Modeling Survival After Acute Myocardial Infarction Using Accelerated Failure Time Models and Space Varying Regression

by

Aijun Yang
B.Sc. University of Victoria 2004

A Thesis Submitted in Partial Fullfillment of the
Requirements for the Degree of

MASTER OF SCIENCE

in the Department of Mathematics and Statistics

Modeling Survival After Acute Myocardial Infarction Using Accelerated Failure Time Models and Space Varying Regression

by

Aijun Yang
B.Sc. University of Victoria 2004

Supervisory Committee

Dr. Farouk Nathoo (Department of Mathematics and Statistics)
Supervisor

Dr. Min Tsao (Department of Mathematics and Statistics)
Supervisor

Dr. Laura Cowen (Department of Mathematics and Statistics)
Departmental Member

Supervisory Committee

Dr. Farouk Nathoo (Department of Mathematics and Statistics)
Supervisor

Dr. Min Tsao (Department of Mathematics and Statistics)
Supervisor

Dr. Laura Cowen (Department of Mathematics and Statistics)
Departmental Member

Abstract

Acute Myocardial Infarction (AMI), commonly known as heart attack, is a leading cause of death for adult men and women in the world. Studying mortality after AMI is therefore an important problem in epidemiology. This thesis develops statistical methodology for examining geographic patterns in mortality following AMI. Specifically, we develop parametric Accelerated Failure Time (AFT) models for censored survival data, where space-varying regression is used to investigate spatial patterns of mortality after AMI. In addition to important covariates such as age and gender, the regression models proposed here also incorporate spatial random effects that describe the residual heterogeneity associated with different local health geographical units. We conduct model inference under a hierarchical Bayesian modeling framework using Markov Chain Monte Carlo algorithms for implementation. We compare an array of models and address the goodness-of-fit of the parametric AFT model through simulation studies and an application to a longitudinal AMI study in Quebec. The application of our AFT model to the Quebec AMI data yields interesting findings concerning aspects of AMI, including spatial variability. This example serves as a strong case for considering the parametric AFT model developed here as a useful tool for the analysis of spatially correlated survival data.

Table of Contents

List of Figures

Acknowledgments

I would like to thank Drs. Farouk Nathoo and Min Tsao. They have guided me in the fields of Bayesian statistics and spatial statistics, statistical computing, and provided many ideas, and offered infinite help and patience during my study at the University of Victoria. Without their support, I could not have got to the point of writing this thesis. Also, I want to thank my husband Menghong Gao and my son Richard and all my family members who always encourage and inspire me. I would like to acknowledge my gratitude to Dr. Charmaine Dean for her data and advice.
My research is supported by a Pacific Leaders Graduate Student Fellowship.

List of abbreviations

AMI.....Acute Myocardial Infarction (1)

PH.....Proportional Hazards (2)

AFT.....Accelerated Failure Time (3)

MCMC.....Markov Chain Monte Carlo (4)

CAR.....Conditional Auto-Regression (5)

MCAR.....Multivariate Conditional Auto-Regression (6)

DIC.....Deviance Information Criterion (7)

AIC.....Akaike Information Criterion (8)

i.i.d.....independent and identically distributed (9)

MLE.....Maximized Likelihood Estimator (10)

MSE.....Mean Square Error (11)

Chapter 1

Introduction

In epidemiology and public health, researchers are interested in studying the risk factors associated with mortality following Acute Myocardial Infarction (AMI). In addition, it is also of interest to characterize the impact of certain treatments, such as revascularization therapy, on mortality following AMI. Spatial variability in mortality is also of interest and may arise due to several factors including environmental conditions or accessibility to health services.

Many studies of AMI focus on analyzing the risk factors that affect the survival rate following AMI. Wilkinson et al. [1] investigated gender differences in those receiving treatment in an observational follow-up study. In [1], there were a total of 216 women and 607 men with AMI admitted to a coronary care unit from 1 January 1988 to 31 December 1992. The baseline variables examined were gender, ethnic group, smoker or non-smoker, and the presence of other chronic diseases. The patient survival time over the first six months was used as the primary end-point. The conclusion of their study was that women with AMI have a worse prognosis than men but this excess risk is confined to the first 30 days and is only partly explained by age and other baseline variables. Survival probability throughout follow-up was consistently lower for women than for men. Another study by Tonne et al. [2] explored the association between the socioeconomic position (SEP) of AMI patients and their AMI survival rates. The study looked at the long term survival for AMI patients using the Cox proportional hazards model. After adjusting for covariates, the study showed that each of the indicators of low SEP is significantly associated with lower

survival rates after AMI.

In addition to the survival analysis of AMI, there are a few studies examining spatial variation of AMI incidence. Kousa et al. [3] examined the association of Magnesium (Mg) content in local ground water to the spatial variation of AMI incidence among men and women 35-74 years of age in rural Finland in 1991-2003. The statistical analysis was carried out using Bayesian methods. This study concluded that high AMI incidence in eastern Finland is associated with soft ground water which is low in Mg. Recently, Loughnan et al. [4] studied seasonal and spatial differences in AMI incidence through examining 33,165 AMI admissions to hospitals over 2186 consecutive days in Melbourne, Australia. They found that there is a seasonal pattern in AMI admissions with increased rates during the colder months. The peak month is July. They provided maps of seasonal AMI admissions showing spatial differences and increased spatial clustering during the warmer months. Thus, spatial variation in AMI survival and incidence has been demonstrated and is an important factor in the disease etiology.

In this thesis, we combine survival analysis with spatial analysis to analyze geographically stratified survival rates following AMI in Quebec. The main objective is to extend the regression model to AMI data by including a spatial random effect that is central to the data analysis. To examine covariate and treatment effects, standard survival regression models that are well developed in the literature may be employed. There are many texts covering survival regression models, among them, David Collett [5] covers the Cox regression model, accelerated failure time models, and various analytical techniques applicable to medical survival data. To examine geographic variability, however, the ordinary regression models are insufficient in that they only include fixed effects, ignoring the spatial dependence structure in the data. Ignoring

spatial dependence may result in biased estimates of variations and inefficient statistical inference. This inspires us to develop spatial survival models which make it possible to (1) investigate whether there is spatial dependence or regional clustering in survival time that is due to environmental influences, and (2) examine regional variation in survival that may be the result of differing health policy and administrative structure across local health administrative districts. There is a considerable amount of literature on spatial modeling. In particular, Banerjee et al. [6] contains a discussion of spatial survival models, including parametric and semiparametric models, spatiotemporal models and multivariate models. However, its primary focus is on methods based on proportional hazards. Accelerated failure time models have not been considered within the spatial setting.

To accommodate the spatial structure of the Quebec AMI data, we develop a class of such AFT models with various spatial random effect modeling techniques within a Bayesian hierarchical framework. By incorporating the spatial effects into the AFT model, we not only obtain better models, but also uncover residual patterns in spatial variation which is of interest to policy makers and may give clues on disease etiology. The implementation of our hierarchical modeling makes full use of Markov Chain Monte Carlo (MCMC) methods to make Bayesian inference from the posterior distribution. Model selection for the Quebec AMI data will be based on the Deviance Information Criterion (DIC) which is described in Section 2.4.3. Being adjusted for other risk factors, significant regional variations in AMI mortality rate based on the selected model provides insight to the health system planning units when addressing the issue of hospital performance and accessibility of specialized cardiology care. Although motivated by the AMI data, the use of the methodology developed in this thesis goes beyond the current setting. For example, we can apply this method to study stroke or some type of cancer mortality rate. Further, we can apply our

spatial analysis techniques to BC geographic health data to study spatial clustering. For example, using the BC Linked Database we can measure the extent to which patients from defined geographic areas utilize a health care facility and evaluate the effectiveness of a health initiative on improving chronic diseases or home community care.

1.1 Motivation

The two main aspects that have motivated this thesis are the epidemiological importance of studying mortality after AMI and the statistical innovation in modeling spatial survival data. Acute myocardial infarction, commonly known as heart attack, is a disease state that occurs when the blood supply to some part of the heart is interrupted. The resulting oxygen shortage causes damage and potential death of heart tissue, which is a leading cause of death for adult men and women. For decades, studying mortality after AMI has been an important problem in epidemiology. How risk factors and treatments affect survival time after AMI is widely analyzed using Cox proportional hazards (PH) and accelerated failure time (AFT) models. An exploratory analysis on AMI data shows that the proportional hazards assumption does not hold. Hence, we take advantages of AFT models where the PH assumption is not required. Whereas a typical AFT model contains only fixed effects, this thesis models survival after AMI through a parametric AFT model with spatial random effects and space-varying regression. To the best of our knowledge, this extension of the AFT model has not been explored before. We will illustrate this point when applying the proposed models to the Quebec AMI data.

1.2 Quebec Cardiac Data

During a study period from January 01, 1996 to December 31, 1999, 61107 individuals aged 25 or older in the province of Quebec in Canada were enrolled into a study after an initial episode of AMI. After first hospitalization, these individuals were followed over time for recurrent episodes of AMI. During the study, the time to death after AMI was recorded. When the study period is over, if a subject is alive then the time from the date of entering the study to the termination date of the study is recorded. This type of observation is considered as Type 1 right-censored [7]. Thus a survival time in days for a subject is either exact or right-censored in this data set. In this cohort, there are 61054 individuals with survival days greater than zero who are included into the analysis. The data set contains 6732 complete observations, which account for 11% of the total observations. The remaining 54322 observations are right-censored. We are aware that such a high propotion of censored data could distort the analysis in some way. In addition to the survival time, a number of explanatory variables are available. These include the gender, age, geographical strata (local health units), and treatment information to indicate whether a subject received the treatment of revascularization through angioplasty or aortocoronary bypass. There are 21053 females and 40001 males who resided in one of 139 local health units, depicted in Figure 1.1, which subdivide the province. The local health units serve as the geographical strata in this study and a primary interest of our analysis lies in the identification of spatial heterogeneity in the survival pattern of subjects across the various local health units of the province.

1.3 Preliminary Analysis

We first present some descriptive summaries of the AMI data. An interesting observation of this study cohort is that the overall median age is 66 years but for

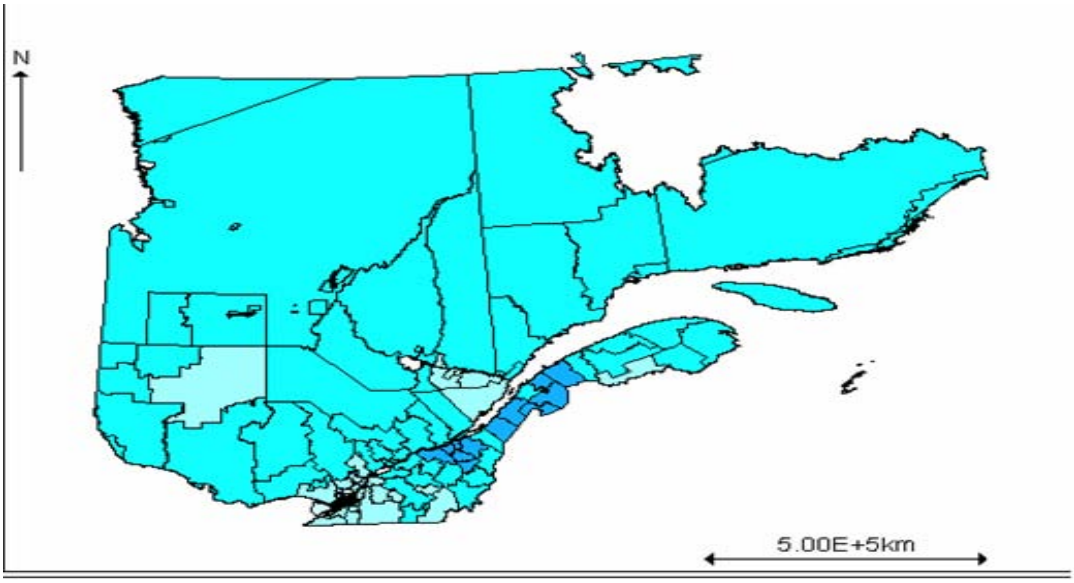


Figure 1.1: Map depicting geographical structure of Quebec divided into 139 local health units

females is 71 and for males is 62 (see Figure 1.2). A substantial age difference by gender is observed in this study. It is commonly believed that female hormones may protect women from heart disease before menopause. Therefore, women tend to develop heart disease at an older age. Also, we observe that the survival time distribution is different between female and male groups with overall median survival time 605 days, female group 584 days and male group 617 days. However, it is unusual that median survival time in the treatment group is 585 days lower than that in the non-treatment group (620 days). We suspect this might be due to a high proportion of censored observations in the treatment group. Table 1.1 is the breakdown of AMI cases by gender and Table 1.2 by treatment. Summary statistics of survival time are shown in Table 1.3. The histograms of the log survival time (T) are shown in Figure 1.3 and density curves are in Figure 1.4. It is seen that the distribution of $\log(T)$ is highly left-skewed.

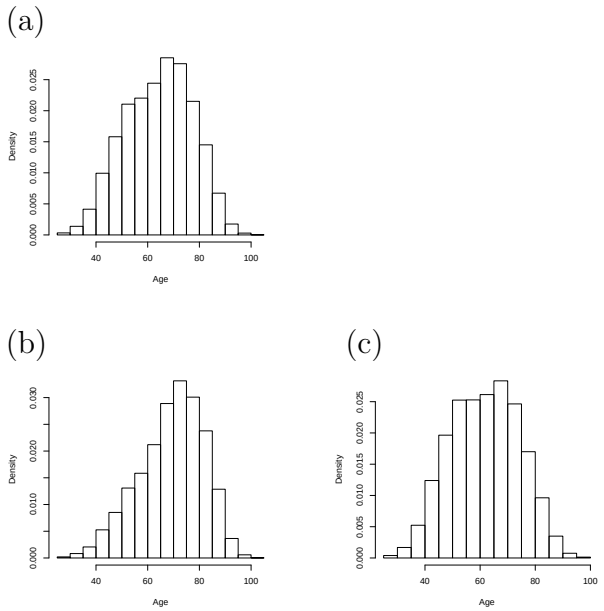


Figure 1.2: Age distribution (a) all cases (b) females, and (c) males.

Table 1.1: Frequency count of AMI cases by gender

gender	N	Observed	Censored	Observed (%)	Overall Observed (%)
female	21053	2902	18151	13.78	11.01
male	40001	3830	36171	9.57	11.01

Table 1.2: Frequency count of AMI cases by treatment

treatment	N	Observed	Censored	Observed (%)	Overall Obs. (%)
treatment=0	37813	5733	32080	15.16	11.01
treatment=1	23241	999	22242	4.30	11.01

Before developing our models for AMI data, we fitted the simple Cox proportional hazards model [8] and checked the model assumption that the effect of covariates is fixed over time. The covariates included in the model are age, gender, and treatment.

Table 1.3: Summary statistics of the survival time T

group	Min.	1st. Qu	Median	Mean	3rd Qu.	Max.
All	1	279	605	645.1	991	1457
Female	1	259	584	628.3	968	1457
Male	1	288	617	654	1002	1457
treatment=0	1	278	620	653.7	1010	1457
treatment=1	1	279	585	631.1	960	1457

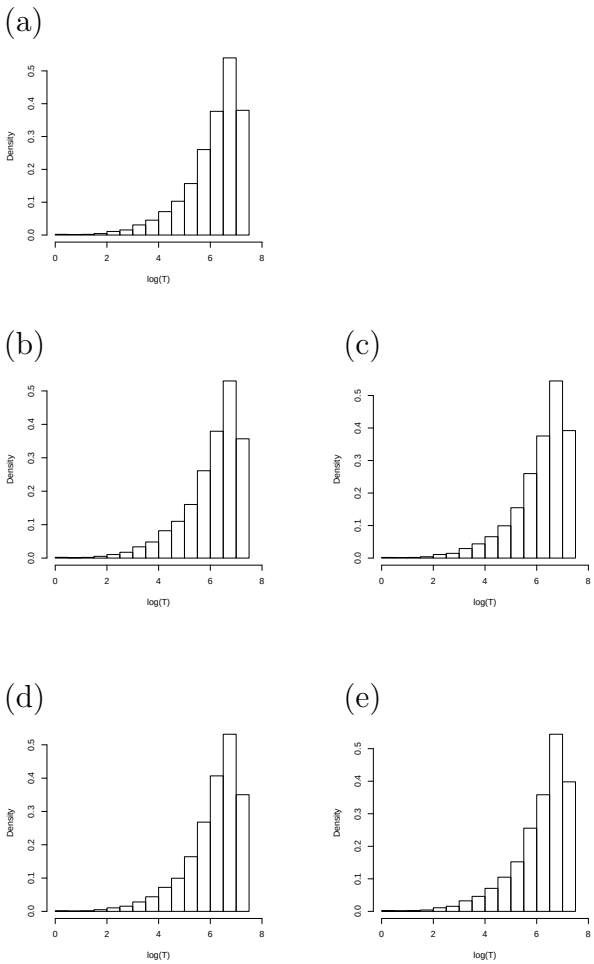


Figure 1.3: Log(T) distribution (a) all cases (b) females, (c) males, (d) treatment, and (e) non-treatment.

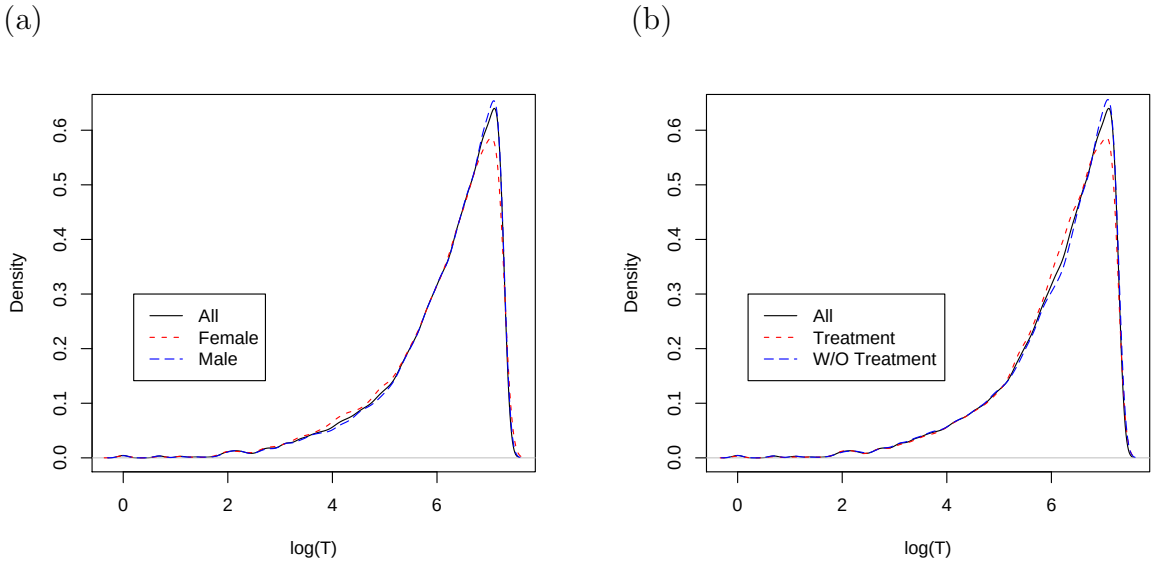


Figure 1.4: (a) Distribution of log(T) by gender (b) Distributions of log(T) by treatment

The p-values for age, treatment and overall covariates from Cox PH model assumption test are significant at $\alpha = 0.05$ level. So, the Cox PH model is not appropriate for the AMI data. This motivated us to seek an alternative model. Further, the AMI data has spatial information available that needs to be utilized. AMI researchers and policy makers are interested in uncovering the residual patterns in spatial variation that may lead to finding new covariates or provide insight to evaluating hospital performance and accessibility of the health care system.

1.4 Literature Review

Multi-center clinical trails and medical studies often take into consideration two types of variability in order to effectively evaluate treatment effects. One type is the known source variability due to covariates that are associated with the patients and center-specific characteristics. Another type is the unknown sources of heterogeneity between centers. The Cox proportional hazards (PH) model [8] is a widely adopted

model to quantify the effects of covariates on the time to event given that the covariates are independent of time. This assumption does not always hold. An alternate model is the accelerated failure time (AFT) model [9] which allows the effects of covariates to act multiplicatively on a time scale.

Researchers are often interested in the unknown sources of heterogeneity, which can happen due to many reasons: geographical differences, difference in availability of treatment facility, and differences in administrative structures. Moreover, the existence of such heterogeneity can be with respect to the baseline characteristics or to the efficacy of the treatment. Kalbfleisch and Prentice [9] adopt a stratified model to address the baseline heterogeneity or treatment effect heterogeneity. In recent years, to deal with such heterogeneity, researchers widely employ a random effects model. For example, Glidden and Vittinghoff [10] have surveyed various methods and suggested that the gamma frailty model is a good choice as it produces estimators with lower MSE than fixed effects or stratified approaches. In addition, this approach out-performs competing methods and appears robust to violation of the assumption of a gamma distribution for the frailty term.

Recently, several authors have extended the proportional hazards frailty model to the spatial setting by incorporating spatial correlation in the frailty distribution. Li and Ryan [11] developed a proportional hazards spatial frailty model in a classical framework. Specifically, they extended the ordinary frailty models by allowing random effects accommodating spatial correlations to enter into the baseline hazard function multiplicatively. Li and Ryan [11] base inference on a marginal rank likelihood approach. Monte Carlo simulations and the Laplace approach are used to tackle the intractable integral in the likelihood function arising from the multivariate spatial frailty distribution. In addition to the methods in the classical framework, Bayesian

methods have also started to emerge. These include Henderson et al. [12], Banerjee and Carlin [13] and Banerjee et al. [14]. In particular, Banerjee and Carlin [13] investigated Cox semiparametric survival modeling approaches, adding spatial and temporal effects through a hierarchical structure. The spatial correlation was handled by placing a particular multivariate generalization of the conditionally autoregressive (CAR) distribution on the region specific frailties. Banerjee et al. [14] developed a Bayesian hierarchical survival models for capturing spatial patterns within the framework of proportional hazards by introducing county-level cancer frailties. The baseline hazard function is modeled semiparametrically using mixture distributions. These Bayesian methods were then further extended by Carlin and Banerjee [15], Jin and Carlin [16], and Nathoo and Dean [17] who develop joint spatial models based on multivariate spatial mixing distributions. All previous work has been based upon the proportional hazards regression framework. The drawbacks of such random effects models include that the PH assumptions may not be satisfied and it could be difficult to interpret the marginal effect after integrating out random effects.

A random effects AFT model, however, can overcome these drawbacks. Models in this framework are appealing due to (a) their ease of interpretability and (b) simple assumption relative to that of proportional hazards. The basic models in this class assume observations are independent and adopt parametric distributional forms. More flexible AFT models adopt a semi-parametric approach and avoid distributional assumptions. In the Bayesian setting, semi-parametric AFT regression models for univariate survival times have been considered by several authors who modeled the underlying baseline survival function or error distribution using a wide range of prior distributions from a Dirichlet process, Polya tree prior, and a normal mixture. For example, Christenson and Johnson [18] modeled the relationship between survival time and covariates through the AFT model with a Dirichlet process

prior assumed for the underlying baseline survival function. Such a model is a good choice when estimating the survival distribution for future individuals with given covariates. They also gave the method for estimating regression coefficients using the marginal distribution for observed censored data. A similar approach was taken by Kuo and Mallick [19] who developed a model consisting of a parametric component for the regression coefficients and a nonparametric component for the unknown error distribution. Bayesian analysis is studied for the case of a parametric prior on the regression coefficients and a mixture-of-Dirichlet-processes prior on the unknown error distribution. More recently, Komarek and Lesaffre [20] developed a semi-parametric approach which models the error distribution as a normal mixture with an unknown number of components, and further allowed for multivariate event-times through the inclusion of a random effect, which was assumed to have an exchangeable correlation structure.

Thanks to recent advances in computing technology, Bayesian approaches to survival models are now computationally feasible and increasingly popular. In this thesis, we consider several parametric AFT models with spatial random effects. We implement our proposed spatial AFT methodology using Monte Carlo methods. The rest of the thesis is organized as follows. In Chapter 2, we review survival data analysis, Bayesian hierarchical modeling, and spatial analysis. These are the key ingredients of our spatial AFT model. Then we review Markov Chain Monte Carlo (MCMC) in Chapter 3, which is required for inference in complex spatial models. In Chapter 4, we introduce the specific model specification, computation and algorithms. We also discuss a simulation study for model assessment and examining the performance of MCMC methods. In Chapter 5, we examine the Quebec data. We start with an exploratory analysis using simple methods from classical survival data analysis. Then we apply and compare Bayesian hierarchical models based on the proposed

spatial modeling framework. We adopt the deviance information criterion (DIC) for model comparison and selection and we explore methods for assessing goodness-of-fit. Summary and future work are given in Chapter 6.

Chapter 2

Background on Statistical Modeling

In this chapter, we provide background material on statistical modeling including survival analysis, frailty models, Bayesian inference and hierarchical modeling, and models used for spatially correlated data. The background material on the related subjects is meant to serve the purpose of better understanding the proposed spatial survival models and their implementations. For a full treatment on survival analysis, spatial statistics, and Bayesian statistics, readers are referred to Klein and Moeschberger [21], Banerjee et al. [22], and Gelman et al. [23], respectively.

2.1 Survival Data Analysis

In this section we consider the basic parameters used in modeling survival data and three major functions characterizing the distribution of survival times. We also provide a short discussion on Kaplan-Meier estimator of the survival function. This section ends with an illustration on the likelihood construction for right censored data.

2.1.1 Basic Concepts in Survival Data Analysis

Survival data appear in various settings. In general, let T be the time until some event of interest occurs. This event can be death, the recurrence of a disease, the occurrence of being unemployed, or committing a second crime, and so forth. Clearly, T is a nonnegative random variable. Three well known functions characterizing the distribution of T are (1) the survival function, (2) the hazard rate (function) and (3) the probability density (or probability mass) function.

The survival function, $S(t)$, is the probability of an individual surviving to time T, which is given by

$$S(t) = Pr(T > t) = 1 - F(t),$$

where $F(t) = Pr(T \leq t)$ is the c.d.f. of T and $f(t) = -dS(t)/dt$ is the density of T. The hazard rate is roughly the probability per time unit that an individual will fail in an interval given that the individual has survived to the beginning of the respective interval. It is denoted by $h(t)$ and defined as

$$\begin{aligned} h(t) &= \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta\, t | T \geq t)}{\Delta\, t} \\ &= f(t)/S(t) = -d \ln[S(t)]/dt. \end{aligned}$$

The cumulative hazard function is defined as $H(t) = \int_0^t h(u)du = -\ln[S(t)]$, which implies that,

$$S(t) = \exp[-H(t)] = \exp[-\int_0^t h(u)du].$$

2.1.2 Kaplan-Meier Estimator of Survival Function

In survival analysis, we often prefer distribution free methods for data analysis. A non-parametric method, the Kaplan-Meier (KM) method, is commonly used to estimate the survival function for a cohort of subjects. The KM estimator of the survival function is also called the Product-Limit estimator. For data without censoring, it is defined for all values of t in the range of the data as:

$$\hat{S}(t) = \begin{cases} 1 & \text{if } t < t_1 \\ \prod_{t_i \leq t}[1 - \frac{d_i}{Y_i}] & \text{if } t_1 \leq t \end{cases}$$

where d_i is the number of events observed at each event time t_i (t_1 is the smallest survival time), and Y_i is the number of individuals or units at risk. The Product-Limit estimator is a step function with jumps at the observed event times. One of

the graphic goodness-of-fit checking tools is to compare the KM survival curve of the raw data to a fitted parametric survival curve.

Although comparing the KM estimates could detect the difference of survival probability among cohorts, it does not give you the magnitude of such difference. Through regression modeling, which is described in the next section, we can get an estimate of the magnitude of such differences for covariates. Further, we can model the impact of multiple covariates on the survival time.

2.1.3 Likelihood Construction for Right Censored Data

Survival data often involve censoring. There are various categories of censoring, such as right censoring, left censoring, and interval censoring. Each type of censoring will lead to a different likelihood function. As our AMI data is type I right censored data, we will only introduce the likelihood construction for the right censored data. For the likelihood construction on the other type of censoring, readers are referred to Klein and Moeschberger [24].

Type I censoring happens when the event is observed only if occurs prior to some prespecified time (e.g. end of study). These censoring times may vary from individual to individual depending on their start times. In right censoring, for a specific individual under study, we assume that there is a lifetime X and a fixed censoring time C_r. The $X's$ are assumed to be independent and identically distributed with probability density $f(x)$ and survival function $S(x)$. Such data set involving right censoring are represented using pairs of (T, δ), where δ indicates whether the lifetime X for a individual corresponds to an event ($\delta = 1$) or is censored ($\delta = 0$). T will be

determined as follow,

$$T = \begin{cases} X & \text{if } \delta = 1 \\ \min(X, C_r) & \text{if } \delta = 0 \end{cases}$$

When constructing likelihood functions, a critical assumption is that the lifetimes and censoring times are independent [24]. An observation with an exact event time provides information on the probability that is approximately equal to the density of X at this time. For a right-censored observation, however, we only know that the event time is larger than this time and only partial information (survival function evaluated at this time) is available to construct the likelihood function. Any information beyond this time is not known. This is called non-informative right censoring. Details of constructing the likelihood function for *type I censoring* are as follows. For $\delta = 0$, it can be seen that [24]

$$\begin{aligned} Pr[T, \delta = 0] &= Pr[T = C_r | \delta = 0] Pr[\delta = 0] \\ &= Pr[\delta = 0] \\ &= Pr(X > C_r) = S(C_r). \end{aligned} \tag{2.1.1}$$

Also, for $\delta = 1$,

$$\begin{aligned} Pr(T, \delta = 1) &= Pr(T = X | \delta = 1) Pr(\delta = 1) \\ &= Pr(X = T | \leq C_r) Pr(X \leq C_r) \\ &= \frac{f(t)}{1 - S(C_r)} (1 - S(C_r)) \\ &= f(t). \end{aligned} \tag{2.1.2}$$

Expressions (2.1.1) and (2.1.2) can be combined into the single expression

$$Pr(t, \delta) = [f(t)^{\delta}][S(t)]^{1-\delta}.$$

If we have a random sample of paris $(T_i, \delta_i), i = 1, ..., n$, the likelihood function is [24]

$$L = \prod_{i=1}^{n} [f(t_i)]^{\delta_i} [S(t_i)]^{1-\delta_i}. \tag{2.1.3}$$

The likelihoods constructed in this section are used primarily for analyzing parametric methods. They also serve as a basis for determining the partial likelihoods used in the semiparametric regression models. For the details on constructing a partial likelihood function, interested readers can refer to the material covered in [25].

2.2 Regression Models for Survival Data

Survival regression models model the survival probability/time as a function of covariates. In the following, we first discuss Cox proportional hazard regression models and then accelerated failure time models (AFT). Important advantages of the AFT models are their complete model specifications and better interpretation of the regression coefficients through a simple log-linear structure.

2.2.1 Cox Proportional Hazards Model

Cox [8] proposed a semiparametric model for the hazard function that allows the addition of explanatory variables, but keeps the baseline hazard as an arbitrary, unspecified, nonnegative function of time. The Cox model specifies the hazard function as

$$h(t;\mathbf{x}) = h_0(t)\exp(\boldsymbol{\beta}'\mathbf{x}), \tag{2.2.1}$$

where $\mathbf{x}$ is a vector of covariates; $\boldsymbol{\beta}$ is the corresponding regression parameter and the baseline hazard $h_0(t)$ in (2.2.1) corresponds to the hazard function when all covariates equal zero. Because the baseline hazard is not assumed to be of a parametric form, Cox's model [7] is referred to as a semiparametric model. The survival function corresponding to model (2.2.1) is

$$S(t;\mathbf{x}) = \exp[-\exp(\boldsymbol{\beta}'\mathbf{x})\int_0^t h_0(u)du]. \tag{2.2.2}$$

One of the restrictions underlying the Cox model with time-fixed covariates is its proportional hazards (PH) assumption. It follows from (2.2.1) that the hazards ratio

between two sets of covariates is constant over time, because the common baseline hazard function cancels out in the ratio of the two hazards. For fixed-time covariates, the exponent of a coefficient describes the relative risk due to the covariate [26]

$$\frac{h(t;\mathbf{x})}{h_0(t)} = \exp(\boldsymbol{\beta}'\mathbf{x}). \tag{2.2.3}$$

There are various methods for assessing the lack-of-fit of a PH model. The simplest and most intuitive method is to plot log[-log(the estimated survival function)] against time for the different cohorts. From the equation (2.2.2), we have

$$S(t;\mathbf{x}) = [S_0(t)]^{\exp(\boldsymbol{\beta}'\mathbf{x})}$$

which implies that

$$\log[-\log S(t;\mathbf{x})] = \boldsymbol{\beta}'\mathbf{x} + \log[-\log S_0(t)].$$

Thus, the plots for different groups defining values of $\mathbf{x}$ should be parallel if the PH assumption holds.

2.2.2 Accelerated Failure Time Models

The AFT model is an alternative to the PH model for the analysis of time-to-event data and it is introduced by Kalbfleisch and Prentice [18], in which the PH assumption is not required. The hazard function in this case can be written as

$$h(t;\mathbf{x}) = \exp(\boldsymbol{\beta}'\mathbf{x})h_0(\exp(\boldsymbol{\beta}'\mathbf{x})t). \tag{2.2.4}$$

The baseline hazard $h_0(t)$ in (2.2.4) corresponds to the hazard function when all covariates equal zero, which can be modeled parametrically or semiparametrically. The survival function corresponding to the hazard function in (2.2.4) is the following:

$$S(t;\mathbf{x}) = S_0[\exp(\boldsymbol{\beta}'\mathbf{x})t]. \tag{2.2.5}$$

The factor $\exp(\boldsymbol{\beta}'\mathbf{x})$ is called an acceleration factor indicating how a change in covariate values changes the time scale from the baseline time scale [27].

The second representation of the relationship between covariate values and survival time in the AFT model is the linear relationship between log time and the covariate values. The usual linear model for log time is,

$$Y = \log(T) = \mu + \mathbf{b}'\mathbf{x} + \sigma W, \tag{2.2.6}$$

where μ is the intercept term, $\mathbf{x}$ is a vector of covariates, σ is a scale parameter, $\mathbf{b}$ is a vector of regression coefficients and the distribution of W is the random error distribution. The interpretation of these regression coefficients is similar to the linear regression [27]. The random error distribution can be assumed to follow one of various distributions including the commonly used extreme-value, normal and logistic distributions. This is equivalent to saying that the corresponding survival time T follows the Weibull, log-normal, and log-logistic distribution, respectively. If we let $S_0(t)$ be the baseline survival function of the random variable $\exp(\mu + \sigma W)$, then the linear log-time model is equivalent to the AFT model (2.2.4) with $\boldsymbol{\beta} = -\mathbf{b}$.

Under the AFT formulation, the effect of treatments and covariates is assumed to act additively on the log time scale and therefore multiplicatively on the time scale itself. Three commonly adopted parametric AFT models are the Weibull, log-normal, and log-logistic in terms of the distribution of survival time. Table 2.1 summarizes the different parametric formulation and the relationship between the error term and its survival function. A plot based on this relationship can be useful in model checking.

Up to this point, we have only considered fixed effects in regression models. The covariates only explain part of the heterogeneity among the subjects. There are

Table 2.1: Different parametric AFT model formulations

$Y = \log T = \mu + \sigma W$ (without covariate)					
Model	$S(t)$	$S(y)$	$S(w)$	$w = f[S(w)]$	λ α vs. μ σ
Weibull	$\exp(-\lambda t^{\alpha})$	$\exp(-\lambda e^{\alpha y})$	$\exp(-e^{w})$	$\log(-\log(S(w))$	$\lambda = \exp(-\frac{\mu}{\sigma})$ $\alpha = \frac{1}{\sigma}$
lognormal	$1 - \Phi(\frac{\log(t)-\mu}{\sigma})$	$1 - \Phi(\frac{y-\mu}{\sigma})$	$1 - \Phi(w)$	$\Phi^{-1}\{1 - S(w)\}$	
loglogistic	$\frac{1}{(1+\lambda t^{\alpha})}$	$\frac{1}{(1+\lambda e^{\alpha y})}$	$\frac{1}{(1+e^{w})}$	$\text{logit}(S(w))$	$\lambda = \exp(-\frac{\mu}{\sigma})$ $\alpha = \frac{1}{\sigma}$

situations where we are interested in investigating the unaccounted heterogeneity, due to missing covariates. This leads to the next section on frailty models, which extend regression models for survival data through the incorporation of random effects.

2.3 Frailty Models

In survival analysis, the notion of frailty provides a convenient way to introduce random effects, association, and unaccounted heterogeneity into survival data. Vaupel et al. [28] coined the phrase *frailty* in survival models. The frailty models are particularly useful when modeling correlated or clustered survival data, as they can be used to model dependence or clustering among survival times in the cluster. Both univariate frailty models and multivariate frailty models can be incorporated into proportional hazards and accelerated failure time survival models, though the former are far more common than the latter.

2.3.1 Univariate Frailty Models

Univariate frailty models take into account residual heterogeneity that cannot be explained through available covariates, and thus accommodate unobserved heterogeneity leading to more general classes of parametric models. The proportional

hazards frailty model and accelerated failure time frailty model are mostly used to account for such heterogeneity. Let ϕ_i denote the random effect, called frailty, for an individual i with covariates $\mathbf{x_i}$. The proportional hazards frailty model assumes

$$h_i(t_i|\mathbf{x_i}, \phi_i) = \exp(\boldsymbol{\beta}'\mathbf{x_i} + \phi_i)h_0(t_i), \tag{2.3.1}$$

whereas the accelerated failure time frailty model assumes

$$h_i(t_i|\mathbf{x_i}, \phi_i) = \exp(\boldsymbol{\beta}'\mathbf{x_i} + \phi_i)h_0(\exp(\boldsymbol{\beta}'\mathbf{x_i} + \phi_i)t_i). \tag{2.3.2}$$

In each model, the frailty ϕ_i is an unobservable random variable varying over the sample, $i = 1, ..., N$. The factor e^{ϕ_i} has a useful interpretation; it increases the individual risk if $\phi_i > 0$ or decreases the individual risk, respectively, if $\phi_i < 0$ by a factor of e^{ϕ_i} for PH; accelerating the survival time by a factor of e^{ϕ_i} for AFT.

Modeling frailty in PH models is relatively simple. Let $Z = e^{\phi}$. It is common to assume that Z has a gamma distribution, an assumption that is made primarily for computational convenience. Because it is easy to derive the closed form expressions of the survival, density, and hazard functions, the usual maximum likelihood estimation method can be implemented. However, modeling frailty in AFT model is not as straightforward through the classic likelihood approach. Hence, Bayesian hierarchical modeling is used, especially for the case when the shared frailty models are adopted. This is further explained in the next section.

2.3.2 Shared Frailty Models

Clustered survival data arise commonly in multi-center clinical trails and medical studies. The clusters can correspond to hospitals or to different sites or geographic regions. Subjects are divided into groups like family, study center or geographic regions. As the individuals in the same group share certain common characteristics

such as genetics in a family setting, it is necessary to accommodate the dependence within groups. To model dependence within groups, it is reasonable to assume that individuals in a cluster are assumed to share the same frailty. That is why it is called shared frailty model, which is introduced by Clayton [29] and extensively studied in Hougaard [30]. The survival times are assumed to be conditionally independent with respect to the shared (common) frailty. A simple version is the constant shared frailty model, in which individuals in a group j share the same frailty ϕ_j. In the regression model, the conditional hazard for the i^{th} individual in the j^{th} group is:

$$h_{ij}(t_{ij}|\mathbf{x_{ij}}, \phi_j) = \exp(\boldsymbol{\beta}'\mathbf{x_{ij}} + \phi_j)h_0(t_{ij}), \tag{2.3.3}$$

where ϕ_j's are independent identically distributed following a chosen distribution, as in the univariate frailty models. The model assumes that the survival times are independent given the values of the frailties. The value of $\boldsymbol{\phi}$ is constant over time and common to all individuals in the same group and this induces a within group dependence. The constant shared frailty, while fairly simple, has the following major limitations: (1) it allows one-dimensional frailty only which implies a positive correlation within group, (2) it constrains the unobserved factors to be the same within a group, and (3) it yields possible confounding of the dependence parameter and the population heterogeneity. There exists a need for more flexibility in modeling correlation in this setting. Multivariate frailty models allow a more general correlation structure. In this regard, Yashin et al. [31] proposed a correlated gamma frailty model to study the influence of genetic and environmental factors on life-related durations. They employed a bivariate survival model based on the concept of correlated individual frailty that can be used for the genetic analysis of durations and applied to Danish twin survival data. Li and Zhong [32] proposed a multivariate survival model for age of onset data of a sibship from an additive genetic gamma frailty model constructed basing on the inheritance vectors. Their approach incorporates both affected

and unaffected sibs, environmental covariates, and age of onset or age at censoring information; therefore, they provide a practical solution for mapping genes for complex diseases with variable age of onset. In such settings, multivariate frailty models are more appropriate than the constant shared frailty models.

If we extend the shared frailty to the spatial setting, then conditional autoregression (CAR) modeling can be applied to spatial survival data. We suspect that frailties corresponding to strata in closer proximity to each other might also be similar in magnitude due to underlying factors that vary spatially. This will be discussed in more details in a later section.

2.4 Bayesian Hierarchical Modeling

Bayesian data analysis relies on practical methods for making inferences from data using probability models for quantities we observe and for quantities about which we wish to learn. The core of Bayesian methods is their explicit use of probability for quantifying uncertainty in statistical inference [33]. Bayesian hierarchical modeling is an essential part of the Bayesian paradigm and allows for specification of complex models with rich features for complex analysis.

2.4.1 Bayesian Inference

Bayesian inference about a parameter θ (or set of parameters $\boldsymbol{\theta}$), or unobserved data $\tilde{y}$, are made in terms of probability statements. These probability statements based on distributions, $p(\theta|y)$ or $p(\tilde{y}|y)$, are conditional on the observed data y. Conditioning on observed data is the essential way in which Bayesian inference departs from frequentist approaches where inference is based on the repeated sampling paradigm.

A Bayesian data analysis is accomplished through the following three steps:

1. Setting up a full probability model – a joint probability distribution for all observable and unobservable quantities in a problem;

2. Conditioning on observed data, calculating and interpreting the appropriate posterior distribution – the conditional probability distribution of the unobserved quantities of ultimate interest, given the observed data;

3. Evaluating the fit of the model and the implications of the resulting posterior distribution. Here, we check the model fit and determine if the substantive conclusions are reasonable, and also examine how sensitive the results are to the model assumptions in step 1. If necessary, we can alter or expand the probability model and thus return to step 1.

Governing the above process is Bayes's rule. As stated in step 1, in order to make probability statements about θ given y, we must begin with a model providing a joint probability distribution for θ and y. The joint probability mass or density function $p(\theta, y)$ can be written as a product of two densities, the prior distribution $p(\theta)$ and the sampling distribution (or data distribution) $p(y|\theta)$, i.e.,

$$p(\theta, y) = p(\theta)p(y|\theta).$$

Conditioning on the known value of the data y, Bayes' rule yields the posterior density:

$$p(\theta|y) = \frac{p(\theta, y)}{p(y)} = \frac{p(\theta)p(y|\theta)}{p(y)} = \frac{p(\theta)p(y|\theta)}{\int p(\theta)p(y|\theta)d\theta}. \tag{2.4.1}$$

These expressions encapsulate the technical core of Bayesian inference: the primary task of any specific application of Bayesian inference is to develop the model $p(\theta, y)$ and perform the necessary computation to summarize $p(\theta|y)$ in appropriate ways. Among them, hierarchical Bayesian modeling coupled with Markov Chain Monte Carlo sampling techniques are widely used.

2.4.2 Hierarchical Modeling

A hierarchical Bayes model is a Bayesian statistical model based on distributions $p(\boldsymbol{y}|\boldsymbol{\theta}), p(\boldsymbol{\theta})$, where the prior $p(\boldsymbol{\theta})$ is decomposed into a sequence of conditional distributions $p_1(\boldsymbol{\theta}|\boldsymbol{\theta_1})p(\boldsymbol{\theta_1}|\boldsymbol{\theta_2})...p_N(\boldsymbol{\theta_{N-1}}|\boldsymbol{\theta_N})$ and $p_{N+1}(\boldsymbol{\theta_N})$. The $\boldsymbol{\theta_i}$'s are called the hyper-parameters of level i, $i = 1...N$ respectively. The hyper-parameters can be estimated from data through the corresponding posterior. We illustrate hierarchical modeling through the following example, describing a model for clustered survival data.

Likelihood at the first level

Suppose we have a survival regression model incorporating a vector of cluster specific random effects $\boldsymbol{\phi}$. Let $\boldsymbol{\theta_1} = \{\boldsymbol{\beta}, \boldsymbol{\phi}\}$ denote the parameters corresponding to the first level of the model, where $\boldsymbol{\beta}$ is the vector of regression coefficients. Assuming uninformative right censoring indicated by δ_{ij} and conditional (on $\boldsymbol{\theta}$) independence of event times t_{ij}, the likelihood can be written as

$$L(\boldsymbol{\theta_1}) = \prod_{j=1}^{J} \prod_{i=1}^{M_j} L_{ij} \tag{2.4.2}$$

where J denotes the number of clusters, M_j is the number of subjects in the j^{th} cluster and

$$\begin{aligned} L_{ij} &= f_{ij}(t_{ij})^{\delta_{ij}} S_{ij}(t_{ij})^{1-\delta_{ij}} \\ &= [h_{ij}(t_{ij})]^{\delta_{ij}} \exp\left\{ - \int_0^{t_{ij}} h_{ij}(u)du \right\}, \end{aligned} \tag{2.4.3}$$

where $h_{ij}(\cdot)$ is the hazard function defining the distribution of t_{ij}. If we take the AFT model for $h_{ij}(\cdot)$, then we can write down the likelihood of (2.4.3) as a function of $\boldsymbol{\theta_1}$ as

$$L_{ij} = \{e^{\boldsymbol{\beta}' \mathbf{x_{ij}} + \phi_j} h_0(e^{\boldsymbol{\beta}' \mathbf{x_{ij}} + \phi_j} t_{ij})\}^{\delta_{ij}} \exp\left\{ - \int_0^{t_{ij}} e^{\boldsymbol{\beta}' \mathbf{x_{ij}} + \phi_j} h_0(e^{\boldsymbol{\beta}' \mathbf{x_{ij}} + \phi_j} u)du \right\},$$

where $\mathbf{x_{ij}}$ are subject specific covariates and ϕ_j is a cluster effect.

Priors at the second level

At the second level of the model, we assign priors to the components of $\boldsymbol{\theta_1}$, for example, we may assign a flat (uniform) prior for $\boldsymbol{\beta}$, and adopt a multivariate normal prior for the random effects $\boldsymbol{\phi}|\Sigma \sim MVN_J(\mathbf{0}, \Sigma)$, where Σ is a $J \times J$, a positive definite symmetric matrix characterizing residual variation between clusters.

Hyperpriors at the third level

Here the parameter $\boldsymbol{\theta_1}$ has a prior density that is written $f(\boldsymbol{\beta}, \boldsymbol{\phi}) = f(\boldsymbol{\beta})f(\boldsymbol{\phi}|\Sigma)$ ($f(\cdot)$ is generic and denotes the density). At the third and final stage of the model we require a prior for the hyper-parameter Σ under the hierarchical framework. A common prior in this setting is the inverse-Wishart

$$\Sigma^{-1} \sim \text{Wishart}(V, S),$$

where $V(> J)$ is a positive scalar degrees of freedom and S is a symmetric, positive semidefinite $J \times J$ matrix. The values of V and S are chosen to reflect prior information on Σ^{-1}.

Posterior distribution

Here, the model unknowns are $\boldsymbol{\theta} = \{\boldsymbol{\theta_1}, \Sigma\}$ and the prior is of the form

$$f(\boldsymbol{\theta}) = f(\boldsymbol{\beta})f(\boldsymbol{\phi}|\Sigma)f(\Sigma).$$

Bayesian inference is based on the posterior distribution $f(\boldsymbol{\theta}|\boldsymbol{y})$, which is given up to normalizing constants as

$$f(\boldsymbol{\theta}|\boldsymbol{y}) \propto L(\boldsymbol{\theta_1})f(\boldsymbol{\beta})f(\boldsymbol{\phi}|\Sigma)f(\Sigma).$$

In most cases, $\boldsymbol{\theta}$ is of high dimension, and the corresponding normalizing constants is thus analytically intractable. In this case, inference is implemented through Monte Carlo procedures that simulate realizations from $f(\boldsymbol{\theta}|\boldsymbol{y})$.

2.4.3 Model Selection

Model selection is the task of selecting a statistical model from a set of potential models for some given data. Within the classical modeling framework, model selection generally relies on measuring two quantities: measure of *fit*, typically a deviance statistic, and *complexity*, the number of free parameters in the model. Since increasing complexity is accomplished by a better fit, models are compared by trading off these two quantities. Burnham and Anderson [34] emphasize the importance of selecting models based on sound scientific principles. It is often the case that more complex models will be better able to adapt their shape to fit the data, but the additional parameters may not represent anything useful. Therefore, a good model selection technique will balance goodness of fit and complexity.

In the Bayesian paradigm, many techniques have been suggested for choosing among competing models. Bayes factor, Bayesian Information Criterion (BIC), Akaike's Information Criterion (AIC), and Deviance Information Criterion (DIC) are among the most often used. Bayes factor arises as a ratio of the marginal likelihoods of the observed data under each model and is defined as

$$BF = \frac{p_1(y)}{p_2(y)} = \frac{\int p_1(y|\theta_1)p_1(\theta_1)d\theta_1}{\int p_2(y|\theta_2)p_2(\theta_2)d\theta_2},$$

where $p(y|\theta_i)$ is the likelihood function under model i and $p_i(\theta_i)$ is the prior specification assigned to model i, $i = 1, 2$. If $BF > 1$, the data favour model 1. However, Bayes factor requires the calculation of the marginal likelihood of the given model, which can be difficult in the complex models we consider in this thesis. The BIC, on

the other hand, is easy to compute and can be used to approximate Bayes factor. For a given model it is defined as

$$BIC = -2\log\{p(y|\tilde{\theta})\} + \log(n)p, \tag{2.4.4}$$

where $\log\{p(y|\tilde{\theta})\}$ is the log-likelihood for the specified model and $\tilde{\theta}$ is MLE of θ, n is the number of observations, and p is the number of parameters. Therefore, it yields criteria for comparison of models that is based on minimizing the right-hand side of equation (2.4.4). Models with lower BIC are preferred. In addition, the difference in the BIC scores between two models asymptotically approaches minus twice the log of the Bayes factor comparing these models, assuming both models are priori equally likely [35]. So the Bayes factor comparing two models can be approximated by calculating the corresponding BIC's. One obvious drawback with Bayes factor is that the marginal likelihood is not defined when the prior is improper. Also, Weakliem [36] pointed out that variations in priors tend to make the BIC inclined to favour excessively simple models in practice. Similar to the BIC, another fit plus penalty model selection tool is AIC, which is given by

$$AIC = -2\log\{p(y|\tilde{\theta})\} + 2p, \tag{2.4.5}$$

which replaces $\log(n)$ in equation (2.4.4) by 2 and is therefore more liberal whenever $n > e^2$. Both BIC and AIC are not appropriate for comparing models with random effects where the parameters are correlated and hence the effective number of parameters will generally be less than p, the raw number of parameters which forms the penalty term used to measure complexity in both BIC and AIC. Indeed, in complex hierarchical models parameters may outnumber observations.

To address this concern, Spiegelhalter et al. [37] proposed a generalization of AIC known as the Deviance Information Criterion (DIC) which is a more appropriate

model selection tool for Bayesian hierarchical models. Spiegelhalter et al. [37] define DIC based on this principle: DIC = 'goodness of fit'+'complexity'. They measure fit via the deviance statistic which is defined as

$$D(\theta) = -\log p(y|\theta),$$

where $p(y|\theta)$ is the likelihood function of the data given parameters under the model. Complexity is measured by an estimate of the *effective number of parameters* and is denoted as

$$\begin{aligned} pD &= E_{\theta|y}[D(\theta)] - D(E_{\theta|y}[\theta]) \\ &= \overline{D(\theta)} - D(\bar{\theta}), \end{aligned}$$

where $\overline{D(\theta)} = E_{\theta|y}[D(\theta)]$ is the posterior mean of the deviance that measures the fit of the model and $D(\bar{\theta}) = D(E_{\theta|y}[\theta])$ is the deviance evaluated at the posterior mean of θ. Hence, pD, the difference between the posterior mean deviance minus deviance evaluated at the posterior mean of the parameters, represents the effect of model fitting and has been used as measure of the *effective number of parameters* of a Bayesian model. For a normal linear fixed model with large sample size, pD is equal to the number of parameters in the model. Whereas, in the correlated random effects model, pD is generally smaller than the number of parameters. The DIC is then defined analogously to AIC as

$$\begin{aligned} DIC &= D(\bar{\theta}) + 2pD \\ &= \overline{D(\theta)} + pD \\ &= \overline{D(\theta)} + \overline{D(\theta)} - D(\bar{\theta}) \\ &= 2\overline{D(\theta)} - D(\bar{\theta}). \end{aligned} \tag{2.4.6}$$

The model with lower DIC is preferred since it balances fit and complexity. The smaller the DIC, the better the fit, and a difference larger than 10 is overwhelming

evidence in favour of the better model [34] [38].

The DIC is a very popular model selection tool primarily because it is easy to calculate and is automatically computed by WinBugs software. Despite its popularity, it seems to lack a thorough theoretical justification and, in addition, is not invariant to model reparametrization. Spiegelhalter et al. [37], however, gave an approximate decision theoretic justification for DIC by mimicking the development of Ripley [39] and Burnham and Anderson [40].

2.4.4 Goodness-of-Fit

Model selection can only serve the purpose of choosing the *best* model among a class of models using defined criteria. Once a model has been chosen, it is important to check the fit of the model to important aspects of the data. When the assumption on the structure of a probability model is not valid or the model fits poorly, statistical inference can be misleading. Gelman et al. [41] discuss Bayesian model checking and improvement.

Scope of the model checking

The term "model" in the Bayesian framework usually encompasses the sampling distribution, the prior distribution, and any hierarchical structure. In practice, it is often the case that more than one reasonable probability model can provide an adequate fit to the data. In addition, sensitivity analysis is also of interest, where we would like to examine to what extent the posterior inferences change when different priors and sampling distributions are used.

In theory, both model checking and sensitivity analysis can be incorporated into the usual prior-to-posterior analysis through a Bayesian model averaging framework

where the model itself is treated as an unknown and each plausible model is assigned a prior probability. Under this perspective, model checking is done by comprehensively incorporating all known substantive information to prior beliefs and all plausible likelihoods. In practice, however, setting up such a super-model is both conceptually difficult and computationally infeasible except in the simplest of problems. Thus, seeking more systematic ways of model checking is necessary and important to applied Bayesian analysis.

Judging models by implications and knowledge

Before diving into formal posterior predictive checking, judging model flaws by their practical implications or examining the inferences from the model using our knowledge is the very first step whenever it is feasible. It is a fact that probability models in most data analysis will not be perfectly true. This is particularly the case when we are making convenient assumptions, and thus it is important to determine whether the model's deficiencies have a noticeable effect on the substantive inferences. This is a central task of Bayesian sensitivity analysis. Checking model adequacy is most easily done through *external validation* using the model to make predictions about future data, and then collecting those data and comparing to the predictions. Often, it will not be possible to collect additional data for the purpose of model checking. In this case, we can at least check whether the model is consistent with the current observed data. If the model fits well, then replicated data generated under the model from the posterior predictive distribution should look similar to the observed data. Such a self-consistency check can reveal any observed discrepancy due to model misfit or chance. However, it is apparent that this is a rather conservative approach depending entirely on the observed data for both model estimation and checking. The basic technique for checking the fit of a model to data is thus to draw simulated values from the posterior predictive distribution of replicated data and compare these samples to

the observed data [41].

Let y denote the observed data, y^{rep} be the replicated data, and θ be the vector of parameters. If the model has explanatory variables, x, y^{rep} will be generated under the same set of x. We assume that y^{rep} is independent of y conditional on θ, so the posterior predictive distribution given the current state of knowledge is $f(Y_{rep}|y) = \int_\Theta f(Y_{rep}|\theta)p(\theta|y)d\theta$. In order to measure the discrepancy between the model and data, we define a test quantity $T(y, \theta)$, which is a scalar summary of parameters and data that is used as standard when comparing data to predictive simulations. Similar to the classical test, Bayesian p-value (posterior predictive p-value) is defined as the probability that the replicated data could be more extreme than the observed data, as measured by the test quantity [41]:

$$pB = Pr(T(y^{rep}, \theta) \geq T(y, \theta)|y).$$

Here the probability is taken over the posterior distribution of θ and the posterior predictive distribution of y^{rep} (that is the joint distribution, $f(\theta, y^{rep}|y)$), i.e.

$$pB = \int_\Theta \int_{y^{rep}} I_{T(y^{rep},\theta)\geq T(y,\theta)} f(Y_{rep}|\theta)\pi(\theta|y)dy^{rep}d\theta, \tag{2.4.7}$$

where I is the indicator function [41]. In this formula, we have used the property of the predictive distribution that $f(y^{rep}|y, \theta) = f(y^{rep}|\theta)$ arising from the conditional independence assumption. In practice, we usually compute the posterior predictive distribution using simulation. For instance, assuming we already have K simulated values from the posterior distribution of θ, we draw one y^{rep} from the distribution $f(y^{rep}|\theta)$ for each simulated θ; we now have K draws from the joint posterior distribution, $f(y^{rep}, \theta|y)$. The posterior predictive check is accomplished by comparing the realized test quantities, $T(y, \theta^k)$, and the predictive test quantities, $T(y^{rep^k}, \theta^k)$. The estimated p-value is just the proportion of these K simulations for

which $T(y^{rep^k}, \theta^k) \geq T(y, \theta^k)$ is true [41].

Without having to calculate posterior predictive p-values, we can use graphical model checking by displaying the data alongside simulated data from the predictive distribution induced by the assumed model, and to look for systematic discrepancies between the observed and simulated data. Three kinds of graphical display often used are direct display of all the data, display of data summaries or parameter inferences, and graphs of residuals or other measures of discrepancy between model predictions and observed data. In addition to the graphical check, we usually adopt numerical posterior predictive check. That is, we specify a test quantity $T(y)$ or $T(y, \theta)$ and an appropriate predictive distribution for the replications y^{rep}. We then compute the pB as given in (2.4.7).

If the test quantity depends on θ and y, then we can obtain $T(y, \theta^k)$ and its replication $T(y^{rep}, \theta^k)$ through K simulations. Thus, the comparison can be displayed either as a scatter plot of the values $T(y, \theta^k)$ vs. $T(y^{rep}, \theta^k)$ or as a histogram of the differences, $T(y, \theta^k) - T(y^{rep}, \theta^k)$. The indication of the model adequacy is that the scatter plot should be symmetric about a 45 degree line or the histogram should include 0. To cover more than one possible model failure in reflecting the process that generated the data, we can compute posterior predictive p-values for a variety of test quantities [41]. Ideally, the test quantities T will be chosen to reflect aspects of the model that are relevant to the scientific purpose to which the inference will be applied. Test quantities are commonly chosen to measure a feature of the data not directly addressed by the probability models; for example, ranks of the sample, or correlation of model residuals with some possible explanatory variable [41].

2.5 Basic Models for Spatial Data

Spatially correlated data arise in many areas of scientific research such as climatology, ecology, epidemiology, and real estate marketing. These type of data are characterized by being highly multivariate and geographically referenced. Spatial data sets are usually classified into one of three basic types: point-referenced data, areal data, and point pattern data. The corresponding models are spatial point-level (geostatistical) models, areal models, and point process models respectively. Our main focus in this review lies with models for areal data as this relates to the data set motivating this thesis. Readers interested in point-level models and point process models are referred to Banerjee et al. [42], and references therein. We provide a brief description of the different data structures here.

2.5.1 Introduction to Spatial Data

In this section, we first introduce the conventional definition for each type of spatial data. Then the characteristic of each type of spatial data is illustrated by examples.

Point-reference data refers to the case where $Y(s)$ is a random vector at a location $s \in R^d$ and s varies *continuously* over D, a fixed subset of R^r that contains an r-dimensional rectangle of positive volume. The fundamental concept underlying the theory is a stochastic process $\{Y(s) : s \in D\}$, where D is a fixed subset of r-dimensional Euclidean space. To capture spatial association, for any s_1, $s_2 \in D$, it is clear that $Y(s_1)$ and $Y(s_2)$ are dependent with strength of their dependence determined by the relative location of s_1 and s_2. To avoid the need of specifying the joint distribution for $Y(s)$ for all $s \in D$, it is typical to assume the spatial process to be Gaussian. In this case, all that is required is a specification for the mean and a valid covariance function. A commonly used and intuitive specification for

the covariance is the exponential model. Here the covariance between measurements at two locations is an exponential function of the interlocation distance, which is $\mathrm{Cov}[\mathrm{Y}(\mathrm{s}_1), \mathrm{Y}(\mathrm{s}_2)] = \sigma^2 \exp\{-\phi \|\mathrm{s}_1 - \mathrm{s}_2\|\}$, where σ^2 and ϕ are positive parameters [42].

(a) (b)

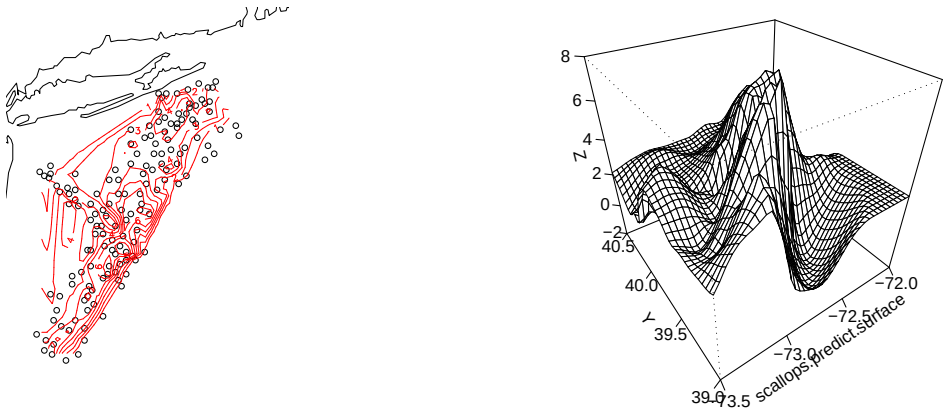


Figure 2.1: Point-level process: (a) Map of observed scallop sites and contours of raw log catch scallop data (b) Perspective plot of the kriged prediction surface for the scallop catch data [42].

In practice we will only observe $Y(s)$ at a finite set of locations, say $s_1, ..., s_n$, and we seek to infer about the mean, variability, and association structure of the process. We also seek to predict $Y(s)$. This process is a classical spatial prediction known as *kriging*. The problem is one of optimal spatial prediction: given observations of a random field $Y = (Y(s_1), ..., Y(s_n))'$, we would like to predict the variable Y at a site s_0 where it has not been observed. In other words, we seek for the best predictor of the value of $Y(s_0)$ based upon the data y. As illustrated in an exploratory analysis on spatially referenced fisheries data, scallop data in [42], $Y(s)$ may represent the density

of the log catch of scallops at site s. While it is conceptually sensible to assume the existence of scallops at all possible sites in the domain which is the Atlantic waters off the coasts of New Jersey and Long Island, New York, in practice the data will be a partial realization of that spatial process. So, predicting the density of scallop catch at a new site will give guidance to the fishing industry. Figure 2.1 (a) is a map of observed scallop sites and contours of raw log catch scallop data. Adding contours lines is necessary to carry out an interpolation, which is to fill in the gaps in the data over a regular grid (where there are no actual observed data) using a bivariate linear interpolation. The dots are the locations of the sites and the number of raw log scallop catch along each contour line is constant. The number of log catch is more easily seen in Figure 2.1 (b), a perspective plot of the kriged prediction surface for the scallop catch data. We can easily find out the number of log catch from the prediction surface and the corresponding longitude (X-axis) and latitude (Y-axis).

The second category of spatial data is *Point pattern data*, where D itself is random; its index set gives the locations of random events that are the spatial point pattern. $Y(s)$ itself can simply equal 1 for all $s \in D$ (including occurrence of the event), or possibly give some additional information on process values at random locations (producing a *marked point pattern process*) [42]. Data that can be represented as point patterns occur frequently in spatial statistics. In forestry, for example, the positions of trees in a forest forms a point pattern in the plane. There may be variables associated with points; the height of a tree may have been recorded along with the position of the tree. A point process is a model for the stochastic process generating the spatial distribution of the points in a point pattern. The positions of the points may be completely independent of each other or the points may appear in spatial clusters. Clustered point processes are relevant for the modeling of positions of weed plants or of disease infected plants which typically appear in clusters in the field.

Clustered point process data also often appear in epidemiological studies. For example, spatial cluster detection is an important tool in cancer surveillance to identify areas of elevated risk and to generate subsequent hypotheses about cancer etiology. A spatial disease cluster may be defined as an area with an unusually elevated disease incidence rate. There are several cluster detection methods used in spatial epidemiology to investigate apparent suspicious groupings of cancer occurrences in both regional count data and case-control data, where the controls are often sampled from the at-risk population and are used to estimate local relative risk or local rates, depending on the method utilized. Wheeler [43] discussed the comparison of spatial clustering and cluster detection techniques for childhood leukemia incidence in Ohio 1996 - 2003. To illustrate a spatial point pattern, Figure 2.2 from Wheeler [43] reveals the spatial pattern of the childhood leukemia cases (unfilled circles) and controls (filled circles). The controls sampled, based on 10 to 1 ratio of controls to cases, approximate the general distribution of population in Ohio. After visually accounting for the distribution of population, as represented by the controls, Figure 2.2 appears no clear overall clustering in the cases and no obvious clusters of cases. However, it seems that there might exist local clusters of irregular shape, in areas of central, southern, and eastern Ohio. Spatial point process models can be used to investigate the existence of such clusters and to quantify the specific nature of the clustering.

Areal data are spatial data corresponding to observations taken over a pre-defined sets of geographical units. Areal data is often referred to as *lattice* data. Such data often arise from agricultural field trials where the plots cultivated form a regular lattice or from image restoration. However, in practice most areal data are summaries over an *irregular* lattice, like a collection of county or other regional boundaries. This is the spatial structure characterizing the Quebec AMI data, and so we discuss mod-

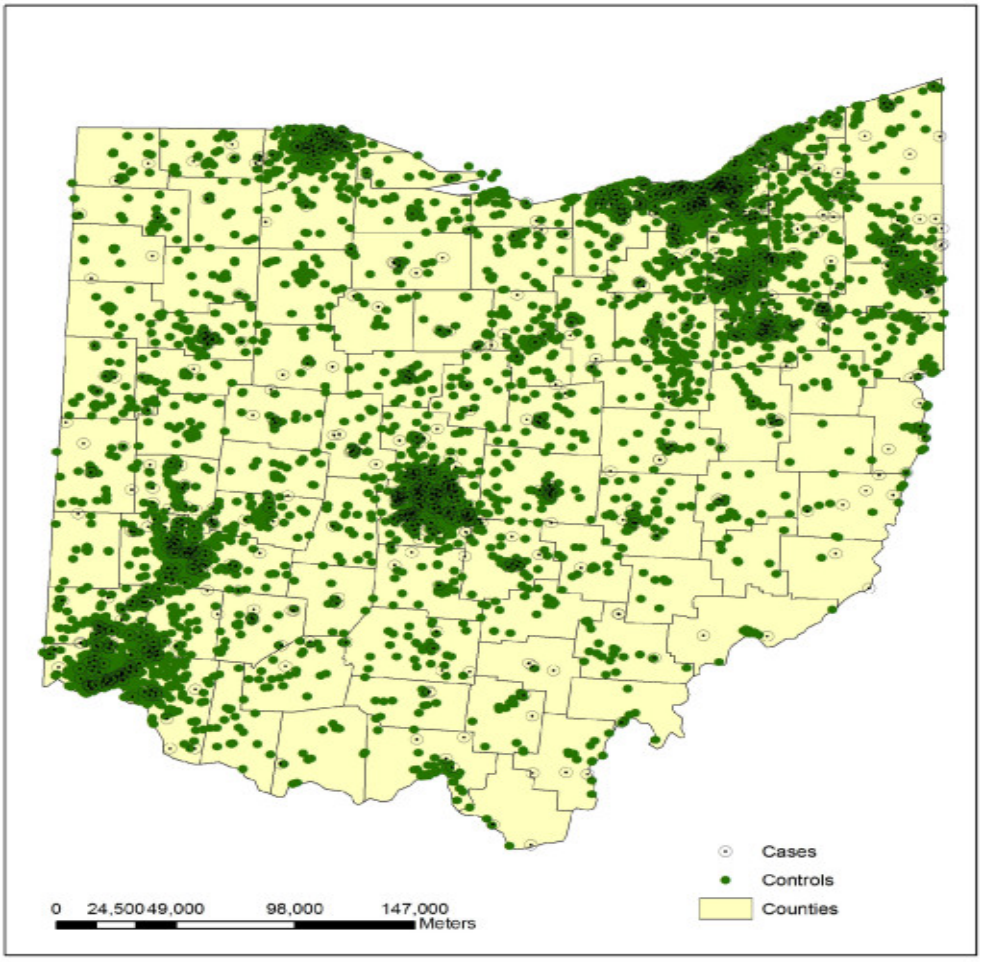


Figure 2.2: Point pattern process: Childhood leukemia cases (unfilled circles) and controls (filled circles) for years 1996 - 2003, Ohio [43].

eling of areal data in more detail.

The primary concept used to characterize the spatial structure of areal data is a proximity (or adjacency) matrix, $\mathbf{W}=(w_{ij})_{n\times n}$. Given measurements $Y_1, Y_2, ..., Y_n$ associated with areal (geographic) units $1, 2, ..., n$, the constant w_{ij} represents some measure of spatial connection between units i and j. Possibilities include binary choices, where $w_{ij} = 1$ if geographic units i and j share a common boundary (neighbours in a regional map). Alternatively, w_{ij} could reflect *distance* between units, for example, a decreasing function of intercentroidal distance between the units (as in a county or other regional map). In general, $\mathbf{W}$ provides the mechanism for introducing spatial structure into the formal modeling of spatial dependence.

A common application of spatial modeling of areal data arises in disease mapping, which is a very active area of biostatistical and epidemiological interest. Mapping of the geographical distribution of cancer incidences or mortality rates can help us understand spatial patterns of disease and identify differences in disease burden across an area. These maps can be used by those involved in the planning of services or in cancer prevention and control programs, both locally and nationally, and can provide useful background information for academics, the government, and the general public. For example, Ferrándiz et al. [44] did a spatial analysis of the relationship between mortality from cardiovascular and cerebrovascular disease in Comunidad Valenciana of Spain. In their study, they also performed disease mapping of standardized mortality ratios to detect clusters of municipalities with high risk. The standardized mortality ratio for region i is $SMR_i = O_i/E_i$, where O_i is the observed and E_i is the expected mortality count, which is indirectly standardized by age groups for each sex, as well as by levels of a deprivation index. From a surveillance perspective, disease mapping before and after standardization by levels of a covariate can determine if

removing the effects of such a covariate changes the geographical pattern of relative risks. By comparing the resulting maps, one can verify whether high-risk regions move to lower levels of risk or if they remain high, indicating that factors other than this covariate are still affecting population health status. This will lead to finding hidden factors not included in the study. The geographic pattern can help surveillance analysts determine the nature of these hidden factors.

For illustration, we also include the maps from their study [44] to explore spatial distribution of risk. Figure 2.3 A and C represent smoothed municipal relative risks, B and D distinguish between significantly high, low and non significant 95% confidence intervals of SMR_s. In Figure 2.3 A and B, SMR_s have been standardized by age, sex, and deprivation index. The standardization in C and D has included one more covariate, magnesium levels as well. No spatial trend is apparent from maps, but several clusters of different sizes are scattered over the entire region as circled. The upper circle shows a cluster of municipalities with high SMR and significant confidence intervals. This is obvious in the map of smoothed $SMRs$ standardized by age, sex, and deprivation index (Figure 2.3 A) and in the map with significance of confidence intervals (Figure 2.3 B). The lower circle shows another less extreme cluster. When the standardizaton also includes magnesium levels, it is done to remove the effect of this covariate in some sense. Comparing Figure 2.3 A and B with C and D, we can see the effect of removing this covariate. For example, the upper circle shows that this change produces even more municipalities with significant confidence intervals than before. In the lower circles, the opposite is true. Some municipalities show a decrease in the significance of their SMR_s (so that their previous high relative risk has been partically explained by their level of magnesium).

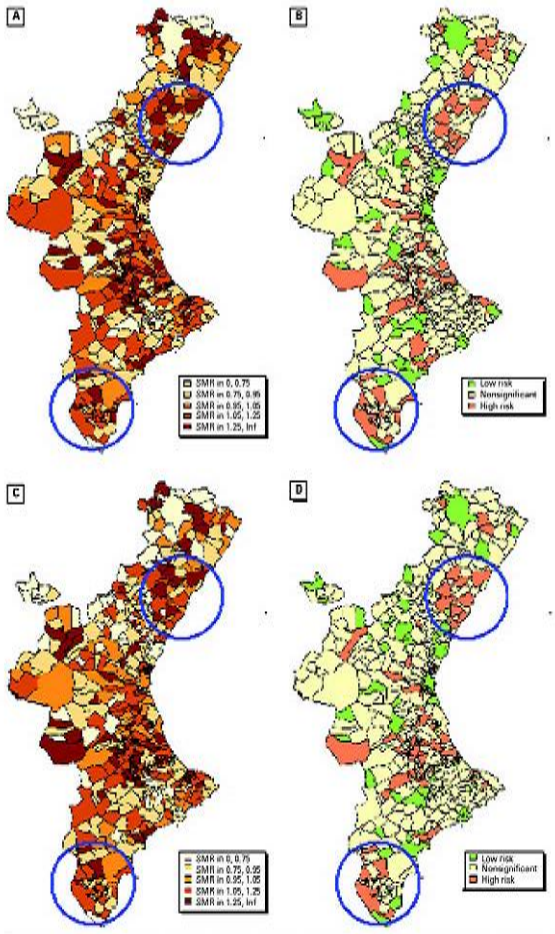


Figure 2.3: Areal data in disease mapping of total cerebrovascular mortality: (A) smoothed SMRs after standardization by age, sex, and deprivation index, (C) smoothed SMRs after further standardization by Mg, (B) and (D) are significance of 95% confidence interval corresponding to (A) and (C)

2.5.2 Brook's Lemma and Markov Random Fields

In the spatial modeling of areal data, it is common to adopt a conditional specification approach, where the joint distribution for a random vector $\boldsymbol{y} = (y_1, .., y_n)$ is specified through a set of conditional distributions. Brook's Lemma [45] is a useful technical result in this regard. It is clear that given the joint distribution $p(y_1, .., y_n)$, the full conditional distributions, $p(y_i|y_j, j \neq i), i = 1, ..., n$, are uniquely determined. But the converse is not always true. In general, a set of conditional distributions are compatible if they can determine a unique and valid joint distribution. If we have a set of compatible full conditional distributions $p(y_i|y_j, j \neq i), i = 1, ..., n$, the form of the resulting joint distribution can be constructed using Brook's Lemma, which is

$$p(y_1, ..., y_n) = \frac{p(y_1|y_2, ..., y_n)}{p(y_{10}|y_2, ..., y_n)} . \frac{p(y_2|y_{10}, y_3, ..., y_n)}{p(y_{20}|y_{10}, y_3, ..., y_n)} ... \frac{p(y_n|y_{10}, ..., y_{n-1,0})}{p(y_{n0}|y_{10}, ..., y_{n-1,0})} . p(y_{10}, ..., y_{n0}). \tag{2.5.1}$$

Here, $\mathbf{y_0} = (y_{10}, .., y_{n0})$ is any fixed point in the support of $p(y_1, ..., y_n)$. Hence the form of $p(y_1, ..., y_n)$ is obtained, up to a normalizing constant, through the full conditional distributions. If $p(y_1, ..., y_n)$ is improper, then this is the best we can do; if $p(y_1, ..., y_n)$ is proper, then the fact that it integrates to 1 determines the constant. The constructive nature of (2.5.1) enables us to create $p(y_1, ..., y_n)$ simply by calculating the product of ratios. This sheds light to specifying the spatial areal model. Usually, when the number of areal units is very large, we do not seek to write down the joint distribution of the Y_i. Rather we prefer to work (and model) exclusively with the corresponding n full conditional distributions. In fact, when specifying spatial models in this way, we would think that the full conditional distribution for Y_i should depend only upon the geographical neighbours of region i. Let ∂_i denote the set of neighbours of region i, then a typical specification based on full conditional distributions arises through the form

$$p(y_i|y_j, j \neq i) = p(y_i|y_j, j \in \partial_i), \tag{2.5.2}$$

which is commonly referred to as a Markov random field (MRF). Detailed explorations of MRF models may be found in Cressie [46].

2.5.3 Univariate CAR Modeling

Before introducing the CAR model, we will review three important definitions: clique, potential function, and Gibbs distribution [42]. A clique is a set of cells such that each element in the set is a neighbour of every other element in the set. In terms of a graph, a clique represents a graph having M nodes and there is an edge connecting each pair of i and j in M. A potential function of order k has k exchangeable arguments that typically operates on the variable values y_{s_1}, y_{s_2}, ..., y_{s_k} associated with a clique $\{s_1, s_2, ..., s_k\}$ of size k. Examples of potential functions for $k = 2$ are: $y_i y_j$, $(y_i - y_j)^2$ and $y_i y_j + (1 - y_i)(1 - y_j)$. A joint distribution $p(y_1, ..., y_n)$ is said to be a Gibbs distribution if it is a function of Y_i only through potentials on cliques induced by the neighbourhood structure. The Gibbs distribution takes the form

$$p(y_1, ..., y_n) \propto \exp\left\{\gamma \sum_k \sum_{\alpha \in M_k} \phi^{(k)}(y_{\alpha_1}, ..., y_{\alpha_k})\right\},$$

where $\phi^{(k)}$ is a potential of order k, M_k is the collection of all cliques of size k and $\gamma > 0$ is a scale parameter. The Hammersley-Clifford Theorem [47] demonstrates that if we have a MRF then the corresponding joint distribution is a Gibbs distribution.

Among the class of MRF models, Gaussian conditionally autoregressive (CAR) model [47], a special case of MRF, has been the most widely used. Let $\Phi = (\phi_1, ..., \phi_n)'$ be a random vector with its joint distribution specified through a set of conditional distributions

$$[\phi_i | \phi_j, j \neq i] \quad \sim \quad N\left(\sum_j \frac{W_{ij}}{W_{i+}} \phi_j, \frac{\tau^2}{W_{i+}}\right), i = 1, ..., n, \tag{2.5.3}$$

where $W_{i+} = \sum_{j=1}^{n} W_{ij}$ denotes the sum of i^{th} row in W and τ^2 is the variance component. The mean of $[\phi_i|\phi_j, j \neq i]$, denoted as $\bar{\phi_i}$ $(\sum_j \frac{W_{ij}}{W_{i+}}\phi_j)$, is a weighted average of neighbouring regions with region i and the variance depends on the number of neighbours that region i has. The more neighbours that i has, the more information we can get and the less variability involved. These full conditionals are compatible. So by Brook's Lemma, the joint distribution can be written as

$$p(\phi_1, ..., \phi_n) \propto \exp\left\{-\frac{1}{2\tau^2}\boldsymbol{\phi}'(D_W - W)\boldsymbol{\phi}\right\}, \tag{2.5.4}$$

or (2.5.4) can be rewritten as

$$p(\phi_1, ..., \phi_n) \propto \exp\left\{-\frac{1}{2\tau^2}\sum_{i\neq j} W_{ij}(\phi_i - \phi_j)^2\right\}, \tag{2.5.5}$$

where $D_W = \text{diag}\{W_{1+}, W_{2+}, ..., W_{n+}\}$ is a diagonal matrix. Note that since $(D_W - W)\mathbf{1} = \mathbf{0}$, (2.5.4) is a singular multivariate normal distribution. There does not exist a valid normalizing constant. So it is an improper distribution and is often referred to as an intrinsically autoregressive model (IAR). Because data could not arise under an improper stochastic mechanism, we cannot use (2.5.4) as a model. Hence, we can only use such an improper CAR model as a prior distribution when introducing random spatial effects at the second stage of a hierarchical specification. In practice, the ϕ_i are sampled using the full conditional distributions in (2.5.3) and a linear constraint such as $\sum_{i=1}^{n} \phi_i = 0$ is imposed in order to recenter each sampled ϕ_i [48].

To ensure $\Sigma_\phi^{-1}(\frac{1}{\tau^2}(D_W - W))$ is non-singular, we can introduce a parameter ρ into the mean specification in (2.5.3), i.e. $E(\phi_i|\phi_j, j \neq i) = \rho\bar{\phi_i}$ [46]. This results in the precision matrix $\frac{1}{\tau^2}(D_W - \rho W)$ in (2.5.4) which is nonsingular if $\rho \in (\lambda_{min}^{-1}, \lambda_{max}^{-1})$ where λ_{min} and λ_{max} are smallest and largest eigenvalues of the scaled adjacency matrix $D_W^{-1}W$ [42]. Although it is not always true, it is typically the case for standard W that $\lambda_{max} = 1$ and $\lambda_{min} < 0$. We denote this distribution by $CAR(\rho, \tau^2)$, where

ρ can be interpreted as a coefficient which measures spatial association and $\frac{1}{\tau^2}$ plays the role of a *conditional clustering* parameter quantifying the *shrinkage* due to the average number of neighbours [49]. Moreover, $\rho = 0$ implies that the ϕ_i's are independent but with individual variances which depend upon the number of neighbours. In practice, we usually take $0 < \rho < 1$ as negative smoothness parameters are not desirable [50].

In general, there are two different representations of proper CAR models. We denote the one just introduced as $CAR(\rho, \tau^2)$ and it follows directly from (2.5.4)

$$p(\phi_1, ..., \phi_n) \propto \exp\left\{-\frac{1}{2\tau^2}\boldsymbol{\phi}'(D_W - \rho W)\boldsymbol{\phi}\right\}, \tag{2.5.6}$$

where $E(\boldsymbol{\phi}) = \mathbf{0}$. In practice, we often work with linear regression models where there is a intercept term μ. If we center $\boldsymbol{\phi}$ at μ (ie. $E(\boldsymbol{\phi}) = \boldsymbol{\mu}$), then we obtain the centered version of CAR

$$p(\phi_1, ..., \phi_n) \propto \exp\left\{-\frac{1}{2\tau^2}(\boldsymbol{\phi} - \boldsymbol{\mu})'(D_W - \rho W)(\boldsymbol{\phi} - \boldsymbol{\mu})\right\}. \tag{2.5.7}$$

We denote it as $CAR(\mu, \rho, \tau^2)$ and this representation will appear in our model implementation in Section 4.2.1.

2.6 Space-Varying Regression

Space-varying regression models are generalizations of standard regression models where the regression coefficients are allowed to change spatially. When working with areal data, the spatial structure of regression coefficients can be modeled using a proper CAR model, thus enabling incorporation of neighbouring structures when modeling the effects of covariates. Different from the ordinary regression models, space-varying coefficient models can introduce spatial information in the model to allow the parameters of interest, such as regression coefficients to vary in space. That

is, the effect of a covariate varies depending on the geographical region. In this way, the spatial configuration of the regions is associated not only with random effects operating as intercepts, but also with the impact of explanatory variables on the response variable. It is reasonable to assume a spatial continuity on the geographical and social environment. Thus we expect that any variation in covariate effects should occur smoothly over space, and the regression parameters should exhibit a spatial structure. In this modeling approach, the coefficient of a covariate varies from region to region.

2.6.1 Recent Applications of Space-Varying Regression

Several recent studies applied space-varying regression models to important scientific applications. Assuncao [51] gives a comprehensive review in this regard. The following examples found in that review have similar settings as our study. Brunsdon et al. [52] applied the method of allowing linear regression coefficients to vary in space to model levels of limiting long-term illness measured for people 45-65 years old using 1991 UK census data. Their study shows considerable evidence that some predictor variables had different impacts on the response variable depending on the census ward location in the north of England. Pavlov [53] used a similar method to estimate home values allowing the parameters of the covariates to vary in space. He finds a substantial spatial variation of the marginal values of hedonic characteristics and his results provide an insight into the segmentation of the market that otherwise would be difficult to obtain. Congdon [54] studied the geographical variation at the local authority level of infant health outcomes in North East Thames Health Region from 1991 to 1993. Modeling still-births and perinatal deaths incidence rates, he allows the risk factors' coefficients to vary depending on the mother's residence place. There is considerable evidence that maternal risk factors, such as mother's age, have different effects on the outcomes depending on the District Health Authority region the mother

lives in. The results of these studies are in favour of considering the space-varying regression to our AMI data, as we suspect the effect of revascularization may vary spatially due to the fact that each local health unit has its own unique health human resource and capability that depend upon its cardiac specialists, clinic practice, and medical equipment.

2.6.2 Formulation of Space-Varying Regression

As space-varying regression models are generalizations of standard generalized linear models, the formulation is straightforward. Suppose we have random variables Y_{ij} representing the observation obtained from the j^{th} subject in region i (e.g., survival time after heart attack) for $i = 1, .., n$, with $j = 1, ..., J_i$. The corresponding covariates represented by $\mathbf{x_{ij}} = (x_{ij0}, x_{ij1}, ..., x_{ijp})'$, where x_{ij0} equals 1 representing an intercept. The space-varying model assumes that regression coefficients are regionally varying and can be written as $\eta_{ij} = \boldsymbol{\beta_i}' \mathbf{x_{ij}}$, where η_{ij} is related to the mean $\mu_{ij} = E(y_{ij})$ via the link function $\eta_{ij} = g(\mu_{ij})$. Here, each covariate in the regression model is associated with, not one, but an entire vector of regression coefficients allowing for region specific associations. In the case of a single covariate and a log-linear Gaussian model for y_{ij} (assuming $y_{ij} > 0$) we obtain

$$\log y_{ij} = \beta_{0_i} + \beta'_{1_i} \mathbf{x_{ij}} + \epsilon_{ij},$$

where $\epsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2)$ represents measurement error, and $(\beta_{0_i}, \beta_{1_i})$ are the region specific intercept and slope parameters with $E(\beta_{0_i}) = \beta_0$ and $E(\beta_{1_i}) = \beta_1$ representing population averages. The vectors $\boldsymbol{\beta_0} = (\beta_{0_1}, ..., \beta_{0_n})'$ and $\boldsymbol{\beta_1} = (\beta_{1_1}, ..., \beta_{1_n})'$ are treated as random effects, and are assigned a spatially structured joint distribution, for example, the proper CAR model of Section 2.5.3.

Chapter 3
Background on Computational Methods

Advanced computing technology has reshaped many approaches to statistics. Much work has been done on algorithmic approaches such as the EM algorithm, resampling techniques, and Markov chain Monte Carlo (MCMC) methods. In particular, MCMC has enjoyed a great deal of successes thanks in large part to its applications in Bayesian inference. It also has applications in frequentist inference. In essence, MCMC methods are a class of algorithms for sampling from a probability distribution based on constructed ergodic Markov chain that has the desired distribution as its equilibrium distribution. The posterior distribution, for example, can be viewed as one type of equilibrium distribution. The state of the chain after a large number of iterations is then used as a sample from the desired distribution. The quality of the sample improves as a function of the number of iterations.

In this chapter, we discuss two simple and common MCMC methods which we will use in our analysis. The rest of this chapter is organized as follows. In Section 3.1, we give a brief introduction to Markov Chains. In Section 3.2 and 3.3, we present the Gibbs sampler and Metropolis-Hastings (M-H) algorithm. Convergence issues of MCMC is discussed in Section 3.4. For a more detailed discussion of MCMC, see Gamerman and Lopes [55].

3.1 Markov Chains

A Markov chain is a stochastic process with the Markov property that, given the present state, future states are independent of the past states. In other words, the

description of the present state fully captures all the information that could influence the future evolution of the process. Future states will be reached through a probabilistic process instead of a deterministic one. At each step the system may change its state from the current state to another state, or remain in the same state, according to a certain probability distribution. The changes of state are called transitions, and the probabilities associated with various state-changes are called transition probabilities. An example of a Markov chain is a simple random walk where the state space is a set of vertices of a graph and the transition steps involve moving to any of the neighbours of the current vertex with equal probability (regardless of the history of the walk). The following discussion on Markov chains focus on the limit theorems and ergodic theory that govern the iterative simulation techniques which will be used in later sections.

3.1.1 Basic Concepts of Markov Chains

A Markov chain is a special type of stochastic process which deals with sequences of random variables. A stochastic process can be defined as a collection of random quantities $\{\boldsymbol{\theta}^{(t)} : t \in T\}$ for some set T, where $\boldsymbol{\theta}^{(t)} \in S$. The set $\{\boldsymbol{\theta}^{(t)} : t \in T\}$ is said to be a stochastic process with state space S and index set T. For illustration, the index set T is taken as countable with equal spacing between elements, defining a discrete time stochastic process. Without loss of generality, it will be assumed to be the set of natural numbers $\mathcal{N}$ and in our context will represent the iterations of a simulation scheme. The state space will in general be a subset of R^d representing the support of a parameter vector. However, for simplicity we shall describe the results focusing primarily on discrete state spaces so as to minimize the technical detail of our presentation.

Definition of Markov chain

A Markov chain is a stochastic process where given the present state, past and future states are independent. This property can be more formally stated as

$$Pr(\boldsymbol{\theta}^{(n+1)} \in A|\boldsymbol{\theta}^{(n)} = x, \boldsymbol{\theta}^{(n-1)} \in A_{n-1}, ..., \boldsymbol{\theta}^{(0)} \in A_0) = Pr(\boldsymbol{\theta}^{(n+1)} \in A|\boldsymbol{\theta}^{(n)} = x) \tag{3.1.1}$$

for all sets $A_0, ..., A_{n-1}, A \subset S$ and $x \in S$. The Markov property in (3.1.1) can also be written in the equivalent forms [55]:

1. E$[h(\boldsymbol{\theta}^{(n)})|\boldsymbol{\theta}^{(m)}, \boldsymbol{\theta}^{(m-1)}, ..., \boldsymbol{\theta}^{(0)}]$=E$[h(\boldsymbol{\theta}^{(n)})|\boldsymbol{\theta}^{(m)}]$ for all bounded functions h and $n > m \geq 0$;

2. $Pr(\boldsymbol{\theta}^{(n+1)} = y|\boldsymbol{\theta}^{(n)} = x, \boldsymbol{\theta}^{(n-1)} = x_{n-1}, ..., \boldsymbol{\theta}^{(0)} = x_0) = Pr(\boldsymbol{\theta}^{(n+1)} = y|\boldsymbol{\theta}^{(n)} = x)$ for all $x_0, ..., x_{n-1}, x, y \in S$.

The above form is clearly appropriate only for discrete state spaces. In general, the probabilities in (3.1.1) depend on x, A, and n. When they do not depend on n, the chain is said to be homogeneous. In this case, a transition function or kernel $P(x, A)$ can be defined as:

1. for all $x \in S$, $P(x, \cdot)$ is a probability measure over S,

2. for all $A \subset S$, the function $x \longmapsto P(x, A)$ can be evaluated.

It is also useful when dealing with a discrete state space to identify $P(x, \{y\}) = P(x, y)$. In this context, the function is called a *transition probability* and satisfies:

1. $P(x, y) \geq 0, \forall x, y \in S$;

2. $\sum_{y \in S} P(x, y) = 1, \forall x \in S$;

as any probability measure $P(x, \cdot)$ should.

Transition probabilities

In the case of a discrete state space $S = \{x_1, x_2, ...\}$, a transition matrix P with $(i, j)th$ element given by $P(x_i, x_j)$ can be defined. Suppose S is finite with k elements, the

transition matrix P is given by

$$P = \begin{pmatrix} P(x_1, x_1) & \cdots & P(x_1, x_k) \\ \vdots & & \vdots \\ P(x_k, x_1) & \cdots & P(x_k, x_k) \end{pmatrix}$$

Transition matrices have all rows summing to one and such matrices are called stochastic. The transition probabilities from state x to state y over m steps, denoted by $P^m(x, y)$, is given by the probability of a chain moving from state x to state y in exactly m steps. It can be obtained for $m \geq 2$ as

$$\begin{aligned} & P^m(x, y) = Pr(\boldsymbol{\theta}^{(m)} = y | \boldsymbol{\theta}^{(0)} = x) \\ &= \sum_{x_1} \cdots \sum_{x_{m-1}} Pr(\boldsymbol{\theta}^{(m)} = y, \boldsymbol{\theta}^{(m-1)} = x_{m-1}, \cdots, \boldsymbol{\theta}^{(1)} = x_1 | \boldsymbol{\theta}^{(0)} = x) \\ &= \sum_{x_1} \cdots \sum_{x_{m-1}} Pr(\boldsymbol{\theta}^{(m)} = y | \boldsymbol{\theta}^{(m-1)} = x_{m-1}) \cdots Pr(\boldsymbol{\theta}^{(1)} = x_1 | \boldsymbol{\theta}^{(0)} = x) \\ &= \sum_{x_1} \cdots \sum_{x_{m-1}} P(x, x_1) P(x_1, x_2) \cdots P(x_{m-1}, y) \end{aligned} \tag{3.1.2}$$

where the second equality holds due to the Markovian property of the process given in (3.1.1) [55]. The last equality means that the matrix containing elements $P^m(x, y)$ is also a stochastic matrix and is given by P^m. Also, for completeness, we let $P^1(x, y) = P(x, y)$ and $P^0(x, y) = I(x = y)$. The above derivation can be used to established that

$$\begin{aligned} P^{n+m}(x, y) &= \sum_z Pr(\boldsymbol{\theta}^{(n+m)} = y | \boldsymbol{\theta}^{(n)} = z, \boldsymbol{\theta}^{(0)} = x) Pr(\boldsymbol{\theta}^{(n)} = z | \boldsymbol{\theta}^{(0)} = x) \\ &= \sum_z P^n(x, z) P^m(z, y). \end{aligned} \tag{3.1.3}$$

Equations (3.1.3) are usually called Chapman-Kolmogorov equations where all summations are with respect to the elements of the state space S, and results are valid for any stage of the chain due to the assumed homogeneity [55]. Higher order transition

matrices can be formed with these higher transition probabilities and it can be shown that they satisfy the relation $P^{n+m} = P^n P^m$ and, in particular, $P^{n+1} = P^n P$ [55].

Marginal Distribution

The marginal distribution of the n^{th} state of the chain $\boldsymbol{\theta}^{(n)}$ can be defined by the row vector $\boldsymbol{\pi}^{(n)}$ with components $\boldsymbol{\pi}^{(n)}(x_i)$, for all $x_i \in S$. For finite state spaces, this is a k-dimensional vector

$$\boldsymbol{\pi}^{(n)} = (\boldsymbol{\pi}^{(n)}(x_1), ..., \boldsymbol{\pi}^{(n)}(x_k)).$$

When $n = 0$, $\boldsymbol{\pi}^{(0)}$ is the initial distribution of the chain. It follows that

$$\begin{aligned} \boldsymbol{\pi}^{(n)}(y) &= Pr(\boldsymbol{\theta}^{(n)} = y) \\ &= \sum_{x \in S} Pr(\boldsymbol{\theta}^{(n)} = y | \boldsymbol{\theta}^{(0)} = x) Pr(\boldsymbol{\theta}^{(0)} = x) \\ &= \sum_{x \in S} P^n(x, y) \boldsymbol{\pi}^{(0)}(x). \end{aligned} \tag{3.1.4}$$

The above equation can be written in matrix notation as

$$\begin{aligned} \boldsymbol{\pi}^{(n)} &= \boldsymbol{\pi}^{(0)} P^n \\ &= \boldsymbol{\pi}^{(0)} P^{n-1} P \\ &= \boldsymbol{\pi}^{(n-1)} P. \end{aligned} \tag{3.1.5}$$

Thus we have a relationship between consecutive marginal distributions $\boldsymbol{\pi}^{(n-1)}$ and $\boldsymbol{\pi}^{(n)}$ [55].

3.1.2 Properties of Markov Chains

The following quantities of interest are important in describing the properties of a Markov chain with a state space S and transition matrix P:

(i) The hitting time of $A \subset S$ is defined as $T_A = \min\{n \geq 0 : \boldsymbol{\theta}^{(n)} \in A\}$ if $\boldsymbol{\theta}^{(n)} \in A$ for some $n > 0$. Otherwise, $T_A = \infty$;

(ii) The probability of the chain starting from state x hitting state y at any step is $\rho_{xy} = Pr_x(T_y < \infty)$;

(iii) And it can be shown that if the initial state of a chain is x, then the expected value of the hitting time of state y is [55]

$$E(T_y|\boldsymbol{\theta}^{(0)} = x) = \sum_{n=0}^{\infty} Pr_x(T_y > n);$$

(iv) The number of visits of a chain to a state y is [55]

$$N(y) = \sum_{n=1}^{\infty} I(\boldsymbol{\theta}^{(n)} = y);$$

(v) And the expected number of visits of a chain to a state y is [55]

$$E(N(y)|\boldsymbol{\theta}^{(0)} = x) = \sum_{n=1}^{\infty} P^n(x, y).$$

Recurrence

A state $y \in S$ is said to be recurrent if the Markov chain, starting in y, returns to y with probability 1 (i.e. $\rho_{yy} = 1$) and is said to be transient if it has positive probability of not returning to y (i.e. $\rho_{yy} < 1$). If a Markov chain starts at a recurrent state y, the hitting time of y, T_y, is a finite random quantity whose mean μ_y can be evaluated. If this mean is finite, the state y is said to be positive recurrent and otherwise the state is said to be null recurrent. Positive recurrence is a very important property for establishing limiting results, as will be seen in the next section.

Reducibility

A state y is said to be accessible from a different state x (written $x \rightarrow y$) if there is a non-zero probability ($\rho_{xy} > 0$) that starting from state x the process will visit state y at some time in the future. Formally, state y is accessible from state x if there exists an integer $n \geq 0$ such that $p^n(x, y) > 0$. A state x is said to communicate with state y

(written $x \longleftrightarrow y$) if it is true that both x is accessible from y and that y is accessible from x. A set of states $C \subset S$ is a communicating class if every pair of states in C communicates with each other, and no state in C communicates with any state not in C. (It can be shown that communication in this sense is an equivalence relation). A communicating class is closed if the probability of leaving the class is zero, namely that if x is in C but y is not, then y is not accessible from x. Finally, a Markov chain is said to be irreducible if the entire state space is a communicating class.

Periodicity

A state $x \in S$ has period d_x if any return to state x must occur in multiples of d_x time steps. For example, if it is only possible to return to state x in an even number of steps, then x is periodic with period 2. Formally, the period of a state is defined as

$$d_x = \gcd\{n \geq 1 : P^n(x, x) > 0\},$$

where $\gcd\{\cdot\}$ is the greatest common divisor of the set $\{\cdot\}$. Note that $P(x, x) > 0$ implies that $d_x = 1$ and furthermore if $x \longleftrightarrow y$ then $d_x = d_y$. Therefore, the states of an irreducible chain have the same period. A state x is aperiodic if $d_x = 1$. A chain is periodic with period d if all its states are periodic with period $d > 1$ and aperiodic if all its states are aperiodic.

Ergodicity

A state x is said to be ergodic if it is aperiodic and positive recurrent. If all states in a Markov chain are ergodic, then the chain is said to be ergodic.

3.1.3 Stationary Distribution and Limiting Theorems

The stationary distribution π is a key concept when studying asymptotic behaviour of the Markov chain as the number of steps or iterations $n \to \infty$. A distribu-

tion π is said to be a stationary distribution of a chain with transition probabilities $P(x, y)$ if

$$\sum_{x \in X} \pi(x) P(x, y) = \pi(y), \quad \forall y \in S. \tag{3.1.6}$$

Equation (3.1.6) can be written in matrix notation as $\pi = \pi P$. If the marginal distribution at any given step n is π then the distribution at the next step is $\pi P = \pi$. Once the chain reaches a stage where π is the distribution of the chain, the chain retains this distribution for all subsequent stages. This distribution is also known as the invariant or equilibrium distribution. If the stationary distribution π exists and $\lim_{n \to \infty} P^n(x, y) = \pi(y)$, then $\pi^{(n)}$ will approach π as $n \to \infty$ regardless of the initial distribution of the chain. In this sense, the distribution is also referred to as the limiting distribution.

An irreducible chain has a stationary distribution if and only if all of its states are positive-recurrent. In that case, π is unique and is related to the expected return time ($M_x = E(T_x)$) [55]:

$$\pi(x) = \frac{1}{M_x}.$$

Further, if the chain is both irreducible and aperiodic, then for any x and y [55],

$$\lim_{n \to \infty} p^n(x, y) = \pi(y) = \frac{1}{M_y}.$$

If a chain is not irreducible, its stationary distributions will not be unique (consider any closed communicating class in the chain; each one will have its own unique stationary distribution. Any of these will extend to a stationary distribution for the overall chain, where the probability outside the class is set to zero). However, if a state y is aperiodic, then it can be shown that [55]

$$\lim_{n \to \infty} P^n(y, y) = \frac{1}{M_y}$$

and for any other state x, let ρ_{xy} be the probability that the chain ever visits state y if it starts at x, then it can also be shown that [55]

$$\lim_{n\to\infty} P^n(x,y) = \frac{\rho_{xy}}{M_y}.$$

Ergodic Theorem

The first important limiting theorem is the ergodic theorem. The ergodic average of a real-valued function $t(\boldsymbol{\theta})$ is the average $\bar{t}_n = \frac{1}{n}\sum_{i=1}^{n} t(\theta^{(i)})$. If the chain is ergodic and $E_\pi(t(\boldsymbol{\theta})) < \infty$ with respect to the unique limiting distribution π then [55]

$$\bar{t}_n \overset{a.s.}{\to} E_\pi(t(\boldsymbol{\theta})) \text{ as } n \to \infty. \tag{3.1.7}$$

This result is a Markov chain equivalent of the strong law of large numbers. It states that averages of chain values also provide strongly consistent estimates of expectations taken with respect to the limiting distribution π despite their dependence. It is this theorem that justifies the use of iterative simulation methods based on Markov chains for Bayesian inference.

3.1.4 Reversible Markov Chains

The idea of a reversible Markov chain comes from the ability to "invert" a conditional probability using Bayes' Rule. Let $\{\boldsymbol{\theta}^{(n)}\}_{n\geq 0}$ be a homogeneous Markov chain with transition probabilities $P(x,y)$ and stationary distribution π. The reversed sequence of states $\boldsymbol{\theta}^{(n)}, \boldsymbol{\theta}^{(n-1)}, ...$ satisfies

$$Pr(\boldsymbol{\theta}^{(n)} = y | \boldsymbol{\theta}^{(n+1)} = x, \boldsymbol{\theta}^{(n+2)} = x_2, ...) = Pr(\boldsymbol{\theta}^{(n)} = y | \boldsymbol{\theta}^{(n+1)} = x)$$

and therefore defines a Markov chain. The transition probabilities for such a reversed chain would be

$$\begin{aligned} P_n^{\star}(x,y) &= Pr(\boldsymbol{\theta}^{(n)} = y | \boldsymbol{\theta}^{(n+1)} = x) \\ &= \frac{Pr(\boldsymbol{\theta}^{(n+1)} = x | \boldsymbol{\theta}^{(n)} = y) Pr(\boldsymbol{\theta}^{(n)} = y)}{Pr(\boldsymbol{\theta}^{(n+1)} = x)} \\ &= \frac{\pi^{(n)}(y) P(y,x)}{\pi^{(n+1)}(x)}. \end{aligned}$$

Thus, a Markov chain is said to be reversible if there is a π such that

$$\pi(x)P(x,y) = \pi(y)P(y,x), \text{ for all } x, y \in S. \tag{3.1.8}$$

This condition is also known as the detailed balance condition. Reversible chains are useful: if there is a distribution π satisfying (3.1.8) for an irreducible chain, then the chain is positive recurrent, reversible with stationary distribution π [55]. Summing over x gives

$$\sum_x \pi(x)P(x,y) = \pi(y).$$

Hence, construction of Markov chains with a given stationary distribution π (which is the posterior distribution in the context of Bayesian inference) reduces to finding transition probabilities $P(x,y)$ satisfying detailed balance equation (3.1.8). The Metropolis algorithm is an example of applying reversible Markov chains.

For a Markov chain in a continuous state space, there are few changes required with respect to the discrete case but the main results are still valid. In particular, convergence to the limiting distribution and the ergodic theorem need basically technical changes in the conditions of the chain to hold. Interested readers may refer to the Section 4.7 in the book [55].

3.2 Gibbs Sampler

The Gibbs sampler is an MCMC scheme that draws samples from the full conditional distributions. This is especially useful when direct sampling from the target distribution is costly or impossible, but the full conditional distributions are known exactly and can be easily sampled. It seems very straightforward to describe, but the mechanism that drives this scheme may be not so obvious. Hence we devote the following subsections to describe how the Gibbs sampler works. The examples here follow Casella and George [56] and Gamerman and Lopes [57].

3.2.1 Illustrating the Gibbs Sampler

Suppose we are given a joint density $f(x, y_1, ..., y_p)$, and are interested in obtaining characteristics of the marginal density

$$f(x) = \int ... \int f(x, y_1, ..., y_p) dy_1 ... dy_p, \tag{3.2.1}$$

such as the mean or variance. There are many cases where the integrations in (3.2.1) are extremely difficult to obtain either analytically or numerically. In such cases the Gibbs sampler provides a way for obtaining $f(x)$ and quantities related to $f(x)$.

The Gibbs sampler allows us to generate a sample $X_1, ..., X_m \sim f(x)$ without knowing the exact form of $f(x)$. By simulating a large enough sample, the mean, variance, or any other quantities related to $f(x)$ can be calculated to the desired degree of accuracy. It is easy to show that the limiting results of the calculations based on simulations are the population quantities. For example, to calculate the mean of $f(x)$, we could use $\bar{x} = \frac{1}{m} \sum_{i=1}^{m} X_i$. By the strong law of large numbers (or the Ergodic theorem in the case of Markov chains), we have [56]

$$\frac{1}{m} \sum_{i=1}^{m} X_i \overset{a.s.}{\to} \int x f(x) dx = E(X). \tag{3.2.2}$$

Thus, by taking m large enough, any population characteristic, even the density itself, can be obtained to any degree of accuracy.

The Gibbs sampler is first illustrated in the two-variable case. Starting with a pair of random variables (X, Y), the Gibbs sampler generates a sample from $f(x)$ by sampling instead from the conditional distributions $f(x|y)$ and $f(y|x)$, which are often known distributions. This generates a "Gibbs sequence" of random variables

$$Y_0^{'}, X_0^{'}, Y_1^{'}, X_1^{'}, Y_2^{'}, X_2^{'}, ..., Y_k^{'}, X_k^{'}. \tag{3.2.3}$$

The initial value $Y_0^{'} = y_0^{'}$ is specified, and the rest of (3.2.3) is obtained iteratively by alternately generating values from the full conditional distributions

$$\begin{aligned} X_j^{'} &\sim f(x|Y_j^{'} = y_j^{'}), \\ Y_{j+1}^{'} &\sim f(y|X_j^{'} = x_j^{'}). \end{aligned} \tag{3.2.4}$$

Such generation of (3.2.3) is referred to as Gibbs sampling. Under reasonably general conditions, $\lim_{k\to\infty} X_k^{'} \xrightarrow{d} X \sim f(x)$. Thus, for k large enough, the final observation in (3.2.3), namely $X_k^{'} = x_k^{'}$, is effectively a sample point from the target distribution $f(x)$.

Example 1. For the following joint distribution of X and Y,

$$f(x, y) \propto \binom{n}{x} y^{x+\alpha+1}(1-y)^{n-x+\beta-1}, \quad x = 0, 1, ..., n \quad 0 \le y \le 1, \tag{3.2.5}$$

suppose we are interested in calculating some characteristics of the marginal distribution $f(x)$ of X. The Gibbs sampler allows us to generate a sample from this marginal as follows. The two full conditional distributions can be easily shown to be

$$\begin{aligned} f(x|y) &\sim \text{Binomial}(n, y) \\ f(y|x) &\sim \text{Beta}(x + \alpha, n - x + \beta). \end{aligned} \tag{3.2.6}$$

If we now apply the iterative scheme (3.2.4) to the distributions (3.2.6), we can generate a sample $X_1, X_2, ..., X_m$ from $f(x)$ and use this sample to estimate any desired characteristic. For the purpose of checking whether Gibbs sampler performs well in this particular case, we can compare the histograms of samples to the exact density function, as the exact marginal density of X can be shown to be

$$f(x) = \binom{n}{x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(x+\beta)\Gamma(n-x+\beta)}{\Gamma(\alpha+\beta+n)}, x = 0, 1, ..., n, \tag{3.2.7}$$

which is a Beta-binomial distribution. Figure 3.1(a) displays histograms of two samples $x_1, ..., x_m$ of size $m = 500$ from the Beta-binomial distribution of (3.2.7) with $n = 16$, $\alpha = 2$, and $\beta = 4$. One of the two samples is obtained through Gibbs Sampling and the other through the exact marginal distribution. Figure 3.1(b) is the density plot for the exact marginal probability and the estimated density based on the Gibbs samples. We will come to this plot again in the next example. The two histograms are very similar, providing evidence that the Gibbs scheme for random number generation is indeed generating (at least approximately) values from the desired marginal distribution.

Example 2. Suppose X and Y have conditional distributions that are both exponential distributions restricted to the interval $(0, B)$, that is

$$\begin{aligned} f(x|y) &\propto ye^{-yx}, 0 < x < B < \infty \\ f(y|x) &\propto xe^{-xy}, 0 < y < B < \infty, \end{aligned} \tag{3.2.8}$$

where B is a known positive constant. The restriction to the interval $(0, B)$ ensures that the marginal $f(x)$ exits. Although the form of this marginal is not easily calculable, by applying the Gibbs sampler to the conditionals in (3.2.8) any characteristic of $f(x)$ can be obtained. Gibbs sampling can be used to estimate the density itself by averaging the final conditional densities from each Gibbs sequence. From (3.2.3), the

(a) (b)

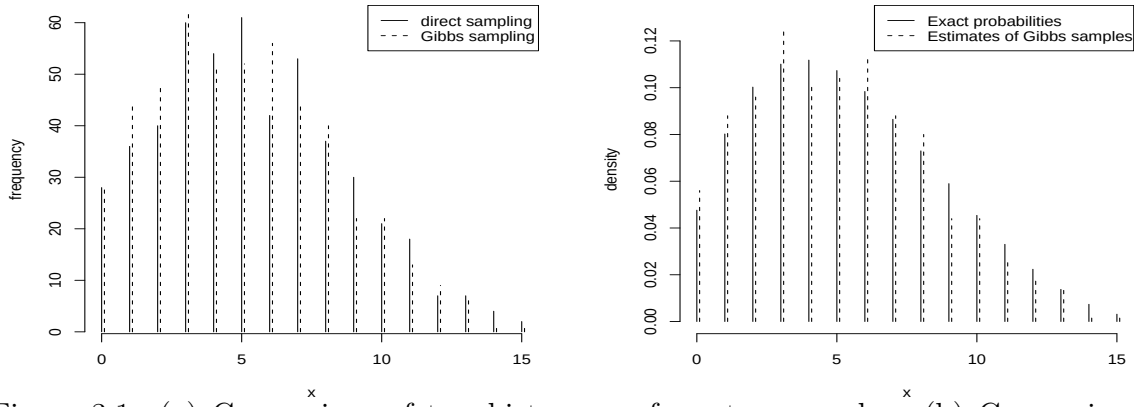


Figure 3.1: (a) Comparison of two histograms from two samples. (b) Comparison between exact probabilities and estimated probabilities

values $X_k^{'} = x_k^{'}$ yields a realization of $X_1, ..., X_m \sim f(x)$ and the values of $Y_k^{'} = y_k^{'}$ yields a realization of $Y_1, ..., Y_m \sim f(y)$. Moreover, the average of the conditional densities $f(x|Y_k^{'} = y_k^{'})$ will be a close approximation to $f(x)$, and we can estimate $f(x)$ with

$$\hat{f}(x) = \frac{1}{m} \sum_{i=1}^{m} f(x|y_i), \tag{3.2.9}$$

where $y_1, ..., y_m$ is the sequence of realized values of final Y observations from each Gibbs sequence. The theory behind the calculation in (3.2.9) is that the expected value of the conditional density is

$$E(f(x|Y)) = \int f(x|y)f(y)dy = f(x). \tag{3.2.10}$$

It is clear that the calculation in (3.2.9) is an approximation to (3.2.10), since $y_1, ..., y_m$ approximate a sample from $f(y)$. This estimating method can also be used in discrete distributions, such as the Beta-binomial of Example 1. Using the observations generated to construct Figure 3.1(a), we can, analogous to (3.2.9), estimate the marginal probabilities of X using

$$\hat{P}(X = x) = \frac{1}{m} \sum_{i=1}^{m} P(X = x|Y_i = y_i).$$

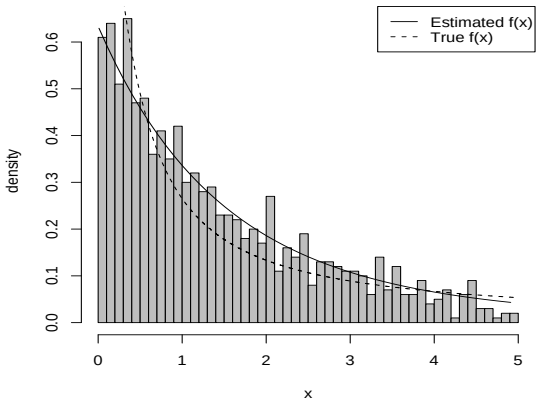


Figure 3.2: Histogram for x from the pair of conditional truncated exponential distributions.

Figure 3.1(b) displays the estimated and exact marginal probabilities. In Figure 3.2 we display a histogram of a sample of size $m = 1000$ from $f(x)$ with $B = 5$ obtained by using the final values from Gibbs sequences. The solid line is the estimate of the marginal density obtained from (3.2.9) and the dashed line is the true marginal density.

Note that in the case where the joint distribution $f(x, y)$ can be calculated the Gibbs sampler may not be needed. However, as was shown in the second example, Gibbs sampling is necessary in situations where $f(x, y)$, $f(x)$, or $f(y)$ cannot be easily calculated.

3.2.2 A Simple Convergence Proof

The two examples demonstrate empirically that the Gibbs sampler works. However, without any formal proof, it is not immediately obvious that X_k' sequences in (3.2.3) converge to $X \sim f(x)$. We now demonstrate the convergence for the simple

case of a 2×2 table with multinomial sampling.

Suppose X and Y are each (marginally) Bernoulli random variables with joint distribution

		X	
		0	1
Y	0	p_1	p_2
	1	p_3	p_4

where $p_i \geq 0, \quad p_1 + p_2 + p_3 + p_4 = 1$. The joint probability mass function (pmf) $f_{x,y}$ is then

$$\begin{bmatrix} f_{x,y}(0,0) & f_{x,y}(1,0) \\ f_{x,y}(0,1) & f_{x,y}(1,1) \end{bmatrix} = \begin{bmatrix} p_1 & p_2 \\ p_3 & p_4 \end{bmatrix}.$$

For this distribution, the marginal pmf of x is given by

$$f_x = [f_x(0), \quad f_x(1)] = [p_1 + p_3, \quad p_2 + p_4], \tag{3.2.11}$$

which is a Bernoulli distribution with success ($x = 1$) probability $p_2 + p_4$.

The conditional distributions of $X|Y = y$ and $Y|X = x$ are also straightforward to find. For example, the distribution of $X|Y = 1 \sim \text{Bernoulli}(\frac{p_4}{p_3+p_4})$. All of the conditional probabilities can be expressed in two matrices,

$$A_{y|x} = \begin{bmatrix} \frac{p_1}{p_1+p_3} & \frac{p_3}{p_1+p_3} \\ \frac{p_2}{p_2+p_4} & \frac{p_4}{p_2+p_4} \end{bmatrix} = \begin{bmatrix} f_{y|x}(0|0) & f_{y|x}(1|0) \\ f_{y|x}(0|1) & f_{y|x}(1|1) \end{bmatrix}$$

and

$$A_{x|y} = \begin{bmatrix} \frac{p_1}{p_1+p_2} & \frac{p_2}{p_1+p_2} \\ \frac{p_3}{p_3+p_4} & \frac{p_4}{p_3+p_4} \end{bmatrix} = \begin{bmatrix} f_{x|y}(0|0) & f_{x|y}(1|0) \\ f_{x|y}(0|1) & f_{x|y}(1|1) \end{bmatrix},$$

which specify the conditional distributions of $Y|X$ and $X|Y$, respectively.

Given these conditional distributions, the iterative sampling scheme applied to this distribution yields (3.2.3) as a sequence of $0'$s and $1'$s. The matrices $A_{x|y}$ and $A_{y|x}$ can be thought of as transition matrices giving the probabilities of getting to x states from y states and vice versa, that is, $P(Y = y|X = x) = P(x, y)$, the probability of going from state x to state y.

If we are only interested in generating from the marginal distribution of X, we are mainly concerned with the $X^{'}$ sequence from (3.2.3). To go from $X_0^{'} \rightarrow X_1^{'}$ we need to go through $Y_1^{'}$, so the iteration sequence is $X_0^{'} \rightarrow Y_1^{'} \rightarrow X_1^{'}$, and $X_0^{'} \rightarrow X_1^{'}$ forms a Markov chain with transition probability

$$P(X_1^{'} = x_1|X_0^{'} = x_0) = \sum_y P(X_1^{'} = x_1|Y_1^{'} = y) \times P(Y_1^{'} = y|x_0^{'} = x_0). \qquad (3.2.12)$$

The transition probability matrix of the $X^{'}$ sequence, $A_{x|x}$ is given by

$$A_{x|x} = A_{y|x}A_{x|y},$$

and now we can easily calculate the probability distribution of any $X_k^{'}$ in the sequence. That is, the transition matrix that gives $P(X_k^{'} = x_k|X_0^{'} = x_0)$ is $(A_{x|x})^k$. Furthermore, if we write $f_k = [f_k(0) \quad f_k(1)]$ to denote the marginal probability distribution of $X_k^{'}$, then for any k,

$$f_k = f_0 A_{x|x}^k = (f_0 A_{x|x}^{k-1})A_{x|x} = f_{k-1}A_{x|x}. \qquad (3.2.13)$$

This is exactly the same form as in (3.1.5). If all the entries of $A_{x|x}$ are positive, then (3.2.13) implies that for any initial probability f_0, as $k \rightarrow \infty, f_k$ converges to the unique distribution f that is a stationary distribution of (3.2.13), and satisfies

$$fA_{x|x} = f. \qquad (3.2.14)$$

Thus, if the Gibbs sequence converges, then f satisfying (3.2.14) must be the marginal distribution of X. It is straightforward to check that (3.2.14) is satisfied by f_x of

(3.2.11), that is,

$$f_x A_{x|x} = f_x A_{y|x} A_{x|y} = f_x.$$

As $k \to \infty$, the distribution of $X_k^{'}$ gets closer to f_x. So if we take a large enough value of k when applying the iteration scheme (3.2.3), we can assume that the distribution of $X_k^{'}$ is approximately f_x. Moreover, the larger the value of k, the better the approximation.

The algebra for the 2×2 case immediately works for any $n \times n$ joint distribution of X's and Y's. We can analogously define the $n \times n$ transition matrix $A_{x|x}$ whose stationary distribution will be the marginal distribution of X. In a continuous state space for X and Y, with suitable assumptions, all of the theory still goes through. So the Gibbs sampler still produces a sample from the marginal distribution of X, though the technical details become far more complicated.

3.2.3 Gibbs Sampling Schema in Multivariate Cases

Assume that the distribution of interest is $\pi(\boldsymbol{\theta})$ where $\boldsymbol{\theta} = (\theta_1, ..., \theta_d)^{'}$. Then the full conditional distributions $\pi_i(\theta_i) = \pi(\theta_i|\boldsymbol{\theta_{-i}}), i = 1, ..., d$ are available, where $\theta_{-i} = (\theta_1, \theta_2, ..., \theta_{i-1}, \theta_{i+1}, ..., \theta_d)$. The Gibbs sampling can be described in the following way:

1. Initialize the iteration counter of the chain $j = 1$ and set initial values $\boldsymbol{\theta^{(0)}} = (\theta_1^{(0)}, ..., \theta_d^{(0)})'$;

2. Obtain a new value $\boldsymbol{\theta^{(j)}} = (\theta_1^{(j)}, ..., \theta_d^{(j)})'$ from $\theta^{(j-1)}$ through successive generation of values
$\theta_1^{(j)} \sim \pi(\theta_1|\theta_2^{(j-1)}, ..., \theta_d^{(j-1)})$,
$\theta_2^{(j)} \sim \pi(\theta_2|\theta_1^{(j)}, \theta_3^{(j-1)} ..., \theta_d^{(j-1)})$,
...
$\theta_d^{(j)} \sim \pi(\theta_d|\theta_1^{(j)}, ..., \theta_{d-1}^{(j)})$;

3. Change counter j to $j+1$ and return to step 2 until convergence is reached.

When convergence is reached, the resulting value $\boldsymbol{\theta}^{(j)}$ is a draw from the target distribution π. The ergodic theorem ensures that empirical averages from the chain converge to the corresponding expectations under $\pi(\boldsymbol{\theta})$ (assuming these are finite). We can then base inference on these samples taken after certain burn-in iterations.

3.3 Metropolis-Hastings Algorithm

The Metropolis-Hastings (M-H) algorithm is a sampling scheme that simulates a (ergodic) Markov chain such that its stationary distribution coincides with the target distribution using a proposal density and a method for rejecting proposed moves. In recent years this Markov chain Monte Carlo (MCMC) method has been widely used to simulate complex, nonstandard multivariate distributions. This algorithm is extremely versatile and gives rise to the Gibbs sampler as a special case. The implementation of our model also heavily relies on this method. Therefore, in this section we provide background on this algorithm to prepare for its implementation in Chapter 5. A paper by Chib and Greenberg [58] provides a detailed introductory exposition of the Metropolis-Hastings algorithm. The following sections recapitulate some of the highlights in their paper, starting with an introduction on acceptance-rejection sampling, then the M-H algorithm, and ending with examples illustrating the application of the M-H algorithm by examples.

3.3.1 Acceptance-Rejection Sampling

The M-H algorithm employs the acceptance-rejection (A-R) sampling techniques which we briefly describe here. Suppose we want to generate samples from the continuous *target density* $\pi(x) = f(x)K$, where $x \in R^d$, $f(x)$ is the unnormalized density, and K is the (possibly unknown) normalizing constant. Let $h(x)$ be a density that

can be easily simulated from, by some known method, and suppose there is a known constant C such that $f(x) \leq Ch(x)$ for all x. Then, the following steps are repeated to obtain a random observation from $\pi(\cdot)$:

1. Generate a candidate X^* from $h(\cdot)$ and a value μ from uniform$(0, 1)$;

2. If $\mu \leq \frac{f(X^*)}{Ch(X^*)}$ then return $X = X^*$; Otherwise go to step 1.

It is easily shown that the accepted value x is a random variate from $\pi(\cdot)$. For this method to be efficient, C must be carefully selected to optimize the acceptance rate. Because the expected number of iterations of step 1 and 2 to obtain a draw is given by C^{-1}, the rejection method is optimized by setting

$$C = \sup_x \frac{f(x)}{h(x)}.$$

3.3.2 Metropolis-Hastings Algorithm

In Section 3.1, we discussed that a major concern of Markov chain theory is to determine conditions under which there exists an invariant distribution π^* and conditions under which iterations of the transition kernel converge to the invariant distribution. It has been proved that if we can find a transition kernel with reversibility then the Markov chain converges to the invariant distribution as the number of iterations $n \to \infty$. We now show how the M-H algorithm finds a $p(x, y)$ with the reversibility such that

$$\pi(x)p(x, y) = \pi(y)p(y, x). \tag{3.3.1}$$

As in the A-R method, suppose we have a density that can generate candidates. Since the generated sequence is a Markov chain, the density depends on the current state of the process. Let $q(x, y)$ denote the *candidate-generating density*, where $\int q(x, y)dy = 1$. If it happens that $q(x, y)$ itself satisfies the reversibility condition in

(3.3.1) for all x, y, then our search for function $p(x, y)$ is over. But most likely it will be

$$\pi(x)q(x, y) > \pi(y)q(y, x). \tag{3.3.2}$$

We need to correct this condition to reduce the probability of move from x to y by introducing a probability $\alpha(x, y) < 1$ that the move is made. The probability $\alpha(x, y)$ is referred to as the *acceptance probability*. If the move is not made, the process again returns x as a value from the target distribution. Thus transitions from x to y ($y \neq x$) are modified to

$$P_{MH}(x, y) = q(x, y)\alpha(x, y),$$

where the upper limit of $\alpha(x, y)$ is 1. When the movement from y to x is not made often enough, we should therefore define $\alpha(y, x)$ to be as large as possible (say 1). The purpose of introducing $\alpha(x, y)$ is to make $P_{MH}(x, y)$ satisfy the reversibility condition (see Section 3.1.4), that is

$$\begin{aligned} \pi(x)q(x, y)\alpha(x, y) &= \pi(y)q(y, x)\alpha(y, x) \\ &= \pi(y)q(y, x). \end{aligned} \tag{3.3.3}$$

This yields $\alpha(x, y) = \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}$. On the other hand, if the inequality in (3.3.2) is reversed, we set $\alpha(x, y) = 1$ and derive $\alpha(y, x)$ as above. Hence, we should define the acceptance probability as follows in order for P_{MH} to be reversible,

$$\begin{aligned} \alpha(x, y) &= \min[\frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}, 1], \text{ if } \pi(x)q(x, y) > 0 \\ &= 1, \qquad\qquad \text{otherwise.} \end{aligned} \tag{3.3.4}$$

To complete the definition of the transition kernel for the Metropolis-Hastings chain, we must consider the probability of the process when the x does not move. That is

$$P_x = 1 - \int_{R^d} q(x, y)\alpha(x, y)dy.$$

Consequently, the transition kernel of the M-H chain, denoted by $P_{MH}(x, dy)$, is given by

$$P_{MH}(x, dy) = q(x, y)\alpha(x, y)dy + P_x\delta_x(dy),$$

where $\delta_x(dy) = 1$ if $x \in dy$ and 0 otherwise. Because $P_{MH}(x, y)$ is reversible by construction, M-H kernel has $\pi(x)$ as its invariant density.

We now make the following remarks about the algorithm. First, the calculation of $\alpha(x, y)$ does not require knowledge of the normalizing constant of equilibrium distribution $\pi(\cdot)$ because it appears both in the numerator and denominator of the acceptance probability. Second, the chain does not necessarily change its state at each iteration. Third, if the candidate-generating density (for example, normal density) is symmetric, then the acceptance probability reduces to $\frac{\pi(y)}{\pi(x)}$.

The simulation of a draw from π can be set up as follows:

1. Initialize the iteration counter $j = 1$ and set an arbitrary initial value $x^{(0)}$.

2. Move the chain to a new value x^* generated from the density $q(x^{(j-1)}, x^*)$.

3. Evaluate the acceptance probability of the move $\alpha(x^{(j-1)}, x^*)$ given by (3.3.4). Generate an independent uniform quantity μ. If $\mu \leq \alpha$, the move is accepted and $x^{(j)} = x^*$. If it is rejected, $x^{(j)} = x^{(j-1)}$.

4. Change the counter from j to $j + 1$ and return to step 2 until approximate convergence is reached.

In practice, the Metropolis-Hastings algorithms requires a tuning phase to achieve the optimal acceptance rate. The value of the tunning parameters, spread and scale, of the candidate distribution $q(x, \cdot)$ affects the acceptance rate and sample space

that is traversed. Consider a random walk chain where the candidate distribution is a normal distribution centered at the current value of the chain with variance ϵ. A large value for the variance allows a move that is very distant from the current value, but the proposed value is likely to fall in the tails of the posterior distribution resulting in a very low acceptance rate. On the other hand, a small value for the variance allows only small moves, it may give a high acceptance rate but the chain will takes many iterations to traverse the entire support of the posterior and hence it takes longer time to converge. Autocorrelations across sample values are likely high in both cases. So far, studies on optimal acceptance rate are not conclusive. Most suggested acceptance rates are empirical. For example, Besag et al. [59] and other authors give a range of acceptance rate of between 0.2 to 0.5. In a specific theoretical context, however Gelman et al. [60] obtained an optimal acceptance rate of 0.24 for high-dimensional problems with both target density and candidate density being normal. Following these guiding rules, we can run the algorithm for relatively small number of iterations in a preliminary tunning phase and determine the acceptance rate. If the acceptance rate is too low or too high, we will decrease or increase the variance of the candidate distribution accordingly. Then we repeat the process until a reasonable acceptance rate is obtained and use this adjusted candidate distribution to run the M-H algorithm and generate simulated data.

3.3.3 Examples

Example 1. Simulating a Bivariate Normal

To illustrate the M-H algorithm we consider the simulation of the bivariate normal distribution $MVN(\boldsymbol{\mu}, \Sigma)$, where $\boldsymbol{\mu} = (1, 2)^{'}$ is the mean vector and $\Sigma = (\sigma_{ij})_{2\times 2}$ is the covariance matrix given by

$$\Sigma = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}.$$

Because of the high correlation the contours of this distribution are elliptical. This distribution can be simulated directly in the Choleski approach by letting $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{P}'\boldsymbol{\mu}$, where $\boldsymbol{\mu} \sim MVN(\mathbf{0}, \mathbf{I_2})$ and $\boldsymbol{P}$ satisfies $\boldsymbol{P}'\boldsymbol{P} = \Sigma$, or via R built-in function $MVRNORM(\cdot)$ (multivariate normal random sample generator).

From the expression of the multivariate normal density, the acceptance probability (for a symmetric candidate-generating density) is

$$\alpha(x, y) = \min \left\{ \frac{\exp[-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{y} - \boldsymbol{\mu})]}{\exp[-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{x} - \boldsymbol{\mu})]}, 1 \right\}, x, y \in R^2. \tag{3.3.5}$$

The following candidate-generating densities can be used and the parameters are adjusted in tuning phase to achieve an acceptance rate of 40% to 50%:

1. Random walk generating density ($y = x + z$), where the increment random variable $z \sim \text{uniform}_2(\boldsymbol{-\delta, \delta})$; that is, the ith component is of z uniform on the interval $(-\delta_i, \delta_i)$. Note that δ_1 and δ_2 control the spread along the the first and second coordinate axis respectively. To avoid excessive moves we let $\delta_1 = .75$ and $\delta_2 = 1$.

2. Random walk generating density ($\boldsymbol{y = x + z}$) with $\boldsymbol{z} \sim MVN(\mathbf{0}, D)$, where $D = \text{diag}(0.6, 0.4)$.

3. The autoregressive generating density $\boldsymbol{y = \mu - (x - \mu) + z}$, where $\mathbf{z}$ is independent uniform with $\delta_1 = 1 = \delta_2$. Thus values of $\boldsymbol{y}$ are obtained by reflecting the current point around $\boldsymbol{\mu}$ and then adding the increment.

Note that the probability of move in the above three cases is given by (3.3.5). In addition, the first two generating densities do not make use of the known value of $\boldsymbol{\mu}$, although the values of the δ_i are related to Σ. Each of these three candidate-generating densities reproduces the shape of the bivariate normal distribution being simulated, but the best result is obtained from the third generating density. To illustrate the characteristics of the output, the left panel of Figure 3.3 contains the

scatter plot of $N = 5,000$ simulated values using R built-in function $MVRNORM(\cdot)$ and the right panel is the scatter plot of $N = 10,000$ simulated values using the third candidate-generating density after $10,000$ burn-in iterations. More observations are taken from the M-H algorithm to make the two plots comparable. The plots of the simulated data generated by the other candidate-generating densities are similar to this and are therefore omitted.

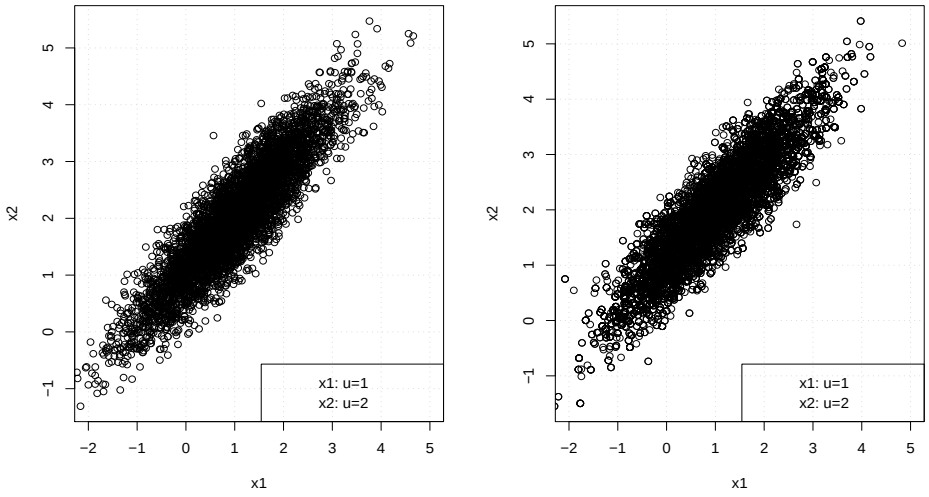


Figure 3.3: Scatter plots of simulated draws: the left is generated by MVRNORM(.) and the right is by M-H method.

Example 2. Simulating a Bayesian Posterior Distribution

This example illustrates the use of the M-H algorithm to sample an intractable distribution that arises in a stationary second-order autoregressive [AR(2)] time series model. The main idea is to generate some observations with known parameters and then apply M-H sampling techniques to obtain the estimates of these parameters. If the M-H algorithm works, the estimates should be close to the true values used to generate the data.

Suppose n observations are simulated from the following AR(2) model

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \epsilon_t, \quad t = 3, .., n, \tag{3.3.6}$$

where $\phi_1 = 1, \phi_2 = -.5$, and $\epsilon_t \sim N(0, 1)$. The values of $\boldsymbol{\phi} = (\phi_1, \phi_2)$ lie in the region $S \subset R^2$ that satisfies the stationary restrictions

$$\phi_1 + \phi_2 < 1; \quad -\phi_1 + \phi_2 < 1; \quad \phi_2 > -1.$$

The likelihood function for (ϕ_1, ϕ_2) based on the observations $\mathbf{Y_n} = (y_1, y_2, .., y_n)'$ is

$$l(\boldsymbol{\phi}, \sigma^2) = \Psi(\boldsymbol{\phi}, \sigma^2) \times (\sigma^2)^{-\frac{n-2}{2}} \times \exp\left[-\frac{1}{2\sigma^2}\sum_{t=3}^{n}(y_t - \boldsymbol{w_t'}\boldsymbol{\phi})^2\right], \tag{3.3.7}$$

where $\boldsymbol{w_t} = (y_{t-1}, y_{t-2})'$,

$$\Psi(\boldsymbol{\phi}, \sigma^2) = (\sigma^2)^{-1}|V^{-1}|^{\frac{1}{2}} \exp\left[-\frac{1}{2\sigma^2}\mathbf{Y_2}'V^{-1}\mathbf{Y_2}\right] \tag{3.3.8}$$

is the density kernel of $\mathbf{Y_2} = (y_1, y_2)'$,

$$V^{-1} = \begin{pmatrix} 1-\phi_2^2 & -\phi_1(1+\phi_2) \\ -\phi_1(1+\phi_2) & 1-\phi_2^2 \end{pmatrix},$$

and the third term in (3.3.7) is proportional to the density of the observations $(y_3, ..., y_n)$ given $\mathbf{Y_2}$.

If the only prior information available is that the process is stationary, then the posterior distribution of the parameters is

$$\pi(\boldsymbol{\phi}, \sigma^2|\mathbf{Y_n}) \propto l(\boldsymbol{\phi}, \sigma^2) I[\boldsymbol{\phi} \in S],$$

where $I[\boldsymbol{\phi} \in S]$ is 1 if $\boldsymbol{\phi} \in S$ and 0 otherwise.

How can this posterior density be simulated? The answer lies in recognizing two facts. First, the blocking strategy is useful for this problem by taking $\boldsymbol{\phi}$ and σ^2 as blocks. Second, from the regression ANOVA decomposition, the exponential term of (3.3.7) is proportional to

$$\exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{\phi}-\hat{\boldsymbol{\phi}})'G(\boldsymbol{\phi}-\hat{\boldsymbol{\phi}})\right],$$

where $\hat{\boldsymbol{\phi}} = G^{-1}\sum_{t=3}^{n}(\boldsymbol{w_t}y_t)$ and $G = \sum_{t=3}^{n}(\boldsymbol{w_t}\boldsymbol{w_t'})$. This is the kernel of the normal density with mean $\hat{\boldsymbol{\phi}}$ and covariance matrix $\sigma^2 G^{-1}$. These observations immediately lead to the following full conditional densities for σ^2 and $\boldsymbol{\phi}$:

$$\pi(\sigma^2|\boldsymbol{\phi}, \boldsymbol{Y_n}) = \text{inv}-\text{gamma}\left\{\frac{1}{2}(n-2), \frac{1}{2}(Y_2V^{-1}Y_2 + \sum_{t=3}^{n}(y_t - \boldsymbol{w_t}\boldsymbol{\phi})^2)\right\}$$

$$\pi(\boldsymbol{\phi}|\boldsymbol{Y_n}, \sigma^2) \propto \Psi(\boldsymbol{\phi}, \sigma^2) \times \{f_{nor}(\boldsymbol{\phi}|\hat{\boldsymbol{\phi}}, \sigma^2 G^{-1})I[\boldsymbol{\phi} \in S]\} \tag{3.3.9}$$

where f_{nor} is the normal density function.

A sample of draws from the joint density $\pi(\boldsymbol{\phi}, \sigma^2|\boldsymbol{Y_n})$ can now be obtained by successively sampling $\boldsymbol{\phi}$ from $\pi(\boldsymbol{\phi}|\boldsymbol{Y_n}, \sigma^2)$, and given this value of $\boldsymbol{\phi}$, simulating σ^2 from $\pi(\sigma^2|\boldsymbol{Y_n}, \boldsymbol{\phi})$. The latter simulation is simply a Gibbs sampling step. For the former, we can apply M-H algorithm by choosing a candidate-generating density. One choice is to exploit the known form of $\pi(\cdot)$ to specify a candidate-generating density. For example, if $\pi(t)$ can be written as $\pi(t) \propto \psi(t)h(t)$, where $h(t)$ is density that can be sampled and $\psi(t)$ is uniformly bounded, then set candidate-generating density $q(x, y) = h(y)$. In this case, the probability of acceptance requires only the computation of the ψ function and is given by

$$\alpha(x, y) = \min\{\frac{\psi(y)}{\psi(x)}, 1\}.$$

Because it can be shown that $|V^{-1}|^{\frac{1}{2}}$ is bounded for all values of $\boldsymbol{\phi}$ in the stationary region, we use the normal density in curly braces of (3.3.9) as the candidate-generating

density. Then, the value of $\boldsymbol{\phi}$ at the $(j+1)^{th}$ iteration is simulated as: draw a candidate $\boldsymbol{\phi}^{(j+1)}$ from $N(\hat{\boldsymbol{\phi}}, \sigma^{2(j)}G^{-1})$; if it satisfies stationarity, move to this point with probability

$$\min\{\frac{\Psi(\boldsymbol{\phi}^{(j+1)}, \sigma^{2(j)})}{\Psi(\boldsymbol{\phi}^{(j)}, \sigma^{2(j)})}, 1\}$$

and otherwise set $\boldsymbol{\phi}^{(j+1)} = \boldsymbol{\phi}^{(j)}$, where $\Psi(.,.)$ is defined in (3.3.8).

Table 3.1: Summaries of the posterior distribution for simulated AR(2) model

		Posterior					
Param.	True	Mean	SD	Median	Lower	Upper	Corr.
ϕ_1	1.000	0.941	0.090	0.944	0.765	1.117	0.010
ϕ_2	-0.500	-0.428	0.092	-0.430	-0.608	-0.248	0.012
σ^2	1.000	0.932	0.137	0.918	0.663	1.201	0.028

The posterior distributions are summarized in Table 3.1 for a sample of 10000 after the 10000 iterations of burn-in. The estimated means are close to the true values and the sample first-order serial correlation in the simulated values is low and not of concern. From these results it is clear that the M-H algorithm has accurately produced a posterior distribution concentrated on the values that generate the data.

3.4 MCMC Convergence Diagnostics

The validity of MCMC methods requires that a sampled value is from the distribution of interest π. This is only obtained when the number of iterations of the chain approaches infinity. In practice this is not attainable. Instead a value obtained at a sufficiently large number of iteration is treated as being drawn from π. The question is how large this iteration number should be. There is no simple answer to this question and most efforts have been devoted to studying as close as possible the convergence characteristics of the MCMC chain.

3.4.1 MCMC Convergence

Convergence refers to the idea that the Gibbs Sampler or other MCMC technique that we choose will eventually reach a stationary distribution. From this point on it stays in this distribution and moves about (or *mixes*) throughout the subspace forever.

There are two main ways to approach the study of convergence. The first one is more theoretical and tries to measure distances and establish bounds on distribution functions generated from a chain [61]. In particular, one can study the total variation distance between the distribution of the chain at iteration j and the limiting distribution P. Special aspects derived from the probabilistic structure of the chain can also be studied. At the moment, however, the results from the study have had little impact on practical work. This is still an open problem.

We can also approach the study of convergence of the chain from a statistical perspective by analyzing the properties of the observed output from the chain. Although this is an empirical as opposed to a theoretical treatment, it is obviously more practical. The difficulty with this approach is that we can only identify if the chain has not converged but we can never guarantee any specified level of precision. Some of the statistical methods include graphical checks of convergence, time series analysis, and Gelman-Rubin multiple chains diagnostic [61].

3.4.2 Informal Convergence Monitors

Gelfand and Smith [62] suggested a few informal checks of convergence based on graphical techniques. After m iterations in n parallel chains, a histogram of the n values of the m^{th} iterations of a given function of $\boldsymbol{\theta}$ can be plotted. This function can be one of the components of $\boldsymbol{\theta}$ and the histogram may be smoothed if desired. The procedure is repeated after a further k times are obtained in the chains. Typically

we take k between 10 and 50. Convergence is accepted if the histograms cannot be distinguished.

We can apply this idea to the graphical representations of the simulated values of a few chosen (transformations of) parameters. The resulting plots provide a rough indication of stationarity behaviour when the sequence of values tends to concentrate around the same pattern. This visual impression can be reinforced when chains started at different values oscillate in the same region.

Another intuitive and easily implemented diagnostic tool is a trace plot (or history plot) which plots the parameter value at time t against the iteration number. If the Markov chain sampler has converged, the trace plot will move around the mode of the distribution. A clear sign of non-convergence with a trace plot occurs when we observe some trending in the sample trace. The problem with trace plots is that sometimes it may appear that the chains have converged, but the chain is just trapped (for a finite time) in a local region (a mode) rather than exploring the full posterior. These techniques must be used with caution and should always be accompanied by some theoretical reasoning. There are many chains that exhibit every indication of convergence without actually achieving it. If the autocorrelation among the parameters of interest is high, then a trace plot will be a poor diagnostic tool for convergence.

3.4.3 Gelman-Rubin Multiple Chain Diagnostic

Another simple method to check for convergence is to use parallel chains started at different points of the parameter space. Use of multiple chains would reveal chains that are trapped in regions around local modes. Also, slow convergence may give rise to metastable behaviour of the chains and this can be easily detected through parallel chains.

Gelman and Rubin [63] explored the observation that the chain trajectories should be the same after approximate convergence using analysis of variance techniques. The basic idea is to test whether dispersion within chains is larger than dispersion between chains. Consider m parallel chains and a real function $\psi = t(\boldsymbol{\theta})$. There are m trajectories $\psi_i^{(1)}, \psi_i^{(2)}, ..., \psi_i^{(n)}$, $i = 1, ..., m$, for ψ. The variances between chains B and within chains W are given by

$$B = \frac{n}{m-1}\sum_{i=1}^{m}(\bar{\psi}_i - \bar{\psi})^2 \text{ and } W = \frac{1}{m(n-1)}\sum_{i=1}^{m}\sum_{j=1}^{n}(\psi_i^{(j)} - \bar{\psi}_i)^2$$

where $\bar{\psi}_i$ is the average of observations of chain $i, i = 1, ..., m$, and $\bar{\psi}$ is the average of these averages. Under convergence of all chains, the mn values are drawn from the posterior. Hence the variance of ψ (σ_ψ^2), can be consistently estimated by W, B, and the weighted average $\hat{\sigma}_\psi^2 = (1 - \frac{1}{n})W + \frac{1}{n}B$.

Prior to the convergence of all chains, W underestimates σ_ψ^2 because not all chains have fully explored the target distribution. The value of B, on the other hand overestimates σ_ψ^2 because the starting points are over-dispersed relative to the target. Following this reasoning, an indicator of convergence can be formed by the estimator of potential scale reduction given by

$$\hat{R} = \sqrt{\frac{\hat{\sigma}_\psi^2}{W}}.$$

Obviously, $\hat{R}$ is always larger than 1. But as $n \to \infty$, both estimators converge to σ_ψ^2 by the ergodic theorem and $\hat{R} \to 1$. Convergence can be evaluated by the proximity of $\hat{R}$ to 1. Gelman [64] suggested accepting convergence when the value of $\hat{R}$ is below 1.2.

Chapter 4

Hierarchical Bayesian AFT Spatial Model

We propose a hierarchical Bayesian approach to extend AFT models to include spatial random effects. There are three major benefits afforded by the Bayesian paradigm. First, it is able to account for uncertainty in parameter estimates when evaluating prediction uncertainty. Second, Bayesian inference does not rely on the asymptotic properties of estimators because results are based entirely on simulated samples from posterior distributions. Third, the Bayesian paradigm allows the incorporation of prior information that may not be as easily incorporated in a frequentist approach. Another benefit is the process decomposition model. A Bayesian framework provides a parametric method for decomposing the observed process into a trend surface (fixed effect of regression components), a baseline error process (error distribution of log survival time), and a multivariate conditional autoregressive process (spatial random effect and space-varying treatment effect). This process decomposition will be illustrated in Section 4.1, which is followed by a discussion of computation and implementation in Section 4.2. The last section of this chapter presents the simulation study to address model assessment and MCMC convergence.

4.1 Model Specification

Consider AFT modeling for individuals in different geographic regions such as local health units, health service delivery areas, or health authorities. For purposes of illustration, suppose there are J local health units and n_j patients in the j^{th} local health unit (LHU). Let t_{ij} be the time to event (e.g, death) of the i^{th} person in the j^{th} LHU, where $i = 1, 2, ..., n_j$ and $j = 1, 2, ..., J$. Thus, each individual can

be unambiguously referred to as the (i, j)-th individual and $\mathbf{x_{ij}}$ is the associated vector of covariates. We will explore various ways of modeling spatial random effects incorporated into different parametric AFT models.

4.1.1 Univariate Spatial CAR Model

To extend the log linear AFT model (2.2.6), we include a spatial random effect ϕ_j for the j^{th} LHU (geographical unit) in the right-hand side,

$$Y_{ij} = \log(t_{ij}) = \mu + \mathbf{b}^{'}\mathbf{x_{ij}} + \phi_j + \sigma w_{ij}, \tag{4.1.1}$$

where μ is the intercept term, $\mathbf{b}^{'}$ is a vector of coefficients corresponding to covariates $\mathbf{x_{ij}}$, ϕ_j is the spatial random effect for the j^{th} LHU and σ is the scale parameter for the error distribution of w_{ij}. Further, let λ be the sum of all the terms except the error term in (4.1.1), i.e.

$$\lambda_{ij} = \mu + \mathbf{b}^{'}\mathbf{x_{ij}} + \phi_j. \tag{4.1.2}$$

If we assume w_{ij} (error term) follows a logistic distribution, the corresponding survival function and probability density function for the survival time t_{ij} based on this log linear representation are, respectively,

$$\begin{aligned} S(t_{ij}; \lambda_{ij}, \sigma) &= \frac{1}{1 + \exp[-\frac{\lambda_{ij}}{\sigma}] t_{ij}^{\frac{1}{\sigma}}}, \\ f(t_{ij}; \lambda_{ij}, \sigma) &= \frac{\frac{1}{\sigma}\exp[-\frac{\lambda_{ij}}{\sigma}] t_{ij}^{(\frac{1}{\sigma}-1)}}{(1 + \exp[-\frac{\lambda_{ij}}{\sigma}] t_{ij}^{\frac{1}{\sigma}})^2} \end{aligned} \tag{4.1.3}$$

Letting δ_{ij} be the death indicator for the (i, j)-th individual, following the likelihood function for right censored data in (2.1.3), we can write down the likelihood as follows

$$L(\lambda_{ij}, \sigma; t_{ij}, \delta_{ij}) = \prod_{j=1}^{J}\prod_{i=1}^{n_j}(f(t_{ij}; \lambda_{ij}, \sigma))^{\delta_{ij}}(S(t_{ij}; \lambda_{ij}, \sigma))^{1-\delta_{ij}}, \tag{4.1.4}$$

where for the log-logistic model $f(t_{ij};\lambda_{ij},\sigma)$ and $S(t_{ij};\lambda_{ij},\sigma)$ are specified in (4.1.3). For different error distributions, (4.1.4) is still valid but the density and survival function in (4.1.3) will be different. Here we assume that the censoring is non-informative - an assumption we make in all subsequent modeling. AFT models based on Weibull and log-normal distributions are also popular. The Weibull AFT model survival and density function are, respectively,

$$S(t_{ij};\lambda_{ij},\sigma) = \exp[-\exp(-\frac{\lambda_{ij}}{\sigma})t_{ij}^{\frac{1}{\sigma}}],$$

$$f(t_{ij};\lambda_{ij},\sigma) = \frac{1}{\sigma}\exp(-\frac{\lambda_{ij}}{\sigma})t_{ij}^{(\frac{1}{\sigma}-1)}\exp[-\exp(-\frac{\lambda_{ij}}{\sigma})t_{ij}^{\frac{1}{\sigma}}] \tag{4.1.5}$$

and the log-normal AFT model has survival and density function

$$S(t_{ij};\lambda_{ij},\sigma) = 1-\Phi(\frac{\log(t_{ij})-\lambda_{ij}}{\sigma}),$$

$$f(t_{ij};\lambda_{ij},\sigma) = \frac{\exp[-\frac{1}{2}(\frac{\log(t_{ij})-\lambda_{ij}}{\sigma})^2]}{t_{ij}\sqrt{2\pi}\sigma}. \tag{4.1.6}$$

A Bayesian hierarchical model is completed by assigning priors to the parameters μ, $\mathbf{b}$, $\boldsymbol{\phi}$ and scale parameter σ of the error distribution. Under the Bayesian hierarchical framework, vague or non-informative prior distributions that play a minimal role in the posterior distribution are adopted. For example, often we adopt normal priors with large variance for μ and the regression coefficients $\mathbf{b}$, and an inverse gamma (inv-gamma) prior for σ. A normal distribution with large variance is vague and the inverse gamma distribution with large mean and variance (or ∞) is vague too. A CAR prior presented in Section 2.5.3 is used to model spatial random effects and approriate priors will be further assigned to the hyper-parameters associated with CAR priors. The hierarchy of assigning priors can be illustrated as follows. The priors at the first level are specified as:

Prior for intercept μ

$$p(\mu) : \mu \sim N(0,\epsilon^2)$$

Prior for regression coefficients $\mathbf{b}$

$$p(\mathbf{b}) : \mathbf{b} \sim MVN(\mathbf{0}, \epsilon^2 \mathbf{I})$$

Prior for σ

$$p(\sigma) : \sigma \sim \text{inv-gamma}(a_0, b_0)$$

CAR (ρ, τ^2) *prior for* $\boldsymbol{\phi}$

$$p(\phi_1, ..., \phi_J) \propto \exp\left\{-\frac{1}{2\tau^2}\boldsymbol{\phi}'(D_W - \rho W)\boldsymbol{\phi}\right\},$$

where ρ is the coefficient of spatial association, τ^2 is the variance parameter associated with the full conditional distribution, D_W is a diagonal matrix defined in Section 2.5.3, and W is the adjacency matrix. This is a proper CAR model discussed in Section 2.5.3 with $E(\boldsymbol{\phi}) = \mathbf{0}$. The second level priors of the hyper-parameters τ^2 and ρ take forms

Prior for the conditional variance of $\boldsymbol{\phi}$

$$p(\tau^2) : \tau^2 \sim \text{inv-gamma}(c_0, d_0)$$

Prior for the spatial association of $\boldsymbol{\phi}$

$$p(\rho) : \rho \sim \text{uniform}(0, 1)$$

The p.d.f of $x \sim \text{inv-gamma}(\alpha, \beta)$ is

$$f(x) = \frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)} e^{-\frac{\beta}{x}},$$

where $E(x) = \frac{\beta}{\alpha - 1}$ when $\alpha > 1$ and $Var(x) = \frac{\beta^2}{(\alpha-1)^2(\alpha-2)}$ when $\alpha > 2$. For the hyper-parameters in the above priors, we choose $\epsilon^2 = 10,000$ to make the variance of the priors for μ and $\mathbf{b}$ so large that the posteriors will be dominated by the likelihood. Also, we set a_0, b_0, c_0, d_0 to 0.01 so that the expectation and variance are ∞. As we do not have much information on ρ except that it lies between 0 and 1, a uniform$(0, 1)$ prior is used.

4.1.2 Two Independent Spatial CAR Models

We can extend the univariate CAR model to include both spatial random effect and space-varying regression coefficients at the same time. In particular, we are interested in determining whether the impact of revascularization therapy is spatially-varying. Employing two i.i.d. spatial CAR models is a straightforward way to model these two spatial processes. The AFT model now takes the form

$$Y_{ij} = \log(t_{ij}) = \mu + \mathbf{b}^{'}\mathbf{x_{ij}} + \phi_{1j} + \phi_{2j}R_{ij} + \sigma w_{ij}, \tag{4.1.7}$$

where $\mathbf{x_{ij}}$ is a vector of covariates including the treatment R_{ij} which is the indicator variable of treatment (revascularization), ϕ_{1j} is the spatial random effect for LHU j ($E(\phi_{1j}) = 0$), and ϕ_{2j} is the coefficient of treatment effect for LHU j ($E(\phi_{2j}) = 0, j = 1, ..., J$). In this formulation, total treatment effect has been split into two parts: (1) the fixed effect that appears as one of the coefficients of covariates, and (2) random effect (ϕ_{2j}) that is associated with the jth LHU. Hence, λ is still defined as the sum of all the terms except the error term in (4.1.7), which now becomes

$$\lambda_{ij} = \mu + \mathbf{b}^{'}\mathbf{x_{ij}} + \phi_{1j} + \phi_{2j}R_{ij}. \tag{4.1.8}$$

The formulation of the likelihood is exactly the same as in (4.1.4). The prior specifications are also the same as in the univariate case, except here we have two CAR models, $\boldsymbol{\phi_1} \sim CAR(\rho_1, \tau_1^2)$ and $\boldsymbol{\phi_2} \sim CAR(\rho_2, \tau_2^2)$, which are assumed to be independent.

4.2 Computation and Implementation

The model specifications feature spatial association and space-varying coefficients for the effect of revascularization. We illustrate the implementation of the parametric AFT model with different spatial modelings.

4.2.1 Univariate Spatial CAR Model Implementation

One big advantage of using MCMC method to draw a sample from the posterior distribution is that we only need to know the kernel of the posterior distribution, which is the product of all the priors and the likelihood. Following our earlier notations in (4.1.4) and using a proper CAR prior with associated density $p(\rho, \tau^2)$, the joint posterior distribution of the parameters is, up to a normalizing constant,

$$\begin{aligned}
&\pi(\mu, \mathbf{b}, \sigma, \phi_j, \rho, \tau^2 | \{t_{ij}\}, \{\mathbf{x_{ij}}\}, \{\delta_{ij}\}) \\
&\propto L(\lambda_{ij}, \sigma; t_{ij}, \delta_{ij}) p(\mu) p(\mathbf{b}) p(\sigma) p(\rho, \tau^2) p(\rho) p(\tau^2) \\
&\propto \prod_{j=1}^{J} \prod_{i=1}^{n_j} (f(t_{ij}; \mathbf{x_{ij}}, \mu, \mathbf{b}, \sigma, \phi_j))^{\delta_{ij}} (S(t_{ij}; \mathbf{x_{ij}}, \mu, \mathbf{b}, \sigma, \phi_j))^{1-\delta_{ij}} \\
&p(\mu) p(\mathbf{b}) p(\sigma) p(\rho, \tau^2) p(\rho) p(\tau^2).
\end{aligned} \tag{4.2.1}$$

Gelfand, Sahu and Carlin [65] suggest hierarchical centering of spatial random effects to improve MCMC convergence. This leads to the reparametrization of (4.1.2) as:

$$\lambda_{ij} = \mathbf{b}' \mathbf{x_{ij}} + \phi_j, \tag{4.2.2}$$

where the intercept term μ is pushed to the second level so that $E(\phi_j) = \mu, j = 1, .., J$ and it does not directly appear in the likelihood. In other word, the ϕ_j's are centered at μ.

Let $\boldsymbol{\theta}$ be a vector of all the parameters and D be the data. The notation of $\boldsymbol{\theta_{-para.i}}$ means all parameters except parameter i. Let $L(\boldsymbol{\theta}|D)$ be the likelihood function in (4.2.1). We adopt the following MCMC algorithm for sampling the parameters. Assuming that $(\mu, \mathbf{b}, \boldsymbol{\phi}, \sigma, \rho, \tau^2)$ is the current state of the chain, we sample the next state by following:

1. Sampling $\mathbf{b}|\sigma, \boldsymbol{\phi}$*:*

We take an approach of updating each component of $\mathbf{b}$, b_i, one by one. Assume b_i follows the prior distribution $b_i \sim N(0, \eta)$ where η is a large value, say 10^4, which gives a vague prior. Sample $b_i^\star \sim N(b_i, \epsilon)$ and accept this proposal with probability

$$\min\left\{1, \frac{L(b_i^\star|\boldsymbol{\theta}_{-\boldsymbol{b_i^\star}}, D)}{L(b_i|\boldsymbol{\theta}_{-\boldsymbol{b_i}}, D)} \frac{p(b_i^\star, \eta)}{p(b_i, \eta)}\right\},$$

where $p(x, \eta)$ is the prior density evaluated at x and ϵ is a tunning parameter that is adjusted to yield an adequate acceptance rate. We will use ϵ as a tunning parameter for the rest of the updating schema. Note that ϵ varies with parameters. During the tunning phase, we decrease ϵ if the acceptance rate is too low (say < 0.2) and increase ϵ if the acceptance rate is too high (say > 0.5) until an empirically reasonable acceptance rate is obtained.

2. Sampling $\sigma|\mathbf{b}, \boldsymbol{\phi}$*:*

In practice, we work with $\frac{1}{\sigma}$. We choose a vague prior, say $\frac{1}{\sigma} \sim$ gamma$(.01, .01)$ so that $E(\frac{1}{\sigma}) = 1$ and $\text{Var}(\frac{1}{\sigma}) = 100$. Sampling $\frac{1}{\sigma}$ is essentially the same as updating b_i except that we need to take a log transformation of $\frac{1}{\sigma}$ by letting $\nu = \log(\frac{1}{\sigma})$. Draw a ν^* $(\log(\frac{1}{\sigma}^\star))$ from $N(\nu, \epsilon)$ and accept it with probability

$$\min\left\{1, \frac{L(\sigma^\star|\boldsymbol{\theta}_{-\boldsymbol{\sigma^\star}}, D)}{L(\sigma|\boldsymbol{\theta}_{-\boldsymbol{\sigma}}, D)} \frac{p(\frac{1}{\sigma}^\star, \eta)}{p(\frac{1}{\sigma}, \eta)} \frac{\frac{1}{\sigma}^\star}{\frac{1}{\sigma}}\right\},$$

where the term $\frac{1}{\sigma}^\star/\frac{1}{\sigma}$ comes from Jacobian due to the log transformation of $\frac{1}{\sigma}$.

3. Sampling $\boldsymbol{\phi}|\mu, \mathbf{b}, \sigma, \rho, \tau^2$*:*

Let D_j be the data in jth LHU and let $p(\boldsymbol{\phi})$ be the density of $N_J(\boldsymbol{\mu}, \Sigma)$, where $\Sigma = \tau^2(D_W - \rho W)^{-1}$ is the covariance matrix given in Section 2.5.3 for $\boldsymbol{\phi} \sim CAR(\boldsymbol{\mu}, \rho, \tau^2)$, a centered version of $CAR(\rho, \tau^2)$. We use a Metropolis-Hastings random walk algo-

rithm for each $\phi_j, j = 1, ..., J$. Sample $\phi_j^\star \sim N(\phi_j, \epsilon)$ and accept $\phi_j^\star$ with probability

$$\min\left\{1, \frac{L(\phi_j^\star|\boldsymbol{\theta}_{-\boldsymbol{\phi}_j^\star}, D_j)}{L(\phi_j|\boldsymbol{\theta}_{-\boldsymbol{\phi}_j}, D_j)} \frac{MVNPDF(\boldsymbol{\phi}^\star, \mu, \Sigma)}{MVNPDF(\boldsymbol{\phi}, \mu, \Sigma)}\right\}, \text{ for } j = 1, ..., J,$$

where $MVNPDF(\cdot)$ is a built-in function in Matlab for computing the density of the mutilvariate normal distribution. For simplicity of implementation, each ϕ_j shares the same tunning parameter ϵ. During the turning phase, we monitor the maximum and minimum acceptance rates of all ϕ_j decrease or increase ϵ to achieve an optimal acceptance rate.

4. Sampling $\mu|\mathbf{b}, \sigma, \boldsymbol{\phi}$:

Sample a $\mu^\star$ from $N(\mu, \epsilon)$ and accept it with probability

$$\min\left\{1, \frac{MVNPDF(\boldsymbol{\phi}, \mu^\star, \Sigma)}{MVNPDF(\boldsymbol{\phi}, \mu, \Sigma)} \frac{p(\mu^\star, \eta)}{p(\mu, \eta)}\right\}.$$

5. Sampling $\rho|\boldsymbol{\phi}, \tau^2$:

As ρ is a hyper-parameter, sampling ρ does not involve likelihood evaluation. We use a flat prior $\rho \sim$ uniform$(0, 1)$ and a logistic transformation of ρ (say $logit(\rho)$) to map ρ to the whole real line where random walk methods perform well. We sample $logit(\rho^\star) \sim N(logit(\rho), \epsilon)$, compute the $\Sigma^\star$ based on the $\rho^\star$ and accept $\rho^\star$ with probability

$$\min\left\{1, \frac{MVNPDF(\boldsymbol{\phi}, \mu, \Sigma^\star)}{MVNPDF(\boldsymbol{\phi}, \mu, \Sigma)} \frac{\rho^\star(1-\rho^\star)}{\rho(1-\rho)}\right\},$$

where the first part of the acceptance ratio is the ratio of density of $CAR(\boldsymbol{\mu}, \rho^\star, \tau^2)$ and $CAR(\boldsymbol{\mu}, \rho, \tau^2)$, and the second part is the ratio of Jacobian for $\rho^\star$ and ρ from the logistic transformation.

6. Sampling $\tau^2|\boldsymbol{\phi}, \rho$:

Again, τ^2 is a hyper-parameter and we employ the conjugate prior inv-gamma(c_0, d_0)

(say $c_0 = 0.01, d_0 = 0.01$, which results in a vague prior with $E(\tau^2)$ and $Var(\tau^2)$ being ∞). Sampling τ^2 only needs a Gibbs step which is

$$\tau^2 \sim \text{inv-gamma}(A, B),$$

where $A = c_0 + 0.5J$ and $B = d_0 + 0.5(\boldsymbol{\phi} - \boldsymbol{\mu})^{'}(D_W - \rho W)(\boldsymbol{\phi} - \boldsymbol{\mu})$. However, in the real implementation, we find its equivalent, univariate inverse-Wishart(2B,2A), more numerically stable.

In general, when updating each parameter, we need to adjust the tunning parameters of the proposal density to achieve the optimal acceptance rates and then run multiple Markov chains starting at different values for each parameter. We closely monitor the trace plot of each chain and decide how many iterations are necessary for a chain to converge.

4.2.2 Two Independent Spatial CAR Models Implementation

The implementation of two independent CAR models follows directly from the univariate case except that there are $J+2$ extra parameters for the spatial treatment random effect. Such spatial treatment random effect will be centered at the fixed part of the treatment effect (the coefficient of treatment) in (4.1.8). This leads to a modified expression for λ_{ij}:

$$\lambda_{ij} = \mathbf{b}^{'}\mathbf{x_{ij}} + \phi_{1j} + \phi_{2j}R_{ij}, \tag{4.2.3}$$

where the covariate vector $\mathbf{x_{ij}}$ excludes treatment R since the fixed treatment effect (say $b.trt$) has been pushed to the center of ϕ_{2j} (i.e. $\boldsymbol{\phi_2} \sim CAR(\boldsymbol{b.trt}, \rho_2, \tau_2^2)$). With two proper CAR priors $p(\boldsymbol{\phi_i}|\rho_i, \tau_i^2)$ where $i = 1, 2$, the joint posterior distribution of

the parameters is,

$$\pi(\mathbf{b}, \sigma, \phi_{1j}, \mu, \rho_1, \tau_1^2, \phi_{2j}, b.trt, \rho_2, \tau_2^2 | \{t_{ij}\}, \{\mathbf{x_{ij}}\}, \{\delta_{ij}\})$$
$$\propto L(\lambda_{ij}, \sigma; t_{ij}, \delta_{ij}) p(\mathbf{b}) p(\sigma) p(\boldsymbol{\phi_1} | \mu, \rho_1, \tau_1^2) p(\boldsymbol{\phi_2} | b.trt, \rho_2, \tau_2^2) p(\mu) p(\rho_1) p(\tau_1^2) p(b.trt) p(\rho_2) p(\tau_2^2)$$
$$\propto \prod_{j=1}^{J} \prod_{i=1}^{n_j} (f(t_{ij}; \mathbf{x_{ij}}, \mu, \mathbf{b}, \sigma, \phi_{1j}, \phi_{2j}))^{\delta_{ij}} (S(t_{ij}; \mathbf{x_{ij}}, \mu, \mathbf{b}, \sigma, \phi_{1j}, \phi_{2j}))^{1-\delta_{ij}}$$
$$p(\mathbf{b}) p(\sigma) p(\boldsymbol{\phi_1} | \mu, \rho_1, \tau_1^2) p(\boldsymbol{\phi_2} | b.trt, \rho_2, \tau_2^2) p(\mu) p(\rho_1) p(\tau_1^2) p(b.trt) p(\rho_2) p(\tau_2^2), \tag{4.2.4}$$

where $p(\boldsymbol{\phi_1} | \mu, \rho_1, \tau_1^2) \propto \exp\left\{-\frac{1}{2\tau_1^2}(\boldsymbol{\phi_1} - \boldsymbol{\mu})'(D_W - \rho_1 W)(\boldsymbol{\phi_1} - \boldsymbol{\mu})\right\}$,
$p(\boldsymbol{\phi_2} | b.trt, \rho_2, \tau_2^2) \propto \exp\left\{-\frac{1}{2\tau_2^2}(\boldsymbol{\phi_2} - \boldsymbol{b.trt})'(D_W - \rho_2 W)(\boldsymbol{\phi_2} - \boldsymbol{b.trt})\right\}$.

MCMC algorithms similar to that used for sampling the parameters in the univariate CAR model can be applied here.

4.3 Simulation Study on Model Assessment

Before applying the proposed hierarchical Bayesian AFT spatial models to our AMI data, we will test the implementation using simulated data. Two simulation studies are carried out, which serve two different purposes. The first is to examine the estimation accuracy of the MCMC algorithm when implementing the complicated models proposed here. The second is to assess model performance.

Simulation Study on Estimation Accuracy

We use the same neighbourhood structure given by the adjacency matrix W (derived from the map of Quebec 139 LHUs in Figure 1.1) to simulate data when spatial random effects are considered. The centered version of two i.i.d CAR models with a logistic error distribution is used to simulate the data. The model specification is

$$Y_{ij} = \log(t_{ij}) = \mathbf{b}' \mathbf{x_{ij}} + \phi_{1j} + \phi_{2j} R_{ij} + \sigma w_{ij},$$

where covariates included in $\mathbf{x_{ij}}$ are age and gender for the $(ij)^{th}$ individual, and R_{ij} is the treatment indicator, $\boldsymbol{\phi_1}$ and $\boldsymbol{\phi_2}$ are two spatial random effects, and σw_{ij} is the white noise that is assumed to have a logistic$(0, \sigma)$ distribution. For convenience, we simulate 50 observations for each local health unit and the total number of observations are $50 \times 139 = 6950$. Age is generated for all 6950 individuals from a normal distribution with mean 50 and variance 10, and gender (1=Female and 0=Male) and treatment are randomly sampled from 0 and 1 with replacement. This set of covariates is used for all the following data simulations.

Assume the error follows logistic distribution with location parameter 0 and scale parameter σ. We simulated three sets of data from log-logistic AFT with two i.i.d spatial CAR random effects. Take first data set for example, the simulation is carried out as follows: first, the spatial random effects $\boldsymbol{\phi_1}$ centered at $\mu = 5$ is generated from a CAR model with $\tau_1^2 = 0.1$ and $\rho_1 = 0.90$; second, we center space-varying treatment coefficients at $b.trt = 0.90$ and generate $\boldsymbol{\phi_2}$ from a CAR model with $\tau_2^2 = 0.03$ and $\rho_2 = 0.80$; third, we generate the error term σw_{ij} from a logistic$(0, \sigma)$ where $\sigma = 1.2$; we center the covariate age at its mean and set the regression coefficients $b.age = -0.09$ and $b.gender = -0.2$; then the logarithm of survival time $\log(t)$ is easily calculated as

$$\log(t_{ij}) = \mathbf{b}^{'}\mathbf{x_{ij}} + \phi_{1j} + \phi_{2j}R_{ij} + \sigma w_{ij},$$

where $b = (b.age, b.gender)$, $\mathbf{x_{ij}}$=(age and gender for the $(ij)^{th}$ individual), $E(\phi_{1j}) = \mu$ and $E(\phi_{2j}) = b.trt$. By keeping the same values of $\mu = 5, b.age = -0.09, b.gender = -0.20$ but replacing the "true values" of $\tau_1^2, \rho_1, \tau_2^2, \rho_2$ with those shown in Table 4.1 for Samples 2 and 3, we generate another two sets of data.

We fit the two i.i.d spatial CAR models as described in Section 4.1.2 to the simu-

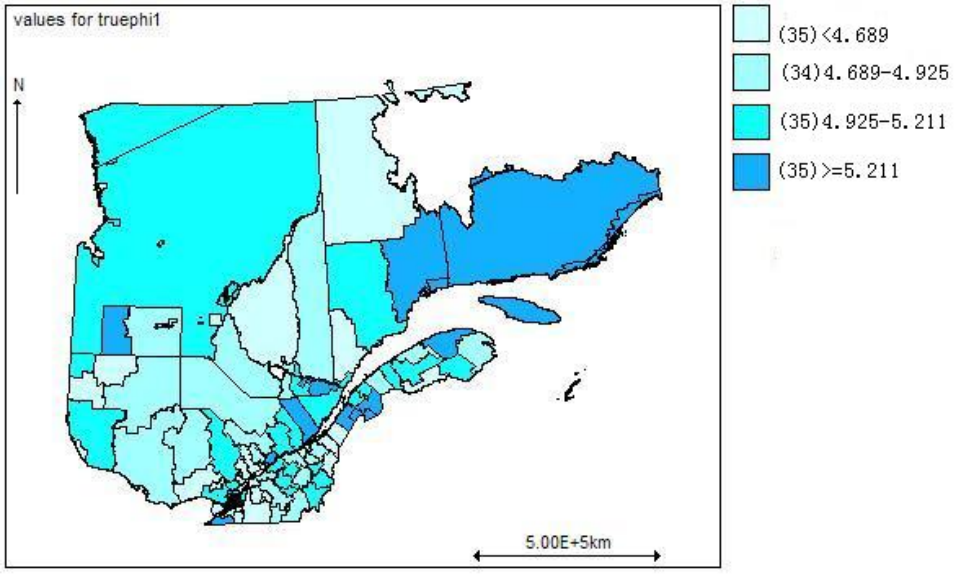
lated data. As mentioned in Section 4.1, we chose a vague normal prior ($N(0, 10000)$) for b.age, b.gender, μ, and b.trt, and an inv-gamma(0.01,0.01) for σ. The priors for $\boldsymbol{\phi_1}$ and $\boldsymbol{\phi_2}$ are $CAR(\boldsymbol{\mu}, \rho_1, \tau_1^2)$ and $CAR(\mathbf{b.trt}, \rho_2, \tau_2^2)$, respectively. For the hyper-parameters associated with two CAR models, we use a uniform$(0, 0.999)$ for ρ_1 and ρ_2, and inv-gamma$(0.01, 0.01)$ for τ_1^2 and τ_2^2. If our MCMC implementation works, posterior estimations of parameters should be close to the true values used for generating the data. We implemented the MCMC algorithm in Matlab. We first ran a few initially overdispersed parallel MCMC chains, and monitored them using measurements of sample autocorrelations within the chains and plots of sample traces. From these, we decided to use a sample of 20000 iterations after burn-in for posterior summaries. Table 4.1 is a summary of the estimated values of model parameters including mean, standard deviation and 95% credible intervals.

The estimated fixed effects and the centered mean of two spatial random effects are accurate in all three cases. The estimated hyper-parameters are less accurate. Nevertheless, the 95% credible intervals of estimators still contain the true values. The estimated smoothing parameter for the space-varying treatment coefficient is not as promising as others. Possible reasons are lack of enough information and potential weak identifiability between two spatial processes when there are a small number of (possibly none) patient(s) who are treated in some regions. This is not unexpected since the parameters of spatial covariance structures are typically weakly identifiable [42]. This is also evident from the maps and scatter plots of the posterior means of the two vectors of spatial random effects (results from Sample 3) and the true values of the random effects shown in Figure 4.1, Figure 4.2, Figure 4.3, and Figure 4.4. By comparing the graphs for the true and estimated values, we see the estimation of the spatial random effects is better than that of the space-varying treatment effects.

Table 4.1: Estimation summary of simulated samples

	Parameter	true value	mean	SD	2.5%	median	97.5%
Sample 1	$E(\phi_{1j}) = \mu$	5.00	4.944	0.048	.851	4.945	5.037
	b.age	-0.09	-0.091	0.003	-0.096	-0.091	-0.086
	b.gender	-0.20	-0.198	0.050	-0.295	-0.199	-0.100
	$E(\phi_{2j}) = b.trt$	0.90	0.921	0.047	0.828	0.923	1.001
	σ	1.20	1.200	0.012	1.175	1.200	1.224
	τ_1^2	0.10	0.071	0.049	0.008	0.061	0.183
	ρ_1	0.90	0.575	0.289	0.038	0.612	0.980
	τ_2^2	0.03	.049	0.050	0.004	0.031	0.185
	ρ_2	0.80	0.452	0.277	0.022	0.436	0.945
Sample 2	$E(\phi_{1j}) = \mu$	5.00	5.110	0.120	4.875	5.112	5.344
	b.age	-0.09	-0.087	0.003	-0.092	-0.087	-0.082
	b.gender	-0.20	-0.202	0.051	-0.301	-0.202	-0.101
	$E(\phi_{2j}) = b.trt$	0.90	0.903	0.052	0.792	0.905	1.011
	σ	1.20	1.217	0.013	1.193	1.218	1.242
	τ_1^2	0.50	0.340	0.092	0.190	0.330	0.549
	ρ_1	0.90	0.910	0.077	0.701	0.929	0.993
	τ_2^2	0.10	0.074	0.065	0.007	0.053	0.253
	ρ_2	0.80	0.452	0.280	0.022	0.435	0.950
Sample 3	$E(\phi_{1j}) = \mu$	5.00	4.954	0.077	4.801	4.955	5.106
	b.age	-0.09	-0.090	0.003	-0.095	-0.090	-0.084
	b.gender	-0.20	-0.171	0.051	-0.270	-0.172	-0.072
	$E(\phi_{2j}) = b.trt$	0.90	0.825	0.053	0.728	0.823	0.931
	σ	1.20	1.233	0.013	1.208	1.233	1.258
	τ_1^2	0.50	0.549	0.133	0.322	0.539	0.839
	ρ_1	0.80	0.705	0.180	0.264	0.742	0.952
	τ_2^2	0.10	0.141	0.115	0.018	0.105	0.440
	ρ_2	0.60	0.472	0.274	0.023	0.467	0.949

(a)



(b)

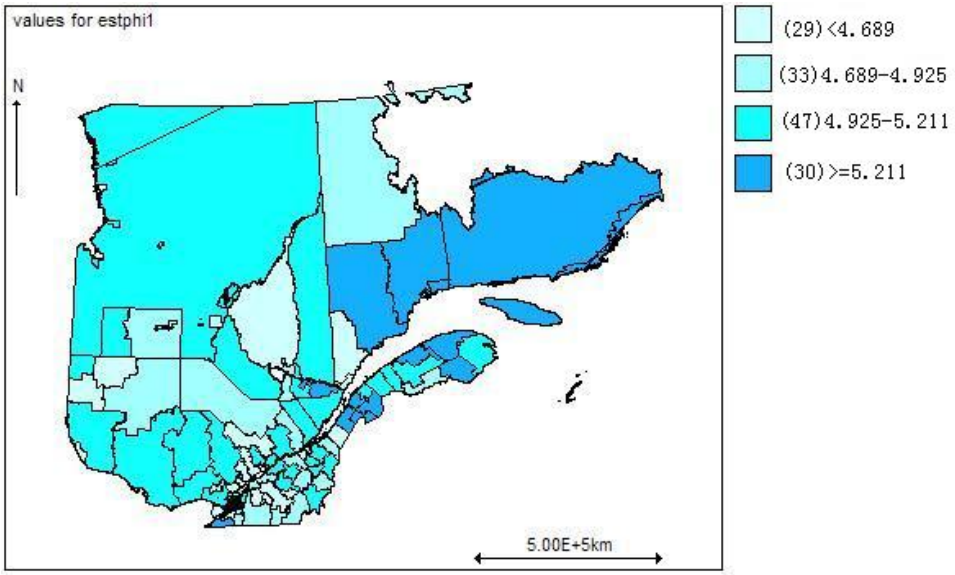
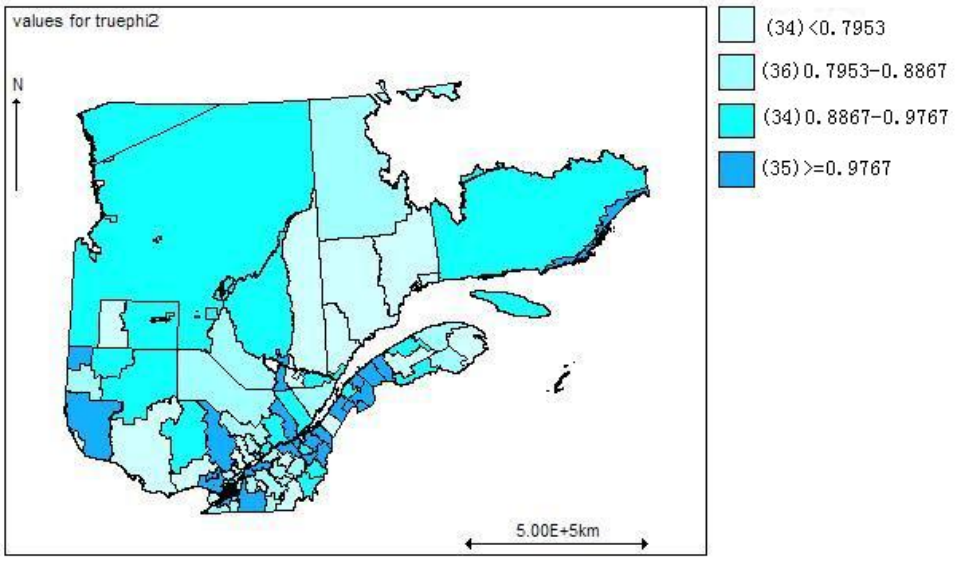


Figure 4.1: (a) Maps of true spatial random effects ($\boldsymbol{\phi_1}$) (b) Maps of estimated spatial random effects ($\hat{\boldsymbol{\phi_1}}$).

(a)



(b)

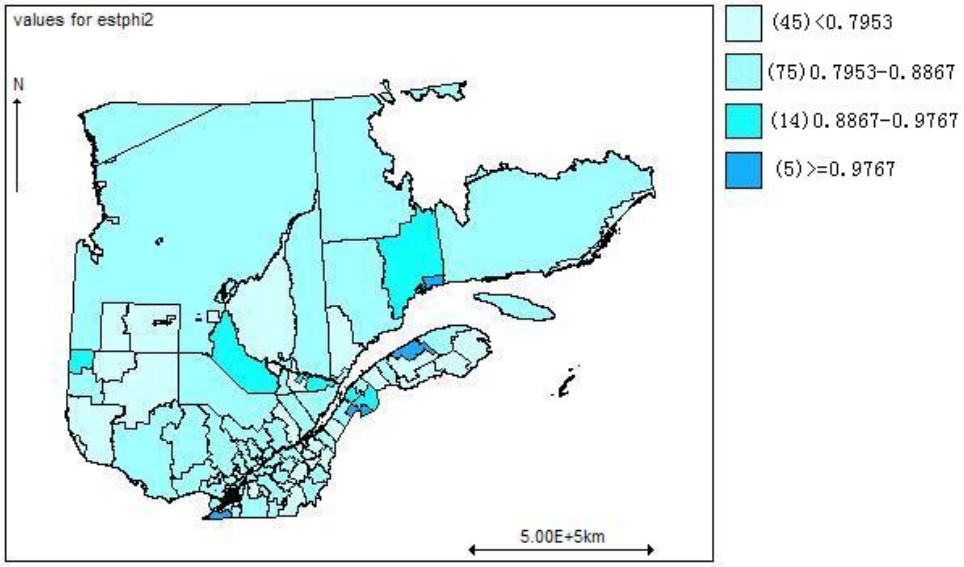
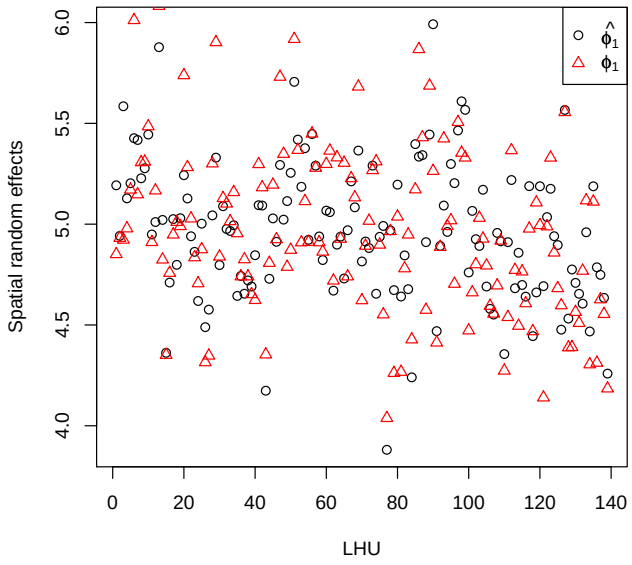


Figure 4.2: (a) Maps of true space-varying treatment effects ($\boldsymbol{\phi_2}$)(b) Maps of estimated space-varying treatment effects ($\hat{\boldsymbol{\phi_2}}$)

(a)



(b)

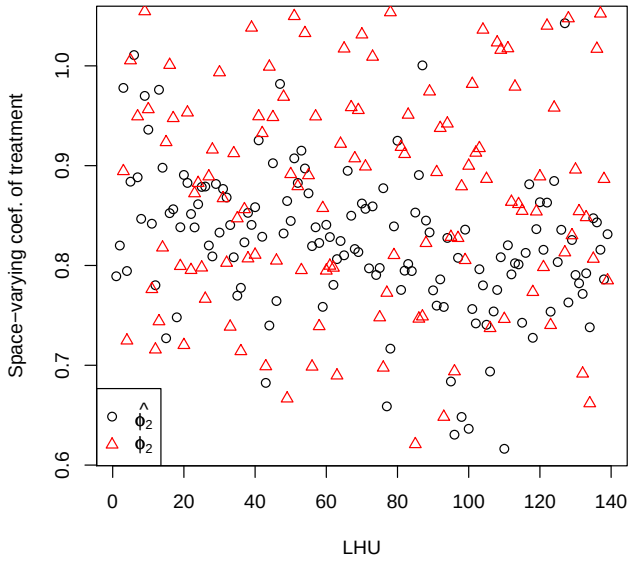


Figure 4.3: (a) Scatter plots of true ($\boldsymbol{\phi_1}$) and estimated ($\hat{\boldsymbol{\phi}}_{\mathbf{1}}$) spatial random effects (b) Scatter plots of true ($\boldsymbol{\phi_2}$) and estimated ($\hat{\boldsymbol{\phi}}_{\mathbf{2}}$) space-varying treatment effects.

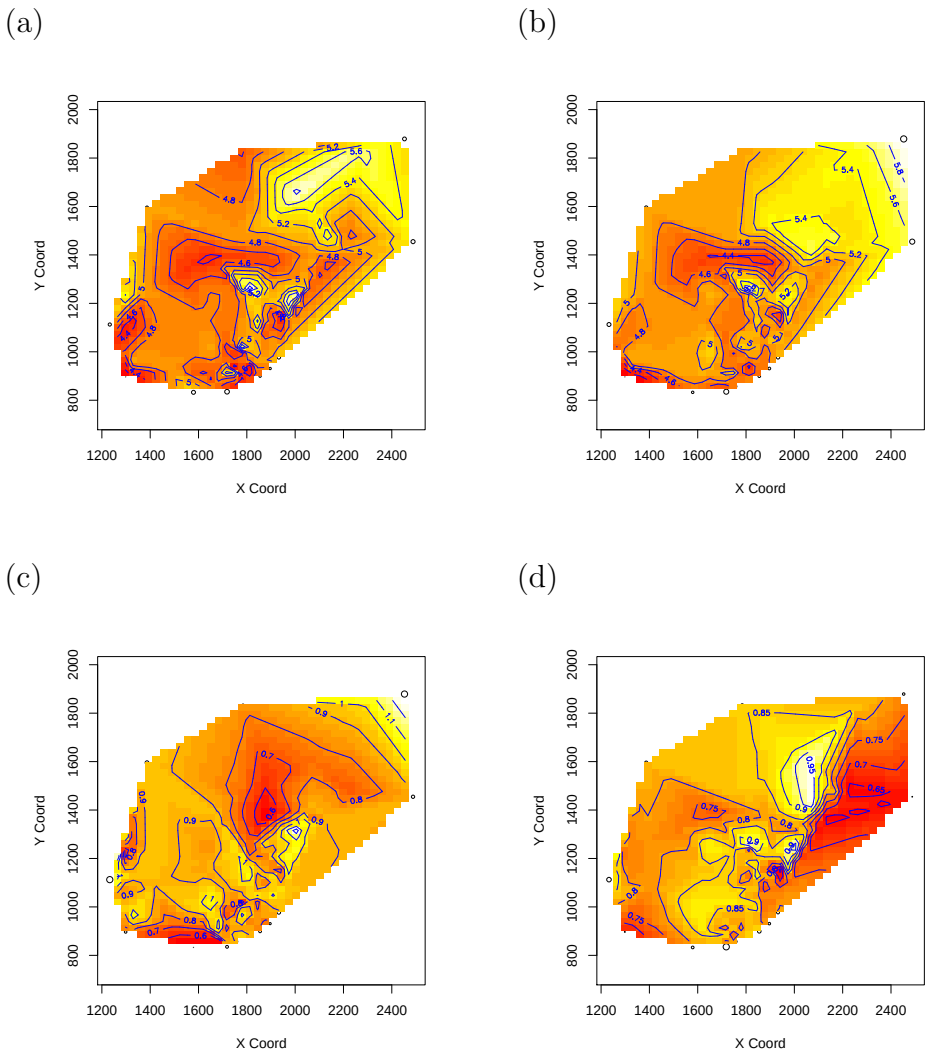


Figure 4.4: (a) Contour maps of true spatial random effects ($\boldsymbol{\phi_1}$) (b) Contour maps of estimated spatial random effects ($\hat{\boldsymbol{\phi}}_{\mathbf{1}}$) (c) Contour maps of true space-varying treatment effects ($\boldsymbol{\phi_2}$) (d) Contour maps of estimated space-varying treatment effects ($\hat{\boldsymbol{\phi}}_{\mathbf{2}}$).

Simulation Study on Model Selection

As before, we assume a logistic error distribution and examine the performance among the various models: (1) fixed effects model ($\log(t_{ij}) = \mu + \mathbf{b}'\mathbf{x_{ij}} + \sigma w_{ij}$), (2) i.i.d random effects model ($\log(t_{ij}) = \mathbf{b}'\mathbf{x_{ij}} + \phi_{1j} + \sigma w_{ij}$, where $\phi_{1j} \sim N(\mu, \tau_1^2)$), (3) CAR model ($\log(t_{ij}) = \mathbf{b}'\mathbf{x_{ij}} + \phi_{1j} + \sigma w_{ij}$, where $\phi_{1j} \sim CAR(\mu, \rho_1, \tau_1^2)$, and (4) two i.i.d CAR models ($\log(t_{ij}) = \mathbf{b}'\mathbf{x_{ij}} + \phi_{1j} + \phi_{2j} R_{ij} + \sigma w_{ij}$, where $\boldsymbol{\phi_1} \sim CAR(\boldsymbol{\mu}, \rho_1, \tau_1^2)$ and $\boldsymbol{\phi_2} \sim CAR(\boldsymbol{b.trt}, \rho_2, \tau_2^2)$). In Model (1) we only consider the fixed effects for covariates age, gender, and treatment. Whereas in Model (2) we take spatial random effects into consideration and assume that they are independent and identically distributed (i.i.d.) with normal density centered at the intercept μ. In Model (3) we remove the i.i.d. assumption and model spatial random effects correlatedly through a CAR model. In Model (4), we add one more spatial process that models space-varying treatment effects. We now discuss four studies, each involves simulating data from a true model given by the true parameters values in Table 4.2. For each set of the simulated data, we fit the four different models, including the true one used to generate this data. Table 4.3 summarizes the DIC scores of each model fit. In Study 1 and 3, the two i.i.d. CAR models actually performed as well as the true model with slightly smaller DIC scores than that of the true models. In Study 2, DIC values for the true model, CAR, and two i.i.d. CAR models are very close. Only in study 4 is the true model better than the other 3 models. In all 3 studies, we see that the two i.i.d. CAR models are competitive compared to the true model. This might also suggest that DIC prefers more complicated models.

Table 4.2: The true values of parameters in Studies 1-4

Study	True model	μ	b.age	b.gender	$b.trt$	σ	τ_1^2	ρ_1	τ_2^2	ρ_2
1	fixed	5.0	-0.09	-0.20	0.90	1.2	–	–	–	–
2	i.i.d.	5.0	-0.09	-0.20	0.90	1.2	0.50	–	–	–
3	CAR	5.0	-0.09	-0.20	0.90	1.2	0.50	0.80	–	–
4	two i.i.d. CAR	5.0	-0.09	-0.20	0.90	1.2	0.50	0.80	0.10	0.60

Table 4.3: DIC scores of the fitted models in Studies 1-4

Fitted model	Study 1	Study 2	Study 3	Study 4
fixed	104,495	104,614	104,556	104,436
i.i.d.	104,499	104,373	104,424	104,299
CAR	104,498	104,374	104,414	104,288
two i.i.d. CAR	$\mathbf{104,494}$	$\mathbf{104,372}$	$\mathbf{104,412}$	$\mathbf{104,278}$

Chapter 5

Application to a Study of Acute Myocardial Infarction in Quebec

In this chapter we illustrate our method developed in Chapter 4 with an application to the Acute Myocardial Infarction (AMI) data in Section 1.2. The study enrolled 61107 individuals aged 25 or older across the province of Quebec, who experienced an initial episode of AMI. After first hospitalization, these individuals were followed over time for recurrent episodes of AMI. During the study period of January 01, 1996 to December 31, 1999, the time to death after AMI for an individual was recorded (this is a complete observation). If an individual survived at the end of study period, then the time from the date entering the study to the termination time is recorded (this is a right censored survival time). The response variable in this study is survival time T (either complete or censored) after AMI. In addition to the survival time, a number of explanatory variables are available. These include the gender, age and geographical strata (local health units) and treatment information to indicate whether an individual received the treatment of revascularization therapy. We only include 61054 individuals with survival days greater than zero in the analysis. These individuals resided in 139 local health units (see map in Figure 1.1) and hence the neighbourhood relationship among these LHUs yields the spatial structure applied in our data modeling. The number of individuals in each LHU varies greatly from the smallest of 6 to the largest of more than 3000. There are about 50% of LHUs with number of individuals under 350, 25% of LHUs with sample size under 205, and 75% of LHUs with sample size under 536. Such big variations in sample size among LHU is not desirable and it might affect our estimation when spatial random effects are

considered. A preliminary analysis of this data set was in Section 1.3. In this chapter, we explain some of the results in Section 1.3 in more detail and more importantly, we apply our AFT model to analyze the data set.

The rest of this chapter is organized as follows. In Section 5.1, we start with classical survival analysis methods without spatial consideration. In Section 5.2, we move to the Bayesian survival analysis and illustrate how the spatial random effects are handled through Bayesian hierarchical modeling. We apply different parametric AFT models and present the model comparison and selection. We end this chapter with assessing the goodness-of-fit of the preferred models.

5.1 Classical Survival Analysis

In Chapter 2, we discussed non-parametric Kaplan-Meier (KM) estimator for estimating the survival function, Cox proportional hazards model, and accelerated failure time models. This section presents the results of applying these models which motivated us to explore the Bayesian methods, incorporating spatial effects into the AFT model.

5.1.1 Fitting Kaplan-Meier Method

Most statistical survival packages have a built-in function to implement the KM estimator. We used the $survfit()$ function in R to get KM estimates of the survival function. The following figures are plots of the KM estimates of survival by gender and by treatment group, respectively. Figure 5.1 (a) shows that a male subject has higher estimated survival probability than a female. Similarly, Figure 5.1 (b) shows that the treatment group has higher estimated survival probability than its counterpart. We need to test whether the differences are significant. In Table 5.1, We test the null hypothesis that the hazard rate of patients in the treatment group is the same

as the hazard rate of patients in the without treatment group. The alternative is that the two hazard rates are different. The test statistic is $\chi^2 = 1473$ with one degree of freedom. Here the p-value of this test is close to zero, and we reject that null hypothesis that the hazard rates in the treatment and non-treatment groups are equal. Similarly, in Table 5.2 the test statistic is $\chi^2 = 264$ with one degree of freedom. Since the p-value of this test is close to zero, we also reject the null hypothesis that hazard rates are equal between female and male.

(a) (b)

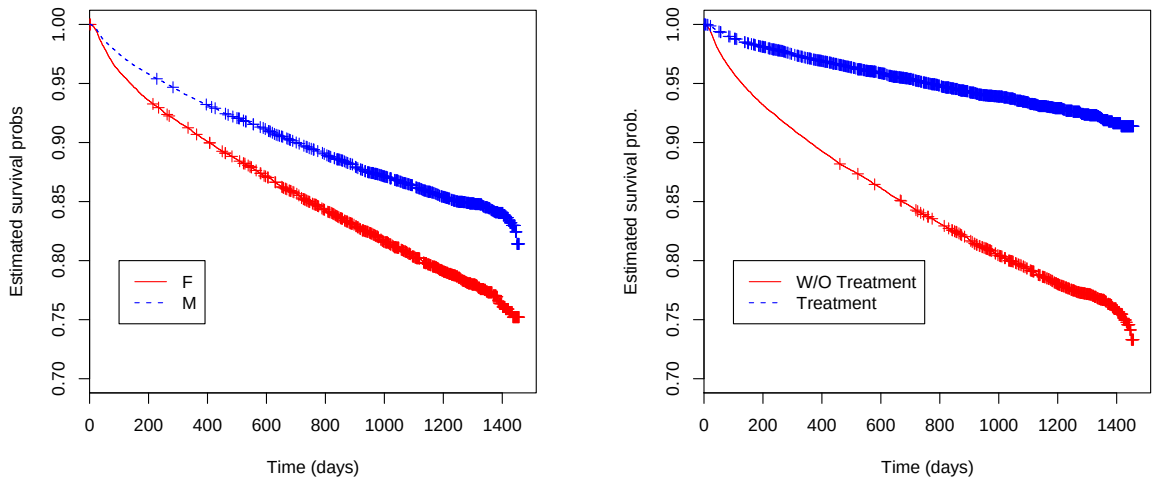


Figure 5.1: (a) KM estimate of the survivor function by gender (b) KM estimate of the survivor function by treatment group

Table 5.1: Log rank (Mantel-Haentzel) test on treatment groups

treatment	N	Observed (O)	Expected (E)	$(O-E)^{2/E}$	$(O-E)^{2/V}$
non-treatment	37813	5733	4209	552	1473
treatment	23241	999	2523	921	1473
Chisq= 1473 on 1 degrees of freedom, p-value ≈ 0					

However, these statistical tests did not provide an estimate of the magnitude of

Table 5.2: Log rank (Mantel-Haentzel) test on gender groups

Gender	N	Observed (O)	Expected (E)	$(O-E)^{2/E}$	$(O-E)^{2/V}$
Female	21053	2902	2271	175.0	264
Male	40001	3830	4461	89.1	264
Chisq= 264 on 1 degrees of freedom, p-value ≈ 0					

the difference in survival for the cohorts being compared. Further, the tests allowed for comparisons based on one grouping factor at a time. We can get an estimate of the magnitude of the survival covariate relationship of interest and can consider multiple factors simultaneously for each subject in a time-to-event study through regression modeling which is described in the following sections.

5.1.2 Fitting Cox Proportional Hazards (PH) Model

The Cox PH model specified in (2.2.1) is fitted to the AMI data using partial likelihood estimation. We fit Cox PH model using the *coxph*() function in R and checked the PH assumption, which is the hazards ratio between two sets of covariates is constant over time. The model includes the covariates age, gender, and treatment. Table 5.3 shows the estimates of the coefficients and the standard errors. We interpret

Table 5.3: Output of Cox PH model

variable	coefficient	exp(coefficient)	se(coefficient)	z	p-value
Age	0.072	1.075	0.001	59.90	0.0e+00
Gender	0.163	1.177	0.026	6.37	1.9e-10
Treatment	-0.804	0.448	0.035	-22.75	0.0e+00
Likelihood ratio test statistics=6070 on 3 df, p-value$\approx$ 0 n= 61054					

the results in Table 5.3 as follows. First of all, the p-values, associated with null hypothesis test that the coefficients of covariates are equal to zero, are significant at

significance level $\alpha = 0.01$. These estimates can be linked to equation (2.2.3) when we think of relative risks. For example, the hazard ratio (relative risk) of death for females relative to males is

$$\frac{h(t; x = F)}{h_0(t)} = \exp(\beta_{\text{gender}}) = \exp\{0.163\} = 1.177 \text{ with s.e. } (0.045).$$

The computed relative risk means that females have a higher hazard (risk) of death than males in this case and the relative risk quantifies this. Similarly, the relative risk of death for patients with treatment relative to patients without is 0.448 (with s.e. 0.016). In other words, the treatment group is associated with a 55 % lower risk of death than the without treatment group.

To assess the lack-of-fit of a PH model, we plotted log-log of the estimated survival function against time for the different cohorts. The log-log plot is presented in Figure 5.2. The curves cross at the early event times and are not parallel for both graphs in (a) and (b). The plotting of age groups shows that curves cross over between age groups 25-40 and 41-50. **This indicates that the PH assumption may not be appropriate for these data.** Grambsch and Therneau [66] present Schoenfeld residuals based on the χ^2 test statistics. Table 5.4 shows the test result. The global test p-value is smaller than 0.05, which is consistent with the observation from Figure 5.2.

One further method for assessing the PH model fit is the Cox-Snell residual plot [67]. Figure 5.3 shows the plot of the residuals versus the estimated cumulative hazard of the residuals. If the Cox model fits the data, the plot should follow the 45° line. The plot suggests that this model does not fit the data well.

(a) (b)

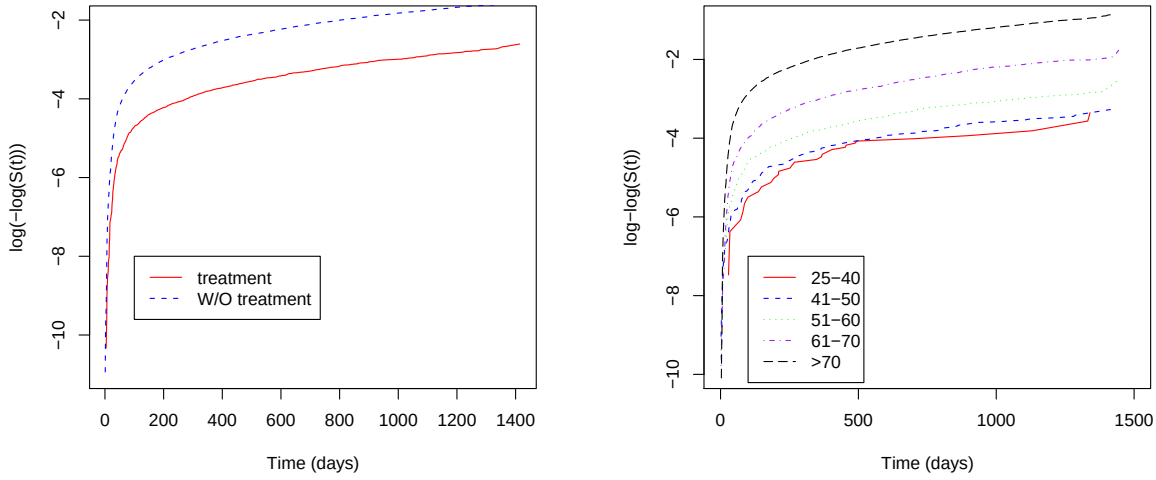


Figure 5.2: (a) log[-log] survivor function by treatment group (b) log[-log] survivor function by stratified age group

Table 5.4: Cox PH model assumption test (reject PH at α=0.05)

variable	ρ (correlation)	χ^2 (test statistics)	p-value
Age	0.030	6.033	0.014
Gender	0.005	0.191	0.662
Treatment	0.024	3.933	0.047
GLOBAL	NA	8.210	0.042

5.1.3 Fitting Accelerated Failure Time Models

Recall that for the AFT model, the linear relationship between log time T and the covariate values is

$$Y = \log(T) = \mu + \mathbf{b}^{'}\mathbf{x} + \sigma W.$$

Under a parametric AFT model, the survival times are assumed to follow either Weibull, log-normal, or log-logistic distributions. By including covariates, we fit these three AFT parametric models listed in Table 2.1 in Section 2.2.2 using full maximum

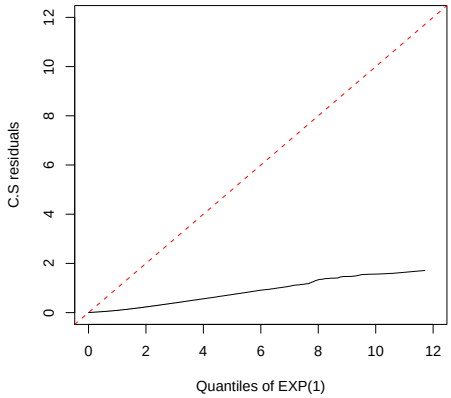


Figure 5.3: Cox-Snell residual plot (solid line–Cox-Shell residual plot, the dash line–45° reference line).

likelihood estimation. The results are presented in Table 5.5.

Table 5.5: Output of AFT different parametric models

	Estimates of Coefficients (SD)					
Dist. of T	σ	μ	b.age	b.gender	b.trt	AIC
Weibull	1.21	9.41(.042)	-0.088(.002)	-0.199(.031)	0.972(.044)	123657
Lnormal	2.30	9.45(.044)	-0.089(.002)	-0.169(.035)	0.982(.042)	123765
Loglogit	1.13	9.12(.041)	-0.090(.002)	-0.195(.033)	0.974(.043)	123636

It is obvious that the estimates of the regression coefficients are different from the output from the PH model. For example, the signs of the coefficients of covariates are different. Take the log-logistic AFT model as an example. If we are interested in the acceleration factor related to treatment, we need to compute β.trt by using the minus regression coefficient corresponding to treatment divided by the scale (σ), which is $\beta.\text{trt} = -0.974/1.13 = -0.862$. Then we say that the treatment effect accelerates time to failure by a factor of $\exp(-0.862) = 0.422$ (s.e. 0.016) and hence slows the

time to death compared to that without treatment. Similarly, we can compute that being female accelerates time to failure by a factor of $\exp(\frac{0.195}{1.13}) = 1.188$ (s.e. 0.034) compared to being male.

5.1.4 Model Checking and Selection

A quantile-quantile or q-q plot is a graphic diagnostic method for the AFT model [27]. The plot is based on the fact that the AFT approach assumes that $S_1(t)$ in one group is equal to $S_0(\theta t)$ in the other. Let t_{0p} and t_{1p} be the p^{th} percentiles of groups 0 and 1, respectively, that is

$$t_{kp} = S_k^{-1}(1-p), k = 0, 1. \tag{5.1.1}$$

Using the relationship of the survival function for two groups, we have $S_0(t_{0p}) = 1 - p = S_1(t_{1p}) = S_0(\theta t_{1p})$ for all t. If the accelerated failure time model holds, then $t_{0p} = \theta t_{1p}$. To check this assumption we compute the Kaplan-Meier estimators of the two groups and estimate the percentiles t_{1p}, t_{0p} for various values of p. The plot of t_{1p} versus t_{0p} should be a straight line through the origin if the AFT model holds. Figure 5.4 shows the q-q plot for treatment versus non-treatment group and the graph is nearly a straight line. This suggests that the AFT model is a good choice.

Another method that can be used for model checking and selection is the standardized residuals, which are based on the log linear model representation in (2.2.6). They are defined as

$$r_i = \frac{\log(t_i) - \hat{\mu} - \hat{\mathbf{b}}\mathbf{x_i}}{\hat{\sigma}},$$

where $\hat{\mu}$, $\hat{\mathbf{b}}$, and $\hat{\sigma}$ are the maximum likelihood estimates of the parameters. We expect that the distribution of the standardized residuals should be close to that of white noise drawn from the assumed error distribution. Then we could easily get KM survival function estimates for these r_i's. In order to fit all the residuals, each residual

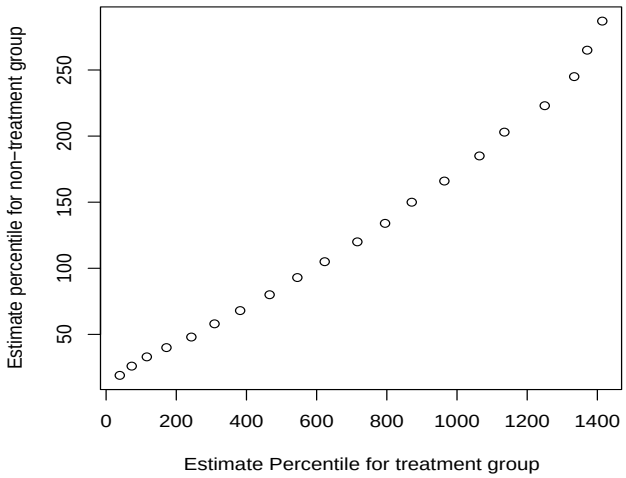


Figure 5.4: q-q plot to check the AFT model (treatment vs. non-treatment).

has been adjusted by subtracting the minimum value and adding a very small value. Under different parametric error term assumptions defined in Table 2.1, plotting a suitably transformed KM estimate $\hat{S}(r)$ against r should give a straight line if the model assumption holds. The plots are presented in Figure 5.5. The three graphs all look approximately like a straight line with exception of the most extreme values. For example, Figure 5.5 (b) is the plot for the log-lognormal. The upper part shows a curve trend. It might be an indication that the log-lognormal model is not robust to the extreme value. Weibull and log-logistic models seem competitive with each other. Overall, there are no strong indications of departures from AFT models.

The most appropriate model may be selected from the family of AFT models based on Akaike's Information Criterion (AIC). The model with the smallest AIC best describes the data. It seems that the log-logistic model in Table 5.5 has the smallest AIC and hence it is the best model for the given data when spatial random effects have not been considered.

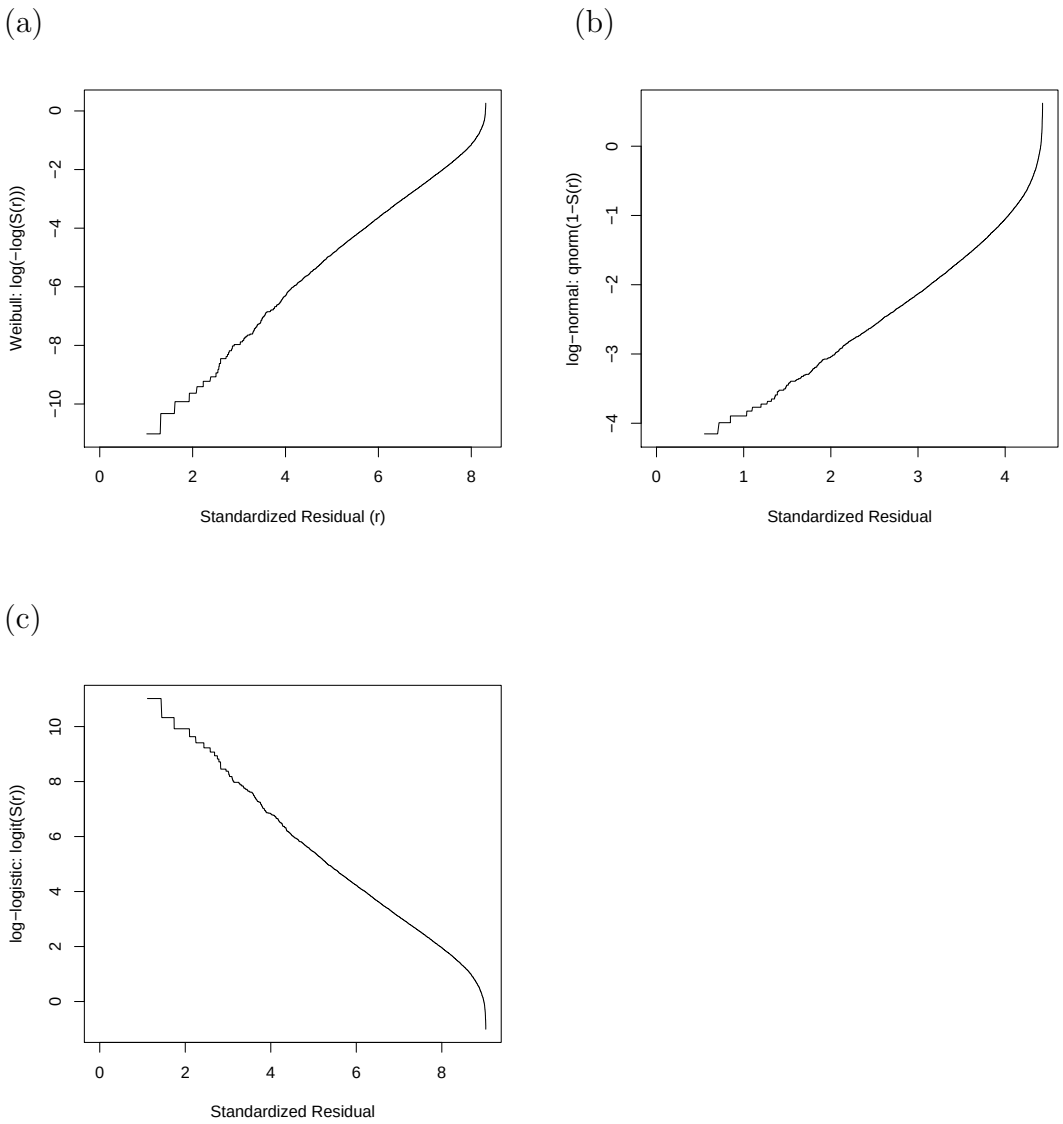


Figure 5.5: Diagnostic standardized residual plots to evaluate the Weibull (a), log-normal(b), and log-logistic (c) AFT models

5.2 Bayesian Survival Analysis

In the above classical survival analyese, we did not take the spatial characteristic of AMI data into consideration. Since AMI data are clustered survival data where the clusters correspond to local health units, we now consider models with spatially distributed random effects. These models begin with spatially oriented frailty terms, but may also include further region-level terms in the parametrization of various covariate effects, such as in a spatially-varying coefficients model. Previous sections discuss parametric AFT survival models with spatial random effects under hierarchical Bayesian framework and show these models may be efficiently implemented using Markov chain Monte Carlo (MCMC) computational techniques. In this section, we present the results from the various models.

5.2.1 Implementation Strategy and Model Convergence Checking

We wrote our program in Matlab to implement the models following the algorithm in Section 4.2. Our program is flexible in that we could easily adapt it to a semi-parametric setting by replacing the likelihood function. We start with replicating the results from *survreg*() using our program under AFT parametric setting, then we work through the implementation from the simplest model to the complicated ones by adding spatial random components. For each of models, we adopt the approach by running multiple initially overdispersed parallel MCMC chains and monitoring them using trace plots and sample autocorrelations within the chains. By tuning the jumping steps of the proposed distribution, we achieve the optimal acceptance rate between 0.2 to 0.5 most of the time. As we discuss in earlier section, if the jumping step (tuning parameter) is too big then the acceptance rate would be too low. Hence the chain needs longer time to converge. On the other hand, if the jumping step is too small, then the acceptance rate would be too high and the chain will not fully traverse the parameter space causing high auto-correlation.

Besides tuning jumping steps, it is also important to check the convergence of the chains in order to decide the length of the "burn-in" period. As mentioned in the previous section, we monitor the trace plots and sample auto-correlation within chains to decide if the MCMC chain converges. We illustrate this approach by taking log logistic AFT with two i.i.d. CAR models as an example, as this model has 4 hyper-parameters and it tends to converge slowly. We run six chains with different initial values spread over the parameter space for each parameter. Each chain is thinned by 40 for reducing the autocorrelations as well as saving computer memory space. As shown in Figure 5.6, the auto-correlations for hyper-parameter $\tau_1^1, \rho_1, \tau_2^2$, and ρ_2 drop quickly after thinning. Further, the trace plots (Figure 5.7) and the ergodic plots (Figure 5.8) indicate that the MCMC chains for regression coefficients and hyper-parameters converge after 400×40 iterations. Hence it is safe to use the second half of the saved 5000 posterior draws thinned by 40 to compute the Gelman and Rubin diagnostic statistics (see Section 3.4.3). In this case, such statistics for all parameters are close to 1, which is a good indication of convergence. Using such techniques, we observe that 20000 iterations for the preconvergence "burn-in" period are sufficient and a further 20000 iterations are used for posterior summarization and DIC calculations.

5.2.2 Model Comparison and Selection

We use the DIC criterion described in Section 2.4.3 to compare various parametric spatial models for the AMI data. Table 5.6 lists the pD and DIC scores for 3 parametric models applied in the different modeling categories. If we confine ourself to the same parametric model but different categories, as we expected, the fixed models with no random effect has the smallest pD value which is close to the actual number of parameters $(\mu, b.age, b.gender, b.trt, \sigma)$ in the model; but the highest DIC score

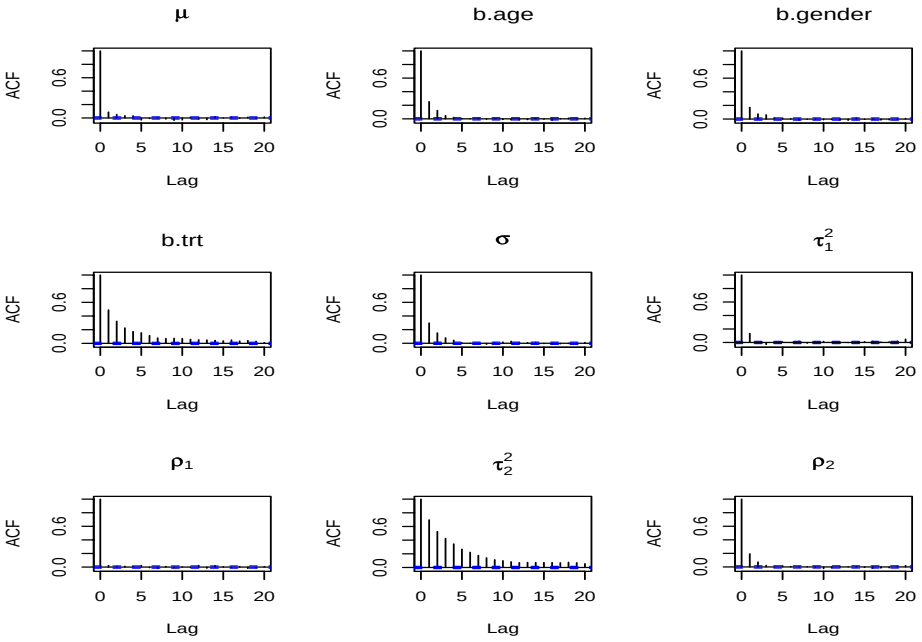


Figure 5.6: Autocorrelation plots of the parameters.

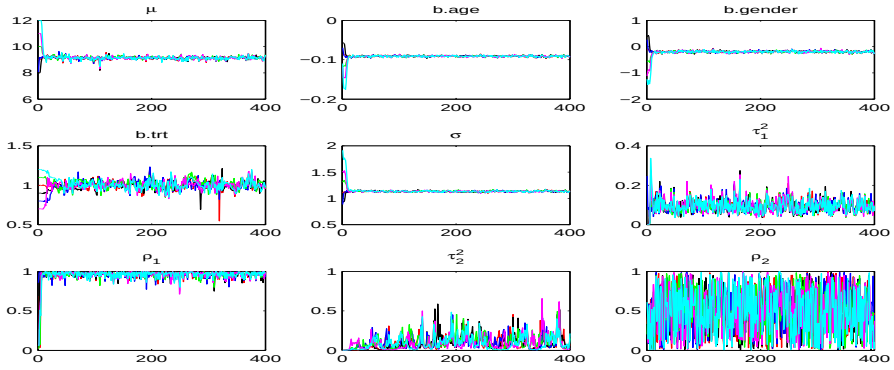


Figure 5.7: Trace plots of the parameters for tmultiple chains.

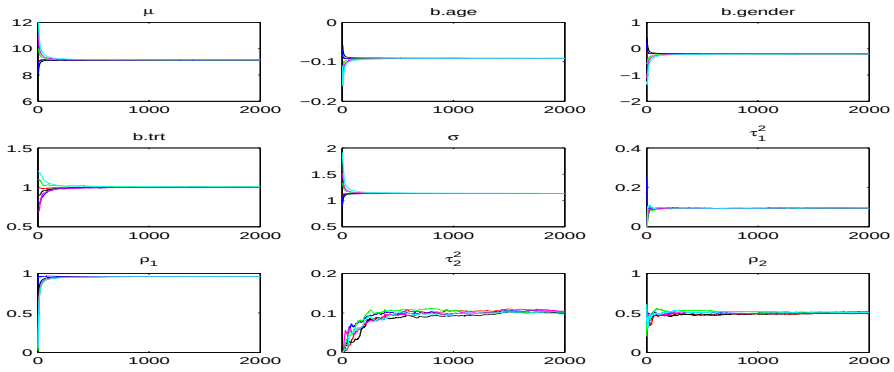


Figure 5.8: Ergodic average plots of the parameters for the multiple chains.

indicates that this simple model does not fit the data well. By adding random effects onto the fixed model, the effective number of parameters pD increases greatly but the DIC scores decrease substantially. Also, pD values for the models with CAR random effects are smaller than that for the i.i.d. random effects model, since random effects are correlated in CAR models. However, DIC values for the CAR and the two i.i.d. CAR models do not substantially differ. Based on DIC, it seems as though there is no need for space-varying regression. The evidence here suggests that while there is residual spatial structure in the data, the treatment effect does not appear to be spatially-varying or there is not enough information in the data. Among these three parametric families, log-logistic performs the best.

5.2.3 Results from the Fitted Models

This section presents the results from fitting four different types of models (fixed effect, i.i.d. spatial random effect, CAR, and two i.i.d. CAR) under three AFT parametric settings (log-logistic, Weibull, and log-normal). A relatively detailed analysis

Table 5.6: DIC comparison of different parametric AFT models

parametric	log-logistic		Weibull		log-normal	
models	pD	DIC	pD	DIC	pD	DIC
fixed	4.8	123,636.8	4.9	123,658.0	4.8	123,764.7
i.i.d.	74.3	123,543.1	71.9	123,574.1	77.5	123,663.4
CAR	56.5	123,519.1	54.7	123,550.8	57.3	123,639.9
two i.i.d. CAR	68.3	123,518.6	68.4	123,549.6	69.6	123,638.5

is provided to the results of the two i.i.d. CAR model from the log-logistic AFT setting. There are three reasons for this detailed analyses: (1) log-logistic AFT is the best among three parametric families, (2) the other three types of models are the sub-model of the two i.i.d. CAR model, and (3) the CAR model and two i.i.d. CAR model perform equally well for the AMI data.

Fixed Effect Models

If we remove the spatial random effect ϕ_j from (4.1.1), we have the log linear survival model formulation for the fixed effects, which is

$$Y_{ij} = \log(t_{ij}) = \mu + \mathbf{b}^{'}\mathbf{x_{ij}} + \sigma w_{ij}. \tag{5.2.1}$$

The λ_{ij} in (4.1.2) hence becomes

$$\lambda_{ij} = \mu + \mathbf{b}^{'}\mathbf{x_{ij}}, \tag{5.2.2}$$

which is a standard AFT regression model. The rest of the Bayesian model specification for the fixed effects is similar to the univariate CAR model in Section 4.1.1 with spatial random effects being removed. The prior specifications for the parameters of the fixed effects and MCMC sampling techniques are the same as described in Section (4.2). Thus details on the implementation will not be repeated here. Table

5.7 presents the results for W in (2.2.6) following logistic, extreme-value, and normal distributions, which corresponds to the AFT log-logistic, Weibull, and log-normal models. The interpretations of the output are the same as in the classical survival

Table 5.7: AFT different parametric fixed effects models

	Model					
	log-logistic		Weibull		log-normal	
Parameter	mean	SD	mean	SD	mean	SD
μ	9.149	0.038	9.431	0.041	9.464	0.043
b.age	-0.0905	0.002	-0.088	0.002	-0.090	0.002
b.gender	-0.198	0.032	-0.200	0.031	-0.171	0.034
b.trt	0.985	0.044	0.982	0.045	0.985	0.041
σ	1.139	0.012	1.225	0.012	2.307	0.022

analysis (see Section 5.1.3). We also notice that results in Table 5.7 are very close to that in Table 5.5. This provides evidence that MCMC sampling techniques perform well in estimation. Thus we are confident about adding random effects in the following models.

Independent Identically Distributed Spatial Random Effect Models

Taking the variations among the LHUs into consideration, we add the i.i.d. spatial random effects for each LHU to the model. The model specification and implementation follow closely to the univariate CAR model. The only difference appears in prior specification for ϕ_j. In i.i.d. spatial random effects models, $\phi_j|\mu, \tau^2 \overset{iid}{\sim} N(\mu, \tau^2)$. Hence, the neighbouring structure for each LHU is irrelevant to this model as the spatial structure is not considered explicitly. Again, we try three different AFT models. The results are shown in Table 5.8. Compared to the fixed effect model, τ^2, the variance of the spatial random effects, is the only parameter added for i.i.d. random

effects models.

Table 5.8: AFT different parametric i.i.d. random effects models

	Model					
	log-logistic		Weibull		log-normal	
Parameter	mean	SD	mean	SD	mean	SD
μ	9.145	0.049	9.435	0.050	9.462	0.055
b.age	-0.091	0.002	-0.089	0.002	-0.091	0.002
b.gender	-0.198	0.034	-0.201	0.032	-0.173	0.036
b.trt	0.987	0.044	0.984	0.045	0.990	0.042
σ	1.134	0.012	1.223	0.014	2.295	0.024
τ^2	0.048	0.012	0.040	0.010	0.059	0.014

Spatial CAR Random Effect Models

It is reasonable to assume the variations exist among different LHUs and such random effects for each LHU is influenced by its neighbours (only first order neighbouring is considered). This yields the CAR model and implementation follows the notes in Section 4.2.1. The results are shown in Table 5.9. The estimated ρ from three models are bigger than 0.95, indicating strong spatial associations among local health units.

Two i.i.d. Spatial CAR Random Effect Models

The key part of this thesis is to study the spatial variation of the treatment effect. If we assume a spatial continuity on the geographical region, we expect that such variation on the treatment effect should occur smoothly over space, and hence the treatment effect should exhibit a spatial structure. This model allows two spatial random effects: one explains the random effects associated with the residual spatial structure; space varying treatment random effects, however, are more related to the

Table 5.9: AFT different parametric spatial CAR random effect models

	Model					
	log-logistic		Weibull		log-normal	
Parameter	mean	SD	mean	SD	mean	SD
μ	9.141	0.112	9.429	0.103	9.450	0.126
b.age	-0.091	0.002	-0.089	0.002	-0.090	0.002
b.gender	-0.196	0.032	-0.198	0.032	-0.171	0.036
b.trt	0.997	0.044	0.995	0.045	1.004	0.042
σ	1.134	0.012	1.223	0.014	2.294	0.023
τ^2	0.096	0.029	0.082	0.026	0.110	0.034
ρ	0.964	0.033	0.964	0.034	0.971	0.029

difference of health care utilization and administration among local health units. Table 5.10 summarizes the results of the fitting two i.i.d. spatial CAR models. The estimated parameters for the second CAR model, which is associated with the spatial variation of treatment, exhibit a high degree of posterior uncertainty, as the standard errors for τ_2^2 and ρ_2 are large relative to their estimates. This is similar to what we observed in the simulation studies.

Analysis on Log-logistic AFT Two i.i.d. CAR models

Table 5.11 summarizes the posterior mean, standard error, and 95% credible intervals for each parameter. All parameter 95% credible intervals do not include zero. This suggests that all of the predictors are significant at the 0.05 level. The interpretations of the regression coefficients are the same as in the classical survival analysis on AFT survival models. A positive estimated value of a regression coefficient indicates a positive relationship between the covariate and the logarithm of survival time. Regression coefficients of age, gender (Male=0 and Female=1) are negative. For the patients in

Table 5.10: AFT different parametric two i.i.d. spatial CAR random effect models

	Model					
	log-logistic		Weibull		log-normal	
Parameter	mean	SD	mean	SD	mean	SD
μ	9.141	0.105	9.427	0.100	9.45	0.122
b.age	-0.091	0.002	-0.089	0.002	-0.090	0.002
b.gender	-0.197	0.034	-0.199	0.032	-0.171	0.035
b.trt	1.003	0.052	0.998	0.052	1.003	0.048
σ	1.135	0.012	1.223	0.014	2.294	0.023
τ_1^2	0.095	0.031	0.080	0.026	0.105	0.034
ρ_1	0.964	0.034	0.962	0.037	0.969	0.029
τ_2^2	0.091	0.082	0.118	0.088	0.102	0.074
ρ_2	0.502	0.283	0.536	0.288	0.487	0.279

the same LHU, increasing age one year or being female changes the logarithm survival time by a factor of -0.091 with 95 % credible intervals $(-0.095, -0.088)$,or -0.197 with 95% CI $(-0.264, -0.131)$, respectively.

The histogram of the posterior draws after burn-in for each parameter is presented in Figure 5.9. It is clear that all histograms of posteriors except for ρ_2 greatly depart from the corresponding vague priors, which are specified in Section 4.2. It means that the Bayesian learning from priors to posteriors after observing data are significant. As the likelihoods dominate the posteriors, priors play negligible roles. However, there is not much learning for ρ_2, as the posterior distribution differs little from the prior distribution, which is uniform$(0, 1)$. In fact, the histogram of posterior simulation of ρ_2 looks similar to a uniform distribution.

Table 5.11: Parameter estimates of Log-logistic AFT two i.i.d. CAR models

parameter	mean	SD	2.50%	50.0%	97.5%
μ	9.141	0.1053	8.9277	9.1396	9.3498
b.age	-0.0912	0.0017	-0.0947	-0.0911	-0.0878
b.gender	-0.1969	0.0336	-0.2636	-0.1967	-0.1313
b.trt	1.0030	0.0519	0.9053	1.0015	1.1087
σ	1.1347	0.0121	1.1116	1.1344	1.1587
τ_1^2	0.0947	0.0309	0.0461	0.0906	0.1660
ρ_1	0.9635	0.0340	0.8712	0.9734	0.9977
τ_2^2	0.0911	0.0821	0.0074	0.0664	0.3050
ρ_2	0.5017	0.2825	0.0288	0.5076	0.9642

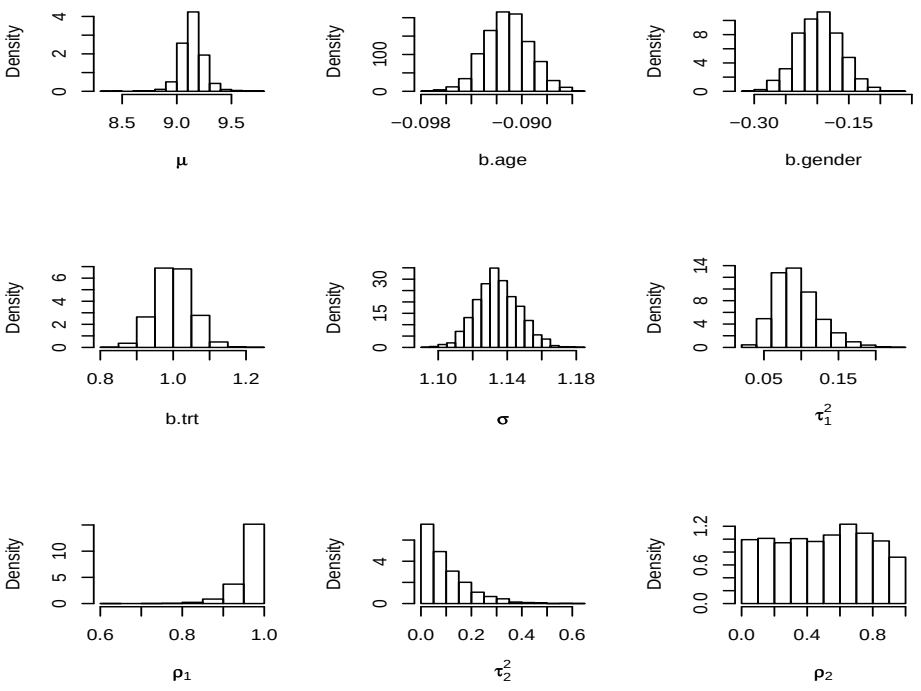


Figure 5.9: Histograms of the parameters.

Figure 5.10 shows the spatial distribution of the posterior means of frailties in the two i.i.d. CAR models. We look at the first CAR spatial random effects. The posterior learning for ρ_1 moves it toward 1, indicating the presence of spatial associations. Visually, clustering of the frailties is shown in Figure 5.10 (a) and (e) by mapping the posterior means of $\boldsymbol{\phi}$. The spatial smoothing in Figure 5.10 (a) is observed compared to Figure 5.10 (e), largely attributable to borrowing of strength from neighbours. High clustering of elevated frailties in the central and south-west corner of Quebec are spotted, with some in the middle-east side as well. It is important to note that these are not maps of elevated marginal effects which would incorporate the regression estimates and space varying treatment coefficients as well. This map also shows the spatial variations that might lead to the direction of finding new covariates. In fact, maps like these which reveal the spatial patterns of the data are extensively used by health science professionals and epidemiologists to identify "hot spots". The hot spots represent areas for further investigation. Figure 5.10 (b) shows the space-varying treatment coefficients. The histogram of ρ_2 in Figure 5.9 indicates there is not enough information of posterior learning for this spatial association parameter of space-varying treatment coefficients when residual spatial structure is included.

Brief Note on Prior Sensitivity Analysis

In our analysis, choosing a vague or non-informative prior is to make a prior play a minimal role in the posterior distribution. Such distributions are sometimes called 'reference prior distributions'. The rational for using non-informative prior distributions is often said to be 'to let the data speak for themselves', so that inferences are unaffected by information external to the current data [68]. However, if there is not information available in data, one should put relevant information into the prior distribution. In the case of analyzing AMI data, it might worth to try a more informative prior (e.g. Beta(18,2)) on smoothing parameter ρ_2 rather using a uniform.

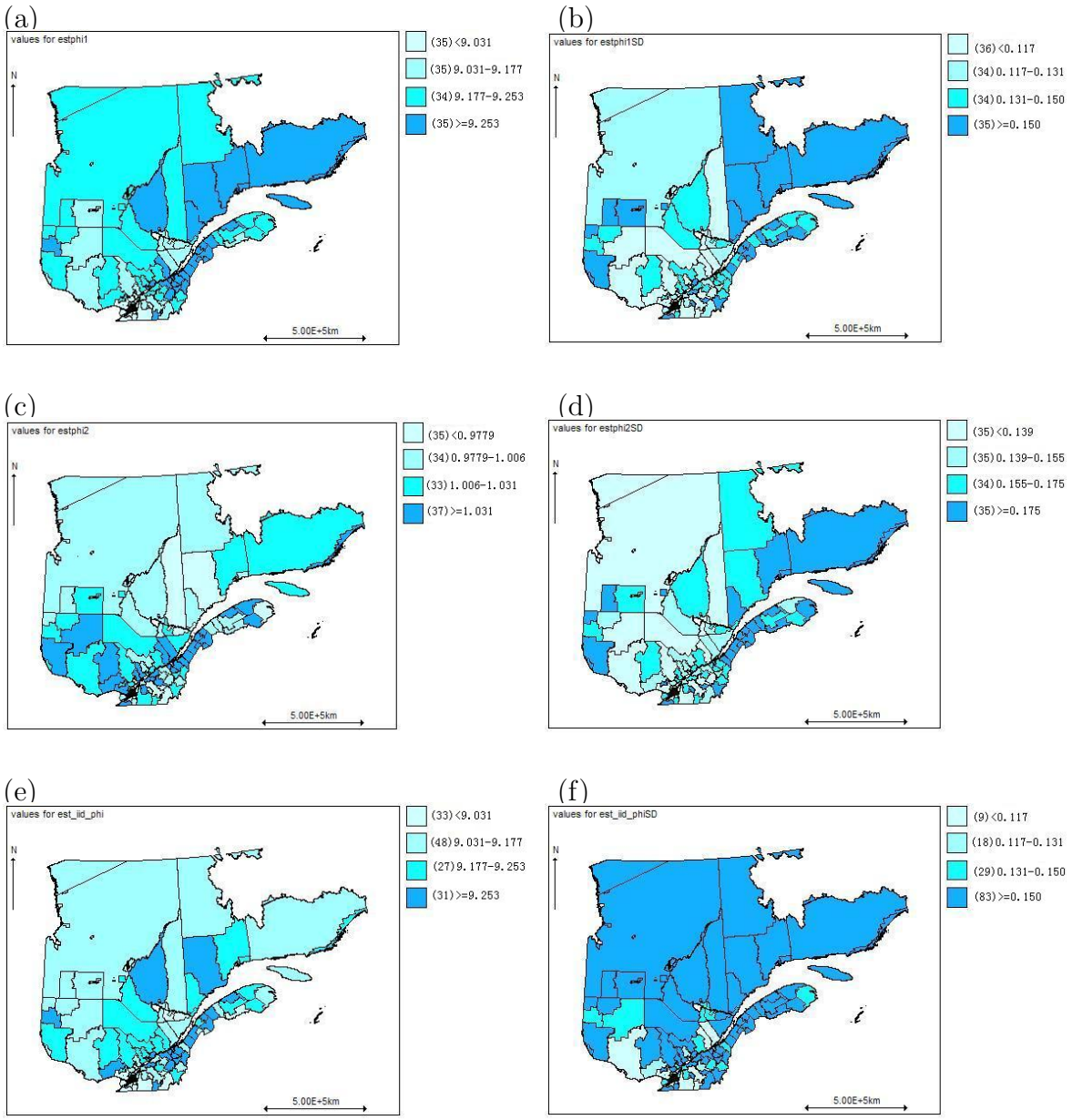


Figure 5.10: (a)Maps of estimated spatial random effects (standard error in (b)) (c) Maps of estimated space-varying treatment effects (standard error in (d)) (e) Maps of estimated i.i.d. random effects (standard error in (f)).

This way, the prior belief on ρ_2 is that it has Beta distribution with mean 0.90 and the variance about 0.004. In our future work, seeking an informative prior for ρ_2 may be an important avenue for further investigation and it may lead to more spatial smoothing on the treatment effect.

5.2.4 Model Checking

In order to assess the goodness-of-fit of our proposed model, we perform a posterior predictive comparison. Following data simulation mechanism described in Section 4.3, we replace the fixed set of parameters with the posterior draws of the parameters for which inference is desired. Suppose we use post burn-in sample of size 5000 from posterior distribution. In a non-censored setup, we generate 5000 sets of (future) data replicates from the posterior predictive distribution using the same covariate values as the observed data. Under the above scheme, we obtain samples from the posterior predictive distribution of the data and a set of percentiles (say 2.5%, 50% and 97.5%) of the replicated data is obtained by retaining the p^{th} percentile of each data-point distribution. Since the replicated data are marginalized over the parameters, fitting a Kaplan-Meier (KM) curve through this set of percentiles of the replicated data is superposed to the KM plot of the observed AMI data. A visual comparison of the goodness-of-fit of our model is then possible.

Figure 5.11 provides a visual assessment of the goodness-of-fit of the log-logistic AFT two i.i.d. CAR model. Overall, the KM-curve of AMI data is very close to that of the p_{50}^{th} of the replicated data with exception to the early survival times. There is no indication of lack of fit in general. By examining the plots within 100 days, we do discover some lack of fit in the early days. This suggests we may use a more flexible error distribution to obtain a better fit. A mixture or other semi-parametric form might be a solution. Alternatively, we could employ different error distributions

for different segments of survival times. In other word, we may assume the baseline hazard function (or survival function) varies over the various time segments. This will be our future work and it will be discussed more in the coming chapter.

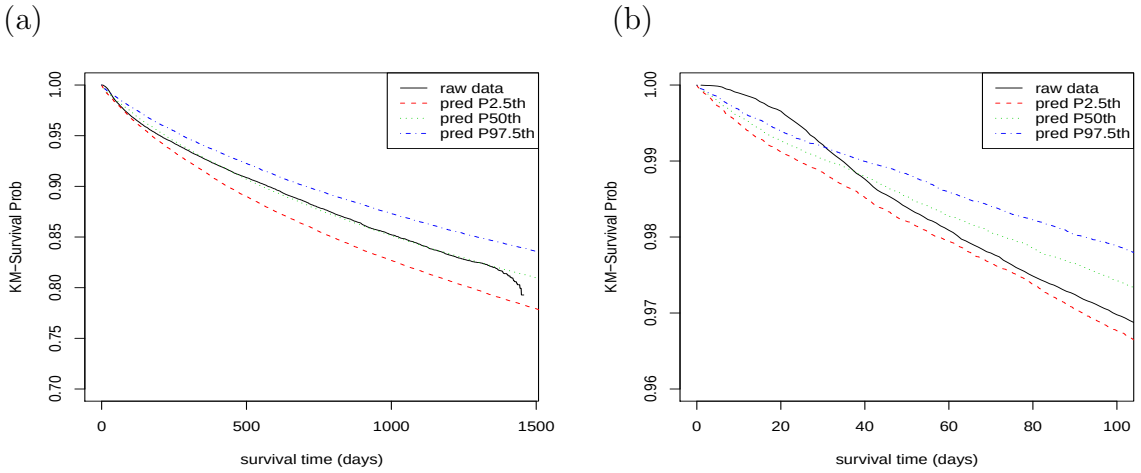


Figure 5.11: (a) Kaplan-Meier fits for the predictive data replicates from the log logistic AFT two i.i.d. CAR model (b) Kaplan-Meier fits for the survival time within 100 days (part of plot (a)).

Chapter 6

Summary and Future Work

In this thesis, we have developed a spatial parametric AFT survival model with space varying coefficients for geographically indexed time to event data. Our model uses two independent CAR distributions to capture correlations across geographical regions and the spatially varying coefficients for a given region. Various competing parametric AFT models are implemented using MCMC computational techniques under a hierarchical Bayesian framework. We illustrated our approach with the analysis of survival data after Acute Myocardial Infarction, which is rather challenging due to its high dimension and spatial nature. We found the approach to be reasonably easy to use and it produced results that should be helpful for further investigating the identified "hot spots". This methodology can also be applied to other spatially distributed data.

Future work on addressing lack of fit in early survival days includes the following. As the model checking indicates that the model seems to be less accurate for early survival days, we need to seek more flexible modeling on the error distribution. First, we could model the error distribution using a mixture of parametric distributions. Take the log-logistic AFT model for example, we can express the density of error terms in the AFT model (2.2.6) as: $f(w|c) = \sum_{j=1}^{K} c_j g_{\sigma_j}(w)$, where $g_{\sigma_j}(w)$ is the density of the logistic distribution with location 0 and scale parameter σ_j and $\mathbf{c} = (c_1, c_2, ..., c_K)$ are a mixture of coefficients that have to be estimated. Values of $\sigma_1, \sigma_2, ..., \sigma_K$ are fixed by design. Second, we may try applying different logistic distributions (or other parametric distributions) to the different segments of survival times. For example,

following an onset of acute myocardial infraction, a patient faces increasing hazard of death over the first couple of hours while waiting for medical interventions and adjunctive therapies and as well as first 30 days after surgeries. The hazard then typically decreases with time. An example of unimodal hazard functions is the log-logistic, which is $h(t) = \frac{\lambda \frac{1}{\sigma} t^{\frac{1}{\sigma}-1}}{1+\lambda t^{\frac{1}{\sigma}}}$ and its corresponding survival function $S(t)$ is listed in Table 2.1. This hazard function decreases monotonically if $\sigma \geq 1$, but if $\sigma \leq 1$, it has a single mode. So, choosing appropriate values of σ can model the changes of the hazard function over time. Third, we could switch from a fully parametric to a semiparametric AFT model. Similar to the parametric AFT model, the hazard function now becomes

$$h_{ij}(t;\mathbf{x}) = \exp(\boldsymbol{\beta}'\mathbf{x_{ij}} + \phi_{0_j} + \phi_{1_j} R_{ij}) h_0(\exp(\boldsymbol{\beta}'\mathbf{x_{ij}} + \phi_{0_j} + \phi_{1_j} R_{ij})t),$$

where the subscript ij represents the i^{th} subject in the j^{th} region, and $\mathbf{x_{ij}}$ are the covariates except treatment which is modeled separately and denoted by R_{ij}. Again, h_0 is a baseline hazard function. We can then model $\log h_0(t)$ using a penalized spline, $\log(h_0(t)) = \sum_{j=1}^{k} a_k B_k(t)$, where $\{B_1(t), ..., B_k(t)\}$ is a set of known cubic B-Spline basis functions based on $k-4$ inner knots which we place along the time axis in a liberal fashion. In addition to the various models on error distribution, we could extend the two i.i.d. CAR to a MCAR model, which appears in various papers, such as Banerjee and Carlin [13] and Jin, Carline and Banerjee [16]. The latter discusses a generalized MCAR model which captures correlations across both geographic regions and multiple random effects, such as frailties and spatially varying coefficients for a given region.

To address the estimation issues of the parameters involved in modeling space-varying treatment effect, in particular regarding ρ_2, we may want to try the followings. First, we will seek an informative prior for ρ_2. A comprehensive sensitivity analysis

on priors will be carried and we will be happy to see under which prior the estimation accuracy of ρ_2 will be greatly improved. Second, we will adopt an MCAR prior for the spatial effects which would allow additional smoothing across the intercept and treatment coefficients. Further investigation into space-varying regression models in general is also required.

Finally, we would like to tack on an open question that how much difference on DIC is significant when selecting a competitive model. In general, DIC is just a rough rule of thumb for model selection. As mentioned in [38], a difference larger than 10 should provide enough evidence in favour of the better model. However, it will be nice to have some formal justification on this. We may achieve this goal by applying a bootstrap method. This will involve a large volume of simulations, through which we obtain the distribution of DIC for various models. Then we can test the significance of the difference of DIC.

References

[1] P. Wilkinson, K. Laji, K. Ranjadayalan, L. Parsons, and A. D. Timmis, Acute myocardial infarction in women: survival analysis in first six months, *British Medical Journal*, **309**(1994) 566-569.

[2] C. Tonne, J. Schwartz, M. Mittleman, S. Melly, H. Suh, and R. Goldberg, Long-term survival after acute myocardial infarction is lower in more deprived neighbourhoods, *Circulation*, **111**(2005) 3063-3070.

[3] C. Tonne, J. Schwartz, M. Mittleman, S. Melly, H. Suh, and R. Goldberg, Magnesium in well water and the spatial variation of acute myocardial infarction incidence in rural Finland, *Applied Geochemistry*, **23**(2008) 632-640.

[4] M. E. Loughnan, N. Nicholls, and N. J. Tapper, Demographic, seasonal, and spatial differences in acute myocardial infarction admissions to hospital in Melbourne Australia, *International Journal of Health Geographics*, **7:42**(2008) doi:10.1186/1476-072X-7-42.

[5] D. Collett, *Modelling Survival Data in Medical Research*, second ed., Chapman & Hall/CRC, New York, 2003.

[6] S. Banejee, B.P Carlin, and A. E. Gelfand, Chapter 9 Spatial survival models, in: *Hierarchical Modeling and Analysis for Spatial Data*, Chapman & Hall/CRC, FL, 2004, pp. 301-334.

[7] J. Lawless, *Statistics Models and Methods for Lifetime data*, Wiley, New York, 1982.

[8] D. R. Cox, Regression models and life tables (with discussion), *J. Royal Statistical Society. B* **34**(1972) 187-220.

[9] J. D. Kalbfleisch and R. L. Prentice, *The Statistical Analysis of Failure Time Data*, second ed., John Wiley Sons, Chichester, 2002.

[10] D. V. Glidden and E. Vittinghoff, Modeling clustered survival data from multicenter clinical trials, *Statistics in Medicine* **23** (2004) 369-388.

[11] Y. Li and L. Ryan, Modeling spatial survival data using the semiparametric frailty models, *Biometrics* **58** (2002) 287-297.

[12] R. Henderson, S. Shimakura, and D. Gorst, Modeling spatial variation in leukemia survival data, *Journal of the American Statistical Association* **97** (2002) 965-972.

[13] S. Banerjee and B. P. Carlin, Hierarchical multivariate CAR models for spatio-temporally correlated survival data (with discussion), *Bayesian Statistics* **7** (2003) 45-63.

[14] S. Banerjee and M. M. Wall, and B. P. Carlin, Frailty modeling for spatially correlated survival data, with application to infant mortality in Minnesota, *Biostatistics* **4** (2003) 123-142.

[15] S. Banerjee and B. P. Carlin, Semiparametric spatiotemporal frailty modeling, *Environmetrics* **14** (2003) 523-535.

[16] X. Jin, B. P. Carlin and S. Banerjee, Generalized hierarchical multivariate CAR models for areal data, *Biometric* **61** (2005) 950-961.

[17] F. Nathoo and C. B. Dean, A mixed mover-stayer model for spatiotemporal two-state process, *Biometric* **63** (2007) 881-891.

[18] R. Christensen and W.O. Johnson, Modeling accelerated failure time with a Dirichlet process, *Biometrika*, **75** (1988) 693-704.

[19] L. Kuo and B. Mallick, Bayesian semiparametric inference for the accelerated failure time model, *The Canadian Journal of Statistics*, **25** (1998) 457-472.

[20] A. Komarek, and E. Lesaffre, Bayesian accelerated failure time model for correlated censored data with a normal mixture as an error distribution, *Statistica Sinica*, **17** (2004) 549-569.

[21] J. P. Kelin and M. L. Moeschberger, *Survival Analysis Techniques for Censored and Truncated Data*, second ed., Springer-Verlag, New York, 2003.

[22] S. Banejee, B.P Carlin, and A. E. Gelfand, *Hierarchical Modeling and Analysis for Spatial Data*, Chapman & Hall/CRC, FL, 2004.

[23] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, *Bayesian Data Analysis*, second ed., Chapman & Hall/CRC, FL, 2003.

[24] J. P. Kelin and M. L. Moeschberger, Chapter 3 Censoring and truncation, in: *Survival Analysis Techniques for Censored and Truncated Data*, second ed., Springer-Verlag, New York, 2003, pp. 63-90.

[25] J. P. Kelin and M. L. Moeschberger, Chapter 8 Semiparametric proportional hazards regression with fixed covariates, in: *Survival Analysis Techniques for Censored and Truncated Data*, second ed., Springer-Verlag, New York, 2003, pp. 243-293.

[26] J. P. Kelin and M. L. Moeschberger, Semiparametric proportional hazards regression with fixed covariates, in: K. Dietz, M. Gail, K. Kricheberg, J. Samet, A. Tsiatis (Eds.), *Survival Analysis Techniques for Censored and Truncated Data*, second ed., Springer-Verlag, New York, 2003, pp. 244-245.

[27] J. P. Kelin and M. L. Moeschberger, Inference for parametric regression models, in : K. Dietz, M. Gail, K. Kricheberg, J. Samet, A. Tsiatis (Eds.), *Survival Analysis Techniques for Censored and Truncated Data*, second ed., Springer-Verlag, New York, 2003, pp. 393-423.

[28] J.W. Vaupel, K.G. Manton, and E. Stallard, The impact of heterogeneity in individual frailty on the Dynamics of Mortality, *Demography* 16, 439-454.

[29] D.G. Clayton, A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence, *Biometrika*, **65** (1978) 141-151.

[30] P. Hougaard, *Analysis of Multivariate Survival Data*, Springer, New York, 2000.

[31] A.I. Yashin. and I.A. Iachine, Genetic analysis of durations: correlated frailty model applied to survival of Danish twins, *Genetic Epidemiology*, **12** (1995) 529-538.

[32] H. Li and X. Zhong, Multivariate survival models induced by genetic frailties, with application to linkage analysis, *Biostatistics*, **3**(2002) 57-75.

[33] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, Chapter 1 Background, in: *Bayesian Data Analsysis*, second ed., Chapman & Hall/CRC, FL, 2003 pp. 3-29.

[34] K. P. Burnham and D. R. AndersonGelman, *Model Selection and Multimodel Inference: A Practical-Theoretic Approach*, second ed., Springer-Verlag, New York, 2002.

[35] R. Kass and L. Wasserman, A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion, *Journal of Amercian Statistics Association*, **90** (1995) 928-934.

[36] D. L. Weakliem, A Critique of the Bayesian Information Criterion for Model Selection, *Sociological Methods & Research*, **27** No. 3 (1999) 359-397.

[37] D. Spiegelhalter, N. Best, B. Carlin, and A. Van der Linde, Bayesian measures of model complexity and fit, *Journal of the Royal Statistical Society, Series B*, **64** (2002) 583-639.

[38] P. Ghosh and G. L. Rosner, A semi-parametric Bayesian approach to average bioequivalence, *Statistics in Medicine*, **26** (2007) 1224-1236.

[39] B. D. Ripley, *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge, 1996, pp. 33.

[40] K.P. Burnham and D.R. Anderson, Chapter 6 in: *Model Selection and Inference*, Springer, New York, 1998.

[41] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, Chapter 6 Model checking and improvement, in: *Bayesian Data Analsysis*, second ed., Chapman & Hall/CRC, FL, 2003 pp. 157-192.

[42] S. Banejee, B.P Carlin, and A. E. Gelfand, Chapter 1, 2 and 3 in: *Hierarchical Modeling and Analysis for Spatial Data*, Chapman & Hall/CRC, FL, 2004, pp. 1-95.

[43] D. C. Wheeler, A comparison of spatial clustering and cluster detection techniques for childhood leukemia incidence in Ohio, 1996 - 2003, *International Journal of Health Geographics*, **6:13** (2007) doi:10.1186/1476-072X-6-13.

[44] J. Ferrándiz, J. J. Abellán, V. Gómez-Rubio, A. López-Quílez, P. Sanmartín, C. Abellá, M. A. Martínez-Beneito, I. Melchor, H. Vanaclocha, Ó. Zurriaga, F. Ballester, J. M. Gil, S. Pérez-Hoyos, and R. Ocaña, Spatial analysis of the relationship between mortality from cardiovascular and cerebrovascular disease and drinking water hardness, *Environmental Health Perspectives*, **112(9)** (2004) 1037-1044, doi: 10.1289/ehp.6737.

[45] D. Brook, On the distinction between the conditional probability and the joint probability approaches in the specification of nearest-neighbour system, *Biometrika*, **51**(1964) 481-483.

[46] N. A. C. Cressie, *Statistics for Spatial Data*, Wiley, New York, 1993.

[47] J. Besag, Spatial interaction and the statistical analysis of lattice systems (with discussion), *Journal of Royal Statistical Society, Series B*, **36**(1974) 192-236.

[48] B. P. Carlin and T. A. Louis, Section 7.8.2 in: *Bayes and Empirical Bayes Methods for Data Analysis*, second ed., Chapman & Hall/CRC, FL, 2004.

[49] L. Bernardinelli and C. Montomoli, Empirical Bayes versus fully Bayesian analysis of geographical variation in disease risk, *Statistics Medicine*, **11** (1992) 983-1007.

[50] M. M. Wall, A close look at the spatial structure implied by the CAR and SAR models, *Journal of Statistical Planning and Inference*, **121**(2004) 311-324.

[51] R. M. Assuncao, Space varying coefficient models for small area data, *Environmetrics*, **14**(2003) 453-473.

[52] C. Brunsdon, S. Fotheringham, M. Charlton, Geographically weighted regression-modeling spatial non-stationarity, *Journal of the Royal Statistical Society, Series D*, **47**(1998) 431-443.

[53] A. D. Pavlov, Space-varying regression coefficients: a semi-parametric approach applied to real estate markets, *Real Estate Economics*, **28**(2000) 249-283.

[54] P. Congdon, A multilevel model for infant health outcomes: maternal risk factors and geographical variation, *the Statistican*, **47**(1998) 159-182.

[55] D. Gamerman and H. F. Lopes, Chapter 4 in: *Markov Chain Monte Carlo: stochastic simulation for Bayesian inference*, second ed., Chapman & Hall/CRC, FL, 2006, pp. 113-139.

[56] G. Casella and E.I. George, Explaining the Gibbs Sampler, *The American Statistician*, **46** (1992) 167-174.

[57] D. Gamerman and H. F. Lopes, Section 5.2 in: *Markov Chain Monte Carlo: stochastic simulation for Bayesian inference*, second ed., Chapman & Hall/CRC, FL, 2006, pp. 142-143.

[58] S. Chib and E. Greenberg, Understanding the Metropolis-Hastings Algorithm, *The American Statistician*, **49** (1995) 327-335.

[59] J. Besag, P. Green, D. Higdon, and K. Mengersen, Bayesian computation and stochastic systems (with discussion), *Statistical Science*, **10**(1995) 3-66.

[60] A. Gelman, G. O. Roberts, and W. R. Gilks, Efficient Metropolis jumping rules, *Bayesian Statistics*, **5**(1996) 599-607.

[61] D. Gamerman and H. F. Lopes, Section 5.4 in: *Markov Chain Monte Carlo: stochastic simulation for Bayesian inference*, second ed., Chapman & Hall/CRC, FL, 2006, pp. 157-169.

[62] A. E. Gelfand and A. F. M. Smith, Sampling-based approaches to calculating marginal densities, *Journal of the American Statistical Association*, **85** (1990) 398-409.

[63] A. Gelman and D. R. Rubin, A single series from the Gibbs sampler provides a false sense of security, *Bayesian Statistics*, **4** (1992a) 625-31.

[64] A. Gelman, Inference and monitoring convergence, in: W. R. Gilks, S. Richardson and D.J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in Practice*, Chapman & Hall, London, 1996, pp. 131-43.

[65] A. E. Gelfand, S. K. Sahu, and B. P. Carlin, Efficient parametrisation for normal linear mixed models, *Biometrika* **82** (1995) 479-88.

[66] P. Grambsch, T. Therneau, Proportional hazards tests and diagnostics based on weighted residuals, *Biometrika*, **81** (1994) 515-526.

[67] J. P. Kelin, M. L. Moeschberger, Regression diagnostics, in: K. Dietz, M. Gail, K. Kricheberg, J. Samet, A. Tsiatis (Eds.), *Survival Analysis Techniques for Censored and Truncated Data*, second ed., Springer-Verlag, New York, 2003, pp. 353-392.

[68] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin, Section 2.9 Noninformative prior distribution, in: *Bayesian Data Analsysis*, second ed., Chapman & Hall/CRC, FL, 2003 pp. 61-65.

Appendix

R Code for Gibbs Sampler Examples

Example 1: Beta-Binomial distribution

```
library(VGAM)
library(DAAG)
#parameter values
n=16
alpha=2
beta=4

M =500 #number of chains
K=10 #size of each chain

#sampling from Beta-binormial distribution using R built-in function
x1= rbetabin.ab (n=M, size=16, alpha=2, beta=4)

#sampling using Gibbs sampler
#initial value
y= 0.5
x=8
sim.x=matrix(0, M, K)
sim.y=matrix(0, M, K)
for (i in 1: M){
    for (j in 2:K){
```

```
        x=rbinom(1, 16, y)
        y=rbeta(1, x+2, 16-x+4)
        sim.x[i,j]=x
        sim.y[i,j]=y
    }
}

#only use the last value of each chain
x2=rep(0,M)
y2=rep(0, M)
for (j in 1:M){
    x2[j]=sim.x[j,K]
    y2[j]=sim.y[j,K]
}

# overlay two histograms for x1 and x2
par(mfrow=c(1, 1))
breaks=seq(0, 16, 0.5)
ind =sort(unique(x1))
freq1=table(x1)
freq2=table(x2)

plot(ind, freq1[ind+1], type="h", xlab="x", ylab="frequency",
    main="", axes=F); axis(1); axis(2)

for (i in 1:length(ind))
    segments(ind[i]+0.1, 0.0, ind[i]+0.1, freq2[i], lty=2)
```

```
legend("topright", legend=c("direct sampling", "Gibbs sampling"), lty=1:2)

#compute the true marginal probability f(x)
sims=M
probm=rep(0,16)
for (i in 1: 16){
    probm[i]=dbetabin.ab(i-1, size=16, alpha=2, beta=4)
}

#compute the estimated marginal probability from the Gibbs samples
ind=sort(unique(x2))
probmhat=table(x2)/sims

#overlay two plots
plot(ind, probm[ind+1], type="h", ylim=c(0,0.13), xlab="x", ylab="density",
    main=" ", axes=F); axis(1); axis(2)

for (i in 1:length(ind))
    legend("topright", legend=c("Exact probabilities", "Estimates of Gibbs samples"),
    lty=1:2)
```

Example 2: truncated exponential distribution

```
# parameter value
B=5
y=2.5 #initial value
M=1000
K=15
sim.x=matrix(0,M,K)
sim.y=matrix(0,M,K)
for (i in 1: M){
    for (j in 1:K){
        u = runif(1)
        x = -log(1-(1-exp(-y*B))*u)/y
        y = -log(1-(1-exp(-x*B))*u)/x
        sim.x[i,j]=x
        sim.y[i,j]=y
    }
}

x=rep(0,M)
y=rep(0,M)
for (j in 1:M){
    x[j]=sim.x[j,K]
    y[j]=sim.y[j,K]
}

#histogram of the simulation using Gibbs sampler
```

```
breaks=seq(0, B, 0.1)
hist(x, breaks, freq=FALSE, col="grey", xlab="x", ylab="density", main=" ")

#estimated marginal density
x.co =seq(0.01, B, 0.1)
estfx = rep(0, length(x.co))
for(i in 1:length(x.co))
    for(j in 1:M){
        estfx[i] = estfx[i] + y[j]*exp(-y[j]*x.co[i])/(1- exp(-y[j]*B))
    } estfx = estfx / M
lines(x.co, estfx)

#True marginal distribution is proportional to (1-exp(-BX))/x
#compute the true density of f(x)
w =0.0001
x1=seq(0.01, B, w)
fx=rep(0,length(x1))
fx=(1-exp(-B*x1))/x1
probx==fx/(sum(fx)*w)

lines(x1,probx,lty=2)
leg.txt =c("Estimated f(x)", "True f(x)")
legend("topright", legend = leg.txt, lty=1:2)
```

R Code for Metropolis-Hasting Algorithms Examples

Example 1: simulating a bivariate normal

```
library(MASS)
library(car)

#sample from R built-in function MVRNORM
Sigma=matrix(c(1, 0.9, 0.9, 1), 2, 2) #variance matrix
Rmvn=mvrnorm(5000,mu=c(1,2),Sigma)

#parameter values same as the above
mu=c(1,2)
inv.Sig=inv(Sigma)

#sampling using M-H algorithm
draw.x=function(x,sigma){
    #z=c(runif(1,1), runif(1)) #use different candidate
    z=c(rnorm(1, 0, sigma), rnorm(1, 0, sigma))
    y=mu-(x-mu)+z
    ratio=exp(-1/2*t(y-mu)%*%inv.Sig%*%(y-mu))/
        exp(-1/2*t(x-mu)%*%inv.Sig%*%(x-mu))
    u=runif(1)
    if (u ≤ ratio) x1=y
    else x1=x
    x.jump= min(ratio,1)
    return(list(x=x1, x.jump=x.jump))
```

```
}

#function to generate M-H sample
metrop.sample =function(x0, nsim, sigma){
sample =matrix(NA, ncol=2, nrow=nsim)
p.jump =matrix(NA, ncol=1, nrow=nsim)
sample[1,]=x0
p.jump[1]=0
for(t in 2:nsim){
    temp=draw.x(sample[t-1,], sigma)
    sample[t,]=temp$x
    p.jump[t]=temp$x.jump
}
print(mean(p.jump))
return(sample)
}
#########Main Program Start Here#########
#initial values for different distributions
#these initial values need to be tuned during the testing

x0=c(0.5,1.5)
#change the initial values corresponding the distribution used
chain1=metrop.sample(x0, 20000, 1)
chain=chain1[10001:20000,] #use burn-in 10000

#scatter plot of data generated from R using Mvrnorm and from M-H
par(mfrow=c(1,2))
```

```
plot(Rmvn, xlim=c(-2,5), xlab='x1', ylab='x2', panel.first=grid())
leg.txt = c("x1: u=1", "x2: u=2")
legend("bottomright", legend = leg.txt)

plot(chain[,1], chain[,2], xlim=c(-2,5), xlab="x1", ylab="x2", panel.first = grid())
leg.txt =c("x1: u=1", "x2: u=2")
legend("bottomright", legend = leg.txt)
```

Example 2: simulating a Bayesian posterior

```
#Simulating a Bayesian Posterior #M-H algorithm to sample an intractable distribution arises in a stationary 2-nd autogressive time series model

library(MASS)
library(car)
library(pscl)

# initial values
phi1=1
phi2=-0.5
phi=c(phi1, phi2)
n=100
y=rep(0, n)
wt=matrix(NA, nrow=2, ncol=n)

sigma=1 #sigma is variance
```

```
#compute the precision matrix
inv.V=matrix(NA, nrow=2,ncol=2)
inv.V[1,1]=1-phi2^2
inv.V[1,2]=-phi1*(1+phi2)
inv.V[2,1]=inv.V[1,2]
inv.V[2,2]=inv.V[1,1]

inv2=inv(inv.V)
y[1:2]=mvrnorm(1, c(0, 0), inv2)

#generate y(t)
for (t in 3: n) {
    e=rnorm(1,0,1)
    y[t]=phi1*y[t-1]+phi2*y[t-2]+e
    wt[,t]=rbind(y[t-1], y[t-2])
}

det=det(inv.V)
Y2=rbind(y[1],y[2])
sum.y2=t(Y2)%*%inv.V%*%Y2

#compute the Psi which used in likelihood function
Psi=1/sigma*det*exp(-1/(2*sigma)*sum.y2)

# likelihood function
lik=function(Psi, y, wt, n, sigma, phi){
    like=Psi*(sigma)^(-(n-2)/2)*exp(-(1/sigma)*sum.yn)
```

```
    return(like)
}

temp=lik(Psi, y, wt, n,sigma, phi)

#phi is the vector of p × 2
psi=function(phi,yt,sigma){
inv=matrix(NA, nrow=2,ncol=2)
inv[1,1]=1-phi[2]^2
inv[1,2]=-phi[1]*(1+phi[2])
inv[2,1]=inv[1,2]
inv[2,2]=inv[1,1]
det=det(inv)
sum.y=t(yt)%*%inv%*%yt
Psi=1/sigma*det*exp(-1/(2*sigma)*sum.y)
return (Psi)
}

temp=psi(phi,Y2,sigma)
G=matrix(rep(0, 4), nrow=2, ncol=2)
sum.wy=matrix(rep(0, 2), nrow=2, ncol=1)

for (t in 3:n){
    G=G+wt[,t]%*%t(wt[,t])
    sum.wy=sum.wy+wt[,t]*y[t]
}
```

```
phi.hat=inv(G)%*%sum.wy

#update phi
phiupdate=function(sigma){
inv.Sig=sigma*inv(G)
I=0
while(I==0){
    phistar=mvrnorm(1, phi.hat, inv.Sig)
#check the stationary condition
    ck1=phistar[1]+phistar[2]
    ck2=-phistar[1]+phistar[2]
    ck3=phistar[2]
    if (ck1< 1 & ck2<1 & ck3>-1){
        I=1
    }
}
ratio=psi(phistar, Y2, sigma)/psi(phi, Y2, sigma)
u=runif(1)
if (u ≤ ratio) {phi=phistar}
jump=min(ratio,1)
return(list(phi=phi,jump=jump))
}

#update sigma
sigmaupdate=function(phi){
inv=matrix(NA, nrow=2, ncol=2)
inv[1,1]=1-phi[2]^2
```

```
inv[1,2]=-phi[1]*(1+phi[2])
inv[2,1]=inv[1,2]
inv[2,2]=inv[1,1]
sum.y2=t(Y2)%*%inv%*%Y2
sum.yn=0
for (t in 3:n){
    sum.yn=sum.yn+(y[t]-t(wt[,t])%*%phi)^2
}

#directly update sigma using inverse gamma
A=49
B=0.5*(sum.y2+sum.yn)
sigma=rigamma(1,A,B)
return (sigma)
}

# initial values
phi.0=c(0.5, -0.2)
sigma.0=0.8

#simulation to draw samples using M-H algorithm
nsim=20000
sample=matrix(NA, ncol=3, nrow=nsim)
p.jump=rep(0, nsim)
sample[1,1:2]= phi.0
sample[1,3]=sigma.0
for (t in 2:nsim){
```

```
    temp=phiupdate(sample[t-1,3])
    sample[t,1:2]=temp$phi
    p.jump[t]=temp$jump
    sample[t,3]=sigmaupdate(sample[t,1:2])
}

#burn-in 10000
post=sample[10001:20000,]
summary(post)

#trace plot and histogram
para = c("phi1","phi2","sigma")
par(mfrow=c(3,2))
for (i in 1:3){
    ts.plot(sample[,i], xlab="iteration", ylab=para[i])
    hist(post[,i], xlab="", main=para[i], prob=T)
}
```

Matlab Code for Log-logistic AFT with Two i.i.d. CAR Models

%**Main Program: gibb.m**

```
clear all
close all
global covar survt cen lhu ind W DW;
% 61054 is the number of observation and 139 is the number of LHUs
[covar,survt,cen,lhu,ind]=getdata(61054,139);
%read the adajacency matrix and number of neighbour;
W =load ('adjmat.txt');
DW=load ('numneigh.txt');
niter=40000;
burn=20000;

%M-H the jump rate of candidates used in M-H steps;
sigmaJumpmu=0.17;
%jump rate for beta;
sigmaJumpb=[0.003, 0.06];
sigmaJumptrt=0.1;
sigmaJumpscale=0.04;

%jump rate for random effect b0;
sigmaJumpb0=0.17;
sigmaJumprho=2.2;

    %jump rate for space varying coefficients;
```

```
sigmaJumpb0t=0.35;
sigmaJumprhotrt=2.2;
sigmaJump=[sigmaJumpmu, sigmaJumpb, sigmaJumptrt, sigmaJumpscale, sigma-
Jumpb0, sigmaJumprho, sigmaJumpb0t, sigmaJumprhotrt];

%initial values
init=[8, -0.09, -0.2, 0.8, 1.4, 0.05, 0.5, 0.05, 0.5];
para=2*139+9;
sims=zeros(niter,para);
logpi=zeros(niter,1);

pararate=7+2*139;
jumprate=zeros(niter,pararate);

a0=init(1);
beta=[init(2:3)];
trtmu=init(4);
scale=init(5);
tau=init(6);
rho=init(7);
tautrt=init(8);
rhotrt=init(9);
%initial values for random effect;
b0=ones(139,1)*init(1);
%initial values for trt coefficient ;
b0trt=ones(139,1)*init(4);
sims(1,:)=[init, b0', b0trt'];
```

```
%M-H inside Gibbs Step for updating the parameter one-by-one;

for k=1:niter
    [a0, pjumpmu]=phimuupdate(a0, b0, tau, rho, sigmaJump(1));
    [beta, pjumpb]=betaupdate(beta, scale, b0, b0trt, sigmaJump(2:3));
    [trtmu, pjumptrt]=phimuupdate(trtmu, b0trt, tautrt, rhotrt, sigmaJump(4));
    [scale, pjumpscale]=scaleupdate(beta, scale, b0, b0trt, sigmaJump(5));

    %update tau directly using gamma distribution
    tau=tauupdate(b0,a0,tau,rho);
    [b0, pjumpb0]=b0update(a0,beta,scale,b0,b0trt,tau,rho,sigmaJump(6));
    [rho, pjumprho]=rhoupdate(b0,a0,tau,rho,sigmaJump(7));

    tautrt=tauupdate(b0trt,trtmu,tautrt,rhotrt);
    [b0trt, pjumpb0trt]=b0trtupdate(trtmu, beta, scale, b0, b0trt,
                tautrt, rhotrt, sigmaJump(8));
    [rhotrt,pjumprhotrt]=rhoupdate(b0trt, trtmu, tautrt, rhotrt, sigmaJump(9));

    sims(k,:)=[a0,beta, trtmu, 1/scale, tau, rho, tautrt, rhotrt, b0', b0trt'];
    jumprate(k,:)=[pjumpmu, pjumpb, pjumptrt, pjumpscale,
            pjumprho, pjumprhotrt, pjumpb0, pjumpb0trt];
    logpi(k)=loglik(beta, scale, b0, b0trt, 0);
end
sumplot(sims, jumprate, logpi, burn, niter);
```

%Function: getdata.m

```
function [covar,survt,cen,lhu,ind]=getdata(N,numLHU)
global covar survt cen lhu ind;
covar=zeros(N,3);

[covar(:,1),covar(:,2),covar(:,3),survt,cen,lhu]=textread('thesisdatalhu.txt','%d %d %d
%f %d %d','headerlines',1);
meanAge=mean(covar(:,1));
for i=1:length(covar(:,1))
    covar(i,1)=covar(i,1)-meanAge;
end
%find the start and end index for each LHU
ind=zeros(numLHU,2);
for k=1:numLHU
    ind(k,1)=min(find(lhu==k));
    ind(k,2)=max(find(lhu==k));
end
```

%**Function: phimuupdate.m**

```
function [phimuin,pjump]=phimuupdate(phimuin, b0in, tauin, rhoin, sigmaJ)
global DW W;
bstar=normrnd(phimuin,sigmaJ);
mu=ones(139,1)*phimuin;
mustar=ones(139,1)*bstar;
Sigma=tauin*(DW-rhoin*W)^(-1);
r=normpdf(bstar,0,100)*mvnpdf(b0in,mustar,Sigma)/normpdf(phimuin,0,100)
        /mvnpdf(b0in,mu,Sigma);
```

```
if unifrnd(0,1)≤ r
    phimuin=bstar;
end
pjump=min(r,1);
```

% **Function: betaupdate.m**

```
function [beta,pjump]=betaupdate(beta,scale,b0,b0trt,sigmaJumpb)
pjump=zeros(1,length(beta));
for i=1:length(beta)
    newb=beta;
    bstar=normrnd(beta(i),sigmaJumpb(i));
    logpostold=loglik(beta,scale,b0,b0trt,0)+log(normpdf(beta(i),0,100));
    newb(i)=bstar;
    logpoststar=loglik(newb,scale,b0,b0trt,0)+log(normpdf(bstar,0,100));
    r=exp(logpoststar-logpostold);
    if unifrnd(0,1) ≤ r
        beta=newb;
    end
pjump(i)=min(r,1);
end
```

% **Function: scaleupdate.m**

```
function [scale,pjump]=scaleupdate(beta, scale, b0, b0trt, sigmaJumpscale)
lscalestar=normrnd(log(scale),sigmaJumpscale);
scalestar=exp(lscalestar);
logpostold=loglik(beta,scale,b0,b0trt,0)+log(gampdf(scale,0.01,0.01));
logpoststar=loglik(beta,scalestar,b0,b0trt,0)+log(gampdf(scalestar,0.01,0.01));
```

```
r=exp(logpoststar-logpostold)*scalestar/scale;
if unifrnd(0,1)≤r
    scale=scalestar;
end
pjump=min(r,1);
```

%**Function: tauupdate.m**

```
tauin=tauupdate(b0in,phimuin,tauin,rhoin)
global DW W;
A=0.01+0.5*length(b0in);
mu0=ones(139,1)*phimuin;
B=0.01+0.5*(b0in-mu0)'*(DW-rhoin*W)*(b0in-mu0);
tauin=iwishrnd(2*B,2*A); %use the inverse Wishard here
```

%**Function: rhoupdate.m**

```
function [rhoin,pjump]=rhoupdate(b0in, phimuin, tauin, rhoin,sigmarho)
global DW W;
%change 1 to 0.999 avoiding division by 0;
max=0.999;
logitrho=log(rhoin/(max-rhoin)); %logistic transform for rho
logitRhostar=normrnd(logitrho,sigmarho);
rhostar=max*(1-1.0/(1+exp(logitRhostar)));
Sigmaold=tauin*(DW-rhoin*W)^(-1);
Sigmastar=tauin*(DW-rhostar*W)^(-1);
mu0=ones(139,1)*phimuin;
logpostold=log(mvnpdf(b0in,mu0,Sigmaold))+log(rhoin*(max-rhoin));
logpoststar=log(mvnpdf(b0in,mu0,Sigmastar))+log(rhostar*(max-rhostar));
```

```
r=exp(logpoststar-logpostold);
if unifrnd(0,1)≤r
    rhoin=rhostar;
end
pjump=min(r,1);
```

%**Function: loglik.m**

```
function Loglik=loglik(beta,scale,b0,b0trt,likall)
global covar survt cen lhu ind;
%two kind of random effect b0–spatial b0.trt–spatial trt coeff;
%the parameter likall is the indicator likall=0 return overall likehood
% otherwise return the likelihood for that LHU only
if likall==0
    L=zeros(139,1);
    for k=1:139
        %find the start and end index where LHU==K
        I1=ind(k,1);
        I2=ind(k,2);
        ck=cen(I1:I2);
        cok=covar(I1:I2,1:2);
        trt=covar(I1:I2,3);
        sur=survt(I1:I2);
        mu=zeros(1,length(sur));
        mu=(cok*beta'+b0(k)+trt*b0trt(k)).*scale;
    L(k)=sum(ck.*(log(scale)-mu+log(sur).*(scale-1)-log(1+exp(-mu).*sur.^scale))
                -log(1+exp(-mu).*sur.^scale));
    end
```

```
    Loglik=sum(L);
elseif likall> 0
    k=likall;
    I1=ind(k,1);
    I2=ind(k,2);
    ck=cen(I1:I2);
    %pick up covar for LHU==k
    cok=covar(I1:I2,1:2);
    trt=covar(I1:I2,3);
    sur=survt(I1:I2);
    mu=zeros(1,length(sur));
    mu=(cok*beta'+b0(k)+trt*b0trt(k)).*scale;
Loglik=sum(ck.*(log(scale)-mu+log(sur).*(scale-1)-log(1+exp(-mu).*sur.^scale))
            -log(1+exp(-mu).*sur.^scale));
end
```

%**Function: sumplot.m**

```
function sumplot(sims,jumprate,logpi,burn,niter)
figure(1)
subplot(3,3,1);
plot(sims(:,1))
title ('Trace plot of mu')
subplot(3,3,2);
plot(sims(:,2))
title ('Trace plot of age')
subplot(3,3,3);
plot(sims(:,3))
```

```
title ('Trace plot of gender')
subplot(3,3,4);
plot(sims(:,4))
title ('Trace plot of fix effect trt')
subplot(3,3,5);
plot(sims(:,5))
title ('Trace plot of sigma')
subplot(3,3,6);
plot(sims(:,6))
title ('Trace plot of tau')
subplot(3,3,7);
plot(sims(:,7))
title ('Trace plot of rho')
subplot(3,3,8);
plot(sims(:,8))
title ('Trace plot of tau.trt')
subplot(3,3,9);
plot(sims(:,9))
title ('Trace plot of rho.trt')
print -dpdf 'coefftracechain.pdf';

sumstat=zeros(9,5);
sumb0=zeros(139,5);
sumb0trt=zeros(139,5);
sample=sims(burn:niter,:);
thin=10;
num=(niter-burn)/thin;
```

```
savedsim=zeros(num,9+139*2);

for j=1:(niter-burn)
    if mod(j,thin)==0
        savedsim(j/thin,:)=sample(j,:);
    end
end

accprate=zeros(1,7);
b0accprate=zeros(139,1);
b0trtaccprate=zeros(139,1);

for i=1:7
    accprate(i)=mean(jumprate(:,i));
end
disp('############Acceptance Rate ############');
disp(' a0 b1.age b2.gender b3.trt scale rho rho.trt') ;
disp(accprate);

for i=1:139
    b0accprate(i)=mean(jumprate(:,i+7));
    b0trtaccprate(i)=mean(jumprate(:,i+7+139));
end
minb0rate=min(b0accprate);
maxb0rate=max(b0accprate);

minb0trtrate=min(b0trtaccprate);
```

```
maxb0trtrate=max(b0trtaccprate);
disp('### MIN and MAX b0 , MIN and MAX b0.trt Acceptance Rate ###');
disp([minb0rate,maxb0rate, minb0trtrate, maxb0trtrate]) ;
p=[0.025,0.5,0.975];
for i=1:9      sumstat(i,1)=mean(sample(:,i));
    sumstat(i,2)=std(sample(:,i));
    sumstat(i,3:5)=quantile(sample(:,i),p,3);
end
for i=1:139
    sumb0(i,1)=mean(sample(:,9+i));
    sumb0(i,2)=std(sample(:,9+i));
    sumb0(i,3:5)=quantile(sample(:,9+i),p,3);

    sumb0trt(i,1)=mean(sample(:,9+i+139));
    sumb0trt(i,2)=std(sample(:,9+i+139));
    sumb0trt(i,3:5)=quantile(sample(:,9+i+139),p,3);
end

%saved the simulation draws
save sumstat.txt sumstat -ascii;
save sumb0.txt sumb0 -ascii;
save sumb0trt.txt sumb0trt -ascii;
save savedsim.txt savedsim -ascii;

% posterior mean of each parameter
postbeta=zeros(2,1);
postb0=zeros(139,1);
```

```
postb0trt=zeros(139,1);
posta0=sumstat(1,1);
postbeta=sumstat(2:3,1);
posttrtmu=sumstat(4,1);
postsigma=sumstat(5,1);
posttau=sumstat(6,1);
postrho=sumstat(7,1);
posttautrt=sumstat(8,1);
postrhotrt=sumstat(9,1);
postb0=sumb0(:,1);
postb0trt=sumb0trt(:,1);
totsum=loglik(postbeta',1/postsigma,postb0,postb0trt,0);
dhat=-2*totsum;
dbar=-2*sum(logpi(burn:niter))/(niter-burn+1);

% compute the DIC pD=dbar-dhat;
DIC=dbar+pD;
DICOUT=zeros(1,4);
DICOUT=[dbar,dhat,pD,DIC];
save DICOUT.txt DICOUT -ascii;

%histogram
figure(2)
subplot(3,3,1);
hist(sample(:,1)) title ('histogram of est of mu')
subplot(3,3,2);
hist(sample(:,2)) title ('histogram of coeff of age')
```

```
subplot(3,3,3);
hist(sample(:,3))
title ('histogram of coeff of gender')
subplot(3,3,4);
hist(sample(:,4))
title ('histogram of est of trt')
subplot(3,3,5);
hist(sample(:,5))
title ('histogram of est of sigma')
subplot(3,3,6);
hist(sample(:,6))
title ('histogram of est of tau')
subplot(3,3,7);
hist(sample(:,7))
title ('histogram of est of rho')
subplot(3,3,8);
hist(sample(:,8))
title ('histogram of est of tau.trt')
subplot(3,3,9);
hist(sample(:,9))
title ('histogram of est of rho.trt')
print -dpdf 'coeffhistchain.pdf';
```