribution, the blue interval around it is the 95-percent empirical confidence interval and the red points are the original (normalised) data.

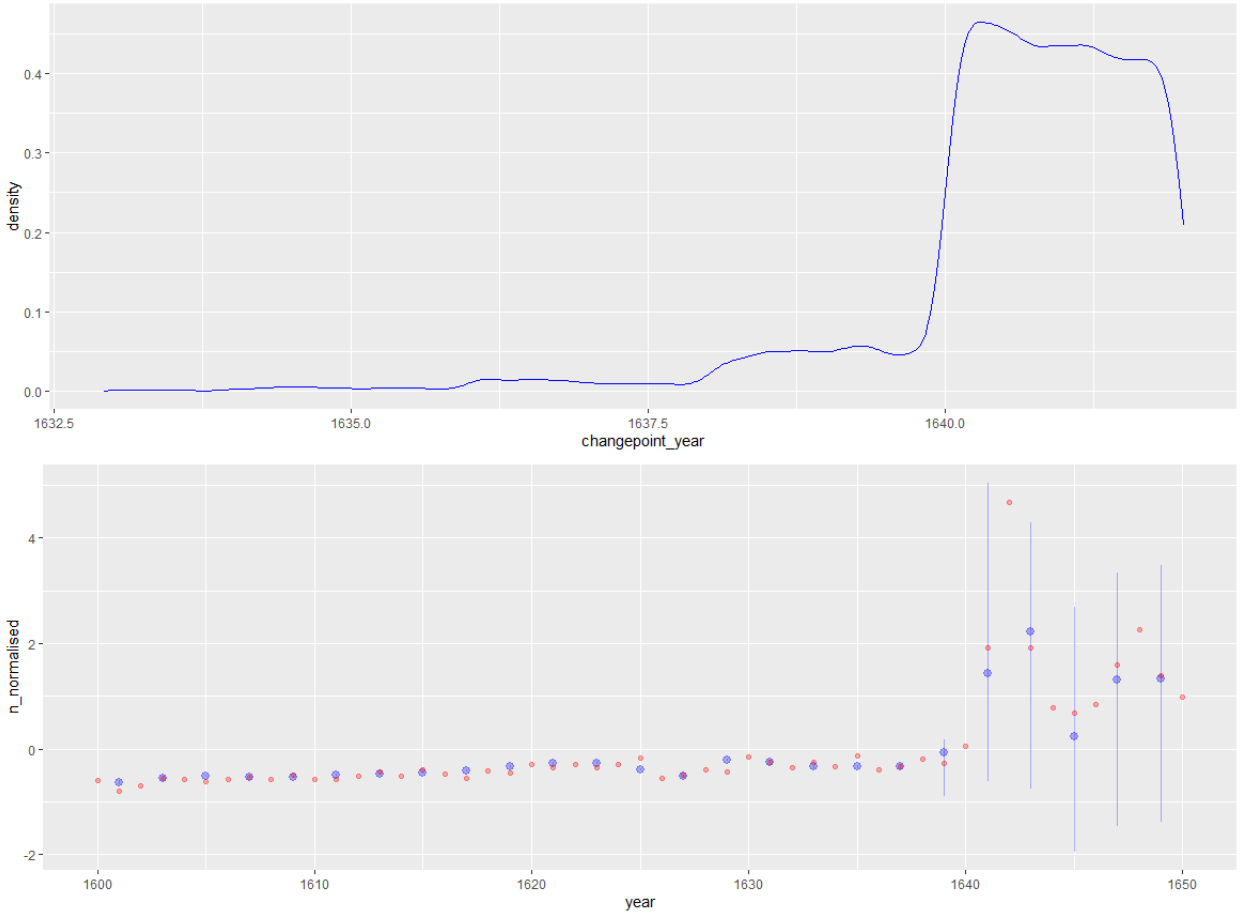

Figure 5.5: The changepoint posterior (upper picture) and the original data and predictions (lower picture) of 17th century book production. The blue predictions are the values of the fitted GP on the time points which were not used to fit the model. ESTC. The red points are the original (normalised) data.

Figure 5.6: Subprocesses of the full additive process, y-axis is plotted w.r.t. to the year of the related feature value. E.g. if average wage was 3 and the related value of bpcn for this subprocess was 0.3 in 1720, then we would plot 0.3 against 1720 for the wage-related subprocess. $f_1 ..., f_6$ from left to right and from top to bottom. 75 percent Confidence interval around the mean

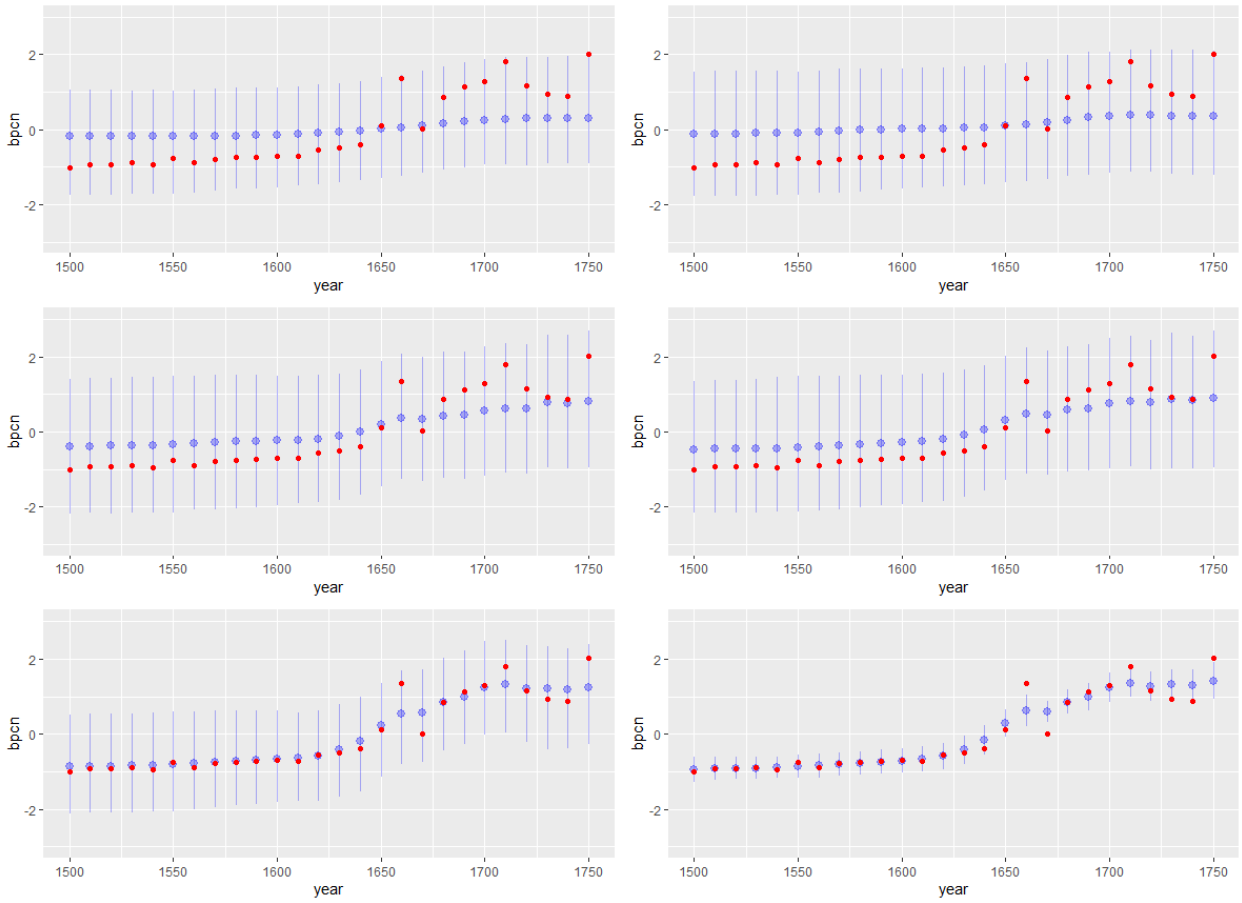

Figure 5.7: Processes $f_1, f_1 + f_2 ..., f_1 + f_2 + f_3 + f_4 + f_5 + f_6$ from left to right and from top to bottom. Same visualisation principle as in the previous figure. 95-persent confidence interval around the mean. In addition, the red points are the original (normalised) data.

Chapter 6

Conclusions

The aims of this thesis were twofold. First, to present the basics and some of the more recent extensions of Bayesian GP models, focusing on developments that have seen usage in the modeling of temporal processes. Second, to explore whether such models could perhaps be used in my other field of academic interest, in the study of History.

The first aim was succesful in the sense that both the basics and some of the more recent advancements were covered. Of course, most of what has been written about time series modeling with GPs had to be left out, but a much greater proportion of that which fulfilled our selection criteria became part of our thesis. The selection criteria was to prefer such techniques of GP modeling that were easy (or at least realistic) to implement in practice and that could be interpreted in a manner that made their use in the modeling of real world phenomenon meaningful. Fortunately, GPs are a very active field of research, also from a practically oriented perspective. Naturally all could not be addressed, but major questions like processes that undergo change in their lifetime, correlated GPs modeled jointly and additive processes were.

Personally, I think that the fact that the tools introduced in this thesis immediately opened me possibilities in my own field of substance – history – and that these possibilities could be demonstrated in practice is a good example of how flexible the GP framework is. It can be used to model processes that are only very vaguely understood, but even then it gives us control over the kind of process we are trying to model, and the Bayesian way of treating parameters allows us to encode our prior knowledge to the model. The more novel contribution of this thesis would be the demonstration of the potential that the GP models have not only in predictive inference, but also in trend detection and in the modeling of phenomenon like historical takeoff moments. The idea of using GPs in either kind of analysis is not new, but in the quantitative analysis of long-term historical developments temporal GPs could be a major tool.

As more and more historical data becomes available for analysis of cultural, economic, demographic and political developments, the possibilities of doing the kind of analysis seen in chapter 5 increase significantly, and it can be done comparatively. It is never wise to say for certain, but hopefully the previous chapter of this thesis will not be the last time GP models are used to model temporal historical phenomenon.

Chapter 7

Bibliography

Chen, Zexun and Bo Wang (2018). "How priors of initial hyperparameters affect Gaussian process regression models". In: *Neurocomputing* 275, pp. 1702–1710.

Cheng, Lu et al. (2020). "An Additive Gaussian Process Regression Model for Interpretable Non-Parametric Analysis of Longitudinal Data". In: *Nature Communications. 10:1798*.

Ding, Shanshan and R. Dennis Cook (2014). "Dimension Folding PCA and PFC for Matrix-Valued Predictors". In: *Statistica Sinica* 24, pp. 463–492.

Douglas Leith, Martin Heidl and John Ringwood (2004). "Gaussian Process Prior Models for Electrical Load Forecasting". In: *8th International Conference on Probabilistic Methods Applied to Power Systems, Iowa State University, Ames, Iowa.*, pp. 109–120.

Duvenaud, David (2014). *Automatic Model Construction with Gaussian Processes*. University of Cambridge.

Duvenaud, David, James Lloyd, et al. (2013). "Structure Discovery in Nonparametric Regression through Compositional Kernel Search". In: *Proceedings of the 30th International Conference on Machine Learning*.

Duvenaud, David, Hannes Nickisch, and Carl Edward Rasmussen (2011). “Additive Gaussian Processes”. In: *arXiv:1112.4394v1 [stat.ML]*.

Garnett, Roman, Michael Osborne, and Stephen Roberts (2009). “Sequential Bayesian Prediction in the Presence of Changepoints”. In: *Proceedings of the 26th International Conference on Machine Learning*, pp. 345–352.

Gelfland, Alan and Erin Schliep (2016). “Spatial statistics and Gaussian processes: A beautiful marriage”. In: *Spacial Statistics*, pp. 86–104.

Gelman, Andrew et al. (2014). *Baeysian Data Analysis*. Boca Raton, Florida.

Gupta, A. and D.Nagar (1999). *Matrix Variate Distributions*. Chapman and Hall.

Harjulehto, Petteri, Riku Klén, and Mika Koskenoja (2017). *Analyysiä Reaaliluvuilla*. Unigrafia.

Hartikainen, Jouni and Simo Särkkä (2010). “Kalman Filtering and Smoothing Solutions To Temporal Gaussian Process Regression Models”. In: *2010 IEEE International Workshop on Machine Learning for Signal Processing*.

Hartmann, Marcelo (2019). *Approximate Bayesian Inference in Multivariate Gaussian Process Regression and Applications to Species Distribution Models*. University of Helsinki.

Hill, Christopher (1987). *The Collected Essays of Christopher Hill: Writing and Revolution in 17th Century England*. University of Massachusetts.

Hörmander, Lars (1983). *The analysis of linear partial differential operators I*. Springer.

Horn, Roger and Johnson Charles (2012). *Matrix Analysis*. Cambridge University Press.

Kurtz, Vince and Hai Lin (2019). “Kalman Filtering with Gaussian Processes Measurement Noise”. In: *arXiv:1909.10582v1 [stat.ML]*.

Lahti, Leo and ”et.al.” (2019). “Bibliographic Data Science and the History of the Book (c. 1500–1800)”. In: *Cataloging & Classification Quarterly* 57.1, pp. 5–23.

Lahti, Leo, Niko Ilomäki, and Mikko Tolonen (2015). “A Quantitative Study of History in the English Short-Title Catalogue (ESTC), 1470-1800”. In: *LIBER Quarterly* 25.2, pp. 87–116.

Lebanon, Guy (2020). *The Analysis of Data*. Publicly available internet book.

Lifshits, M.A (2014). *Random Processes by Example*. World Scientific.

Loyd, James, David Duvenaud, and Roger Grosse (2014). "Automatic Construction and Natural Language Description of Nonparametric Regression Models". In: *arXiv:1402.4304v3 [stat.ML]*.

McKay, David (1998). "Introduction to Gaussian Processes". In: *Neural Networks and Machine Learning*.

McKenzie, Donald (2002a). "Printing and publishing 1557–1700: Constraints on the London book Trades". In: *The Cambridge History of the Book in Britain*. Ed. by John Barnard and Donald McKenzie. Vol. 4. The Cambridge History of the Book in Britain. Cambridge University Press, pp. 553–567.

— (2002b). "The London Book Trade In 1644". In: *Making Meaning. 'Printers of the Mind' and Other Essays'*. Ed. by Peter McDonald and Michael Suarez. University of Massachusetts Press, pp. 126–143.

Murphy, Kevin (2012). *Machine Learning: A Probabilistic Perspective*. The MIT Press.

O'Hagan, Tony (1978). "Curve Fitting and Optimal Design for Prediction". In: *Journal of the Royal Statistical Society. Series B*, pp. 439–468.

Osborne, Michael, Roman Garnett, and Stephen Roberts (2010). "Active Data Selection for Sensor Networks with Faults and Changepoints". In: *24th IEEE International Conference on Advanced Information Networking and Applications*.

Parker, Geoffrey (2013). *Global Crisis: War, Climate Change and Catastrophe in the Seventeenth Century*. Yale University Press.

Pedregosa, Fabian et al. (2011). "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12, pp. 2825–2830.

Peltonen, Markku (2011). "The Development of the Book Trade In Britain". In: *The Oxford History of Popular Print Culture*. Ed. by Joad Raymond and Gary Kelly. Oxford University Press, pp. 252–262.

Pinker, Steven (2011). *The Better Angels of Our Nature: Why Violence Has Declined*. Viking Books.

Rasmussen, Carl and Christopher Williams (2006). *Gaussian Processes for Machine Learning*. The MIT Press.

Raymond, Joad (2003). *Pamphlets and Pamphleteering in Early Modern Britain*. Cambridge University Press.

Reece, Steven and Stephen Roberts (2010). "The Near Constant Acceleration Gaussian Process Kernel for Tracking". In: *IEEE Signal Processing Letters* 17. No. 8, pp. 707–710.

— (2013). "An Introduction to Gaussian Processes for the Kalman Filter Expert". In: *Information Fusion (FUSION), 2010 13th Conference*.

Roberts, S. et al. (2012). "Gaussian Processes for Timeseries Modelling". In: *Philosophical Transactions of the Royal Society A 371: 20110550*.

Scheidel, Walter (2019). *Escape from Rome: The Failure of Empire and the Road to Prosperity*. Princeton University Press.

Siegrist, Kyle (2020). *Libretexts: The Multivariate Normal Distribution*. Available at 'https://stats.libretexts.org/Bo

Stan (n.d.). *Stan User's Guide*. Available at https://mc-stan.org/docs/$2_23/stan-users-guide$.

Teckentrup, Aretha (2019). "Convergence of Gaussian Process Regression with Estimated Hyperparameters and Applications in Bayesian Inverse Problems". In: *arXiv:1909.00232 [math.NA]*.

Thobar, Felibe, Thang Bui, and Richard Turner (2015). "Learning Stationary Time Series using Gaussian Processes with Nonparametric Kernels". In: *Advances in Neural Information Processing Systems*, pp. 3501–3509.

Vanhatalo, Jarno (2019). "Spatial Modelling and Bayesian Inference: Lecture notes 2019". Unpublished course lecture note.

Vanhatalo, Jarno, Marcelo Hartmann, and Lari Veneranta (June 2020). "Additive Multivariate Gaussian Processes for Joint Species Distribution Modeling with Heterogeneous Data". In: *Bayesian Anal.* 15.2, pp. 415–447. URL: https://doi.org/10.1214/19-BA1158.

Vanhatalo, Jarno, Zitong Li, and Mikko Sillanpää (2017). “Gaussian Process Framework For Temporal Dependence and Disperancy Functioncs in Ricker-type Population Growth Models”. In: *The Annals of Applied Statistics*, pp. 1375–1402.

Vehtari, Aki and Janne Ojanen (2012). “A Survey of Bayesian Predictive Methods For Model Assessment, Selection and Comparison”. In: *Statistics Surveys* 6, pp. 142–228.

Vries, Jan de (1984). *European urbanization, 1500-1800*. Harvard University Press.

Wang, Ruye (2020). *Machine Learning e176, online lecture notes (to be published as a book by Cambridge University Press)*. Available at ’http://fourier.eng.hmc.edu/e176’. Book Mathematical Statistics and Stochastic Processes.

Williams, Christopher (1998). “Prediction with Gaussian Processes: From Linear Regression to Linear Prediction and Beyond”. In: *Learning in Graphical Models*. Ed. by Michael Jordan. Springer, pp. 599–621.

Zaret, David (2000). *Origins of Democratic Culture: Printing, Petitions and the Public Sphere in Early-Modern England*. Princeton University Press.

Zwicker, Steven (2003). “Habits of reading and early modern literary culture”. In: *The Cambridge History of Early Modern English Literature*. Ed. by David Loewenstein and Janel Mueller. The New Cambridge History of English Literature. Cambridge University Press, pp. 170–198.

Chapter 8

Appendix

8.1 The Data

The data used in this thesis at Chapter 5 was aggergated from various sources. The figures for the annual number of books were derived from the English Short Title Catalogue (ESTC), given to me for this purpose by the Computational History group of the University of Helsinki. Unfortunately, the ESTC as such is not publicly available, but it can be queried at http://estc.bl.uk. The data regarding early modern urbanisation and literacy is based on publicly available data sets (https://ourworldindata.org) that were extrapolated as part of the ”historical kriging” demo.

The data sets as found from the site are based on previous historical research on urbanisation and literature, and the site explains its method reasonably well (https://ourworldindata.org/literacyhistorical-change-in-literacy and https://ourworldindata.org/grapher/urbanization-last-500-years). The data regarding the average wage was likewise extrapolated with GP model with a EQ kernel, and its source is the data provided by the Historian Gregory Clarke at
http://faculty.econ.ucdavis.edu/faculty/gclark/data.html. All datasets were downloaded between

the May and July of 2020. The books per capita -measure was derived by multiplying the number of distinct print products per year by 1000 and dividing the result with the population number obtained from the kriging demo for population.

8.2 Model Priors and Diagnostics

Observations and features (disregarding time, that was measured in years) were normalised to have mean 0 and standard deviation (sd) 1. This makes assigning priors and interpreting posterior distributions easier. The primary diagnostic tool to check for the convergence of the MCMC algorithm was $\hat{R}$, that can be characterised as a measure of relative similarity of variance between MCMC chains, 1 being the target value that indicates convergence. There is is no exact threshold of acceptable $\hat{R}$ value - and the similarity of chains doesn't actually prove convergence - but one recently used threshold is 1.1.[1] while in the standard work Bayesian Data Analysis 1.05 is used.[2] The final versions of the models of chapter 5 all satisfied this convergence check. In all cases, the priors were independent. Note that some of the screenshots at the end of this chapter include the simulated draws from the posterior predictive distributions of parameters as well. This is because in some cases sampling from the posterior predictive distribution was done within STAN as well, but the relevant parameters are the ones described below.

Stan allows the setting of lower limit for MCMC sampling of parameter values. As we often wanted to limit ourselves to non-negative values, this lower limit was sometimes set to 0 for distributions that by default cover also negative values with positive density. Distributions handled in this manner have a subindex $+$, e.g $N(0,1)_+$ for a zero-centered normal distribution that is only

[1] See the supplementary method 3 document of Cheng et al. 2020
[2] Appendix C of Gelman et al. 2014

sampled for its non-negative half.

8.2.1 Historical Kriging Models

In all the three cases, the kernel was a EQ-kernel with the following notation: $k(t_i,t_j) = \alpha^2 \exp\left(-\frac{1}{2}\frac{(t_i-t_j)^2}{\rho^2}\right)$. Three MCMC chains with 2500 warmup and 2500 real samples each were run for the posterior distribution. In three of the kriging cases, the priors for the parameters were:

$$\alpha \sim N(30,10)_+$$
$$\rho \sim N(30,10)_+$$

And for the literacy-rate kriging they were

$$\alpha \sim N(50,10)_+$$
$$\rho \sim N(50,10)_+$$

The hyperparameter priors highlight how the assumption of strong connections between observations was encoded to the priors.

8.2.2 Changepoint Model

Both the kernel before and after the changepoint were EQ kernels, and here we use the following notation: $k(t_i,t_j) = \alpha_v^2 \exp\left(-\frac{1}{2}\frac{(t_i-t_j)^2}{\rho_v^2}\right)$, where $v \in (1,2)$ marks the time before (1) and after (2) the changepoint. The covariance between the kernels was set to 0. The changepoint parameter is here

marked with c. IG stands for Inverse-Gamma distribution. Four MCMC chains with 4000 warmup and 4000 real samples each were run for the posterior distribution. The priors were:

$$\alpha_1 \sim N(0,1)_+$$
$$\alpha_2 \sim N(0,1)_+$$
$$\rho_1 \sim IG(5,5)$$
$$\rho_2 \sim IG(5,5)$$
$$c \sim U(1600,1650)$$

8.2.3 Additive GP Model

All the subprocesses f_j of $g = f_1 ... + f_6 + \varepsilon$ had EQ kernels, for which we use similar notation as in the previous subchapters. The variable ε stands for the "white noise", meant to represent change in printing that is not part of the long-term trend. Four MCMC chains with 500 warmup and 1500 real samples each were run for the posterior distribution. This model could also be interpreted as the marginalised version of the model with a normal outcome, mean defined by a latent Gaussian

Process and skewed by Gaussian noise, which was discussed at chapter 3.

$$\alpha_1 \sim N(0,1)_+$$
$$\vdots$$
$$\alpha_6 \sim N(0,1)_+$$
$$\rho_1 \sim IG(5,5)$$
$$\vdots$$
$$\rho_6 \sim IG(5,5)$$
$$\varepsilon \sim N(0,1)$$

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
rho	37.590886782	1.909139e-01	11.029048822	15.127797822	30.262732020	3.777999e+01	45.140562741	58.571020539	3337.345	1.0007681
alpha	1.117057112	1.253285e-02	0.602298314	0.548914715	0.783102352	9.759859e-01	1.274486771	2.440537830	2309.529	1.0012778
y2[1]	-1.010860403	1.155320e-05	0.001004550	-1.012860460	-1.011531119	-1.010860e+00	-1.010179290	-1.008933088	7560.297	1.0003131
y2[2]	-1.017017591	4.995640e-04	0.043158410	-1.090446033	-1.029762182	-1.016234e+00	-1.004159089	-0.944376119	7463.603	0.9997103

Figure 8.1: Summary of Posterior Convergence and distributions (alpha and rho being the relevant variables, the table omits most simulated y's) as given by STAN for the final urban population proportion kriging model

parameter	mean	sd	2.5%	25%	50%	75%	97.5%
rho	35.129494561	1.027749e+01	12.12820635	29.0104262258	35.752003265	41.88267396	53.66145887
alpha	1.169669720	5.982654e-01	0.57535265	0.8134220469	1.035515832	1.33923956	2.54789783
y2[1]	-1.387538591	1.017633e-03	-1.38953760	-1.3882303350	-1.387532201	-1.38687045	-1.38562087
y2[2]	-1.391649364	4.980341e-02	-1.47241947	-1.4081769253	-1.391523281	-1.37690069	-1.30414138
y2[3]	-1.394345976	9.706044e-02	-1.55147343	-1.4277873397	-1.394542054	-1.36540151	-1.21568781
y2[4]	-1.396123522	1.372965e-01	-1.62408232	-1.4450241754	-1.396604348	-1.35340783	-1.12797263

Figure 8.2: Summary of Posterior Convergence and distributions (alpha and rho being the relevant variables, the table omits most simulated y's) as given by STAN for the final total population of England and Wales kriging model

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
rho	50.878347005	2.748476e-01	1.043165e+01	30.5342322	43.56651813	51.041422653	57.95139895	71.5026863	1440.5311	1.0002342
alpha	31.001583820	1.337082e+00	1.824881e+01	0.7168339	18.75584075	34.812883226	44.57194031	59.9031589	186.2744	1.0068411
y2[1]	-0.952064262	1.511426e-01	1.331923e+01	-28.9673171	-6.02446822	-1.065764278	3.85958540	28.2164269	7765.7703	0.9998025
y2[2]	-0.944341294	1.545188e-01	1.360733e+01	-29.4762590	-6.10110454	-1.059957309	3.98140270	28.8195185	7755.0319	0.9998038

Figure 8.3: Summary of Posterior Convergence and distributions (alpha and rho being the relevant variables, the table omits most simulated y's) as given by STAN for the final literacy kriging model

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
rho	14.894156212	1.220484e-02	0.6233158498	13.59775363	14.495501870	14.922377133	15.321969632	16.051199141	2608.267	0.9999464
alpha	0.879899239	3.680989e-03	0.1762973354	0.61381353	0.753391453	0.853968819	0.976037972	1.304288438	2293.836	1.0022036
y2[1]	-1.593041356	1.150913e-05	0.0010081598	-1.59500277	-1.593710844	-1.593062187	-1.592351734	-1.591048555	7673.150	0.9997075
y2[2]	-1.600829472	1.050965e-04	0.0090670585	-1.61916497	-1.606646612	-1.600763447	-1.594841985	-1.583297302	7443.149	1.0001333
y2[3]	-1.605509015	1.733000e-04	0.0149375367	-1.63579738	-1.615048436	-1.605296987	-1.595680076	-1.576590010	7429.518	1.0001724

Figure 8.4: Summary of Posterior Convergence and distributions (alpha and rho being the relevant variables, the table omits most simulated y's) as given by STAN for the final wage kriging model

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
a1	0.3477037	0.003176142	0.06788722	0.2464481	0.2982823	0.3373359	0.3859138	0.503936	456.8527	1.002861
a2	1.7911844	0.015514018	0.39598915	1.1789412	1.5069320	1.7428551	2.0125337	2.711734	651.5053	1.004447
rho1	2.3663144	0.007978907	0.17511676	1.9717713	2.2637431	2.3828640	2.4872935	2.663571	481.6911	1.005951
rho2	1.1418851	0.015640476	0.42411579	0.4938175	0.8244304	1.0860288	1.4063858	2.100584	735.3075	1.005094
c1	1640.6516437	0.095575482	1.12631124	1637.3234654	1640.2682384	1640.8247855	1641.4009571	1641.942562	138.8749	1.017696
lp__	6.3582881	0.095446055	1.71654880	1.8871651	5.5335628	6.8262916	7.6165629	8.389247	323.4419	1.010346

Figure 8.5: Summary of Posterior Convergence and distributions as given by STAN for the final changepoint-model

	mean	se_mean	sd	2.5%	25%	50%	75%	97.5%	n_eff	Rhat
rho[1]	1.2814748	0.0095870213	0.68536820	0.51483080	0.8432998	1.1070980	1.5098235	3.0303697	5110.701	0.9999712
rho[2]	1.3439165	0.0114769628	0.77619446	0.53401923	0.8765021	1.1534505	1.5976534	3.1952944	4573.905	1.0007859
rho[3]	1.3005947	0.0096437683	0.69058505	0.52518647	0.8517909	1.1276335	1.5369529	3.0815716	5127.915	1.0001958
rho[4]	1.2882029	0.0115701804	0.76248075	0.49203716	0.8000676	1.1005253	1.5508454	3.2734489	4342.877	1.0007248
rho[5]	1.1474699	0.0105070067	0.67480503	0.44920589	0.7124574	0.9847417	1.3780303	2.7976167	4124.759	0.9999345
rho[6]	1.3692964	0.0099326481	0.71593669	0.54554152	0.9065269	1.1988646	1.6360319	3.1676721	5195.402	0.9999338
alpha[1]	0.6707903	0.0064849215	0.49252379	0.02627325	0.2837595	0.5779604	0.9572982	1.8836801	5768.262	0.9997078
alpha[2]	0.5682275	0.0054701697	0.45537136	0.02370920	0.2183981	0.4600752	0.8066714	1.6945950	6929.945	0.9995683
alpha[3]	0.7280988	0.0069032097	0.54072422	0.03019045	0.3036512	0.6253816	1.0389349	2.0236581	6135.494	1.0000630
alpha[4]	0.5826174	0.0055257308	0.45679245	0.02220678	0.2307680	0.4796153	0.8306907	1.6995059	6833.739	1.0000944
alpha[5]	0.8212288	0.0066012919	0.50179192	0.06264662	0.4482609	0.7631865	1.1237774	1.9834967	5778.159	0.9998032
alpha[6]	0.5557698	0.0061541035	0.47798359	0.01870225	0.1915250	0.4280878	0.7857895	1.8429970	6032.487	1.0003337
sigma	0.3392789	0.0008521898	0.06813559	0.21863517	0.2937649	0.3335357	0.3796978	0.4864981	6392.569	0.9995031
lp__	-35.5820347	0.0577598118	2.72914149	-41.85860717	-37.1658554	-35.2693734	-33.6036110	-31.2852969	2232.547	0.9997818

Figure 8.6: Summary of Posterior Convergence and distributions as given by STAN for the final Additive GP model.