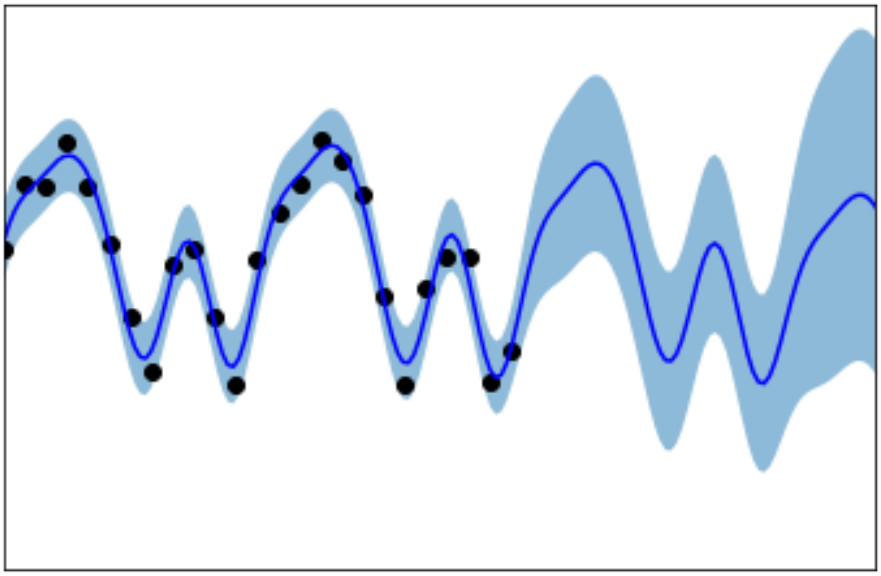

Gaussian Process Regression for Modelling Blood Glucose Dynamics

A study based on clinical data from subjects with type 2 diabetes mellitus

Master's thesis in Engineering Mathematics and Computational Science

Noel Waters

DEPARTMENT OF MATHEMATICS

CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2021
www.chalmers.se

Master's thesis 2021

Gaussian Process Regression for Modelling Blood Glucose Dynamics

A study based on clinical data from subjects with type 2 diabetes mellitus

NOEL WATERS

CHALMERS
UNIVERSITY OF TECHNOLOGY

Department of Mathematical Sciences
Chalmers University of Technology
Gothenburg, Sweden 2021

Gaussian Process Regression for Modelling Blood Glucose Dynamics
A study based on clinical data from subjects with type 2 diabetes mellitus
NOEL WATERS

Supervisors at AstraZeneca: Michail Doulis, Principal Data Scientist at Data Science & AI, BioPharmaceuticals R&D, AstraZeneca.
Magnus Rattray, Senior Data Science Director at Data Science & AI, BioPharmaceuticals R&D, AstraZeneca.
Supervisor at Chalmers: Umberto Picchini, Associate Professor at Department of Mathematical Sciences, Chalmers
Examiner: Moritz Schauer, Associate Senior Lecturer at Department of Mathematical Sciences, Chalmers

Master's Thesis 2021
Department of Mathematical Sciences
Chalmers University of Technology
SE-412 96 Gothenburg
Telephone +46 31 772 1000

Cover: Posterior predictive distribution of a Gaussian process regression model with a locally periodic kernel

Gaussian Process Regression for Modelling Blood Glucose Dynamics
A study based on clinical data from subjects with type 2 diabetes mellitus
NOEL WATERS
Department of Mathematical Sciences
Chalmers University of Technology

Abstract

Type 2 diabetes is a disease characterized by poor control of blood glucose levels. Continuous Glucose Monitoring (CGM) is an increasingly popular technology for studying glucose levels and evaluating treatment effects. Although CGM technology gives potential for granular insights into disease characteristics, more can be done in terms of exploiting this rich and dense data source to the fullest. This study aims to investigate the usefulness of Gaussian process regression as a framework for modelling blood glucose dynamics. The CGM data were collected from a previous clinical trial on a cohort of overweight and obese type 2 diabetes patients. Gaussian process modelling tools were used to capture short-term and recurring trends while adjusting for long-term changes in glucose control. Results indicate that structure such as periodicity can be successfully modelled. Interpreting specific modelling results showed to be challenging due to a high degree of uncertainty in the model hyperparameters. Non-stationary models should be considered to better account for the irregular occurrence of meal-related glucose spikes and differences between day and night glycemic variability. Finally, the periodic properties of blood glucose dynamics should be further explored.

Keywords: Gaussian Processes, Continuous Glucose Monitoring, Periodograms, Type 2 Diabetes

Acknowledgements

First of all, I'd like to thank Michail Doulis for his unwavering support and for working tirelessly to keep me on track throughout this project. I'm also grateful for the mathematical and practical advice provided by Umberto Picchini and the deep insights regarding Gaussian Processes provided by Magnus Rattray. You all gave me the tools and confidence necessary to move forward. Finally, I extend thanks to my friends and family without whom I could not have powered through. I've learnt a ton during this project, nonetheleast the value of having such great people around.

Noel Waters, Gothenburg, June 2021

Contents

List of Figures xi

1 Introduction 1
- 1.1 Background . . . 1
  - 1.1.1 Modelling blood glucose dynamics with Gaussian processes . . . 2
- 1.2 Aim . . . 3
  - 1.2.1 Research questions . . . 3
- 1.3 Scope . . . 3

2 Theory 5
- 2.1 Gaussian process regression . . . 5
- 2.2 Kernels . . . 7
  - 2.2.1 The RBF kernel . . . 7
  - 2.2.2 Matérn kernels . . . 8
  - 2.2.3 Periodic kernels . . . 9
  - 2.2.4 Constructing new kernels from old . . . 10
  - 2.2.5 Spectral mixture kernels and stationarity . . . 11
  - 2.2.6 Example predictions . . . 12
- 2.3 Training Gaussian processes . . . 13
  - 2.3.1 Choosing between kernels . . . 14
  - 2.3.2 Non-convexity of optimization . . . 14
- 2.4 Decomposing Gaussian processes . . . 15
- 2.5 Periodograms & Lomb-Scargle . . . 16

3 Methods 17
- 3.1 Data characteristics and pre-processing . . . 17
  - 3.1.1 Study material and data considerations . . . 18
- 3.2 Software . . . 18
- 3.3 Optimization . . . 18
  - 3.3.1 Fixating the noise parameter . . . 19
- 3.4 Detecting and quantifying periodicities . . . 19
- 3.5 Model selection and comparisons . . . 20
- 3.6 Detrending and decomposing signals . . . 21
- 3.7 Predictions via Spectral mixture kernels . . . 22
  - 3.7.1 Initializing the Spectral mixture kernel hyperparameters . . . 23

4 Results 25

4.1 Matérn kernels to evaluate treatment effects . . . 25
4.1.1 Comparisons on detrended data . . . 28
4.2 Incorporating structure: periodicity . . . 30
4.2.1 Hypotheses for Locally periodic models . . . 32
4.2.2 Analysing Locally periodic models . . . 32
4.3 Extrapolations with Spectral mixture kernels . . . 37
4.3.1 Spectral mixture kernels on simulated data . . . 37
4.3.2 Spectral mixture kernels on CGM data . . . 39

5 Discussion **43**
5.1 Study limitations . . . 45

6 Conclusions **49**

Bibliography **51**

A Appendix 1 **I**

List of Figures

2.1 Covariance function and samples from GPs with RBF kernels with three different hyperparameter setups. 7
2.2 Sample paths from GPs with Matérn covariance functions with different values of the roughness parameter ν. All kernel outputscales are set to 1. 8
2.3 Covariance functions and sample paths from GPs with Cosine and Periodic kernels. All kernel outputscales are set to 1. 9
2.4 Covariance functions and sample paths from GPs with Locally periodic kernels, given by multiplying the Periodic and Matern_{25} kernel. All kernel outputscales are set to 1. 10
2.5 Characteristics of a four component Spectral mixture kernel with means μ=[1/11, 1/4, 1/7, 1/82], weights w=[1.6, 1.1, 1.6, 0.4] and variances v^2=[0.0031^2, 0.0064^2, 0.0479^2, 0.0017^2]. 12
2.6 Predictions using four different GP models. The confidence region covers two standard deviations. 13
2.7 Decomposition of a process into three separate processess: 1: Aperiodic process given my Matérn_{15} kernel. 2: Repeating pattern process given by Periodic kernel. 3: Sinusoidal process given by Cosine kernel. Only the means of the three posterior distributions are shown. 15
2.8 Lomb-Scargle periodograms of sample traces from GPs with Locally periodic kernels. The colors match the corresponding traces in figure 2.4. 16

3.1 Left pane: Fragment of a reported CGM trace. All observations land precisely on the horizontal lines. Right pane: Count of all unique blood glucose values for patient TB. 20
3.2 Illustration of the detrending procedure. Left pane: Original trace of patient TO. Right pane: Detrended trace of patient TO 22

4.1 Difference in NLML between Matérn_{05} and Matérn_{15} models. Almost all values are negative, meaning that for each patient, the Matérn_{15} model gave a better fit. 26
4.2 Left pane: Fitted outputscales and lengthscales. Right pane: Quotients of fitted lengthscales/outputscales. Parameters are taken from the top ten Matérn_{15} models for each patient in the treatment group based on the final 14 day phase. 27

4.3 Fold changes in Lengthscale/Outputscale (LO) quotients between the final and initial phase of all patients, colored by study group. Fold changes are computed as the LO-quotients in the final phase over the LO-quotients in the initial phase. 28

4.4 Glucose traces of most improved subjects with respect to the LO-quotient throughout their final and initial phase. Both display long term trends in the initial phase. 29

4.6 Fold changes in Lengthscale/Outputscale quotients between the final and initial phase of all patients, colored by study group and computed on detrended data. Fold changes are computed as the LO-quotients in the final phase over the LO-quotients in the initial phase. 29

4.5 Left pane: Fitted outputscales and lengthscales. Right pane: Quotients of fitted Lengthscales/Outputscales. Parameters are taken from the top ten Matérn$_{15}$ models for each patient in the treatment group, based on the final 14 day phase on detrended data. 30

4.7 Four typical periodograms for the CGM data. The bottom left pane shows a patient who was deemed to display no periodicity. 31

4.8 Difference in NLML between the top ten Rough and Smoother models for each patient in the final phase. A negative value indicates the Smoother model gave a better fit. 33

4.9 Detrended and decomposed signal for both the initial and final phase of patient TB. The decomposition was based on the best Smooth model with respect to NLML. 34

4.10 Detrended and decomposed signal for both the initial and final phase of patient TS. The decomposition was based on the best Smooth model with respect to NLML. 34

4.11 Optimized periodic lengthscales of the top ten Rough and Smoother models for each patient. Each dot represents a model and the blue color indicates that particular model had its Locally periodic outputscale parameter optimized to below 0.1. 35

4.12 % Changes from the initial guess to optimized values of the Decay lengthscale parameter among the top ten models for each patient in the final phase. Left pane: Results for Rough model. Right pane: Results for Smoother model. 36

4.13 True signal, long term component and decomposition based on the best Rough and Smoother model for the initial phase of patient TD. Left Pane: Rough model, Right Pane: Smoother model. 36

4.14 Comparison of NLML between top ten Matérn$_{15}$ and Smoother models for each patient throughout the final phase. A negative value means the Smoother model is favored for that patient. A blue dot indicates the Locally periodic component outputscale in the Smoother model was optimized to a value below 0.1. 37

4.15 SM modelling of data simulated from Periodic + Cosine kernel. Top pane: True and estimated covariance function. Middle pane: Initial and optimized analytic power spectrum of the SM-kernel. Bottom pane: The simulated signal and SM-kernel prediction. The confidence region covers two standard deviations 38
4.16 SM modelling of data simulated from the Periodic+Cosine+Matérn$_{15}$ + RBF kernel. Top pane: True and estimated covariance function. Middle pane: Initial and optimized analytic power spectrum of the SM-kernel. Bottom pane: The simulated signal and SM-kernel prediction. The confidence region covers two standard deviations 39
4.17 SM modelling of data simulated from the Periodic+Matérn$_{15}$ kernel. Top pane: True and estimated covariance function. Middle pane: Initial and optimized analytic power spectrum of the SM-kernel. Bottom pane: The simulated signal and SM-kernel prediction. The confidence region covers two standard deviations 40
4.18 Spectral mixture modelling results on detrended initial 12 days of patient PC. The model depicted had the lowest NLML on training data. 41
4.19 Spectral mixture modelling results on detrended initial 12 days of patient PC. The model depicted had the smallest mean squared error of prediction on the evaluation data. 41

5.1 Example trace of blood glucose values, colored by night and day time. Night time was defined as times between 00:00 and 06:00. The horizontal lines denote the limits for hypo- and hyperglycemia. 45
5.2 Histogram of ratios of variance of blood glucose levels between night- and day time on detrended data, for all patients. A value above 1 indicates the day time variance is higher. Night time was defined as times between 00.00 and 06.00. 46
5.3 Illustration of the skewness of CGM data. Histograms of glucose values for all patients across the full study period are shown, and the skewness computed for each patient separately. The left panes regard original data and the right panes regard detrended data. 47

1
Introduction

Diabetes is a growing worldwide health challenge. There are two types of diabetes. Type 1 diabetes, in which the body's capability to produce insulin is severely diminished or absent, and type 2 diabetes in which an impaired insulin production is also coupled with low responsiveness to the hormone. For both diabetes types, a characteristic feature is poor control of blood glucose levels. In recent years, advancements in technology for Continuous Glucose Monitoring (CGM) has enabled diabetes patients to track their glucose level trajectories in close to real time via providing measurements of blood glucose levels as often as once every 5 or 15 minutes. The technology aids patients in avoiding severe short term issues as well as long term complications associated to hyper- and hypoglycemia, for example via signalling when blood glucose levels reach dangerous levels [1].

Due to its potential for giving granular insights, CGM is used in clinical trials to evaluate treatment effects and better understand the disease symptomatology. It is of interest to uncover non-trivial patterns in CGM data that might not be captured or revealed by the conventional measures. Potential insights into sources of glucose variability could help researchers evaluate treatment efficacy, and in the long run improve the situation for subjects suffering from diabetes. In addition, a mathematical model with strong predictive power could be used to guide subjects or machines regarding when to e.g. consume food or inject insulin, so as to avert future complications. Such a model must be flexible enough to handle the many sources of variability in blood glucose levels, for example meals and exercise.

1.1 Background

In recent years, a new class of medications, the Glucagon-like peptide-1 receptor agonists - also knows as GLP-1 receptor agonists - has been introduced for the treatment of type 2 diabetes. The D5670C00011 study is a phase 2, randomised, double-blind, placebo-controlled study to evaluate the efficacy, safety, tolerability, and pharmacokinetics of different doses of MEDI0382 in overweight and obese subjects with type 2 diabetes mellitus [2]. MEDI0382 is a synthetic peptide with both glucagon-like peptide-1 (GLP-1) and glucagon receptor co-agonist activity [3]. The primary objective of this study was to assess the effects of MEDI0382 on glucose control and body weight versus placebo after 49 days of treatment. Other secondary objectives of that study included the characterization of the safety profile and tolerability of MEDI0382, while one of the exploratory objectives focused on

the assessment of the effect of MEDI0382 on glucose lowering during different meals and times of the day as measured by CGM. The study was divided in two cohorts. 39 subjects were randomized for Cohort 1 divided between two arms, one receiving MEDI0382 and the other placebo. Cohort 1 evaluated the efficacy, safety, tolerability, and pharmacokinetics of MEDI0382. The drug was titrated up, i.e given in increasing doses, in weekly intervals from 50 to 300 μg and administered over 49 days. Cohort 2 consisted of 24 subjects and explored an alternative 2-week titration schedule. Each dose of MEDI0382 was administered as a subcutaneous injection each morning. Cohort 1 subjects used a Freestyle Libre® Pro CGM device to measure interstitial glucose levels during the study. The Freestyle Libre® Pro CGM device measures interstitial glucose levels every 15 minutes for 2 weeks continuously. The study results and mechanistic insights of MEDI0382 are discussed in detail in *Efficacy, Safety, and Mechanistic Insights of Cotadutide, a Dual Receptor Glucagon-Like Peptide-1 and Glucagon Agonist* [4]. Details around the study protocol, study population and statistical methods for analysis can be found in the statistical analysis plan [2].

A distinction is sometimes made between glycemic control, which relates to the long-term levels of blood or glucose values, and glycemic variability (GV), which relates to variations seen on a shorter time-frame. Examples of popular measures of GV are the Coefficient of Variation (CV) which is the standard deviation of blood glucose values divided by its mean, and "% time spent in hyper- or hypoglycemia" which counts the minutes spent in these phases [5]. Typical signals of hyperglycemia are spikes in glucose levels following a meal, called post-prandial peaks. A measure of these, and other glucose spikes, is the Mean Amplitude of Glycemic Excursion (MAGE). Due to difficulties in determining what constitutes such an excursion and exactly what is meal-related, there is no consensus on the best way to quantify these peaks, or in fact GV in general [6]. Despite the lack of consensus on what the best measure is, a major proportion of GV measures are concerned either with amplitudes of variations or times in different states, whereas few explicitly take the rate of change, or frequency of oscillations, into account. A notable exception is the recently proposed measure Glycemic Variability Percentage, which compares the piecewise length of an observed glucose trace to an ideal straight line, thus accounting for rates of change in blood glucose levels as well as amplitude [7]. However, the authors mention caveats of the method such as its dependency of the units for glucose concentration and gaps in data, leaving room for the development of other measures. Therefore, it is necessary to explore new methods to analyse dense and information-rich CGM data to obtain granular insights that will help to better understand the mechanistic action of various treatments in diabetes.

1.1.1 Modelling blood glucose dynamics with Gaussian processes

The fluctuating blood glucose levels of a person can be viewed as a continuous time stochastic process and CGM devices as a means of sampling from this latent process, with noise. Gaussian processes (GPs) are a flexible class of stochastic processes

which via a Bayesian approach can be used in a regression framework. Therefore, they might be ideal to model GV. Generally, any GP is completely specified by its mean and covariance function and much flexibility of this class comes from the freedom in choosing these functions and their hyperparameters. The covariance function, also known as the kernel, can for example encode smoothness, periodicity and general noise level [8]. Such properties are in many cases clearly visible from the kernel structure, which gives GPs the benefit of interpretability and has inspired e.g. automatic GP modelling and translation of modelling results to plain English [9]. In terms of predictive power, a class of kernels called Spectral mixture kernels have been shown to extrapolate well, while simultaneously uncovering and highlighting patterns such as periodicity in data [10]. Moreover, GPs have been used before as a part of developing an artificial pancreas, albeit in a simulated environment where subject meal-times were known [11].

1.2 Aim

This project aims to propose methods for, and investigate the usefulness of, Gaussian process regression modelling of blood glucose levels in patients with type 2 diabetes. This in an effort to obtain clinical insights and complement conventional methods for evaluating treatment effects and gain granular insights on the mechanism of action of specific medications of diabetes.

1.2.1 Research questions

1. What combinations of kernels can be used to analyse CGM data and which aspect of a CGM-trace is captured by each part of the GP model?
2. To what extent can specific combinations of kernels and their hyperparameters unveil features that can be interpreted in clinical/biological terms?
3. Regarding Spectral mixture kernels: To what extent does the spectral density of such stationary kernels unveil hidden periodicities in blood glucose levels?

1.3 Scope

This project is an initial attempt of using GPs to model CGM data and does not aim to develop methods for GP modelling in general. Thus, the project will not cover numerical methods to train GP models in detail, but rather rely on present computer programs for this task. Therefore, the project is limited to using kernel functions that are reasonable to construct within these programs.

In terms of relating patient characteristics to GP-modelling, it is conceivable to incorporate metadata, such as medical history or occurrences of adverse events, into the mean and covariance functions directly. However, such an approach will not be attempted in this work, since it could lead to less generalisable models and it would require knowledge of relevant features to incorporate a-priori. Moreover, the project will not address non-Gaussian likelihood choices when training models. This means

that the measurement errors associated with each observation are assumed to be i.i.d. mean zero Gaussians.

Finally, GP modelling is sometimes performed in a fully Bayesian manner where priors on kernel hyperparameters are set up. That way, via for example Markov chain Monte Carlo methods one can obtain estimates of the uncertainty of hyperparameters to incorporate when making predictions. Such a procedure is beyond the scope of this work and we resort to point estimates of kernel hyperparameters, obtained primarily via Maximum Likelihood estimation.

2

Theory

A stochastic (random) process can be seen as a collection of random variables $\{X(t), t \in T\}$, where T is an index set. The possible values of $X(t), t \in T$ is called the *state space*. If both the state space and parameter set are continuous, the process is called a continuous-state process. A Gaussian process (GP) is an example of such a continuous state process, and specifies a distribution over functions. We write

$$f \sim GP(m(t), k(t, t')) \tag{2.1}$$

if f is a GP with mean function $m(t) = E[f(t)]$ and covariance function $k(t, t') = cov(f(t), f(t'))$ for $t, t' \in \mathbb{R}^P$. For the purposes of this thesis, $P = 1$ since time is the only measured dimension that the blood glucose variations takes place over. A GP is a generalization of the Gaussian probability distribution, in the sense that for any finite collection of n points $[t_1, ..., t_n]$,

$$[f(t_1), ..., f(t_n)] \sim N(m(t_1), ...m(t_n), K(t_1, ...t_n)) \tag{2.2}$$

is an n-dimensional multivariate Gaussian distribution. Its mean vector $\mathbf{m}$ has entries $\mathbf{m}_i = m(t_i)$ and its covariance matrix K has entries $K_{i,j} = k(t_i, t_j)$ [12]. An important note is that the covariance is only a function of the inputs in $t_i, i \in [1, ...n]$, and thus does not take into account any process values at any particular point. Just like Gaussian distributions are probabilistically determined by their mean and covariance, GPs are completely specified by their mean and covariance functions, the latter often being called a *kernel* or *kernel function*. It is precisely the choice of these functions that determines what type of structure any GP regression model can capture [8].

2.1 Gaussian process regression

Throughout this chapter, $f(T)$ will denote the Gaussian distribution given by the GP f viewed at a finite collection of points T. f may also denote a Gaussian distribution, in which case it should be clear from the context.

Equation (2.2) shows that specifying a mean- and covariance function is a way to put a prior over functions. Consider evaluating the GP at a finite collection of points. That is equivalent do drawing samples from a multivariate normal distribution, and each resulting trace will be a realization of that prior over functions.

Now, consider two Gaussian distributions taken from a GP with a zero mean function

and shared covariance function , that is

$$f(T) \sim N(0, K(T,T)), \tag{2.3}$$
$$f(T^*) \sim N(0, K(T^*,T^*)) \tag{2.4}$$

where T and T* denote vectors of points where the GP is queried and K is the covariance function. Then, Gaussianity gives that the joint distribution of $[f = f(T), f^* = f(T^*)]$ is also Gaussian and takes the form:

$$p\begin{pmatrix} f \\ f* \end{pmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} K(T,T) & K(T,T^*) \\ K(T^*,T) & K(T^*,T^*) \end{pmatrix}\right]. \tag{2.5}$$

where the off-diagonal matrices $K(T^*,T)$ and $K(T,T^*)$ are cross covariances between all time points in T and T^*. Now, for inference purposes, it is of interest to use observations to update prior beliefs and make predictions. By use of Bayes rule and normal-normal conjugacy, a closed form expression for the posterior distribution of $f^* = f(T^*)$ given $f(T)$ is obtained as:

$$\begin{aligned} p(f^* \mid T^*, T, f) \sim N\Big(&K(T^*,T)K(T,T)^{-1}f, \\ &K(T^*,T^*) - K(T^*,T)K(T,T)^{-1}K(T,T^*)\Big). \end{aligned} \tag{2.6}$$

Now, making predictions of values for f^* is simply a matter of sampling from the Gaussian distribution given in equation (2.6) or evaluating its posterior mean. The posterior covariance conveniently gives a direct measure of uncertainty for all such points. Although the prior mean was zero, the posterior mean in equation (2.6) can adapt to the data given a suitable choice of covariance function. This emphasizes the importance of that choice and justifies the common principle of assigning zero-mean priors when fitting GP models[8].

Equations (2.5) and (2.6) can be thought of as prior and posterior distributions of GPs where observations f are taken without noise. However, we often assume process values are observed with some noise as in the following hierarchical model:

$$\epsilon \sim \mathcal{N}(0, \sigma^2) \tag{2.7}$$
$$f \sim GP(m(t), k(t,t')), [t,t'] \in \mathbb{R} \tag{2.8}$$
$$y(T) = f(T) + \epsilon \tag{2.9}$$

That is, we are assuming our observations y are realizations of a GP f observed at times T with centered, i.i.d, Gaussian noise. σ^2 is referred to as the noise parameter. For such a model, the joint prior distribution can instead be written as

$$p\begin{pmatrix} y \\ f^* \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} K(T,T) + \sigma^2 \cdot \mathbf{I} & K(T,T^*) \\ K(T^*,T) & K(T^*,T^*) \end{pmatrix}\right], \tag{2.10}$$

and the posterior distribution can be written as

$$\begin{aligned} p(f^* \mid T^*, T, \mathbf{y}) \sim \mathcal{N}\Big(&K(T^*,T)[K(T,T) + \sigma^2 \cdot \mathbf{I}]^{-1}\mathbf{y}, \\ &K(T^*,T^*) - K(T^*,T)[K(T,T) + \sigma^2 \cdot \mathbf{I}]^{-1}K(T,T^*)\Big), \end{aligned} \tag{2.11}$$

where $\mathbf{I}$ is an identity matrix with dimensions the same as $K(T, T)$. This added uncertainty for example allows for the mean of the posterior predictive distribution at points $[T_i^* \in T]$ to deviate from the observed values.

2.2 Kernels

In this study, all GP models are given a prior mean of zero. Thus, differences in properties between investigated models are completely determined by their kernels, or covariance functions. By restricting the properties of the covariance function, one introduces inductive biases that are necessary for a GP-model to learn anything from data via application of equation (2.11). Each kernel itself has a number of parameters, which modify the shape of the covariance function and thus also the GP itself. These parameters are often referred to as *hyperparameters*.

2.2.1 The RBF kernel

The Radial Basis Function (RBF) kernel, also known as the Squared Exponential, can be written as

$$k_{\text{RBF}}(\mathbf{t_1}, \mathbf{t_2}) = \sigma^2 \exp\left(-\frac{|\mathbf{t_1} - \mathbf{t_2}|^2}{l^2}\right) \tag{2.12}$$

where the outputscale σ^2 controls the overall amplitude of the signal variations, and l is a lengthscale parameter, which determines how quickly the functions are allowed to vary in time, or equivalently how fast the covariance decays with distance.
An illustration of the effect of varying the hyperparameters σ and l on the kernel function itself and its corresponding GP-prior over functions is shown in figure 2.1.

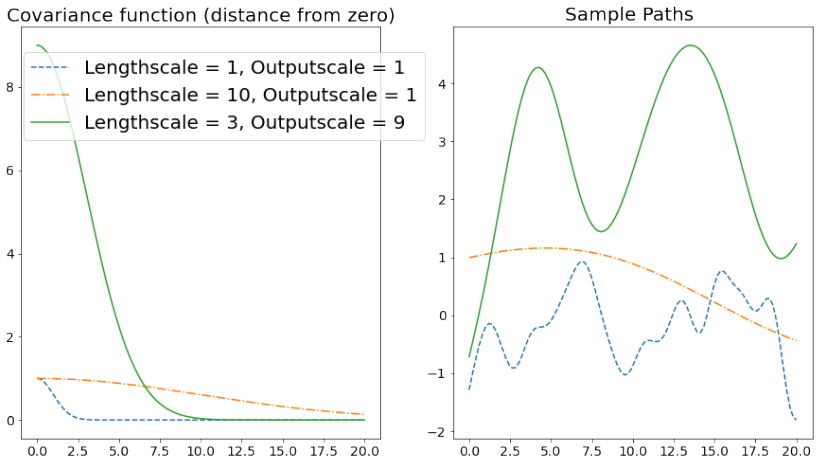

Figure 2.1: Covariance function and samples from GPs with RBF kernels with three different hyperparameter setups.

Notice how smooth the sample paths are in figure 2.1. Indeed, the RBF kernel implies the process to be modelled is infinitely differentiable, an assumption that

makes the kernel ill fit to model processes that can not be assumed to vary smoothly over time [8].

2.2.2 Matérn kernels

The Matérn family of kernels can be written as,

$$k_{\text{Matern}}(\mathbf{t_1}, \mathbf{t_2}) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\sqrt{2\nu} \frac{|\mathbf{t_1} - \mathbf{t_2}|}{l} \right)^{\nu} K_\nu \left(\sqrt{2\nu} \frac{|\mathbf{t_1} - \mathbf{t_2}|}{l} \right),$$

where K_ν is a modified Bessel function, $\Gamma()$ is the gamma function, ν is a positive parameter determining differentiability , l is a lengthscale parameter and the outputscale σ^2 determines the overall variance. For ν=0.5,1.5 and 2.5 the expressions for Matérn covariances simplify greatly to

$$\nu = 0.5 : k_{\text{Matern05}}(\mathbf{t_1}, \mathbf{t_2}) = \sigma^2 \exp\left(-\frac{|\mathbf{t_1} - \mathbf{t_2}|}{l} \right) \tag{2.13}$$

$$\nu = 1.5 : k_{\text{Matern15}}(\mathbf{t_1}, \mathbf{t_2}) = \sigma^2 \left(1 + \sqrt{3}|\mathbf{t_1} - \mathbf{t_2}| \right) \exp\left(\frac{\sqrt{3}|\mathbf{t_1} - \mathbf{t_2}|}{l} \right) \tag{2.14}$$

$$\nu = 2.5 : k_{\text{Matern25}}(\mathbf{t_1}, \mathbf{t_2}) = \sigma^2 \left(1 + \frac{\sqrt{5}|\mathbf{t_1} - \mathbf{t_2}|}{l} + \frac{5|\mathbf{t_1} - \mathbf{t_2}|^2}{3l^2} \right) \exp\left(\frac{\sqrt{5}|\mathbf{t_1} - \mathbf{t_2}|}{l} \right). \tag{2.15}$$

Depending on the choice of ν, the corresponding GP will be very rough or more smooth [8]. As with the RBF-kernel, the lengthscale l determines how quickly the functions are allowed to vary in time, and the outputscale σ^2 the overall variance or amplitude of the process variability. An illustration of the dependence on ν is shown in figure 2.2.

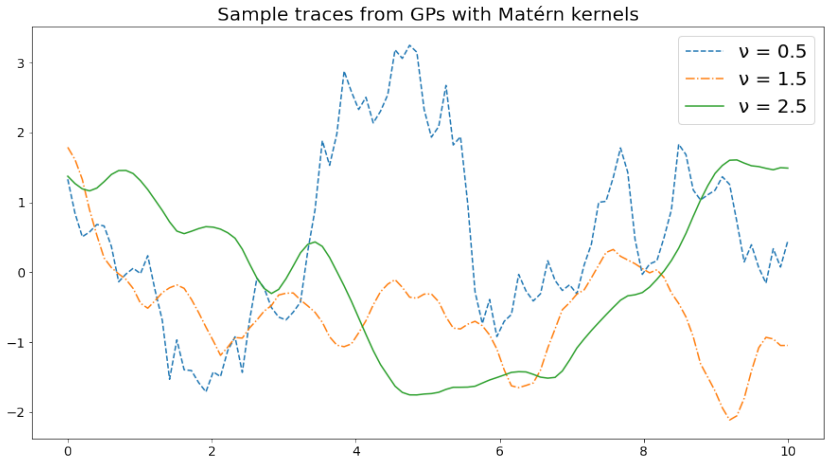

Figure 2.2: Sample paths from GPs with Matérn covariance functions with different values of the roughness parameter ν. All kernel outputscales are set to 1.

Notice in figure 2.2 the increasingly squiggly or rough behaviour as ν decreases, suggesting the Matérn class of kernels are superior to the RBF kernel when modelling rougher signals.

2.2.3 Periodic kernels

The Matérn and RBF kernels share the trait that the covariance between two points decreases monotonically with their distance. However, if the process of interest is known to co-vary at specific time intervals, periods, it may be of interest to incorporate such a property into the GP-prior. Two examples of kernels that place periodic behaviour over the GP priors are the Cosine kernel, given as

$$k_{\text{Cosine}}(\mathbf{t_1}, \mathbf{t_2}) = \sigma^2 \cos\left(2\pi|\mathbf{t_1} - \mathbf{t_2}|/p\right), \tag{2.16}$$

where p is the period length and σ^2 controls the overall variance/amplitude, and the Periodic kernel, given as

$$k_{\text{Periodic}}(\mathbf{t_1}, \mathbf{t_2}) = \sigma^2 \exp\left(-\frac{2\sin^2\left(\pi|\mathbf{t_1} - \mathbf{t_2}|/p\right)}{\ell^2}\right), \tag{2.17}$$

which also has a lengthscale parameter l. While the Cosine kernel generates and models perfectly sinusoidal signals, the Periodic kernel can be thought of as modelling perfectly repeating patterns. The hyperparameter l in the Periodic kernel controls the flexibility with which a function can vary within a period. This is analogous to the lengthscale's influence on the RBF and Matérn kernels [9].
Figure 2.3 illustrates the behaviour of both these kernels, along with the impact of varying the lengthscale parameter for the Periodic kernel. The shorter lengthscale allows for more within-period fluctuations.

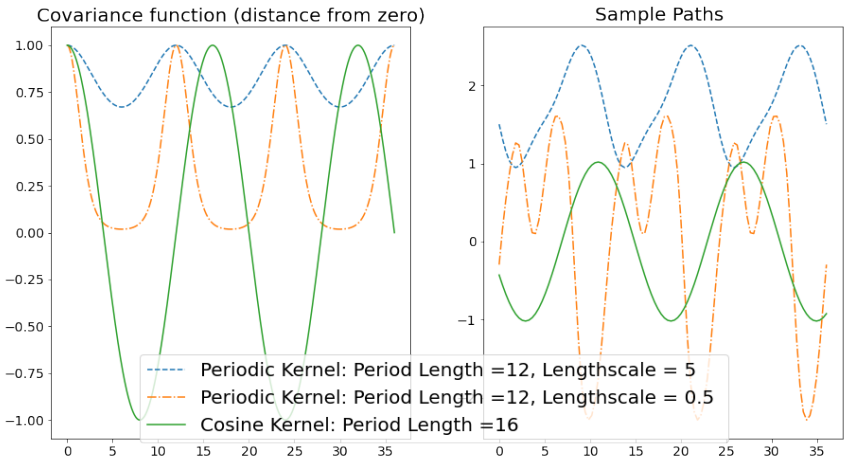

Figure 2.3: Covariance functions and sample paths from GPs with Cosine and Periodic kernels. All kernel outputscales are set to 1.

Important to note is that these kernel functions encode exact periodicity in the GP-prior, which is a very strong assumption to place on a stochastic process. However, as shall be seen in section 2.2.4, there are simple ways to mitigate this strong assumption.

2.2.4 Constructing new kernels from old

Any valid covariance function must be positive semi-definite. This restricts the space of functions that can be considered valid kernels and can make it hard to construct valid kernels. However, it is well known that for any two valid kernels, both their sum and their product are valid covariance functions. This flexibility allows the user to construct, train and test expressive kernels based on using simple kernels as building blocks. Further instructive examples of this are provided by Duvenaud et al. [13].

A commonly used product kernel is the so called Locally periodic kernel, which is constructed by multiplying the Periodic kernel with a kernel parametrized by a lengthscale, for example from the Matérn family. That lengthscale is called the Decay lengthscale. This multiplication relaxes the exact periodicity enforced by the Periodic kernel so as to allow for variations in patterns and amplitudes across periods. This is illustrated in figure 2.4, where the covariance function and samples drawn from GPs constructed as Periodic Kernel*Matérn$_{25}$ kernel, are shown. Clearly, the shorter the Decay lengthscale, the faster the covariance function tends to zero, and the less periodic are samples from the Locally periodic GP-prior.
Using the Matérn$_{25}$ kernel to make the periodicity local also adds a certain squiglyness to the GP-prior, allowing such a kernel to model rougher processes.

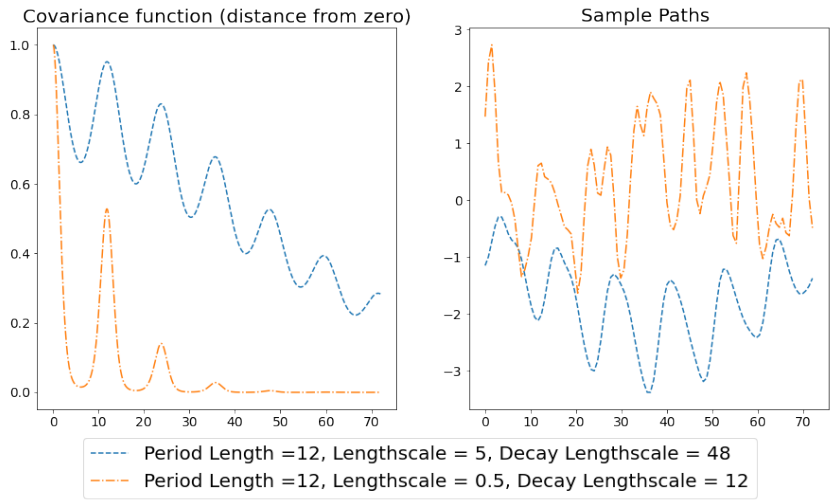

Figure 2.4: Covariance functions and sample paths from GPs with Locally periodic kernels, given by multiplying the Periodic and Matern$_{25}$ kernel. All kernel outputscales are set to 1.

2.2.5 Spectral mixture kernels and stationarity

All kernels covered in this thesis are stationary, as can be seen from the fact that their function values are only dependent on the distance between two evaluated points, that is $K(t_1, t_2) = K(|t_1 - t_2|)$. In terms of GPs over the time domain, this means to assume that the statistical properties of the signal do not change over time. As an example, the Locally periodic kernel can accommodate slight differences in variability between periods, but it does not accommodate for the disappearance of periodicity altogether. Although stationarity is a strong assumption to make, it also has a number of favorable properties, one of them being that every valid stationary kernel has a well defined spectral density, as given by its Fourier transform. The spectral density quantifies the frequency content of a signal. The covariance function of a stationary process is completely determined by its spectral density, which many times can be more interpretable than the signal itself [10]. This idea rests at the core of the Spectral mixture kernels class.

Spectral mixture kernels (SM-kernels) are a class of stationary kernel functions that can be thought of as a special type of composite kernels. In the one dimensional case, they are parametrized as

$$k_{SM}(\tau) = \sum_{q=1}^{Q} w_q \cos\left(2\pi\mu_q\tau\right) \cdot \exp\left(-2\pi^2\tau^2 v_q^2\right), \tag{2.18}$$

where $\tau = |t - t'|$. This makes them equivalent to a weighted sum of products between Cosine- and RBF-kernels where $\frac{1}{\mu_q}$ is each components period. The SM-kernel's Fourier dual, or spectral density, is a mixture of Q Gaussians, symmetrized around zero where the q_{th} Gaussian has mean μ_q, variance v_q^2 and is scaled by the weight w_q. That means the spectral density at frequency s can be computed analytically as

$$S(s) = \sum_{q=1}^{Q} w_q \frac{\phi_q(s) + \phi(-s)}{2}, \tag{2.19}$$

where $\phi_q(s) = \frac{1}{\sqrt{2\pi v_q^2}} \exp\left(\frac{(s-\mu_q)^2}{2v_q^2}\right)$ is the probability density function of the normal distribution with mean μ_q and variance v_q^2. This allows the SM-kernel's spectral density to take on many different forms, meaning the SM-kernel can approximate other stationary kernels, since approximating a kernel's spectral density means that the kernel function itself will be approximated.

This flexibility allows for a seemingly automatic selection of relevant kernel properties and the researcher can ideally gain insights into the data by studying the fitted hyperparameters of a Spectral mixture model. Indeed, its inventor showcases the capability of the SM-kernels to pick up various patterns and extrapolate well, along with being interpretable via its power spectrum [10]. Figure 2.5 illustrates the relationship between the analytic power spectrum of a SM-kernel and its covariance function. It may be hard to see from the kernel function itself that it encapsulates a 4- and 11-periodic sinusoidal component, along with a decaying 7-periodic compo-

nent. That is more clear the peaks in the analytic power spectrum, where the width of each Gaussian bell indicates how fast the corresponding component is decaying.

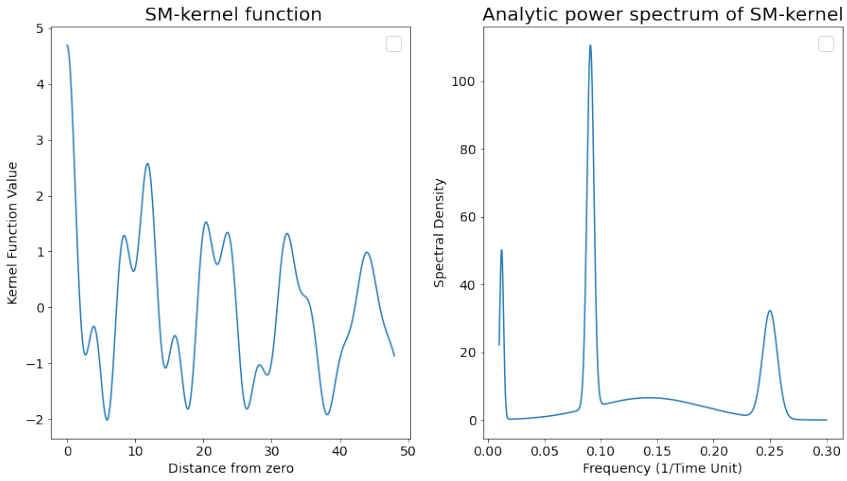

Figure 2.5: Characteristics of a four component Spectral mixture kernel with means μ=[1/11, 1/4, 1/7, 1/82], weights w=[1.6, 1.1, 1.6, 0.4] and variances v^2=[0.0031^2, 0.0064^2, 0.0479^2, 0.0017^2].

Note that the SM-kernels' flexibility comes at a cost. Drawing conclusions from models with such kernels requires a great confidence in that the best hyperparameters and number of mixture components are properly inferred via the optimization method employed by the researcher. Such an optimization has been shown to be difficult for the SM-kernels and methods that go beyond this thesis's scope have been proposed to tackle this issue [14]. Without such methods, it is of great importance to have a good starting point for optimizing the SM-kernel parameters.

It is suggested by Wilson [10] that one could initialize the SM-kernel components starting from an estimation of the power spectral density of the signal. In particular, since the SM-kernel's Fourier dual is a mixture of Gaussians, one can fit a Gaussian mixture model (GMM) with the desired number of mixture components to the power spectral density of the data, and then initialize the SM kernel components based on the means, variances and weights of those GMM components. A more detailed account of GMMs is found in Appendix A.

2.2.6 Example predictions

To illustrate the impact of kernel choice on the predictive performance of a GP, four of the discussed kernel types were fitted to the same data, by applying equation (2.11). Data was simulated from a GP with a Periodic kernel with lengthscale 5 and period length 12, whereafter zero- mean i.i.d noise with variance 0.01 was added. The observation noise parameter was set to 0.01, allowing for some uncertainty of predictions at observed points.

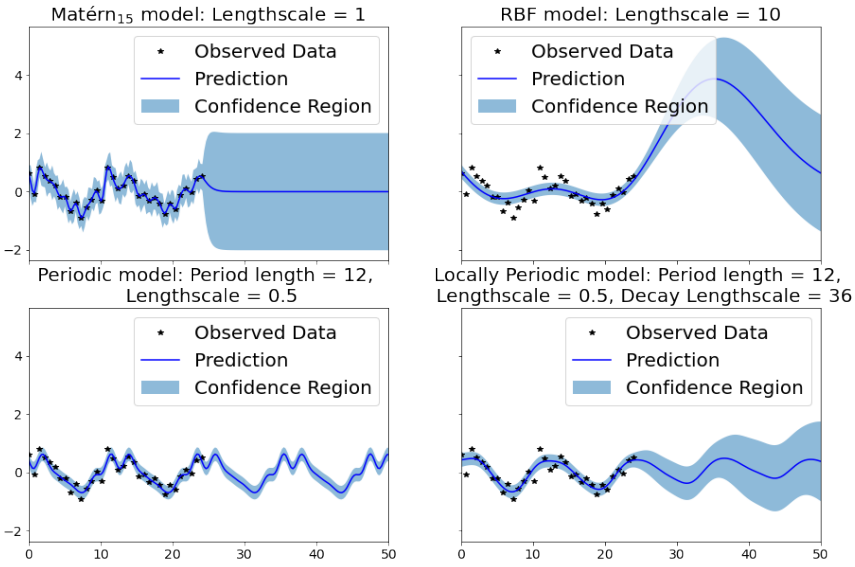

Figure 2.6: Predictions using four different GP models. The confidence region covers two standard deviations.

Figure 2.6 shows the predictions from the four models. The Matérn model, which had a very short lengthscale, quickly reverts back to predicting a zero mean, whereas the periodic kernel predicts the same pattern indefinitely. The RBF model varies too slowly to capture the observations, due to its very long lengthscale. To make better predictions than this, one should turn to optimizing the hyperparameters.

2.3 Training Gaussian processes

The problem of training or optimizing a GP model mainly consists of two tasks: Choosing a kernel and tuning its hyperparameters.
A method to evaluate a GP model's fit is to study the negative log marginal likelihood (NLML) of the observations, given the model. To understand why this is, first consider the marginal likelihood, also known as evidence, given as equation (2.20),

$$p(y \mid T, \theta, \sigma) = \int p(y|f, \sigma)p(f|K_\theta(T))df, \tag{2.20}$$

where θ denotes the kernel hyperparameters for the kernel K_θ which is the covariance function for the GP prior over functions f, y are the observations and σ is the standard deviation of observation noise at the observation points T. The name "marginal" comes from the fact that the latent process f is integrated out, meaning that $p(y \mid T, \theta, \sigma)$ is simply the likelihood of your observations at time points T being generated from a process with parameters θ and σ.
Since both factors in the integral in equation (2.20) are Gaussian, Normal-Normal conjugacy allows for obtaining an exact expression for the marginal likelihood, and

therefore its logarithm as well. The latter is more useful to work with since the likelihood itself will take on extremely small values when multiple observations are made. The NLML can be expressed as

$$-\log(p(y \mid T, \theta)) = \frac{1}{2} y^T [K_\theta(T) + \sigma^2 \cdot \mathbf{I}]^{-1} y + \frac{1}{2} \log(\mid [K_\theta(T) + \sigma^2 \cdot \mathbf{I}] \mid) + \frac{n}{2} \log(2\pi), \tag{2.21}$$

where $\mid * \mid$ denotes the determinant and $\mathbf{I}$ the identity matrix [12]. The terms of the NLML in equation (2.21) are interpretable, in the sense that the the first addend determines the data fit, while the second addend determines model complexity (and is independent of the measurements). Complexity here refers to the range of probable functions that may be generated from the GP-prior, and is not necessarily tied to the number of model hyperparameters [10]. The property of the NLML to penalize overly flexible/complex models while favoring data fit is sometimes referred to as an automatic Occams razor, and mitigates the risk of over-fitting when training the model. It is one of the reasons NLML optimization is popular [8]. Now, to train the model amounts to taking the derivative of the NLML with respect to all parameters $[\theta,\sigma]$ and seek a minimum where that gradient is zero.
Training GPs can be computationally expensive as it naively scales cubically with the number of input points due to the necessity of computing the inverse of a large covariance matrix when evaluating the NLML in equation (2.21). There are a number of proposed ways to tackle this, such as inducing point methods which aim to effectively reduce the number of training points. One such method called KISS-GP, works for stationary kernels in particular which makes it suitable for this project [15].

2.3.1 Choosing between kernels

Hyperparameter optimization via minimizing the NLML requires that the kernel itself, with its unique set of hyperparamters, is already specified when optimization begins. The final choice between kernels can however be made after optimization of multiple models, via comparing the NLML corresponding to the optimal hyperparameter setup for each kernel under investigation. The difference in NLML between models has for example been used to detect degrees of periodicity in oscillating gene expressions, in a study where a Locally periodic model and Matérn$_{05}$ model were compared [16]. Naturally, such a comparison is only reasonable if the data fed into the different models is identical.

2.3.2 Non-convexity of optimization

A known issue with with optimizing kernel hyperparameters via minimizing the NLML is the multi-modality of the NLML loss surface. In other words, the optimization problem is not convex and one can not guarantee any obtained set of hyperparameters is in fact corresponding to a unique global optimum [17]. This is particularly expressed for kernels expressing periodic behaviour, or composite kernels formed by adding and multiplying together simpler kernels, such as the Spectral mixture kernels [14].

This issue must be minded at all times when one wishes to interpret the optimized hyperparameters of a GP model, or use them in some down-stream statistical analysis procedure. For such purposes, and in terms of obtaining good predictive performance, it is therefore key that well motivated initialization procedures are in place throughout the search for appropriate kernels and hyperparameters.

2.4 Decomposing Gaussian processes

As stated in the research questions, it is of interest to understand what aspect of a trace is captured by which part of a GP model. Part of the strong interpretability of GPs stem from the fact that if the kernel is a sum of covariance functions, its corresponding GP can be viewed as a sum of separate GPs, each having one of the kernel terms as a covariance function. For example, if we consider the GP $f = f_1 + f_2$ where $f_1(T) \sim \mathcal{N}(0, K_1(T,T)$ and $f_2(T) \sim \mathcal{N}(0, K_2(T,T)$, the distribution for $f_1(T^*)$ conditioned on observing $f(T) = f_1(T) + f_2(T)$ can be written as

$$p\left(f_1|(f_1(T)+f_2(T)\right) \sim \mathcal{N}\Big(K_1^{*\mathbf{T}}(K_1+K_2)^{-1}[f_1(T)+f_2(T)], \\ K_1^{**} - K_1^{*\mathbf{T}}(K_1+K_2)^{-1}K_1^*\Big), \tag{2.22}$$

where $K_1^{**} = K(T,T^*), K_1 = K_1(T,T), K_2 = K_2(T,T)$ and ^{T} denotes a transpose [13]. The formula for decomposing the process generalizes to sums of $i > 2$ kernels if $K_1 + K_2$ is replaced with $\sum_i K_i$. An example of such a decomposition, where the kernel used to model the signal consisted of three components, is shown in figure 2.7.

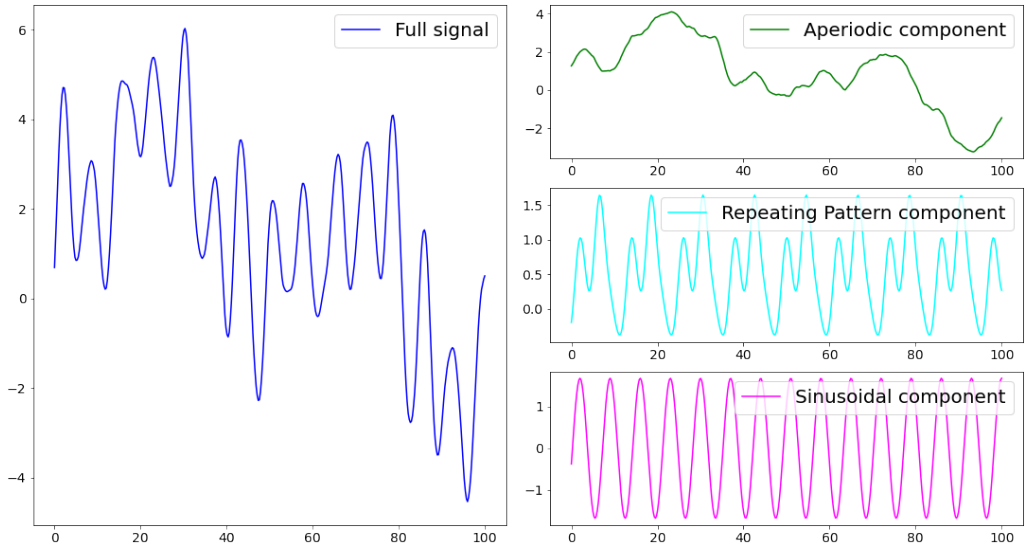

Figure 2.7: Decomposition of a process into three separate processess: 1: Aperiodic process given my Matérn$_{15}$ kernel. 2: Repeating pattern process given by Periodic kernel. 3: Sinusoidal process given by Cosine kernel. Only the means of the three posterior distributions are shown.

2.5 Periodograms & Lomb-Scargle

Biological processes often vary periodically. A periodogram is an estimate of the power spectral density of a signal. If there are periodicities in the signal being observed, these ideally show up as peaks in such a periodogram. Lomb-Scargle's method of estimating signal power spectral densities is useful since it's robust to gaps or unevenly sampled data [18].
As an example, figure 2.8 shows the power spectral densities of the sample paths given by the Locally periodic kernels shown in the right pane of figure 2.4. Recall that these samples come from processes with an approximate 12 hour periodicity and correspondingly there are strong peaks around 12 in the Lomb-Scargle periodograms.

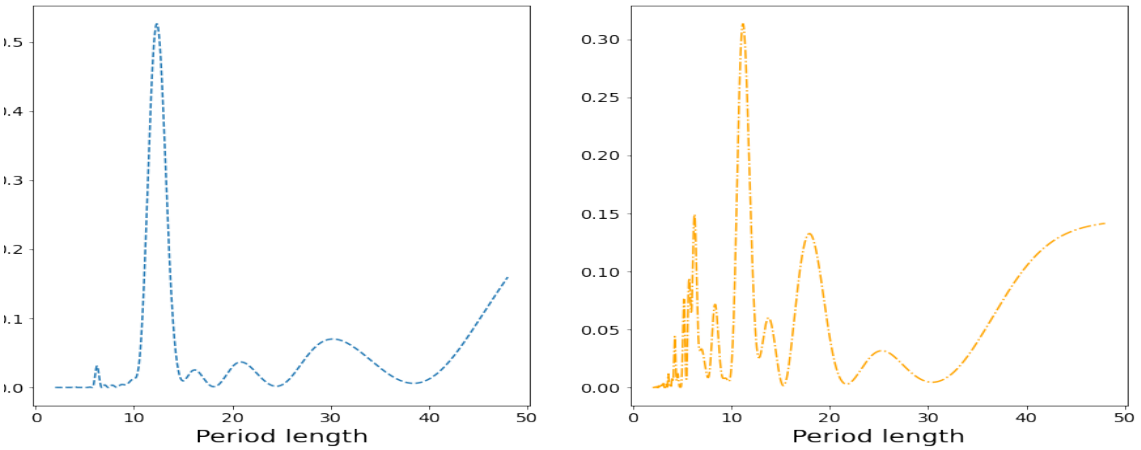

Figure 2.8: Lomb-Scargle periodograms of sample traces from GPs with Locally periodic kernels. The colors match the corresponding traces in figure 2.4.

3
Methods

For the purposes of this work clinical trial data as described in the Introduction were accessed. This chapter covers the main aspects of the workflow. Given the exploratory nature of this study, several secondary modelling aspects were developed ad hoc according to interim analysis results. The majority of the details pertaining to each GP model are therefore covered in the Results chapter.

3.1 Data characteristics and pre-processing

The data made available for this project consisted of blood glucose measurements from overweight and obese subjects suffering from type 2 diabetes. The study participants were randomly assigned to be either in a placebo or treatment group. 13 subjects were originally placed in the placebo group and 26 subjects were placed in the treatment group. Throughout the study, each participant in the treatment group was given an increasing dose of the drug under investigation over the first three weeks, whereafter a constant and high dose was administered daily for the remainder of the study period. The placebo group received none of the drug, but rather a pharmacologically inactive substitute. The objective of the study was to assess both drug tolerability and treatment effects [2]. Blood glucose values were measured over a period of about two months, and sampled every 15 minutes. For all subjects, there were multiple gaps in the data throughout this time period, likely corresponding to device malfunctions or bi-weekly scheduled sensor changes.

Since the data was given in the format of YYYY-MM-DD Hour:Minute:Second, for example 2017-10-04 12:28:00 with different initial start times for different subjects, it was decided that each subjects times should be converted to hours, such that the first reading of their CGM device would correspond to time zero. The conversion was facilitated via the R function **fasttime::fastPOSIXct**.

To characterize treatment effects, the idea was to train GP models on both the initial and final phase of the study and compare modelling results. This would require largely complete data during these phases. During explorations of the raw data, it was observed that most subjects had largely complete CGM readings during the first 12 day phase. In five cases, subjects had gaps larger than 24 hours throughout this phase, wherefore they were excluded from all subsequent analysis. It was also decided that any subject with any gap exceeding 72 hours, or having less that three weeks of readings in total throughout the final four week phase of the study, should

be excluded from the analysis. After filtering the data based on both these criteria, 30 subjects remained, 20 from the treatment arm and 10 from the placebo arm.

3.1.1 Study material and data considerations

For the purposes of this work participants in the placebo arm of the study are identified with letter P preceding a capital letter and similarly patients in the treatment arm are identified by the letter T preceding a capital letter. In addition, no metadata such as age, medical history, weight etc. is reported, since it was determined this would not substantially improve the quality of this thesis report but merely be an unnecessary exposure of such patient characteristics.

3.2 Software

Python 3.7.9 was the programming language used for GP modelling throughout this thesis project. The software used for fitting GP-models was GPyTorch version 1.3.1 [19]. It was selected based on its extensive documentation available online, an implementation of Spectral mixture Kernels at present and being comparatively fast. Since all kernels used throughout the thesis work were stationary, the scalable inference method KISS-GP was used to speed up hyperparameter optimization in cases where GPyTorch's standard methods were severely time consuming [15]. This includes all results up until section 4.3. It should be noted that when making predictions, the KISS-GP models were converted back to regular and more exact GP models, since time was never an issue for such inference purposes.

3.3 Optimization

Optimization here refers to finding a set of hyperparameters for a certain kernel on a given data set that minimises the NLML, given in equation (2.21). GPyTorch is built on PyTorch [20] functionality, meaning it offers the same type of options for optimizers. PyTorch version 1.7.1 was used in this project. Although a number of optimizers including L-BFGS were explored, all results shown in the report were obtained by use of the Adam optimizer with a learning rate set to 0.4. GPyTorch reports the NLML divided by number of training points, which in practice meant this loss gave magnitudes around 0-10. The reason for choosing Adam was partly its superior speed but also the tendency for other optimizers to crash due to various numerical issues. All computations were made with double (64 bit) precision to avoid an otherwise commonly occurring problem with near-singular matrices. Based on the heuristic that optimization can halt whenever the loss has stopped decreasing, the optimization scheme was implemented such that optimization ended if the variance of the losses through the last five epochs was below 10^{-5}, or the number of iterations exceeded 100. The minimum number of epochs was set to 21.

An interesting observation concerns the impact of the scale of the CGM-data. Values of hyperparameters did not seem to change over optimization epochs when the input

data was on its original scale. However, once data was normalized to have a variance of one, optimization became more flexible. This is clearly a desirable feature, wherefore all optimization was done in this way. The only affected kernel hyperparameters are the scale parameters σ, which can at any time be re-scaled via multiplying with the original standard deviation of the CGM-trace under investigation.

3.3.1 Fixating the noise parameter

The GP models under consideration can be expressed as:

$$\begin{aligned} \text{Hyperparameters: } &= \theta \\ \text{GP Prior : } p(f \mid T, \theta) &\sim \mathcal{N}(0, K_\theta(T, T)) \\ \text{Likelihood : } p(y \mid f) &\sim \mathcal{N}(f, \sigma^2) \end{aligned}$$

The likelihood noise parameter σ^2 is what connects the latent GP to the observed blood glucose levels and models the uncertainty of observed values.
In theory, since this is a parameter that enters the kernel matrix, as seen in equation (2.11), and the log marginal likelihood, it can be inferred via optimization along with other hyperparameters.

Throughout the early analyses made on this particular type of CGM data, an observation was that this parameter was always optimized to values close to its theoretical limit at zero. These results are in conflict with information from the CGM device manufacturers, which report a Mean absolute Relative Difference (MARD) of around 10%. MARD is calculated by comparing estimated blood glucose values between the CGM device and a more refined reference method which is assumed to give the true value. A closer look at how blood glucose values are reported, as exemplified in figure 3.1, revealed that observations are confined to a discrete set of values.
Therefore it was deemed unreasonable to try to optimize the noise parameter and it was fixed to 10% of the signal variance. Exact modelling results, e.g optimized hyperparameter values, will vary depending on where this parameter is fixed at, but in this investigation, the conclusions were not altered when the value was set to a lower level of 1%, wherefore such results are omitted in this report. In addition, this "binning" of glucose values as a consequence of sensor technology likely makes the observed process very rough. With this in mind, the RBF kernel was avoided when modelling short term variations, since that covariance function models very smooth, infinitely differentiable, functions.

3.4 Detecting and quantifying periodicities

To detect periodicities in the CGM-traces, Lomb-Scargle's method for estimating power spectral densities was used. The peaks of the periodograms dictated the use of kernels encoding periodicity and guided the choice of where to initialize such kernels. The computation of normalized periodograms was performed via the Python implementation **scipy.signal.lombscargle**. Normalization amounts to dividing all

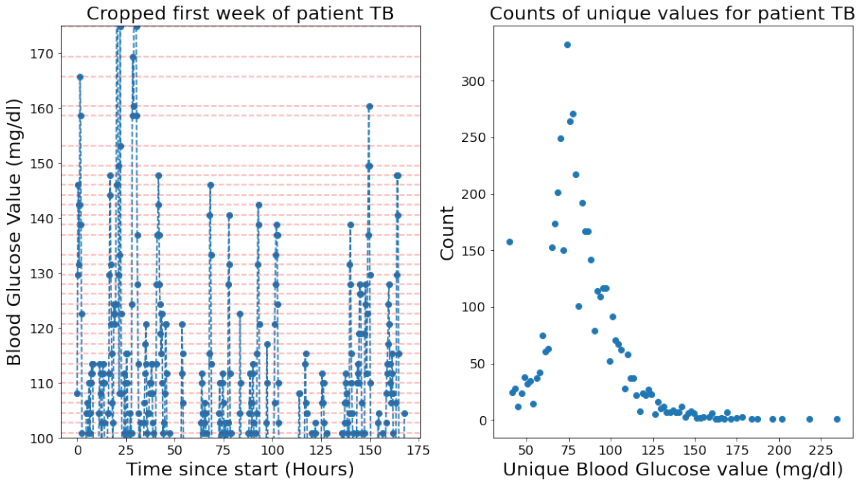

Figure 3.1: Left pane: Fragment of a reported CGM trace. All observations land precisely on the horizontal lines. Right pane: Count of all unique blood glucose values for patient TB.

powers by the sum of squared residuals of a mean-zero constant model, and multiplying by two. This makes PSD values end up in the range 0 to 1, which is convenient when comparing data sets.

Regarding the range of frequencies for the method to try, the upper limit was set to 1/hour and the lower limit to 1/48 hours. The lower limit was set after trying lower value and seeing no further peaks in the periodograms. This corresponds to the method not indicating any long-term periodic trends in the data. The number of points to evaluate frequencies at between the upper and lower limit was set to 10000.

3.5 Model selection and comparisons

Model selection refers to the construction of kernels as sums and products of simple kernel functions. Model comparisons refer to comparing NLML fit for a given patient and time-frame, and comparing fitted hyperparameters of models with training data either belonging to the first 12 days (initial phase) of the study, or the final 14 days (final phase). Increasingly complex models to evaluate and compare between subjects were primarily selected based on emerging research questions in relation to CGM data. The models discussed in detail throughout the Results, Discussion and Conclusions are all mean-zero models, with either of the following kernels:

1. Matérn_{05}
2. Matérn_{15}
3. $\text{Matérn}_{05} + \text{Periodic} \cdot \text{Matérn}_{25}$
4. $\text{Matérn}_{15} + \text{Periodic} \cdot \text{Matérn}_{25}$

To tackle the problem of getting stuck in local optima, multiple random restarts

were performed based on re-initializing the model hyperparameters with draws from pre-defined distributions. For reproducibility, these draws were made with a fixed random seed so as to make the initialization sites identical across different patients, when using the same kernel. The details of these initializations are described in the Results section.

To evaluate whether any model outperformed another on the data, the NLML was computed between the ten best fitting models for each kernel choice and patient. For example, when comparing the Matérn_{05} and Matérn_{15} models throughout the final 14 days of treatment, the NLML of the top ten Matérn_{05} models were subtracted from the NLML of the top ten Matérn_{15} models, for each patient individually.

An idea was to compare fitted hyperparameters between models based on the initial and final phase, respectively, for each patient individually. This in an effort to find systematic differences between the treatment and placebo group. The methods with which to compare fitted hyperparameters across the final and initial phase were highly dependent on the model under investigation, wherefore they are covered in detail in the Results section. An overarching idea was that one must verify that the best models with respect to NLML for a given data set had similar hyperparameters, before any attempts at comparisons of hyperparameters between data sets could be made.

3.6 Detrending and decomposing signals

Using GPs, a long-term trend can be estimated jointly with short-term variations by summing kernels with variance at very different lengthscales. In practice it was found that such attempts mostly ended up in any long-term component getting a variance of zero, even in cases of obvious trends. Therefore, in cases where it was of interest to capture or remove long-term variations separately, a detrending step was included.

To avoid the problem of specifying a functional form for the long-term trends, a GP consisting of a single RBF kernel with a lengthscale l of four days was fitted to the data. The lengthscale parameter was chosen so as to mitigate the risk of the GP capturing any short-term or daily variations. The noise parameter σ^2 was set to 10% of the total signal variance. The posterior predictive mean of this GP model was subtracted from the initial blood glucose values to perform the detrending. A variant of this method has been implemented before in a study that aimed to identify oscillating gene expressions masked by potential long-term trends [16]. An illustration of this method is shown in figure 3.2

In cases where sums of kernels were used, posterior decompositions were made in order to evaluate what aspects of the signal, for example the commonly occurring strong peaks, were captured by the different components.

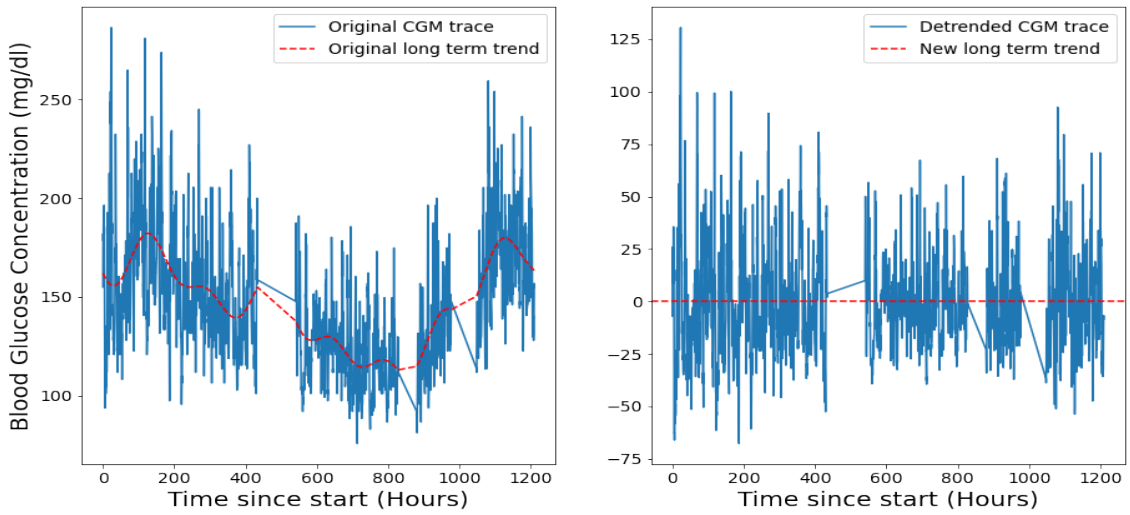

Figure 3.2: Illustration of the detrending procedure. Left pane: Original trace of patient TO. Right pane: Detrended trace of patient TO

3.7 Predictions via Spectral mixture kernels

The Spectral mixture kernels were explored with respect to how well they could predict future blood glucose values after being trained on detrended data. If predictions were successful, it would indicate that the SM-models had picked up hidden patterns in the data, which would warrant a further inspection of their fitted hyperparameters and analytic power spectra.

Before evaluating the prediction power of GPs with SM-kernels on CGM data, they were evaluated on simulated data. Data was simulated by sampling from GPs with three different kernel functions, after which Gaussian noise with a variance of 10% of the simulated signal variance was added to the data. Since the covariance functions of these three stationary processes were known, it also allowed for an investigation into the capability of SM-kernels to approximate other stationary kernels. The kernels used to simulate the data were:

1. Periodic kernel with period 12 and lengthscale 0.2 plus Cosine kernel with period 4.
2. Periodic kernel from 1 plus Cosine kernel with period 7 plus RBF kernel with lengthscale 20 plus Matérn$_{15}$ kernel with lengthscale 1.
3. Periodic Kernel with period 12 and lengthscale 0.03 + Matérn$_{15}$ kernel with lengthscale 1.

These kernels have defining characteristics such as repeating patterns, sinusoids and short-term fluctuations, as inspired by insights into CGM data characteristics.

The number of Spectral mixture components to use in each Spectral mixture GP model was selected by counting the number of peaks in the periodogram of the sim-

ulated signals. The GP models were fitted to the first 80% of the simulated data and evaluated at the final 20%, and the noise parameter σ^2 was fixed to 10% of the signal variance for all investigated data sets.

Throughout the simulation study, the learning rate of the Adam optimizer was set to 0.1, since a larger value resulted in the optimized kernel mimicking a Short-term kernel with no extrapolation capabilities. When evaluating the performance of SM-kernels for predictions, both a visual inspection of the predicted trace and computations of the mean squared error of predictions were used.

3.7.1 Initializing the Spectral mixture kernel hyperparameters

For the spectral mixture kernel, initializations were made via fitting a Gaussian Mixture Model (GMM) to the subjects normalized power spectral density as computed by Lomb-Scargle's method. GMMs are described in appendix A. The method was adapted from existing code in GPyTorch that employed Fast Fourier Transforms to estimate the PSD, a technique which assumes all observations are evenly spaced apart in time. This motivated the change to Lomb-Scargle's method, since the CGM data often contained multiple gaps.

Since the present implementation is based on a normalized power spectrum, the weights of the fitted GMM were multiplied with the variance of the data prior to assigning these as the Spectral mixture component weights w_q in equation (2.18). That way, the prior variance at any time would be equivalent to the variance of the data under consideration. It was assumed that due to the inherent stochasticity in fitting GMMs to data, several restarts of this initialization method would lead to different starting points being explored.

4

Results

The main focus of this work was to evaluate if certain GP models can provide insights into Glycemic Variability. In section 4.1, models based on Matérn kernels are investigated in terms of model fit and whether their hyperparameters can be interpreted in clinical terms. In section 4.2.2, composite kernels accounting for periodicity are investigated similarly. In section 4.3, Spectral mixture kernels are investigated as a means to make predictions of future glucose values, and provide insights into periodic trends in CGM data.

4.1 Matérn kernels to evaluate treatment effects

A review of common measures of Glycemic Variability, revealed that many measures are concerned with the amplitude of signal variations (Standard Deviation, Coefficient of Variation) and don't take into account how rapidly glucose values swing between high and low levels [1]. It is of desire to control both these phenomena [7]. The family of kernels that are parametrised by a lengthscale and outputscale parameter, as for example the Matérn kernels, have the potential to characterize both these aspects. As is seen in figure 2.1, a shorter lengthscale corresponds to a process with rapid fluctuations, whereas a longer lengthscale corresponds a slowly varying process. The outputscale σ^2 instead controls the over all amplitude of the variability. With this in mind, it was hypothesized that by fitting a GP model with a simple Matérn kernel to all patients, differences in Glycemic Variability between the initial and final phase of the study that might not be revealed by measures such as Coefficient of Variation, could be uncovered. Both the Matérn_{15} and Matérn_{05} kernels were investigated.

The workflow for each of the 30 patients can be summarized as follows:

1. Split data into initial 12 days and final 14 days
2. Normalize data according to standard deviation of the initial 12 days
3. Fit 50 GP-models starting from different initial conditions, according to predetermined initializaton schemes.
4. Check that the best models with respect to NLML have similar hyperparameter values.
5. Compare fitted hyperparameters across the two time frames and treatment groups.

The initialization scheme for these kernels was to sample outputscales uniformly from the range 0.5-1.5, and lengthscales uniformly from the range 2-8. The rationale behind this was that outputscales should largely match the variance of the

normalized data, and that short term variations should not have strong correlations to glucose values more than 8 hours away.

To begin with, it was observed that the Matérn$_{15}$ kernel gave a better fit with respect to NLML than the rougher Matérn$_{05}$ kernel for virtually all patients, both throughout the initial and final phase. To illustrate this, the differences in NLML between the 10 best Matérn$_{15}$ and Matérn$_{05}$ models for each patient over the final 14 day phase are shown in figure 4.1. The Matérn$_{05}$ model NLMLs were subtracted from the Matérn$_{15}$ model NLMLs. Thus, a negative value means the Matérn$_{15}$ is favored.

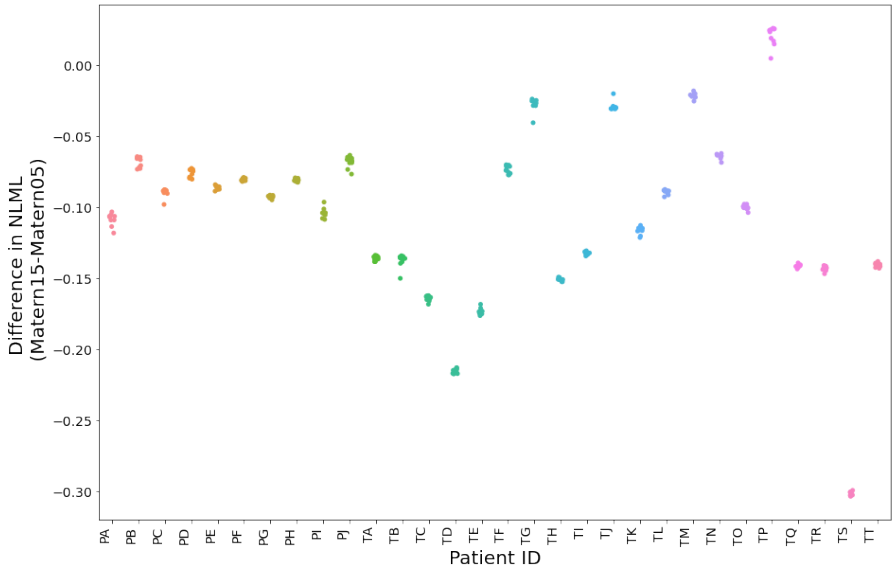

Figure 4.1: Difference in NLML between Matérn$_{05}$ and Matérn$_{15}$ models. Almost all values are negative, meaning that for each patient, the Matérn$_{15}$ model gave a better fit.

Therefore, only results from models based on the Matérn$_{15}$ kernels were analysed further.

In figure 4.2, the fitted hyperparameters for Matérn$_{15}$ models, computed from the final 14 days of the treatment group, are shown. Only the hyperparameters of the top ten models with respect to NLML for each patient are shown. The quotients between the lengthscale and outputscale parameters for every patients top ten models are shown in the right pane.
There is great uncertainty in the optimized hyperparameters as can be seen by the spreads for all patients in figure 4.2. For a given patient, the optimized values of lengthscale and outputscales seem correlated, making analysis of one parameter in isolation of the other problematic. As illustrated in the right pane of figure 4.2, the LO-quotient $\frac{Lenghthscale}{Outputscale}$ between the two measures is more concentrated. This was

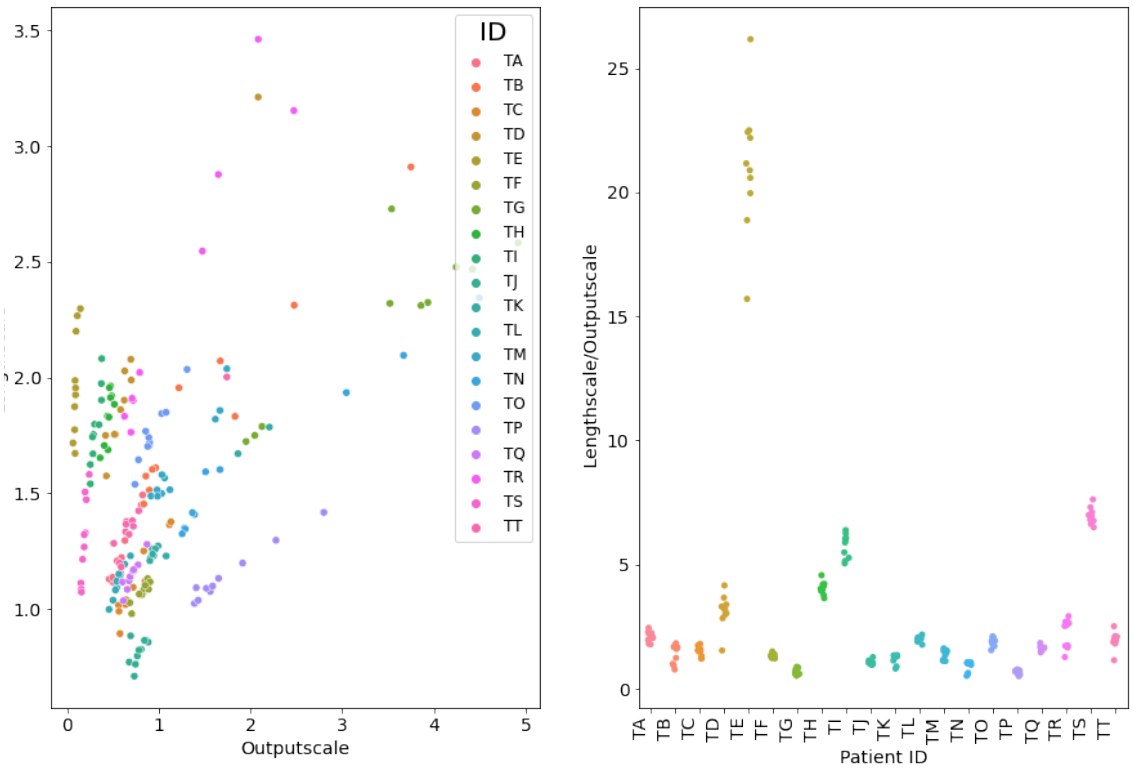

Figure 4.2: Left pane: Fitted outputscales and lengthscales. Right pane: Quotients of fitted lengthscales/outputscales. Parameters are taken from the top ten Matérn$_{15}$ models for each patient in the treatment group based on the final 14 day phase.

particularly true in the final phase of the treatment group. For the other three data sets, the spread of quotients across the top ten optimized models were significantly larger.

This LO-quotient has a neat interpretation in terms of glycemic variability, as it is desirable to both have a low variance and slowly fluctuating levels, the latter corresponding to long lengthscales. To compare between the initial and final phases of this study, the LO-quotient for both phases was computed for the two phases respectively. The quotient corresponding to the best model for each phase and patient was selected. Then the fold change, as defined by the quotient of LO quotients between the final and the initial phase, was computed. Results are shown in figure 4.3.

Figure 4.3 indicates a number of subjects in the treatment group have strong improvements between the initial and final phase, as evidenced by a higher LO-quotient in the final phase, corresponding to a value above 1 on the y-axis in the figure. Recall that the input data for fitting models in both phases was normalized to the standard deviation of the initial 12 days. Thus, any strong long-term trend superimposed on the Short-term fluctuations would skew the results with respect to

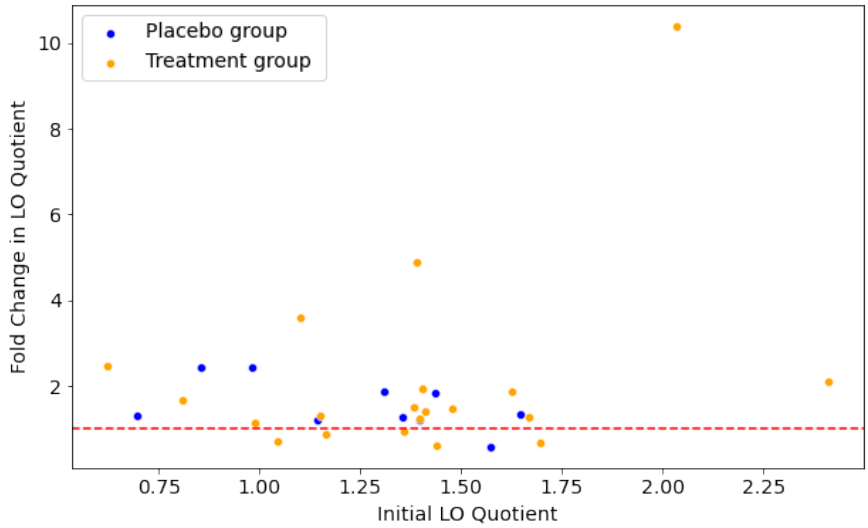

Figure 4.3: Fold changes in Lengthscale/Outputscale (LO) quotients between the final and initial phase of all patients, colored by study group. Fold changes are computed as the LO-quotients in the final phase over the LO-quotients in the initial phase.

estimating the relative outputscale between the two phases. Such a trend would also violate the assumptions of stationarity associated with using a Matérn kernel, making its hyperparameter estimates unrepresentative. Figure 4.4 illustrates that this is exactly the case for the two subjects that showed the strongest improvements in LO-quotient. This motivated the next step, which was to remove the long-term trend via detrending prior to looking for differences among the model hyperparameters.

4.1.1 Comparisons on detrended data

Recall the distinction between glucose control and Glycemic Variability (GV), where the former relates to the overall level of glucose values whereas the latter relates to temporary fluctuations [5]. Detrending can be thought of as an attempt to remove variability due to changing glucose control, to better estimate GV only.

After detrending all subject traces according to the procedure detailed in section 3.6, models were fitted with a Matérn_{15} kernel as described in 4.1, with the difference that values were normalized by the standard deviation of the detrended trace of the initial 12 days. This way, any long-term trend would not affect the comparison between phases. Figure 4.5 illustrates how the strong correlation between fitted outputscales and lengthscales persists for the detrended models, hindering any direct comparisons between such parameters across groups.
The LO-quotient shown in 4.5 was yet again only stable for the final phase of the treatment group. In the other three conditions, the spread of quotients across the top ten optimized models were significantly larger. Thus, figure 4.6 which illustrates

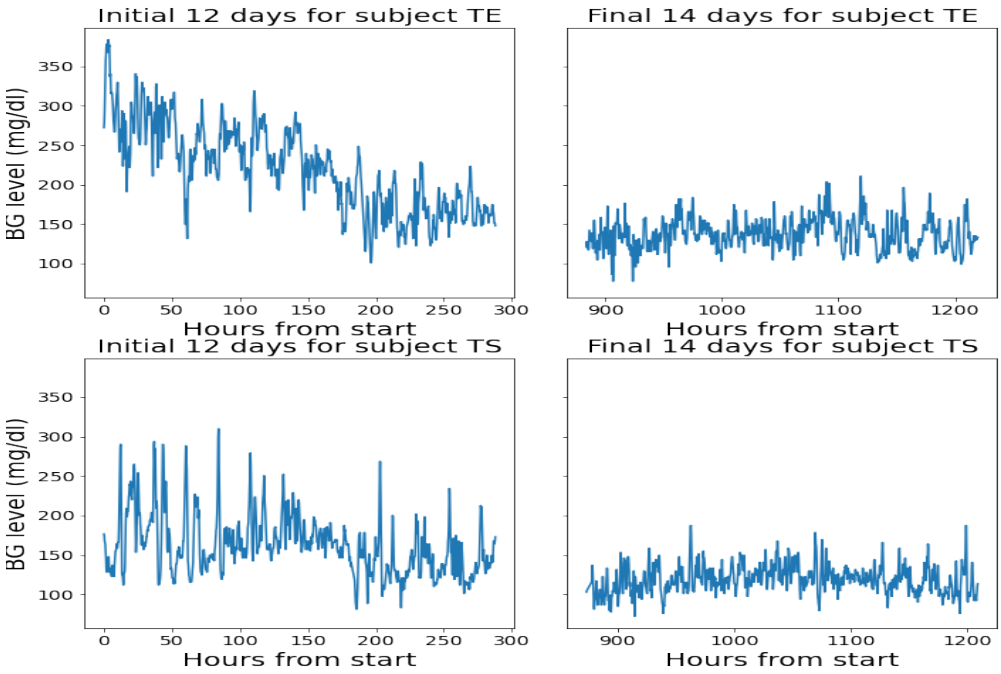

Figure 4.4: Glucose traces of most improved subjects with respect to the LO-quotient throughout their final and initial phase. Both display long term trends in the initial phase.

changes in LO-quotients between the best models across the two phases on the detrended data, can not be said to provide reliable evidence of changes in GV.

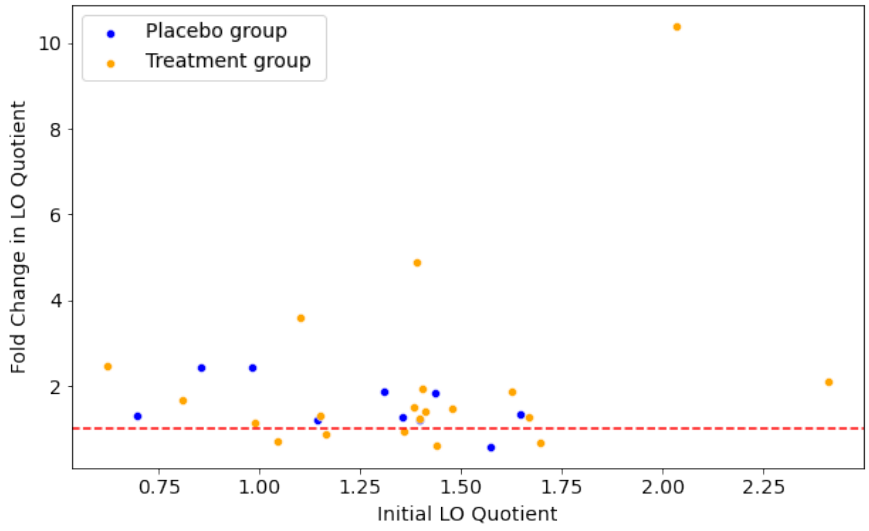

Figure 4.6: Fold changes in Lengthscale/Outputscale quotients between the final and initial phase of all patients, colored by study group and computed on detrended data. Fold changes are computed as the LO-quotients in the final phase over the LO-quotients in the initial phase.

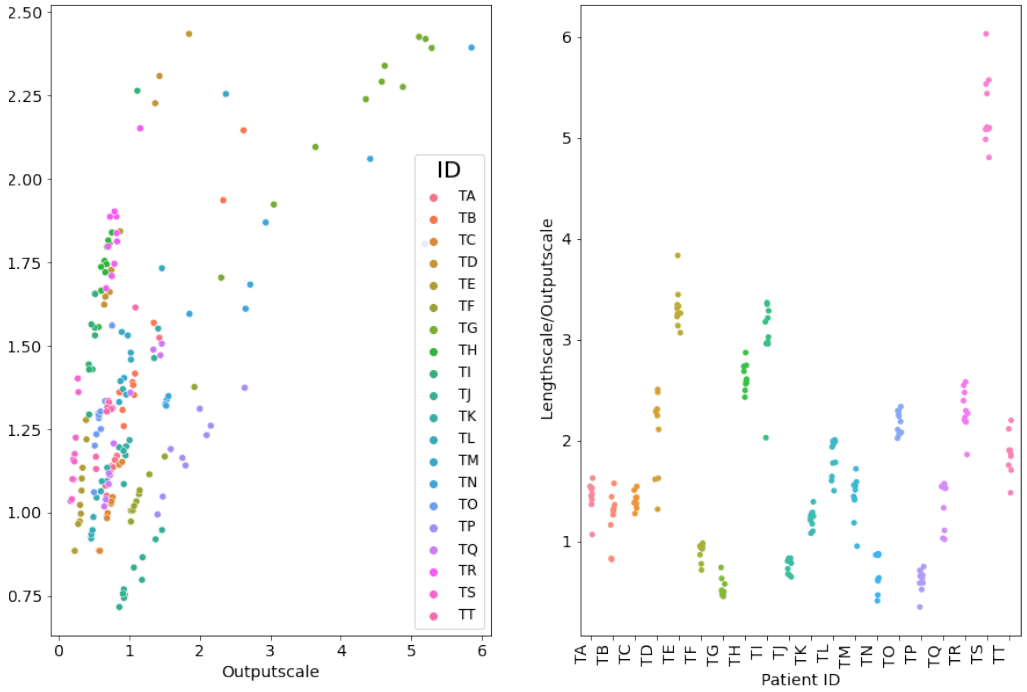

Figure 4.5: Left pane: Fitted outputscales and lengthscales. Right pane: Quotients of fitted Lengthscales/Outputscales. Parameters are taken from the top ten Matérn_{15} models for each patient in the treatment group, based on the final 14 day phase on detrended data.

4.2 Incorporating structure: periodicity

One potential shortcoming of the simple Matérn model is that it assumes no repeating structure of the CGM traces. However, a circadian rhythm is well documented for blood glucose variations. Thus, adding periodicity assumptions into the GP-priors could be a means of capturing such rhythms. To guide the choice of which periodicities to encode into the GP-priors, periodograms were computed from the full two months of data for all patients.

Figure 4.7 illustrates four typical periodograms of the CGM data. A defining feature of most patients is a very strong peak around at 24, often accompanied with one at 12 as well. This was the case both in the treatment and placebo group. 26 out of 30 patients displayed an approximate 24 hour peak, 19 displayed a 12 hour peak and only 3 lacked these peaks completely. There were a few occurrences of 6 and 8 hour peaks. Based on these periodograms, it was decided that a kernel component with an almost 24 hour periodicity should be included in the models.

To model the periodicity, the Periodic kernel which models repeating patterns was selected. It was multiplied by a Matérn_{25} kernel so as to reduce both its smoothness and exact periodicity assumptions. In an attempt to also capture variability that was not associated with any periodic behaviour, a Short-term component was added to this Locally periodic model. Two variants of the Short-term component

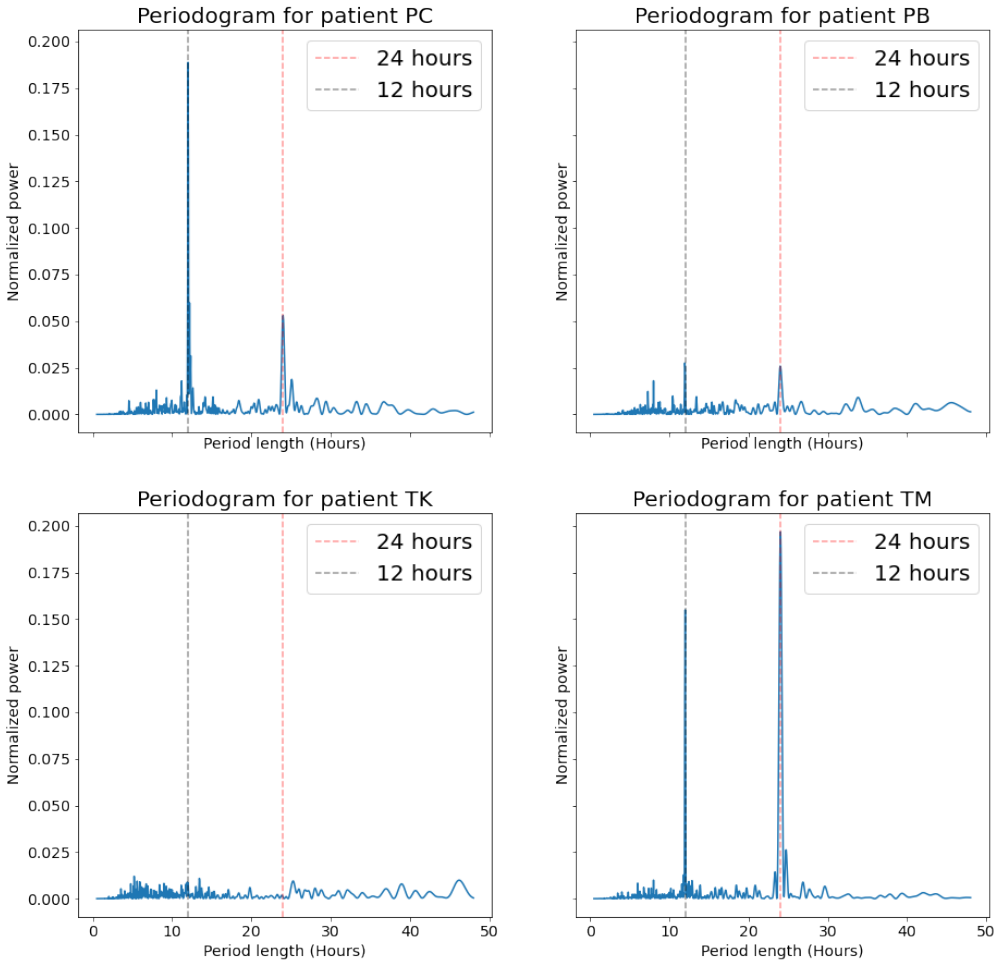

Figure 4.7: Four typical periodograms for the CGM data. The bottom left pane shows a patient who was deemed to display no periodicity.

were investigated, the Matérn$_{05}$ kernel and the Matérn$_{15}$ kernel. The former is characterized by being rougher than the latter, and as discussed in section 4.1 was shown to give a worse fit across virtually all patients, when used on itself.

To summarize, the two kernels under investigation were:

$$\text{Rough Short-term+Locally periodic: } Matérn_{05} + Periodic \cdot Matérn_{25} \quad (4.1)$$

$$\text{Smoother Short-term + Locally periodic: } Matérn_{15} + Periodic \cdot Matérn_{25} \quad (4.2)$$

In order to avoid a potential masking effect of long-term trends, an identical detrending and normalization procedure to that of section 4.1.1 was performed on each individual data set. Therefore, a simple model comparison via computing the difference in NLML between these composite models and the previously studied Matérn$_{15}$ was enabled.

When optimizing the models, 50 different initial conditions were tested. They were identical for the two models under investigation. The initialization scheme for the

Short-term components was to sample its outputscale uniformly from 0.5-1.5 and the lengthscale uniformly from 2-8. As for the Locally periodic component, the period lengths were sampled from a normal distribution with mean 24 and variance 0.8, since the intention was to model a circadian rhythm. The periodic lengthscale controlling for within period flexibility was sampled uniformly between 1/8 and 1/4 of the period length, and the Decay lengthscale was sampled uniformly from between 3 and 4 times the period length, allowing for slight between day variations. The Locally periodic kernel outputscale parameter was sampled uniformly between 0.1 and 0.9.

4.2.1 Hypotheses for Locally periodic models

Four hypotheses in relation to CGM variability were set up to be evaluated via these models:

1. Can the Locally periodic component reliably capture the regularly occuring, presumably meal related spikes in the data, also known as post-prandial peaks?
2. If the Locally periodic component can capture the meal related peaks, can the Short-term component model between-meal variability or Fasting Glucose Levels.
3. Amplitudes of post-prandial peaks are known to vary across days. Can this difference in amplitude between days be captured by the Short-term component, while the Locally periodic component captures what's similar on each day?
4. Are Locally periodic models favored over Matérn_{15} models?

4.2.2 Analysing Locally periodic models

Throughout this section, the two Locally periodic models based on the kernel functions in equation (4.1) are referred to multiple times. To distinguish between the two, the one set up with a Matérn_{05} kernel for its Short-term component will be referred to the Rough model. The other will be referred to as the Smoother model to reflect these models' relative smoothness.

The Smoother model gave a lower NLML than the Rough model for virtually every patient among its top ten models, as shown in figure 4.8. Only the results from the final phase are shown but the same holds true for the initial 12 day phase.
To quantify to which extent the periodic component captures any variability, one could think to study differences in fitted outputscales between a model's Locally periodic and Short-term components. Before doing so, the decomposition property of GPs given in equation (2.22), was used to split the CGM-traces into two processes, based on the Short-term and Locally periodic components respectively.

In figure 4.9, such a decomposition is shown for a patient, both for the initial 12 days of the study and the final 14 days. The original trace plus three and the long-term trend are also included in the illustration, along with the variance of the respective traces and components.

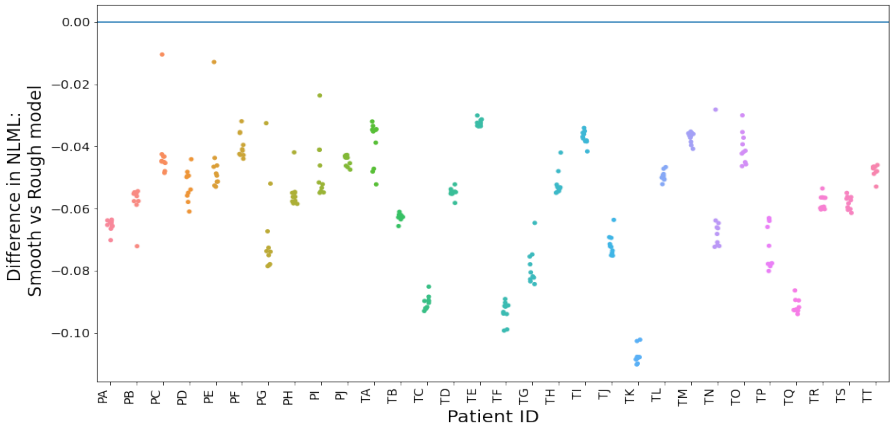

Figure 4.8: Difference in NLML between the top ten Rough and Smoother models for each patient in the final phase. A negative value indicates the Smoother model gave a better fit.

In figure 4.9 a strong regularity in peak occurrences is visible, e.g in the final phase there are roughly three peaks per day at similar times apart. It can be seen that the periodic component captures much of the peaks in the data, but that the Short-term component also captures some of that variability. This suggests it may be useful to compare the outputscale parameter of the Short-term component between the initial and final phases, to get an estimate of changes in how variable the peaks are between days. However, such an analysis could be problematic, based on two factors.

Firstly, the tendency for the top ten models within each patient and phase to have variable optimized hyperparameters holds true for the Smooth and Rough models as was the case with the simple Matérn$_{15}$ models investigated in section 4.1.
Secondly, as is illustrated in figure 4.10, there are examples of patients where the Locally periodic component has negligible variance in one of the phases. This is reflected in the hyperparameters with an extremely small outputscale for the Locally periodic component of that patient and phase. In such cases, the short-term components of the different phases can not be said to model the same type of variability.

If it were uncommon for the Locally periodic component outputscale to be set to zero, perhaps the corresponding patients could simply be ruled out of the analysis. However, as is illustrated in figure 4.11, where each blue dot represents a model where the outputscale of the Local Periodic component has been optimized to below 0.1, this is not the case. The illustration covers the top ten models fitted to the final 14 days for each patient for both the Rough and Smoother models, but results are similar for the initial phase. Figure 4.11 also shows that in cases where the outputscales/variances of the Locally periodic components have been estimated near zero, their periodic lengthscale hyperparameters take on particularly variable values. This is unsurprising, since if there is little contribution from the Locally periodic kernel to the overall covariance function, changing its hyperparameters will have little or no effect on the model's performance. On the other hand, in virtually

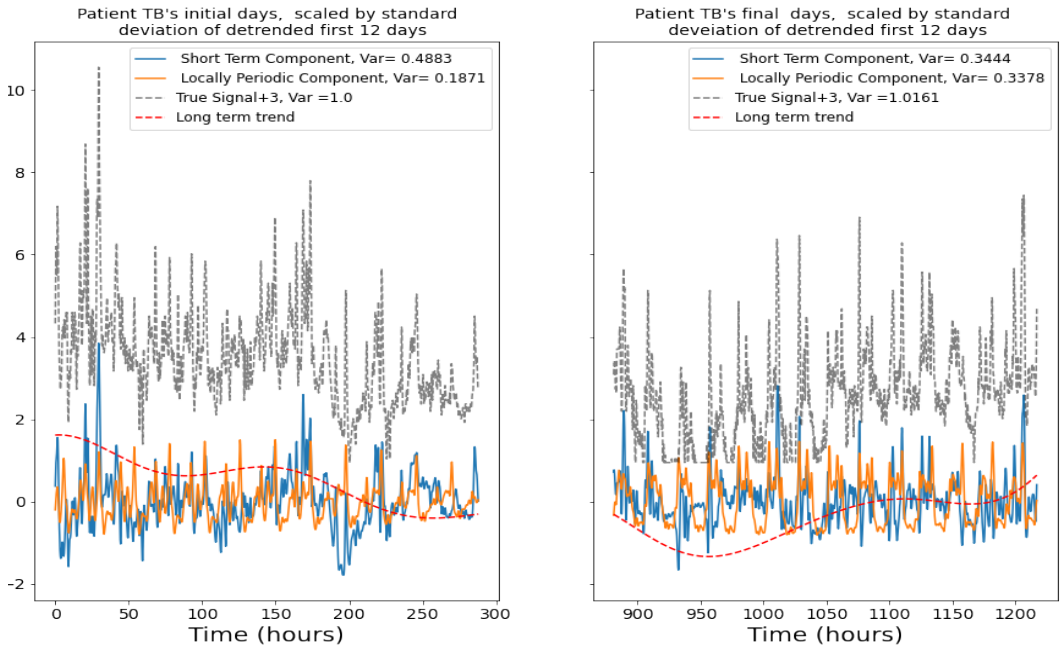

Figure 4.9: Detrended and decomposed signal for both the initial and final phase of patient TB. The decomposition was based on the best Smooth model with respect to NLML.

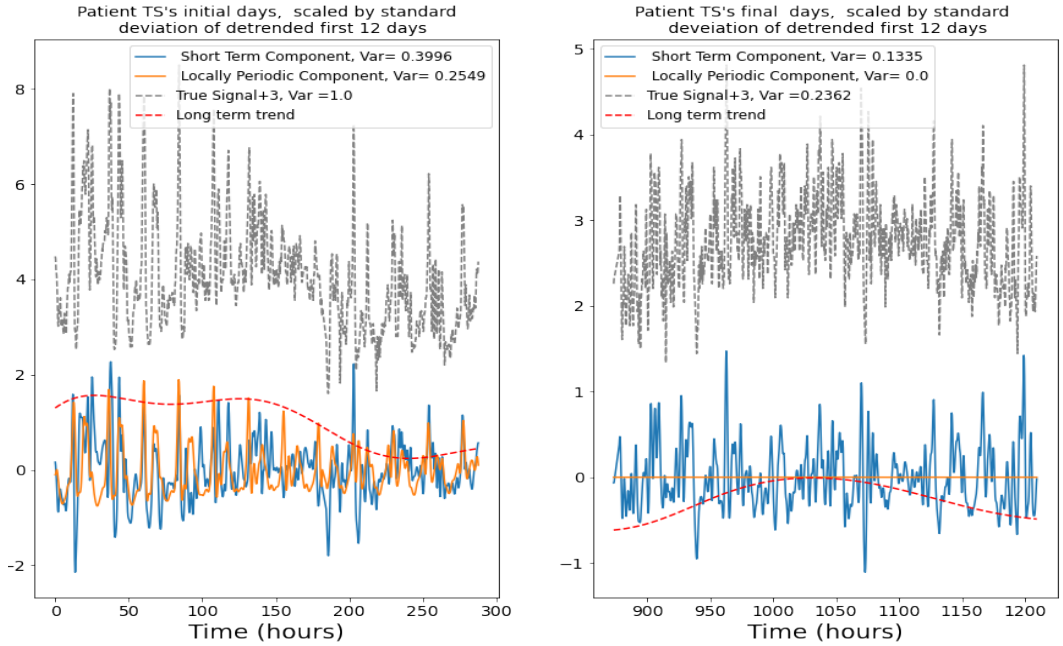

Figure 4.10: Detrended and decomposed signal for both the initial and final phase of patient TS. The decomposition was based on the best Smooth model with respect to NLML.

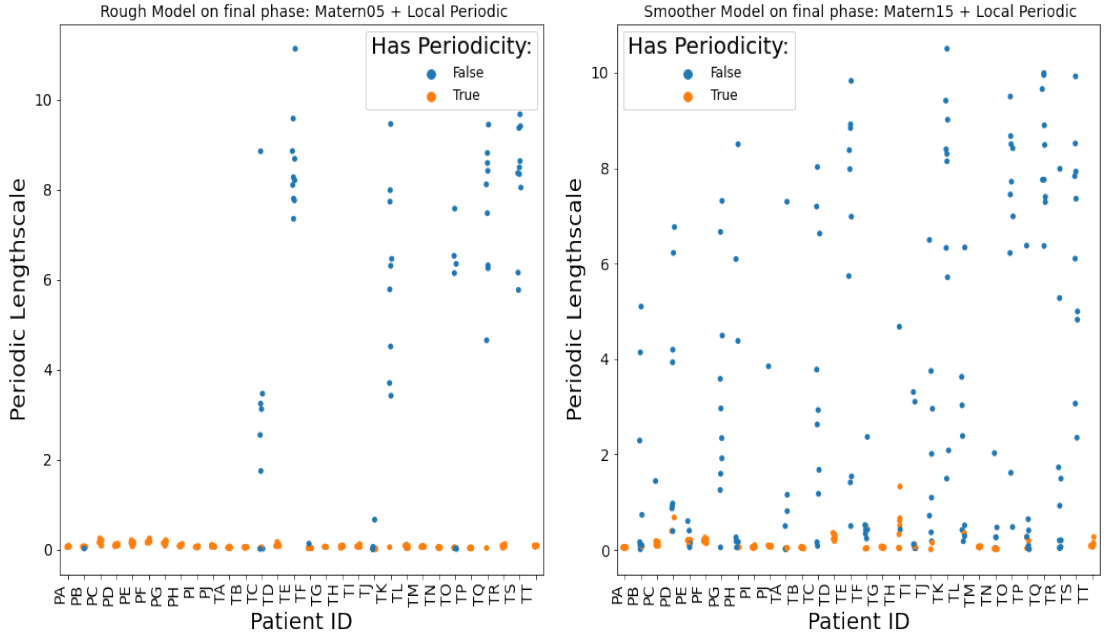

Figure 4.11: Optimized periodic lengthscales of the top ten Rough and Smoother models for each patient. Each dot represents a model and the blue color indicates that particular model had its Locally periodic outputscale parameter optimized to below 0.1.

all cases where the Locally periodic component outputscales are larger than 0.1, the periodic lengthscale is set to a very small value, between 0.01 and 0.5. Recall that this hyperparameter controls the flexibility with which the process can vary within a period. With values as small as 0.01, the correlation between points as little as 15 minutes apart is minimal.

To investigate why the Rough model had a tendency to support periodicity more often than the Smooth model, their fitted hyperparameters were investigated. A marked difference is illustrated in figure 4.12, which displays histograms over the % changes from initial to optimized values of the Decay lengthscale among the top ten models for each patient in the final phase. The histograms show there is a tendency for the Decay lengthscale to be optimized to be shorter in the Rough models, and longer in the Smoother models. Recall that they were initialized to the same values.

Figure 4.13 illustrates an example where only the Rough model captured variations via its Locally periodic component. Here,the Rough model has a much lower Decay lengthscale at 41 compared to its Smoother counterpart at 96, and note the trace of its Locally periodic component does vary considerably in both amplitude and overall pattern across different days.

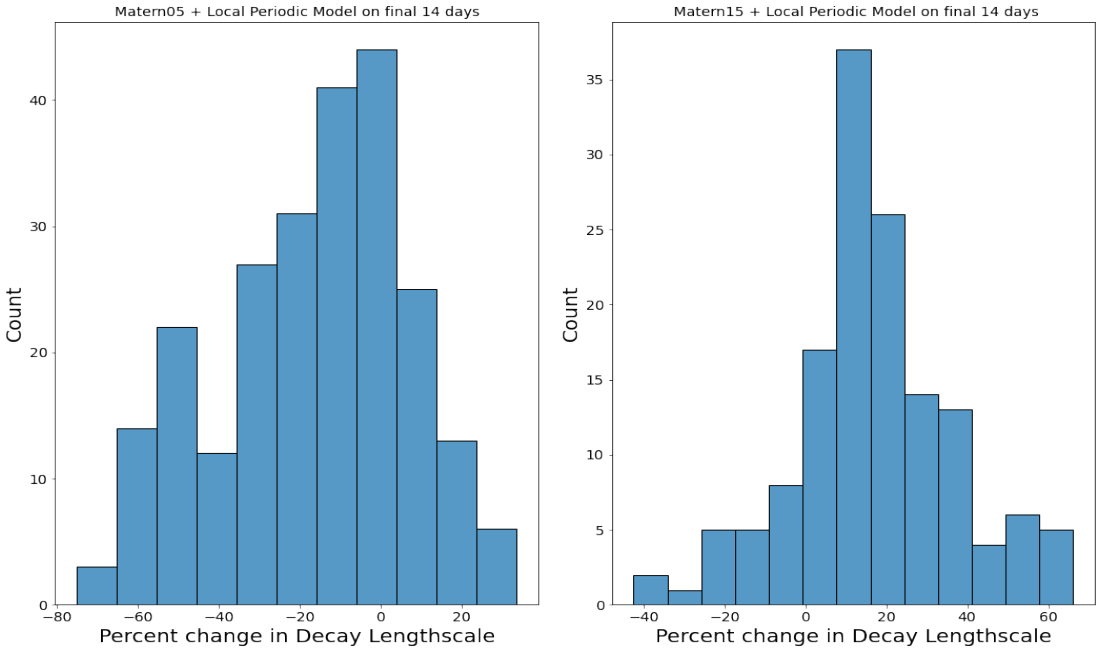

Figure 4.12: % Changes from the initial guess to optimized values of the Decay lengthscale parameter among the top ten models for each patient in the final phase. Left pane: Results for Rough model. Right pane: Results for Smoother model.

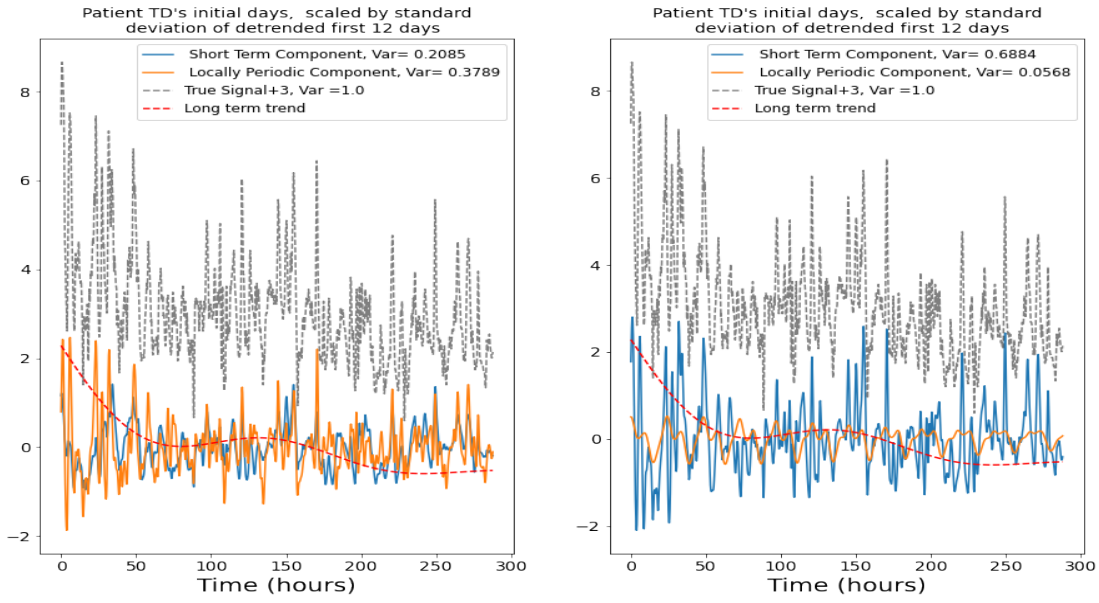

Figure 4.13: True signal, long term component and decomposition based on the best Rough and Smoother model for the initial phase of patient TD. Left Pane: Rough model, Right Pane: Smoother model.

To investigate the hypotheses regarding periodic models being favored over aperiodic

models, the difference in NLML between the aperiodic Matérn$_{15}$ model and the ocally periodic Smoother model on was computed. This is illustrated in figure 4.14, where the top ten Smoother and Matérn$_{15}$ models of all patients throughout the final phase are compared. This result indicates the model incorporating periodic structure is favored for many patients. In cases where the outputscale of the Locally periodic component was optimized to near-zero, the difference is consistently smaller, in line with the idéa that in such cases, the two models encode very similar structure.

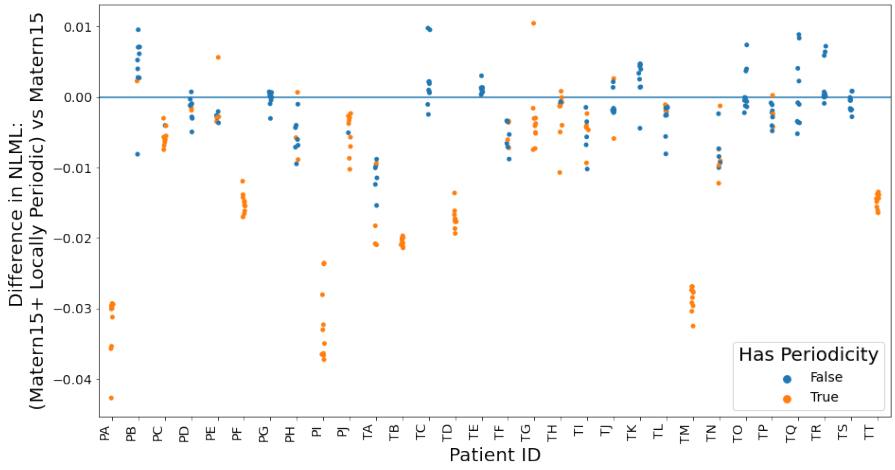

Figure 4.14: Comparison of NLML between top ten Matérn$_{15}$ and Smoother models for each patient throughout the final phase. A negative value means the Smoother model is favored for that patient. A blue dot indicates the Locally periodic component outputscale in the Smoother model was optimized to a value below 0.1.

4.3 Extrapolations with Spectral mixture kernels

For GP-models to extrapolate, they must have covariance functions that do not decay fast toward zero, since otherwise they can not learn from past values. This is illustrated in figure 2.6, which compares kernels in terms of predictions. For short-lengthscale Matérn models, predictions quickly revert back to its mean-zero prior, whereas periodic models predict the same pattern indefinitely. Models based on Spectral mixture kernels, as defined by sums of locally periodic kernels, can ideally find a good trade off between the behaviour of those models.

4.3.1 Spectral mixture kernels on simulated data

The kernels used to simulate the data for this study are listed in section 3.7. These kernel functions are referred to as the ground truth in this section.
Figure 4.15 shows the modelling results for the Periodic+Cosine kernel simulation. The model effectively manages to discover the source of variability, as evidenced by the estimated covariance function very closely matching the true covariance function.

The initialization based on Lomb-Scargle's method was also fairly close to the ground truth.

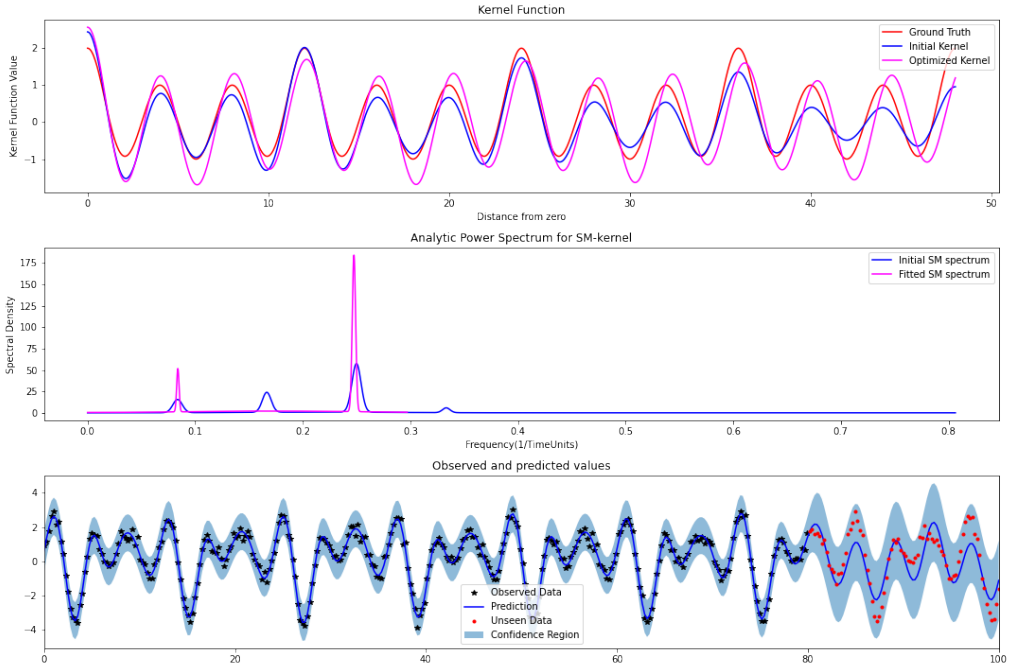

Figure 4.15: SM modelling of data simulated from Periodic + Cosine kernel. Top pane: True and estimated covariance function. Middle pane: Initial and optimized analytic power spectrum of the SM-kernel. Bottom pane: The simulated signal and SM-kernel prediction. The confidence region covers two standard deviations

Figure 4.16 shows the results for the Periodic+Cosine+Matérn$_{15}$+RBF kernel simulation, where a 10 component SM-model was fitted to the data. The estimated covariance function is slightly off, but the extrapolation is fairly accurate. Again, the optimized kernel function is not very far off from the initialization based on Lomb-Scargle's method. Notice that the analytic spectrum corresponding to the optimized hyperparameters has a "bell"-shape starting from about zero, which is how the Fourier transform of an RBF-kernel looks. This indicates the model has found the hidden long-term trend simulated via an RBF kernel with a lengthscale of 20.

Figure 4.17 shows the results for the simulation from a Periodic kernel with period 12 and lengthscale 0.03 plus Matérn$_{15}$ with lengthscale 1. 15 mixture components were used, since the observed spectra contained that many peaks. This kernel choice was made based on the observation that on the CGM data, the Locally periodic kernels tended to get such short periodic lengthscales, as evidenced in figure 4.11.

For the Periodic + Matérn$_{15}$ simulation, the SM-model failed to accurately estimate the ground truth covariance function, as seen in figure 4.17. In particular, it failed to obtain a near-zero covariance in between the 12 units spaced-out peaks. Correspondingly, the extrapolation is poor as can be seen from the prediction line not coinciding with the recurring peaks in the simulated data.

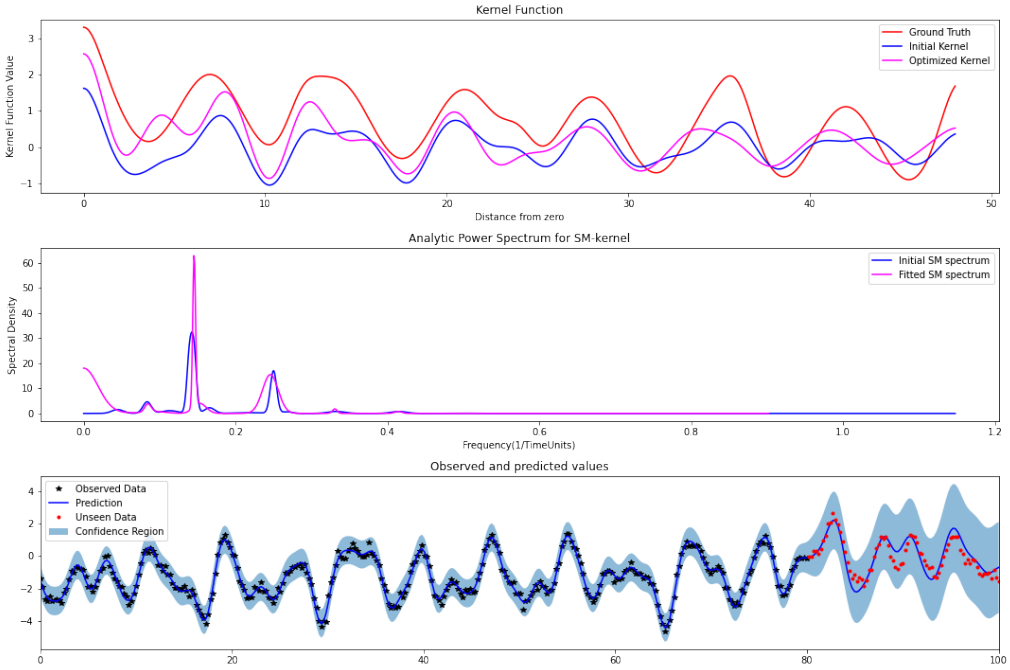

Figure 4.16: SM modelling of data simulated from the Periodic+Cosine+Matérn$_{15}$ + RBF kernel. Top pane: True and estimated covariance function. Middle pane: Initial and optimized analytic power spectrum of the SM-kernel. Bottom pane: The simulated signal and SM-kernel prediction. The confidence region covers two standard deviations

4.3.2 Spectral mixture kernels on CGM data

Just like in sections 4.1 and 4.2.2, the CGM-data was detrended prior to being fitted with an SM-kernel, so as to mitigate the effects of non-stationary long term trends. To incorporate the long-term trend in predictions, one could simply add its predictive mean after predicting the shorter-term variations separately.

As an example of the SM-kernels predictive performance on CGM data, a 10 component model was fitted to the the initial 12 days of patient PC, who's periodogram showed a very strong 12 hour peak, as seen in figure 4.7. The two following days were used for evaluating the extrapolation. For completeness, both the model with the lowest NLML on the train data, and the model with the lowest mean squared error of prediction on the evaluation data, are displayed.
Figure 4.18 shows the results for the minimum NLML model. The initialization method pushes the model toward a kernel with a 12 hour periodicity, as we can see a "hill" around 12 in the kernel function. This "hill" is still present in the optimized model, and the covariance function resembles that of a cosine kernel. With respect to predictions, the model fails to capture the magnitude of the peaks and valleys in the data.
Figure 4.19 shows the results for the model with the minimum mean squared error of prediction. For this model, the optimized parameters suggest an approximately 8-periodic model as opposed to the 12-periodicity given by the initialization. In terms

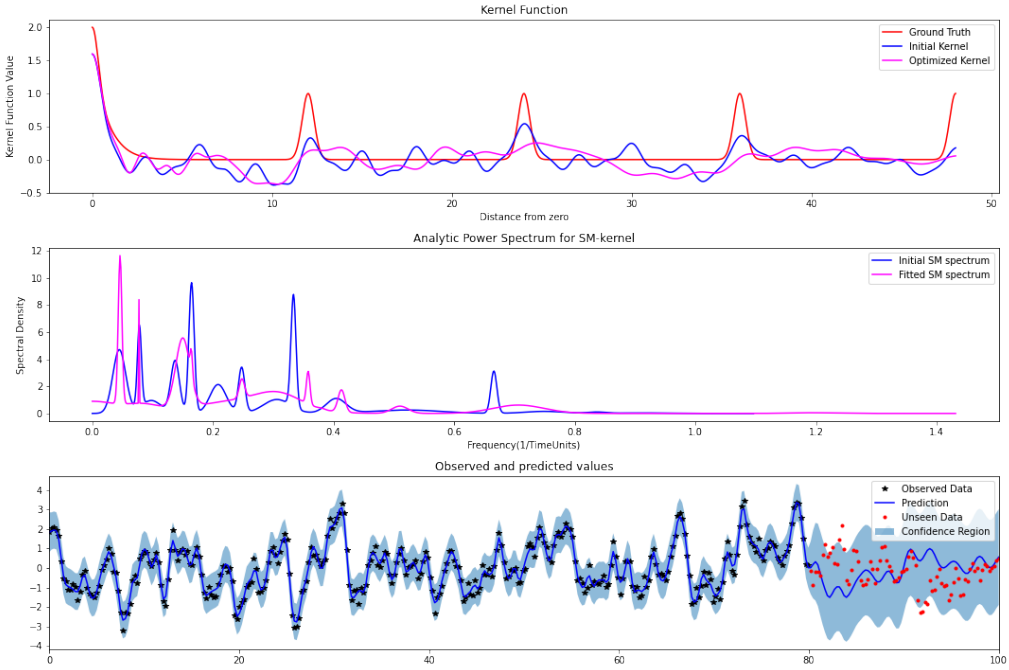

Figure 4.17: SM modelling of data simulated from the Periodic+Matérn$_{15}$ kernel. Top pane: True and estimated covariance function. Middle pane: Initial and optimized analytic power spectrum of the SM-kernel. Bottom pane: The simulated signal and SM-kernel prediction. The confidence region covers two standard deviations

of predictions, this model also fails to capture the magnitude of the CGM trace peaks altogether. Figures 4.19 and 4.18 are merely examples from a single patient, but throughout testing on other subjects, no case was observed when the SM-models accurately predicted the magnitude of the peaks, even on a short timescale.

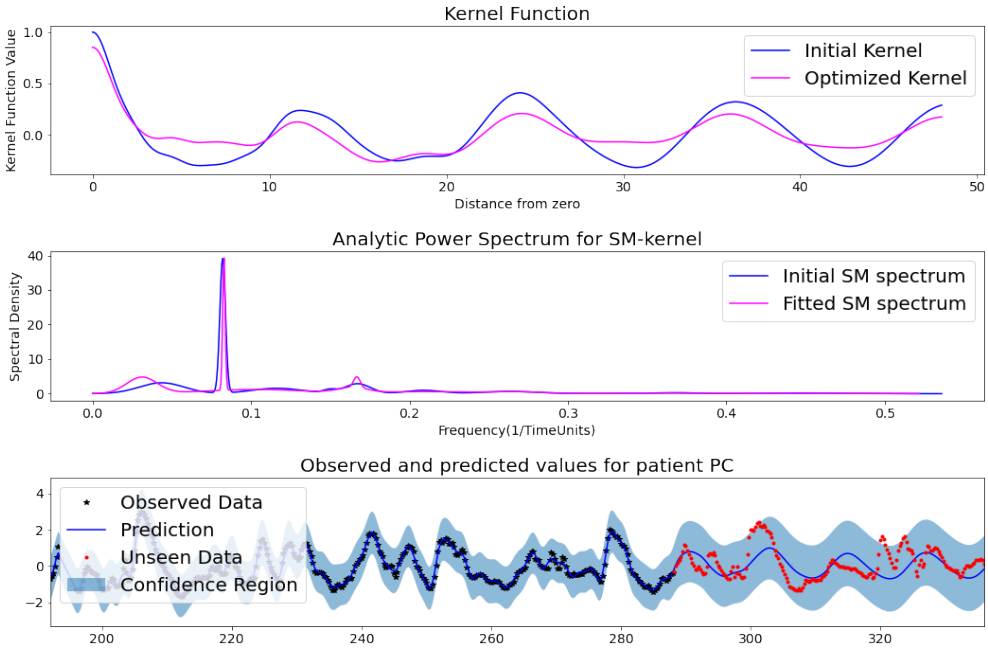

Figure 4.18: Spectral mixture modelling results on detrended initial 12 days of patient PC. The model depicted had the lowest NLML on training data.

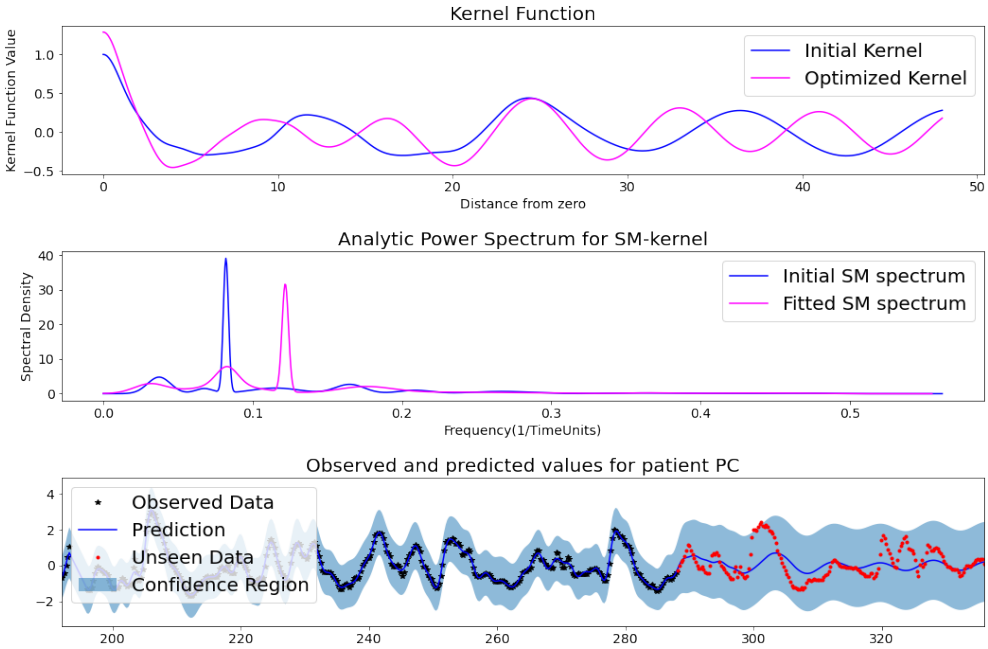

Figure 4.19: Spectral mixture modelling results on detrended initial 12 days of patient PC. The model depicted had the smallest mean squared error of prediction on the evaluation data.

5

Discussion

In section 4.1, GP models based on Matérn kernels were analysed, in an attempt to provide a reliable, interpretable and robust measure of glycemic variability (GV) that is more granular than measures such as the Coefficent of Variation. It was observed that the optimized hyperparameters for each subject were highly variable throughout random restarts, as exemplified in figure 4.2. This variability made comparisons of fitted hyperparameters, e.g the quotient $\frac{Lengthcale}{Outputscale}$ between the final and initial phase of the study period, unfeasible. It was hypothesised that observed long-term trends could be a cause of the uncertainties in hyperparameter optimization, since such trends contradict the assumptions of stationarity associated with the Matérn kernels. However, computations on detrended data yielded hyperparameter variability similar to that on the original scale, thus the models based on Matérn kernels did not yield any measures that were considered useful to characterize GV.

In section 4.2, the value of incorporating periodicity into the GP models was investigated. The approach was motivated by strong peaks in Lomb-Scargle periodograms of the CGM data, as exemplified in figure 4.7, and executed via addition of a Locally periodic kernel to the simpler Matérn models investigated in the previous section. Four hypotheses, as stated in section 4.2.1, were investigated.

The hypothesis that the Locally periodic component could "control" for the frequently occuring peaks, to allow analysis of the short-term noise as a proxy for Fasting Glucose Levels, was not supported by the results, as exemplified in figure 4.9, where the peaks in the true data are captured both by the Locally periodic and Short-term components. The hypothesis that the Locally periodic components could reliably capture the common peaks is not supported by the results either. As exemplified in figure 4.10 there are cases where that component has a negligible variance. The example in figure 4.10 also suggests comparisons between the Short-term outputscale of the final and initial phases would be inadequate, since sometimes the signal variance is shared between components and sometimes it's carried completely by the Short-term component.

Nevertheless, figure 4.11 shows that in virtually all cases where the Locally periodic outputscale component was not forced toward having zero variance, its periodic lengthscale parameter was optimized to a very small value. An explanation for this phenomenon could be that the periodic component is indeed trying to fit to the presumably meal related peaks. Consider that close to a peak, we expect large changes in glucose values. A correlation function that assumes high correlations to nearby

points, such as a long-lengthscale kernel, would assign a high likelihood of staying at the peak. This is unprecedented in the data and therefore, the lengthscale of the Locally periodic component is forced toward small values.

The study of differences in optimized hyperparameters between models parameterized with Matérn_{05} kernels as opposed to s Matérn_{15} kernels for their Short-term components, revealed that the Decay lengthscale was often optimized to be much shorter in the Matérn_{05} case, as seen in figure 4.12. This tendency coincided with fewer Short-term Matérn_{05} models having a Locally periodic component outputscale pushed toward zero, and might be an explanation to this phenomenon. To see why, recall that the Decay lengthscale governs to which extent the periodic pattern can vary across periods, with a long Decay lengthscale forcing the model to have near exactly repeating patterns and a short one allowing for both changes in amplitude and overall pattern across nearby days, as seen in figure 2.4. With a short enough Decay lengthscale, the Locally periodic kernel might not encode any clearly distinguishable periodicity at all, as appears to be the case in figure 4.13. In such scenarios it's conceivable that the Locally periodic component would not be forced toward zero, even when the trace being modeled is not particularly periodic.

Overall, the strong dependency on the Decay lengthscale regarding what type of variability the Locally periodic kernel can capture, suggests it is not ideal for modelling what is similar in peaks across different days. Accordingly, the Short-term component in the investigated models is likely not ideal for modelling differences in peaks across different days. Still, the comparison of NLML between the Locally periodic models and completely a-periodic Matérn models, shown in figure 4.14 indicated that the Locally periodic models were favored in many cases, in particular when their periodic component was not pushed toward zero. This result suggests incorporating structure such as periodicity in CGM-models may be favorable. However, results do not support that single Locally periodic kernels are immediately useful to characterize clinically relevant features of CGM data, such as post-prandial peaks. It appears more sophisticated models are needed to reliably account for these strong and variable peaks.

Regarding the use of Spectral mixture kernels for extrapolation, simulations indicate the current initialization and optimization method works well for certain types of traces, such as those seen in figures 4.15 and 4.16. However, figure 4.17 indicated that for data simulated from a simple periodic kernel with a short periodic lengthscale, it can be difficult for the SM-kernels to approximate its covariance function and extrapolate well, given the current optimization procedure. Recall that the analysis of Locally periodic models suggested many CGM data sets are best modeled with such a short periodic lengthscale, as shown in figure 4.11. These observations might explain the poor predictive performance of the Spectral mixture models on the CGM data, as exemplified in figures 4.19 and 4.18. This highlights the need for more advanced optimization methods than those implemented in the scope of this project, if the SM-kernels and their spectra are to be interpreted in clinical terms.

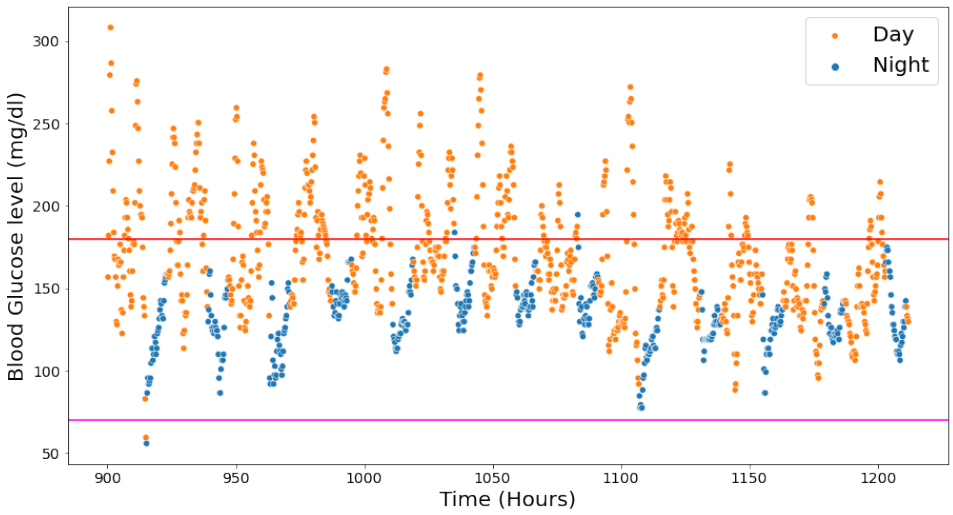

Figure 5.1: Example trace of blood glucose values, colored by night and day time. Night time was defined as times between 00:00 and 06:00. The horizontal lines denote the limits for hypo- and hyperglycemia.

Resorting to Lomb-Scargle's method for analysing the frequency content of CGM traces could also be considered, since that on itself indicated there are strong periodic tendencies in blood glucose dynamics of most patients, and heterogeneity between patients. Perhaps the lack of periodicity is associated with relevant clinical features, such as adverse events affecting food intake?

5.1 Study limitations

None of the attempts at modelling blood glucose dynamics detailed in sections 4.1,4.2.2 and 4.3 could be said to provide immediately useful insights for clinicians. As the search over models has not been exhaustive, one can not conclude GPs are not useful for modelling CGM traces.

Importantly, all the investigated kernels throughout this project were stationary, which means the CGM-traces were assumed to be realizations of stationary processes. A violation of this assumption was apparent when the occurrence of long-term trends was observed. To remedy this, data was detrended prior to being modelled. However, long-term trends are not the only sources of non-stationarity. It is for example known that blood glucose levels are different between nights and days, with hypoglycemia being more common at night [21]. An example of this tendency from the CGM data is shown in figure 5.1.
In addition to lower levels, overall GV during nights might be different than during days. Figure 5.2 illustrates this, where the ratios of blood glucose variance between night and day were computed for both the final and initial weeks, across both treatment groups. The analysis was made on detrended data so as not to obscure the

variability of interest with any long-term trends.

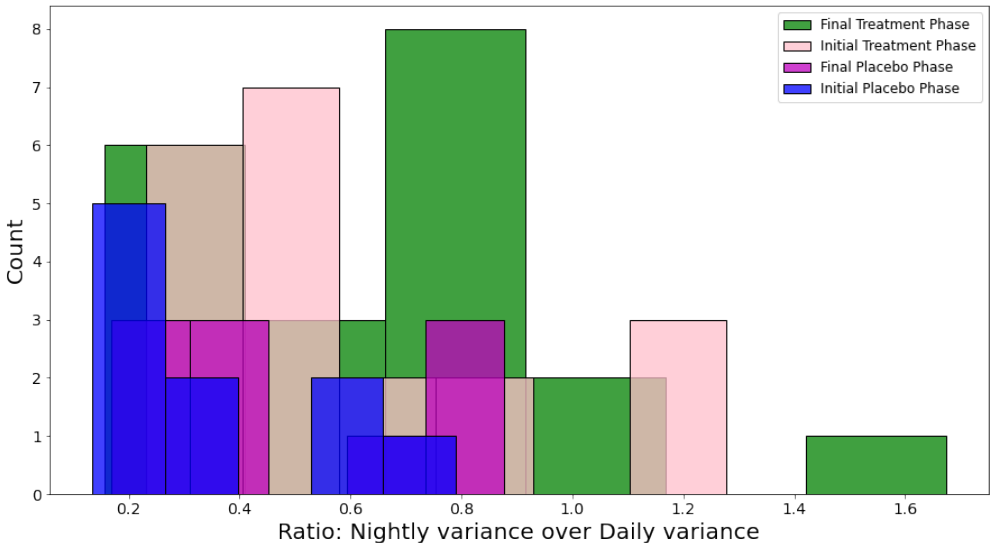

Figure 5.2: Histogram of ratios of variance of blood glucose levels between night- and day time on detrended data, for all patients. A value above 1 indicates the day time variance is higher. Night time was defined as times between 00.00 and 06.00.

Figure 5.2 supports the claim that for most patients, the night time variance is lower than its day time counterpart. The few exceptions may be of interest to study more closely. Perhaps some events occurred during the study which dramatically changed those patients daily routines. This result suggests that a model which somehow incorporates the time-of-day in its covariance function, might be favorable. Naturally, this would be a non-stationary model.

Another cause of non-stationarity concerns meal times. If it is a-priori known when meals occur, it's also known when peaks are more likely to occur. Thus, the statistical properties of the signal are different throughout different times of day, certainly for patients with regular meal times. In a study where GPs were used as part of an artificial pancreas system, data-points following a meal were for example excluded from the training data set, so as not to worsen predictions throughout the rest of the day [11]. They were possible to exclude because the meal times were simulated and known, and this allowed the researchers to use the stationary Periodic $\cdot$ Exponential kernel to model the blood glucose dynamics. It was beyond the scope of this investigation to figure out how to infer meal times and potentially incorporate that information into the kernel function. A starting point for future attempts at building non-stationary GP-models could be to consider the approaches detailed by Cheng et al. [22], which tailor specifically toward clinical data.

As mentioned in section 3.3.1, the assumption of Gaussian observation noise was already violated for this data. However, there are also other causes for concern in terms of Gaussianity of the CGM traces. Due to the common occurrence of several-standard-deviation peaks, CGM data is naturally skewed to the right, and heavy tailed. The skewness issue is accentuated by the fact that corresponding dips in

glucose levels are physiologically impossible. This is illustrated in figure 5.3 where histograms of the glucose values of all individual patients are shown, both on original and detrended data. In addition, the distribution of skewnesses across all subjects in this study are shown. For reference, the skewness of a Gaussian distribution is 0, since its mean and median coincide.

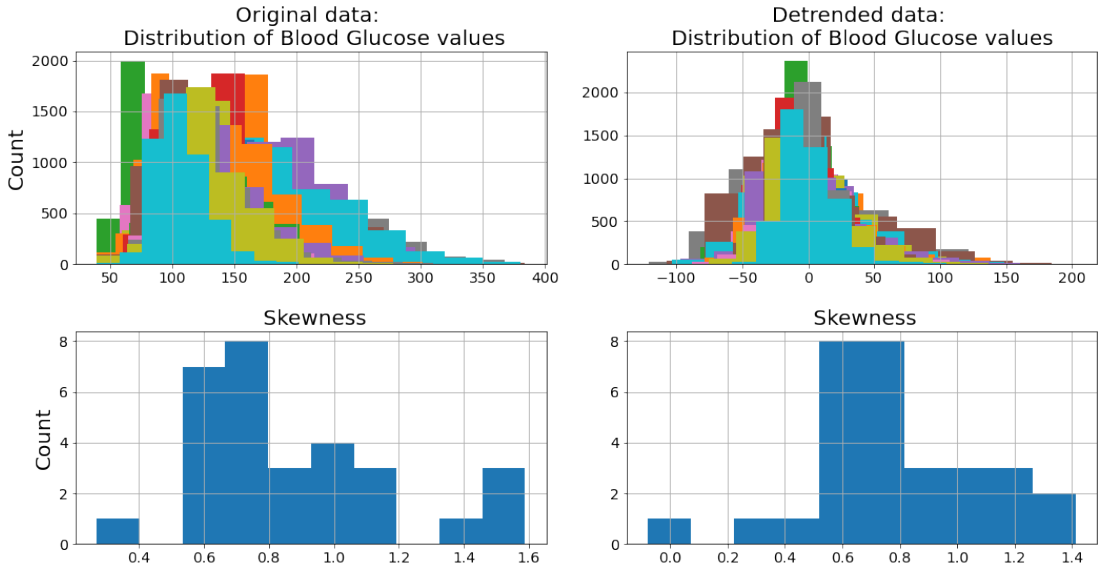

Figure 5.3: Illustration of the skewness of CGM data. Histograms of glucose values for all patients across the full study period are shown, and the skewness computed for each patient separately. The left panes regard original data and the right panes regard detrended data.

Skewness is known to be problematic in GP regression [23]. As an example of why, consider that all priors over the latent process being observed are symmetric, since they are all Gaussian themselves. Thus, any such prior over functions will have a low probability of generating asymmetric traces such as those observed in this study, making the prior a poor guess. Recall also that the marginal likelihood used to evaluate model fit, given in equation 2.20, is itself Gaussian. Thus, the extreme values corresponding to blood glucose peaks will have a strong impact on the value of this likelihood, potentially making it an unreliable measure and hyperparameter optimization based on minimizing the NLML, uncertain.
Certain transformations, such as the natural log or square root, would reduce the right-skewedness of the CGM traces and potentially make them more Normal-like. An effect of such a transform would be that the impact on regression of hypoglycemic episodes would be weighted up while hyperglycemic episodes would be weighted down.

Finally, it is strongly advised to consider the technical details of the CGM device used with regards to how it reports blood glucose values. It might be the case, as was observed in this study, that observations are rounded off or confined to a discrete set of values. If so it might be best to fix the noise parameter to a high

value so as to avoid inferring an underestimated noise level.

6
Conclusions

For this work, several GP models were fitted to clinical trial CGM data, with the aim to interpret their optimized hyperparameters in clinical/biological terms. Results showed support for locally periodic GP models on CGM data in favor of completely aperiodic models. Moreover, results indicate that Locally Periodic kernels with an approximately 24-hour periodicity can to some extent model the frequently occurring peaks in blood glucose levels. It remains to extend these models so as to more reliably capture such peaks, which are a defining characteristic of CGM traces.

Due to the observed uncertainty in the optimized hyperparameters of the investigated models, results indicated that the hyperparameters do not reliably reflect features that can be interpreted in clinical or biological terms. A potential cause for the observed uncertainty in optimized hyperparameters may be the frequently occurring peaks in blood glucose levels. These peaks make the distributions of CGM data right-skewed, which contradicts the Gaussianity assumptions of GP models. To handle this deviation from Gaussianity in future studies, one could try square root- or log transforming the data prior to fitting any GP models.

A limitation of the investigated locally periodic models is their assumption of stationarity. Although long term changes in glucose control were handled via detrending, there are other sources of non-stationarity in CGM data, including irregular meal times and differences in glycemic variability between night and day. This motivates development of non-stationary GP models for CGM data. To that end, methods proposed by Cheng et al. [22] may provide a good starting point.

Finally, GP models with Spectral mixture kernels were investigated in terms of how well they could predict future blood glucose values, and how their power spectral density could reveal hidden periodicities in CGM data. Results indicate that these models can not accurately predict the characteristic peaks in blood glucose values, given the present implementation. However, periodograms obtained by Lomb-Scargle's method as such provided insights into numerous periodic tendencies of glycemic variability, and a closer study into the spectral properties of CGM traces might prove useful for characterising disease symptomatology or treatment effects. A starting point could be to investigate the cause for lack of an approximately 24 hour peak in some patients' periodograms.

Bibliography

[1] Grazia Aleppo. Approaches for Successful Outcomes with Continuous Glucose Monitoring. *Role of Continuous Glucose Monitoring in Diabetes Treatment. Arlington, VA: American Diabetes Association*, pages 13–18, 2018. URL https://www.ncbi.nlm.nih.gov/books/NBK538974/.

[2] Statistical Analysis Plan for Protocol D5670C00011. A Phase 2, Randomised, Double-Blind, Placebo-Controlled Study to Evaluate the Efficacy, Safety, Tolerability, and Pharmacokinetics of Different Doses of MEDI0382 in Overweight and Obese Subjects with Type 2 Diabetes Mellitus. https://clinicaltrials.gov/ProvidedDocs/00/NCT03244800/SAP_001.pdf, 2017. [Online; accessed 16-May-2021].

[3] Philip Ambery, Parker Victoria, Stumvoll Michael, Posch Maximilian, Heise Tim, Plum-Moerschel Leona, Tsai Lan-Feng, Robertson Darren, Jain Meena, Petrone Marcella, Rondinone Cristina, Hirshberg Boaz, and Jermutus Lutz. Medi0382, a GLP-1 and glucagon receptor dual agonist, in obese or overweight patients with type 2 diabetes: a randomised, controlled, double-blind, ascending dose and phase 2a study. *The Lancet*, 391:2607–18, 2018.

[4] Rajaa Nahra, Wang Tao, Oscarsson Jan, Repetto Enrico, Gadde Kishore, Stumvoll Michael, Jermutus Lutz, Hirshberg Boaz, and Ambery Philip. Effects of Cotadutide (MEDI0382) on Biomarkers of Nonalcoholic Steatohepatitis in Overweight or Obese Subjects with Type 2 Diabetes Mellitus: A 26-Week Analysis of a Randomized Phase 2b Study, 2019.

[5] Guillermo E Umpierrez and Boris P Kovatchev. Glycemic Variability: How to Measure and Its Clinical Implication for Type 2 Diabetes. *The American journal of the medical sciences*, 356(6):518–527, 2018. URL https://doi.org/10.1016/j.amjms.2018.09.010.

[6] David Rodbard. Glucose Variability: A Review of Clinical Applications and Research Developments. *Diabetes technology & therapeutics*, 20(S2):S2–5, 2018. URL https://www.liebertpub.com/doi/10.1089/dia.2018.0092.

[7] Thomas A Peyser, Andrew K Balo, Bruce A Buckingham, Irl B Hirsch, and Arturo Garcia. Glycemic Variability Percentage: A Novel Method for Assessing Glycemic Variability from Continuous Glucose Monitor Data. *Diabetes technology & therapeutics*, 20(1):6–16, 2018. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5846572/.

[8] C. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006. URL http://www.gaussianprocess.org/gpml/chapters/RW.pdf.

[9] David Duvenaud. *Automatic model construction with Gaussian processes*. PhD thesis, University of Cambridge, 2014. URL https://www.cs.toronto.edu/~duvenaud/thesis.pdf.

[10] Andrew Gordon Wilson. *Covariance kernels for fast automatic pattern discovery and extrapolation with Gaussian processes*. PhD thesis, University of Cambridge Cambridge, UK, 2014. URL https://www.cs.cmu.edu/~andrewgw/andrewgwthesis.pdf.

[11] Lukas Ortmann, Dawei Shi, Eyal Dassau, Francis J Doyle, Berno JE Misgeld, and Steffen Leonhardt. Automated Insulin Delivery for Type 1 Diabetes Mellitus Patients using Gaussian Process-based Model Predictive Control. In *2019 American Control Conference (ACC)*, pages 4118–4123. IEEE, 2019.

[12] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian Data Analysis, Third Edition*. CRC press, 2013.

[13] David Duvenaud, James Lloyd, Roger Grosse, Joshua Tenenbaum, and Ghahramani Zoubin. Structure Discovery in Nonparametric Regression through Compositional Kernel Search. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1166–1174. PMLR, 2013. URL http://proceedings.mlr.press/v28/duvenaud13.html.

[14] Fergus Simpson, Vidhi Lalchand, and Carl Rasmussen. Marginalised Spectral Mixture Kernels with Nested Sampling. 2020. URL https://arxiv.org/pdf/2010.16344.pdf.

[15] Andrew Wilson and Hannes Nickisch. Kernel Interpolation for Scalable Structured Gaussian Processes (KISS-GP). In *International Conference on Machine Learning*, pages 1775–1784. PMLR, 2015. URL http://proceedings.mlr.press/v37/wilson15.pdf.

[16] Nick E Phillips, Cerys Manning, Nancy Papalopulu, and Magnus Rattray. Identifying stochastic oscillations in single-cell live imaging time series using Gaussian processes. *PLoS computational biology*, 13(5):1–30, 2017. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005479.

[17] Subhasish Basak, Sébastien Petit, Julien Bect, and Emmanuel Vazquez. Numerical issues in maximum likelihood parameter estimation for Gaussian process regression. *arXiv preprint arXiv:2101.09747*, 2021. URL https://arxiv.org/pdf/2101.09747.pdf.

[18] Tanya L Leise. Analysis of Nonstationary Time Series for Biological Rhythms Research. *Journal of Biological Rhythms*, 32(3):187–194, 2017. URL https://doi.org/10.1177/0748730417709105.

[19] Jacob R Gardner, Geoff Pleiss, David Bindel, Kilian Q Weinberger, and Andrew Gordon Wilson. GPyTorch: Blackbox Matrix-Matrix Gaussian Process Inference with GPU Acceleration. In *Advances in Neural Information Processing Systems*, 2018. URL https://arxiv.org/abs/1809.11165.

[20] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[21] Long Vu, Sarah Kefayati, Tsuyoshi Idé, Venkata Pavuluri, Gretchen Jackson, Lisa Latts, Yuxiang Zhong, Pratik Agrawal, and Yuan-Chi Chang. Predicting Nocturnal Hypoglycemia from Continuous Glucose Monitoring Data with Extended Prediction Horizon. In *AMIA Annual Symposium Proceedings*, volume 2019, pages 874–882. American Medical Informatics Association, 2019. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7153099/.

[22] Li-Fang Cheng, Bianca Dumitrascu, Michael Zhang, Corey Chivers, Michael Draugelis, Kai Li, and Barbara Engelhardt. Patient-Specific Effects of Medication Using Latent Force Models with Gaussian Processes. In *International Conference on Artificial Intelligence and Statistics*, pages 4045–4055. PMLR, 2020. URL http://proceedings.mlr.press/v108/cheng20c.html.

[23] Alessio Benavoli, Dario Azzimonti, and Dario Piga. Skew Gaussian processes for classification. *Machine Learning*, 109(9):1877–1902, 2020.

[24] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Science & Business Media, 2009.

A

Appendix 1

Gaussian Mixture Models (GMMs) were used to initialize the Spectral Mixture Kernel hyperparameters. In the one dimensional case, fitting a GMM amounts to finding Q components, weights w_q, means μ_q and variances σ_q^2 such that the function

$$f(x) = \sum_{q=1}^{Q} w_q \phi_q(x),$$

where $\phi_q(x) = \frac{1}{\sqrt{2\pi\sigma_q^2}} \exp\left(\frac{(x-\mu_q)^2}{2\sigma_q^2}\right)$ is the probabilty density function of the Normal distribution, matches the data under investigation closely. It is usually accomplished by use of the EM-algorithm [24].

DEPARTMENT OF MATHEMATICAL SCIENCES
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden
www.chalmers.se

CHALMERS
UNIVERSITY OF TECHNOLOGY