Convolved Gaussian Process Regression Models for Multivariate Non-Gaussian Data

A'yunin Sofro

Thesis submitted for the degree of
Doctor of Philosophy

School of Mathematics & Statistics
Newcastle University
Newcastle upon Tyne
United Kingdom

November 2016

Acknowledgements

First and foremost, praise be to Allah s.w.t for giving me His blessings throughout my studies. I would like to express my gratitude to those who had extended their help and support throughout my studies; towards the completion of this thesis in particular.

Thank you to my supervisor, Dr Jian Qing Shi for his guidance and patience during this research. The opportunity to learn from him is priceless and and will not be forgotten. I would also like to thank my second supervisor Professor John Matthews for his encouragement and suggestions throughout this research.

I am indebted to Ministry of Education, Indonesia for sponsoring my study through the scholarship. My special thanks to my husband Sony Febrianto, my children Zaky Athallah Muhammad and Natasya Azzahra for their great patience and support. Not to be forgotten, my parents and other family members for their continuous moral support. Last but not least, to all my friends who have helped me undergo the hard times together, which have all been worthwhile.

Abstract

Multivariate regression analysis has been developed rapidly in the last decade for dependent data. The most difficult part in multivariate cases is how to construct a cross-correlation between response variables. We need to make sure that the covariance matrix is positive definite which is not an easy task. Several approaches have been developed to overcome the issue. However, most of them have some limitations, such as it is hard to extend it to the case involving high dimensional variables or capture individual characteristics. It also should point out that the meaning of the cross-correlation structure for some methods is unclear. To address the issues, we propose to use convolved Gaussian process ($\mathcal{CGP}$) priors (Boyle & Frean, 2005).

In this dissertation, we propose a novel approach for multivariate regression using $\mathcal{CGP}$ priors. The approach provides a semiparametric model with multi-dimensional covariates and offers a natural framework for modelling common mean structures and covariance structures simultaneously for multivariate dependent data. Information about observations is provided by the common mean structure while individual characteristics also can be captured by the covariance structure. At the same time, the covariance function is able to accommodate a large-dimensional covariate as well.

We start to make a broader problem from a general framework of $\mathcal{CGP}$ proposed by Andriluka *et al.* (2006). We investigate some of the stationary covariance functions and the mixed forms for constructing multiple dependent Gaussian processes to solve a more complex issue. Then, we extend the idea to a multivariate non-linear regression model by using convolved Gaussian processes as priors.

We then focus on an applying the idea to multivariate non-Gaussian data, i.e. multivariate Poisson, and other multivariate non-Gaussian distributions from the exponential family. We start our focus on multivariate Poisson data which are found in many problems relating to public health issues. Then finally, we provide a general framework for a multivariate binomial data and other multivariate non-Gaussian data.

The definition of the model, the inference, and the implementation, as well as its asymptotic properties, are discussed. Comprehensive numerical examples with both simulation studies and real data are presented.

Key Words : Convolved Gaussian process, Cross-correlation, Multivariate dependent data, Multivariate nonlinear regression, Multivariate non-Gaussian regression, Stationary covariance functions

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Aims

Regression analysis for multivariate non-Gaussian data has been developed rapidly in the last decade. One of the well-known models is multivariate Poisson regression which can be used for analysing count data. The observed data become more complex when they are dependent and when there is more than one response variable. One of the examples is dengue fever and malaria data that we will discuss in this thesis, where the outputs are the number of cases of dengue fever and malaria. We begin the investigation by exploring the areas at risk of dengue and malaria transmission which are presented in Figures 1.1 and 1.2. From those figures, it is clearly seen that the transmission of two diseases have a similar structure mapping where tropical regions of developing countries are the most high-risk areas in the world, such as Indonesia.

Several methods have been proposed to investigate the issue. We start analysing the data separately by using a simple method, i.e. Poisson regression. This approach assumes independent observations. In fact, this assumption is unrealistic in many problems. Some research on Poisson regression for dependent data has been done, such as the intrinsic conditional autoregressive model. Other methods include using kriging or Gaussian process regression where Gaussian processes are the backbone of the method and its performance shows promising results. Although those approaches are able to handle correlation within observations for each response variable, they could not accommodate the cross-correlation between two response variables .

In the dengue fever and malaria data, we highlight that it is more sensible if we analyse them at the same time since the diseases involve a similar correlation structure. Those diseases, which are transmitted by a virus from mosquitoes, have similar signs and symptoms. It is also worth noting that the study of spatial correlation is an important issue since the geographical patterns of the diseases are similar to each other. The outputs

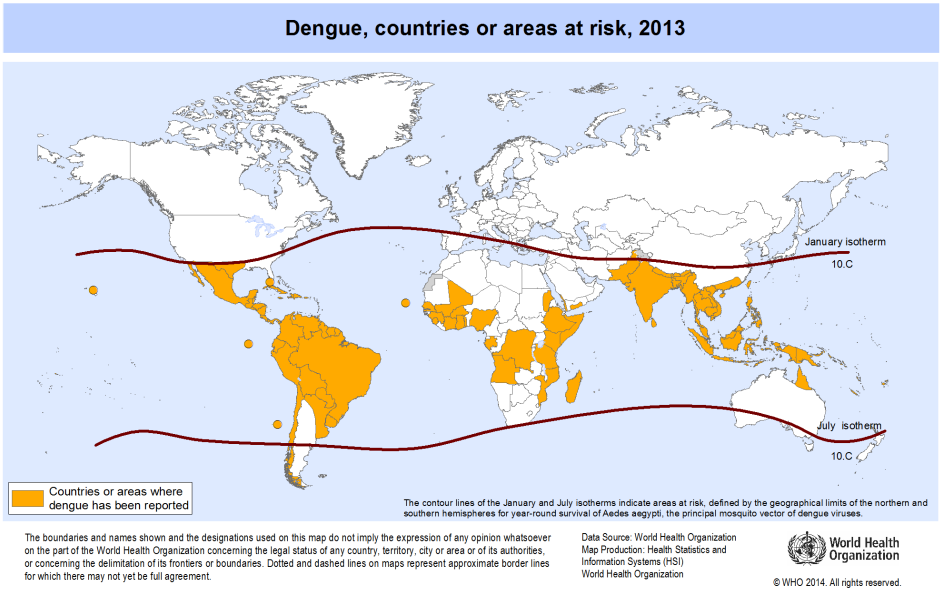

Figure 1.1: Countries at risk of dengue transmission in 2013 (Source: World Health Organization, 2013)

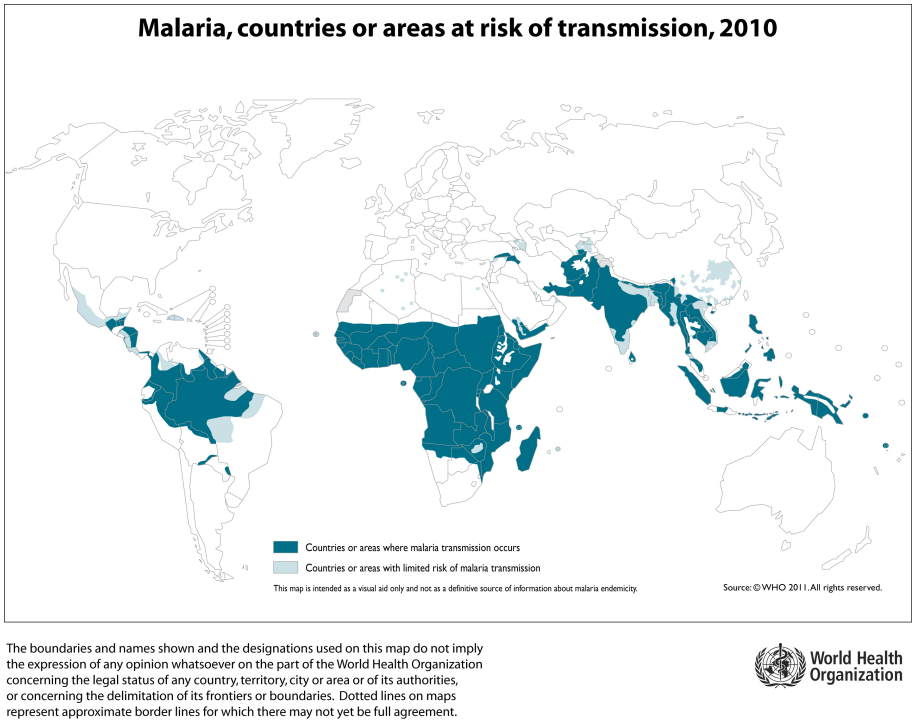

Figure 1.2: Countries at risk of malaria transmission in 2010 (Source: World Health Organization, 2010)

of the data are the number of cases of dengue fever and malaria in East Java, Indonesia. Thus, in this case, we are interested in a single set of composite boundaries for multiple disease outcomes. As a consequence, the statistical model needs to account for correlation between diseases and locations, see Ma & Carlin (2007).

Constructing a cross-correlation structure for each response, where we should guarantee that the covariance matrix is positive definite, is another major problem in analyzing dependent data. This issue makes it difficult since ensuring that the covariance matrix is positive definite is not an easy job. Several methods have been developed to overcome the problem, such as using multivariate conditional autrogressive (MCAR) or multivariate conditional kriging. However, all the approaches present some limitations, such as unclear covariance structure of the cross-corrrelation, difficulty in extending to high dimensional variables, see Martinez-Beneito (2013) or hard to capture individual characteristics for each response variable, see Crainiceanu *et al.* (2012).

We are going to address these difficult cases by using convolved Gaussian process regression models in this thesis. It is an extension of a general framework for constructing dependent outputs proposed by Andriluka *et al.* (2006) which offers flexibility and robustness.

1.2 A Brief Literature Review

A univariate Poisson regression with correlation structure has been developed in the last decade. Some research has been done regarding the handling of this issue. The intrinsic conditional autoregressive (ICAR) model introduced by Besag & Kooperberg (1995) is one of the popular methods. This method has been extended into a spatial or temporal correlated generalized linear mixed model (Sun *et al.* (2000); MacNab & Dean (2001); Martinez-Beneito *et al.* (2008); Silva *et al.* (2008)). A generalized linear mixed model using prior distribution for spatially structured random effect is a common alternative way, see Carlin & Gelfand (2004). But they limited the correlation structure into intrinsic conditional autoregressive model (ICAR), Rue & Held (2005) and Mohebbi (2011) demonstrated an application on how to apply it to analyse cancer data.

However, based on extensive studies by Wall (2004), the spatially correlated structure of ICAR approach is too complicated, involving complex implementation and lack of physical explanation. Renato & Krainski (2004) have also pointed out that preliminary knowledge and a good understanding are needed in determining and investigating the effect of the choice of precision for the covariance matrix. This approach is therefore less efficient for practitioners. Thus, it is essential to develop a more efficient method to model the spatial correlation, see Haining (1990); Bernadinelli *et al.* (1995).

An alternative way is using kriging under spatial statistics which is also used in geo-

statistics, see Diggle *et al.* (2007*a*). The Gaussian process is the backbone of the method and its performance shows promising results. The Gaussian processes have been widely use in the field of geostatistics since the 1960's. However the Gaussian process in this approach was limited to two or three dimensions with a known covariance kernel.

To provide a more flexible model, we extend the idea by proposing to use Gaussian process regression ($\mathcal{GPR}$) as the correlation structure. Recently $\mathcal{GPR}$ and related methods have been developed quickly and have a wide application in machine learning and other areas, see Rasmussen & Williams (2006). Some recent development can be found in Shi & Choi (2011) and Wang & Shi (2014). The covariance structure of $\mathcal{GPR}$ is defined by a covariance kernel which depends on a set of covariates. Consequently it provides a very flexible covariance structure coping with large dimensional covariates with a known covariance kernel and a variety of variables such as geographic position, distance among the spatial areas and other variables related to culture, lifestyle and even the previous observations.

However, in practice, there are many cases where it is more sensible to analyse the multivariate data together, for example, the dengue fever and malaria data discussed in the previous section or the cancer data (lung and oesophageal cancer) which will be discussed in Chapter 5. Research on a bivariate case is still an active area in statistics. Many researchers have studied and explored this topic, for example, Kocherladota & Kocherlakota (1992) have investigated the inference of bivariate Poisson regression; Karlis & Ntzoufras (2003) have developed an EM algorithm for bivariate Poisson distribution and related models; and Jung & Winkelmann (1993), Karlis & Ntzoufras (2003), and Karlis & Bermudez (2011); Vernic (1997) have applied a bivariate Poisson for two aspects of labour mobility, football champions, and insurance respectively. All these methods have some limitations, one major problem is that they did not model the cross-correlation between response variables.

Constructing a cross-correlation structure for each response is not an easy task because we need to ensure that the covariance matrix is positive definite. A variety of alternative approaches with special model structures have been proposed to overcome this problem. Kim *et al.* (2001) have presented a twofold CAR for two different diseases. However, this approach is only applicable to bivariate cases and it is difficult to extend to the larger dimensional case. Gelfand & Vounatsou (2003) have introduced a multivariate conditional autoregressive model (MCAR) based on Mardia (1988). The restriction of this approach is that it is not easy to understand the structure of the cross-correlation. Jin *et al.* (2005*b*) have tried to improve the method by providing a general framework for the multivariate conditional autoregressive model(GMCAR). The joint distribution of GMCAR is

$$p(z_1, z_2) = p(z_1 \mid z_2)p(z_2),$$

where z_1 and z_2 are the first and the second response variable respectively. This approach has one big problem, i.e. its performance depends on the ordering of response variables. Another restriction, if we increase the dimension of response variables to more than two, is that it leads to a very complex problem since it has many potential orderings. Thus, Jin *et al.* (2007) have further improved their method called co-regionalized model.

One big restriction on the above approaches is that they are not allowed to include other correlations apart from a spatial neighbourhood, see Martinez-Beneito (2013). The kriging method provides flexibility. Crainiceanu *et al.* (2012) have developed an extension of the kriging method. The approach has presented a natural method of smoothing the dependent bivariate data involving stationary Gaussian processes and it provides promising results for some problems. However, this approach has difficulty in capturing the individual characteristics of each response variable since the method uses a conditional dependency correlated structure.

As an alternative way, we can parameterize covariance functions directly and treat Gaussian processes as a convolution of a white noise processes with a smoothing kernel function. This approach is called a convolution method and performs well for multivariate dependent processes; it was introduced firstly by Boyle & Frean (2005). It also has provided huge flexibility and robustness since Andriluka *et al.* (2006) proposed a general framework for constructing dependent outputs using some of the stationary covariance structures.

We propose using the convolved Gaussian process priors to model the multivariate dependent Gaussian and non-Gaussian data. The convolution method has been used in the last decade, e.g. a multivariate spatial process have been presented by Paciorek (2003) and Majumdar *et al.* (2010). The computational efficiency of multiple outputs and its implementation in the dynamic model have also been investigated by Alvarez & Lawrence (2011) and Alvarez (2011).

There are some points worth noting regarding the advantages of our proposed model: (1) it offers a semiparametric regression model for Gaussian and non-Gaussian data with multivariate responses and multi-dimensional covariates; (2) it provides a natural framework for capturing the individual features on modelling mean structure and covariance structure simultaneously; (3) it enables us to handle a large dimensional covariate in covariance functions as priors.

1.3 Structure of the Thesis

The thesis is organized as follows. In Chapter 2, we review the development of Poisson regression for dependent univariate response variable. We focus on the spatial correlation structure of the count regression. We investigate several correlation structures for Poisson

data. In this chapter, we provide the details of the model and describe how we estimate unknown parameters. Simulation studies have been presented to help understand the performance of the approaches.

Chapter 3 still focuses on the Poisson regression problem with a univariate response variable but having a spatial correlation structure. We propose to use a Gaussian process prior to model correlated structure. We will provide details on how to define the model, how to estimate unknown parameters and hyper-parameters and how to calculate predictions. A comprehensive simulation study is conducted to illustrate the performance of the new method. It is also compared with some existing methods. The method has also been used to analyse our real data : Dengue fever data in Indonesia.

In Chapter 4, we extend the $\mathcal{GPR}$ approach to construct the cross-correlation of dependent Gaussian processes by using a convolution method. Several stationary covariance functions and a mixed form of covariance structure are investigated. Then, we propose the convolved Gaussian process ($\mathcal{CGP}$) priors for multivariate non-linear regression. We will extend the general framework proposed by Andriluka *et al.* (2006) and use a mixed covariance function to construct the covariance structure of the model. We also provide details on how to define the model, how to estimate parameters and hyper-parameters and how to calculate the predictions. Some simulation studies are presented to investigate the performance of the proposed model including how the model reacts from misspecified covariance structures schemes. We also compare the proposed model with an existing approach.

We apply the idea of using $\mathcal{CGP}$ priors to multivariate non-Gaussian data in Chapter 5. We focus on bivariate Poisson regression. We explain the details including the definition of the model, and explain how to estimate the unknown parameters and hyper-parameters and calculate the predictions. The asymptotic consistency is also reported in this chapter. We provide extensive simulation studies including a scenario which illustrates how the proposed model is able to offer high flexibility on the choice of covariance functions and to accommodate a large dimensional covariance function. We also apply this proposed model to two real data sets, i.e. dengue fever and malaria data and cancer data. The proposed method is also compared with existing methods.

Then, we provide a general model using $\mathcal{CGP}$ priors to other non-Gaussian data in an exponential family, such as binomial and ordinal data in Chapter 6. We present a natural framework of semiparametric multivariate regression for data following an exponential family distribution. Similar to the previous chapter, we also report the details of the model, inference for estimating unknown parameters and hyper-parameters and calculating the predictions. Consistency is explained in this chapter as well. Several simulation studies and real data, e.g. adverse birth outcome data are provided to investigate the performance of the proposed model including robustness and how it handles the difficulty of

large dimensional covariance functions. Finally, we conclude in Chapter 7 with comments regarding further research.

Chapter 2

Correlation Structures of Poisson Regression

2.1 Introduction

Regression analysis is a well established statistical method which is used to show the relationship between independent variables and dependent variables. Thus, it can be said that Poisson regression is analysing the relation between independent variables and a dependent variable which follows Poisson distribution, see e.g., Cameron & Trivedi (1998). The approach typically deals with count data which is very common in the area of public health.

One of our examples concerns dengue fever data which will be discussed in later chapters of the thesis. The output of the data is the number of dengue fever cases. We have began the investigation by exploring the transmission of the disease in the previous chapter through Figure 1.1. It has shown that there is strong evidence that the spread of the disease displays a certain geographical pattern. In fact, there is also considerable correlation between observed data since the disease is transmitted by a virus from mosquitoes. As a consequence, it seems more sensible if we include a spatial effect when analysing the data.

The problem now is that spatially correlated observations do not satisfy the independence assumption of Poisson regression as a part of a generalized linear model. However, to accommodate this issue, a generalized linear mixed model (GLMM) offers flexibility by including a correlated effect in the model. Thus, in this chapter, we will review the development of the correlation structures under GLMM and focus on dealing with spatial count data.

This chapter is organized as follows. Section 2 reviews the conventional Poisson method for modelling disease incidence rate. Some natural frameworks for capturing a spatial

correlation in Poisson regression are discussed in Sections 3 and 4.

2.2 Poisson Regression Model

In modelling disease incidence rates, there are a definition which needs to understand. The most common indicator used for comparing regional death rates is the standardized mortality ratio (SMR) in place i which is defined by

$$SMR_i = \frac{z_i}{E_i},$$

where z_i is observed number of deaths in place i and E_i is the expected number of deaths based on reference mortality rates applied to the regional demographic structure. Also E_i is defined as

$$E_i = \sum_g r_g \times p_{g,i},$$

where r_g is a standard mortality rate (e.g. national mortality rate) of demographic group g and $p_{g,i}$ is regional population size specific to demographic group g in place i. The demographic group g is usually determined by age or sex-age attributes.

In a generalized linear model (GLM), there is a link function h connecting mean from the dependent variable with independent variables ($\mathbf{U}$), i.e.

$$\begin{aligned} E\left(\frac{z_i}{E_i}\right) &= h\left(\frac{\mu_i}{E_i}\right) \\ &= \mathbf{U}_i^T\boldsymbol{\beta} \\ &= \beta_0 + \beta_1 U_{i1} + \beta_2 U_{i2} + ... + \beta_k U_{ik}. \end{aligned}$$

So we can write the above equation as follows

$$h\left(\frac{\mu_i}{E_i}\right) = \beta_0 + \beta_1 U_{i1} + \beta_2 U_{i2} + ... + \beta_k U_{ik}, \quad i = 1, ..., N, \tag{2.1}$$

where N is the number of observations and k is the number of predictors (covariates). Function h is called the link function. We can also rewrite (2.1) as

$$\frac{\mu_i}{E_i} = h^{-1}(\mathbf{U}_i^T\boldsymbol{\beta})$$

There are two link functions that can be used in Poisson regression. The first is the identity link and the second is the log link. The form of identity link function as follows :

$$\frac{\mu_i}{E_i} = \mathbf{U}_i^T\boldsymbol{\beta}.$$

The log link is

$$\log\left(\frac{\mu_i}{E_i}\right) = \mathbf{U}_i^T\boldsymbol{\beta}.$$

If we use the log link, the relationship between mean and independent is described below.

$$\mu_i \quad = \quad E_i e^{\mathbf{U}_i^T\boldsymbol{\beta}}.$$

The log link function is a more popular one because the link guarantees that the value of independent variable is non negative. In term of Poisson regression, the common link function is the log link function and the model is :

$$\mu_i \quad = \quad E_i \exp(\beta_0 + \beta_1 U_{i1} + \beta_2 U_{i2} + ... + \beta_k U_{ik}) = E_i e^{\mathbf{U}_i^T\boldsymbol{\beta}}. \tag{2.2}$$

Hence, the probability density function of Poisson regression can be written as follows

$$P_r(z_i;\boldsymbol{\beta}) = \frac{e^{-\mu_i}\mu_i^{z_i}}{z_i!}$$

where $\mu_i = E_i \exp(\mathbf{U}_i^T\boldsymbol{\beta})$ is the mean and $\boldsymbol{\beta}$ are unknown parameters. Meanwhile, the mean and the variance for Poisson regression model is defined as

$$\mu_i = E_i \exp(\mathbf{U}_i^T\boldsymbol{\beta}) = Var(z_i).$$

The likelihood function of Poisson regression is

$$L(\boldsymbol{\beta} \mid \mathbf{z}) = \prod_{i=1}^{N} \frac{e^{-\mu_i}\mu_i^{z_i}}{z_i!}. \tag{2.3}$$

And, the log likelihood function of equation (2.3) is therefore given by

$$\begin{aligned} \log(L(\boldsymbol{\beta} \mid \mathbf{z})) \quad &= \quad \sum_{i=1}^{N} (-\mu_i + z_i \log(\mu_i) - \log(z_i!)) \\ &= \quad \sum_{i=1}^{N} z_i \log(\mu_i) - \sum \mu_i + constant. \end{aligned}$$

Here μ_i is given in equation (2.2) and then unknown parameters $\boldsymbol{\beta}$ can be estimated by maximizing the above log likelihood, ignoring the constant.

2.3 Non-Spatial Correlation

Now, for the ith observation, let z_i denote the response variable, τ_i denote a random effect and U_i denote covariates for the fixed effects, where $i = 1, ...N$ and N is the sample size. Typically in a spatial GLMM for Poisson distribution, observations are assumed to be conditional independent and follow a Poisson distribution. Then, a Poisson log linear mixed model with non spatial random effect can be expressed as follows

$$\begin{aligned} z_i \mid \tau_i &\sim Poisson(\mu_i) \\ \log(\mu_i) &= \log(E_i) + \mathbf{U}_i^T\boldsymbol{\beta} + \tau_i \end{aligned} \tag{2.4}$$

where $\tau_i \sim \mathcal{N}(0, \sigma^2)$. The parameter θ, i.e. σ^2 indicates the variances in the population distribution and these random effects represents the influences of area i that are not captured by the observed covariates, Mohebbi (2011).

The random effects $\boldsymbol{\tau}$ are unknown , thus the marginal density of $\mathbf{z}$ does not have a convenient closed-form representation. To estimate parameters, we can do so by maximizing the following marginal density $\mathbf{z} = (z_1, ..., z_N)^T$

$$\begin{aligned} p(\mathbf{z}|\boldsymbol{\beta}, \theta) &= \int p(\mathbf{z}|\boldsymbol{\beta}, \boldsymbol{\tau})p(\boldsymbol{\tau})d\boldsymbol{\tau} \\ &= \int p(\mathbf{y}, \boldsymbol{\tau})d\boldsymbol{\tau} \end{aligned} \tag{2.5}$$

where $\boldsymbol{\tau} \sim \mathcal{N}(0, \sigma^2)$. Obviously, the above marginal density function is analytically intractable. One of the methods used to address this issue is to use a Laplace approximation. Note that

$$\begin{aligned} \Phi(\boldsymbol{\tau}) &= \log p(\mathbf{y}, \boldsymbol{\tau}) = \log p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\tau}) + \log p(\boldsymbol{\tau}) \\ &= \log p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\tau}) - \frac{N}{2}\log 2\pi - \frac{N}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{N}\tau_i^2 \end{aligned} \tag{2.6}$$

where $\log p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\tau}) = \sum_{i=1}^{N}(y_i \log(\mu_i) - \mu_i)$ with $\mu_i = \exp(\mathbf{U}_i^T\boldsymbol{\beta} + \tau_i)$ and then the log likelihood of equation (2.5) can be written as

$$l(\boldsymbol{\beta}, \boldsymbol{\theta}) = \sum_{i=1}^{N}\log\int \exp(\Phi(\boldsymbol{\tau}))d\boldsymbol{\tau}. \tag{2.7}$$

Let $\boldsymbol{\tau}_0$ be the maximiser of $\Phi(\boldsymbol{\tau})$, then a Laplace approximation is

$$\int \exp(\Phi(\boldsymbol{\tau}))d\boldsymbol{\tau} = \exp\left\{\Phi(\boldsymbol{\tau}_0) + \frac{N}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{H}|\right\} \tag{2.8}$$

where $\mathbf{H}$ is the negative of the second derivative of $\Phi(\boldsymbol{\tau})$ respect to $\boldsymbol{\tau}$ and evaluated at $\boldsymbol{\tau}_0$, see Wood (2012). We have $\mathbf{H} = \mathbf{C} - diag(\frac{1}{\sigma^2}, ..., \frac{1}{\sigma^2})$ and $\mathbf{C}$ is a diagonal matrix,

$$\mathbf{C} = Diag\left(\exp(\mathbf{U}_1^T\hat{\boldsymbol{\beta}} + \tau_{01}), ..., \exp(\mathbf{U}_N^T\hat{\boldsymbol{\beta}} + \tau_{0N})\right).$$

In order to estimate the parameters, we maximize the likelihood function with Laplace approximation in equation (2.7).

2.3.1 Simulation Study

In this section, we provide an illustration to investigate the performance of the model in equation (2.4). The procedure is as follows :

i. The true value parameters of $\theta = \sigma^2$ and $\boldsymbol{\beta} = (\beta_0, \beta_1)$ are $1, 2$ and 3 respectively

ii. Generate $U \sim \mathcal{N}(0, 1)$ and the random effect $\tau_i \sim \mathcal{N}(0, \sigma^2)$

iii. Calculate $\mu_i = \exp(\beta_0 + \beta_1 U_i + \tau_i)$

iv. Generate response response $z_i \sim Poisson(\mu_i)$

v. Estimate parameters by maximizing the log likelihood of the marginal distribution from equation (2.7) and (2.8).

In order to measure model performance, we use root mean square error (RMSE) which can be defined as follows.

$$RMSE_j = \left(\frac{\sum_{i=1}^{r}\left(\beta_{ij} - \hat{\beta}_{ij}\right)^2}{r}\right)^{\frac{1}{2}} \tag{2.9}$$

where $\hat{\beta}$ is an estimated parameter and r is the number of replication. Table 2.1 shows the sample mean and the average of RMSE based on one hundred replications for two different sample sizes. From Table 2.1, it is clear that the estimated parameters tend to

	Sample Mean			RMSE		
Sample size	σ^2	β_0	β_1	σ^2	β_0	β_1
200	0.9856	2.0030	2.9984	0.0508	0.0197	0.0096
600	0.9827	2.0000	2.9999	0.0311	0.0097	0.0050

Table 2.1: Sample mean and RMSE with different sample size based on one hundred replications.

the true values and the RMSE is decreasing as the sample size increases as we expected.

2.4 Neighbourhood-based Spatial Correlation

In this section, we use a model similar to equation (2.4). The Poisson regression with neighbourhood-based spatial effect can be specified as follows

$$\begin{aligned} z_i \mid \tau_i &\sim Poisson(\mu_i) \\ \mu_i &= \exp(U^T\boldsymbol{\beta} + \tau_i) \\ \log(\mu_i) &= U_i^T\boldsymbol{\beta} + \tau_i, \quad i = 1..., N \end{aligned}$$

where the random effect τ_i is assumed to have general conditional autoregressive structure which can be defined as

$$\tau_i|\tau_j \sim \mathcal{N}(\sum_{j\neq i} b_{ij}\tau_j, \sigma_i^2).$$

The component b_{ij} is the weight of each other observation on the mean of τ_i and σ_i^2 is a variance for τ_i. As an example, if state i has M neighbours and $b_{ij} = \frac{1}{M}$ for every state that is a neighbour and 0 otherwise, then the conditional mean of a state's observation is the mean of all neighbours' observations.

In order to estimate the parameters involved in the model, we need to derive the joint distribution of all observations. Unfortunately, the above assumption just gives us full conditional distributions. But, by applying Brook's lemma with Gaussian conditional, we can define the following

$$p(\tau_1, ..., \tau_n) \propto \exp\left\{\frac{-1}{2}\boldsymbol{\tau}'\mathbf{D}^{-1}(\mathbf{I} - \mathbf{B})\boldsymbol{\tau}\right\} \tag{2.10}$$

where $\mathbf{B} = b_{ij}$ and $D_{ii} = \sigma_i^2$. This implies that a joint multivariate normal distribution for $\boldsymbol{\tau}$ with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma} = (\mathbf{I} - \mathbf{B})^{-1}\mathbf{D}$, Banerjee *et al.* (2011).

There are two problems that we need to consider before we estimate parameters of the model using a joint distribution for all observations given the conditional neighbour. The first problem is symmetry. It can be addressed by making some definition, i.e. $b_{ij}\sigma_j^2 = b_{ji}\sigma_i^2$ or $\frac{b_{ij}}{\sigma_i^2} = \frac{b_{ji}}{\sigma_j^2}$, Monogan (2012). However, $\mathbf{B}$ is still not symmetric. The proximity matrix $\mathbf{W}$ where entries w_{ij} (with $w_{ii} = 0$) and the choices for w_{ij}. For example w_{ij} is equal to 1 if i, j share a common boundary. Now, we assume that $\mathbf{W}$ is symmetric when we suppose the $b_{ij} = \frac{w_{ij}}{w_{i+}}$ and $\sigma_i^2 = \frac{\sigma^2}{w_{i+}}$. Thus, we can rewrite the joint distribution in equation (2.10) as following

$$p(\tau_1, ..., \tau_n) \propto \exp\left\{\frac{-1}{2\sigma^2}\boldsymbol{\tau}'(\mathbf{D}_w - \mathbf{W})\boldsymbol{\tau}\right\}$$

where $\mathbf{D}_w$ is diagonal with $(D_w)_{ii} = w_{i+}$ and $\mathbf{W} = w_{ij}$. Here w_{ij} is a measure of the asso-

ciation between two observations (typically equal to 1 if the observations are neighbours and 0 otherwise) and w_{i+} is the sum of all w_{ij} for observation i.

The second problem is that we need to ensure that $(\mathbf{D}_w - \mathbf{W})$ is positive definite to provide a valid joint distribution. This does not seem easy since in practice, it might not be positive semi definite. Moreover, $(\mathbf{D}_w - \mathbf{W})\mathbf{1} = \mathbf{0}$, if we define matrix $\mathbf{W}$ as w_{ij} which is equal to 1 if the observations are neighbours and 0 otherwise. As consequence the inverse of covariance matrix $\mathbf{\Sigma}^{-1}$ is singular meaning that $\mathbf{\Sigma}$ does not exist.

As a consequence, in this case, it provides an improper distribution which is called intrinsic conditional autoregressive model (ICAR) model or intrinsic Gaussian Markov random field, Rue & Held (2005). We can rewrite the model ICAR as follows

$$p(\tau_1, ..., \tau_n) \propto \exp\left\{\frac{-1}{2\sigma^2}\sum_{i\neg j} w_{ij}(\tau_i - \tau_j)^2\right\}.$$

The distribution does not change if we add any constant to a τ_i, thus it is considered that p is an improper distribution. To address this issue, we need a constraint in the model such as

$$\sum_i \tau_i = 0$$

and can redefine the inverse of the covariance matrix as follows.

$$\mathbf{\Sigma}^{-1} = \mathbf{D}_w - \rho\mathbf{W}$$

and choose ρ to make $\mathbf{\Sigma}^{-1}$ non singular. This is guaranteed if

$$\rho \in (\frac{1}{\lambda_1}, \frac{1}{\lambda_n}),$$

where $\lambda_1 < ... < \lambda_n$ are ordered eigenvalues of $\mathbf{D}_w^{-\frac{1}{2}}\mathbf{W}\mathbf{D}_w^{-\frac{1}{2}}$. The bounds can be simplified, by replacing $\mathbf{W}$ with $\tilde{\mathbf{W}} = Diag(\frac{1}{w_{i+}})\mathbf{W}$. Then,

$$\mathbf{\Sigma}^{-1} = \mathbf{D}_w(\mathbf{I} - \alpha\tilde{\mathbf{W}})$$

where $\mathbf{D}_w$ is diagonal and if $|\alpha| < 1$, then $\mathbf{I} - \alpha\tilde{\mathbf{W}}$ is nonsingular, see Carlin & Banerjee (2003).

2.4.1 Correlation in the Neighbourhood GLMM

In this subsection, we provide an illustration in order to understand how α of a covariance structure influences the correlation matrix. As we have explained in the previous subsection, we can write that $\tau_1, ...\tau_N \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = (\sigma^{-2}\mathbf{Q})^{-1}$ and $\mathbf{Q} = \mathbf{D}_w(\mathbf{I} - \alpha\tilde{\mathbf{W}})$.

Figure 2.1: Map of East Java province, Indonesia (http://mapsof.net/indonesia/east-java-province)

In practice, it does not seem easy to investigate the correlation for each observation. Thus, we need to investigate how the precision matrix and other parameters can affect the covariance matrix. We will show how α plays an important role in terms of the performance of correlation matrix. Specially, we use dengue data cases in East Java, Indonesia in 2010 which will discuss in the next chapter. The output is the number of dengue fever cases in every city and there are 38 cities in total. We left two cities out since they have missing data. Figure 2.1 shows the real map based on present circumstances. In this investigation, we focus on the 1st and 2nd city, i.e. Ponorogo and Trenggalek respectively. From Figure 2.1, it is clear that between Ponorogo and Trenggalek there could be a high correlation because they share a common boundary.

From Figure 2.2 and 2.3 we see the different performance of the correlation matrix based on two choices of α. From Figure 2.2, there is considerable evidence to show that by taking $\alpha = 0.9$, the correlation between two cities is sensible. It is because both areas, which are represented by 1 and 2, have a dark color which means that the areas are highly correlated. We now compare the performance by taking another value of α, say 0.75 which can be found in Figure 2.3. As we expected, the choice of the value of α has a big impact in terms of correlation matrix performance. From Figure 2.3, it shows that the correlation between the two areas (1 and 2) has changed into a low correlation which is not acceptable based on the actual situation.

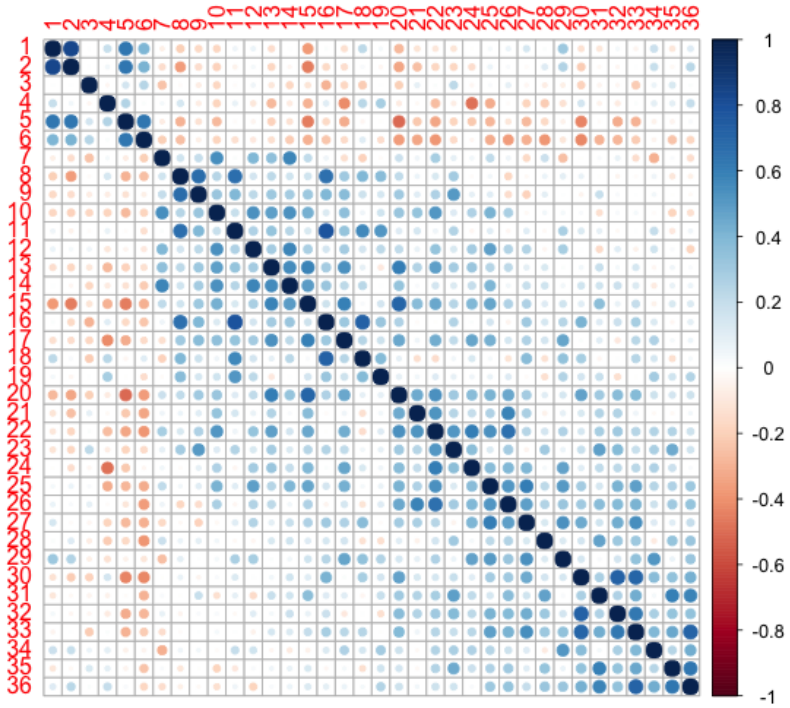

Figure 2.2: The correlation matrix from 36 cities in East Java province, Indonesia based on $\alpha = 0.9$.

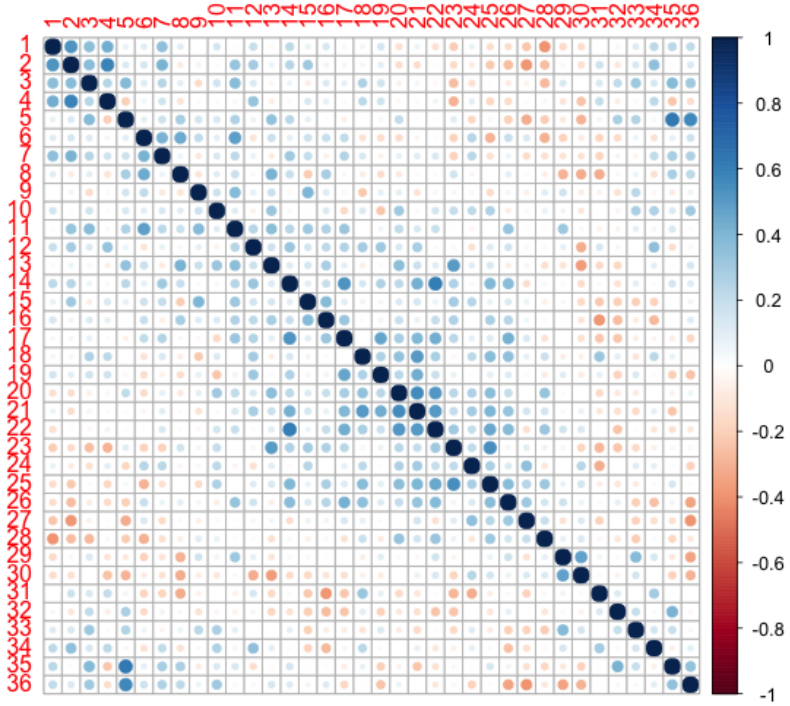

Figure 2.3: The correlation matrix from 36 cities in East Java province, Indonesia based on $\alpha = 0.75$

2.4.2 Simulation Study

Similar to the previous simulation study, the aim of this subsection is to understand the performance of the spatially based neighbourhood model. The procedure used to simulate the data is as follows :

i, Give the true value as parameters of theta β_0, β_1 and σ^2, i.e. $6.7, -0.002$ and 0.8 respectively

ii, Generate random effects; e.g. $\tau_1, ...\tau_N \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, where $\mathbf{\Sigma} = (\sigma^{-2}\mathbf{Q})^{-1}$ and $\mathbf{Q} = \mathbf{D}_w(\mathbf{I} - \alpha\tilde{\mathbf{W}})$. The bounds can be simplified, by replacing $\mathbf{W}$ by $\tilde{\mathbf{W}} = Diag(\frac{1}{w_{i+}})\mathbf{W}$. Define $\mathbf{D}_w$ as a diagonal matrix with $(D_w)_{ii} = w_{i+}$ and $\mathbf{W} = w_{ij}$. Here, w_{ij} is a measure of the association between two observations (typically equal to 1 if the observations are neighbours and 0 otherwise) and w_{i+} is the sum of all w_{ij} for observation i. For α, if $|\alpha| < 1$, then $\mathbf{I} - \alpha\tilde{\mathbf{W}}$ is non singular and here, we define $\alpha = 0.5$. Hence, we can generate $\tau_1, ...\tau_N \sim \mathcal{MVN}(\mathbf{0}, (\sigma^{-2}(\mathbf{D}_w(\mathbf{I} - 0.5\tilde{\mathbf{W}}))^{-1})$.

iii, Generate the covariates $U_i \sim \mathcal{N}(0, 1)$ and calculate the mean $\mu_i = \exp(\beta_0 + \beta_1 U_i + \tau_i)$

iv, Generate response variable $z_i \sim Poisson(\mu_i)$

v, Estimate parameters by redefining equation (2.6) as follows

$$\Phi(\boldsymbol{\tau}) \quad = \quad \log p(\mathbf{z}|\boldsymbol{\beta}, \boldsymbol{\tau}) - \frac{1}{2}\log|\mathbf{\Sigma}| - \frac{1}{2}\boldsymbol{\tau}^T\mathbf{\Sigma}^{-1}\boldsymbol{\tau} - \frac{N}{2}\log 2\pi \tag{2.11}$$

where N is the number of observations.

	σ	β_0	β_1
True values	0.8	6.7	-0.002
Sample mean	0.79989	6.70175	-0.002016
Average SD	0.00034	0.01018	0.000132
RMSE	0.00018	0.01485	0.000133

Table 2.2: Sample mean, the average of standard deviation (Average SD) taken from the information matrix and the RMSE based on one hundred replications.

Table 2.2 shows that estimated parameters (Sample mean) and the value of the average standard deviation (Average SD) which can be calculated from information matrix or the inverse of negative Hessian matrix. Meanwhile, the Hessian matrix is the second derivative of marginal likelihood respect to parameters. Another measurement is the value of root mean square (RMSE) between estimated parameters and true values as it is defined in equation (2.9) based on one hundred replications. The estimated parameters perform reasonably well. Although one of the parameters have a slightly different value of the

average SD and RMSE, the measurement performance values of other parameters tend to be similar each other as we would expect.

2.5 Chapter Summary

In this chapter, we have reviewed the development of the correlation structure in Poisson regression. We noticed that a spatial correlation structure based on neighbourhood is more realistic in capturing spatially correlated compared with non-spatial structure or conventional Poisson regression. The performance of the spatial approach provided a promising result.

However, there are some points worth noting in terms of the neighbours method, i.e. in practice, we need to ensure that $\mathbf{Q} = \mathbf{D}_w(\mathbf{I} - \alpha\tilde{\mathbf{W}})$ is at least semi-definite positive by taking some number for $|\alpha| < 1$. If we take $\alpha = 0$, it can be interpreted that z_i become independent. Therefore, this seems to offer a less convenient approach for practitioners since we know that the performance of the correlation matrix might change significantly if we take different values. We then need to find a more flexible and efficient way of handling spatial correlated count data. This is the aim of the next chapter.

Chapter 3

Spatial Poisson Modelling using Gaussian Process Regression

3.1 Introduction

In recent years, modelling for count data has developed quickly, particularly for health data. The conventional way to handle this situation is using a Poisson regression model by assuming independence among the data. However, in practice this is a strong assumption since the majority of health data are dependent. One popular approach is to solve the problem by using a conditional autoregressive model to model the correlation structure, which is first proposed in Besag & Kooperberg (1995).

In the previous chapter, we discussed the performance of the ICAR model. It provided a promising result, but we need to have a basic knowledge of how to choose the precision matrix and other parameters in order to define the covariance matrix. Therefore, further development is needed. Kriging is another well known geostatistical, approach based on stationary Gaussian processes, see Diggle *et al.* (2007*a*). However, this approach has limitations, i.e. it is not allowed to involve more than two or three dimensions of covariates in the covariance structure. Thus, the aim of this chapter is to propose a model which can describe the spatial covariance flexibly. To overcome the problem, we propose a method which enables us to accommodate a large dimensional covariance structure. To achieve this we extend the method by using a Gaussian process regression ($\mathcal{GPR}$) model.

This chapter will be organized as follows. Sections 3.2 and 3.3 will describe the details on how we use a $\mathcal{GPR}$ to model dependency and provide the technical details for inference and implementation respectively. comprehensive numerical investigations including simulation studies and a real example will be presented to demonstrate the performance of the proposed method in Section 3.4.

3.2 Gaussian Process Regression ($\mathcal{GPR}$)

Let y_i be a response variable and $\mathbf{x}_i$ be P-dimensional covariates. A Gaussian process regression model can be specified as follows :

$$y_i = f(\mathbf{x}_i) + \epsilon_i, \quad i = 1, ..., N \tag{3.1}$$

where $\epsilon_i \sim i.i.d \quad \mathcal{N}(0, \sigma^2)$, σ^2 unknown and

$$f(.) \sim \mathcal{GP}(\mu(\cdot), k(\cdot, \cdot)),$$

if

$$\mathbf{f} = (f_1, f_2, .., f_N) \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{K}),$$

where the i-th element of $\boldsymbol{\mu}$ is $\mu(\mathbf{x}_i)$, the (i, j)-th element of $\mathbf{K}$ is $k(\mathbf{x}_i, \mathbf{x}_j)$ and $k(\cdot, \cdot)$ is a covariance function.

It is common to assume a zero mean function in the Gaussian process prior, i.e. $\mu(.) = 0$. The most popular choice for the covariance function is the following squared exponential kernel,

$$k(\mathbf{x}_i, \mathbf{x}_j) = v_0 \exp\left\{ -\frac{1}{2} \sum_{p=1}^{P} w_p (x_{ip} - x_{jp})^2 \right\} \tag{3.2}$$

where $v_0, w_p, p = 1, ...P$ denotes the set of hyper-parameters and are defined as $\boldsymbol{\theta}$. Making a suitable choice of a kernel covariance function and its hyper-parameters can improve the prediction accuracy. One of the popular methods is using empirical Bayesian estimation to select hyper-parameters, Shi & Choi (2011).

3.2.1 Empirical Bayesian Estimates

As in the previous section the problem of estimating hyper-parameters is an important topic. In this case, rather than making assumptions on the probability structure for the hyper-parameters, the empirical Bayesian approach uses the observed data to estimate them.

Now, we focus on the empirical Bayesian approach. Let us assume that we have observed a set of data $\mathcal{D} = (y_i, \mathbf{x}_i), i = 1, ..., N, \mathbf{x}_i \in T \subset \mathcal{R}^P$, P is the dimension of the input vector $\mathbf{x}_i$. A Gaussian process regression model is generally formulated as

$$y_i | f_i \sim g(f_i) \tag{3.3}$$

$$(f_1, ..., f_N) \sim \mathcal{GP}(\mathbf{0}, k(\cdot, \cdot)) \tag{3.4}$$

where $k(\mathbf{x}_i, \mathbf{x}_j) = cov(f(\mathbf{x}_i), f(\mathbf{x}_j))$. Here $g(f_i)$ may be a normal distribution with mean f_i and variance σ^2. In this case (3.3) can be rewritten as

$$y_i|f_i \sim \mathcal{N}(f_i, \sigma^2). \tag{3.5}$$

The observations y_i may have some other distributions, for example Poisson distribution which will be discussed in the next section. From (3.3) and (3.4) we can obtain the marginal likelihood of $\mathbf{y} = (y_1, ..., y_N)^T$ given $\boldsymbol{\theta}$, i.e.

$$p(\mathbf{y}|\boldsymbol{\theta}) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\boldsymbol{\theta})d\mathbf{f} \tag{3.6}$$

where $p(\mathbf{y}|\mathbf{f}) = \prod_{i=1}^{n} g(f_i)$ and $\mathbf{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$. Consequently for the continuous response with normal distribution as given in (3.5), the marginal (3.6) has an analytical form as a multivariate normal. Thus, the marginal distribution of $\mathbf{y}$ is a normal distribution $N(\mathbf{0}, \boldsymbol{\Psi})$ where $\boldsymbol{\Psi} = \mathbf{K} + \sigma_\epsilon^2\mathbf{I}$. Hence, the marginal log likelihood of $\boldsymbol{\theta}$ is written as

$$L(\boldsymbol{\theta}|\mathcal{D}) = -\frac{1}{2}\log|\boldsymbol{\Psi}(\boldsymbol{\theta})| - \frac{1}{2}\mathbf{y}^T\boldsymbol{\Psi}(\boldsymbol{\theta})^{-1}\mathbf{y} - \frac{N}{2}\log 2\pi \tag{3.7}$$

Now, $\boldsymbol{\theta} = (w_p, v_0)$ is estimated by maximizing the above log likelihood. In practice, we often estimate σ^2 and $\boldsymbol{\theta}$ at the same time. For other distributions, it is not straightforward to calculate the marginal likelihood in equation (3.6). The model with Poisson data will be discussed in the next section. Other types of data will be discussed in Chapter 6.

3.2.2 Fitted Values and Predictions

From the model structure defined in (3.3), we can calculate posterior probabilities of $\mathbf{f} = (f(\mathbf{x}_1), ..., f(\mathbf{x}_N))^T$. Let $\mathbf{K}$ be a $N \times N$ covariance matrix evaluated at all pairs of the N design points. Furthermore, the observed variables $\mathbf{y} = (y_1, ..., y_N)^T$ given $\mathbf{f}$ and noise variance σ^2 have a multivariate normal distribution with mean $\mathbf{f}$ and covariance $\sigma^2\mathbf{I}$ in which $\mathbf{f}$ also follows a multivariate normal with mean $\mathbf{0}$ and covariance $\mathbf{K}$, see (Shi & Choi, 2011). We can express it as follows :

$$(y_1, ..., y_N|\mathbf{f}, \sigma^2) \sim \mathcal{N}(\mathbf{f}, \sigma^2\mathbf{I})$$

where $\mathbf{f} \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$. Thus, the posterior distribution of $\mathbf{f}$, $p(\mathbf{f}|\mathcal{D}, \sigma^2)$ from a set of data $(\mathcal{D})$ is proportional to the product of two normal distributions.

$$p(\mathbf{f}|\mathcal{D}, \sigma^2) \propto \varphi_N(\mathbf{y}|\mathbf{f}, \sigma^2\mathbf{I})\varphi_N(\mathbf{f}|\mathbf{0}, \mathbf{K})$$

where $\varphi_N(.|\boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the density function of N-variate normal distribution with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. Hence, $\varphi_N(\mathbf{y}|\mathbf{f}, \sigma^2\mathbf{I})$ denotes as a multivariate normal distribution with mean $\mathbf{f}$ and covariance matrix $\sigma^2\mathbf{I}$. Also $\varphi_N(\mathbf{f}|\mathbf{0}, \mathbf{K})$ defines as a multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $\mathbf{K}$.

When $\boldsymbol{\theta}$ is given (or estimated), the posterior distribution $p(\mathbf{f}|\mathcal{D}, \sigma^2)$ is a multivariate normal distribution with

$$E(\mathbf{f}|\mathcal{D}, \sigma^2) = \mathbf{K}(\mathbf{K} + \sigma^2\mathbf{I})^{-1}\mathbf{y},$$

$$Var(\mathbf{f}|\mathcal{D}, \sigma^2) = \sigma^2\mathbf{K}(\mathbf{K} + \sigma^2\mathbf{I})^{-1}.$$

The marginal distribution of $\mathbf{y}$, $p(\mathbf{y})$ is also given by a multivariate normal distribution,

$$\mathbf{y} = (y_1, ..., y_N) \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Psi}) \tag{3.8}$$

where $\boldsymbol{\Psi}$ is a $N \times N$ matrix, of which the (i, j) th element is defined as

$$\boldsymbol{\Psi}(i, j) = Cov(y_i, y_j) = k(\mathbf{x}_i, \mathbf{x}_j) + \sigma^2\delta_{ij} \tag{3.9}$$

where δ_{ij} is the Kronecker delta.

In terms of prediction, let $\mathbf{x}^*$ be a new input and $f(\mathbf{x}^*)$ be the related nonlinear function. Thus, $(f(\mathbf{x}_1), ..., f(\mathbf{x}_N), f(\mathbf{x}^*))$ can be assumed to be the same Gaussian process as the training data. Hence, the posterior distribution of $f(\mathbf{x}^*)$ given the training data $\mathcal{D}$ is a Gaussian distribution with

$$E(f(\mathbf{x}^*)|\mathcal{D}, \sigma^2) = \boldsymbol{\Psi}^T(\mathbf{x}^*)\boldsymbol{\Psi}^{-1}\mathbf{y} \tag{3.10}$$

$$Var(\mathbf{f}(\mathbf{x}^*)|\mathcal{D}, \sigma^2) = k(\mathbf{x}^*, \mathbf{x}^*) - \boldsymbol{\Psi}^T(\mathbf{x}^*)\boldsymbol{\Psi}^{-1}\boldsymbol{\Psi}(\mathbf{x}^*) \tag{3.11}$$

where $\Psi(\mathbf{x}^*) = (k(\mathbf{x}^*, \mathbf{x}_1), ..., k(\mathbf{x}^*, \mathbf{x}_N))^T$ is covariance between $f(\mathbf{x}^*)$ and $\mathbf{f} = (f(\mathbf{x}_1, ..., f(\mathbf{x}_N))$. Also $\boldsymbol{\Psi}$ is covariance matrix of $\mathbf{y} = (y_1, ..., y_N)$.

The predictive distribution of $y^* = f(x^*) + \epsilon^*$ is also a Gaussian distribution with the same mean as equation 3.10 and variance given by

$$k(\mathbf{x}^*, \mathbf{x}^*) - \boldsymbol{\Psi}^T(\mathbf{x}^*)\boldsymbol{\Psi}^{-1}\boldsymbol{\Psi}(\mathbf{x}^*) + \sigma^{*2}.$$

3.3 Spatial Poisson Model with $\mathcal{GPR}$

We now use a $\mathcal{GPR}$ to model the correlated structure for non-Gaussian data; it can be called as a spatial generalized linear mixed model (SGLMM $\mathcal{GPR}$). In this chapter , we focus on Poisson data, i.e. the model is defined by

$$z_i|\tau_i \sim Poisson(\mu_i) \tag{3.12}$$

$$\log(\mu_i) = \mathbf{U}_i^T \boldsymbol{\beta} + \tau(\mathbf{x}_i), i = 1, ..., N$$

where N is the sample size and $\tau(\cdot) \sim \mathcal{GP}(\mathbf{0}, k(\cdot, \cdot)$ has been explained in the previous section.

3.3.1 Empirical Bayesian Estimates

As in the previous chapter, the problem of estimating the hyper-parameters $\boldsymbol{\theta}$ and coefficient parameters ($\boldsymbol{\beta}$) is a significant topic. We use an empirical Bayesian approach to estimate hyper-parameters $\boldsymbol{\theta}$. In practice, $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ can be estimated at the same time.

We recall how we obtain the marginal likelihood of $\mathbf{z} = (z_1, ..., z_N)^T$ given $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ from (3.6) as follows.

$$\begin{aligned} p(\mathbf{z}|\boldsymbol{\beta}, \boldsymbol{\theta}) &= \int p(\mathbf{z}|\boldsymbol{\beta}, \boldsymbol{\tau}) p(\boldsymbol{\tau}|\boldsymbol{\theta}) d\boldsymbol{\tau} \\ &= \int \Phi(\boldsymbol{\tau}) d\boldsymbol{\tau} \end{aligned} \tag{3.13}$$

where $\boldsymbol{\tau} \sim \mathcal{N}(\mathbf{0}, \mathbf{K})$ and

$$\Phi(\boldsymbol{\tau}) \quad = \quad \log p(\mathbf{z}|\boldsymbol{\beta}, \boldsymbol{\tau}) - \frac{1}{2} \log |\mathbf{K}(\boldsymbol{\theta})| - \frac{1}{2} \boldsymbol{\tau}^T \mathbf{K}^{-1}(\boldsymbol{\theta}) \boldsymbol{\tau} - \frac{N}{2} \log 2\pi. \tag{3.14}$$

It is impossible to calculate (3.13) directly since the dimension of the integration is N, which is usually a large number. We use a Laplace approximation as discussed around equation (2.8).

3.3.2 Predictions

It is of interest to predict z^* at a new test input $\mathbf{x}^*$. The main purpose in this section is to calculate $E(z^*|\mathcal{D})$ and $Var(z^*|\mathcal{D})$.

Let $\tau(\mathbf{x}^*) = \tau^*$ be the underlying latent variable at $\mathbf{x}^*$. The expectation of z^*, conditional on τ^* is given by

$$E(z^*|\tau^*, \mathcal{D}) = \exp(\mathbf{U}^{*T} \hat{\boldsymbol{\beta}} + \tau^*). \tag{3.15}$$

It follows that

$$E(z^*|\mathcal{D}) = E[E(z^*|\tau^*, \mathcal{D})] = \int \exp(\mathbf{U}^{*T} \hat{\boldsymbol{\beta}} + \tau^*) p(\tau^*|\mathcal{D}) d\tau^*. \tag{3.16}$$

One method to calculate the above expectation is to approximate $p(\tau^*|\mathcal{D})$ using a Laplace

approximation. We can rewrite it as

$$\begin{aligned} p(\tau^*|D) &= \int p(\tau^*|\boldsymbol{\tau}, D)p(\boldsymbol{\tau}|D)d\boldsymbol{\tau} \\ &= \int p(\tau^*, \boldsymbol{\tau}|D)d\boldsymbol{\tau} \\ &= \frac{1}{p(\mathbf{z})}\int p(\mathbf{y}|\boldsymbol{\tau})p(\tau^*, \boldsymbol{\tau})d\boldsymbol{\tau}. \end{aligned} \tag{3.17}$$

For convenience, we denote $(\boldsymbol{\tau}, \tau^*)^T$ and its covariance matrix $K_{N+1,N+1}$ by $\boldsymbol{\tau}_+$ and $\mathbf{K}_+$ respectively. Thus, the equation (3.16) can be written as

$$\frac{1}{p(\mathbf{y})}\int \left[\exp(\mathbf{U}^{*T}\hat{\boldsymbol{\beta}} + \tau^*)\prod_{i=1}^{N} p(z_i|\hat{\boldsymbol{\beta}}, \tau_i)\right]\left[(2\pi)^{-\frac{(N+1)}{2}}|\mathbf{K}_+|^{-\frac{1}{2}}\exp(-\frac{1}{2}\boldsymbol{\tau}^T\mathbf{K}_+^{-1}\boldsymbol{\tau}_+)\right]d\tau_+ \tag{3.18}$$

The calculation of the integral is not tractable, since the dimension of $\boldsymbol{\tau}_+$ is usually very large. We still use a Laplace approximation. Note that

$$\tilde{\boldsymbol{\Phi}}(\boldsymbol{\tau}_+) = \log(\exp(\mathbf{U}^{*T}\hat{\boldsymbol{\beta}}+\tau^*))+\sum_{i=1}^{N}\log p(z_i|\hat{\boldsymbol{\beta}}, \tau_i)-\frac{N+1}{2}\log(2\pi)-\frac{1}{2}\log|\mathbf{K}_+|-\frac{1}{2}\boldsymbol{\tau}_+^T\mathbf{K}_+^{-1}\boldsymbol{\tau}_+$$

where $\log p(\mathbf{z}|\boldsymbol{\beta}, \boldsymbol{\tau}) = \sum_{i=1}^{N}(z_i\log(\mu_i) - \mu_i)$ with $\mu_i = \exp(\mathbf{U}^{*T}\hat{\boldsymbol{\beta}} + \tau_i)$. Equation (3.17) can be expressed as

$$p(z^*|\mathcal{D}) = \frac{1}{p(\mathbf{z})}\int \exp(\tilde{\boldsymbol{\Phi}}(\boldsymbol{\tau}_+))d\boldsymbol{\tau}_+.$$

Let $\hat{\boldsymbol{\tau}_+}$ be the maximiser of $\tilde{\boldsymbol{\Phi}}(\boldsymbol{\tau}_+)$, then by using Laplace approximation we have

$$\int \exp(\tilde{\boldsymbol{\Phi}}(\boldsymbol{\tau}_+))d\boldsymbol{\tau}_+ = \exp(\tilde{\boldsymbol{\Phi}}(\hat{\boldsymbol{\tau}_+}) + \frac{N+1}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{K}_+^{-1} + \mathbf{C}_+|) \tag{3.19}$$

where $\mathbf{C}_+$ is the negative of the second derivative of

$$(\mathbf{U}^{*T}\hat{\boldsymbol{\beta}} + \tau^*) + \sum_{i=1}^{N}\left[z_i(\mathbf{U}_i^T\hat{\boldsymbol{\beta}} + \tau_i) - \exp(\mathbf{U}_i^T\hat{\boldsymbol{\beta}} + \tau_i)\right]$$

with respect to $\boldsymbol{\tau}_+$ and evaluated at $\hat{\boldsymbol{\tau}}_+$. Here $\mathbf{C}_+$ is a diagonal matrix, i.e.

$$\mathbf{C}_+ = Diag\left(\exp(\mathbf{U}_1^T\hat{\boldsymbol{\beta}} + \hat{\tau_1}), ..., \exp(\mathbf{U}_N^T\hat{\boldsymbol{\beta}} + \hat{\tau_N}), 0\right).$$

Since $p(\mathbf{z})$ has already been investigated, the predictive mean (3.16) can be calculated.

To calculate $Var(z^*|\mathcal{D})$, we use the formula :

$$Var(z^*|\mathcal{D}) = E[Var(z^*|\tau^*, \mathcal{D})] + Var[E(z^*|\tau^*, \mathcal{D})]. \tag{3.20}$$

Because $Var(z^*|\tau^*, \mathcal{D}) = E(z^*|\tau^*, \mathcal{D})$, therefore

$$E[Var(z^*|\tau^*, \mathcal{D})] = E(z^*|\mathcal{D}). \tag{3.21}$$

From the model definition, we have

$$\begin{aligned} Var(E[z^*|\tau^*, D]) &= E[E(z^*|\tau^*, D)]^2 - [E[E(z^*|\tau^*, \mathcal{D})]]^2 \\ &= \int (\exp(\mathbf{U}^{*T}\hat{\boldsymbol{\beta}} + \tau^*))^2 p(\tau^*|\mathcal{D}) d\tau^* - [E(z^*|\tau^*, \mathcal{D})]^2. \end{aligned} \tag{3.22}$$

The first item in (3.22) can be obtained by Laplace approximation similar to $E(z^*|\mathcal{D})$ in (3.16). The second item is the square of (3.17).

3.4 Numerical Examples

In this section we demonstrate several numerical examples including a simulated study and a real application using dengue fever data from East Java, Indonesia.

3.4.1 Simulation study: Scenario 1

We generate random numbers from a Poisson distribution using the algorithm below. The first issue is to estimate parameters using empirical Bayesian estimates and below is the procedure for generating data and estimating parameters. The simulation example is similar to our Dengue fever data. The correlated structure depends on the location (latitude and longitude) of each area.

i. Give the true values as the parameters $\boldsymbol{\beta} = (\beta_0, \beta_1)$ are 1 and 2 respectively and the hyper-parameters of covariance kernel $\boldsymbol{\theta} = (w, v)$ are 0.04 and 1 respectively.

ii. Generate covariate $U \sim \mathcal{N}(0, 1)$

iii. Generate random effect, we generate $\tau_i \sim \mathcal{GP}(\mathbf{0}, k(\mathbf{x}_i, \mathbf{x}_j))$. Here $k(\mathbf{x}_i, \mathbf{x}_j)$ is squared exponential covariance function based on previous section where $\mathbf{x}$ is latitude and longitude from each city in Dengue fever data

iv. Generate the response variable , i.e Poisson distribution

$$\mu_i = \exp(\beta_0 + \beta_1 U_i + \tau_i)$$

$$z_i \sim Poisson(\mu_i)$$

v. Estimate parameters by maximizing equation (3.14) with Laplace aproximation.

We generate 30 observations as the training data, the remaining cities are used as the test data. Table 3.1 shows the sample mean and the values of RMSE (RMSE) of the estimated parameters β_0 and β_1, based on one hundred replications.

Parameters($\boldsymbol{\beta}$)	β_0	β_1
True values	1	2
Sample mean	0.99475	2.00275
RMSE	0.00035	0.14478

Table 3.1: Sample mean and RMSE for estimated parameters based on one hundred replications

To evaluate the performance of the model, we use root-mean-square error (RMSE). This measurement calculates the differences between estimated parameters and the true values using equation (2.9). Now, we recall the formula of RMSE from the equation (2.9) for measuring performance of prediction as follows.

$$RMSE_j = \sqrt{\frac{\sum_{i=1}^{r}(y_{ij} - \hat{y_{ij}})^2}{r}},$$

where $\hat{y}$ is predicted value, y is actually observed values and r is the number of replications. Table 3.1 shows that the value of sample mean of estimated parameters are close to the true values and the RMSE values are close to zero. It means that the Laplace approximation works well. The estimates of the parameters are very close to their true values.

We are also interested in predicting the response at a new point. The estimated parameters can be used to predict the performance of test data. The average values of the root mean squared error (average RMSE) between $\hat{\boldsymbol{\mu}}$ and $\boldsymbol{\mu}$ for test data calculated based on one hundred replications is 0.2620. Here, the average true values of $\boldsymbol{\mu}$ is 18.19. Thus, It shows that the predictive performance is acceptable.

3.4.2 Simulation study: Scenario 2

In this scenario, we will present numerical results of a simulation study comparing several models. Here, the three models considered are SGLMM $\mathcal{GPR}$, the existing approach SGLMM ICAR and the conventional Poisson regression model which assumes the data are independent.

In the simulation study, we first generate random data from a Poisson distribution (3.12) with true values $\beta_0 = 1$ and $\beta_1 = 2$, using the structure similar to that used for the thirty six cities specified in the Dengue fever data example. The spatial covariates $\mathbf{x}$ are longitude and latitude of each city. We select 30 of them as training data and the remaining as test data. To measure performance, we calculate the sample mean of the estimated parameters and the root mean squared error (RMSE) between the estimated

and the true values of the parameters, and also calculate the average of RMSE between the predicted values and the actual observed data.

We suppose that $\mathbf{z}$ are observed data and the model can be specified as

$$\begin{aligned} z_i &\sim Poisson(\mu_i) \\ \log(\mu_i) &= U_i^T \boldsymbol{\beta} + \tau_i \end{aligned}$$

where the vector of random effects τ_i is assumed to have a general conditional autoregressive structure. In the intrinsic conditional autoregressive (ICAR) method, the conditional distribution of the random effect is given by :

$$\tau_i | \tau_j \sim \mathcal{N}(\sum_{j \neg i} b_{ij} \tau_j, \sigma_i^2), i = 1, ..., n$$

where $b_{ij} = \frac{w_{ij}}{w_{i+}}$ and $\sigma_i^2 = \frac{\sigma^2}{w_{i+}}$ and w_{ij} is a measure of the association between two observations (equal to 1 if the observations are neighbours and 0 otherwise) and w_{i+} is the sum of all w_{ij} for observation i.

In SGLMM-$\mathcal{GPR}$ model, we assume that τ_i follows a Gaussian process regression model as defined in equations (3.1) and (3.2). Here, the covariance kernel is squared exponential which $\mathbf{x}$ are the location of each area measured by its longitude and latitude.

Table 3.2 shows the sample mean from different methods and also the value of RMSE (RMSE) based on fifty replications. All the methods show good performance in terms

	Sample Mean		RMSE	
Method	β_0	β_1	β_0	β_1
True value	1	2		
Poisson	0.9991	1.9984	0.1504	0.2295
SGLMM ICAR	0.9953	2.0013	0.1503	0.2281
SGLMM $\mathcal{GPR}$	0.9991	1.9984	0.1504	0.2296

Table 3.2: Sample mean of the estimated parameters and RMSE for several models based on fifty replications

of estimating the parameters. The value of average RMSE (average RMSE) between μ and $\hat{\mu}$ can be seen in Table 3.3. It shows clearly that the performance of SGLMM $\mathcal{GPR}$ is the best. SGLMM ICAR show improvement compared with the conventional Poisson regression which assumes the data are independent.

3.4.3 Dengue Fever Data

An increasingly important topic in epidemiology is dengue fever. This disease is a mosquito-borne infection found in tropical and sub-tropical regions around the world. In recent

Method	Average RMSE
Poisson	0.4579
SGLMM ICAR	0.1388
SGLMM $\mathcal{GPR}$	0.0239

Table 3.3: The average of RMSE between $\boldsymbol{\mu}$ and $\hat{\boldsymbol{\mu}}$ using three different models based on fifty replications

years, transmission has increased predominantly in urban and semi-urban areas and has become a major international public health concern.

Firstly, we begin by exploring the spreading of dengue fever in Indonesia which can be found in Figure 3.1. It is clear that the disease displays a geographical pattern and is highly spatially correlated. Its prevalence in Indonesia is the highest in the world and had around 50 thousand new cases in 2011. It means that for every 1000 people living in an affected area there are two people who become infected. This is a big issue and needs to be analysed to find out the risk and identify preventable factors in order to reverse the trend. Moreover, research will give a better understanding of how the community can achieve good quality of life. Therefore, this example will analyse dengue fever to determine the risk and significant factors of the diseases in East Java, Indonesia. The conventional and two Poisson regression models, SGLMM ICAR and SGLMM $\mathcal{GPR}$, are used and compared.

The data are taken from Health Office East Java Indonesia in 2010. In this example, the response variable (z_i) is the number of dengue fever cases in 2010 at city i and the covariates include the percentage of households that can access healthy water (x_1), healthy waste disposal (x_2), waste water disposal facilities (x_3), clean and healthy behaviour (x_4) and healthy housing (x_5). Table 3.4 shows summaries of the Dengue Fever data in East Java, Indonesia i 2010.

Variable	Min	Mean	Max
Healthy water	9.15	71.57	401.10
Healthy waste disposal	0.00	55.51	100.00
Waste water disposal facilities	1.64	51.79	97.37
Clean and healthy behaviour	7.00	37.36	79.86
Healthy housing	13.84	65.39	99.69
Dengue fever cases	18.00		3379.0

Table 3.4: Summaries of Dengue Fever data

Tables 3.5 to 3.7 show the estimated parameters using three different models, i.e. conventional Poisson regression, SGLMM $\mathcal{GPR}$ and SGLMM ICAR. Using Poisson regression model, each covariate has a significant effect to this disease. One of the reasons is because standard error (SE) is calculated by assuming they are independent and then SE is extremely small, resulting in vary small p-value. In conclusion, p-value might be

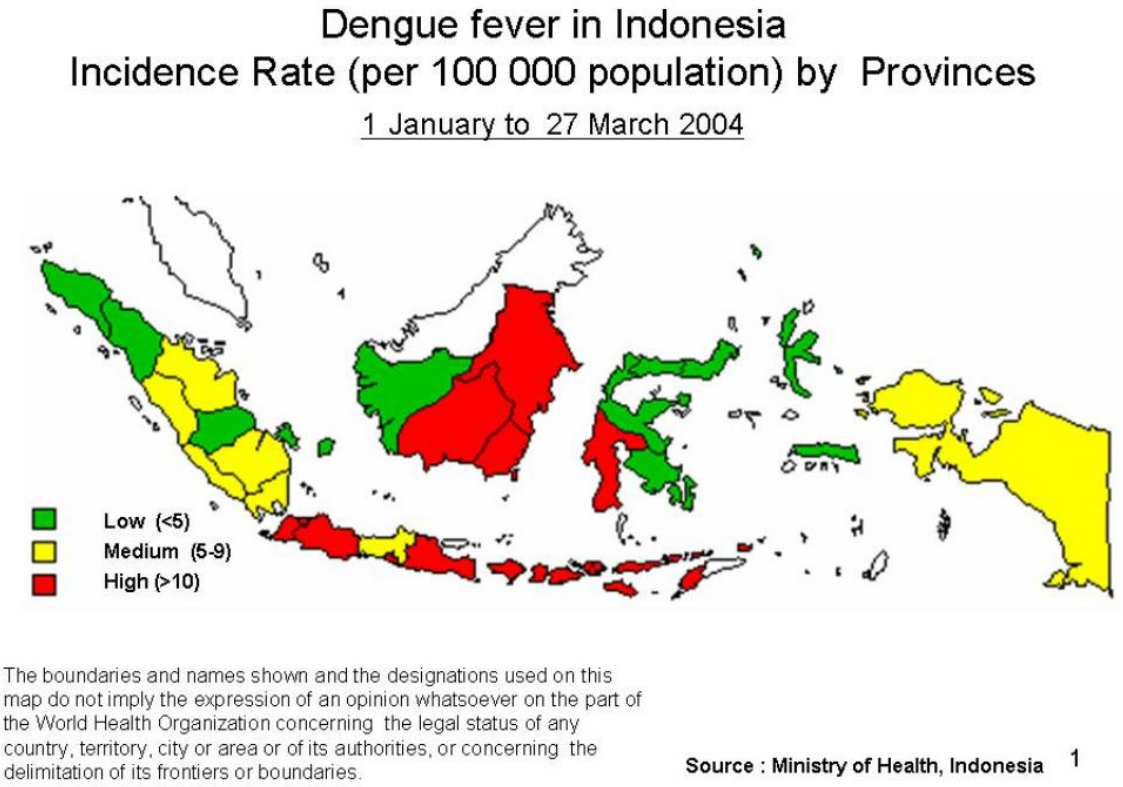

Figure 3.1: Incidence rate of dengue fever in Indonesia by provinces from 1 January to 27 March 2004 (Source: World Health Organization, 2004).

	Poisson regression		
Coefficients	Estimate	SE	Pvalue
Intercept	5.6258	0.0296	0.00
Healthywater	-0.003	0.0001	0.00
Trashcan	-0.001	0.0002	0.00
Wastewaterdis	0.0513	0.0003	0.00
Cleanhealthybehav	0.0014	0.0004	0.00
Healthyhouse	0.0057	0.0005	0.00

Table 3.5: Estimated parameters (coefficient of covariates) using conventional Poisson regression model

	SGLMM ICAR		
Coefficients	Estimate	SE	Pvalue
Intercept	5.6015	0.011	0.00
Healthywater	-0.005	0.001	0.70
Trashcan	-0.003	0.004	0.87
Wastewaterdis	0.0053	0.005	0.44
Cleanhealthybehav	0.0071	0.006	0.29
Healthyhouse	0.0112	0.006	0.04

Table 3.6: Estimated parameters (coefficient of covariates) using spatial GLMM with ICAR

	SGLMM $\mathcal{GPR}$		
Coefficients	$\hat{\boldsymbol{\beta}}$	SE	P-value
Intercept	5.6005	0.012	0.0000
Healthywater	-0.005	0.005	0.0000
Trashcan	-0.007	0.003	0.0274
Wastewaterdis	0.0035	0.003	0.9578
Cleanhealthybehav	0.0077	0.008	0.3060
Healthyhouse	0.0052	0.002	0.0565

Table 3.7: Estimated parameters (coefficient of covariates) using spatial GLMM with $\mathcal{GPR}$

questionable.

However, a more realistic model is to assume dependence between neighbouring areas. The disease is caused by a virus and can be spread quickly to the neighbouring areas. The geographic position is a very significant factor regarding transmission of the disease. From Table 3.6 and Table 3.7, it can been seen that the estimates of the parameters using SGLMM ICAR and SGLMM $\mathcal{GPR}$ are quite similar to the ones obtained by the conventional model. But the P-values are quite different to the previous model. However we cannot give an over-interpretation of the p-value. The non-significant part may be caused by correlation among covariates.

To compare the predictive performance of the models, we randomly select 30 cities as

training data and the remaining 6 cities as test data. After obtaining the estimation of the parameters, we calculate the prediction for the test data.

Method	RE
Poisson regression	0.0544
SGLMM ICAR	0.0238
SGLMM GPR	0.0143

Table 3.8: The relative error (RE) for different methods

Table 3.8 gives the values of relative error (RE), which is defined as the percentage of the difference between the prediction and the observation with respect to the observation. From Table 3.8, we see that SGLMM GPR is the best among these models.

3.5 Chapter Summary

In this chapter, we proposed a Gaussian process regression model for the covariance structure. It allows the use of geographic position and other variables to define the spatial correlation structure, and thus it provides a very flexible model. The simulation study and the real data example shows its good performance. The computation of the method is quite efficient.

We discussed the model for a univariate response variable in this chapter. In practice, it might be better to consider multivariate response variables together by borrowing information from each other. For example, other diseases such as malaria are also spread by mosquitoes. Therefore, it is better if we analyse both dengue fever and malaria at the same time. The main problem here is how we construct a cross-correlation between response variables. We will discuss this problem in the next chapters.

Chapter 4

A Convolved Gaussian Process for Multiple Dependent processes

4.1 Introduction

In the previous chapter, we discussed non Gaussian regression with a Gaussian process as our prior and with particular reference to spatial Poisson regression. This model provides sufficient flexibility to define the spatial correlation structure. The benefit of using a Gaussian process is that it enables us to extend the idea to address multiple dependent Gaussian processes. During the last decade, many researchers have been developing this method for multiple processes. The vast majority of approaches have attempted to handle the difficulty of capturing inter-dependence between two processes and to ensure that the covariance matrix is positive definite.

Now, we can estimate the parameters from observed data to model multiple dependent processes instead of specifying and controlling the parameters of the positive covariance structure. An alternative way is to use a convolution method. In this chapter we will focus on constructing multiple dependent Gaussian processes, i.e. the definition of the cross-correlation structure. We extend the method proposed by Andriluka *et al.* (2006) and consider a broader range of problems by comparing models with different covariance functions. Therefore, we are able to test sensitivity when choosing different covariance structures.

We first investigate briefly in Section 4.2 the convolved Gaussian process for a single process and then look at the problem of constructing a convolved Gaussian process for multiple dependent processes using different covariance functions in Section 4.3. The last section will explore how we analyse multivariate nonlinear data with convolved Gaussian process as priors. Both Sections 4.3 and 4.4 will provide technical details and comprehensive simulation studies.

4.2 Convolved Gaussian Process for a Single Process

This section recalls the definition of a Gaussian process which has been discussed in Chapter 3 and extends it to a convolved Gaussian process. A Gaussian process can be defined as follows. Let $\mathbf{f} = (f(\mathbf{x}_1), ..., f(\mathbf{x}_N))$, then we can define $f(\cdot)$ as a Gaussian process ($\mathcal{GP}$) with mean function $\mu(\cdot)$ and a covariance function $k(\cdot, \cdot)$ and will write the Gaussian process as

$$f(.) \sim \mathcal{G}P(\mu(\cdot), k(\cdot, \cdot)),$$

if

$$\mathbf{f} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{K}),$$

where the i-th element of $\boldsymbol{\mu}$ is $\mu(\mathbf{x}_i)$, the (i, j)-th element of $\mathbf{K}$ is $k(\mathbf{x}_i, \mathbf{x}_j)$ and $k(\cdot, \cdot)$ is a covariance function. It is common to set the mean function to be zero for simplicity.

Now we define a Gaussian process in a different way, namely a convolved Gaussian process. Let $\tau(\mathbf{x})$ be Gaussian white noise $\tau(\mathbf{x}) \sim^{iid} \mathcal{N}(0, \sigma^2)$ and $h(\mathbf{x})$ be a smoothing kernel for $\mathbf{x} \in \mathcal{R}^P$. We can construct a convolved Gaussian process $\eta(\mathbf{x})$ as

$$\begin{aligned} \eta(\mathbf{x}) &= h(\mathbf{x}) \star \tau(\mathbf{x}) \\ &= \int h(\mathbf{x} - \boldsymbol{\alpha})\tau(\boldsymbol{\alpha})d\alpha = \int h(\boldsymbol{\alpha})\tau(\mathbf{x} - \boldsymbol{\alpha})d\alpha, \end{aligned} \tag{4.1}$$

where $\star$ denotes convolution. If we suppose a smooth kernel $h(\mathbf{x})$ is given by

$$h(\mathbf{x}) = v \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{A}(\mathbf{x} - \boldsymbol{\mu})\right\}$$

then the $\eta(\mathbf{x})$ defined in (4.1) is a convolved Gaussian process ($\mathcal{CGP}$). It is equivalent to a GP with zero mean and the following covariance function

$$k(\mathbf{x}_i, \mathbf{x}_j) = \pi^{\frac{P}{2}} v^2 |\mathbf{A}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{4}(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{A}(\mathbf{x}_i - \mathbf{x}_j)\right\}, \tag{4.2}$$

where v and $\mathbf{A}$ are parameters. We denote a convolved Gaussian process by

$$\eta(\mathbf{x}) \sim \mathcal{CGP}(h(\mathbf{x}), \tau(\mathbf{x})). \tag{4.3}$$

Examining the exponential and its quadratic form of covariance structure in equation (4.2), we can see that $(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{A}(\mathbf{x}_i - \mathbf{x}_j)$ can be replaced by a squared Mahalanobis distance (m^2). This relationship will be used to construct a convolved Gaussian process from a Gaussian process. For more explanation we will provide specific examples in the

next section.

4.3 A Convolved Gaussian Process for Multiple Dependent Processes

The convolved Gaussian process defined in the previous section can be extended to define multiple dependent processes. We first define three independent Gaussian white noises, namely $\tau_0(\mathbf{x}), \tau_1(\mathbf{x})$ and $\tau_2(\mathbf{x})$. We start the discussion by constructing four $\mathcal{CGP}$s as follows :

$$\begin{aligned} \eta_1(\mathbf{x}) &\sim \mathcal{CGP}(g_1(\mathbf{x}), \tau_1(\mathbf{x})) \\ \eta_2(\mathbf{x}) &\sim \mathcal{CGP}(g_2(\mathbf{x}), \tau_2(\mathbf{x})) \end{aligned} \tag{4.4}$$

where $g_1(\mathbf{x})$ and $g_2(\mathbf{x})$ are two smoothing kernels. Here, $\eta_1(\mathbf{x})$ and $\eta_1(\mathbf{x})$ are two independent $\mathcal{CGP}$s since they are constructed using independent Gaussian white noises $\tau_1(\mathbf{x})$ and $\tau_2(\mathbf{x})$. Another two $\mathcal{CGP}$s can be specified as

$$\begin{aligned} \xi_1(\mathbf{x}) &\sim \mathcal{CGP}(h_1(\mathbf{x}), \tau_0(\mathbf{x})) \\ \xi_2(\mathbf{x}) &\sim \mathcal{CGP}(h_2(\mathbf{x}), \tau_0(\mathbf{x})) \end{aligned} \tag{4.5}$$

where different kernels $h_1(\mathbf{x})$ and $h_2(\mathbf{x})$ are used to define the different covariance structures and the same white noise $\tau_0(\mathbf{x})$ implies the dependency between $\xi_1(\mathbf{x})$ and $\xi_2(\mathbf{x})$. It is clear that $\xi_1(\mathbf{x})$ and $\xi_2(\mathbf{x})$ are dependent but are independent from $\eta_1(\mathbf{x})$ and $\eta_2(\mathbf{x})$. Therefore, from four $\mathcal{CGP}$s we can define bivariate dependent Gaussian processes as

$$f_a(\mathbf{x}) = \xi_a(\mathbf{x}) + \eta_a(\mathbf{x}), \quad a = 1, 2. \tag{4.6}$$

Based on equation (4.6), the dependency between $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ is modelled by $\xi_1(\mathbf{x})$ and $\xi_2(\mathbf{x})$, while the individual characteristics are modelled by $\eta_1(\mathbf{x})$ and $\eta_2(\mathbf{x})$.

Let us assume

$$\mathbf{f} = (f_1(\mathbf{x}_{1i}), i = 1, ..., N_1; f_2(\mathbf{x}_{2i}), i = 1, ..., N_2)$$

where $\mathbf{x}_i = (\mathbf{x}_{1i}, \mathbf{x}_{2i}) \in T \subset \mathcal{R}^P$. Then we can define $f(\cdot)$ as a dependent Gaussian processes with zero means and a covariance function $\mathcal{K}(\cdot, \cdot)$ and can be written as follows :

$$f(.) \sim \mathcal{MGP}(0, \mathcal{K}(\cdot, \cdot)), \tag{4.7}$$

where

$$\begin{aligned} \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j) &= \begin{pmatrix} Cov(f_1(\mathbf{x}_i), f_1(\mathbf{x}_j)) & Cov(f_1(\mathbf{x}_i), f_2(\mathbf{x}_j)) \\ Cov(f_2(\mathbf{x}_i), f_1(\mathbf{x}_j)) & Cov(f_2(\mathbf{x}_i), f_2(\mathbf{x}_j)) \end{pmatrix} \\ &=^d \begin{pmatrix} k_{11}(\mathbf{x}_i, \mathbf{x}_j) & k_{12}(\mathbf{x}_i, \mathbf{x}_j) \\ k_{21}(\mathbf{x}_i, \mathbf{x}_j) & k_{22}(\mathbf{x}_i, \mathbf{x}_j) \end{pmatrix}. \end{aligned} \tag{4.8}$$

The dependent Gaussian processes in equation (4.6) can also be defined as follows :

$$\mathbf{f} = (f_1(\mathbf{x}), f_2(\mathbf{x}))^T \sim \mathcal{N}(\mathbf{0}, \mathcal{K}_{N_1 N_2}). \tag{4.9}$$

$\mathbf{f}$ has multivariate normal distribution with zero mean and covariance matrix $\mathcal{K}_{N_1 N_2}$ with $(N_1 + N_2) \times (N_1 + N_2)$ dimensions where the (i, j)-th element of $\mathcal{K}_{N_1 N_2}$ is $\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j)$ and $\mathcal{K}(\cdot, \cdot)$ is a covariance function.

Standard Gaussian process models use a stationary covariance, in which the covariance between any two points is a function of Euclidean distance, Paciorek (2003). Due to the stationarity of kernel functions, we can define a separation vector $\mathbf{d} = \mathbf{x}_i - \mathbf{x}_j$. In other words, the covariance matrix for the model can be specified based on equation (4.8) as follows

$$\begin{aligned} k_{11}(\mathbf{d}) &= k_{11}^{\xi_1}(\mathbf{d}) + k_{11}^{\eta_1}(\mathbf{d}); \qquad k_{12}(\mathbf{d}) = k_{12}^{\xi_{12}}(\mathbf{d}); \\ k_{22}(\mathbf{d}) &= k_{22}^{\xi_2}(\mathbf{d}) + k_{22}^{\eta_2}(\mathbf{d}); \qquad k_{21}(\mathbf{d}) = k_{12}^{\xi_{12}}(-\mathbf{d}). \end{aligned} \tag{4.10}$$

Now we provide a general framework to derive the set of $k_{ab}(\mathbf{d})$ from any stationary covariance kernel. Similar to the previous closed forms in equation (4.10), if $a = b$ it is defined as autocovariance, but if $a \neq b$, it is called cross-covariance between output a and b. We can express it in a closed form by applying the proposition as follows.

Proposition 4.3.1. Assume that $\mathcal{S}(m)$ is an isotropic covariance function on $\mathbb{R}^P$, for any $P \in \mathbb{N}$. Then the covariance $k_{ab}(\mathbf{d})$ in (4.10) is given by

$$k_{ab}(\mathbf{d}) = \frac{v_a v_b (2\pi)^{P/2}}{|A_a + A_b|^{1/2}} \mathcal{S}(\sqrt{Q_{ab}(\mathbf{d})})$$

where

$$Q_{ab}(\mathbf{d}) = \mathbf{d}^T A_a (A_a + A_b)^{-1} A_b \mathbf{d}$$

$v_a, v_b \in \mathbb{R}$ and arbitrary positive matrices $A_a, a = 1, 2$ and $k_{ab}(\mathbf{d})$ is a positive definite function on $\mathbb{R}^p, p = 1, 2, ...$ for $a, b = 1, 2$.

The proof of the proposition uses similar arguments to the one used in Andriluka *et al.*

(2006) and Paciorek (2003) although Proposition 4.3.1 only focuses on two outputs. The details of the proof can be found in Appendix B.

From Proposition 4.3.1, we are able to obtain closed forms of the kernel function for the model and provide flexibility regarding the choice of the stationary covariance structure. We can apply the proposition by taking any isotropic covariance function to build multiple dependent Gaussian processes. There are several stationary kernels which can be extended by Proposition 4.3.1, such as gamma exponential, exponential, Matern and rational quadratics as we will discuss in this chapter. By using Proposition 4.3.1, the closed forms of four kernels in equation (4.8) or (4.10) can be expressed as

$$\begin{aligned} k_{aa}^{\xi_a}(\mathbf{d}) &= \frac{v_a^2(\pi)^{P/2}}{|A_a|^{\frac{1}{2}}}\mathcal{S}(\sqrt{Q_{aa}(\mathbf{d})}) \\ k_{ab}^{\xi_{ab}}(\mathbf{d}) &= \frac{v_a v_b(2\pi)^{P/2}}{|A_a + A_b|^{1/2}}\mathcal{S}(\sqrt{Q_{ab}(\mathbf{d})}) \\ k_{aa}^{\eta_a}(\mathbf{d}) &= \frac{w_a^2(\pi)^{P/2}}{|B_a|^{\frac{1}{2}}}\mathcal{S}(\sqrt{Q_{aa}(\mathbf{d})} \quad ; \quad a, b = 1, 2 \quad \textit{and} \quad a \neq b. \end{aligned} \tag{4.11}$$

Specific examples and implementation of Proposition 4.3.1 will be provided in the remainder of this section.

4.3.1 Example of $\mathcal{CGP}$s for Multiple Dependent Gaussian Processes

The convolution approach provides considerable flexibility in terms of generating the $\mathcal{CGP}$. However, this approach needs an integrable kernel function, and so we focus here on stationary covariance functions. Furthermore, the flexibility in selecting covariance functions as convolved kernels will also be discussed.

i. Squared Exponential

The first example of Proposition 4.3.1 is applying a squared exponential covariance function as a stationary convolved kernel. The squared exponential has the following formula

$$k(\mathbf{d}) = v \exp \frac{-\|\mathbf{d}\|^2}{2l^2}, \tag{4.12}$$

where the parameter l controls the vertical scale of the process. Now, we explain the relationship between equation (4.12) and Proposition 4.3.1. From the previous explanation, we can replace m^2 (the squared Mahalanobis distance) with Q_{ab}. As we know, if the covariance matrix is the identity matrix, the Mahalanobis distance reduces to the Euclidean distance. So, we can replace $\frac{\mathbf{d}}{l}$ in equation 4.12 with $\sqrt{Q_{ab}}$, Paciorek (2003).

Our aim is to develop multiple dependent Gaussian processes which are constructed from four convolved Gaussian processes with squared exponentials as the covariance structure. Based on Proposition 4.3.1, the covariance matrix structure for convolved Gaussian processes using a square exponential covariance structure can be defined as

$$k_{ab}(\mathbf{d}) = \frac{v_a v_b (2\pi)^{P/2}}{|A_a + A_b|^{\frac{1}{2}}} e^{-\frac{1}{2}Q_{ab}(\mathbf{d})}. \tag{4.13}$$

where $a, b = 1, 2$, $Q_{ab}(\mathbf{d}) = \mathbf{d}^T A_a (A_a + A_b)^{-1} A_b \mathbf{d}$ and $k_{ab}(\mathbf{x})$ is a positive definite function. We highlight that the closed forms are also provided by Boyle and Frean, (2005).

Thus we are able to develop bivariate dependent Gaussian processes which follow a multivariate normal distribution with zero mean and closed forms of four covariance functions in equation (4.10) with equation (4.13). The details of the closed form kernels are as follows

$$\begin{aligned} k_{aa}^{\xi_a}(\mathbf{d}) &= \frac{v_a^2 (\pi)^{P/2}}{|A_a|^{\frac{1}{2}}} e^{-\frac{1}{2}Q_{aa}(\mathbf{d})} \\ k_{ab}^{\xi_{ab}}(\mathbf{d}) &= \frac{v_a v_b (2\pi)^{P/2}}{|A_a + A_b|^{\frac{1}{2}}} e^{-\frac{1}{2}Q_{ab}(\mathbf{d})} \\ k_{aa}^{\eta_a}(\mathbf{d}) &= \frac{w_a^2 (\pi)^{P/2}}{|B_a|^{\frac{1}{2}}} e^{-\frac{1}{2}Q_{aa}(\mathbf{d})} \quad ; a, b = 1, 2 \quad \textit{and} \quad a \neq b. \end{aligned} \tag{4.14}$$

In this model, we consider estimating parameters by using a simple method, such as maximum likelihood estimation. Now, we recall equation (4.9). Let us consider $\mathbf{f} = (f_1, ..., f_{N_1}, f_{N_1+1}, ..., f_{N_1+N_2})^T$. Then we have

$$\mathbf{f} \sim \mathcal{N}(\mathbf{0}, \mathcal{K}_{N_1 N_2}); \qquad \mathcal{K}_{N_1 N_2} = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}, \tag{4.15}$$

where for $a, b = 1, 2$

$$C_{ab} = \begin{pmatrix} k_{ab}(\mathbf{x}_{a1}, \mathbf{x}_{b1}) & \cdots & k_{ab}(\mathbf{x}_{a1}, \mathbf{x}_{bN_b}) \\ \vdots & \vdots & \vdots \\ k_{ab}(\mathbf{x}_{aN_a}, \mathbf{x}_{b1}) & \cdots & k_{ab}(\mathbf{x}_{aN_a}, \mathbf{x}_{bN_b}) \end{pmatrix}.$$

Based on the proof in Appendix B, k_{ab} is non-negative. From the covariance functions in equation (4.11), we have defined that $\boldsymbol{\Theta} = (v_a, A_a, w_a, B_a)$ for $i = 1, 2$ as parameters. Inference for this model is actually an extension from the single $\mathcal{GP}$ which is explained in the third chapter. Hence, we can calculate the log likelihood function for

parameters $\boldsymbol{\Theta}$ as follows :

$$L(\boldsymbol{\Theta}|\mathcal{D}) = -\frac{1}{2}\log \mathcal{K}_{N_1N_2}(\boldsymbol{\Theta})| - \frac{1}{2}\mathbf{f}^T\mathcal{K}_{N_1N_2}(\boldsymbol{\Theta})^{-1}\mathbf{f} - \frac{N_1+N_2}{2}\log 2\pi. \tag{4.16}$$

Therefore, we estimate parameters by maximising the log likelihood function in equation (4.16).

Regarding the first investigation of the multiple processes, we explore its performance by examining the sample outputs. We generate the latent process by using the model in equation (4.6) with closed form kernel functions as in equation (4.14). Three different curves, each containing 30 observations, are taken to be samples of multiple output processes. Figure 4.3 shows three samples of two process components f_1 and

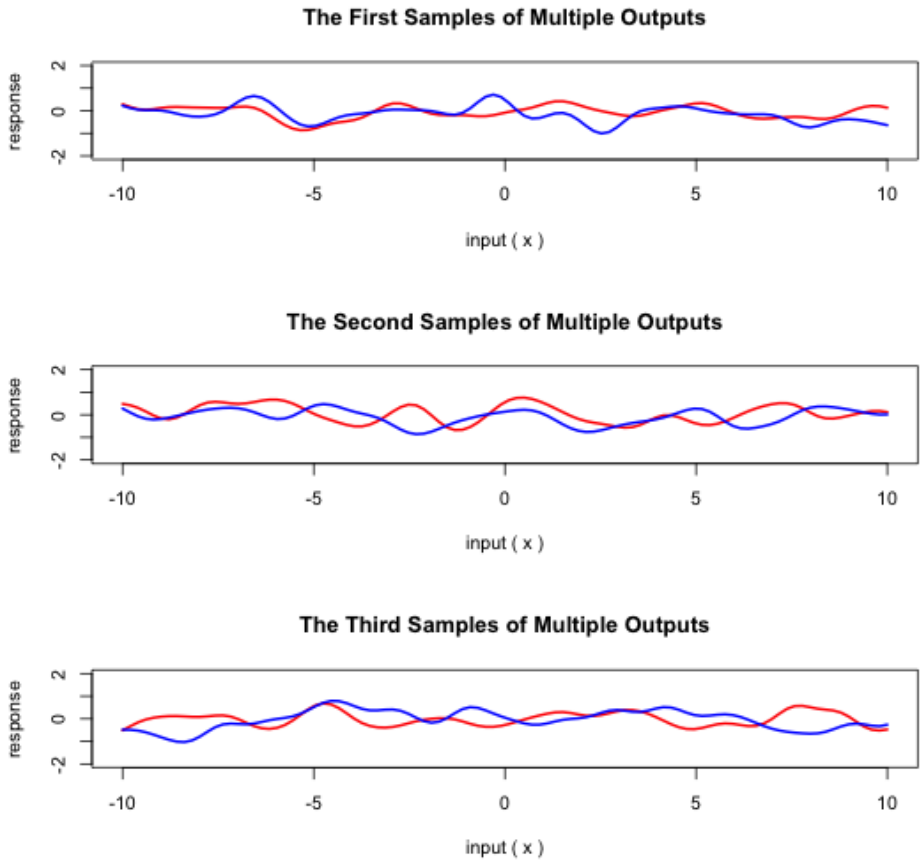

Figure 4.1: Multiple dependent output process samples generated from squared exponential covariance functions as convolved kernels with parameters $\boldsymbol{\Theta} = (v_1, v_2, A_1, A_2, w_1, w_2, B_1, B_2)$ are $0.2, 0.2, 1, 1, 0.2, 0.2, 1, 1$ respectively, and $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$ as equally spaced points in $[-5, 5]$. Red curves are defined as f_1 and blue curves are f_2.

f_2. Here, f_1 and f_2 are dependent in all the samples although each of them has its own characteristics. Moreover, we have tried to explore further these curves regarding

dependency issues. In Figure 4.2, we provide a correlation map between two processes. It shows that the correlation between points in each output is high, i.e. almost 1. Furthermore, there is also large correlation (cross correlation) between $\mathbf{f}_1$ and $\mathbf{f}_2$. The samples also show that the performance of each response offers slightly different figures and that the processes are dependent on each other. Choosing a suitable covariance function is allowed for multiple output processes. Thus, this is the nature of the data source which can be used for a priori knowledge to select it. Further investigation will be discussed in the next section.

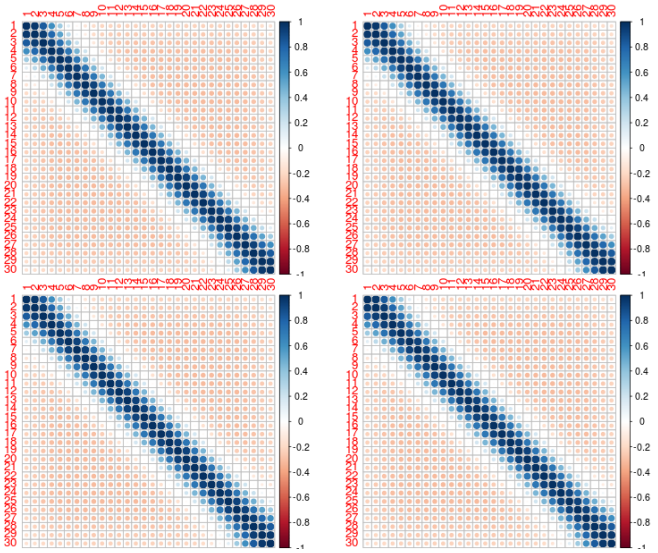

Figure 4.2: Correlation map of multiple dependent Gaussian processes with squared exponential convolved kernels; top left: correlation map of $\mathbf{f}_1$ at 30 equally spaced points; top right: correlation map between $\mathbf{f}_1$ and $\mathbf{f}_2$; and bottom right: correlation map of $\mathbf{f}_2$

A maximum likelihood estimator is used to estimate parameters by maximizing the likelihood function in (4.16). Table 4.1 lists the average estimates of parameters, $\boldsymbol{\Theta}$ (sample mean) using a maximum likelihood estimator for one hundred replications. It has been mentioned previously that $\boldsymbol{\Theta} = (v_1, v_2, A_1, A_2, w_1, w_2, B_1, B_2)$. The maximum likelihood estimates seem acceptable when compared to the true values. Also, it can be seen that the values of root mean square (RMSE) between the estimated parameters (A_1, A_2, B_1, B_2) and true values close to zero. It means that the differences between estimated parameters (A_1, A_2, B_1, B_2) and true values are small. Meanwhile, for estimated parameters v_1, v_2, w_1 and w_2 are quite far from the true values. It might because we do not add any noise variable in this modelling. In conclusion, this issue needs to investigate more.

ii. Gamma Exponential Family

Another stationary covariance function can be constructed using the Gamma expo-

Parameters	True value	Sample Mean	RMSE
v_1	0.04	0.00171	0.04190
v_2	0.04	0.00177	0.04209
A_1	1	1.000182	0.00144
A_2	1	1.000156	0.00124
w_1	0.04	0.000921	0.04013
w_2	0.04	0.000980	0.04022
B_1	1	1.000766	0.00077
B_2	1	1.000723	0.00083

Table 4.1: Sample mean and RMSE estimated parameters from multiple dependent Gaussian processes with square exponential convolved kernels

nential family covariance structure.

By applying Proposition 4.3.1, the closed form of the covariance structure of the model can be defined as

$$k_{ab}(\mathbf{d}) = \frac{\nu_a \nu_b (2\pi)^{P/2}}{|A_a + A_b|^{\frac{1}{2}}} e^{(-\sqrt{Q_{ab}(\mathbf{d})})^{\gamma}}, \quad 0 < \gamma \leqslant 2.$$

where $a, b = 1, 2$ and $Q_{ab}(\mathbf{d}) = (\mathbf{d})^T A_a (A_a + A_b)^{-1} A_b(\mathbf{d})$. We are also able to construct bivariate dependent processes with Gamma exponential covariance functions which follow multivariate normal distribution with zero means and covariance structures given in equation (4.8). The covariance can be calculated from equation (4.10) and the closed forms of kernel functions can be defined as

$$\begin{aligned} k_{aa}^{\xi_a}(\mathbf{d}) &= \frac{v_a^2 (\pi)^{p/2}}{|A_a|^{\frac{1}{2}}} e^{(-\sqrt{Q_{aa}(\mathbf{d})})^{\gamma}}; \quad a, b = 1, 2 \quad \textit{and} \quad a \neq b \\ k_{ab}^{\xi_{ab}}(\mathbf{x}) &= \frac{v_a v_b (2\pi)^{p/2}}{|A_a + A_b|^{\frac{1}{2}}} e^{(-\sqrt{Q_{ab}(\mathbf{d})})^{\gamma}} \\ k_{aa}^{\eta_a}(\mathbf{x}) &= \frac{w_a^2 (\pi)^{p/2}}{|B_a|^{\frac{1}{2}}} e^{(-\sqrt{Q_{aa}(\mathbf{d})})^{\gamma}}. \end{aligned} \tag{4.17}$$

iii. Rational Quadratic

The model can also be generated by using a rational quadratic covariance function. According to proposition 4.3.1, the closed forms of the covariance structure for the model with rational quadratic covariance structure can be defined as

$$k_{ab}(\mathbf{d}) = \frac{v_a v_b (2\pi)^{P/2}}{|A_a + A_b|^{\frac{1}{2}}} \left(\frac{1}{1 + \frac{1}{2\alpha} Q_{ab}(\mathbf{d})} \right)^{\alpha}, \quad \alpha > 0. \tag{4.18}$$

iv. Matern

Furthermore, the model can also use Matern covariance functions as convolved kernel and the general closed forms of four kernels can be specified as

$$\begin{aligned} k_{ab}(\mathbf{d}) &= \frac{v_a v_b (2\pi)^{P/2}}{|A_a + A_b|^{\frac{1}{2}}} \frac{1}{\Gamma(\nu) 2^{\nu-1}} \left(\sqrt{2\nu Q_{ab}(\mathbf{d})}\right)^{\nu} \\ & K_{\nu}\left(\sqrt{2\nu Q_{ab}(\mathbf{d})}\right), \quad \nu > 0, \end{aligned} \tag{4.19}$$

if $\nu = \frac{3}{2}$, so the closed forms of kernel functions from equation (4.10) can be defined as

$$k_{ab}^{\xi_{ab}}(\mathbf{d}) = \frac{v_a v_b (2\pi)^{P/2}}{|A_a + A_b|^{\frac{1}{2}}} \left(1 + \sqrt{3Q_{ab}(\mathbf{d})}\right) \exp\left(\sqrt{-3Q_{ab}(\mathbf{d})}\right) \tag{4.20}$$

where $a, b = 1, 2$ and $Q_{ab}(\mathbf{d}) = (\mathbf{d})^T A_a (A_a + A_j)^{-1} A_b(\mathbf{d})$.

v. Mixed Covariance functions as a Convolved Kernel

Different covariance functions can be used to model specific characteristics for each component in convolved Gaussian processes. The idea here is to set up dependent Gaussian processes based on equation (4.6) by using mixed kernels. We now use an example to illustrate this idea. The shared convolved Gaussian processes ($\boldsymbol{\xi}$) have squared exponential covariance functions. The independent convolved Gaussian process η_1 is constructed by using a squared exponential kernel and η_2 by using a gamma exponential with ($\gamma = 0.5$). The closed forms for four kernels are :

$$\begin{aligned} k_{aa}^{\xi_a}(\mathbf{d}) &= \frac{v_a^2 (\pi)^{p/2}}{|A_a|^{\frac{1}{2}}} e^{-\frac{1}{2}Q_{aa}(\mathbf{d})} \qquad k_{11}^{\eta_1}(\mathbf{d}) = \frac{w_1^2 (\pi)^{p/2}}{|B_1|^{\frac{1}{2}}} e^{-\frac{1}{2}Q_{11}(\mathbf{d})} \\ k_{ab}^{\xi_{ab}}(\mathbf{d}) &= \frac{v_a v_b (2\pi)^{p/2}}{|A_a + A_b|^{\frac{1}{2}}} e^{-\frac{1}{2}Q_{ab}(\mathbf{d})} \qquad k_{22}^{\eta_2}(\mathbf{d}) = \frac{w_2^2 (\pi)^{p/2}}{|B_2|^{\frac{1}{2}}} e^{(-\sqrt{Q_{22}(\mathbf{d})})^{\gamma}} \\ & a, b = 1, 2 \quad \textit{and} \quad a \neq b. \end{aligned} \tag{4.21}$$

The characteristics of the two dependent processes can be seen from Figure 4.3. From the figure, it can be seen that there are significantly different performances between two process components but they are clearly dependent. This is also shown by the correlation map in Figure 4.4. The panels on the top left and bottom right clearly show the difference in the covariance structures between $\mathbf{f}_1$ and $\mathbf{f}_2$.

Table 4.2 shows the result of estimated parameters and the RMSE based on one hundred replications. We still use two sample sizes of 30. Similar to the Table 4.1, some of estimated parameter shows good performance although the data look very different for the two components. This shows the flexibility of using the convolved $\mathcal{GP}$

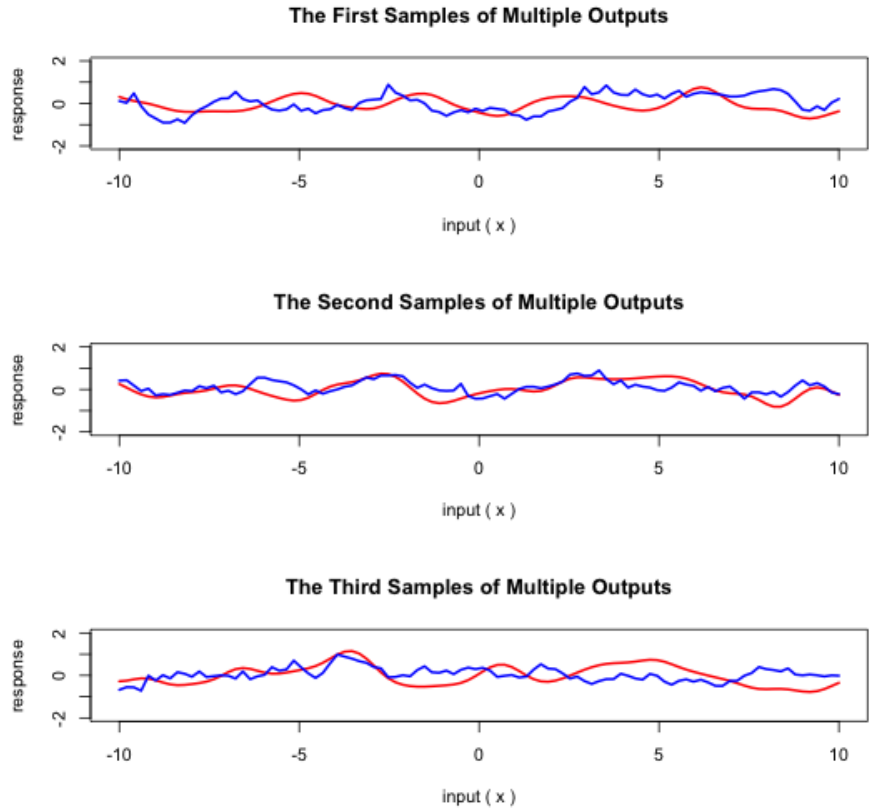

Figure 4.3: Multiple dependent output process samples generated from squared exponential covariance functions and a Gamma exponential covariance function ($\gamma = 1$) as convolved kernels. Red curves are defined as f_1 and blue curves are f_2.

approach. Meanwhile for v_1, v_2, w_1 and w_2 are fairly far form the true values. Again, we need to have more work to investigate, such as adding "jitter" (noise). In the next section, we will discuss nonlinear model which is added some noises. Inference and simulation examples will also provide in the next section.

4.4 Convolved Gaussian Process Priors for Multivariate Nonlinear Regression Analysis

Multivariate nonlinear regression is developing rapidly. The approach using Gaussian process priors is becoming more and more popular due to its flexibility, particularly in the field of machine learning and geostatistics. In the multivariate case, we have showed in the previous sections that $\mathcal{CGP}$ is a flexible and good way to model the dependency and individual characteristics of multiple processes. In this section we will briefly explain the model of multivariate nonlinear regression using convolved Gaussian process ($\mathcal{CGP}$)

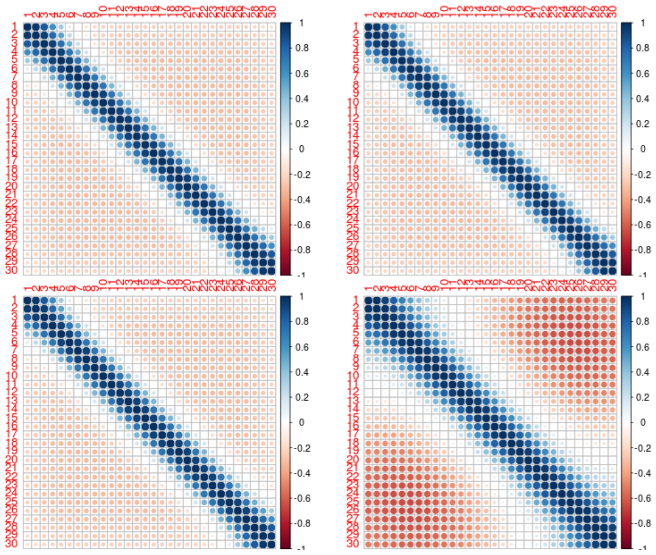

Figure 4.4: Correlation map of multiple dependent Gaussian processes using squared exponentials and a Gamma exponential as convolved kernels between f_1 and f_2; top left: correlation map of $\mathbf{f}_1$ at 30 equally spaced points; top right: correlation map between $\mathbf{f}_1$ and $\mathbf{f}_2$; and bottom right: correlation map of $\mathbf{f}_2$

Parameters	True value	sample Mean	RMSE
v_1	0.04	0.00029	0.03981
v_2	0.04	0.00049	0.03981
A_1	1	0.99972	0.00031
A_2	1	0.99974	0.00027
w_1	0.04	0.00026	0.03997
w_2	0.04	-0.00026	0.04003
B_1	1	1.00041	0.00042
B_2	1	1.00037	0.00037

Table 4.2: The estimated parameters and the RMSE from multiple Gaussian process with square exponential convolution and gamma exponential ($\gamma = 0.5$) as mixed convolved kernel based on 100 replications.

priors.

The general model can be written as

$$\begin{aligned} \mathbf{y}_1 &= f_1(\mathbf{x}) + \epsilon_1 \\ \mathbf{y}_2 &= f_2(\mathbf{x}) + \epsilon_2, \end{aligned} \tag{4.22}$$

where $(\mathbf{y}_1, \mathbf{y}_2)$ are the two responses, $\mathbf{x}$ is a P-dimensional covariate and two independent error items $\epsilon_1 \sim \mathcal{N}(0, \sigma_1^2)$ and $\epsilon_2 \sim \mathcal{N}(0, \sigma_2^2)$. Here, $f_1(\mathbf{x})$ and $f_2(\mathbf{x})$ are two unknown nonlinear regression function. We will use $\mathcal{CGP}$'s defined in (4.6) to (4.10) as priors of f_1

and f_2. Similar to the univariate case, the performance of Bayesian inference for the above model often depends on the choice of the hyper-parameters. Thus, rather than assuming the probability structure of hyper-parameters, we can use observed data to estimate them by using an empirical Bayesian approach.

4.4.1 Empirical Bayesian Estimates

Let us consider that we have observed the following set of data

$$\mathcal{D} = \left\{ \begin{pmatrix} y_{1i} \\ x_{1i} \end{pmatrix}; i = 1, ..., N_1, \begin{pmatrix} y_{2i} \\ x_{2i} \end{pmatrix}; i = 1, ..., N_2 \right\}.$$

It comprises N_1 and N_2 observations, each consisting of a $\mathcal{P}$-dimensional input vector $\mathbf{x}_{1i}$ and $\mathbf{x}_{2i}$ and scalar output y_{1i} and y_{2i} respectively. A discrete multivariate nonlinear with $\mathcal{CGP}$ priors model is given by

$$\begin{aligned} y_{1i} &= f_1(\mathbf{x}_{1i}) + \epsilon_{1i}, \quad y_{2i} = f_2(\mathbf{x}_{2i}) + \epsilon_{2i}, \\ \epsilon_{1i} &\sim \mathcal{N}(0, \sigma_1^2); \qquad \epsilon_{2i} \sim \mathcal{N}(0, \sigma_2^2) \end{aligned} \tag{4.23}$$

where $(f_1(\cdot), f_2(\cdot))^T$ follows a multivariate normal prior with zero mean and a covariance function which is explained in equations (4.8) and (4.10). A dependent Gaussian process regression with $\mathcal{CGP}$ priors is an extension of the single output model which has been explained in Chapter 3. As a result, inference for Gaussian process regression can be applied to this model by replacing the number of observations and the size of covariance structure.

Now, we can rewrite $\mathbf{y} = (y_1, ..., y_{N_1}, y_{N_1+1}, ..., y_{N_1+N_2})^T$, then we have

$$\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Psi}); \qquad \mathbf{\Psi} = \begin{pmatrix} C_{11} + \sigma_1^2 I_{N_1} & C_{12} \\ C_{21} & C_{22} + \sigma_2^2 I_{N_2} \end{pmatrix}. \tag{4.24}$$

The log likelihood function with hyper-parameters $\boldsymbol{\theta} = (v_1, v_2, A_1, A_2, w_1, w_2, B_1, B_2)$ is

$$L(\boldsymbol{\theta}|\mathcal{D}) = -\frac{1}{2} \log |\mathbf{\Psi}(\boldsymbol{\theta})| - \frac{1}{2} \mathbf{y}^T \mathbf{\Psi}(\boldsymbol{\theta})^{-1} \mathbf{y} - \frac{N_1 + N_2}{2} \log 2\pi. \tag{4.25}$$

We will estimate the hyper-parameters by maximising the likelihood function in (4.25).

4.4.2 Predictions

Regarding the predictions, we can refer to and extend the inference from Gaussian regression around equation (3.10) which is explained in the third chapter. Suppose that we want to calculate the predictive distribution of $\mathbf{y}^* = (y_1^*, y_2^*)^T$ at a new points $\mathbf{x}^*$. It is also a

bivariate Gaussian distribution with

$$\hat{\mathbf{y}}^* = \mathbf{\Psi}_{\mathbf{y}}^{*T}\mathbf{\Psi}^{-1}\mathbf{y} \tag{4.26}$$

$$\hat{\mathbf{\Psi}}_{y^*} = \mathbf{\Psi}^* - \mathbf{\Psi}_{\mathbf{y}}^{*T}\mathbf{\Psi}^{-1}\mathbf{\Psi}^*. \tag{4.27}$$

where $\mathbf{\Psi}_{\mathbf{y}}^{*T}$ is a $2 \times (N_1 + N_2)$ covariance matrix between $\mathbf{y}^*$ and $\mathbf{y}$ which can be expressed as

$$\begin{pmatrix} k_{11}(\mathbf{x}^*, \mathbf{x}_{11})...k_{11}(\mathbf{x}^*, \mathbf{x}_{1N_1}) & k_{12}(\mathbf{x}^*, \mathbf{x}_{21})...k_{12}(\mathbf{x}^*, \mathbf{x}_{2N_2}) \\ k_{21}(\mathbf{x}_{11}, \mathbf{x}^*)...k_{21}(\mathbf{x}_{1N_1}, \mathbf{x}^*) & k_{22}(\mathbf{x}^*, \mathbf{x}_{21})...k_{22}(\mathbf{x}^*, \mathbf{x}_{2N_2}) \end{pmatrix},$$

$\mathbf{\Psi}^*$ is a 2×2 covariance matrix of $\mathbf{y}^*$

$$\mathbf{\Psi}^* = \begin{pmatrix} k_{11}(\mathbf{x}^*, \mathbf{x}^*) + \sigma_1^2 I_{N_1} & k_{12}(\mathbf{x}^*, \mathbf{x}^*) \\ k_{21}(\mathbf{x}^*, \mathbf{x}^*) & k_{22}(\mathbf{x}^*, \mathbf{x}^*) + \sigma_2^2 I_{N_2} \end{pmatrix},$$

and $\mathbf{\Psi}$ is defined in (4.24).

4.4.3 Numerical Examples

In this subsection, two different examples are considered. We first generate a set of data as training and test data, $\mathbf{y}$ given the latent variable $\mathbf{x}$, by assuming that $\mathbf{y}$ and $\mathbf{x}$ are nonlinearly related. We estimate hyper-parameters using training data with an empirical Bayesian approach and then calculate the prediction mean and variance. As a measure of goodness of fit, the values of root mean squared error (RMSE) between predictions and their true values are used.

Scenario 1

The aim of the first scenario is to test the sensitivity of the proposed model with different covariances functions, considering mixed squared exponential and gamma exponential covariance functions. The true model used to generate the process is

$$\begin{pmatrix} \mathbf{y}_{1i} \\ \mathbf{y}_{2i} \end{pmatrix} = \begin{pmatrix} Sin(6\mathbf{x}_i) \\ Cos(6\mathbf{x}_i) \end{pmatrix} + \begin{pmatrix} \tau_{1i}(\mathbf{x}_i) \\ \tau_{2i}(\mathbf{x}_i) \end{pmatrix} + \begin{pmatrix} \epsilon_{1i} \\ \epsilon_{2i} \end{pmatrix}$$

where $\tau(\cdot) \sim \mathcal{MGP}(0, \mathcal{K}(\cdot,\cdot))$ and $\mathcal{K}(\cdot,\cdot)$ is defined by equations (4.8) and (4.10). Here $\mathbf{x}$ are equally spaced points in $[0, 1]$. We consider a mixed squared exponential and gamma exponential covariance functions which have been discussed in the previous section. The covariance structure is the same as the example discussed in part(v) in Section 4.3.1, i.e η_1 is generated from a squared exponential covariance function with the true values of $(w_1 = 0.04, B_1 = 1)$; η_2 follows from a Gamma exponential covariance functions with true values of the hyper-parameters $(w_2 = 0.04, B_2 = 1)$. The shared processes, $\boldsymbol{\xi}$ have squared

exponential covariance functions with true values of $(v_1 = 0.04, A_1 = 1, v_2 = 0.04, A_2 = 1)$. The true values of hyper-parameters $\boldsymbol{\theta} = (v_1, v_2, A_1, A_2, w_1, w_2, B_1, B_2, \sigma_1, \sigma_2)$ are 0.04, 0.04, 1, 1, 0.04, 0.04, 1, 1, 0.1 and 0.1 respectively. Two component processes, each containing 30 data points, are generated and used as training data. We generate also the same number of data points for test data.

We use the nonlinear regression model (4.22) and the convolved Gaussian process priors as discussed in Sections 4.4.1 and 4.4.2. To test the sensitivity of the choice of covariance functions, several models are considered and compared.

1. *Model 1*
   We assume that it has the same covariance structure as the true model, i.e. ξ_1, ξ_1, η_1 have a squared exponential covariance function while η_2 has a Gamma exponential covariance function.

2. *Model 2*
   Similar to *Model 1*, but η_2 has a rational quadratic covariance function with the value of α is 0.5

3. *Model 3*
   Similar to *Model 1*, but η_2 has a Matern covariance function with $\nu = \frac{3}{2}$

4. *Model 4*
   Similar to *Model 1*, but η_2 has a squared exponential covariance function

5. *Model 5*
   We also compare with exist model proposed by Crainiceanu *et al.* (2012), namely the $\mathcal{CD}$ model.

Now, we define $\mathcal{CD}$ model. Suppose that $\mathbf{y}_1$ and $\mathbf{y}_2$ are two dependent processes that can be written as follows

$$\mathbf{y}_2 \mid \mathbf{y}_1 \quad \sim \quad \mathcal{N}(\alpha \mathbf{y}_1, \sigma_\epsilon^2) \tag{4.28}$$

where $\mathbf{y}_1$ is a Gaussian process with zero mean and any stationary covariance functions. In this setting, we define the model in (4.28) with squared exponential covariance function in $\mathbf{y}_1$. Figures 4.5 and 4.6 show the prediction mean curves with the different models.

The figures show that the prediction mean curves given by all the proposed models are quite close to the true curves. It also can be seen from the figures that *Model 2*, *Model 3* and *Model 4* also perform reasonably well, even though misspecified covariance functions are used. This is possibly because the empirical Bayesian approach can select the best member from each covariance function family respectively for η_1, η_2 and the shared processes $\boldsymbol{\xi}$. Although it cannot beat the true model, the flexibility can still guarantee

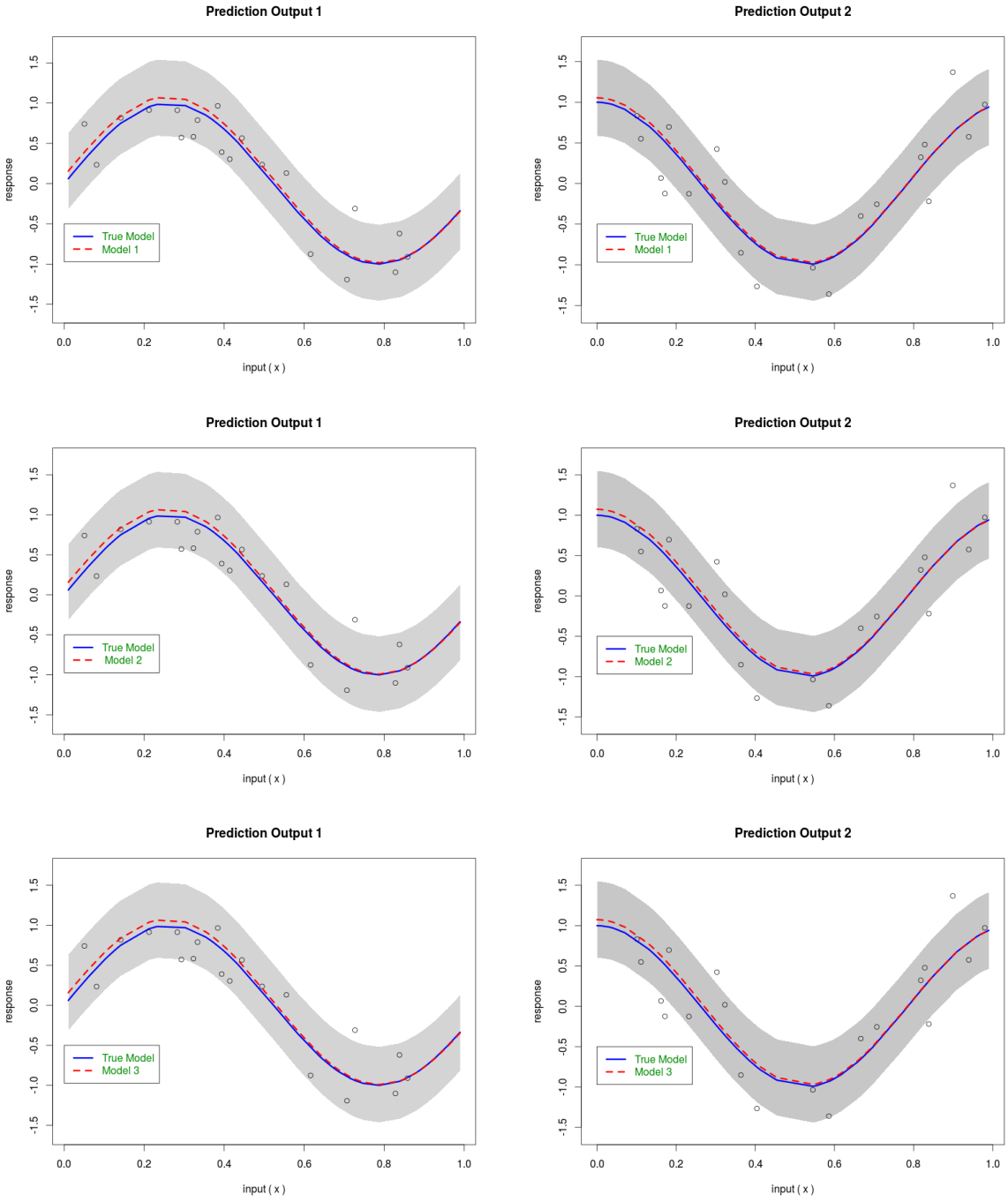

Figure 4.5: The predictions from two strongly dependent outputs. The dots denote test data (sample), the red dashed lines represent the predictions by three different covariance functions (*Model 1*, *Model 2* and *Model 3*) and the blue solid lines are the true curves with 95% confidence intervals (the shaded regions)

a reasonably good solution. In contrast, the dependency across the component processes in *Model 5* is determined by a covariance structure. It fails to capture the individual

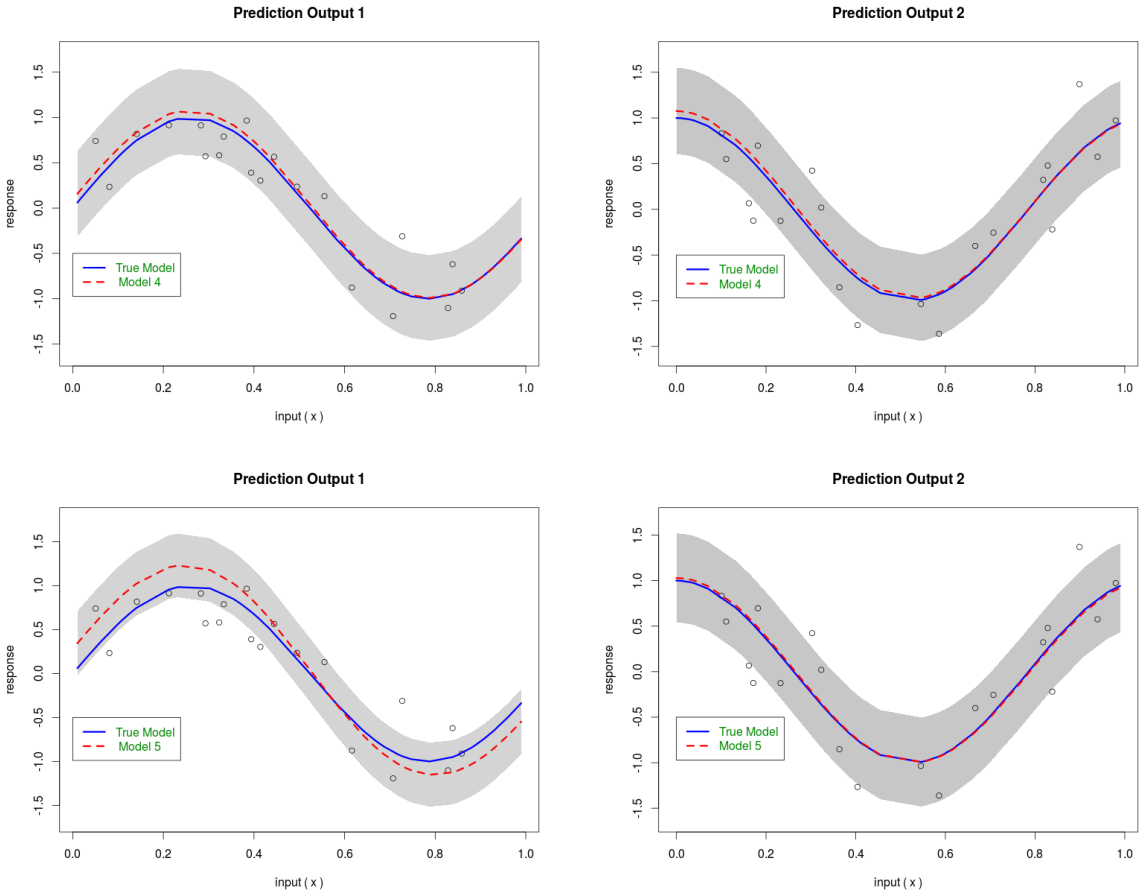

Figure 4.6: The predictions from two strongly dependent outputs. The dots denote test data (sample), the red dashed lines represent the predictions by two different covariance functions (*Model 4* and *Model 5*) and the blue solid lines are the true curves with 95% confidence intervals (the shaded regions)

characteristics.

The above findings are confirmed by a simulation study. The average values of the root of mean squared errors (RMSE) between $y_a(\mathbf{x})$ at test data points and $\hat{y}_a(\mathbf{x})$ are calculated based on one hundred repetitions for different models and are reported in Table 4.3.

Type of Models	Average RMSE
Model 1	0.01145
Model 2	0.01220
Model 3	0.01185
Model 4	0.01222
Model 5	0.16470

Table 4.3: Average of RMSE prediction between y and $\hat{y}$ from various models for one hundred replications.

Scenario 2

In the second scenario, we provide another setting to test the performance of the proposed model for data with a more general covariance structure.

The true model for generating random process as follows :

$$\begin{aligned} y_{ai}(\mathbf{x}_i) &= sin(6\mathbf{x}_i) + \tau_{ai}(\mathbf{x}_i) + \epsilon_{ai}(\mathbf{x}_i) \\ \tau_{ai}(\cdot) &\sim \mathcal{MGP}(0, \mathcal{K}(\cdot,\cdot)); \quad \epsilon_a(\mathbf{x}) \sim \mathcal{N}(0, \sigma_a^2), \quad a = 1, 2 \end{aligned}$$

where $\mathcal{K}(\cdot,\cdot)$ is defined by equations (4.8) and (4.10). We use a Matern class with $\nu = \frac{3}{2}$ as covariance functions for $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$. Then two output processes, each containing 30 equally spaced points in $[0, 1]$ are simulated from the same initial values of hyper-parameters $\boldsymbol{\theta}$ in previous scenario, i.e. $(v_1 = 0.04, v_2 = 0.04, A_1 = 1, A_2 = 1, w_1 = 0.04, w_2 = 0.04, B_1 = 1, B_2 = 1)$.

In this scenario, we compare the performance of different models, i.e. *Model 6* which is similar to the true model that $\boldsymbol{\eta}$ and $\boldsymbol{\xi}$ have Matern class with $\nu = \frac{3}{2}$ covariance function, *Model 4* and *Model 5*. The procedure for constructing the last two models are the same as in the previous scenario. In *Model 4*, we define that η_1, η_2, ξ_1 and ξ_2 have a squared exponential covariance function. For *Model 5* has the same covariance structure as $\mathcal{CD}$ model.

The prediction mean functions for different models are presented in Figure 4.7. We use Bayesian information criterion (BIC) to select models and use the value of RMSE between the actual values of $y_a(\mathbf{x})$ and the prediction mean function $\hat{y}_a(\mathbf{x})$ for test data to compare the models. The results are also compared to the $\mathcal{CD}$ approach in *Model 5*.

Type of Models	BIC	RMSE
Model 6	30.668	0.001938
Model 4	165.984	0.004572
Model 5	194.166	0.005425

Table 4.4: The value of BIC from Three Different Models for one replication.

The results for one replication are presented in Figure 4.7 and Table 4.4. The results show that the prediction mean functions for *Model 6*, with different covariance functions, are similar and all close to the true mean function and that the performance for prediction of the individual processes are comparable to each other. Also we obtain the smallest value of BIC among all of the models. From Table 4.5, we see that *Model 6* provides the best model performance as expected. Although *Model 4* is not the best choice, the performance is only slightly worse than the best model. This shows the flexibility and robustness of the proposed model by using $\mathcal{CGP}$ priors. As with the first scenario, *Model*

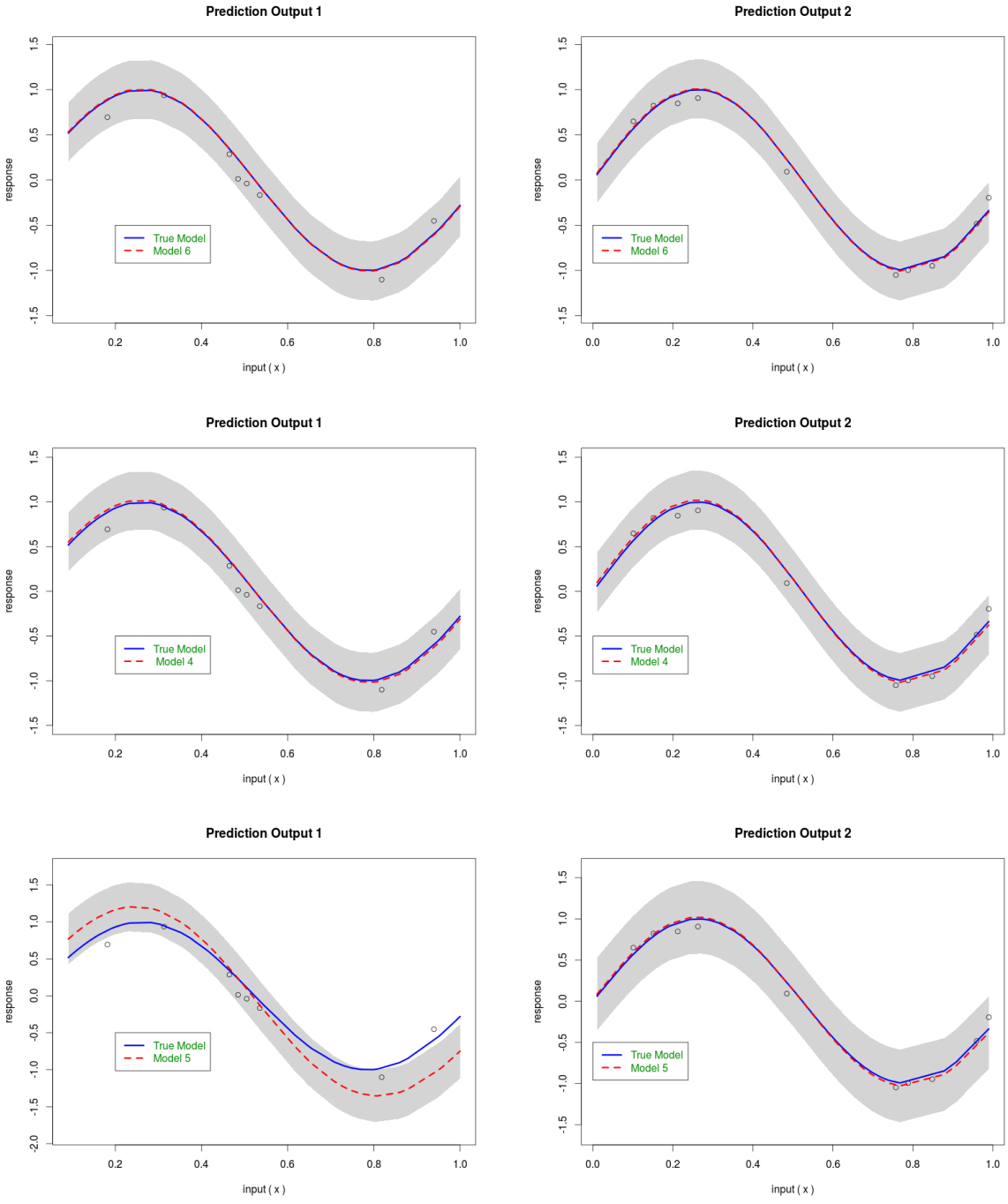

Figure 4.7: The predictions from two strongly dependent outputs. The dots are test data (sample), the red dashed lines represent the predictions by three different covariance functions (*Model 6*, *Model 4* and *Model 5*) and the blue solid lines are the true curves with 95% confidence intervals (the shaded regions)

5 fails to provide a good fit.

Table 4.5 shows the summary statistics based on a simulation study with 100 replica-

tions, where Prop-BIC is the proportion achieving the smallest value of BIC and Average RMSE is the average of RMSE. Table 4.5 has shown that the proposed model with differ-

Type of Models	Prop-BIC	Average RMSE
Model 6	0.75	0.01382
Model 4	0.25	0.01411
Model 5	0	0.28946

Table 4.5: Average RMSE of prediction between y and $\hat{y}$ and the Proportion of the Smallest Values of BIC (Prop-BIC) from three different models for one hundered replications.

ent covariance gives comparable results in term of average RMSE values. Similar to the previous results, the value of the average RMSE for all models is quite close to the true mean function although the true model (*Model 6*) is the best one as expected. Regarding the proportion of the smallest values of BIC, *Model 6*, as the true model, has shown good performance as well. Meanwhile, *Model 6* would be selected in most of the cases although *Model 4* would also be selected occasionally. The performance of *Model 5* is much worse than the other two.

4.5 Chapter Summary

In this chapter, we proposed the extension of convolved Gaussian processes for multivariate nonlinear regression analysis by investigating many stationary covariance functions and their mixed forms as convolved kernels. We considered covariance functions such as squared exponential, gamma exponential, rational quadratic and Matern. Furthermore, we also explored a way to apply mixed covariance functions for constructing multiple dependent Gaussian processes and used them as priors in a multivariate nonlinear model. During our investigation, we were able to identify several advantages using a convolved Gaussian process for multiple dependent processes. One of the most significant advantages is that the proposed model provided huge flexibility regarding the choice of covariance functions. From the first scenario, it has been shown that the model is very robust because it still provides a reasonably good result even if a misspecified covariance function is used. Furthermore, this proposed method is also able to perfectly capture the main features of each processes. Strong evidence has been delivered both in Scenario 1 and 2. In all scenarios, we have compared several models with the $\mathcal{CD}$ approach in Crainiceanu *et al.* (2012). As a result, the extension model with convolved Gaussian process priors provides a way to tackle individual characteristics of each response component in multivariate nonlinear regression. But their model fails to do so. However, there are some limitations to the model. It seems not every covariance function has a closed from in terms of $k_{ab}(\mathbf{d})$. Thus in this chapter we mainly focussed on stationary covariance functions. We will extend the

model to non-Gaussian multivariate models in the next two chapters.

Chapter 5

Convolved Gaussian Process Priors for Multivariate Poisson Regression Analysis

5.1 Introduction

The previous chapter sets out the framework of Convolved $\mathcal{GP}$s for multiple dependent processes. We now extend the model to multivariate non-Gaussian data, such as classification and count data. In this chapter, we will focus on bivariate Poisson regression which is still an active topic in the recent statistics literature.

The vast majority of approaches assume that observed data are independent. In practice, however, many bivariate count data sets contain dependent observations, especially in medical data which have a spatial correlation related to area or region. In recent years, many researchers have explored this widely, such as the conditional dependency ($\mathcal{CD}$) model proposed by Crainiceanu *et al.* (2012). Unfortunately, it fails to tackle the individual characteristics of each response component as we explained in Chapter 4. The Convolved Gaussian process ($\mathcal{CGP}$) model has performed well and offers flexibility in choosing covariance structures. It enables us to propose a novel approach to multivariate Poisson regression analysis. The main advantage of the model is that it can offer flexibility in handling cross-correlations between two responses and capture the important features of each response component. At the same time, the model is also able to ensure that the spatial correlation structure is modelled properly and the covariance matrix for the combined responses is positive definite.

Based on Section 4.4 in Chapter 4, we emphasize that our proposed model for multivariate nonlinear regression is very flexible and robust. We extend the idea to use $\mathcal{CGP}$ priors for multivariate Poisson regression, namely multivariate convolved Gaussian pro-

cess Poisson regression analysis ($\mathcal{MCGPPR}$). This model has some features worth noting, i.e. it offers a semiparametric regression model for multivariate Poisson data. It means that the model is able to combine the regression relationship between multivariate Poisson responses and multidimensional covariates which consist of both linear and nonlinear models. The prior specification of the covariance kernel also enables us to accommodate a nonlinear model involving large dimensional covariates.

This chapter is organized as follows. We will explain the details of the $\mathcal{MCGPPR}$ model in Section 5.2. The details of inference including estimation and prediction will be provided in Sections 5.3 and 5.4. In Section 5.5, we will explain an asymptotic theory of information consistency. Comprehensive simulation studies and real data applications will be discussed in the last section.

5.2 The Model

In this section we will explain the multivariate Poisson regression model using convolved Gaussian processes($\mathcal{CGP}$) as priors. The general model can be written as

$$\begin{aligned} \mathbf{z_a} \mid \boldsymbol{\tau}_a &\sim Poisson(\boldsymbol{\mu}_a) \\ \log(\boldsymbol{\mu}_a) &= \mathbf{U}_a^T \boldsymbol{\beta}_a + \boldsymbol{\tau}_a, \quad a = 1, 2. \end{aligned} \tag{5.1}$$

Here $\mathbf{z_a}, a = 1, 2$ stands for two correlated response variables, for example the number of dengue fever and number of malaria cases in our Dengue fever and Malaria data. As we discussed in Chapter 3, the observations are spatially correlated. We used a Gaussian process prior to define such correlation, but we analyse the dengue fever data and malaria data separately. In practice, those diseases are certainly correlated. It would, therefore, be better to analyse those two data sets together by defining a cross-correlation among the different observations for the two response variables. Crainiceanu *et al.* (2012) proposed a conditional dependency approach.

As we have pointed out in Chapter 4, $\mathcal{CGP}$ is a more flexible model in terms of defining cross-correlation for multivariate dependent data. We extend the idea to the above model. The cross-correlation of $\mathbf{z}_a$ is modelled via a dependent latent variable $\boldsymbol{\tau}_a, a = 1, 2$. We assume a $\mathcal{CGP}$ prior for $\boldsymbol{\tau}_a$ depending on $\mathbf{x}$ and an unknown hyper-parameter $\boldsymbol{\theta}$. Specifically we define a nonparametric $\mathcal{CGP}$ by equations (4.6) to (4.10). In this case the regression relationship between the bivariate Poisson regression $\mathbf{z}_a$ and the covariates $\mathbf{x}_a$ is modelled by the covariance structure of $\boldsymbol{\tau}(\mathbf{x})$.

The $\mathbf{z}_a$'s follow a Poisson distribution with $(\boldsymbol{\mu}_a)$ where $a = 1, 2$. If we use the log link

function, the density function is given by

$$p(\mathbf{z}_a \mid \boldsymbol{\tau}_a) = \frac{e^{-(\mathbf{U}_a^T\boldsymbol{\beta}_a+\boldsymbol{\tau}_a)}(\mathbf{U}_a^T\boldsymbol{\beta}_a+\boldsymbol{\tau}_a)^{\mathbf{z}_a}}{\mathbf{z}_a!}. \tag{5.2}$$

The marginal density function of $\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2)$ is therefore given by

$$p(\mathbf{z}) = \int p(\mathbf{z} \mid \boldsymbol{\tau}, \boldsymbol{\beta}) p(\boldsymbol{\tau} \mid \boldsymbol{\theta}) d\boldsymbol{\tau},$$

where $p(\boldsymbol{\tau} \mid \boldsymbol{\theta})$ is the density function of $\boldsymbol{\tau} = (\boldsymbol{\tau}_1, \boldsymbol{\tau}_2)$. Similar to the multivariate non-linear regression case, the performance of Bayesian inference for the above model often depends on the choice of the hyper-parameters $\boldsymbol{\theta}$. Thus, in order to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, we also use an empirical Bayesian analogue of the approach used in the previous chapter.

5.3 Empirical Bayesian Estimates

Suppose that we have observed the following data,

$$\mathcal{D} = \left\{ \begin{pmatrix} z_{1i} \\ U_{1i} \\ \mathbf{x}_{1i} \end{pmatrix}; i = 1, ..., N_1, \begin{pmatrix} z_{2i} \\ U_{2i} \\ \mathbf{x}_{2i} \end{pmatrix}; i = 1, ..., N_2 \right\} \tag{5.3}$$

where z_{1i}, z_{2i} are observations of the two variable responses, U_{1i}, U_{2i} are bivariate covariates and $\mathbf{x}_i = (\mathbf{x}_{1i}, \mathbf{x}_{2i})$ are covariates modelled covariance structure τ_{i1}, τ_{2i}. N_1 and N_2 are the number of the first and the second responses respectively. A proposed discrete multivariate Poisson regression model with $\mathcal{CGP}$ priors ($\mathcal{MCGPPR}$) is therefore given by

$$\begin{pmatrix} z_{1i}(\mathbf{x}_{1i}) \\ z_{2i}(\mathbf{x}_{2i}) \end{pmatrix} \sim \begin{pmatrix} Poisson(\mu_{1i}(\mathbf{x}_{1i})), & i = 1, ..., N_1 \\ Poisson(\mu_{2i}(\mathbf{x}_{2i})), & i = 1, ..., N_2 \end{pmatrix} \tag{5.4}$$

where

$$\begin{pmatrix} \mu_{1i}(\mathbf{x}_{1i}) = \exp(\mathbf{U}_{1i}^T\boldsymbol{\beta}_1 + \tau_{1i}(\mathbf{x}_{1i})) \\ \mu_{2i}(\mathbf{x}_{2i}) = \exp(\mathbf{U}_{2i}^T\boldsymbol{\beta}_2 + \tau_{2i}(\mathbf{x}_{2i})) \end{pmatrix} \quad and \quad \begin{pmatrix} \tau_{1i}(\cdot) \\ \tau_{2i}(\cdot) \end{pmatrix} \sim \mathcal{MGP}(\mathbf{0}, \mathcal{K}(\cdot,\cdot)),$$

where $\mathcal{K}(\cdot,\cdot))$ is defined by equation (4.8) and (4.10). We can say also that $(\tau_{1i}(\cdot), \tau_{2i}(\cdot))^T$ follow a multivariate normal prior with zero mean and covariance function based on equation (4.11).

The idea of empirical Bayesian learning is to choose the value of the hyper-parameter $\boldsymbol{\theta}$ by maximizing the marginal density function. Thus, $\boldsymbol{\theta}$ as well as the unknown parameter $\boldsymbol{\beta}$ can be estimated at the same time by maximizing the following marginal density of

$\mathbf{z} = (z_{11}, ..., z_{1N_1}, z_{21}, ..., z_{2N_2})^T$ given $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, i.e.

$$\begin{aligned} p(\mathbf{z}|\boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{x}) &= \int p(\mathbf{z} \mid \boldsymbol{\tau}, \boldsymbol{\beta}) p(\boldsymbol{\tau}|\boldsymbol{\theta}) d\boldsymbol{\tau} \\ &= \int \prod_{a=1}^{2} p(\mathbf{z}_a \mid \boldsymbol{\tau}, \boldsymbol{\beta}) p(\boldsymbol{\tau}|\boldsymbol{\theta}) d\boldsymbol{\tau} \\ &= \int \left\{ \prod_{a=1}^{2} \prod_{i=1}^{N_a} p(z_{a_i} \mid \tau_{ai}, \boldsymbol{\beta}) \right\} p(\boldsymbol{\tau}|\boldsymbol{\theta}) d\boldsymbol{\tau} \end{aligned}$$

or the marginal log-likelihood

$$\begin{aligned} l(\boldsymbol{\beta}, \boldsymbol{\theta}) &= \log \{ p(\mathbf{z} \mid \boldsymbol{\beta}, \boldsymbol{\theta}, \mathbf{x}) \} \\ &= \log \int \prod_{a=1}^{2} \prod_{i=1}^{N_a} p(z_{a_i} \mid \tau_{ai}, \boldsymbol{\beta}) (2\pi)^{\frac{-\sum_{a=1}^{2} N_a}{2}} \mid \mathcal{K}_{N_1 N_2} \mid^{\frac{-1}{2}} \exp \left\{ \boldsymbol{\tau}^T \mathcal{K}_{N_1 N_2}^{-1} \boldsymbol{\tau} \right\} d\boldsymbol{\tau} \end{aligned} \tag{5.5}$$

where $p(z_{a_i} \mid \tau_{ai}, \boldsymbol{\beta})$ is derived from the Poisson distribution. Obviously the integral involved in the above marginal density is analytically intractable unless $p(z_{a_i} \mid \tau_{ai}, \boldsymbol{\beta})$ has a special form such as the density function of the normal distribution. One method to address this problem is to use a Laplace approximation. We denote

$$\Phi(\boldsymbol{\tau}) = \sum_{a=1}^{2} \sum_{i=1}^{N_a} \{ \log p(z_{a_i} \mid \tau_{ai}, \boldsymbol{\beta}) \} - \frac{1}{2} \log |\mathcal{K}_{N_1 N_2}| - \frac{1}{2} \boldsymbol{\tau}^T \mathcal{K}_{N_1 N_2}^{-1} \boldsymbol{\tau} - \frac{N_1 + N_2}{2} \log 2\pi, \tag{5.6}$$

where $\sum_{a=1}^{2} \sum_{i=1}^{N_a} \log p(z_{a_i} \mid \tau_{ai}, \boldsymbol{\beta}) = \sum_{i=1}^{N_1} (z_{1i} \log(\mu_{1i}) - \mu_{1i}) + \sum_{i=1}^{N_2} (z_{2i} \log(\mu_{2i}) - \mu_{2i})$ with $\mu_{1i} = \exp(\mathbf{U}_{1i}^T \boldsymbol{\beta}_1 + \tau_{1i})$ and $\mu_{2i} = \exp(\mathbf{U}_{2i}^T \boldsymbol{\beta}_2 + \tau_{2i})$. Then the log likelihood of equation (5.5) can be written as

$$l(\boldsymbol{\beta}, \boldsymbol{\theta}) = \log \int \exp(\Phi(\boldsymbol{\tau})) d\boldsymbol{\tau}. \tag{5.7}$$

Let $\boldsymbol{\tau}_0$ be the maximiser of $\Phi(\boldsymbol{\tau})$, then by Laplace approximation we have

$$\int \exp(\Phi(\boldsymbol{\tau})) d\boldsymbol{\tau} = \exp \left\{ \Phi(\boldsymbol{\tau}_0) + \frac{N_1 + N_2}{2} \log(2\pi) - \frac{1}{2} \log |\mathbf{H}| \right\} \tag{5.8}$$

where $\mathbf{H}$ is the second derivative of $\log p(\mathbf{z}, \boldsymbol{\tau})$ respect to $\boldsymbol{\tau}$ and evaluated at $\boldsymbol{\tau}_0$ values

(Wood, 2012). Thus, $\mathbf{H} = \mathbf{C} + \mathcal{K}^{-1}_{N_1 N2}(\boldsymbol{\theta})$ and $\mathbf{C}$ is a diagonal matrix,

$$\begin{aligned}\mathbf{C} &= Diag(\exp(\mathbf{U}_{11}^T\hat{\boldsymbol{\beta}}_1 + \tau_{011}), ..., \exp(\mathbf{U}_{1N_1}^T\hat{\boldsymbol{\beta}}_1 + \tau_{01N_1},\\ &\quad exp(\mathbf{U}_{21}^T\hat{\boldsymbol{\beta}}_2 + \tau_{021}), ..., \exp(\mathbf{U}_{2N_2}^T\hat{\boldsymbol{\beta}}_2 + \tau_{02N_2})).\end{aligned}$$

In order to estimate hyper-parameters, we maximize the likelihood function with Laplace approximation in equation (5.7) and (5.8).

5.4 Predictions

In term of prediction, it is of interest to predict $\mathbf{z} = (z_1^*, z_2^*)^T$ at a new point $\mathbf{U} = (U_1^*, U_2^*)$ and $\mathbf{x} = (x_1^*, x_2^*)$. We use $\mathcal{D}$ to denote all the training data and assume that the model itself has been trained (i.e. all unknown parameters have been estimated) by the method discussed in the previous section. The main purpose of this section is to calculate $E(\mathbf{z}^*|\mathcal{D})$ and $Var(\mathbf{z}^*|\mathcal{D})$.

Let $\mathbf{x}^* = (x_1^*, x_2^*)$ be a new input and $\boldsymbol{\tau}^* = \boldsymbol{\tau}(\mathbf{x}^*) = (\tau_1^*, \tau_2^*)$ be the underlying latent variable at $\mathbf{x}^*$. The expectation of $\mathbf{z}$ conditional on $\boldsymbol{\tau}^*$ is given by

$$E(\mathbf{z}^*|\boldsymbol{\tau}^*, \mathcal{D}) = \begin{pmatrix} E(z_1^*|\tau_1^*, \mathcal{D}) \\ E(z_2^*|\tau_2^*, \mathcal{D}) \end{pmatrix} = \begin{pmatrix} \exp(\mathbf{U}_1^{*T}\hat{\boldsymbol{\beta}}_1 \\ \exp(\mathbf{U}_2^{*T}\hat{\boldsymbol{\beta}}_2 \end{pmatrix} = \exp(\mathcal{U}^{*T}\hat{\beta} + \boldsymbol{\tau}^*)$$

It follows that

$$E(\mathbf{z}^* \mid \mathcal{D}) = E[E(\mathbf{z}^* \mid \boldsymbol{\tau}^*, \mathcal{D})] = \int \exp(\mathcal{U}^{*T}\hat{\beta} + \boldsymbol{\tau}^*)p(\boldsymbol{\tau}^*|D)d\boldsymbol{\tau}^*. \tag{5.9}$$

We denote $\boldsymbol{\tau}^* = (\tau_1^*, \tau_2^*)$ and, to calculate the above expectation (5.9), we can approximate using a Laplace approximation where $p(\boldsymbol{\tau}^*|\mathcal{D})$ can be written as

$$\begin{aligned} p(\boldsymbol{\tau}^*|\mathcal{D}) &= \int p(\boldsymbol{\tau}^*|\boldsymbol{\tau}, \mathcal{D})p(\boldsymbol{\tau}|\mathcal{D})d\boldsymbol{\tau} \\ &= \int p(\boldsymbol{\tau}^*, \boldsymbol{\tau}|\mathcal{D})d\boldsymbol{\tau} \\ &= \frac{1}{p(\mathbf{z})}\int p(\mathbf{z}|\boldsymbol{\tau})p(\boldsymbol{\tau}^*, \boldsymbol{\tau})d\boldsymbol{\tau}. \end{aligned} \tag{5.10}$$

Hence, equation (5.9) above can be rewritten as

$$E(\mathbf{z}^*|\mathcal{D}) = \frac{1}{p(\mathbf{z})}\int\int \exp(\mathcal{U}^{*T}\hat{\beta} + \boldsymbol{\tau}^*)p(\mathbf{z}|\boldsymbol{\tau})p(\boldsymbol{\tau}^*, \boldsymbol{\tau})d\boldsymbol{\tau}d\boldsymbol{\tau}^*. \tag{5.11}$$

For convenience we denote $(\boldsymbol{\tau}, \boldsymbol{\tau}^*)^T$ and its covariance matrix $\mathcal{K}_{N_1 N_2}$ with dimension $(N_1 + N_2 + 2) \times (N_1 + N_2 + 2)$ by $\boldsymbol{\tau}_+$ and $\mathcal{K}_+$ respectively. Thus, the equation (5.10)

can be written as

$$E(\mathbf{z}^*|\mathcal{D}) = \frac{1}{p(\mathbf{z})}\int \exp(\mathcal{U}^{*T}\hat{\beta}+\boldsymbol{\tau}^*)\left[\prod_{i=1}^{N_1} p(y_{1i}|\hat{\boldsymbol{\beta}}_1,\tau_{1i})\right]\left[\prod_{i=1}^{N_2} p(y_{2i}|\hat{\boldsymbol{\beta}}_2,\tau_{2i})\right] \left[(2\pi)^{-\frac{(N_1+N_2+2)}{2}}|\mathcal{K}_+|^{-\frac{1}{2}}\exp(-\frac{1}{2}\boldsymbol{\tau}_+^T\mathcal{K}_+^{-1}\boldsymbol{\tau}_+)\right] d\tau_+.$$

The calculation of the integral is not tractable, since the dimension of $\boldsymbol{\tau}_+$ is usually very large. We now use Laplace approximation and denote

$$\tilde{\boldsymbol{\Phi}}(\boldsymbol{\tau}_+) = \log\exp(\mathcal{U}^{*T}\hat{\beta}+\boldsymbol{\tau}^*)+\sum_{i=1}^{N_1}\log p(z_{1i}|\hat{\boldsymbol{\beta}}_1,\tau_{1i})+\sum_{i=1}^{N_2}\log p(z_{2i}|\hat{\boldsymbol{\beta}}_2,\tau_{2i})- \frac{N_1+N_2+2}{2}\log(2\pi)-\frac{1}{2}\log|\mathcal{K}_+|-\frac{1}{2}\boldsymbol{\tau}_+^T\mathcal{K}_+^{-1}\boldsymbol{\tau}_+,$$

where $\log p(\mathbf{z}_1|\boldsymbol{\beta}_1,\boldsymbol{\tau}_1)=\sum_{i=1}^{N_1}(z_{1i}\log(\mu_{1i})-\mu_{1i})$ with $\mu_{1i}=\exp(\mathbf{U}_{1i}^T\boldsymbol{\beta}_1+\tau_{1i})$ and $\log p(\mathbf{z}_2|\boldsymbol{\beta}_2,\boldsymbol{\tau}_2)=\sum_{i=1}^{N_2}(z_{2i}\log(\mu_{2i})-\mu_{2i})$ with $\mu_{2i}=\exp(\mathbf{U}_{2i}^T\boldsymbol{\beta}_2+\tau_{2i})$. Equation (5.2) can be expressed as

$$E(\mathbf{z}^*|\mathcal{D})=\frac{1}{p(\mathbf{z})}\int\exp(\tilde{\boldsymbol{\Phi}}(\boldsymbol{\tau}_+))d\boldsymbol{\tau}_+. \tag{5.12}$$

Let $\hat{\boldsymbol{\tau}_+}$ be the maximiser of $\tilde{\boldsymbol{\Phi}}(\boldsymbol{\tau}_+)$, then by using Laplace approximation we have

$$\int\exp(\tilde{\boldsymbol{\Phi}}(\boldsymbol{\tau}_+))d\boldsymbol{\tau}_+ = \exp(\tilde{\boldsymbol{\Phi}}(\hat{\boldsymbol{\tau}_+})+\frac{N_1+N_2+2}{2}\log(2\pi)+ -\frac{1}{2}\log|\mathcal{K}_+^{-1}+\mathbf{C}_+|) \tag{5.13}$$

where $\mathbf{C}_+$ is the second derivative of

$$\log\exp(\mathcal{U}^{*T}\hat{\beta}+\boldsymbol{\tau}^*)+\sum_{i=1}^{N_1}\left[z_{1i}(\mathbf{U}_{1i}^T\hat{\boldsymbol{\beta}}_1+\tau_{1i})-\exp(\mathbf{U}_{1i}^T\hat{\boldsymbol{\beta}}_1+\tau_{1i})\right]$$

$$+\sum_{i=1}^{N_2}\left[z_{2i}(\mathbf{U}_{2i}^T\hat{\boldsymbol{\beta}}_2+\tau_{2_i})-\exp(\mathbf{U}_{2i}^T\hat{\boldsymbol{\beta}}_2+\tau_{2i})\right]$$

with respect to $\boldsymbol{\tau}_+$, evaluated at $\hat{\boldsymbol{\tau}}_+$. Therefore $\mathbf{C}_+$ becomes a diagonal matrix which can be written as follows.

$$\mathbf{C}_+ = Diag(\exp(\mathbf{U}_{11}^T\hat{\boldsymbol{\beta}}_1+\hat{\tau_{11}}),...,\exp(\mathbf{U}_{1N_1}^T\hat{\boldsymbol{\beta}}_1+\tau_{1\hat{N}_1}), \exp(\mathbf{U}_{21}^T\hat{\boldsymbol{\beta}}_2+\hat{\tau_{21}}),...,\exp(\mathbf{U}_{2N_2}^T\hat{\boldsymbol{\beta}}_2+\tau_{2\hat{N}_2}),0,0).$$

In terms of $Var(\mathbf{z}^*|\mathcal{D})$, we can evaluate it from

$$Var(\mathbf{z}^*|\mathcal{D}) = \begin{pmatrix} Var(z_1^*|\mathcal{D}) & Cov(z_1^*, z_2^*|\mathcal{D}) \\ Cov(z_1^*, z_2^*|\mathcal{D}) & Var(z_2^*|\mathcal{D}), \end{pmatrix} \tag{5.14}$$

and to calculate $Var(\mathbf{z}^*|\mathcal{D})$, we use the formula :

$$Var(z^*|\mathcal{D}) \quad = \quad E[Var(z^*|\boldsymbol{\tau}^*, \mathcal{D})] + Var[E(z^*|\boldsymbol{\tau}^*, \mathcal{D}), \tag{5.15}$$

where z could be either z_1 or z_2. From the model definition, we have

$$\begin{aligned} Var[E(\mathbf{z}^*|\boldsymbol{\tau}^*, \mathcal{D}) &= E[E(\mathbf{z}^*|\boldsymbol{\tau}^*, \mathcal{D})]^2 - [E[E(\mathbf{z}^*|\boldsymbol{\tau}^*, \mathcal{D})]]^2 \\ &= \int (\exp(\mathcal{U}^{*T}\hat{\beta} + \boldsymbol{\tau}^*))^2 p(\boldsymbol{\tau}^*|D) d\boldsymbol{\tau}^* - [E(\mathbf{z}^*|\boldsymbol{\tau}^*, \mathcal{D})]^2. \end{aligned} \tag{5.16}$$

The first equation in (5.16) can be obtained by Laplace approximation similar to $E(\mathbf{z}^*|\mathcal{D})$ in (5.12). Because $Var(\mathbf{z}^*|\boldsymbol{\tau}^*, \mathcal{D}) = E(\mathbf{z}^*|\boldsymbol{\tau}^*, \mathcal{D})$ for Poisson distribution, we can write

$$E[Var(\mathbf{z}^*|\boldsymbol{\tau}^*, \mathcal{D})] = E(\mathbf{z}^*|\mathcal{D}). \tag{5.17}$$

We can calculate the equation 5.17 exactly similar to equation (5.12).

We apply the formula to evaluate $Cov(z_1^*, z_2^*|\mathcal{D})$:

$$Cov(z_1^*, z_2^*|\mathcal{D}) = E[z_1^* z_2^*|\mathcal{D}] - E[(z_1^*|\mathcal{D})]E[(z_2^*|\mathcal{D})]. \tag{5.18}$$

The procedure for calculating $E[\mathbf{z}^*|\boldsymbol{\tau}^*, \mathcal{D}]$ can be applied to evaluate $E[z_1^* z_2^*|\mathcal{D}]$ and the second equation in (5.18) is exactly the same as (5.9).

5.5 Consistency

The prediction based on a $\mathcal{GPR}$ model is consistent when the sample size of the data collected from a certain curve is sufficiently large and the covariance function satisfies certain regularity conditions. The consistency does not depend on the common mean structure or the choice of the values of hyper-parameters involved in the covariance function, see Shi & Choi (2011).

In this section, we will discuss information consistency and extend it to a more general context than the result of Wang & Shi (2014). We focus on $\tilde{\mathbf{z}}$ to $\mathbf{z}$, where $\tilde{\mathbf{z}} = (\tilde{z}_{11}, ...\tilde{z}_{1N_1}, \tilde{z}_{21}, ..., \tilde{z}_{2N_2})$ are predicted observations and $\mathbf{z} = (z_{11}, ..., z_{1N_1}, z_{21}, ..., z_{2N_2})$ are actual observations. N_1 and N_2 are the number of observations of the first input and the second input respectively. We can rewrite the data as $\mathbf{z} = (z_{1i}, z_{2i})$ $i = 1, ..., N_1$ or N_2 at the points $\boldsymbol{\tau} = (\tau_{1i}, \tau_{2i})$ and corresponding covariate values $\mathbf{X}_{N_1 N_2} = (\mathbf{x}_{1i}, \mathbf{x}_{2i})$ where

$(\mathbf{x}_{1i}, \mathbf{x}_{2i}) \in \mathcal{X} \subset \mathbb{R}^P$ are independently drawn from distribution $\mathbf{u}(\mathbf{x})$.

We assume that z_{1i} and z_{2i} are sets of samples and follow a bivariate Poisson distribution with $\mu_{1i}(\mathbf{x}_i) = \exp(\mathbf{U}_{1i}^T\boldsymbol{\beta}_1 + \tau_{1i}(\mathbf{x}_i))$ and $\mu_{2i}(\mathbf{x}_i) = \exp(\mathbf{U}_{2i}^T\boldsymbol{\beta}_2 + \tau_{2i}(\mathbf{x}_i))$ respectively and $(\tau_{1i}(\cdot), \tau_{2i}(\cdot)) \sim \mathcal{MGP}(\mathbf{0}, \mathcal{K}(\cdot,\cdot))$ was discussed in the previous chapter. Therefore, the stochastic process $\tau_1(\cdot)$ and $\tau_2(\cdot)$ induces a measure on space $\mathcal{F} : \{f(\cdot) : \mathcal{X} \to \mathbb{R}\}$. For convenience, we can rewrite

$$\begin{aligned}\mathbf{z} &= (z_{11}, ..., z_{1N_1}, z_{21}, ..., z_{2N_2}) \\ &= (z_1, ..., z_{N_1}, z_{N_1+1}, ..., z_{N_1+N_2})\end{aligned}$$

and covariate values $\mathbf{X}_{N_1N_2} = (\mathbf{x}_1, ...\mathbf{x}_{N_1}, \mathbf{x}_{N_1+1}, ..., \mathbf{x}_{N_1+N_2})$. Let

$$\mathcal{D}_{N_1N_2} = \{(\mathbf{x}_i, z_i), i = 1, ..., N_1 + N_2)\},$$

where

$$E(\mathbf{z}|\boldsymbol{\tau}) = \begin{pmatrix} E(z_{1i}|\tau_{1i}) \\ E(z_{2i}|\tau_{2i}) \end{pmatrix} = \begin{pmatrix} \exp(\mathbf{U}_{1i}^T\boldsymbol{\beta}_1 + \tau_{1i}(\mathbf{x}_{1i})) \\ \exp(\mathbf{U}_{2i}^T\boldsymbol{\beta}_2 + \tau_{2i}(\mathbf{x}_{2i})) \end{pmatrix} = \exp(\mathcal{U}^T\hat{\beta} + \tau_i(\mathbf{x}_i)).$$

Suppose that the hyper-parameters $\boldsymbol{\theta}$ in the covariance function are estimated by an empirical Bayesian method and the estimator is denoted by $\tilde{\boldsymbol{\theta}}$. Let τ_0 be the true underlying function, i.e. the true mean of z_i given by $\mu_i = \exp(\mathcal{U}_i^T\boldsymbol{\beta} + \tau_0(\mathbf{x}_i))$. Denote

$$p_{mgp}(\mathbf{z}) = \int p(z_1, ..., z_{N_1}, z_{N_1+1}, ..., z_{N_1+N_2}|\tau(\mathbf{x})dp_{N_1+N_2}(\tau)$$

and

$$p_0(\mathbf{z}) = p(z_1, ..., z_{N_1}, z_{N_1+1}, ..., z_{N_1+N_2}|\tau_0(\mathbf{x})),$$

then $p_{mgp}(\mathbf{z})$ is the Bayesian predictive distribution of $\mathbf{z}$ based on $\mathcal{CGP}$ model. Note that $dp_{N_1+N_2}(\tau)$ depends on the $N_1 + N_2$ since the hyper-parameters of τ are estimated from the data. We say that p_{mgp} achieves information consistency if

$$\frac{1}{N_1 + N_2} E_{\mathbf{X}_{N_1N_2}} \left(D[p_0(\mathbf{z}), p_{mgp}(\mathbf{z})]\right) \to 0 \quad as \quad N_1 \to \infty \quad and \quad N_2 \to \infty, \tag{5.19}$$

where $E_{\mathbf{X}_{N_1N_2}}$ denotes the expectation under the distribution of $\mathbf{X}_{N_1N_2}$ and $D[p_0(\mathbf{z}), p_{mgp}(\mathbf{z})]$ is the Kullback-Leibler divergence between $p_0(.)$ and $p_{mgp}(.)$, i.e. ,

$$D[p_0(\mathbf{z}), p_{mgp}(\mathbf{z})] = \int p_0(\mathbf{z}) \log \frac{p_0(\mathbf{z})}{p_{mgp}(\mathbf{z})} d\mathbf{z}$$

Theorem 5.1. *Under the $\mathcal{MCGPPR}$ model* (5.1) *and the condition given in Lemma D.1,*

the prediction $\hat{\mathbf{z}}$ *is information consistent to the true curve* $\mathbf{z}_0$ *if the RKHS norm* $\|\tau_0\|^2_{K_{N_1N_2}}$ *is bounded and the expected regret term* $E_{\mathbf{X}_{N_1N_2}}(\log|\mathbf{I}+\delta\mathcal{K}_{N_1N_2}|) = o(N_1+N_2)$. *The error bound is specified in* (D.8).

The proof of the theorem is also given in Appendix D.

Remark. The condition $|b''(\alpha)| \leqslant e^{\kappa\alpha}$ in Lemma D.1 is satisfied for a bivariate Poisson model with a convolved GPR where $b(\alpha) = \log p(z_i|\tau(\mathbf{x}_i)) = e^{\alpha}$ with $\alpha = \mathcal{U}_i^T\boldsymbol{\beta} + \tau_i$

Remark. The regret term $R = \log|\mathbf{I} + \delta\mathcal{K}_{N_1N_2}|$ depends on the covariance function $(\mathbf{x}_i, \mathbf{x}_j)$ for a convolved bivariate Gaussian Process and covariate distribution $\mathbf{u}(\mathbf{x})$. The convolved Gaussian process still belongs stationary Gaussian processes, see Choi (2005). We can use it to identify the upper bounds of the expected regret for some widely used covariance functions by extending some specific results in Seeger *et al.* (2008) and Wang & Shi (2014). The details of discussion is in Appendix D.

5.6 Numerical Results

In this section, we demonstrate the proposed method by comprehensive simulation studies with several scenarios and also present results from some real data analysis.

5.6.1 Simulation Studies

i. Scenario 1

In the first scenario, we use a discrete bivariate Poisson regression model in 5.4 as the true model to generate random data from bivariate Poisson distribution.

$$\begin{pmatrix} z_{1i}(\mathbf{x}_i) \\ z_{2i}(\mathbf{x}_i) \end{pmatrix} \sim \begin{pmatrix} Poisson(\mu_{1i}(\mathbf{x}_i)), & i = 1, ..., N_1 \\ Poisson(\mu_{2i}(\mathbf{x}_i)), & i = 1, ..., N_2 \end{pmatrix} \tag{5.20}$$

where

$$\begin{pmatrix} \mu_{1i}(\mathbf{x}_i) = \exp(\mathbf{U}_{1i}^T\boldsymbol{\beta}_1 + \tau_{1i}(\mathbf{x}_i)) \\ \mu_{2i}(\mathbf{x}_i) = \exp(\mathbf{U}_{2i}^T\boldsymbol{\beta}_2 + \tau_{2i}(\mathbf{x}_i)) \end{pmatrix}, \qquad \begin{pmatrix} \tau_{1i}(\cdot) \\ \tau_{2i}(\cdot) \end{pmatrix} \sim \mathcal{MGP}(\mathbf{0}, \mathcal{K}(\cdot,\cdot)),$$

and $\mathcal{K}(\cdot,\cdot)$ is defined by (4.8) and (4.10) with the following hyper-parameter. $\boldsymbol{\beta}$ are $\beta_{10} = 1$, $\beta_{11} = 2$, $\beta_{20} = 1$ and $\beta_{21} = 2$.

Random processes τ_{1i} and τ_{2i} are generated by a mixed covariance structure which is the same as the one used in Scenario 1, Section 4.4.3, i.e. the combination of the squared exponential covariance function and the Gamma exponential covariance function. We recall the setting of the covariance structure for true model as follows, i.e. the

covariance structure, which contains η_1, is generated from a squared exponential covariance function with true values of $(w_1 = 0.04, B_1 = 1)$ and η_2 is built from a Gamma exponential family covariance function with the true values of $(w_2 = 0.04, B_2 = 1)$. Also, the shared processes $\boldsymbol{\xi}$ follow the squared exponential covariance function with true values $(v_1 = 0.04, v_2 = 0.04, A_1 = 1, A_2 = 1)$. Here, $\mathbf{x}_i$'s are equally spaced in $[-5, 5]$ and $\boldsymbol{\tau} = \{\tau_{1i}, \tau_{2i}\}$ are dependent Gaussian processes which are simulated with the above true values. The observations z_{ai} follow a Poisson distribution with $\mu_{ai} = \exp(\mathbf{U}_{ai}^T \boldsymbol{\beta}_a + \tau_{ai})$ where $a = 1, 2$ and $i = 1, ..., N_1$ or N_2 and each response variable contains 20 observations.

We set two different covariance structures of $\boldsymbol{\tau}$ as priors for our $\mathcal{MCGPPR}$ models and compare them with an existing approach. We recall again equation (4.6) to define random processes $(\boldsymbol{\tau})$. The details are as follows

$$\begin{aligned} \tau_1 &= \xi_1 + \eta_1 \\ \tau_2 &= \xi_2 + \eta_2 \end{aligned}$$

where $\eta_1(\mathbf{x})$ and $\eta_1(\mathbf{x})$ are two independent $\mathcal{CGP}$s since they are constructed by independent Gaussian white noises. Different kernels $h_1(\mathbf{x})$ and $h_2(\mathbf{x})$ are used to define the different covariance structures and the same white noise defines the dependency between $\xi_1(\mathbf{x})$ and $\xi_2(\mathbf{x})$, see equations (4.4) and (4.5) in Chapter 4.

1. *Model 1*
   Here, ξ_1, ξ_2 and η_1 have squared exponential covariance functions and η_2 has a Gamma exponential covariance function. It is also the same as the above true model.
2. *Model 4*
   All η_1, η_2, ξ_1 and ξ_2 have squared exponential covariance functions.
3. *Model 5*
   For comparison, we use the $(\mathcal{CD})$ model developed in Crainiceanu *et al.* (2012). We recall the model in equation (4.28), it assumes that $\boldsymbol{\tau}_1$ and $\boldsymbol{\tau}_2$ are two dependent processes that can be written as follows

$$\boldsymbol{\tau}_2 \mid \boldsymbol{\tau}_1 \sim \mathcal{N}(\alpha \boldsymbol{\tau}_1, \sigma_\epsilon^2) \tag{5.21}$$

   where $\boldsymbol{\tau}_1$ is a Gaussian process with zero mean and a squared exponential covariance function.

Table 5.1 shows the results of the sample mean of the estimated parameters $\hat{\boldsymbol{\beta}}$ us-

ing an empirical Bayesian approach for three different models based on one hundred replications. Table 5.2 provides the values of RMSE (RMSE) based on one hundred replications.

	Sample Mean			
Model	β_{11}	β_{12}	β_{21}	β_{22}
True values	1	2	1	2
Model 1	0.99918	1.99568	0.99467	1.99344
Model 4	0.99698	2.00071	0.99615	2.00055
Model 5	0.98020	2.02617	0.98334	2.00496

Table 5.1: The sample mean of estimated parameters ($\boldsymbol{\beta}$) for bivariate Poisson model with Convolved $\mathcal{GPR}$ based on one hundred times replications.

	RMSE			
Model	β_{11}	β_{12}	β_{21}	β_{22}
Model 1	0.03496	0.04547	0.03967	0.03739
Model 4	0.04833	0.04560	0.04106	0.04066
Model 5	0.13076	0.17025	0.13972	0.15640

Table 5.2: The values of RMSE (RMSE) from estimated parameters $\boldsymbol{\beta}$ for bivariate Poisson model with Convolved $\mathcal{GPR}$ based on one hundred times replications.

From Table 5.1, it can be seen that empirical Bayesian estimation provides good estimate of parameters and all three different models have a similar result, although *Model 1* gives the closest value of estimated parameters to the true values as expected.

It is not surprising that *Model 1* gives the best results since it has the same covariance structure as the true model. However, *Model 4* also gives fairly good results although the covariance structure is not the same as the true model. It shows the great flexibility and robustness of the $\mathcal{MCGPPR}$ model proposed in this chapter. The results of both *Model 1* and *Model 4* are much better than those of *Model 5*.

ii. Scenario 2

In the second scenario, we set up several different $\mathcal{MCGPPR}$ models with mixed covariance functions to illustrate how sensitively the results depend on the choice of covariance structures. We use the same true model and true values as the first scenario, i.e. we use the mixed the covariance structure of Gamma exponential and squared exponential covariance functions.

Several different models have been set up previously based on scenario 1 in Chapter 4. Firstly, we redefine four different priors of $\mathcal{MCGPPR}$ models and comparison model. In this scenario, we still apply the same priors as the first scenario, i.e. *Model 1*, *Model*

4 and *Model 5*. Also for the remaining models namely, *Model 2* and *Model 3*, we recall the set up as follows

a. *Model 2*
   η_2 has a rational quadratic covariance function and other Gaussian processes (ξ_1, ξ_2, η_1) have covariance functions similar to *Model 1*, i.e. squared exponential covariance function

b. *Model 3*
   η_2 has a Matern covariance function with $\nu = \frac{3}{2}$ and the remaining convolved Guassian processess (ξ_1, ξ_2, η_1)have squared exponential covariance functions.

We randomly generate 40 observations as training data and test data separately for each component response. After estimating parameters using empirical Bayesian estimates for all training data, we then use estimated parameters to predict test data. Table 5.3 shows the average values of RMSE (Average RMSE) for different models between μ and $\hat{\mu}$ based on one hundred replications.

Model	Average RMSE
Model 1	0.02627
Model 2	0.03841
Model 3	0.03028
Model 4	0.03459
Model 5	0.10920

Table 5.3: Average RMSE prediction between μ and $\hat{\mu}$ from various models based on one hundred replications.

From the results in Table (5.3), the average RMSE's are quite similar to each other apart from *Model 5*. *Model 1* also provides the best result among all of the models as expected. For other models, the results are still reasonable, even though we used misspecified covariance functions in those models. That means that the $\mathcal{MCGPPR}$ model is robust. Also, similar to the result of Scenario 1 in Chapter 3, the comparison model *Model 5* fails again to achieve a good performance.

iii. Scenario 3
In the third scenario, we will show another important feature of the proposed model, i.e. dealing with multidimensional covariates in the covariance structure. We set the true model as follows

1. Generate random processes

The following is the true model used to generate multiple processes.

$$\begin{aligned} y_{1i}(\mathbf{x}_i) &= 0.2\mathbf{x}_{1i} \times |\mathbf{x}_{1i}|^{\frac{1}{3}} + \log(\mathbf{x}_{2i}) + \tau_{1i}(\mathbf{x}_i) \quad i = 1, ..., N_1 \\ y_{2i}(\mathbf{x}_i) &= sin(\mathbf{x}_{2i}) + 0.4\mathbf{x}_{2i} \times |\mathbf{x}_{1i}|^{\frac{1}{4}} + \tau_{2i}(\mathbf{x}_i) \quad i = 1, ..., N_2 \end{aligned}$$

where $(\tau_{1i}(\cdot), \tau_{2i}(\cdot)) \sim \mathcal{MGP}(\mathbf{0}, \mathcal{K}(\cdot, \cdot))$ and $\mathcal{K}(\cdot, \cdot)$ is explained in Chapter 4.
$\mathbf{x} = \{\mathbf{x}_{1i}, \mathbf{x}_{2i}\}$ are equally spaced points $[-5, 10]$ and $[1, 2]$ respectively and $\boldsymbol{\tau} = \{\tau_{1i}, \tau_{2i}\}$ is dependent Gaussian processes which can be formed in a similar way to the first scenario in *Model 1*, i.e. a mixture squared exponential covariance function and a Gamma exponential covariance function were used. Also the true values are the same as those used in Scenario 1. We recall the setting of the mixed form of covariance structures based on Scenario 1. The structure of the covariance function is such that η_1 has a squared exponential covariance function and η_2 has a Gamma exponential covariance functions with the true values $(w_1 = 0.04, B_1)$ and $(w_2 = 0.04, B_2 = 1)$ respectively. The shared processes $\boldsymbol{\xi} = (\xi_1, \xi_2)$ have squared exponential covariance functions with the true values $(v_1 = 0.04, A_1 = 1, v_2 = 0.04, A_2 = 1)$. Each generated process contains 20 observations. We generate training data and test data separately.

2. Generate random multivariate Poisson data
   In order to generate randomly multivariate Poisson data, we need to calculate the mean of each response component as following

$$\begin{aligned} \boldsymbol{\mu}_{1i}(\mathbf{x}_i) &= \exp(\mathbf{y}_{1i}(\mathbf{x}_i)); \quad \mathbf{z}_{1i}(\mathbf{x}_i) \sim Poisson(\boldsymbol{\mu}_{1i}(\mathbf{x}_i)), \quad i = 1, ..., N_1 \\ \boldsymbol{\mu}_{2i}(\mathbf{x}_i) &= \exp(\mathbf{y}_{2i}(\mathbf{x}_i)); \quad \mathbf{z}_{2i}(\mathbf{x}_i) \sim Poisson(\boldsymbol{\mu}_{2i}(\mathbf{x}_i)), \quad i = 1, .., N_2 \end{aligned}$$

In this setting, we consider two different models, i.e. *Model 1* assuming a similar covariance structure to the true model and *Model 5*, a comparison model formed based on equation (5.21).

Figures 5.1 and 5.2 show a comparison between the predicted means of underying process $\mathbf{y}$ using *Model 1*, *Model 5* and the true mean curve for the first and the second output.

In Figure 5.1, there are some points worth noting. Although the estimated curve of *Model 5* is acceptable, it still is not better than the best model, i.e. *Model 1* which provides the smallest difference between estimated mean curve and mean true curve. Another important feature is that our proposed model is able to provide good performance which deals with multidimensional covariance structures.

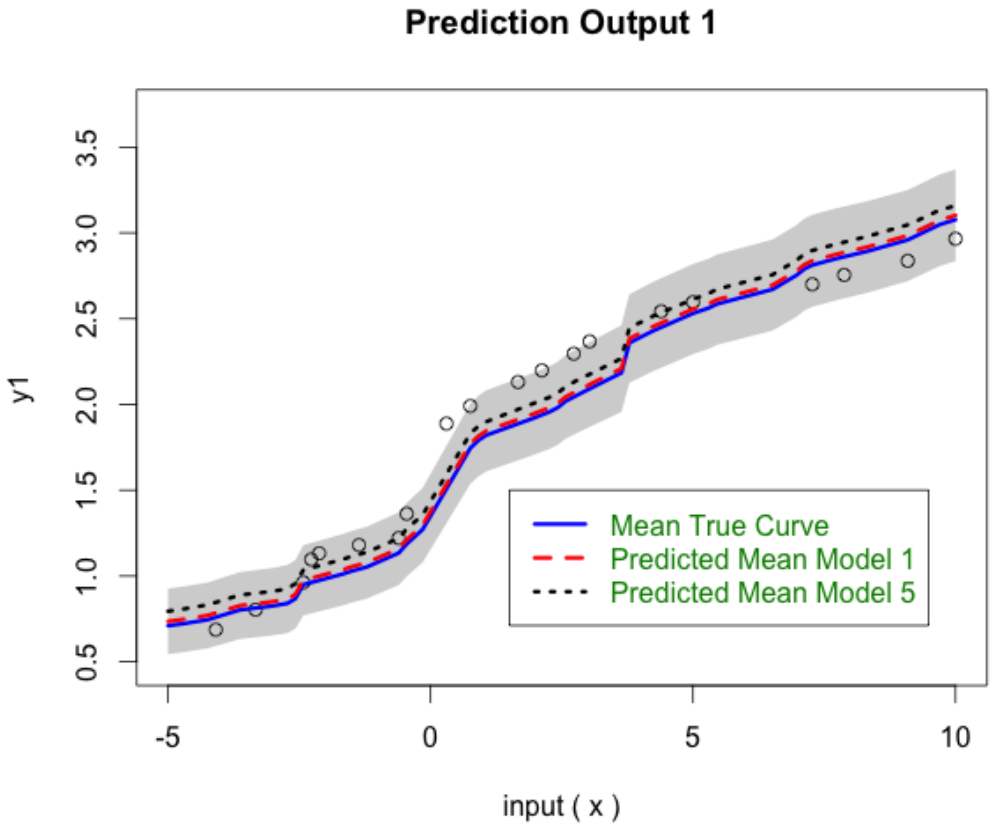

Figure 5.1: The predictions of the first outputs of bivariate Poisson regression where the dots are test data (sample), the red and black dashed lines represent the predicted mean of underlying process y_1 using *Model 1* and *Model 5* respectively. The blue solid lines are the true mean curve with 95% confidence intervals (the shaded regions).

Table 5.4 shows the average values of RMSE (average RMSE) between the estimated $\boldsymbol{\mu}$ and $\hat{\boldsymbol{\mu}}$ for the two models based on one hundred replications. Table 5.4 shows that

Model	Average RMSE
Model 1	0.00304
Model 5	0.00579

Table 5.4: The average of RMSE (Average RMSE) from prediction between μ and $\hat{\mu}$ with two different models based on one hundred replications.

Model 1 performs much better than *Model 5*.

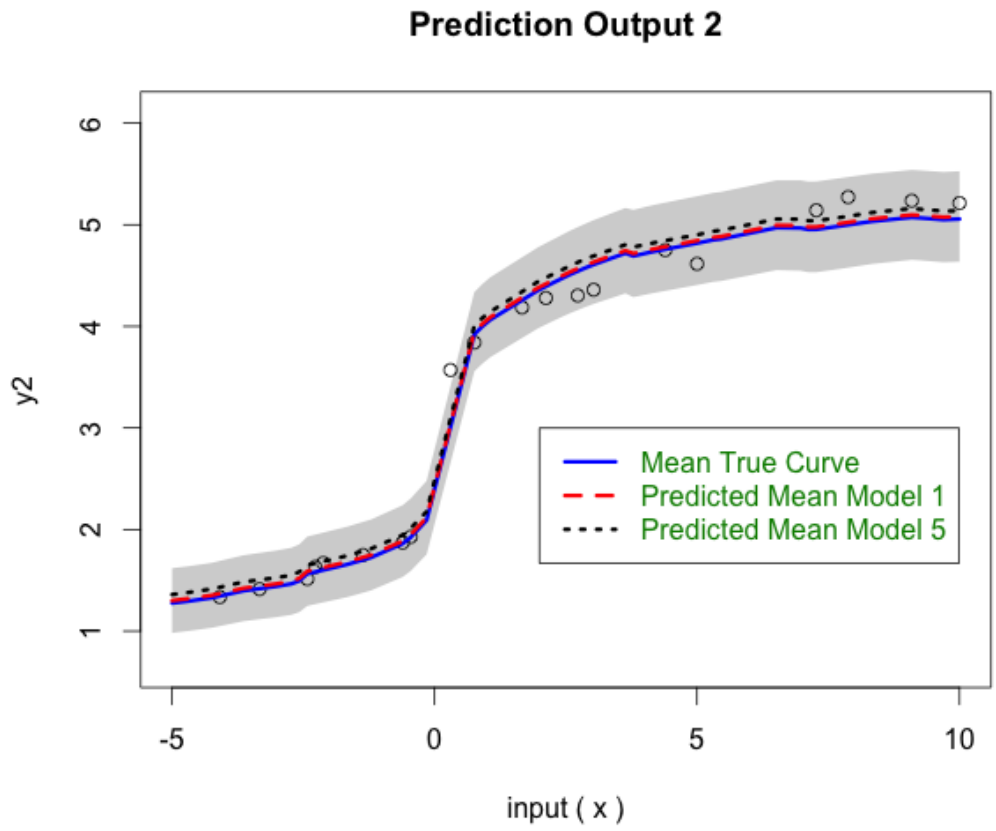

Figure 5.2: The predictions of the second outputs of bivariate Poisson regression where the dots are test data (sample), the red and black dashed lines represent the predicted mean of underlying process y_1 using *Model 1* and *Model 5* respectively. The blue solid lines are the true mean curve with 95% confidence intervals (the shaded regions).

5.6.2 Real Data Analysis

We will present results for two real sets of data. The first one is data relating to two type of cancers in Minnesota, USA. The second data concern Dengue fever and Malaria in Indonesia.

1. Lung and Oesophageal Cancer data
   From information on the NHS web site (www.nhs.uk), we believe that one of the most dangerous and common types of cancer is lung cancer. Every year there are around 44,500 people diagnosed with this condition in the UK. The symptoms usually do not always appear in the early stages, although some symptoms develop in many people, such as blood or persistent coughing, breathlessness and weight loss. In over 85 percent of cases, the main cause of lung cancer is cigarette smoking although people who have never smoked can be diagnosed with this cancer. Smoking can cause other cancers,

such as oesophageal cancer and mouth cancer.

There are more than 8,500 new cases of oesophageal cancer diagnosed each year in the UK which means that this cancer is uncommon but is not rare. As with lung cancer, smoking and drinking alcohol are the risk factors associated with this cancer.

In 2005, Jin, et al. analyzed the relationship between lung cancer and oesophageal cancer using a generalized intrinsics autoregressive model which was based on neighbourhood for each region as the main effect of the model. Unfortunately, the existing model has some difficulty in prediction when determining of the neighbourhood for each area. When applying our proposed model, we expect that it will overcome these difficulties and provide an easy way to predict cases in future.

We present the number of cases for each cancer in Minnesota, US. The map presented in 5.3 shows clearly that the county-level maps of the age-adjusted standardized mortality ratios (SMRs) between lung and oesophageal have a positive correlation across region or area. This strong evidence motivates us to use the $\mathcal{MCGPPR}$ model. Summaries

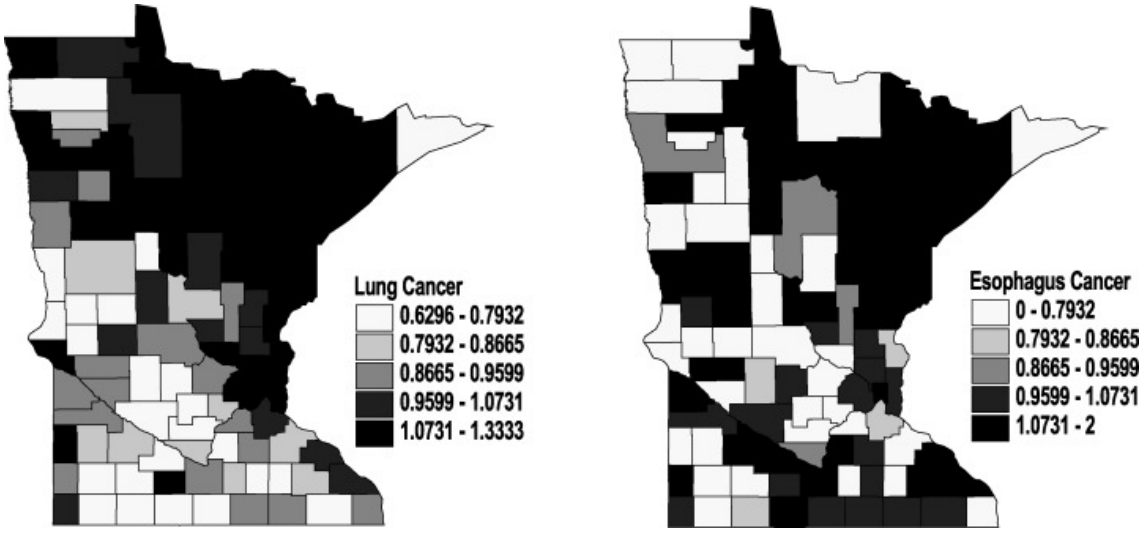

Figure 5.3: Maps of age-adjusted SMR for lung and esophagus cancer in Minnesota (Source : Biometrics , 61(4) : 950-61, December 2005)

of Lung and Oesophageal Cancer data in Minnesota also can be seen in Table 5.5.

Variable (cases)	Min	Max
Lung cancer	15.0	3797.0
Esophagus cancer	0.0	319.0

Table 5.5: Summaries of Lung and Oesophageal Cancer data

The model can be written as

$$z_{ia} \sim Poisson(E_{ia} e^{\tau_{ia}(\mathbf{x}_i)}), \quad i = 1, ..., 87, \quad a = 1, 2 \tag{5.22}$$

where z_{ia} is the observed number of deaths due to cancer a in county i, E_{ia} is the corresponding expected number of deaths (assumed known) and $\tau_{ia}(\cdot) \sim \mathcal{MGP}(\mathbf{0}, \mathcal{K}(\cdot, \cdot))$ which is explained in equations (4.8) and (4.10). Here, all component of $\mathcal{K}(\cdot, \cdot)$ have a squared exponential covariance function. Also, $\mathbf{x}$ are defined as latitude and longitude for each county. Then we compare our proposed models with the $\mathcal{CD}$ model.

To measure the goodness of fit of the model, we use the Akaike information criteria (AIC) which are shown in Table 5.6. From Table 5.6 it shows clearly that using our

Method	AIC
$\mathcal{CD}$	1640.202
$\mathcal{MCGPPR}$	1399.822

Table 5.6: AIC's values for different methods

model is better than $\mathcal{CD}$ model.

We now select data randomly from the whole data set to form training data consisting of two thirds of the data and the remainder is used for test data. Then we estimate parameters by an empirical Bayesian approach using training data and after that, we consider the problem of prediction for test data. To measure predictive performance of the models, we compare the predicted responses with true values based on ten replications. Then we calculate the average of values of realtive error (Average RE) which can be defined as the average of the difference between the predicted values and the actual observation defined by actual observation. In Table 5.7 shows the average of the relative error of actual observation and $\hat{\boldsymbol{\mu}}$. The performance of $\mathcal{MCGPPR}$ is much better than $\mathcal{CD}$.

Method	Average RE
$\mathcal{CD}$	0.0149
$\mathcal{MCGPPR}$	0.0080

Table 5.7: The average of relative error for different methods based on ten replications

2. Dengue Fever and Malaria data

   In Chapter 2, we analysed dengue fever and malaria data separately. We believe that the two of response variables have spatial correlation. We compared several methods to deal with this spatial effect, i.e. using an intrinsic autoregressive model, a Gaussian process prior and a conventional Poisson regression model. From the results, we have noted that using a Gaussian process prior provides more flexibility in defining the

covariance structure. However, both of the diseases can be spread by two different types mosquitoes which are hard to distinguish from each other. Dengue fever and malaria are from infected Aedes Aegypti and the plasmodium parasite of Anopheles mosquitoes respectively. Therefore, it is more sensible if we analyse them together as those diseases are correlated.

We consider three type of models and include the different sets of multidimensional covariates in the covariance structure. The first model involves location (latitude and longitude) and all covariates (health water (x_1), healthy rubbish bin (x_2), waste water disposal facilities (x_3), clean and healthy behaviour (x_4) and healthy house (x_5)). The second model contains location and some of the covariates, x_1, x_2, x_3, i.e. healthy water, healthy rubbish bin and waste water disposal facilities respectively . The third model for the covariance structure uses only distance measured by geographical position (latitude and longitude). Summaries Dengue Fever data can be seen in Table 3.4. For Malaria data, the minimum cases is zero and the maximum cases is 428.

Models	Average ER	
	$\mathcal{MCGPPR}$	$\mathcal{CD}$
Full (location and all covariates)	0.000994	0.001374
Location and x_1, x_2, x_3	0.001018	0.002000
Location	0.001137	0.002252

Table 5.8: The average of error rate for different set of covariates in covariance structures between $\mathcal{MCGPPR}$ model and $\mathcal{CD}$ approach based on fifteen replications

To compare the performance of the models, we randomly selected two thirds of the cities as training data and the remaining cities as test data. After obtaining estimates of the parameters, we then predict the test data . Table 5.8 gives the average values of relative error (Average RE) for each model, which is defined as the average of the difference between the prediction and the observation with respect to the observation based on fifteen replications. We also compare them with the $\mathcal{CD}$ approach. Table 5.8 shows clearly that the model with location and all covariates involved in the covariance structure of our proposed model give the best results with an average percentage error 0.09 % only. Although the two other models from the proposed approach are not a good choice, the results are still acceptable. It also shows that the $\mathcal{MCGPPR}$ model performs better than the $\mathcal{CD}$ model.

5.7 Chapter Summary

In this chapter, we proposed a novel method for multivariate Poisson regression analysis by applying different types of stationary covariance functions and also investigating a mixed

form using convolved Gaussian process priors. One of the most important features of the model is that it enables us to deal with the relationship between multidimensional Poisson covariates and a Poisson dependent variable. The multiple dependent processes formed by a convolved Gaussian process $\mathcal{CGP}$ model can be treated as nonlinear random effects.

We provided a framework for constructing a model for multivariate Poisson regression using convolved Gaussian process priors. The procedures for inference and implementation in simulation studies and real data analysis are established. We also presented the asymptotic theory based on information consistency.

There are several advantages of the proposed model that are worth noting. Similar to the previous chapter, the model provides a robust approach and offers flexibility in choosing covariance structures and this is confirmed in our comprehensive simulation studies. The $\mathcal{MCGPPR}$ model performs very well in term of prediction.

Another important feature of $\mathcal{MCGPPR}$ model is that it is able to address the problem of large dimensional covariates. Based on the third simulation study, the proposed model offers good performance with a multidimensional covariates scenario. In all scenarios, we compare $\mathcal{MCGPPR}$ model with the $\mathcal{CD}$ model. Our model performs consistently better than the existing approach.

In this chapter, even though our framework model focuses on multivariate Poisson regression, it is not difficult to extend the model to other multivariate non-Gaussian data in the exponential families. In the next chapter, we will discuss the general model.

We offered a general framework model to use $\mathcal{CGP}$ priors in multivariate semiparameteric regression analysis for response variables from Gaussian and non-Gaussian distributions in the exponential family of distribution. The model, and its implementation, including the technical details of the inference, are provided. We also reported asymptotic theory based on information consistency of the general model for multivariate non-Gaussian. The natural general extension can be applied for any distribution which assumes data from the exponential family distribution. Comprehensive simulation studies and applications with real data are also explored. The performance of the proposed models are usually better than other methods and show good flexibility and robustness.

Related to the topics discussed in this thesis, some interesting problems are worth further attention. For example, large dimensional integration. We used Laplace approximation. It provides reasonably good results. However, the approximation error will increase when the dimension increases (bear in mind that the dimension of the integration is equal to the sample size). We need to develop more efficient algorithms; see e.g Wang & Shi (2014).

We also encounter restriction sin terms of building mixture covariance functions. It seems that not every covariance function is integrable; thus, here we just focused on some stationary covariance functions, such as exponential squared, rational quadratic,

Matern and Gamma exponential covariance functions. Therefore, a further investigation is needed to provide a very flexible model although the proposed models have offered a good performance. We focused on bivariate response variables in the thesis. Although there is no difficulty in extending the method to the multivariate case in theory, the implementation may be challenging. More research in this area is essential.

Chapter 6

Convolved Gaussian Process Priors for Multivariate Non-Gaussian Regression Analysis

The aim of this chapter is to extend the $\mathcal{MCGPPR}$ model (5.1) to situations where the response variable, denoted by $\mathbf{z} = (z_1, z_2)$, is known to be non-Gaussian data. In the previous chapter we discussed Poisson data; here we also focus on other bivariate non-Gaussian data. The work is motivated by the following example which concerns data collected during adverse birth outcomes. Length of pregnancy and birth weight are crucial factors that can determine infant health and survival rates for years to come. The huge risk for mortality and a variety of health and developmental issues come from preterm and low birth weight infants. Preterm birth (PTB) is defined as the length of gestation being less than 37 weeks, and weighing less than 2,500 grams is the definition of a low birth weight (LBW). The measures z_1 and z_1 take a value of either 0 or 1, each component corresponding to the normal term pregnancies or preterm birth and the normal or low weight respectively. Whilst there is considerable evidence that those factors are correlated and correlation also exists within geographic regions as demographic characteristics, socioeconomic characteristics can vary geographically. The aim of the example is to develop a general framework to investigate the regression relationship between the multivariate binary response variables $\mathbf{z}$ and a set of covariates. The general model also can define a cross-correlation between multivariate dependent non-Gaussian data.

Neelon *et al.* (2014) have attempted using a bivariate conditional autoregressive model. Unfortunately, the existing approach has a limitation, i.e. they assigned a uniform prior restricted to interval $(-1, 1)$ for parameter α which is used to control the within-subject correlation between PTB and LBW. Some authors, such as Crainiceanu *et al.* (2012) have provided another model using conditional dependency via a Gaussian process to deal

with bivariate binomial spatial data. But, again this approach also has a limitation, i.e. it has failed to accommodate individual characteristics since the conditional dependency structure is not able to capture them.

To overcome the issues, we will investigate an alternative method using convolution Gaussian process priors with our proposed model. This method has been discussed in the last two chapters and has some advantages worth noting , i.e. our proposed model via $\mathcal{CGP}$ priors ($\mathcal{MCGPPR}$) provided a promising result and good performance. It also offered flexibility in the choice of covariate functions. Therefore, it enables us to extend the idea to consider $\mathcal{CGP}$ priors for multivariate non-Gaussian regression, namely multivariate convolved Gaussian process non-Gaussian regression analysis ($\mathcal{MCGPNGR}$).

Similar to $\mathcal{MCGPPR}$, the advantages of this model include : (1) it provides a general framework on how to use a convolved Gaussian process to define a model for multivariate semiparametric regression analysis for multivariate response variables from exponential families with multi-dimensional covariates; (2) a cross-correlation between two responses is captured by modelling the mean and covariance structures simultaneously; and (3) it is able to accommodate a large multidimensional nonlinear function involving covariates since the priors are formed by covariance functions.

This chapter is organized as follows. The general framework of the $\mathcal{MCGPNGR}$ model, how we estimate hyper-parameters and calculate prediction and the asymptotic properties based on information consistency are provided in Section 6.1. Comprehensive numerical examples including simulation studies and real data applications are considered in Section 6.2.

6.1 The Model

Let $\mathbf{z}_a$ be a non-Gaussian response variable where a is a-th response. We assume that $\mathbf{z}_a$ are dependent and also within the response to z_{ai} and z_{aj} are dependent at different observations. We suppose that $\mathbf{z}_a$ has a distribution from an exponential family with the following density function

$$f(\mathbf{z}_a \mid \alpha_a, \phi_a) = \exp\left\{\frac{\mathbf{z}_a\alpha_a - b(\alpha_a)}{c(\phi_a)} + d(\mathbf{z}_a, \phi_a)\right\} \tag{6.1}$$

where α_a and ϕ_a are the canonical and dispersion parameter respectively. We have $E(\mathbf{z}_a) = b'(\alpha_a)$ and $Var(\mathbf{z}_a) = b''(\alpha_a)c(\phi_a)$, where $b'(\alpha_a)$ and $b''(\alpha_a)$ are the first and two derivatives of $b(\alpha)$ with respect to α.

Multivariate convolved Gaussian process non-Gaussian regression analysis ($\mathcal{MCGPNGR}$)

models are defined as follows :

$$E(\mathbf{z}_a \mid \boldsymbol{\tau}_a) \quad = \quad h(\boldsymbol{\mu}_a + \boldsymbol{\tau}_a) \quad a = 1, 2 \tag{6.2}$$

where $h^{-1}(\cdot)$ is a given link function. Similar to $\mathbf{z}_a$ in Chapter 4, here $\mathbf{z_a}$ (where $a = 1, 2$) also define two dependent response variables, for example the number of preterm and low weight infant in adverse birth outcomes data. Many researchers believe that the observations are spatially correlated. Thus, rather than analysing separately, it is more sensible to investigate those data together.

Some important features of $\mathcal{MCGPPPR}$ are also true for $\mathcal{MCGPNG}$. For example, it is a robust model; it provides the flexibility to construct a cross-correlation between two components. Here, we also assume that the latent variables $\boldsymbol{\tau}_a$ which depend on $\mathbf{x}$ and unknown hyper-parameters $\boldsymbol{\theta}$ define a cross-correlation of $\mathbf{z}_a$. A nonparametric $\mathcal{CGP}$ can be used as priors for $\boldsymbol{\tau}_a$ which is explained by equations (4.6) to (4.10). The regression relationship between the bivariate non-Gaussian regression $\mathbf{z}_a$ and the covariates $\mathbf{x}_a$ is modelled by the covariance structure of $\boldsymbol{\tau}(\mathbf{x})$.

If we use a linear mean function which depends on a set of p scalar covariates $\mathbf{U}_a$, then equation (6.2) can be written as

$$E(\mathbf{z}_a \mid \boldsymbol{\tau}_a) \quad = \quad h(\mathbf{U}_a^T\boldsymbol{\beta} + \boldsymbol{\tau}_a). \tag{6.3}$$

As an example, we consider a special case of dependent binary data (e.g. classification problem with two classes). Convolved Gaussian process priors for bivariate binomial regression can be specified as follows

$$\mathbf{z}_a|\boldsymbol{\tau}_a \sim Bin(1, \boldsymbol{\pi}_a).$$

If we use the logit link function, given by

$$\boldsymbol{\pi}_a = \frac{\exp(\mathbf{U}_a^T\boldsymbol{\beta}_a + \boldsymbol{\tau}_a)}{1 + \exp(\mathbf{U}_a^T\boldsymbol{\beta}_a + \boldsymbol{\tau}_a)}, \quad a = 1, 2, \tag{6.4}$$

where (τ_1, τ_2) follows a $\mathcal{CGP}$ as defined in equation (4.6) to (4.10). We can get the marginal density function of $\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2)$ using

$$p(\mathbf{z}) \quad = \quad \int p(\mathbf{z} \mid \boldsymbol{\tau}, \boldsymbol{\beta}) p(\boldsymbol{\tau} \mid \boldsymbol{\theta}) d\boldsymbol{\tau}, \tag{6.5}$$

where $p(\boldsymbol{\tau}|\boldsymbol{\theta})$ is the density function of $\boldsymbol{\tau} = (\boldsymbol{\tau}_1, \boldsymbol{\tau}_2)$. The density functions for other distributions from the exponential families can be obtained similarly.

As an analogue with the $\mathcal{MCGPPPR}$ model, to estimate $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$, we also use an

empirical Bayesian approach. In general, the technical details for inference regarding parameters and prediction using the $\mathcal{MCGPNGR}$ model are quite similar to the case of $\mathcal{MCGPPR}$ model discussed briefly in Chapter 5. Therefore, we can rewrite a general marginal log-likelihood function based on equation (5.5) as follows

$$\begin{aligned} l(\boldsymbol{\beta}, \boldsymbol{\theta}) &= \log \int \prod_{a=1}^{2} \prod_{i=1}^{N_a} p(z_{a_i} \mid \tau_{ai}, \boldsymbol{\beta}) (2\pi)^{\frac{-\sum_{a=1}^{2} N_a}{2}} \mid \mathcal{K}_{N_1 N_2} \mid^{-\frac{1}{2}} \\ &\quad \exp\left\{ \boldsymbol{\tau}^T \mathcal{K}_{N_1 N_2}^{-1} \boldsymbol{\tau} \right\} d\boldsymbol{\tau} \end{aligned} \tag{6.6}$$

where $p(z_{a_i} \mid \tau_{ai}, \boldsymbol{\beta})$ is derived from the exponential family as defined in equation (6.1) and (6.2).

Meanwhile, in terms of prediction, we focus on how we calculate $E(\mathbf{z}^* \mid \mathcal{D})$ and $Var(\mathbf{z}^* \mid \mathcal{D})$ which are used as the prediction mean and predictive variance of $\mathbf{z}^*$. In order to make a general form, we rewrite the expectation of $\mathbf{z}^*$ conditional on $\boldsymbol{\tau}^*$ based on equation (5.9) as follows

$$E(\mathbf{z}^* \mid \mathcal{D}) = E[E(\mathbf{z}^* \mid \boldsymbol{\tau}^*, \mathcal{D})] = \int h(\mathcal{U}^{*T} \hat{\beta} + \boldsymbol{\tau}^*) p(\boldsymbol{\tau}^* | \mathcal{D}) d\boldsymbol{\tau}^*. \tag{6.7}$$

The form of equation (6.7) is analytically intractable. We again a Laplace approximation. Therefore, the remaining inference is the same as in the previous chapter.

To calculate $Var(\mathbf{z}^* \mid \mathcal{D})$, we redefine equation (5.16) and (5.17) and the variance of $\mathbf{z}^*$ conditional on $\boldsymbol{\tau}^*$ is given by

$$Var(\mathbf{z}^* | \mathcal{D}) = E[Var(\mathbf{z}^* | \boldsymbol{\tau}^*, \mathcal{D})] + Var[E(\mathbf{z}^* | \boldsymbol{\tau}^*, \mathcal{D}). \tag{6.8}$$

Based on the model definition, we have

$$\begin{aligned} Var[E(\mathbf{z}^* | \boldsymbol{\tau}^*, \mathcal{D}) &= E[E(\mathbf{z}^* | \boldsymbol{\tau}^*, \mathcal{D})]^2 - [E[E(\mathbf{z}^* | \boldsymbol{\tau}^*, \mathcal{D})]]^2 \\ &= \int (h(\mathcal{U}^{*T} \hat{\beta} + \boldsymbol{\tau}^*))^2 p(\boldsymbol{\tau}^* | \mathcal{D}) d\boldsymbol{\tau}^* - [E(\mathbf{z}^* | \boldsymbol{\tau}^*, \mathcal{D})]^2, \end{aligned} \tag{6.9}$$

and

$$\begin{aligned} E[Var(\mathbf{z}^* \mid \boldsymbol{\tau}^*, \mathcal{D})] &= \int Var(\mathbf{z}^* \mid \boldsymbol{\tau}^*, \mathcal{D})) p(\boldsymbol{\tau}^* \mid \mathcal{D}) d\boldsymbol{\tau}^* \\ &= \int b''(\hat{\alpha}^*) c(\phi) p(\boldsymbol{\tau}^* \mid \mathcal{D}) d\boldsymbol{\tau}^* \end{aligned} \tag{6.10}$$

where $\hat{\alpha}^*$ is a function of $h(\mathcal{U}^{*T} \hat{\beta} + \boldsymbol{\tau}^*)$. Thus (6.8) and (6.10) can be also calculated by numerical integration.

In terms of information consistency of the $\mathcal{MCGPNGR}$ model, we provide a general theorem from Theorem 5.1 and Lemma D.1 as follows

Theorem 6.1. *Under the $\mathcal{MCGPNGR}$ model* (6.1) *and* (6.2) *and the condition given in Lemma 6.1, the prediction $\hat{\mathbf{z}}$ is information consistent to the true curve $\mathbf{z}_0$ if the RKHS norm $\|\tau_0\|^2_{K_{N_1N_2}}$ is bounded and the expected regret term $E_{\mathbf{X}_{N_1N_2}}(\log|\mathbf{I}+\delta\mathcal{K}_{N_1N_2}|) = o(N_1+N_2)$. The error bound is specified in Appendix D.*

The proof of the theorem is similar to that given in Appendix D.

Lemma 6.1. *Suppose z_{1i} and z_{2i} are conditionally independent samples from exponential family given* (6.1) *and $\tau_0 \in F$ has a multivariate convolved Gaussian prior with zero mean and bounded covariance function $\mathcal{K}_{N_1N_2}$ for any covariate values in $\mathcal{X}$. Suppose that $\mathcal{K}_{N_1N_2}$ is continuous in $\boldsymbol{\theta}$ and the estimator $\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta}$ almost surely as $N_1 \to \infty$ and $N_2 \to \infty$. If there exists a positive number κ such that $|b''(\alpha)| \leqslant e^{\kappa\alpha}$, then*

$$\begin{aligned} & -\log p_{mgp}(z_1,...,z_{N_1},z_{N_1+1},..,z_{N_1+N_2}) + \log p_0(z_1,...,z_{N_1},z_{N_1+1},..,z_{N_1+N_2}) \\ \leqslant \quad & \frac{1}{2}\|\tau_0\|^2_{\mathcal{K}_{N_1N_2}} + \frac{1}{2}\log|\mathbf{I}+\delta\mathcal{K}_{N_1N_2}| + C \end{aligned} \tag{6.11}$$

where $\|\tau_0\|^2_{\mathcal{K}_{N_1N_2}}$ is the reproducing kernel Hilbert space (RKHS) norm of τ_0 associated with $\mathcal{K}_{N_1N_2}$, $\mathbf{I}$ is the $(N_1+N_2)\times(N_1+N_2)$ identity matrix, δ and C are some positive constants.

Remark. The condition $|b''(\alpha)| \leqslant e^{\kappa\alpha}$ in Lemma 6.1 is satisfied for a wide range of exponential family distribution in (6.1), see Wang & Shi (2014) for the detailed discussion.

6.2 Numerical Results

In this section we demonstrate the proposed method with several examples involving comprehensive simulation studies and a real data application. Various distributions in the exponential family will be discussed, such as the bivariate binomial and ordinal distributions.

6.2.1 Bivariate Binomial Data

We use a $\mathcal{MCGPNGR}$ model with logit link function to generate bivariate Binomial data. The following equation defines a data set with a classification case,

$$\mathcal{D} = \left\{ \begin{pmatrix} z_{1i} \\ U_{1i} \\ \mathbf{x}_{1i} \end{pmatrix}; i = 1,...,N_1, \begin{pmatrix} z_{2i} \\ U_{2i} \\ \mathbf{x}_{2i} \end{pmatrix}; i = 1,...,N_2 \right\} \tag{6.12}$$

where z_{1i}, z_{2i} are observations of the two responses, U_{1i}, U_{2i} are bivariate binomial covariates and $\mathbf{x}_i = (\mathbf{x}_{1i}, \mathbf{x}_{2i})$ are P-dimensional covariates involved in covariance structure

τ_{i1}, τ_{2i}. Here, N_1 and N_2 are the sample size of each responses. A discrete multivariate binomial regression model with $\mathcal{CGP}$ priors ($\mathcal{MCGPNGR}$) is therefore given by

$$\begin{pmatrix} z_{1i}(\mathbf{x}_{1i}) \mid \tau_{1i}(\mathbf{x}_{1i}) \\ z_{2i}(\mathbf{x}_{2i}) \mid \tau_{2i}(\mathbf{x}_{2i}) \end{pmatrix} \sim \begin{pmatrix} Bin(1, \pi_{1i}), & i = 1, ..., N_1 \\ Bin(1, \pi_{2i}), & i = 1, ..., N_2 \end{pmatrix} \tag{6.13}$$

$$\begin{pmatrix} \pi_{1i} = \frac{\exp(\mathbf{U}_{1i}^T \boldsymbol{\beta}_1 + \tau_{1i}(\mathbf{x}_{1i}))}{1+\exp(\mathbf{U}_{1i}^T \boldsymbol{\beta}_1 + \tau_{1i}(\mathbf{x}_{1i}))} \\ \\ \pi_{2i} = \frac{\exp(\mathbf{U}_{2i}^T \boldsymbol{\beta}_2 + \tau_{2i}(\mathbf{x}_{2i}))}{1+\exp(\mathbf{U}_{2i}^T \boldsymbol{\beta}_2 + \tau_{2i}(\mathbf{x}_{2i}))} \end{pmatrix}$$

and

$$\begin{pmatrix} \tau_{1i}(\cdot) \\ \tau_{2i}(\cdot) \end{pmatrix} \sim \mathcal{MGP}(\mathbf{0}, \mathcal{K}(\cdot, \cdot)),$$

where $\mathcal{K}(\cdot, \cdot)$ is defined by equation (4.8) and (4.10). The density function for $\mathbf{z} = (z_{11}, ..., z_{1N_1}, z_{21}, ..., z_{2N_2})$ is given by

$$\begin{aligned} p(\mathbf{z} \mid \boldsymbol{\tau}) &= \prod_{a=1}^{2} p(\mathbf{z}_a \mid \boldsymbol{\tau}, \boldsymbol{\beta}) \\ &= \prod_{a=1}^{2} \prod_{i=1}^{N_a} p(z_{ai} \mid \tau_{ai}, \boldsymbol{\tau}) \\ &= \prod_{a=1}^{2} \prod_{i=1}^{N_a} \pi_{ai}^{z_{ai}} (1 - \pi_{ai})^{1 - z_{ai}} \end{aligned}$$

The marginal log-likelihood is calculated by equation (6.6) and the empirical Bayesian estimates of hyper-parameters ($\boldsymbol{\theta}$) and parameters ($\boldsymbol{\beta}$) can be estimated.

Scenario1

The aim of this scenario is to show the performance of proposed model by using different sample sizes. We use a discrete model in equation (6.13) as the true model and $\mathcal{K}(\cdot, \cdot)$ is given by equations (4.6) to (4.10). The true values of $\boldsymbol{\beta}$ are $\beta_{10} = 1$, $\beta_{11} = 2$, $\beta_{20} = 1$ and $\beta_{21} = 2$. In this scenario, latent variables $\boldsymbol{\tau}$ are generated similarly to scenario 1 in Chapter 4. i.e. via a mixed form covariance structure. We recall the setting of how we generate random latent variables based on equation (4.6) and also Scenario 1 in the previous chapter as follows

$$\begin{aligned} \tau_1 &= \xi_1 + \eta_1 \\ \tau_2 &= \xi_2 + \eta_2. \end{aligned} \tag{6.14}$$

The mixed form contains η_1 which is generated from a squared exponential covariance function with true value of $(w_1 = 0.04, B_1 = 1)$ and η_2 is formed from a Gamma exponential family covariance function with the true values of $(w_2 = 0.04, B_2 = 1)$. Also, the shared processes $\boldsymbol{\xi}$ follow the squared exponential covariance function with true values $(v_1 = 0.04, v_2 = 0.04, A_1 = 1, A_2 = 1)$. In this scenario, we investigate several different sample sizes, i.e. 15, 20 and 25 for each response. The generated data are equally spaced in $[-5, 5]$. The observations z_{ai} follow a binary distribution $Bin(1, \pi_{ai})$ with

$$\pi_{ai} = \frac{\exp(\mathbf{U}_{ai}^T \boldsymbol{\beta}_a + \tau_{ai}(\mathbf{x}))}{1 + \exp(\mathbf{U}_{ai}^T \boldsymbol{\beta}_a + \tau_{ai}(\mathbf{x}))}$$

where $a = 1, 2$ and $i = 1, ..., N_1$ or N_2. N_1 and N_2 are the number of first response and second response variables respectively.

Similar to Scenario 1 in the previous chapter, we generate new data as training data and test data. In order to investigate the performance of the estimated parameters, we calculate the sample mean of the estimates $\boldsymbol{\beta}$ and the root mean squared error (RMSE) between the estimated and the true values.

Table 6.1 shows the result of mean estimated parameters (sample mean) for $\boldsymbol{\beta}$ using the empirical Bayesian approach for three different sample sizes based on one hundred replications. Table 6.2 provides RMSE between estimated values and the true values for three different sample sizes based on one hundred replications.

		Sample Mean		
Parameter	True	*Sample 15*	*Sample 20*	*Sample 25*
β_{11}	1	1.00857	1.01319	1.00138
β_{12}	2	1.98954	1.99088	2.00983
β_{21}	1	1.00118	1.00519	0.99543
β_{22}	2	2.00487	2.00221	2.00326

Table 6.1: The sample mean of estimated parameters ($\boldsymbol{\beta}$) for bivariate Poisson model with Convolved GPR for three different sample size based on one hundred replications.

	RMSE		
Parameter	*Sample 15*	*Sample 20*	*Sample 25*
β_{11}	0.17885	0.16853	0.05417
β_{12}	0.21904	0.14082	0.06442
β_{21}	0.10635	0.15774	0.07657
β_{22}	0.10544	0.11192	0.07019

Table 6.2: RMSE between estimated parameters ($\boldsymbol{\beta}$) and true values for bivariate the binomial model with convolved GPR for three different sample sizes based on one hundred replications.

Table 6.1 and 6.2 show that estimated parameters are acceptable and closer to the true

values with sample size increasing as we would expect.

In order to explore the performance of the prediction, for convenience we use the estimated mean ($\hat{\boldsymbol{\mu}} = \hat{\boldsymbol{\pi}}$) of the model in equation (6.13) to predict the actual response. Later, we consider calculating the difference between $\boldsymbol{\mu}$ and $\hat{\boldsymbol{\mu}}$ by using RMSE criterion as the measurement of predictive performance. Table 6.3 shows the average values of RMSE (Average RMSE) based on one hundred replications. From Table 6.3, it is clearly seen that the performance is reasonably good and the accuracy improves as the sample size increases.

Sample Size	Average RMSE
15	0.013999
20	0.013304
25	0.013034

Table 6.3: The average values of RMSE between $\boldsymbol{\mu}$ and $\hat{\boldsymbol{\mu}}$ for bivariate Binomial model with Convolved $\mathcal{GPR}$ for three different sample size based on one hundred replications.

Scenario 2

The aim of this scenario is to investigate how sensitive the results are to the choice of covariance structures. The method of generating random processes is the same as the one used in Section 2 Chapter 4. We design several different priors for $\mathcal{MCGPNGR}$ model as following

i. Generate random processes
   The true model used to generate a multiple process is :

$$\begin{aligned} y_{1i}(\mathbf{x}_i) &= 0.8sin(0.5\mathbf{x}_i)^3 + \tau_{1i}(\mathbf{x}_i), \quad i = 1, ..., N_1 \\ y_{2i}(\mathbf{x}_i) &= 0.8cos(0.5\mathbf{x}_i)^3 + \tau_{2i}(\mathbf{x}_i) \quad i = 1, ..., N_2 \\ (\tau_{1i}(\cdot), \tau_{2i}(\cdot))^T &\sim \mathcal{MGP}(\mathbf{0}, \mathcal{K}(\cdot, \cdot)) \end{aligned}$$

   where for each response, $x_{ai}, a = 1, 2$ is equally spaced in $[-4, 4]$ and covariance function $\mathcal{K}(\cdot, \cdot)$ based on equations (4.8) and (4.10) in Chapter 4. For this scenario, we set the covariance structure (mixed form) and initial values the same as the ones used in the first scenario.

   We recall that the structure of the covariance function is based on equation (6.14), i.e. ξ_1, ξ_2 are generated from squared exponential covariance functions with the true values ($v_1 = 0.04, A_1 = 1$) and ($v_2 = 0.04, A_2 = 1$) respectively. Meanwhile, η_1 has a squared exponential covariance function with the true values ($w_1 = 0.04, B_1 = 1$) and η_2 has a Gamma exponential covariance function with the true values ($w_2 = 0.04, B_2 = 1$).

In order to generate random bivariate binomial data, we first calculate the mean of each response as follows :

$$\begin{pmatrix} \pi_{1i} = \frac{\exp(y_{1i})}{1+\exp(y_{1i})} \\ \pi_{2i} = \frac{\exp(y_{2i})}{1+\exp(y_{2i})} \end{pmatrix}$$

and

$$\mathbf{z} \mid \boldsymbol{y} = \begin{pmatrix} z_{1i} \mid y_{1i} \\ z_{2i} \mid y_{2i} \end{pmatrix} \sim \begin{pmatrix} Bin(1, \pi_{1i}), & i = 1, ..., N_1 \\ Bin(1, \pi_{2i}), & i = 1, ..., N_2 \end{pmatrix}.$$

A sample containing 20 data observations for each component is generated and used as training data. We also generate another 20 observations as test data to investigate the performance of the prediction.

ii. Model comparison Now, we use four different priors of $\mathcal{MCGPNGR}$ and equation (5.21) ($\mathcal{CD}$ method). The covariance structure is given by equation (6.14) with the following specific assumptions :

1. *Model 1*
   This model is similar to the true model i.e. ξ_1, ξ_2 and η_1 have a squared exponential covariance function while η_2 has a Gamma exponential covariance function
2. *Model 2*
   Similar to *Model 1* but η_2 has a rational quadratic covariance function
3. *Model 3*
   Similar to *Model 1* but η_2 has a Matern covariance function
4. *Model 4*
   η_2 has a squared exponential covariance function while the rest of processes are the same as *Model 1*
5. *Model 5*
   We compare the proposed model with existed model, i.e. $\mathcal{CD}$ approach based on equation (5.21).

Figure 6.1 and 6.2 show a comparison between prediction means of underlying process $\mathbf{y}$ by using different models which compare the true mean and the prediction mean using the $\mathcal{CD}$ approach. The average of RMSE (Average RMSE) is also calculated based on one hundred replications and is given in Table 6.4.

In Figure 6.1 and Figure 6.2, there are some points worth noting. Although the predicted mean of *Model 5* is acceptable, it does not do better than the other models, i.e. *Model 1* which provides the smallest difference between predicted mean and mean true curve.

From Table 6.4, *Model 1* is the best as we would expect. *Models 2* to *4* are not the best choices but their results are very close to the best model and perform much better

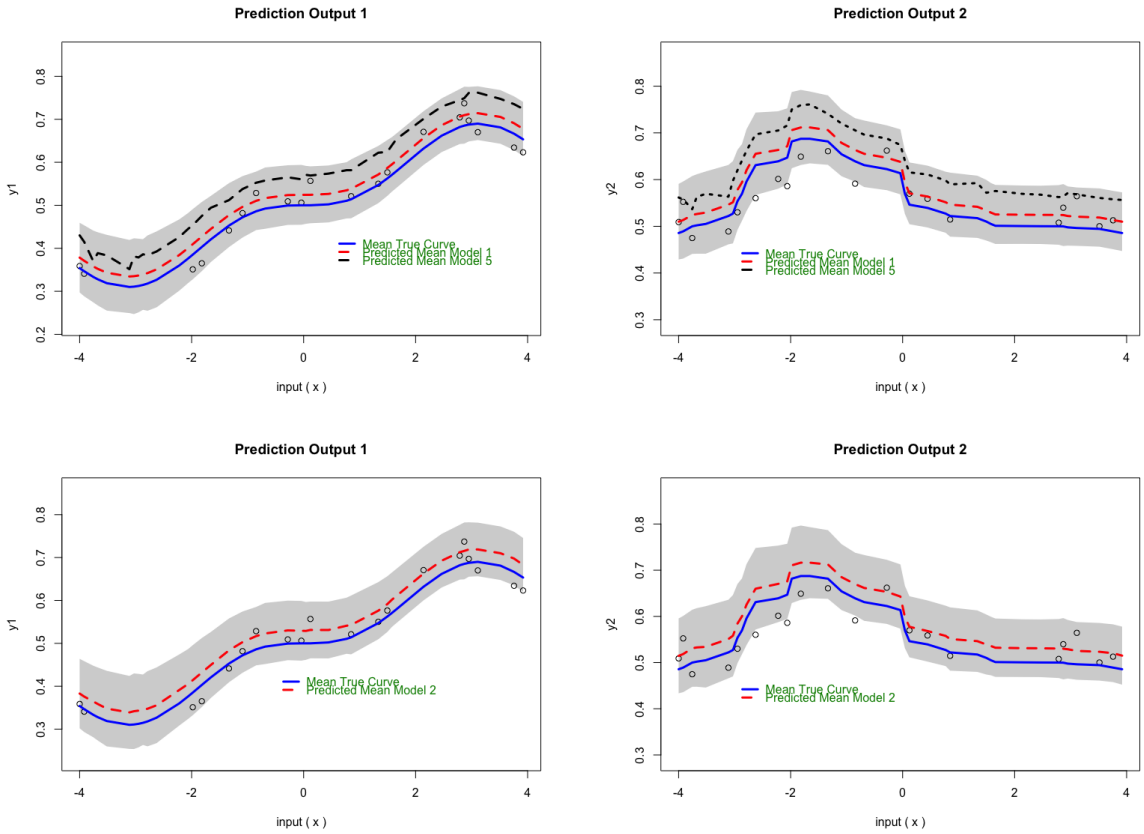

Figure 6.1: Example 1 : The predictions from two strongly dependent bivariate binomial data with three different models (*Model 1*, *Model 2* and *Model 5*). The dots are test data (sample), the red dashed lines represent the predictions using three different covariance functions and the blue solid lines are the true curves with 95% confidence intervals (the shaded regions)

Model	Average RMSE
Model 1	0.0267220
Model 2	0.0268342
Model 3	0.0268087
Model 4	0.0267539
Model 5	0.0468069

Table 6.4: Average RMSE prediction between μ and $\hat{\mu}$ of bivariate binomial data from various models based on one hundred replications.

than *Model 5* although we used misspecified covariance functions in those models. This shows the robustness of our approach.

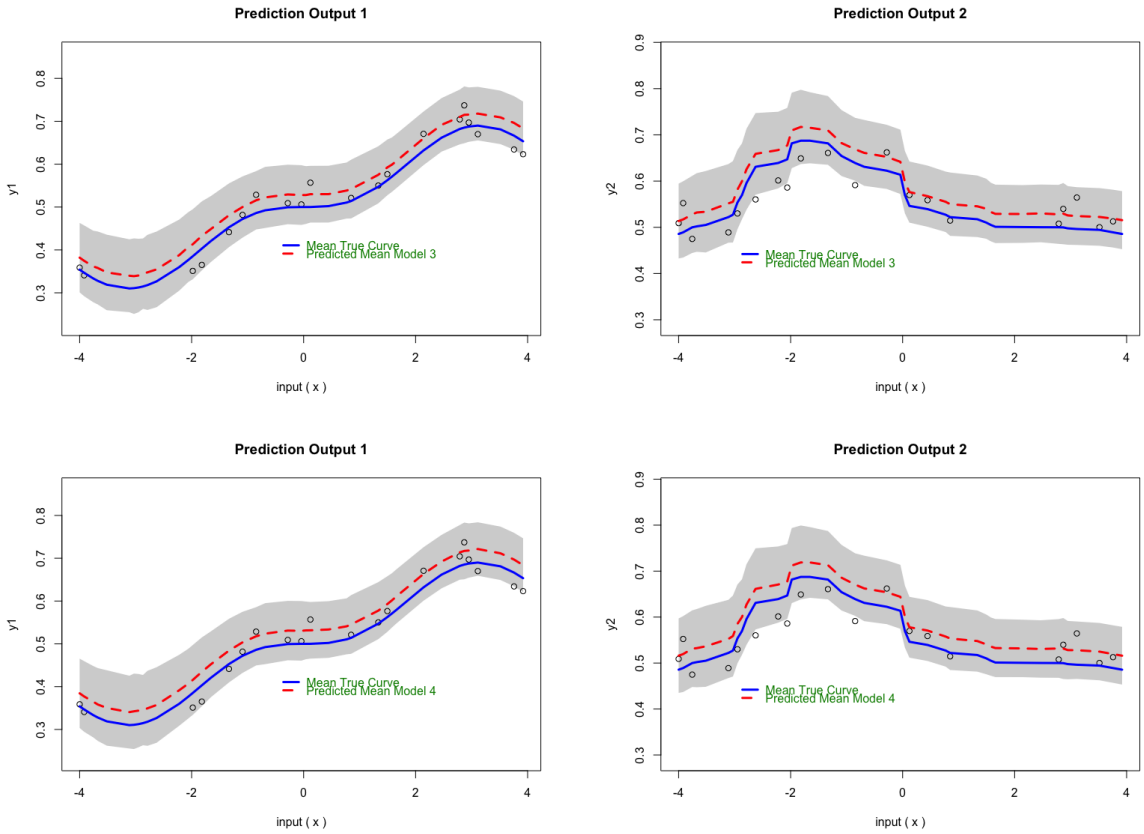

Figure 6.2: Example 2 : The predictions from two strongly dependent bivariate binomial data with two different models (*Model 3* and *Model 4*). The dots are test data (sample), the red dashed lines represent the predictions using two different covariance functions and the blue solid lines are the true curves with 95% confidence intervals (the shaded regions)

6.2.2 Bivariate Ordinal Data

We further demonstrate the proposed method using ordinal data. We use a $\mathcal{MCGPNGR}$ model with probit link function. Similar to bivariate binomial data, we redefine the procedure. The observed data are

$$\mathcal{D} = \left\{ \begin{pmatrix} z_{1i} \\ U_{1i} \\ \mathbf{x}_{1i} \end{pmatrix}; i = 1, ..., N_1, \begin{pmatrix} z_{2i} \\ U_{2i} \\ \mathbf{x}_{2i} \end{pmatrix}; i = 1, ..., N_2 \right\} \tag{6.15}$$

where z_{1i}, z_{2i} are observations of the two responses, U_{1i}, U_{2i} are bivariate ordinal covariates and $\mathbf{x}_{1i}, \mathbf{x}_{2i}$ are P-dimensional covariates involved in the covariance structure for τ_{i1}, τ_{2i}. Here, N_1 and N_2 are the sample size of each responses. A discrete multivariate ordinal

regression model with $\mathcal{CGP}$ priors ($\mathcal{MCGPNGR}$) with r ordered categories is given by

$$\begin{pmatrix} z_{1i}(\mathbf{x}_{1i}) \mid y_{1i}(\mathbf{x}_{1i}) \\ z_{2i}(\mathbf{x}_{2i}) \mid y_{2i}(\mathbf{x}_{2i}) \end{pmatrix} = \begin{pmatrix} j & if & b_j < y_{1i}(\mathbf{x}_{1i}) \leqslant b_{j+1} \\ j & if & b_j < y_{2i}(\mathbf{x}_{2i}) \leqslant b_{j+1} \end{pmatrix} \tag{6.16}$$

$$\begin{pmatrix} y_{1i}(\mathbf{x}_{1i}) = \mathbf{U}_{2i}^T \boldsymbol{\beta}_1 + \tau_{1i}(\mathbf{x}_{1i}) \\ y_{2i}(\mathbf{x}_{2i}) = \mathbf{U}_{2i}^T \boldsymbol{\beta}_2 + \tau_{2i}(\mathbf{x}_{2i}) \end{pmatrix}$$

and

$$\begin{pmatrix} \tau_{1i}(\cdot) \\ \tau_{2i}(\cdot) \end{pmatrix} \sim \mathcal{MGP}(\mathbf{0}, \mathcal{K}(\cdot,\cdot)),$$

where $i = 1, ..., N_1$ or N_2, $b_0 = -\infty$ $b_r = \infty$ and b_j for $j = 1, ..., r-1$ are the thresholds to be estimated. Meanwhile, $\mathcal{K}(\cdot,\cdot))$ is defined by equations (4.8) to (4.10). The density function for $\mathbf{z} = (z_{11}, .., z_{1N_1}, z_{21}, ..., z_{2N_2})$ is given by

$$\begin{aligned} p(\mathbf{z} \mid \boldsymbol{y}) &= \prod_{a=1}^{2} \prod_{i=1}^{N_a} p(z_{ai} \mid y_{ai}) \\ &= \prod_{a=1}^{2} \prod_{i=1}^{N_a} p(b_{z_{ai}} < y_{ai} < b_{z_{ai+1}}) \end{aligned}$$

The marginal log-likelihood is calculated by (6.6). The empirical Bayesian estimates of $\boldsymbol{\beta}$, the hyper-parameter and the thresholds are estimated by maximizing the marginal likelihood.

Scenario 1

The true model used to generate the latent process is based on a discete model in equation (6.16) and is specified as follows :

$$\begin{aligned} y_{ai}(\mathbf{x}_{ai}) &= U_{ai}^T \beta(\mathbf{x}_{ai}) + \tau_{ai}(\mathbf{x}_{ai}), \quad a = 1, 2 \quad i = 1, ..., N_a \\ z_{ai}(\mathbf{x}_{ai}) &= \begin{cases} 0 & if \quad y_{ai}(\mathbf{x}_{ai}) \leq 2 \\ 1 & if \quad 2 < y_{ai}(\mathbf{x}_{ai}) \leq 3 \\ 2 & if \quad y_{ai}(\mathbf{x}_{ai}) > 3 \end{cases} \end{aligned}$$

and

$$\begin{pmatrix} \tau_{1i}(\cdot) \\ \tau_{2i}(\cdot) \end{pmatrix} \sim \mathcal{MGP}(\mathbf{0}, \mathcal{K}(\cdot,\cdot)).$$

The random data of bivariate ordinal type are generated by using the true values of $\beta_{10} = 1, \beta_{11} = 2, \beta_{20} = 1$ and $\beta_{21} = 2$. Latent variables $\boldsymbol{\tau}$ are generated using the same procedure described in subsection 6.2.1. The training data and test data are generated

separately. Each response contains 20 observations and is equally spaced in $[-5, 5]$. In this scenario, we assume $r = 2$ and thresholds b_1 and b_2 are unknown parameters. We compare two different priors of $\mathcal{MCGPNGR}$, i.e. *Model 1* and *Model 4* with $\mathcal{CD}$ approach (*Model 5*) defined in Scenario 2 in subsection 6.2.1.

Table 6.5 shows the result of mean estimated parameters (sample mean) and the value of RMSE (RMSE) for $\boldsymbol{\beta}$ using empirical Bayesian approach for three different models based on one hundred replications. Table 6.6 shows the results for estimated threshold and the value of RMSE's (RMSE).

	Sample Mean				RMSE			
Parameter	β_{10}	β_{11}	β_{20}	β_{21}	β_{10}	β_{11}	β_{20}	β_{21}
True Values	1	2	1	2				
Model 1	1.003723	2.000646	1.003581	2.000532	0.003842	0.000734	0.003709	0.000637
Model 4	1.003756	2.000696	1.003606	2.000548	0.003886	0.000786	0.003707	0.000639
Model 5	1.021499	1.997239	1.011650	2.001238	0.077500	0.010801	0.041669	0.004394

Table 6.5: The sample mean and the value of RMSE (RMSE) of estimated parameters ($\boldsymbol{\beta}$) for the bivariate ordinal model with several models based on one hundred replications.

	Sample Mean		RMSE	
Parameter	b_1	b_2	b_1	b_2
True Values	2	3		
Model 1	1.98107	3.01162	0.01894	0.011634
Model 4	1.98091	3.01172	0.01909	0.011733
Model 5	1.97604	3.06227	0.07645	0.224627

Table 6.6: The sample mean of estimated thresholds ($\boldsymbol{b}$) and the value of RMSE (RMSE) for bivariate ordinal data with several models based on one hundred replications.

It is not surprising that the result of *Model 1* has the best performance since *Model 1* uses the same covariance structure as the true model. However, *Model 4* also offers a promising result although the covariance structure is different to the true model. We can conclude that the proposed model provides good robustness in the choice of covariance function.

Scenario 2

We further consider all the models defined in Scenario 2 in subsection 6.2.1, where *Model 5* is based on $\mathcal{CD}$ approach. Table 6.7 shows the average of RMSE (average RMSE) of the prediction of $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$ based on one hundred replictions. From Table 6.7, three models from *Model 2* to *Model 4* show good performance although they can not be better than *Model 1*. It means that there is great robustness in $\mathcal{MCGPNGR}$. Meanwhile, similar

Model	Average RMSE
Model 1	0.001105
Model 2	0.001905
Model 3	0.001905
Model 4	0.001120
Model 5	0.003864

Table 6.7: Average RMSE prediction between $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$ of bivariate ordinal data from various models based on one hundred replications.

again to other scenarios, *Model 5* fails.

Scenario 3

In the third scenario, we show another promising feature of the proposed model, i.e. the capacity to accommodate large dimensional covariates in the covariance function. We design the true model as follows.

i. Generate random processes
The true model used to generate multiple processes is follows

$$\begin{aligned} y_{1i}(\mathbf{x}_i) &= 0.2\mathbf{x}_{1i} \times |\mathbf{x}_{1i}|^{\frac{1}{4}} + \frac{1}{(\mathbf{x}_{2i})} + \tau_{1i}(\mathbf{x}_i) \\ y_{2i}(\mathbf{x}_i) &= \frac{1}{1+\exp(-1.5 \times \mathbf{x}_{1i})} + 0.01\mathbf{x}_{2i} \times |\mathbf{x}_2 i|^{\frac{1}{4}} + \tau_{2i}(\mathbf{x}_i) \\ i &= 1, ..., N_1 \quad or \quad N_2 \end{aligned}$$

where $(\tau_{1i}(\cdot), \tau_{2i}(\cdot)^T \sim \mathcal{MGP}(\mathbf{0}, \mathcal{K}(\cdot,\cdot))$, where is $\mathcal{K}(\cdot,\cdot)$ defined by equation (4.10) and (4.8). Meanwhile, the latent variables $\boldsymbol{\tau}$ are generated the same as *Model 1* via a mixed form covariance function including squared exponential and Gamma exponential covariance functions.

We recall the setting of the covariance structure as follows. Here, η_1 has a squared exponential covariance function with the true values $(w_1 = 0.04, B_1 = 1)$, ξ_1 and ξ_2 are generated also from squared exponential covariance functions with the true values $v_1 = 0.04, A_1 = 1$ and $v_2 = 0.04, A_2 = 1$ respectively. But η_2 are formed from a Gamma exponential covariance function with true values $(v_2 = 0.04, A_2 = 1)$. Each generated process of training data contains 20 observations where $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2)$ are equally spaced in $[1, 5]$ and $[1, 4]$ respectively. Further we generate the same number of observations as test data.

ii. Generate random bivariate ordinal data

In order to generate random bivariate ordinal data. We use the following formula :

$$z_{ai}(\mathbf{x}_i) = \begin{cases} 0 & if \quad y_{ai}(\mathbf{x}_i) \leq 0.2 \\ 1 & if \quad 0.2 < y_{ai}(\mathbf{x}_i) \leq 0.7 \\ 2 & if \quad y_{ai}(\mathbf{x}_i) > 0.7 \end{cases}$$

where $a = 1, 2$ and $i = 1, .., N_1$ or N_2.

We consider two different models, i.e. *Model 1* assuming similar a covariance structure to the true model and *Model 5*, a comparison model formed based on equation (5.21).

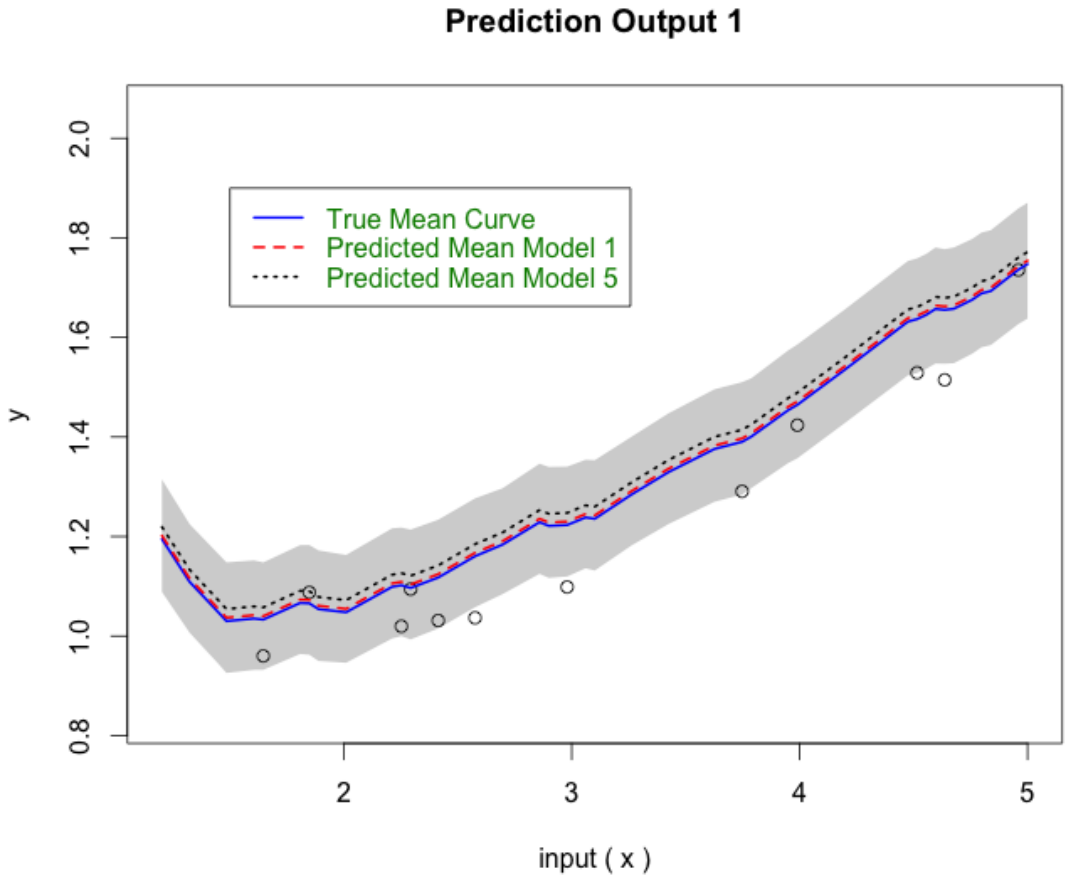

Figure 6.3: Example 1: The predictions of the first outputs of bivariate ordinal regression where the dots are test data (sample), the red and black dashed lines represent the predicted mean using *Model 1* and *Model 5* respectively. The blue solid lines are the true mean curve with 95% confidence intervals (the shaded regions).

Figures 6.3 to 6.4 show the performance of the prediction mean and the true mean curve. Table 6.8 shows the results of average values of RMSE (Average RMSE) between the estimated $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$ based on one hundred replications. From those results, there are

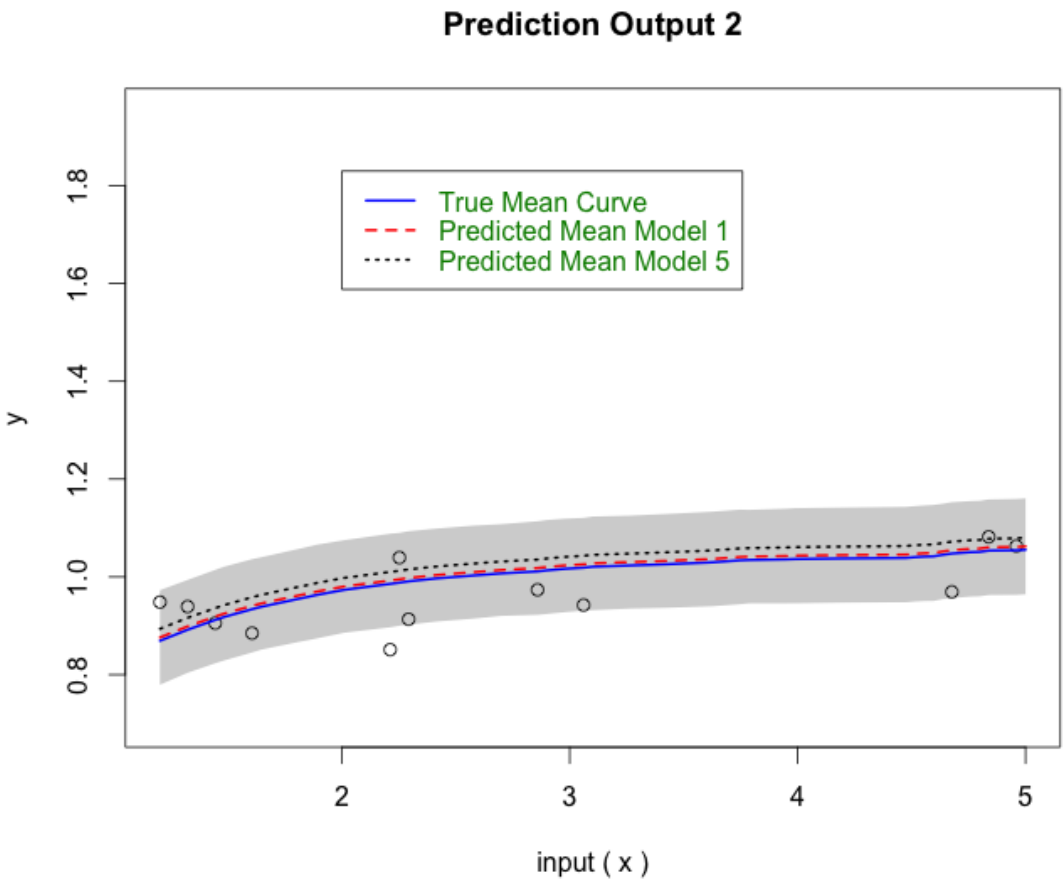

Figure 6.4: Example 2: The predictions of the second outputs of bivariate ordinal regression where the dots are test data (sample), the red and black dashed lines represent the predicted mean using *Model 1* and *Model 5* respectively. The blue solid lines are the true mean curve with 95% confidence intervals (the shaded regions).

Model	Average RMSE
Model 1	0.0007813
Model 5	0.0012219

Table 6.8: The average of RMSE (Average RMSE) from prediction between $\boldsymbol{y}$ and $\hat{\boldsymbol{y}}$ with two different models based on one hundred replications.

some findings that we can highlight. *Model 1* has the best performance as we expect. We can also say that our proposed models are better than *Model 5* for handling multi-dimensional covariates in the covariate structure.

6.2.3 Adverse Birth Outcome Data

We now consider the example of bivariate binomial data. This application involves the analysis of the period of gestation which is the most important indicator to determine an infants health. We consider two important measures : the length of pregnancy and the birth weight.

The normal term of gestation is around 37 weeks to 41 weeks, while preterm birth means a live birth before 37 completed weeks of gestation. The results of a preterm birth are more likely to include complications such as acute respiratory, gastrointestinal, immunologic, and central nervous system problems. Meanwhile, birth weight also influences an infant's health for years to come. A live birth which has a weight of less than 2500 grammes is considered a low birth weight. It is also more likely to be associated with a range of neuro-developmental disorders, underdeveloped lungs, breathing and heart.

Neelon *et al.* (2014) attempted to analyse the problem between preterm and low birth weight in North Carolina, USA using a bivariate conditional autoregressive model. The data is taken from the North Carolina Center for Health Statistics from 2007 to 2008. Also, they considered demographic and socioeconomic information. Table 6.9 shows summaries of Adverse Birth Outcome data in North Carolina, USA in 2007-2008

Variable (cases)	Min	Max
Low birth weight	3.0	1373.0
Preterm	49.0	14050.0

Table 6.9: The summaries of Adverse Birth Outcome data in North Carolina, USA in 2007-2008.

In most cases, the low birth weight is caused by the preterm birth. It is therefore better to model them as dependent multivariate response variables. This motivated us to use $\mathcal{MCGPNGR}$ model. The model can be defined as following

$$z_{ai} \sim Bin(n_{ai}, \pi_{ai}), \quad i = 1, ..., 100 \quad a = 1, 2. \tag{6.17}$$

and

$$\pi_{ai} = \frac{\exp(\tau_{ai}(\mathbf{x}_i))}{1 + \exp(\tau_{ai}(\mathbf{x}_i))}$$

where z_{1i} denotes the numbers of positive indications of a preterm birth in county i and z_{2i} denotes the numbers of positive indications of a low birth weight in county i, and n_{1i} and n_{2i} denote the corresponding numbers of people giving birth in county i. Whereas, $\mathbf{x}_i$ are defined from spatial point values of latitude and longitude of each county. Also, $\boldsymbol{\tau} = \{\tau_{1i}, \tau_{2i}\}$ are dependent Gaussian processes with zero mean and covariance function based on equations (4.8) to (4.10).

We select two thirds of the whole data set randomly as training data and the remaining

as test data. We estimate hyper-parameters via an empirical Bayesian approach using training data. The estimated hyper-parameters are used to predict the test data. We then calculate the average of relative error (average RE) for measuring predictive performance with ten replications. We also compare the proposed method with $\mathcal{CD}$ approach.

Table 6.10 shows the average error rate for the predicted values and the actual observations. The performance of $\mathcal{MCGPNGR}$ is much better than $\mathcal{CD}$ approach.

Model	*Average RE*
$\mathcal{MCGPNGR}$	0.00170
$\mathcal{CD}$	0.00448

Table 6.10: The average value of relative error (average RE) from prediction between the predicted values and the actual observations with two different models based on ten replications.

6.3 Chapter Summary

In this chapter, we proposed a $\mathcal{MCGPNGR}$ for multivariate dependent non-Gaussian data. The general framework is able to capture the relationship between multidimensional covariates and multivariate non-Gaussian dependent data. The novel approach enables us to handle cross-correlation between response variables which can be defined from convolved Gaussian process priors.

We provided a natural framework on how to use $\mathcal{CGP}$ to define multivariate generalized semiparametric regression analysis for response variable from exponential families. We offered a procedure of inference and its implementation and also an asymptotic theory based on information consistency. $\mathcal{MCGPNGR}$ assumes that the response variable follows a distribution from the exponential family. Therefore, it is easy to apply them to other distributions, although the examples reported only include binomial data, ordinal data and Poisson data.

One of the important features of the $\mathcal{CGP}$ method is that it can deal with multidimensional covariates with a known covariance kernel. A promising result has been provided and the results are consistently better than $\mathcal{CD}$ method.

Chapter 7

Conclusion and Future Work

In this thesis, we started the investigation with a univariate model for count data using a $\mathcal{GPR}$ since this method is more flexible in modelling the correlated structure than the existing ICAR model. However, if the data set is dependent, such as in the cases of dengue fever and malaria, it might be better if we use multivariate analysis. Constructing the cross-correlation structure is one of the biggest issues in dependent data analysis since we need to ensure that the correlation matrix is positive definite. Therefore, we proposed using convolved Gaussian process ($\mathcal{CGP}$) priors and implemented this method to analysis multivariate non-Gaussian data. This is the main purpose of this thesis.

An existing method has been developed by Crainiceanu *et al.* (2012). This method provided a natural framework to smooth dependent binomial data using a stationary Gaussian process. Unfortunately, this approach has failed to cover individual characteristics of each response because the conditional dependent structure is not allowed to capture the features. A convolution method is an alternative way to address this issue since it provides huge flexibility and robustness. Andriluka *et al.* (2006) have also investigated a framework of constructing general convolved Gaussian processes for a stationary Gaussian process. We extended the idea with mixed covariance functions to build the cross-correlation between two variable responses.

We began the discussion with a multivariate nonlinear model using $\mathcal{CGP}$ priors. We used mixture covariance functions to solve a complex problem. We have also determined that the proposed model is very robust since good performance is consistent even when there is a misspecification of the covariance structure. Meanwhile the comparison model, i.e. conditionally dependent approach in Crainiceanu *et al.* (2012), can not beat the achievement of the proposed model.

In this thesis, we extended the idea to multivariate non-Gaussian data, which are provided in Chapter 5 focusing on mutivariate Poisson data. Similar to the Gaussian data, the performance of the proposed ($\mathcal{MCGPPR}$) model offered a robust model as we

also expected. It means that the proposed model provides good result consistently even though we set a misspecification of the covariance structure. Another advantage is that using the specified covariance functions as priors also enables us to accommodate a large dimensionality of covariates in the covariance matrix. We further extend the model to cover the distributions in the exponential family in Chapter 6.

We offered a general framework model to use $\mathcal{CGP}$ priors in multivariate semiparameteric regression analysis for response variables from Gaussian and non-Gaussian distributions in the exponential family of distributions. The model, and its implementation, including the technical details of the inference, are provided. We also reported asymptotic theory based on information consistency of the general model for multivariate non-Gaussian. The natural general extension can be applied for any distribution which assumes data from the exponential family distribution. Comprehensive simulation studies and applications with real data are also explored. The performance of the proposed models are usually better than other methods and show good flexibility and robustness.

Related to the topics discussed in this thesis, some interesting problems are worth further attention. For example, large dimensional integration. We used Laplace approximation. It provides reasonably good results. However, the approximation error will increase when the dimension increases (bear in mind that the dimension of the integration is equal to the sample size). We need to develop more efficient algorithms; see e.g Wang & Shi (2014).

We also encounter restriction in terms of building mixture covariance functions. It seems that not every covariance function is integrable; thus, here we just focused on some stationary covariance functions, such as exponential squared, rational quadratic, Matern and Gamma exponential covariance functions. Further investigation is needed to provide a very flexible model although the proposed models have offered a good performance. We focused on bivariate response variables in the thesis. Although there is no difficulty in extending the method to the multivariate case in theory, the implementation may be challenging. More research in this area is essential.

Appendices

Appendix A

A solution of the symmetry problem in the covariance matrix of the CAR model

We can rely on Brook's Lemma to derive a unique joint distribution from the full conditionals. For $\mathbf{y_0} = (y_{10}, ..., y_{n0})'$ any fixed point in the support of p, Brook's Lemma informs us :

$$p(y_1, ..., y_n) = \frac{P(y_1|y_2, ..., y_n)}{P(y_{10}|y_2, ..., y_n)} \frac{P(y_2|y_{10}, y_3, ..., y_n)}{P(y_{20}|y_{10}, y_3, ..., y_n)} ... \frac{P(y_n|y_{10}, ..., y_{n-1}, 0)}{P(y_{n0}|y_{10}, ..., y_{n-1}, 0)} P(y_{10}, ..., y_{n0}).$$

To show the implementation of Brook's Lemma in CAR model, first, notice that the product or difference of two gaussian density function can be easily derived :

$$\frac{1}{\sqrt{2\pi}\sigma} exp\left\{\frac{-1}{2\sigma^2}(x-\mu)^2\right\} \times \frac{1}{\sqrt{2\pi}\tau} exp\left\{\frac{-1}{2\tau^2}(y-\nu)^2\right\}$$

$$= \frac{1}{2\pi\sigma\tau} exp\left\{\frac{-1}{2\sigma^2}(x-\mu)^2 - \frac{1}{2\tau^2}(y-\nu)^2\right\}$$

and we can easily find :

$$\frac{1}{\sqrt{2\pi}\sigma} exp\left\{\frac{-1}{2\sigma^2}(y-\mu)^2\right\} \div \frac{-1}{\sqrt{2\pi}\tau} exp\left\{\frac{-1}{2\tau^2}(y-\nu)^2\right\}$$

$$= \frac{1}{2\pi\sigma\tau} exp\left\{\frac{-1}{2\sigma^2}(y-\mu)^2 + \frac{1}{2\tau^2}(x-\nu)^2\right\}$$

Since we are going to take τ and σ to be known, the determination of the product comes down to working with the terms in the exponential, which are of the form :

$$\frac{-1}{2}\left\{\frac{(x-\mu)^2}{\sigma^2} \pm \frac{(y-\nu)^2}{\tau^2}\right\}.$$

We now apply Brook's Lemma. Let's work with case where we have $n = 3$ for convinience. Also, for convenience, let us choose $y_{i,0} = 0$ for $i = 1, 2, 3$. Using what we know about Gaussian densities, we know it is good enough to determine the term in the exponential if all of the parameters are known. Therefore, to determine

$$\frac{P(y_1, y_2, y_3)}{P(y_1 = 0, y_2 = 0, y_3 = 0)}$$

all we have to do is determine what term are in the exponential and what signs they have. The term in the exponential is :

$$\frac{-1}{2}\left\{\frac{(y_1 - b_{12}y_2 - b_{13}y_3)^2}{\tau_1^2} - \frac{(b_{12}y_2 + b_{13}y_3)^2}{\tau_1^2} + \frac{(y_2 - b_{23}y_3)^2}{\tau_2^2} - \frac{(b_{23}y_3)^2}{\tau_2^2} + \frac{(y_3)^2}{\tau_3^2}\right\}$$

$$= \frac{-1}{2}\left(\frac{y_1^2}{\tau_1^2} - \frac{2y_1(b_{12}y_2 + b_{12}y_3)}{\tau_1^2} + \frac{y_2^2}{\tau_2^2} - \frac{2y_2(b_{23}y_3)}{\tau_2^2} + \frac{y_3^2}{\tau 3^2}\right).$$

This suggest that we can rewrite this a quadratic form :

$$= \frac{-1}{2}\begin{bmatrix} y_1 & y_2 & y_3 \end{bmatrix}\begin{bmatrix} \tau_1^{-2} & c_{12} & c_{13} \\ c_{12} & \tau_2^{-2} & c_{23} \\ c_{13} & c_{23} & \tau_3^{-2} \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}.$$

We get :

$$c_{12} = \frac{b_{12}}{\tau_1^2}; c_{13} = \frac{b_{13}}{\tau_1^2}; c_{23} = \frac{b_{23}}{\tau_2^2}.$$

Since we could have performed Brooks Lemma in any order, we necessarily get that : $\frac{b_{ij}}{\tau_i^2} = \frac{b_{ji}}{\tau_j^2}$ if we are indeed to get a compatible joint distribution.

Appendix B

Proof of Proposition 4.3.1

The proof is a application from Schoenberg's theorem which states on Hilbert space, there is positive definite function, such as

$$\mathcal{S}(m) = \int_0^\infty \exp(-m^2 s) d\mathcal{H}(s)$$

where $\mathcal{H}(\cdot)$ is non-decreasing, bounded and $s \geq 0$. We can express $\mathcal{S}(\sqrt{Q_a(\mathbf{d})})$ as

$$\mathcal{S}(\sqrt{Q_{ab}(\mathbf{d})}) = \int_0^\infty \exp(-Q_{ab}(\mathbf{d}) s) d\mathcal{H}(s). \tag{B.1}$$

Meanwhile, we have a convolved Gaussian process with covariance function $k_{ab}(\mathbf{x}_i, \mathbf{x}_j)$ which can be constructed by convolving gaussian white noise and a smoothing kernel as follows

$$k_{ab}(\mathbf{x}_i, \mathbf{x}_j) = \int k_a \mathbf{x}_i(\alpha) k_b \mathbf{x}_j(\alpha) d\alpha \tag{B.2}$$

where $k_{\mathbf{x}}$ is a kernel function centered at $\mathbf{x}$ and $\mathbf{x}_i, \mathbf{x}_j$ and α are location in R^2. Now, let us assume a set of data $D = ((f_{1i}, \mathbf{x}_{1i}), i = 1, ..., N; (f_{2i}, \mathbf{x}_{2i}), i = 1, ..., N)$ where $(\mathbf{x}_{1i}, \mathbf{x}_{2i}) \in T \subset R^P$ or we can also consider $\mathcal{D} = (\mathbf{f}_a, \mathbf{x}_a)$. We write t_a and $\mathbf{x}_a$ for the task and the input of $\mathbf{f}_a$. Therefore, we need to show that the covariance function is positive

definite in every Euclidean space, $R^p, p = 1, 2, ...$ as follows

$$\begin{aligned}
\sum_{i=1}^{N}\sum_{j=1}^{N} a_i a_j k(\mathbf{x}_i, \mathbf{x}_j) &= \sum_{i=1}^{N}\sum_{j=1}^{N} a_i a_j \int k_{t_a\mathbf{x}_i}(\alpha) k_{t_a\mathbf{x}_j}(\alpha) d\alpha \\
&= \int \sum_{i=1}^{N}\sum_{j=1}^{N} a_i a_j k_{t_a\mathbf{x}_i}(\alpha) k_{t_a\mathbf{x}_j}(\alpha) d\alpha \\
&= \int \sum_{i=1}^{N} a_i k_{t_a\mathbf{x}_i}(\alpha) \sum_{j=1}^{N} a_j k_{t_a\mathbf{x}_j}(\alpha) d\alpha \\
&= \int \left(\sum_{i=1}^{N} a_i k_{t_a\mathbf{x}_i}(\alpha) d\alpha \right)^2 \\
&\geq 0.
\end{aligned}$$

Hence, it is clear that covariance function is positive definite because the value of integral is non negative. In term of the definition of convolution, Boyle (2005) and Shi & Choi (2011) have provided great details and proofs. First, let us consider $k_a(\mathbf{x}) = v \exp(-\frac{1}{2}(\mathbf{x} - \mu)^T \mathbf{A}_a(\mathbf{x} - \mu))$ and $k_{ab}(\mathbf{x}_i, \mathbf{x}_j)$ can be defined as

$$k_{ab}(\mathbf{x}_i, \mathbf{x}_j) = \pi^{\frac{P}{2}} v_a v_b |\mathbf{A}_a + \mathbf{A}_b|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2}(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{\Sigma} (\mathbf{x}_i - \mathbf{x}_j) \right\}, \tag{B.3}$$

where $\boldsymbol{\Sigma} = \mathbf{A}_a(\mathbf{A}_a + \mathbf{A}_b)^{-1}\mathbf{A}_b$. Therefore, we can say that equation (B.1) can be defined based on (B.3) and (B.2) as

$$\begin{aligned}
\mathcal{S}(\sqrt{Q_{ab}(\mathbf{d})}) &= \int_0^\infty \exp\left\{ -\frac{1}{2}(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{\Sigma} (\mathbf{x}_i - \mathbf{x}_j) s \right\} d\mathcal{H}(s) \\
&= \int_0^\infty \int k_a^s \mathbf{x}_i(\alpha) k_b^s \mathbf{x}_j(\alpha) d\alpha d\mathcal{H}(s).
\end{aligned} \tag{B.4}$$

It has been shown that the value of integral is non negative value since $s \leq 0$. Hence, the covariance function of $k_{ab}(\mathbf{d})$ is positive definite. As a result, covariance function $\mathcal{K}$ based on equation (4.8) is also positive definite.

Appendix C

Convolved Covariance Functions

C.1 Gamma Exponential Covariance Function

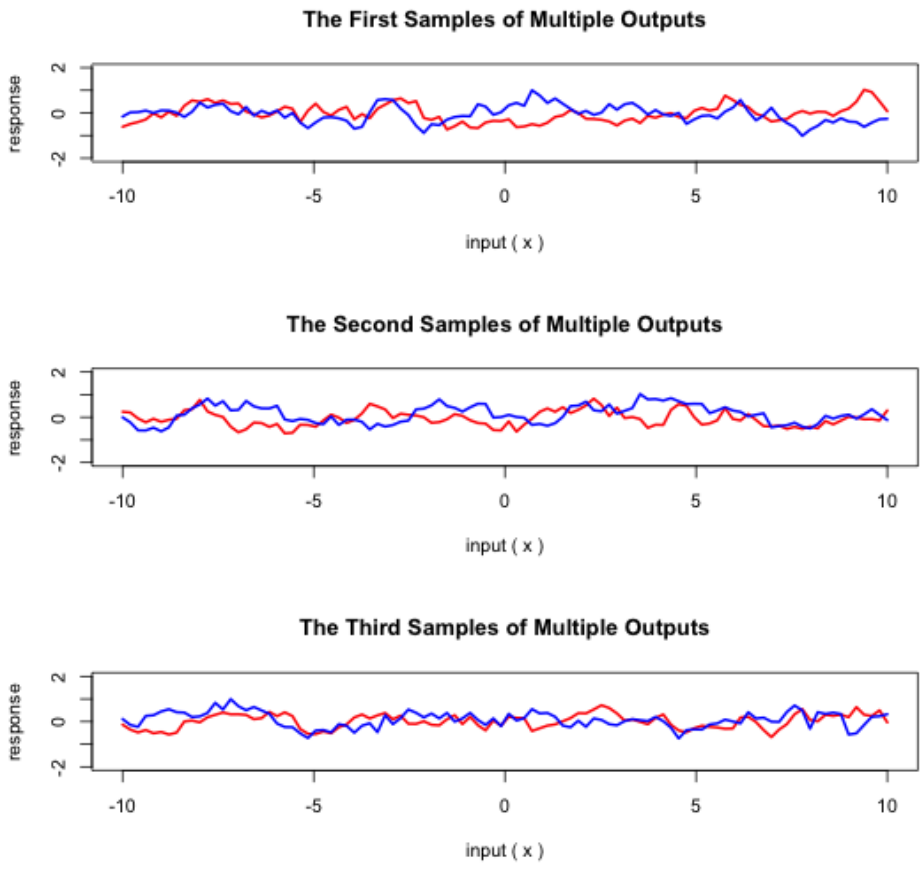

Figure C.1: Multiple dependent output process samples generated from a Gamma exponential covariance function ($\gamma = 1$) as convolved kernels. Red curves are defined as f_1 and blue curves are f_2.

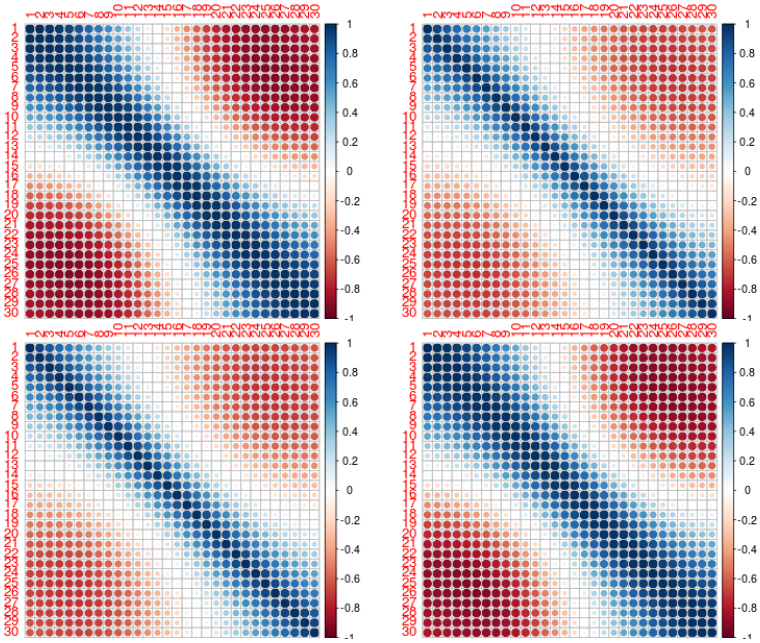

Figure C.2: Correlation map of multiple dependent Gaussian processes using a Gamma exponential as convolved kernels between f_1 and f_2; top left: correlation map of $\mathbf{f}_1$ at 30 equally spaced points; top right: correlation map between $\mathbf{f}_1$ and $\mathbf{f}_2$; and bottom right: correlation map of $\mathbf{f}_2$

Parameters	True value	Average Mean	Average RMSE
v_1	0.04	-0.000015	0.04000
v_2	0.04	-0.000015	0.04000
A_1	1	1.000257	0.00025
A_2	1	1.000257	0.00025
w_1	0.04	0.000099	0.03999
w_2	0.04	0.000099	0.03999
B_1	1	1.000427	0.00042
B_2	1	1.000427	0.00042

Table C.1: Average RMSE of the estimated parameters from multiple Gaussian process with Gamma exponential covariance function with $\gamma = 1$ based on 100 replications.

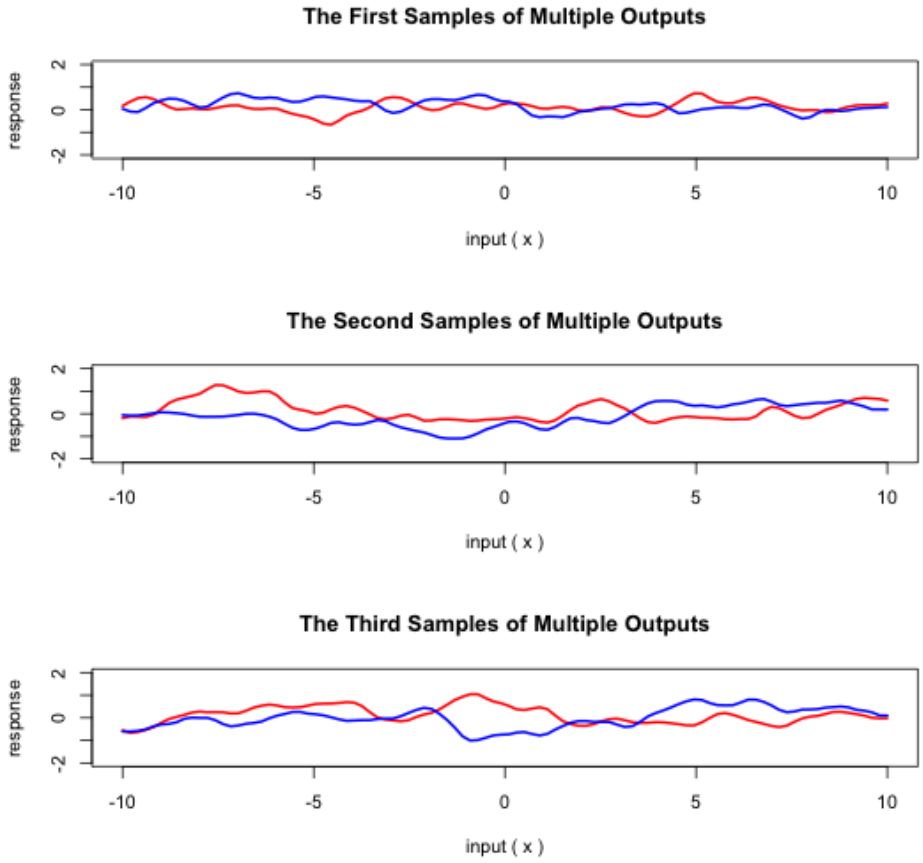

Figure C.3: Multiple dependent output process samples generated from rational quadratic covariance function with $\alpha = 0.5$ as convolved kernels . Red curves are defined as f_1 and blue curves are f_2.

C.2 Rational Quadratic Covariance Function

The closed forms of kernel functions from equation (4.10) can be defined as

$$\begin{aligned} k_{aa}^{\xi_a}(\mathbf{d}) &= \frac{v_a(\pi)^{P/2}}{|A_a|^{\frac{1}{2}}}\left(\frac{1}{1+\frac{1}{2\alpha}Q_{aa}(\mathbf{d})}\right)^{\nu} \\ k_{ab}^{\xi_{ab}}(\mathbf{d}) &= \frac{v_a v_b(2\pi)^{P/2}}{|A_a+A_b|^{\frac{1}{2}}}\left(\frac{1}{1+\frac{1}{2\alpha}Q_{ab}(\mathbf{d})}\right)^{\nu} \\ k_{aa}^{\eta_a}(\mathbf{x}) &= \frac{w_a(\pi)^{P/2}}{|B_a|^{\frac{1}{2}}}\left(\frac{1}{1+\frac{1}{2\alpha}Q_{aa}(\mathbf{d})}\right)^{\nu}, \quad \nu > 0 \end{aligned} \tag{C.1}$$

where $a, b = 1, 2$ and $Q_{ab}(\mathbf{d}) = (\mathbf{d})^T A_a(A_a + A_j)^{-1}A_b(\mathbf{d})$.

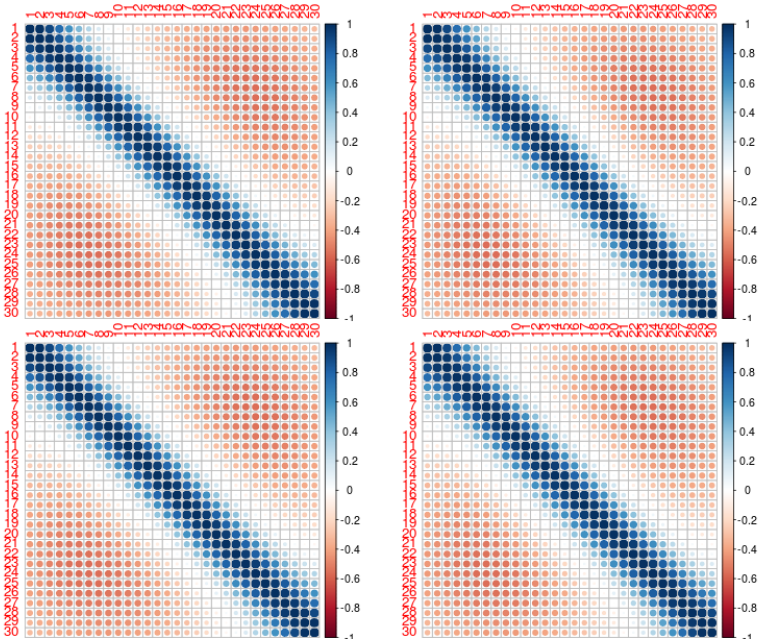

Figure C.4: Correlation map of multiple dependent Gaussian processes using a rational quadratic as convolved kernels between f_1 and f_2; top left: correlation map of $\mathbf{f}_1$ at 30 equally spaced points; top right: correlation map between $\mathbf{f}_1$ and $\mathbf{f}_2$; and bottom right: correlation map of $\mathbf{f}_2$

Parameters	True value	Average Mean	Average RMSE
v_1	0.04	0.000017	0.03999
v_2	0.04	0.000027	0.03999
A_1	1	1.000089	0.00008
A_2	1	1.000089	0.00008
w_1	0.04	0.000027	0.03747
w_2	0.04	0.000025	0.03999
B_1	1	1.00034 7	0.00034
B_2	1	1.000342	0.00034

Table C.2: Average RMSE of the estimated parameters from multiple Gaussian process with rational quadratic covariance function with $\alpha = 0.5$ based on 100 replications.

C.3 Matern Covariance Function

The closed forms of kernel functions from equation (4.10) can be defined as

$$k_{aa}^{\xi_a}(\mathbf{d}) = \frac{v_a(\pi)^{P/2}}{|A_a|^{\frac{1}{2}}} \frac{1}{\Gamma(\nu)2^{\nu-1}} \left(\sqrt{2\nu Q_{aa}(\mathbf{d})}\right)^{\nu} K_{\nu}\left(\sqrt{2\nu Q_{aa}(\mathbf{d})}\right), \quad \nu > 0,$$

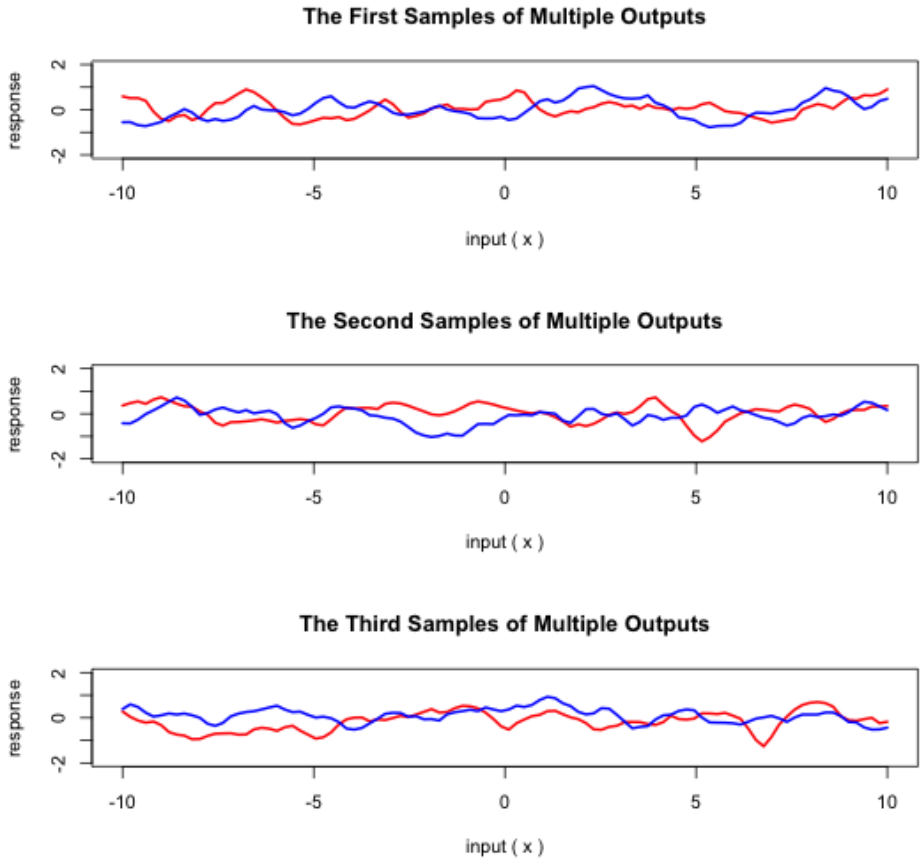

Figure C.5: Multiple dependent output process samples generated from Matern covariance function as convolved kernels. Red curves are defined as f_1 and blue curves are f_2.

$$k_{ab}^{\xi_{ab}}(\mathbf{d}) = \frac{v_a v_b (2\pi)^{P/2}}{|A_a + A_b|^{\frac{1}{2}}} \frac{1}{\Gamma(\nu) 2^{\nu-1}} \left(\sqrt{2\nu Q_{ab}(\mathbf{d})}\right)^{\nu} K_\nu \left(\sqrt{2\nu Q_{ab}(\mathbf{d})}\right), \quad \nu > 0,$$

$$k_{aa}^{\eta_a}(\mathbf{d}) = \frac{w_a (\pi)^{P/2}}{|B_a|^{\frac{1}{2}}} \frac{1}{\Gamma(\nu) 2^{\nu-1}} \left(\sqrt{2\nu Q_{aa}(\mathbf{d})}\right)^{\nu} K_\nu \left(\sqrt{2\nu Q_{aa}(\mathbf{d})}\right), \quad \nu > 0, \tag{C.2}$$

where $a, b = 1, 2$ and $Q_{ab}(\mathbf{d}) = (\mathbf{d})^T A_a (A_a + A_j)^{-1} A_b(\mathbf{d})$. If $\nu = \frac{3}{2}$, so the closed forms of

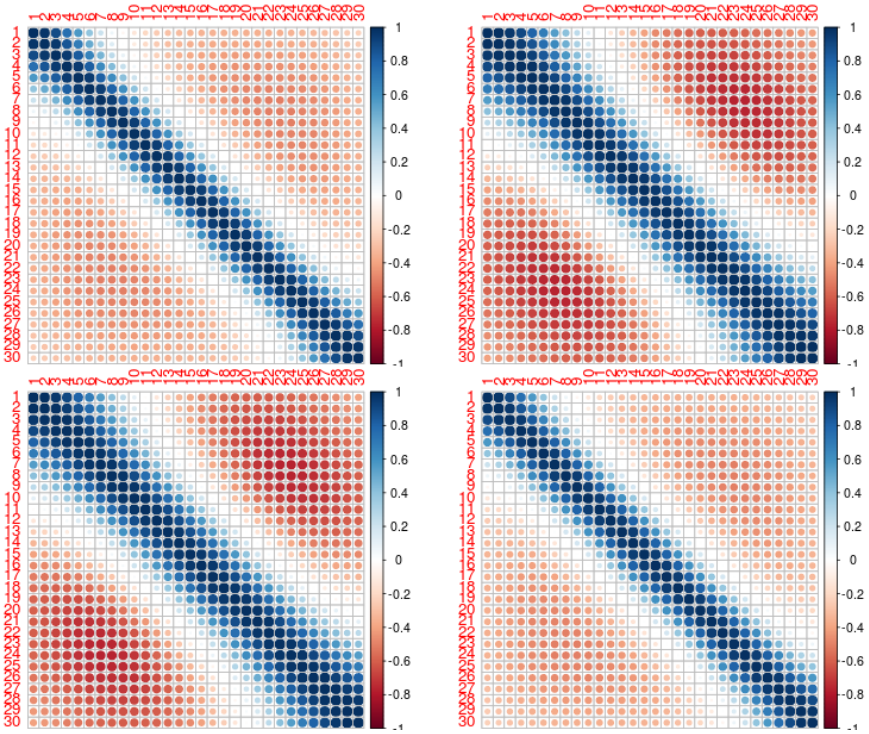

Figure C.6: Correlation map of multiple dependent Gaussian processes using a Matern as convolved kernels between f_1 and f_2; top left: correlation map of $\mathbf{f}_1$ at 30 equally spaced points; top right: correlation map between $\mathbf{f}_1$ and $\mathbf{f}_2$; and bottom right: correlation map of $\mathbf{f}_2$

Parameters	True value	Average Mean	Average RMSE
v_1	0.04	0.000015	0.03999
v_2	0.04	0.000015	0.03999
A_1	1	1.000122	0.00012
A_2	1	1.0000122	0.00012
w_1	0.04	-0.000004	0.04000
w_2	0.04	-0.000004	0.04000
B_1	1	1.000384	0.00038
B_2	1	1.000384	0.00038

Table C.3: Average RMSE of the estimated parameters from multiple Gaussian process with Matern covariance function with $\nu = \frac{3}{2}$ based on 100 replications.

kernel functions from equation (4.10) can be defined as

$$\begin{aligned} k_{aa}^{\xi_a}(\mathbf{d}) &= \frac{v_a(\pi)^{P/2}}{|A_a|^{\frac{1}{2}}}\left(1+\sqrt{3Q_{aa}(\mathbf{d})}\right)\exp\left(\sqrt{-3Q_{aa}(\mathbf{d})}\right) \\ k_{ab}^{\xi_{ab}}(\mathbf{d}) &= \frac{v_a v_b(2\pi)^{P/2}}{|A_a+A_b|^{\frac{1}{2}}}\left(1+\sqrt{3Q_{ab}(\mathbf{d})}\right)\exp\left(\sqrt{-3Q_{ab}(\mathbf{d})}\right) \\ k_{aa}^{\eta_a}(\mathbf{d}) &= \frac{w_a(\pi)^{P/2}}{|B_a|^{\frac{1}{2}}}\left(1+\sqrt{3Q_{aa}(\mathbf{d})}\right)\exp\left(\sqrt{-3Q_{aa}(\mathbf{d})}\right) \end{aligned} \tag{C.3}$$

where $a, b = 1, 2$ and $Q_{ab}(\mathbf{d}) = (\mathbf{d})^T A_a (A_a + A_j)^{-1} A_b(\mathbf{d})$.

The maximum likelihood estimates seem acceptable when compared to the true values. Also, it can be seen that the values of root mean square (RMSE) between the estimated parameters (A_1, A_2, B_1, B_2) and true values close to zero. It means that the differences between estimated parameters (A_1, A_2, B_1, B_2) and true values are small. Meanwhile, for estimated parameters v_1, v_2, w_1 and w_2 are quite far from the true values. It might because we do not add any noise variable in this modelling. In conclusion, this issue needs to investigate more.

Appendix D

Some Technical Details for Consistency

This technical details is extension from consistency theorem in (Wang & Shi, 2014).

Lemma D.1. *Suppose z_{1i} and z_{2i} are conditional independent samples from a bivariate Poisson distribution given* (5.1) *and $\tau_0 \in \mathcal{F}$ has a multivariate convolved Gaussian prior with zero mean and bounded covariance function $\mathcal{K}(\cdot,\cdot)$ for any covariate values in $\mathcal{X}$. Suppose that $\mathcal{K}(\cdot,\cdot)$ is continuous in $\boldsymbol{\theta}$ and the estimator $\hat{\boldsymbol{\theta}} \to \boldsymbol{\theta}$ almost surely as $N_1 \to \infty$ and $N_2 \to \infty$. If there exists a positive number κ such that $|b"(\alpha)| \leqslant e^{\kappa\alpha}$, then*

$$\begin{aligned} & -\log p_{mgp}(z_1, ..., z_{N_1}, z_{N_1+1}, ..., z_{N_1+N_2}) + \log p_0(z_1, ..., z_{N_1}, z_{N_1+1}, ..., z_{N_1+N_2}) \\ \leqslant \quad & \frac{1}{2}\|\tau_0\|^2_{\mathcal{K}_{N_1N_2}} + \frac{1}{2}\log|\mathbf{I} + \delta\mathcal{K}_{N_1N_2}| + C \end{aligned} \tag{D.1}$$

where $\|\tau_0\|^2_{\mathcal{K}_{N_1N_2}}$ is the reproducing kernel Hilbert space (RKHS) norm of τ_0 associated with $\mathcal{K}(\cdot,\cdot)$, $\mathcal{K}_{N_1N_2}$ is the covariance matrix of τ_0 over the covariance $\mathbf{X}_{N_1N_2}$, $\mathbf{I}$ is the $(N_1+N_2)\times(N_1+N_2)$ identity matrix, δ and C are some positive constants.

Proof. In this proof, we use these covariance functions to define functions on $\mathcal{X}$. The space of such functions is known as a reproducing kernel Hilbert space (RKHS) . Let $\mathcal{H}$ be RKHS associated with covariance function $\mathcal{K}(\cdot,\cdot)$ defined as the previous section. Consider the linear space of all finite kernel expansions and $\mathcal{H}_{N_1+N_2}$ the span of $\{\mathcal{K}(\cdot,\cdot)\}$ i.e.

$$\mathcal{H}_{N_1+N_2} = \left\{ f(\cdot) : f(\mathbf{x}) = \sum_{i=1}^{N_1+N_2} \alpha_i \mathcal{K}(\mathbf{x}, \mathbf{x}_i), \alpha_i \in \mathbb{R} \right\}.$$

We first assume the true underlying function $\tau_0 \in \mathcal{H}_{N_1+N_2}$ then $\tau_0(\cdot)$ can be expressed as

$$\tau_0(.) = \sum_{i=1}^{N_1+N_2} \alpha_i \mathcal{K}(\cdot, \mathbf{x}_i) \triangleq \mathbf{K}_{N_1N_2}(.)\boldsymbol{\alpha}.$$

where $\mathbf{K}_{N_1N_2} = (\mathcal{K}(\cdot, \mathbf{x}_i), ..., \mathcal{K}(\cdot, \mathbf{x}_{N_1+N_2}))$ and $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_{N_1+N_2})$. By the properties of RKHS, $\|\tau_0\|^2_{\mathcal{K}_{N_1N_2}} = \boldsymbol{\alpha}^T \mathcal{K}_{N_1N_2}\boldsymbol{\alpha}$, and $(\tau(\mathbf{x}_1), ..., \tau(\mathbf{x}_{N_1+N_2}))^T = \mathcal{K}_{N_1N_2}\boldsymbol{\alpha}$ where $\mathcal{K}_{N_1N_2} = (\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j))$ is the covariance matrix over $\mathbf{x}_i, \quad i = 1, ..., (N_1 + N_2)$.

Let P and $\hat{P}$ be any two measures on F, then it yields by the Fenchel-Legendre duality relationship that, for any functional $g(.)$ on F,

$$E_{\hat{p}}[g(\tau)] \leqslant \log E_p[e^{g(\tau)}] + D[\bar{P}, P] \tag{D.2}$$

Now in the above inequality let

1. $g(\tau)$ be $\log p(z_1, ..., z_{N_1}, z_{N_1+1}, ..., z_{N_1+N_2})$ for any $z_1, ..., z_{N_1+N_2}$ in $\mathcal{Z}$ and $\tau \in \mathcal{F}$

2. P be the measure induced by $\mathcal{MGP}(\mathbf{0}, \mathcal{K}(\cdot, \cdot))$, hence $\tilde{p}(z_1, ..., z_{N_1}, z_{N_1+1}, ..., z_{N_1+N_2}) = \mathcal{N}(\mathbf{0}, \hat{\mathcal{K}_{N_1N_2}})$ and

$$\begin{aligned} E_p[e^{g(\tau)}] &= E_p[p(z_1, ..., z_{N_1}, z_{N_1+1}, ..., z_{N_1+N_2})] \\ &= \int p(z_1, ..., z_{N_1}, z_{N_1+1}, ..., z_{N_1+N_2} \mid \tau) dp_{N_1+N_2} \\ &= p_{mgp}(\mathbf{z}) \end{aligned}$$

   where $\hat{\mathcal{K}}_{N_1N_2}$ is defined in the same way as $\mathcal{K}_{N_1N_2}$ but the $\boldsymbol{\theta}$ being replaced by its estimator $\hat{\boldsymbol{\theta}}$.

3. $\bar{P}$ be the posterior distribution of $\tau(\cdot)$ on $\mathcal{F}$ which has a prior distribution $\mathcal{MGP}(0, \mathcal{K}(\cdot, \cdot))$ and normal likelihood $\prod_{i=1}^{N_1+N_2} N(\hat{z}(i); \tau(\mathbf{x_i}), \sigma^2)$, where

$$\hat{\mathbf{z}} \triangleq \begin{pmatrix} \hat{z}_1 \\ \vdots \\ \hat{z}_{N_1+N_2} \end{pmatrix} = (\mathcal{K}_{N_1N_2} + \sigma^2\mathbf{I})\boldsymbol{\alpha} \tag{D.3}$$

   and σ^2 is a constant to be specified. In other words, we assume a model $z = \tau(\mathbf{x}) + \epsilon$ with $\epsilon \sim N(0, \sigma^2)$ and $\tau(\cdot) \sim \mathcal{MGP}(0, \mathcal{K}(\cdot, \cdot))$, and $\hat{\mathbf{z}}$ defined by equation (D.3) is a set of observations at $\mathbf{x}_1, ..., \mathbf{x}_{N_1+N_2}$. Thus, $\bar{P}(\tau) = p(\tau \mid \hat{\mathbf{z}}, \mathbf{X}_{N_1N_2})$ is a probability measure on $\mathcal{F}$. Therefore, by bivariate convolved Gaussian process regression, the

posterior of $(\tau_1, ..., \tau_{N_1+N_2}) \triangleq (\tau(\mathbf{x}_1, ..., \tau(\mathbf{x}_{N_1+N_2})))$ is

$$\begin{aligned}
\bar{p}(\tau_1, ..., \tau_{N_1+N_2}) &\triangleq p(\tau_1, ..., \tau_{N_1+N_2} \mid \hat{\mathbf{z}}, \mathbf{X}_{N_1N_2}) \\
&= N(\mathcal{K}_{N_1N_2}(\mathcal{K}_{N_1N_2} + \sigma^2\mathbf{I})^{-1}\hat{\mathbf{z}}, \mathcal{K}_{N_1N_2}(\mathcal{K}_{N_1N_2} + \sigma^2\mathbf{I})^{-1}\sigma^2) \\
&= N(\mathcal{K}_{N_1N_2}\boldsymbol{\alpha}, \mathcal{K}_{N_1N_2}(\mathcal{K}_{N_1N_2} + \sigma^2\mathbf{I})^{-1}\sigma^2)) \\
&= N(\mathcal{K}_{N_1N_2}\boldsymbol{\alpha}, \mathcal{K}_{N_1N_2}\mathbf{B}^{-1}) \qquad \text{(D.4)}
\end{aligned}$$

where $\mathbf{B} = (\mathbf{I} + \sigma^{-2}\mathcal{K}_{N_1N_2})$

It follows that

$$\begin{aligned}
D[\bar{P}, P] &= \int_{\mathcal{F}} \log \frac{d\bar{P}}{dP} d\bar{P} \\
&= \int_{\mathcal{R}^{N_1+N_2}} \bar{p}(\tau_1, ..., \tau_{N_1+N_2}) \log \frac{\bar{p}(\tau_1, ..., \tau_{N_1+N_2})}{\tilde{p}(\tau_1, ..., \tau_{N_1+N_2})} \\
&= \frac{1}{2}[\log|\hat{\mathcal{K}}_{N_1N_2}| - \log|\mathcal{K}_{N_1+N_2}| + \log|\mathbf{B}| + tr(\mathcal{K}^{-1}_{N_1+N_2}\mathcal{K}_{N_1N_2}\mathbf{B}^{-1}) + (\mathcal{K}_{N_1N_2}\boldsymbol{\alpha})^T \\
&\quad \hat{\mathcal{K}}^{-1}_{N_1N_2}(\mathcal{K}_{N_1N_2}\boldsymbol{\alpha}) - n] \\
&= \frac{1}{2}[-\log|\hat{\mathcal{K}}^{-1}_{N_1N_2}\mathcal{K}_{N_1+N_2}| + \log|\mathbf{B}| + tr(\mathcal{K}^{-1}_{N_1+N_2}\mathcal{K}_{N_1N_2}\mathbf{B}^{-1}) + \|\tau_0\|^2_{K_{N_1N_2}} \\
&\quad \boldsymbol{\alpha}^T\mathcal{K}_{N_1N_2}(\hat{\mathcal{K}}^{-1}_{N_1N_2}\mathcal{K}_{N_1N_2} - \mathbf{I})\boldsymbol{\alpha} - n]
\end{aligned}$$

On the other hand,

$$E_{\bar{p}}[g(\tau)] = E_{\hat{p}}[\log p(\tau_1, ..., \tau_{N_1+N_2})|\tau] = \sum_{i=1}^{N_1+N_2} E_{\bar{p}}[\log p(z_i)|\tau(\mathbf{x}_i)].$$

By Taylor's expansion, expanding $\log p(z_i|\tau(\mathbf{x}_i))$ to the second order $\tau_0(\mathbf{x}_i)$ yields

$$\begin{aligned}
\log p(z_i|\tau(\mathbf{x}_i)) &= \log p(z_i|\tau_0(\mathbf{x}_i)) + \frac{d[\log p(z_i|\tau(\mathbf{x}_i))]}{d\tau(\mathbf{t}_i)}\Big|_{\tau(\mathbf{x}_i)=\tau_0(\mathbf{x}_i)}(\tau(\mathbf{x}_i) - \tau_0(\mathbf{x}_i)) \\
&\quad + \frac{1}{2}\frac{d^2[\log p(z_i|\tau(\mathbf{x}_i))]}{[d\tau(\mathbf{x}_i)]^2}\Big|_{\tau(\mathbf{x}_i)=\tilde{\tau}(\mathbf{x}_i)}(\tau(\mathbf{x}_i) - \tau_0(\mathbf{x}_i))^2
\end{aligned}$$

where $\tilde{\tau}(\mathbf{x}_i) = \tau_0(\mathbf{x}_i) + \lambda(\tau(\mathbf{x}_i) - \tau_0(\mathbf{x}_i))$ for some $0 \leqslant \alpha \leqslant 1$.

The canonical link function with Convolved GPR , we have

$$p(z_i|\tau(\mathbf{x}_i)) = \exp\left\{\frac{z_i\tau(\mathbf{x}) - b(\tau(\mathbf{x}))}{c(\phi_i)} + d(z_i, \phi_i)\right\}, \qquad \text{(D.5)}$$

thus

$$\frac{d^2[\log p(z_i|\tau(\mathbf{x}_i))]}{[d\tau(\mathbf{x}_i)]^2}\Big|_{\tau(\mathbf{x}_i)=\tilde{\tau}(\mathbf{x}_i)}(\tau(\mathbf{x}_i) - \tau_0(\mathbf{x}_i))^2 = -\frac{b''\tilde{\tau}(\mathbf{x}_i)}{c(\phi_i)}$$

It follows that

$$\begin{aligned} E_{\bar{p}}[\log p(z_i|\tau(\mathbf{x}_i))] &= \log p(z_i|\tau_0(\mathbf{x}_i)) + \frac{d[\log p(z_i|\tau(\mathbf{x}_i))]}{d\tau(\mathbf{x}_i)}\Big|_{\tau(\mathbf{x}_i)=\tau_0(\mathbf{x}_i)}(\tau(\mathbf{x}_i)-\tau_0(\mathbf{x}_i)) \\ &\quad + \frac{1}{2c(\phi_i)}E_{\bar{p}}[b''\tilde{\tau}(\mathbf{x}_i)(\tau(\mathbf{x}_i)-\tau_0(\mathbf{x}_i)^2]. \end{aligned}$$

Since $\bar{P(.)}$ is the posterior of $\tau(.)$ which has prior $\mathcal{MGP}(\mathbf{0}, \mathcal{K}(\cdot,\cdot))$ and normal likelihood $\prod_{i=1}^{N_1+N_2} N(z_i; \tau(\mathbf{x}_i), \sigma^2)$, where $\tau(\mathbf{x}_i)$ is normally distibuted under $\bar{P}$ and it follows from (D.4) that

$$\begin{aligned} \tau(\mathbf{x}_i) &\sim \mathcal{N}(\mathcal{K}^{(i)}_{N_1N_2}, (\mathcal{K}_{N_1N_2}\mathcal{B}^{-1})_{ii}) \\ &= \mathcal{N}(\tau_0(\mathbf{x}_i), (\mathcal{K}_{N_1N_2}\mathbf{B}^{-1})_{ii})) \triangleq \mathcal{N}(\tau_{0i}, \mathcal{K}_{ii}) \end{aligned}$$

where $\mathcal{K}^{(i)}_{N_1N_2}$ denotes the ith the row of $\mathcal{K}_{N_1N_2}$ and $(\mathcal{K}_{N_1N_2}\mathbf{B}^{-1})_{ii}$ is the ith diagonal element of $(\mathcal{K}_{N_1N_2}\mathbf{B})^{-1}$. Therefore, $E_{\bar{p}}[\tau(\mathbf{x}_i)-\tau_0(\mathbf{x}_i)] = 0$ and

$$\begin{aligned} E_{\bar{p}}[b''\tilde{\tau}(\mathbf{x}_i)(\tau(\mathbf{x}_i)-\tau_0(\mathbf{x}_i))^2] &\leqslant E_{\bar{p}}[e^{\kappa\tilde{\tau}(\mathbf{x}_i)}(\tau(\mathbf{x}_i)-\tau_0(\mathbf{x}_i))^2] \\ &= \int_{-\infty}^{+\infty}(\tau_i-\tau_{0i})^2 e^{\kappa\tau_{0i}+\kappa\lambda(\tau_i-\tau_{0i})N(\tau_{0i},K_{ii})d\tau_i} \\ &= e^{\kappa\tau_{0i}+\frac{1}{2}\kappa^2\lambda^2\mathcal{K}_{ii}}(\kappa^2\lambda^2\mathcal{K}_{ii}+1)\mathcal{K}_{ii} \leqslant \tilde{\delta}\mathcal{K}_{ii} \end{aligned}$$

since the covariance function is bounded. Here $\tilde{\delta}$ is a generic positive constant. Thus, we have

$$-E_{\bar{P}}[\log p(z_i|\tau(\mathbf{x}_i))] \leqslant -\log p(z_i|\tau_0(\mathbf{x}_i)) + \frac{\tilde{\delta}}{2}tr(\mathcal{K}_{N_1N_2}\mathbf{B}^{-1})_{ii}.$$

and

$$-\sum_{i=1}^{N_1+N_2} E_{\bar{P}}[\log p(z_i|\tau(\mathbf{x}_i))] \leqslant -\sum_{i=1}^{N_1+N_2}\log p(y_i|\tau_0(\mathbf{x}_i)) + \frac{\tilde{\delta}}{2}tr(\mathcal{K}_{N_1N_2}\mathbf{B}^{-1}).$$

i.e.

$$\log p_0(z_1,...,z_{N_1},z_{N_1+1},...,z_{N_1+N_2}) \leqslant E_{\bar{p}[g(\tau)]} + \frac{\tilde{\delta}}{2}tr(\mathcal{K}_{N_1N_2}\mathbf{B}^{-1})$$

Combining the bounds give

$$\begin{aligned} & -\log p_{mgp}(z_1,...,z_{N_1},z_{N_1+1},...,z_{N_1+N_2}) + \log p_0(z_1,...,z_{N_1},z_{N_1+1},...,z_{N_1+N_2}) \\ \leqslant\ & \log E_p[e^{g(\tau)}] + E_{\bar{p}}[g(\tau)] + \frac{\tilde{\delta}}{2}tr(\mathcal{K}_{N_1N_2}\mathbf{B}^{-1}) \\ \leqslant\ & D[\bar{P},P] + \frac{\tilde{\delta}}{2}tr(\mathcal{K}_{N_1N_2}\mathbf{B}^{-1}) \\ =\ & \frac{1}{2}[-\log|\hat{\mathcal{K}}^{-1}_{N_1N_2}\mathcal{K}_{N_1+N_2}| + \log|\mathbf{B}| + tr(\mathcal{K}^{-1}_{N_1+N_2}\mathcal{K}_{N_1N_2}\mathbf{B}^{-1} + \tilde{\delta}\mathcal{K}_{N_1N_2}\mathbf{B}^{-1}) + \|\tau_0\|^2_{\mathcal{K}_{N_1N_2}} \\ & \boldsymbol{\alpha}^T\mathcal{K}_{N_1N_2}(\hat{\mathcal{K}}^{-1}_{N_1N_2}\mathcal{K}_{N_1N_2}-\mathbf{I})\boldsymbol{\alpha} - n] \end{aligned} \tag{D.6}$$

Since the covariance function is continuous in $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}_{N_1+N_2} \to \boldsymbol{\theta}$ and we have $\hat{\mathcal{K}_{N_1N_2}}\mathcal{K}_{N_1N_2} - \mathbf{I} \to 0$ as $N_1 \to \infty$ and $N_2 \to \infty$, hence $N_1 + N_2 \to \infty$. Therefore there exist some positive constants C and ϵ such that

$$-\log|\hat{\mathcal{K}}^{-1}_{N_1N_2}\mathcal{K}_{N_1+N_2}| < C, \quad \boldsymbol{\alpha}^T\mathcal{K}_{N_1N_2}(\hat{\mathcal{K}}^{-1}_{N_1N_2}\mathcal{K}_{N_1N_2} - \mathbf{I})\boldsymbol{\alpha} < C,$$

$$tr(\mathcal{K}^{-1}_{N_1+N_2}\mathcal{K}_{N_1N_2}\mathbf{B}^{-1} < tr((\mathbf{I} + \epsilon\mathcal{K}_{N_1N_2})\mathbf{B}^{-1}),$$

since the covariance function is bounded.

Thus RHS of (D.6)

$$< \frac{1}{2}\|\tau_0\|^2_{K_{N_1N_2}} + \frac{1}{2}[2C + \log|\mathbf{B}| + tr((\mathbf{I} + (\epsilon + \tilde{\delta})\mathcal{K}_{N_1N_2})\mathbf{B}^{-1}) - n]$$

Note that thee above inequality holds for all $\sigma^2 > 0$, thus letting $\sigma^2 = (\epsilon + \tilde{\delta})^{-1}$ and $\delta = \epsilon + \tilde{\delta}$ yields that the RHS of (D.6) becomes

$$\frac{1}{2}\|\tau_0\|^2_{K_{N_1N_2}} + \frac{1}{2}\log(\mathbf{I} + \delta\mathcal{K}_{N_1N_2}) + C$$

Thus we have

$$\begin{aligned} -\log p_{mgp}(z_1, ..., z_{N_1}, z_{N_1+1}, ..., z_{N_1+N_2}) \quad \leqslant \quad & -\log p_0(z_1, ..., z_{N_1}, z_{N_1+1}, ..., z_{N_1+N_2}) + \frac{1}{2}\|\tau_0\|^2_{K_{N_1N_2}} + \\ & \frac{1}{2}\log(\mathbf{I} + \delta\mathbf{K}_{N_1N_2}) + C \end{aligned} \tag{D.7}$$

for any $\tau_0(\cdot) \in H_{N_1+N_2}$.

Taking infimum on RHS of (D.7) over τ_0 and applying *Representer Theorem* (see Lemma 2 in Seeger *et al.* (2008)), we obtain

$$\begin{aligned} \log p_{mgp}(z_1, ..., z_{N_1}, z_{N_1+1}, ..., z_{N_1+N_2}) + \log p_0(z_1, ..., z_{N_1}, z_{N_1+1}, ..., z_{N_1+N_2}) \\ \leqslant \frac{1}{2}\|\tau_0\|^2_{K_{N_1N_2}} + \frac{1}{2}\log(\mathbf{I} + \delta\mathcal{K}_{N_1N_2}) + C \end{aligned}$$

for all $\tau_0(\cdot) \in H_{N_1+N_2}$. The proof is complete □

Proof of Theorem 5.1. It follows from the definition of information consistency that

$$\begin{aligned} D[p_0(\mathbf{z}), p_{mgp}(\mathbf{z})] \quad = \quad & \int p_0(z_1, ..., z_{N_1}, z_{N_1+1}, ..., z_{N_1+N_2})\frac{p_0(z_1, ..., z_{N_1}, z_{N_1+1}, ..., z_{N_1+N_2})}{p_{mgp}(z_1, ..., z_{N_1}, z_{N_1+1}, ..., z_{N_1+N_2})} \\ = \quad & \int p_0(z_1, ..., z_{N_1}, z_{N_1+1}, ..., z_{N_1+N_2}) \\ & [-\log p_{mgp}(z_1, ..., z_{N_1}, z_{N_1+1}, ..., z_{N_1+N_2}) + \log p_0(z_1, ..., z_{N_1}, z_{N_1+1}, ..., z_{N_1+N_2})] \\ & d(z_1, ..., z_{N_1}, z_{N_1+1}, ..., z_{N_1+N_2}). \end{aligned}$$

Applying Lemma D.1 we obtain that

$$\frac{1}{N_1+N_2}E_{\mathbf{X}_{N_1N_2}}\left(D[p_0(\mathbf{z}),p_{mgp}(\mathbf{z})]\right) \leqslant \frac{1}{2(N_1+N_2)}\left\|\tau_0\right\|^2_{K_{N_1N_2}} + \frac{1}{2(N_1+N_2)}E_{\mathbf{X}_{N_1N_2}}\log(\mathbf{I}+\delta\mathcal{K}_{N_1N_2}) + \frac{C}{N_1+N_2} \tag{D.8}$$

where δ and C are some positive constants. Theorem 1 follows from (D.8). □

Remark. Lemma D.1 requires that the estimator of the coefficients ($\boldsymbol{\beta}$) and hyper-parameters ($\boldsymbol{\theta}$) are consistent. Yi et al. (2011), provided that the empirical Bayesian estimator of hyper-parameters $\boldsymbol{\theta}$ as $N \to \infty$ under certain regularity. The estimator $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ for bivariate Poisson regression with convolved Gaussian process priors are consistent under certain regularity, if $N = N_1 + N_2$, where $N_1 \to \infty \quad and \quad N_2 \to \infty$.

Remark. Some specific results of the regret term $R = E_{\mathbf{X}_{N_1N_1}}(\log|\mathbf{I}+\delta\mathcal{K}_{N_1N_2}|)$ as follows :

i. if $\mathcal{K}(\mathbf{x}_i,\mathbf{x}_j) = \mathbf{x}_i^T\mathbf{x}_j$, i.e. a linear covariance kernel, and the covariate distribution $\mathbf{u}(\mathbf{x})$ has bounded support, then

$$E_{\mathbf{X}_{N_1N_1}}(\log|\mathbf{I}+\delta\mathcal{K}_{N_1N_2}|) = O(\log(N_1+N_2));$$

ii. if $\mathbf{u}(\mathbf{x})$ is normal and the covariance functions are the squared exponential

$$E_{\mathbf{X}_{N_1N_1}}(\log|\mathbf{I}+\delta\mathcal{K}_{N_1N_2}|) = O((\log(N_1+N_2))^{P+1});$$

iii. if $\mathbf{u}(\mathbf{x})$ is bounded support and the covariance functions are Matern, then

$$E_{\mathbf{X}_{N_1N_1}}(\log|\mathbf{I}+\delta\mathcal{K}_{N_1N_2}|) = O((N_1+N_2)^{P/(2v+P)}(\log(N_1+N_2)^{2v/(2v+P)}));$$

iv. if covariance functions are mixed between squared exponential and Matern, then

$$E_{\mathbf{X}_{N_1N_1}}(\log|\mathbf{I}+\delta\mathcal{K}_{N_1N_2}|) = O((N_1+N_2)^{P/(2v+P)}(\log(N_1+N_2)^{2v/(2v+P)}));$$

It is obvious that for all of the above cases the information consistency in the proposed model is achieved.

Remark. The consistency considered in Theorem 1 assumes the mean function is known. If the mean function is unknown and is estimated from the observations, its uncertainty

needs to be taken into account. We denote by $\hat{\mu(t)}$ the estimator of the mean function $\mu(t)$ and let

$$\hat{p}_{gp}(\mathbf{y}) = \int_{\mathcal{F}} \hat{p}(z_1, ..., z_{N_1}, z_{N_1+1}, ..., z_{N_1+N_2}|\tau(\mathbf{X}_{N_1N_2}))dp_{N_1+N_2}$$

where $\hat{p}(z_1, ..., z_{N_1}, z_{N_1+1}, ..., z_{N_1+N_2}|\mathbf{X}_{N_1N_2}))$ is the conditional distribution of

$$(z_1, ..., z_{N_1}, z_{N_1+1}, ..., z_{N_1+N_2})$$

with the estimated mean function $\hat{\mu(t)}$. It follow from Lemma D.1 that

$$\begin{aligned}
& -\log p_{mgp}(z_1, ..., z_{N_1}, z_{N_1+1}, ..., z_{N_1+N_2}) + \log p_0(z_1, ..., z_{N_1}, z_{N_1+1}, ..., z_{N_1+N_2}) \\
= \;& \log p_{mgp}(z_1, ..., z_{N_1}, z_{N_1+1}, ..., z_{N_1+N_2}) - \log \hat{p}(z_1, ..., z_{N_1}, z_{N_1+1}, ..., z_{N_1+N_2}) \\
& -\log p_{mgp}(z_1, ..., z_{N_1}, z_{N_1+1}, ..., z_{N_1+N_2}) + \log p_0(z_1, ..., z_{N_1}, z_{N_1+1}, ..., z_{N_1+N_2}) \\
\leqslant \;& -\log p_{mgp}(z_1, ..., z_{N_1}, z_{N_1+1}, ..., z_{N_1+N_2}) + \log p_0(z_1, ..., z_{N_1}, z_{N_1+1}, ..., z_{N_1+N_2}) + \\
& \|\tau_0\|^2_{K_{N_1N_2}} + \log(\mathbf{I} + \delta\mathcal{K}_{N_1N_2}) + C.
\end{aligned}$$

For the canonical link function,we have

$$\begin{aligned}
& \hat{p}(z_1, ..., z_{N_1}, z_{N_1+1}, ..., z_{N_1+N_2}|\tau(\mathbf{X}_{N_1N_2})) \\
= \;& \exp\left\{\sum_{i=1}^{N_1+N_2} \frac{z_i(\hat{\mu}+\tau(\mathbf{x})) - b(\hat{\mu}+\tau(\mathbf{x}))}{c(\phi_i)} + \sum_{i=1}^{N_1+N_2} d(z_i, \phi_i)\right\} \\
\triangleq \;& e^{g(\hat{\mu}+\tau)}
\end{aligned}$$

and

$$\begin{aligned}
& p(z_1, ..., z_{N_1}, z_{N_1+1}, ..., z_{N_1+N_2}|\tau(\mathbf{X}_{N_1N_2})) \\
= \;& \exp\left\{\sum_{i=1}^{N_1+N_2} \frac{z_i(\mu+\tau(\mathbf{x})) - b(\mu+\tau(\mathbf{x}))}{c(\phi_i)} + \sum_{i=1}^{N_1+N_2} d(z_i, \phi_i)\right\} \\
\triangleq \;& e^{g(\mu+\tau)}
\end{aligned}$$

If z_i has finite two moments and its variance is bounded away from zero, there exist positive constants C_1, C_2 and C_3 such that $|b'(\cdot)| < C_1$ and $C_2 < c(\cdot) < C_3$. For a bivariate Poisson distribution the dispersion value $c(\cdot)$ in canonical link function is one. It follows that

$$b(\hat{\mu}+\tau) - b(\mu+\tau) \leqslant C_1 \|\hat{\mu}-\mu\|, \quad or, \quad -b(\mu+\tau) \leqslant C_1 \|\hat{\mu}-\mu\| - b(\hat{\mu}+\tau).$$

Hence,

$$\begin{aligned} g(\mu+\tau) &\leqslant \sum_{i=1}^{N_1+N_2} \frac{(|z_i|+C_1)\,\|\hat{\mu}-\mu\|}{c(\phi_i)} + g(\hat{\mu}-\tau) \\ &= \sum_{i=1}^{N_1+N_2} z_i(\mu-\hat{\mu}) + C_1\,\|\hat{\mu}-\mu\| + g(\hat{\mu}+\tau). \end{aligned}$$

It yields that

$$\begin{aligned} \log p_{mgp}(z_1,...,z_{N_1},z_{N_1+1},...,z_{N_1+N_2}) &- \log \hat{p}_{mgp}(z_1,...,z_{N_1},z_{N_1+1},...,z_{N_1+N_2}) \\ &= \log \frac{\int_{\mathcal{F}} e^{g(\mu+\tau)} dp_{N_1+N_2}}{\int_{\mathcal{F}} e^{g(\hat{\mu}+\tau)} dp_{N_1+N_2}} \\ &\leqslant \sum_{i=1}^{N_1+N_2} (|z_i|+C_1)\,\|\hat{\mu}-\mu\|\,. \end{aligned}$$

Therefore, following the same argument as in (D.8) we obtain

$$\begin{aligned} \frac{1}{N_1+N_2} E_{\mathbf{X}_{N_1N_2}}\left(D[p_0(\mathbf{z}), p_{mgp}(\mathbf{y})]\right) \leqslant\ & \tilde{C}\,\|\hat{\mu}-\mu\|\,\frac{1}{2(N_1+N_2)}\,\|\tau_0\|^2_{K_{N_1N_2}} + \\ & \frac{1}{2(N_1+N_2)}\log(\mathbf{I}+\delta\mathcal{K}_{N_1N_2}) + \frac{C}{N_1+N_2}, \end{aligned}$$

where $\tilde{C}$, δ and C are some positive constants.

Bibliography

Alam, M. M. 2009 An efficient algorithm for the pseudo likelihood estimation of the generalized linear mixed models (glmm) with correlated random effects. Working paper, Dalarna University.

Alvarez, M. A. 2011 Convolved gaussian process prior for multivariate regression with applications to dynamical sysstems. PhD thesis, School of Computer Science, University of Manchester.

Alvarez, M. A. & Lawrence, N. D. 2011 Computationally efficient convolved multiple output gaussian processes. *Journal of Machine learning research* **12**, 1459:1500.

Andriluka, M., Weizsacker, L. & Hofmann, T. 2006 Multi-class classification with dependent gaussian process. Darmstadt University of Technology.

Babaei, A. & Jabbari, B. 2010 Distance distribution of bivariate poisson network nodes. *EEE Communications Letters* **14**, 9.

Banerjee, S., Carlin, B. P. & Gelfand, A. E. 2011 *Hierarchical Modeling and Analysis for Spatial Data*. Taylor and Francis Ltd.

Berkhout, P. & Plug, E. 2004 A bivariate poisson count data model using conditional probabilities. Department of Economics, University of Amsterdam.

Bernadinelli, L., Clyton, D. & Montomoli, C. 1995 Bayesian estimates of disease maps: how important are priors? *Statistics in Medicine* **14**, 2411–2431.

Besag, J. & Kooperberg, C. 1995 On conditional and intrinsics autoregressions. *Biometrika* **82**, 733–746.

Besag, J., York, J. C. & Mollie, A. 1991 Bayesian image restoration with two applications in spatial statistics (with discussion). *Annals of the Institute of Statistical Mathematics* **43**, 1–59.

Boyle, P. & Frean, M. 2005 Multiple output guassian process regression. Technical Report, Victoria University of wellington, April 2005.

Breslow, N. E. & Clayton, D. G. 1993 Approximate inference in generalized linear mixed models. *The American Statistical Association* **88**, 9–25.

Brunsdon, C., Fotheringham, S. & Charlton, M. 1998 Geographically weighted regression-modelling spatial non stationarity. *Journal of the Royal Statistical Society* **47**.

Cameron, A. C. & Trivedi, P. K. 1998 *Regression analysis of count data*. Cambridge University Pres.

Carlin, B. & Banerjee, S. 2003 Hierarchical multivariate car models for spatio temporally correlated survival data. *Bayesian Statistics* .

Carlin, S. B. B. P. & Gelfand, A. E. 2004 *Hierarchical Modeling and Analysis for Spatial Data*. Boca Rotan, Taylor and Francis.

Choi, T. 2005 Posterior consistency in nonparametric regression problem under gaussian process priors. PhD thesis, Carnegie Mellon Universitu, Pittsburgh, PA.

Christensen, F. 2004 Monte carlo maximum likelihood in model-based geostatistics. *Computational and Graphical Statistics* **13**, 702–718.

Crainiceanu, C. M., Diggle, P. J. & Rowlingson, B. 2008 Bivariate binomial spatial modeling of loa loa prevalence in tropical africa. *Journal of the American Statistical Association* **103**, 21–37.

Diggle, P., Ribeiro & Paulo, J. 2007*a* *Model Based Geostatics*. Springer.

Diggle, P., Thomson, M. C., Chirstensen, O. F., Rowlingsin, B., Obsomer, V., Wanji, S., Takougang, I., Enyong, P., Kamqro, J., Remne, J. H., Boussinesq, M. & Molyneux, D. H. 2007*b* Spatial modeling and the prediction of loa loa risk : decision making under uncertainty. *Ann Trop Med Parasidol* **101(6)**, 499:509.

Dormann, C., McPherson, J. & Araujo, M. 2007 Methods to account for spatial autocorrelation in the analysis of species distributional data : a review. *Ecography* **30**, 609–628.

Elyazar, I., Hay, S. & Baird, J. 2011 Malaria distribution, prevalance, drug resistance and control in indonesia. *Adv Parasitol* **74**, 41–175.

Fortheringham, S., Brunsdon, C. & Charlton, M. 2003 *Geographically Weighted Regression*. The Willey, Hoboken.

Gelfand, A. & Vounatsou, P. 2003 Proper multivariate conditional autoregressive models for spatial data analysis. *BioStatistics* **4**, 11–25.

Genton, M. G. & Kleiber, W. 2015 Cross-covariance functions for multivariate geostatistics. *Statistics Science* **30(2)**, 147–163.

Haining, R. 1990 *Spatial Analysis in the Social and Environmental Sciences*. Cambridge, Cambridge University Press.

Hu, X., Simpson, D., Lindgren, F. & Rue, H. 2013 Multivariate gaussian random field using system of stochastic partial differential equations. *arXiv (Stat.ME)* **2**, 1307–1379.

Jiang, J. 2007 *Linear and generalized linear mixed models and their application*. New York, London: Springer.

Jin, X., Banerjee, S. & Carlin, B. P. 2005*a* Multivariate lattice models for areal data with application to multiple disease mapping. Division of Biostatistics, University of Minnesota.

Jin, X., Banerjee, S. & Carlin, B. P. 2007 Order-free co-regionalized areal data models with application to multi-disease mapping. *J. R Stat Soc Series B Stat methodol* **69(5)**, 817–838.

Jin, X., Carlin, B. & Banerjee, S. 2005*b* Generalized hierarchical multivariate car model for areal data. *Biometrics* .

Jung, C. & Winkelmann, R. 2001 Two aspect of labor mobility: A bivariate poisson regression approach. *Empirical Economics* **3**, 543–556.

Jung, R. & Winkelmann, R. 1993 Two aspects of labor mobility : A bivariate poisson regression approach. *Empirical Economics* **18**, 543–556.

Kano, K. & Kawamura, K. 1991 On recurrence relations for the probability function of multivariate generalized poisson distribution. *Communications in Statistics-Theory and Methods* **20**, 165178.

Karlis, D. & Bermudez, D. 2011 Mixture of bivariate poisson regression models with an application to insurance. Kavala, Athens.

Karlis, D. & Ntzoufras, J. 2003 Bayesian and non bayesian analysis for soccer using bivariate poisson regression models. Kavala, Athens.

Khoiriyah, I. & Sofro, A. 2012 Bivariate poisson regression (study case in malaria and dengue haemorrhagic fever in east java, indonesia). Thesis, Mathematics Department, State University of Surabaya, Indonesia.

Kim, H., Sun, D. & Tsutakawa, R. K. 2001 A bivariate bayes method for improving the estimates of mortality rates with a twofold conditional autoregressive model. *Journal of the American Statistical Association* **96**, 1506–1521.

Kocherladota, S. & Kocherlakota, K. 1992 Bivariate discrete distributions. Marcel Dekker: New Delhi.

Lang, W. 2009 *Mixed effects models*. Champman and Hall.

Lekdee, K. & Ingsrisawang, L. 2013 Generalized linear mixed models with spatial random effects for spatio-temporal data : An application to dengue fever mapping. *Journal of Mathmematics and Statistics* **9**, 137–143.

Lindgren, F., Rue, H. & Lindstrom, J. 2011 An explicit link between gaussian fields and gaussian markov random fields. *Journal of the Royal Statistical Society* **73**, 423–498.

Liu, Q. & Pierce, D. 1993 Heterogeneity in mantel-haenszel-type models. *Biometrika* **80**, 543–56.

Ma, H. & Carlin, B. 2007 Bayesian multivariate areal wombling for multiple disease boundary analysis. *International Society for Bayesian Analysis* **2**, 281–302.

MacNab, Y. & Dean, C. 2001 Autoregressive spatial smoothing and temporal spline smoothing for mapping rate. *Biometrics* **57**, 73–85.

MacNab, Y. C. 2011 On gaussian markov random fields and bayesian disease mapping. *Statistical Method in Medical Research* **20**, 49–68.

Majumdar, A., D.Paul & Bautista, D. 2010 Multivariate nonstationary spatial processes. *Statistics Sinica* **20**, 675–695.

Mardia, K. 1988 Multi dimensional multivariate gaussian markov random fields with application to image processing. *Journal of Multivariate Analysis* **24**, 265–284.

Martinez-Beneito, M. 2013 A general modelling framework for multivariate disease mapping. *Biometrika* **100**, 539–553.

Martinez-Beneito, M., Lopez-Quilez, A. & Botella-Rocamora, P. 2008 An autoregressive approach to spatio-temporal disease mapping. *Statistics in Medicine* **27**, 2874–2889.

Mohebbi, M. 2011 A poisson regression approach for modelling spatial autocorrelation between geogrphically referenced observations. BMC MEdical Research Methodology.

Monogan, J. 2012 The fifty american state in space and time : Applying conditionally autoregressive models to state politics. University of Georgia, USA.

Monogan, J. 2013 Areal data models. University of Georgia, Spring 2013.

Neelon, B., Anthopolos, R., L, M. & Miranda 2014 A spatial bivariate probit model for correlated binary data with application to adverse birth outcomes. *Statistical Methods in Medical Research* **23(2)**, 119–133.

Ntzourfras, I. 2009 *Bayesian Modelling with WinBugs*. The Willey, Hoboken.

Paciorek, C. 2009 Understanding instrinsic gaussian markov random field spatial models, including instrinsic conditional. Department of Statistics, University of California and Department of Biostatistics, Havard School of Public Health, USA.

Paciorek, C. J. 2003 Nonstationary gaussian processes for regression and spatial modelling. Dissertation, Carnegie Mellon University.

Rasmussen, C. E. & Williams, C. K. 2006 *Gaussian Process for Machine Learning*. Cambridge, England: MIT Press.

Renato & Krainski, E. 2004 Neighborhood dependence in bayesian spatial models. *Biometrical Journal* **2009**, 851869.

Rue, H. & Held, L. 2005 *Gaussian Markov Random Fields, Theory and Applications*. Chapman and Hall, Taylor and Francis Group.

Seeger, M., Kakade, S. & Foster, D. 2008 Information consistency of nonparamteric gaussian process methods. *IEEE Transactions on Information Theory* **54**, 2376–2382.

Serradilla, J. 2012 Gaussian process models for process monitoring and control. PhD thesis, School of Mathematics and Statistics, Newcastle University, UK.

Shi, J. & Choi, T. 2011 *Gaussian Process Regression Analysis for Functional Data*. Boca Rotan, Taylor and Francis.

Shin, K. & Raghu, P. 2007 A method for fast generation of bivariate poisson random vectors. Proceeding of 2007, Winter Simulation Conference, USA.

Silva, G., Dean, C., Niyonsenga, T. & Vanessa, A. 2008 Hierarchical bayesian spatio-temporal anaysis of revascularization odds using smoothing splines. *Statistics in Medicine* **27**, 2381–2401.

Sun, D., Speckman, P. & Tsutakawa, R. K. 1999 Random effects in generalized linear mixed models. Technical Report, National Institute of Statistical Sciences, USA.

Sun, T., Kim, H. & He, Z. 2000 Spatio-temporal interaction with disease mapping. *Statistics in Medicine* **19**, 2015–2035.

Vernic, R. 1997 On the bivariate generalized poisson distribution. *Astin* **27**, 1.

Vonesh, E. F. 1996 A note on the use of laplace's approximation for nonlinear mixed effects models. *Biometrika* **83**, 447–452.

Wall, M. 2004 A close look at the spatial structure implied by the car and sar models. *Journal of Statistical Planning and Inference* **121**, 311 324.

Wang, B. & Shi, J. 2014 Generalized gaussian process regression models for non-gaussian functional data. *JASA* **4**, 11–25.

Wood, D. 2013 Statistical modelling. APTS course, Nottingham University, United Kingdom.

Wood, S. N. 2012 Apts statistical computing. APTS course, Cambridge University, United Kingdom.

Yi, G., Shi, J. Q. & Choi, T. 2011 Penalized gaussian process regression and classification for high-dimensional nonlinear data. *Biometrics* **67**, 1285–1294.