Statistical Modelling with Additive Gaussian Process Priors

Sahoko Ishida

Department of Statistics

London School of Economics and Political Science

This dissertation is submitted for the degree of

Doctor of Philosophy

February 2024

To my family and friends

Declaration

I certify that the thesis I have presented for examination for the MPhil/PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it). The copyright of this thesis rests with the author. Quotation from it is permitted, provided that full acknowledgement is made. This thesis may not be reproduced without my prior written consent. I warrant that this authorisation does not, to the best of my belief, infringe the rights of any third party. I declare that my thesis consists of 25,355 words.

Sahoko Ishida

February 2024

Acknowledgements

Writing this thesis would not have been possible without the help, support and encouragement of many people. First and foremost, I would like to thank my supervisor, Wicher Bergsma. His depth of knowledge and enthusiasm for research inspired me to pursue PhD in the first place. I am profoundly thankful for his remarkable patience and generosity with time. Throughout these years, even when it has been stressful, I have always left our meetings feeling encouraged and motivated. I could not have asked for a more considerate and inspiring supervisor, whom I will always look up to not only as a researcher but also as a person.

I would also like to convey my gratitude to the London School of Economics and Political Science for their financial support and to the Department of Statistics, particularly Penny Montague, Sarah McManus, and Imelda Noble Andolfo, for their unwavering assistance and support throughout both my master's and PhD studies.

I am grateful for the teaching opportunities I have had at LSE. I have gained invaluable insights from Jouni Kuha, Yunxiao Chen, Kostas Kalogeropoulos, Christine Yuen, Anastasia Kakou, James Abdey, and Sara Geneletti. I extend my heartfelt gratitude to Irini Moustaki and the members of the Psychometrics lab for kindly inviting me to their social gatherings many times. I am also grateful for the chance I have had to work with Francesca Panero.

In completing this thesis, I wish to express my sincere appreciation to the examiners, Kostas Kalogeropoulos and Emanuele Giorgi, for their invaluable feedback, insightful comments, and engaging discussions throughout the evaluation process.

My first year would have been considerably more challenging without the camaraderie of my office mates on the seventh floor: Anica Kostic, JingHan Tee, José Manuel Pedraza-Ramírez, Gianluca Giudice, Davide de Santis and KaiFang Zhou. I consider myself extremely fortunate to have commenced my studies in such a warm, welcoming, and enjoyable office environment. My appreciation extends to Camilo Cárdenas-Hurtado, Shakeel Gavioli-Akilagun, Eduardo Ferioli-Gomes, Alexandros Pavlis, Motonori Oka, Xinyi Liu and Yudong Chen for always being a fun company. To Anica Kostic, my office mate, flatmate, friend, and occasional tutor– there is no one else with whom I would have chosen to spend the lockdown period in our tiny shoe-box house. I am incredibly fortunate to have you as a friend.

Having 5,955 miles to my home (yes, I just checked) was not always easy. To my family and friends back home – thank you for always making sure that I feel safe, calm, and happy whenever I return home. My special thanks go to my parents, who always encouraged me to pursue what I found most interesting and meaningful, even when my choices seemed unconventional to them at times.

Finally to Gabriel Wallin - thank you for always being a good role model and supporting me by listening to my presentation practices, reviewing my drafts, and dragging me out of bed in the morning with the best cup of coffee, especially during the last stretch of this journey. I am truly thankful.

Abstract

Regression with Gaussian process (GP) priors has become increasingly popular due to its ability to model complex relationships between variables and handle auto-correlation in the data through the covariance function of the process, called *kernel*. Despite its popularity, the statistical modelling aspect of GP regression has received relatively limited attention.

In this thesis, we explore a regression model where the regression function can be decomposed into a sum of lower-dimensional functions, akin to the principles of Generalised Additive Models (Hastie and Tibshirani, 1990). We propose additive interaction modelling using a class of *hierarchical ANOVA decomposition kernel*. This flexible statistical modelling framework naturally accommodates interaction effects of any order without increasing the number of model parameters. Our approach facilitates straightforward assessment and comparison of models with different interaction structures through the model marginal likelihood. We also demonstrate how this framework enhances the interpretability of complex data structures, especially when combined with the concept of kernel centring.

The second segment of the thesis focuses on the computational aspects of implementing the proposed additive models for handling large-scale data structured in multidimensional grids. Such structured data often arise in scenarios involving multi-level repeated measurements, as commonly seen in spatio-temporal analysis or medical, behavioural, and psychological studies. Leveraging the Kronecker product structure

within the covariance matrix, we reduce the time complexity to $O(n^3)$ and storage requirements to $O(n^2)$. We extend existing work in the GP literature to encompass all models under hierarchical ANOVA decomposition kernels. Additionally, we address issues related to incomplete grids and various missingness mechanisms.

We illustrate the practical application of our proposed methodologies using both simulated and real-world spatio-temporal and longitudinal data.

Table of contents

List of figures

List of tables

Chapter 1

Introduction

With data at hand, a statistical model aims to provide a simple summary of the data and describe the relationship between a variable y, called the response or dependent variable, and a set of variables x, called predictors or independent variables. This relationship can be decomposed into a systematic and random component. For a real-valued response y, a typical statistical model thus takes the following form:

$$y = \underbrace{f(x)}_{\text{Systematic variation}} + \underbrace{\epsilon}_{\text{Random variation}}$$

where x belongs to some set $\mathcal{X}$, e.g. $\mathbb{R}^p$ if x is p-dimensional. The systematic effect of x on y is captured through regression function f, which is estimated based on the sample of observed y and x, denoted by $\mathcal{D} = (y_i, x_i)_{i=1}^n$. The effect may be assumed linear, in which case a standard linear regression can be used. For modelling non-linear relationships, popular approaches include polynomial regression as an extension to the standard linear regression, smoothing spline (Wahba, 1990) and Gaussian process (GP) regression (Rasmussen and Williams, 2006). GP regression has gained popularity due to its ability to account for various relationships between y and x, including but not limited to linear, by assuming different covariance structures of $k(x, x') = \text{cov}(f(x), f(x'))$ for

$x, x' \in \mathcal{X}$. Through this function k, GPs can inherently incorporate auto-correlation in the data, rendering it particularly useful for spatial, temporal and spatio-temporal data where accounting for auto-correlation is crucial. Moreover, employing a GP as a prior on the regression function f provides the possibility of quantifying the uncertainty in the inferred systematic effect.

1.1 Additive Gaussian process models

Constructing a good statistical model requires a thorough investigation of data and is often a challenging task. This thesis concerns statistical modelling with Gaussian process (GP) priors. More specifically, we focus on additive GP models, which assume that the regression function f is decomposed into e.g., $f(x) = a + f_1(x_1) + f_2(x_2) + \ldots + f_d(x_d)$, where a is a constant term. This formulation is useful when the predictor x is divided into several sets $x = (x_1, x_2, \ldots, x_d)$, and each set of variables has a different relationship to y. The effect of each predictor may also interact with each other, in which case, additional terms modelling two-way interaction effects, e.g.,$f_{12}(x_1, x_2)$, $f_{23}(x_2, x_3)$ can be added to f. Higher-ordered interaction effects may also be considered. For instance, with $d = 3$, the regression function with the highest-ordered interaction effect consists of $2^3 = 8$ terms and is given by

$$\begin{aligned} f(x) = a + &\underbrace{f_1(x_1) + f_2(x_2) + f_3(x_3)}_{\text{main effect}} \\ &+ \underbrace{f_{12}(x_1, x_2) + f_{23}(x_2, x_3) + f_{31}(x_3, x_1)}_{\text{two-way interaction effect}} + \underbrace{f_{123}(x_1, x_2, x_3)}_{\text{three-way interaction effect}} . \end{aligned} \tag{1.1}$$

In this thesis, we explore additive GP models with diverse interaction structures. Our contributions to the existing literature encompass the following key aspects:

- Providing a comprehensive framework for additive interaction modelling with GP priors, employing the *ANOVA decomposition kernel*,
- Enhancing the interpretability of the proposed additive GP models for higher-order interaction models, which is made possible due to the centring of kernels, and
- Providing computationally efficient implementations of additive GP model accommodating various interaction structures, particularly tailored for analysing large-scale datasets often encountered in spatio-temporal and longitudinal data.

We summarise the methodological challenges in the current additive GP literature and our main contributions as follows.

Statistical modelling with ANOVA decomposition kernel

With GP models, statistical modelling typically involves specifying the structure of its covariance k. While there are various ways to specify interaction models, such as (1.1), we adopt *ANOVA decomposition kernel* introduced by Stitson et al. (1999) in the context of Support Vector Machines. The concept of ANOVA decomposition kernel can be traced back to functional ANOVA (f-ANOVA) decomposition (Huang, 1998; Stone, 1994), which aims to decompose a d-variate function $f(x)$ where $x = (x_1, \ldots, x_d)$ into the form:

$$f(x) = f_0 + \sum_{l=1}^{d} f(x_l) + \sum_{1 \le l < l' \le d} f_{ll'}(x_l, x_{l'}) + \ldots.$$

In the literature on splines and reproducing kernel Hilbert space, Smoothing-Spline ANOVA (SS-ANOVA) proposed by Wahba (1990) and Gu and Wahba (1993) gained popularity. Similarly to classical analysis of variance, f-ANOVA models have been used to identify significant differences in the functional mean of different groups. The usual approach, e.g. in Wahba (1990), has more model parameters for larger (more interaction terms) models; hence, applying this directly to GP models poses considerable challenges in estimation and model comparison.

This thesis uses an ANOVA decomposition kernel construction proposed by Bergsma and Jamil (2023) in the I-prior methodology (Bergsma, 2020), which has a close connection to Gaussian process regression. This formulation is more parsimonious, requiring only $d+1$ scale parameters to be estimated, regardless of the number of interaction terms involved in the model. This facilitates a straightforward model selection procedure. Moreover, we select interaction terms in a hierarchical manner using *hierarchical ANOVA decomposition kernel.* As with any statistical modelling problem, interaction terms are included along with main and lower-order interaction terms. This differs from the approach in Stitson et al. (1999), where the l-th order model includes all l-th interactions but no higher or lower ones.

Interpretability

Interpretation of a regression model is an essential part of analysing real-world data. Plate (1999) is one of the first to discuss the importance of modelling interactions and a trade-off between interpretability and accuracy in flexible modelling using GP, which can be done through kernel construction. Duvenaud et al. (2011) is a more recent attempt at kernel-based modelling of main and interaction effects. However, the proposed methodology does not guarantee certain common practices in statistical modeling, such as a hierarchical structure in interaction terms. In the presence of higher-order interactions, further discussion is warranted regarding the interpretation of main and lower-order interaction effects within the proposed model. In our approach, we utilise *centring* of kernels when constructing the hierarchical ANOVA decomposition kernel. This is an analogue to the centring of the predictors in standard linear regression, which provides a meaningful interpretation of the constant term as the expected value of the response when all predictors are set to zero. Within the proposed additive GP models, after centring the kernels, all terms in the model, including the main and lower-order interaction effects, possess an intuitive interpretation.

Computation

Implementing such flexible GP models poses computational challenges due to time complexity $O(n^3)$ and storage requirement $O(n^2)$. Consequently, standard GP regression becomes impractical for many real-world datasets of substantial scale. Various strategies have been developed to alleviate the computational burden associated with GP models (for a comprehensive overview, see Liu et al. (2020)). One particularly effective approach is to make use of a *Kronecker* product structure in a covariance matrix, which is applicable when the data has a multidimensional grid structure. This concept is pertinent to scenarios such as multi-level panel data commonly encountered in spatio-temporal and longitudinal datasets, where measurements are repeatedly collected at each sample location or individual. As demonstrated in prior work by Flaxman et al. (2015); Gilboa et al. (2013); Saatçi (2012); Wilson et al. (2014), the Kronecker method can reduce the computational complexity and storage requirements to a best-case scenario of $O(n)$ without resorting to approximations. However, these methods were limited to specific submodels, such as saturated models or those with only the highest-ordered interaction terms. This implies that other additive models, such as main effect or non-saturated interaction effect models, still face constraints with $O(n^3)$ computations and $O(n^2)$ storage or necessitate adopting alternative scalable methods relying on approximations. While these alternative methods avoid imposing constraints on the kernel structure, they typically do not achieve the same level of scalability and accuracy as the Kronecker method.

This thesis introduces an extension of the Kronecker method, specifically designed to handle any models constructed using the hierarchical ANOVA kernel, wherein the centring of kernels plays a crucial role. This novel Kronecker approach enables the modelling of large-scale data with many structures, including, but not limited to, saturated models. Another common challenge associated with the Kronecker method is its reliance on complete datasets without missing values. To address this limitation, an approximation method for handling incomplete data has been proposed, as outlined in the works of Gilboa et al. (2013) and Wilson et al. (2014). We evaluate the effectiveness of this approach under various missingness mechanisms, providing insights into its performance in different scenarios.

1.2 The outline of the thesis

The thesis is organised as follows.

- Chapter 2 offers an overview of Gaussian process regression. The chapter provides a summary of the estimation and inference of GP models for both Gaussian and non-Gaussian likelihoods. We explore the relationships between GP models and classical methods such as kernel ridge regression and Kriging. Additionally, we introduce the concept of kernel centring with illustrative examples.

- Chapter 3 revolves around additive GP models, emphasising interaction modelling utilizing the hierarchical ANOVA decomposition kernel. We study the properties of the posterior mean function of each term consisting of an additive model and discuss how it facilitates the interpretation of the result.

- In Chapter 4, we present our approach for the efficient implementation of the proposed model, particularly tailored for multi-dimensional grid-structured data. This method capitalizes on the Kronecker product structure within the covariance matrix of the model. The chapter begins with an introduction to this special data structure, the Kronecker product, and its inherent properties. We then delve into how the model's covariance matrix can be effectively decomposed into a sum of Kronecker products, leading to efficient evaluations of the likelihood function and the posterior of additive Gaussian process models.

- Chapter 5 addresses the challenge of handling missing grids within the Kronecker product approach. We highlight the limitations of the proposed approach in the literature and discuss the direction for future research.

Furthermore, for readers seeking additional technical details pertaining to Gaussian processes, kernels, and the Kronecker product, we have provided Appendices A and B. Throughout the thesis, we underscore the practical application of the proposed model through analyses of real-world spatio-temporal and longitudinal data, as well as through simulation studies. Additional results and illustrations related to data analysis can be found in Appendix C.

Chapter 2

Regression with Gaussian process prior

This chapter provides the basics of regression models with a Gaussian process (GP) prior. Consider a regression model for a real-valued response y and a set of predictors $\mathbf{x}$ which belongs to a set $\mathcal{X}$. The set $\mathcal{X}$ can be, for example, $\mathbb{R}^p$ for p-dimensional real-valued predictors. For $i = 1, \ldots, n$, the model is expressed as

$$y_i = f(\mathbf{x}_i) + \epsilon_i \tag{2.1}$$

where the error terms $(\epsilon_1, \ldots, \epsilon_n) \sim \mathbf{MVN}(\mathbf{0}, \Sigma)$. For i.i.d errors, we can write $\Sigma = \sigma^2 \mathbf{I}_n$ where $\mathbf{I}_n$ is $n \times n$ identity matrix. The main idea here is to put a prior on the function f directly. Specifically, we shall assume f to follow a GP. By assuming a different covariance function, called a kernel, GP regression can model various relationships between a response variable and predictors.

This chapter is structured as follows. In Section 2.1, we provide an introduction to GPs and kernels. This section offers a list of common kernels frequently employed in the fields of machine learning and spatio-temporal analysis. Additionally, we introduce the concept of kernel centring, which holds significant importance in the subsequent chapters. Section 2.2 is dedicated to the estimation and inference aspects of regression with Gaussian process priors. We focus on scenarios where the response variable is real-valued, and we assume a Gaussian likelihood. Furthermore, in Section 2.3, we delve into the extension of GP regression to accommodate various types of response variables, including categorical and count data. We emphasize that while GP regression excels in modelling nonlinear relationships and is often regarded as a non-parametric approach, it readily accommodates the inclusion of parametric relationships. This adaptability is exemplified in the introduction of semi-parametric GP models in Section 2.4. Lastly, in Section 2.5, we briefly explore other statistical methods that bear a connection to GP regression.

2.1 Gaussian processes and kernels

The key component of a GP is its covariance function, known as the *kernel.* We provide definitions of kernels and Gaussian processes.

Definition 1 (Kernel)**.** *Let $\mathcal{X}$ be a nonempty set. A kernel is defined as a symmetric positive definite function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ satisfying, for all $n = 1, 2, ...,$ $a_1, \dots a_n \in \mathbb{R}$ and $\mathbf{x}_1, \dots \mathbf{x}_n \in \mathcal{X}$, $\sum_{i=1}^n \sum_{j=1}^n a_i a_j k(\mathbf{x}_i, \mathbf{x}_j) \geq 0$.*

Definition 2 (Gaussian Process)**.** *Let $\mathcal{X}$ be a non-empty set, $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a (positive definite) kernel, and $m : \mathcal{X} \to \mathbb{R}$ be a real-valued function. A random function $f : \mathcal{X} \to \mathbb{R}$ is a Gaussian Process with mean function m and kernel k, if, for any finite set $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathcal{X}$, the random vector $\mathbf{f} = (f(\mathbf{x}_1), \dots, f(\mathbf{x}_n))^\top$ follows a multivariate normal distribution $\mathbf{MVN}(\mathbf{m}, \mathbf{K})$ where $\mathbf{m} = (m(\mathbf{x}_1), \dots, m(\mathbf{x}_n))^\top$ and $\mathbf{K}$ is a $n \times n$ matrix with (i, j)-th element given by $k(\mathbf{x}_i, \mathbf{x}_j)$.*

Throughout the thesis, we denote a random function f following a GP with a mean function m and a kernel k by $f \sim \mathrm{GP}(m, k)$. The matrix $\mathbf{K}$ is referred to as the Gram matrix. The GP is fully specified by its mean function and kernel, allowing us to incorporate our beliefs about the regression function. In the absence of prior beliefs, a common choice is to set the mean function to zero, i.e., $m(x) = 0$ for all $x \in \mathcal{X}$. Under this assumption, the kernel k defines a unique GP, and selecting a specific kernel corresponds to making different assumptions about the underlying process. For instance, the squared exponential (SE) kernel is widely used in machine learning and associated with a very smooth process that possesses mean-square derivatives of all orders (Adler, 1981, Chapter 2). The next section introduces some of the popular kernels, including the SE kernel. For more comprehensive reviews on different classes of kernels, see Rasmussen and Williams (2006, Chapter 4) and Genton (2001) for example.

2.1.1 Common kernels

The squared exponential kernel (2.2) is arguably the most frequently used kernel in the machine learning literature.

Example 1 (Squared exponential kernel)**.** *Let $\mathcal{X} \subset \mathbb{R}^p$. For $\alpha > 0$ and $\rho > 0$, a squared exponential (SE) kernel $k_{se} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is defined by*

$$k_{se}(\mathbf{x}, \mathbf{x}') = \alpha^2 \exp\left(-\frac{||\mathbf{x} - \mathbf{x}'||^2}{2\rho^2}\right), \quad \mathbf{x}, \mathbf{x}' \in \mathcal{X}. \tag{2.2}$$

We refer to unknown parameters in a kernel as hyperparameters. In the example above, we have a scale parameter α, which is common to all kernels, and a length-scale parameter ρ. The values of these hyperparameters also carry assumptions on the underlying process. See Figure 2.1 for a further demonstration of this.

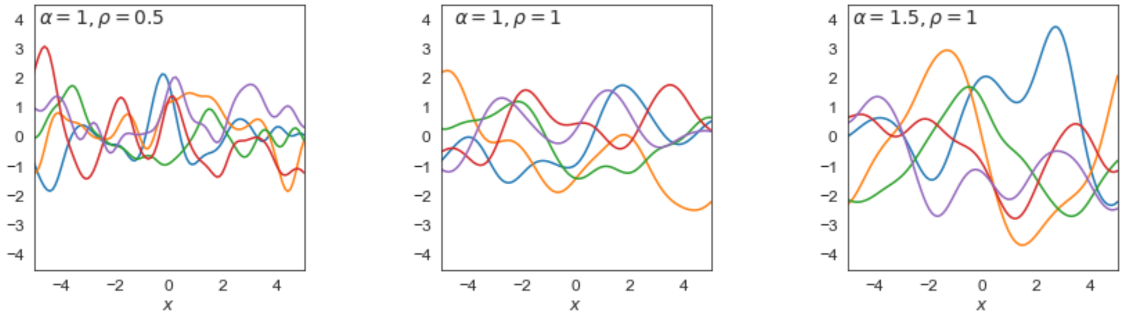


Fig. 2.1 Sample paths from a zero-mean Gaussian process with a one-dimensional squared exponential kernel. The y axis is $y = f(x)$. Taking the middle panel as a reference, the left panel has a smaller length scale which makes the process more wiggly. Having a larger sample scale parameter (right panel) makes the average distance from the mean (0) bigger.

The importance of SE kernel is also attributed to its relation to other popular kernels, such as the Matérn class kernel (Matérn, 1960; Stein, 1999) and the squared exponential periodic kernel, introduced below.

Example 2 (Matérn kernel kernel)**.** *Let $\mathcal{X} \subset \mathbb{R}^p$. For $\alpha > 0$, $\rho > 0$ and $\nu > 0$, the the Matérn kernel $k_{mat} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is defined by*

$$k_{mat}(\mathbf{x}, \mathbf{x}') = \alpha^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}||\mathbf{x} - \mathbf{x}'||}{\rho} \right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}||\mathbf{x} - \mathbf{x}'||}{\rho} \right), \quad \mathbf{x}, \mathbf{x}' \in \mathcal{X}, \tag{2.3}$$

where Γ is the gamma function and K_ν is the modified Bessel function of a second kind.

Example 3 (Periodic squared exponential kernel)**.** *Let $\mathcal{X} \subset \mathbb{R}^p$. For $\alpha > 0$, $\rho > 0$ and $p > 0$, the periodic squared exponential kernel $k_{pr} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is defined by*

$$k_{pr}(\mathbf{x}, \mathbf{x}') = \alpha^2 \exp\left(-\frac{2\sin^2(\frac{\pi|\mathbf{x}-\mathbf{x}'|}{p})}{\rho^2} \right), \quad \mathbf{x}, \mathbf{x}' \in \mathcal{X}.$$

With Matérn class kernels, the smoothness of the process can be controlled by parameter ν. With $\nu = 1.5$, the process is “rough” compared with $\nu = 2.5$ (see figure 2.2). The smoothness can be discussed in terms of mean square differentiability. A sample path with the Matérn class kernel with ν is $k = \lceil \nu \rceil$ times mean square differentiable. It is worth noting that for $\nu \to \infty$, it equals the SE kernel. The periodic SE kernel is useful when handling, e.g., the continuous-time process that has a regular cycle, e.g. daily, weekly or annual. The period parameter p can be treated as known or unknown, and the GP path from the periodic kernel is a periodic function of period p. It can be derived from the SE kernel; we have $k_{se}(\mathbf{u}, \mathbf{u}') = k_{pr}(\mathbf{x}, \mathbf{x}')$ where $\mathbf{u} = (\sin(\frac{p}{2\pi}\mathbf{x}), \cos(\frac{p}{2\pi}\mathbf{x}))^\top$. In fact, any kernel k can be made periodic with this formulation.

A constant kernel $k_{const} : \mathcal{X} \times \mathcal{X}$, which is given by

$$k_{const}(\mathbf{x}, \mathbf{x}') = \alpha^2, \quad \mathbf{x}, \mathbf{x}' \in \mathcal{X} \tag{2.4}$$

is usually used in combination with other kernels. This yields random constant functions, in particular, functions of the form $f(x) = c$, where $c \sim N(0, \alpha^2)$.

The SE kernel, Matérn kernel, periodic SE kernel and constant kernel are all in the class of *stationary* kernels, more specifically *isotropic* kernels. A stationary kernel is a function of a lag vector $\tau = \mathbf{x} - \mathbf{x}'$ of two inputs. When the value of the function depends only on the norm of the two inputs $r = ||\tau||$, the kernel is said to be isotropic, and the corresponding process is invariant under a shift in time or space. While the assumption of isotropy or stationarity gives a nice interpretation of correlation structure, we need a class of non-stationary kernels in the case where this assumption does not hold. A few simple examples of non-stationary kernels include the linear kernel and polynomial kernel.

Example 4 (Polynomial kernel)**.** *Let $\mathcal{X} \subset \mathbb{R}^p$. For $\alpha > 0$ and positive integer d, the polynomial kernel of degree d $k_{pol} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is defined by*

$$k_{pol}(\mathbf{x}, \mathbf{x}') = \alpha^2 \left(\mathbf{x}^\top \mathbf{x}' + c\right)^d,$$

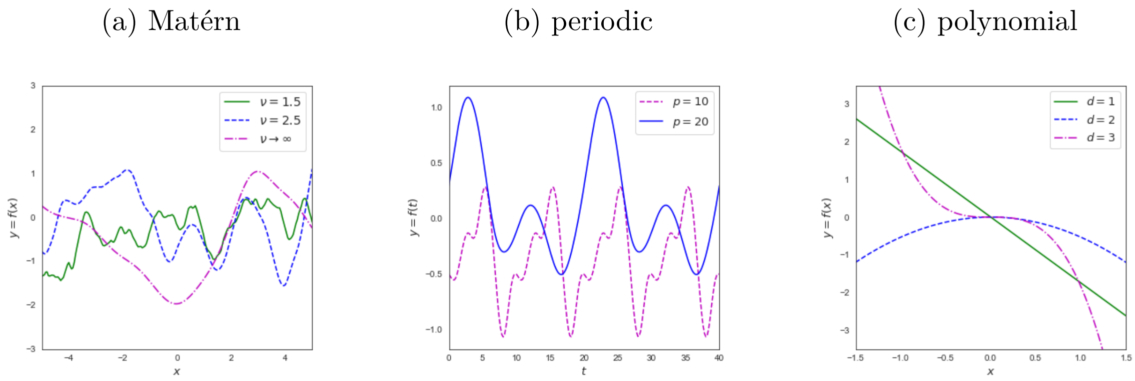

(a) Matérn (b) periodic (c) polynomial

Fig. 2.2 Sample paths from a zero-mean GP with different kernels. For all panels, the scale parameter α is set to be 1, and the length-scale parameter $\rho = 1$ for (a) and (b). For the polynomial kernel, $c = 0$. For the additional parameters, see the legend of each panel.

With $d = 1$, we get the linear kernel. Using the linear kernel or the polynomial kernel in GP regression corresponds with Bayesian linear or polynomial regression. Another useful non-stationary kernel is the fractional Brownian Motion kernel and kernels that are constructed from this kernel, such as its centred version. We will introduce this in Section 2.1.3.

Finally, we give an example of a kernel that can be used for categorical variables. Let $\mathcal{X}$ now be a finite set, e.g., $\mathcal{X} = \{1, 2, \ldots, J\}$ where J is a positive integer and each integer in the set represents a different category. We define the kernel for categorical variable $k_{cat} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ by

$$k_{cat}(\mathbf{x}, \mathbf{x}') = \alpha^2 \delta_{\mathbf{x}\mathbf{x}'}, \quad \mathbf{x}, \mathbf{x}' \in \mathcal{X}, \tag{2.5}$$

where δ is the Kronecker delta. This kernel assumes no inter-category correlation. Using this kernel to account for individual variability in multi-level models corresponds with the random effect model. We give an example in Section 3.4.2.

2.1.2 Kernel sums and products

Given two kernels $k_1 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $k_2 : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ constructed as their sum or product

$$k(\mathbf{x}, \mathbf{x}) = k_1(\mathbf{x}, \mathbf{x}') + k_2(\mathbf{x}, \mathbf{x}')$$
$$k(\mathbf{x}, \mathbf{x}) = k_1(\mathbf{x}, \mathbf{x}')k_2(\mathbf{x}, \mathbf{x}')$$

is also a kernel. That is, in this sense, the set of kernels is closed with respect to addition and multiplication.

The kernel k_1 or k_2 can be the constant kernel (2.4). Hence, adding a positive constant or multiplying by a positive constant gives a positive definite kernel.

The addition and multiplication operations above can be extended to kernels on different sets, as we illustrate next. Consider two kernels $k_1 : \mathcal{X}_1 \times \mathcal{X}_1 \to \mathbb{R}$ and $k_2 : \mathcal{X}_2 \times \mathcal{X}_2 \to \mathbb{R}$, where $\mathcal{X}_1$ and $\mathcal{X}_2$ are two different sets. Then $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ where $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$ given by

$$k((\mathbf{x}_1, \mathbf{x}_2), (\mathbf{x}_1', \mathbf{x}_2')) = 1 + k_1(\mathbf{x}_1, \mathbf{x}_1') + k_2(\mathbf{x}_2, \mathbf{x}_2') + k_1(\mathbf{x}_1, \mathbf{x}_1')k_2(\mathbf{x}_2, \mathbf{x}_2'), \quad \mathbf{x}_l, \mathbf{x}_l' \in \mathcal{X}_l$$

is a positive definite kernel.

2.1.3 Centring of kernels

GP paths may be arbitrarily positioned. For example, all paths of a GP with a polynomial kernel pass through the fixed point $(0, \alpha^2 c^d)$, which may be undesirable. In this case, centring of kernels can be applied. A positive definite kernel can be centred by the following.

Definition 3 (Centred kernel)**.** *Let P be a probability distribution over a non-empty set $\mathcal{X}$ and $X, X' \sim P$ are independent. Any kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ may be centred by*

$$k_{cent}(\mathbf{x}, \mathbf{x}') = \mathbb{E}_P \left[k(\mathbf{x}, \mathbf{x}') - k(\mathbf{x}, X) - k(\mathbf{x}', X') + k(X, X') \right], \quad \mathbf{x}, \mathbf{x}' \in \mathcal{X}. \tag{2.6}$$

A kernel centred by the above retains positive definiteness (see Appendix A.1.2), and GPs with such a kernel have centred paths in the sense that $\mathbb{E}_{X \sim P}[f(X)] = 0$. In practice, we take P to be the empirical distributions of $\mathbf{x}_1, \ldots \mathbf{x}_n$. An empirically centred kernel is given by:

$$k_{cent}(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') - \frac{1}{n} \sum_{i=1}^{n} k(\mathbf{x}, \mathbf{x}_i) - \frac{1}{n} \sum_{j=1}^{n} k(\mathbf{x}', \mathbf{x}_j) + \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} k(\mathbf{x}_i, \mathbf{x}_j). \tag{2.7}$$

This ensures the function f evaluated at $\mathbf{x}_1, \ldots \mathbf{x}_n$ sums to zero, i.e., $\sum_{i=1}^{n} f(\mathbf{x}_i) = 0$. The empirically centred Gram matrix for a given $\mathbf{K}$, can be computed using a centring matrix $\mathbf{C} = \mathbf{I}_n - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^\top$, by $\mathbf{K}^{(c)} = \mathbf{CKC}$. This means that all columns and rows of $\mathbf{K}^{(c)}$ sum to 0. Centring of a kernel plays a key role in the interpretation and efficient computation, as shown in Section 3.2.1 and 4.4.3.

Centring example with fractional Brownian motion kernel

We will illustrate the centring of kernels using the fractional Brownian Motion (fBM) kernel.

Example 5 (fractional Brownian motion kernel)**.** *Let $\mathcal{X} \subset \mathbb{R}^p$ be a set. For $\alpha > 0$ and $0 < \gamma < 1$, the fractional Brownian motion kernel $k_{fbm_\gamma} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is given by*

$$k_{fbm_\gamma}(\mathbf{x}, \mathbf{x}') = \frac{\alpha^2}{2} \left(||\mathbf{x}||^{2\gamma} + ||\mathbf{x}'||^{2\gamma} - ||\mathbf{x} - \mathbf{x}'||^{2\gamma} \right), \quad \mathbf{x}, \mathbf{x}' \in \mathcal{X} \tag{2.8}$$

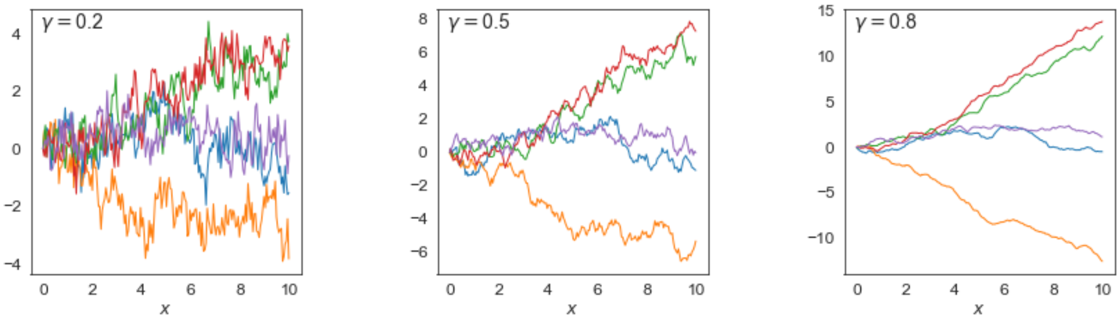


Fig. 2.3 Sample paths ($y = f(x)$) from the fBM kernel given in (2.8) with different values for the Hurst coefficient γ.

The Hurst coefficient γ in (2.8) determines the roughness of the process. A smaller value of γ is associated with rougher sample paths (see Figure 2.3). With $\gamma = 0.5$, we have the standard Brownian motion. From Figure 2.3, we notice $f(0) = 0$ for all paths, which may be undesirable for the problems we consider in the paper. To avoid this, a fBM kernel can be centred.

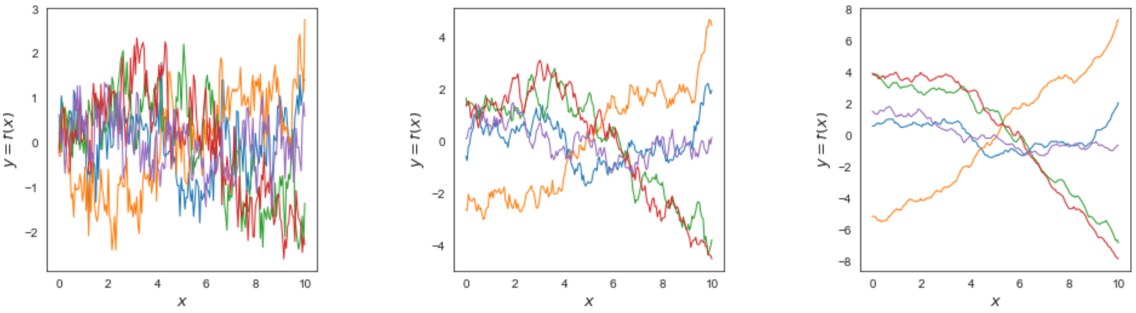
Example 6 (Centred fractional Brownian Motion kernel)**.** *Applying (2.7) to the fBM kernel (2.8), we have*

$$k^{(c)}_{fbm_\gamma}(\mathbf{x}, \mathbf{x}') = \frac{\alpha^2}{2n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} \left(||\mathbf{x} - \mathbf{x}_i||^{2\gamma} + ||\mathbf{x}' - \mathbf{x}_j||^{2\gamma} - ||\mathbf{x} - \mathbf{x}'||^{2\gamma} - ||\mathbf{x}_i - \mathbf{x}_j||^{2\gamma} \right) \tag{2.9}$$

For the rest of the thesis, we always centre fBM kernels with respect to the empirical distribution.

Another issue with the fBM kernel in (2.8) is its roughness. In contrast to a GP with SE kernel, fractional Brownian motion paths are differentiable nowhere; hence they may be too rough to be used in many examples. To remedy this, we introduce a squared kernel.

Definition 4 (Squared kernel)**.** *Let $\mathcal{X}$ be a non-empty set. Given a kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and data $\mathbf{x}_i \in \mathcal{X}$ for $i = 1, \dots, n$, a* squared kernel *is a function $k_{sq} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$*



(a) $\gamma = 0.2$ (b) $\gamma = 0.5$ (c) $\gamma = 0.8$

Fig. 2.4 Sample paths from centred fBM kernels (2.9) with different values for the Hurst coefficient γ.

given by

$$k_{sq}(\mathbf{x}, \mathbf{x}') = \sum_{i=1}^{n} k(\mathbf{x}, \mathbf{x}_i) k(\mathbf{x}', \mathbf{x}_i), \quad \mathbf{x}, \mathbf{x}' \in \mathcal{X}. \tag{2.10}$$

As this kernel consists of products and sums of positive definite kernels, it is a positive definite kernel. The corresponding Gram matrix is given by $\mathbf{K}_{sq} = \mathbf{K}\mathbf{K}$. If we use the empirically centred kernel, the resulting Gram matrix $\mathbf{K}_{sq}^{(c)} = \mathbf{K}^{(c)}\mathbf{K}^{(c)}$ is also empirically centred, as we have

$$\mathbf{K}_{sq}^{(c)}\mathbf{1} = \mathbf{K}^{(c)}\left(\mathbf{K}^{(c)}\mathbf{1}\right) = \mathbf{K}^{(c)}\mathbf{0} = \mathbf{0}.$$

This kernel has been shown to be useful in previous work by Bergsma (2020); Jamil (2018); Jamil and Bergsma (2020) in the context of I-priors, which has a close connection to GP regression and kernel methods. As shown in Figure 2.5, GP paths with squared (and centred) γ-fBM kernels are much smoother. In fact, the smoothness properties of GP paths with fBM kernel and squared fBM kernel can be discussed in terms of Hölder condition. While the realisations from the former are known to be a.s. Hölder of any order less than γ (see Embrechts and Maejima (2002) for example), the latter is Hölder of order 2γ. Bergsma (2020) discusses the smoothness properties of fBM

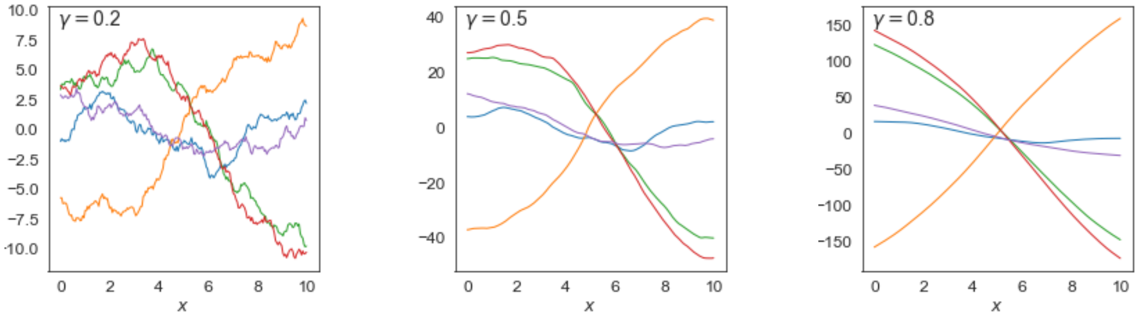


Fig. 2.5 Sample paths ($y = f(x)$) from centred and squared fBM kernels constructed using (2.9) and (2.10) with different values for the Hurst coefficient γ.

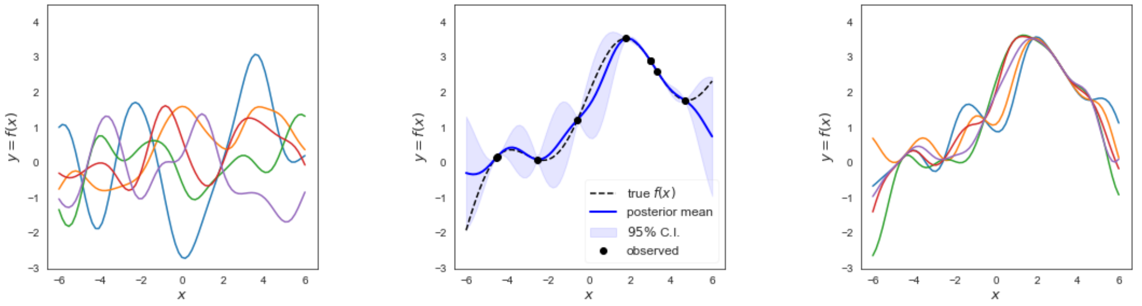
paths and squared fBM paths using different concepts of smoothness, including Hölder condition and regularity.

2.2 Estimation and inference for Gaussian likelihood

Let us revisit the regression model in (2.1) where we assume a zero-mean GP prior on f and i.i.d error. With $\mathbf{y} = (y_1, \ldots, y_n)^\top$ and $\mathbf{X}$ being a data matrix gathering covariates from all observations, the marginal distribution of $\mathbf{y}$ given $\mathbf{X}$, obtained by integrating out the prior, is:

$$\mathbf{y}|\mathbf{X} \sim \mathrm{MVN}_n(\mathbf{0}, \mathbf{K} + \sigma^2\mathbf{I}). \tag{2.11}$$

The following sections discuss the posterior distribution, the estimation of the hyperparameters and model selection.



(a) Sample paths from prior (b) Posterior mean (c) Sample paths from posterior

Fig. 2.6 Example of one-dimensional regression/curve fitting with a GP prior. Consider that we have a sample of size 7 (from a true function, $f(x) = 2 + \frac{2}{5}x - \frac{1}{20}x^2 + \sin x$). The prior is a GP with SE kernel with $\alpha = 1$ and $\rho = 1$. The panel (a) shows sample paths from the prior GP and (b) shows the posterior mean with 95% confidence interval together with the observations and the true function. Random sample paths drawn from the posterior are shown in (c).

2.2.1 Posterior

For a Gaussian likelihood, the GP is a conjugate prior, i.e., the posterior is also a GP. Specifically, we have $f|\mathbf{y}, \mathbf{X} \sim \mathrm{GP}(\bar{m}, \bar{k})$ with the mean function $\bar{m} : \mathcal{X} \to \mathbb{R}$ and the kernel $\bar{k} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ given by

$$\bar{m}(\mathbf{x}) = \mathbf{k}(\mathbf{x})^\top (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{y}, \qquad \mathbf{x} \in \mathcal{X} \tag{2.12}$$

$$\bar{k}(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') - \mathbf{k}(\mathbf{x})^\top (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{k}(\mathbf{x}'), \quad \mathbf{x}, \mathbf{x}' \in \mathcal{X} \tag{2.13}$$

where $\mathbf{k}(\mathbf{x}) = (k(\mathbf{x}_1, \mathbf{x}), \ldots, k(\mathbf{x}_n, \mathbf{x}))^\top$ for $\mathbf{x} \in \mathcal{X}$. See Appendix A.2 for the derivation and Figure 2.6 for an example of GP regression with a one-dimensional covariate $x \in \mathbb{R}$. The posterior given by (2.12) and (2.13) may more commonly be discussed in connection to prediction problems. In fact, given a new point $\mathbf{x}^*$ the posterior distribution of $f(\mathbf{x}^*)$ is Gaussian with mean $\bar{m}(\mathbf{x}^*)$ and variance $\bar{k}(\mathbf{x}^*, \mathbf{x}^*)$. In this thesis, we mainly use a zero-mean GP prior. Nonetheless, it is important to highlight that incorporating a non-zero mean function is a straightforward extension. For a GP prior with a non-zero mean function, denoted as $m : \mathcal{X} \to \mathbb{R}$, the posterior mean is expressed as:

$$\bar{m}(\mathbf{x}) = m(\mathbf{x}) + \mathbf{k}(\mathbf{x})^\top (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} (\mathbf{y} - \mathbf{m}), \quad \mathbf{x} \in \mathcal{X} \tag{2.14}$$

where $\mathbf{m} = (m(\mathbf{x}_1), \ldots, m(\mathbf{x}_n))^\top$. The kernel of the posterior remains unchanged from the expression given in (2.13).

2.2.2 Hyperparameter estimation

The fully Bayesian approach to estimating f is obtained by assigning priors (also known as hyperpriors) to the hyperparameters denoted by $\boldsymbol{\theta}$. This includes parameters in the kernels and the variance in the error term σ^2. In such a case, we most typically resort to Markov chain Monte Carlo (MCMC) to obtain samples from the posterior. The posterior mean, mode or median may be used as parameter estimates.

Alternatively, we can use the empirical Bayes approach (maximum marginal likelihood (MML) estimation), where the parameters are estimated by maximising the log marginal likelihood denoted by $p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = \int p(\mathbf{y}|\mathbf{f}) p(\mathbf{f}|\mathbf{X}, \boldsymbol{\theta}) d\mathbf{f}$. With Gaussian likelihood, this equals

$$\log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = -\frac{1}{2}\mathbf{y}^\top (\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{y} - \frac{1}{2} \log |\mathbf{K} + \sigma^2 \mathbf{I}_n| - \frac{n}{2} \log 2\pi. \tag{2.15}$$

Here the hyperparameters are involved in the Gram matrix $\mathbf{K}$ and the error term. The value of hyperparameters can also be estimated using cross-validation (CV).

Estimating hyperparameters with either CV or MML estimation may be costly. To reduce the computational complexity, it may be necessary to use a grid search. For example, the values $\nu = 1.5, 2.5$ and 3.5 for the parameter ν for the Matérn class kernel

(2.3), or $0.1, 0.2, \ldots, 0.9$ for the Hurst coefficient $\gamma \in (0, 1)$ of the fractional Brownian motion kernel (2.8) are common choices.

2.2.3 Model selection

In a real-world application, a model selection problem may involve choosing an appropriate type of kernel. It is worth noting that a more complex model does not necessarily result in a larger marginal likelihood (MacKay, 1995; Murray and Ghahramani, 2005), making it a suitable criterion for comparing models with different kernels. The marginal likelihood, as presented in the previous section (2.15), depends on the hyperparameters $\boldsymbol{\theta}$. In a fully Bayesian approach, $\boldsymbol{\theta}$ should be integrated out, giving:

$$p(\mathbf{y}|\mathbf{X}, \mathcal{M}) = \int p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \mathcal{M}) p(\boldsymbol{\theta}|\mathcal{M}) d\boldsymbol{\theta}, \tag{2.16}$$

where $p(\boldsymbol{\theta}|\mathcal{M})$ represents the prior for the hyperparameters of a given model $\mathcal{M}$, and $p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \mathcal{M})$ denotes its density function. The Bayes factor allows us to compare two different models, $\mathcal{M}$ and $\mathcal{M}'$:

$$\mathrm{BF}(\mathcal{M}, \mathcal{M}') = \frac{p(\mathbf{y}|\mathbf{X}, \mathcal{M})}{p(\mathbf{y}|\mathbf{X}, \mathcal{M}')}.$$

A Bayes factor greater than one indicates stronger support for $\mathcal{M}$ than for $\mathcal{M}'$. The strength of the evidence favouring one model over another depends on the value of the Bayes factor; for guidelines, refer to Kass and Raftery (1995). In practice, evaluating the integral in (2.16) can be done via, e.g. bridge sampling (Bennett, 1976; Meng and Schilling, 2002; Meng and Wong, 1996). This can be computationally expensive, and one may resort to the Laplace approximation or the estimated marginal likelihood, which involves plugging in the parameter estimates $\hat{\boldsymbol{\theta}}$ into (2.15). The Laplace approximation is effective when the posterior of $\boldsymbol{\theta}$ is approximately multivariate normal around the mode of the posterior.

A disadvantage of estimating the marginal likelihood is that then the Bayes factor is not a valid model selector in general, and there is a risk of overfitting. This problem can be avoided when comparing models with different interaction structures; see Section 3.2.2 for further details.

If models are compared in terms of out-of-sample predictive accuracy, one can attempt to quantify it with cross-validation using appropriate scoring rules, such as log predictive density. The drawback is the necessity to repeat the model fitting procedure for, e.g., k times for a k-fold cross-validation. Leave-one-out cross-validation (LOO-CV) is the special case where k equals the number of observations n. The Bayesian LOO-CV is given by $\frac{1}{n}\sum_{i=1}^{n}\log p(y_i|\mathbf{y}_{-i},\mathbf{X},\mathcal{M})$ where

$$p(y_i|\mathbf{y}_{-i},\mathbf{X},\mathcal{M}) = \int p(y_i|\mathbf{y}_{-i},\mathbf{X},\boldsymbol{\theta},\mathcal{M})p(\boldsymbol{\theta}|\mathbf{y}_{-i},\mathbf{X}_{-i},\mathcal{M})d\boldsymbol{\theta}. \tag{2.17}$$

Note that $\mathbf{y}_{-i}$ or $\mathbf{X}_{-i}$ denotes $\mathbf{y}$ or $\mathbf{X}$ after removing the i-th observation. With Gaussian likelihood, $p(y_i|\mathbf{y}_{-i},\boldsymbol{\theta},\mathbf{x}_i,\mathcal{M})$ can be computed analytically. See e.g. Rasmussen and Williams (2006, Section 5) and Vehtari et al. (2016). To bypass fitting the model n times, we can use importance sampling to sample from $p(\boldsymbol{\theta}|\mathbf{y}_{-i},\mathbf{X}_{-i},\mathcal{M})$ with the posterior with full data $p(\boldsymbol{\theta}|\mathbf{y},\mathbf{X},\mathcal{M})$ as the candidate distribution and approximate the LOO-CV as described in Vehtari et al. (2017); Vehtari and Lampinen (2002); Vehtari et al. (2016).

Alternatively, information criteria, such as DIC (Spiegelhalter et al., 2002) or WAIC (Watanabe and Opper, 2010) can be considered as approximately unbiased estimates of the expected log predictive density for new, i.e. out of sample, data. In a manner akin to AIC (Akaike, 1973), they consist of two terms: the deviance, i.e., the negative of the log predictive density for existing data, which is a biased estimate of the target quantity, and a bias correction term, which relates to the effective number of parameters. DIC is simpler to compute but has known disadvantages, such as the possibility of producing a negative effective number of parameters and its limitation to regular models. WAIC overcomes these issues and is fully Bayesian, using log pointwise posterior predictive density, $\sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\theta}|\mathbf{y},\mathbf{X},\mathcal{M}}[p(y_i|\mathbf{x}_i, \boldsymbol{\theta}, \mathcal{M})]$. This contrasts DIC, which uses the log joint density evaluated at the posterior mean $\log p(\mathbf{y}|\mathbf{X}, \hat{\boldsymbol{\theta}}_{\text{Bayes}})$. The bias correction term for DIC or WAIC requires approximating $\mathbb{E}_{\boldsymbol{\theta}|\mathbf{y},\mathbf{X},\mathcal{M}}[\log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \mathcal{M})]$ or $\mathbb{E}_{\boldsymbol{\theta}|\mathbf{y},\mathbf{X},\mathcal{M}}[\log p(y_i|\mathbf{x}_i, \boldsymbol{\theta}, \mathcal{M})]$ respectively.

WAIC is asymptotically equal to Bayesian LOO-CV (Watanabe and Opper, 2010), but for finite n or hierarchical model, there are noticeable differences as shown in Gelman et al. (2014). Vehtari et al. (2017) introduces Pareto-smoothed importance sampling (PSIS) LOO-CV and shows that this is more robust in the finite n with weak priors or influential observations.

2.3 Estimation and inference for Non-Gaussian likelihood

In section 2.2, we saw that the posterior is analytically tractable with a Gaussian likelihood. This is due to the Gaussian distribution being a conjugate prior for a Gaussian mean. For non-conjugate priors, a closed-form expression for the posterior is typically unavailable.

Given a sample $(y_i, \mathbf{x}_i)$ for $i = 1, \ldots, n$, where $y_i \in \mathcal{Y}$ follows non-Gaussian distribution, we consider a model

$$g(\mu_i) = f(\mathbf{x}_i) \tag{2.18}$$

Response (y_i)	$\mathcal{Y}$	Distribution	link
Binary	$\{0,1\}$	Bernoulli	logit
Binary	$\{0,1\}$	Bernoulli	probit
Categorical	$\{1,2,\ldots,C\}$	Categorical	softmax
Count	$\{0,1,2,\ldots\}$	Poisson	log

Table 2.1 Distributions and link functions for different types of response variables.

where $\mu_i = \mathbb{E}[y_i]$ is the mean of y_i, $g(\mu_i)$ is a link function, and the function f follows a GP, i.e., $f \sim \text{GP}(m,k)$. For example, if the response is counts, $y_i \in \{0,1,2,...\}$ and we assume a Poisson distribution with rate parameter λ_i, we have $\mathbb{E}[y_i] = \lambda_i$. The log-link function $g(\lambda_i) = \log(\lambda_i)$ is commonly used. See Table 2.1 for other types of response variables. This section primarily focuses on the univariate case, such as the Poisson model and Bernoulli-logit/probit model; however, the generalisation to multivariate problems, e.g., (multi-class) categorical with soft-max link function, is straightforward. See Appendix A.3 for the details.

The procedure of deriving the posterior predictive distribution $p(y^*|\mathbf{y}, \mathbf{X}, \mathbf{x}_{new})$ given a new input $\mathbf{x}^*$ is three-fold and requires the evaluation of the integrals:

1. the posterior distribution of $\mathbf{f} = (f(x_1), \ldots, f(x_n))^\top$

$$p(\mathbf{f}|\mathbf{y}, \mathbf{X}) = \frac{p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{X})}{\int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{X})d\mathbf{f}} \tag{2.19}$$

   where $p(\mathbf{f}|\mathbf{X})$ is the prior distribution, in this case, multivariate Gaussian, and $p(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^n p(y_i|f_i)$ is the likelihood function;

2. the conditional distribution for $f^* = f(\mathbf{x}^*)$

$$p(f^*|\mathbf{y}, \mathbf{X}, \mathbf{x}^*) = \int p(f^*|\mathbf{f}, \mathbf{X}, \mathbf{x}^*)p(\mathbf{f}|\mathbf{y}, \mathbf{X})d\mathbf{f} \tag{2.20}$$

where $p(f^*|\mathbf{f}, \mathbf{X}, \mathbf{x}^*)$ is the probability density function of the normal distribution with its mean and variance given by

$$\begin{aligned}\mathbb{E}[f^*|\mathbf{f}, \mathbf{X}, \mathbf{x}^*] &= k(\mathbf{x}^*)^\top \mathbf{K}^{-1}\mathbf{f} \\ \mathrm{Var}[f^*|\mathbf{f}, \mathbf{X}, \mathbf{x}^*] &= k(\mathbf{x}^*, \mathbf{x}^*) - k(\mathbf{x}^*)^\top \mathbf{K}^{-1} k(\mathbf{x}^*);\end{aligned}$$

3. the predictive distribution for the response y^*

$$p(y^*|\mathbf{y}, \mathbf{X}, \mathbf{x}^*) = \int p(y^*|f^*)p(f^*|\mathbf{y}, \mathbf{X}, \mathbf{x}^*)df^*. \tag{2.21}$$

Due to the non-Gaussian likelihood function, the integrals (2.19)-(2.21) typically lack an analytical solution, hence necessitating the use of approximations. The next section discusses different approximation methods.

2.3.1 Approximation methods

There is a rich literature on approximation methods for GP regression with non-Gaussian likelihood, many of which focus on approximating the posterior distribution of $\mathbf{f}$ given by (2.19). Despite the computational cost, utilising MCMC to sample from the posterior is considered the gold standard. Once we obtain the posterior sample from $p(\mathbf{f}|\mathbf{y}, \mathbf{X})$, evaluating (2.20) and (2.21) is straight-forward. The samples from the conditional distribution and the predictive distribution can be obtained with a simple sequential Monte Carlo algorithm.

An alternative to the numerical approximation is analytical approximation methods such as Laplace approximation (Rasmussen and Williams, 2006, Chapter 3), variational inference (Chai, 2012; Khan et al., 2012; Khan and Lin, 2017; Opper and Archambeau, 2009), or expectation propagation (Kim and Ghahramani, 2003, 2006; Minka, 2001), which also target approximating the posterior $p(\mathbf{f}|\mathbf{y}, \mathbf{X})$ by $q(\mathbf{f})$.

Laplace approximation (LA) aims to approximate the posterior with a multivariate normal $\text{MVN}(\boldsymbol{\mu}, \mathbf{V})$ where the mean vector is given by the mode of the log of the (un-normalised) posterior

$$\Psi(\mathbf{f}) := \log p(\mathbf{y}|\mathbf{f}) - \frac{1}{2}\mathbf{f}^\top \mathbf{K}^{-1}\mathbf{f} - \frac{1}{2}\log|\mathbf{K}| - \frac{n}{2}\log 2\pi \tag{2.22}$$

and the covariance matrix is given by the negative inverse Hessian of $\Psi(\mathbf{f})$ evaluated at the mode, i.e., $\boldsymbol{\mu} = \hat{\mathbf{f}} := \arg\max_{\mathbf{f}} \Psi(\mathbf{f})$ and $\mathbf{V} = (-\nabla^2\Psi(\hat{\mathbf{f}}))^{-1}$. Differentiating (2.22) with respect to $\mathbf{f}$, we have the gradient and the Hessian:

$$\nabla\Psi(\mathbf{f}) = \nabla \log p(\mathbf{y}|\mathbf{f}) - \mathbf{K}^{-1}\mathbf{f} \tag{2.23}$$

$$\nabla^2\Psi(\mathbf{f}) = \nabla^2 \log p(\mathbf{y}|\mathbf{f}) - \mathbf{K}^{-1}. \tag{2.24}$$

As $\nabla \log p(\mathbf{y}|\mathbf{f})$ is a non-linear function of $\mathbf{f}$, we cannot solve $\nabla\Psi(\mathbf{f}) = 0$ for $\mathbf{f}$ directly, but we can use Newton's method to find a mode with the iteration update:

$$\boldsymbol{\mu}^{k+1} = \boldsymbol{\mu}^k - \nabla^2\Psi(\boldsymbol{\mu}^k)^{-1}\nabla\Psi(\boldsymbol{\mu}^k). \tag{2.25}$$

The Laplace approximation retains popularity due to its simplicity and low computational cost. Nevertheless, it can yield inadequate approximations of the true posterior, particularly when the Hessian matrix fails to accurately represent the width and skewness of the distribution peak (Rasmussen and Williams, 2006, Chapter 3). Consequently, this deficiency leads to poor approximations of both the predictive distribution and the marginal likelihood, with the latter playing a crucial role in hyperparameter estimation, as evidenced in the works of Kuss and Rasmussen (2005) and Nickisch and Rasmussen (2008).

Variational inference has become increasingly popular as a useful alternative to Laplace approximation, in which we aim to find $q(\mathbf{f})$ that minimises the Kullback–Leibler (KL) divergence of the approximated from the true posterior,

$$q(\mathbf{f}) := \underset{q^*(\mathbf{f})}{\arg\min}\, \mathbf{D}_{\mathrm{KL}}[q^*(\mathbf{f})||p(\mathbf{f}|\mathbf{y},\mathbf{X})] \tag{2.26}$$

where $\mathbf{D}_{\mathrm{KL}}[q(\mathbf{f})||p(\mathbf{f}|\mathbf{y},\mathbf{X})] = \mathbb{E}_{q(\mathbf{f})}[\log \frac{q(\mathbf{f})}{p(\mathbf{f}|\mathbf{y},\mathbf{X})}]$. This is equivalent to maximising the Evidence Lower Bound (ELBO):

$$\begin{aligned}\mathrm{ELBO}(q(\mathbf{f})) &= -\mathbf{D}_{\mathrm{KL}}[q(\mathbf{f})||p(\mathbf{f}|\mathbf{X})] + \mathbb{E}_{q(\mathbf{f})}\left[\log p(\mathbf{y}|\mathbf{f})\right] \\ &= \mathbb{E}_{q(\mathbf{f})}\left[\log p(\mathbf{y},\mathbf{f}) - \log q(\mathbf{f})\right], \end{aligned} \tag{2.27}$$

or minimising the Variational Free Energy (VFE), which is simply the negative ELBO. In practice, we restrict $q(\mathbf{f})$ to have a certain form or distribution. With GP, it is common to assume multivariate Gaussian $\mathrm{MVN}(\boldsymbol{\mu},\mathbf{V})$ on $q(\mathbf{f})$, similar to the Laplace approximation. We then treat $\boldsymbol{\mu}$ and $\mathbf{V}$, (or more commonly, the natural parameters of the multivariate Gaussian distributions: $\eta^{(1)} = \mathbf{V}^{-1}\boldsymbol{\mu}, \eta^{(2)} = \frac{1}{2}\mathbf{V}^{-1}$) as variational parameters, and optimise them using (stochastic) gradient descent. See e.g. Khan and Lin (2017) for more details.

Expectation Propagation (EP) also leads to the posterior, approximated by the Gaussian distribution, but through a local likelihood approximation. Noting that the likelihood factorises, $p(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^{n} p(y_i|f_i)$, we have the true posterior,

$$p(\mathbf{f}|\mathbf{y},\mathbf{X}) \propto p(\mathbf{f}|\mathbf{X})p(\mathbf{y}|\mathbf{f}) = p(\mathbf{f}|\mathbf{X})\prod_{i=1}^{n} p(y_i|f_i). \tag{2.28}$$

With each likelihood $p(y_i|f_i)$ approximated by $t(f_i)$, $q(\mathbf{f})$ has the form:

$$q(\mathbf{f}) \propto p(\mathbf{f}|\mathbf{X}) \prod_{i=1}^{n} t(f_i). \tag{2.29}$$

The function $t(f_i)$ is given by e.g., un-normalised Gaussian $t(f_i) = s_i \exp\Big(-\frac{1}{2\sigma_i^2}(f_i - m_i)^2\Big)$, where the parameters s_i, m_i and σ_i^2 are optimised so that the $q(\mathbf{f})$ becomes as close as possible to the true posterior $p(\mathbf{f}|\mathbf{y}, \mathbf{X})$. This is done by individual s_i, m_i, σ_i^2 and hence also $t(f_i)$ being updated sequentially. At the convergence, we have $\tilde{s}_i, \tilde{m}_i$, $\tilde{\sigma}_i^2$ and

$$\prod_{i=1}^{n} t(f_i) = \mathrm{MVN}(\tilde{\mathbf{m}}, \tilde{\Sigma}) \prod_{i=1}^{n} Z_i$$

where $Z_i = s_i\sqrt{2\pi\tilde{\sigma}_i^2}$, $\tilde{\mathbf{m}} = (\tilde{m}_1, \ldots, \tilde{m}_n)^\top$ and $\tilde{\Sigma}$ is diagonal with $\tilde{\Sigma}_{i,i} = \tilde{\sigma}_i^2$. Then the approximated posterior is given by $\mathrm{MVN}(\boldsymbol{\mu}, \mathbf{V})$ where

$$\begin{aligned} \boldsymbol{\mu} &= \mathbf{V}\tilde{\Sigma}^{-1}\tilde{\mathbf{m}} \\ \mathbf{V} &= (\mathbf{K}^{-1} + \tilde{\Sigma}^{-1})^{-1}. \end{aligned}$$

See e.g. Minka (2001) and Rasmussen and Williams (2006, Chalpter 3) for the detail of the algorithm and its implementation.

Once the posterior $p(\mathbf{f}|\mathbf{y}, \mathbf{X})$ is approximated by the respective multivariate Gaussian with mean $\boldsymbol{\mu}$ and the covariance matrix $\mathbf{V}$, the rest of the predictive inference is straightforward. The conditional distribution $p(f^*|\mathbf{y}, \mathbf{X}, \mathbf{x}^*)$ is now analytically tractable. More specifically, it is a multivariate normal distribution with

$$\begin{aligned} \mathbb{E}[f^*|\mathbf{y}, \mathbf{X}, \mathbf{x}^*] &= k(\mathbf{x}^*)^\top\mathbf{K}^{-1}\boldsymbol{\mu} \\ \mathrm{Var}[f^*|\mathbf{y}, \mathbf{X}, \mathbf{x}^*] &= k(\mathbf{x}^*, \mathbf{x}^*) - k(\mathbf{x}^*)^\top\mathbf{K}^{-1}k(\mathbf{x}^*) + k(\mathbf{x}^*)^\top\mathbf{K}^{-1}\mathbf{V}\mathbf{K}^{-1}k(\mathbf{x}^*). \end{aligned} \tag{2.30}$$

The last step, computing posterior predictive distribution (2.21), can be done by Monte Carlo sampling.

Variational inference and expectation propagation often provide superior approximations compared to the Laplace method, all while maintaining computational efficiency when contrasted with MCMC-based approaches. However, a well-recognized issue with these techniques is that they may not guarantee a positive semi-definite (PSD) covariance matrix for the approximated posterior distribution. The work by Wilkinson et al. (2023) delves into the intricacies of various approximation methods, including Laplace approximation, expectation propagation, variational inference, and posterior linearization, offering insights and connections among them. They introduce innovative algorithms that ensure a PSD covariance matrix, applicable across a spectrum of these approximation techniques.

Variational inference, along with many other analytical approximations, presents an additional challenge in the need for customizing model-specific algorithms involving derivatives of the target function—such as the ELBO or marginal likelihood—depending on the chosen kernel and its structure. This can pose a significant hurdle for practitioners. Automatic Differentiation Variational Inference (ADVI) by Kucukelbir et al. (2017) provides an automated implementation of variational inference that extends beyond GP models to more general model classes. It is readily available in the probabilistic programming language Stan, where users specify a probabilistic model (likelihood and prior), and the program generates the variational algorithm and posterior samples. Notably, Stan can leverage the same code used for MCMC-based posterior sampling to implement ADVI.

However, it is crucial to highlight that these approximation techniques do not alleviate the cubic computational complexity inherent to GP models. Larger datasets require more computationally efficient methods. Variational inference, in conjunction with inducing point techniques, can tackle this challenge, leading to the development of sparse variational GPs (Hensman et al., 2015; Matthews et al., 2016; Titsias, 2009). Further exploration of this computational issue is provided in Chapter 4.

2.3.2 Hyperparameter estimation

We discuss the estimation of the hyperparameters for GP regression with a non-Gaussian likelihood. In this context, the hyperparameters to be estimated, denoted as $\boldsymbol{\theta}$, are the parameters associated with the chosen kernel. When employing MCMC for posterior approximation, it is often most convenient to pursue fully Bayesian inference, introducing hyper-priors on $\boldsymbol{\theta}$ and obtaining posterior samples. With analytical approximation methods, it is more customary to optimize the marginal likelihood and obtain Maximum A Posteriori (MAP) estimates. It is worth noting that due to the non-Gaussian nature of the likelihood, the marginal likelihood typically lacks a closed-form expression.

In the case of Laplace approximation to the posterior, approximating the marginal likelihood is a sensible approach. By conducting a first-order Taylor expansion of (2.22), evaluated at its mode, $\hat{\mathbf{f}} = \arg\max_{\mathbf{f}} \Phi(\mathbf{f})$, and substituting it into $p(\mathbf{y}|\mathbf{X}) = \int \exp(\Phi(\mathbf{f}))d\mathbf{f}$, we arrive at the following expression:

$$p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}) = -\frac{1}{2}\hat{\mathbf{f}}^\top \mathbf{K}^{-1}\hat{\mathbf{f}} + \log p(\mathbf{y}|\hat{\mathbf{f}}) - \frac{1}{2}\log|\mathbf{K}| - \frac{1}{2}\log|\mathbf{V}| \tag{2.31}$$

Here, $\mathbf{V} = (-\nabla^2\Psi(\hat{\mathbf{f}}))^{-1}$. See Rasmussen and Williams (2006, Chapter 3) for the details. When employing variational inference, the ELBO (2.27) itself serves as a lower bound on the marginal likelihood (evidence). This can be utilized to select optimal values for $\boldsymbol{\theta}$. While not all terms within the ELBO possess closed-form expressions, they can be approximated numerically using methods such as quadrature techniques. It is important to recognize that (2.31) for Laplace approximation depends on the current value of $\hat{\mathbf{f}}$, whereas the ELBO (2.27) depends on $q(\mathbf{f})$, or more precisely, its variational parameters. As a result, iterating between hyperparameter estimation and the posterior approximation algorithm becomes necessary.

2.3.3 Model selection

The model selection criteria discussed in Section 2.2.3, including marginal likelihood, LOO-CV and WAIC, can be considered for model comparison. However, non-Gaussian likelihood makes the computation of the key quantities analytically intractable, hence additional approximation is required. For example, with LOO-CV, the predictive density for the holdout observation y_i is given by (2.17) and unlike the Gaussian case, $p(y_i|\mathbf{y}_{-i}, \mathbf{X}, \boldsymbol{\theta}, \mathcal{M})$ does not have a closed-form expression. Vehtari et al. (2016) proposed an approximation of LOO-CV using EP or LA. Noting that

$$p(y_i|\mathbf{y}_{-i}, \mathbf{X}, \boldsymbol{\theta}, \mathcal{M}) = \int p(y_i|f_i, \boldsymbol{\theta}) p(f_i|\mathbf{y}_{-i}, \mathbf{X}, \boldsymbol{\theta}, \mathcal{M}) df_i,$$

we consider approximating $p(f_i|\mathbf{y}_{-i}, \mathbf{X}, \boldsymbol{\theta}, \mathcal{M})$ by q_{f_i}in the same manner as EP or LA algorithm described in Section 2.3.1. For each case, the approximated distribution q_{f_i} is already obtained as a by-product of the algorithm or easy to compute using the result at hand; therefore, the additional cost of computing LA-LOO or EP-LOO, if any, is negligible. If the approximation is poor, k-fold cross-validation is recommended (Vehtari et al., 2016).

2.4 Semi-parametric GP models

Let us assume that we have a d-dimensional covariate $\mathbf{x} \in \mathcal{X}$ with e.g. $\mathcal{X} \subset \mathbb{R}^d$, some of which we expect a complex, non-linear relationship with the response y and the other having a linear effect on the response. This is common in, for example, spatial analysis, where we have location information represented by 2− or 3−dimensional geo-coordinates and some additional information available, such as population or average income in the area/location of interest. The spatial information can be used to account for auto-correlation and non-linear relationships, but the effect of other variables may be linear, and its interpretation may be of scientific interest. This linear part can easily be incorporated into the GP model.

Let us revisit the regression model given by (2.1). We assume that $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^d$, and is separated into two components, $\mathbf{x} = (\mathbf{z}^\top, \mathbf{s}^\top)^\top$, where $\mathbf{z} \in \mathcal{Z} \subset \mathbb{R}^p$ and $\mathbf{s} \in \mathcal{S} \subset \mathbb{R}^{d-p}$. Note that $\mathcal{X} = \mathcal{Z} \times \mathcal{S}$. Then, for $i = 1, \ldots, n$, we consider a regression model:

$$y_i = f(\mathbf{z}_i, \mathbf{s}_i) + \epsilon_i \tag{2.32}$$

$$f(\mathbf{z}_i, \mathbf{s}_i) = a + \mathbf{z}_i^\top \boldsymbol{\beta} + f_s(\mathbf{s}_i) \tag{2.33}$$

where we assume an i.i.d error $\epsilon_i \sim N(0, \sigma^2)$, $a \sim N(0, \sigma_a^2)$ and $f_s \sim \mathrm{GP}(0, k_s)$ with a kernel $k_s : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$. Then, the linear part can be incorporated into the mean function of the prior and $\boldsymbol{\beta}$ in the semi-parametric GP model can be estimated along with other hyperparameters. See (2.14) for the posterior with such non-zero mean function.

Instead of assuming a fixed mean function, we can place a prior on the regression coefficient $\boldsymbol{\beta}$, as $\boldsymbol{\beta} \sim \mathrm{MVN}(\mathbf{0}, \mathbf{B})$. After integrating out $\boldsymbol{\beta}$, the prior on the overall function f is $\mathrm{GP}(0, k)$ where $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is given by

$$k(\mathbf{x}, \mathbf{x}') = \sigma_a^2 + \mathbf{z}^\top \mathbf{B} \mathbf{z}' + k_s(\mathbf{s}, \mathbf{s}'). \tag{2.34}$$

This is equivalent to re-write (2.33) as $f(\mathbf{z}_i, \mathbf{s}_i) = a + f_z(\mathbf{z}_i) + f_s(\mathbf{s}_i)$ and assume $f_z \sim \mathrm{GP}(0, k_z)$ where the kernel $k_z : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ is given by $k_z(\mathbf{z}, \mathbf{z}') = \mathbf{z}^\top \mathbf{B} \mathbf{z}'$. The constant term a can be included in the linear term by writing $\mathbf{z}_i = (1, z_{1i}, \ldots, z_{pi})^\top$

and $\boldsymbol{\beta} = (\alpha, \beta_1, \ldots, \beta_p)^\top$. The matrix $\mathbf{B}$ is then a $(p+1) \times (p+1)$ matrix, and we re-write (2.34) by

$$k(\mathbf{x}, \mathbf{x}') = \mathbf{z}^\top \mathbf{B} \mathbf{z}' + k_s(\mathbf{s}, \mathbf{s}'). \tag{2.35}$$

Equivalently, we have $\mathbf{f} \sim \mathrm{MVN}(\mathbf{0}, \mathbf{K})$ with the Gram matrix given by

$$\mathbf{K} = \mathbf{Z}\mathbf{B}\mathbf{Z}^\top + \mathbf{K}_s \tag{2.36}$$

where $\mathbf{Z}$ is $n \times (p+1)$ matrix with i-th row given by $\mathbf{z}_i^\top$, $\mathbf{K}_s = \{k_{s,ij}\}_{1 \leq i,j \leq n}$. Incorporating the parametric component into the kernel k simplifies the parameter estimation. For instance, assuming $\mathbf{B} = \sigma_\beta \mathbf{I}_{p+1}$, we only need to estimate the scale parameter σ_β instead of $p+1$ regression coefficients. Another advantage is the flexibility in modelling. As we see in Chapter 3, the interaction between different predictors, e.g., $\mathbf{s}$ and $\mathbf{z}$, can be modelled by taking the tensor product of the two specified kernels k_s and k_z.

Note that we can replace the linear part in (2.33) with $\phi(\mathbf{z}_i)^\top \boldsymbol{\beta}$, where $\phi : \mathbb{R}^p \rightarrow \mathbb{R}^q$ are some fixed basis functions. The following section focuses on the simple linear case; however, they can easily be generalised to semi-parametric models. See Rasmussen and Williams (2006, Chapter 2) for the details. It is also worth noting that the parametric function and non-parametric function in (2.33) can be defined on the same set, i.e., we consider a regression model specified by (2.1) and $f(\mathbf{x}_i) = a + \mathbf{x}_i^\top \boldsymbol{\beta} + f_x(\mathbf{x}_i)$ where the prior over f is specified in the same manner as (2.34).

2.4.1 Posterior

Following Rasmussen and Williams (2006, Chapter 2), the posterior of f for the model given by (2.32) - (2.34) is a GP with mean and covariance given by

$$\begin{aligned}\bar{m}(\mathbf{x}) &= \mathbf{k}_s(\mathbf{s})^\top(\mathbf{K}_s+\sigma^2\mathbf{I}_n)^{-1}\mathbf{y}+\mathbf{r}^\top\bar{\boldsymbol{\beta}} && (2.37)\\ \bar{k}(\mathbf{x},\mathbf{x}') &= k_s(\mathbf{s},\mathbf{s}')-\mathbf{k}_s(\mathbf{s})^\top(\mathbf{K}_s+\sigma^2\mathbf{I}_n)^{-1}\mathbf{k}_s(\mathbf{s})+\\ &\qquad \mathbf{r}(\mathbf{x})^\top\left(\mathbf{B}^{-1}+\mathbf{Z}^\top(\mathbf{K}_s+\sigma^2\mathbf{I}_n)^{-1}\mathbf{Z}\right)^{-1}\mathbf{r}(\mathbf{x}') && (2.38)\end{aligned}$$

where

$$\begin{aligned}\mathbf{r}(\mathbf{x}) &= \mathbf{z}-\mathbf{Z}^\top\left(\mathbf{K}_s+\sigma^2\mathbf{I}_n\right)^{-1}\mathbf{k}_s(\mathbf{s}) && (2.39)\\ \bar{\boldsymbol{\beta}} &= \left(\mathbf{B}^{-1}+\mathbf{Z}^\top(\mathbf{K}_s+\sigma^2\mathbf{I}_n)^{-1}\mathbf{Z}\right)^{-1}\mathbf{Z}^\top(\mathbf{K}_s+\sigma^2\mathbf{I}_n)^{-1}\mathbf{y} && (2.40)\end{aligned}$$

for $\mathbf{s},\mathbf{s}'\in\mathcal{S}$ and $\mathbf{z},\mathbf{z}'\in\mathcal{Z}$. Notice that the first term in (2.37) and the first two terms in (2.38) corresponds with the posterior mean and kernel given in (2.12) and (2.13) for a simple GP regression. The additional terms can be seen as contributions of the parametric term. The equation (2.40) can be interpreted as the posterior mean of the regression coefficient. Moreover, we can derive the posterior distribution of $\boldsymbol{\beta}$, which is multivariate normal with mean given by (2.40) and the covariance matrix given by

$$\boldsymbol{\Sigma}_\beta = \mathbf{B}-\mathbf{B}\mathbf{Z}^\top\left(\mathbf{K}_s+\sigma^2\mathbf{I}_n\right)^{-1}\mathbf{Z}\mathbf{B} \qquad (2.41)$$

In the case of a vague prior, e.g. $\mathbf{B}=\sigma_\beta^2\mathbf{I}_p$ where $\sigma_\beta^2\to\infty$, the expression

$$\left(\mathbf{B}^{-1}+\mathbf{Z}^\top(\mathbf{K}_s+\sigma^2\mathbf{I}_n)^{-1}\mathbf{Z}\right)^{-1}$$

involved in (2.38) and (2.40) reduces to $\left(\mathbf{Z}^\top(\mathbf{K}_s+\sigma^2\mathbf{I}_n)^{-1}\mathbf{Z}\right)^{-1}$.

For a non-Gaussian likelihood, although we do not know exact posterior, we have expressions similar to (2.37) and (2.40), but $\mathbf{y}$ replaced with $\mathbb{E}[\mathbf{f}|\mathbf{y}, \mathbf{X}]$ and $\mathbf{K}_s + \sigma^2\mathbf{I}_n$ replaced with $\mathbf{K}_s$. For example, for the mean of the regression coefficient, we have:

$$\bar{\beta} = \left(\mathbf{B}^{-1} + \mathbf{Z}^\top\mathbf{K}_s^{-1}\mathbf{Z}\right)^{-1} \mathbf{Z}^\top\mathbf{K}_s^{-1}\mathbb{E}[\mathbf{f}|\mathbf{y}, \mathbf{X}].$$

The variance can be computed by

$$\mathbf{\Sigma}_\beta = \mathbf{B} - \mathbf{B}\mathbf{Z}^\top\mathbf{K}_s^{-1}\mathbf{Z}\mathbf{B} + \mathbf{B}\mathbf{Z}^\top\mathbf{K}_s^{-1}\mathbf{V}\mathbf{K}_s^{-1}\mathbf{Z}\mathbf{B},$$

where $\mathbf{V}$ is the covariance matrix of the posterior of $\mathbf{f}$.

2.4.2 Marginal likelihood

The marginal likelihood for the discussed model with Gaussian likelihood can be computed by

$$\begin{aligned}\log p(\mathbf{y}|\mathbf{X}, \mathbf{B}) = & -\frac{1}{2}\mathbf{y}^\top(\mathbf{K}_s + \sigma^2\mathbf{I}_n)^{-1}\mathbf{y} + \frac{1}{2}\mathbf{y}^\top\mathbf{C}\mathbf{y} \\ & -\frac{1}{2}\log|\mathbf{K}_s + \sigma^2\mathbf{I}_n| - \frac{1}{2}\log|\mathbf{B}| - \frac{1}{2}\log|\mathbf{A}| - \frac{n}{2}\log 2\pi \qquad (2.42)\end{aligned}$$

where $\mathbf{A} = \mathbf{B}^{-1} + \mathbf{X}^\top(\mathbf{K}_s + \sigma^2\mathbf{I}_n)^{-1}\mathbf{X}$ and $\mathbf{C} = (\mathbf{K}_s + \sigma^2\mathbf{I}_n)^{-1}\mathbf{X}\mathbf{A}^{-1}\mathbf{X}^\top(\mathbf{K}_s + \sigma^2\mathbf{I}_n)^{-1}$. If we have a vague prior on $\mathbf{B}$, evaluating the log determinant term becomes an issue as it approaches minus infinity. In such a case, we have to discard the log determinant term involving $\mathbf{B}$ and re-scale the last term by $1 - \frac{m}{n}$ (Ansley and Kohn, 1985).

2.5 Related methods

In this section, we discuss the connection between GP regression and other well-established statistical methods, namely kernel ridge regression in Section 2.5.1, and Kriging and conditional auto-regressive models in 2.5.2.

2.5.1 Kernel ridge regression

In this section, we consider a regression problem for the response y_i and the predictors $\mathbf{x}_i \in \mathcal{X}$ for $i = 1, \dots, n$, where the aim is to find a function $f : \mathcal{X} \to \mathbb{R}$ that best explains the relationship between y_i and $\mathbf{x}_i$. With kernel ridge regression, the function f is estimated as a solution to the following problem:

$$\operatorname*{arg\,min}_{f \in \mathcal{H}_k} \left(\frac{1}{n} \sum_{i=1}^{n} (y_i - f(\mathbf{x}_i))^2 + \lambda ||f||_{\mathcal{H}_k} \right) \tag{2.43}$$

where $\mathcal{H}_k$ is a Reproducing kernel Hilbert Space (RKHS, see Appendix A.1.1) induced by the kernel k. Let us now consider a similar problem, but the first term replaced by a loss function $l : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$:

$$\operatorname*{arg\,min}_{f \in \mathcal{H}_k} \left(l\left(y_i, f(\mathbf{x}_i)\right) + \lambda ||f||_{\mathcal{H}_k} \right). \tag{2.44}$$

By the representer theorem (Kimeldorf and Wahba, 1970; O'sullivan et al., 1986; Schölkopf et al., 2001), the solution to the above (2.44) has the form

$$f^*(\mathbf{x}) = \sum_{i=1}^{n} \alpha_i k(\mathbf{x}, \mathbf{x}_i), \quad \mathbf{x} \in \mathcal{X} \tag{2.45}$$

where $\alpha_i \in \mathbb{R}$ for $i = 1, \dots, n$ are some coefficient that depends on the sample $(y_i, \mathbf{x}_i)_{i=1}^n$. The equation above (2.45) can also be written as

$$f^*(\mathbf{x}) = \mathbf{k}(\mathbf{x})^\top \boldsymbol{\alpha}, \quad \mathbf{x} \in \mathcal{X} \tag{2.46}$$

where $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_n)^\top$ and $\mathbf{k}(\mathbf{x}) = (k(\mathbf{x}, \mathbf{x}_1), \dots, k(\mathbf{x}, \mathbf{x}_n))^\top$. For a mean squared loss, we recover the problem given in (2.43). Substituting (2.46) into (2.43) with $\lambda = \sigma^2$, and minimising with respect to $\boldsymbol{\alpha}$, we have $\hat{\boldsymbol{\alpha}} = (\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1}\mathbf{y}$. We see that this corresponds with the posterior mean of GP regression, as given in (2.12).

2.5.2 Methods in spatial statistics

Kriging

It is widely recognized that the posterior (predictive) mean in GP regression has a close relationship with Kriging prediction, a classical method in geostatistics. Kriging was first introduced for a mining problem by Krige (1951), with formal mathematical theories developed by Matheron (1963). In this section, we focus on the regression problem for a continuous response and see the connection between universal Kriging, also called regression Kriging, to the semi-parametric GP models discussed in Section 2.4. Following common practices in geostatistics, we introduce slight changes in notation and formulation. The model is defined as follows:

$$y(\mathbf{s}_i) = \underbrace{\mu(\mathbf{s}_i)}_{\text{mean component}} + \underbrace{e(\mathbf{s}_i)}_{\text{error component}} \tag{2.47}$$

Here, the mean and error components are specified as:

$$\mu(\mathbf{s}_i) = \mathbf{Z}(\mathbf{s}_i)^\top \boldsymbol{\beta} \tag{2.48}$$

$$e(\mathbf{s}_i) = f(\mathbf{s}_i) + \epsilon(\mathbf{s}_i). \tag{2.49}$$

This model bears a resemblance to (2.32) and (2.33), with the terms y_i, $\mathbf{z}_i$, and ϵ_i replaced by $y(\mathbf{s}_i)$, $\mathbf{Z}(\mathbf{s}_i)$, and $\epsilon(\mathbf{s}_i)$, respectively. We also assume that the constant term is included in (2.48). The error component (2.49) is assumed to be a zero-mean random process, with its first term f accounting for spatial variation and the second term ϵ accounting for i.i.d. measurement error. Specifically, we assume that $\mathbb{E}[\epsilon(\mathbf{s})] = 0$ and $\mathrm{Var}[\epsilon(\mathbf{s})] = \sigma^2$ for $\mathbf{s} \in \mathcal{S}$. As we are dealing with spatially referenced data, $\mathcal{S}$ is commonly considered to be either $\mathbb{R}^2$ or $\mathbb{R}^3$.

The primary objectives in this context may involve estimating the regression coefficient $\boldsymbol{\beta}$, understanding the spatial auto-correlation structure modelled by f, or making predictions at locations where samples were not taken. The latter step, involving prediction, is commonly referred to as Kriging.

In classical geostatistics, the initial step is to compute the ordinary least square estimator for $\boldsymbol{\beta}$. To address auto-correlation, the estimation of the covariance function (kernel) becomes a crucial subsequent stage. In geostatistics, when the error process exhibits second-order stationarity (i.e., when the covariance function is stationary), it is common to focus on a semi-variogram, defined as

$$\gamma(\mathbf{s}, \mathbf{s}') = \frac{1}{2}\mathrm{Var}\left(e(\mathbf{s}) - e(\mathbf{s}')\right). \tag{2.50}$$

Another widely used stationarity assumption is intrinsic stationarity, characterized by:

$$\gamma(\mathbf{s}, \mathbf{s}') = \gamma(\mathbf{s} - \mathbf{s}'), \quad \mathbf{s}, \mathbf{s}' \in \mathcal{S}.$$

Under this assumption, there exists a relationship between the covariance function k and the semi-variogram γ:

$$\gamma(\mathbf{h}) = k(\mathbf{0}) - k(\mathbf{h}) \tag{2.51}$$

where $\mathbf{h} = \mathbf{s} - \mathbf{s}' \in \mathcal{S}$. The empirical estimation of the semi-variogram can be performed using the residual errors, $\hat{e}(\mathbf{s}_i)$ for $i = 1, \ldots, n$, obtained from the initial step. This empirical semi-variogram is typically smoothed using parametric semi-variogram models, such as Gaussian (SE), exponential, or Matérn.

If the primary objective is to estimate and interpret the regression coefficients, then the coefficients should be re-estimated using estimated generalized least squares (EGLS). The EGLS estimator aligns with the results from GP regression (2.40) when a vague prior is placed on $\boldsymbol{\beta}$.

For prediction at an arbitrary location $\mathbf{s}^* \in \mathcal{S}$, Kriging aims to interpolate from surrounding measurements using the following equation:

$$\hat{Y}(\mathbf{s}^*) = \sum_{i=1}^{n} w_i \mathbf{Y}(\mathbf{s}_i) = \mathbf{w}^\top \mathbf{Y}. \tag{2.52}$$

where $\mathbf{w} = (w_1, \ldots, w_n)^\top$ is a fixed vector and $\mathbf{Y} = (Y_(\mathbf{s}_1), \ldots, Y_(\mathbf{s}_n))^\top$. This is often referred to as a linearity condition. Kriging provides the best linear unbiased prediction, where the estimator minimizes:

$$\mathrm{Var}[Y(\mathbf{s}^*) - \hat{Y}(\mathbf{s}^*)]$$

subject to $\mathbb{E}[\hat{Y}(\mathbf{s})] = \mathbb{E}[Y(\mathbf{s})]$. This is equivalent to minimizing: $\mathbb{E}([Y(\mathbf{s}^*) - \hat{Y}(\mathbf{s}^*))^2]$ with the same bias constraint. This leads to a minimization problem of the Lagrangian function:

$$L(\mathbf{w}, \boldsymbol{\lambda}) = -\mathbf{w}^\top \boldsymbol{\Gamma} \mathbf{w} + 2\mathbf{w}^\top \boldsymbol{\gamma} - 2\boldsymbol{\lambda} \left(\mathbf{Z}(\mathbf{s}^*)^\top - \mathbf{w}^\top \mathbf{Z} \right) \tag{2.53}$$

where $\mathbf{\Gamma} = [\gamma(\mathbf{s}_i, \mathbf{s}_j)]_{1\leq i,j,\leq n}$, $\boldsymbol{\gamma} = (\gamma(\mathbf{s}^*, \mathbf{s}_1), \dots, \gamma(\mathbf{s}^*, \mathbf{s}_n))^\top$ and $\mathbf{Z}$ is the data matrix with i-th row given by $\mathbf{Z}(\mathbf{s}_i)$. Minimizing (2.53) with respect to $\mathbf{w}$ (and $\boldsymbol{\lambda}$), we obtain:

$$\hat{\mathbf{w}} = \left(\boldsymbol{\gamma} + \mathbf{Z}\left(\mathbf{Z}^\top\mathbf{\Gamma}^{-1}\mathbf{Z}\right)^{-1}\mathbf{r}_\gamma\right)^\top \mathbf{\Gamma}^{-1}. \tag{2.54}$$

where $\mathbf{r}_\gamma = \left(\mathbf{Z}(\mathbf{s}^*) - \mathbf{Z}^\top\mathbf{\Gamma}^{-1}\boldsymbol{\gamma}\right))$. Hence, the Kriging equation for prediction and variance becomes:

$$\hat{Y}(\mathbf{s}^*)s = \boldsymbol{\gamma}^\top\mathbf{\Gamma}^{-1}\mathbf{Y} + \mathbf{r}_\gamma^\top\left(\mathbf{Z}^\top\mathbf{\Gamma}^{-1}\mathbf{Z}\right)^{-1}\mathbf{Z}^\top\mathbf{\Gamma}^{-1}\mathbf{Y} \tag{2.55}$$

$$\hat{\sigma}^2(\mathbf{s}^*) = \boldsymbol{\gamma}^\top\mathbf{\Gamma}^{-1}\boldsymbol{\gamma} - \mathbf{r}_\gamma^\top\left(\mathbf{Z}^\top\mathbf{\Gamma}^{-1}\mathbf{Z}\right)^{-1}\mathbf{r}_\gamma \tag{2.56}$$

By exploiting the relationship between the semi-variogram and covariance function under the intrinsic stationarity assumption (2.51), we can establish the equivalence between Kriging equations (2.55) and (2.56) and the GP posterior mean and variance for the semi-parametric model (2.37) and (2.34). For a detailed derivation of Kriging equations, see e.g., Cressie (1993, Chapter 3) or Zimmerman and Stein (2010).

Conditional auto-regressive model

When dealing with areal data, such as crime counts or disease cases within regions of a city, one approach to model spatial variation is to compute the centroid of each region as a measure of location and apply standard GP regression. However, a more commonly employed technique in spatial statistics is the use of Conditional Auto-Regressive (CAR) models. This class of models originated in the work of Besag (1974), who introduced models for analysing spatially discrete data. In CAR models, the specification of auto-correlation structure involves the weighted adjacency matrix $\mathbf{W}$, where the i, j-th element is positive only when regions i and j share a border and zero otherwise. It is also symmetric, $w_{ij} = w_{ji}$.

Let us consider a regression model similar to (2.47)-(2.49), given by:

$$y_i = \mu_i + e_i,$$

where $\mu_i = \mathbf{z}_i^\top \boldsymbol{\beta}$. In CAR models, the prior for the (spatial) error component is specified through a set of conditional distributions of e_i given all other error components, denoted as $\mathbf{e}_{-i}$. The simplest of such distributions is defined as follows:

$$e_i|\mathbf{e}_{-i} \sim N\left(\frac{1}{d_i}\sum_{j\sim i} e_j, \frac{\tau^2}{d_i}\right),$$

where d_i denotes the number of neighboring regions, and $j \sim i$ indicates that region j shares a border with region i. This prior specification corresponds to the intrinsic auto-regressive model (Besag et al., 1991). The joint distribution of $\mathbf{e} = (e_1, \ldots, e_n)$ is then given by:

$$\mathbf{e} \sim \mathbf{MVN}(0, \tau^2\mathbf{Q}^{-1}),$$

where $\mathbf{Q} = \mathbf{D} - \mathbf{W}$ and $\mathbf{D}$ is a degree matrix, i.e., d$n \times n$ diagonal matrix with diagonal elements $d_1, \ldots, d_n$, and $\mathbf{W}$ is the adjacency matrix. Note that the precision matrix $\mathbf{Q}$ in this case is centred and thus singular. Besag et al. (1991) also proposed adding a random effect to the error term, which is called the convolution model. In this case, the individual terms consisting of the error term are not identifiable; hence additional assumptions are needed. The matrix $\mathbf{D} - \mathbf{W}$ is also known as the Graph Laplacian or un-normalized Graph Laplacian when $\mathbf{W}$ is a weighted adjacency matrix. Other common choices for $\mathbf{Q}$ include $\mathbf{D} - \rho\mathbf{W}$ with $0 < \rho < 1$, corresponding to the Cressie model (Cressie, 1993; Stern and Cressie, 2000), $\mathbf{Q} = \rho\mathbf{W} + (1-\rho)\mathbf{I}$, which is known as Leroux model (Leroux et al., 1999; MacNab, 2003), and:

$$\mathbf{Q} = \mathbf{I}_n - \phi\mathbf{W}, \tag{2.57}$$

where $\lambda_1^{-1} \leq \phi \leq \lambda_n^{-1}$ and λ_1 and λ_n are the smallest and largest eigenvalues of $\mathbf{W}$. The prior specification given by (2.57) is particularly significant, as many CAR priors can be reformulated in this form when the response and covariates are appropriately scaled (Cressie et al., 2005; De Oliveira, 2012). This covariance matrix also corresponds to a kernel on graphs, known as the Katz kernel (Katz, 1953) or Von Neumann Diffusion Kernel (Kandola et al., 2002). Hence, for areal data, we can consider a multivariate Gaussian prior on the spatial error component with its covariance matrix determined by different kernels on graphs. For other kernels on graphs, see e.g. Avrachenkov et al. (2019); Fouss et al. (2012); Smola and Kondor (2003).

CAR models find extensive use in cases involving non-Gaussian likelihoods, such as the Poisson distribution with a log link function. For an overview, see Waller and Carlin (2010). A comprehensive comparison of various CAR models for disease mapping can be found in Lee (2011).

Chapter 3

Additive interaction modelling with Gaussian process priors

Chapter 2 is primarily dedicated to regression with Gaussian process (GP) priors with a single, possibly multidimensional covariate but also covers models involving another set of covariates with a presumed linear relationship to the response variable (a semi-parametric model, see Section 2.4). We add a linear kernel to complement an existing kernel to capture this linear component. This is an example of an additive GP model (Duvenaud et al., 2013; Plate, 1999).

As the name implies, additive GP models assume an additive structure of GPs for the regression function, akin to generalised additive models (Hastie and Tibshirani, 1990). In the semi-parametric model discussed in the preceding chapter, we incorporate two parts: one that models a potentially non-linear and complex relationship and another in the form of a linear combination of covariates and regression coefficients. This effectively models the main effects of $\mathbf{x}_1$ and $\mathbf{x}_2$. In many real-world scenarios, we encounter more than two sets of covariates, where the effects may also interact. For instance, let us consider a case with a response variable $y_i \in \mathbb{R}$ and three sets of covariates: $\mathbf{x}_{1i} \in \mathcal{X}_1$, $\mathbf{x}_{2i} \in \mathcal{X}_2$, and $\mathbf{x}_{3i} \in \mathcal{X}_3$ for $i = 1, \ldots, n$. It is important to note that each of $\mathbf{x}_l$ for $l = 1, 2, 3$ does not necessarily have to be one-dimensional; for instance, $\mathcal{X}_1 \subset \mathbb{R}^2$ if $\mathbf{x}_1$ represents geographical coordinates. We can model the relationships between the response and covariates as follows:

$$y_i = f(\mathbf{x}_{1i}, \mathbf{x}_{2i}, \mathbf{x}_{3i}) + \epsilon_i \tag{3.1}$$

where ϵ_i represents the error term. The specific form of $f(\mathbf{x}_{1i}, \mathbf{x}_{2i}, \mathbf{x}_{3i})$ depends on the model assumptions. One possible example is:

$$\begin{aligned} f(\mathbf{x}_{1i}, \mathbf{x}_{2i}, \mathbf{x}_{3i}) = a + \underbrace{f_1(\mathbf{x}_{1i}) + f_2(\mathbf{x}_{2i}) + f_3(\mathbf{x}_{3h})}_{\text{main effect}} \\ + \underbrace{f_{13}(\mathbf{x}_{1i}, \mathbf{x}_{3i}) + f_{23}(\mathbf{x}_{2i}, \mathbf{x}_{3i})}_{\text{two-way interaction effect}}, \end{aligned} \tag{3.2}$$

which includes a constant term, main effect terms, and two-way interaction effect terms among all three sets of covariates. If this model seems to overfit the data, one may consider eliminating some two-way interaction terms. Furthermore, suppose all two-way interaction effects are present. In that case, it may be worth investigating the addition of a three-way interaction term $f_{123}(\mathbf{x}_{1i}, \mathbf{x}_{2i}, \mathbf{x}_{3i})$ to potentially provide a better explanation of the data structure.

This chapter explores how GP models approach this statistical modelling problem through kernel functions. We place special emphasis on the construction and selection of interaction terms and the interpretation of both main and interaction effects. We primarily focus on models involving two sets of predictors as a motivating example since the concepts and advantages can be effectively illustrated in this simple setting. Each section also provides a generalisation to higher-dimensional cases.

3.1 Statistical modelling with kernels

Let $\mathcal{X}_1$ and $\mathcal{X}_2$ be nonempty sets. Consider a regression model for a real-valued response y and two (sets of) predictors $\mathbf{x}_1 \in \mathcal{X}_1, \mathbf{x}_2 \in \mathcal{X}_2$. We denote the sample (of size n) as $(y_i, \mathbf{x}_{1i}, \mathbf{x}_{2i})_{i=1}^n$. Similar to (3.1) we have:

$$y_i = f(\mathbf{x}_{1i}, \mathbf{x}_{2i}) + \epsilon_i \tag{3.3}$$

where the error terms $(\epsilon_1, \ldots, \epsilon_n) \sim \mathbf{MVN}(\mathbf{0}, \Sigma)$. For i.i.d errors, we write $\Sigma = \sigma^2 \mathbf{I}_n$ where $\mathbf{I}_n$ is the $n \times n$ identity matrix. Given two predictors and a constant a, it is natural to consider the following additive models:

$$f(\mathbf{x}_{1i}, \mathbf{x}_{2i}) = a + f_1(\mathbf{x}_{1i}) + f_2(\mathbf{x}_{2i}) \tag{3.4}$$

$$f(\mathbf{x}_{1i}, \mathbf{x}_{2i}) = a + f_1(\mathbf{x}_{1i}) + f_2(\mathbf{x}_{2i}) + f_{12}(\mathbf{x}_{1i}, \mathbf{x}_{2i}). \tag{3.5}$$

While the first represents the main effect model, the second assumes an additional interaction effect between the two predictors. The main idea of additive GP models is to put a GP prior on each function f_1, f_2 and f_{12} in (3.4) and (3.5).

3.1.1 Sum kernels and main effect terms

Consider the main effect model with two predictors specified in (3.3) and (3.4). Given two kernels $k_1 : \mathcal{X}_1 \times \mathcal{X}_1 \to \mathbb{R}$ and $k_2 : \mathcal{X}_2 \times \mathcal{X}_2 \to \mathbb{R}$, we assume $f_1 \sim \mathrm{GP}(0, k_1)$ and $f_2 \sim \mathrm{GP}(0, k_2)$. A nice property of kernels is that the sum of the given two valid kernels constitutes a new kernel (See section 2.1.2). With the additional assumption that the constant term $a \sim N(0, 1)$, the overall function $f : \mathcal{X} \to \mathbb{R}$ with $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$

follows $\mathrm{GP}(0,k)$, where $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is given by

$$k((\mathbf{x}_1, \mathbf{x}_2), (\mathbf{x}_1', \mathbf{x}_2')) = 1 + k_1(\mathbf{x}_1, \mathbf{x}_1') + k_2(\mathbf{x}_2, \mathbf{x}_2').$$

Note that the first term 1 is also a positive definite kernel, known as a constant kernel. Hence, the function k can be seen as a sum of three positive definite kernels. To avoid constraining the prior variance of the constant term a, we multiply the overall kernel k by α_0^2 where $\alpha_0 > 0$, i.e.,

$$k((\mathbf{x}_1, \mathbf{x}_2), (\mathbf{x}_1', \mathbf{x}_2')) = \alpha_0^2 \left(1 + k_1(\mathbf{x}_1, \mathbf{x}_1') + k_2(\mathbf{x}_2, \mathbf{x}_2')\right).$$

With this formulation, we assume $a \sim N(0, \alpha_0^2)$, and the scale parameters of k_1 and k_2, (denoted by α_1^2 and α_2^2), are re-scaled by the factor of α_0^2. The corresponding Gram matrix is given by

$$\mathbf{K} = \alpha_0^2 \left(\mathbf{1}_n \mathbf{1}_n^\top + \mathbf{K}_1 + \mathbf{K}_2\right) \tag{3.6}$$

where $\mathbf{1}_n \mathbf{1}_n^\top$ is a $n \times n$ matrix with all elements equal to 1, and $\mathbf{K}_l$ is a Gram matrix with i,j-th element given by $k_l(\mathbf{x}_l, \mathbf{x}_l')$. We can generalise this to the case with d predictors $\mathbf{x}_1, \ldots \mathbf{x}_d$.

Example 7 (Main effect kernel)**.** *Let $\mathcal{X}_l$ be a nonempty set and k_l be a positive definite kernel on $\mathcal{X}_l \times \mathcal{X}_l$ for $l = 1, \ldots, d$. Let $\mathcal{X} = \mathcal{X}_1, \ldots, \mathcal{X}_d$. Given d predictors, the kernel that gives a main effect GP model is defined on $\mathcal{X} \times \mathcal{X}$ and given by*

$$k_{main}((\mathbf{x}_1, \ldots \mathbf{x}_d), (\mathbf{x}_1', \ldots \mathbf{x}_d')) = \alpha_0^2 \left(1 + \sum_{l=1}^{d} k_l(\mathbf{x}_l, \mathbf{x}_l')\right), \quad \mathbf{x}_l, \mathbf{x}_l' \in \mathcal{X}_l. \tag{3.7}$$

3.1.2 Tensor product kernels and interaction effect terms

The interaction effect model (3.5) has an additional function term f_{12}. We put a zero mean GP prior on this, with its kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ given by a tensor product of two kernels k_1 and k_2. Formally, we have $k = k_1 \otimes k_2$ defined by $k((\mathbf{x}_1, \mathbf{x}_2), (\mathbf{x}'_1, \mathbf{x}'_2)) = k_1(\mathbf{x}_1, \mathbf{x}'_1)k_2(\mathbf{x}_2, \mathbf{x}'_2)$ where $\otimes$ is a tensor product operator. Generalising this kernel to the case involving d predictors, we have the following definition.

Definition 5 (Tensor product kernel)**.** *Let $\mathcal{X}_l$ be a nonempty set and k_l be a positive definite kernel on $\mathcal{X}_l \times \mathcal{X}_l$ for $l = 1, \dots, d$. With $\mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_d$, a tensor product of $\{k_l\}_{l=1}^d$ is a kernel on $\mathcal{X} \times \mathcal{X}$ defined as*

$$(\otimes_{l=1}^d k_l)((\mathbf{x}_1, ..., \mathbf{x}_d), (\mathbf{x}'_1, ..., \mathbf{x}'_d)) = \prod_{l=1}^d k_l(\mathbf{x}_l, \mathbf{x}'_l), \quad \mathbf{x}_l, \mathbf{x}'_l \in \mathcal{X}_l. \tag{3.8}$$

Using the sum kernel for the main effect terms, and the tensor product kernel for the interaction effect term, we can specify the prior on the regression function f in (3.5) as $\text{GP}(0, k)$, where $k = \alpha_0^2(1 + k_1 + k_2 + k_1 \otimes k_2)$ is defined by

$$k((\mathbf{x}_1, \mathbf{x}_2), (\mathbf{x}'_1, \mathbf{x}'_2)) = \alpha_0^2 \left(1 + k_1(\mathbf{x}_1, \mathbf{x}'_1) + k_2(\mathbf{x}_2, \mathbf{x}'_2) + k_1(\mathbf{x}_1, \mathbf{x}'_1)k_2(\mathbf{x}_2, \mathbf{x}'_2)\right). \tag{3.9}$$

The corresponding Gram matrix can be written as

$$\mathbf{K} = \alpha_0^2 \left(\mathbf{1}_n\mathbf{1}_n^\top + \mathbf{K}_1 + \mathbf{K}_2 + \mathbf{K}_1 \circ \mathbf{K}_2\right) \tag{3.10}$$

where $\circ$ is the element-wise product operator.

3.1.3 ANOVA decomposition kernel

When including interaction terms in a model, it is common practice to include both the main terms and all lower-order interaction terms. In this section, we introduce a special class of additive kernels, known as the ANOVA decomposition kernel, that can naturally take this into consideration.

Following Bergsma and Jamil (2023), for a model with d predictors, we define the *saturated ANOVA decomposition kernel* as follows.

Definition 6 (Saturated ANOVA decomposition kernel)**.** *Let $\mathcal{X}_l$ be a nonempty set, k_l be a positive definite kernel on $\mathcal{X}_l \times \mathcal{X}_l$ for $l = 1, \ldots, d$ and α_0 be a positive constant. With $\mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_d$, the saturated ANOVA decomposition kernel $k_{s\text{-}anova} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is given by:*

$$k_{s\text{-}anova}((\mathbf{x}_1, ..., \mathbf{x}_l), (\mathbf{x}'_1, ..., \mathbf{x}'_l)) = \alpha_0^2 \prod_{l=1}^{d} (1 + k_l(\mathbf{x}_l, \mathbf{x}'_l)), \quad \mathbf{x}_l, \mathbf{x}'_l \in \mathcal{X}_l. \tag{3.11}$$

The ANOVA decomposition kernel was first introduced by Stitson et al. (1999) in the context of Support Vector Machines. It borrows the idea of an ANOVA decomposition of a function in a Reproducing Kernel Hilbert Space (RKHS), as used in smoothing spline ANOVA models (Gu, 2002; Gu and Wahba, 1993; Wahba et al., 1995). The formulation in (3.11) is different from the standard ANOVA decomposition kernel used in the spline models, where a separate scale parameter is used for the individual interaction term. The proposed specification is more parsimonious; for d-dimensional covariates, the number of scale parameters is $d + 1$. It is also different from the kernel used in Stitson et al. (1999), which only included the interaction terms of a fixed order, i.e., the three-way interaction model involves only three-way interaction effect terms and not the lower-order terms.

The saturated ANOVA decomposition kernel has 2^d terms, including a constant term, all of the base kernels, the d-th order interaction term and any lower order interaction terms. The kernel given by (3.9) is, in fact, this class of kernel. It can also be seen as a special case of tensor product kernels by treating $\tilde{k}(\mathbf{x}_l, \mathbf{x}'_l) = 1 + k_l(\mathbf{x}_l, \mathbf{x}'_l)$ as one kernel. In contrast to the saturated ANOVA decomposition kernel, a GP model with a tensor product kernel in its simplest form, given by Definition 5 with each k_l not being a sum kernel involving the constant kernel, only includes the highest interaction term. As we will see in the following sections, this may lead to a poor fit and cause difficulty in interpreting the fitted model. The saturated ANOVA kernel, however, assumes the highest complexity given a set of base kernels k_l, which may be an overfit to the data.

To address this issue, we introduce another class of ANOVA decomposition kernels, called the *hierarchical ANOVA decomposition kernel.* This kernel includes a constant term, all main terms, and any interaction terms of any orders constructed using tensor product kernels. This means that no new kernels are introduced for interaction terms, and for given data, all models under this class of kernels share the same hyperparameters. Additionally, this kernel must have a hierarchical structure. If we include any p-th order interaction terms, any lower order interaction terms involving any covariates used in the p-th order interaction terms must also be included. The smallest kernel in this class is the main effect kernel given by (3.7), and the largest kernel is simply the saturated ANOVA decomposition kernel. Figure 3.1 illustrates the differences between the kernels discussed in this section with $d = 4$. It is worth noting that there are many hierarchical ANOVA kernels, and Figure 3.1b shows one such example. We may use a simpler term, the ANOVA kernel, to refer to the ANOVA decomposition kernel.

3.1.4 Separable, sum of separable and non-separable kernels

Structured kernels can be categorized as either “separable” or “non-separable”. A *separable* kernel, denoted as $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, takes the form of a tensor product kernel: $k = \otimes_{l=1}^{d} \tilde{k}_l$, where each component kernel $\tilde{k}_l$ contributing to the construction of k can itself be a combination of multiple kernels, such as a sum or product. For instance, in the saturated ANOVA decomposition kernel, each $\tilde{k}_l = 1 + k_l$ for a given k_l. Essentially,

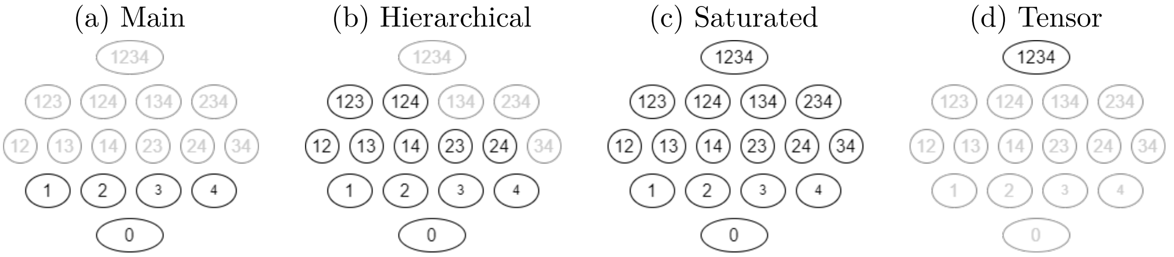

(a) Main (b) Hierarchical (c) Saturated (d) Tensor

Fig. 3.1 Visualisation of ANOVA decomposition kernels (Panel (a),(b) and (c)) and a tensor product kernel (Panel (d)) with $d = 4$ dimensional covariates. The term 0 in the panels refers to the constant term, and the terms 1 to 4 refer to the main effect term that corresponds to k_l for $l = 1, \ldots, 4$. The remaining terms are interaction effect terms, e.g., the term 123 models the three-way interaction effect involving the covariates $\mathbf{x}_1$,$\mathbf{x}_2$ and $\mathbf{x}_3$. Panel (b) is an example of a hierarchical ANOVA decomposition kernel. Adding the term 34 to this example gives another example of such a kernel. If we are to include the term 134 and/or 234, the term 34 should also be added to ensure a hierarchical structure.

separable kernels entail that the tensor product separates the kernel for each dimension l.

A more general class is the "sum of separable" kernel, which has the form $k = \sum_{q=1}^{Q} \otimes_{l=1}^{d} \tilde{k}_{lq}$, where for a given $l \in \{1, \ldots, d\}$, $\tilde{k}_{l1}, \tilde{k}_{l2}, \ldots, \tilde{k}_{lQ}$ are all defined on $\mathcal{X}_l \times \mathcal{X}_l$. Hierarchical ANOVA decomposition kernels are special cases of this class of kernel, where some of $\tilde{k}_{lq}$ is a constant kernel (2.4), with the scale parameter $\alpha = 1$, i.e., $\tilde{k}_{lq}(x_l, x_l') = 1, \forall x_l, x_l' \in \mathcal{X}_l$, and for all non-constant kernels, we have $\tilde{k}_{lq} = \tilde{k}_l$ for $q = 1, \ldots, Q$ and $l = 1, \ldots, d$.

On the other hand, *non-separable* kernels encompass those that do not adhere to the separable structure. For interested readers, see, e.g. Cressie and Huang (1999); Gneiting (2002) for the stationary case, and Fonseca and Steel (2011); Wang et al. (2020) for the non-stationary case.

3.2 Posterior and model comparison

3.2.1 Decomposition of the posterior under additive models

For inferring the posterior distribution or predictive distribution of the overall function f and estimating model hyperparameters, the methods presented in Chapter 2 remain directly applicable. However, if our interest shifts towards the individual component function that constitutes the function f, e.g., (3.4) or (3.5), deriving posterior distributions for these individual functions becomes essential.

In particular, suppose the regression function satisfies $f_{all} = \sum_{j=1}^{J} f_j \sim \mathrm{GP}(0, \sum_{j=1}^{J} k_j)$ where the component functions are uncorrelated, i.e., $f_j \sim \mathrm{GP}(0, k_j)$. Note that the ANOVA decomposition kernels used in this thesis and described in the previous section have this structure. Then the posterior distribution of the jth component is $f_j|y \sim \mathrm{GP}(\bar{m}_j, \bar{k}_j)$, where

$$\bar{m}_j(\mathbf{x}_j) = \mathbf{k}_j(\mathbf{x}_j)^\top(\mathbf{K}_{all} + \sigma^2\mathbf{I}_n)^{-1}\mathbf{y}, \qquad \mathbf{x}_j \in \mathcal{D}_j \tag{3.12}$$

$$\bar{k}_j(\mathbf{x}_j, \mathbf{x}'_j) = \bar{k}_j(\mathbf{x}_j, \mathbf{x}'_j) - \mathbf{k}_j(\mathbf{x}_j)^\top(\mathbf{K}_{all} + \sigma^2\mathbf{I}_n)^{-1}\mathbf{k}_j(\mathbf{x}'_j), \quad \mathbf{x}_j, \mathbf{x}'_j \in \mathcal{D}_j \tag{3.13}$$

for $j = 1, 2..., J$ and J denoting the number of terms. Note that a kernel may have tensor product form, e.g. the j-th term can be $k_j(\mathbf{x}_j, \mathbf{x}'_j) = k_l(\mathbf{x}_l, \mathbf{x}'_l)k_{l'}(\mathbf{x}_{l'}, \mathbf{x}'_{l'})$ with $\mathbf{x}_j = (\mathbf{x}_l, \mathbf{x}_{l'})^\top$, $\mathbf{x}_l \in \mathcal{X}_l$, $\mathbf{x}_{l'} \in \mathcal{X}_{l'}$ and $\mathcal{D}_j = \mathcal{X}_l \times \mathcal{X}_{l'}$.

3.2.2 Comparing models under hierarchical ANOVA kernels

In this section, we focus on model comparison and selection among models with different interactions, i.e., given a set of predictors and kernels, we aim to select a model incorporating the appropriate interaction terms.

Section 2.2.3 and 2.3.3 discuss different options for model comparison and selection in broader contexts, such as marginal likelihood, Watanabe Akaike information criteria (WAIC) and leave-one-out cross-validation (LOO-CV). When comparing models with different interaction structures, we can adhere to the principles outlined in these preceding sections in theory. However, as discussed, computing these quantities can pose a formidable computational challenge associated with integrating out hyperparameters $\boldsymbol{\theta}$, especially in the case of non-Gaussian likelihood. For example, the marginal likelihood for a model $\mathcal{M}$, as given by (2.16),

$$p(\mathbf{y}|\mathbf{X}, \mathcal{M}) = \int p(\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \mathcal{M}) p(\boldsymbol{\theta}|\mathcal{M}) d\boldsymbol{\theta},$$

is typically numerically approximated by e.g. employing bridge sampling (Bennett, 1976; Meng and Schilling, 2002; Meng and Wong, 1996). The choice of hyper-priors on $\boldsymbol{\theta}$ may also play a role, and sensitivity analysis is recommended. The alternative plug-in marginal likelihood, or best-fit predictive density $p(\mathbf{y}|\mathbf{X}, \hat{\boldsymbol{\theta}}, \mathcal{M})$ is much simpler to evaluate but often susceptible to overfitting.

Our parsimonious interaction model specification circumvents this issue. Specifically, irrespective of the order or number of interaction terms within a model, the number of hyperparameters remains constant with the ANOVA kernel construction. Bergsma and Jamil (2023) demonstrate the favourable performance of plug-in marginal likelihood compared to methods such as Lasso, spike and slab prior (Mitchell and Beauchamp, 1988), or g-prior (Zellner, 1986), in selecting the correct interactions within the I-prior framework employing the same ANOVA decomposition kernel specification for interaction terms. I-prior is closely associated with Gaussian process regression, as discussed in Bergsma (2020). The spike and slab approach is utilized in GP regression (Dance and Paige, 2022) for variable selection by placing the spike and slab prior on the length scale parameters of the kernels; however, its applicability is limited to kernels with a length scale parameter and is not directly applicable to the task of selecting interactions.

An additional challenge arises in the case of non-Gaussian likelihood due to the absence of a closed-form expression for the marginal likelihood (even for the plug-in marginal likelihood), the pointwise predictive density for WAIC, or the hold-out predictive density for LOO-CV. If the posterior is approximated via Laplace approximation (LA) or expectation propagation (EP), then it is natural to approximate LOO-CV or marginal likelihood similarly using LA or EP. Nonetheless, the accuracy of such approximations warrants careful investigation. In cases where the chosen approximation is inadequate, it is customary in the literature to compare models using k-fold cross-validation error, employing a suitable scoring rule such as the Brier score (Brier, 1950) for categorical variables. We take this approach in our data example in Section 3.4.1.

3.3 ANOVA kernel, centring and interpretation

As discussed in Plate (1999) and Duvenaud et al. (2011), one effective way to interpret an additive GP model is to visualise each function contributing to the model. If a kernel k_j is centred, the corresponding posterior mean function sums to zero over each input, i.e., $\sum_{i=1}^{n} \bar{m}_j(\mathbf{x}_{j,i}) = 0$. Furthermore, if we denote the posterior mean of the interaction effect term between $\mathbf{x}_l$ and $\mathbf{x}_{l'}$ by $\bar{m}_{ll'}(\mathbf{x}_l, \mathbf{x}_{l'})$ and use centred k_l and $k_{l'}$, we have $\sum_{i=1}^{n} \bar{m}_{ll'}(\mathbf{x}_{li}, \mathbf{x}_{l'}) = \sum_{i=1}^{n} \bar{m}_{ll'}(\mathbf{x}_l, \mathbf{x}_{l'i}) = 0$ for $\mathbf{x}_l \in \mathcal{X}_l$ and $\mathbf{x}_{l'} \in \mathcal{X}_{l'}$. This indicates that all terms in the additive regression function, including main effects and lower-order interaction effects, have intuitive interpretations and can be understood as averaged effects. Consider a regression model for a real-valued response y and two predictors $\mathbf{x}_1 \in \mathcal{X}_1$ and $\mathbf{x}_2 \in \mathcal{X}_2$ as specified by (3.3) and (3.5). Repeating the latter, the regression function is

$$f(\mathbf{x}_{1i}, \mathbf{x}_{2i}) = a + f_1(\mathbf{x}_{1i}) + f_2(\mathbf{x}_{2i}) + f_{12}(\mathbf{x}_{1i}, \mathbf{x}_{2i})$$

for $i = 1, \ldots, n$. The prior over the regression function f is an additive GP with saturated ANOVA decomposition kernel (3.9),

$$k((\mathbf{x}_1, \mathbf{x}_2), (\mathbf{x}_1', \mathbf{x}_2')) = 1 + k_1(\mathbf{x}_1, \mathbf{x}_1') + k_2(\mathbf{x}_2, \mathbf{x}_2') + k_1(\mathbf{x}_1, \mathbf{x}_1')k_2(\mathbf{x}_2, \mathbf{x}_2'),$$

where we assume that each base kernel k_l for $l = 1, 2$ is empirically centred and $\alpha_0 = 1$ for simplicity. We write the Gram matrix $\mathbf{K} = \mathbf{1}_n\mathbf{1}_n^\top + \mathbf{K}_1 + \mathbf{K}_2 + \mathbf{K}_1 \circ \mathbf{K}_2$. Due to centring, we have $\mathbf{K}_1\mathbf{1}_n = \mathbf{K}_2\mathbf{1}_n = \mathbf{0}$. Using (3.12), the posterior mean function is the sum of $\bar{m}_a = \mathbf{1}_n^\top\mathbf{w}$ which corresponds with the constant term, and

$$\begin{aligned} \bar{m}_1(\mathbf{x}_1) &= \mathbf{k}_1(\mathbf{x}_1)^\top\mathbf{w} & \mathbf{x}_1 &\in \mathcal{X}_1 \\ \bar{m}_2(\mathbf{x}_2) &= \mathbf{k}_2(\mathbf{x}_2)^\top\mathbf{w} & \mathbf{x}_2 &\in \mathcal{X}_2 \\ \bar{m}_{12}(\mathbf{x}_1, \mathbf{x}_2) &= \left(\mathbf{k}_1(\mathbf{x}_1) \circ \mathbf{k}_2(\mathbf{x}_2)\right)^\top \mathbf{w} & \mathbf{x}_1 &\in \mathcal{X}_1, \mathbf{x}_2 \in \mathcal{X}_2 \end{aligned}$$

where $\mathbf{w} = (\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{y}$ and $\mathbf{k}_l(\mathbf{x}_l) = (k_l(\mathbf{x}_l, \mathbf{x}_1), \ldots, k_l(\mathbf{x}_l, \mathbf{x}_n))^\top$ for $l = 1, 2$. Each $\bar{m}_j$ corresponds with the term f_j for $j \in \{1, 2, 12\}$.

3.3.1 Properties of posterior mean functions

Centring kernels implies that each of the posterior mean functions for main terms sums to zero over each input, i.e., $\sum_{i=1}^n \bar{m}_1(\mathbf{x}_{1i}) = \sum_{i=1}^n \bar{m}_2(\mathbf{x}_{2i}) = 0$. We can see this by

$$\sum_{i=1}^n \bar{m}_l(\mathbf{x}_{li}) = \underbrace{\mathbf{1}_n^\top\mathbf{K}_l}_{=\mathbf{0}^\top} \mathbf{w} = 0$$

for $l = 1, 2$. For interaction terms, we have a similar property, but the summation is over one input, e.g., $\sum_{i=1}^{n} \bar{m}_{12}(\mathbf{x}_{1i}, \mathbf{x}_2) = 0$. We show this by

$$\begin{aligned}\sum_{i=1}^{n} \bar{m}_{12}(\mathbf{x}_{1i}, \mathbf{x}_2) &= \mathbf{1}_n^\top (\mathbf{K}_1 \bullet \mathbf{k}_2(\mathbf{x}_2))^\top \mathbf{w} = \mathbf{1}_n^\top (\mathbf{k}_2(\mathbf{x}_2) \bullet \mathbf{K}_1)^\top \mathbf{w} \\ &= \mathbf{1}_n^\top (\mathbf{D}_{k_2} \mathbf{K}_1)^\top \mathbf{w} = \underbrace{\mathbf{1}_n^\top \mathbf{K}_1}_{=\mathbf{0}^\top} \mathbf{D}_{k_2} \mathbf{w} = 0\end{aligned}$$

where $\bullet$ is the row-wise Kronecker product (see Appendix B.1) and $\mathbf{D}_{k_2} = \operatorname{diag}(\mathbf{k}_2(\mathbf{x}_2))$. We can show that $\sum_{i=1}^{n} \bar{m}_{12}(\mathbf{x}_1, \mathbf{x}_{2i}) = 0$ in a similar manner. Interestingly, the summation over both inputs does not equal 0 in general, i.e., $\sum_{i=1}^{n} \bar{m}_{12}(\mathbf{x}_{1i}, \mathbf{x}_{2i}) \neq 0$. The exception is when we have multi-dimensional grid data, as we will introduce in Chapter 4. The multi-dimensional grid data can be seen as multi-level panel / balanced longitudinal data, where each unit in the upper level (e.g., individual) has the measurements at the same time points. For a detailed explanation of such data structure, refer to Section 4.2. Let us consider a simple 2-level panel data with $\mathbf{x}_{1i}$ and $\mathbf{x}_{2j}$ representing level-specific covariates where $i = 1, \dots, n_1$ and $j = 1, \dots, n_2$ and assume interaction model. Then we have $\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \bar{m}_{12}(\mathbf{x}_{1i}, \mathbf{x}_{2j}) = 0$ where $\bar{m}_{12}$ is the posterior mean of the interaction term.

3.3.2 Interpretation

These properties of posterior mean functions allow for meaningful interpretation of each main and interaction term. To see this, we start with the interaction term. We can understand how the effect of $\mathbf{x}_1$ changes depending on the level of $\mathbf{x}_2$ by plotting

$$g(\mathbf{x}_1 | \mathbf{x}_2) := \bar{m}_1(\mathbf{x}_1) + \bar{m}_{12}(\mathbf{x}_1, \mathbf{x}_2) \tag{3.14}$$

as a function of $\mathbf{x}_1$ for a fixed value of $\mathbf{x}_2 \in \mathcal{X}_2$. Let us now evaluate this function at each observed value $\mathbf{x}_{2,i}$ for $i = 1, \ldots n$, and take the average. We have that

$$\frac{1}{n}\sum_{i=1}^{n} g(\mathbf{x}_1|\mathbf{x}_{2i}) = \bar{m}_1(\mathbf{x}_1) + \underbrace{\frac{1}{n}\sum_{i=1}^{n} \bar{m}_{12}(\mathbf{x}_1, \mathbf{x}_{2i})}_{=0} = \bar{m}_1(\mathbf{x}_1).$$

This allows for interpreting the main effect term $\bar{m}_1(\mathbf{x}_1)$ in the interaction model as the effect of $\mathbf{x}_1$ averaged over each input of $\mathbf{x}_2$. We can generalise this to a higher-order interaction case. For example, if we have the third predictor $\mathbf{x}_3 \in \mathcal{X}_3$, and include the three-way interaction term in the regression model, we can show that $\sum_{i=1}^{n} \bar{m}_{123}(\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_{3i}) = 0$, for $\mathbf{x}_1 \in \mathcal{X}_1$, $\mathbf{x}_2 \in \mathcal{X}_2$. Then the interpretation of the two-way interaction effect $\bar{m}_{12}(\mathbf{x}_1, \mathbf{x}_2)$ stays the same as the example above (3.14), but now averaged over each input of $\mathbf{x}_3$. The posterior mean of the constant term, $\bar{m}_a$, has an interpretation as the average of the posterior mean of the response $\mathbf{y}$ after averaging over all (main and interaction) effects, under two circumstances: 1) when we have multi-dimensional grid data, or 2) when we have main effects only model with non-structured data.

3.4 Application

In this section, we illustrate the proposed method using two datasets. The first dataset, bovine tuberculosis (BTB) data in Cornwall, UK, is multivariate spatial/spatio-temporal patterns, where the location, time and genotypes responsible for each outbreak of the disease are recorded. For this dataset, we aim to conduct spatio-temporal analysis using the ANOVA decomposition kernel in order to produce a segregation map over space and time. We see that a proper treatment of the main and interaction effects is important. The second data is balanced longitudinal data, where the subjects (cattle) are randomly divided into two treatment groups, and the weight is measured over the course of the study period. We use this dataset to illustrate the interpretability of the proposed ANOVA decomposition kernel. More specifically, we aim to estimate the average growth curve for different treatment groups.

3.4.1 Bovine Tuberculosis in Cornwall

Bovine tuberculosis (BTB) is an infectious disease of cattle caused by Mycobacterium bovis. The disease has been endemic in the UK, especially from the late 1980s to 2010s. In this period of time, there was a widespread in the south-west of England and Wales. The BTB dataset from Diggle et al. (2005) contains 919 confirmed cases of outbreak from 1989 to 2002 in Cornwall, UK. The data has been examined by using non-parametric methods (Diggle et al., 2005) and by using log-Gaussian Cox Process (LGCP) models (Diggle et al., 2013; Taylor et al., 2015). For each observation, the location of the herd where the BTB breakdown was detected, the year of the breakdown, and the genotype that was responsible for the outbreak were recorded. The location information is based on 2-dimensional spatial coordinates. During the period of 14 years, a total of 12 genotypes were identified for 899 cases. We limit the data to the four most prevalent types (genotypes 9, 12, 15 and 20), which totals 873 cases. Figure 3.2 shows the locations of these cases. From the plot, we can observe a spatial pattern of genotypes over the region. For example, more cases are detected in the northeast and southwest of Cornwall, but it is genotype 9 that prevails in the former, while genotypes 15 and 20 are dense in the latter. From Figure 3.3, we notice that the relative prevalence of each genotype changes over time. In this section, we focus on the classification/segregation of different genotypes over space and time.

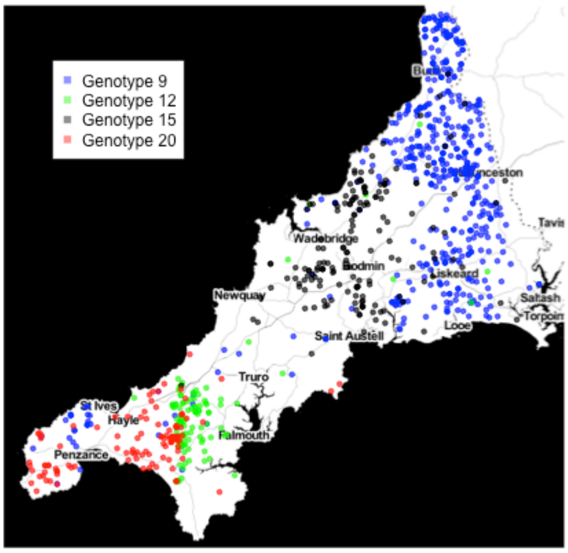


Fig. 3.2 The location of BTB outbreak

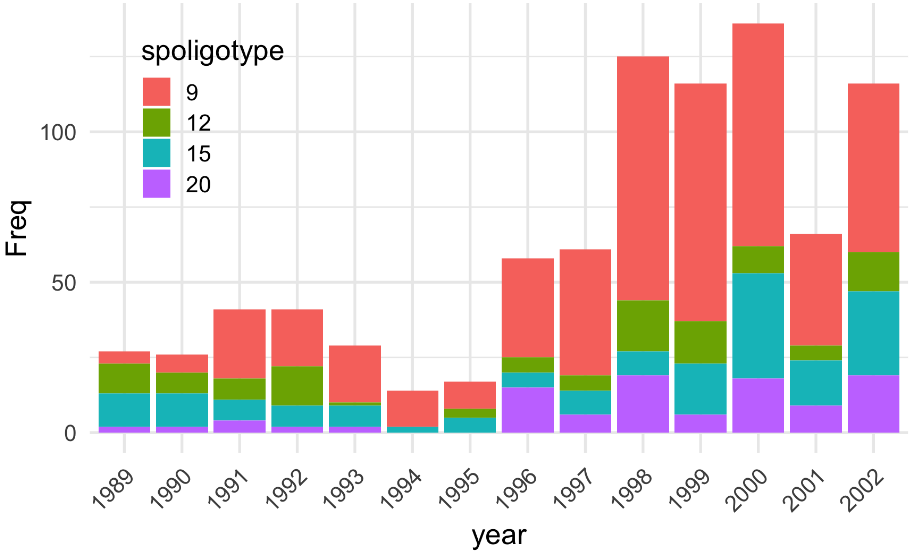


Fig. 3.3 Frequency by genotypes

Model

Let $y_i \in \mathcal{Y}$ be the categorical response variable, recording the genotypes of i-th observation, and $\mathbf{s}_i \in \mathcal{S}, t_i \in \mathcal{T}$ be the location and time of the outbreak of the disease for the i observation for $i = 1, \ldots, 873$. Note $\mathcal{Y} = \{9, 12, 15, 20\}$, $\mathcal{T} \subset \mathbb{R}$ and the location information is given by two-dimensional geographical coordinate, i.e., $\mathcal{S} \subset \mathbb{R}^2$. For the **spatial model**, we assume $y_i \sim \text{Categorical}(\pi_{9,i}, \pi_{12,i}, \pi_{15,i}, \pi_{20,i})$ where the probability $\pi_{j,i} = p(y_i = j)$ for $j \in \mathcal{Y}$ and $i = 1, \ldots, 873$ is given by

$$\pi_{j,i} = \frac{\exp f(j, \mathbf{s}_i)}{\exp \sum_{j' \in \mathcal{Y}} f(j', \mathbf{s}_i)}. \tag{3.15}$$

We put a zero mean GP prior on f and assume no inter-class correlation. See Appendix A.3 for a further explanation of this. Equivalently, we can specify the prior on the vector of $f(j, \mathbf{s}_i)$ given by

$$\mathbf{f} = \left(f(9, \mathbf{s}_1), \ldots, f(9, \mathbf{s}_{873}), \ldots\ldots, f(20, \mathbf{s}_1), \ldots, f(20, \mathbf{s}_{873})\right)^\top.$$

Then we assume $\mathbf{f} \sim \text{MVN}(\mathbf{0}, \mathbf{K})$ where $\mathbf{K}$ is a block diagonal matrix given by $\mathbf{I}_4 \otimes \mathbf{K}_s$. Note $\otimes$ here is the Kronecker product (see Section 4.3) and $\mathbf{K}_s$ is a 873×873 matrix of which the i, i'-th element is given by $\alpha_0^2\,(1 + k(\mathbf{s}_i, \mathbf{s}_i'))$ for a positive constant α_0^2. It is possible to assume different covariance structures for different classes; however, in the absence of prior knowledge, we used the same kernel for all classes. For all $f(j, \mathbf{s}_i)$ to be identifiable, we can e.g., assume $f(j, \mathbf{s}_i) = 0$ for one class, or ensure $\sum_{j \in \mathcal{Y}} f(j, \mathbf{s}_i) = 0$ for $i = 1, \ldots, 873$. Here we take the former approach and set $f(9, \mathbf{s}_i) = 0$.

For the **spatio-temporal models**, we can replace $f(j, \mathbf{s}_i)$ in (3.15) with $f(j, \mathbf{s}_i, t_i)$ to incorporate the temporal variation. The prior on $\mathbf{f}$ is zero-mean multivariate Gaussian with the covariance matrix given by $\mathbf{K} = \mathbf{I}_4 \otimes \mathbf{K}_{st}$. We consider three different structures for $\mathbf{K}_{st}$ corresponding to the main effect, the main effect + spatio-temporal interaction

Model	Description	$\mathbf{K}_{st}$
1	Main effect	$\alpha_0^2(\mathbf{1}\mathbf{1}^\top + \mathbf{K}_s + \mathbf{K}_t)$
2	Main + space-time interaction	$\alpha_0^2(\mathbf{1}\mathbf{1}^\top + \mathbf{K}_s + \mathbf{K}_t + \mathbf{K}_s \circ \mathbf{K}_t)$
3	Space-time interaction only	$\alpha_0^2(\tilde{\mathbf{K}}_s \circ \tilde{\mathbf{K}}_t)$

Table 3.1 The list of spatio-temporal models for BTB dataset and their covariance structures. Note for model 3, we omit the scale parameters of matrix $\mathbf{K}_s$ and $\mathbf{K}_t$ as they are not identifiable. The unscaled matrices are denoted by $\tilde{\mathbf{K}}_s$ and $\tilde{\mathbf{K}}_t$.

effect, and the spatio-temporal interaction effect only model. See Table 3.1 for the list of the model and the structure of $\mathbf{K}_{st}$. Note that for Model 1 and Model 2, we use non-saturated and saturated hierarchical ANOVA decomposition kernels. We look at a 5-fold cross-validation error to compare models with different structures. For this purpose, we used the Automatic Differentiation Variational Inference (ADVI) algorithm in Stan. Once a model is selected, it is re-estimated using MCMC.

We use squared exponential kernel (2.2) for both space k_s and time k_t, hence there are length-scale parameters ρ_s and ρ_t, and scale parameters α_s and α_t respectively. Additionally, all models have α_0 as overall scale parameters. For priors of length-scale parameters, we use Inverse Gamma distribution, which puts negligible mass on values close to zero and has a heavy right tail. In contrast, for the scale parameters, we use half-normal, which puts non-negligible mass around zero.

Result

Table 3.2 shows the misclassification rate and Brier score (Brier, 1950) for each model. Brier score is a strictly proper scoring rule for a categorical response variable, and it is computed by $\frac{1}{873}\sum_{i=1}^{873}\sum_{j\in\mathcal{Y}}(\hat{\pi}_{j,i} - r_{j,i})^2$ where $r_{j,i}$ is the re-coded response variable, which takes the value 1 if $y_i = j$ and 0 otherwise, and $\hat{\pi}_{j,i}$ is the mean of the posterior predictive distribution. Comparing the spatial model and spatio-temporal models (Model 1 and Model 2), we see that incorporating the temporal information improves the prediction accuracy. Furthermore, as Model 2 (main + space-time interaction) has a lower classification error and Brier score, we conclude that the spatial segregation changes over time. Interestingly, Model 3 (the interaction-only model) performs worse than the simple main effect model, which highlights the importance of properly modelling interaction effects in a hierarchical manner. We conclude that the spatio-temporal model with both the main and interaction effects (Model 2) is the best fit for the data. We note, however, that only the squared exponential kernel is considered here, and further analysis of the choices of different kernels should be conducted. We briefly discuss this in Appendix C.1.

Figure 3.4 shows the maps of the posterior (conditional) probabilities, $p(y = j|\mathbf{s}, t)$ for different genotypes and years. It is conditional in the sense that it is the probability of the outcome being in the class j given we observe an event at that point and time. We see the patterns shifting with time. For example, genotype 9, which was mainly prevalent in the western area in the beginning, also has higher conditional probabilities in the central part over the years. For the segregation map, following Diggle et al. (2013), we compute $P_{j,c} = Pr\{p(y = j) > c|\mathbf{s}, t\}$ where $0 < c < 1$ is some threshold. For a chosen c, let $A_{j,q}$ denote a set of locations satisfying $P_{j,c} > q$ where $0 < q < 1$. In this way, the uncertainty of the point prediction can be incorporated, compared to plotting the set satisfying $\mathbb{E}[p(y = j)|\mathbf{s}, t] > q$ for some q. Figure 3.5 shows $A_{j,q}$ for

Table 3.2 The 5-fold CV errors for spatial and spatio-temporal GP models

Model	kernel	Misclassification rate	Brier score
Spatial			
	SE	0.142	0.224
Spatio-temporal			
Model 1	1+ SE + SE	0.135	0.214
Model 2	(1+SE)(1+SE)	0.12	0.204
Model 3	SE*SE	0.434	0.615

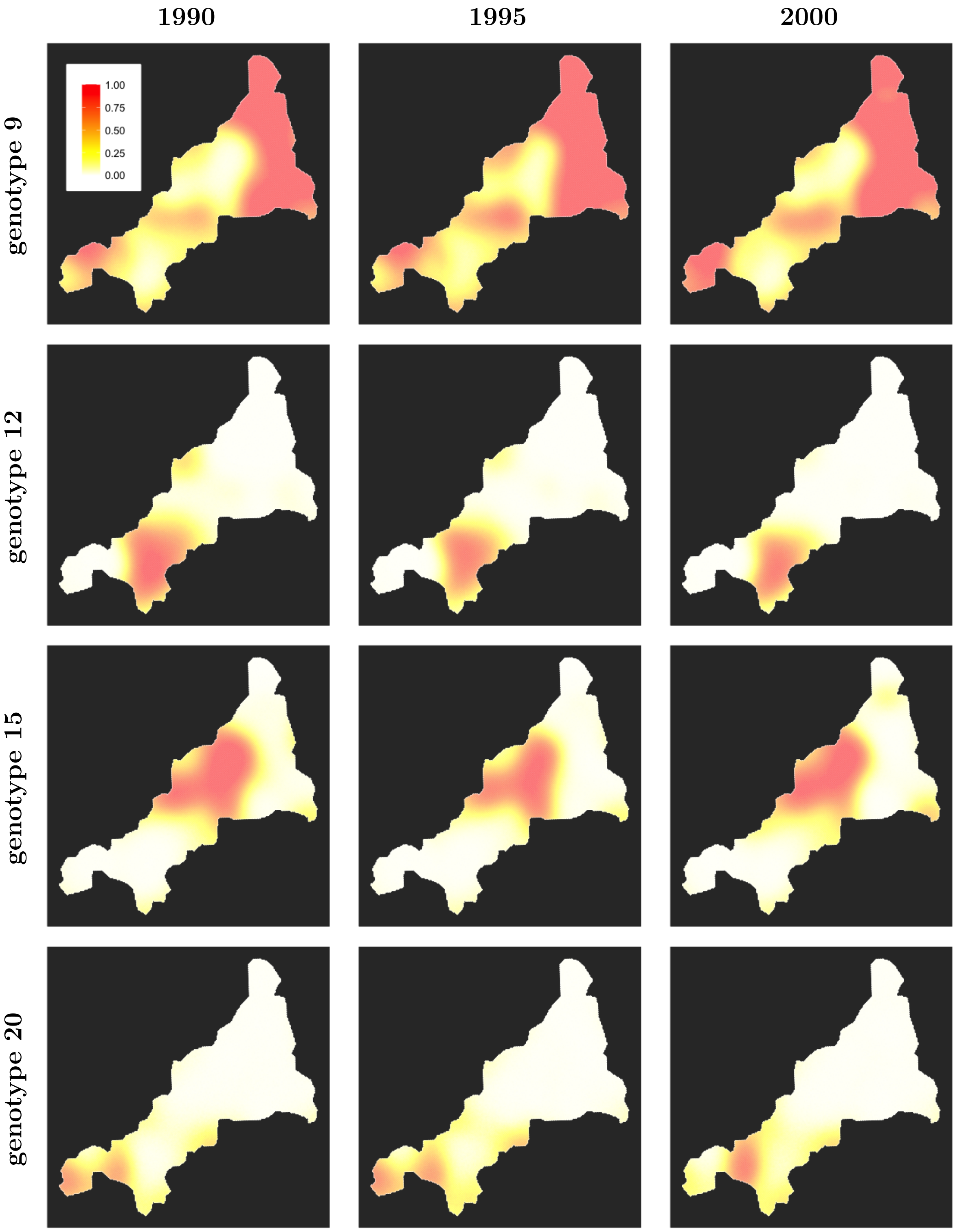
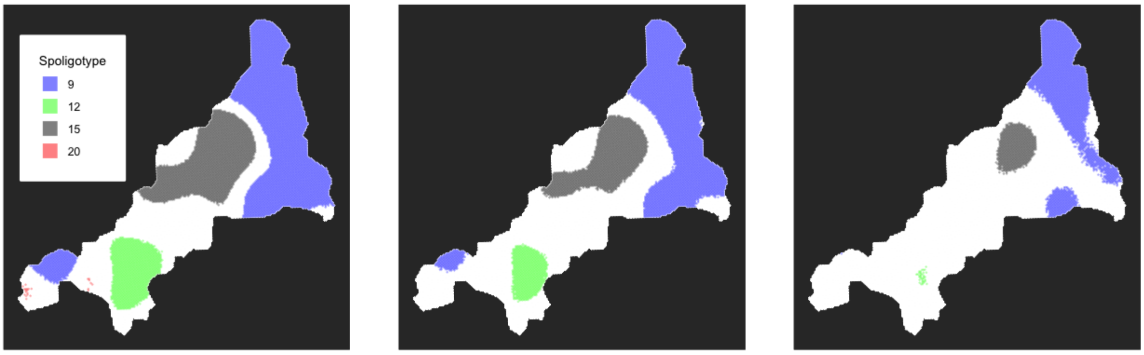
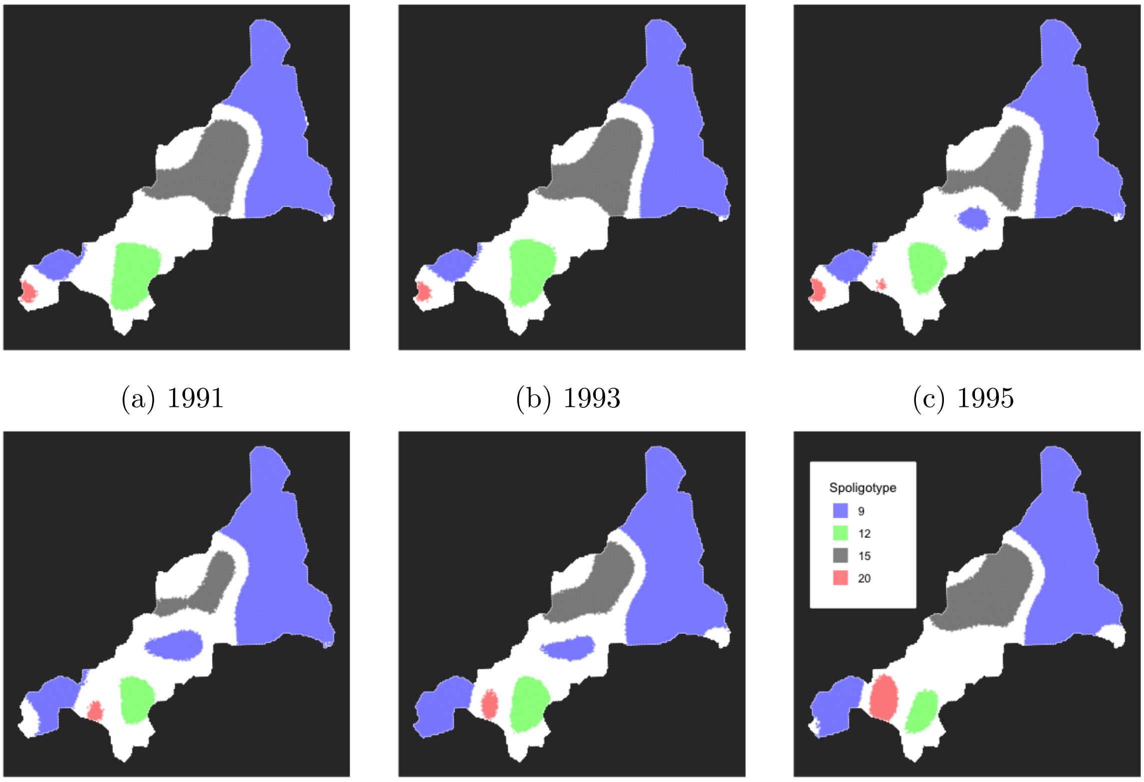


Fig. 3.4 Maps of conditional probabilities $\hat{\pi}_{j,i} = \mathbb{E}[p(y_i = j|\mathbf{s}, t)]$ for different genotypes j and time t.



(a) $q = 0.5$ (b) $q = 0.7$ (c) $q = 0.9$

Fig. 3.5 Segregation map for $t = 1989$. We plot a set of locations $\sim$ that satisfy $Pr(p(y = j) > 0.8|\mathbf{s}, t = 1989) > q$. The value of q is $0.5, 0.7$ and 0.9 respectively.



(a) 1991 (b) 1993 (c) 1995

(d) 1997 (e) 1999 (f) 2001

Fig. 3.6 Segregation map for different years. Each colour represents different genotypes and the coloured area is a set of location for which $Pr(p(y = j) > 0.8|\mathbf{s}, t) > 0.5$

$q = 0.5, 0.7, 0.9$. We see in the left panel of Figure 3.5 that genotype 9 is dominant in one large area in the east and in one small area in the west. Genotype 12 and 15 are dominant in the upper-middle and lower-middle part of the region, while only a very small area where genotype 20 is dominant is seen. These dominant areas become smaller with a higher threshold q, but genotype 9 is still dominant in the east area at $q = 0.9$. We can also observe how this segregation pattern changes over time in Figure 3.6. In these maps, we set $c = 0.8$ and $q = 0.5$. An example of a noticeable change is with genotype 9, which stays dominant in the east and in a small area in the west, but with additional area appearing in the central part of the region from the mid-1990s.

3.4.2 Longitudinal data analysis

The second dataset we analyse is balanced longitudinal data consisting of repeated measurements of weights from 60 cattle. The cattle were randomly assigned to either Treatment Group A or Treatment Group B. Each group is of size 30. For each cattle, the weight is measured 11 times over the course of 133 days. The data is analysed in Kenward (1987) and is available in the R package `jmcm` (Pan and Pan, 2017). The model formulation we used for this analysis has similarity to Cheng et al. (2019), but some notable differences are the inclusion of the (global) constant term and the centring of kernels.

Given the available information (treatment group, cattle identification number, and measurement time), we consider the following model:

$$y_{i,j,k} = f(t_i, j, k) + \epsilon_{i,j,k} \tag{3.16}$$

where t_i is the day that the measurement is taken, j is the indicator of the treatment group, and k is the id of each cattle. Therefore, $y_{i,j,k}$ represents the weight of cattle k which belongs to Treatment Group j, taken at time t_i. We assume an iid error. As the presence of three-way interaction was confirmed in Jamil (2018, Chapter 4), in this section, we focus on the interpretation of the results. With the three-way interaction model, the function f above has the additive structure:

$$\begin{aligned} f(t_i, j, k) &= \alpha + f_1(t_i) + f_2(j) + f_3(k) \\ &\qquad + f_{12}(t_i, j) + f_{13}(t_i, k) + f_{23}(j, k) + f_{123}(t_i, j, k). \end{aligned} \tag{3.17}$$

Following Jamil (2018, Chapter 4), we specify the prior on f as $GP(0, k)$ where k is given by the saturated ANOVA decomposition kernel $\alpha_0^2((1 + k_1) \otimes (1 + k_2) \otimes (1 + k_3))$. A squared-centred standard BM kernel is employed for k_1, while the centred categorical kernel (2.5) is utilised for both k_2 and k_3. All three kernels are empirically centred. This model postulates the existence of both a treatment effect and random effect on cattle growth, with these effects interacting with each other; for instance, the treatment effect may vary over time and differ across individual cattle.

The model includes four hyperparameters: an overall scale parameter α_0 alongside a scale parameter for each k_l where $l = 1, 2, 3$. We adopt a half-normal distribution as the prior distribution for these parameters. It is important to note that the main aim of this section is to illustrate the interpretability of the proposed additive interaction GP model. In many real-world data applications, careful consideration of the selection of kernels and priors is imperative.

Figure 3.7 shows the posterior mean of the growth curve from the specified regression model. We see that the curve is different for different cattle; however, it is difficult to generalise how the growth curve differs in the treatment group on average. Additive GP models allow for the decomposition of posterior mean as discussed in Section 3.2.1, and with the use of ANOVA decomposition kernel, and centring, the main effect and two-way interaction effect can be interpreted as "average" effects. Figure 3.8a shows the posterior mean of the main effect term $f_1(t)$, denoted by $\bar{m}_1(t)$. This can be seen

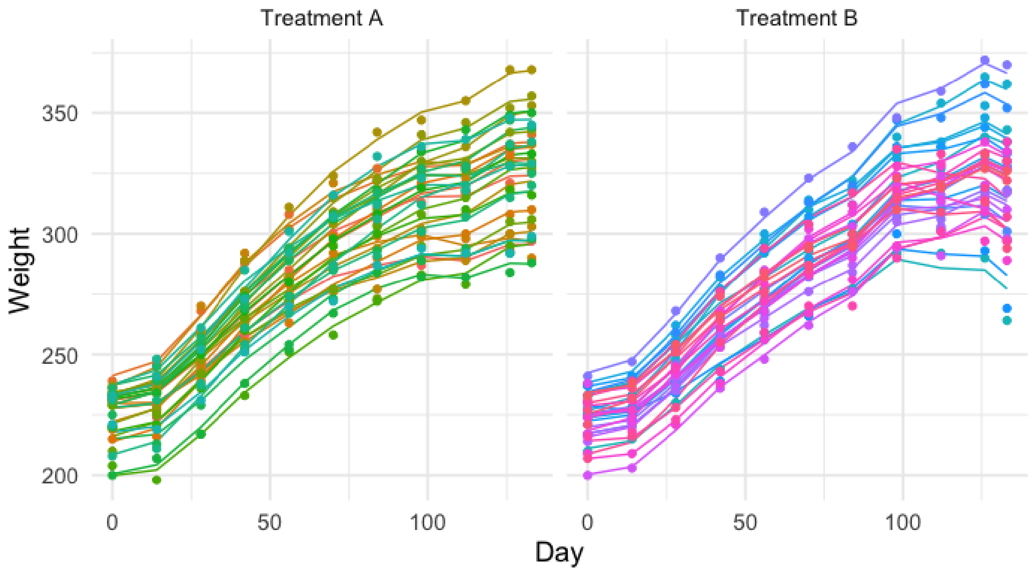
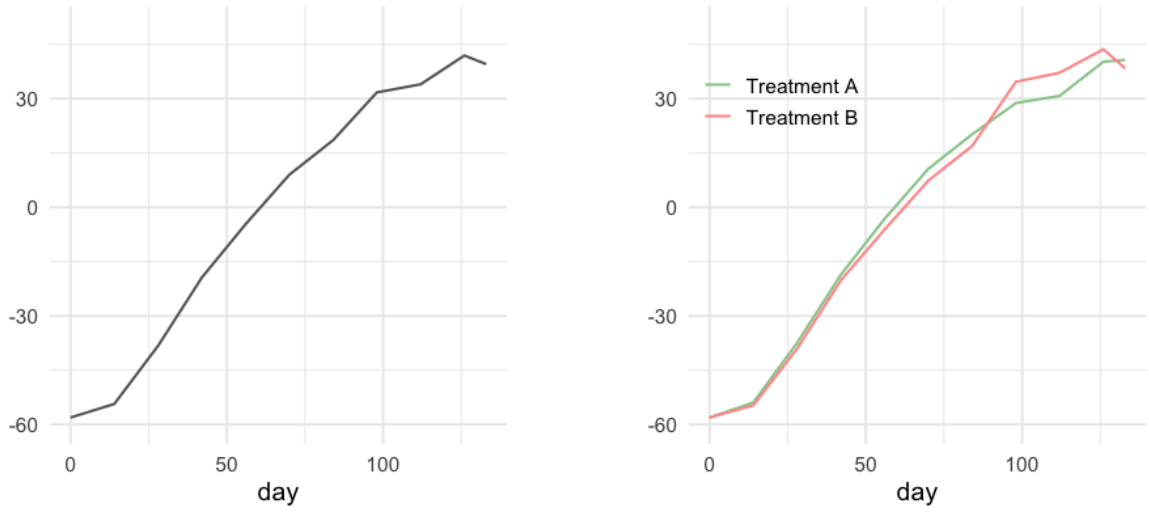


Fig. 3.7 The observed and fitted growth curve of 60 cattle by treatment group

as the growth curve (after centring the response) or the effect of *time*, averaged over the treatment effect and the random effect of the individual cattle. The regression model assumes that the growth of cattle depends on the treatment group. Figure 3.8b illustrates the interaction effect between *time* and *treatment group*. This is the plot of $\bar{m}_1(t) + \bar{m}_{12}(t, j)$ for $j = \{\text{Treatment A}, \text{Treatment B}\}$, which can be interpreted as the average growth curve for the two treatment groups. These curves are after the individual cattle variability is taken into account. We see that the growth curves are similar in the beginning, with group A, on average, growing quicker. This changes at a later stage when the growth curve for Group B surpasses that of Group A. The curve for Group B also captures the weight decrease that was observed for many cattle belonging to the group. Other main effects and interaction effects can also be visualised and interpreted in the same manner. We include them in Appendix C.2



(a) Main effect: $\bar{m}_1(t)$

(b) Inteaction effect: $\bar{m}_1(t) + \bar{m}_{12}(t, j)$

Fig. 3.8 Average centred growth curve

3.5 Discussion

This section explored statistical modelling with a GP prior. We particularly focused on modelling interaction effects of different degrees. We introduced Additive GP models with hierarchical ANOVA decomposition kernels for this purpose. Not only does this approach comply with standard practice in statistical modelling, such as hierarchical inclusion interaction effects, but it also has some attractive properties, such as intuitive interpretation of main and lower-order interaction effects in the presence of higher-order interaction effects. Additionally, ANOVA decomposition kernels offer a parsimonious model in the sense that, given a set of predictors, higher-order interaction models share the same number (and set) of model parameters, which facilitates model selection and identification of appropriate interaction effects. However, the computational cost still remains the main hurdle. The next chapter addresses this issue for large datasets with special structure applicable, for example, to balanced longitudinal data considered in this chapter.

Chapter 4

Kronecker method for additive Gaussian process models

In this chapter, we address the computational challenges posed by Gaussian Process (GP) regression. The primary challenge, as we will see in Section 4.1, revolves around the demanding nature of evaluating, storing, and operating on the Gram matrix of order n, where n represents the sample size. Implementing GP regression involves $O(n^3)$ operations and $O(n^2)$ storage, making it increasingly impractical as n grows. We address this issue specifically for datasets exhibiting a particular structure referred to as a *multidimensional grid* (Section 4.2). When such a structured dataset is at hand, the Gram matrix can be efficiently represented using the Kronecker product (Section 4.3). Section 4.4 then delves into how we exploit the Kronecker product and its associated properties to facilitate the evaluation of each component essential for GP model implementation. To illustrate the efficacy of this approach, we present a practical demonstration using large-scale real-world data.

4.1 Computational challenge

Let us revisit the regression model with GP priors, considered in the previous chapters. Here, we denote the response by $y \in \mathbb{R}$, the (set of) covariates by $\mathbf{x} \in \mathcal{X}$ and the sample $(y_i, \mathbf{x}_i)_{i=1}^n$ with $\mathbf{y} = (y_1, \ldots, y_n)^\top$. The evaluation of the log marginal likelihood (2.15), the posterior mean (2.12) and covariance (2.13) involves the $n \times n$ marginal covariance matrix $\mathbf{K} + \sigma^2\mathbf{I}$. The main bottlenecks are:

1. Matrix-vector multiplication involving inverse of the Gram matrix,

$$\left(\mathbf{K} + \sigma^2\mathbf{I}\right)^{-1}\mathbf{v}$$

   where $\mathbf{v} = \mathbf{y}$ for the log marginal likelihood and $\mathbf{v} = \mathbf{k}(\mathbf{x}) = (k(\mathbf{x}, \mathbf{x}_1), \ldots, k(\mathbf{x}, \mathbf{x}_n))^\top$ for the posterior mean and kernel; and

2. Log-determinant of the Gram matrix,

$$\log|\mathbf{K} + \sigma^2\mathbf{I}|.$$

Taking the inverse of a $n \times n$ matrix and computing its determinant have cubic time complexity if naively done. Furthermore, matrix-vector multiplication between the resulting inverted matrix and a vector has $O(n^2)$ scaling. Storing the (Gram) matrix of order n has $O(n^2)$ memory requirement.

4.2 Multidimensional grid structure

While there are various approaches to reduce the computational burden of GP models (see Liu et al. (2020) for an overview), we focus on a method that is applicable to data with multidimensional grid structure. Multidimensional grid data, which can also be

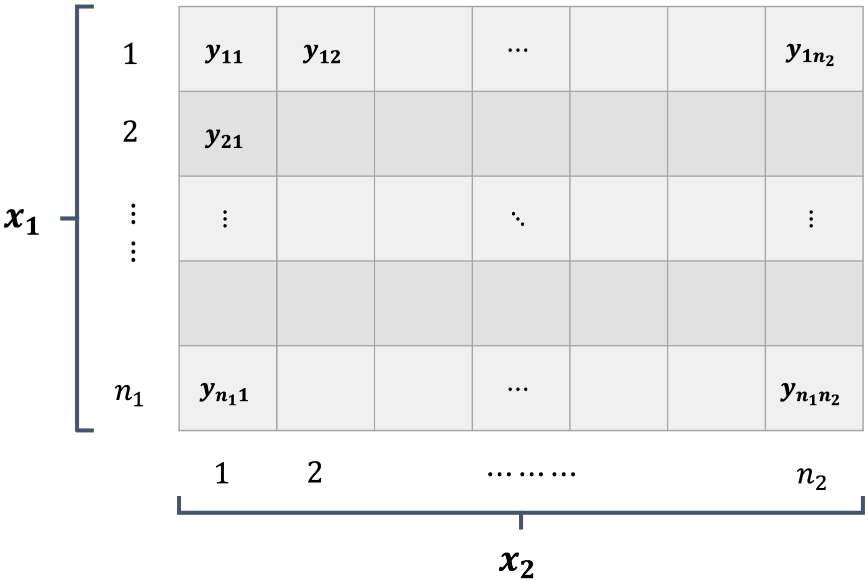


Fig. 4.1 Illustration of two-dimensional grid

seen as a multi-level panel or balanced longitudinal data, is common in image analysis, spatial and spatio-temporal data, and repeated measurement in medical or behavioural science.

Formally we say that the data has a multi-dimensional grid structure when the predictors form a *Cartesian grid*,

$$\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_d$$

where $\mathcal{X}_l$ represents a set of observed values for the input dimension l and $\times$ is the Cartesian product. We let n_l represent the cardinality of set $\mathcal{X}_l$, i.e., $n_l = |\mathcal{X}_l|$. The total number of observations is therefore $n = \prod_{l=1}^{d} n_l$.

Figure 4.1 illustrates two-dimensional grid data. In image analysis, the pair $\mathbf{x}_1, \mathbf{x}_2$ typically represents the $X - Y$ coordinate of a pixel, while with spatial analysis of e.g. meteorological measurements over a rectangular region, it may be a set of geographical coordinates. In environmental monitoring, it is common that the measurements are taken from multiple monitoring stations over some periods. Then $\mathbf{x}_1$ denotes the

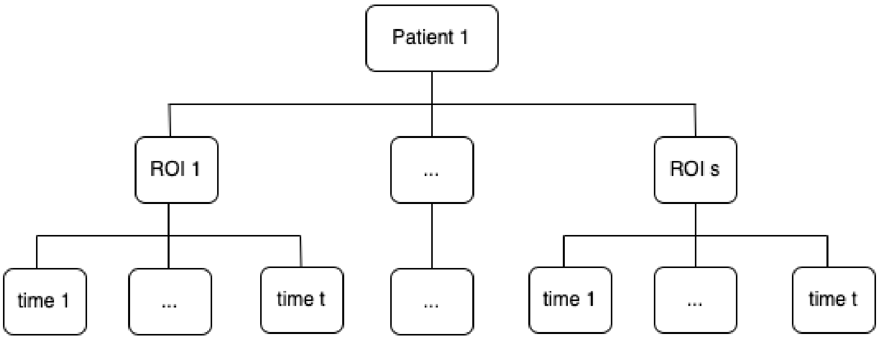


Fig. 4.2 Example of three-dimensional grid

geographical coordinate of a monitoring station, and $\mathbf{x}_2$ is a timestamp of when each measurement is recorded, with $\mathcal{X}_1 \subset \mathbb{R}^2$, a set of coordinates for all monitoring stations, and $\mathcal{X}_2 \subset \mathbb{R}$, a set of timestamps.

For higher dimensions, we give an example in brain imaging (Figure 4.2), such as functional Magnetic Resonance Imaging (fMRI) studies. The dataset for each patient may consist of measurements at different Region of Interest (ROI) recorded for a certain period of time. This constitutes a three-dimensional grid as illustrated in Figure 4.3.

When a multidimensional grid structure is present in the data, the computations of the key components listed in 4.1 can be made efficient using Kronecker methods. In the following section, we introduce the Kronecker product and its key properties.

4.3 Kronecker product

Consider two matrices $\mathbf{A} = \{a_{i,j}\}_{1\leq i\leq n, 1\leq j\leq m}$ and $\mathbf{B} = \{b_{i,j}\}_{1\leq i\leq p, 1\leq j\leq q}$. The Kronecker product of the two matrices, $\mathbf{A} \otimes \mathbf{B}$, is the matrix of size $np \times mq$ given by

$$\mathbf{A} \otimes \mathbf{B} = \begin{bmatrix} a_{1,1}\mathbf{B} & \dots & a_{1,m}\mathbf{B} \\ \vdots & \ddots & \vdots \\ a_{n,1}\mathbf{B} & \dots & a_{n,m}\mathbf{B}. \end{bmatrix}$$

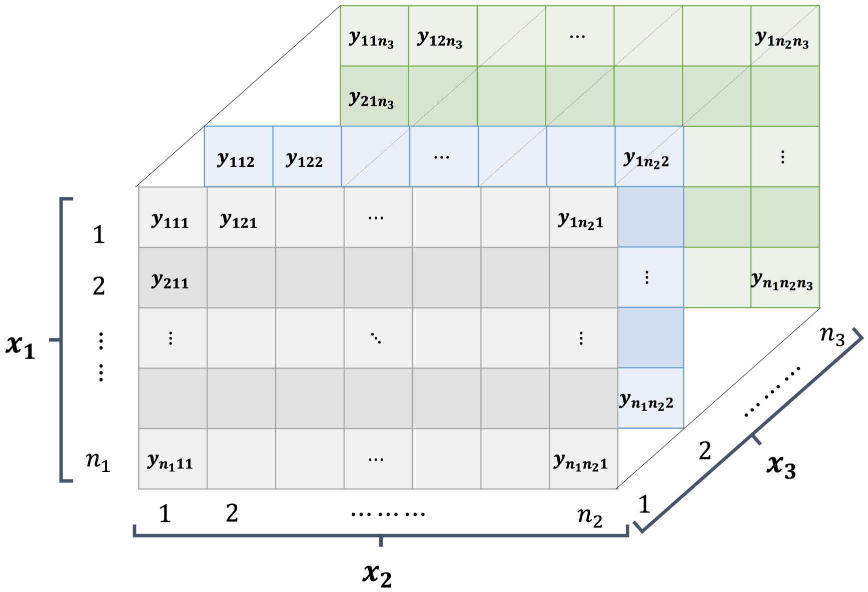


Fig. 4.3 Illustration of three-dimensional grid

More generally, we denote the Kronecker product of $d \geq 2$ matrices, $\mathbf{A}_l$ where $l = 1, \ldots d$ by

$$\mathbf{A} = \bigotimes_{l=1}^{d} \mathbf{A}_l = \mathbf{A}_1 \otimes \mathbf{A}_2 \otimes \ldots \otimes \mathbf{A}_d$$

If each matrix $\mathbf{A}_l$ is size $n_l \times m_l$, the resulting Kronecker product matrix $\mathbf{A}$ has size $\prod_{l=1}^{d} n_l \times \prod_{l=1}^{D} m_l$.

4.3.1 Kronecker product properties

We list some of the properties of Kronecker product that we use in this paper. In addition to $\mathbf{A}_l$ defined above, let us assume we have, for $l = 1, \ldots, d$, $\mathbf{B}_l$ of size $p_l \times q_l$, $\mathbf{B}'_l$ of size $p_l \times q_l$, $\mathbf{C}_l$ of size $h_l \times k_l$ and $\mathbf{D}_l$ of size $m_l \times p_l$. The size of matrices is given so that the operations $\mathbf{B}_l + \mathbf{B}'_l$ and $\mathbf{A}_l\mathbf{D}_l\mathbf{B}_l$ are allowed.

1. Bilinearity:

$$\mathbf{A}_l \otimes (\mathbf{B}_l + \mathbf{B}'_l) = \mathbf{A}_l \otimes \mathbf{B}_l + \mathbf{A}_l \otimes \mathbf{B}'_l$$

2. Associativity:

$$\mathbf{A}_l \otimes (\mathbf{B}_l \otimes \mathbf{C}_l) = (\mathbf{A}_l \otimes \mathbf{B}_l) \otimes \mathbf{C}_l$$

$$\alpha(\mathbf{A}_l \otimes \mathbf{B}_l) = (\alpha \mathbf{A}_l) \otimes \mathbf{B}_l = \mathbf{A}_l \otimes (\alpha \mathbf{B}_l)$$

   where α is a scalar.

3. Transpose:

$$\left(\bigotimes_{l=1}^{d} \mathbf{A}_l \right)^{\top} = \bigotimes_{d=l}^{d} \mathbf{A}_l^{\top} \tag{4.1}$$

4. Inverse:

$$\left(\bigotimes_{l=1}^{d} \mathbf{A}_l \right)^{-1} = \bigotimes_{d=l}^{d} \mathbf{A}_l^{-1}$$

5. The mixed product properties:

$$\bigotimes_{l=1}^{d} (\mathbf{A}_l \mathbf{D}_l) = \left(\bigotimes_{l=1}^{d} \mathbf{A}_l \right) \left(\bigotimes_{l=1}^{d} \mathbf{D}_l \right) \tag{4.2}$$

6. Matrix vector product

$$(\mathbf{A}_l \otimes \mathbf{B}_l)\, \mathbf{v} = \text{vec}\left(\mathbf{B}_l \mathbf{V} \mathbf{A}_l^{\top} \right) \tag{4.3}$$

   where $\mathbf{V} = \text{vec}^{-1}(\mathbf{v})$ is the inverse of the vectorization operator and $\mathbf{v}$ is a vector of length $m_l q_l$.

For the proceeding sections, it becomes useful to consider a generalisation of (4.2). For example, for triplets of matrices $\mathbf{A}_l, \mathbf{D}_l, \mathbf{B}_l$ we have the following:

$$\bigotimes_{l=1}^{d} (\mathbf{A}_l \mathbf{D}_l \mathbf{B}_l) = \left(\bigotimes_{l=1}^{d} \mathbf{A}_l \right) \left(\bigotimes_{l=1}^{d} \mathbf{D}_l \right) \left(\bigotimes_{l=1}^{d} \mathbf{B}_l \right). \tag{4.4}$$

The property on the matrix-vector product provides a computational advantage. Let us assume matrices $\mathbf{A}_l$ and $\mathbf{B}_l$ involved in (4.3) are a squared matrix of order n_1 and n_2. Naively evaluating the left-hand side of (4.3) takes $(n_1 n_2)^2$ operations, while exploiting the right-hand side reduces the cost to $(n_1 n_2)(n_1 + n_2) \leq (n_1 n_2)^2$. We extend this property to evaluate e.g., $(\bigotimes_{l=1}^{d} \mathbf{A}_l)\mathbf{v}$ where $\mathbf{v}$ is a vector of an appropriate length, in Section 4.4.1.

4.3.2 Eigendecomposition of a matrix with Kronecker product

A key operation for the efficient methods proposed in this section is the eigendecomposition of a Gram matrix involving the Kronecker product. Let $\mathbf{A}_l$ be a $n_l \times n_l$ diagonalizable matrix for $l = 1, \dots, d$, and

$$\mathbf{A} = \bigotimes_{l=1}^{d} \mathbf{A}_l.$$

We write the eigendecomposition of a matrix $\mathbf{A}_l$ by $\mathbf{A}_l = \mathbf{Q}_l \mathbf{\Lambda}_l \mathbf{Q}_l^\top$, where $\mathbf{\Lambda}_l$ is a diagonal matrix of which the diagonal elements are eigenvalues of $\mathbf{A}_l$ in non-decreasing order, which we denote by $\boldsymbol{\lambda}_l = (\lambda_{1,l}, \dots, \lambda_{n_l,l})^\top$, and $\mathbf{Q}_l$ is an orthonormal matrix with its i-th column $\mathbf{q}_i$ being the eigenvector which corresponds to the i-th eigenvalue. Using the mixed product properties of the Kronecker product (4.4), the eigendecomposition of the matrix $\mathbf{A}$ is the following:

$$\begin{aligned} \mathbf{A} &= \bigotimes_{l=1}^{d} \left(\mathbf{Q}_l \mathbf{\Lambda}_l \mathbf{Q}_l^\top \right) \\ &= \bigotimes_{l=1}^{d} \mathbf{Q}_l \bigotimes_{l=1}^{d} \mathbf{\Lambda}_l \bigotimes_{l=1}^{d} \mathbf{Q}_l^\top \\ &= \left(\bigotimes_{l=1}^{d} \mathbf{Q}_l \right) \left(\bigotimes_{l=1}^{d} \mathbf{\Lambda}_l \right) \left(\bigotimes_{l=1}^{d} \mathbf{Q}_l \right)^\top \qquad (4.5) \end{aligned}$$

From the second to the third line, we used the property of Kronecker product involving matrix transpose (4.1). Let $\mathbf{Q} \equiv \bigotimes_{l=1}^{d} \mathbf{Q}_l$ and $\mathbf{\Lambda} \equiv \bigotimes_{l=1}^{d} \mathbf{\Lambda}_l$. Note that $\mathbf{Q}$ is orthonormal, i.e., $\mathbf{Q}\mathbf{Q}^\top = \mathbf{I}_n$. We can confirm this by

$$\begin{aligned}
\mathbf{Q}\mathbf{Q}^\top &= \left(\bigotimes_{l=1}^{d} \mathbf{Q}_l\right)\left(\bigotimes_{l=1}^{d} \mathbf{Q}_l\right)^\top \\
&= \left(\bigotimes_{l=1}^{d} \mathbf{Q}_l\right)\left(\bigotimes_{l=1}^{d} \mathbf{Q}_l^\top\right) \\
&= \bigotimes_{l=1}^{d} \mathbf{Q}_l\mathbf{Q}_l^\top = \bigotimes_{l=1}^{d} \mathbf{I}_{n_l} = \mathbf{I}_n.
\end{aligned}$$

Furthermore, the matrix $\mathbf{\Lambda}$ is a diagonal matrix, whose diagonal elements are given by the Kronecker product of the eigenvalue vectors $\boldsymbol{\lambda}_l$. We have:

$$\begin{aligned}
\mathbf{\Lambda} &= \bigotimes_{l=1}^{d} \mathbf{\Lambda}_l \\
&= \bigotimes_{l=1}^{d} (\operatorname{diag}(\boldsymbol{\lambda}_l)) \\
&= \operatorname{diag}\left(\bigotimes_{l=1}^{d} \boldsymbol{\lambda}_l\right). \qquad (4.6)
\end{aligned}$$

We use these results in the following sections for the case where $\mathbf{A}_l$ is a Gram matrix.

4.4 Efficient computation using Kronecker method

The primary component of this scalable method is a Kronecker product structure in a Gram matrix, applicable when the predictors form a multidimensional grid. To give an example, let us revisit the models with two predictors considered in Chapter 3. If the two predictors form a two-dimensional grid, we can rewrite the model equation (3.3) as

$$y_{i,j} = f(\mathbf{x}_{1i}, \mathbf{x}_{2j}) + \epsilon_{i,j} \qquad (4.7)$$

where $y_{i,j}$ is the response from the i,j-th location of the grid with $i = 1, \ldots, n_1$, $j = 1, \ldots, n_2$ and $n = n_1 n_2$. We write $\mathbf{y} = (y_{1,1}, \ldots, y_{1,n_2}, \ldots, y_{n_1,1} \ldots, y_{n_1,n_2})$ and define $\boldsymbol{\epsilon}$ similarly. The main effect and interaction effect models in (3.4) and (3.5) then can be specified by

$$f(\mathbf{x}_{1i}, \mathbf{x}_{2j}) = a + f_1(\mathbf{x}_{1i}) + f_2(\mathbf{x}_{1j}) \tag{4.8}$$

$$f(\mathbf{x}_{1i}, \mathbf{x}_{2j}) = a + f_1(\mathbf{x}_{1i}) + f_2(\mathbf{x}_{1j}) + f_{12}(\mathbf{x}_{1i}, \mathbf{x}_{2j}). \tag{4.9}$$

If we use the same prior as the previous model with $\alpha_0 = 1$, we have $\mathbf{y}|\mathbf{X} \sim \text{MVN}_n(\mathbf{0}, \mathbf{K} + \sigma^2 \mathbf{I}_n)$ with the Gram matrix for each model given by

$$\mathbf{K} = \mathbf{1}_{n_1}\mathbf{1}_{n_1}^\top \otimes \mathbf{1}_{n_2}\mathbf{1}_{n_2}^\top + \mathbf{K}_1 \otimes \mathbf{1}_{n_2}\mathbf{1}_{n_2}^\top + \mathbf{1}_{n_1}\mathbf{1}_{n_1}^\top \otimes \mathbf{K}_2. \tag{4.10}$$

$$\mathbf{K} = (\mathbf{1}_{n_1}\mathbf{1}_{n_1}^\top + \mathbf{K}_1) \otimes (\mathbf{1}_{n_2}\mathbf{1}_{n_2}^\top + \mathbf{K}_2)$$

where $\otimes$ is a Kronecker product operator. We relax the assumption on α_0 later in section 4.4.3. In what follows, we show how we can exploit this structured Gram matrix for efficient computation. In Section 4.4.1, we outline the general procedures of the Kronecker approach, with the key operation being the eigendecomposition of the Gram matrix. We illustrate how this can be achieved for models employing saturated and non-saturated hierarchical ANOVA kernels in Sections 4.4.2 and 4.4.3, respectively.

4.4.1 General procedure of Kronecker approach

For d dimensional grid data, the main idea of Kronecker methods is to decompose the Gram matrix in the form:

$$\mathbf{K} = \left(\bigotimes_{l=1}^{d} \mathbf{Q}_l \right) \mathbf{D} \left(\bigotimes_{l=1}^{d} \mathbf{Q}_l \right)^\top \tag{4.11}$$

where each $\mathbf{Q}_l$, and hence also $\mathbf{Q} := \bigotimes_{l=1}^{d} \mathbf{Q}_l$, is orthonormal and $\mathbf{D}$ is diagonal with non-negative entries. Once we obtain this decomposition, the log determinant of the marginal covariance matrix can be computed by

$$\log|\mathbf{K} + \sigma^2 \mathbf{I}| = \sum_{i=1}^{n} \log\Big(\mathbf{D}_{i,i} + \sigma^2\Big)$$

where $\mathbf{D}_{i,i}$ is the i-th diagonal element of $\mathbf{D}$. This costs $O(n)$ operations. The multiplication of the inverted matrix and a vector $\mathbf{v}$ can be expressed as

$$\Big(\mathbf{K} + \sigma^2 \mathbf{I}_n\Big)^{-1} \mathbf{v} = \left(\bigotimes_{l=1}^{d} \mathbf{Q}_l\right) \Big(\mathbf{D} + \sigma^2 \mathbf{I}_n\Big)^{-1} \left(\bigotimes_{l=1}^{d} \mathbf{Q}_l\right)^{\top} \mathbf{v}. \tag{4.12}$$

The inversion of the middle diagonal matrix can be done by simply inverting its diagonal elements. Evaluating the above also requires matrix-vector multiplication $(\bigotimes_{l=1}^{d} \mathbf{Q}_l)^{\top}\mathbf{v}$. Let us rewrite this expression by

$$\left(\bigotimes_{l=1}^{d} \mathbf{Q}_l\right)^{\top} \mathbf{v} = \left(\bigotimes_{l=1}^{d} \mathbf{Q}_l^{\top}\right) \mathbf{v} = \left(\bigotimes_{l=1}^{d-1} \mathbf{Q}_l^{\top}\right) \mathbf{v}_d \tag{4.13}$$

where $\mathbf{v}_d = \mathrm{vec}(\mathbf{Q}_d^{\top}\mathbf{V})$ with $\mathrm{vec}(\mathbf{A})$ being a vectorisation operator transforming a $p \times q$ matrix $\mathbf{A}$ to a vector of length pq by stacking the columns of the matrix, and $\mathbf{V}$ is a $n_d \times \frac{n}{n_d}$ matrix whose elements are filled with elements of vector $\mathbf{v}$ in column-major order. Computing $\mathbf{v}_d$ takes $O(n_l^2 \frac{n}{n_l}) = O(nn_l)$. Iteratively applying this to get $\mathbf{v}_{d-1} = \mathrm{vec}(\mathbf{Q}_{d-1}^{\top}\mathbf{V}_d), \mathbf{v}_{d-2} = \mathrm{vec}(\mathbf{Q}_{d-2}^{\top}\mathbf{V}_{d-1}), \ldots \mathbf{v}_1 = \mathrm{vec}(\mathbf{Q}_1^{\top}\mathbf{V}_2)$ thus requires $O(n\sum_{l=1}^{d} n_l)$ operations, and the final vector $\mathbf{v}_1$ equals $(\bigotimes_{l=1}^{d} \mathbf{Q}_l)^{\top}\mathbf{v}$. The complete algorithm is described in Saatçi (2012, Chapter 5) and Wilson et al. (2014).

The Kronecker method has been used and proven useful for an efficient implementation of GP models; see e.g. Saatçi (2012, Chapter 5), Groot et al. (2014) Wilson et al. (2014) and Flaxman et al. (2015). However, the existing method is only applicable to limited sub-models with so-called separable kernel structures (see Section 3.1.4). This includes models with a tensor-product kernel or a saturated ANOVA decomposition kernel and is not capable of handling the non-saturated models that involve the addition of matrices in a Kronecker product form, such as (4.10). As mentioned previously, using a tensor product kernel implies including only the interaction term of the highest order. This may be problematic in many applications where assessing the effect of each predictor is needed. On the other hand, using the saturated ANOVA kernel means that we assume a saturated model, which could often overfit the data.

4.4.2 Eigendecomposition for saturated ANOVA kernel

Now let us assume that we have d-dimensional grid structure in the predictors. We have a response vector $\mathbf{y}$ of length n. If we use the saturated ANOVA decomposition kernel (3.11), the Gram matrix can be written as

$$\mathbf{K} = \bigotimes_{l=1}^{d} \tilde{\mathbf{K}}_l$$

where $\tilde{\mathbf{K}}_l = (\mathbf{1}_n \mathbf{1}_{n_l}^\top + \mathbf{K}_l)$ and $\mathbf{K}_l = \{k_{l,(i,j)}\}_{n_l \times n_l}$ with $k_{l,(i,j)} = k_l(\mathbf{x}_{l,i}, \mathbf{x}_{l,j})$. We write the eigendecomposition of each matrix $\tilde{\mathbf{K}}_l$ by $\tilde{\mathbf{K}}_l = \tilde{\mathbf{Q}}_l \tilde{\mathbf{\Lambda}}_l \tilde{\mathbf{Q}}_l^\top$. It follows from the result in (4.5),

$$\mathbf{K} = \left(\bigotimes_{l=1}^{d} \tilde{\mathbf{Q}}_l \right) \left(\bigotimes_{l=1}^{d} \tilde{\mathbf{\Lambda}}_l \right) \left(\bigotimes_{l=1}^{d} \tilde{\mathbf{Q}}_l \right)^\top . \tag{4.14}$$

Note that $\tilde{\mathbf{Q}} := \left(\bigotimes_{l=1}^{d} \tilde{\mathbf{Q}}_l \right)$ is orthonormal and $\tilde{\mathbf{\Lambda}} := \left(\bigotimes_{l=1}^{d} \tilde{\mathbf{\Lambda}}_l \right)$ is diagonal with non-negative entries as each $\tilde{\mathbf{K}}_l$ is positive semi-definite.

4.4.3 Eigendecomposition for hierarchical ANOVA kernel

We now show that the Kronecker product structure can be exploited for efficient computation even when we have a more general structure in the kernel, such as (4.10) if each (sub-)Gram matrix $\mathbf{K}_l$ is empirically centred by (2.7). To show this, we first establish the following results on an empirically centred Gram matrix.

Eigendecomposition of a centred Gram matrix

Recall that a Gram matrix $\mathbf{K}$ of order n can be empirically centred using centring matrix $\mathbf{C} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$ by $\mathbf{CKC}$. We denote the resulting matrix by $\mathbf{K}^{(c)}$. For a centred Gram matrix $\mathbf{K}^{(c)}$, we have the following result.

Lemma 1. *Any eigenvector $\mathbf{q}_i$ of a $n \times n$ centred Gram matrix $\mathbf{K}^{(c)}$ associated with non-zero eigenvalue λ_i is orthogonal to $\mathbf{1}_n$.*

Proof. Using $\mathbf{K}^{(c)}\mathbf{q}_i = \lambda_i\mathbf{q}_i$, we have

$$\mathbf{q}_i^\top\mathbf{1}_n = \frac{1}{\lambda_i}\mathbf{q}_i^\top\mathbf{K}^{(c)}\mathbf{1}_n = \mathbf{0}.$$

The last equality is due to the fact that all rows and columns of a centred matrix sum to 0. □

Lemma 2. *Any $n \times n$ centred Gram matrix $\mathbf{K}^{(c)}$ has the following eigendecomposition:*

$$\mathbf{K}^{(c)} = \mathbf{Q}^{(c)}\mathbf{\Lambda}^{(c)}\mathbf{Q}^{(c)\top} \tag{4.15}$$

where

$$\mathbf{\Lambda}^{(c)} = \mathbf{diag}\left((0, \lambda_2, \ldots, \lambda_n)^\top\right), \lambda_j \geq 0 \ \forall j \in \{2, \ldots, n\} \tag{4.16}$$

and

$$\mathbf{Q}^{(c)} = \begin{bmatrix} \frac{1}{\sqrt{n}} & & & \\ \vdots & \mathbf{q}_2 & \cdots & \mathbf{q}_n \\ \frac{1}{\sqrt{n}} & & & \end{bmatrix}. \tag{4.17}$$

Proof. Let k denote the number of zero eigenvalues of $\mathbf{K}^{(c)}$. Due to the centring, $\operatorname{rank}(\mathbf{K}^{(c)}) \leq n-1$, i.e., we have $k \geq 1$.

For $k = 1$, we have $\lambda_j > 0$, $\forall j \in \{2\ldots, n\}$ and the eigenvectors $\mathbf{q}_2, \ldots, \mathbf{q}_n$ are orthogonal to $\mathbf{1}_n$ from Lemma 1. Normalising the vector $\mathbf{1}_n$ completes an orthonormal basis, hence the first column of $\mathbf{Q}^{(c)}$ is given by $\frac{1}{\sqrt{n}}\mathbf{1}$.

For $k \geq 2$, the first k columns of $\mathbf{Q}^{(c)}$, $(\mathbf{q}_1, \ldots, \mathbf{q}_k)$, are not uniquely determined. Using $\mathbf{q}_j^\top(\frac{1}{\sqrt{n}}\mathbf{1}_n) = \frac{1}{\sqrt{n}}\mathbf{q}_j^\top\mathbf{1}_n = \mathbf{0}$ for $j = k+1, \ldots, n$, we set $\mathbf{q}_1 = \frac{1}{\sqrt{n}}\mathbf{1}_n$ and find $(\mathbf{q}_2, \ldots, \mathbf{q}_k)$ to complete an orthonormal system. □

In practice, we may use a computer program to obtain a initial set of normalised eigen-vectors denoted by $\mathbf{v}_1, \ldots, \mathbf{v}_n$. For $k \geq 2$, $\mathbf{v}_1, \ldots, \mathbf{v}_k$ may not contain a vector $\frac{1}{\sqrt{n}}\mathbf{1}_n$ but $\operatorname{span}(\mathbf{v}_1, \ldots, \mathbf{v}_k)$ contains $\mathbf{1}_n$. To have orthonormal bases $\mathbf{q}_1, \ldots, \mathbf{q}_n$ specified above, we take $\mathbf{q}_1 = \frac{1}{\sqrt{n}}\mathbf{1}_n$ and $\mathbf{q}_j = \mathbf{v}_j$ for $j = k+1, \ldots, n$. The rest of the vectors $\mathbf{q}_2, \ldots, \mathbf{q}_k$ can be computed using for example a Gram–Schmidt process.

Remark 1. *The $n \times n$ matrix $\mathbf{1}_n\mathbf{1}_n^\top$ has the following decomposition:*

$$\mathbf{1}_n\mathbf{1}_n^\top = \mathbf{Q}^{(c)}\mathbf{A}_n\mathbf{Q}^{(c)\top} \tag{4.18}$$

where $\mathbf{Q}^{(c)}$ is given by (4.17) and $\mathbf{A}_n$ is a $n \times n$ matrix with i, i-th element n and 0 everywhere else, i.e.,

$$\mathbf{A}_n = diag\Big((n, 0, \ldots, 0)^\top\Big). \tag{4.19}$$

Eigendecomposition with a hierarchical ANOVA decomposition kernel

Consider a hierarchical decomposition ANOVA kernel for data with a d-dimensional grid structure. Let us now assume that we have m terms in our additive kernel where $1 + d \leq m \leq 2^d$. The corresponding Gram matrix $\mathbf{K}$ then also involves addition of m

matrices $\mathbf{M}_p$ for $p = 1, \dots, m$:

$$\mathbf{K} = \sum_{p=1}^{m} \mathbf{M}_p \tag{4.20}$$

where

$$\mathbf{M}_p = \bigotimes_{l=1}^{d} \mathbf{B}_l \quad \text{where} \quad \mathbf{B}_l = \begin{cases} \mathbf{K}_l^{(c)}, & \text{if } \ k_l^{(c)}\text{is involved in the } p\text{-th term} \\[1em] \mathbf{1}_{n_l}\mathbf{1}_{n_l}^\top, & \text{otherwise.} \end{cases}$$

For this class of Gram matrices, we have the following lemma.

Lemma 3. *A matrix* $\mathbf{K}$ *of the form given by (4.20) has the following decomposition:*

$$\mathbf{K} = \left(\bigotimes_{l=1}^{d} \mathbf{Q}_l^{(c)}\right) \mathbf{D} \left(\bigotimes_{l=1}^{d} \mathbf{Q}_l^{(c)}\right)^\top$$

where $\mathbf{Q}_l^{(c)}$ *is orthonormal matrix whose columns consist of eigenvectors of* $\mathbf{K}_l^{(c)}$*, and* $\mathbf{D}$ *is diagonal with non-negative entries.*

Proof. From Lemma 2 (see 4.15 and 4.18), for $l = 1, \dots, d$, we have:

$$\begin{aligned} \mathbf{K}_l^{(c)} &= \mathbf{Q}_l^{(c)} \boldsymbol{\Lambda}_l^{(c)} \mathbf{Q}_l^{(c)\top} \\ \mathbf{1}_{n_l}\mathbf{1}_{n_l}^\top &= \mathbf{Q}_l^{(c)} \mathbf{A}_l \ \mathbf{Q}_l^{(c)\top} \end{aligned} \tag{4.21}$$

where (4.21) is an eigendecomposition of a centred matrix $\mathbf{K}_l^{(c)}$ i.e., $\mathbf{Q}_l^{(c)}$ is orthonormal, $\boldsymbol{\Lambda}_l^{(c)}$ is diagonal with non-negative eigenvalues in the diagonal (see 4.16 and 4.17) and $\mathbf{A}_l$ is a $n_l \times n_l$ matrix with $\mathbf{A}_{1,1} = n_l$ and 0 everywhere else. Then using the mixed product property of Kronecker products, we can decompose $\mathbf{M}_p$ as

$$\mathbf{M}_p = \bigotimes_{l=1}^{d} \mathbf{Q}_l^{(c)} \bigotimes_{l=1}^{d} \mathbf{D}_{pl} \bigotimes_{l=1}^{d} \mathbf{Q}_l^{(c)\top}$$

where

$$\mathbf{D}_{pl} = \begin{cases} \boldsymbol{\Lambda}_l^{(c)} & \text{if } k_l^{(c)} \text{is involved in the } p\text{-th term} \\ \\ \mathbf{A}_l & \text{otherwise.} \end{cases}$$

Let $\mathbf{D}_p = \bigotimes_{l=1}^{d} \mathbf{D}_{pl}$. We have

$$\begin{aligned} \mathbf{K} = \sum_{p=1}^{m} \mathbf{M}_p &= \sum_{p=1}^{m} \left(\bigotimes_{l=1}^{d} \mathbf{Q}_l^{(c)} \mathbf{D}_p \bigotimes_{l=1}^{d} \mathbf{Q}_l^{(c)\top} \right) \\ &= \bigotimes_{l=1}^{d} \mathbf{Q}_l^{(c)} \left(\sum_{p=1}^{m} \mathbf{D}_p \right) \bigotimes_{l=1}^{d} \mathbf{Q}_l^{(c)\top}. \end{aligned} \tag{4.22}$$

It is easy to see that each $\mathbf{D}_p$ and hence also the matrix $\mathbf{D} := \sum_{p=1}^{m} \mathbf{D}_p$ is diagonal with non-negative diagonal entries. □

Example 8 (Example of the Kronecker method for a hierarchical ANOVA kernel with $d = 2$)**.** *Consider a model specified by (4.7) and (4.8), i.e., the main effect model for two-dimensional grid data. If we use centred kernels, the Gram matrix given by (4.10) can be written as*

$$\mathbf{K} = \mathbf{1}_{n_1}\mathbf{1}_{n_1}^\top \otimes \mathbf{1}_{n_2}\mathbf{1}_{n_2}^\top + \mathbf{K}_1^{(c)} \otimes \mathbf{1}_{n_2}\mathbf{1}_{n_2}^\top + \mathbf{1}_{n_1}\mathbf{1}_{n_1}^\top \otimes \mathbf{K}_2^{(c)}$$

and can be decomposed as

$$\begin{aligned} \mathbf{K} = \mathbf{Q}_1^{(c)}\mathbf{A}_1\mathbf{Q}_1^{(c)\top} \otimes \mathbf{Q}_2^{(c)}\mathbf{A}_2\mathbf{Q}_2^{(c)\top} + \mathbf{Q}_1^{(c)}\boldsymbol{\Lambda}_1\mathbf{Q}_1^{(c)\top} \otimes \mathbf{Q}_2^{(c)}\mathbf{A}_2\mathbf{Q}_2^{(c)\top} \\ + \mathbf{Q}_1^{(c)}\mathbf{A}_1\mathbf{Q}_1^{(c)\top} \otimes \mathbf{Q}_2^{(c)}\boldsymbol{\Lambda}_2\mathbf{Q}_2^{(c)\top} \\ = \left(\mathbf{Q}_1^{(c)} \otimes \mathbf{Q}_2^{(c)}\right)\left(\mathbf{A}_1 \otimes \mathbf{A}_2 + \boldsymbol{\Lambda}_1 \otimes \mathbf{A}_2 + \mathbf{A}_1 \otimes \boldsymbol{\Lambda}_2\right)\left(\mathbf{Q}_1^{(c)} \otimes \mathbf{Q}_2^{(c)}\right)^\top . \end{aligned}$$

Thus far, we have restricted the prior variance of the constant term a in (4.8) and (4.9) to be 1. Lifting this assumption is straightforward. If the prior variance is α_0^2, the decomposition of the corresponding Gram matrix can be expressed in the same way as (4.14) and (4.22) with all elements of the middle diagonal matrix multiplied by α_0^2.

4.4.4 Computational complexity and space requirement

Kronecker methods significantly reduce the cost of computing the log determinant of the matrix $\mathbf{K}+\sigma^2\mathbf{I}$, and solving the linear system $(\mathbf{K}+\sigma^2\mathbf{I})^{-1}\mathbf{v}$, which usually has $O(n^3)$ when $\mathbf{K}$ is an $n \times n$ Gram matrix. The key operations are eigendecomposition $\mathbf{K}$ and matrix-vector multiplication involving Kronecker products (4.13). With a Kronecker product structure, eigendecomposition is applied to each $\mathbf{K}_l$ of size $n_l \times n_l$ individually, which has $O(n_l^3)$ complexity. The total cost for the eigendecomposition of $\mathbf{K}$ then reduces to $O(\sum_{l=1}^d n_l^3)$, which is dominated by the largest of n_l. The second component is a matrix-vector multiplication in $(\bigotimes_{l=1}^d \mathbf{Q}_l^\top)\mathbf{v}$. A matrix-vector multiplication of an $n \times n$ matrix and a vector of length n usually requires $O(n^2)$ operations. Using the algorithm provided in Saatçi (2012, Chapter 5) and Wilson et al. (2014), this Kronecker product matrix-vector multiplication takes $O(n \sum_{l=1}^d n_l)$ which is much less than the usual $O(n^2)$. Once we have eigenvalues of all sub-Gram matrices $\mathbf{K}_l$, computing the log-determinant has an additional cost of $O(n)$. The storage requirement reduces from $O(n^2)$ to $O(\sum_{l=1}^d n_l^2)$ which is associated with storing matrices $\mathbf{Q}_1, \dots, \mathbf{Q}_d$. Previous work by Saatçi (2012) and Wilson et al. (2014) explored the use of the Kronecker method in GP regression and demonstrated improved computational time through simulation studies. Our approach, which shares the same key factors determining computational cost (namely, eigendecomposition of Gram matrices and matrix-vector multiplication), is expected to yield similar computational gains.

Summary of Computational Complexity

- Time complexity
  - Eigendecomposition of Gram matrices: $O(n^3)$ to $O(\sum_{l=1}^{d} n_l^3)$
  - Matrix-vector multiplication: $O(n^2)$ to $O(n \sum_{l=1^d} n_l)$
  - Log-determinant: $O(n)$
- Memory requirement: $O(n^2)$ to $O(\sum_{l=1}^{d} n_l^2)$

4.4.5 Other scalable approaches

A number of methods have been proposed to enhance the scalability of GP models. As summarized by Liu et al. (2020), one mainstream approach involves approximating the Gram matrix $\mathbf{K}$. This can be achieved by utilizing a subset of data, typically of size $m \ll n$ (subset-of-data), or by exploiting sparsity in the Gram matrix. This is based on the assumption that the covariance between distant points is zero, resulting in sparse kernels (Melkumyan and Ramos, 2009). A particularly popular technique is the low-rank approximation using inducing points (e.g., Hensman et al. (2013); Titsias (2009)), inspired by Nystrom's method (Williams and Seeger, 2001). In the spatio-temporal setting, Datta et al. (2016) introduced dynamic nearest neighbour GP that induces a sparse structure in the inverse of the covariance matrix with wide range of kernel structures, including non-separable kernels. This was used to analyse air pollution data similar in size to the data in our study provided in Section 4.5. Unlike the Kronecker approach, which necessitates a multidimensional grid structure for the data, these methods can be applied to broad data structures. However, the Kronecker approach offers the advantage of avoiding approximation, as it rather exploits the structure of the data to efficiently evaluate and store the key components required for estimation and inference. In fact, the Kronecker approach and other scalable methods can complement each other, as exemplified by Wilson and Nickisch (2015), who incorporated a grid structure into inducing points. Although their work focused on the tensor product kernel, the method can be extended to handle additive kernels using the decomposition discussed in Section 4.4.3.

4.5 Application to NO2 concentrations in London

We applied our method to analyse a dataset of hourly nitrogen dioxide (NO_2) concentrations in London, covering the period before and after the first COVID-19 lockdown in the UK. NO_2 is a harmful air pollutant that adversely affects human health and the environment. Short-term exposure to high levels of NO_2 can irritate the respiratory system, exacerbating conditions like asthma, while prolonged exposure is linked to lung and cardiovascular diseases. Additionally, NO_2 contributes to environmental damage through acid rain, smog, and ozone formation. According to the World Health Organization, recommended ambient NO_2 concentration levels are below $10\mu g/m^3$ for the annual mean and at most 3-4 exceedance days per year with a 24-hour mean of $25\mu g/m^3$. Investigating the direct and indirect effects of policy interventions and regulation changes that reduce NO_2 emissions is crucial. One example of such indirect policy intervention is the lockdown measures taken across the world. In urban areas like London, vehicle emissions are the main source of NO_2. A number of studies have shown that restricted mobility due to the lockdown has contributed to a drop in NO_2 concentration in the air on a global scale (See Dutta et al. (2021) and Cooper et al. (2022) for example). By examining measurement data from monitoring stations across the country, a comparable conclusion was drawn at the national level in the UK (Higham et al., 2021; Jephcote et al., 2021; Lee et al., 2020). The majority of such studies analyse the average daily concentrations. While daily mean data are suitable when the research question revolves around assessing long-term changes in the concentration level, interesting findings may be reached by analysing hourly-recorded data. It is known that the concentration of NO2 in the air varies over the course of a day, with a distinct daily cycle. If we are to investigate the effect of lockdown, in addition to studying the average downtrend, which is typically done by analysing daily or weekly average data, one may ask if the daily cycle changed over time during this period. With measurement data from multiple sites, it is also possible to study spatial patterns, and with that pattern identified, we can conduct further research on spatio-temporal interaction. That is, we can let the daily cycle or global time trend be different at different locations. Answering these questions requires analysis of hourly-measured data over a number of days at different locations, which easily results in massive data. However such data typically have a balanced panel structure as described in the next section, and can be analysed efficiently using the proposed Kronecker approach. Although our data analysis is exploratory, we aim to show that flexible additive GP models combined with this Kronecker method make it possible to investigate important research questions that would have been otherwise infeasible.

4.5.1 Dataset

We used a dataset of NO2 concentrations (measured in $\mu g/m^3$) collected at various sites in the London Air Quality network[1] during the period from January 6th, 2020 to May 30th, 2020. We excluded sites with more than 30% missing values or more than 48 consecutive missing values, resulting in a total of 59 sites and 208,152 observations. As seen in Figure 4.5, the data has a three-level structure: site location (Easting and Northing), day, and hour of the day. To accommodate the Kronecker method, which usually requires a complete grid structure, we imputed the remaining missing values following the steps outlined in Appendix C.3.1. Although the method used

[1] https://www.londonair.org.uk

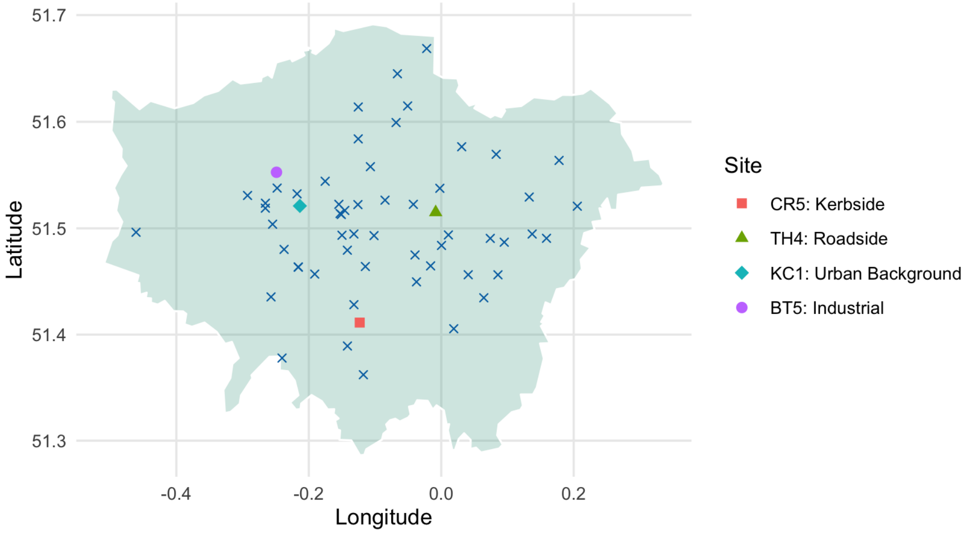


Fig. 4.4 The location of NO2 measurement sites included in the dataset. Four sites are selected and labelled for illustration purposes. Site CR5 is in the Borough of Croydon, while TH4, KC1 and BT5 are in Tower Hamlets, Kensington and Chelsea, and Brent. The sites in the dataset are classified into 5 categories: Kerbside, Roadside, Urban Background, Suburban and Industrial.

for imputation is simple, due to a very small proportion (0.62%) of missing values, we believe that any bias introduced does not have significant impact. Another issue to consider is the transformation of the response, NO_2 concentrations. While they are typically modeled on a log-scale to address the right skewness commonly found in air pollutant concentration data, due to the presence of zero and negative records (accounting for measurement equipment uncertainty), we retained the response variable in its original scale. The observed NO_2 levels range from -2.9 to 320.6. Additionally, we adjusted for the transition from winter time to summer time during the study period (details provided in Appendix C.3.1).

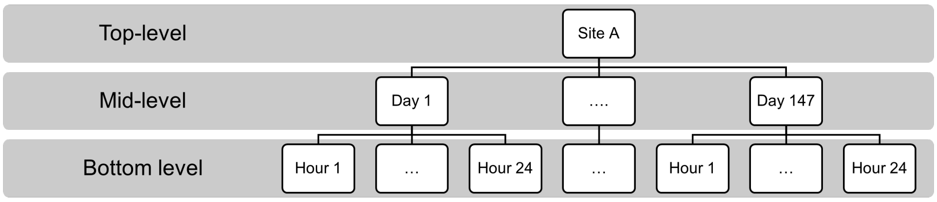


Fig. 4.5 The structure of the NO_2 concentrations data

4.5.2 Model formulation

As shown in Figure 4.5, the dataset has a three-dimensional grid structure, with location, day and hour as predictors, denoted by $\mathbf{x}_1 \in \mathcal{X}_1$, $x_2 \in \mathcal{X}_2$ and $x_3 \in \mathcal{X}_3$ where $\mathcal{X}_1 \subset \mathbb{R}^2$, $\mathcal{X}_2$ represents the set of calendar dates numbered $1, 2, \ldots$, and $\mathcal{X}_3$ represent hour of the day indexed by $1, 2, ..., 24$. Let $y_{s,d,h}$ denote the observed NO2 concentration from monitoring station s, on day d at hour h where $s = 1, \ldots, n_1, d = 1, \ldots, n_2, h = 1, \ldots, n_3$ and $n_1 = 59, n_2 = 147, n_3 = 24$. To model the response, we consider a function of three variables $f : \mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \times \mathcal{X}_3 \to \mathbb{R}$ and assume a zero mean GP prior. The model is given by:

$$y_{s,d,h} = f(\mathbf{x}_{1s}, x_{2d}, x_{3h}) + \epsilon_{s,d,h}$$

with $f \sim \mathrm{GP}(0, k)$ where a covariance kernel $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is modelled according to our prior belief. We assume i.i.d error with $\epsilon_{s,d,h} \sim N(0, \sigma^2)$. We consider several kernels that belong to a class of hierarchical ANOVA decomposition kernels and the tensor product kernel. Note that for $l = 1, 2, 3$, we denote the base kernel for each predictor by k_l defined on $\mathcal{X}_l \times \mathcal{X}_l$ and assume $f_l \sim \mathrm{GP}(0, k_l)$. The interaction terms are constructed using the tensor product kernel (3.8). For the model specification, we drop the subscripts s, d, h to simplify the expression.

List of models

We list the kernels for the models we consider for this dataset. Note that the prior over the overall function f follows zero mean GP with kenel $k = \alpha_0^2 k_{m_j}$ for $j = 1, \dots, 5$.

Model 1: main effect

The first model we consider is the main effect model, where $f(\mathbf{x}_1, x_2, x_3)$ and the kernel are given by

$$\begin{aligned} f(\mathbf{x}_1, x_2, x_3) &= a + f_1(\mathbf{x}_1) + f_2(x_2) + f_3(x_3) \\ k_{m1} &= 1 + k_1 + k_2 + k_3. \end{aligned}$$

The Gram matrix with this kernel and all other kernels under consideration are given in Appendix C.3.2. This model does not involve any interaction effects, meaning that the effect of the location is constant throughout the whole period. Or equivalently, the global time trend captured using the day predictor x_2 and the daily cyclical effect captured by the hour predictor x_3 are both assumed the same for all sites. We use Model 1 as the baseline model and extend it to include two-way or three-way interaction effects.

Model 2: space-time interaction

If we assume a space and time interaction, i.e., both the global time trend and the daily cycle are different at different location, we have Model 2 specified by

$$\begin{aligned} f(\mathbf{x}_1, x_2, x_3) &= a + f_1(\mathbf{x}_1) + f_2(x_2) + f_3(x_3) + f_{12}(\mathbf{x}_1, x_2) + f_{13}(\mathbf{x}_1, x_2) \\ k_{m2} &= k_{m1} + k_1 \otimes k_2 + k_1 \otimes k_3. \end{aligned}$$

Model 3: all two-way interaction

The model with all two-way interactions extends Model 2 by

$$f(\mathbf{x}_1, x_2, x_3) = a + f_1(\mathbf{x}_1) + f_2(x_2) + f_3(x_3) + f_{12}(\mathbf{x}_1, x_2) + f_{13}(\mathbf{x}_1, x_2) + f_{23}(x_2, x_3)$$

$$k_{m3} = k_{m2} + k_2 \otimes k_3$$

and adds further assumption that the daily cycle changes over time.

Model 4: saturated/three-way interaction model

If we consider the saturated model with a three-way interaction, the model equals

$$\begin{aligned} f(\mathbf{x}_1, x_2, x_3) &= a + f_1(\mathbf{x}_1) + f_2(x_2) + f_3(x_3) + \\ &\qquad f_{12}(\mathbf{x}_1, x_2) + f_{13}(\mathbf{x}_1, x_2) + f_{23}(x_2, x_3) + f_{123}(\mathbf{x}_1, x_2, x_3) \\ k_{m4} &= k_{m3} + k_1 \otimes k_2 \otimes k_3. \end{aligned}$$

Note that this is the saturated ANOVA decomposition kernel for three-dimensional grid data.

Model 5: three-way interaction only

Finally, we also fit a model with only the three-way interaction term, of which the model function f and the kernel are given by

$$f(\mathbf{x}_1, x_2, x_3) = f_{123}(\mathbf{x}_1, x_2, x_3)$$

$$k_{m5} = k_1 \otimes k_2 \otimes k_3.$$

Although this separable kernel is widely used in machine learning applications, with this kernel construction, the interpretation of the effects of each predictor is difficult. The result shows that it fits the data poorly compared to other models under consideration.

Kernel choice

The choice of baseline kernels k_l is also an important factor that reflects our prior belief about the underlying process. As discussed in Section 2.1.3, a GP with the squared-centred standard Brownian motion kernel ($\gamma = 0.5$) has good smoothness properties. This kernel, as well as other fractional Brownian motion-based kernels, has a computational advantage over other common kernels, such as squared exponential kernels, as we do not have to perform eigendecomposition or matrix-vector multiplication at each iteration of a chosen optimisation algorithm (See Appendix B.2). We use this as a starting point and explore different options. We found that the spatial process f_1 is rougher than the temporal processes f_2, f_3. To determine the optimal values for the Hurst coefficient γ in k_1, we conducted a grid search, which led to the choice of $\gamma = 0.3$. For k_2 and k_3, squared-centred standard Brownian motion kernels ($\gamma = 0.5$) produced a good fit. We use a half normal distribution with its scale parameter set to 1 as priors on the rest of the hyper-parameters, including the scale parameters of the overall kernel and each base kernel, $\alpha_0, \alpha_1, \alpha_2, \alpha_3$, and the variance of the error term σ^2 and estimated them by the posterior mean from MCMC samples. The specified models are implemented using the programming language Stan[2]. Changing the hyper-prior to, e.g., a log-normal distribution made a little difference in the result.

4.5.3 Results

The main results are shown in Table 4.1, including the point estimates of hyper-parameters, log marginal likelihood improvement compared to the baseline model and computational time. We use the best-fit log marginal likelihood to compare different models. The marginal likelihood has a closed-form expression since we assume Gaussian likelihood on the response y, and different interaction models share the same

[2] The code is provided at github.com/sahokoishida/Additive-GP-Kronecker

Table 4.1 Results from fitting Model 1 to Model 5 to London NO_2 data. The difference of the log marginal likelihood in comparison to that of the baseline model (Model 1) is shown as Δmloglik. The log marginal likelihood for Model 1 is -857,889. The average time (in minutes) taken to obtain 2 chains of 300 MCMC samples after the 200 warm-up phase is also displayed. For model 5, we only need one scale parameter α_1 due to identifiability issues.

Model	α_0	α_1	α_2	α_3	σ
1: main	6.87	5.08	2.23	0.34	14.85
2: spatio-temporal interaction	14.36	1.29	0.67	0.24	12.54
3: all two-way interaction	10.67	1.86	1.32	0.31	8.37
4: saturated	52.05	0.48	0.94	0.051	6.51
5: three-way interaction only	-	0.0	-	-	36.43
Model	Δmloglik	Time (m)			
1: main	-	20.9			
2: spatio-temporal interaction	25,834	14.3			
3: all two-way interaction	98,070	17.5			
4: saturated	104,109	17.1			
5: three-way interaction only	-186,450	1.96			

number of hyper-parameters (see Section 3.2.2). From the log marginal likelihood values for Model 1 to Model 4, we see that the fit to the data improves as the model becomes more complex. It is apparent that Model 3 (all two-way interactions) offers significant improvement compared to Models 2 (spatio-temporal interaction), indicating the importance of including the additional interaction term between daily cycle and global time effect. The improvement in the log-marginal likelihood by including the three-way interaction is not as large, however, as we still sees a major improvement, we conclude that Model 4 (saturated) is the best model. It is important to note, however, the saturated model does not always offer the best fit, and in order to confirm the presence of higher order interactions, a comparison to simpler models is essential. MCMC sampling took less than 20 minutes on average for all models. Note that Model 5, which took 1.96 minutes, only has two model parameters. We also estimated hyper-parameters by finding maximiser of the log marginal likelihood (2.15), which took a few seconds for each model as shown in Appendix C.3.3. It can also be seen that the interaction-only model performs worse than the simple main effect model. In Figure 4.6, we show the posterior mean and the 2.5% and 97.5% quantiles of the posterior predictive distribution derived by Model 3, at selected sites (see Table 4.4) from the 23rd of March for two weeks together with the corresponding observed values. We see that the posterior predictive mean captures the overall trend well, but smoother than the observed. The model comparison implies that, on top of the space-time interaction, the effect of the hours of the day interacts with the effects of global time. That is, the global time trend and the daily cycle of NO_2 concentration level are different for different locations, and the identified daily cycle also changes over time. Furthermore, the fact that the additional three-way interaction term led to further improvement in log marginal likelihood suggests that this change in daily pattern over time depends on the measurement sites.

4.5.4 Visualisation and interpretation of each effect

To interpret the chosen model (Model 4), we can visualise each effect using the mean decomposition of additive GP models seen in (3.12). We denote the posterior mean function of the main effect terms by $\bar{m}_l$ for $l = 1, 2, 3$, two-way interaction effect terms by $\bar{m}_{ll'}$ for $1 \leq l < l' \leq 3$, and the three-way interaction effect term by $\bar{m}_{123}$. In addition, we have the mean of the constant term $\bar{m}_a = 28.988$. As discussed in Section 3.2.1, due to centering of kernels, the mean of the main effect, as well as mean (over each covariate) of interaction effect is zero, e.g., we have $\sum_{s=1}^{59} \bar{m}_1(\mathbf{x}_{1,s}) = 0$, $\sum_{d=1}^{147} \bar{m}_{23}(x_{2,d}, x_{3,h}) = 0$ and $\sum_{h=1}^{24} \bar{m}_{123}(\mathbf{x}_{1,s}, x_{2,d}, x_{3,h}) = 0$. This allow for interpreting the lower order interaction effects, including the main effects, as averaged effects. In what follows, we visualise and interpret the main effects, selected two-way interaction effects and the three-way interaction effects.

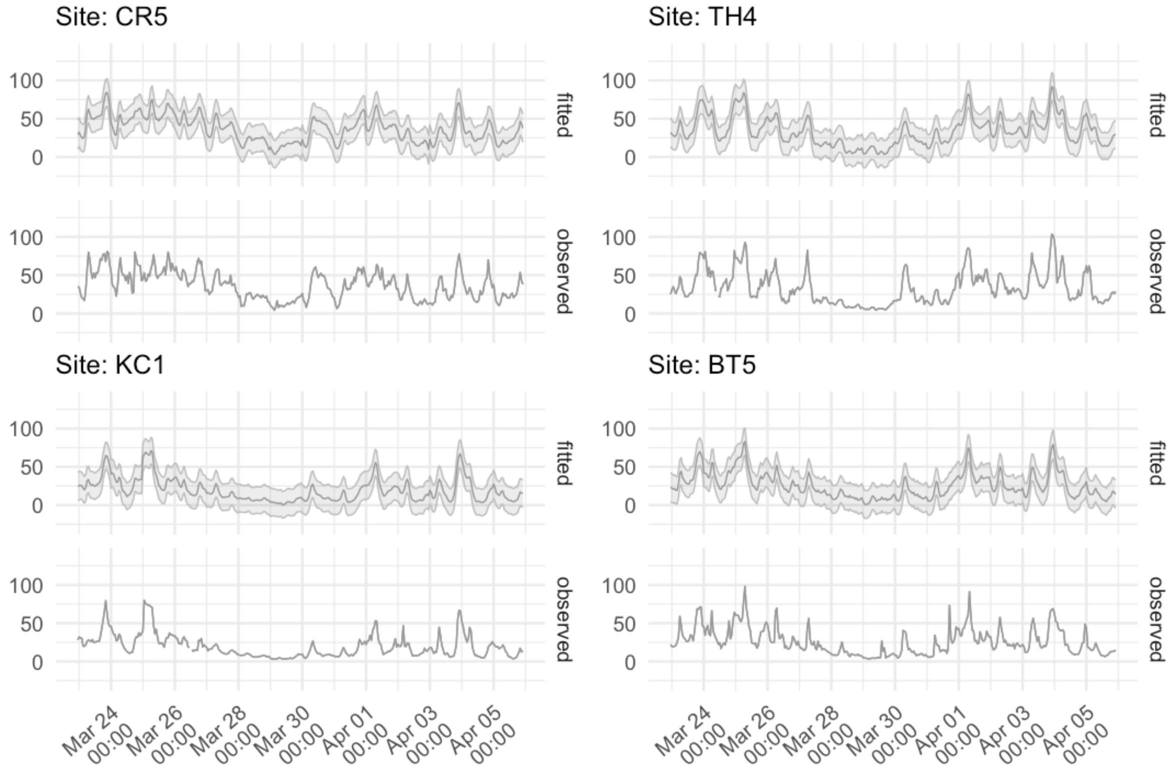
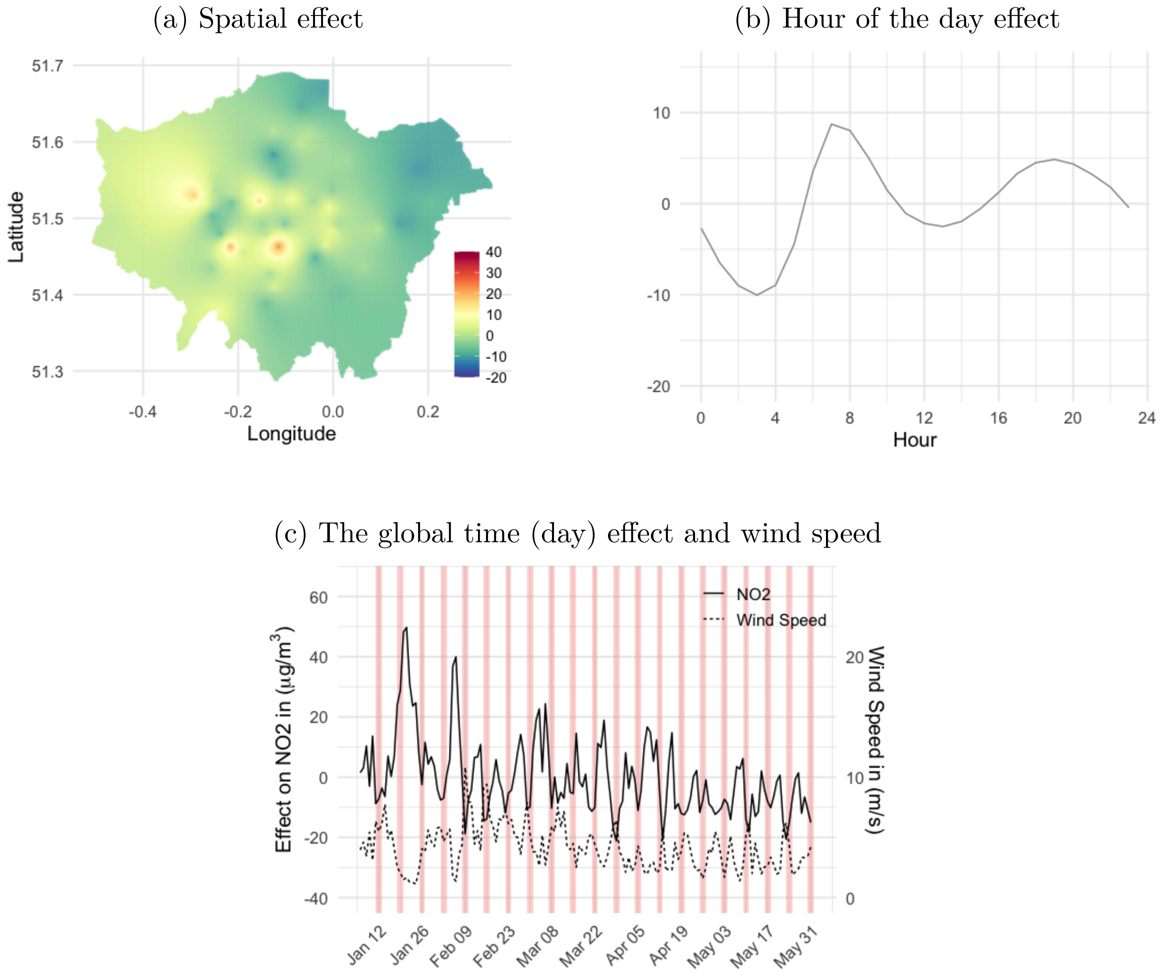


Fig. 4.6 Observed and fitted (with 95% predictive bands) NO2 concentrations (in $\mu g/m^3$) at 4 different sites

Figure 4.7 shows the three main effects $\bar{m}_1(\mathbf{x}_1), \bar{m}_2(x_2)$ and $\bar{m}_3(x_3)$. From Figure 4.7a showing spatial effect averaged over hour of the day and calendar date, we see a few hot spots in central London and a negative effect in the outskirts of London, especially towards the east. However, as we did not take into account the types of sites in modelling, we need to be careful with the interpretation. The average global time effect is visualised in Figure 4.7c. We notice a small downward trend over the period. Whether this is the effect of the lockdown or the result of lower emission from e.g. heating source due to warmer weather needs further investigation. For example, we can compare the records from previous years, or incorporate information such as temperature in the model. Short-term variation is also apparent. There are many factors to consider in explaining this variation. One possibility is the weekend effect associated with less traffic. Looking at the figure, we see that some troughs in the


(a) Spatial effect

(b) Hour of the day effect

(c) The global time (day) effect and wind speed

Fig. 4.7 Main effects on NO_2 in $\mu g/m^3$. In panel (c), weekends are highlighted. We see low wind speed corresponding to a higher level of pollution.

figure seem to fall into weekend although there are some deviations from this pattern. This may be due to some meteorological factors affecting the concentration level of the pollutant. For example, it is well-known that lower wind speeds are associated with higher concentrations. The two-week period starting on the 12th of January sees one large peak across the sites, which corresponds with the trough in the wind speed. Other weather conditions such as temperature, precipitation or sunlight also affect the levels of NO_2 concentrations in the air. Including meteorological data as covariates can

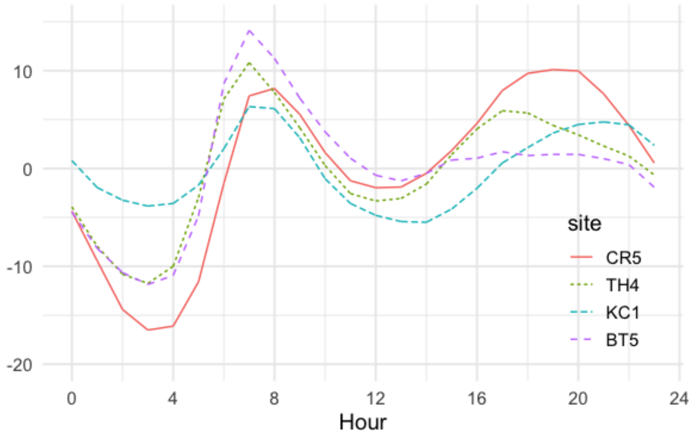


Fig. 4.8 Hour of the day effect on NO$_2$ in μg/m^3 by site

help improve the fit, as well as estimate and visualise the time trend after removing the effect of the weather conditions. If the weekly pattern is still apparent, we can introduce another level in the data structure, i.e., a four-dimensional grid structure, and gain further computational efficiency. The hour of the day effect shown in 4.7b is averaged over monitoring stations and calendar date. It shows two clear peaks in the morning and in the evening with the former is bigger in magnitude.

As two-way interaction effects between x_3 and both $\mathbf{x}_1$ and x_2 are present, the hour of the day effect changes over location and time. To see this we can plot $\bar{m}_3(x_3) + \bar{m}_{13}(\mathbf{x}_1, x_3)$ and $\bar{m}_3(x_3) + \bar{m}_{23}(x_2, x_3)$ as a function of x_3 for different $\mathbf{x}_1$ and x_2. Figure 4.8 shows the estimated daily cycle at selected sites. It is noticeable that Station CR5 and Station TH4 which are located in kerbside and roadside respectively, have two peaks corresponding to rush hours. While the latter (TH4) shows a similar pattern to that of the main effect, the former is distinctively different with a larger peak in the evening. The station in the background, KC1, also shows two peaks but with much smaller magnitude of the effect. Station BT5 is close to an industrial site and with only one peak in the morning. Figure 4.9 shows the interaction effect between the global time and the hour of the day for selected weeks. During the week commencing on the 27th of January (week 5 of 2020), there are two clear peaks in the beginning of the week, but this pattern becomes less clear over the course of the week. During the weekend, the magnitude of the effect becomes smaller. The week starting on the 23rd of March (week 14 of 2020), where the first COVID-19 lockdown in the UK took place, shows more irregular patterns. The Monday of this week is the day that the British Prime Minister announced the plan to introduce the measure which legally came into force 3 days later. The first national lockdown lasted the next few months with parts of the restriction being lifted from mid-May. The last two plots show the effect in the week starting on the 20-th of April (week 17 of 2020, week 4 of lockdown) and on the 25-th of May (week 22 of 2020, week 11 of lockdown). There is a distinct peak in the morning for many of the days, but the evening peak is much less apparent compared with week 5. Appendix C.3.4 includes a few more examples of how to visualise and interpret two-way interaction effects.

The chosen model indicates that the change of the cyclical effect over time is different for different location. This can be visualised by plotting $\bar{m}_3(x_3) + \bar{m}_{23}(x_2, x_3) + \bar{m}_{123}(\mathbf{x}_1, x_2, x_3)$ as a function of x_3 (shown in Figure 4.10). In week 5 of 2020, the pattern in Station CR5 and KC1 are similar especially in the beginning of the week. On the eleventh week into the lockdown, the pattern for Station KC1 is similar to the last panel in Figure 4.9, while this is not the case for Station CR5, where two daily peaks are still observed.

4.6 Discussion

This section proposed an efficient method to implement additive GP regression for multidimensional grid data, by exploiting the Kronecker product structure in the Gram matrix. This approach has been used in the literature but has had limited use, as it can only handle models with kernels constructed as tensor products, which include

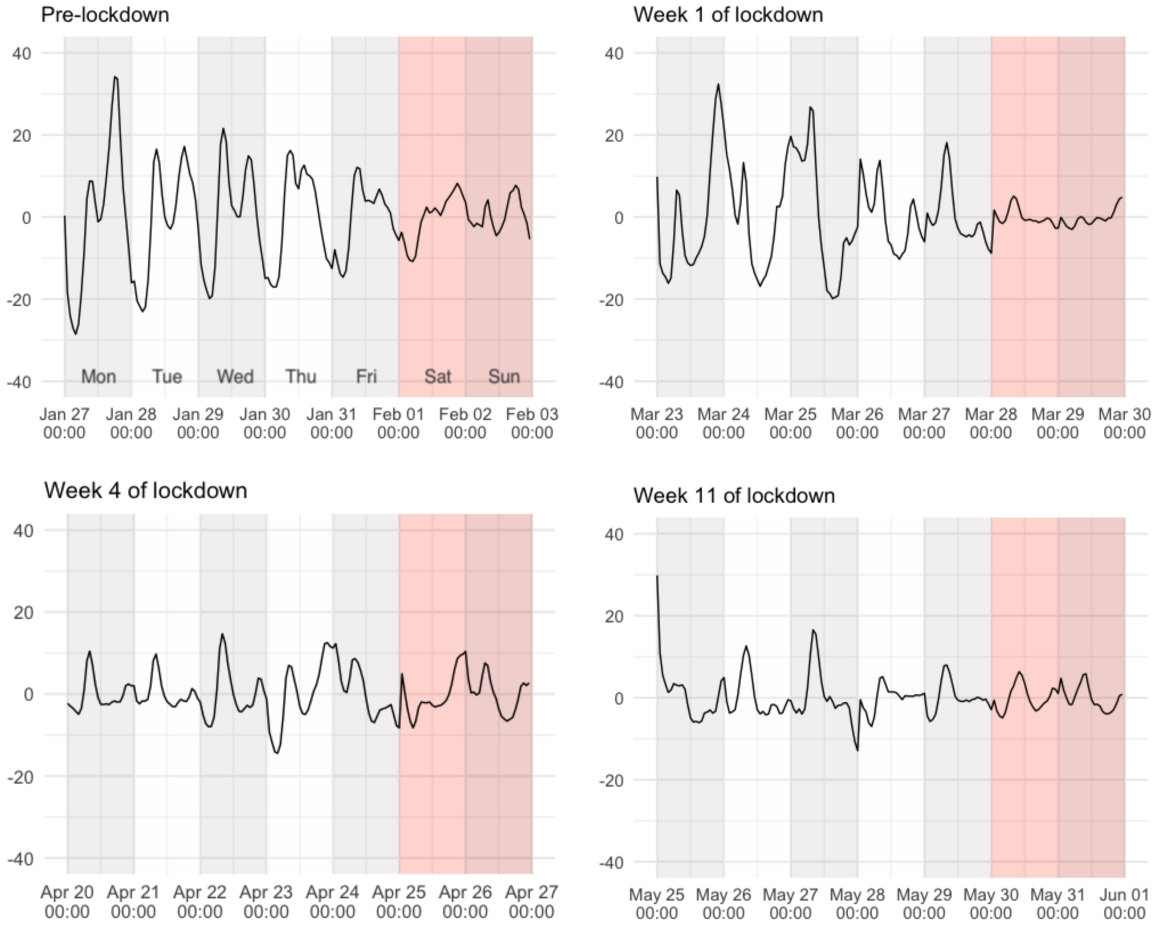


Fig. 4.9 Change in hour of the day effect (in μg/m^2) over time averaged over different monitoring stations.

saturated interaction model, and the model with the highest-ordered interaction. Our contribution is to extend the Kronecker approach to handle additive interaction effect models of various structures, by making use of the ANOVA decomposition kernel and centring of kernel. This allows for the analysis of large-scale multidimensional grid data without being constrained to the saturated model.

The proposed method is applied to efficiently analyse hourly-recorded ambient NO2 concentrations in London for the period covering both before and after the COVID-19 lockdown measure was introduced. We treated the data as three-dimensional grid data, with the location of the monitoring stations, day, and hour of the day as three

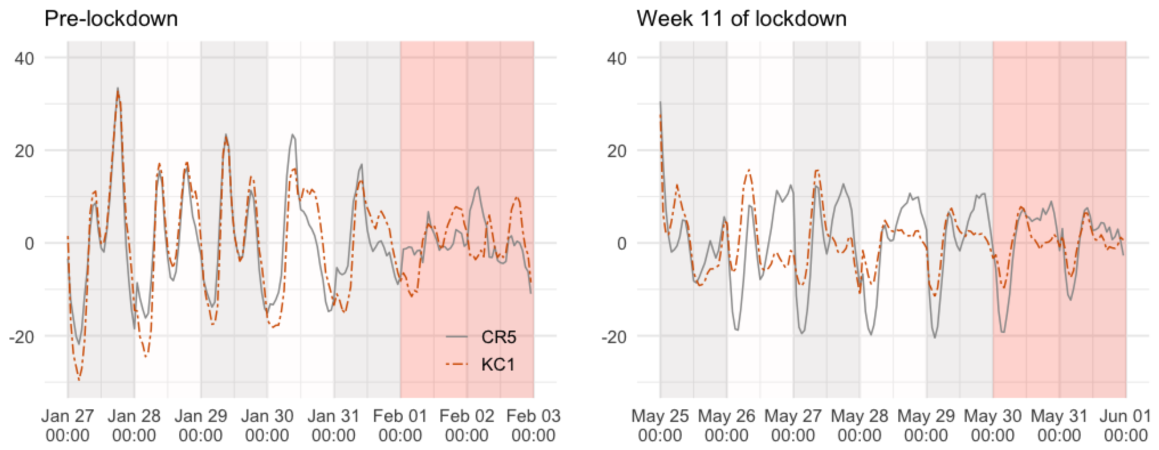


Fig. 4.10 Change in hour of the day effect (in μg/m^2) over time at two different monitoring stations, CR5 and KC1

predictors. The effect of three variables can be seen as the spatial effect, global time effect and daily cyclical effect respectively. We considered five regression models, including the main effect model, two kinds of two-way (hierarchical) interaction models, the three-way (hierarchical) interaction model, and the three-way interaction only model. We compared the models in terms of the marginal likelihood and found that the three-way interaction model is the best fit for the data, suggesting that the global time trend and the daily cycle are different for different locations, with the latter changing over time. How daily cycles changes over the course of the period differs across various monitoring stations. The proposed Kronecker approach enabled efficient implementation of all models considered. This allowed us to compare different models and confirm the presence of interaction terms. It is also important to note that the success of our approach does not reduce the significance of other scalable approaches to GP regression that can be used for more general data structures such as a subset of data, Nyström approximation, inducing points methods, sparse variational methods, or vice versa. In fact, the Kronecker approach, including our proposal, can be combined with such methods, which can then facilitate the analysis of even larger datasets.

We would like to point out that, while it is not considered in this section, the inclusion of other covariates (e.g. such as information related to monitoring stations, or meteorological variables for our data example) is possible under the proposed model. If the covariates are level-specific, e.g. types of the monitoring stations (roadside, background etc.) at the top level, or daylight time in London at the middle level, we can simply add the centred covariance function for each covariate to the kernel given at each level, which in this example, are the kernels for location and day, respectively. However, many meteorological variables are observed at a cross-level. For example, we may want to use hourly-recorded wind speed or precipitation at different locations. If the effect of these meteorological variables is assumed linear (including polynomial), we can incorporate them into a model while still avoiding cubic time complexity, by combining the idea of semi-parametric GP model, discussed in Section 2.4, with our Kronecker approach. Using such information in the model is especially important if prediction is the main interest.

Another worthwhile aspect to investigate is the challenge of handling missing values. In our data analysis, we assume a complete grid structure, i.e., no missing values. This may not always be a reasonable assumption for real-world applications. For example, in air quality monitoring, there may be some periods where the data is not available due to malfunctioning of the monitoring devices. Repeated measurements in social, psychological, or medical science commonly suffer from dropout. In this data analysis example, we excluded some monitoring sites that have many missing records and imputed the rest of the missing values by a simple procedure. If the proportion of missing values in the data is large, more sophisticated approaches should be considered to avoid potential bias. Gilboa et al. (2013); Wilson et al. (2014) proposed an approximation to the likelihood in the presence of missing values in multidimensional grid structure data. The next section explore this further.

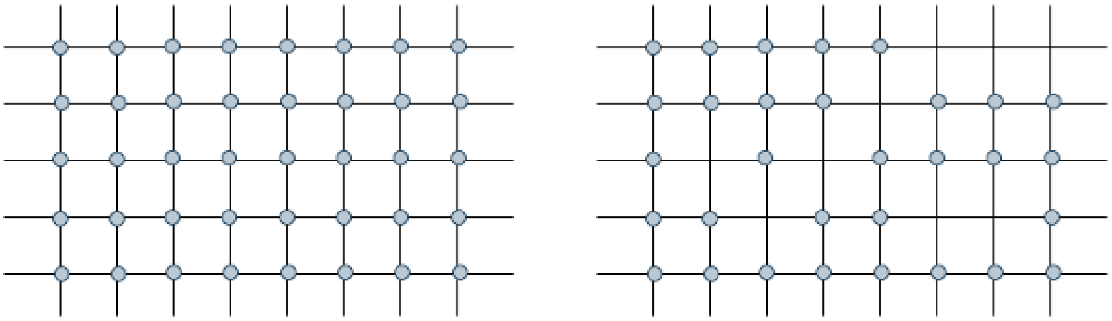
Chapter 5

Kronecker Gaussian process models with incomplete grid

In the previous chapter, we discussed how Kronecker product structures in the Gram matrix facilitate computations in Gaussian process (GP) regression. While the advantage of this approach, scalability without approximation, is clear, what limits its use in practice is the fact that it generally requires a complete grid structure. This is not the case for many real-world data. In this chapter, we consider a regression model with a hierarchical ANOVA decomposition kernel for data where the response vector $\mathbf{y}$ is on an incomplete d-dimensional grid $\mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_d$. We denote each predictor by $\mathbf{x}_l \in \mathcal{X}_l$ for $l = 1, \ldots, d$ and a set of d-dimensional predictors by $\mathbf{x} = (\mathbf{x}_1^\top, \ldots, \mathbf{x}_d^\top) \in \mathcal{X}$. Figure 5.1 shows an example with $d = 2$.

5.1 Issues with incomplete grid

In the previous chapter, a critical assumption we operated under was the completeness of the grid, signifying the absence of missing values within the response vector $\mathbf{y}$. Nonetheless, incomplete grids are a common occurrence in real-world datasets. For



(a) Complete grid

(b) Incomplete grid

Fig. 5.1 Illustration of complete and incomplete grid for two-dimensional grid structured data.

example, in balanced longitudinal data, missing values may result from subject dropouts. In the realm of spatio-temporal data, as illustrated in the previous chapter's air quality monitoring dataset, missing records can be attributed to malfunctioning monitoring devices. It is worth noting that in these scenarios, we are primarily concerned with missing values in the response variable y rather than the predictors.

We denote the observed portion of the response vector $\mathbf{y}$ as $\mathbf{y}_{obs}$ and the missing portion as $\mathbf{z}$. We denote the length of $\mathbf{y}_{obs}$ by n, and the complementary set of observations is denoted as $m = N - n$. Similarly, we partition the matrix $\mathbf{X}$ into $\mathbf{X}_{obs}$ and $\mathbf{X}_{ms}$. Notably, in the context of an incomplete grid, $\mathbf{X}_{ms}$ is not missing; instead, it represents a matrix that aggregates the predictors corresponding to the missing responses.

Handling missingness in the response variable within multidimensional grid data poses distinct challenges compared to non-grid data (Jafrasteh et al., 2023; Smola et al., 2005), or datasets with missing values in the predictor variables (Liu et al., 2018). One potential solution is to utilize only the observed part (complete-case analysis). However, when dealing with a large N, the value of n often remains substantial, making the naive implementation of GP models a challenging endeavour. The marginal distribution of $\mathbf{y}_{obs}$, given the portion of $\mathbf{X}$ that aligns with the observed section of the response, follows a multivariate normal distribution. However, it lacks a Kronecker product structure in the Gram matrix that could be leveraged for analysis.

5.2 Kronecker method with incomplete grid

To address this challenge, Gilboa et al. (2013) introduced the concept of imaginary observations within the missing grid. They demonstrated that these imaginary observations have no detrimental impact on posterior inference (including predictions) under certain prior assumptions about the imaginary observations. To estimate hyperparameters, an approximation to the marginal likelihood for the observed $\mathbf{y}_{obs}$ was employed. This approximation allows efficient computation. This approach has been demonstrated for Gaussian likelihood in applications like image inpainting and spatio-temporal temperature forecasting (Wilson et al., 2014), as well as for non-Gaussian likelihood in spatio-temporal crime rate forecasting (Flaxman et al., 2015). The kernel structures considered in these papers are separable. Note that the kernel is called separable when it can be separated by a tensor product. This includes the saturated ANOVA decomposition kernel. The proposed method is directly applicable to models featuring sum of separable kernels as well, provided the Gram matrix possesses an eigendecomposition of the form:

$$\mathbf{K} = \left(\bigotimes_{l=1}^{d} \mathbf{Q}_l \right) \mathbf{D} \left(\bigotimes_{l=1}^{d} \mathbf{Q}_l \right)^\top , \tag{5.1}$$

Here, $\mathbf{Q}_l$ represents an orthonormal matrix whose columns constitute the eigenvectors of the sub Gram matrix $\mathbf{K}_l$, and $\mathbf{D}$ is diagonal with positive entries. As discussed in Section 4.4.3, this condition holds for hierarchical ANOVA decomposition kernels, where each kernel k_l is empirically centred by (2.7).

5.2.1 Posterior distribution with missing grid

Even with incomplete grid data, Wilson et al. (2014) and Gilboa et al. (2013) show that it is possible to perform (asymptotically) exact predictive inference while still benefiting from a Kronecker product structure in the Gram matrix. The first step is to fill the missing part of the response $\mathbf{z}$ with imaginary observation $\mathbf{y}_{ms}$ and assume $\mathbf{y}_{ms} \sim N(0, w^{-1}\mathbf{I}_m)$. Given the observed ($\mathbf{y}_{obs}$, $\mathbf{X}_{obs}$) and hyper-parameter $\boldsymbol{\theta}$, we aim to obtain the posterior of f, which is a GP with mean and kernel given by:

$$\bar{m}(x) = \mathbf{k}_n(\mathbf{x})^\top \left(\mathbf{K}_{nn} + \sigma^2\mathbf{I}_n\right)^{-1} \mathbf{y}_{obs}, \qquad \mathbf{x} \in \mathcal{X} \tag{5.2}$$

$$\bar{k}(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') - \mathbf{k}_n(\mathbf{x})^\top \left(\mathbf{K}_{nn} + \sigma^2\mathbf{I}_n\right)^{-1} \mathbf{k}_n(\mathbf{x}'), \quad \mathbf{x}, \mathbf{x}' \in \mathcal{X}. \tag{5.3}$$

where the Gram matrix $\mathbf{K}_{nn}$ and the vector $\mathbf{k}_n(\mathbf{x})$ are evaluated only at $\mathbf{X}_{obs}$. However, it is costly to compute and store $\mathbf{K}_{nn}$ and evaluate $(\mathbf{K}_{nn} + \sigma^2\mathbf{I}_n)^{-1}\mathbf{y}_{obs}$. Due to the assumed Kronecker product structure (5.1), it is much more efficient to handle $\mathbf{K}_{NN}$. To utilise this, we consider the vector $\tilde{\mathbf{y}} = (\mathbf{y}_{obs}^\top, \mathbf{y}_{ms}^\top)^\top$. We have

$$\tilde{\mathbf{y}} \sim \text{MVN}\left(\mathbf{0}_N, \quad \mathbf{K}_{NN} + \mathbf{D}\right) \tag{5.4}$$

where

$$\mathbf{K}_{NN} = \begin{pmatrix} \mathbf{K}_{nn} & \mathbf{K}_{nm} \\ \mathbf{K}_{nm}^\top & \mathbf{K}_{mm} \end{pmatrix}$$

and

$$\mathbf{D} = \begin{pmatrix} \sigma^2\mathbf{I}_n & \mathbf{0}_{nm} \\ \mathbf{0}_{nm}^\top & w^{-1}\mathbf{I}_m \end{pmatrix}.$$

Note that $\mathbf{K}_{mm}$ is the covariance (Gram) matrix evaluated at each row of $\mathbf{X}_{ms}$ and $K_{nm} = K_{mn}^\top$ is the cross-covariance matrix with its i, j-th row given by $k(x_i^{(obs)}, x_j^{(ms)})$ for $i = 1, \ldots, n$ and $j = 1, \ldots, m$. Then the posterior mean and covariance kernel of f

given $\tilde{\mathbf{y}}$ and $\mathbf{X}$ are

$$\bar{m}^*(\mathbf{x}) = \mathbf{k}_N(\mathbf{x})^\top(\mathbf{K}_{NN} + \mathbf{D})^{-1}\tilde{\mathbf{y}}, \qquad \mathbf{x}, \in \mathcal{X} \tag{5.5}$$

$$\bar{k}^*(\mathbf{x}, \mathbf{x}') = k(\mathbf{x}, \mathbf{x}') - \mathbf{k}_N(\mathbf{x})^\top(\mathbf{K}_{NN} + \mathbf{D})^{-1}\mathbf{k}_N(\mathbf{x}'), \quad \mathbf{x}, \mathbf{x}' \in \mathcal{X}. \tag{5.6}$$

Using a standard formula for the inverse of a block matrix, it can be shown that as $w \to 0$,

$$\bar{m}^*(\mathbf{x}) \to \bar{m}(\mathbf{x}), \quad \bar{k}^*(\mathbf{x}, \mathbf{x}') \to \bar{k}(\mathbf{x}, \mathbf{x}'), \quad \mathbf{x}, \mathbf{x}' \in \mathcal{X}$$

See Gilboa et al. (2013); Wilson et al. (2014) for the proof. Note that due to the matrix $\mathbf{D}$ not being the multiple of the identity matrix, the method discussed in the previous chapter to compute $(\mathbf{K}_{NN} + \mathbf{D})^{-1}\tilde{\mathbf{y}}$ or $(\mathbf{K}_{NN} + \mathbf{D})^{-1}\mathbf{k}_N(x')$ does not directly apply; however we can efficiently evaluate them using a conjugate gradient method with the preconditioning matrix $\mathbf{D}^{-1/2}$ (PCG)[1]. If the eigendecomposition of $\mathbf{K}_{NN}$ is readily available[2], evaluating $(\mathbf{K}_{NN} + \mathbf{D})^{-1}\tilde{\mathbf{y}}$ costs $O(JN\sum_{l=1}^{d} n_l)$ for a $d-$dimensional grid data, where J denotes the number of iterations required for PCG. Note for a complete grid, this was $O(N\sum_{l=1}^{d} n_l)$. The missing grids can be filled using e.g., posterior mean (5.5), evaluated at each grid.

5.2.2 Hyper-parameter estimation

To estimate the hyper-parameters, $\boldsymbol{\theta}$, we consider maximum marginal likelihood estimation. Let $p(\mathbf{y}_{obs}|\boldsymbol{\theta})$ be the marginal likelihood based on all observed responses.

[1] Instead of solving these directly, we can solve the system of linear equations $(\mathbf{K}_{NN} + \mathbf{D})\mathbf{v} = \mathbf{b}$ for $\mathbf{v}$. Here $\mathbf{b}$ is either $\mathbf{y}$ or $\mathbf{k}_N(x')$. Using preconditioning matrix and solve $\mathbf{D}^{-1/2}(\mathbf{K}_{NN} + \mathbf{D})\mathbf{D}^{-1/2}\mathbf{v} = \mathbf{D}^{-1/2}\mathbf{b}$ improves convergence.

[2] This is the case if we use saturated ANOVA decomposition kernel, or non-saturated hierarchical ANOVA decomposition kernel with centred kernel to construct $\mathbf{K}_{NN}$.

We treat the predictors as constant here. The log marginal likelihood is given by

$$\log p(\mathbf{y}_{obs}|\boldsymbol{\theta}, \mathbf{X}_{obs}) = -\frac{1}{2}\mathbf{y}_{obs}^\top \left(\mathbf{K}_{nn} + \sigma^2\mathbf{I}_n\right)^{-1} \mathbf{y}_{obs} - \frac{1}{2}\log|\mathbf{K}_{nn} + \sigma^2\mathbf{I}_n| - \frac{n}{2}\log 2\pi.$$

The first term can be replaced by $-\frac{1}{2}\tilde{\mathbf{y}}^\top(\mathbf{K}_{NN} + \mathbf{D})^{-1}\tilde{\mathbf{y}}$ with the same reasoning in the previous section. The second term involving the log determinant is more complicated. Gilboa et al. (2013) and Wilson et al. (2014) proposed an approximation by

$$\log|\mathbf{K}_{nn} + \sigma^2\mathbf{I}_n| \approx \sum_{i=1}^{n} \log\left(\tilde{\lambda}_i^n + \sigma^2\right) \tag{5.7}$$

where $\tilde{\lambda}_i^n = \frac{n}{N}\lambda_i^N$ for $i = 1, \ldots, n$, and $\lambda_1^N, \ldots, \lambda_n^N$ are the n largest eigenvalues of the Gram matrix $\mathbf{K}_{NN}$. The approximation is asymptotically consistent (Baker and Taylor, 1979). The advantage of this method is its relatively low computational cost. For each update of $\boldsymbol{\theta}$, in addition to other computational costs required for the standard Kronecker GP approach discussed in Section 4.4.4, evaluating the first term takes $O(JN\sum_{l=1}^d n_l)$ operations as seen in Section 5.2.1. The second term is simply rescaling $n < N$ eigenvalues, thus is linear in N.

5.2.3 Missing data mechanism

When dealing with missing data, it is also important to consider missingness mechanisms. Rubin (1976) described three missingness mechanisms: Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing Not at Random (MNAR). Under the assumption of MCAR, the probability of missing does not depend on any observed or missing variables and is equal to all units, and a complete case analysis does not lead to a biased estimator. If MAR, the missing mechanism can be explained by other recorded variables. In such cases, a complete case analysis can be used if these covariates are included in the model. However, unlike the MCAR assumption, the MAR assumption is generally challenging to test, and the missing mechanism in many real-world cases is MNAR. This includes when the missingness depends on the unobserved predictors or missing value itself. For example, in environmental monitoring of e.g., temperature, the malfunctioning of the recording device leads to missing records, but the cause of malfunctioning may be related to unobserved/unrecorded factors such as the age of the device at each location or extreme temperature. Failing to account for missing data mechanisms can produce biased estimates. The simple method discussed in Section 5.2.1 and 5.2.2 only works under MCAR or MAR assumption. We see this in Section 5.3.

5.2.4 Considerations for other methods

Another possible approach is to use an iterative algorithm, such as the expectation-maximization algorithm (Dempster et al., 1977) or one of its variants, that sequentially imputes the missing grid $\mathbf{z}$ given and updates the value of hyper-parameters using complete-data marginal likelihood. With the EM algorithm, while the E-step have a closed form solution, it requires computing the conditional mean and covariance of $p(\mathbf{z}|\mathbf{y}_{obs}, \boldsymbol{\theta}^{(t-1)})$ at each iteration. The entries of the mean vector of length m and the $m \times m$ covariance matrix have to be computed using the PCG algorithm. While the mean vector can be computed in $O(JN\sum_{l=1}^{d} n_l)$ time, for the covariance matrix, this operation has to be repeated $\frac{m(m+1)}{2}$ times. Here m denotes the number of missing grids. With EM, the covariance matrix also has to be inverted for each update of the hyperparameters. Stochastic variants, such as Monte-Carlo EM (Wei and Tanner, 1990), stochastic EM (Celeux and Diebolt, 1985; Diebolt and Ip, 1995) and the stochastic approximation (Robbins and Monro, 1951) version of EM (Delyon et al., 1999) may be considered. On a related topic, Kim and Leskovec (2011) proposed an MCEM algorithm for Kronecker graphs with missing nodes and edges; however, the sampling step for the stochastic variants still involves sampling from the conditional distribution of the missing response $\mathbf{z}$, which is m dimensional. If naively done, this requires the computation of the $m \times m$ covariance matrix. Applying these iterative algorithms to a dataset with a large number of missing responses is challenging.

The EM algorithm has been used for GP models with multivariate responses in Bonilla et al. (2007) where the values of some response variables are missing for some observations. This has a close connection to GP models for multi-dimensional grid data, as both approaches involve covariance matrices expressed using the Kronecker product. In their approach, the vector $\mathbf{f}$ evaluated at each row of $\mathbf{X}$ is treated as missing. However, the derived algorithm still involves taking expectations with regard to the conditional distribution of $\mathbf{y}_{obs}$, which poses a computational issue similar to the one discussed.

Nevertheless, the EM algorithm or its variants have an advantage over the method discussed in Section 5.2.1 and 5.2.2, as more complex missingness mechanism can be modelled and incorporated into the algorithm. This can be an attractive option if the sampling scheme in MCEM or stochastic approximation EM can be improved.

Other approaches in the GP literature for addressing missing values in multidimensional grids include Imani et al. (2019), who proposed the adoption of nested GP models within this context, and Hori et al. (2016), who introduced an iterative algorithm employing regularized Principal Component Analysis. It is worth noting that both of these studies (implicitly or explicitly) assume the presence of saturated interactions. Therefore, the potential applicability of these methodologies to non-saturated models warrants further investigation.

5.3 Simulation

We investigate the performance of the discussed method with synthetic data. We consider 2-dimensional grid data on two predictors $x_1 \in \mathcal{X}_1$ and $x_2 \in \mathcal{X}_2$. We generate data by the following additive structure:

$$y = f_1(x_1) + f_2(x_2) + \epsilon,$$

where $\epsilon \sim N(0, 1.5^2)$. The functions $f_1 : \mathcal{X}_1 \to \mathbb{R}$ and $f_2 : \mathcal{X}_2 \to \mathbb{R}$ are given by

$$\begin{aligned} f_1(x_1) &= -10 + 33\phi(x_1|1, 0.8^2) + 2\phi(x_1|2.5, 0.25^2) \\ &\quad + 65\phi(x_1|4, 1.5^2) + (\exp(1.25(x_1 - 4.5)) - 1)\, I(x_1 > 4.5) \\ f_2(x_2) &= 5 + 43\phi(x_2|0, 1.2^2) + 2\phi(x_2|2.5, 0.25^2) \\ &\quad + 55\phi(x_2|4, 1.2^2) + (\exp(-1.25x_2) - 1)\, I(x_2 < 0) \end{aligned}$$

where $\phi(x|\mu, \sigma^2)$ is the probability density function of a normal distribution with mean μ and variance σ^2. See Figure 5.2 for illustration. We let $\mathcal{X}_1$ and $\mathcal{X}_2$ be a set of $\sqrt{N}$ equispaced points on an interval $\{-1, 6\}$, hence the data are on a 2-dimensional grid $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$. We drop m observations denoted by $\mathbf{z} = (z_1, \ldots, z_m)^\top$ from the total of N observations (i.e., the proportion of the missing values is $\frac{m}{N}$), based on the (1) MCAR, (2) MAR and (3) MNAR assumptions. For (2) we drop the response y depending on the value of x_1 and x_2. For (3), the largest $100\frac{m}{N}\%$ y-values are treated as missing. We vary the size of the grid N and the proportion of missing data $\frac{m}{N}$ and check the root mean squared error (RMSE) of the predicted value. We also check the RMSE for σ. See Figure 5.3 and 5.4 for an example of the simulated data and the three missing mechanisms with $N = 70^2$. We assume a zero mean GP prior on f with

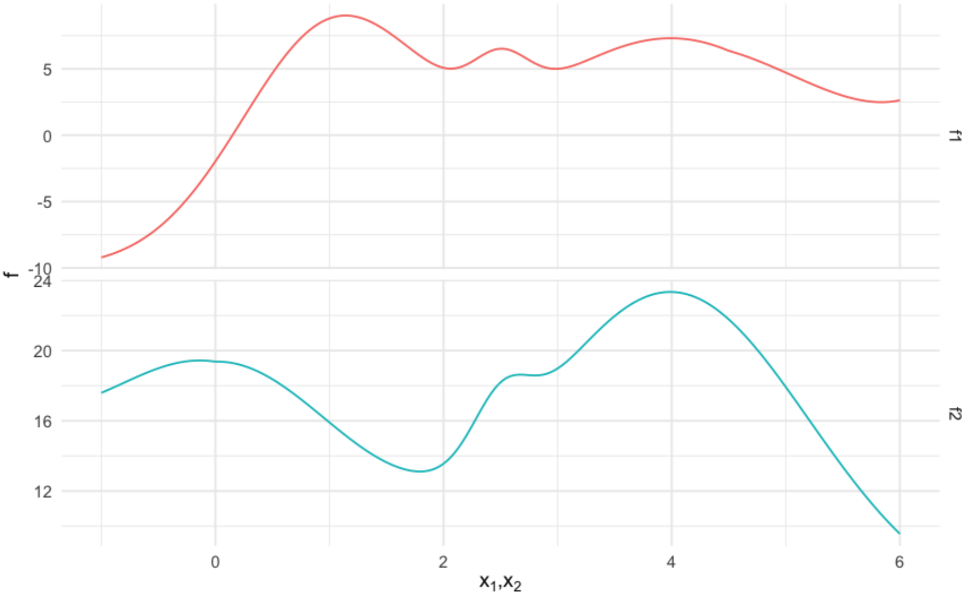


Fig. 5.2 Functions $f_1(x_1)$ and $f_2(x_2)$ for data generation

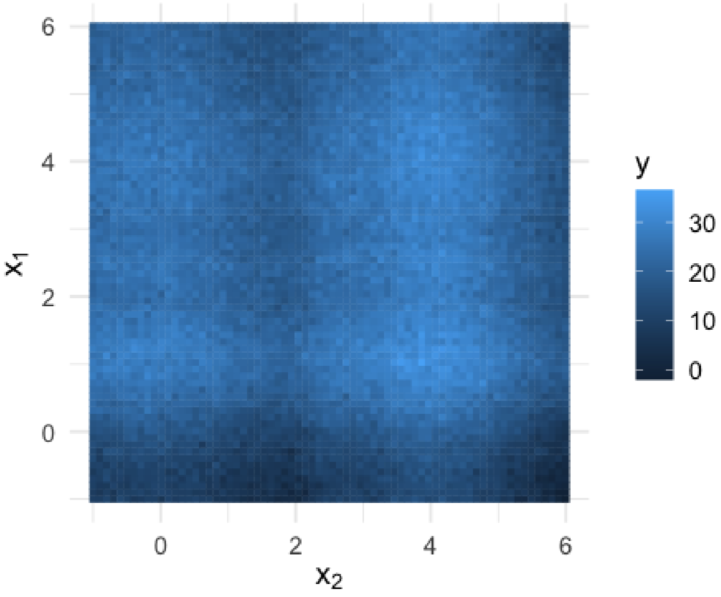
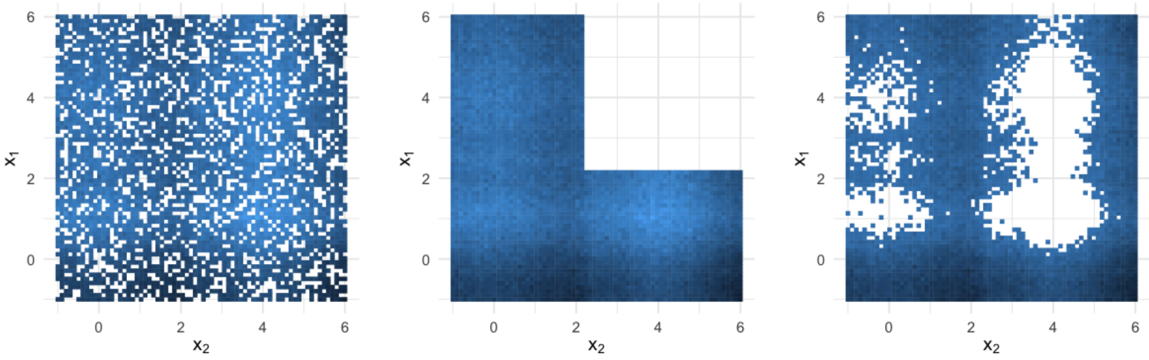


Fig. 5.3 Synthetic data, $N = 70^2$



(a) MCAR (b) MAR (c) MNAR

Fig. 5.4 Three missing data mechanisms for the synthetic data with the grid size 70 × 70 and the missing proportion 30%.

hierarchical ANOVA kernel

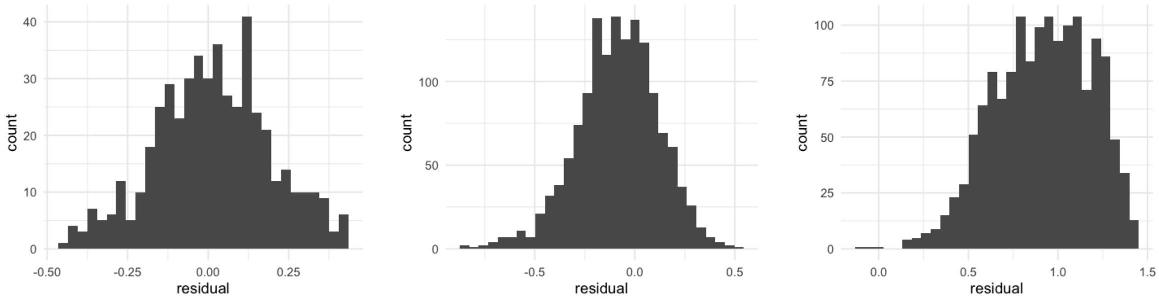
$$k((x_1, x_2)^\top, (x'_1, x'_2)^\top) = \alpha_0^2(1 + k_1(x_1, x'_1) + k_2(x_2, x'_2))$$

with squared and centred Brownian motion kernel ($\gamma = \frac{1}{2}$) for k_1 and k_2. Therefore we have hyperparameters $\boldsymbol{\theta} = (\alpha_0, \alpha_1, \alpha_2, \sigma_\epsilon)^\top$ where α_1, α_2 are scale parameters for k_1 and k_2, and σ_ϵ is the standard deviation of the error term.

Table 5.1 shows the mean of the parameter estimate for σ, the RMSE for σ, the mean of the RMSE for $f = f_1 + f_2$ and the average running time under the three different missing mechanisms and with the varying missing proportions (10%,20% and 30%) for a 70 × 70 grid. These are based on 20 replications. Note that the true values of scale parameters are unknown, but we provide their estimates in Table C.4 in the Appendix C.4. We notice that under MCAR and MAR, the RMSE of σ is very small. Under MNAR, the standard deviation of the error term is underestimated, as well as the imputed values for $\mathbf{z}$. See Figure 5.5 for the residuals $\hat{z}_i - y_i^{true}$ where $\hat{z}_i$ is the posterior predictive mean of a missing response y_i^{true} given by (5.5). We can see that for the MNAR case, the distribution of the residuals shown in Figure 5.5c is not centred around zero and is skewed to the left. In terms of the computational cost,

Table 5.1 RMSEs for the parameters and for missing grid. Running time is measured in seconds. The synthetic data with 70 × 70 grid size. For each scenario, the experiment is repeated 20 times. See Table C.4 for more information.

	MCAR			MAR			MNAR		
	10%	20%	30%	10%	20%	30%	10%	20%	30%
$\bar{\sigma}$	1.5	1.49	1.49	1.5	1.5	1.5	1.45	1.43	1.42
RMSE-σ	0.02	0.02	0.023	0.018	0.017	0.019	0.051	0.074	0.085
RMSE-f	0.16	0.17	0.18	0.17	0.19	0.22	0.73	0.89	1.01
Time(s)	138	146	141	111	110	104	155	147	141



(a) MCAR (b) MAR (c) MNAR

Fig. 5.5 The histograms of residuals for the synthetic data of grid size 70 × 70 and the missing proportion 30%.

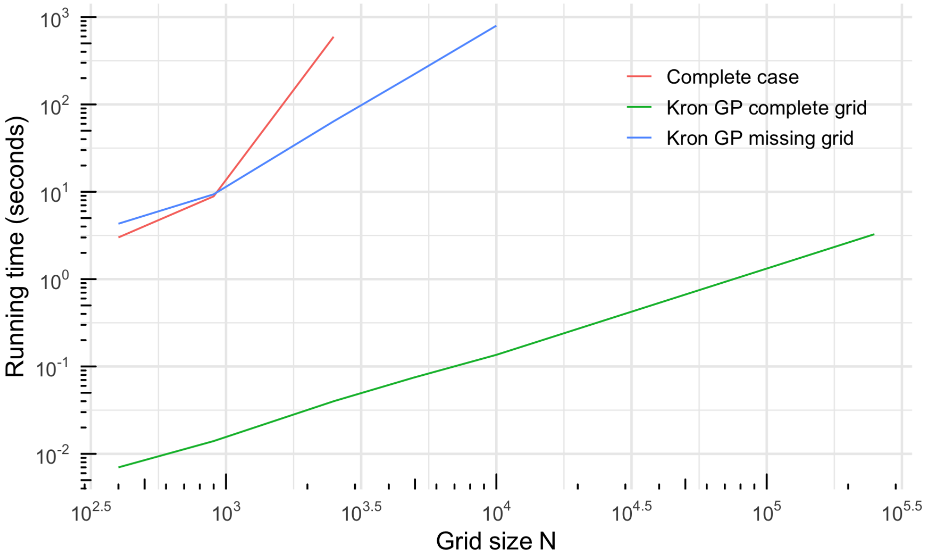


Fig. 5.6 Running time of Kronecker GP methods and complete case analysis. For complete case analysis and the Kroncker method with a missing grid, the missing proportion is 10%, and the missingness mechanism is MAR.

although we do not see the same scalability of the Kronecker method for a complete grid structure, exploiting the Kronecker structure with the use of imaginary observations on the missing grid is more efficient compared to naively implementing a complete case analysis (See Figure 5.6).

5.4 Discussion

In this chapter, we delve into the challenge posed by incomplete grids in the Kronecker method for GP regression. The approach proposed by Gilboa et al. (2013); Wilson et al. (2014) represents an approximation akin to complete case analysis, holding an advantage over a naive implementation. For datasets with a grid size of N and m missing values, the time complexity is $O(JN\sum_{l=1}^{d} n_l)$ for the former and $O(N(1-r)^3)$ for the latter, where J stands for the number of iterations required for the PCG algorithm, and $r = m/N$ denotes the proportion of missing values on the grid. The complexity of the approximated Kronecker method remains unaffected by the number or proportion of missing grids directly, though they may influence the number of iterations J in practice. This is in stark contrast to iterative methods like the EM algorithm, which also scales with the number of missing values m.

While the discussed Kronecker method has primarily been applied to models featuring separable kernels, it can be readily extended to accommodate models constructed using a non-saturated hierarchical ANOVA kernel. We have assessed the method's performance for such a model under varying assumptions regarding the mechanisms of missingness. Our simulation study, conducted on 2-dimensional grid-structured data, demonstrates that the method performs admirably for Missing Completely at Random (MCAR) and Missing at Random (MAR) situations. In the case of MAR, we assume that the missingness can be elucidated by the covariates, which, in this instance, form a multidimensional grid and are thus incorporated into the model. However, since the method approximates complete case analysis by employing likelihood that ignores the missing data mechanism, it falters under the Missing Not at Random (MNAR) assumption. It is important to note that the simulation setting is limited in scope, with one synthetic data and assuming simple missing data mechanisms. In many real-world examples, missing mechanisms are often more complex. For instance, in environmental monitoring, if machine malfunctions are the cause of the missingness, it is likely that the measurements are not recorded until the machine is fixed. The length of such consecutive missing values may depend on various factors and also needs to be taken into account. Further studies investigating different missing data mechanisms are needed.

As outlined in the preceding chapter, the Kronecker method exhibits significant utility across various fields, including environmental, psychological, medical, and behavioural studies where repeated measurements or longitudinal data are common. In these contexts, a prevalent challenge arises with missing data, often accompanied by an

MNAR mechanism. In such situations, it becomes imperative to integrate the missing data mechanism into the model. Consequently, future work should prioritize addressing this challenge within the framework of the Kronecker method. More specifically, incorporating the missingness mechanism into the EM algorithm will be an attractive option if the computational efficiency can be improved.

Chapter 6

Conclusions

In this chapter, we present an overview of the thesis, highlighting its contributions, and outlining potential directions for future research.

Within this thesis, our primary focus centred on developing a statistical model for a response variable y and a set of predictors $x = (x_1, x_2, \ldots, x_d)^\top$ where the regression function $f(x)$ can be decomposed into the addition of several functions. Each x_l belongs to a set $\mathcal{X}_l$, and we write $\mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_d$. The simplest form of this model is known as the *main effect* model, wherein we make the following assumption:

$$f(x) = a + \sum_{l=1}^{d} f_l(x_l),$$

This assumption implies that each predictor x_l possesses an individual and distinct relationship with the response variable y. We assumed an additive Gaussian process (GP) prior. Specifically, we assume that each component function f_l follows a GP with prior distribution $\mathrm{GP}(0, k_l)$, where the kernel function $k_l : \mathcal{X}_l \times \mathcal{X}_l \to \mathbb{R}$ plays a pivotal role in characterizing the relationship between predictor x_l and response y. Notably, the sum of these individual functions, $\sum_{l=1}^{d} f_l$, also follows a GP with a kernel determined by the summation of individual kernels, as expressed by: $k = \sum_{l=1}^{d} k_l$.

Chapter 2 serves as a foundational introduction within the thesis, elucidating key concepts related to GPs, kernels, and regression with GP priors. Additionally, it establishes a contextual framework by exploring the connections between the GP regression and other classical statistical methodologies, including kernel ridge regression, Kriging (Krige, 1951), and conditional autoregressive models. This chapter primarily concerned the regression model with $d = 1$ or $d = 2$. In the latter case ($d = 2$), we introduced a specific example of an additive GP model, called the *semi-parametric GP* model, where a linear kernel is added to another kernel of choice, typically of a non-linear nature. This combination yields a model that can capture both linear and non-linear relationships, thereby enhancing its utility for various practical applications.

Chapter 3 then generalised this framework to $d \geq 2$, and also considered *interaction effect* models. In these models, functions representing the interaction effects between different variables, denoted as $f_{ll'}(x_l, x_{l'})$, are incorporated into the regression function alongside the main effect model. Instead of introducing new kernels for these interaction effect functions, a more parsimonious approach is adopted using the tensor product kernel $k_{ll'} = k_l \otimes k_{l'}$. This approach draws inspiration from the field of smoothing splines, particularly the ANOVA decomposition of functions in a Reproducing Kernel Hilbert Space (Gu, 2002; Gu and Wahba, 1993; Wahba, 1990). By adapting this concept into the decomposition of kernels, Stitson et al. (1999) introduced the *ANOVA decomposition kernel*, also known simply as the *ANOVA kernel*. Within the context of GP regression, we employed a more parsimonious specification of the ANOVA kernel that also adheres to common practices of statistical modelling to model interaction effects. This offers several advantages, such as maintaining the same number of model parameters for both main and interaction models, and its inherent additive nature enhances interpretability (Buja et al., 1989).

In addition to the saturated interaction model with the saturated ANOVA kernel, which encompasses all possible interactions, including the highest-ordered and all lower-ordered interactions, we consider a parsimonious model structure that includes only interactions with substantive effects while preserving a hierarchical interaction structure, using the *hierarchical ANOVA decomposition kernel* (Bergsma and Jamil, 2023). The data analysis of spatio-temporal disease classification highlighted the importance of hierarchical modelling of spatial and temporal effects.

The use of the ANOVA kernel, when centred, further enhances interpretability. This is particularly attractive when investigating lower-order interaction effects. To illustrate this, we examined a case involving longitudinal data, where subjects are assigned to different treatment groups, and their weights are measured. Understanding the variations in growth curves among different treatment groups, considering individual variability and its interaction with treatment effects, is often intricate. however, the additive GP model with centred (hierarchical) ANOVA kernels yields interpretable outputs and provides the average growth curves for each treatment group.

In Chapter 4, we shift our focus to the computational challenges inherent in GP models. These challenges primarily stem from the extensive operations involving an $n \times n$ covariance matrix, resulting in a time complexity of $O(n^3)$ and a space requirement of $O(n^2)$. We considered large-scale multidimensional grid data, where the predictors $x_1, x_2, \ldots, x_d$ collectively form a Cartesian grid $\mathcal{X}_1 \times \mathcal{X}_2 \times \ldots \times \mathcal{X}_d$. Within this structured data, the covariance matrix can be expressed using the Kronecker product. By exploiting the properties of the Kronecker product, we can evaluate the marginal likelihood and the posterior mean and covariance without explicitly forming the full covariance matrix. It is sufficient to evaluate only the covariance matrix on each level, which is of order $n_l = |\mathcal{X}_l|$. In the literature, the kernels called *separable*, i.e. the kernels that have the form: $k = \tilde{k}_1 \otimes \ldots \otimes \tilde{k}_p$ where each $\tilde{k}_l$ is defined on $\mathcal{X}_l$, were considered (Flaxman et al., 2015; Gilboa et al., 2013; Saatçi, 2012; Wilson et al., 2014). We extend the work in the literature to the special case of the sum of separable kernels, which are constructed using centred (non-saturated) hierarchical ANOVA kernels. This expansion allows us to efficiently compare models with varying interaction structures, thereby aiding in the selection of a more parsimonious model, provided that such a model is adequate.

One of the primary limitations associated with the Kronecker product approach is its reliance on a comprehensive grid, necessitating the recording of the response variable for every input point within the Cartesian grid $\mathcal{X}_1 \times \ldots \times \mathcal{X}_p$. Chapter 5 addresses the issue by applying and assessing a methodology introduced by Gilboa et al. (2013) and Wilson et al. (2014) to models utilizing non-saturated ANOVA kernels. The proposed algorithm allows us to benefit from the Kronecker product structure even with an incomplete grid. We considered various missingness mechanisms and concluded that for Missing completely at Random and Missing at Random scenarios, the method performs admirably. Moreover, it preserved a notable computational advantage when compared to the naive implementation of complete case analysis.

The key contribution of the thesis to the field of Gaussian process models is to provide a statistical modelling framework with GP priors, especially for additive interaction models, to enhance the interpretability of additive GP models, and to offer a computationally efficient implementation of this methodology tailored for multi-dimensional grid data, which is frequently encountered in the realms of spatio-temporal analysis and longitudinal data analysis. We conclude the thesis by listing a set of open questions and possible future research.

- **Kronecker product approach with non-Gaussian likelihood**. Chapter 4 concerned the Kronecker product approach with Gaussian likelihood. For non-Gaussian likelihood and *separable* kernels, Flaxman et al. (2015) proposed the Kronecker method with Laplace approximation. It is worth noting that the Kronecker product structure can also facilitate likelihood evaluation within the context of Markov Chain Monte Carlo (MCMC) methods.

While the application of this approach may initially appear straightforward, a noteworthy complication arises when dealing with the centred (sum of separable, non-saturated) ANOVA decomposition kernel. The crucial property we leverage for the covariance matrix decomposition is that each sub-covariance matrix possesses, at most, $n_l - 1$ non-zero eigenvalues. With non-Gaussian likelihoods, this property implies that the model covariance matrix is always singular. Taking the inverse and log-determinant of this matrix can be avoided for some steps in Laplace approximation or expectation propagation (Section 2.3) as discussed in Rasmussen and Williams (2006, Chapter 3). For the remaining parts, using a pseudoinverse may be a viable option. Nevertheless, it is imperative to thoroughly investigate the performance and numerical stability of algorithms involving additive Kronecker products. Furthermore, exploring the possibility of employing more advanced algorithms, such as variational inference, presents another path for future research.

- **Kronecker product approach with additional predictors**. Within this thesis, our primary assumption has been that only one set of predictors is observed at each level of multi-dimensional grid data. However, in many real-world datasets, additional covariates may be available, introducing an added layer of complexity. For instance, when monitoring air-pollutant concentrations in a spatio-temporal context alongside geographical coordinates and timestamps, we may possess supplementary information, such as the station type or the proximity to a road. Notably, in this scenario, the additional covariates typically vary only at one level of the data hierarchy.

As outlined in Section 4.6, we have the option to incorporate a distinct kernel, such as a linear kernel, to capture the effects of these additional covariates. This supplementary kernel can be combined with the existing kernel, which, in this instance, accounts for spatial variation. If both kernels are centred, the resulting sum kernel remains centred as well. However, if we assume interactions between these two kernels, necessitating the introduction of an additional tensor product kernel, the resulting composite kernel is not centred in general. Consequently, this limitation restricts the applicability of the Kronecker approach to separable kernels in such scenarios.

Moreover, when dealing with cross-level covariates, we encounter a similar issue. If the effects of these covariates are linear and do not interact with the predictors that form the multidimensional grid, the model can still be efficiently evaluated by exploiting the Kronecker product structure. However, complications arise when interactions between covariates and predictors are present, rendering the model more complex and challenging to analyze using the Kronecker approach.

- **Gaussian process models with a missing value under Missing not at Random scenario**. In Chapter 5.2.2, we arrive at the conclusion that the existing Kronecker method, when applied to an incomplete grid, falls short when confronted with a Missing Not at Random scenario. In cases where the missingness mechanism is non-random, it becomes imperative to incorporate this mechanism into the modeling process. One viable avenue for achieving this is through methods like the Expectation-Maximization (EM) algorithm. However, it's worth noting that the scalability of the EM algorithm becomes problematic as the number of missing values increases.

  Even when employing more computationally efficient alternatives, such as stochastic approximation EM, the need to sample from the conditional distribution of the missing responses at each iteration outweighs the advantages offered by the Kronecker product approach. This underscores the need for a computationally efficient sampling approach that can better accommodate the complexities introduced by missing data in such contexts.

- **Multivariate response**. The proposed methodology, employing the ANOVA decomposition kernel, along with its efficient implementation method via the Kronecker product, naturally lends itself to handling multivariate responses. Consider, for instance, a scenario where measurements at each monitoring station encompass the concentrations of multiple air pollutants. When dealing with data involving multivariate responses, we can conceptually view each class of measurement as a categorical predictor, effectively treating the data as multidimensional grid data. Under this perspective, the model's covariance matrix readily involves the Kronecker product.

  By adopting the ANOVA kernel, we can incorporate variations in the means attributed to different classes (main effect) and explore how the influence of other covariates, such as temporal or spatial patterns, varies across these different classes (interaction effect). However, addressing multivariate responses introduces a unique challenge—selecting and estimating an appropriate kernel for each class. This becomes particularly complex when the covariance structure itself must be estimated, as it may not always be straightforward to apply centring to the kernel in this context.

References

Adler, R. J. (1981). *The geometry of random fields*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Chichester.

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle–in: Second international symposium on information theory (eds) bn petrov, f. *Csaki. BNPBF Csaki Budapest: Academiai Kiado*.

Alvarez, M. A., Rosasco, L., Lawrence, N. D., et al. (2012). Kernels for vector-valued functions: A review. *Foundations and Trends® in Machine Learning*, 4(3):195–266.

Ansley, C. F. and Kohn, R. (1985). Estimation, filtering, and smoothing in state space models with incompletely specified initial conditions. *The Annals of Statistics*, 13(4):1286–1316.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404.

Avrachenkov, K., Chebotarev, P., and Rubanov, D. (2019). Similarities on graphs: Kernels versus proximity measures. *European Journal of Combinatorics*, 80:47–56.

Baker, C. T. and Taylor, R. (1979). The numerical treatment of integral equations. *Journal of Applied Mechanics*, 46(4):969.

Bennett, C. H. (1976). Efficient estimation of free energy differences from Monte Carlo data. *Journal of Computational Physics*, 22(2):245–268.

Bergsma, W. (2020). Regression with I-priors. *Econometrics and Statistics*, 14:89–111.

Bergsma, W. and Jamil, H. (2023). Additive interaction modelling using I-priors. *arXiv preprint arXiv:2007.15766*.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225.

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics*, 43:1–20.

Bonilla, E. V., Chai, K., and Williams, C. (2007). Multi-task Gaussian process prediction. *Advances in neural information processing systems*, 20.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly weather review*, 78(1):1–3.

Buja, A., Hastie, T., and Tibshirani, R. (1989). Linear smoothers and additive models. *The Annals of Statistics*, 17(2):453–510.

Carlin, B. P., Banerjee, S., et al. (2003). Hierarchical multivariate CAR models for spatio-temporally correlated survival data. *Bayesian statistics*, 7(7):45–63.

Celeux, G. and Diebolt, J. (1985). The SEM algorithm: a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational statistics quarterly*, 2:73–82.

Chai, K. M. A. (2012). Variational multinomial logit Gaussian process. *The Journal of Machine Learning Research*, 13(1):1745–1808.

Cheng, L., Ramchandran, S., Vatanen, T., Lietzén, N., Lahesmaa, R., Vehtari, A., and Lähdesmäki, H. (2019). An additive Gaussian process regression model for interpretable non-parametric analysis of longitudinal data. *Nature communications*, 10(1):1798.

Cooper, M. J., Martin, R. V., Hammer, M. S., Levelt, P. F., Veefkind, P., Lamsal, L. N., Krotkov, N. A., Brook, J. R., and McLinden, C. A. (2022). Global fine-scale changes in ambient NO2 during COVID-19 lockdowns. *Nature*, 601(7893):380–387.

Cressie, N. and Huang, H.-C. (1999). Classes of nonseparable, spatio-temporal stationary covariance functions. *Journal of the American Statistical association*, 94(448):1330–1339.

Cressie, N., Perrin, O., and Thomas-Agnan, C. (2005). Likelihood-based estimation for Gaussian MRFs. *Statistical Methodology*, 2(1):1–16.

Cressie, N. A. (1993). *Statistics for spatial data.* Wiley Series in Probability and Statistics. John Wiley & Sons, Inc, New York.

Dance, H. and Paige, B. (2022). Fast and scalable spike and slab variable selection in high-dimensional Gaussian processes. In *International Conference on Artificial Intelligence and Statistics*, pages 7976–8002. PMLR.

Datta, A., Banerjee, S., Finley, A. O., Hamm, N. A., and Schaap, M. (2016). Nonseparable dynamic nearest neighbor Gaussian process models for large spatio-temporal data with an application to particulate matter analysis. *The annals of applied statistics*, 10(3):1286–1316.

De Oliveira, V. (2012). Bayesian analysis of conditional autoregressive models. *Annals of the Institute of Statistical Mathematics*, 64:107–133.

Delyon, B., Lavielle, M., and Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *Annals of statistics*, 27(1):94–128.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

Diebolt, J. and Ip, E. H. (1995). Stochastic em: method and application. In Fagerberg, J., Mowery, D. C., and Nelson, R. R., editors, *Markov chain Monte Carlo in practice*, chapter 15, pages 259–273. CRC press, London.

Diggle, P., Zheng, P., and Durr, P. (2005). Nonparametric estimation of spatial segregation in a multivariate point process: bovine tuberculosis in Cornwall, UK. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(3):645–658.

Diggle, P. J., Moraga, P., Rowlingson, B., and Taylor, B. M. (2013). Spatial and spatio-temporal log-Gaussian Cox processes: extending the geostatistical paradigm. *Statistical Science*, 28(4):542–563.

Dutta, V., Kumar, S., and Dubey, D. (2021). Recent advances in satellite mapping of global air quality: evidences during COVID-19 pandemic. *Environmental Sustainability*, 4:469–487.

Duvenaud, D., Lloyd, J., Grosse, R., Tenenbaum, J., and Zoubin, G. (2013). Structure discovery in nonparametric regression through compositional kernel search. In *International Conference on Machine Learning*, volume 28, pages 1166–1174.

Duvenaud, D., Nickisch, H., and Rasmussen, C. (2011). Additive Gaussian processes. In *Advances in Neural Information Processing Systems*, volume 24.

Embrechts, P. and Maejima, M. (2002). Selfsimilar processes, princeton ser. *Appl. Math., Princeton University Press, Princeton, NJ.*

Flaxman, S., Wilson, A., Neill, D., Nickisch, H., and Smola, A. (2015). Fast Kronecker inference in Gaussian processes with non-Gaussian likelihoods. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 607–616.

Fonseca, T. C. and Steel, M. F. (2011). Non-Gaussian spatiotemporal modelling through scale mixing. *Biometrika*, 98(4):761–774.

Fouss, F., Francoisse, K., Yen, L., Pirotte, A., and Saerens, M. (2012). An experimental investigation of kernels on graphs for collaborative recommendation and semisupervised classification. *Neural networks*, 31:53–72.

Gelfand, A. E. and Vounatsou, P. (2003). Proper multivariate conditional autoregressive models for spatial data analysis. *Biostatistics*, 4(1):11–15.

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Statistics and computing*, 24(6):997–1016.

Genton, M. G. (2001). Classes of kernels for machine learning: a statistics perspective. *Journal of Machine Learning Research*, 2:299–312.

Gilboa, E., Saatçi, Y., and Cunningham, J. P. (2013). Scaling multidimensional inference for structured Gaussian processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):424–436.

Gneiting, T. (2002). Nonseparable, stationary covariance functions for space–time data. *Journal of the American Statistical Association*, 97(458):590–600.

Goovaerts, P. (1997). *Geostatistics for natural resources evaluation.* Oxford University Press, New York.

Groot, P., Peters, M., Heskes, T., and Ketter, W. (2014). Fast Laplace approximation for Gaussian processes with a tensor product kernel. In *The 26th Benelux Conference on Artificial Intelligence.*

Gu, C. (2002). *Smoothing spline ANOVA models.* Springer Series in Statistics. Springer, New York.

Gu, C. and Wahba, G. (1993). Semiparametric analysis of variance with tensor product thin plate splines. *Journal of the Royal Statistical Society: Series B (Methodological)*, 55(2):353–368.

Hastie, T. and Tibshirani, R. (1990). *Generalized additive models.* CRC Press, Boca Raton.

Hensman, J., Fusi, N., and Lawrence, N. D. (2013). Gaussian processes for big data. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence*, pages 282–290.

Hensman, J., Matthews, A., and Ghahramani, Z. (2015). Scalable variational Gaussian process classification. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, volume 38, pages 351–360.

Higham, J., Ramírez, C. A., Green, M., and Morse, A. (2021). UK COVID-19 lockdown: 100 days of air pollution reduction? *Air Quality, Atmosphere & Health*, 14:325–332.

Hori, T., Montcho, D., Agbangla, C., Ebana, K., Futakuchi, K., and Iwata, H. (2016). Multi-task Gaussian process for imputing missing data in multi-trait and multi-environment trials. *Theoretical and Applied Genetics*, 129:2101–2115.

Huang, J. Z. (1998). Projection estimation in multiple regression with application to functional ANOVA models. *The annals of statistics*, 26(1):242–272.

Imani, F., Cheng, C., Chen, R., and Yang, H. (2019). Nested Gaussian process modeling and imputation of high-dimensional incomplete data under uncertainty. *IISE Transactions on Healthcare Systems Engineering*, 9(4):315–326.

Jafrasteh, B., Hernández-Lobato, D., Lubián-López, S. P., and Benavente-Fernández, I. (2023). Gaussian processes for missing value imputation. *Knowledge-Based Systems*, 273:110603.

Jamil, H. (2018). *Regression modelling using priors depending on Fisher information covariance kernels (I-priors).* PhD thesis, London School of Economics and Political Science.

Jamil, H. and Bergsma, W. (2020). Bayesian variable selection for linear models using I-priors. In *Theoretical, Modelling and Numerical Simulations Toward Industry 4.0*, pages 107–132. Springer.

Jephcote, C., Hansell, A. L., Adams, K., and Gulliver, J. (2021). Changes in air quality during COVID-19 'lockdown' in the United Kingdom. *Environmental Pollution*, 272:116011.

Journel, A. G. and Huijbregts, C. J. (1976). Mining geostatistics.

Kandola, J., Cristianini, N., and Shawe-taylor, J. (2002). Learning semantic similarity. *Advances in neural information processing systems*, 15.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795.

Katz, L. (1953). A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43.

Kenward, M. G. (1987). A method for comparing profiles of repeated measurements. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 36(3):296–308.

Khan, E., Mohamed, S., and Murphy, K. P. (2012). Fast Bayesian inference for non-conjugate Gaussian process regression. In *Advances in Neural Information Processing Systems*, volume 25.

Khan, M. and Lin, W. (2017). Conjugate-computation variational inference : Converting variational inference in non-conjugate models to inferences in conjugate models. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54, pages 878–887. PMLR.

Kim, H. and Ghahramani, Z. (2003). The EM-EP algorithm for gaussian process classification. In *Proceedings of the Workshop on Probabilistic Graphical Models for Classification at ECML*.

Kim, H. and Ghahramani, Z. (2006). Bayesian Gaussian process classification with the EM-EP algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(12):1948–1959.

Kim, M. and Leskovec, J. (2011). The network completion problem: Inferring missing nodes and edges in networks. In *Proceedings of the 2011 SIAM international conference on data mining*, pages 47–58. SIAM.

Kimeldorf, G. S. and Wahba, G. (1970). A correspondence between bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41(2):495–502.

Krige, D. G. (1951). A statistical approach to some basic mine valuation problems on the witwatersrand. *Journal of the Chemical, Metallurgical and Mining Society of South Africa*, 52:119–139.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. (2017). Automatic differentiation variational inference. *Journal of Machine Learning Research*, 18:1–45.

Kuss, M. and Rasmussen, C. E. (2005). Assessing approximate inference for binary Gaussian process classification. *Journal of Machine Learning Research*, 6:1679–1704.

Lee, D. (2011). A comparison of conditional autoregressive models used in Bayesian disease mapping. *Spatial and spatio-temporal epidemiology*, 2(2):79–89.

Lee, J. D., Drysdale, W. S., Finch, D. P., Wilde, S. E., and Palmer, P. I. (2020). UK surface NO_2 levels dropped by 42% during the covid-19 lockdown: impact on surface O_3. *Atmospheric Chemistry and Physics*, 20(24):15743–15759.

Leroux, B. G., Lei, X., and Breslow, N. (1999). Estimation of disease rates in small areas: a new mixed model for spatial dependence. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 135–178. Springer.

Lewandowski, D., Kurowicka, D., and Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of multivariate analysis*, 100(9):1989–2001.

Liu, H., Ong, Y.-S., Shen, X., and Cai, J. (2020). When Gaussian process meets big data: A review of scalable GPs. *IEEE transactions on neural networks and learning systems*, 31(11):4405–4423.

Liu, T., Wei, H., and Zhang, K. (2018). Wind power prediction with missing data using Gaussian process regression and multiple imputation. *Applied Soft Computing*, 71:905–916.

MacKay, D. J. (1995). Probable networks and plausible predictions-a review of practical Bayesian methods for supervised neural networks. *Network: computation in neural systems*, 6(3):469.

MacNab, Y. C. (2003). Hierarchical Bayesian modeling of spatially correlated health service outcome and utilization rates. *Biometrics*, 59(2):305–315.

Matheron, G. (1963). Principles of geostatistics. *Economic Geology*, 58:1246–1266.

Matthews, A. G. d. G., Hensman, J., Turner, R., and Ghahramani, Z. (2016). On sparse variational methods and the Kullback-Leibler divergence between stochastic processes. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51, pages 231–239.

Matérn, B. (1960). *Spatial variation.* Allmänna Förlaget.

Melkumyan, A. and Ramos, F. T. (2009). A sparse covariance function for exact Gaussian process inference in large datasets. In *Proceedings of the Twenty-First International Joint Conference on Artificial Intelligence.*

Meng, X.-L. and Schilling, S. (2002). Warp bridge sampling. *Journal of Computational and Graphical Statistics*, 11(3):552–586.

Meng, X.-L. and Wong, W. H. (1996). Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, pages 831–860.

Minka, T. P. (2001). *A family of algorithms for approximate Bayesian inference.* PhD thesis, Massachusetts Institute of Technology.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032.

Murray, I. and Ghahramani, Z. (2005). A note on the evidence and Bayesian Occam's razor. Technical report, Gatsby Computational Neuroscience Unit.

Nickisch, H. and Rasmussen, C. E. (2008). Approximations for binary Gaussian process classification. *Journal of Machine Learning Research*, 9:2035–2078.

Opper, M. and Archambeau, C. (2009). The variational Gaussian approximation revisited. *Neural computation*, 21(3):786–792.

O'sullivan, F., Yandell, B. S., and Raynor Jr, W. J. (1986). Automatic smoothing of regression functions in generalized linear models. *Journal of the American Statistical Association*, 81(393):96–103.

Pan, J. and Pan, Y. (2017). jmcm: An R package for joint mean-covariance modeling of longitudinal data. *Journal of Statistical Software*, 82:1–29.

Plate, T. A. (1999). Accuracy versus interpretability in flexible modeling: Implementing a tradeoff using Gaussian process models. *Behaviormetrika*, 26(1):29–50.

Rasmussen, C. E. and Williams, C. K. (2006). *Gaussian processes for machine learning.* MIT press, Cambridge, Mass.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3):581–592.

Saatçi, Y. (2012). *Scalable inference for structured Gaussian process models.* PhD thesis, University of Cambridge.

Schölkopf, B., Herbrich, R., and Smola, A. J. (2001). A generalized representer theorem. In *International conference on computational learning theory*, pages 416–426. Springer.

Seeger, M. and Jordan, M. I. (2004). Sparse Gaussian process classification with multiple classes. Technical report, University of California at Berkeley, Berkeley, CA.

Smola, A. J. and Kondor, R. (2003). Kernels and regularization on graphs. In *Learning theory and kernel machines*, pages 144–158. Springer.

Smola, A. J., Vishwanathan, S., and Hofmann, T. (2005). Kernel methods for missing variables. In *International Workshop on Artificial Intelligence and Statistics*, pages 325–332. PMLR.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van Der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 64(4):583–639.

Stein, M. L. (1999). *Interpolation of spatial data: some theory for Kriging.* Springer Science & Business Media, New York.

Stern, H. S. and Cressie, N. (2000). Posterior predictive model checks for disease mapping models. *Statistics in Medicine*, 19(17-18):2377–2397.

Stitson, M., Gammerman, A., Vapnik, V., Vovk, V., Watkins, C., and Weston, J. (1999). Support vector regression with ANOVA decomposition kernels. In Schölkopf, B., Burges, C. J., and Smola, A. J., editors, *Advances in Kernel methods: Support Vector Learning.* The MIT Press, Cambridge, Massachusetts.

Stone, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. *The annals of statistics*, 22(1):118–171.

Taylor, B., Davies, T., Rowlingson, B., and Diggle, P. (2015). Bayesian inference and data augmentation schemes for spatial, spatiotemporal and multivariate log-Gaussian Cox processes in R. *Journal of Statistical Software*, 63:1–48.

Titsias, M. (2009). Variational learning of inducing variables in sparse Gaussian processes. In *Proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, volume 5, pages 567–574. PMLR.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and computing*, 27:1413–1432.

Vehtari, A. and Lampinen, J. (2002). Bayesian model assessment and comparison using cross-validation predictive densities. *Neural computation*, 14(10):2439–2468.

Vehtari, A., Mononen, T., Tolvanen, V., Sivula, T., and Winther, O. (2016). Bayesian leave-one-out cross-validation approximations for Gaussian latent variable models. *The Journal of Machine Learning Research*, 17(1):3581–3618.

Ver Hoef, J. M. and Cressie, N. (1993). Multivariable spatial prediction. *Mathematical Geology*, 25:219–240.

Wahba, G. (1990). *Spline models for observational data.* Society for Industrial and Applied Mathematics, Philadelphia.

Wahba, G., Wang, Y., Gu, C., Klein, R., and Klein, B. (1995). Smoothing spline ANOVA for exponential families, with application to the Wisconsin epidemiological study of diabetic retinopathy: the 1994 Neyman Memorial Lecture. *The Annals of Statistics*, 23(6):1865–1895.

Waller, L. A. and Carlin, B. P. (2010). Disease mapping. In Gelfand, A. E., Diggle, P., Guttorp, P., and Fuentes, M., editors, *Handbook of spatial statistics*, chapter 3, pages 217–243. CRC press, Boca Raton, FL.

Wang, K., Hamelijnck, O., Damoulas, T., and Steel, M. (2020). Non-separable non-stationary random fields. In *International Conference on Machine Learning*, pages 9887–9897. PMLR.

Watanabe, S. and Opper, M. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of machine learning research*, 11(12).

Wei, G. C. and Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85(411):699–704.

Wilkinson, W. J., Särkkä, S., and Solin, A. (2023). Bayes-Newton methods for approximate Bayesian inference with PSD guarantees. *Journal of Machine Learning Research*, 24(83):1–50.

Williams, C. and Seeger, M. (2001). Using the Nyström method to speed up kernel machines. In *Advances in Neural Information Processing Systems*, volume 13.

Wilson, A. and Nickisch, H. (2015). Kernel interpolation for scalable structured Gaussian processes (KISS-GP). In *International conference on machine learning*, pages 1775–1784. PMLR.

Wilson, A. G., Gilboa, E., Nehorai, A., and Cunningham, J. P. (2014). Fast kernel learning for multidimensional pattern extrapolation. *Advances in Neural Information Processing Systems*, 27.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques.*

Zhang, H. (2007). Maximum-likelihood estimation for multivariate spatial linear coregionalization models. *Environmetrics: The official journal of the International Environmetrics Society*, 18(2):125–139.

Zimmerman, D. L. and Stein, M. (2010). Classical geostatistical methods. In Gelfand, A. E., Diggle, P., Guttorp, P., and Fuentes, M., editors, *Handbook of spatial statistics*, chapter 3, pages 29–44. CRC press, Boca Raton, FL.

Appendix A

Kernel and Gaussian processes

A.1 Centring of kernels

A.1.1 Reproducing kernel Hilbert space

Recall that Hilbert space is a complete inner product space equipped with a positive definite inner product. Let $\mathcal{H}$ be a Hilbert space of functions over a set $\mathcal{X}$ with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. The Hilbert space $\mathcal{H}$ is called a reproducing kernel Hilbert space (RKHS) if and only if there exists a function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ satisfying

1. $k(x, \cdot) \in \mathcal{H}$ for all $x \in \mathcal{X}$

2. $f(x) = \langle f, k(x, \cdot) \rangle_{\mathcal{H}}$ for all $f \in \mathcal{H}$ and $x \in \mathcal{X}$

The function k is called reproducing kernel. Note that using the two properties, we have that $k(x, x') = \langle k(x, \cdot), k(x', \cdot) \rangle_{\mathcal{H}}$, hence k is positive definite. It can be shown by the Moore–Aronszajn theorem (Aronszajn (1950)) that a kernel defines a unique RKHS and vice versa. We write the norm of a function in f in $\mathcal{H}$ as $\|f\|_{\mathcal{H}} = \sqrt{\langle f, f \rangle}$.

A.1.2 Centring of kernel and functions in RKHS

Let P be distribution over a non-empty set $\mathcal{X}$ and $X, X' \in \mathcal{X}$ are independent and follow P. We consider a kernel k on $\mathcal{X}$ and let $\mathcal{H}_k$ denote the RKHS induced by k. We can centre this kernel by,

$$k_{cent}(x, x') = \langle k(x, \cdot) - \mu_P, k(x', \cdot) - \mu_P \rangle_{\mathcal{H}_k} \tag{A.1}$$

where μ_P is the kernel mean given by

$$\mu_P := \mathop{\mathbb{E}}_{X \sim P}[k(X, \cdot)] = \int_{\mathcal{X}} k(x, \cdot) dP(x).$$

Note that the expectation of any function $f \in \mathcal{H}_k$ can be computed as an inner product with μ_P:

$$\begin{aligned}
\mathop{\mathbb{E}}_{X \sim P}[f(X)] &= \int_{\mathcal{X}} f(x) dP(x) \\
&= \int_{\mathcal{X}} \langle k(x, \cdot), f \rangle_{\mathcal{H}_k} dP(x) \\
&= \langle \int_{\mathcal{X}} k(x, \cdot) dP(x), f \rangle_{\mathcal{H}_k} = \langle \mu_P, f \rangle_{\mathcal{H}_k}.
\end{aligned}$$

The centred kernel (A.1) is positive definite by construction. We can see that this corresponds with (2.6) by

$$\begin{aligned}
\langle k(x, \cdot) - \mu_P, k(x', \cdot) - \mu_P \rangle_{\mathcal{H}_k} &= \langle k(x, \cdot), k(x', \cdot) \rangle_{\mathcal{H}_k} - \langle \mu_P, k(x', \cdot) \rangle_{\mathcal{H}_k} \\
&\quad - \langle k(x, \cdot), \mu_P \rangle_{\mathcal{H}_k} + \langle \mu_P, \mu_P \rangle_{\mathcal{H}_k} \\
&= k(x, x') - \mathop{\mathbb{E}}_{X \sim P}[k(x', X)] - \mathop{\mathbb{E}}_{X' \sim P}[k(X', x)] + \mathop{\mathbb{E}}_{X, X' \sim P}[k(X, X')].
\end{aligned}$$

Note that

$$\begin{aligned}
\underset{X,X'\sim P}{\mathbb{E}}[k(X,X')] &= \int_{\mathcal{X}}\int_{\mathcal{X}} k(x,x')dP(x)dP(x') \\
&= \int_{\mathcal{X}}\int_{\mathcal{X}} \langle k(x,\cdot),k(x',\cdot)\rangle_{\mathcal{H}_k} dP(x)dP(x') \\
&= \langle \int_{\mathcal{X}} k(x,\cdot)dP(x), \int_{\mathcal{X}} k(x',\cdot)dP(x')\rangle_{\mathcal{H}_k} = \langle \mu_P, \mu_P\rangle_{\mathcal{H}_k}.
\end{aligned}$$

Given a sample $x_1, \dots x_n$ drawn from P, the kernel mean μ_P can be estimated empirically, by

$$\hat{\mu}_P = \frac{1}{n}\sum_{i=1}^{n} k(x_i, \cdot).$$

By replacing μ_P with $\hat{\mu_P}$, we obtain (2.7).

A.2 Posterior in GPR

Let us consider the regression model considered in (2.1) with normal i.i.d error, i.e., we have

$$\mathbf{y} = (y_1, \dots, y_n)^\top \sim \text{MVN}_n(\mathbf{0}, \mathbf{K} + \sigma^2\mathbf{I}).$$

Let us assume that we have $\mathbf{x}_j^* \in \mathcal{X}$ for $j = 1, \dots, m$. This set of $\mathbf{x}_j^*$ can include $\mathbf{x}_i$ in a sample (training set). We denote $\mathbf{f}^* = (f(\mathbf{x}_1^*), \dots, f(\mathbf{x}_m^*))$. From our prior we know

$$\mathbf{f}^* \sim \text{MVN}_m\left(\mathbf{0}, \mathbf{K}_{**}\right) \tag{A.2}$$

where $\mathbf{K}_{**} = \{k_{i,j}^{**}\}_{m\times m}$ and $k_{i,j}^{**} = k(\mathbf{x}_i^*, \mathbf{x}_j^*)$. Hence we have

$$\begin{bmatrix}\mathbf{y}\\ \mathbf{f}^*\end{bmatrix} \sim \text{MVN}_{n+m}\left(\begin{bmatrix}\mathbf{0}\\ \mathbf{0}\end{bmatrix}, \begin{bmatrix}\mathbf{K}+\sigma^2\mathbf{I} & \mathbf{K}_*\\ \mathbf{K}_*^\top & \mathbf{K}_{**}\end{bmatrix}\right)$$

where $\mathbf{K}_* = \{k_{*i,j}\}_{n\times m}$ and $k_{*i,j} = k(\mathbf{x}_i, \mathbf{x}_j^*)$. By conditional distribution of multivariate normal distribution,

$$\mathbf{f}^*|\mathbf{X}, \mathbf{X}^*, \mathbf{y} \sim \mathrm{MVN}_m(\boldsymbol{\mu}^*, \mathbf{V}^*)$$

where

$$\begin{aligned} \boldsymbol{\mu}^* &= \mathbf{K}^{*\top}(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{y} \\ \mathbf{V}^* &= \mathbf{K}^{**} - \mathbf{K}^{*\top}(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{K}^*. \end{aligned}$$

It is easy to see that we have $\boldsymbol{\mu}^* = (\bar{m}(\mathbf{x}_1^*), \ldots, \bar{m}(\mathbf{x}_1^*))$ and $\mathbf{V}^* = \{v_{i,j}^*\}_{m\times m}$ with $v_{i,j}^* = \bar{k}(\mathbf{x}_i^*, \mathbf{x}_j^*)$ where $\bar{m} : \mathcal{X} \to \mathbb{R}$ and $\bar{k} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ are given by

$$\begin{aligned} \bar{m}(\mathbf{x}) &= \mathbf{k}(\mathbf{x})^\top(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{y}, && \mathbf{x} \in \mathcal{X} \\ \bar{k}(\mathbf{x}, \mathbf{x}') &= k(\mathbf{x}, \mathbf{x}') - \mathbf{k}(\mathbf{x})^\top(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{k}(\mathbf{x}'), && \mathbf{x},\mathbf{x}' \in \mathcal{X}. \end{aligned}$$

By Kolmogorov extension theorem and Definition 2, this implies that the posterior is the Gaussian process $\mathrm{GP}(\bar{m}, \bar{k})$.

A.3 Multi-class categorical and multivariate response

In this section, we first show how the GP model for a binary response can be extended to a multi-class categorical response and discuss its connection to the GP models for multivariate response.

Consider a sample $(\tilde{y}_i, \mathbf{x}_i)_{i=1}^n$ where $\tilde{y}_i \in \mathcal{Y}$ is a response variable which takes value j if the i-th instance in the dataset belongs to category j for $j = 1, \ldots, c$ and $\mathbf{x}_i \in \mathcal{X}$ is a set of predictors. We have $\mathcal{Y} = \{1, \ldots, c\}$. We assume $\tilde{y} \sim \mathit{Categorical}(p_1, \ldots, p_c)$ where $p_j = p(\tilde{y} = j)$ subject to $\sum_{j=1}^c p_j = 1$. It is common to re-code the response variable by

$$y_{j,i} = \begin{cases} 1 & \text{if } \tilde{y}_i = j \\ 0 & \text{otherwise.} \end{cases}$$

Note that only one element in the vector of a response for i-th instance, $\mathbf{y}_i = (y_{1,i}, \ldots, y_{c,i})^\top$, can take the value 1. We model the relationship between the response and the predictors $\mathbf{x}_i \in \mathcal{X}$ using the soft-max function:

$$p_{i,j} = p(y_{j,i} = 1) = \frac{\exp\left(f(j, \mathbf{x}_i)\right)}{\sum_{j'=1}^{c} \exp\left(f(j', \mathbf{x}_i)\right)}. \tag{A.3}$$

Let $\mathcal{D} = \mathcal{Y} \times \mathcal{X}$. We assume the function $f : \mathcal{D} \to \mathbb{R}$ to follow zero mean GP, $f \sim GP(0, k)$ where the kernel $k : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ is, for example, given by

$$k((j, x)^\top, (j', x')^\top) = k_c(j, j')k_x(\mathbf{x}, \mathbf{x}'), \quad \mathbf{x}, \mathbf{x}' \in \mathcal{X}, j, j' \in \mathcal{Y}.$$

Through kernel k_c, we can model class dependence. For instance, if $k_c(j, j') = 0$ for all $j \neq j'$, no inter-class correlation is assumed. If we define

$$\mathbf{f} = (f(1, \mathbf{x}_1), \ldots, f(1, \mathbf{x}_n), \ldots\ldots, f(c, \mathbf{x}_1), \ldots, f(c, \mathbf{x}_n))^\top$$

and similarly,

$$\mathbf{y} = (y_{1,1}, \ldots, y_{1,n}, \ldots\ldots, y_{c,1}, \ldots, y_{c,n})^\top,$$

we have:

$$\mathbf{f}|\mathbf{X} \sim \mathrm{MVN}(\mathbf{0}, \mathbf{K})$$

where $\mathbf{K} = \mathbf{K}_c \otimes \mathbf{K}_x$ and $\{\mathbf{K}_c\}_{1 \le j,j' \le c}$ and $\{\mathbf{K}_x\}_{1 \le i,i' \le n}$ are Gram matrices with each element given by $k_c(j, j')$ and $k_x(\mathbf{x}_i, \mathbf{x}_i')$. If we assume no inter-class correlation, then $\mathbf{K}_c$ is a diagonal matrix with positive diagonal entries $\mathbf{D}$ and the matrix $\mathbf{K}$ is block diagonal:

$$\mathbf{K} = \begin{bmatrix} d_{1,1}\mathbf{K}_x & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & d_{2,2}\mathbf{K}_x & \dots & \mathbf{0} \\ \vdots & & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & d_{c,c}\mathbf{K}_x \end{bmatrix}.$$

where $d_{j,j}$ is the j-th diagonal element of $\mathbf{D}$. The log-likelihood for this model is given by

$$\log p(\mathbf{y}|\mathbf{f}) = \mathbf{y}^\top \mathbf{f} - \sum_{i=1}^{n} \log \left(\sum_{j=1}^{c} \exp f(j, \mathbf{x}_i) \right).$$

Once the prior and likelihood is specified, the procedures described in Section 2.3 can be used for model estimation and inference. For the detail of each approximation method in the multi-class categorical response setting, see Rasmussen and Williams (2006, Chapter 3) for Laplace approximation, Chai (2012) for Variational Inference and Seeger and Jordan (2004) for Expectation propagation. It is worth noting that multi-class classification with GP in this formulation has an identifiability issue. For $\mathbf{f}$ to be identifiable, we have to e.g. set $f(1, \mathbf{x}_i) = 0$ or ensure $\sum_{j=1}^{c} f(j, \mathbf{x}_i) = 0$ for all $i = 1, \dots, n$. The latter can be satisfied by empirically centring the k_c. See Section 2.1.3.

This model formulation is closely connected to GP models for multivariate response. This class of models may also be called multi-task GP (Bonilla et al., 2007) or GP models for a vector output (Alvarez et al., 2012). The multivariate response is common in spatio-temporal analysis. For example, in environmental monitoring, multiple meteorological variables (wind speed, temperature, etc.) are recorded at each monitoring station. The crime cases in a region are also often available in different categories.

Using the same notations for $\mathbf{y}$ and $\mathbf{f}$ we specify the likelihood by e.g. $y_{j,i} \sim \text{Poisson}(\lambda_{j,i})$ with $\log \lambda_{j,i} = f(j, \mathbf{x}_i)$. The specification on k_c plays an important part in this context. One popular approach is putting prior on $\mathbf{K}_c$ using, e.g. Inverse

Wishart distribution, or LDK prior (Lewandowski et al., 2009). Note that the latter can be used as a prior on the correlation matrix. The entries of the covariance matrix or matrices that account for inter-class correlation can also be estimated using the EM algorithm see, e.g., (Zhang, 2007), which considered the spatial linear model of coregionalization (LCM). The LCM (Goovaerts, 1997; Journel and Huijbregts, 1976) is a popular model for a multivariate response, which has a more complex structure in $\mathbf{K}$ e.g., $\mathbf{K} = \sum_{q=1}^{Q} \mathbf{K}_{cq} \otimes \mathbf{K}_{xq}$.

It is also possible to consider the ANOVA decomposition kernel introduced in Section 3.1.3, in which case the Gram matrix can be expressed as $(\mathbf{1}\mathbf{1}^\top + \mathbf{K}_c) \otimes (\mathbf{1}\mathbf{1}^\top + \mathbf{K}_x)$. This formulation allows us to understand how the effect of $\mathbf{x}$ is different for different classes (interaction effect) but also the effect of $\mathbf{x}$ averaged over all classes (main effect).

The Kriging framework and conditional auto-regressive model for areal data, discussed in Section 2.5, can be extended to multivariate spatial variables. The former is known as *co-kriging*. See, e.g. Ver Hoef and Cressie (1993) for the details. For the multivariate conditional auto-regressive models (MCAR), see Carlin et al. (2003); Gelfand and Vounatsou (2003).

Appendix B

Kronecker product

B.1 Row-wise Kronecker product

Consider two matrices $\mathbf{A} = \{a_{i,j}\}_{1\leq i\leq n, 1\leq j\leq m}$ and $\mathbf{B} = \{b_{i,j}\}_{1\leq i\leq n, 1\leq j\leq q}$. Let $\mathbf{A}_i$ and $\mathbf{B}_i$ be the $i-$th row of the matrices $\mathbf{A}$ and $\mathbf{B}$ respectively. The row-wise Kronecker product of the two matrices, $\mathbf{A} \bullet \mathbf{B}$, is the matrix of size $n \times mq$ given by

$$\mathbf{A} \bullet \mathbf{B} = \begin{bmatrix} \mathbf{A}_1 \otimes \mathbf{B}_1 \\ \mathbf{A}_2 \otimes \mathbf{B}_2 \\ \vdots \\ \mathbf{A}_n \otimes \mathbf{B}_n \end{bmatrix}$$

where $\otimes$ is the Kronecker product (see Section 4.3). The row-wise Kronecker product may also be called the face-splitting product. Let $\mathbf{v}$ be a vector of length n. Then we have

$$\mathbf{A} \bullet \mathbf{v} = \mathbf{v} \bullet \mathbf{A} = \mathbf{V}_d \mathbf{A} \tag{B.1}$$

where $\mathbf{V}_d = \text{diag}(\mathbf{v})$, a diagonal matrix with its diagonal elements given by $\mathbf{v}$.

B.2 Eigendecomposition of Gram matrix with fBM kernel

Assume a tensor product kernel

$$k(\mathbf{x}, \mathbf{x}') = \prod_{l=1}^{d} k_d(\mathbf{x}_l, \mathbf{x}'_l)$$

over a multidimensional grid $\mathcal{X} = \mathcal{X}_1 \times \ldots \times \mathcal{X}_d$ where $\mathbf{x}_l \in \mathcal{X}_l$ and $k_l : \mathcal{X}_l \times \mathcal{X}_l$. Let $n_l = |\mathcal{X}_l|$. Then, the associated Gram matrix can be written as

$$\mathbf{K} = \bigotimes_{l=1}^{d} \mathbf{K}_l$$

where $\mathbf{K}_l$ is a $n_l \times n_l$ gram matrix for l input dimension, with i, j-th element given by $k_l(\mathbf{x}_{(l),i}, \mathbf{x}_{(l),j})$. Let $\mathbf{K}_l = \mathbf{Q}_l \mathbf{\Lambda}_l \mathbf{Q}_l^\top$ be the eigendecomposition of each matrix. Then the eigendecomposition of the matrix K is the following:

$$\begin{aligned} \mathbf{K} &= \bigotimes_{l=1}^{d} \left(\mathbf{Q}_l \mathbf{\Lambda}_l \mathbf{Q}_l^\top \right) \\ &= \bigotimes_{l=1}^{d} \mathbf{Q}_l \bigotimes_{l=1}^{d} \mathbf{\Lambda}_l \bigotimes_{l=1}^{d} \mathbf{Q}_l^\top . \end{aligned} \tag{B.2}$$

This decomposition leads to a particularly efficient algorithm when using an fBM kernel or a squared and centred fBM kernel with a known Hurst coefficient γ_l. Let each k_l be a fBM$_{\gamma_l}$ kernel. This means we have only one hyper-parameter (scale parameter) to estimate for each dimension l. We denote the corresponding gram matrix by $\mathbf{K}_l = \alpha_l \mathbf{K}'_l$ where $\mathbf{K}'_l$ is un-scaled gram matrix. Let $\mathbf{K}'_l = \mathbf{Q}_l \mathbf{\Lambda}_l \mathbf{Q}_l^\top$ be the eigendecomposition of the un-scaled matrix. Then eigendecomposition of $\mathbf{K}_l$ is

$$\mathbf{K}_l = \alpha_l \mathbf{K}'_l = \mathbf{Q}_l \left(\alpha_l \mathbf{\Lambda}_l \right) \mathbf{Q}_l^\top .$$

Using (B.2), we can write

$$\mathbf{K} = \bigotimes_{l=1}^{d} \mathbf{K}'_l = \bigotimes_{l=1}^{d} \mathbf{Q}_l \bigotimes_{l=1}^{d} (\alpha_l \mathbf{\Lambda}_l) \bigotimes_{l=1}^{d} \mathbf{Q}_l^\top .$$

This means that we do not have to apply eigendecomposition at each iteration when estimating the hyper-parameters by maximising the marginal likelihood or by MCMC. The inverse and the determinant can be updated by simply multiplying each eigenvalue by the scale parameters.

Appendix C

Illustration and data analysis

C.1 Bovine Tuberculosis in Cornwall

In Section 3.4.1, we used squared exponential kernels to compare different model structures for Bovine Tuberculosis data. The posterior mean function, in this case, is infinitely differentiable. This may be too smooth. We run the spatial model with Matérn class kernels (See Table C.1) and see that the models with Matérn class kernels achieve higher prediction accuracy. Similarly, for spatio-temporal models, it is important to consider and evaluate models with different kernels in the future.

Table C.1 CV (5-folds) errors for spatial GP models with different kernels. Matérn(1.5) and Matérn(2.5) refer to Matérn kernel with the κ parameter1.5 and 2.5.

Model	kernel	Misclassification rate	Brier score
Spatial			
	SE	0.1421	0.2242
	Matérn(1.5)	0.1305	0.2209
	Matérn(2.5)	0.1237	0.2159

C.2 Longitudinal data analysis

The model considered in Section 3.4.2 given by (3.16) and (3.17) has eight terms in total. These include a constant term a, and three main effect terms f_1, f_2, f_3, three two-way interaction effect terms f_{12}, f_{23}, f_{31} and one three-way interaction effect term f_{123}. The posterior mean $\bar{m}$ consists of 8 terms of constant, main and interaction effect terms. We write $\bar{m} = \sum_{h \in \mathcal{H}} \bar{m}, h$ where $\mathcal{H} = \{0, 1, 2, 3, 12, 23, 31, 123\}$. Here $\bar{m}_0$ corresponds with the constant term. Figure C.1 shows each $\bar{m}_h$ against *day*.

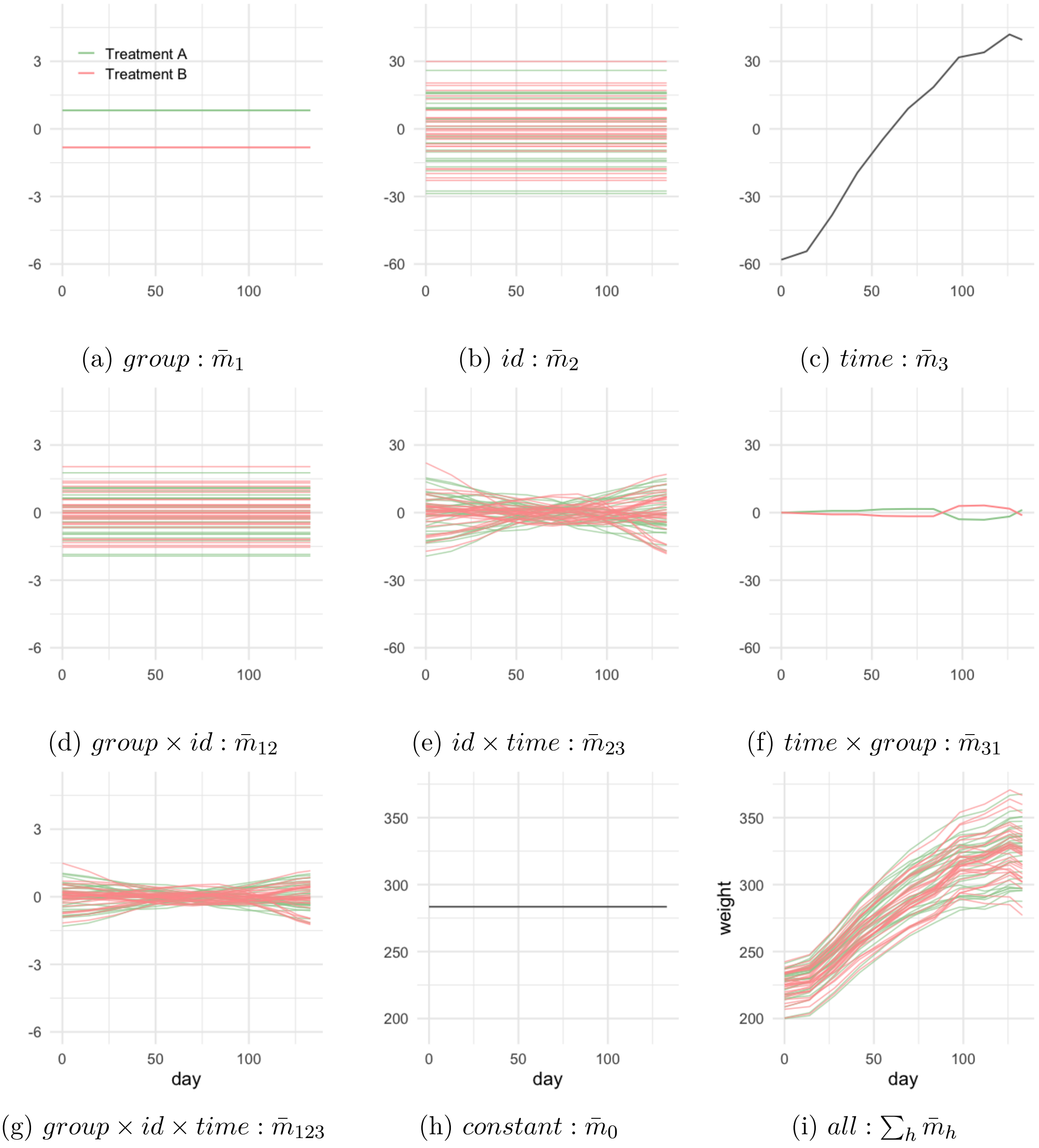
The main effects (Figure C.1a-C.1c) can simply be interpreted as the average effect of *group*, *id* and *time* respectively. For two-way interaction effects, it is natural to consider e.g., $\bar{m}_3 + \bar{m}_{13}$ which shows how the effect of *time* (or the growth curve) is different for different treatment groups, as we saw in Section 3.4.2. Here, we give another example. To understand how the growth curve differs among different cattle, we show the plot of $\bar{m}_3 + \bar{m}_{23}$ in Figure C.2. This can be seen as the effect of **time** for different cattle (*id*) after the effect of treatment is averaged out. We see that there is more variability in the growth attributed to individuals at the beginning and the end of the study period.

C.3 NO$_2$ concentrations in London

C.3.1 Data manipulation

Missing value imputation

In the dataset used in this paper, we had $1,290$ missing values out of $208,152$. For each missing value at a measurement site, we created a small subset of the data consisting of the observations collected from the same location from 24 hours before to 24 hours after the missing values were observed. A simple one-dimensional Gaussian process



(a) $group : \bar{m}_1$ (b) $id : \bar{m}_2$ (c) $time : \bar{m}_3$

(d) $group \times id : \bar{m}_{12}$ (e) $id \times time : \bar{m}_{23}$ (f) $time \times group : \bar{m}_{31}$

(g) $group \times id \times time : \bar{m}_{123}$ (h) $constant : \bar{m}_0$ (i) $all : \sum_h \bar{m}_h$

Fig. C.1 The posterior mean function $\bar{m}_h$ for $h = \{0, 1, 2, 3, 12, 23, 31, 123\}$ together with the added posterior mean $\bar{m} = \sum_h \bar{m}_h$. Each $\bar{m}_h$ can also be understood as the effect on the weight (y-axis). The scale of the y-axis is adjusted for each effect.

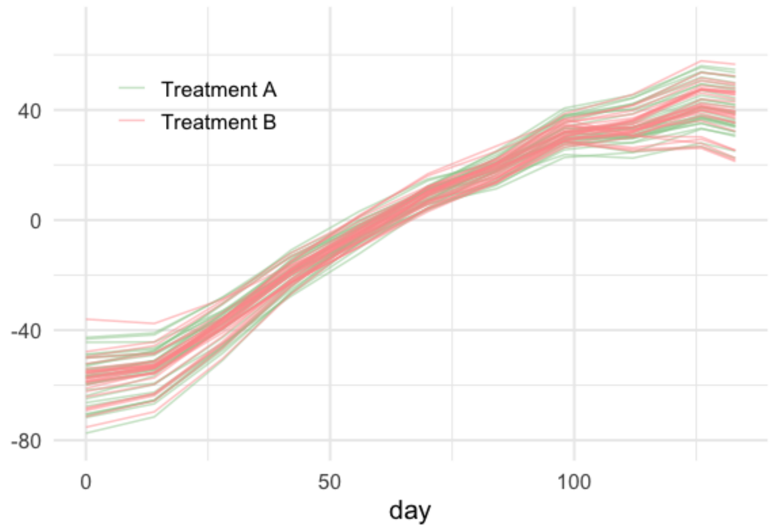


Fig. C.2 The plot of $time + time \times id : \bar{m}_3 + \bar{m}_{31}$

regression with squared and centred standard Brownian Motion kernel ($\gamma = \frac{1}{2}$) is then fitted. We replace the missing value with the posterior predictive mean given by (2.12).

Adjustment of Summer time

The study period, from January 6 2020, to May 31, 2020, includes the clock change to British Summer Time (BST), starting at 1:00 AM on March 29. The timestamps in the original data are all in Greenwich Mean Time (GMT). We converted the timestamp to match BST from 1:00 AM (in GMT). This resulted in an hour gap without any record at 1:00 AM of the adjusted timestamp. We filled the gap with a mean of the m before and after. The procedure is summarised in Table C.2.

C.3.2 Gram matrix for each models

Let $\mathbf{K}_l$ denote the Gram matrix that corresponds with the kernel k_l in Section 4.5.2, i.e., $\mathbf{K}_l = \{k_{i,j}^{(l)}\}_{1\leq i,j,\leq n_l}$. The Gram matrices for the models under consideration in Section 4.5.2 are listed as follows.

Table C.2 Adjustment from GMT to BST

Time in GMT	Original y	Time in BST	Adjusted time	Adjusted y
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
2020-03-29 0:00	y_{t-1}	-	2020-03-29 0:00	y_{t-1}
			2020-03-29 1:00	$(y_{t-1}+y_t)/2$
2020-03-29 1:00	y_t	2020-03-29 2:00	2020-03-29 2:00	y_t
2020-03-29 2:00	y_{t+1}	2020-03-29 3:00	2020-03-29 3:00	y_{t+1}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

- Model 1: main effect

$$\mathbf{K}_{m1} = \bigotimes_{l=1}^{3} \mathbf{1}_{n_l}\mathbf{1}_{n_l}^\top + \mathbf{K}_1 \otimes \mathbf{1}_{n_2}\mathbf{1}_{n_2}^\top \otimes \mathbf{1}_{n_3}\mathbf{1}_{n_3}^\top + \mathbf{1}_{n_1}\mathbf{1}_{n_1}^\top \otimes \mathbf{K}_2 \otimes \mathbf{1}_{n_3}\mathbf{1}_{n_3}^\top + \mathbf{1}_{n_1}\mathbf{1}_{n_1}^\top \otimes \mathbf{1}_{n_2}\mathbf{1}_{n_2}^\top \otimes \mathbf{K}_3$$

- Model 2: space-time interactions

$$\mathbf{K}_{m2} = \mathbf{K}_{m1} + \mathbf{K}_1 \otimes \mathbf{K}_2 \otimes \mathbf{1}_{n_3}\mathbf{1}_{n_3}^\top + \mathbf{K}_1 \otimes \mathbf{1}_{n_2}\mathbf{1}_{n_2}^\top \otimes \mathbf{K}_3$$

- Model 3: all two-way interactions

$$\mathbf{K}_{m3} = \mathbf{K}_{m2} + \mathbf{1}_{n_1}\mathbf{1}_{n_1}^\top \otimes \mathbf{K}_2 \otimes \mathbf{K}_3$$

- Model 4: saturated

$$\mathbf{K}_{m4} = \mathbf{K}_{m3} + \mathbf{K}_1 \otimes \mathbf{K}_2 \otimes \mathbf{K}_3 = \bigotimes_{l=1}^{3} \left(\mathbf{1}_{n_l}\mathbf{1}_{n_l}^\top + \mathbf{K}_l\right)$$

- Model 5: three-way interaction only

$$\mathbf{K}_{m5} = \mathbf{K}_1 \otimes \mathbf{K}_2 \otimes \mathbf{K}_3$$

Table C.3 The parameter estimates obtained by maximising the log-marginal likelihood using the L-BFGS algorithm provided in Stan. The time taken (in seconds) until convergence is also provided.

Model	α_0	α_1	α_2	α_3	σ	Time(s)
1: main	6.86	5.02	2.2	0.32	14.85	3.4
2: spatio-temporal interaction	14.4	1.28	0.67	0.24	12.54	4.1
3: all two-way interaction	10.63	1.87	1.32	0.31	8.37	4.2
4: saturated	52.02	0.48	0.94	0.051	6.51	3.4
5: three-way interaction only	-	0.0008	-	-	36.43	1.4

C.3.3 Hyper-parameter estimation by Naive Bayes

In addition to obtaining samples from the posterior of the hyper-parameters, we also estimated the hyper-parameters by finding the maximiser of the log marginal likelihood (2.15). We used the optimisation algorithm provided by Stan, which can be run with the same code used for MCMC sampling. The convergence was achieved within a few seconds for all models. The values obtained (Table C.3) are close to the MCMC sample means.

C.3.4 Visualisation of interaction effects

The selected model implies that the effect of x_2 (the global time effect) takes on different forms for different values of $\mathbf{x}_1$ (location). This is shown in Figure C.3, which plot $\bar{m}_2(x_2) + \bar{m}_{12}(\mathbf{x}_1, x_2)$ as a function x_2 given a set of coordinates $\mathbf{x}_1$ at selected stations. Figure C.4 shows how the spatial effect changes over the course of a day. These are the plots of $\bar{m}_1(\mathbf{x}_1) + \bar{m}_{13}(\mathbf{x}_1, x_3)$ as a function of $\mathbf{x}_1$ evaluated at different hours of the day.

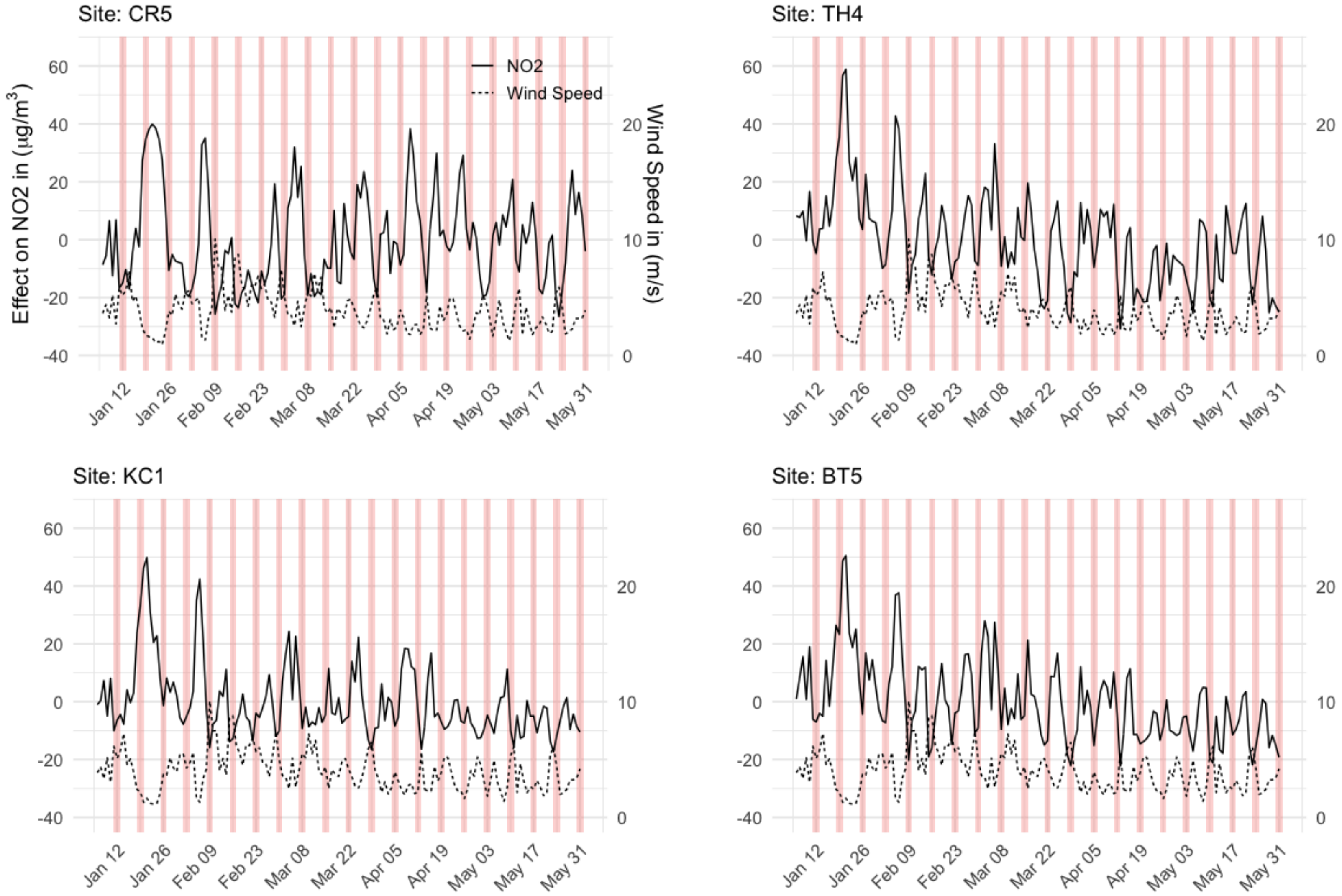


Fig. C.3 The global time (day) effect and wind speed by site with weekends highlighted. The effect is averaged over the hour of the day.

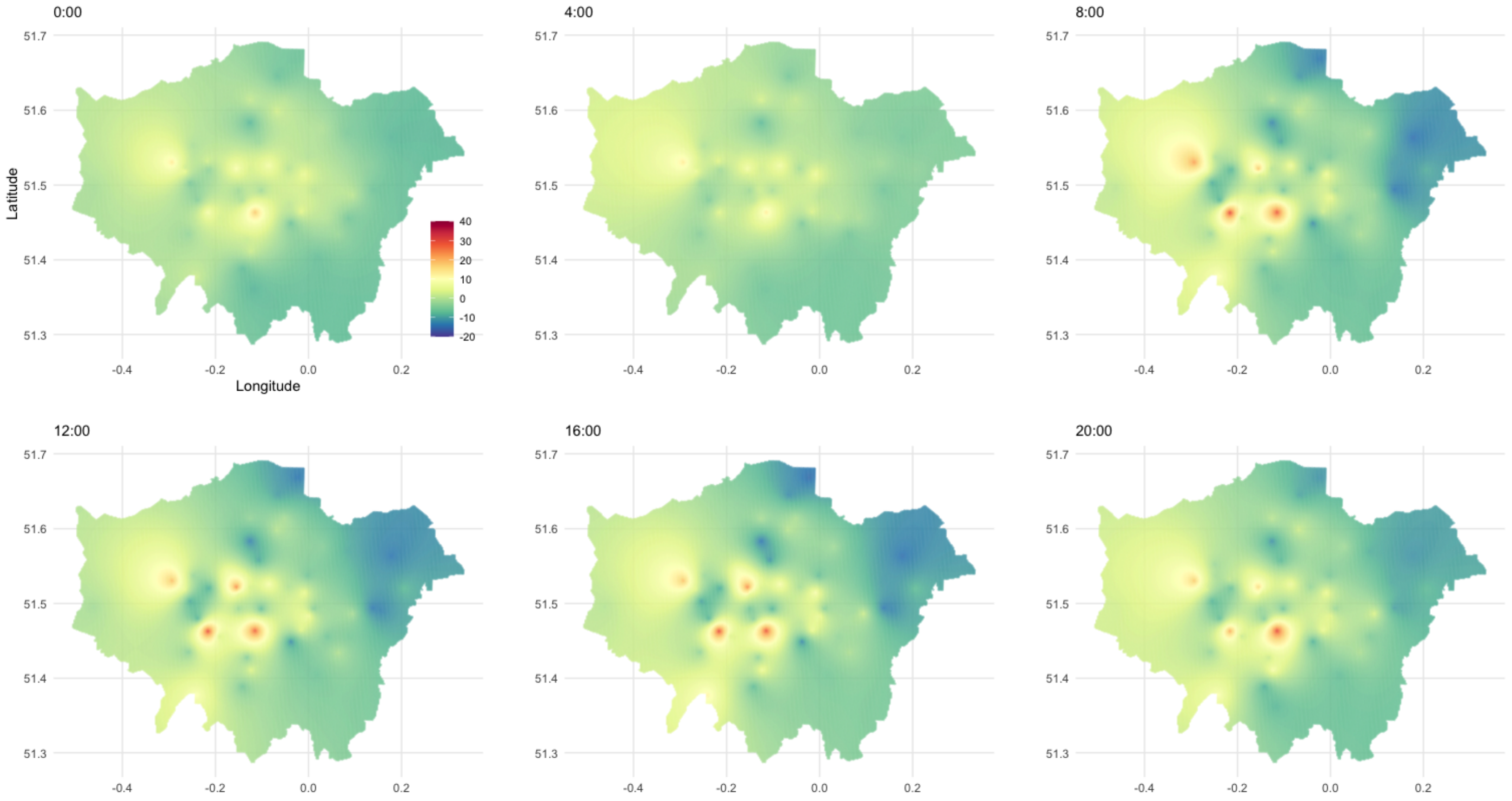


Fig. C.4 The spatial effect over London at different hours of the day averaged over different calendar dates. The effect is measured in μg/m^3.

C.4 Additional results for the simulation study with incomplete grid

We provide in Table C.4 the mean of the parameter estimates and standard error for the simulation study considered in Section 5.3. The simulation setting considered is: grid size 70 × 70, three missingness mechanisms considered in the section; Missing at Completely Random(MCAR), Missing at Random(MAR) and Missing not at Random(MNAR), the missing proportions varying from 10% to 30%, and the number of experiments for each scenario 20.

		α_0	α_1	α_2	σ
Complete data		21.47 (0.02)	0.11 (0.004)	0.11 (0.006)	1.5 (0.02)
MCAR	10%	21.45 (0.07)	0.12 (0.005)	0.11 (0.007)	1.5 (0.02)
	20%	21.49 (0.06)	0.14 (0.006)	0.13 (0.008)	1.49 (0.02)
	30%	21.45 (0.04)	0.16 (0.01)	0.13 (0.01)	1.49 (0.02)
MAR	10%	21.47 (0.07)	0.11 (0.004)	0.11 (0.005)	1.5 (0.02)
	20%	21.45 (0.07)	0.11 (0.004)	0.11 (0.006)	1.5 (0.02)
	30%	21.46 (0.05)	0.11 (0.005)	0.11 (0.005)	1.5 (0.02)
MNAR	10%	21.29 (0.07)	0.11 (0.01)	0.11 (0.01)	1.45 (0.01)
	20%	21.16 (0.05)	0.11 (0.01)	0.11 (0.01)	1.43 (0.02)
	30%	21.02 (0.07)	0.11 (0.01)	0.11 (0.01)	1.42 (0.02)

Table C.4 The parameter estimates and its standard error (in parentheses) for simulation study with 70 × 70 grid size, based on 20 replications.