University of Cincinnati

Date: 3/10/2022

I, Jieyan Zhang, hereby submit this original work as part of the requirements for the degree of Doctor of Philosophy in Mathematical Sciences.

It is entitled:

Bayesian Hierarchical Modeling for Dependent Data with Applications in Disease Mapping and Functional Data Analysis

Student's name: **Jieyan Zhang**

This work and its defense approved by:

Committee chair: Emily Kang, Ph.D.

Committee member: Won Chang, Ph.D.

Committee member: Bledar Konomi, Ph.D.

Committee member: Seongho Song, Ph.D.

41698

Bayesian Hierarchical Modeling for Dependent Data with Applications in Disease Mapping and Functional Data Analysis

A dissertation submitted to the

Graduate School

of the University of Cincinnati

in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

in the Department of Mathematical Sciences

of the College of Arts and Sciences

by

Jieyan Zhang

M.S. University of Cincinnati

April 2022

Committee:

Emily L. Kang, Ph.D., Chair

Won Chang, Ph.D.

Bledar A. Konomi, Ph.D.

Seongho Song, Ph.D.

Abstract

Bayesian hierarchical modeling has a long history but did not receive wide attention until the past few decades. Its advantages include flexible structure and capability of incorporating uncertainty in the inference. This dissertation develops two Bayesian hierarchical models for the following two scenarios: first, spatial data of time to disease outbreak and disease duration, second, large or high dimensional functional data that may cause computational burden and require rank reduction. In the first case, we use cucurbit downy mildew data, an economically important plant disease data recorded in sentinel plot systems from 23 states in the eastern United States in 2009. The joint model is established on the dependency of the spatially correlated random effects, or frailty terms. We apply a parametric Weibull distribution to the censored time to disease outbreak data, and a zero-truncated Poisson distribution to the disease duration data. We consider several competing process models for the frailty terms in the simulation study. Given that the generalized multivariate conditionally autoregressive (GMCAR) model, which contains correlation and spatial structure, provides a preferred DIC and LOOIC results, we choose the GMCAR model for the real data. The proposed joint Bayesian hierarchical model indicates that states in the mid-Atlantic region tend to have a high risk of disease outbreak, and in the infected cases, they tend to have a long duration of cucurbit downy mildew. The second Bayesian hierarchical model smooths functional curves simultaneously and nonparametrically with improved computational efficiency. Similar to the frequentist counterpart, principal analysis by conditional expectation,

the second model reduces rank through the multi-resolution spline basis functions in the process model. The proposed method outperforms the commonly used B-splines basis functions by providing a slightly better estimation within a much shorter computing time. The performance of this model is also examined using two real data sets, a sleeping energy expenditure data from an obesity study conducted in Baylor College of Medicine, and a human mortality data.

Acknowledgments

This dissertation would not be possible without support from many people. First of all, I would like to express my deepest appreciation to my advisor, Dr. Emily L. Kang, for her guidance throughout my Ph.D. study. She always responses in a timely manner, and her advice is inspirational. Whenever I feel stuck in my research, she listens to my concerns, looks into the issues, and offers alternative solutions with her experience and expertise. Not only working on theoretical knowledge and coding skills, she encourages me to present my research by bringing opportunities from on campus poster presentations to international conferences. She is also open to different research topics, which introduces me to new subjects outside my dissertation area. Under her guidance, I have become a curious, independent, and self-motivated Ph.D. student.

I would also like to thank Dr. Peter S. Ojiambo from the Department of Entomology and Plant Pathology, North Carolina State University. He provided the botanical epidemic data set that is used in the second chapter of this dissertation. He also made significant contribution to the data description and manuscript revision.

I am grateful to Dr. Won Chang, Dr. Bledar A. Konomi, and Dr. Seongho Song, who served on my advance exam and dissertation committee. Their insightful suggestions help me improve this dissertation from different perspectives and shed light on the direction of future research.

I sincerely thank my Procter & Gamble mentor, Dr. A. Narayanan. He showed me the

role that statisticians play in the industry. Not only working on the methodology, we also created web tools, so that the analysis results are accessible to all users regardless of their knowledge of programming. We have established quite a few web tools to meet different clients' needs and in different languages, R, JMP, SAS, and so on. It is such a valuable and delightful experience to work with him.

Last but not least, I would like to dedicate this dissertation to my beloved family, my parents, my husband and my son. Thank you for your unconditional support and endless love.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

Bayesian hierarchical modeling has been studied extensively over the recent few decades and has been applied to a broad range of science fields. A large number of models have been built to fulfill different purposes. This dissertation proposes two Bayesian hierarchical models. The first model is motivated by a real data from a botanical epidemiology study. The data has two parts, time to disease outbreak, or the survival data, and time that disease lasts, or the duration data. The goal of this study is to explore the effects of plant type and location on the disease. Unlike the first model that focuses on a specific data, the second model aims to smooth and estimate functional data curves with improved computational efficiency in general. The computational feasibility has growing importance nowadays due to the larger size and higher dimension of the data. The rank reduction in the second model is carried out through approximations using multi-resolution spline basis functions.

Both models in this dissertation has a three-level Bayesian hierarchical structure. It con-

sists of data model, process model, and parameter model. At the first level, the data model is specified based on the underlying process. This process is then modeled at the second level with estimated parameters. Finally, the parameters can be estimated either using empirical methods or from hyper prior distributions. Because the data model is conditional on the process model, and the process model is conditional on the parameter model, the nested structure refers to the Bayes' theorem,

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}, \tag{1.1}$$

with B be the fixed observation, and A from the underlying levels. Bayesian hierarchical model uses the fact that the posterior probability is proportional to the product of likelihood and prior probability. It is therefore able to incorporate uncertainty in the statistical inferences and also capture complicated model specification in a hierarchy of manageable model layers.

This chapter starts with a brief introduction of the related terms of the two models that are needed for the succeeding chapters.

1.1 Disease Mapping

For the plant pathology data, we model spatially distributed dependent variable in the random effects term. The spatial information is usually included in the model through either the two-dimensional coordinates of the location or the adjacency pattern. The first method is called geostatistical modeling by Cressie (1993). The spatial correlation is based

on the Euclidean distance between two geographic locations. While the first method assumes continuous space, the second method treats the space as discrete lattice, where neighbouring locations have a higher correlation. For both cases, the spatial autocorrelation has the assumption that

$$\boldsymbol{\phi} \sim \mathcal{N}\Big(\boldsymbol{\mu}, \boldsymbol{\Sigma}\Big), \tag{1.2}$$

where $\boldsymbol{\Sigma}$, a positive definite covariance matrix, can be derived from a distance decay function or a spatial weight matrix, respectively. Since state information is available in the plant pathology data, we apply the lattice modeling method in the process model.

When a neighbouring structure is exhibited in the data, conditional autoregressive (CAR) models are frequently used to show the spatial correlation. Let $\boldsymbol{\phi} = (\phi_1, ..., \phi_I)'$ be the spatial random variable from I areal units, each ϕ_i is conditional on a weighted sum of its neighbouring locations with unknown variance τ_i^2,

$$\phi_i \mid \boldsymbol{\phi}_{(-i)} \sim \mathcal{N}(\sum_{l=1}^{I} w_{il}\phi_l, \tau_i^2), \tag{1.3}$$

$\boldsymbol{\phi}_{(-i)} = (\phi_1, ..., \phi_{i-1}, \phi_{i+1}, ..., \phi_I)'$. The spatial relationship is an $I \times I$ binary matrix with 1 for neighbours and 0 otherwise, and notated as $\boldsymbol{W}$ for weights, or sometimes $\boldsymbol{A}$ for adjacency. In CAR models, the matrix is symmetric with diagonal elements 0, which means that an areal unit is not a neighbour of itself. A diagonal $I \times I$ matrix $\boldsymbol{D}$ is also available with element $\{i, i\}$ be the number of neighbours of location i and off-diagonal entries be 0. The CAR model will be generalized to a bivariate case to fit the two parts of the plant pathology data, where spatial correlation exists both within and between the two parts.

1.2 Functional Data Modeling

Functional data is used in the second model. It takes the form of functions or curves, $Z_1(t), ..., Z_n(t)$, and is continuous over a time interval $\mathcal{T}$. The functions can be treated as the realizations of one-dimensional stochastic process in a Hilbert space L^2 with mean function $\mu(t) = E(Z(t))$ and covariance function $\Sigma(t, t^*) = cov(Z(t), Z(t^*))$. Because real data is usually observed with measurement errors, a random noise on the trajectory is included in the model. We have

$$\begin{aligned} &Y_i(t) = Z_i(t) + \epsilon_i(t), \quad t \in \mathcal{T}, \\ &E(\epsilon_i(t)) = 0, \quad Var(\epsilon_i(t)) = \sigma^2. \end{aligned} \tag{1.4}$$

Functional data analysis (FDA) has been widely applied to various subjects, such as economics trends, environmental monitoring, medical science, and much more. The challenges include but not limited to, 1) the data is sometimes sparse or irregularly observed, 2) the data with measurement errors requires pre-smoothing step, and 3) the data can go up to infinite dimensions which brings theoretically difficulties and is computationally expensive. 84 FDA application articles were studied in a research by Ullah and Finch (2013). Smoothing and rank reducing methods were used in the majority of the papers. 72 studies (85.7%) applied some types of smoothing methods, with 25 of them were B-spline. 51 studies (60.7%) used functional principal component analysis (FPCA), a common tool for dimension reduction and data imputation when data is sparse.

B-spline basis functions form a basis for a function space in which any function can be

represented by a linear combination of the basis functions. The domain of the basis functions is divided by a nondecreasing knot vector. Each B-spline basis function is described as "local" since it is only non-zero on a subinterval of the domain. There are other basis functions available for smoothing and interpolation, including Fourier series basis functions, radial basis functions, and wavelets basis functions. Different methods are suitable under different circumstances. Because cubic B-spline basis functions are frequently used for Gaussian process data, they will be used to compare with the proposed automatically adaptive multi-resolution spline basis functions in the second half of this dissertation.

FPCA is performed to estimate initial values of the mean and covariance functions in the second model. It is a functional version of principal component analysis of multivariate data. It reduces dimensions for high dimensional data and interpolates for sparse data. FPCA is executed through the spectral decomposition of covariance function

$$\Sigma(t, t^*) = \sum_{k=1}^{\infty} \lambda_k \phi_k(t) \phi_k(t^*), \tag{1.5}$$

where λ_k are eigenvalues in descending order and ϕ_k are the corresponding orthogonal eigenfunctions. The functional curve can be expressed as

$$Z_i(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_{ik} \phi_k(t), \tag{1.6}$$

where the ξ_{ik} are uncorrelated random variables with $E(\xi_{ik}) = 0$ and $var(\xi_{ik}) = \lambda_k$. The infinity in the summation term is reduced to a large enough K, so that the first K terms provide estimation close enough to the infinite sum. With an increasing K, the estimation bias decreases, and the variance explained increases. FPCA through conditional expectation

(Yao et al., 2005) is available in R package `fdapace`. A suggested cutoff of K is to explain 99% variance of the singular value decomposition of the fitted covariance function. The estimated mean and covariance functions will be used in the second Bayesian hierarchical model.

1.3 Outline

The remainder of the dissertation proceeds as follows.

Chapter 2 starts with a description of the cucurbit downy mildew (CDM) data. CDM is an economically important disease that affects plants in the family of *Cucurbitaceae*. It exhibits annual extinction-colonization cycles and significant long distance spread at the continental scale in the United States. Time to disease outbreak and epidemic duration are often associated in some ways and understanding the nature of this association has important implications for risk assessment and managing plant disease epidemics. In this chapter, we develop a joint Bayesian hierarchical model with a parametric Weibull distribution for the censored time to disease outbreak data, and a zero-truncated Poisson distribution for the disease duration data in the data model, and a generalized multivariate conditionally autoregressive (GMCAR) model for the spatially correlated random effects in the process model. The proposed model is shown to have a smaller bias and outperform other three Bayesian hierarchical models with respect to model selection criteria, DIC and LOOIC, in the simulation study. We then apply the model to the real data and conclude that the mid-

Atlantic region tends to have a higher risk of CDM outbreak and a longer CDM duration than the South Central states.

Inspired by the multi-resolution spline basis functions for fixed rank kriging, Chapter 3 setup a Bayesian hierarchical framework with approximations by basis functions. The proposed model estimates mean and covariance functions simultaneously and nonparametrically with enhanced computing efficiency. The methodology is described both theoretically and practically at the beginning. It is then examined through two simulation studies, for one-dimensional and two-dimensional settings, respectively. Comparing with the the model using widely accepted B-splines basis functions, the results of our model show improvement in terms of root mean square error and computing time and memory. Two real data sets are used in this chapter. The first one is a one-dimensional sleeping energy expenditure (SEE) data that was collected for an obesity study in the Children's Nutrition Research Center of Baylor College of Medicine. The smoothed data from the proposed model is proven to have a smaller misclassification rate and is therefore beneficial for follow-up studies. The second one is a two-dimensional mortality data from 12 countries, which is recorded in grid of age and year. We fit the model with half of the data and save the other half to test the performance of the model through root mean square error.

Chapter 4 concludes this dissertation with a brief summary and discussion.

Chapter 2

Joint Spatial Modeling of Time to Disease Outbreak and Epidemic Duration in Risk Assessment of Botanical Epidemics

2.1 Introduction

Long-distance spread of invasive plant pathogens negatively impacts ecosystem function (Crowl et al., 2008) and influences policy decisions for managing resultant disease epidemics. Growth in global trade has been cited as one of the factors responsible for the increase in

the frequency of these invasive pathogens. If and when such epidemics occur, containment and eradication programs are implemented to contain these invasions and in some cases, specific measures may also be developed to prevent potential incursions. Understanding processes and factors that affect biological invasions in space and time is thus important to establish and predict the risk of biological invasions (Madden et al., 2007). Knowledge generated thereof can facilitate planning and designing effective measures to eradicate and contain these invasions (Zadoks and Van den Bosch, 1994). Additionally, spatial effects may identify areas that require detailed epidemiological research to inform disease intervention.

Disease surveillance programs are routinely used to monitor outbreak of epidemics caused by new or invasive pathogens that are aerially transmitted and fundamental to these programs are data collected from sentinel surveillance systems (Edmond et al., 2011; Randrianasolo et al., 2010). In botanical epidemiology, sentinel surveillance systems typical consist of fixed plots, generally planted early across the landscape and designated a priori for regular surveillance within the disease monitoring network (Christiano and Scherm, 2007; Ojiambo and Holmes, 2011). Systematic information is collected from the sentinel sites over time, while disease characteristics are assessed in all locations within the monitoring network. Consequently, records of disease outbreaks in these sentinel surveillance systems result in time to disease outbreak data that can be used to model subject-specific fixed effects due to various covariates (Ojiambo and Kang, 2013). Time to disease outbreak in sentinel sites closer together may be more similar than in sites that are farther apart, which results in spatial autocorrelation within the data. Thus, random effects are incorporated in time-to-event

models within a spatial framework to account for unobserved heterogeneity in the event time (Ojiambo and Kang, 2013). Using time to disease data from a cucurbit downy mildew (CDM) surveillance in the United States, we observed that a Bayesian spatially structured frailty conditionally autoregressive (CAR) model provided a better fit to the data than either the unstructured frailty model or a model without frailty (Ojiambo and Kang, 2013). In addition, regions with low or high risk of disease outbreak were identified whereby states in the mid-Atlantic region were usually associated with high risk of CDM outbreak (Ojiambo and Kang, 2013).

Time to disease outbreak at a sentinel site and epidemic duration in the neighbouring locations within the monitoring network affect each other. A later outbreak implies less time for disease duration due to seasonal cycle of plant growth, while an earlier outbreak makes a longer duration of the epidemic possible. Conversely, sentinel sites with a longer disease duration have an increased opportunity for pathogen reproduction, multiplication and inoculum dispersal and will increase the risk of disease outbreak in disease-free sentinel sites in surrounding locations compared to sentinel sites with a shorter epidemic duration. While time to disease outbreak and disease duration data can be modelled separately, classic models such as standard hazards models for time to disease outbreak and linear mixed models for disease duration data do not consider dependencies between these two different types of data which may lead to inefficient or biased results. Joint models for time to disease outbreak and disease duration are robust methods that takes into consideration the dependency and association between these two different types of data. A joint modeling approach confers

several advantages including allowing simultaneous investigation of the effects of covariates on these two different types and avoiding overestimation or underestimation of the impact of an intervention in disease control by providing valid and efficient inferences.

Unlike on longitudinal measurement (He and Luo, 2016; Ibrahim et al., 2010; Wu et al., 2012; Zhang et al., 2017), joint modeling method has rarely been applied to survival and duration data in botanical epidemiology to estimate risk of outbreak in plant disease surveillance systems (Nathoo, 2010). In this chapter, we look into a considerable amount of Bayesian hierarchical modeling literature (Lawson et al., 2014; Nathoo, 2010; Zhou et al., 2008) and develop a joint modeling framework for time to disease outbreak and disease duration to estimate the risk of disease in sentinel surveillance plant disease systems. Our joint modeling approach consists of a parametric survival model for the time-to-event outcomes and a truncated Poisson model for the disease duration. Both models have a hierarchical structure that incorporate individual-specific covariates and state-specific spatially correlated random effects since random effects for states in closer proximity to each other tend to be similar. CAR model (Besag, 1974; Carlin et al., 2003; Gelfand and Vounatsou, 2003; Kim et al., 2001) is frequently used to describe the spatial dependence (Banerjee et al., 2003; Neelon et al., 2013, 2014; Ojiambo and Kang, 2013). Because regions with short time to disease outbreak are likely to have long disease duration, we model the spatial random effects in both the survival and the truncated Poisson model components jointly via a generalized multivariate CAR (GMCAR) model (Jin et al., 2005), which is able to describe dependence between these two model components. We then consider several competing process models

and select among them the best fitting model using two criteria: the deviance information criterion, DIC (Spiegelhalter et al., 2002) and leave-one-out cross validation information criterion, LOOIC (Vehtari et al., 2017).

The reminder of the chapter proceeds as follows. In Section 2.2, we describe the source of data used in the project and highlight key attributes and spatial extent of the data set. In Section 2.3, we outline the proposed joint modeling development in terms of time-to-disease outbreak and disease duration data and compare a few possible process models for the frailty terms. The Bayesian implementation details and model selection criteria are given in Section 2.4. We apply the final model to the simulated data in Section 2.5 and the 2009 CDM data in Section 2.6. Section 2.7 concludes with a brief discussion and summary of our major findings as they potential relate to policy intervention for botanical epidemics exhibiting long distance spread at a landscape level.

2.2 Data Description

Cucurbit downy mildew is caused by the obligate oomycete *Pseudoperonospora cubensis*, an economically important pathogen that affects plants within the family *Cucurbitaceae* (Ojiambo et al., 2015). The pathogen exhibits significant long distance dispersal at the landscape level (Ojiambo and Holmes, 2011), and since its hosts are sensitive to frost, incursions of the pathogen into northern latitudes in the United States occur annually from subtropical overwintering areas in southern Florida (Ojiambo et al., 2015). These annual extinction-

colonization cycles of the pathogen in northern latitudes provides a useful framework to examine the spatial and temporal dynamics of a disease resulting from a pathogen that exhibits long distance dispersal (Ojwang et al., 2021).

Disease epidemics recorded in the United States in 2009 were analyzed using data obtained from the CDM ipmPIPE program (Ojiambo et al., 2011). The CDM ipmPIPE is part of the United States Department of Agriculture Pest Information Platform for Extension and Education (PIPE) program. Records of confirmed outbreaks of CDM were collected as part of the sentinel and non-sentinel plot monitoring network designed to alert growers on the risk of CDM outbreak in the fields. Sentinel plots were fixed plots, planted early and designated for weekly surveillance, while non-sentinel plots consisted of commercial fields, research plots, and home gardens. A total of 85 sentinel plots from 23 states in the eastern United Sates were monitored in 2009. A total 107 counties were affected by CDM in the 2009 epidemic. Besides spatial information and time of disease outbreak, information on cucurbit host types was also included in the data. The host types in both sentinel and non-sentinel plots were classified into three groups: cucumber, squash and other cucurbit species. The latter host type category was composed of watermelon, cantaloupe, and pumpkin. The total observation period for disease monitoring was 204 days. Data is recorded as censored when disease outbreak is not observed within the range. In this study, epidemic duration (days) refers to the length of time from when CDM is first reported on a host to when the host becomes completely necrotic and is not able to produce inoculum to infect hosts in neighbouring fields.

The bar graph in Figure 2.1 shows the number of censored and uncensored records of each cucurbit host type in each state. Both censored and uncensored data were highest in Florida and North Carolina, and lowest in Massachusetts and Wisconsin. Further, disease outbreaks were also not observed for certain host type and state combinations. For example, uncensored data is not available in cucumber fields in Alabama, and in fields of other host species in Louisiana. Six of the 23 states are chosen to show more details of the censoring data in Figure 2.2. The upper panel is the box plot of survival time of the uncensored data, which is days until disease outbreak within the 204 days period. When CDM is observed, the corresponding duration days is recorded as well. Figure 2.2 lower panel is the box plot of the disease duration. The count of uncensored survival and duration data match in most cases, for instance, one CDM outbreak is observed in other species fields in Delaware in Figure 2.2 upper panel, and therefore, one duration record is found in the same place in Figure 2.2 lower panel. However, there are very few special cases, where disease outbreak is observed, but epidemic duration information is not provided. For example, there were five CDM outbreaks in Alabama for which the epidemic duration is not available due to lack of information on the length over which these epidemics remained infective. Consequently, fields in duration data is a subset of the fields in uncensored survival data.

A total of 551 possible cases (i.e., host type by sentinel or non-sentinel plot combination) are monitored for 204 days in the 2009 CDM impPIPE data. Downy mildew outbreaks are observed in 30.1% of the cases examined. The outbreak rates are 28.5% for the 158 cucumber cases, 47.4% for the 154 squash cases, and 20.1% for the 239 other hosts cases.

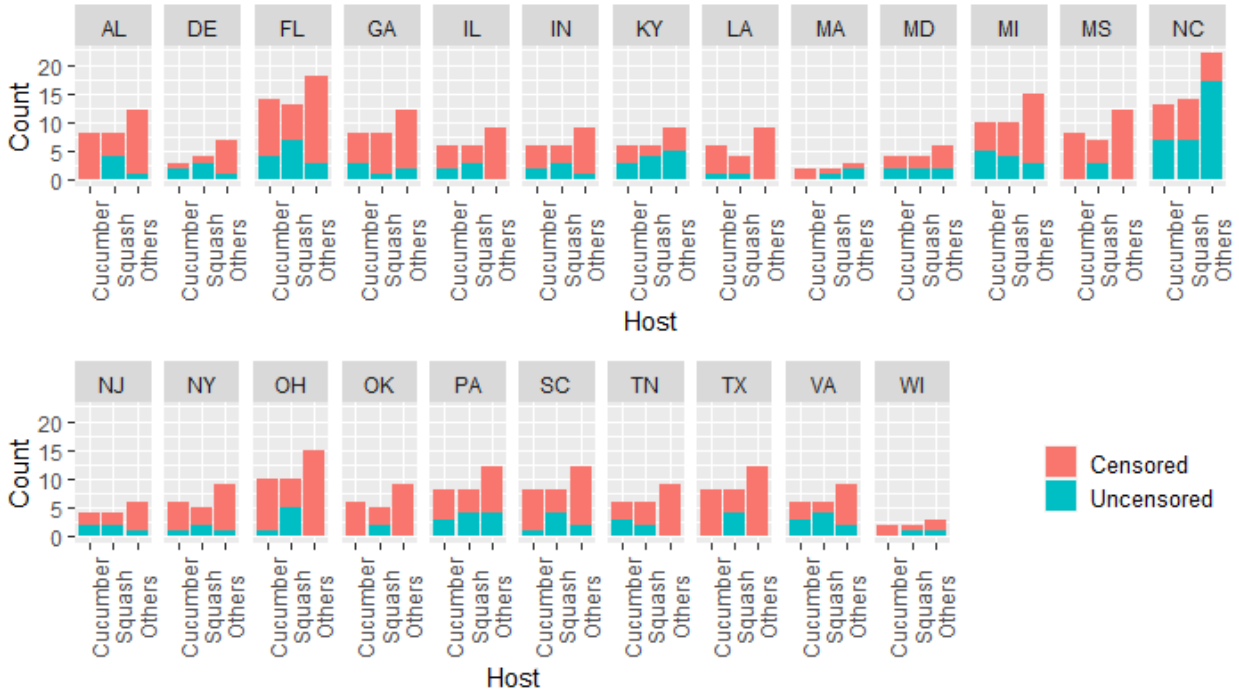


Figure 2.1: Bar graph of count of the censoring data (204 days)

Disease duration is recorded for 75.3% of the infected cases. Daily infected cases and number for each host are shown in Figure 2.3. Seasonal epidemics development, featured by a low and sporadic start, rapid increase in the middle and a gradual decrease toward the end (Ojiambo and Kang, 2013), is shown in the plot. Outbreaks of CDM increased rapidly for squash cases after day 100. Numbers of infected cucumber and other cases reached a climax at around 140 and 160, respectively. The epidemic peak lasted about 50 days for squash, and around 20-30 days for the rest hosts. Outbreak of new disease cases gradually declined and stopped after 190 days, due to the start of unfavorable weather for the CDM pathogen late in the fall season..

With censored data recorded as 204, the average survival and duration days of each

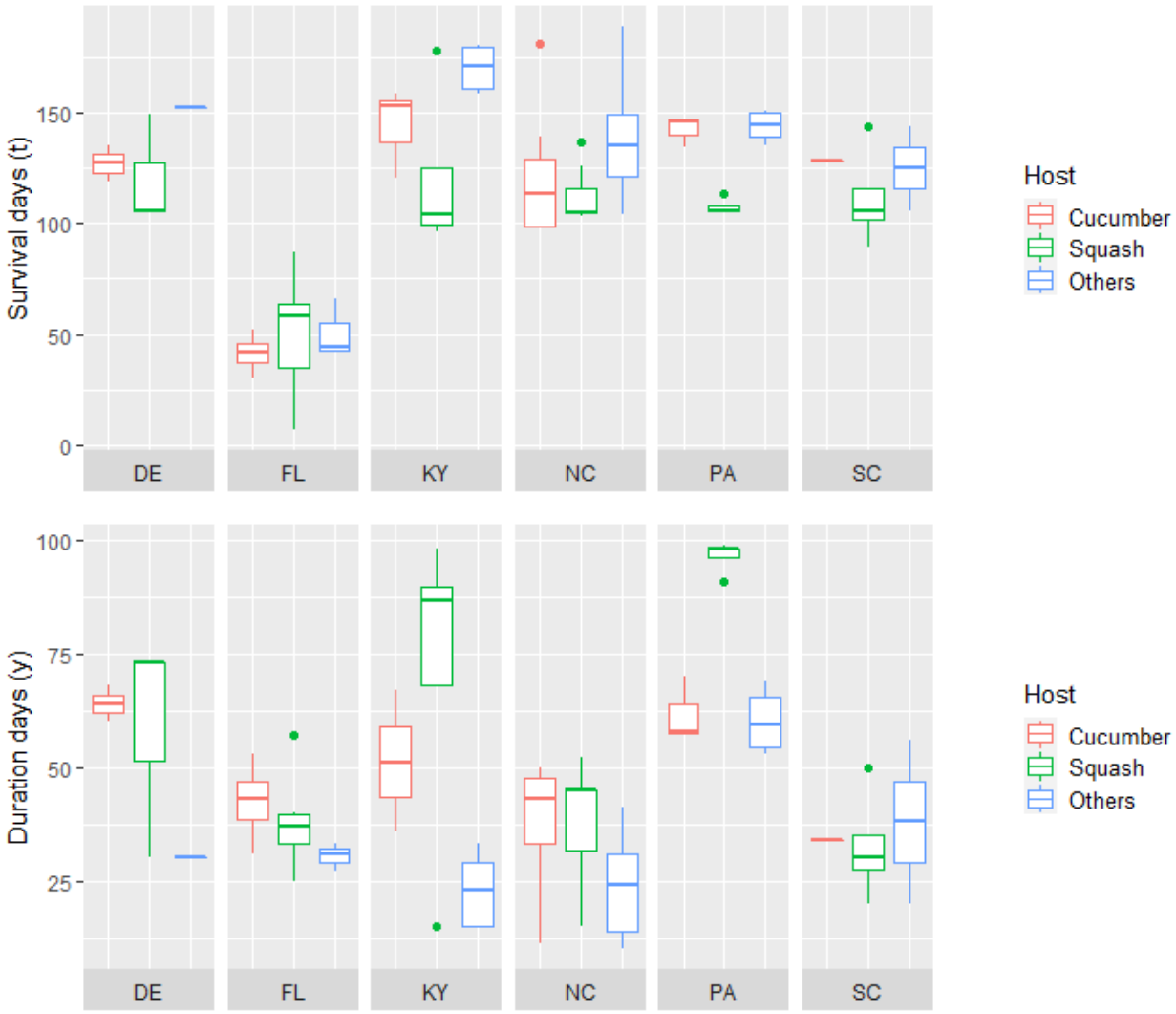


Figure 2.2: Box plots of uncensored survival days (upper panel) and disease duration days (lower panel)

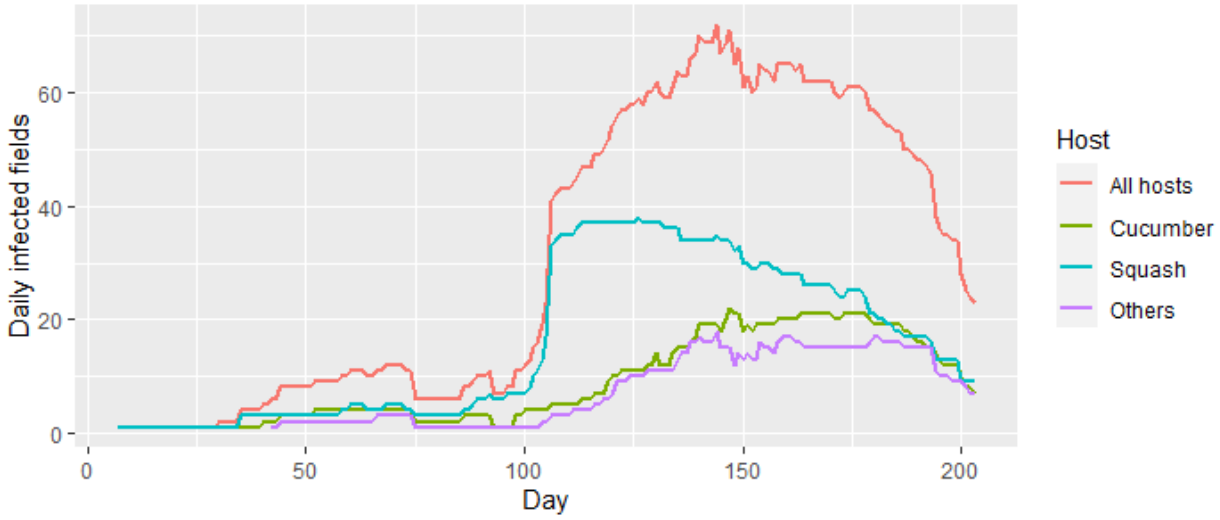


Figure 2.3: 2009 daily number of cucurbit downy mildew infected fields

plant host are shown in the maps in Figure 2.4. The three rows indicate the plant hosts, cucumber, squash and other hosts. Disease cases on squash plants, in general, have a shorter survival time and longer epidemic duration, compared with the other two cucurbit hosts. For example, squash cases in Florida are among the earliest outbreak at around 121.1 days, and squash cases in Illinois have the longest duration of 110.7 days. For cucumber and other hosts, several states have no disease outbreak recorded during the observation period and these states include Texas, Louisiana, Wisconsin, and Illinois. As a result, survival data in these states is shown as 204 in the survival map in the left panel, and the corresponding epidemic duration is missing in map in the right panel. Due to the large amount of states with censoring data only, the geographic pattern is not easily discernible. However, we can still see that the outbreak of CDM in east coast and especially Mid-Atlantic region where the CDM tends to have longer duration, is more severe than along the Gulf of Mexico.

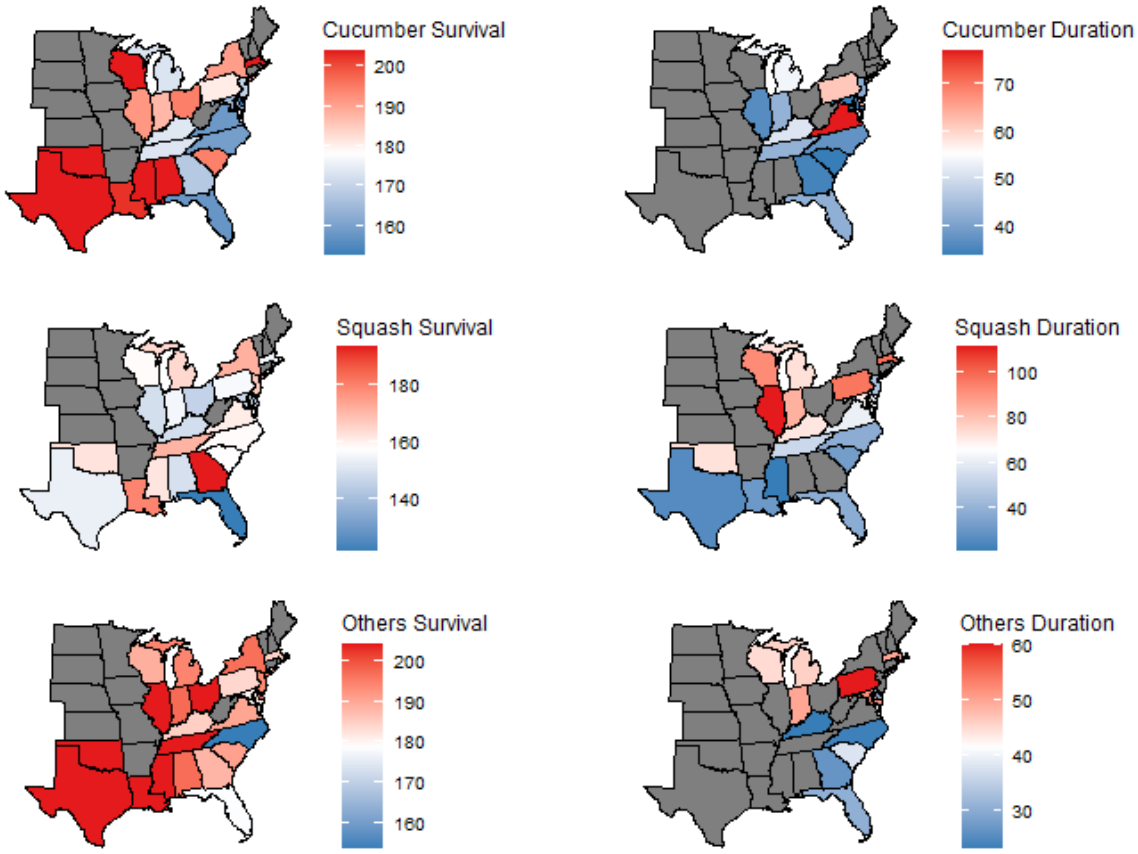


Figure 2.4: Map of average survival days with censored data recorded as 204 (left), and average duration days (right)

2.3 Model

In this section, we build a Bayesian hierarchical model to analyze the two components of the CDM impPIPE data, i.e., the time until disease outbreak (or survival time), t_{ij}, and where it occurs, the duration of the epidemic, y_{ij}. $i = 1, ..., I$ is the index of state, $j = 1, ..., n_i$ is the index of subject from state i. A two-part data model is used to fit these two types of information. Covariates in the data model indicates different host types, while random effects are spatial frailties defined at the state level. Four process models are used to fit

the random effects and will be compared in Section 2.5. Parameter model lists the prior distributions for the parameters. The three levels of the hierarchical structure, i.e., data model, process model, and parameter model, are presented below in the same order.

2.3.1 Data Model

Survival Model

To fit the survival time until disease outbreak, we adopted the widely used proportional hazard model. Let t_{ij} be the survival time or censoring for subject j in state i, $f(t)$ and $F(t)$ are the probability density function (p.d.f.) and the cumulative distribution function (c.d.f) of the random variable t_{ij}, respectively, the hazard at time t_{ij}, defined as $h(t_{ij}) = f(t)/(1 - F(t))$, has the multiplicative form

$$h(t_{ij} \mid \boldsymbol{x}_{ij}) = h_0(t_{ij}) exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \phi_i), \tag{2.1}$$

where $\boldsymbol{x}_{ij}$ is a vector of individual-specific covariates, $\boldsymbol{\beta}$ are the corresponding coefficients, ϕ_i are the state-specific random effects. The first term on the right hand side, $h_0(t_{ij})$, represents the baseline hazard. The second term, $exp(\boldsymbol{x}'_{ij}\boldsymbol{\beta} + \phi_i)$, is relative risk, the increase/decrease of hazard for each subject. The relative risk depends on subject and state only, thus it remains the same throughout the epidemic period.

We adopt the Weibull model from Banerjee et al. (2003). The hazard function of this model is parametric and can be written as

$$h(t_{ij} \mid \boldsymbol{x}_{1ij}) = \rho t_{ij}^{\rho-1} exp(\boldsymbol{x}'_{1ij}\boldsymbol{\beta}_1 + \phi_{1i}). \tag{2.2}$$

The subscript $_1$ indicates the first part of the data model, the survival model. We save $\boldsymbol{x}_{2ij}$, $\boldsymbol{\beta}_2$ and ϕ_{2i} for the duration model in the second part. $\rho > 0$ is the shape parameter of the Weibull distribution. The log hazard function is a linear regression on log time with slope $\rho - 1$. In other words, the hazard monotonically increases when $\rho > 1$, decreases when $0 < \rho < 1$, and is constant when $\rho = 1$. Let γ_{ij} be the disease indicator for subject j in state i, $\gamma_{ij} = 0$ if not infected before censoring time, $\gamma_{ij} = 1$ if infected. Then the likelihood function is

$$\begin{aligned} \boldsymbol{L}(t_{ij}, \boldsymbol{x}_{1ij}, \gamma_{ij} \mid \rho, \boldsymbol{\beta}_1, \phi_{1i}) &= \prod_{i=1}^{I} \prod_{j=1}^{n_i} h(t_{ij})^{\gamma_{ij}} (1 - F(t_{ij})) \\ &\propto \prod_{i=1}^{I} \prod_{j=1}^{n_i} \Big\{ \rho t_{ij}^{\rho-1} exp(\boldsymbol{x}'_{1ij}\boldsymbol{\beta}_1 + \phi_{1i}) \Big\}^{\gamma_{ij}} exp\Big\{ - t_{ij}^{\rho} exp(\boldsymbol{x}'_{1ij}\boldsymbol{\beta}_1 + \phi_{1i}) \Big\}. \end{aligned} \tag{2.3}$$

Duration Model

Disease duration is not available for the censored data. For the other cases, when disease occurs before censoring, or to say, the disease indicator $\gamma_{ij} = 1$, the data also has information about the duration of the epidemic. Let y_{ij} be the disease duration days for subject j in state i, y_{ij} does not exist if $\gamma_{ij} = 0$. When disease outbreak is observed, the disease duration has to be positive, and therefore, $y_{ij} = 0$ is excluded from the duration model. A zero-truncated Poisson distribution is applied to y_{ij},

$$p(y_{ij} \mid \lambda_{ij}) = \begin{cases} 0, & \text{if } y_{ij} = 0, \\ \frac{\lambda_{ij}^{y_{ij}} exp(-\lambda_{ij})}{y_{ij}!(1-exp(-\lambda_{ij}))}, & \text{if } y_{ij} = 1, 2, ... \end{cases} \tag{2.4}$$

The parameter λ_{ij} is further modeled using log link function,

$$log(\lambda_{ij}) = \boldsymbol{x}'_{2ij}\boldsymbol{\beta}_2 + \phi_{2i}. \tag{2.5}$$

$\boldsymbol{x}_{2ij}$ is a vector of individual-specific covariates. We have $\boldsymbol{x}_{1ij} = \boldsymbol{x}_{2ij}$ in the real data, which means that the plant host covariate in the survival and the duration model are identical, although this is in general not necessary. Similar as in the survival model above, $\boldsymbol{\beta}_2$ is the corresponding coefficients, ϕ_{2i} is the state-specific random effects of the duration model. Recall that γ_{ij} is the disease indicator, likelihood function therefore is

$$\boldsymbol{L}(y_{ij}, \boldsymbol{x}_{2ij}, \gamma_{ij} \mid \boldsymbol{\beta}_2, \boldsymbol{\phi}_{2i}) = \prod_{i=1}^{I}\prod_{j=1}^{n_i}\Big(\mathbb{1}_0(\gamma_{ij}) + \mathbb{1}_1(\gamma_{ij})p(y_{ij})\Big), \tag{2.6}$$

where $\mathbb{1}$ is the indicator function defined as $\mathbb{1}_c(\gamma_{ij}) = 1$ if $\gamma_{ij} = c$, and 0 otherwise.

2.3.2 Process Model

Generalized multivariate conditional autoregressive (GMCAR) model is used as the process model for the state-specific random effects, or frailty terms, $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$, in the second stage of the hierarchical structure. In addition, three other models, univariate conditional autoregressive (UniCAR) model, multivariate normal (MvNorm) model, and independent normal (IndNorm) model, are also applied to the frailty terms as comparison models. The classification of the four models is shown in Table 2.1. Models in the first row of the table have no geographical information. Models in the first column assume independence between $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$. Therefore, IndNorm model in the upper left corner is the simplest model and has the

	Independent	Dependent
Non-spatial	IndNorm Model	MvNorm Model
Spatial	UniCAR Model	GMCAR Model

Table 2.1: Four process models comparison

fewest parameters, while GMCAR model in the lower right corner is relatively complicated and has the most parameters among the four models. The GMCAR model is chosen because of its capability to incorporate more information without overfitting. Detailed results of model comparison are discussed in Section 2.5.2 for the simulated data and in Section 2.6 for the 2009 CDM ipmPIPE data.

GMCAR Model

Conditional autoregressive (CAR) model was first introduced by Besag (1974) and has received increasing attention over the past decades. Taking $\boldsymbol{\phi} = (\phi_1, ..., \phi_I)$, the local dependence is shown in the fully conditional distribution,

$$\phi_i \mid \boldsymbol{\phi}_{(-i)} \sim \mathcal{N}(\sum_{l=1}^{I} c_{il}\phi_l, \tau_i^2), \tag{2.7}$$

where $i, l = 1, ..., I$, $\boldsymbol{\phi}_{(-i)} = (\phi_1, ..., \phi_{i-1}, \phi_{i+1}, ..., \phi_I)'$, $c_{ii} = 0$, and $c_{il} = 0$ if state i and l are not next to each other.

A popular implementation of CAR is the intrinsic autoregressive (IAR) model (Besag et al., 1991). The IAR model takes $c_{il} = w_{il}/w_{i.}$ and $\tau_i^2 = \tau^2/w_{i.}$, where $w_{i.} = \sum_{l=1}^{I} w_{il}$. As before, $w_{ii} = 0$, $w_{il} = 0$ if state i and l are not next to each other. For the IAR model, the fully conditional distribution in (2.7) becomes $\phi_i \mid \boldsymbol{\phi}_{(-i)} \sim \mathcal{N}(\sum_{l=1}^{I} \frac{w_{il}}{w_{i.}}\phi_l, \frac{\tau^2}{w_{i.}})$. By

Brook's Lemma (Besag, 1974), a unique multivariate normal distribution with mean $\mathbf{0}$ and covariance matrix $[\tau(\boldsymbol{D}-\boldsymbol{W})]^{-1}$ can be obtained from the fully conditional distribution. Here, $\boldsymbol{W}=\{w_{il}\}_{i,l=1}^{I}$, $\boldsymbol{D}$ is an $I\times I$ diagonal matrix with diagonal elements $w_{i.}$. Typically, w_{il} is set to be 1, if state i and l are neighbors, and 0 otherwise. $\boldsymbol{W}$ is the adjacency matrix of the map. $w_{i.}$ is the total number of neighbors of state i. Since $\boldsymbol{D}-\boldsymbol{W}$ is singular, and thus non-invertible, the multivariate normal distribution is in fact improper. A smoothing parameter, α, is then introduced to remedy this issue. Taking $[\tau(\boldsymbol{D}-\alpha\boldsymbol{W})]^{-1}$ as the new variance-covariance matrix, a proper joint distribution is guaranteed if $|\alpha|<1$ (Carlin et al., 2003).

Similar to the univariate fully conditional distribution in (2.7), the multivariate counterpart is introduced by Mardia (1988) and takes the form

$$\boldsymbol{\nu}_i \mid \boldsymbol{\nu}_{(-i)} \sim \mathcal{N}(\sum_{l=1}^{I}\boldsymbol{B}_{il}\boldsymbol{\nu}_l, \boldsymbol{\Gamma}_i). \tag{2.8}$$

Note that $\boldsymbol{\nu}_i$ denotes a p dimension vector of $(\phi_{1i},\phi_{2i},...,\phi_{pi})'$, $i=1,...,I$, whereas $\boldsymbol{\phi}_k=(\phi_{k1},...,\phi_{kI})'$, $k=1,...,p$. The two vectors both consist of frailty terms, but are in different dimensions. $\boldsymbol{\nu}_{(-i)}=(\boldsymbol{\nu}_1',...,\boldsymbol{\nu}_{i-1}',\boldsymbol{\nu}_{i+1}',...,\boldsymbol{\nu}_I')'$. $\boldsymbol{B}_{il}$ and $\boldsymbol{\Gamma}_i$ are $p\times p$ matrices. The joint multivariate CAR model, denoted as MCAR$(\alpha,\boldsymbol{\Lambda})$ (Carlin et al., 2003), can then be generalized from the univariate case,

$$\boldsymbol{\nu}\sim\mathcal{N}(\mathbf{0},[(\boldsymbol{D}-\alpha\boldsymbol{W})\otimes\boldsymbol{\Lambda}]^{-1}), \tag{2.9}$$

$\boldsymbol{\nu}=(\boldsymbol{\nu}_1',\boldsymbol{\nu}_2',...,\boldsymbol{\nu}_I')'$, α is again the smoothing parameter, $\boldsymbol{\Lambda}$ is a $p\times p$ symmetric and positive definite matrix. The Kronecker product of $\boldsymbol{D}-\alpha\boldsymbol{W}$ and $\boldsymbol{\Lambda}$ in the precision matrix ensures

the propriety of the distribution, and shows a separable structure of the model. Taking $p = 2$ as an example, the bivariate CAR model can be written as

$$\begin{pmatrix} \boldsymbol{\phi}_1 \\ \boldsymbol{\phi}_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} (\boldsymbol{D} - \alpha \boldsymbol{W})\Lambda_{11} & (\boldsymbol{D} - \alpha \boldsymbol{W})\Lambda_{12} \\ (\boldsymbol{D} - \alpha \boldsymbol{W})\Lambda_{12} & (\boldsymbol{D} - \alpha \boldsymbol{W})\Lambda_{22} \end{pmatrix}^{-1} \right). \tag{2.10}$$

$\boldsymbol{\phi}_1 = (\phi_{11}, \phi_{12}, ..., \phi_{1I})'$, $\boldsymbol{\phi}_2 = (\phi_{21}, \phi_{22}, ..., \phi_{2I})'$. $\mathbf{0}$ is an I dimensional vector of all 0's.

In reality, the correlation between two frailties and the correlation between two states are very likely to be different. As a result, more than one smoothing parameter, α, may be needed to show the difference in correlation within and between $\boldsymbol{\phi}$'s. Several multivariate CAR models are developed to include more information in the model, such as 2fCAR $(\alpha_0, \alpha_1, \alpha_2, \alpha_3, \tau_1, \tau_2)$ (Kim et al., 2001), MCAR $(\alpha_1, \alpha_2, \Lambda_{11}, \Lambda_{12}, \Lambda_{22})$ (Carlin et al., 2003; Gelfand and Vounatsou, 2003) and so on. Jin et al. (2005) proposed a generalized MCAR (GMCAR) model to directly specify marginal and conditional distribution,

$$\begin{aligned} \boldsymbol{\phi_1} &\sim \mathcal{N}\Big(\mathbf{0}, [\tau_1(\boldsymbol{D} - \alpha_1 \boldsymbol{W})]^{-1}\Big), \\ \boldsymbol{\phi_2} \mid \boldsymbol{\phi_1} &\sim \mathcal{N}\Big((\eta_0 \boldsymbol{I} + \eta_1 \boldsymbol{W})\boldsymbol{\phi_1}, [\tau_2(\boldsymbol{D} - \alpha_2 \boldsymbol{W})]^{-1}\Big). \end{aligned} \tag{2.11}$$

$\boldsymbol{I}$ is an $I \times I$ identity matrix. $\boldsymbol{\phi}_1$ is the univariate CAR model as mentioned before. η_0 and η_1 are called bridging parameters. Conditional on $\boldsymbol{\phi}_1$, $\boldsymbol{\phi}_2$ has a conditional mean of weighted average of $\boldsymbol{\phi}_1$, where η_0 is the weight from the same state, η_1 is the weight from neighborhood. Spatial information is thus included in, not only the precision matrix, but also the conditional mean. α_1 and α_2 are the smoothing parameters and are restricted to be between $(0, 1)$ to ensure the positive definiteness of the variance-covariance matrix. τ_1 and τ_2 controls the precision scale of the marginal and conditional distribution, respectively.

We choose the conditional order $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi_2} \mid \boldsymbol{\phi_1}$, rather than $\boldsymbol{\phi}_2$ and $\boldsymbol{\phi_1} \mid \boldsymbol{\phi_2}$ for this particular real data for a couple of reasons. First, the causality is more obvious in the real life this way than the other way around. Intuitively, the survival time to disease outbreak ($t_{ij} \mid \phi_{1i}$) affects disease duration ($y_{ij} \mid \phi_{2i}$). Disease outbreaks that occur early in a state increase opportunity for pathogen reproduction and lead to a longer duration time in this state and its neighbouring states. While outbreaks that occur later limit the time for disease outbreak in neighboring states due to unfavorable weather condition. Second, as mentioned in Section 2.2, duration information from certain state/host combination is not available in 2009 CDM ipmPIPE data. Therefore, fitting survival model first is reasonable with respect to computational efficiency. However, if another data set is used to fit the hierarchical model, the two conditional orders can perform equally well, which will be shown in Section 2.5.2.

Alternative Process Models

Three other models are applied to the frailty terms. The first one consists of two separate univariate CAR (UniCAR) models. $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$ are assumed to be independent and can be written as

$$\begin{aligned} \boldsymbol{\phi}_1 &\sim \mathcal{N}(\mathbf{0}, [\tau_1(\boldsymbol{D} - \alpha_1 \boldsymbol{W})]^{-1}), \\ \boldsymbol{\phi}_2 &\sim \mathcal{N}(\mathbf{0}, [\tau_2(\boldsymbol{D} - \alpha_2 \boldsymbol{W})]^{-1}). \end{aligned} \tag{2.12}$$

It is actually a special case of GMCAR with $\eta_0 = \eta_1 = 0$.

The second model has no geographic information, but it takes the correlation of the two random effects from the same state, ϕ_{1i} and ϕ_{2i}, into account. It has a bivariate normal

distribution setting, $\begin{pmatrix} \phi_{1i} \\ \phi_{2i} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12}^2 \\ \sigma_{12}^2 & \sigma_2^2 \end{pmatrix} \right)$. Putting the frailty terms together, the multivariate normal (MvNorm) model is denoted as,

$$\begin{pmatrix} \boldsymbol{\phi}_1 \\ \boldsymbol{\phi}_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_{12}^2 \\ \sigma_{12}^2 & \sigma_2^2 \end{pmatrix} \otimes \boldsymbol{I} \right). \tag{2.13}$$

Taking $\sigma_{12}^2 = 0$, we have the most straightforward setting of the random effects. ϕ_{ki} is independent and identically distributed (i.i.d.). $\phi_{ki} \sim \mathcal{N}(0, \sigma_k^2)$, $k = 1, 2$, $i = 1, ..., I$. Equivalently,

$$\begin{aligned} \boldsymbol{\phi}_1 &\sim \mathcal{N}(\mathbf{0}, \sigma_1^2 \boldsymbol{I}), \\ \boldsymbol{\phi}_2 &\sim \mathcal{N}(\mathbf{0}, \sigma_2^2 \boldsymbol{I}). \end{aligned} \tag{2.14}$$

2.3.3 Parameter Model

The prior distributions of all the parameters are specified to complete the Bayesian hierarchical model. Typically, weakly informative priors are assumed to allow the data to play the principal role in parameter estimation. Starting from the data model, a vague and proper $Gamma(1, 0.001)$ prior (Zhou et al., 2008), with mean 1000 and variance 10^6, is chosen for the shape parameter ρ in the survival model. Weakly informative Gaussian priors are assigned to the coefficients, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ (Neelon et al., 2013). For the spatial process models (GMCAR and UniCAR), we adopt the prior setting by Jin et al. (2005), Gaussian prior with mean 0 and variance 10^4 for bridging parameters η_0 and η_1, and $Gamma(1, 0.1)$ distribution with mean 10 and variance 100 for the scale parameters τ_1 and τ_2. Smoothing parameters,

α_1 and α_2, have $Unif(0,1)$ distribution to ensure the propriety of the distribution. In the non-spatial process models (MvNorm and IndNorm), standard deviation parameters have flat prior, $\sigma_1, \sigma_2 > 0$. Covariance matrix is constrained to be symmetric and positive definite.

2.4 Model Fitting

2.4.1 Bayesian Implementation

The joint posterior distribution of the Bayesian hierarchical model, from Bayes' theorem, is

$(processes,parameters \mid data)$

$$\propto (data \mid processes, parameters) \times (process \mid parameters) \times (parameters)$$

(Kang and Cressie, 2011). For the hierarchical model with the proposed GMCAR process model, we have

$$\begin{aligned} &p(\boldsymbol{\phi}_1, \boldsymbol{\phi}_2,\mu_1, \mu_2, \tau_1, \tau_2, \alpha_1, \alpha_2, \eta_0, \eta_1, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \rho \mid \boldsymbol{t}, \boldsymbol{y}) \\ &\quad \propto L(\boldsymbol{t} \mid \rho, \boldsymbol{\beta}_1, \boldsymbol{\phi}_1) L(\boldsymbol{y} \mid \boldsymbol{\beta}_2, \boldsymbol{\phi}_2) p(\boldsymbol{\phi}_1, \boldsymbol{\phi}_2 \mid \mu_1, \mu_2, \tau_1, \tau_2, \alpha_1, \alpha_2, \eta_0, \eta_1) \\ &\quad \times p(\mu_1)p(\mu_2)p(\tau_1)p(\tau_2)p(\alpha_1)p(\alpha_2)p(\eta_0)p(\eta_1)p(\boldsymbol{\beta}_1)p(\boldsymbol{\beta}_2)p(\rho). \end{aligned} \tag{2.15}$$

Recall that $\boldsymbol{t} = \{t_{ij}\}$, $\boldsymbol{y} = \{y_{ij}\}$, $i = 1, ..., I$, $j = 1, ..., n_i$ are the time to disease and disease duration data. The first two terms on the right-hand side are the likelihood functions from

(2.3) and (2.6). The joint distribution of $\boldsymbol{\phi}_1$ and $\boldsymbol{\phi}_2$ is

$$
\begin{aligned}
&p(\boldsymbol{\phi}_1, \boldsymbol{\phi}_2 \mid \mu_1, \mu_2, \tau_1, \tau_2, \alpha_1, \alpha_2, \eta_0, \eta_1) \propto \\
&\qquad \tau_1^{n/2} |\boldsymbol{D} - \alpha_1 \boldsymbol{W}|^{1/2} \times exp\{-\frac{\tau_1}{2} \boldsymbol{\phi}_1' (\boldsymbol{D} - \alpha_1 \boldsymbol{W}) \boldsymbol{\phi}_1\} \times \tau_2^{n/2} |\boldsymbol{D} - \alpha_2 \boldsymbol{W}|^{1/2} \times \\
&\qquad exp\{-\frac{\tau_2}{2} [\boldsymbol{\phi}_2 - (\eta_0 \boldsymbol{I} + \eta_1 \boldsymbol{W}) \boldsymbol{\phi}_1]' (\boldsymbol{D} - \alpha_2 \boldsymbol{W}) [\boldsymbol{\phi}_2 - (\eta_0 \boldsymbol{I} + \eta_1 \boldsymbol{W}) \boldsymbol{\phi}_1]\}.
\end{aligned} \tag{2.16}
$$

The remaining terms in (2.15) are from the prior distributions defined in Section 2.3.3.

The comparison of four process models is implemented in R (version 3.5.2) using package `RStan` (version 2.18.2), the R interface of `Stan`. `Stan` generates samples from the posterior distribution for inference by the No-U-Turn Sampler (NUTS), an extension to Hamiltonian Monte Carlo (HMC) algorithm. HMC is a Markov chain Monte Carlo (MCMC) method that introduces an auxiliary momentum variable and leapfrog updates to suppress the local random walk behavior. Therefore, HMC is able to reach the target distribution more rapidly than random walk Metropolis or Gibbs sampling. NUTS further avoids the need of choosing leapfrog step parameter in HMC, so that less hand-tuning is involved in model fitting (Hoffman and Gelman, 2014).

We run four parallel chains for each of the four models and within each chain, warm up iterations are used to estimate the mass matrix of the posterior and allow adjustment for correlated parameters. Trace plots, which is available in package `bayesplot`, are used to examine convergence visually. In addition, split $\hat{R}$ statistic is a generic diagnostic for consistency of the Markov chains. `RStan` splits each chain into two halves, and checks consistency by calculating the potential scale reduction statistic (Gelman et al., 1992). Let N be the

number of simulation draws after warm up in each chain, $\hat{R} = \sqrt{(\frac{N-1}{N}W + \frac{1}{N}B)/W}$, where W and B are within-chain and cross-chain sample variances of the parameters, respectively. $\hat{R}$ will be close to 1, if all chains are at equilibrium. Values greater than 1.1 indicates problems in convergence. Besides, autocorrelation is examined by effective sample size. Effective sample size ($\hat{n}_{eff}$) is an estimate of n_{eff}, the approximate effective number of independent draws of each unknown parameter (Stan Development Team, 2019a). Let M be the number of chains ($M = 4$ in this case), if the draws are independent, $n_{eff} = MN$. However, the simulation draws are typically positively correlated when sampled using MCMC method. n_{eff} is then defined as $MN/(1 + 2\sum_{t=1}^{\infty} \rho_t)$, where ρ_t is the autocorrelations at lag t. The estimation error is proportional to $1/\sqrt{n_{eff}}$. Larger $\hat{n}_{eff}$ indicates more reliable Bayesian inference and therefore is preferred.

2.4.2 Model Selection Criteria

Two statistics are calculated from the `RStan` results to compare the performance of different competing process models. The first one is the deviance information criterion (DIC). Similar to Akaike information criterion (AIC) and Bayesian information criterion (BIC), DIC is a penalized likelihood criterion that consists of the posterior expectation and effective number of parameters (Spiegelhalter et al., 2002). It is based on the Bayesian deviance, $D(\boldsymbol{\theta}) = -2 \log f(\boldsymbol{y} \mid \boldsymbol{\theta}) + 2 \log h(\boldsymbol{y})$, where the first term is likelihood function of $\boldsymbol{y}$ given parameters $\boldsymbol{\theta}$, and the second term is a standardizing function of $\boldsymbol{y}$ and thus irrelevant to model parameters.

The effective number of parameter is $p_D = E_{\boldsymbol{\theta}|\boldsymbol{y}}(D(\boldsymbol{\theta})) - D(E_{\boldsymbol{\theta}|\boldsymbol{y}}(\boldsymbol{\theta})) = \overline{D}(\boldsymbol{\theta}) - D(\overline{\boldsymbol{\theta}})$. DIC is then defined as

$$DIC = \overline{D}(\boldsymbol{\theta}) + p_D. \tag{2.17}$$

When comparing several models, smaller $\overline{D}$ indicates larger likelihood and a better fit, smaller p_D indicates smaller effective number of parameters and a simpler model. Therefore, a smaller DIC score is preferred. Since DIC is scale-free, only the differences between DIC scores are meaningful in model comparison.

Besides DIC, Pareto smoothed importance sampling (PSIS) leave-one-out (LOO) cross validation has also frequently been used to compare models in recent years. The PSIS estimate of the LOO expected log pointwise predictive density (elpd) is defined by Vehtari et al. (2017) as

$$\widehat{elpd}_{PSIS-LOO} = \sum_{i=1}^{n} \log\Big(\frac{\sum_{s=1}^{S} w_{is} p(y_i \mid \boldsymbol{\theta}_s)}{\sum_{s=1}^{S} w_{is}}\Big), \tag{2.18}$$

where $i = 1, .., n$ is the index of data points, $s = 1, ..., S$ is the index of iterations. $\boldsymbol{\theta}_s$ is the posterior estimate of parameter from the s^{th} iteration. w_{is} is the weight of data point i in iteration s. While leaving one data point out each time, the calculation of the weight w_{is} involves importance ratio, generalized Pareto distribution fitting, and truncation. The calculation are implemented in the package `loo` in R. `loo` returns $LOOIC = -2 \times \widehat{elpd}_{PSIS-LOO}$. As with DIC, smaller LOOIC score indicates a preferred model.

2.5 Simulation

2.5.1 Simulation Setting

The geographical information of this simulation study is based on the $I = 23$ states map from the CDM impPIPE data. We take $\boldsymbol{D} = Diag(m_i)$, where m_i is the number of neighbors of state i. $\boldsymbol{W}$ is the adjacency matrix with $w_{ij} = 1$ if state i and j are neighbors, $w_{ij} = 0$ otherwise. Four studies are designed, where the state-specific random effects $\boldsymbol{\phi}$'s are generated from the four process models, GMCAR model (2.11), UniCAR model (2.12), MvNorm model (2.13), and IndNorm model (2.14), respectively. True values of the parameters are listed in Table 2.2. We choose these values in order to create survival and duration data that is close to the real data, as shown in Figure 2.2. 50 Monte Carlo run are conducted in each of the four studies. Within each run, we assign dummy variable $\boldsymbol{x}_{kij}$ to imitate the three plant hosts in the real data. The collections of covariate vectors, $\boldsymbol{X}_1 = (\boldsymbol{x}'_{111}, ..., \boldsymbol{x}'_{1In_{1I}})'$ in the survival model and $\boldsymbol{X}_2 = (\boldsymbol{x}'_{211}, ..., \boldsymbol{x}'_{2In_{2I}})'$ in the duration model, are two three-column binary matrices, with row $(1, 0, 0)$ indicates plant host I, $(0, 1, 0)$ indicates plant host II, and $(0, 0, 1)$ indicates plant host III. Again, $\boldsymbol{x}_{1ij}$ and $\boldsymbol{x}_{2ij}$, $i = 1, ..., I$, $j = 1, ..., n_{ki}$, $k = 1, 2$, are not required to be the same.

For each state, we randomly generate $n_{1i} = 6$ uncensored survival data (t_{ij}) with 2 for each plant host and $n_{2i} = 3$ non-zero disease duration data (y_{ij}) with 1 for each plant host

Model	ρ	β_1	β_2	α_1	α_2	τ_1	τ_2	η_0	η_1	σ_1^2	σ_2^2	σ_{12}^2
Study 1 (GMCAR)	2	$(-12, -10, -8)$	$(3, 4, 5)$	0.7	0.9	0.5	2.5	0.5	0.2	.	.	.
Study 2 (UniCAR)	2	$(-12, -10, -8)$	$(3, 4, 5)$	0.7	0.9	0.5	2.5	.	.	.	.	.
Study 3 (MvNorm)	2	$(-12, -10, -8)$	$(3, 4, 5)$	.	.	.	.	.	.	2	0.2	0.5
Study 4 (IndNorm)	2	$(-12, -10, -8)$	$(3, 4, 5)$	.	.	.	.	.	.	2	0.2	.

Table 2.2: True value of parameters

from the data model,

$$
\begin{aligned}
t_{ij} &\sim Weibull\Big(\rho, exp(\phi_{1i} + \boldsymbol{x}_{1ij}\boldsymbol{\beta}_1)\Big), \quad j = 1, ..., n_{1i}, \quad i = 1, ..., I, \\
y_{ij} &\sim Poisson\Big(exp(\phi_{2i} + \boldsymbol{x}_{2ij}\boldsymbol{\beta}_2)\Big), \quad j = 1, ..., n_{2i}, \quad i = 1, ..., I \quad y_{ij} > 0.
\end{aligned}
\tag{2.19}
$$

Top 10% of the uncensored t_{ij} are treated as censored data. Other censoring percentage can be used as well. A possible guideline for choosing this percentage is that at least one uncensored t_{ij} should be observed for each state.

The generated data is then fit to five different models. In addition to the four models mentioned before, GMCAR model, UniCAR model, MvNorm model, and IndNorm model, we also include a reverse ordered GMCAR to see if the conditional order actually matters in the GMCAR model,

$$
\begin{aligned}
\boldsymbol{\phi_2} &\sim \mathcal{N}\Big(\mathbf{0}, [\tau_2(\boldsymbol{D} - \alpha_2\boldsymbol{W})]^{-1}\Big), \\
\boldsymbol{\phi_1} \mid \boldsymbol{\phi_2} &\sim \mathcal{N}\Big((\eta_0\boldsymbol{I} + \eta_1\boldsymbol{W})\boldsymbol{\phi_2}, [\tau_1(\boldsymbol{D} - \alpha_1\boldsymbol{W})]^{-1}\Big).
\end{aligned}
\tag{2.20}
$$

For each model fitting, four parallel chains run simultaneously. 3000 iterations are used as warm up period in each chain. Fast convergence to stationary is observed within this period. 5000 more iterations after warm up give posterior estimation size of $5000 \times 4 = 20,000$. Because `Stan` algorithm has many leapfrog steps within each iteration, it needs more time for each iteration comparing with Gibbs or Metropolis sampling. However, the leapfrog steps provide lower autocorrelated iterations. It takes much fewer iterations to get good mixing for `Stan` than other methods, such as `BUGS` (Stan Development Team, 2019b). Vehtari et al. (2017) states that 4 chains run for 1000 post warm up iterations is not a large number of

simulation draws, but is more than sufficient in many real-world settings for `Stan`. Therefore, more iterations for both warm up and posterior estimation periods in this simulation study are not necessary.

2.5.2 Simulation Results

We introduce $z_{kij} = \phi_{ki} + \boldsymbol{x}_{kij}\boldsymbol{\beta}_k$. Since 50 Monte Carlo runs are executed with the true value of the parameters known, we are able to compute average mean square error (AMSE) of z_{kij}. Let t be the index of the $T = 50$ generated data sets, $\hat{z}_{kij}^{(t)}$ be the posterior mean of the tth data set, AMSE is estimated as

$$\widehat{AMSE} = \frac{1}{T\sum_{k=1}^{2}\sum_{i=1}^{I} n_{ki}} \sum_{t=1}^{T}\sum_{k=1}^{2}\sum_{i=1}^{I}\sum_{j=1}^{n_{ki}} (\hat{z}_{kij}^{(t)} - z_{kij}^{(t)})^2. \tag{2.21}$$

The associated Monte Carlo standard error is

$$\widehat{se}(\widehat{AMSE}) = \sqrt{\frac{1}{T\sum_{k=1}^{2}\sum_{i=1}^{I} n_{ki}(T\sum_{k=1}^{2}\sum_{i=1}^{I} n_{ki} - 1)} \sum_{t=1}^{T}\sum_{k=1}^{2}\sum_{i=1}^{I}\sum_{j=1}^{n_{ki}} \left[(\hat{z}_{kij}^{(t)} - z_{kij}^{(t)})^2 - \widehat{AMSE}\right]^2}. \tag{2.22}$$

Because true values of the parameters remain unknown in the real world, AMSE and its standard error are usually not accessible. They provide an overview of how well the hierarchical model parameters are estimated in this simulation study, but will not be used for model selection in Section 2.6.

The AMSE, standard error (SE) and percentage change relative to the true model (Δ) are shown in Table 2.3. Column indicates which process model the data is generated from.

Row indicates which model the simulated data is fit to. "-" shows that data fits to the true model, for example, when the simulated data in Study 1 fits back to GMCAR model, a "-" will be shown in the first row, first Δ column. The percentage change of AMSE is defined as $\Delta_{model} = (AMSE_{model} - AMSE_{true})/AMSE_{true}$. All the values in the table are relatively small. AMSE are of the order of 10^{-1}, which means that $\hat{\boldsymbol{z}}$ are close enough to the real $\boldsymbol{z}$. Standard error are of the order of 10^{-2}, which means that the computed AMSE is reliable. When the data is generated from GMCAR and UniCAR models (Study 1 and 2), the two Norm models have larger AMSE, about $23\% - 43\%$ more than the AMSE of the true model, due to the missing geographic information. GMCAR model consistently performs better, even in Study 2 where the true frailty terms are not correlated. The reverse ordered GMCAR provides the second smallest AMSE. Besides, when the data is generated from MvNorm and IndNorm models (Study 3 and 4), the three CAR models all outperform the true models. They have AMSE about $7\% - 32\%$ less than the AMSE of the true model. It suggests that some noises may be picked up by the CAR models in the process of parameter estimation. Overfitting will be checked next with DIC and LOOIC results.

Taking $\log h(\boldsymbol{y}) = 0$ in (2.17), $D(\boldsymbol{\theta}) = -2 \log f(\boldsymbol{y}, \boldsymbol{t} \mid \boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \rho)$. Because the difference among the five models is in the process model, instead of the likelihood function of data model, $D(\boldsymbol{\theta})$ is identical for all five models. For each set of the simulated data, we computed the differences of DIC and LOOIC among the five models. Table 2.4 summarizes the median of differences between each fit and the true model. While smaller values are preferred, a negative value indicates that a model outperforms the true model. For example,

	Study 1 (GMCAR Data)		Study 2 (UniCAR Data)		Study 3 (MvNorm Data)		Study 4 (IndNorm Data)	
Model	AMSE (SE)	Δ	AMSE (SE)	Δ	AMSE (SE)	Δ	AMSE (SE)	Δ
GMCAR	8.13 (1.58)	-	7.88 (1.39)	-8.94	8.58 (2.00)	-31.85	7.45 (1.43)	-24.18
(reversed)	9.50 (2.06)	16.90	9.53 (1.88)	10.23	10.23 (2.57)	-18.70	9.10 (1.83)	-7.37
UniCAR	8.69 (1.84)	6.89	8.65 (1.67)	-	9.83 (2.51)	-21.89	8.09 (1.60)	-17.60
MvNorm	11.61 (2.56)	42.92	12.16 (2.41)	40.58	12.59 (3.16)	-	11.09 (2.20)	12.94
IndNorm	10.33 (2.25)	27.09	10.65 (2.11)	23.16	11.51 (2.95)	-8.58	9.82 (1.97)	-

"-" indicates fit of the true model

Table 2.3: AMSE ($\times 10^{-1}$), standard error ($\times 10^{-2}$) and percentage change (%)

	Study 1 (GMCAR Data)		Study 2 (UniCAR Data)		Study 3 (MvNorm Data)		Study 4 (IndNorm Data)	
Model	DIC	LOOIC	DIC	LOOIC	DIC	LOOIC	DIC	LOOIC
GMCAR	-	-	2.50	1.81	3.97	2.09	2.91	1.82
(reversed)	-1.40	-0.32	0.30	0.79	2.94	2.91	1.36	1.30
UniCAR	-0.45	0.93	-	-	4.23	3.74	1.61	0.87
MvNorm	-2.47	-0.35	-0.42	1.75	-	-	-0.33	0.50
IndNorm	0.30	2.38	0.08	1.46	2.58	3.06	-	-

"-" indicates fit of the true model

Table 2.4: Median of DIC and LOOIC difference of the simulated data

when the data is generated from IndNorm model (Study 4), fitting to MvNorm model sometimes gives small negative DIC differences (-0.33). Note that IndNorm is in fact a special case of MvNorm with $\sigma_{12} = 0$. Therefore, we are able to get similar or even better DIC scores when fitting IndNorm data to MvNorm model. In contrast, when data is generated from MvNorm model (Study 3), fitting to IndNorm gives DIC 2.58 larger than the true model. We noticed that the largest difference median, 4.23, can be observed when fitting MvNorm data to UniCAR model. It may be caused by the missing correlation between frailty terms. However, since the DIC and LOOIC scores in all four studies are around $1800 - 1900$, the differences in the table are negligible. It is safe to conclude that the five models have comparable DIC and LOOIC results.

Comparing with the differences of AMSE, the differences of DIC and LOOIC are subtle. This is because of the setting of the simulation study, where each pair of β and ϕ has very limited data, either one duration data or two survival data that can be censored. As a result, the log likelihood can be similar for the true parameter and a not so well-estimated parameter.

Although the data is sometimes not ample in the real world, the design of GMCAR model makes it possible to borrow information from spatial structure and correlation between the two data model parts. It has advantages when estimating parameters in all four scenarios. Recall that in the real data, fixed effects coefficients $\boldsymbol{\beta}$ and frailty terms $\boldsymbol{\phi}$ indicate plant type and location, respectively, which are of interest to plant pathologists. Therefore, the proposed GMCAR model is preferred among the five process models.

2.6 Real Data

We now fit CDM impPIPE data to the five hierarchical models with different process models for the frailty terms. To be conservative, 5000 iterations are included in the warm up period of each chain. It is more than sufficient to reach the state of convergence. A further $10,000 * 4 = 40,000$ post warm up iterations are then drawn to estimate the parameters. Again, because the data models have the same likelihood function in all five scenario, and the parameters have quite vague priors, DIC and LOOIC are fair in model comparison. While the actual number of parameters in the data model is 53 (1 Weibull shape parameter ρ, 2×3 covariates coefficient $\boldsymbol{\beta}$, and 2×23 state-specific random effects $\boldsymbol{\phi}$), the effective numbers of parameters p_D are much smaller than that. As shown in Table 2.5, the GMCAR model has the smallest p_D, 33.33, and thus allows more random effects smoothing. GMCAR also has the smallest DIC, 3676.98. Following GMCAR, reverse ordered GMCAR, UniCAR, and MvNorm models have similar DIC values, about 1 more than the GMCAR model. IndNorm

	GMCAR	(reversed)	UniCAR	MvNorm	IndNorm
$\overline{D}(\boldsymbol{\theta})$	3643.65	3640.81	3642.19	3638.82	3641.14
$D(\overline{\boldsymbol{\theta}})$	3610.33	3604.14	3606.58	3599.88	3603.11
p_D	33.33	36.67	35.61	38.94	38.03
DIC	3676.98	3677.48	3677.80	3677.76	3679.18
LOOIC	3755.83	3759.71	3760.05	3761.49	3765.46
$\widehat{elpd}_{PSIS-LOO}$ difference	-	-1.9	-2.1	-2.8	-4.8
SE of $\widehat{elpd}_{PSIS-LOO}$ difference	-	2.2	2.5	3.0	3.2

Table 2.5: Real data DIC and LOOIC comparison

model performs less well, about 2-3 more than the GMCAR model.

The information criterion difference gets more obvious in LOOIC (Table 2.5). LOOIC of GMCAR model is 3755.83. It outperforms the reverse ordered GMCAR, UniCAR, MvNorm, and IndNorm models by 4, 4, 6, and 10, respectively. The $\widehat{elpd}_{PSIS-LOO}$ of each data point are computed as well. Similar to the DIC results, the three CAR models are close to each other. Although GMCAR provides the smallest absolute value of the $\widehat{elpd}_{PSIS-LOO}$ among the three CAR models, the differences, 1.9 and 2.1, are not significant, taking into account that the estimated standard error of the difference is 2.2 and 2.5. On the other hand, the $\widehat{elpd}_{PSIS-LOO}$ difference between GMCAR and IndNorm model is larger than the estimated standard error, which indicates that the GMCAR model has better predictive performance than the IndNorm model for the CDM impPIPE data.

Posterior medians of the Weibull parameter ρ of the five models are about 1.98, with 95% credible interval (1.71, 2.27). It suggests that the baseline hazard increases over time, which is consistent with the characteristic of seasonal epidemics development. Figure 2.5 shows

posterior median and 95% credible interval of the covariates coefficients $\boldsymbol{\beta}$ of five models. The estimate of the parameters are rather similar across the models. The credible intervals have similar width in the survival model. In the duration model, the two Norm models have the narrowest width due to the lack of spatial information in the frailty terms. The two GMCAR models have narrower credible interval width than the UniCAR model because of the correlated survival and duration parts. A similar contour can be found in both survival model and duration model that squash has the largest value among the three plant hosts and other hosts has the smallest value, or $\beta_{k2} > \beta_{k1} > \beta_{k3}$, $k = 1, 2$. While higher β_1 indicates a higher hazard, which leads to less survival time, and higher β_2 indicates a longer duration time, it confirms the connection between the two parts of the data model that shorter survival time comes with longer duration, and vice versa. In this case, squash has the highest hazard of disease outbreak and is expected to have the longest duration time, which is coincident with the count in Figure 2.3.

We further looked into the differences between estimated coefficients $\boldsymbol{\beta}$ of the three host groups. Since the five models have rather similar estimates of the coefficients, we list 2.5, 50 and 97.5 percentiles of coefficients differences from the proposed GMCAR model in Table 2.6. Cucumber v.s. squash in the survival model denotes $\beta_{11} - \beta_{12}$, cucumber v.s. others in the survival model denotes $\beta_{11} - \beta_{13}$, and so on. Because 0 is not included in any of the six 95% interval, the differences between plant host groups are all significant at 0.05 level. Taking the "Others" host type as a reference group, cucumber increases the hazard rate by a factor of $e^{0.03} = 1.03$ to $e^{0.85} = 2.34$, squash increases the hazard rate by $e^{0.78} = 2.18$ to

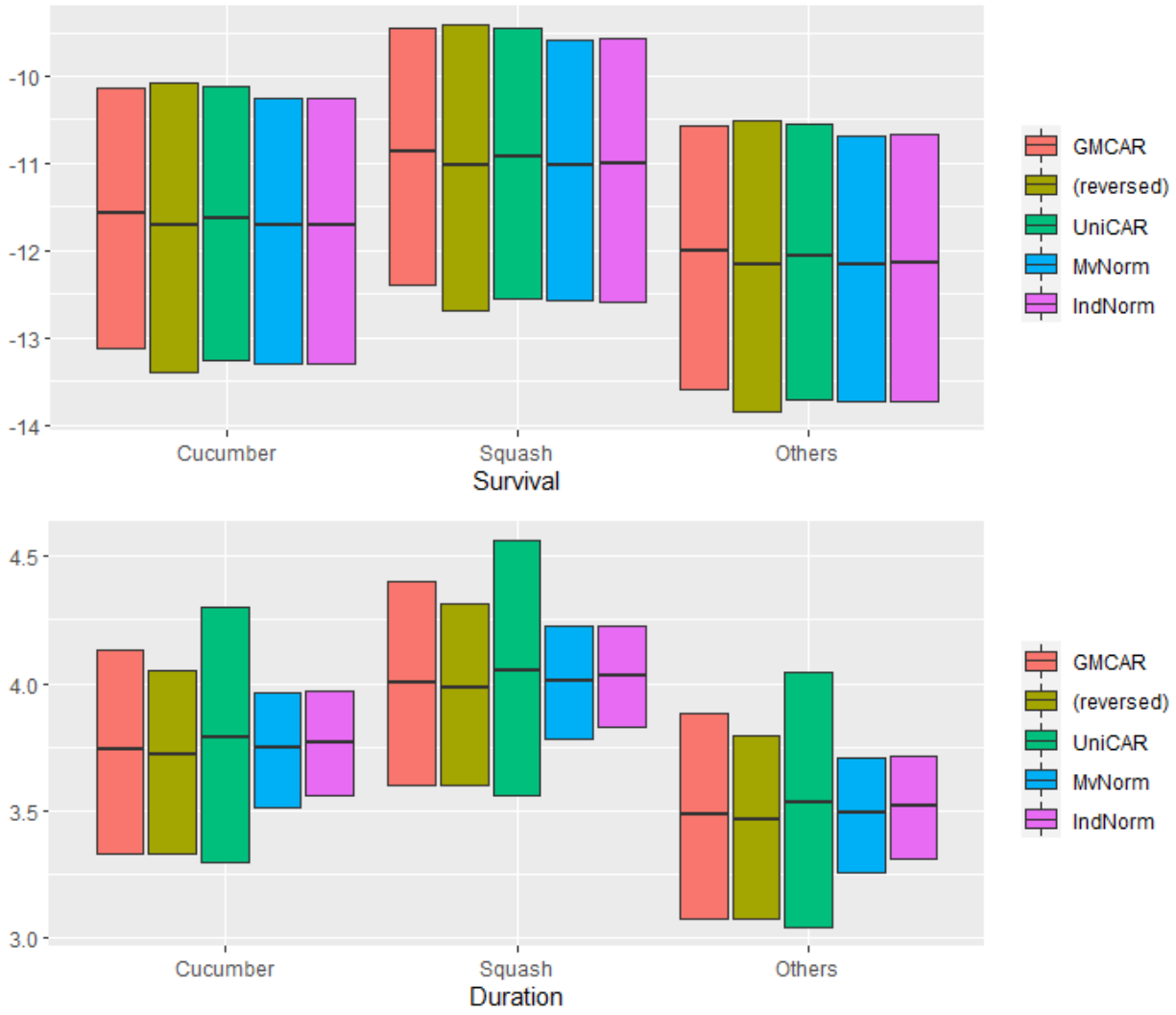


Figure 2.5: Posterior median (black line) and 95% credible interval (box) of the covariates coefficients: $\boldsymbol{\beta}_1$ (upper panel) and $\boldsymbol{\beta}_2$ (lower panel)

		2.5%	50%	97.5%
Survival	Cucumber v.s. Squash	-1.08	-0.70	-0.33
	Cucumber v.s. Others	0.03	0.44	0.85
	Squash v.s. Others	0.78	1.14	1.51
Duration	Cucumber v.s. Squash	-0.33	-0.27	-0.20
	Cucumber v.s. Others	0.18	0.25	0.33
	Squash v.s. Others	0.45	0.52	0.59

Table 2.6: Coefficients difference of GMCAR model

$e^{1.51} = 4.53$ and has the highest CDM outbreak hazard among the three host groups. In the duration model, cucumber increases the estimated duration days by a factor of $e^{0.18} = 1.20$ to $e^{0.33} = 1.39$, squash increases by $e^{0.45} = 1.57$ to $e^{0.59} = 1.80$ and therefore is expected to have the longest epidemic duration time among the three host groups examined.

We obtained the posterior median and 95% equal-tail interval of the rest GMCAR related parameters. The point and interval estimate of the smoothing parameters are 0.75 and $(0.09, 0.98)$ for α_1 and 0.70 and $(0.06, 0.99)$ for α_2. Because the parameter is set to be between $(0, 1)$, the wide credible intervals, caused by large censoring rate and sparse duration data, denotes limited spatial association between the state-specific random effects. The estimate of the precision parameters are 3.26 and $(1.05, 14.64)$ for τ_1 and 7.74 and $(2.96, 19.42)$ for τ_2. These values suggest larger covariances between the survival frailty than the duration frailty, which can be from the setting that τ_1 is a marginal precision parameter while τ_2 is a conditional precision parameter. The estimated bridging parameters are 0.05 and $(-0.63, 0.69)$ for η_0, and 0.38 and $(0.11, 0.92)$ for η_1. Since 0 is included in the 95% interval of η_0 but not the interval of η_1, it suggests that the connection between ϕ_{2i} and ϕ_{1i} is weaker

than the connection between ϕ_{2i} and ϕ_{1j}, where i and j are neighbouring states. Besides, η_0 and η_1 take mostly positive values, which proves that the survival frailty and duration frailty are positively correlated. The posterior median of $\boldsymbol{\phi}$ using the GMCAR process model is shown in the map in Figure 2.6. The positive correlation can be observed in many of the states. For example, Louisiana has a smaller posterior median of ϕ_1, -0.79, which implies smaller hazard and longer survival, and a smaller value of ϕ_2, -0.56, which indicates shorter duration, while Massachusetts has a larger posterior median of ϕ_1, 0.16, which implies larger hazard and shorter survival, and a larger value of ϕ_2, 0.50, which indicates longer epidemic duration. Overall, states along the mid-Atlantic region of the United States (e.g., North Carolina, New York, Pennsylvania) have larger values of the survival and duration frailty terms, and states in southwest of the United States (e.g., Texas, Louisiana) have smaller values of the survival and duration frailty terms, which matches the conclusions from Figure 2.4.

2.7 Conclusion

In this chapter, we have developed a joint model that consists of a censored Weibull model for survival data and a zero truncated Poisson model for duration data. The two models are linked together by spatially correlated random effects. GMCAR and a few other process models are applied to the random effects. The simulated data demonstrates the advantage of GMCAR model through AMSE, DIC and LOOIC criteria when the data has spatial

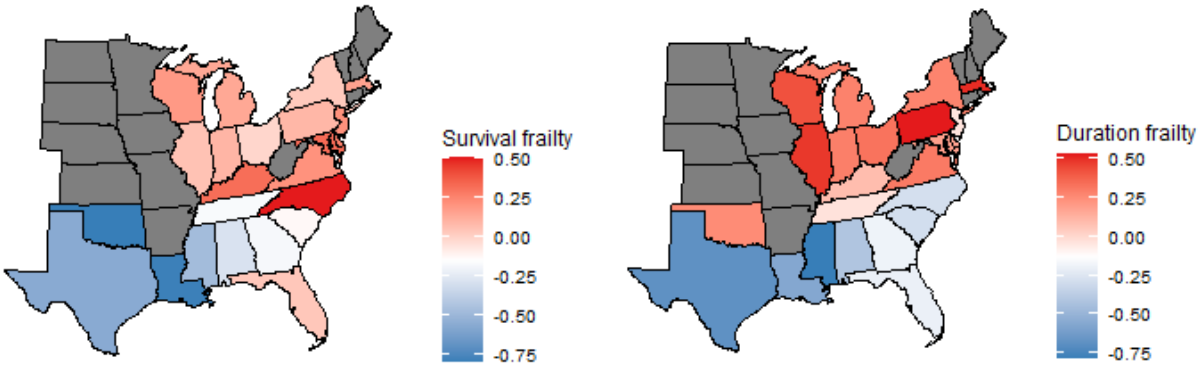


Figure 2.6: Posterior median of state-specific random effects, ϕ_1 (left) and ϕ_2 (right)

correlation. We further illustrated this hierarchical joint model with the 2009 CDM ipmPIPE data from 23 states in the eastern United States. Bayesian methods are implemented using `Stan`'s no-U-turn sampling (NUTS). Full posterior inference is available, including plant host type coefficients and state level frailties. The positive correlation between the survival and duration part of the model is confirmed by the posterior inference. The joint model brings the two pieces of information together and helps botanical epidemiologists explain the pattern of the disease and predict when certain data is missing. The 2009 CDM ipmPIPE data indicates risk of disease outbreak is high in states along the mid-Atlantic region in the United States as reported in previous studies (Ojiambo and Kang, 2013). Thus, disease surveillance efforts should be concentrated in these states once initial outbreaks are reported in areas close to overwintering sources in southern Florida and along the Gulf of Mexico. The GMCAR model can be applied to other plant disease systems whose pathogens exhibit

significant long-distance dispersal at the landscape level to asses the risk of disease outbreak during the disease monitoring season. While the importance of joint models has increasingly been recognized in medical literature (e.g., He and Luo 2016; Papageorgiou et al. 2019; Zhang et al. 2017), joint models are still not widely used in botanical epidemiology due to computational limitations. The availability of softwares within open source packages that can fit joint models (Rizopoulos et al., 2016) should further facilitate the application of these models in botanical epidemiology.

We are also aware of the possibility of data model extension. Comparing with the parametric Weibull survival model, nonparametric counterparts are available for the hazard function (2.1). Carlin and Hodges (1999) investigate parametric Weibull and semiparameteric model with mixture of monotone baseline hazard functions on highly stratified data. Mixture model of logistic regression and proportional hazard function is also possible and studied by Kuk and Chen (1992) and Peng and Dear (2000). For the duration part of the data model, a point mass at zero and the truncated Poisson distribution for the non-zero observations can be combined to form a Poisson hurdle model that accounts for zero inflation in the data (Neelon et al., 2013). These extensions are suitable for different real data situations and will be explored in our future investigation.

Finally, the study of parameters' distribution and model comparison of a Bayesian hierarchical framework with latent Gaussian model can also be implemented using integrated nested Laplace approximations (INLA) (Rue et al., 2009). INLA is a computationally efficient alternative to MCMC and is available to researchers in `R` package `INLA`. Instead of

drawing samples from multivariate joint posterior distribution, INLA focuses on estimating univariate marginal posterior distribution of the model parameters. Details of the method can be found in Lindgren et al. (2011) and Lindgren and Rue (2015). Several spatial and spatial-temporal model applications of INLA include: counts of salmonellosis in cattle from 184 regions of Switzerland (Schrödle and Held, 2011), suicide mortality from 32 boroughs of London, and counts of low birth weight from 159 counties in Georgia (Blangiardo et al., 2013).It is worth to explore the alternative implementation of joint Bayesian hierarchical model in the following studies.

Chapter 3

Bayesian Hierarchical Functional Data Analysis with Automatically Adaptive Multi-Resolution Spline Basis Functions

3.1 Introduction

Functional data analysis (FDA) has been applied to a wide range of data, such as spatial, longitudinal, and image data, and has therefore received increasing attention over the past decades. Since the functional data is recorded on a discrete grid t over a continuum $\mathcal{T}$ with

measurement error, revealing the underlying continuous function free from noise, or smoothing, is crucial for FDA. A vast amount of literature exists regarding to FDA methods. One class of methods extends linear regression to functional data and regresses the functional response on a set of predictors (Guo, 2002; Morris and Carroll, 2006; Zhu et al., 2011); another class, called functional principal component analysis (FPCA), aims to characterize variation and reduce dimension (Silverman, 1996; Yao et al., 2005; Chiou, 2012). However, many of the FDA approaches work on point estimation and lack reliable uncertainty quantification. Bayesian methods, on the other hand, handle uncertainty quantification with full probability models. Yang et al. (2016) proposed a Bayesian counterpart of functional principal analysis by conditional expectation. Instead of treating each curve independently, the Bayesian hierarchical model smooths functional curves simultaneously and nonparametrically. The spontaneity borrows information from all curves and is therefore capable to keep the systematic patterns. The nonparametrical characteristic of the model ensures the estimation flexibility.

As real functional data has been collected in increasingly larger size and higher dimensions, the Bayesian hierarchical model shows greater computational complexity. Several well-developed kriging methods are available to reduce rank, including approximation using predictive process (Banerjee et al., 2008; Finley et al., 2009; Banerjee et al., 2010), fixed rank kriging (Cressie and Johannesson, 2008; Cressie et al., 2010; Kang and Cressie, 2011), and imposing parametric assumptions on the random process precision matrix (Lindgren et al., 2011; Nychka et al., 2015). Some combination and comparison of the existing kriging

methods can be found in Bradley et al. (2015, 2016). Because many of the kriging methods rely heavily on the tuning parameters, which, when selected inappropriately, lead to biased or unstable results, Tzeng and Huang (2018) improved the fixed rank kriging (FRK) by defining a class of basis function from thin-plate splines that avoids knot allocation and scale selection.

Motivated by this novel FRK moethod, we applied the so-called multi-resolution spline basis functions to the Bayesian hierarchical function data analysis model. The proposed model has the following advantages. First, it keeps the characteristics of the Bayesian hierarchical model that estimates mean and covariance functions simultaneously and nonparametrically. Second, because the multi-resolution spline basis functions are sorted in the order of smoothness and the number of basis functions is selected by Akaike information criterion (AIC), it reduces computational burden without sacrificing the precision. The conclusion is confirmed by the comparison of root mean square error (RMSE) and the number of basis functions between the proposed model and the same Bayesian hierarchical structure using B-splines basis functions approximations (Yang et al., 2017). Third, it can be expanded directly to higher dimensions, which will be shown in the numerical studies.

The reminder of the chapter proceeds as follows. Section 3.2 consists of two parts. The first half outlines the Bayesian hierarchical model and the steps of Markov chain Monte Carlo (MCMC) procedure. The second half illustrates the FRK multi-resolution spline basis function, both theoretically and visually, and how it is executed in the hierarchical model. The proposed model is then applied to the simulated data in Section 3.3 and the real data

in Section 3.4. The simulated data is generated in one-dimensional and two-dimensional t settings. Different combinations of stationarity and observation grid are examined in the one-dimensional t setting. The selection of number of basis functions is studied in the two-dimensional t setting. The results from the FRK multi-resolution spline basis functions are compared with the ones from the commonly used B-splines basis functions. Two real data sets are used to test the model prediction performance. Section 3.4.1 is sleeping energy expenditure (SEE) data, a one-dimensional functional data that collected for obesity study by the Children's Nutrition Research Center (CNRC) of Baylor College of Medicine. Section 3.4.2 is mortality data, which is in two-dimensional grid of age and year. Section 3.5 summarizes our major findings and concludes with a brief discussion.

3.2 Method

3.2.1 Bayesian Hierarchical Model

The Bayesian hierarchical model consists of data model, process model, and parameter model. In the data model, we use the widely accepted functional data with general measurement error model. Let $Y_i(t)$ be the observed functional data, $i = 1, ..., n$ is the index of subjects, t is the observation time points over interval $\mathcal{T}$, the model with measurement error

is given by

$$
\begin{aligned}
&Y_i(t) = Z_i(t) + \epsilon_i(t), \quad t \in \mathcal{T}, \\
&Z_i(t) \sim GP\big(\mu_z(t),\ \Sigma_z(t,t)\big), \quad \epsilon_i(t) \sim N(0,\ \sigma_\epsilon^2).
\end{aligned}
\tag{3.1}
$$

$Z_i(t)$ indicates the true functional curve with shared mean function $\mu_z(t)$ and covariance function $\Sigma_z(t,t)$. $\epsilon_i(t)$ is the measurement error with mean 0 and variance σ_ϵ^2 that follows inverse Gamma distribution, $IG(a_\epsilon, b_\epsilon)$. t usually represents time points, and therefore is one-dimensional. But it can also be in higher dimensions, for example, two-dimensional latitude and longitude, or a vector of covariates. Yang et al. (2016) proposed a Gaussian-Wishart process model for the shared mean and covariance functions of $Z_i(t)$. It uses Gaussian process (GP) for $\mu_z(t)$ to smooth functional curves simultaneously and inverse Wishart process (IWP) (Dawid, 1981) for $\Sigma_z(t,t)$ to estimate covariance nonparametrically. It is denoted by

$$
\begin{aligned}
&\mu_z(t)|\Sigma_z(t,t) \sim GP\big(\mu_0(t),\ c\Sigma_z(t,t)\big), \\
&\Sigma_z(t,t) \sim IWP\big(\delta,\ \sigma_s^2 A(t,t)\big).
\end{aligned}
\tag{3.2}
$$

$\{\mu_0, c, \delta, A(t,t)\}$ are hyper-prior parameters and are determined empirically. μ_0 is set to be the smoothed sample mean. $c = 1$, $\delta = 5$ are uninformative priors for the mean-covariance functions. We use Matérn covariance function (Matérn, 1960) for $A(t,t)$ in this chapter, while smoothed covariance estimate is another possible choice. $\sigma_s^2 \sim Gamma(a_s, b_s)$ is the scale parameter when estimating the covariance structure $A(t,t)$. The inverse Gamma and Gamma parameters, $\{a_\epsilon, b_\epsilon, a_s, b_s\}$, can be decided using heuristic Bayesian approach (Yang et al., 2016).

Yang et al. (2017) suggested a rank reduction method to solve the computational complexity when the functional data is observed on random grid or in high dimension. With the data model in (3.1), approximations by basis functions were introduced to the process model of the true signal $Z_i(t)$ in (3.2). Let $\boldsymbol{\tau}$ be a working grid with $\boldsymbol{\tau} = (\tau_1, \tau_2, ..., \tau_L)^T \subset \mathcal{T}$ and $L << p$ pooled observation grid points, and $\boldsymbol{B}(\boldsymbol{\tau}) = [b_1(\boldsymbol{\tau}), b_2(\boldsymbol{\tau}), ..., b_K(\boldsymbol{\tau})]$ is K selected basis functions with coefficients $\boldsymbol{\zeta}_i = (\zeta_{i1}, \zeta_{i2}, ..., \zeta_{iK})^T$, the true signal on the working grid $Z_i(\boldsymbol{\tau})$ can be written as

$$Z_i(\boldsymbol{\tau}) = \sum_{k=1}^{K} \zeta_{ik} b_k(\boldsymbol{\tau}) = \boldsymbol{B}(\boldsymbol{\tau})\boldsymbol{\zeta}_i. \tag{3.3}$$

$\boldsymbol{B}(\boldsymbol{\tau})^{-1}$ is the inverse of the basis matrix $\boldsymbol{B}(\boldsymbol{\tau})$ when $K = L$ and is the generalized inverse of $\boldsymbol{B}(\boldsymbol{\tau})$ when $K \neq L$ or when $\boldsymbol{B}(\boldsymbol{\tau})$ is singular. By multiplying $\boldsymbol{B}(\boldsymbol{\tau})^{-1}$ to both sides of (3.3), we have $\boldsymbol{\zeta}_i = \boldsymbol{B}(\boldsymbol{\tau})^{-1} Z_i(\boldsymbol{\tau})$. $\boldsymbol{\zeta}_i$ is therefore a linear transformation of $Z_i(\boldsymbol{\tau})$ and

$$\boldsymbol{\zeta}_i \sim GP\big(\boldsymbol{\mu}_\zeta = \boldsymbol{B}(\boldsymbol{\tau})^{-1}\mu_z(\boldsymbol{\tau}),\ \boldsymbol{\Sigma}_\zeta = \boldsymbol{B}(\boldsymbol{\tau})^{-1}\Sigma_z(\boldsymbol{\tau},\boldsymbol{\tau})\boldsymbol{B}(\boldsymbol{\tau})^{-T}\big). \tag{3.4}$$

The process model in (3.2) can be rewritten in terms of $\boldsymbol{\zeta}$,

$$\begin{aligned} &\boldsymbol{\mu}_\zeta|\boldsymbol{\Sigma}_\zeta \sim GP\big(\boldsymbol{B}(\boldsymbol{\tau})^{-1}\mu_0(\boldsymbol{\tau}),\ c\boldsymbol{\Sigma}_\zeta\big), \\ &\boldsymbol{\Sigma}_\zeta \sim IWP\big(\delta,\ \sigma_s^2\boldsymbol{B}(\boldsymbol{\tau})^{-1}A(\boldsymbol{\tau},\boldsymbol{\tau})\boldsymbol{B}(\boldsymbol{\tau})^{-T}\big). \end{aligned} \tag{3.5}$$

The updated Bayesian hierarchical model gives the following joint posterior distribution

$$\begin{aligned} &f(\boldsymbol{\zeta},\boldsymbol{\mu}_\zeta, \boldsymbol{\Sigma}_\zeta, \sigma_\epsilon^2, \sigma_s^2|Y) \propto \\ &\quad f(Y|\boldsymbol{\zeta}, \sigma_\epsilon^2) f(\sigma_\epsilon^2) f(\boldsymbol{\zeta}|\boldsymbol{\mu}_\zeta, \boldsymbol{\Sigma}_\zeta) f(\boldsymbol{\mu}_\zeta|\boldsymbol{\Sigma}_\zeta) f(\boldsymbol{\Sigma}_\zeta|\sigma_s^2) f(\sigma_s^2). \end{aligned} \tag{3.6}$$

$f(Y|\boldsymbol{\zeta}, \sigma_\epsilon^2)$, $f(\boldsymbol{\zeta}|\boldsymbol{\mu}_\zeta, \boldsymbol{\Sigma}_\zeta)$ and $f(\boldsymbol{\mu}_\zeta|\boldsymbol{\Sigma}_\zeta)$ on the left hand side are probability density functions of GP in (3.1), (3.4), and (3.5). $f(\boldsymbol{\Sigma}_\zeta|\sigma_s^2)$ is the inverse Wishart distribution. $f(\sigma_s^2)$ and $f(\sigma_\epsilon^2)$ are Gamma and inverse Gamma distribution, respectively. The MCMC procedure then samples from the full conditional distribution.

Step 0: After setting initial values of the hyper-prior parameters in the parameter model and selecting basis functions, details will be covered in Section 3.2.2, we get $\boldsymbol{\mu}_\zeta$ and $\boldsymbol{\Sigma}_\zeta$ from (3.5).

Step 1: From $f(\boldsymbol{\zeta}|Y, \boldsymbol{\mu}_\zeta, \boldsymbol{\Sigma}_\zeta) \propto f(Y|\boldsymbol{\zeta}, \sigma_\epsilon^2) f(\boldsymbol{\zeta}|\boldsymbol{\mu}_\zeta, \boldsymbol{\Sigma}_\zeta)$, we update $\boldsymbol{\zeta}$ based on the full conditional multivariate normal distribution using $\boldsymbol{\mu}_\zeta$, $\boldsymbol{\Sigma}_\zeta$, σ_ϵ^2 and the observed data $Y(t)$,

$$\begin{aligned}&\boldsymbol{\zeta}_i|(Y_i(\boldsymbol{t}_i), \boldsymbol{\mu}_\zeta, \boldsymbol{\Sigma}_\zeta, \sigma_\epsilon^2) \sim \\ &MVN\big((\frac{\boldsymbol{B}(\boldsymbol{t}_i)^T\boldsymbol{B}(\boldsymbol{t}_i)}{\sigma_\epsilon^2} + \boldsymbol{\Sigma}_\zeta^{-1}\big)^{-1}(\frac{\boldsymbol{B}(\boldsymbol{t}_i)^T Y_i(\boldsymbol{t}_i)}{\sigma_\epsilon^2} + \boldsymbol{\Sigma}_\zeta^{-1}\boldsymbol{\mu}_\zeta\big),\ (\frac{\boldsymbol{B}(\boldsymbol{t}_i)^T\boldsymbol{B}(\boldsymbol{t}_i)}{\sigma_\epsilon^2} + \boldsymbol{\Sigma}_\zeta^{-1}\big)^{-1}),\end{aligned} \tag{3.7}$$

where $\boldsymbol{t}_i$ is the observed time points of the ith curve.

Step 2: From $f(\boldsymbol{\mu}_\zeta, \boldsymbol{\Sigma}_\zeta|\boldsymbol{\zeta}, \sigma_s^2) \propto \prod_{i=1}^n f(\boldsymbol{\zeta}_i|\boldsymbol{\mu}_\zeta, \boldsymbol{\Sigma}_\zeta) f(\boldsymbol{\mu}_\zeta|\boldsymbol{\Sigma}_\zeta) f(\boldsymbol{\Sigma}_\zeta|\sigma_s^2)$, we update $\boldsymbol{\mu}_\zeta$ and $\boldsymbol{\Sigma}_\zeta$ conditioning on $\boldsymbol{\zeta}$ in (3.7),

$$\begin{aligned}&\boldsymbol{\mu}_\zeta|(\boldsymbol{\Sigma}_\zeta, \boldsymbol{\zeta}_1, ..., \boldsymbol{\zeta}_n) \sim NM\big(\frac{1}{n+c}(\sum_{i=1}^n \boldsymbol{\zeta}_i + c\boldsymbol{B}(\boldsymbol{\tau})^{-1}\mu_0(\boldsymbol{\tau})),\ \frac{1}{n+c}\boldsymbol{\Sigma}_\zeta\big) \\ &\boldsymbol{\Sigma}_\zeta|(\boldsymbol{\mu}_\zeta, \boldsymbol{\zeta}_1, ..., \boldsymbol{\zeta}_n) \sim IW\big(n+1+\delta,\ \sum_{i=1}^n(\boldsymbol{\zeta}_i - \boldsymbol{\mu}_\zeta)(\boldsymbol{\zeta}_i - \boldsymbol{\mu}_\zeta)^T + \\ &\qquad c(\boldsymbol{\mu}_\zeta - \boldsymbol{B}(\boldsymbol{\tau})^{-1}\mu_0(\boldsymbol{\tau}))(\boldsymbol{\mu}_\zeta - \boldsymbol{B}(\boldsymbol{\tau})^{-1}\mu_0(\boldsymbol{\tau}))^T + \sigma_s^2\boldsymbol{B}(\boldsymbol{\tau})^{-1}A(\boldsymbol{\tau}, \boldsymbol{\tau})\boldsymbol{B}(\boldsymbol{\tau})^{-T}\big).\end{aligned} \tag{3.8}$$

Step 3: From $f(\sigma_\epsilon^2|Y,Z) \propto \prod_{i=1}^n f(Y_i(\boldsymbol{t}_i)|Z_i(\boldsymbol{t}_i),\sigma_\epsilon^2) f(\sigma_\epsilon^2)$, we update the noise term σ_ϵ^2 using the observed data $Y_i(t)$ and the estimated true signal $Z_i = \boldsymbol{B}(\boldsymbol{t}_i)\boldsymbol{\zeta}_i$

$$\sigma_\epsilon^2|(Y,Z) \sim IG\big(a_\epsilon + \frac{1}{2}\sum_{i=1}^n p_i,\ b_\epsilon + \frac{1}{2}\sum_{i=1}^n (Y_i(\boldsymbol{t}_i) - Z_i(\boldsymbol{t}_i))^T (Y_i(\boldsymbol{t}_i) - Z_i(\boldsymbol{t}_i))\big), \quad (3.9)$$

where p_i is the number of observed time points of the ith signal.

Step 4: From $f(\sigma_s^2|\Sigma_Z(\boldsymbol{\tau},\boldsymbol{\tau})) \propto f(\Sigma_Z(\boldsymbol{\tau},\boldsymbol{\tau})|\sigma_s^2) f(\sigma_s^2)$, we update σ_s^2 by

$$\sigma_s^2|\Sigma_Z(\boldsymbol{\tau},\boldsymbol{\tau}) \sim Gamma\big(a_s + \frac{(\delta+K-1)K}{2},\ b_s + \frac{1}{2} trace(A(\boldsymbol{\tau},\boldsymbol{\tau})\Sigma_Z(\boldsymbol{\tau},\boldsymbol{\tau})^{-1})\big), \quad (3.10)$$

where $\Sigma_Z(\boldsymbol{\tau},\boldsymbol{\tau}) = \boldsymbol{B}(\boldsymbol{\tau})\boldsymbol{\Sigma}_\zeta \boldsymbol{B}(\boldsymbol{\tau})^T$.

MCMC then loops over posterior distributions in Step 1 - 4 with updated $\{\boldsymbol{\zeta}, \boldsymbol{\mu}_\zeta, \boldsymbol{\Sigma}_\zeta, \sigma_\epsilon^2, \sigma_s^2\}$ from previous iteration. The posterior means of the iterations after warm up period for $\{Z_i(t), \mu_z(t), \Sigma_z(t,t), \sigma_\epsilon^2\}$ are used for model comparison and data prediction in the numerical studies.

3.2.2 Basis Function Selection

For the basis function approximation in the process model in Section 3.2.1, quite a few methods are available to select the basis functions, for example, B-splines (De Boor et al., 1978; Ramsay et al., 1988; Meyer et al., 2008) is frequently used for Gaussian process data, and Fourier series is usually chosen for periodic data. In this chapter, we apply resolution adaptive fixed rank kriging (FRK) method to the basis function selection because of the following advantages: first, resolution adaptive FRK uses multi-resolution spline (MRS)

basis functions, which sorts basis functions in descending order with regard to smoothness, second, instead of using t only as in B-splines, resolution adaptive FRK takes both t and the observed functional data Y into account and selects the optimal number of basis functions K automatically by AIC, third, resolution adaptive FRK is able to extend to higher dimensional t directly. We start with some introduction of the FRK method.

FRK is developed from the spatial random effects model,

$$\begin{aligned} Z_i(\boldsymbol{t}) &= \mu_i(\boldsymbol{t}) + \boldsymbol{w}_i^T \boldsymbol{f}(\boldsymbol{t}) + \xi_i(\boldsymbol{t}) \\ &= \mu_i(\boldsymbol{t}) + \Sigma_{k=1}^{K} w_{ik} f_k(\boldsymbol{t}) + \xi_i(\boldsymbol{t}), \quad \boldsymbol{t} \in D, \quad i = 1, ..., n. \end{aligned} \tag{3.11}$$

We use $\boldsymbol{t}$ to keep notation consistent. But $\boldsymbol{s}$ is usually used in the spatial random effects model. $Z_i(\boldsymbol{t}) = (Z_i(\boldsymbol{t}_1), ..., Z_i(\boldsymbol{t}_p))$ is an independent spatial process observed at p locations on d dimensional domain $D \subset \mathbb{R}^d$. $n \geq 1$, with $n = 1$ for single realization in geostatistics. $Z_i(\boldsymbol{t})$ has mean $\mu_i(\boldsymbol{t})$ and spatial covariance $C(\boldsymbol{t}, \boldsymbol{t}^*) = cov(Z_i(\boldsymbol{t}), Z_i(\boldsymbol{t}^*))$. $\boldsymbol{f}(\boldsymbol{t}) = (f_1(\boldsymbol{t}), ..., f_K(\boldsymbol{t}))^T$ are K prespecified basis functions with $K \leq p$. $\boldsymbol{w}_i = (w_{i1}, ..., w_{iK})^T$ are unobservable random weights that follows $N(\boldsymbol{0}, M)$ with M be an unkown nonnegative definite covariance matrix. $\xi_i(\boldsymbol{t}) \sim N(0, \sigma_\xi^2)$ is a white-noise process and is uncorrelated with $\boldsymbol{w}_i$. We get $C(\boldsymbol{t}, \boldsymbol{t}^*) = \boldsymbol{f}(\boldsymbol{t})^T M \boldsymbol{f}(\boldsymbol{t}^*) + \sigma_\xi^2 \boldsymbol{I}$, where $\boldsymbol{I}$ is a $p \times p$ identity matrix with 1 on the diagonal and 0 elsewhere. Let F be the $p \times K$ matrix $[\boldsymbol{f}(\boldsymbol{t}_1), ..., \boldsymbol{f}(\boldsymbol{t}_p)]^T$, Cressie and Johannesson (2008) coined the term *fixed rank*, because the rank of FMF^T is less than or equal to the number of basis functions K.

Resolution adaptive FRK extracts basis functions from thin-plate splines (Green and Silverman, 1993) that balances the distance between noisy data Z and thin-plate splines

functions f and the smoothness penalty J,

$$f = arg\,min \sum_{\boldsymbol{t} \in D} (Z(\boldsymbol{t}) - f(\boldsymbol{t}))^2 + \rho J(f). \tag{3.12}$$

$\rho \geq 0$ is a tuning parameter. $J(f)$ is the penalty with a smaller value indicating a smoother function. Given $\boldsymbol{t} = (x_1, ..., x_d)^T$, the solution of (3.12) can be written in the form of $f(\boldsymbol{t}) = \boldsymbol{\alpha}^T \boldsymbol{\phi}(\boldsymbol{t}) + \beta_0 + \sum_{j=1}^{d} \beta_j x_j$, where $\boldsymbol{\phi}(\boldsymbol{t}) = (\phi_1(\boldsymbol{t}), ..., \phi_p(\boldsymbol{t}))^T$ are functions of spatial distance, $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_d)^T \in \mathbb{R}^{d+1}$, $\boldsymbol{\alpha} \in \mathbb{R}^p$ and $X^T \boldsymbol{\alpha} = \mathbf{0}$ with

$$X = \begin{pmatrix} 1 & x_{11} & ... & x_{1d} \\ \vdots & & \ddots & \\ 1 & x_{p1} & ... & x_{pd} \end{pmatrix}.$$

MRS basis functions are a set of basis functions that requires no knot allocation and is in descending order in terms of degrees of smoothness. Tzeng and Huang (2018) gave the definition of MRS basis functions,

$$f_k(\boldsymbol{t}) = \begin{cases} 1; & k = 1, \\ x_{k-1}; & k = 2, ..., d+1, \\ \lambda_{k-d-1}^{-1} (\boldsymbol{\phi}(\boldsymbol{t}) - \Phi X (X^T X)^{-1} x)^T \boldsymbol{v}_{k-d-1}; & k = d+2, ..., p, \end{cases} \tag{3.13}$$

where x is the row of matrix X, and Φ is a $p \times p$ matrix of the spatial distance functions, $[\boldsymbol{\phi}(\boldsymbol{t}_1), ..., \boldsymbol{\phi}(\boldsymbol{t}_p)]^T$. Let $Q = \boldsymbol{I} - X(X^T X)^{-1} X^T$, λ_k and $\boldsymbol{v}_k$ are the kth eigenvalue and the corresponding eigenvector of $Q\Phi Q$ with $\lambda_1 \geq ... \geq \lambda_p$. In the ordered function set, higher order functions show shape of the data in a larger scale, and lower order functions capture

more details of the data. When certain number of basis functions are selected, ordered function set includes basis functions that have more information of the data first. As a result, it represents data better with a relatively small number of basis functions.

Figure 3.1 and 3.2 are two examples of the MRS basis functions. Figure 3.1 has one-dimensional $(t_1, ..., t_{40})$ from $Uniform(0, 1)$. The irregularly spaced points are shown at the bottom of each subplot. We can see that the first 20 basis functions are in descending order of smoothness. Besides, more structures can be found at the locations where more data points are observed. It is ideal for nonparametric models, because MRS basis functions are able to show more details when there are more data, but also keep simple when there are fewer data in that area. Similarly, Figure 3.2 is the first 20 basis functions of 40 randomly generated points from two-dimensional space $[0, 1] \times [0, 1]$. The descending order of smoothness can be observed from f_1 to f_{20}. Finer structures can be found at locations of more data points, see the black points in the subplots.

Assuming that $\mu(\boldsymbol{t}) = 0$ in (3.11) for simplicity, the maximum likelihood estimator of M and σ_ξ^2 have closed form expressions and can be implemented using R package `autoFRK` (Tzeng et al., 2020). The number of basis functions is selected from $[d+1, K^*]$ automatically using Akaike's information criterion (AIC) (Akaike, 1973), $\hat{K} = arg\,min_{d+1 \le K \le K^*} AIC(K)$. K^* is a user defined sufficiently large value with default $K^* = p$ when $p \le 100$, and $K^* = 10\sqrt{p}$ when $p > 100$. AIC contains both variance and the number of free parameters. Thus, it provides optimal solution that balances both model precision and model simplicity.

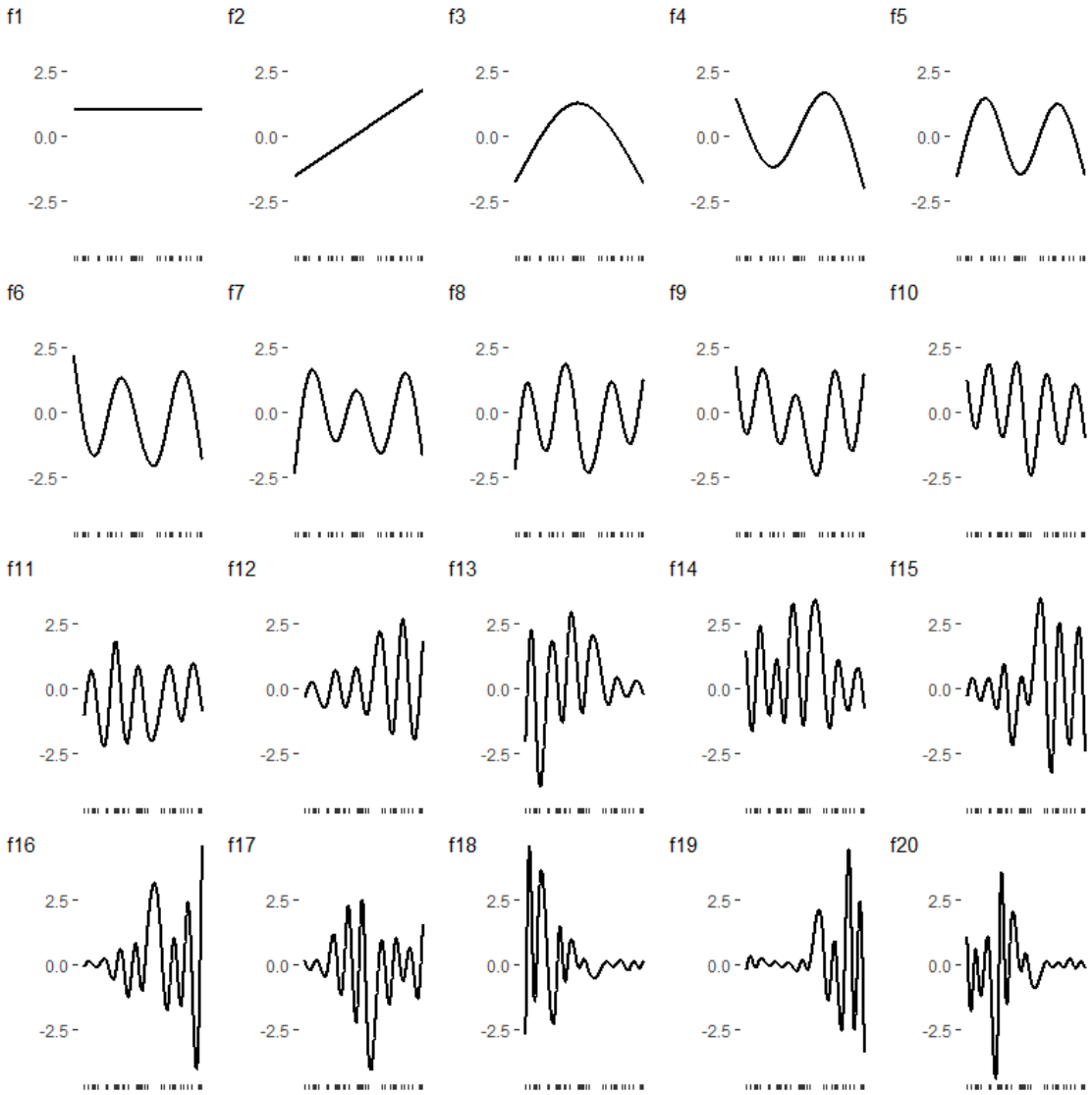


Figure 3.1: First 20 MRS basis functions of one-dimensional $\boldsymbol{t}$ from $Uniform(0, 1)$

Figure 3.2: First 20 MRS basis functions of two-dimensional $\boldsymbol{t}$ randomly generated from $[0,1] \times [0,1]$

3.3 Simulation

Two simulation studies are conducted in this section to compare the Bayesian hierarchical models with approximations by the widely accepted B-splines (BSP) basis functions and by the proposed resolution adaptive fixed rank kriging (FRK) basis functions. Simulation Study I has one-dimensional t_{ij}, $i = 1, ..., n$, $j = 1, ..., n_i$, while Simulation Study II has two-dimensional t_{kij}, $k = 1, 2$ and can be extended to higher dimensions. MCMC samples of

both studies consist of 2,000 burn-in and 10,000 post warm up iterations. Posterior means of BSP and FRK methods are calculated and shown in the plots in this section alongside with the true values. Root mean square error (RMSE) is used to evaluate the performance of the two methods. In Simulation Study I, BSP and FRK produce similar results when raw data is observed on equally-spaced grid. But FRK has smaller RMSE than BSP when data is observed on randomly generated grid. It implies that FRK estimations are closer to the true value than BSP on random grid. In Simulation Study II, BSP and FRK have about the same RMSE value, but FRK selects a smaller number of basis functions and thus has higher computational efficiency.

3.3.1 Simulation Study I

We consider four scenarios in this study, (1) stationary data on common grid, (2) nonstationary data on common grid, (3) stationary data on random grid, and (4) nonstationary data on random grid. Common grid has $p = 40$ equally spaced points over the time interval $\mathcal{T} = (0, \pi/2)$ for all curves. Random gird indicates that each curve has different $p = 40$ randomly generated points from $Uniform(0, \pi/2)$. $n = 30$ functional curves are generated for each scenario. For stationary data, the Gaussian process of the true signal is

$$Z_i(t) \sim GP\big(\mu(t) = 3\sin(4t), \quad \Sigma(t, t^*) = 5M(|t - t^*|; \rho, \nu)\big), \quad i = 1, ..., 30, \tag{3.14}$$

where M is Matérn covariance function $M(d; \rho, \nu) = \frac{1}{\Gamma(\nu)2^{\nu-1}}(\sqrt{2\nu}\frac{d}{\rho})^{\nu}K_{\nu}(\sqrt{2\nu}\frac{d}{\rho})$, with scale parameter $\rho = 0.5$ and order of smoothness parameter $\nu = 3.5$. $\Gamma(.)$ is Gamma function, and

$K_\nu(.)$ is the second kind modified Bessel function. For nonstationary data, the true signal is a nonlinear transformation of the stationary signal and has the form

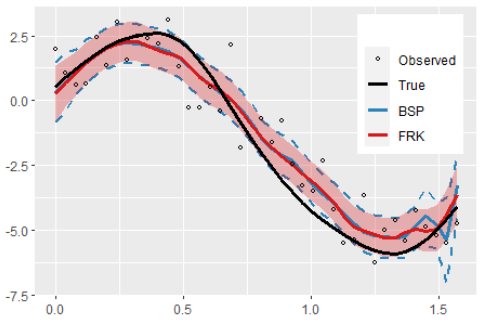
$$\tilde{Z}_i(t) = h(t)Z(s(t)), \quad h(t) = t + 1/2, \quad s(t) = t^{2/3}. \tag{3.15}$$

Equivalently, the nonstationary signal has mean function $\tilde{\mu}(t) = 3(t + 1/2)\sin(4t^{2/3})$ and covariance function $\tilde{\Sigma}(t, t^*) = 5(t + 1/2)(t^* + 1/2)M(|t^{2/3} - t^{*2/3}|; \rho, \nu)$. For both stationary and nonstationary cases, the noise term $\epsilon_i(t)$ in (3.1) follows $N(0, (\sqrt{5}/2)^2)$ and is included to get the raw functional data $Y_i(t) = Z_i(t) + \epsilon_i(t)$.

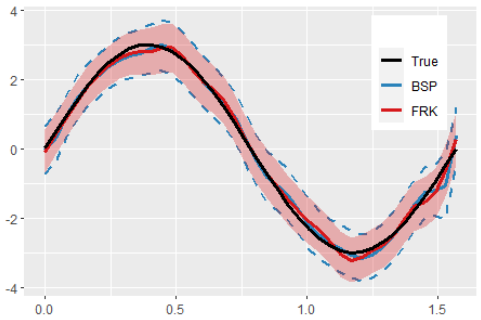
The initial values of mean function $\mu_z(\tau)$ and covariance function $\Sigma_z(\tau, \tau)$ are estimated using Principle Analysis by Conditional Expectation (PACE) (Yao et al., 2005) with the number of principle functions being selected by default to explain 99% fraction of variance. Next, we interpolate the PACE results on a $L = 20$ equally spaced working gird over $\mathcal{T}$ for common grid and a $(\frac{0}{L-1}, ..., \frac{L-1}{L-1})$ percentiles of the pooled observation grid for random grid. For BSP method, we specify the B-spline degree of freedom to be the number of time points on the working gird, $K = L$. The number of basis functions is therefore fixed at 20. For FRK method, K is selected by AIC and changes with the simulated data. For example, we get $K = 7$ for a stationary common grid data set, $K = 10$ for a nonstationary common grid data set, and $K = 18$ for a stationary and a nonstationary random grid data sets. Figure 3.3 shows one example of raw functional data (Y_i) and true signal curve (Z_i) from the 30 curves on the left panel, and mean function (μ) of the 30 curves on the right panel. Plots are arranged into four rows, which correspond to the four scenarios. In each row, the raw

functional data are shown in black dots, the true signal and the true mean of the 30 signals are shown in black lines. Estimations using BSP and FRK methods are in blue and red, respectively, with shaded bands indicating 95% credible intervals. The two methods both produce accurate results as the shaded areas cover most of the true value. BSP and FRK are almost overlapped when data is generated from common grid, while the latter has smoother curves toward the end of the first four plots. The two methods separate a little for the random grid data in the bottom two rows. FRK method provides mean estimates closer to the true mean curves than BSP in the second half of Figure 3.3f and 3.3h.

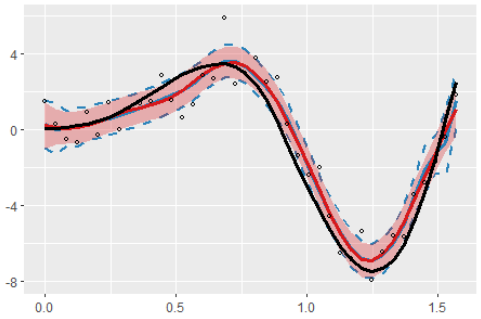
The simulation is repeated 100 times for each of the four scenarios. The mean and standard deviation of the 100 RMSE are calculated and shown in Table 3.1 with standard deviation in parenthesis and difference between average RMSE of BSP and FRK in Δ column. Mean function μ and covariance surface Σ are evaluated on the observation grid for common grid data, and on an equally spaced grid of length 40 over the range of pooled observation time for random grid data. Features that observed in Figure 3.3 can be found in the RMSE table. For the common grid data, basis functions using both methods perform equally well. FRK has RMSE about 0.01 less than BSP for Z_i and μ of both stationary and nonstationary data. The RMSE difference between the two methods is negligible for Σ and the noise term variance σ^2_ϵ. As shown in Figure 3.3, the difference increases when raw data is generated from random grid. FRK outperforms BSP by providing RMSE 0.05 and 0.08 less than BSP for stationary and nonstationary, respectively, for Z_i, and 0.08 and 0.28 less than BSP for μ. Σ of FRK has not only RMSE 0.29 and 0.56 less than BSP, it also has smaller standard



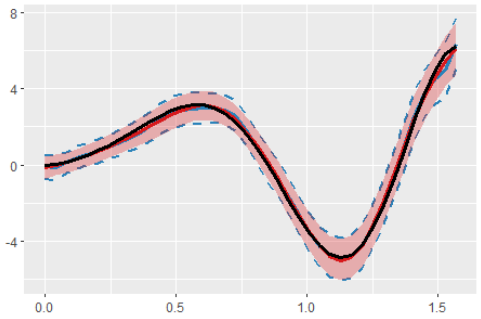
(a) Stationary curve estimates common grid



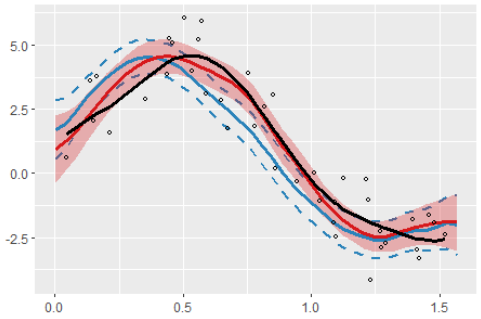
(b) Stationary mean estimates common grid



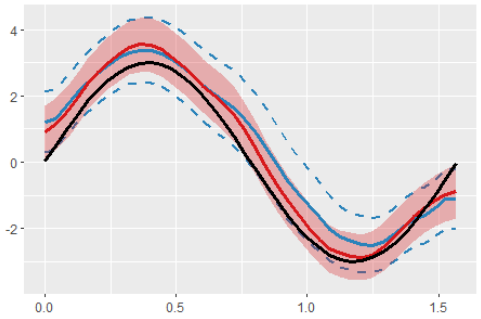
(c) Nonstationary curve estimates common grid



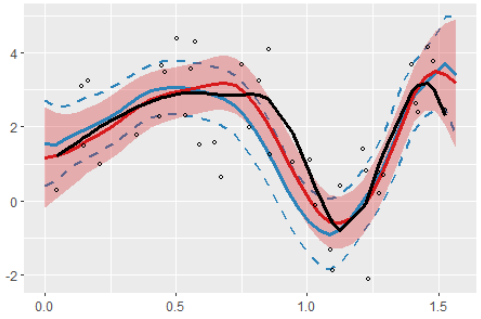
(d) Nonstationary mean estimates common grid



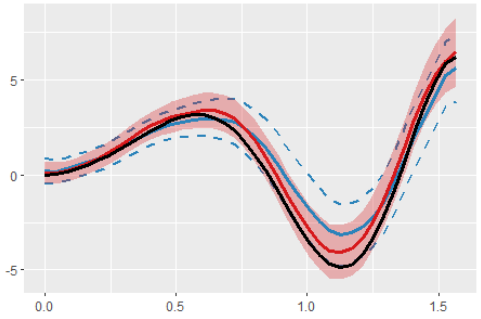
(e) Stationary curve estimates random grid



(f) Stationary mean estimates random grid



(g) Nonstationary curve estimates random grid



(h) Nonstationary mean estimates random grid

Figure 3.3: Example curve and mean estimate plots with 95% credible intervals

deviations, shown in parenthesis in the second last row of the table, and thus more stable performance. Looking at the table vertically, FRK gets similar results no matter the data is observed on common grid or random grid, for example, RMSE of Z_i is around $0.41 - 0.42$ for stationary data, and $0.48 - 0.49$ for nonstationary data, RMSE of Σ increases a little from 0.96 to 0.98 for stationary data, and 1.78 to 1.84 for nonstationary data. However, BSP is more sensitive to the observation grid. Using BSP basis function, RMSE of Z_i increases from 0.43 to 0.46 for stationary data, and 0.50 to 0.56 for nonstationary data, and Σ increases from 0.96 to 1.27 for stationary data, and 1.78 to 2.40 for nonstationary data. We conclude that FRK basis functions provide similar results to BSP basis functions when data is observed on common grid, and have better estimations when data is observed on random grid.

3.3.2 Simulation Study II

In the second simulation study, we generate $n = 10$ two-dimensional $\boldsymbol{t}$ functional data. The observation grid consists of t_1, $p_1 = 30$ equally spaced points over $\mathcal{T}_1 = (0, \pi/2)$, and t_2, $p_2 = 20$ randomly generated points from $\mathcal{T}_2 = (0, 1)$ uniform distribution. The true signal follows

$$\begin{aligned} Z_i(\boldsymbol{t}) \sim GP\Big(&\mu(t_1, t_2) = 3\big(t_2 \sin(4t_1) + (1 - t_2)\cos(4t_1)\big), \\ &\Sigma(t_1, t_1^*, t_2, t_2^*) = \Sigma(t_1, t_1^*) \otimes \Sigma(t_2, t_2^*)\Big), \quad i = 1, ..., 10. \end{aligned} \tag{3.16}$$

Covariance function $\Sigma(t, t^*) = 2M(|t - t^*|; \rho, \nu)$. Same as in Simulation Study I, M is the Matérn covariance function with $\rho = 0.5$ and $\nu = 3.5$. $\Sigma(t_1, t_1^*, t_2, t_2^*)$ is the Kronecher

		Stationary			Nonstationary		
		BSP	FRK	Δ	BSP	FRK	Δ
Common Grid	$Z_i(t)$	0.433 (0.021)	0.419 (0.018)	-0.014	0.503 (0.023)	0.490 (0.022)	-0.013
	$\mu(t)$	0.411 (0.151)	0.399 (0.156)	-0.013	0.536 (0.201)	0.530 (0.205)	-0.007
	$\Sigma(t,t)$	0.963 (0.364)	0.958 (0.351)	-0.004	1.783 (0.542)	1.778 (0.536)	-0.005
	σ^2_ϵ	0.052 (0.038)	0.049 (0.037)	-0.004	0.060 (0.041)	0.061 (0.040)	0.001
Random Grid	$Z_i(t)$	0.456 (0.023)	0.405 (0.022)	-0.052	0.557 (0.027)	0.479 (0.022)	-0.078
	$\mu(t)$	0.463 (0.137)	0.383 (0.139)	-0.079	0.813 (0.208)	0.537 (0.203)	-0.276
	$\Sigma(t,t)$	1.269 (0.687)	0.983 (0.340)	-0.286	2.400 (0.851)	1.840 (0.596)	-0.560
	σ^2_ϵ	0.073 (0.050)	0.053 (0.037)	-0.020	0.122 (0.062)	0.062 (0.046)	-0.060

Table 3.1: Average RMSEs (with standard deviation) and differences (Δ) of BSP and FRK methods

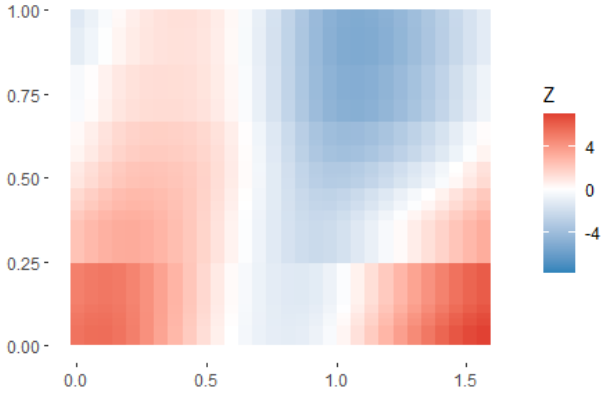
product of the two covariance matrices. Raw data is generated by adding the noise term $\epsilon_i(\boldsymbol{t}) \sim N(0, 1/2)$ to the true signal $Z_i(\boldsymbol{t})$.

Because of the two-dimensional $\boldsymbol{t}$, BSP method used in the Simulation Study I cannot be applied to this data directly. We first estimate mean function $\mu_z(\tau)$ and covariance function $\Sigma_z(\tau, \tau)$ using FRK instead of PACE and predict on a 10×10 working grid based on data density for the initial values. For BSP method, we get a B-spline basis matrix for each dimension, t_1 and t_2, with degree of freedom 10 from the working grid. The Kronecker product of the two basis matrices is used as the basis functions of the two-dimensional $\boldsymbol{t}$. For FRK method, we predict basis functions for the working grid with K selected automatically by AIC. That is to say, K is fixed at 100 when using BSP method, and the value of K depends on the data when using FRK method, for example, $K = 28$ for one of the simulated data sets. Because the computational complexity is $O(nK^3m)$ for n curves, K basis functions, and m MCMC iterations, the difference of K makes great impact on computing time and memory usage. For the data size and number of iterations in this simulation study, the reduction of K from 100 to 28 shortens MCMC sampling time from 21.8 minutes to 3.3 minutes by about 85%.

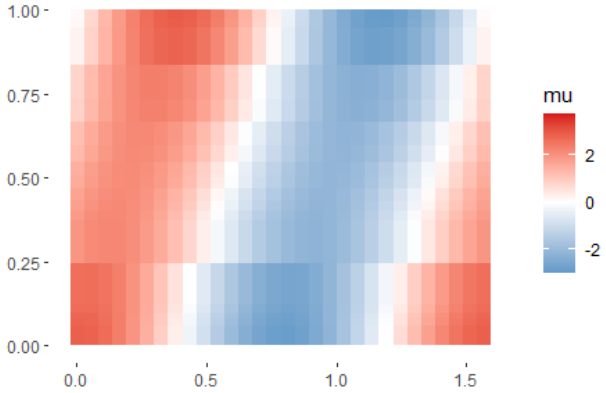
The results estimated using BSP and FRK basis functions are shown in Figure 3.4 on the 30×20 grid. Figure 3.4a is the heat map of one example true signals Z_i from the 10 generated surfaces. Figure 3.4b is the heat map of the true mean function μ of the 10 surfaces. The second row of Figure 3.4 is posterior mean of Z_i and μ using BSP basis function. The last row of Figure 3.4 is posterior mean of Z_i and μ using FRK basis function. The two methods

both give satisfying results for μ, as the three plots on the right hand side are similar. Z_i estimated by BSP and FRK basis functions are both able to reveal the color pattern of the true signal shown in Figure 3.4a. However, the color of BSP method becomes less clear at the edges of Figure 3.4c, or at the ends of the observation grid. It is coincident with what we saw in the one-dimensional counterpart in Figure 3.3, where BSP has less smooth estimates at the end of the curve than FRK.

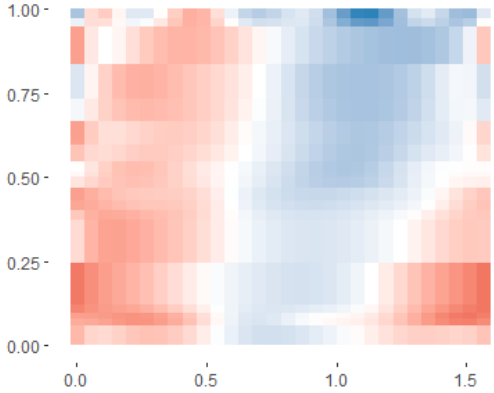
We again replicate the simulation 100 times as we did in Simulation Study I and calculate the average RMSE and its standard deviation of the two methods, shown in Table 3.2. FRK has RMSE 0.07 less than BSP for Z_i and 0.04 less than BSP for σ_ϵ^2. FRK and BSP has similar RMSE that are 0.01 apart for μ. Besides, because the numbers of basis functions are different for the two methods, we include K in Table 3.2 as well. BSP has K fixed at 100. Therefore it has standard deviation 0, shown in parenthesis. FRK has an average K 31.7 with standard deviation 7.9. The average computing time is 21.4 minutes for BSP method and 3.1 minutes for FRK. In the 100 simulations, the maximum number of basis functions for FRK is 53, which shortens the computing time from 21.7 to 6 minutes by 72.3%. The minimum number of basis functions is 3. It shortens the computing time from 18.6 to 1 minute by 94.6%, but it also brings higher RMSE. A smaller number of K without losing precision of estimation is important, especially when the observation grid is large or in higher dimensions.



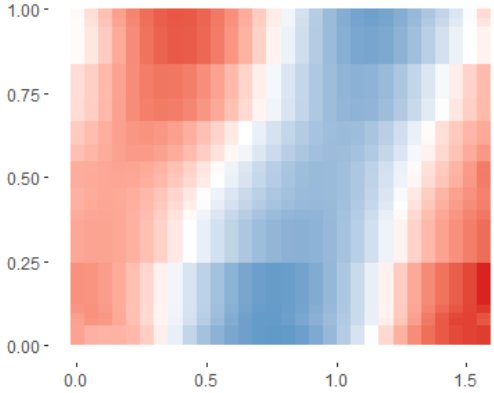
(a) True signal surface



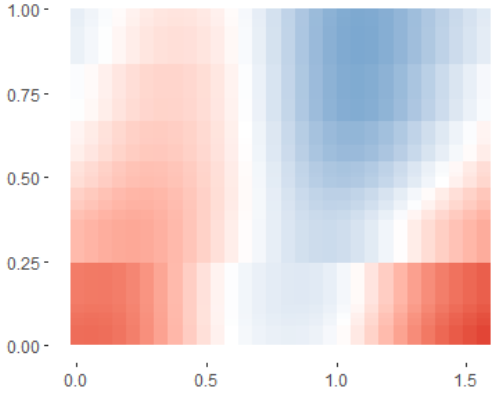
(b) True mean surface



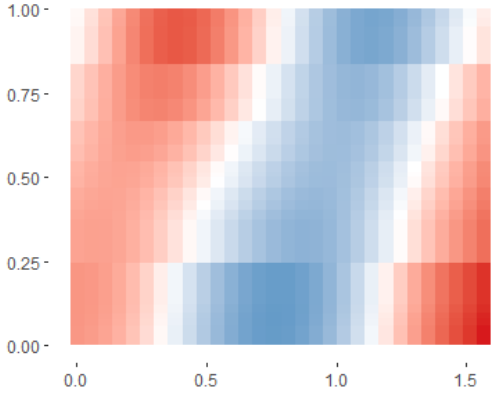
(c) Signal estimated by BSP basis functions



(d) Mean estimated by BSP basis functions



(e) Signal estimated by FRK basis functions



(f) Mean estimated by FRK basis functions

Figure 3.4: True and estimated signal (Z_i) and mean (μ) plots

	BSP	FRK	Δ
K	100 (0)	31.670 (7.897)	-68.330
$Z_i(t)$	0.252 (0.160)	0.187 (0.125)	-0.065
$\mu(t)$	0.604 (0.197)	0.612 (0.214)	0.009
σ^2_ϵ	0.068 (0.168)	0.031 (0.201)	-0.037

Table 3.2: Average number of basis functions (K) and average RMSEs and their differences (Δ)

3.4 Real Data

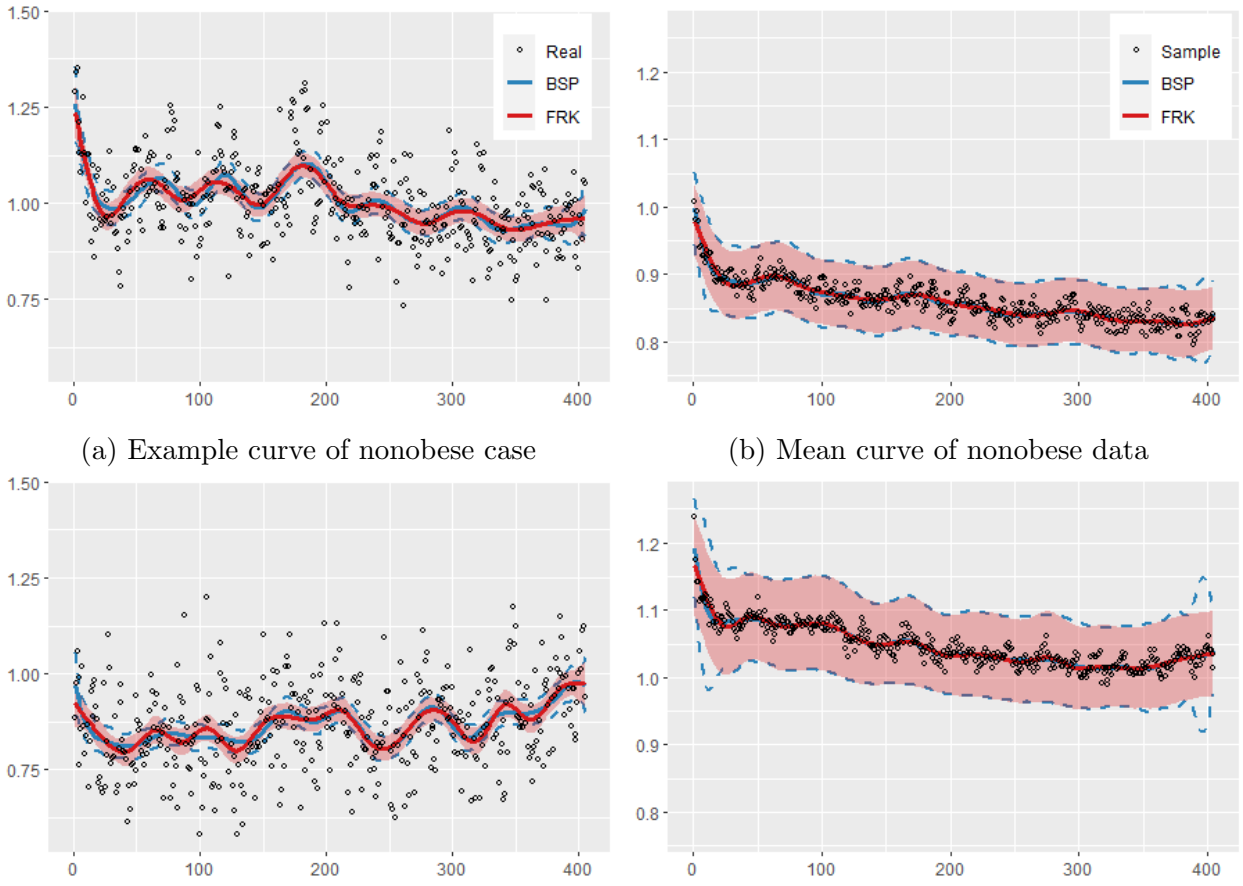
3.4.1 SEE Data

Sleeping energy expenditure (SEE) data set is from an obesity study conducted by the Children's Nutrition Research Center (CNRC) of Baylor College of Medicine in Houston, Texas (Lee et al., 2017). It measures SEE in unit of kcal for 5-19 years old children and adolescents. Data was recorded every minute during 12:00-7:00am using room respiration calorimeters. $n = 106$ subjects, including 44 obese cases and 62 nonobeses cases as control group, were measured at $p = 405$ time points. We take common grid t from $\mathcal{T} = [1 : 405]$, and working grid τ as $L = 20$ equally spaced points over $\mathcal{T}$. The Bayesian hierarchical models with BSP and FRK basis functions are applied to nonobese and obese groups seperately. In this case, $K = 20$ for the two groups when using BSP basis function, K, suggested by AIC, is 15 and 19 for nonobese and obese groups respectively when using FRK basis function.

Figure 3.5 shows one example curve and the sample mean from both groups in black circles. Posterior estimates of BSP and FRK are shown in the plot in blue and red, respectively.

For the example curves in Figure 3.5a and Figure 3.5c, we can see more periodic pattern in the nonobese case, while the points of the obese case is more scattered. FRK gives smoother posterior mean towards the end of the nonobese curve, and keeps more details from the first half of the obese data. For the sample mean curves in Figure 3.5b and Figure 3.5d, BSP and FRK give similar posterior means. However, 95% credible interval of BSP, shown between the blue dashed lines, oscillates a little at the beginning and the end of the mean curves. It is more obvious in the obese cases due to the characteristics of the obese SEE data.

SEE difference between the nonobese and obese groups is also examined in this study. Using the leave-one-out cross-validataion (LOOCV), we compare the misclassification rate among raw data, BSP smoothed data, and FRK smoothed data. For each curve, we train a support-vector machines (SVM) model (Cortes and Vapnik, 1995) with the rest of the curves, and predict if the curve belongs to nonobese or obese group. SVM prediction is implemented using R package `e1071` Version 1.7-4. Raw data gives a misclassification rate of 49.06%. Because the data is highly fluctuanting, as shown in the black circles in Figure 3.5a, it is reasonable that the two groups get mixed up and the classification is almost random. Using the posterior mean of signal from BSP and FRK methods, the misclassification rate drops to 41.51% and 31.13%, respectively. It suggests that, by removing noises, the smoothed data reveals underlying pattern of the obese and nonobese cases and therefore is beneficial to future research.



(a) Example curve of nonobese case

(b) Mean curve of nonobese data

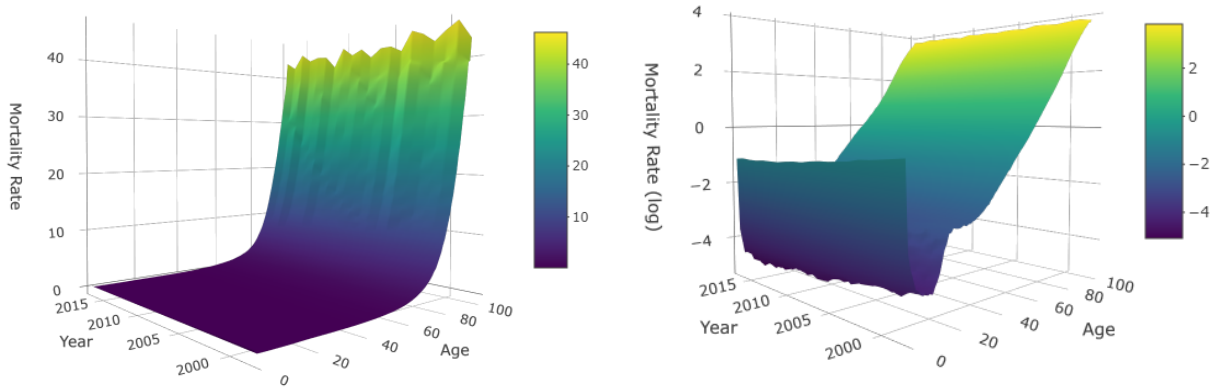
(c) Example curve of obese case

(d) Mean curve of obese data

Figure 3.5: Posterior mean and 95% credible interval of example and mean curves of SEE Data

3.4.2 Mortality Data

The Human Mortality Database (HMD) (University of California, Berkeley, and Max Planck Institute for Demographic Research, 2020) contains original calculations of death rates and life tables from 41 countries and areas. Due to insufficient data, we use mortality rate of age 0-100 between 1998 and 2017 for the following $n = 12$ countries, the United States of America, Japan, Germany, France, Italy, Australia, Netherlands, Belgium, Greece, Czechia,



(a) Original sample mean

(b) Log transformed sample mean

Figure 3.6: Original and log transformed sample mean of mortality data

Sweden, and Portugal, in this study. The sample mean of the 12 countries is shown in Figure 3.6a. Mortality rate increases dramatically after age 60. A small peak happens at age 0. Although it is almost invisible at current scale, it is obvious in the log transformation plot in Figure 3.6b. To avoid abrupt changes in the direction of the surface during the infant stage and to flatten the ascending slope at the senior stage, we use the log transformed data from age 2-100 in this section. Results of age 0-100 will be shown in Appendix B. We therefore have the two-dimensional common grid $\boldsymbol{t}$ with t_1, 99 equally spaced points over $\mathcal{T}_1 = [2, 100]$, and t_2, 20 equally spaced points over $\mathcal{T}_2 = [1998, 2017]$.

To study the application of our proposed method, we fit half of the data, with $t_1 = 2, 4, 6, ..., 100$, to the Bayesian hierarchical model. We set BSP basis functions as Kronecker product of two B-Spline basis matrices with degree of freedom 20 and 10 for t_1 and t_2, respectively. The number of basis functions is $K = 200$ from the dimension of the two basis matrices product. Figure 3.7 shows AIC of FRK for selected Ks. Suggested by the AIC

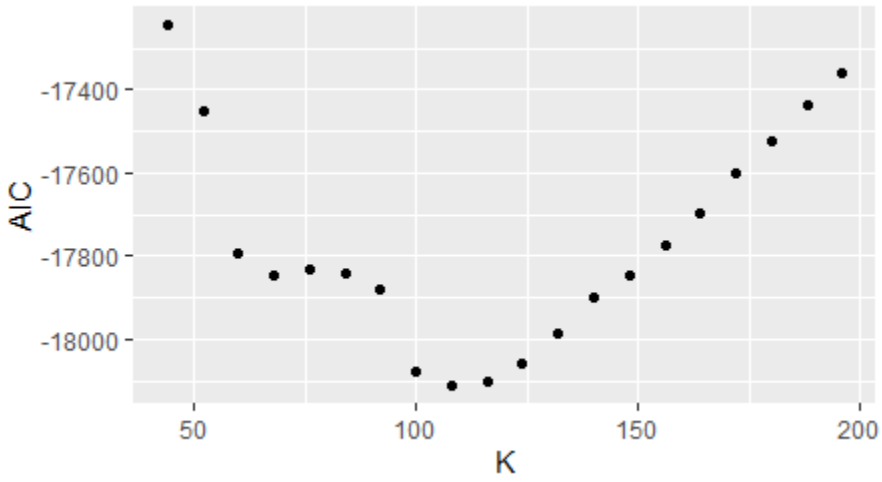


Figure 3.7: AIC of selected Ks of the 12 countries mortality data

results, $K = 108$ for this data. We then estimate the mortality rate at age $3, 5, ..., 99$ between year 1998 and 2017 for the 12 countries using the two methods. The RMSE between real data and the estimated values using BSP basis functions is 0.2333. Decreasing by 38.41%, the RMSE between real data and the estimated values using FRK basis functions is 0.1437. 67.02% of the $49 \times 20 \times 12 = 11760$ testing data points are covered within the 95% credible interval using BSP basis functions, while the rate increases to 81.67% using FRK basis functions. BSP method can be improved by increasing the number of knots that define the splines. It will, however, greatly increase the computational complexity and the required memory size.

Taking the United States as an example, the predicted values are shown in Figure 3.8. In general, both methods predict mortality rates close to the real data. We can see the blue to red color gradient from bottom to top of the three plots, which means that the mortality rate increases with respect to age. Besides, the color are slightly tilted with lower

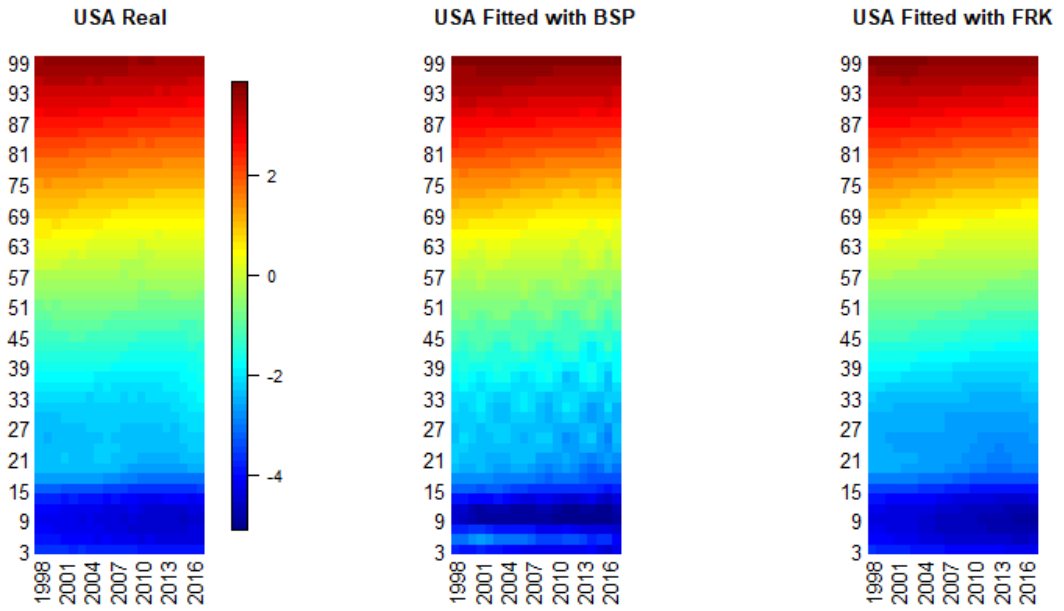


Figure 3.8: USA real and estimated mortality rate of age 2-100

left hand side and higher right hand side, which indicates that the mortality rate decreases with respect to year due to the development of medical science and social welfare. However, BSP (Figure 3.8 center panel) provides higher estimated mortality rate than the real data at the bottom edge of the plot. It does not get similar results until around age 9. BSP also shows a periodic pattern horizontally between age 20-30, which is not observed in the real data. Similar situation happens in mortality rates prediction of other countries. FRK basis functions are preferred for this Bayesian hierarchical model, because it maintains a more stable performance in prediction.

3.5 Conclusion

Inspired by the resolution adaptive fixed rank kriging and its corresponding class of basis functions, this chapter develops a Bayesian hierarchical model that estimates mean and covariance functions simultaneously and nonparametrically with reduced dimension through approximations by multi-resolution spline basis functions. Some features of the basis function include that they are sorted in the descending order of smoothness, and the number of basis functions is suggested based on AIC. The model thus has improved computational efficiency while keeping the precision of the posterior inference. The advantage is demonstrated in the numerical studies through comparison between the proposed model and the model using B-splines basis functions.

In the simulation studies, the model with FRK basis functions not only provides a preferred results for RMSE, it also has a smoother estimate while approaching to the endpoints in the one dimensional case and to the edges in the two dimensional case. When RMSE is not available in the SEE data, the signals estimated by the proposed model decreases misclassification rate of obese and control groups, comparing with the original data and the estimate from model using BSP basis functions. The proposed model once again gets smaller RMSE and covers more true values in 95% credible interval for the two-dimensional mortality data, while using about half of the number of basis functions that is used in BSP method.

As shown in the two-dimensional simulated data example in Section 3.3.2 and the mor-

tality data example that recorded on grid of age and year in Section 3.4.2, the Bayesian hierarchical model with basis functions approximations can be expanded to two or higher dimensions directly. It is noteworthy that Yang et al. (2017)'s model of functional data can be applied to curves in $\mathbb{R}^3$. In a recent brain fiber bundles study by Zhang et al. (2019), each fiber connecting a pair of brain regions is viewed as a parameterized curve from the tractography data set. The fiber curves are decomposed to shape, translation, and rotation components. Each coordinates are then fitted to the Bayesian hierarchical model independently.

Chapter 4

Summary

This dissertation consists of two projects. Each project develops a Bayesian hierarchical model for a certain scenario. The set up of the model and the procedures to get the posterior inference are given in details. The advantages of the models are proved both theoretically and numerically. Simulated data sets are used to compare the proposed models with a few alternative models. The proposed models are also applied to several real data sets for illustration.

We start with a brief introduction of the background of the related topics in Chapter 1. These topics include Bayesian hierarchical modeling, the spatial frailty terms in disease mapping, conditional autoregressive modeling, functional data analysis, functional principal component analysis, and spline basis functions.

Chapter 2 is motivated by the cucurbit downy mildew (CDM) impPIPE data. The data has information about plant type, state, survival days within the 204 days observation pe-

riod, and duration days if outbreak. We, accordingly, develop a joint Bayesian hierarchical model with Weibull distribution for the censored time to disease outbreak data, a zero-truncated Poisson distribution for the disease duration data, and a generalized multivariate conditionally autoregressive (GMCAR) model for the spatial frailty terms. The model is appropriate because of the annual extinction-colonization cycles and significant long distance spread at the continental scale of the CDM disease. It is then compared with three other models that have no dependency or spatial information, namely, univariate conditional autoregressive (UniCAR) model, multivariate normal (MvNorm) model, and independent normal (IndNorm) model. The models are selected using average mean square error (AMSE), for simulated data, and deviance information criterion (DIC) and leave-one-out information criterion (LOOIC). The proposed hierarchical model in general outperforms the other three models by providing smaller bias and preferred DIC and LOOIC results. Applying the model to the 2009 CDM impPIPE data, we conclude that states in the mid-Atlantic region tend to have a higher risk of disease outbreak, and in the infected cases, they are likely to have a longer duration of CDM among the 23 states in the eastern United States.

We propose another Bayesian hierarchical model in Chapter 3. The framework uses approximations by basis functions to estimate underlying functional signals. The Markov chain Monte Carlo (MCMC) steps are therefore conducted in a reduced rank space generated by the basis functions. The data model is the common functional data model with measurement error. The true signals have shared mean and covariance functions that follow Gaussian-Wishart distributions in the process model. The mean and covariance functions

are transformed by the multi-resolution spline basis functions with the full conditional distribution and the algorithm given explicitly in Section 3.2.1. The performance of the proposed model is compared with the model using B-splines basis functions. It provides a smaller root mean square error (RMSE) with a smaller number of basis functions for the simulated data. The improvement in computational efficiency is important in high dimensional functional data analysis. It later shows that using the proposed model, the smoothed data has a smaller misclassification rate for the sleeping energy expenditure (SEE) data. It also has a smaller RMSE and a preferred credible interval coverage for the 12 countries mortality data.

Bibliography

Akaike H (1973) Maximum likelihood identification of gaussian autoregressive moving average models. Biometrika 60(2):255–265

Banerjee S, Wall MM, Carlin BP (2003) Frailty modeling for spatially correlated survival data, with application to infant mortality in minnesota. Biostatistics 4(1):123–142

Banerjee S, Gelfand AE, Finley AO, Sang H (2008) Gaussian predictive process models for large spatial data sets. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70(4):825–848

Banerjee S, Finley AO, Waldmann P, Ericsson T (2010) Hierarchical spatial process models for multiple traits in large genetic trials. Journal of the American Statistical Association 105(490):506–521

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

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

Blangiardo M, Cameletti M, Baio G, Rue H (2013) Spatial and spatio-temporal models with r-inla. Spatial and spatio-temporal epidemiology 4:33–49

Bradley JR, Cressie N, Shi T (2015) Comparing and selecting spatial predictors using local criteria. Test 24(1):1–28

Bradley JR, Cressie N, Shi T (2016) A comparison of spatial predictors when datasets could be very large. Statistics Surveys 10:100–131

Carlin BP, Hodges JS (1999) Hierarchical proportional hazards regression models for highly stratified data. Biometrics 55(4):1162–1170

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

Chiou JM (2012) Dynamical functional prediction and classification, with application to traffic flow prediction. The Annals of Applied Statistics 6(4):1588–1614

Christiano R, Scherm H (2007) Quantitative aspects of the spread of asian soybean rust in the southeastern united states, 2005 to 2006. Phytopathology 97(11):1428–1433

Cortes C, Vapnik V (1995) Support-vector networks. Machine learning 20(3):273–297

Cressie N (1993) Statistics for Spatial Data. Wiley Series in Probability and Statistics, Wiley, URL https://books.google.com/books?id=4SdRAAAAMAAJ

Cressie N, Johannesson G (2008) Fixed rank kriging for very large spatial data sets. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70(1):209–226

Cressie N, Shi T, Kang EL (2010) Fixed rank filtering for spatio-temporal data. Journal of Computational and Graphical Statistics 19(3):724–745

Crowl TA, Crist TO, Parmenter RR, Belovsky G, Lugo AE (2008) The spread of invasive species and infectious disease as drivers of ecosystem change. Frontiers in Ecology and the Environment 6(5):238–246

Dawid AP (1981) Some matrix-variate distribution theory: notational considerations and a bayesian application. Biometrika 68(1):265–274

De Boor C, De Boor C, Mathématicien EU, De Boor C, De Boor C (1978) A practical guide to splines, vol 27. springer-verlag New York

Edmond M, Wong C, Chuang S (2011) Evaluation of sentinel surveillance system for monitoring hand, foot and mouth disease in hong kong. Public Health 125(11):777–783

Finley AO, Sang H, Banerjee S, Gelfand AE (2009) Improving the performance of predictive process modeling for large datasets. Computational statistics & data analysis 53(8):2873–2884

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

Gelman A, Rubin DB, et al. (1992) Inference from iterative simulation using multiple sequences. Statistical science 7(4):457–472

Green PJ, Silverman BW (1993) Nonparametric regression and generalized linear models: a roughness penalty approach. Crc Press

Guo W (2002) Functional mixed effects models. Biometrics 58(1):121–128

He B, Luo S (2016) Joint modeling of multivariate longitudinal measurements and survival data with applications to parkinson's disease. Statistical methods in medical research 25(4):1346–1358

Hoffman MD, Gelman A (2014) The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. Journal of Machine Learning Research 15(1):1593–1623

Ibrahim JG, Chu H, Chen LM (2010) Basic concepts and methods for joint models of longitudinal and survival data. Journal of Clinical Oncology 28(16):2796

Jin X, Carlin BP, Banerjee S (2005) Generalized hierarchical multivariate car models for areal data. Biometrics 61(4):950–961

Kang EL, Cressie N (2011) Bayesian inference for the spatial random effects model. Journal of the American Statistical Association 106(495):972–983

Kim H, Sun D, Tsutakawa RK (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(456):1506–1521

Kuk AY, Chen CH (1992) A mixture model combining logistic regression with proportional hazards regression. Biometrika 79(3):531–541

Lawson AB, Carroll R, Castro M (2014) Joint spatial bayesian modeling for studies combining longitudinal and cross-sectional data. Statistical methods in medical research 23(6):611–624

Lee JS, Zakeri IF, Butte NF (2017) Functional data analysis of sleeping energy expenditure. PloS one 12(5):e0177,286

Lindgren F, Rue H (2015) Bayesian spatial modelling with r-inla. Journal of statistical software 63:1–25

Lindgren F, Rue H, Lindström J (2011) An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation approach. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 73(4):423–498

Madden LV, Hughes G, Van Den Bosch F (2007) The study of plant disease epidemics. Am Phytopath Society

Mardia K (1988) Multi-dimensional multivariate gaussian markov random fields with application to image processing. Journal of Multivariate Analysis 24(2):265–284

Matérn B (1960) Spatial variation-stochastic models and their application to some problems in forest surveys and other sampling investigations. meddelanden fran statens skogsforskningsintitut, almaenna foerlaget, stockholm. (1986), 49 (5)

Meyer MC, et al. (2008) Inference using shape-restricted regression splines. The Annals of Applied Statistics 2(3):1013–1033

Morris JS, Carroll RJ (2006) Wavelet-based functional mixed models. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68(2):179–199

Nathoo FS (2010) Joint spatial modeling of recurrent infection and growth with processes under intermittent observation. Biometrics 66(2):336–346

Neelon B, Ghosh P, Loebs PF (2013) A spatial poisson hurdle model for exploring geographic variation in emergency department visits. Journal of the Royal Statistical Society: Series A (Statistics in Society) 176(2):389–413

Neelon B, Gelfand AE, Miranda ML (2014) A multivariate spatial mixture model for areal data: examining regional differences in standardized test scores. Journal of the Royal Statistical Society: Series C (Applied Statistics) 63(5):737–761

Nychka D, Bandyopadhyay S, Hammerling D, Lindgren F, Sain S (2015) A multiresolution gaussian process model for the analysis of large spatial datasets. Journal of Computational and Graphical Statistics 24(2):579–599

Ojiambo P, Holmes G (2011) Spatiotemporal spread of cucurbit downy mildew in the eastern united states. Phytopathology 101(4):451–461

Ojiambo P, Kang E (2013) Modeling spatial frailties in survival analysis of cucurbit downy mildew epidemics. Phytopathology 103(3):216–227

Ojiambo P, Holmes G, Britton W, Babadoost M, Bost S, Boyles R, Brooks M, Damicone J, Draper M, Egel D, et al. (2011) Cucurbit downy mildew ipmpipe: a next generation web-based interactive tool for disease management and extension outreach. Plant health progress 12(1):26

Ojiambo PS, Gent DH, Quesada-Ocampo LM, Hausbeck MK, Holmes GJ (2015) Epidemiology and population biology of pseudoperonospora cubensis: A model system for management of downy mildews. Annual review of Phytopathology 53:223–246

Ojwang AM, Ruiz T, Bhattacharyya S, Chatterjee S, Ojiambo PS, Gent DH (2021) A general framework for spatio-temporal modeling of epidemics with multiple epicenters: Application to an aerially dispersed plant pathogen. Frontiers in Applied Mathematics and Statistics p 72

Papageorgiou G, Mauff K, Tomer A, Rizopoulos D (2019) An overview of joint modeling of time-to-event and longitudinal outcomes. Annual review of statistics and its application 6:223–240

Peng Y, Dear KB (2000) A nonparametric mixture model for cure rate estimation. Biometrics 56(1):237–243

Ramsay JO, et al. (1988) Monotone regression splines in action. Statistical science 3(4):425–441

Randrianasolo L, Raoelina Y, Ratsitorahina M, Ravolomanana L, Andriamandimby S, Heraud JM, Rakotomanana F, Ramanjato R, Randrianarivo-Solofoniaina AE, Richard V (2010) Sentinel surveillance system for early outbreak detection in madagascar. BMC Public Health 10(1):31

Rizopoulos D, Taylor JM, Van Rosmalen J, Steyerberg EW, Takkenberg JJ (2016) Personalized screening intervals for biomarkers using joint models for longitudinal and survival data. Biostatistics 17(1):149–164

Rue H, Martino S, Chopin N (2009) Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. Journal of the royal statistical society: Series b (statistical methodology) 71(2):319–392

Schrödle B, Held L (2011) Spatio-temporal disease mapping using inla. Environmetrics 22(6):725–734

Silverman BW (1996) Smoothed functional principal components analysis by choice of norm. The Annals of Statistics 24(1):1–24

Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A (2002) Bayesian measures of model complexity and fit. Journal of the royal statistical society: Series b (statistical methodology) 64(4):583–639

Stan Development Team (2019a) Stan reference manual. https://mc-stan.org/docs/2_20/reference-manual-2_20.pdf, accessed: 2019-08-30

Stan Development Team (2019b) Stan user's guide. https://mc-stan.org/docs/2_20/stan-users-guide-2_20.pdf, accessed: 2019-08-30

Tzeng S, Huang HC (2018) Resolution adaptive fixed rank kriging. Technometrics 60(2):198–208

Tzeng S, Huang HC, Wang W, Nychka D, Gillespie C (2020) Automatic fixed rank kriging. https://cran.r-project.org/web/packages/autoFRK/autoFRK.pdf, accessed: 2020-11-03

Ullah S, Finch CF (2013) Applications of functional data analysis: A systematic review. BMC medical research methodology 13(1):1–12

University of California, Berkeley, and Max Planck Institute for Demographic Research (2020) Human mortality database. https://www.mortality.org/, accessed: 2020-05-01

Vehtari A, Gelman A, Gabry J (2017) Practical bayesian model evaluation using leave-one-out cross-validation and waic. Statistics and computing 27(5):1413–1432

Wu L, Liu W, Yi GY, Huang Y (2012) Analysis of longitudinal and survival data: joint modeling, inference methods, and issues. Journal of Probability and Statistics 2012

Yang J, Zhu H, Choi T, Cox DD, et al. (2016) Smoothing and mean–covariance estimation of functional data with a bayesian hierarchical model. Bayesian Analysis 11(3):649–670

Yang J, Cox DD, Lee JS, Ren P, Choi T (2017) Efficient bayesian hierarchical functional data analysis with basis function approximations using gaussian–wishart processes. Biometrics 73(4):1082–1091

Yao F, Müller HG, Wang JL (2005) Functional data analysis for sparse longitudinal data. Journal of the American statistical association 100(470):577–590

Zadoks J, Van den Bosch F (1994) On the spread of plant disease: a theory on foci. Annual review of phytopathology 32(1):503–521

Zhang D, Chen MH, Ibrahim JG, Boye ME, Shen W (2017) Bayesian model assessment in joint modeling of longitudinal and survival data with applications to cancer clinical trials. Journal of Computational and Graphical Statistics 26(1):121–133

Zhang Z, Descoteaux M, Dunson DB (2019) Nonparametric bayes models of fiber curves connecting brain regions. Journal of the American Statistical Association

Zhou H, Lawson AB, Hebert JR, Slate EH, Hill EG (2008) Joint spatial survival modeling for the age at diagnosis and the vital outcome of prostate cancer. Statistics in medicine 27(18):3612–3628

Zhu H, Brown PJ, Morris JS (2011) Robust, adaptive functional regression in functional mixed model framework. Journal of the American Statistical Association 106(495):1167–1179

Appendix A

Appendix for Chapter 2

This section gives some extra details for the joint spatial model study in Chapter 2. Figure A.1 and A.2 are supplements to Figure 2.2. Other than the six states chosen for Figure 2.2 where more observations are available, figure A.1 and A.2 are box plots of the full data set. Figure A.1 shows the uncensored survival data before the period of 204 days ends. Disease outbreak is not observed for certain hosts in a few states, such as Alabama and Massachusetts. Some states have very limited observations, such as Louisiana and Wisconsin. Figure A.2 shows the duration days when the outbreak happens. Less data is available for the duration part. For example, Louisiana, Mississippi, and Texas have only one observation, Alabama, New York, and Ohio have no data recorded at all. In this case, the estimation needs information from the neighbouring states or the survival part of the model.

Apart from AMSE of $\hat{z}_{kij}^{(t)}$, DIC, and LOOIC that evaluate fixed and random effects at the same time, we also examine the fixed effects coefficients only, which corresponds to hosts

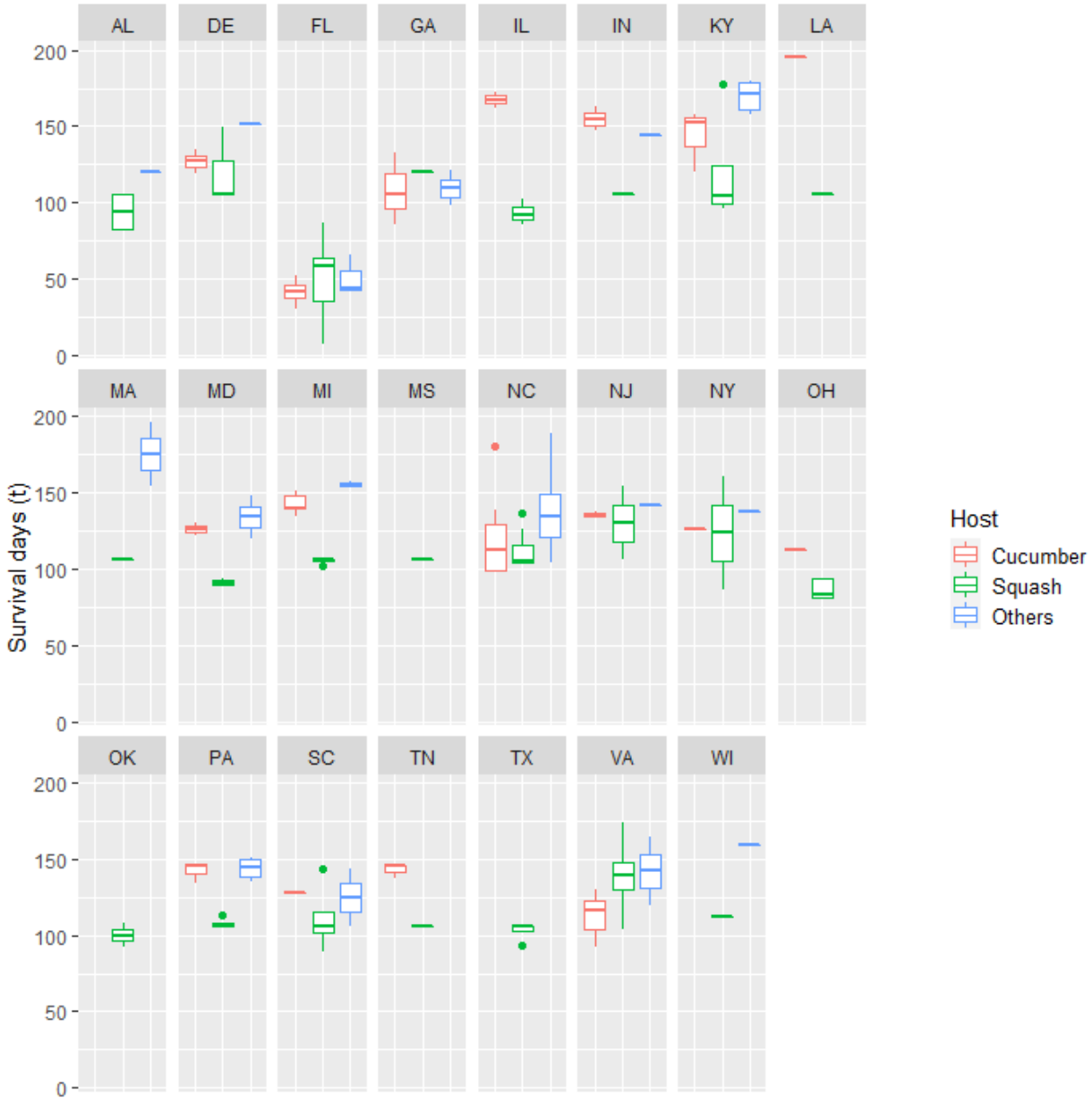


Figure A.1: Box plots of uncensored survival days data from 23 states

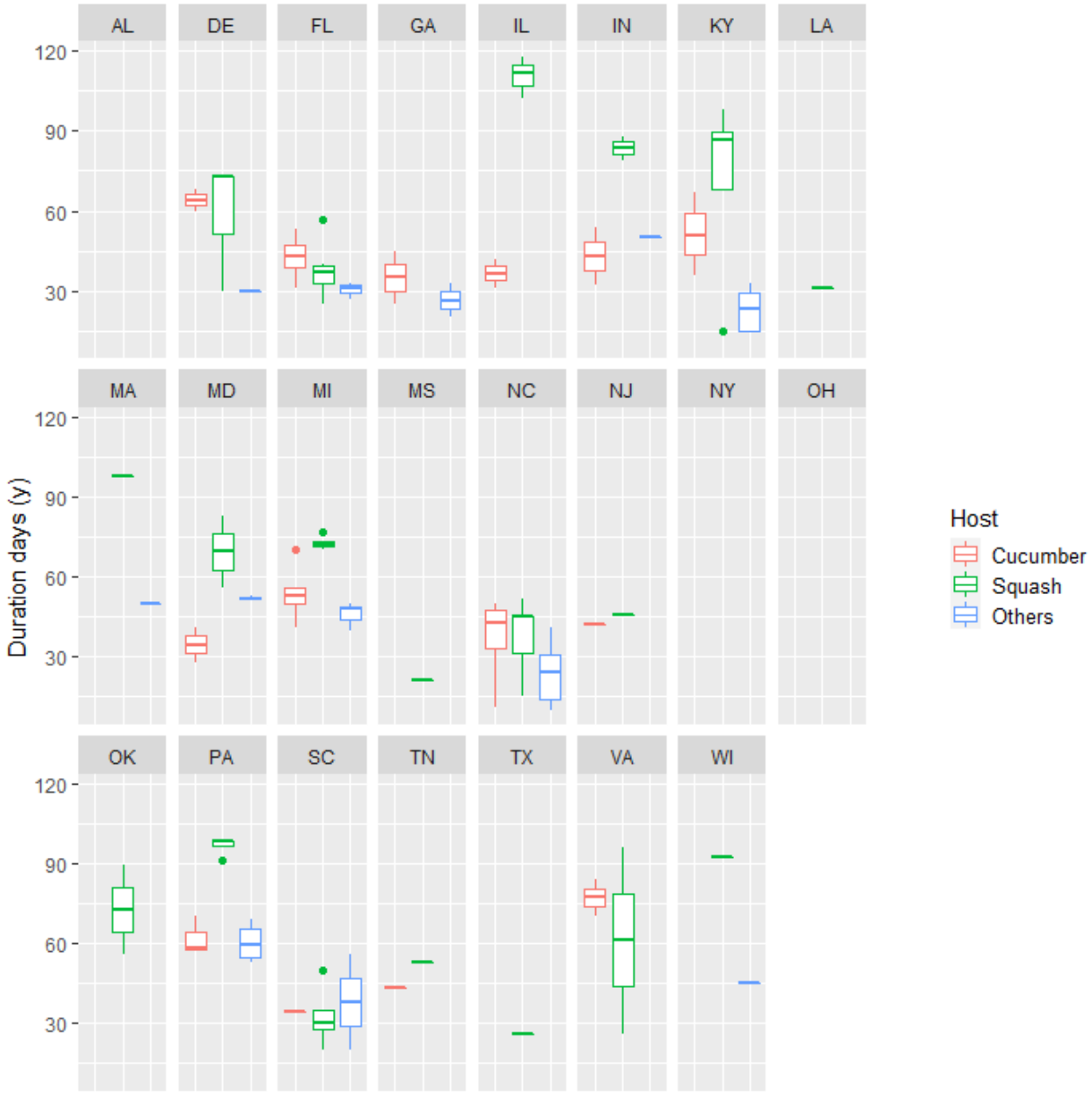


Figure A.2: Box plots of disease duration days data from 23 states

and is of interest to plant pathologists, in the simulation studies. Figure A.3 is a box plot of 50 posterior means of $\boldsymbol{\beta}$ for each of the four studies. X-axis indicates the true model, color indicates the fitted model, and a black horizontal line shows the true value. The left panel is the three coefficients in the survival model, the right panel is the three coefficients in the duration model. Fitting to the five models gives similar results in terms of median and interquartile range. The two Norm models has medians of $\hat{\boldsymbol{\beta}}_1$ slightly farther away from the true value than the ones of the three CAR models. Unlike $\hat{\boldsymbol{\beta}}_1$ in the survival model that is estimated using censoring data, $\hat{\boldsymbol{\beta}}_2$ in the duration model is easier to estimate and has posterior means very close to the true values, which is shown by the narrower ranges and overlapped black lines. We also have shorter boxes for Study 3 and 4 in Figure A.3 right panel due to the simulation parameter setting. When the true model is GMCAR, UniCAR has a little bit wider $\hat{\boldsymbol{\beta}}_2$ interquartile range comparing to the other four models, same as the reverse ordered GMCAR when the true model is UniCAR. However, they are only visible for the interquartile range and become indifferent in terms of the range of the 50 posterior means. Putting $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_2$ together, the posterior means confirm the AMSE conclusion in Table 2.3 that CAR models in general give preferable results in these four scenarios.

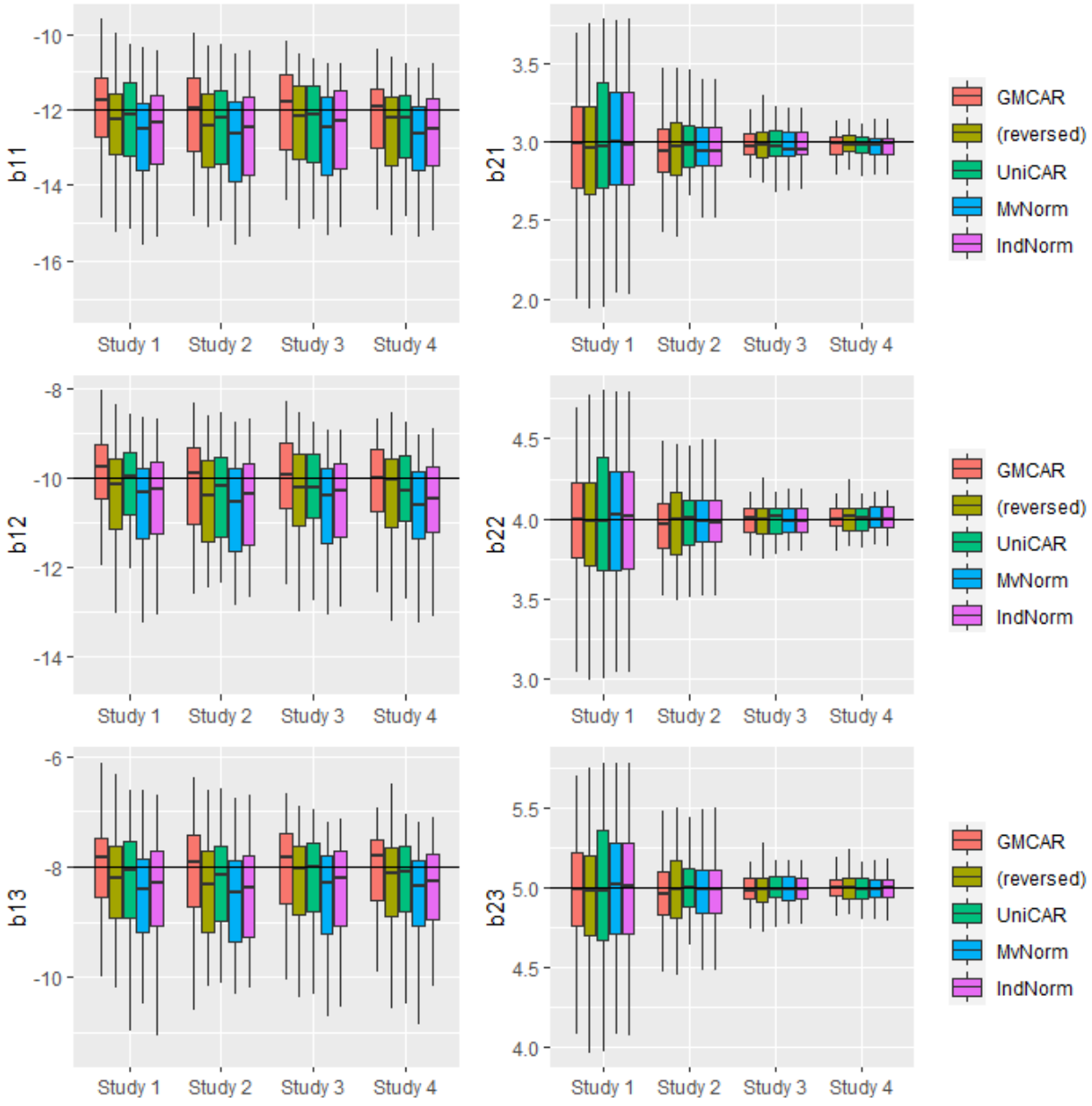


Figure A.3: Box plots of posterior means of the 50 Monte Carlo runs in the simulation study

Appendix B

Appendix for Chapter 3

In Section 3.4.2, we used the log transformation of the mortality data from age 2-100, in order to avoid abrupt changes in the direction of the surface during the infant stage. We now try fitting the Bayesian hierarchical model with age 0-100 log transformed data and compare the performance of BSP and FRK methods when functional surfaces show steep slope and direction change. Again, we train the model with half of the data, $t_1 = 0, 2, 4, ..., 100$. With the same setting as in Section 3.4.2, the number of basis function is $K = 200$ for the BSP method. Selected by AIC, $K = 190$ for the FRK method. The mortality rate is then estimated at age $1, 3, 5, ..., 99$ between year 1998 and 2017 for the 12 countries using the two methods. The RMSE between real data and the estimated values using BSP basis functions is 0.3671. The RMSE between real data and the estimated values using FRK basis functions is 0.1772, less than half of the RMSE of BSP. 58.1% of the 12000 testing data points are covered within the 95% credible interval using BSP basis functions. The rate increases to

74.55% using FRK basis functions.

Figure B.1 shows the real and predicted values of the mortality rate of the United States. Both methods display the pattern from blue (bottom) to red (top), or mortality rate from low to high, which is similar to the real data plot. BSP has a periodic pattern horizontally between age 20-35, which is not observed in the plots of real data and FRK estimate. In contrast to the ascending trend of mortality rate with respect to age, age 1 has a higher mortality rate comparing to the following few years. It is not well represented using both methods. While the true mean of log mortality rate of age 1 in 1998-2017 is -3.09, the average posterior mean is -0.83 for BSP method and -2.07 for FRK method. Figure B.2 shows the real and predicted values of the mortality rate of Japan as another example. In the left panel, a lighter blue color, indicating a higher mortality rate, can be observed in 2011 in the lower half of the plot. However, both BSP and FRK smooth the data and are not able to show the characteristic. The RMSE between fitted value using BSP and the true value in the year of 2011 is 0.086. It decreases to 0.071 using FRK.

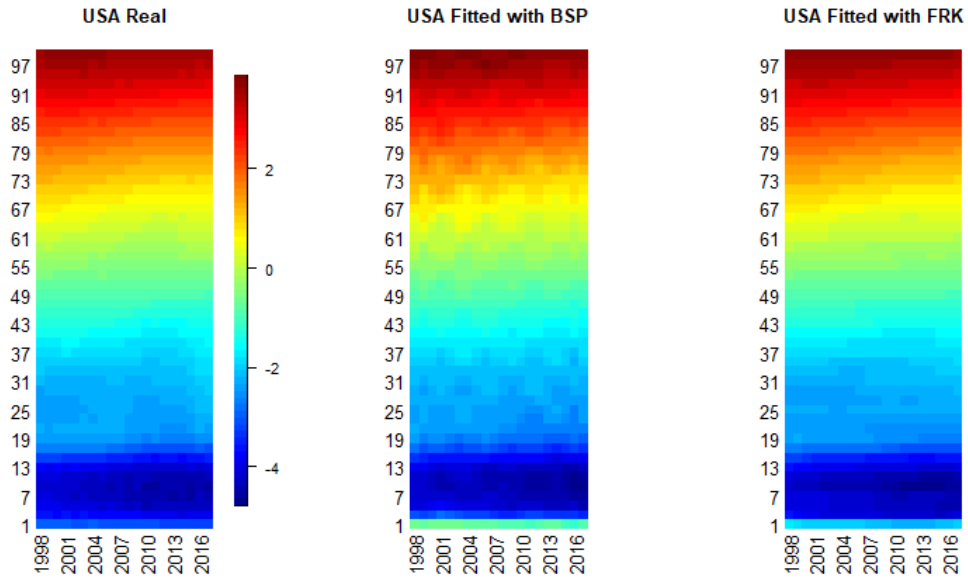


Figure B.1: USA real and estimated mortality rate of age 0-100

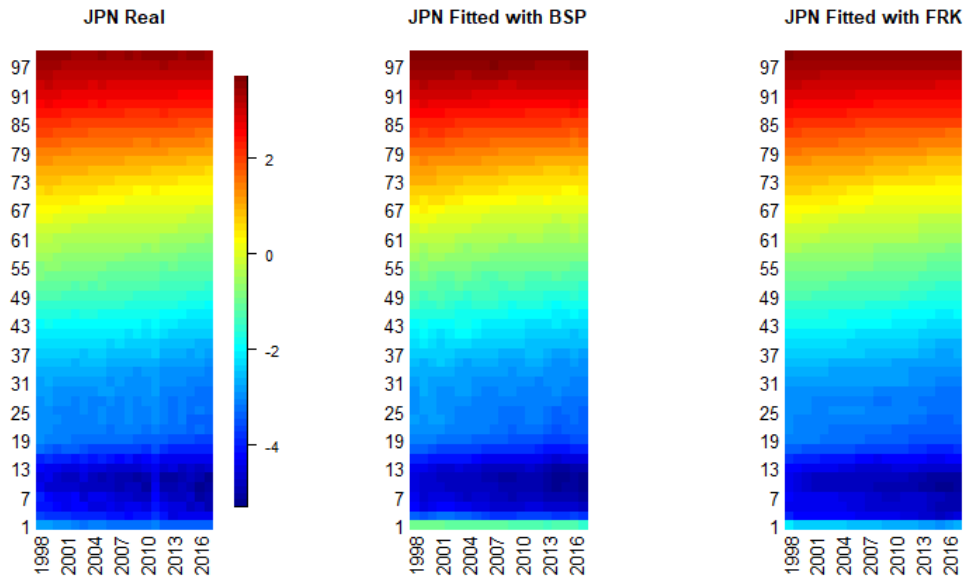


Figure B.2: Japan real and estimated mortality rate of age 0-100