UNIVERSITY OF CALIFORNIA,
IRVINE

Statistical Methods for Quantifying Spatial Effects on Disease Incidence Using Individual-Level Data

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Statistics

by

Lu Bai

Dissertation Committee:
Professor Daniel L. Gillen, Chair
Professor Yaming Yu
Professor Zhaoxia Yu

2016

DEDICATION

For my husband, Siming, who gave me unlimited support and encouragement during exciting and frustrated times; and for my daughter, Iris, who has made me stronger, better and more fulfilled than I could have ever imagined. I love you to the moon and back.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank the members of my dissertation committee, Professor Yaming Yu and Professor Zhaoxia Yu, for their service and insights. I would also like to thank Professor Viera Veronica and Professor Scott Bartell for their support and allowing me to collaborate with them on their scientific projects. Thank you to all the professors in the department for their teaching and and the many discussions provided me with the foundational knowledge of Statistics that paved the way for the research presented here.

I would like to especially thank my advisor, Professor Daniel Gillen, for his support during my time in graduate school. His dedication to Statistics and my education has been marvelous. Without his patience, guidance and encouragement, I cannot imagine how my Ph.D life would have turned out. Our weekly meetings were always insightful and I will truly miss them.

I also thank all the graduate students for their friendship my time as a graduate student. Thanks to Rosemary for administrative support throughout the past years.

Last but not least, I would like to thank my husband, Siming, and my daughter, Iris, who has been by my side through thick and thin these past years.

CURRICULUM VITAE

Lu Bai

EDUCATION

Doctor of Philosophy in Statistics **2016**
University of California, Irvine *Irinve, California*

Master of Science in Electrical Engineering **2011**
Tsinghua University *Beijing, China*

Bachelor of Science in Electrical Engineering **2008**
Tsinghua University *Beijing, China*

RESEARCH EXPERIENCE

Graduate Research Assistant **2007–2012**
University of California, Irvine *Irvine, California*

Biostats Intern **Summer 2015**
Genentech, Inc *South San Fransisco, California*

TEACHING EXPERIENCE

Teaching Assistant **2014–2016**
University of California, Irvine *Irvine, California*

Manuscripts

- Robert E. Bristow, Jenny Chang, Argyrios Ziogas, Daniel L. Gillen, Lu Bai,Veronica M. Vieira, Spatial Analysis of Advanced-Stage Ovarian Cancer Mortality in California, American Journal of Obstetrics and Gynecology, 2015 Jul; 213(1):43.e1-8
- Lu Bai, Daniel L. Gillen, Scott M. Bartell, Verónica Vieira, "Mapping smoothed spatial effect estimates from individual-level data: `MapGAM` "
- Lu Bai, Daniel Gillen, "Spatial Analysis of Disease Outcomes via Bayesian Adaptive Thin-Plate Smoothing Splines"
- Lu Bai, Daniel Gillen, "Spatial Analysis of Right-Censored Survival Data via Bayesian Adaptive Thin-Plate Smoothing Splines"

- Lu Bai, Zhaoxia Yu, Daniel Gillen, Empirical Shrinkage Methods for Family-based Association Tests of Quantitative Traits.
- Lu Bai, Daniel, Gillen, "Flexible Group Sequential Monitoring Methods that Incorporate Covariate Adjustment in Observational Studies"

SOFTWARE

MapGAM `RPackage`

Mapping Smoothed Effect Estimates from Individual-Level Data

ABSTRACT OF THE DISSERTATION

Statistical Methods for Quantifying Spatial Effects on Disease Incidence
Using Individual-Level Data

By

Lu Bai

Doctor of Philosophy in Statistics

University of California, Irvine, 2016

Professor Daniel L. Gillen, Chair

In epidemiologic studies, researchers are commonly interested in quantifying geospatial effects on the incidence of disease to illustrate health disparities potentially attributable to environmental, demographic, and/or socioeconomic factors. The current research we focus on the development of advanced statistical modeling methods to determine how the pattern of disease outcome changes over geographical location.

We first extend and illustrate the utility of `MapGAM`, a user-friendly R package that provides a unified framework for estimating, predicting and drawing inference on covariate-adjusted spatial effects using individual-level data. The package also facilitates visualization of spatial effects via automated mapping procedures. `MapGAM` estimates covariate-adjusted spatial associations with a univariate outcome using generalized additive models that include a non-parametric bivariate smooth term of geolocation parameters. Estimation and mapping methods are implemented for continuous, discrete, and right-censored survival data.

Next, we note that smoothing approaches commonly implemented in generalized additive models and used in spatial analyses assume the amount of smoothing to be equal across geographical regions. The result is that some regions tend to be under-smoothed, while others are over-smoothed. We extend the work of Yue et al. (2010) in the context of brain imaging

analysis and propose a hierarchical Bayesian adaptive thin-plate spline that allows for spatial smoothing of continuous, binary and count outcomes along with the ability to adjust for potential confounding factors. The proposed method allows for the amount of smoothing to flexibly vary depending on the local extent of spatial effect by using nonstationary spatial Gaussian Markov random fields. Performance of the approach is evaluated via simulation and the proposed methodology is applied to an epidemiologic study investigating spatial heterogeneity in the risk of preterm birth among Massachusetts residents.

Finally, we further extend the proposed adaptive smoothing techniques to the case of right-censored survival outcome data. Simulation is used to compare the performance of the proposed method to commonly used non-adaptive smoothing method, and the approach is applied to data from an epidemiology study seeking to quantify spatial variability in the survival times of advanced-stage ovarian cancer patients in California.

Chapter 1

Motivation and Introduction

1.1 Spatial Analysis for Epidemiology studies

Spatial epidemiology is the description and analysis of geographically indexed health data with respect to demographic, environmental, health resource, behavioral, socioeconomic, genetic, and infectious risk factors(Elliott and Wartenberg, 2004). In spatial epidemiology studies, mapping spatial distributions of disease incidence and prevalence has long been used for describing spatial patterns (in both extend and magnitude) of risk, identifying risk factors of public health concern and predicting disease outcomes in different geographical areas (Koch, 2005; Stevens and Pfeiffer, 2011). Among the myriad of uses, spatial epidemiologic studies can provide support and references for public health communities and decision makers to plan intervention strategies, propose relevant policies, reallocate health resource utilities, identify geographical clusters with higher or lower disease risk or rate and design future studies.

For epidemiology studies, cases can not be randomly allocated to different geographical areas. Thus, although the underlying (or crude) geographic pattern of disease can be observed

by public health practitioners, these patterns may reflect a spurious association between the disease outcome and the exposure of interest due to the confounding of other important spatially-varying predictors such as socioeconomic status, race/ethnicity, or environmental exposures. Put more clearly, people tend to cluster in space in systematic ways that may be highly predictive of disease outcome. For example, individuals of high socioeconomic status tend to live near others with high incomes and in areas with better housing and schooling than those in lower-income areas. Individuals with higher incomes are more likely to engage in leisure-time exercise and maintain healthier diets, and less likely to be smokers. As each of these characteristics tend to be favorable risk factor profiles for most health outcomes, a consequence is that individuals residing in more affluent areas till tend to have better health (Smith et al., 1996). Spatial analyses can provide insight in disease risk by simultaneously adjusting for potential confounding variables while accounting for heterogenous spatial distributions. Thus, adjusted disease risk after controlling for potential spatially-varying confounders is important and essential identifying health disparities and furthering the pursuit of increased public health.

Traditional reporting of disease occurrence is usually performed at a national or regional scale. However, advanced data collection systems and analytic methods allow for improved disease map resolution thereby allowing for the quantification of disease risks and/or rates on a smaller-area scale. The end result is that investigators are able to have a better understanding of the relationship between the disease outcome and locally relevant health risk factors such as exposures to local environmental pollution sources, local socioeconomic and behavioral factors, and utilization of local health care resources, to name a few. That being said, higher resolution disease mapping implies that the scale of the study is narrowed to a smaller area and hence a reduction in the size of the underlying population used for estimation. The end result of utilizing a smaller sample size will thus result in higher local variation of statistical estimates, and this local variation may lead to increased bias in estimates and invalid inference if not properly accounted for in the underlying spatial model. As such,

maintaining geographic resolution and providing reliable statistical estimates are at odds with one another in the analysis of spatial effects.

In the age of "Big Data" investigators are now more likely than ever to collect individual-level data from disease registries and other electronic databases. However, due to the lack of available statistical methods to easily and appropriately balance data sparsity that inherently comes with the analysis of individual-level spatial data, it is still common for epidemiologists to produce disease maps at aggregated area levels by county, census contract or other geographical divisions, In these cases, aggregated event counts within administrative districts or regions are often coupled with potentially relevant background and covariate information at summarized at the area level (eg. mean and/or quantile summaries by region). To estimate spatial effect in such settings, mixed effect models are commonly used by introducing a spatially correlated random effect (Clayton and Kaldor, 1987) to the intercept. With appropriate spatial correlation structure, the local estimate is able borrow strength from nearby or neighboring areas to reduce variation of estimates due to small sample sizes. Bayesian formulations of the mixed effect model (Besag et al., 1991) as well as how to structure the covariance among spatial effects (Cressie, 1993; Waller and Gotway, 2004; Diggle et al., 1998; Banerjee et al., 2004; Lu et al., 2007) have been widely discussed in the statistical literature. However, such maps still produce poor spatial resolution by aggregating disease cases, and large between-area differences in precision. Moreover, use of area-level covariates as proxies for individual-level covariates may not control confounding, causing cross-level or aggregation bias(Greenland, 1992; Pardo-Crespo et al., 2013). Finally, registries typically record residence at the time of diagnosis. If the outcome of interest is characterized by long latency periods, important exposures may have occurred many years prior while the case resided at a different location. Due to misclassification in exposure status, maps that ignore latency may tend to be flatter if population movement is random with respect to disease status the non-differential misclassification error will result in spatial effects that are attenuated towards a the null (Polissar, 1980).

In this thesis, we focus on population-based studies with individual-level or point-based data that can address the above problems previously discussed in area-based studies. As will be described with the introduction of the examples that motivate our research, the residential location of each subject is geocoded, and detailed individual-level potential risk factors are collected. Thus, the disease map generated from point-based data will yield higher resolution, and adjusting for individual-level predictors will permit much better control of confounding. In addition, such data can provide detailed information on residential history so that epidemiologists can account for latency by mapping where people lived for specified lengths of time prior to the occurrence or diagnosis of the outcome of interest.

However, estimating and mapping spatial effects from individual-level disease data requires different statistical models from those commonly employed in area-level data analyses. For point-based data, unless the data are aggregated back into areas, simple stratification and standardization methods are unappealing due to inaccuracy in estimating effects over multiple small strata (Webster et al., 2006) In these cases, smoothing provides an efficient method to reduce variability by borrowing strength from adjacent areas while allowing for non-parametric flexibility when estimating the spatial distribution. Generalized additive models (GAMs), originally proposed by Hastie and Tibshirani (1986), are common model-based approaches for mapping point-based epidemiologic data (Webster et al., 2006; Vieira et al., 2008; Baker et al., 2011; Akullian et al., 2014; Bristow et al., 2014; Hoffman et al., 2015). As will be discussed in more detail in Chapter 2, GAMs extend the generalized linear models (GLMs) paradigm by replacing the usual linear predictor in these models with an additive predictor. The result is that the effect of a selected covariate (or a set of covariates) can be nonlinear while the effects of of other covariates can be left additive. In the context of a spatial analysis, GAMs generally include a smooth term of the spatial parameters and a linear predictor for adjustment covariates. The smooth term is then used to describe the (generally) complex relationship between the outcome and the spatial parameters without imposing any specific parametric forms on the relationship. Separation of the smoothed

spatial effect and the assumed additive linear effect of other adjusted covariates, results in a convenient and efficient method confounding adjustment. While GAMs provide a unified statistical framework that allows for smoothing disease risk over geographic areas using individual-level data, the motivating examples presented next highlight deficiencies in these models that lead to the methodologic developments presented in the thesis. Namely, there does not exist an implementation of GAMs for bivariate smoothing terms in the context of censored survival data as considered in the California ovarian cancer study discussed below. Next, our research will demonstrate that the current implementation of GAMs for generalized linear models can results biased and/or highly variable estimates of spatial effects due to the use of a spatially invariant smoother parameter. This is illustrated in the context of both the Massachusetts preterm birth risk study and the California ovarian cancer study. These motivating examples reveal the need for further statistical methods to analyze continuous, discrete, and censored outcomes while incorporating adaptive smoothing techniques for spatial effects and the adjustment of potential confounding factors based on individual-level data.

1.2 Motivating Examples

We begin by briefly introducing two epidemiologic studies seeking to quantify heterogeneity in the incidence of co-mobordities and mortality. The primary objective of both studies is to estimate the effect of location on the risk of clinical outcomes that cannot be explained by other measurable covariates. If achieved, the findings would be useful in identifying potential health disparities among differentially susceptible subpopulations. A common characteristic of both studies is that each involves the collection of population-based data from two different U.S. states resulting in non-uniformly distributed observations over irregular spatial boundaries. As presented throughout the remainder of the thesis, these common features

lead to complex statistical issues regarding the optimal way one can flexibly estimate geospatial effects on the risks of disease, hence motivating the statistical methodology developed and presented here.

1.2.1 Preterm Birth Risk Study

We first consider a population-based study of the risk of preterm birth using data from the state of Massachusetts. Data were obtained on all live and still births from the Massachusetts state birth registry with an estimated conception date from 1/1/01 through 12/31/08. The primary outcome for the study is represented as a binary indicator of whether each observed birth was considered to be preterm, defined as a gestational age of less than 37 weeks. Family and maternal characteristics were obtained for each participant and included maternal age, maternal race (non-hispanic white, non-hispanic black, hispanic, Asian and other), insurance (managed care, medicare, medicaid, other insurance and not insured), maternal education (<12th grade, high school graduation and college), smoking during pregnancy (yes/no), alcohol during pregnancy (yes/no), median income of block group (median household income of census block groups), father education (<12th grade, high school graduation and college), and parity. The primary aim of the study is to identify the effect of living geographical location on preterm birth risk after controlling for the potential confounding factors described above. To this end, geospatial location was defined and collected as the longitude and latitude coordinates of the primary residence of the expectant mother during the time of pregnancy.

To best summarize spatial heterogeneity on the risk of preterm birth it is necessary to flexibly estimate the odds preterm birth as a function of geospatial location while adjusting for those factors *a priori* considered to be be potential confounding factors in the relationship of interest. Further, to communicate the complex relationship between location and the odds of preterm birth to the broader scientific community, spatial disease heat maps will be

necessary. Chapter 2 presents currently implemented generalized additive models (GAMs) that are used to produce the above estimates via non-parametric spatial smoothers. However, as elaborated in Chapter 4, the current methodology assumes a common smoothing parameter across the entire geographic region which, as will be demonstrated, produces biased and/or highly variable estimates of the odds of preterm birth as a function of geospatial location. To overcome this gap in the statistical literature, we propose a novel hierarchical Bayesian generalized linear model in Chapter 4 that produces reliable statistical estimates of confounding-adjusted spatial effects via adaptive smoothing. The end result is a general methodology that can be used to model continuous, binary, or count outcome data in the presence of confounding factors while utilizing spatially adaptive smoothing parameters to reduce bias and variability in estimated geospatial effects.

1.2.2 Ovarian Cancer Study

In the United States, there are 22,000 new cases of ovarian cancer diagnosed and over 14,000 disease-related deaths annually. These cases account for more deaths than all other gynecologic cancers combined (Siegel et al., 2014). Survival determinants are known to be multifactorial but are still not fully understood. Known correlates of patient survival include patient disease history, stage of disease, and the treatment regime used to fight the disease. Among the less understood potential risk factors is spatial location, which may serve as a proxy for patient access to care. To quantify this potential association, we designed and carried out a retrospective population-based study to analyze the effect of geographic variation on advanced-stage invasive epithelial ovarian cancer mortality (Bristow et al., 2014). Cases were obtained from the California Cancer Registry, a comprehensive population-based statewide registry. All participants were reported with International Federation of Gynecology and Obstetrics (FIGO) stage III/IV disease during the time period ranging from 1/1/1996 through 12/31/2006. The primary outcome of interest is represented by patient

survival time, defined as the time from diagnosis to the time of death. Patients that were not observed to expire over the course of followup were censored and contribute partial information to the outcome, represented as the last known time of survival prior to being censored. The study recorded mutlitple known survival determinant factors including patients' demographic characteristics (age, race, insurance type and socioeconomic status score), tumor characteristics (FIGO stage, grade, histology and tumor size) and utilized health care delivery system characteristics (whether the hospital where treatment was received treats more than 20 cases per year and whether the treatment received was adherent to National Comprehensive Cancer Network treatment guidelines for advanced-stage ovarian cancer). The primary goal of the study is to examine whether geolocation has an independent effect on patient survival time after adjustment for the above recorded known correlates of mortality in ovarian cancer patients, and to quantify how the adjusted survival rate changes over different geolocations across California. Ultimately, the adjusted geospatial effect will then reflect the effect of underlying spatially-varying factors that are not adjusted in the model, which may be related to general health disparities including health care resource utilization, environmental exposures, and lifestyle choices among others. For the purpose of the study, geospatial location was defined and collected as the longitude and latitude coordinates of the primary residence of the case at the time of diagnosis.

Typically, the Cox proportional hazards model (Cox, 1972, 1975) is used to examine the covariate-adjusted effect of a predictor of on the risk of mortality for censored survival data. However, the spatial effect considered in the California ovarian cancer data is likely to be complex and the assumption of linearity of the effect or a non-linear effect with a specific functional form would likely lead to bias in estimated spatial associations. As such it is necessary to non-parametrically smooth the effect of geospatial location on the risk of mortality. As noted in Chapter 2 currently implemented GAM models that allow for smoothed estimation of covariate effects do not incorporate bivariate smoothing of predictors for the analysis of censored survival data. To address this gap in the statistical literature and

software, Chapter 3 extends develops a Cox proportional hazards GAM that is capable of simultaneously adjusting for confounders and estimating the complex associations between survival and geospatial location. While the Cox proportional hazards GAM model developed in Chapter 3 provides analysts with easily implementable approach for flexibly estimating bivariate spatial effects, the use of the GAM with a time-invariant smoothing parameter can lead to biased and or highly variable estimates of the relative risk of death as a function of geospatial location, similar to that described in the context of the Massachusetts data above. To address this further gap in statistical methodology, in Chapter 5 we extend the adaptively smoothed Bayesian hierarchical model for GLMs that is presented in Chapter 4 to incorporate censoring and hence the analysis of censored survival times as considered the California ovarian cancer data.

1.3 Overview of the Remainder of the Dissertation

Estimating and mapping the crude and adjusted spatial effects on disease outcomes (such as disease risks and survival rates) based on individual-level data is the main focus of this dissertation. In this introductory chapter, we have introduced the motivation for the methods developed in the thesis. Chapter 2 reviews the relevant but less known background material for the methods developed in the thesis including various types of smoothing procedures that have been developed in the statistical literature, an introduction to GAMs including the fitting algorithms implemented by GAMs and inference for nonparametric smoothers in GAMs, and introduction to the Cox proportional hazards model for analyzing censored survival data. In Chapter 3, we propose and implement algorithms to fit a Cox proportional hazard model for censored survival data that allows for the inclusion of a bivariate spatial smooth term. We also introduce the R package `MapGAM` that was, in part, created through the work in this dissertation in order to provide analysts with convenient and comprehensive

software to analyze and map crude or adjusted spatial effects in epidemiologic studies. In Chapter 4, we introduce a novel hierarchical Bayesian framework for modeling univariate continuous and discrete outcome via adaptive smoothing. These adaptive smoothing methods are shown minimize bias and variance in estimated spatial effects when compared to currently used GAMs. We further extend the methodology presented in Chapter 4 by developing a Bayesian framework for censored survival data via adaptive smoothing in Chapter 5. Taken as a whole, the research developed in Chapters 3, 4, and 5 of the thesis provides a suite of general methodologies for analyzing the most commonly encountered outcomes in spatial epidemiology studies (continuous, binary, count, and censored outcome data) in the presence of confounding factors while utilizing spatially adaptive smoothing parameters to reduce bias and variability in estimated geospatial effects. While the thesis represents a major advance in the analysis of spatial effects on disease incidence, further challenges remain. In Chapter 6 we conclude with a discussion of the advances that have been made here and highlight areas of future work that will continue to propel the field forward.

Chapter 2

Statistical Background

In this chapter, we provide a review of the relevant background material for the methods developed in the thesis. We begin with a review of generalized linear models (GLMs) in Section 2.1, a general class of multiple regression models for continuous and discrete outcomes arising from controlled and observational studies. Specifically, GLMs lay the methodologic foundation for models that will be developed and used for disease mapping when the outcome of interest is a binary response as in the case of the Massachusetts preterm birth data. A review of the semi-parametric Cox proportional hazards model for the analysis of censored time-to-event data, as observed in the California ovarian cancer study, is provided in Section 2.2. In Section 2.3 we review three of the most commonly employed methods for statistical smoothing of a response variable: local-averaging smoother, smoothing splines, and thin-plate splines. In Section 2.4 we consider the incorporation of a smoothing term into the GLM framework, providing a review of GAMs along with an overview of the backfitting algorithm for estimating parameters in GAMs for a univariate outcome. In Section 2.4 we also present an application of the use of a GAM in the context of spatial mapping when the the distribution of the response is a member of the exponential family. The section concludes with a discussion of intrinsic Gaussian Markov random fields (IGMRFs). This

review completes the foundational groundwork for the Bayesian adaptive smoothing methods developed in Chapters 4 and 5 that utilize an IGMRF prior distribution on the underlying spatial effects.

2.1 The Generalized Linear Model

A generalized linear model (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1983) consists of a random component, a systematic component and a link function to line the two components together. The random component identifies the probability distribution of response Y, which is assumed to belong to the exponential family with density function of the form

$$p_Y(y;\theta;\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}. \tag{2.1}$$

In the above density, θ is termed the canonical or natural parameter, and ϕ general represents a nuisance parameter that characterizing the dispersion of response Y. Denote the expectation and variance of the response Y as μ and V,which are related to θ by $\mu = b'(\theta)$, and $V = b''(\theta)a(\phi)$. μ is related to d covariates $X_1, \cdots, X_d$ by

$$g(\mu) = \eta = \beta_0 + \beta_1 X_1 + \cdots + \beta_d X_d,$$

where η represents the linear predictor or systematic component of the model and $g(\cdot)$ is termed the link function, linking the mean response to the linear predictor. $\beta_i,\ \ i = 1, \cdots, d$ represents the coefficient associated with X_i, reflecting the relationship between the outcome and the corresponding covariate.

As an example, consider a binary outcome Y distributed Bernoulli with $P(Y = 1) = p$. It is easy to see that the distribution of Y is a member of the exponential family by rewriting

the probability mass function for Y as

$$p_Y = \exp\left\{Y \log\left(\frac{p}{1-p}\right) + \log(1-p)\right\}, \tag{2.2}$$

and letting $\theta = \log(p/(1-p))$, $b(\theta) = -log(1-p)$, $a(\phi) = 1$, and $c(y, \phi) = 0$. For a binary response, the logit function (as given by the canonical parameter) and probit function (defined as the inverse of the standard Normal cumulative distribution function) are commonly used for link function $g(\cdot)$. When the logit link function is used, it is easily seen that β_i represents the log-odds ratio comparing subpopulations differing in X_i by 1-unit. Hence, e^{β_i} represents the odds ratio comparing subpopulations differing in X_i by 1-unit.

Estimation of model parameters $\theta \equiv (\beta_0, \cdots, \beta_d, \phi)$ in a GLM is typically carried out via the method of maximum likelihood. Briefly, assuming N independent observations, the likelihood function is given by

$$L(Y, X; \theta) = \prod_i^N p_Y(Y_i|X_i, \theta),$$

and the maximum likelihood estimate $\hat{\theta}$ is given by

$$\arg\max_\theta L(Y, X; \theta)$$

Due to monotonicity of the log function, one can equivalently maximize the log-likelihood function given by

$$l(Y, X; \theta = \sum_i^N \log\left(p_Y(Y_i|X_i, \theta)\right).$$

With N observed responses $y_i, i = 1, \cdots, N$ and corresponding covariates $x_{ij}, i = 1, \cdots, N, j = 1, \cdots, d$, the maximum likelihood estimate (MLE) of θ is given by the solution to the score

equation defined as:

$$U(Y, X; \theta) = \partial l / \partial \theta = 0$$

In general, no closed form solution for $\hat{\theta}$ exists, however the score equation can be solved numerically using the Newton-Raphson algorithm. A slight modification of the Newton-Raphson algorithm is obtained by replacing the second-derivative of the log-likelihood function with it's expectation, yielding the Fisher scoring algorithm. It can easily be shown (McCullagh and Nelder, 1983) that the Fisher scoring algorithm as applied to GLMs is equivalent to an iteratively re-weighted least square regression. Specifically, at iteration j of the algorithm let $\mu^{(j-1)}$, $V^{(j-1)}$ and $\eta^{(j-1)}$ denote the estimates resulting from iteration $j-1$. Based on these estimates, one can construct the *working response* for the i^{th} observation, $z_i^{(j)}$, by

$$z_i^{(j)} = \eta_i^{(j-1)} + (y_i - \mu_i^{(j-1)}) \left(\frac{\partial \eta_i}{\partial \mu_i} \right) \Bigg|_{\mu_i^{(j-1)}, \eta_i^{(j-1)}},$$

and weights w_i by

$$w_i^{(j)} = \left(\frac{\partial \eta_i}{\partial \mu_i} \right)^2 \Bigg|_{\mu_i^{(j-1)}, \eta_i^{(j-1)}} V_i^{(j-1)}.$$

Then the estimate $\theta^{(j)}$ is updated by regressing $z_i^{(j)}$ on covariates $x_{i1}, \cdots, x_{id}$ with weights $w_i^{(j)}$. The procedure stops when $||\theta^{(j)} - \theta^{(j-1)}|| < \epsilon$, where ϵ is a specified tolerance threshold.

2.2 Cox Proportional Hazards Regression for Survival Analysis

Censored survival data are commonly analyzed in epidemiology studies (Assael et al., 2002; Ho et al., 1993; Ickovics et al., 2001; Bramball et al., 1995; Steliarova-Foucher et al., 2004; Bristow et al., 2014). In these settings, one wishes to model the time from a specified origin until the occurrence of a well-defined event, commonly referred to as the survival or failure time. Examples include the time to death since disease diagnosis, the time the disease occurrence since exposure to an environmental condition, or the time to disease relapse since remission. In these cases, it is often the case that some observations are *right-censored*, implying that followup of some individuals may prior to the occurrence of the event of interest. Reasons for ending followup may include a subject removing his- or herself from the study or that the event of interest has not occurred at the time the study is closed. It is well known that, in the presence of censoring, standard statistical methods for analyzing continuous times yield inefficient (at best) and biased (at worst) estimates of covariate effects on survival. To address these issues, survival analysis methods consist of a class of procedures for analyzing covariate effects on the time-to-event in the presence of censoring. For the remainder of this section we briefly introduce the common targets of inference in survival studies and the most frequently used regression models utilized in survival studies.

2.2.1 Common Estimands of Interest in Survival Analysis: The Survival and Hazard Functions

Let T_i' denote the *true* time-to-event for observation i, $i = 1, \ldots, N$, and let C_i denote the censoring time for observation i (eg. the time from the study origin to when individual i quit the study or the study is closed). Only one of T_i' or C_i is observed for each observation,

and hence we define $T_i = min(T_i', C_i)$ to be the *observed* time for observation i, $i = 1, \ldots, N$. Further, define δ_i to be the event indicator for subject i such that $\delta_i = 1$ if $T_i \leq C_i$ and $\delta_i = 0$ if $T_i > C_i$.

Let $f(t)$ and $F(t)$ denote the probability density function (pdf) and cumulative distribution function (cdf) for T', respectively (omitting the observation index i for brevity). Then the survival function $S(t)$ is defined as the probability that the true time-to-event is longer than t. That is, the survival function is defined as

$$S(t) = Pr\{T' > t\} = 1 - Pr\{T' <= t\} = 1 - F(t) = \int_t^\infty f(s)ds.$$

The hazard function, $\lambda(t)$, is defined as the instantaneous rate of failure at time t given survival up to time t. Thus the hazard function is given by

$$\lambda(t) = lim_{\Delta t \to 0^+} \frac{Pr\{t \leq T' < t + \Delta t | T' \geq t\}}{\Delta t}.$$

It is easy to see that the hazard function can be written in terms of the pdf and survival distribution by noting that

$$\begin{aligned} \lambda(t) &= lim_{\Delta t \to 0^+} \tfrac{Pr\{t \leq T' < t+\Delta t | T' \geq t\}}{\Delta t} \\ &= lim_{\Delta t \to 0^+} \tfrac{Pr\{t \leq T' < t+\Delta t, T' \geq t\}/Pr\{T' \geq t\}}{\Delta t} \\ &= lim_{\Delta t \to 0^+} \tfrac{Pr\{t \leq T' < t+\Delta t\}}{\Delta t} \Big/ Pr\{T' \geq t\} = \tfrac{f(t)}{S(t)}. \end{aligned}$$

Further, noting that $f(t)$ is the negative derivative of $S(t)$, ie.

$$\lambda(t) = -\frac{d}{dt} log(S(t)),$$

and we also have that

$$S(t) = \exp\left\{-\int_0^t \lambda(s)ds\right\}.$$

The above relationship gives rise to the cumulative hazard function, $\Lambda(t)$, defined as

$$\Lambda(t) = \int_0^t \lambda(s)ds,$$

and hence the relationship,

$$S(t) = \exp\left\{-\Lambda(t)\right\}.$$

Thus, both the hazard function and survival can be used to fully characterize the distribution of the survival time, T'.

In many cases it is justifiable to assume that the true failure time T_i' and censoring time C_i for subject i are independent. In this case, the likelihood function incorporated both censored and fully observed survival times is obtained as

$$L = \prod_{i=1}^{N} [f(T_i)]^{\delta_i} [S(T_i)]^{1-\delta_i}.$$

Thus the above likelihood function reduces to the usual likelihood function for independent data when all subjects have fully observed survival times (ie. $\delta_i = 1$ for $i = 1, \ldots, N$), but allows for partial contributions up to the censoring times for those individuals that were censored prior to failure.

Given the above specification of the likelihood function in the presence of censoring, it is relatively straightforward to obtain maximum likelihood estimates when one is will to assume a parametric probability model for the the true survival times. For example, the simplest

survival function is the exponential distribution which is characterized by a constant hazard function $\lambda(t) = \lambda$ for $t > 0$. Using the relationship between the survival function and the hazard function, the corresponding survival function for the exponential distribution is then given by

$$S(t) = \exp\{-\lambda t\}.$$

In this case, covariate effects on the survival distribution can easily be incorporated into the model by writing the hazard function as a linear combination of the covariates and unknown parameters of the covariate:

$$\log(\lambda) = \beta_0 + \beta_1 X_1 + \cdots + \beta_d X_d.$$

By the above model specification, the parameter β_j represents the difference in the log-hazard comparing subpopulations differing in X_j by 1-unit that are similar with respect all other adjustment covariates. Alternatively, e^{β_j} represents the ratio of hazard functions comparing subpopulations differing in X_j by 1-unit that are similar with respect all other adjustment covariates. Because this ratio is constant with respect to time (recall that $\lambda(t) = \lambda$ for $t > 0$), the above model is said to fall with the family of proportional hazards survival models.

2.2.2 Cox Proportional Hazards model

The assumption of a mis-specified parametric survival distribution can lead to biased and/or inefficient estimation, as is true with mis-specified parametric models for uncensored data. This lack of robustness in parametric models motivated the use of semi-parametric regression model ss in the context of censored survival data. Indeed the most widely used regression model for relating covariates to censored survival times is the semi-parametric Cox proportional hazards model (Cox, 1972, 1975). Briefly, let X be a $d \times 1$ vector of covariates that

may be related to the survival times. The Cox proportional hazard assumes

$$\lambda(t) = \lambda_0(t) \exp\{X^T \beta\}, \tag{2.3}$$

where $\lambda(t)$ represents the hazard at time t and $\lambda_0(t)$ is a non-specified baseline hazard (i.e. the hazard for subjects with covariates values of $X = 0$). β is a $d \times 1$ vector of regression coefficients associated with covariates X. As noted with the exponential survival model previously presented, the Cox model specification provided in equation (2.3) is also a member of the family of proportional hazards survival models. This is easily seen by noting that the ratio of hazard function comparing subpopulations differing in, say X_j, by 1-unit but similar with respect to all other covariate values is constant as a function of time, and is given by e^{β_j}, $j = 1, \ldots, d$.

While the exponential model and the Cox model are both members of the family of proportional hazards survival models, there is a key distinction between the two approaches. This distinction comes in the form of the specification of the baseline hazard function, $\lambda_0(t)$. While the baseline hazard function is assumed to be constant and hence can be estimated via a single parameter using maximum likelihood estimation in the exponential survival model, estimation of β in the Cox model can be carried out without any assumption on the baseline hazard. Because of this, the Cox proportional hazards is termed a *semi-parametric*, as the baseline hazard may be infinitely dimensional yet the the relative covariate effects are specified by a finite number of parameters (ie. $\beta_1, \ldots, \beta_d$)

The desire to estimate the regression coefficients β in the Cox model without any assumptions on the baseline hazard function implies that estimation cannot be carried out via usual maximum likelihood. Instead, Cox Cox (1975) proposed and justified under the assumption of independence between censoring and failure times that estimation of β be carried out by maximizing the *partial likelihood*. Following Cox's construction of the partial likelihood,

suppose that one and only one subject (with covariate $X_i = x_i$) fails at time $T_i = t_i$, then the partial likelihood contribution for subject i is given by the probability that a subject with covariate value $X = x_i$ fails at time $T = t_i$ given that some subject failed at time $T = t_i$. More specifically, letting $R_i \equiv \{j | t_j \leq t_i\}$ denote the *risk set* or set of all subjects not censored or observed to to fail by time t_i, the contribution of subject i to the partial likelihood is given by

$$\begin{aligned}
PL_i &= Pr\{\text{subject with covariate } x_i \text{ fails at } t_i \mid \text{ some subject failed at } t_i\} \\
&= \frac{Pr\{\text{subject with } x_i \text{ fails at } t\}}{Pr\{\text{ some subject in } R_i \text{ failed at } t\}} \\
&= \frac{\lambda_i(t_i)(\Delta t)}{\sum_{j \in R_i} \lambda_j(t_i)(\Delta t)} \\
&= \frac{\lambda_0(t_i)\exp\{x_i^T\beta\}}{\sum_{j \in R_i} \lambda_0(t_i)\exp\{x_j^T\beta\}} \qquad \text{(from (2.3))} \\
&= \frac{\exp\{x_i^T\beta\}}{\sum_{j \in R_i} \exp\{x_j^T\beta\}}
\end{aligned}$$

Under this specification, only subjects with observed failure times contribute terms to the partial likelihood. However, the partial likelihood still incorporates information from right-censored observations by considering the probability that any subject that is still at risk at time $T = t_i$ (whether eventually censored or not) fails at time $T = t_i$. Perhaps more importantly, it can be seen from the specification of PL_i that the partial likelihood does not depend upon the baseline hazard function, and hence there is no need to assume any parametric form for the baseline hazard function. Supposing no tied failure times, the partial likelihood is defined as

$$PL = \prod_{i \in D} PL_i, \tag{2.4}$$

where D denotes the set of indices of the failures. Although the partial likelihood does not correspond to a fully parametric likelihood, it has been shown that many of the asymptotic properties of traditional maximum likelihood methods including analogous asymptotic dis-

tributions of the Wald, score, and likelihood ratio test hold for partial maximum likelihood estimators (Cox, 1975; Andersen and Gill, 1982).

It should be noted that the above derivation of the partial likelihood assumed that no two observations were observed to have the same failure time. In the event of tied failure times there are two ways to calculate the partial likelihood. As an illustration, suppose two patients, subject 1 and subject 2, both failed at time t. The *exact* partial likelihood (Peto, 1972; Kalbfleisch and Prentice, 1973) assumes that survival times are truly continuous in nature, and hence the probability of two subjects having the exact same survival time is zero. In this case, these two subjects have the same recorded survival time because the system of measurement used for recording time does not have enough accuracy or information lost (eg. death times may be measured in days as opposed to seconds). Without any knowledge of the true ordering of the survival times of the two subjects, it is necessary to take into account all possible ordering of the ties when calculating the denominator in (2.4). If one assumes that subject 2 failed before subject 1, then the risk set of subject 2 would include subject 1 while the risk set of subject 1 would not include subject 2. Similarly, if subject 1 failed before subject 2, then subject 1 would be included in the risk set of subject 2. Notationally, let D denote the set of indices of the *distinct* failure times, then the exact partial likelihood taking into account all possible orderings of tied failure times is given by

$$PL_M = \prod_{i \in D} \left\{ \prod_{j \in F_i} e^{X_j^T \beta} \sum_{P \in Q_i} \prod_{r=1}^{d_i} \left[\sum_{l \in R(i,P,r)} e^{X_l^T \beta} \right]^{-1} \right\},$$

where F_i denotes the set of failures at time t_i, d_i is the number of elements in F_i, Q_i is the set of $d_i!$ possible permutations of the corresponding observations in F_i, $P = (p_1, \cdots, p_{d_i})$ is an element of Q_i, and$R(i, P, r)$ is the risk set R_i excluding the elements $p_1, \cdots, p_{r-1}$.

An alternative approach to handling tied failure times in the partial likelihood is to assume that time is truly discrete (eg. the number of visits prior to testing positive for a particular

illness) so that there is no true ordering of any tied observations (Cox, 1972). As before suppose two patients, subject 1 and subject 2, both failed at time t. Under the *discrete* method, the partial likelihood contribution for this failure time is constructed as the probability that the two subjects both failed at time t given that two failures were observed at time t. Notationally, the partial likelihood assuming discrete ties is given by

$$PL_D = \prod_{i \in D} \frac{\prod_{j \in F_i} e^{X_j^T \beta}}{\sum_{l \in R_{d_i}} \prod_{j \in l} e^{X_j^T \beta}},$$

where R_{d_i} is the collection of all sets of d_i labels chosen from the risk set R_i without replacement, and l is an element of R_{d_i}.

Both the exact and discrete methods are computationally complex when maximizing the likelihood in the presence of many tied failure times. As such, approximations to the partial likelihood are often used. The first approximation due to Breslow (Breslow, 1974) modifies the partial likelihood as follows:

$$PL_B = \prod_{i \in D} \frac{\prod_{j \in F_i} e^{X_j^T \beta}}{\left[\sum_{l \in R_i} e^{X_l^T \beta}\right]^{d_i}}.$$

In the Breslow approximation, the dominator of the partial likelihood is inflated as it assumes that all tied observations are in the risk set without taking into account that one may have occurred before another. Noting that the Breslow approximation can lead to severe bias in coefficients estimates when survival data contain many tied events, Efron proposed alternative approximation Efron (1977) as follows:

$$PL_E = \prod_{i \in D} \frac{\prod_{j \in F_i} e^{\eta_j}}{\prod_{j=1}^{d_i} \left[\sum_{l \in R_i} e^{\eta_l} - \sum_{l \in F_i} e^{\eta_l}(j-1)/d_i\right]}.$$

As can be seen, Efron's approximation uses an average of the risk of the tied failures when considering ties in the risk set. While still resulting in some bias when multiple ties are

present, the Efron approximation tends to performs better than the Breslow approximation and is generally recommended when the exact or discrete partial likelihood cannot be computed due to computational constraints.

2.3 Smoothing Methods

To this point, the regression methods presented in the context of GLMs and Cox's proportional hazards model have assumed a parametric relationship between covariates and the response as defined in the linear predictor. In order to relax the parametric assumptions on these associations smoothing may be utilized. However, before introducing smoothing in the context of a regression model, we first present a basic background on the most common types of smoothers that are utilized by analysts.

A smoother is a tool for summarizing the trend of a response measurement Y as a function of one or more predictors $X \in \mathbb{R}^d$, where d denotes the number of predictors. The high flexibility of a smoother produces an estimate of the trend that is less variable than Y itself and hence must be weighed against potential increases in bias due to model overfitting. In the context of a linear smoother we consider a continuous response modeled with a smooth function of covariates so that

$$Y = f(X_1, \cdots, X_d) + \epsilon, \tag{2.5}$$

where ϵ denotes a random error term such that $E[\epsilon] = 0$ and $var(\epsilon) = \sigma^2$ (the assumption of constant variance can be relaxed when we discuss weighted smoothers). Here $f(\cdot)$ denotes the smoothing function, and the goal is to estimate $f(\cdot)$ based upon observed data $[(Y_1, X_1), \ldots, (Y_N, X_N)]$ where Y_i is the univariate response for observation i and X_i is the $d \times 1$ vector of observed covariates for subject i, $i = 1, \ldots, N$. An important property of the

smoother is that it does not assume a rigid form for the smooth function.

For the remainder of this section we review three commonly employed choices of smoothing methods. Local-averaging smoothers and smoothing splines are discussed in Section 2.3.1 and Section 2.3.2, respectively. These two smoothing methods are the mostly commonly used techniques in the context of regression modeling and are implemented in most standard statistical software packages. We conclude this section with a review of thin-plate splines as an alternative to local-average smoothers and smoothing splines. While less commonly used, thin-plate splines have advantages in that they allow the degree of smoothness to vary over the support of predictors and hence will be utilized for the Bayesian adaptive smoothing methods developed in Chapters 4 and 5.

2.3.1 Local-Averaging Smoother

Perhaps the most intuitive of all smoothing techniques, local-averaging smoothers use the the weighted average of observations across the neighborhood of the prediction point as an estimate. Local-averaging smoothers date back at least to Ezekiel (1941), who suggested a smoother similar to the running-mean. As a multi-dimensional generalization of a running mean, K-nearest-neighbor methods (Aha, 1997) use average value of the neighborhood as an estimate, given by

$$\hat{f}(X_0) = \frac{1}{k} \sum_{X_i \in N_k(X_0)} Y_i,$$

where the neighborhood $N_k(X_0)$ is defined by the k closest points to X_0 in the data. Thus k is a smoothing parameter determining the amount of smoothing to be performed. Figure 2.1 shows the estimated smooth functions after applying a K-nearest-neighbor smoother to a data set with different choice of k. The K-nearest neighbor estimator can be written as a

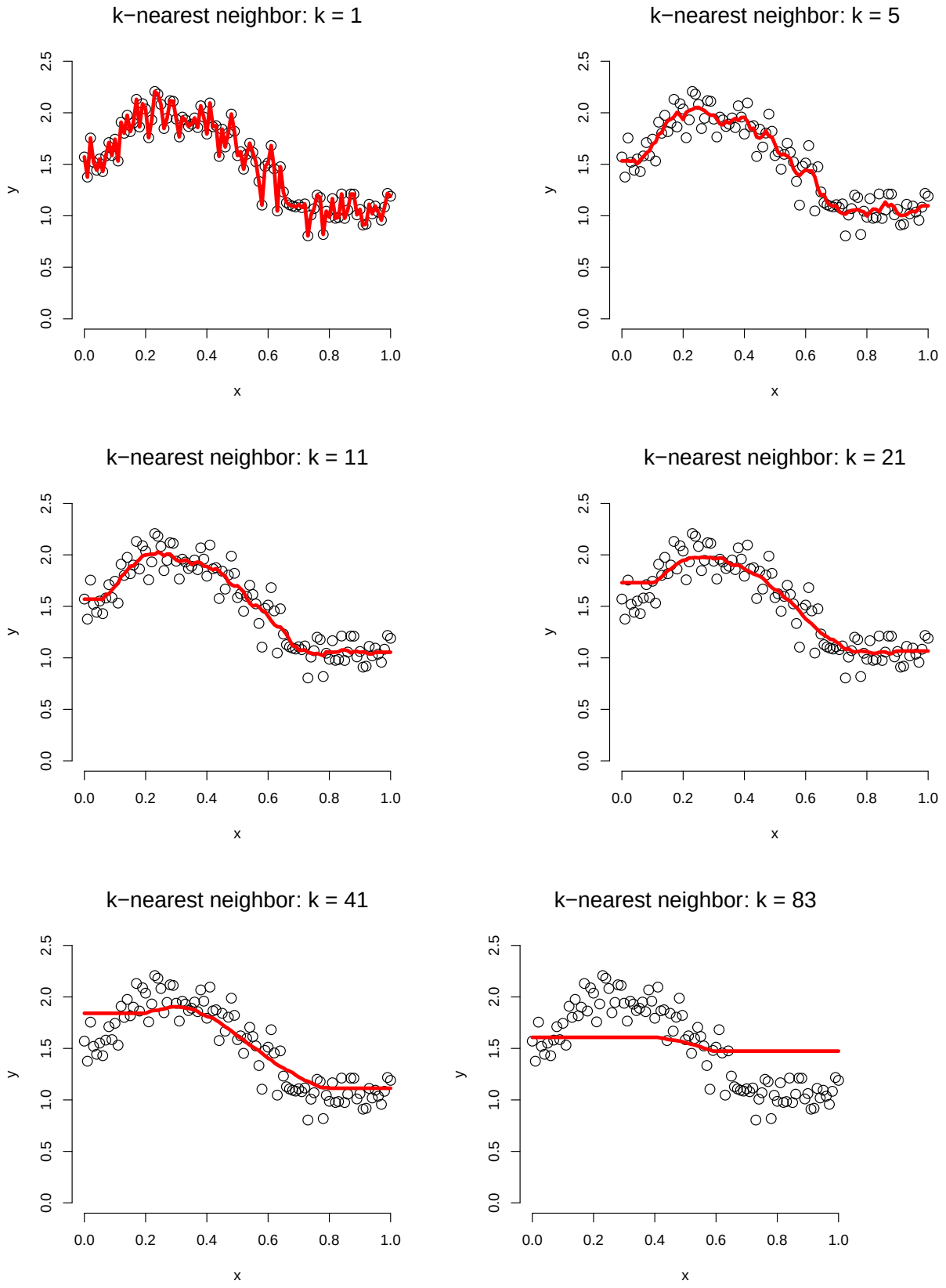


Figure 2.1: K-nearest neighbor smoothers applied to response y over covariate x with different choices of smoothing parameter, k. Dots: observed data; Line: estimated smooth function.

weighted summation of the observations:

$$\hat{f}(X_0) = \sum_{i=1}^{N} K_k(X_0, X_i) Y_i,$$

where the function

$$K_k(X_0, X) = \begin{cases} 1/k & \text{if} X_i \in N_k(X_0) \\ 0 & \text{otherwise} \end{cases}$$

assigns weights to each Y_i. The weight function is dependent on the target location X_0 and choice of smoothing parameter, k. Figure 2.2 illustrates effect of the weighting function when estimating the smooth at different locations. In later work, extensions of the kernel smoother were proposed and developed by Nadaraya (1964) and Priestley and Chao (1972). These authors suggested that the weights assigned to observations be based on a kernel function, $K_{h_\lambda}(X_0, X)$, of the form

$$K_{h_\lambda}(X_0, X) = D\left(\frac{||X - X_0||}{h_\lambda(X_0)}\right),$$

where $X, X_0 \in \mathbb{R}^d$ are two vectors of covariates, and $||\cdot||$ defines the Euclidean distance between them. Here, $h_\lambda(X_0)$ is a *smoothing parameter* determining the kernel radius of the smoothing neighborhood and hence partly determining the level of resulting smoothing. Finally, the *kernel* function $D(\cdot)$ is a positive valued function, with values decreasing as the distance between X and X_0 increases. Popular choices of kernel functions include:

- The tri-cube kernel : $D(x) = (1 - |x|^3)^3$
- The Gaussian kernel : $D(x) = e^{\left\{-\frac{x^2}{2}\right\}}$.

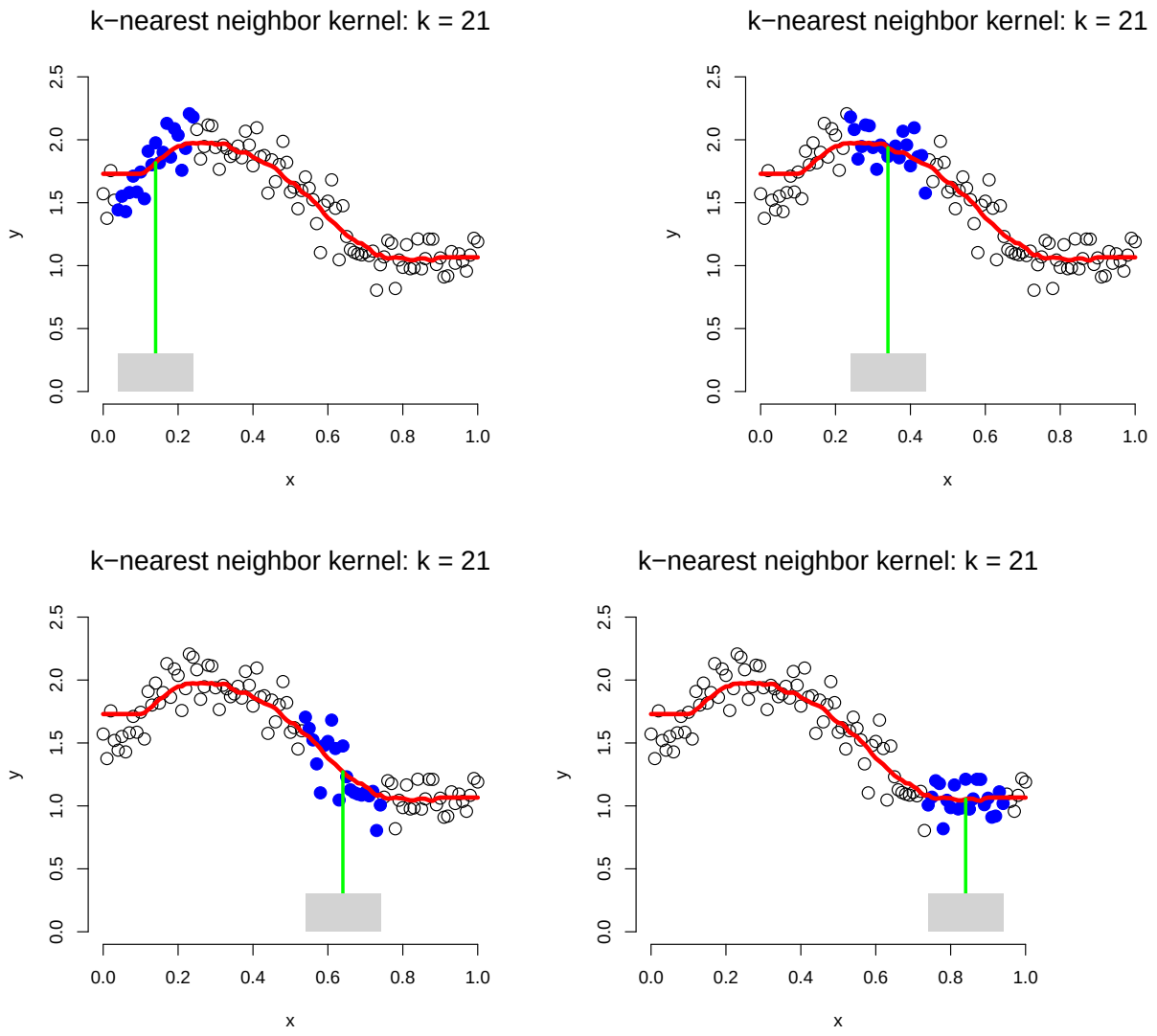


Figure 2.2: Kernel representation of a K-nearest neighbor smoother. Black dots: observed data; Green line: vertical line indicating the target location; Blue dots: k closest points; Gray box: shape of weight function when estimating the smooth function at the target location.

After specification of $h_\lambda(X_0)$ and $D(\cdot)$, the smooth function at X_0 can be estimated using a kernel-weighted average given by

$$\hat{f}(X_0) = \frac{\sum_{i=1}^{N} K_{h_\lambda}(X_0, X_i) Y_i}{\sum_{i=1}^{N} K_{h_\lambda}(X_0, X_i)}.$$

Figure 2.3 shows an example of the Gaussian kernel smoother with $h_\lambda(X_0) = \lambda$. λ is the standard deviation of the weights. Figure 2.3 illustrates the kernel representation of four choices of λ.

The kernel estimates described above can suffer from bias at the boundary and where observed locations are not evenly spaced. Therefore, instead of taking a weighted average of observations as kernel smoothers do, locally-weighted scatterplot smoothers (LOESS) (Cleveland, 1979, 1981; Clark, 1977; Tibshirani and Hastie, 1987; Devlin, 1988; Cleveland and Devlin, 1988; Cleveland et al., 1988) fit a weighted linear or polynomial regression model over the observations as an estimate. For any given point X_0, the smooth function can be estimated by a polynomial function of X_0, i.e.

$$\hat{f}(X_0) = P^p_{\beta(X_0)}(X_0) = \sum_{j_1 + j_2 + \cdots + j_d \leq p} \beta_{j_1 j_2 \cdots j_d}(X_0) X_{0j_1}^{j_1} \cdots X_{0j_d}^{j_d},$$

where $P^p_{\beta(X_0)}(X_0)$ is a p degree polynomial function of $X_0 \in \mathbb{R}^d$ and $\beta(X_0)$ is a set of model coefficients $\beta(X_0) = \{\beta_{j_1 j_2 \cdots j_d}(X_0), j_1 + j_2 + \cdots + j_d \leq p\}$ to be estimated. The coefficients are obtained by solving a weighted least squares problem of the form

$$\min_{\beta(X_0)} \sum_{i=1}^{N} K_{h_\lambda}(X_0, X_i)(Y_i - P^p_{\beta(X_0)}(X_i)).$$

Building on previous work, the weights used in local regression can be assigned using the kernel function. Most commonly, the tri-cube weighting function is used to provide weights for LOESS method.

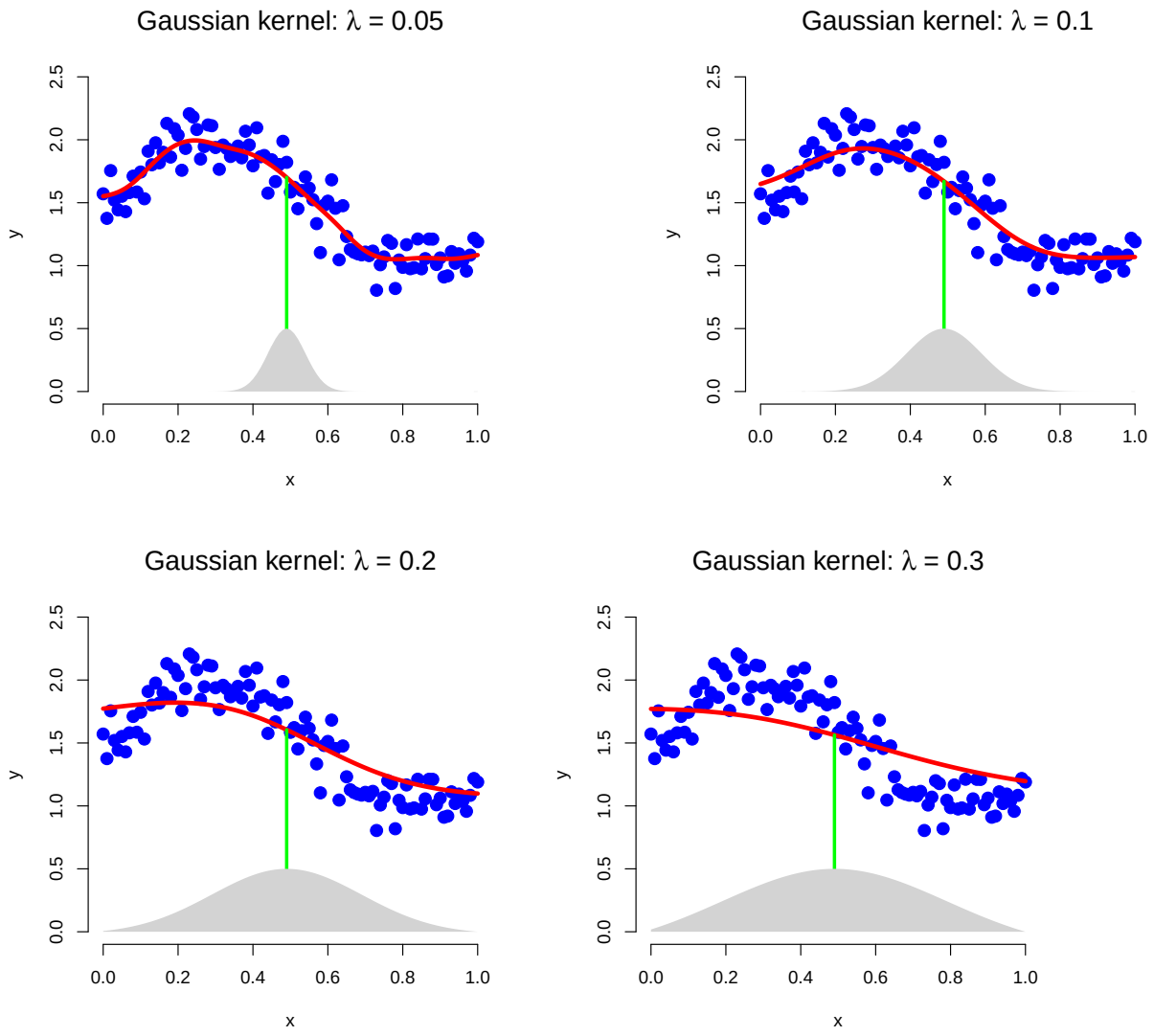


Figure 2.3: Kernel representation of a Gaussian kernel smoother. Black dots: observed data; Green line: vertical line indicating the target location; Blue dots: k closest points; Gray box: shape of weight function when estimating the smooth function at the target location.

The LOESS requires a choice of span, or window size, w for the weighting function. The span size w is the proportion of the total number of observations that are contained in each neighborhood. Suppose the sample size of the study is N, then for each point, LOESS will choose the $[wN]$ observed points with smallest Euclidean distance from the target point as the neighborhood. Then the tri-cube weighting function for one observation X_0 will be

$$w(X_0) = \left(1 - \left(\frac{||X - X_0||}{h_w(X_0)}\right)^3\right)^3,$$

where $h_w(X_0)$ is the longest Euclidean distance among the distances from neighborhood points to X_0, and the choice of neighborhood points are related to the span size w. In the assignment of weights, observations with covariate value X_0 receive the highest weight, 1, and observation of the longest distance in the neighborhood receive the lowest weight, 0.

Since the estimate, $\hat{f}(x)$, is obtained from a locally weighted regression on the response it is linear in y_i and can be written as

$$\hat{f}(x_i) = \sum_{j=1}^{N} l_j(x_i) y_j. \tag{2.6}$$

Let L be the matrix whose (i, j)th element is $l_j(x_i)$ and let

$$\hat{L} = I - L,$$

where I is the $N \times N$ identity matrix. For $k = 1$ and 2, let

$$\delta_k = tr(\hat{L}^T \hat{L})^k.$$

Then δ_1 can be regarded as the degrees of freedom for the smoother and the standard deviation of Y can be estimated by

$$\hat{\sigma} = \sqrt{\frac{SSE}{\delta_1}}.$$

Based on equation (2.6), the standard deviation of $\hat{f}(x_i)$ can be estimated by

$$s(\hat{f}(x_i)) = \hat{\sigma}\sqrt{\sum_{j=1} N l_j^2(x_i)}.$$

Let

$$\rho = \delta_1^2/\delta^2.$$

Then the distribution of

$$\frac{\hat{f}(x_i) - f(x_i)}{s(\hat{f}(x_i))}$$

can be approximated by a t-distribution with ρ degrees of freedom, which can be used to obtain pointwise confidence intervals for the estimated smooth function.

2.3.2 Smoothing Splines

As an alternative to the kernel smoothers discussed above, smoothing splines represent the smooth function as an element in the space defined by a basis based on a set of knots. If $b_i(x)$ is the ith basis function, then the smooth function $f(x)$ is assumed to have a representation given by

$$f(x) = \sum_{i=1}^{q} b_i(x)\beta_i.$$

Then (2.5) is a linear model and can be estimated using least squares.

The choice of basis is based on the penalized least-squares criterion (Reinsch, 1967). Suppose the smooth function is with respect to a univariate predictor, then the estimated smooth function is obtained by minimizing

$$\sum_{i=1}^{N} [y_i - f(x_i)]^2 + \lambda \int f''(x)^2 dx, \tag{2.7}$$

where λ is a tuneable paramter that is used to control the relative weight, chosen to balance the conflicting goals of fitting the observed data well and producing a smooth estimate. A cubic spline basis arises naturally from the specification of the smoothing objective in (2.7). Consider defining a cubic spline function, $f(x)$, with k knots, $x_1, \cdots, x_k$. Let $\beta_j = f(x_j)$ and $\delta_j = f''(x_j)$. Then, for $x_j \leq x \leq x_{j+1}$ the spline can be written as

$$f(x) = a_j^-(x)\beta_j + a_j^+(x)\beta_{j+1} + c_j^-(x)\delta_j + c_j^+\delta_{j+1},$$

where

$$a_j^-(x) = \frac{x_{j+1}-x}{x_{j+1}-x_j}; c_j^- = \frac{(x_{j+1}-x)^3/(x_{j+1}-x_j)-(x_{j+1}-x_j)(x_{j+1}-x)}{6}$$
$$a_j^+(x) = \frac{x-x_j}{x_{j+1}-x_j}; c_j^- = \frac{(x-x_j)^3/(x_{j+1}-x_j)-(x_{j+1}-x_j)(x-x_j)}{6}.$$

Based on the penalized criterion, the spline must be continuous to second derivative at x_j and have zero second derivative at x_1 and x_k. Thus we have the constraint that

$$B\delta^- = D\beta,$$

where $\delta^- = (\delta_2, \cdots, \delta_{k-1})^T$ (since $\delta_1 = \delta_k = 0$) and $\beta = (\beta_1, \cdots, \beta_k)^T$. The matrix B is defined as

$$B_{i,i} = (x_{i+2} - x_i)/3; B_{i,i+1} = B_{i+1,i} = (x_{j+2} - x_{j+1})/6,$$

and matrix D is defined as

$$D_{i,i} = 1/(x_{j+1} - x_j); D_{i,i+2} = 1/(x_{j+2} - x_{j+1}); D_{i,i+1} = -D_{i,i} - D_{i,i+2}.$$

Defining the matrices $F^- = B^{-1}D$ and $F = [0, (F^-)^T, 0]^T$, we have that $\delta = F\beta$. As such, the spline can be re-written entirely in terms of β as

$$f(x) = a_j^-(x)\beta_j + a_j^+(x)\beta_{j+1} + c_j^-(x)F_j\beta_j + c_j^+F_j\beta_{j+1},$$

which can in-turn be re-written as an element in the basis-defined space:

$$f(x) = \sum_{i=1}^{k} b_i(x)\beta_i,$$

where $b_1(x) = 1$, $b_2(x) = x$, and $b_{i+2} = R(x, x_i)$ for $i = 1, \cdots, k-2$ with

$$R(x,z) = \left[\left(z - \frac{1}{2}\right)^2 - \frac{1}{12}\right]\left[\left(x - \frac{1}{2}\right)^2 - \frac{1}{12}\right]/4 - \left[\left(|x-z| - \frac{1}{2}\right)^4 - \frac{1}{2}\left(|x-z| - \frac{1}{2}\right)^4 + \frac{7}{240}\right]/24.$$

Lancaster and Salkauskas (1986) showed that the penalty term in (2.7) can be written as

$$J(f) = \int_{x_1}^{x_k} f''(x)^2 dx = \beta^T D^T B^{-1} D\beta \tag{2.8}$$

and $S \equiv D^T B^{-1} D$ is termed the penalty matrix for the basis.

The cubic smoothing spline was developed for smoothing univariate predictors. However, spatial analysis requires a bivariate smoother to smooth over two geographical parameters. For this, tensor product spline bases (De Boor, 1978; Wood, 2006) can be constructed to smooth several variables. This section will introduce the construction of a smooth function of two covariates, u and v. The generalization to more predictors is trivial.

Suppose the low rank bases are available, then the smooth functions f_u and f_v of each of the

covariates can be written as:

$$f_u(u) = \sum_{i=1}^{I} \alpha_i a_i(u); f_v(v) = \sum_{l=1}^{L} \delta_i d_l(v),$$

where $a_i(u)$ and $d_l(v)$ are the basis functions for the two covariates, and the α_i and δ are corresponding model parameters. To convert the function f_u into a smooth function of u and v, the parameters α_i can be represented by a expansion of the basis for $f(v)$ as :

$$\alpha_i(v) = \sum_{l}^{L} \delta_{il} d_l(v),$$

which gives the tensor product basis representation:

$$f_{uv}(u, v) = \sum_{i=1}^{I} \sum_{l=1}^{L} \delta_{il} d_l(v) a_i(u).$$

Having constructed the tensor product basis, it is necessary to build the penalty term for the spline, which can be written as

$$J(f) = \lambda_u \int_v J_u(f_{u|v}) dv + \lambda_v \int_u J_v(f_{v|u}) du,$$

where $J_u(f_{u|v}) = \int \left(\partial^2 f_{u|v}/\partial u^2\right)^2 du$, with $f_{u|v} = \sum_{i=1}^{I} \alpha(v) a_i(u)$. From (2.8) we have

$$J_u(f_{u|v}) = \beta^T M_v^T S_u M_v \beta,$$

where matrix M_v is the matrix such that $\alpha(v) = M_v \beta$, and S_u is the penalty matrix for the basis. Thus

$$\int_v J_u(f_{u|v}) dv = \beta^T \int_v M_v^T S_u M_v dv \beta.$$

This integral is easily computed via numerical methods. A similar form can be obtained for

$\int_u J_v(f_{v|u})du$. Then the parameters can be estimated by solving the minimization problem in (2.7) by the Newton-Raphson algorithm (Wood, 2006).

2.3.3 Thin-Plate Splines

There are two drawbacks when using regression splines. First, the number and location of knots must be selected for each basis, which introduces extra subjectivity into the estimates. Second, the bases are chosen for smooths of one predictor, and even for tensor product, the basis is constructed from the two low rank basis functions. Thin-plate splines (Duchon, 1977) discussed in this section allow for estimating a smooth function of multiple covariates, so they are prominantly used for spatial analysis (Cressie, 1993). Another elegant property of thin-plate splines is that they do not require the selection of knots positions or basis functions.

We illustrate thin-plate splines as a smooth function of two predictors in this section. Let u and v denote the two predictors to be smoothed. The mean of response Y can be modeled by a smooth function $f(u, v)$ as

$$Y = f(u, v) + \epsilon;\ \epsilon \sim_{iid} \mathbf{N}(0, \sigma^2). \tag{2.9}$$

With N observations (Y_i, u_i, v_i), $i = 1, \ldots, N$, the thin-plate spline smoothing method estimates $f(\cdot)$ by finding the function $\hat{f}$ that minimizes

$$\sum_{i=1}^{N} (y_i - f(u_i, v_i))^2 + \lambda J_2(f). \tag{2.10}$$

$J_2(f)$ is a penalty function measuring the smoothness of the function f, and λ is a smoothing parameter, controlling the tradeoff between the data fitting and smoothness of f, which is

defined as

$$J_2(f) = \int\int \left(\frac{\partial^2 f}{\partial u^2}\right)^2 + \left(\frac{\partial^2 f}{\partial u \partial v}\right)^2 + \left(\frac{\partial^2 f}{\partial v^2}\right)^2 dudv.$$

The solution to the minimization problem in (2.10) has the form,

$$\hat{f}(u,v) = \sum_{i=1} N a_i \varphi(||(u,v) - (u_i,v_i)||) + b_1 + b_2 u + b_3 v. \tag{2.11}$$

The estimator in (2.11) consists of two parts. The first part is a local non-affine component and the second part is a global affine component. The global affine component, $b_1 + b_2 u + b_3 v$, is an element in the function space spanned by linearly independent polynomials, for which J_2 is zero. The local non-affine component is a representation in terms of radial basis functions, $\varphi(\dot{)}$, defined by

$$\varphi(r) = r^2 log(r).$$

The coefficients $a_i, i = 1, \cdots, N$, b_1, b_2 and b_3 are pararmeters to be estimated. Let a be the vectors of $a_i, i = 1, \cdots, N$, and vector $b = [b_1, b_2, b_3]^T$. Define matrix Φ by $\Phi_{ij} = \varphi(||(u_i,v_i) - (u_j,v_j)||)$ and matrix X by $X_i = [1, u_i, v_i]$, where X_i is the ith row of X. Coefficient a is subject to linear constrains that $X^T a = 0$. Then the thin plate splin fitting problem becomes to minimize:

$$||Y - \Phi a - Xb||^2 + \lambda a^T \Phi a \text{ subject to} X^T a = 0, \tag{2.12}$$

A drawback of thin-plate splines is that they are computationally expensive. The number of parameters estimated is based on the number of observations, though the degrees of freedom for the model is usually much smaller than the number of observations. A low rank approximation may be used to reduce the computational cost of the model while maintaining

the performance of the smoother. PCA-based dimension reduction (Van Der Linde, 2003) is one approach commonly employed as the approximation.

To truncate the space of the components with respect to parameter a, we first find the eigen-decomposition of the matrix Φ so that $\Phi = UDU^T$, where the diagonal values in D (eigenvalues of Φ) are ordered by absolute values in descending order and the columns of U are the corresponding eigenvectors. The subspace spanned by the eigenvectors related to the k eigenvalues with largest absolute value provide an approximation of the space spanned by all the eigenvectors. Let D_k denote the top right $k \times k$ submatrix of D and let U_k be the first k columns of U. Then a can be approximated by $U_k a_k$, where $a_k \in \mathbb{R}^k$, and the resulting minimization problem in (2.12) is reduced to minizing

$$||Y - U_k D_k a_k - Xb||^2 + \lambda a_k^T D_k a_k \text{ subject to} X^T U_k a_k = 0. \tag{2.13}$$

Additionally, the linear constraints for a_k can be absorbed (Wood, 2006). First, a matrix Z_k with orthogonal columns can be found such that $X^T U_k Z_k = 0$ (for example, based on the QR decomposition of $U_k^T X$). Then let $a_k = Z_k \tilde{a}$, the constrained minimization problem in (2.13) results in an unconstrained minimization of:

$$||Y - U_k D_k Z_k \tilde{a} - Xb||^2 + \lambda \tilde{a}^T Z_k^T D_k Z_k \tilde{a}.$$

with respect to $\tilde{a}$ and b. The Lanczos iteration method (Lanczos, 1950) can be employed to find U_k and D_k at a lower cost of $O(n^2 k)$ operations.

2.4 Generalized Additive Models (GAMs) for Disease Mapping

In this section we review the general form of a GAM as well as the fitting algorithm used to implement the model. We conclude the section by illustrating the specification of a GAM for disease mapping in the context of an epidemiologic study.

2.4.1 Model Specification

Generalized additive models provide an extension of the generalized linear models introduced in Section 2.1 by replacing the parametric linear predictor with an additive predictor that incorporates non-parametric smoothing terms. To develop the model, suppose that the probability distribution of response Y belongs to the exponential family (see Section 2.1) and that there exists d covariates $X_1, \cdots, X_d$ that are to be used to model the response. In the absence of interaction effects among the covariates, the GAM assumes that

$$g(\mu) = \eta = \beta_0 + \sum_{j=0}^{d} s_j(X_j), \tag{2.14}$$

where $\mu = E[Y]$, $g(\cdot)$ is the link function, η is the additive predictor, and $s_j(\cdot)$ is a smooth term for covariate X_j, which can be estimated by any of the smoothing methods previously discussed in Section 2.3. In the trivial case, $s_j(\cdot)$ can also be a linear function reducing the model to the usual GLM. For identifiability, we constrain the model such that $\sum_{i=1}^{N} s_j(X_{ij}) = 0$ for each covariate X_j, $j = 1, \ldots, d$, with i representing the i^{th} observation, $i = 1, \ldots, N$.

For ease of exposition, we will focus on the model specification provided in (2.14) to illustrate the fitting procedure for a GAM in the following sections. While this specification excludes potential interaction terms among the covariates, it should be noted that an interaction effect

among two or more covariates can be included in a single smoothing term, thus maintaining the basic structure of the model. However, in practice, since the smooth term is to be non-parametrically estimated via a defined smoothing technique one must consider the increased variance that is likely to result when smoothing over multiple covariate interactions. Of course the tolerance for variance inflation, and ultimate performance of the model, will depend upon the size of data used for model fitting, but in general it is commonly suggested that no more than three covariates should be included in a single smooth term.

2.4.2 Backfitting Algorithm to Estimate the Additive Effects

In this section, we illustrate the backfitting algorithm for fitting a GAM. Starting from the simplest case and omitting the observation index for clarity, we begin by assuming that response Y follows the Gaussian distribution (a member of the exponential family of distributions) and that the model utilizes the identity link function. Under this specification, the mean model is given by

$$E[Y] = \mu = \beta_0 + \sum_{j=0}^{d} s_j(X_j).$$

For some s_j, $j = 1, \ldots, d$, if β_0 and $s_{j'}$ for $j' \neq j$ are all known, then s_j can be estimated by regressing the univariate *partial residual*, $R_j \equiv Y - \beta_0 - \sum_{j' \neq j} s_{j'}(X_{j'})$, on the specified smoother for X_j. Motivated from this, the backfitting procedure estimates $s_j, j = 1, \cdots, d$ with following steps:

1. Initialize $\hat{\beta}_0 = E[Y] = \mu$ and $\hat{s}_1 = \cdots, = \hat{s}_d = 0$;

2. Iterate:

   for $j = 1, \cdots, d$:

- Calculate the partial residual:

$$R_j = Y - \hat{\beta}_0 - \sum_{j' \neq j} \hat{s}_{j'}(X_{j'});$$

- Update $\hat{s}_j$ by regressing partial residual R_j on the smoother for X_j;
- Center the $\hat{s}_j$ obtained from last step at zero.
- For each observation, compute fitted value $\hat{Y}_i = \hat{\beta}_0 + \sum_{j=0}^{d} \hat{s}_j(X_{ij}),\ i = 1, \ldots, N.$

Repeat until: The change in $\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2$ from the past iteration to the current iteration is less than a defined convergence criteria.

Breiman and Freidman (1985) has shown that with the above backfitting algorithm, the prediction for $E[Y]$ is unique and is therefore the best additive approximation. For each smooth term, the uniqueness cannot be guaranteed for all classes of smoothers. However it has been shown that uniqueness is guaranteed for two of the most commonly used smoothers in the context of GAMs: cubic spline and kernel smoothers as discussed in Section 2.3.

2.4.3 Local Scoring Procedure

By way of introduction, we first illustrated the backfitting algorithm for the case of a Gaussian distributed response with identity link function. We now consider the more general case of model (2.14) by assuming that the probability distribution of the response belongs to the exponential family with arbitrary link function. In this case, modeling fitting implements a *local scoring* procedure that focuses on the local score function rather than the Fisher score function that commonly used when estimating standard GLMs, as shown in Section 2.1.

To motivate the local scoring procedure, let $l(\eta_i, Y_i)$ denote the log-likelihood function based upon the response and covariates for observation $i,\ i = 1, \ldots, N$. Estimation of η_i via

Fisher scoring as discussed in Section 2.1 maximizes the overall log-likelihood function $l = \sum_{i=1}^{N} l(\eta_i, y_i)$. However, in for a general response this procedure does not force the estimate of η_i to be smooth. For example, in the case of a logistic regression model for a binary outcome, Fisher scoring yields $\hat{\eta}_i = +\infty$ if $Y_i = 1$ and $\hat{\eta}_i = -\infty$ if $Y_i = 0$. To address this concern in the context of a GAM where smoothness of η is required, local scoring instead estimates η by maximizing the expected log-likelihood function based on the observation of corresponding subject:

$$E(l(\hat{\eta}, Y)) = \max_{\eta} E(l(\eta, Y)),$$

where the expectation is taken over the joint distribution of X and Y. Under standard regularity conditions (namely the ability to interchange integration and differentiation), we obtain

$$E[dl/d\eta]_{\hat{\eta}} = 0. \tag{2.15}$$

While there is no general closed for solution to (2.15), a first-order Taylor series expansion leads to an iterative estimating procedure given by

$$\eta^{new} = \eta^{old} - E[dl/d\eta]_{\eta^{old}} / E[d^2 l/d\eta^2]_{\eta^{old}},$$

which is equivalent to

$$\eta^{new} = E\left[\eta - \frac{dl/d\eta}{E[d^2 l/d\eta^2]}\right]_{\eta^{old}}.$$

When the distribution of the response belongs to the exponential family, the first and second

derivatives of the expected log likelihood are given by

$$\frac{dl}{d\eta} = (Y-\mu)V^{-1}\left(\frac{d\mu}{d\eta}\right),$$

and

$$\frac{d^2l}{d\eta^2} = (Y-\mu)V^{-1}\left(\frac{d}{d\eta}\right)\left[V^{-1}\left(\frac{d\mu}{d\eta}\right)\right] - \left(\frac{d\mu}{d\eta}\right)^2 V^{-1}. \tag{2.16}$$

Then taking the expectation (conditional on X) of equation (2.16) we obtain

$$E\left[\left(\frac{d^2l}{d\eta^2}\right)\middle| X\right] = -\left(\frac{d\mu}{d\eta}\right)^2 V^{-1}.$$

Hence η is updated by

$$\eta^{new} = E\left[\eta + (Y-\mu)\left(\frac{d\eta}{d\mu}\right)\Big|_{\eta^{old},\mu^{old}}\right]. \tag{2.17}$$

Further, letting Y_w^{old} denote the working response computed in terms of η^{old} and μ^{old} and given by

$$Y_w^{old} = \eta + (Y-\mu)\left(\frac{d\eta}{d\mu}\right)\Big|_{\eta^{old},\mu^{old}}, \tag{2.18}$$

we obtain from equations (2.14), (2.17), and (2.18),

$$E[Y_w^{old}] = \beta_0^{new} + \sum_{j=1}^{d} s_j^{new}.$$

As such, the coefficients β_0^{new} and s_j must be estimated in order to obtain an updated value of η^{new} in Eq. (2.17). This is achieved via the backfitting algorithm shown in Section 2.4.2.

Specifically, we begin by defining W as

$$W = (d\mu/d\eta)^2 V^{-1}|_{\eta^{old},\mu^{old}} \tag{2.19}$$

and initializing $s = 0$. The backfitting procedure is used to iteratively update each smooth term by regressing the partial residual on the smoother for X_j with weight W defined in equation(2.19) until it converges.

Putting the above together, the overall algorithm for fitting a generalized additive model is as follows:

1. Initialize $\beta_0 = E[Y]$ and $s_j = 0, j = 1, \cdots, d$.
2. Loop:
   (a) Based on the current estimates of β_0 and $s_j,\ j = 1, \cdots, d$, calculate η^{old} as well as the working response Y_w^{old} and corresponding weights W using equation(2.14), (2.18) and (2.19), respectively.
   (b) Update β_0 and $s_j,\ j = 1, \cdots, d$ via the backfitting algorithm.
3. Repeat 1. and 2. until convergence.

While the backfitting algorithm is a relatively efficient method to estimate additive effects, convergence can be slow if the covariates included in the linear predictor are correlated (Chambers and Hastie, 1992). To eliminate most of the problems associated with slow fitting due to multicolinearity when more than one smooth term is included in the linear predictor, it is beneficial for all of the linear terms in the model to be fitted together, treating them as a single term in the iterative procedure above. Moreover, it can be additionally beneficial to decompose each smooth term into a parametric (linear) and non-parametric (smooth)

component such that

$$s_j(X_j) = \beta_j X_j + s'_j(X_j), \tag{2.20}$$

where $s'_j(X_j)$ represents the non-parametric component, and the linear coefficient β_j is fitted together with the remaining parametric linear terms in the model.

2.4.4 Standard Errors

If no smooth terms are included in the specification of GAM the model reduces to a standard GLM where the covariance matrix for the estimated model parameters (MLEs) is given by the inverse of the Fisher information matrix. However, when smooth terms are present in the GAM, variance estimation requires computation of the *operation matrix* G_j for each smooth term s_j, such that $s_j = G_j z$, where z is the working response from the last iteration of the fitting algorithm described above and is asymptotically distributed as a Gaussian random variable. From this, the covariance matrix for the estimated s_j is given by $G_j Cov(z) G_j^T$, which can be estimated by $\hat{\phi} G_j W^{-1} G_j^T$, where W is a diagonal matrix with elements defined by the weights used in the last iteration of the fitting algorithm.

The operation matrix, G_j, tends to be computationally expensive to obtain for non-parametric or semi-parametric smoothing procedures, and hence approximations are often used when estimating $G_j Cov(z) G_j^T$. One approach is to approximate $\hat{\phi} G_j W^{-1} G_j^T$ by $\hat{\phi} G_j W^{-1}$, which is generally conservative for non-projection smoothers (Chambers and Hastie, 1992). In this case, G_j can be orthogonally decomposed into $G_j = H_j + N_j$, where H_j can be obtained as the design matrix corresponding to the parametric portion of equation (2.20), and N_j corresponds to the non-parametric portion. Thus, the variance of the estimated smooth term can be approximated via a decomposition of two variance components: (1) the variance from the parametric portion of (2.20) which captures the correlation all parametric terms that are

fitted together, and (2) the variance from the non-parametric portion of (2.20) reflecting the marginal information obtained in the smoothing terms.

2.4.5 Example: Disease Mapping Using GAMs

To motivate the use of GAMs for spatial disease mapping, we consider modeling observations that are distributed on a map with location coordinates u_i and v_i denoting the geographical parameters for the i^{th} observation, $i = 1, \ldots, N$. Let Y_i denote the outcome of interest and X_i denote a vector of adjustment covariates. Further suppose that the distribution of the outcome belongs to the exponential family. The GAM for a spatial effect analysis can be specified as

$$g(\mu_i) = \eta_i = \beta_0 + X_i^T\beta + s(u_i, v_i), \quad i = 1, \ldots, N,$$

where $g(\cdot)$ is the link function that links the mean of the outcome $\mu_i = E[Y_i]$ to the linear predictor, η_i. Further, because the distribution of the outcome belongs to the exponential family, the variance of the outcome is defined by the assumed probability model and denoted as $V_i \equiv Var(Y_i) = V(\mu_i, \phi)$, a function of the mean and nuisance parameter ϕ. In the above specification, β denotes a vector of coefficients associated with adjustment covariate X_i and $s(u_i, v_i)$ represents the spatial effects of interest, a nonlinear function of location.

As noted in the previous section, when fitting the model, we separate the spatial effect into parametric and non-parametric portions: $s(u_i, v_i) = \gamma_u u_i + \gamma_v v_i + f(u_i, v_i)$, and the model becomes

$$g(\mu_i) = \eta_i = \beta_0 + X_i^T\beta + \beta_u u_i + \beta_v v_i + f(u_i, v_i), \quad i = 1, \ldots, N,$$

In this case, the parametric portion of the spatial effect will be fitted jointly along with

the other adjustment variables, X_i, using least squares. The non-parametric portion will be fitted using a single non-parametric smoother. As previously stated, to ensure identifiability we constrain the model so that $\sum_i^N f(u_i, v_i) = 0$.

Estimating the non-parametric smooth term $f(u_i, v_i)$ requires two decisions: (1) the type of smoother and (2) the size of the neighborhood used by the smoother. As population densities often vary dramatically over geographic space, a locally weighted scatterplot smoother (LOESS) (Cleveland, 1979, 1981; Cleveland and Devlin, 1988) is typically used as the bivariate smoothing function for the two geolocation parameters u and v. LOESS is a natural choice as it is able to adapt the size of the smoothing neighborhood to the local density while maintaining the smoothness features of a kernel. As previously noted, this method defines the neighborhood based on the k nearest subjects, and weights points within the neighborhood using a tri-cube distance function centered at a target point and decreasing to zero at the furthest neighbor (Hastie and Tibshirani, 1986). Other smoothing methods, as discussed in Section 2.3, could also be utilized.

The amount of smoothing performed by loess depends on the number and distribution of points in the neighborhood. A small neighborhood reduces bias but increases variance, while a large neighborhood increases bias and reduces variability. As such, the guiding principle behind the selection of a neighborhood size (also called *span* or *bandwidth*) is to provide a trade-off between bias and variance and is generally operationalized via mean squared error or MSE. Multiple criteria have been proposed for choosing an appropriate span size. Cross-validation for span selection is discussed in (Friedman and Stuetzle, 1984), and provides a computational approach to estimating out-of-sample MSE by cleverly stratifying and resampling from a training dataset. Briefly, K-fold cross-validation randomly divides an existing training dataset into K roughly equal parts, performs model fitting on $K - 1$ of those parts and assesses out-of-sample MSE on the remaining K-th part. This process is repeated until all K strata have served as hold-out datasets, at which time out-of-sample

MSE is averaged over all K hold-outs. Leave-one-out, or $N - fold$, cross-validation takes K equal to the sample size of the training data in the K-fold procedure. By focusing on out-of-sample prediction error, cross-validation seeks to avoid the over-fitting issues (ie. low bias and high variability that results in lack of generalizability) that are encountered when model selection utilizes the full dataset in both a training and testing capacity. Extensions of K-fold cross-validation have also been proposed. Notably, Kelsall and Diggle (1998) used a weighted least squared cross-validation to choose an optimal degree of smoothing for the kernel smoother, while a generalized cross-validation criterion was proposed by Wood (2006) for cubic spline smoothing. For an analytic approximation to out-of-sample prediction error when taking the loss function to be the log-likelihood function, minimization of Akaike's Informaiton Criterion (AIC) was proposed for non-parametric smoother selection in Hurvich et al. (1998)and Webster et al. (2006).

Along with estimation of spatial effects, inference regarding whether location is associated with outcomes is often desired in disease mapping studies. To this end, a global test of spatial effects can be conducted via a likelihood ratio test. This test is formulated by comparing the deviance between a full model (including the spatial smoother) and a reduced model (omitting the spatial smoother). For the full model, the degrees of freedom of the non-parametric portion can be approximated by $tr(S) - 1$ (Chambers and Hastie, 1992), where tr is the trace function and S denotes the smoothing matrix. The degrees of freedom of the parametric portion of the GAM are given by $d + 3$, where d denotes the length of covariate vector X. Thus, the degrees of freedom of the full model is $tr(S) + d + 2$, and the degrees of freedom for the likelihood ratio test statistic is $tr(S) + 1$. Relying on standard likelihood theory, under the null hypothesis of no spatial effect, the likelihood ratio test statistic is approximately distributed as a χ^2 random variable with the degrees of freedom $tr(S) + 1$.

Due to the data-driven nature of choosing the smoothing span size and the approximations involved in calculating the degrees of freedom for the GAM, the likelihood ratio test may

result in anti-conservative inference implying that nominal type I error rates are higher than the desired level of the test. In light of this, a permutation test (Kelsall and Diggle, 1998; Webster et al., 2006) has been proposed to test for a global spatial effect and pointwise significance of particular areas. To conduct the test, the locations of individuals are randomly permuted while preserving the observed response and adjustment covariate structure. Based on each permuted dataset, the deviance statistic corresponding to the spatial effect prediction for each observation is then calculated. The procedure is repeated multiple times resulting in an empirical distribution of permutation deviance statistics. Dividing the rank of the observed deviance statistics by the number of permutations provides a p-value to test the global adjusted spatial effects. Analogously, a p-value for testing the pointwise spatial effect of a particular location can be obtained by dividing the rank of the observed spatial effect at that location by number of permutations.

2.5 Intrinsic Gaussian Markov Random Fields

As background for the research developed in Chapters 4 and 5, in this section we briefly introduce the Intrinsic Gaussian Markov Random Field (IGMRF). The IGMRF is a type of Gaussian Markov Random Field (GMRF), partly characterized by a precision matrix that is not of full rank. It is because of this property that IGMRFs are often utilized as prior distributions in various applications (Rue and Held, 2005), especially for spatial analysis (Assuncao et al., 2002; Gamerman et al., 2003; Lang et al., 2002; Yue and L., 2010; Yue et al., 2010).

2.5.1 Gaussian Markov Random Fields

We start from the Gaussian Markov Random Field (GMRF). Let $z = z_1, \cdots, z_n$ denote a random variable following a multivariate normal distribution with mean μ and covariance matrix Σ. Define an undirected graph $G = (V, E)$, where $V = 1, \cdots, n$ denotes the index for z, and E records edges between nodes such that there is no edge between two nodes provided that the two nodes are independent conditional on all other nodes. Then we say that z is a GMRF with respect to the graph G.

It has been proven by Rue and Held (2005) that for $i \neq j$, and letting $Q = \Sigma^{-1} > 0$ denote the precision matrix for z, we have

$$z_i \perp z_j | z_{-ij} \iff Q_i j = 0.$$

Thus the nonzero elements of Q determine the graph G, and for a given graph G, we know the nonzero elements in Q. Then the GMRF with respect to a graph G has density

$$\pi(z) = (2\pi)^{-N/2} |Q|^{1/2} exp\left(-\frac{1}{2}(z-\mu)^T Q (z-\mu)\right),$$

and

$$Q_{ij} \neq 0 \iff i, j \in E \text{ for all } i \neq j.$$

A useful property of a GMRF is the local Markov property, implying that conditional independence can be extracted from the graph G. For a GMRF with regard to a graph G, the following Markov properties are equivalent (Speed and Kiiveri, 1986):

Definition 2.1. *The pairwise Markov property:*

$$z_i \perp z_j | z_{ij}, \quad \text{if } i, j \notin E \text{ and } i \neq j;$$

Definition 2.2. *The local Markov property:*

$$z_i \perp z_{-\{i,ne(i)\}} | z_{ne(ij)}, \quad \forall i \in E,$$

where $ne(i)$ is the neighbors of z_i;

Definition 2.3. *The global Markov property:*

$$z_A \perp z_B | z_C,$$

where sets A, B and C are disjoint, and every path from A to B passes C, i.e. C separates A and B.

From the above definitions, it is clear that the pairwise Markov property describes the conditional independence between two nodes, the local Markov property describes the conditional independence between a node and its neighboring nodes, and the global Markov property is related to the conditional independence between two disjoint sets.

2.5.2 Intrinsic Gaussian Markov Random Fields

We now turn our attention to IGMRFs. An IGMRF is a type of GMRF conditional on linear constraints. More specifically, let z be a mean-zero GMRF with dimension n and precision matrix $Q > 0$. Now consider a linear constraint on z such that $Az = a$, where A is a $k \times n$ ($k < n$) full rank matrix. Then the conditional probability density of z can be derived as follows.

We being by writing the precision matrix Q as

$$Q = UDU^T,$$

where $D = \text{diag}\{\lambda^1, \cdots, \lambda_n\}$ and $\lambda_i, i = 1, \cdots, n$ are the eigenvalues of Q and the columns of the matrix U are the corresponding eigenvectors. Let U_k denote the first k columns of U so that $U_k^T U_k = I_k$. Further, define $C \equiv AU_k$. Then C is also full rank, and $U_k^T z = C^{-1}a \equiv \hat{a}$ is equivalent to the constraint $Az = a$.

Now, let $y = U^T z$, which also follows a multivariate normal distribution with

$$\begin{aligned} &E[y] = U^T E[z] = 0, \quad \text{and} \\ &Var[y] = U^T Var(z) U = U^T Q^{-1} U = U^T U D^{-1} U^T U = D^{-1}. \end{aligned}$$

Further considering $U_k^T z = (y_1, \cdots, y_k)$, with $U^T x = \hat{a}$, we have $E[y|Az = a] = [\hat{a}^T, 0^T]^T$ and $prec[y|Az = a] = \tilde{D}$, where $\tilde{D} = \text{diag}\{0, \cdots, 0, \lambda_{k+1}, \lambda_n\}$. From this we have that $E[z|Az = a] = V[\hat{a}^T, 0^T]^T \equiv V\tilde{a}$ and $Prec[z|Az = a] = U\tilde{D}U^T \equiv \tilde{Q}$. Therefore, the conditional probability density of z is is given by

$$\begin{aligned} \pi(z|Az = a) &= (a\pi)^{-(n-k)/2} \textstyle\prod_{i=k+1}^{n} \lambda^{1/2} \exp\left\{-\frac{1}{2}(z - V\tilde{a})^T \tilde{Q}(z - V\tilde{a})\right\} \\ &= (a\pi)^{-(n-k)/2} (|\tilde{Q}|^*)^{1/2} \exp\left\{-\frac{1}{2}(z - V\tilde{a})^T \tilde{Q}(z - V\tilde{a})\right\}, \end{aligned} \tag{2.21}$$

where $|\cdot|^*$ denotes the generalized determinant, or product of the non-zero eigenvalues.

An improper GMRF is defined based on the density function in (2.21). Let $Q \geq 0$ be an $n \times n$ matrix with rank $n - k$. Then $z = [z_1, \cdots, z_n]^T$ is an *improper GMRF* of rank $n - k$ with parameters (μ, Q) relative to a graph G if the density of z can be written as

$$\pi(z) = (a\pi)^{-(n-k)/2} (|Q|^*)^{1/2} \exp\left\{-\frac{1}{2}(z - \mu)^T Q(z - \mu)\right\},$$

where

$$Q_{ij} \neq 0 \iff i,j \in E \quad \forall i \neq j.$$

Importantly, $z \in \mathbf{R}^n$ can be decomposed into two parts:

$$z = z^{\|} + z^{\perp},$$

where $z^{\|}$ is the part of z in the subspace spanned by the columns of A, which is the null space of $\tilde{Q}$, and $z^{\perp}$ is the part of z orthogonal to $z^{\|}$. Thus $\pi z = \pi z^{\perp}$, so that $\pi(\dot{)}$ is invariant to the addition of any $x^{\|}$, a critical feature of improper GMRFs.

The above leads to the generation of an intrinsic GMRF (IGMRF). Specifically, an *IGMRF of first order* is an improper GMRF of rank $n-1$ where $Q\mathbf{1} = 0$, which implies that $\sum_j Q_{ij} = 0$, for all i, $i = 1, \ldots, n$. To illustrate the features of the IGMRF, we consider $A^T = \mathbf{1}$ and thereby note that the density for an IGMRF of first order is invariant to the addition of the overall mean level. For example, taking $\mu = 0$ we have

$$E[z_i | z_{-i}] = -\frac{1}{Q_{ii}} \sum_{j:(i,j)\in E} Q_{ij} z_j,$$

with $-\sum_{j:(i,j)\in E} Q_i j / Q_i i = 1$. Therefore, the conditional mean of z_i is the weighted mean of the neighboring nodes, but does not involve the overall mean level. With this feature, many IGMRFs are constructed such that the deviation from the overall level is a smooth surface in space, making them an attractive choice as a prior for Bayesian thin-plate splines as considered in the next section.

2.5.3 An IGMRF prior for Bayesian Thin-Plate Splines

Recall the model described in (2.9) where the thin-plate spline estimator, f, is the solution to the minimization problem:

$$\hat{f} = \arg\min_f \left\{ \sum_i (y_i - f(u_i, v_i))^2 + \lambda J_2(f) \right\},$$

where $J_2(f)$ is the smoothing penalty for function f given by

$$J_2(f) = \int\int_{\mathbb{R}^2} \left(\frac{\partial^2 f(u,v)}{\partial u^2} \right)^2 + \left(\frac{\partial^2 f(u,v)}{\partial u \partial v} \right)^2 + \left(\frac{\partial^2 f(u,v)}{\partial v^2} \right)^2 dudv.$$

For the smooth function f, it is reasonable to assume that the integrable derivatives of order up to 3 vanish at infinity and hence guaranteeing sufficient smoothness. In this case, the second term in the integral of $J_2(f)$ can be written (Yue and L., 2010) as

$$\begin{aligned}
\int\int_{\mathbb{R}^2} \left(\tfrac{\partial^2 f(u,v)}{\partial u \partial v} \right)^2 dudv &= \int\int_{\mathbb{R}^2} \left(\tfrac{\partial^2 f(u,v)}{\partial u \partial v} \right) \left(\tfrac{\partial^2 f(u,v)}{\partial u \partial v} \right) dudv \\
&= \int \left(\tfrac{\partial^2 f(u,v)}{\partial u \partial v} \right) \left(\tfrac{\partial f(u,v)}{\partial u} \right) \Big|_{-\infty}^{\infty} du - \int\int_{\mathbb{R}^2} \left(\tfrac{\partial^3 f(u,v)}{\partial u \partial v^2} \right) \left(\tfrac{\partial f(u,v)}{\partial u} \right) dudv \\
&= -\int \left(\tfrac{\partial^2 f(u,v)}{\partial v^2} \right) \left(\tfrac{\partial f(u,v)}{\partial u} \right) \Big|_{-\infty}^{\infty} dv + \int\int_{\mathbb{R}^2} \left(\tfrac{\partial^2 f(u,v)}{\partial u^2} \right) \left(\tfrac{\partial^2 f(u,v)}{\partial v^2} \right) dudv \\
&= \int\int_{\mathbb{R}^2} \left(\tfrac{\partial^2 f(u,v)}{\partial u^2} \right) \left(\tfrac{\partial^2 f(u,v)}{\partial v^2} \right) dudv
\end{aligned}$$

Thus we can rewrite the penalty term $J_2(f)$ as

$$\begin{aligned}
J_2(f) &= \int\int_{\mathbb{R}^2} \left(\tfrac{\partial^2 f(u,v)}{\partial u^2} \right)^2 + \left(\tfrac{\partial^2 f(u,v)}{\partial u^2} \right) \left(\tfrac{\partial^2 f(u,v)}{\partial v^2} \right) + \left(\tfrac{\partial^2 f(u,v)}{\partial v^2} \right)^2 dudv \\
&= \int\int_{\mathbb{R}^2} \left[\left(\tfrac{\partial^2}{\partial u^2} + \tfrac{\partial^2}{\partial v^2} \right) f(u,v) \right]^2 dudv.
\end{aligned}$$

From this, a thin-plate spline estimator can also be derived as the solution to the following

minimization problem:

$$\arg\min_f \left[\sum_i (y_i - f(u_i, v_i))^2 + \lambda \int\int_{\mathbb{R}^2} \left[\left(\frac{\partial^2}{\partial u^2} + \frac{\partial^2}{\partial v^2} \right) f(u,v) \right]^2 dudv \right]. \tag{2.22}$$

Now consider a modification of Eq. (2.9). Specifically, suppose response y is taken on a regular lattice (u_j, v_k), with

$$y_{jkl} = f(u_j, v_k) + \epsilon, \epsilon \sim^{iid} \mathcal{N}(0, \tau^{-1}),\ j = 1, \cdots, n_u;\ k = 1, \cdots, n_v;\ l = 1, \cdots, r_{jk}$$

where u_j and v_k are equally spaced by distance h in vertical and horizontal directions, respectively, and n_u denotes the number of rows while n_v denotes the number of columns for the lattice. From this specification there are $n = n_u \times n_v$ grid points in total and we let r_{jk} denote the number of observations at location (u_j, v_k). Note that r_{jk} can be 0 since it is not required that there are observations at each grid point. From this, the total number of observations is given by

$$N = \sum_{j=1}^{n_u} \sum_{k=1}^{n_v} r_{jk}.$$

Now, let z denote the $n \times 1$ vector defined by $f(u_j, v_k)$, such that $z_{j+n_u(k-1)} = f(u_j, v_k)$, for $j = 1, \cdots, n_u$, and $k = 1, \cdots, n_v$. Let y and ϵ be the vectors of observed responses and random errors, respectively, and let D be the $N \times n$ *incidence matrix* such that $D_{ij} = 1$ if y_i is observed at location j, which is location $(u_{j \bmod n_u}, v_{\lfloor j/n_u \rfloor + 1})$, in the vectorized ordering of the lattice, and $D_{ij} = 0$, otherwise. Then the vectorized model is written as

$$y = Dz + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \tau^{-1} I_N).$$

Though not explicitly stated, it is worth noting that for non-regular data, the data can be

binned to a regular lattice (Hardle and Scott, 1992; Scott, 2003).

The GMRF prior is motivated by a difference approximation of the penalty term in equation (2.22). The second derivative of f at location (u_j, v_k) can be approximated by second-order backward difference operators:

$$\frac{\partial^2}{\partial u^2} f(u_j, v_k) \approx h^{-2} \bigtriangledown^2_{(1,0)} f(u_j, v_k);$$
$$\frac{\partial^2}{\partial v^2} f(u_j, v_k) \approx h^{-2} \bigtriangledown^2_{(0,1)} f(u_j, v_k),$$

where the second-order backward difference operators, $\bigtriangledown^2_{(1,0)}$ and $\bigtriangledown^2_{(0,1)}$, are defined as

$$\bigtriangledown^2_{(1,0)} f(u_j, v_k) = f(u_{j+1}, v_k) - 2f(u_j, v_k) + f(u_{j-1}, v_k);$$
$$\bigtriangledown^2_{(1,0)} f(u_j, v_k) = f(u_j, v_{k+1}) - 2f(u_j, v_k) + f(u_j, v_{k-1}).$$

Thus the differential operator in $J_2(f)$ at location (u_j, v_k) can be approximated by

$$h^{-4} \left[\left(\bigtriangledown^2_{(1,0)} + \bigtriangledown^2_{(0,1)} \right) f(u_j, v_k) \right]^2$$
$$= h^{-4} [z_{j+1+n_u(k-1)} + z_{j+1+n_u(k-1)} + z_{j+n_u(k)} + z_{j+n_u(k-2)} - 4z_{j-1+n_u(k-1)}]^2.$$

Now, let d be the $(n_u - 2)(n_v - 2) \times 1$ vector of differential operators in $J_2(f)$, with

$$d_{(j-1)+(n_u-2)(k-2)} = z_{j+1+n_u(k-1)} + z_{j+1+n_u(k-1)} + z_{j+n_u(k)} + z_{j+n_u(k-2)} - 4z_{j-1+n_u(k-1)}, \quad (2.23)$$

for $j = 2, \cdots, n_u - 2$ and $k = 2, \cdots, n_v - 2$. Then we have $d = B_0 z$, where B_0 is an $(n_u - 2)(n_v - 2) \times n$ matrix with coefficients in equation (2.23). As a result, an approximation of the thin-plate spline penalty $J_2(f)$ is given by

$$\frac{1}{h^4} \sum_{j=2}^{n_u-1} \sum_{k=2}^{n_v-1} [z_{j+1+n_u(k-1)} + z_{j+1+n_u(k-1)} + z_{j+n_u(k)} + z_{j+n_u(k-2)} - 4z_{j-1+n_u(k-1)}]^2.$$

Defining $A_0 = B_0^T B_0$, we have

$$J(f) \approx d^T d = z^T B_0^T B_0 z = z^T A_0 z.$$

Further letting $\lambda_h = \lambda / h^4$, the estimate $\hat{z}$ is the solution to

$$\hat{z} = \arg\min_z [||y - Dz||^2 + \lambda_h z^T A_0 z],$$

which suggests the prior on z for Bayesian modeling of thin-plate splines should be of the form

$$z|\delta \propto \delta^{(n-m)/2(|A_0|^*)_+^{1/2}} \exp\left(-\tfrac{\delta}{2} z^T A_0 z\right),$$

where $m = 2(n_u + n_v - 2)$ is the dimension of the null space of A_0 and δ is a precision parameter. Note that the random vector z is an improper GMRF because it follows an improper multivariate Gaussian distribution and satisfies the local Markov properties previously specified. Specifically, the conditional distribution of each $z_{j+n_u(k-1)}$ is Gaussian and only depends on its neighbor. It is convenient to use graphical notation to present the the conditional expectation and precisions of an interior $z_{j+n_u(k-1)}$ as

$$E[z_{j+n_u(k-1)}|z_{-[j+n_u(k-1)]}] = \tfrac{1}{20} \left(\begin{array}{ccccc} & \circ & \circ & \circ & \\ \circ & \circ & \bullet & \circ & \circ \\ 8\circ & \bullet & * & \bullet & \circ \\ \circ & \circ & \bullet & \circ & \circ \\ & \circ & \circ & \circ & \end{array} \right) - 2 \left(\begin{array}{ccccc} & \circ & \circ & \circ & \\ \circ & \circ & \bullet & \circ & \circ \\ \circ & \bullet & * & \bullet & \circ \\ \circ & \circ & \bullet & \circ & \circ \\ & \circ & \circ & \circ & \end{array} \right) - 1 \left(\begin{array}{ccccc} & \circ & \circ & \bullet & \\ \circ & \circ & \circ & \circ & \circ \\ \bullet & \circ & * & \circ & \bullet \\ \circ & \circ & \circ & \circ & \circ \\ & \circ & \circ & \bullet & \end{array} \right);$$

$$Prec[z_{j+n_u(k-1)}|z_{-[j+n_u(k-1)]}] = 20\delta,$$

where the conditional expectation at location "$*$" is a weighted summation over all values at locations with "$\bullet$", the number in front of each grid denotes the weight given to the corresponding "$\bullet$" locations, and the "$\circ$"s fix the spatial configuration.

As specified, the improper GMRF prior on z with a precision matrix A_0 is not attractive in a Bayesian thin-plate spline setting because the dimension of the null space of A_0 is too large and hence the basis of the null space is non-interpretable (Kneib, 2006). To address this, boundary terms mjst be added to fix the rank-deficiency. Specifically, with

$$\begin{aligned}
\bigtriangledown_{(1,0)} f(u_j, v_k) &= f(u_j, v_k) - f(u_{j-1}, v_k) \\
\bigtriangledown_{(0,1)} f(u_j, v_k) &= f(u_j, v_k) - f(u_j, v_{k-1}),
\end{aligned}$$

the corrections made at the four corners and edges are

$$\begin{aligned}
\bigtriangledown_1 f(u_1, v_1) &\equiv (\bigtriangledown_{(1,0)} + \bigtriangledown_{(0,1)}) f(u_2, v_2) \\
\bigtriangledown_2 f(u_{n_u}, v_1) &\equiv -\bigtriangledown_{(1,0)} f(u_{n_u}, v_1) + \bigtriangledown_{(0,1)} f(u_{n_u}, v_2) \\
\bigtriangledown_3 f(u_1, v_{n_v}) &\equiv \bigtriangledown_{(1,0)} f(u_2, v_{n_v}) + \bigtriangledown_{(0,1)} f(u_1, v_{n_v}) \\
\bigtriangledown_4 f(u_{n_u}, v_{n_v}) &\equiv -(\bigtriangledown_{(1,0)} + \bigtriangledown_{(0,1)}) f(u_{n_u}, v_{n_v}) \\
\bigtriangledown_5 f(u_j, v_1) &\equiv \bigtriangledown^2_{(1,0)} f(u_{j+1}, v_1) + \bigtriangledown_{(0,1)} f(u_j, v_2) \\
\bigtriangledown_6 f(u_j, v_{n_v}) &\equiv \bigtriangledown^2_{(1,0)} f(u_{j+1}, v_{n_v}) - \bigtriangledown_{(0,1)} f(u_j, v_{n_v}) \\
\bigtriangledown_7 f(u_1, v_k) &\equiv \bigtriangledown_{(1,0)} f(u_2, v_k) + \bigtriangledown^2_{(0,1)} f(u_1, v_{k+1}) \\
\bigtriangledown_8 f(u_{n_u}, v_k) &\equiv -\bigtriangledown_{(1,0)} f(u_{n_u}, v_k) + \bigtriangledown^2_{(0,1)} f(u_{n_u}, v_{k+1})
\end{aligned} \tag{2.24}$$

Letting $\tilde{d}$ be the $n \times 1$ vector of differential operators (including the correctted operators in (2.24)) at ordered locations, we have $\tilde{d} = Bz$. Then an improved approximation of $J_2(f)$ can be expressed as

$$\begin{aligned}
J_2(f) \approx \; & \tfrac{1}{h^4} \Bigg[\sum_{j=2}^{n_u - 1} \sum_{k=2}^{n_v - 1} \Big((\bigtriangledown^2_{(1,0)} + \bigtriangledown^2_{(0,1)}) f(u_j, v_k) \Big)^2 + \left(\bigtriangledown_1 f(u_1, v_1)\right)^2 + \left(\bigtriangledown_2 f(u_{n_u}, v_1)\right)^2 \\
& + \left(\bigtriangledown_3 f(u_1, v_{n_v})\right)^2 + \left(\bigtriangledown_4 f(u_{n_u}, v_{n_v})\right)^2 + \sum_{j=2}^{n_u - 1} \left(\left(\bigtriangledown_5 f(u_j, v_1)\right)^2 + \left(\bigtriangledown_6 f(u_j, v_{n_v})\right)^2 \right) \\
& + \sum_{k=2}^{n_v - 1} \left(\left(\bigtriangledown_7 f(u_1, v_k)\right)^2 + \left(\bigtriangledown_8 f(u_{n_u}, v_k)\right)^2 \right) \Bigg] .
\end{aligned}$$

Further letting $A = B^T B$, the roughness penalty term can be written as

$$J_2(f) \approx= \tilde{d}^T \tilde{d} = z^T B^T B z = z^T A z.$$

It then follows that a random vector z with prior distribution

$$z|\delta \propto \delta^{(n-1)/2(|A|^*)_+^{1/2}} \exp\left(-\tfrac{\delta}{2} z^T A z\right),$$

is an IGMRF of first order since the rank of A is $n-1$, the local Markov property remains satisfied, and $A\mathbf{1} = 0$ after the boundary corrections are implemented. The conditional expectation of z at a border or corner are shown as follows using graphical notation:

$$E[z_{2+n_u(k-1)}|z_{-[2+n_u(k-1)]}] = \frac{1}{20}\left(8\begin{smallmatrix}\circ&\circ&\circ&\circ&\circ\\ \circ&\circ&\bullet&\circ&\circ\\ \circ&\bullet&*&\bullet&\circ\\ \circ&\circ&\circ&\circ&\circ\end{smallmatrix} + 7\begin{smallmatrix}\circ&\circ&\circ&\circ&\circ\\ \circ&\circ&\circ&\circ&\circ\\ \circ&\circ&*&\circ&\circ\\ \circ&\circ&\bullet&\circ&\circ\end{smallmatrix} - 2\begin{smallmatrix}\circ&\circ&\circ&\circ&\circ\\ \circ&\bullet&\circ&\bullet&\circ\\ \circ&\circ&*&\circ&\circ\\ \circ&\bullet&\circ&\bullet&\circ\end{smallmatrix} - 1\begin{smallmatrix}\circ&\circ&\bullet&\circ&\circ\\ \circ&\circ&\circ&\circ&\circ\\ \bullet&\circ&*&\circ&\bullet\\ \circ&\circ&\circ&\circ&\circ\end{smallmatrix}\right);$$

$$E[z_{2+n_u}|z_{-[2+n_u]}] = \frac{1}{20}\left(8\begin{smallmatrix}\circ&\circ&\circ\circ\\ \circ&\bullet&\circ\circ\\ \circ&*&\bullet\circ\\ \circ&\circ&\circ\circ\end{smallmatrix} + 7\begin{smallmatrix}\circ&\circ&\circ\circ\\ \circ&\circ&\circ\circ\\ \bullet&*&\circ\circ\\ \circ&\bullet&\circ\circ\end{smallmatrix} - 2\begin{smallmatrix}\circ&\circ&\circ\circ\\ \bullet&\circ&\bullet\circ\\ \circ&*&\circ\circ\\ \bullet&\circ&\bullet\circ\end{smallmatrix} - 1\begin{smallmatrix}\circ&\bullet&\circ\circ\\ \circ&\circ&\circ\circ\\ \circ&*&\circ\bullet\\ \circ&\circ&\circ\circ\end{smallmatrix}\right);$$

$$E[z_{1+n_u(k-1)}|z_{-[1+n_u(k-1)]}] = \frac{1}{12}\left(7\begin{smallmatrix}\circ&\circ&\circ&\circ&\circ\\ \circ&\circ&\bullet&\circ&\circ\\ \circ&\circ&*&\circ&\circ\end{smallmatrix} + 6\begin{smallmatrix}\circ&\circ&\circ&\circ&\circ\\ \circ&\circ&\circ&\circ&\circ\\ \circ&\bullet&*&\bullet&\circ\end{smallmatrix} - 2\begin{smallmatrix}\circ&\circ&\circ&\circ&\circ\\ \circ&\bullet&\circ&\bullet&\circ\\ \circ&\circ&*&\circ&\circ\end{smallmatrix} - 1\begin{smallmatrix}\circ&\circ&\bullet&\circ&\circ\\ \circ&\circ&\circ&\circ&\circ\\ \bullet&\circ&*&\circ&\bullet\end{smallmatrix}\right);$$

$$E[z_{1+n_u}|z_{-[1+n_u]}] = \frac{1}{12}\left(7\begin{smallmatrix}\circ&\circ&\circ\circ\\ \circ&\bullet&\circ\circ\\ \circ&*&\circ\circ\end{smallmatrix} + 6\begin{smallmatrix}\circ&\circ&\circ\circ\\ \circ&\circ&\circ\circ\\ \circ&*&\bullet\circ\end{smallmatrix} + 5\begin{smallmatrix}\circ&\circ&\circ\circ\\ \circ&\circ&\circ\circ\\ \bullet&*&\circ\circ\end{smallmatrix} - 2\begin{smallmatrix}\circ&\circ&\circ\circ\\ \bullet&\circ&\bullet\circ\\ \circ&*&\circ\circ\end{smallmatrix} - 1\begin{smallmatrix}\circ&\bullet&\circ\circ\\ \circ&\circ&\circ\circ\\ \circ&*&\circ\bullet\end{smallmatrix}\right);$$

$$E[z_{1+n_u}|z_{-[1+n_u]}] = \frac{1}{6}\left(5\begin{smallmatrix}\circ&\circ&\circ\\ \bullet&\circ&\circ\\ *&\bullet&\circ\end{smallmatrix} - 2\begin{smallmatrix}\circ&\circ&\circ\\ \circ&\bullet&\circ\\ *&\circ&\circ\end{smallmatrix} - 1\begin{smallmatrix}\bullet&\circ&\circ\\ \circ&\circ&\circ\\ *&\circ&\bullet\end{smallmatrix}\right).$$

It can be further shown that the posterior distribution of z is multivariate normal with mean $S_{\lambda_h} D^T y$ and covariance $\tau^{-1} S_{\lambda_h}$, where $S_{\lambda_h} = (D^T D + \lambda_h A)^{-1}$, and hence $D^T D + \lambda_h A$ must be invertible for the posterior to exist. From this, it is evident that the posterior mean of a smooth function value at a location is the weighted summation over the observations at the

location and the estimates at neighboring locations. The smoothing parameter λ_h controls the trade-off between the observed data at the target location and that at neighboring locations.

Chapter 3

Spatial Analysis for Censored Survival Data and the R: MapGAM Package

3.1 Introduction

In spatial epidemiology studies, mapping crude and adjusted spatial distributions of disease risk is a useful tool for identifying risk factors of public health concern (Elliott and Wartenberg, 2004). The underlying (or crude) geographic pattern of disease is often what is observed by public health practitioners, but these patterns may be due to important spatially-varying predictors such as socioeconomic status, race/ethnicity, or environmental exposures. Individual-level spatial analyses can provide insight regarding disease risk by adjusting for these variables without aggregation bias (also known as ecological bias). Disease risks often have complex spatial patterns that are subject to high variability due to sparsity. Smoothing provides an efficient method to deal with these issues by borrowing strength from adjacent observations to reduce variability while allowing for non-parametric flexibility when estimating the spatial distribution of risk. Generalized additive models (GAMs), originally

proposed by Hastie and Tibshirani (1986), are common model-based approaches for mapping point-based epidemiologic data(Webster et al., 2006; Vieira et al., 2008; Baker et al., 2011; Akullian et al., 2014; Bristow et al., 2014; Hoffman et al., 2015). GAMs provide a unified statistical framework that allows for the adjustment of individual-level risk factors when evaluating spatial variability in a flexible way.

There has been a number of R (R Core Team, 2015) packages implementing GAMs and related models. The `gam` package (Hastie, 2004) (originally made for S-PLUS) provides an implementation of the GAM framework of Hastie and Tibshirani (1986) by providing two types of commonly used smoothing methods: cubic smoothing splines (Wahba, 1990; Green and Silverman, 1994) for univariate variables and local kernel smoothing (LOESS) (Cleveland, 1979, 1981; Cleveland and Devlin, 1988) for multivariate variables. The `mgcv`(Wood, 2009; Breslow and Clayton, 1993) package implements cubic smoothing splines and tensor product smooths, an extension of cubic splines to multiple dimensions. `mgcv` also provides various criterion to aid in the selection of model complexity via the choice of effective degrees of freedom and provides functions to fit generalized additive mixed effects models (GAMMs) for correlated data. Package `gamlss` (Rigby and Stasinopoulos, 2005; Stasinopoulos and Rigby, 2007) implements an extension of the GAM that incorporates selected distributions outside of the exponential family. Neither the `gam` package nor the `gamlss` package provides standard errors for prediction points, which is necessary when providing point-wise confidence bands for disease maps and identifying significant areas.

With respect to censored survival data, parametric additive models can be fit using either the `gamlss.cens` package (Stasinopoulos et al., 2015) or the `VGAM` package (W., 2007). However, none of the above packages provide an implementation of the Cox proportional hazard additive model for censored survival data that allows for multivariate smoothing of covariates, despite the fact that spatial effect estimation in the context of survival outcomes is of great interest in epidemiology studies (Henderson et al., 2002; Bristow et al., 2014). The `mgcv`

package implements the Cox model with a tensor product smooth, but knots are difficult to choose for population-based study where subjects are usually not uniformly geographically distributed. LOESS smoothing is suggested for the spatial analysis for epidemilogy(Webster et al., 2006) and to the best of our knowledge there is no existing R package that implements a Cox model with a bivariate LOESS smoothing term.

Moreover, displaying spatial predictions on a map with irregular geographic boundaries is a non-trivial effort, often handled by exporting statistical predictions to separate specialized geographic information system (GIS) software such as ArcGIS that requires a paid user license (Webster et al., 2006; Vieira et al., 2008) or by omitting geographic boundaries altogether(Akullian et al., 2014). At best, these limitations and complexities pose a significant barrier to researchers not already well versed in both GAMs and GIS methods and at worst may lead to reporting errors due to the inefficient transfer of estimates between separate software packages. As such, having a unified system for estimating and visualizing covariate-adjusted spatial effects on outcomes arising from the most commonly encountered epidemiologic study designs would greatly improve the ability of data scientists to conduct efficient and reproducible analyses in these settings.

To address the above deficiencies of current software, `MapGAM` was built to provide a single R package that allows for estimating, predicting, and visualizing covariate-adjusted spatial effects using individual-level data. The package estimates covariate-adjusted spatial associations with a univariate outcome via GAMs that include a non-parametric bivariate smooth term of geolocation parameters. Estimation and mapping methods are implemented for continuous, discrete, and right-censored survival data. In addition, support functions for efficient control sampling in case-control studies and inferential procedures for testing global and pointwise spatial effects are implemented. We have found that a unified system for estimating and visualizing covariate-adjusted spatial effects on outcomes arising from the most commonly encountered epidemiologic study designs greatly facilitates efficient and

reproducible analyses in these settings.

This section includes an introduction and illustration of the `MapGAM` package. The remainder of the section is organized as follows: Section 3.2 considers estimating spatial effects on right-censored survival times via a Cox proportional hazards additive model. The estimation procedures implemented in `MapGAM` are provided and a brief simulation study considers the performance of the proposed fitting methods in various settings is discussed in Section 3.3. Section 3.4 provides an overview of the methodology implemented in `MapGAM` for estimating and visualizing spatial effects in the context of generalized additive models. Two illustrative examples using `MapGAM` to analyze hypothetical case-control data from the state of Massachusetts and censored survival data from California State are provided in Section 3.5. Section 3.6 concludes with discussion of the utility of the `MapGAM` package and considers possible extensions of the package in future research.

3.2 Cox Proportional Hazards Addtiive Models

In this section we briefly introduce the methodology implemented in MapGAM as an extension of the GAM methods previously discussed in Section 2.4. We consider modeling censored survival data that are distributed on a map with u and v denoting the geographical parameters. Let t denote the survival time and X be a $d \times 1$ vector of d adjustment covariates. The Cox proportional hazards additive model to analyze the spatial effect incorporates a bivariate smoother into the Cox proportional hazards model (Kelsall and Diggle, 1998) as

$$\lambda(t) = \lambda(0) \exp\{\eta\} \tag{3.1}$$

with

$$\eta = X^T\beta + f(u, v),$$

where $\lambda(t)$ represents the hazard at time t and $\lambda_0(t)$ the baseline hazard. β is a $d \times 1$ vector of regression coefficients associated with covariates X, reflecting the first trend relationship between the hazard and the corresponding covariates. $f(u, v)$ represents the spatial effect on the hazard, which is a nonlinear function of location parameters. We separate the spatial effect $f(u, v)$ into two parts (Details see Section 2.14):

$$f(u, v) = \beta_u u + \beta_v v + s(u, v),$$

where the parametric part $\beta_u u + \beta_v v$ will be fitted jointly with other adjusted variables using least squares. The nonparametric part $s(u, v)$ will be fitted using a smoother. To ensure identifiability, we constrain the model so that the the summation of the nonparametric part over all observations is 0. Let $\tilde{X} = [X^T, u, v]^T$, and $\tilde{\beta} = [\beta^T, \beta_u, \beta_v]^T$ be the corresponding coefficients, then the model becomes

$$\eta = \tilde{X}^T\tilde{\beta} + s(u, v). \tag{3.2}$$

Let t_i, δ_i, u_i, v_i, X_i denote the observed time, censoring status, geographical parameters, covariates for subject i. With $\eta_i = \tilde{X}_i^T\tilde{\beta} + s(u_i, v_i)$, the partial likelihood is

$$PL = \prod_{i \in D} \frac{e^{\eta_i}}{\sum_{j \in R_i} e^{\eta_i}},$$

where D is the indices of observed events. $R_i = \{j : t_j \geq t_i\}$ denotes the risk set just prior to time t_i. In the event of tied failur times, `MapGAM` package uses Efron approxiamtion for

the partial likelihood:

$$PL_E = \prod_{i \in D} \frac{\prod_{j \in F_i} e^{\eta_j}}{\prod_{j=1}^{d_i} \left[\sum_{l \in R_i} e^{\eta_l} - \sum_{l \in F_i} e^{\eta_l} (j-1)/d_i \right]},$$

where F_i is the set of failures at time t_i, and d_i is the number of elements in F_i.

let l denote the log partial likelihood function for the data. To estimate the parameters of the model we seek to maximize the expected local log likelihood:

$$\hat{\eta}_i = \max_{\eta_i} E(l(\eta(\tilde{X}_i, s_i), t_i, \delta_i)),$$

Under standard regularity conditions (namely the ability to interchange integration and differentiation), we have

$$E[dl/d\eta_i]_{\hat{\eta}_i} = 0, \tag{3.3}$$

While there is no general closed form for solution to (3.3), a first-order Taylor series expansion leads to an iterative estimating procedure given by

$$\eta_i^{new} = \eta_i^{old} - E[dl/d\eta]_{\eta_i^{old}} / E[d^2 l/d\eta^2]_{\eta_i^{old}},$$

which is equivalent to

$$\eta_i^{new} = E\left[\eta - \frac{dl/d\eta}{E[d^2 l/d\eta^2]} \right]_{\eta_i^{old}}. \tag{3.4}$$

We can compute the first and second derivatives of the log partial likelihood as

$$\frac{dl}{d\eta_i} = \delta_i - \sum_{j \in C_i} \frac{e^{\eta_i}}{\sum_{k \in R_j} e^{\eta_k}}, \tag{3.5}$$

and

$$\frac{d^2l}{d\eta_i^2} = -\sum_{j\in C_i} \frac{e^{\eta_i}}{\sum_{k\in R_j} e^{\eta_k}} + \sum_{j\in C_i} \frac{e^{2\eta_i}}{\left(\sum_{k\in R_j} e^{\eta_k}\right)^2}, \tag{3.6}$$

where $C_i = j : i \in R_j$ is the sets of subjects whose risk sets contain i. The Cox model is a semi-parametric model without any specification for the distribution of the survival times, so it is not possible to calculate a close form for the expectation of the second derivatives of the log partial likelihood as required in (3.4). So before updating η, a GAM model can be fitted using the second derivatives as responses to estimate the expectation of the second derivatives of log partial likelihood.

To this end, by the equation(3.4), with an estimate η^{old}, the new estimate for η can be obtained using the following two steps:

1. Estimate $E[d^2l/d\eta^2]$ by fitting a generalized additive model using $d^2l/d\eta^2$ as responses, including the linear predictor of X and a bivariate smoother of geolocation parameters.

2. Estimate η^{new} using the backfitting algorithm described in Section 2.4.2 with $-1/\hat{E}[d^2l/d\eta^2]$ as weights and $\eta^{old} - [dl/d\eta]_{\eta^{old}}/\hat{E}[d^2l/d\eta^2]$ as responses.

3.3 Simulation Study

In this section we assess the performance of our proposed method for fitting the Cox proportional hazards additive model using two simulation studies. In both simulation settings, two spatial parameters (u, v) and adjustment covariate x are generated from a uniform distribution with range from -1 to 1. Survival times were then simulated from an exponential distribution with a hazard function. The first simulation example assumes a linear effect of all covariates on the log-hazard and that the effect of adjustment covariate x does not

interact with the effect of the spatial parameters u and v.

$$\lambda = 0.03 \exp\left\{\log(0.7)x + \log(1.2)u + \log(1.5)v\right\}. \tag{3.7}$$

In the second simulation example, the spatial parameters have a nonlinear effect on the log-hazard, while the adjustment covariate x has a linear effect that does not interact with the spatial coordinates. The hazard function used in the second simulation example is

$$\lambda = 0.03 \exp\left\{\log(0.7)x + \log(1.2)u + \log(1.5)v + \log(0.8)u^2 + \log(1.8)uv\right\}. \tag{3.8}$$

The true data-generating heatmaps of the two examples are shown in Figures 3.1a and 3.1c, respectively. When we set a seed of 269, with $N = 5000$ sampled data points. The survival times under the first (second) simulation setting range from 0.0011(0.0011) to 316.5(396.8), and have a median of 22.66(24.16). In both settings, censoring times were randomly sampled from a $Uniform(0, 70)$ distribution and observed times were taken to be the minimum of the true failure time and censoring time for each observation, yielding approximately 41.6% and 43.9% censoring in scenario 1 and 2, respectively. Code for this simulation is provided in the Appendix. Cox proportional hazards additive models were fit and the spatial effect of the points on an equally-spaced grid (201×201) extended across $u \in [-1, 1]$ and $v \in [-1, 1]$ were predicted using the `modgam` function from the `MapGAM` package. Smoothing span sizes of 0.4 and 0.2 were utilized for scenario 1 and 2, respectively. In each case, these values roughly correspond to the automated span size chosen when optimizing AIC.

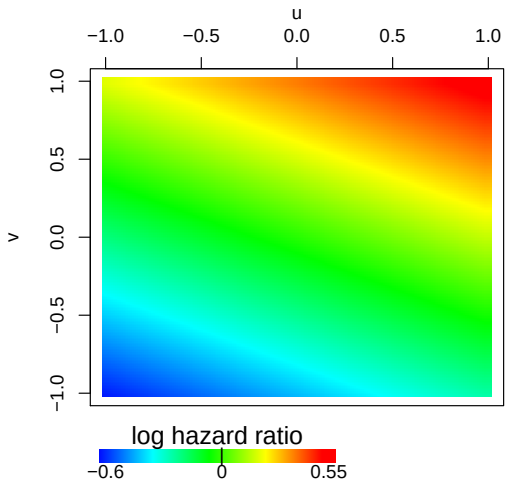
Figure 3.1b and 3.1d display the estimated spatial effects for example data sets using the first (linear relationship) and second (nonlinear relationship) simulation settings, respectively. Comparing the estimated values in Figures 3.1b and 3.1d to the corresponding true data generating values displayed in Figures 3.1a and 3.1d, we can see that the additive proportional hazards model implemented in `MapGAM` accurately recreates the true spatial effects (either

linear or nonlinear) giving rise to the data. In addition, two scatterplots of the estimated versus true spatial effect are provided in Figure 3.2a and 3.2b, again illustrating that the additive proportional hazards method outlined above is able to correctly identify the spatial effects present in the data with minimal bias.

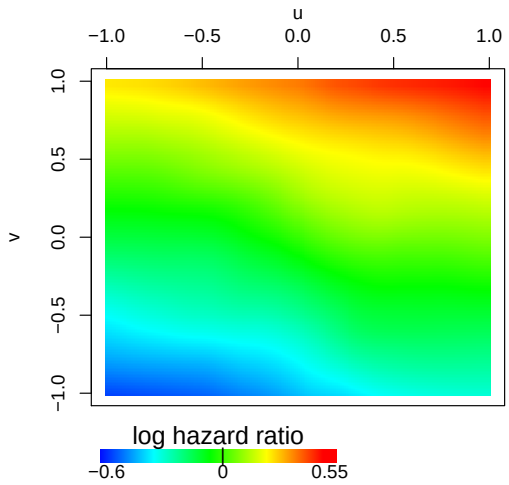
Scatterplots comparing the empirical standard errors and the estimated standard errors for the estimated spatial effects using the first (linear relationship) and second (nonlinear relationship) simulation settings are displayed in Figure 3.3a and 3.3b, respectively. The estimated standard errors are calculated based on several crude approximations, so compared to the empirical standard errors, the estimates are not consistent for all grid points. To make more reliable inference to identify significant clusters, permutation test is suggested to use although computational time cosuming.

3.4 `MapGAM` Package

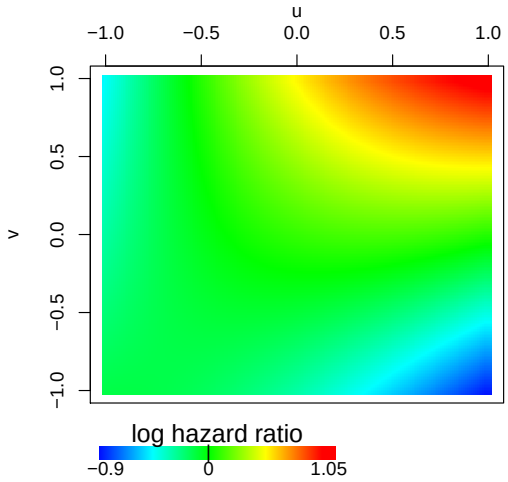
In the `MapGAM` package, typical spatial applications will start with the `predgrid()` function to create a regular grid of points within the study area, potentially restricted to points within optional map boundaries (e.g., a country, state, or regional map obtained from the `maps` package or imported from a shapefile). Crude or covariate-adjusted odds ratios, hazard ratios, or other effect estimates are then obtained for each grid point using the `modgam()` function to smooth by geolocation. `modgam()` provides compatible and flexible interfaces. Specifically, the model can be specified via a formula statement, or for users less familiar with writing model formulas in R, the formula can be omitted in which case the model is specified implicitly by structuring the data so that the first column of the data represents the outcome to be modeled (or the first two columns for survival objects), the next two columns represent the parameters for geolocation, and the remaining columns represent the adjustment covariates to be included in the model. For the univariate outcome, with the model



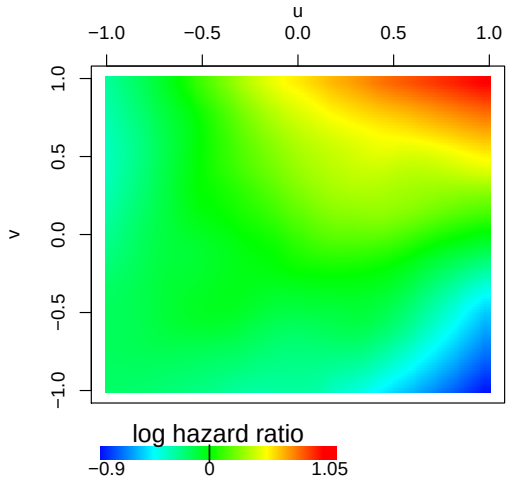
(a) Truths for linear spatial effect



(b) Estimates for linear spatial effect

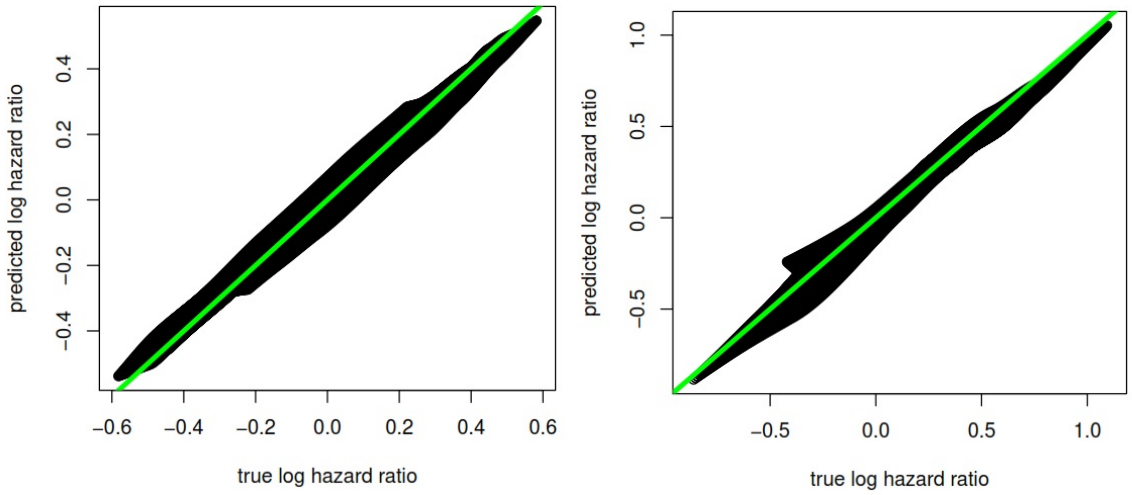


(c) Truths for nonlinear spatial effect



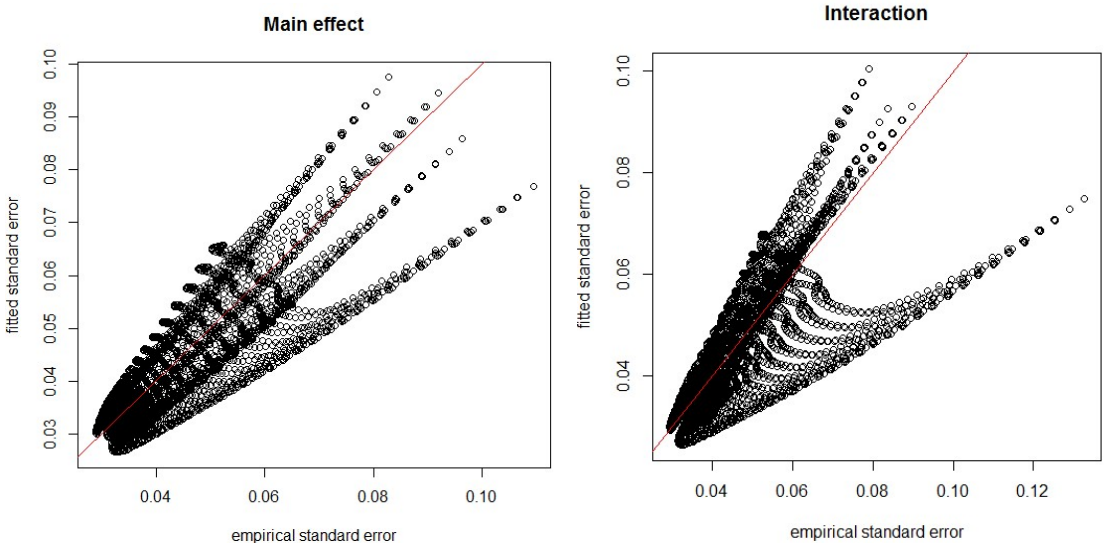
(d) Estimates for nonlinear spatial effect

Figure 3.1: Heatmaps of the the log-hazard ratio comparing the hazard of the location to the median hazard for two simulation examples with 5000 simulated observations. For the first simulation example with linear spatial effect on log-hazards: (a) estimated log-hazard ratio; (b) true log-hazard ratio; For the second simulation example with nonlinear spatial effect on log-hazards: (c) true log-hazard ratio; (d) estimated log-hazard ratio.



(a) Linear spatial effect (b) Nonlinear spatial effect

Figure 3.2: Comparisons of the true log-hazard ratio and the estimated log-hazard ratio for two simulation examples with 5000 simulated observations: (a) result for the first simulation example with linear spatial effect; (b) result for the second simulation example with nonlinear spatial effect.



(a) Linear spatial effect (b) Nonlinear spatial effect

Figure 3.3: Comparisons of the empirical standard errors based on 10000 simulations and the esimated standard errors of the spatial effects for two simulation examples with 5000 simulated observations: (a) result for the first simulation example with linear spatial effect; (b) result for the second simulation example with nonlinear spatial effect.

specified, `modgam()` proceeds by calling the `gam()` function in the `gam` package to estimate model parameters, then calls `mypredict.gam` to generate predictions for the specified grid. For the censored survival data, the Cox proportional hazards additive model is implemented by `gamcox()` function, and `predict.gamcox()` function can make predictions for newdata based on the model fitted by `gamcox()` function. `modgam` deals with the censored survival data by calling the *gamcox*() function to fit the model, then calls the *predict.gamcox*() function to generate predictons for the specified grid. A locally weighted scatterplot smoother (LOESS) (Cleveland, 1979, 1981; Cleveland and Devlin, 1988) is utilized as the bivariate smoothing function for the two geolocation parameters u and v in the `MapGAM` package. The smoothing parameter defining the neighborhood used to select the K nearest observations points for smoothing may be user specified or automatically chosen by minimizing AIC (Webster et al., 2006). The `optspan()` function can be used to find an optimal span size (proportion of data size included in the neighborhood) for the LOESS smoother. Optionally, the `modgam()` function can call `optspan()` to choose the optimal span for fitting the model in an automated fashion.

Considering the estimated spatial effect $f(u_i, v_i)$ for the i^{th} location, researchers are often interested in the spatial effect difference (or ratio, log-ratio) comparing each location to a defined reference. To obtain spatial effect estimates, one can specify `type="spatial"`, then `modgam()` provides three options for the choice of reference: the median of $f(u_i, v_i)$, $i = 1, \ldots, n$, the mean of $f(u_i, v_i)$, $i = 1, \ldots, n$, or an estimated spatial effect value at a user-specified geolocation. Alternatively, specifying `reference="none"` will produce prediction estimates based upon the linear predictor for each covariate combination in the prediction dataset (including the model intercept). To produce estimates of effects for all adjustment covariates, the option `type="all"` may be specified. The result of `modgam()` is an object of class `modgam()` that can be summarized by class-defined printing and plotting methods. Specifically, a heatmap of the predicted values from a fitted model can be generated using either the `colormap()` or `plot()` functions. For tailored plots, the `trimdata()` and

`sampcont()` functions can be used to restrict data to those areas within a specified set of map boundaries and to conduct simple or spatiotemporal stratified sampling from eligible controls–a useful feature for analysis of data from large cohorts. A function structure plot of the package is illustrated in Figure 3.4.

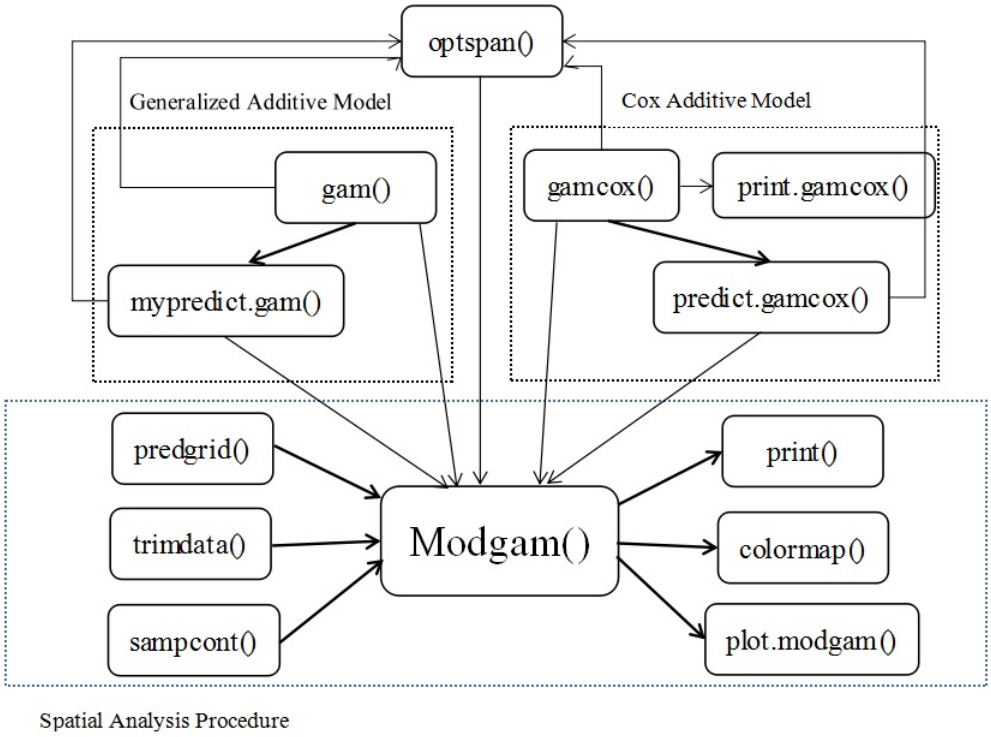


Figure 3.4: The function structure of the `MapGAM` package. the thicker arrows indicate the procedure flow, and the thinner arrows mean that the funtion the arrow points to calls the function at the arrow's origin. `trimdata()` and `sampcont` functions can be used to preprocess the observed data. `predgrid()` function can be used to generate a regular grid based on the map. Then the data and grid are the input for `modgam()` function to estimate the spatial effect. For GAM model, `modgam()` calls `gam()` and `mypredict.gam()`; For Cox additive model, `gamcox()` and `predict.gamcox()` are called by `modgam()`. `optspan()` function can be optionally called by `modgam()` to search for an optimal span size for the spatial smooth term. Finally, the result of `modgam()` can be summarized by `print.gamcox()` and visualized by `colormap()` and `plot.modgam()`.

Besides fitting the model, `MapGAM` provides pointwise standard errors as well as confidence intervals, which are returned by the `modgam()` function by setting `se.fit=TRUE`. The estimated pointwise standard errors for spatial effects are derived from the sum of two variance curves: one from the parametric part, $\beta_u u_i + \beta_v v_i$, and the other from the non parametric

part, s_i (Chambers and Hastie, 1992). The variance curve for the parametric part of the function reflects the joint covariance behavior, and the variance for the nonparametric part reflects the marginal information.

`modgam()` conducts a global test for spatial effects via a likelihood ratio test by comparing the deviance between a full model (including the spatial smoother) and a reduced model (omitting the spatial smoother). For the full model, the degrees of freedom of the non-parametric part is computed as $tr(S)-1$, where S is the smoothing matrix, and the degrees of freedom of the parametric part is $N-d-3$ (or $N-d-2$ for Cox proportional hazards additive model). Thus, the degrees of freedom of the full model is $N-tr(S)-d-2$ ($N-tr(S)-d-1$), and the degrees of freedom for the likelihood ratio test statistic is $tr(S)+1$. The function `modgam()` returns the p-value for the likelihood ratio test automatically. `modgam()` also performs permutation test (Kelsall and Diggle, 1998; Webster et al., 2006) to test the global spatial effect and poitwise significance. The function returns the results of the permutation test by specifying `permute= N.perm`, where `N.permt` is the permutation times.

For visualizing inference for spatial effects, the `plot` function will plot all point estimates along with the associated lower and higher band of confidence intervals provided that `se.fit=TRUE` is specified in the original `modgam()` call. By setting `contours = "intervals"`, areas with confidence intervals excluding 0 (on the log estimated effect scale) will be indicated on the map by plotting the contours of an indicator vector created to indicate whether 0 is below, between or above the confidence intervals at the grid points. By setting `contours = "permrank"`, contours will be added to indicate significant areas that had a pointwise permutation based p value less than a specified threshold (default of .05).

3.5 Application Example

3.5.1 Preterm Risk Study

In this section we present an illustrative example using `MapGAM` to analyze hypothetical case-control data from Massachusetts. `MAdata` is a simulated case-control study dataset in the package. There are 90 cases and 910 controls with random geolocations within Massachusetts, geocoded on a Lambert projection (in meters). `MAmap` is a map of Massachusetts using the same projection. There are also other three covariates in the dataset: smoking, mercury exposure and selenium exposure. All the covariates values are randomly generated. There are 1000 simulated observations in total. Summary of the dataset is as following:

```
R>data(MAdata)
R>data(MAmap)
R>summary(MAdata)
```

```
      Case           Xcoord           Ycoord          Smoking
 Min.   :0.00   Min.   : 35354   Min.   :778430   Min.   :0.000
 1st Qu.:0.00   1st Qu.:111465   1st Qu.:869089   1st Qu.:0.000
 Median :0.00   Median :183100   Median :891067   Median :0.000
 Mean   :0.09   Mean   :175054   Mean   :889081   Mean   :0.177
 3rd Qu.:0.00   3rd Qu.:236826   3rd Qu.:919684   3rd Qu.:0.000
 Max.   :1.00   Max.   :327861   Max.   :954253   Max.   :1.000
    Mercury          Selenium
 Min.   :0.1418   Min.   :0.2049
 1st Qu.:0.7206   1st Qu.:0.8573
 Median :1.0010   Median :1.1836
 Mean   :1.1471   Mean   :1.3590
 3rd Qu.:1.4017   3rd Qu.:1.6844
 Max.   :5.6298   Max.   :5.8963
```

The geolocations of the observations are shown in Figure 3.5, which can be generated using the following code:

```
R> # map participants, cases in red and controls in black
R> plot(MAmap)
```

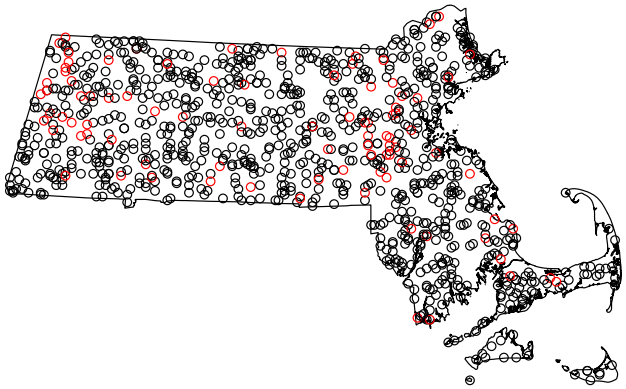


Figure 3.5: MAdata geolocations. Control: black; Case: red.

```
R> points(MAdata.Xcoord,MAdata.Ycoord,col=MAdata.Case+1)
```

We first start with generating a prediction grid for the map using `predgrid` by

```
R>gamgrid <- predgrid(MAdata, map=MAmap)    # requires PBSmapping package
```

After defining a prediction grid, `modgam()` is used to fit a GAM model based on the `MAdata` and generate predictions on the defined grid. A formula expression indicates that the indicator `Case` is specified as the response, and two spatial parameters `Xcoord` and `Ycoord` are included in `lo()` to specify a geospatial smoothing term. In addition, potential confounders `Smoking`, `Mercury` and `Selenium` are also adjusted for in the model as linear terms. Argument `sp` is used to specify the span size for the spatial smoothing term. A specification of `sp = null` (the default) implies that an optimal span will be selected. Note that if the model formula is not supplied, the data must be structuring so that the outcome is in the first

column, the two spatial parameters are in the second and third columns and the adjustment variables are in other columns. In that case, specifying `m="adjusted"` will include all other columns of the data as linear terms in the model and `m="crude"` will fit only the two spatial parameters (the `m` argument is ignored if a model formula is supplied). For this particular example, the resulting call to `modgam()` using the formula statement is given as follows:

```
R> fit1 <- modgam(Case ~ lo(Xcoord, Ycoord) + Smoking + Mercury + Selenium,
+                 data=MAdata, rgrid=gamgrid, sp=NULL, verbose = FALSE)
R> # which is equivalent to:
R> # fit1 <- modgam(data=MAdata,rgrid=gamgrid, m="adjusted", sp=NULL,
R> #                  verbose=FALSE)
R> #
R> fit1

Call:
modgam(formula = Case ~ lo(Xcoord, Ycoord) + Smoking + Mercury +
    Selenium, data = MAdata, rgrid = gamgrid, sp = NULL, verbose = FALSE)

Model:
Case ~ lo(Xcoord, Ycoord, span = 0.3, degree = 1) + Smoking +
    Mercury + Selenium
Family: binomial Link: logit

Coefficients:
                                      (Intercept)
                                    -6.911648e+00
lo(Xcoord, Ycoord, span = 0.3, degree = 1)Xcoord
                                     2.363118e-06
lo(Xcoord, Ycoord, span = 0.3, degree = 1)Ycoord
                                     4.376156e-06
                                          Smoking
                                     1.533433e+00
                                          Mercury
                                     5.729589e-01
                                         Selenium
                                    -6.431932e-01

Degrees of Residual Freedom: 982.9497
Residual Deviance: 500.0045
AIC: 534.1051
p value for testing the global spatial effect: 9.428405e-05
```

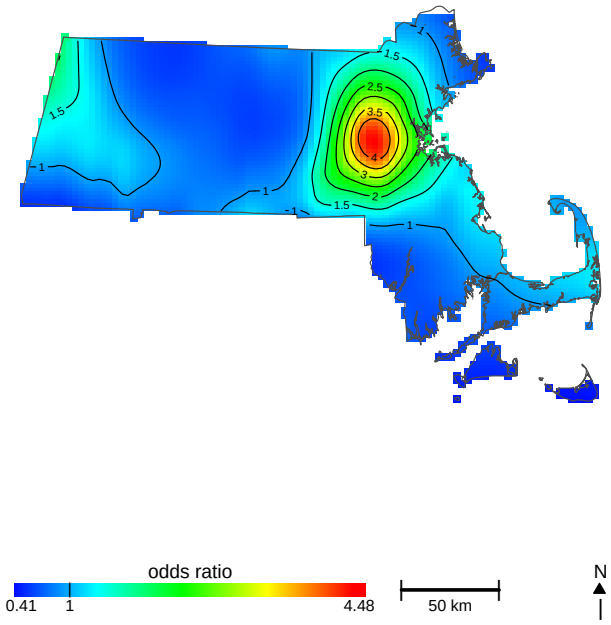


Figure 3.6: Heatmap of the odds ratio of spatial effect predictions compared to the median odds.

```
 Spatial effect predictions:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-0.87980 -0.26250  0.00000  0.08151  0.29030  1.49900
```

A heatmap of the estimated spatial effect predictions (representing the odds ratio comparing the odds at each location to the median odds across all locations) can be generated using the `modgam()` plotting routine via a call to the `plot()` function. This in turn relies upon the `colormap()` function defined within `MapGAM`. The resulting heatmap is displayed in Figure 3.6. The `exp` argument is used to specify whether the heatmap is drawn on the scale of the odds ratio (`exp=TRUE`) or the log odds ratio (`exp=FALSE`).

```
plot(fit1,MAmap,contours="response")
```

3.5.2 Ovarian Cancer Study

In this section we provide an example of applying the `MapGAM` package to estimate spatial effects on censored survival data using hypothetical survival times derived from the state of California. we use the `MapGAM` package to estimate and visualize spatial effects for a dataset simulated from information on censored survival times of California ovarian cancer patients. These are data contained in the object `CAdata` within the `MapGAM` package. The original source is the California advanced-stage invasive epithelial ovarian cancer patients reported to the California Cancer Registry from 1996 to 2006 (Bristow et al., 2014). After removing patients with age < 25 and > 80 for identifiability reasons, and adding random noise to the geolocation parameters, `CAdata` is a random draw of size N=5,000 from the original dataset. Observed times and failure status were simulated based upon the observed distribution found in the original dataset. Potential covariates available in the dataset include age and insurance type (6 categories in total: Managed Care, Medicare, Medicaid, Other Insurance, Not Insured and Unknown) A summary of `CAdata` is as follows.

```
R>data(CAdata)
R> summary(CAdata)
```

```
      time               event              X                Y
 Min.   : 0.004068   Min.   :0.0000   Min.   :1811375   Min.   :-241999
 1st Qu.: 1.931247   1st Qu.:0.0000   1st Qu.:2018363   1st Qu.: -94700
 Median : 4.749980   Median :1.0000   Median :2325084   Median : -60387
 Mean   : 6.496130   Mean   :0.6062   Mean   :2230219   Mean   :  87591
 3rd Qu.: 9.609031   3rd Qu.:1.0000   3rd Qu.:2380230   3rd Qu.: 318280
 Max.   :24.997764   Max.   :1.0000   Max.   :2705633   Max.   : 770658
      AGE         INS
 Min.   :25.00   Mcd: 431
 1st Qu.:53.00   Mcr:1419
 Median :62.00   Mng:2304
 Mean   :61.28   Oth: 526
 3rd Qu.:71.00   Uni: 168
 Max.   :80.00   Unk: 152
```

`CAmap` is the map file for California State. The geolocations of the observations are plotted in Figure 3.7.

```
R> data(CAmap)
R> plot(CAmap)
R> points(CAdata$X,CAdata$Y)
```

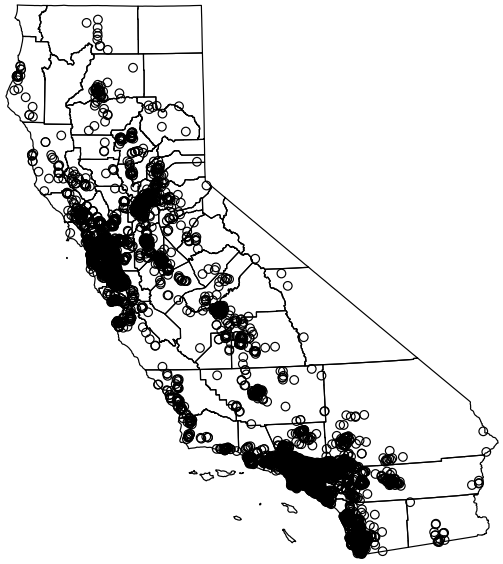


Figure 3.7: Geolocations of the observations in CAdata.

Below we generate the object `CAgrid` for the state of California using the `predgrid()` function and estimate spatial effects on the relative risk of death from a Cox proportional hazards additive model using the `modgam()` function. Finally, a heatmap of the log-hazard ratio comparing the log-hazard of each location to the median log-hazard is plotted using the plotting routines defined for `modgam()` objects via `plot()`. The resulting heatmap is displayed in Figure 3.8.

```
R> CAgrid = predgrid(CAdata[,c("X","Y")],map=CAmap,nrow=186,ncol=179)
```

```
R> fit2 <- modgam(Surv(time,event)~AGE+factor(INS)+lo(X,Y),data=CAdata,
+                rgrid=CAgrid, sp=0.3, verbose =FALSE)
R> # which is equivalent to
R> # fit2 <- modgam(data=CAdata,rgrid = CAgrid,family="survival",sp = 0.3,
R> # verbose = FALSE)
R> plot(fit2,CAmap,exp=T,border.gray=0.5)
R> fit2

Call:
modgam(formula = Surv(time, event) ~ AGE + factor(INS) + lo(X,
    Y), data = CAdata, rgrid = CAgrid, sp = 0.3, verbose = FALSE)

Model:
Surv(time, event) ~ lo(X, Y) + AGE + factor(INS)
span:  0.3

Coefficients:
           AGE factor(INS)Mcr factor(INS)Mng factor(INS)Oth factor(INS)Uni
    0.02657848     0.03657777     0.05251440     0.16770033     0.26790051
factor(INS)Unk
    0.07594159

Degrees of Residual Freedom: 4974.383
Residual Deviance: 46161.06
AIC: 46212.29
p value for testing the global spatial effect: <1e-5

 Spatial effect predictions:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-1.14900 -0.23630  0.00000  0.06566  0.35150  1.27200
```

Spatial inference can also be visualized. Setting `se.fit=TRUE`, `modgam()` returns pointwise standard errors and confidence intervals. The resulting confidence intervals are then plotted via the `plot` function, and are shown in Figure 3.9

```
R> fit3 <- modgam(Surv(time,event)~AGE+factor(INS)+lo(X,Y),data=CAdata,
+         rgrid=CAgrid, sp=0.3, verbose =FALSE, se.fit=TRUE)
R> plot(fit3,CAmap,exp=T,mapmin=0.2,mapmax=5,border.gray=0.7,
+      contours="interval")
R> fit3

Call:
```

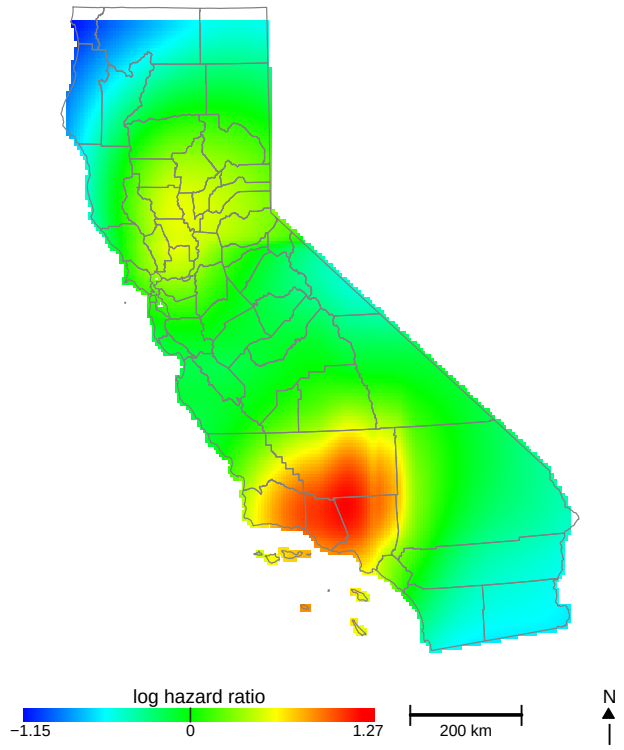


Figure 3.8: Heatmap of the log hazard ratio compared to the median log hazard.

```
modgam(formula = Surv(time, event) ~ AGE + factor(INS) + lo(X,
    Y), data = CAdata, rgrid = CAgrid, sp = 0.3, se.fit = TRUE,
    verbose = FALSE)

Model:
Surv(time, event) ~ lo(X, Y) + AGE + factor(INS)
span:  0.3

Coefficients:
           AGE factor(INS)Mcr factor(INS)Mng factor(INS)Oth factor(INS)Uni
    0.02657848     0.03657777     0.05251440     0.16770033     0.26790051
factor(INS)Unk
    0.07594159

Degrees of Residual Freedom: 4974.383
Residual Deviance: 46161.06
AIC: 46212.29
p value for testing the global spatial effect: <1e-5

 Spatial effect predictions:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
```

```
-1.14900 -0.23630  0.00000  0.06566  0.35150  1.27200

Spatial effect 95 % lower interval:
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
-1.6100 -0.6257 -0.2189 -0.2173  0.1520  1.0400

Spatial effect 95 % higher interval:
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
-0.68810  0.07192  0.29650  0.34860  0.61050  1.73600
```

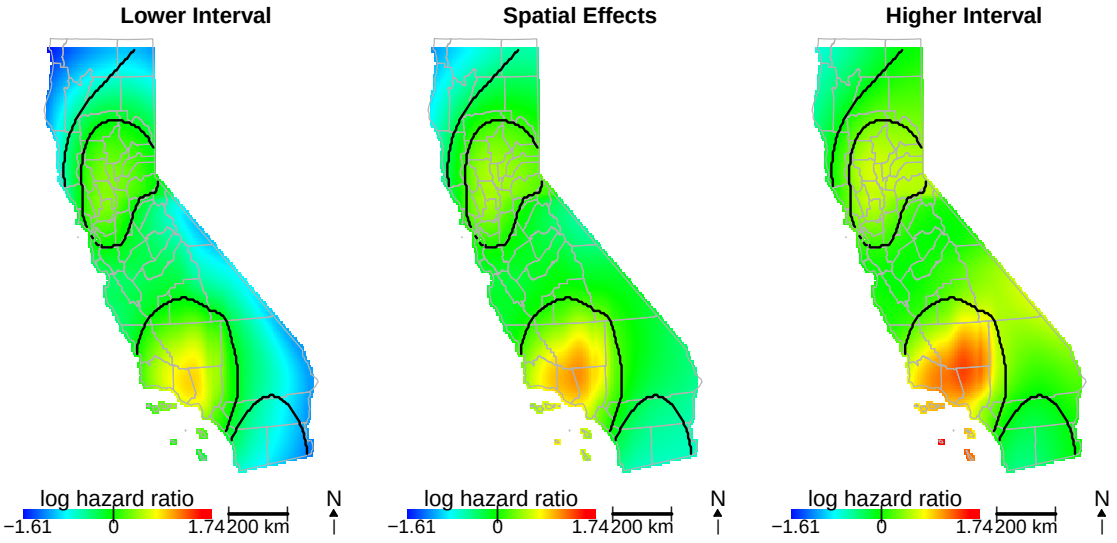


Figure 3.9: Heatmap of the hazard ratio as well as confidence intervals compared to the median hazard with significant areas circled which were identified by confidence intervals.

3.6 Discussion

Bivariate LOESS smoothing with standard error estimation is computationally intensive, especially in the context of GAMs and proportional hazards models. For example, with 5000 observations, a span size of 0.2, and a binomial outcome `modgam()` took about 1 second to provide estimates without standard errors but about 50 seconds with standard error estimates (`se.fit=TRUE`) on recent personal computers. For the same span size and number of observations but with a proportional hazards model, `modgam()` took about 40 seconds without standard errors and about 70 seconds with standard errors. Although slower than

we might like, the run times for `se.fit=TRUE` are much faster than the pointwise permutation test we previously employed which required a 1000-fold increase in run times (Webster et al., 2006).

Estimating and mapping spatial distributions of disease risk is extremely useful for identifying health disparities, and mapping risk surfaces that are adjusted for individual-level confounding variables is of great interest to epidemiologists. By developing and actively maintaining a convenient R package, `MapGAM`, we intend to facilitate mapping crude and covariate-adjusted spatial effects for the most common probability models used to characterize the relationship of disease risk to spatial location and other factors. In the future we hope to improve the flexibility of the package by expanding the incorporated smoothing methods, including the addition of basis expansion and tensor product methods, allowing for smoothing over more than two dimensions, and expanding the `sampcont()` function to include additional sampling methods such as matching. Further research on the development and implementation of adaptive smoothing methods that allow for the amount of smoothing to vary depending on the local extent of a spatial effect is currently in progress, and may be added to the package in a future update.

Chapter 4

Spatial Analysis for Disease Outcomes via Adaptive smoothing

4.1 Introduction

In spatial epidemiology studies with point-based data, spatial (bivariate) smoothing methods are often employed. Three commonly used spatial smoothing methods are local-averaging smoothers, smoothing splines, and thin-plate splines, as discussed in Chapter 2. As previously described, local-averaging smoothers use the weighted average of responses within the neighborhood as an estimate, and a variety of weighting schemes have been proposed in the literature. Kernel smoothers (Nadaraya, 1964; Priestley and Chao, 1972) use kernel functions to define weights, and the radius of the kernel functions determines the size of smoothing neighborhood and hence the amount of smoothing. Locally-weighted scatterplot smoothers (LOESS) (Cleveland, 1979, 1981; Clark, 1977; Tibshirani and Hastie, 1987) calculate the weights using a weighted (tricube weighting function) local linear or polynomial regression with data points residing in the neighborhood of the observation to be predicted. Alternatively, smoothing splines fit a smooth curve to observed responses using spline functions that are estimated by minimizing a penalized likelihood function. Tensor-product smoothing splines (De Boor, 1978; Wood, 2006) construct the spline functions using two sets of knots

on two spatial parameters, where the spline functions can again be estimated by minimizing a penalized likelihood. As an alternative to smoothing splines, thin-plate splines (Duchon, 1977) have gained popularity for spatial analysis because there is no need to specify knots for the form of spline functions (Cressie, 1993; Hutchinson, 1995).

Another prominent approach for geographical smoothing is Kriging (Stein, 1999), which models the values of a function at different points via a Gaussian process governed by prior covariances, along with weights calculated based on the covariances. Kriging can also be justified from a Bayesian perspective (Williams, 1998). This justification starts with a Gaussian process as the prior distribution over functions and the covariance between any two points is the covariance function (or kernel) of the Gaussian process evaluated at the spatial location of the two points. Predicted values are then obtained by combining the Gaussian prior with a Gaussian likelihood function computed using the observed values.

Despite the fairly common use of kernel-based methods and smoothing splines, each of these smoothing methods uses a global smoothing parameter, and hence the amount of smoothing is held constant across all locations. The result is that some regions may under-smoothed, while others are over-smoothed. To address this deficiency, a series of spatially adaptive penalized spline models have been proposed by Ruppert and Carroll (2000); Lang et al. (2002); Baladandayuthapani and Carroll (2005); Crainiceanu and Goodner (2007); Krivobokova and Kauermann (2008). A common feature of all of these methods is the introduction of spatial adaptivity by imposing a functional structure on the smoothing parameters under the context of an ordinary penalized spline again requiring knot specification along with the selection of a penalization parameter.

To improve spatial adaptivity and avoid not specification, Bayesian models have recently been proposed that rely on a class of adaptive Gaussian Markov Random Fields (GMRFs) with stochastic interaction weights in a space-varying coefficient model. Specifically, Brezger and Hennerfeind (2007) extended Intrinsic GMRFs (IGMRFs), introduced in Chapter 2,

by specifying independent gamma priors on the unknown weights. However, this approach failed to provide a spatially-varying covariance as needed when estimating functions with high spatial variability (Lang et al., 2002). ?Brezger and Hennerfeind (2007) later proposed an IGMRF model for the logarithms of the weights in a second hierarchy. More recently, Yue and L. (2010) constructed a particular IGMRF prior with a full-rank factorization property in the first hierarchy then specified another simpler GMRF prior for the adaptive variance function. Their approach results in a two-dimensional extension of the one-dimensional adaptive modeling procedure proposed in the work of Lang et al. (2002).

From the perspective of spatial disease mapping, two drawbacks of the above Bayesian models for adaptive spatial analysis are that (1) they assume a normally distributed outcome, and (2) they do not consider adjustment for additional covariates beyond location. In epidemiologic studies, discrete outcomes such such as disease incidence and count data are common outcomes of interest. Further, adjustment of potential confounders is nearly always necessary when estimating adjusted spatial effects in order to isolate the spatial pathway of interest. In this chapter, we build on the previous work of Yue et al. (2010) and propose a hierarchical Bayesian adaptive thin-plate spline for generalized additive models (GAMs) that allows for spatial smoothing of continuous, binary and count outcomes. Moreover, our proposed method allows for the amount of smoothing to flexibly vary depending on the local extent of spatial effect by using nonstationary spatial Gaussian Markov random fields while accommodating adjustment for potential confounders.

The remainder of this chapter is organized as follows: Nonstationary GMRF priors corresponding to adaptive thin-plate splines for continuous and discrete outcomes with additional covariate adjustment are introduced in Section 4.2. This section also introduces the relevant MCMC techniques required for fitting the proposed model. Performance of the proposed method is investigated by simulation in Section 4.3. An application to the previously introduced (see Chapter 1) epidemiologic study investigating spatial heterogeneity in the risk

of preterm birth among Massachusetts residents is given in Section 4.4. In Section 4.5, we conclude with a discussion of the proposed method and possibilities for future work.

4.2 Methods

Assume there are N observations, and let (u_i, v_i) represent the longitude and latitude of the i^{th} observation's location in a map, $i = 1, \ldots, N$. Further consider the following spatial model for a univariate disease outcome assuming that the outcome follows a distribution that belongs to the exponential family:

$$g(E[Y_i]) = X_i^T \beta + f(u_i, v_i), \tag{4.1}$$

where $g(\dot{)}$ is the link function, $X_i \in \mathbb{R}^p$ is a vector of covariates to be controlled, and $f(\dot{)}$ is an unknown bivariate function representing the spatial effect on the outcome Y_i. Let $L(Y_i, X_i, u_i, v_i)$ represent the i^{th} observation's contribution to the log-likelihood function. Then the thin-plate spline estimator of f is the solution to

$$\hat{f} = \min_f \left\{ -\sum_i L(Y_i, X_i, u_i, v_i)) + \lambda J_2(f) \right\}, \tag{4.2}$$

where $J_2(f)$ is the roughness penalty for f on $\mathbb{R}^2$ and is given by

$$\begin{aligned} J_2(f) &= \iint_{\mathbb{R}^2} \left[\left(\tfrac{\partial^2 f(u,v)}{\partial u^2} \right)^2 + \left(\tfrac{\partial^2 f(u,v)}{\partial u \partial v} \right)^2 + \left(\tfrac{\partial^2 f(u,v)}{\partial v^2} \right)^2 \right] du dv \\ &= \iint_{\mathbb{R}^2} \left[\left(\tfrac{\partial^2}{\partial u^2} + \tfrac{\partial^2}{\partial v^2} \right) f(u,v) \right] du dv. \end{aligned}$$

In the above model, the smoothing parameter λ controls the trade-off between the likelihood of the data and the smoothness of the function from the penalty term $J_2(f)$. As previously noted, a global specification for λ fails to adapt to variable smoothness across different locations. To allow for spatial adaptivity, Yue and L. (2010) proposed a Bayesian hierarchical model by allowing λ to spatially vary according to the local extent of the smooth function. More precisely, the method allows for adaptive smoothing of the spatial effect using a prior

based on a discretized thin-plate spline. However, the approach proposed by Yue and L. (2010) is assumes a continuous normally distributed outcome and a rectangle map without consideration of adjustment for potential confounding factors. In following sections, we extend the work of Yue and L. (2010) by proposing a Bayesian adaptive thin-plate smoothing splines based on the data model given in (4.1) that can be utilized on a map with arbitrary shape.

4.2.1 Thin-Plate Splines Prior

We begin by generating an evenly-spaced grid containing n grid points across the map with distance h between any two grid points. Denote the vectorized longitude and latitude grid points by $\tilde{u} = [\tilde{u}_1, \cdots, \tilde{u}_n]^T$ and $\tilde{v} = [\tilde{v}_1, \cdots, \tilde{v}_n]^T$, respectively. Define $\mathbf{z} = [z_1, \cdots, z_n]^T$ by

$$z_j = f(\tilde{u}_j, \tilde{v}_j),\ j = 1, \ldots, n,$$

and let $D = [d_{ij}]$ be the $N \times n$ incidence matrix with $d_{ij} = 1$ if the jth grid point is the nearest grid point to the i^{th} observation's location, and $d_{ij} = 0$ otherwise. From this, the data model can be vectorized as

$$\mathrm{g}(E[Y]) = X\beta + Dz.$$

If the distance, h, between any two grid points is small enough, the second partial derivative of f in Eq. (5.3) can be approximated by

$$\frac{\partial^2}{\partial u^2} f(\tilde{u}_j, \tilde{v}_j) \approx h^{-2} \nabla^2_{(1,0)} f(\tilde{u}_j, \tilde{v}_j)$$
$$\frac{\partial^2}{\partial u^2} f(\tilde{u}_j, \tilde{v}_j) \approx h^{-2} \nabla^2_{(0,1)} f(\tilde{u}_j, \tilde{v}_j),$$

where $\nabla^2_{(1,0)}$ and $\nabla^2_{(0,1)}$ denote the second order backward difference operators. Let z_{jL} reprensent the function value at the left of the point $(\tilde{u}_j, \tilde{v}_j)$, and z_{jR}, z_{jU}, z_{jD} represent function values at three other locations: Right, Upper, Lower, respectively. The the second

order backward difference operator is defined as

$$\nabla^2_{(1,0)} f(\tilde{u}_j, \tilde{v}_j) = z_{jR} - 2z_j + z_{jL}$$
$$\nabla^2_{(0,1)} f(\tilde{u}_j, \tilde{v}_j) = z_{jD} - 2z_j + z_{jU}$$
.

A resulting approximation of the roughness penalty, $J_2(f)$, is then given by

$$J_2(f) \approx \frac{1}{h^4} \sum_j^n [(z_{jR} + z_{jL} + z_{jD} + z_{jU} - 4z_j)]^2. \tag{4.3}$$

The approximation has a quadratic expression $z' A_0 z$ with each entry in A_0 defined by coefficients in equation (4.3). Specify a prior distribution for z of the form

$$z|\delta \propto \exp\left(-\frac{\delta}{2} z' A_0 z\right). \tag{4.4}$$

Then the random vector z in equation (4.4) is an improper GMRF and satisfies the Markov conditional independence assumption. Further, the posterior distribution of z is given by

$$z|Y, X, \beta, \delta \propto \exp\left(\sum_i L(Y_i, X_i, u_i, v_i) - \frac{\delta}{2} z' A_0 z\right),$$

which suggests that an estimate based on the posterior distribution of z can provide the solution to the optimization problem posed by (4.2). However, because the backward difference operator cannot be applied to boundary points, the null space of the matrix A_0 has dimension equal to the number of boundary points.

To fix the above rank-deficiency issue with A_0, (Yue and L., 2010) proposed adding boundary terms based on a rectangle map. As our interest lies in spatial mapping over arbitrary geographic boundaries, we generalize the boundary terms proposed by (Yue and L., 2010) incorporate maps of any shape. To this end, let I_{jL} denote the indicator that there exists a grid point at the left of the point ($\tilde{u}_j$,$\tilde{v}_j$), and let I_{jR}, I_{jU} and I_{jD} denote indicators at the other three positions: right, upper and lower respectively. The penalty term $J_2(f)$ can then

be approximated after correcting for the boundary terms as:

$$J_2(f) \approx \frac{1}{h^4} \sum_{j}^{n} [(z_{jR}I_{jR} + z_{jL}I_{jL} + z_{jD}I_{jD} + z_{jU}I_{jU}) - z_j(I_{jR} + I_{jL} + I_{jD} + I_{jU})]^2. \quad (4.5)$$

Now let $\tilde{A}$ denote the $n \times n$ structure matrix in the quadratic expression $z'\tilde{A}z$ for the approximation in (4.5). Then $\tilde{A}$ is semi-definite with rank $n-1$ and can be written as $\tilde{A} = \tilde{B}'\tilde{B}$, where $\tilde{B}$ is a $n \times n$ matrix such that

$$[\tilde{B}z]_j = (z_{jR}I_{jR} + z_{jL}I_{jL} + z_{jD}I_{jD} + z_{jU}I_{jU}) - z_j(I_{jR} + I_{jL} + I_{jD} + I_{jU}).$$

Thus, the vector $\tilde{B}z$ is the summation of the second order difference of spatial effect f in vertical and horizontal directions with rank $n-1$. If we define $B_{(n-1)\times n}$ by deleting the first row of $\tilde{B}$, then B is full rank with rank $n-1$. Then letting $A = B'B$, the prior on z is can be specified as

$$z|\delta \propto \exp\left(-\frac{\delta}{2}z'Az\right), \quad (4.6)$$

where the random vector z is an IGMRF of the first order and satisfies the Markov conditional independence assumption.

4.2.2 Spatially Adaptive Thin-Plate Spline Priors

The prior on z shown in equation (4.6) is equivalent to a multivariate normal prior on Bz, with $Bz \sim \mathcal{N}(\mathbf{0}, \delta^{-1}I)$, and utilizes a global δ. As such, smoothness of the model fit is not adaptive based on location. To introduce adaptivity, following Yue and L. (2010), we replace the global δ with locally-varying parameters δ_j, $j = 1, \ldots, n$. Then, a small value of δ_j represents less smoothing of the spatial effect z at grid point j, and vice versa.

To re-paramterize δ_j, set $\delta_j = \delta e^{\gamma_j}$, where δ is a location invariant scale parameter, and γ_j

serves as the adaptive precision for δ_j. The modified prior of z is of the form

$$z|\delta,\gamma \propto \exp\left(-\frac{\delta}{2}z'A_\gamma z\right)$$

where $\gamma = [\gamma_1, \cdots, \gamma_{n-1}]'$, and $A_\gamma = B'\text{diag}\{e^{\gamma_1}, e^{\gamma_2}, \cdots, e^{\gamma_{n-1}}\}B$. This is then equivalent to a Gaussian prior for Bz with

$$Bz \sim \mathcal{N}\left(0, \delta\begin{pmatrix} e^{-\gamma_1} & 0 & \cdots & 0 \\ 0 & e^{-\gamma_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \cdots \\ 0 & 0 & \cdots & e^{-\gamma_{n-1}} \end{pmatrix}\right), \tag{4.7}$$

which shows that the variances of the second order differences of the function at different locations are different indeed different. A smaller γ_j means the variance of the second order difference of the function at location $(\tilde{u}_{j+1}, \tilde{v}_{j+1})$ is larger, and hence the penalty on the smoothness is at grid point j is smaller, and vice versa.

An additional prior must be defined for vector γ. As the elements in γ are also spatially dependent, we specify the prior on γ to be a first order IGMRF on our grid of the form

$$\gamma|\delta,\eta \propto (\delta\eta)^{n-2} exp\left(-\frac{\delta\eta}{2}\gamma' M\gamma\right) I_{\mathbf{1}'\gamma=0},$$

where M is an $(n-1)\times(n-1)$ matrix with rank $n-2$, and entries are based on the quadratic expression

$$\gamma' M\gamma = \sum_j \left[\gamma_{jL}I_{jL} + \gamma_{jU}I_{jU} - \gamma_j(I_{jL} + I_{jU})\right]^2.$$

However, because M is not full rank, we employ the constraint $I_{\mathbf{1}'\gamma=0}$ in order to guarantee identifiability.

To this point we have specified an IGMRF prior on the spatial effect z in the first hierarchy which allows for smoothness to vary adaptively by introducing vector γ, and specifying a first-

order IGMRF prior on γ in the second hierarchy. Prior specification for β, the coefficients representing the association between the outcome and potential adjustment covariates is also required. To this end, we specify a diffuse Gaussian prior on β such that

$$\beta \sim \mathcal{N}(\mu_0, \Sigma_0), \quad \Sigma_0 = \text{diag}\{\tau_1, \cdots, \tau_p\}.$$

4.2.3 Hyperpriors

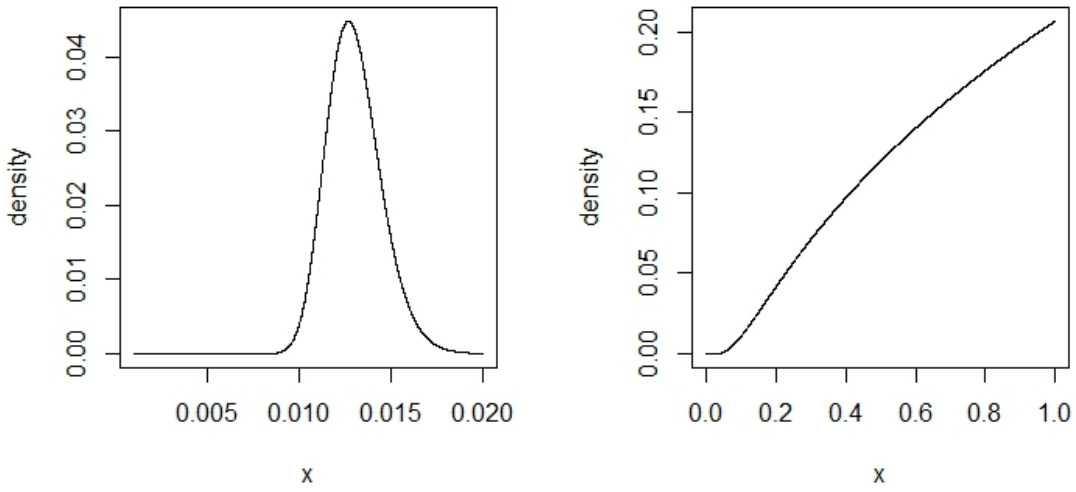
To complete the model, hyperpriors on the precision hyperparamters δ and η are required for a full Bayesian specification. As we discussed in Section 4.2.2, δ is a scale parameter which controls the degree of global smoothing taken on the whole field z. A smaller value of δ yields a less smooth function. The parameter η determines the degree of smoothing put on γ. A smaller value of η implies a more variable precision for γ and therefore more adaptivity applied on the smoothing of z. We set a Pareto prior for δ and an inverse gamma prior for η of the forms:

$$\begin{aligned} &\delta \propto \tfrac{c^d}{(c+\delta)^{d+1}} \\ &\eta \propto \eta^{-a-1} \exp(-b/\eta) \end{aligned} \tag{4.8}$$

The probability density functions for the inverse gamma priors and Pareto priors with different sets of parameters are shown in Figure 4.1 and Figure 4.2, respectively. Because the Pareto distribution can be written as a scale mixture of exponentials by introducing a latent variable θ, we modify the prior on δ by

$$\begin{aligned} &\delta|\theta \sim Exp(\theta) \\ &\theta \sim Gamma(d, c) \end{aligned} \tag{4.9}$$

Full Bayesian inference can be guaranteed using the hyperpriors following Yue et al (2012). The ranges of values of a, b, c and d can be targeted to yield sufficient nonliearity patterns based on the posterior distribution, and the values can be further tuned based on Bayesian model selection criterions such as BIC.



(a) $a = 80; b = 1$.

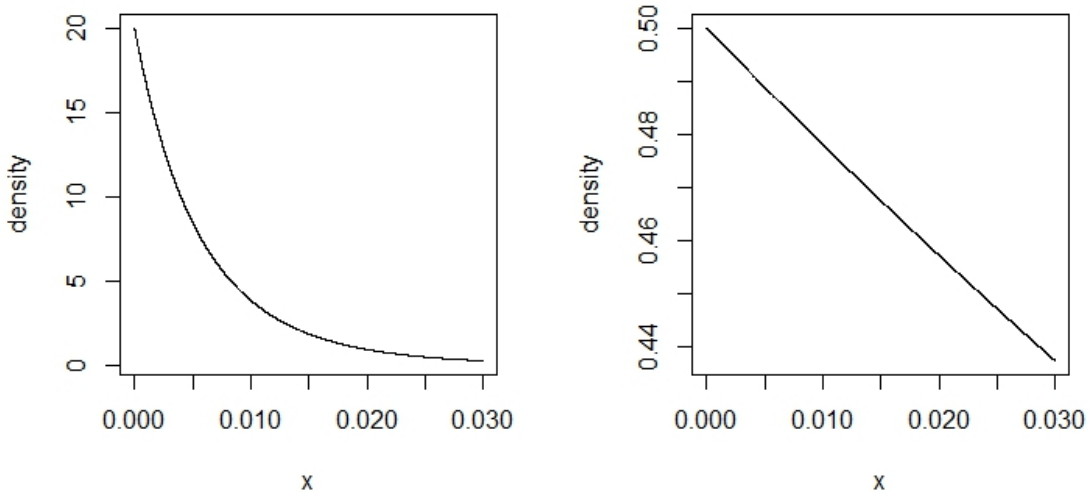
(b) $a = 0.5; b = 0.2$.

Figure 4.1: Probability density function of the inverse gamma distribution with different sets of parameters; (a): $a = 80$ and $b = 1$; (b): $a = 0.5$ and $b = 0.2$.

4.2.4 MCMC Sampling from the Posterior Distribution

As previously noted, we specify a IGMRF prior on z and γ, a diffuse Gaussian distribution for β, a inverse gamma distribution for η, and a Pareto prior for δ. We have also introduced a latent variable θ since the Pareto distribution can be seen to be a scale mixture of exponentials. Based upon this model specification, the conditional posterior distributions are given by

$$
\begin{aligned}
&z|Y, X, u, v, \gamma, \beta, \delta \propto \exp\left\{\textstyle\sum_i L(Y_i, X_i, u_i, v_i) - \frac{\delta}{2} z^T A_\gamma z\right\} \\
&\beta|Y, X, u, v, z \propto \exp\left\{\textstyle\sum_i L(Y_i, X_i, u_i, v_i) - (\beta - \mu_0)^T \Sigma^{-1} (\beta - \mu_0)/2\right\} \\
&\gamma|z, \eta, \delta, \theta \propto \exp\left\{-\delta z^T A_\gamma z/2 - \delta\eta\gamma^T M\gamma/2\right\} I_{\mathbf{1}^T\gamma=0} \\
&\delta|z, \eta \sim \text{Gamma}(n - 1/2, z^T A_\gamma z/2 + \eta\gamma^T M\gamma/2 + \theta) \\
&\eta|\gamma \propto \eta^{(n-2)/2-a-1} \exp\{-\delta\eta\gamma^T M\gamma/2 - b/\eta\} \\
&\theta|\delta \sim \text{Gamma}(d + 1, \delta + c).
\end{aligned}
\tag{4.10}
$$



(a) $c = 0.05; d = 8$. (b) $a = 2; b = 8$.

Figure 4.2: Probability density function of the Pareto distribution with different sets of parameters; (a): $c = 0.05$ and $d = 8$; (b): $c = 2$ and $d = 8$.

From the above, we implement the following posterior MCMC sampling procedure using a Gibbs sampler. We use block sampling (Cater and Kohn, 1996) to sample z by first dividing the vector into several small blocks and then updating block-by-block using a Metropolis-Hastings algorithm. The point locations in each block, and block size, should be chosen so that, based on the GMRF prior, the points within the block are correlated and the number of dependent points outside of the block are as small as possible If the number of dependent points that are outside of the block is too large, convergence will be slowed because of the high dependency among blocks. For example, Figure 4.3 shows two types of block choice denoted by "$\bullet$". The points denoted by "$*$" are dependent with the block. In the first example, points within the block are highly correlated, where the number of dependent points outside of the block is 20. In the second example, the left two points in the block and the right four points in the block are not correlated, but the number of dependent points outside of the block is $24 > 20$. Hence, the first example shown in Figure 4.3a would result in faster convergence than the second example shown in Figure 4.3b.

(a) Example1. (b) Example 2.

Figure 4.3: Examples of blocks and dependent points when performing block sampling procedures for MCMC samples of z. (a): Example 1 ; (b): Example 2.

The proposal for the Metropolis-Hastings algorithm is taken to be the prior on that block conditioned on points outside of the block. We define $z = (z_{bl}, z_{-bl})'$, where z_{bl} represents the values at the points in a block, and z_{-bl} are the values at the points outside the block. Then the posterior distribution of z_{bl} is

$$\pi(z_{bl}|Y, X, u, v, z_{-bl}, \gamma, \beta, \delta) \propto p(Y|X, u, v, z, \beta)\pi^*(z_{bl}|z_{-bl}, \gamma, \delta).$$

Since

$$z_{bl}, z_{-bl}|\gamma, \delta \propto \exp\left\{\frac{\delta}{2}[z_{bl}, z_{-bl}]^T \begin{pmatrix} A_\gamma[bl, bl] & A_\gamma[bl, -bl] \\ A_\gamma[-bl, bl] & A_\gamma[-bl, -bl] \end{pmatrix} [z_{bl}, z_{-bl}]\right\},$$

So the conditional prior for z_{bl} is

$$z_{bl}|z_{-bl}, \gamma, \delta \sim \mathcal{N}\left(-A_\gamma[bl, bl]A_\gamma^{-1}[bl, -bl]z_{-bl}, \delta A_\gamma^{-1}[bl, bl]\right), \tag{4.11}$$

Although A_γ is not full rank (the rank is $n-1$) and the prior on z is an IGMRF of the first order, the prior distribution of a block conditional on the boundary values is a propoer multivariate normal distribution indicating that the conditional covariance matrix $A_\gamma[bl, bl]$ is of full rank (Cater and Kohn, 1996). As such, we use the conditional prior in equation (4.11) as the proposal distribution when updating each block. The candidate z_{bl}^* drawn from

the proposal distribution is then accepted with probability

$$\begin{aligned}\pi_z^{accept} &= \min\left(1, \frac{\pi(z_{bl}^*|Y,X,u,v,z_{-bl},\gamma,\beta,\delta)\pi^*(z_{bl}|z_{-bl},\gamma,\delta)}{\pi(z_{bl}|Y,X,u,v,z_{-bl},\gamma,\beta,\delta)\pi^*(z_{bl}^*|z_{-bl},\gamma,\delta)}\right) \\ &= \min\left(1, \frac{p(Y|X,u,v,z_{bl}^*,z_{-bl},\beta)\pi^*(z_{bl}^*|z_{-bl},\gamma,\delta)\pi^*(z_{bl}|z_{-bl},\gamma,\delta)}{p(Y|X,u,v,z_{bl},z_{-bl},\beta)\pi^*(z_{bl}|z_{-bl},\gamma,\delta)\pi^*(z_{bl}^*|z_{-bl},\gamma,\delta)}\right) \\ &= \min\left(1, \frac{p(Y|X,u,v,z_{bl}^*,z_{-bl},\beta)}{p(Y|X,u,v,z_{bl},z_{-bl},\beta)}\right).\end{aligned}$$

β can be updated using a Metropolis-Hastings algorithm with proposal distribution

$$\beta^* \sim \mathcal{N}\left((\Sigma_0^{-1} + V^{-1}(\hat{\beta}))^{-1}(\Sigma_0^{-1}\mu_0 + V^{-1}(\hat{\beta})\hat{\beta}), (\Sigma_0^{-1} + V^{-1}(\hat{\beta}))^{-1}\right),$$

where $\hat{\beta}$ is the MLE of β conditional on z, and $V(\beta)$ is the asymptotic variance of $\hat{\beta}$ given by the inverse of the corresponding Fisher information.

Similar to the sampling procedure for z, block sampling can again be used to sample from the posterior distribution of γ, but more steps are required because there is a constraint for γ, given by $\mathbf{1}^T\gamma = 0$. To sample γ, we start by dividing the vector into several blocks. For each block, a candidate γ_{bl}^* can be drawn from the prior distribution conditional on γ_{-bl}^*:

$$\gamma_{bl}|\gamma_{-bl},\eta,\delta \sim \mathcal{N}\left(-M[bl,bl]M^{-1}[bl,-bl]\gamma_{-bl}, \eta\delta M^{-1}[bl,bl]\right),$$

then γ_{bl}^* is accepted with probability

$$\pi_\gamma^{accept} = \min\left(1, \frac{p(z|\gamma_{bl}^*,\gamma_{-bl},\delta)}{p(z|\gamma_{bl},\gamma_{-bl},\delta)}\right).$$

Centering γ_{bl}^* to $\mathbf{1}^T\gamma_{bl}$ so that $\mathbf{1}^T\gamma_{bl}^* = \mathbf{1}^T\gamma_{bl}$, it is seen that the constraint $\mathbf{1}^T\gamma = 0$ is satisfied. However, this implies that the sum of the γ_i for each block update will never changes throughout the MCMC iterations. Therefore, two different sets of blocks can be used alternately. For example, in the mth iteration, γ may be updated based on the first set of blocks shown in Figure 4.4a and in the $m+1$th iteration, γ may be updated based on the second set of blocks shown in Figure 4.4b. It should be noted that any two blocks from the two sets cannot be the same.

(a) Block set 1. (b) Block set 2.

Figure 4.4: Examples of two different sets of blocks.

The posterior distributions of δ and θ are both gamma distributions as shown in (5.9), and are easily sampled from. The posterior conditional distribution of η is log-concave, so the adaptive rejection Metropolis sampling (ARMS) method (Gilks and Wild, 1992) can be utilized to sample from the posterior distribution of η. ARMS is implemented in the R package `ars` (Rodriguez and Komarek, 2014). The basic idea of ARMS is that for a concave (or log-concave) distribution, samples can be drawn using the rejection sampling procedure, and the rejection envelop is a (exponential) piecewise function based on adaptively specified knots. Figure 4.5 shows an example for the piecewise function corresponding to a concave probability distribution.

4.3 Simulation Study

We performed a single simulation study using data generated on the map of Massachusetts State. The geographical distribution of the data are shown in Figure 4.6. Consistent with the Massachusetts preterm birth data, we considered a binary outcome generated by

$$y \sim \text{Bern}\left(\text{logit}^{-1}(X\beta + f(u,v))\right)$$

where $\beta = [\log(0.8), \log(1.5)]^T$ and specify a nonlinear spatial effect f such that

$$-1.5 + 3\exp\left(-3\left(u/100000 + v/100000 - 11\right)^2 - 2\left(u/100000 - v/100000 + 6.7\right)^2\right) \quad (4.12)$$

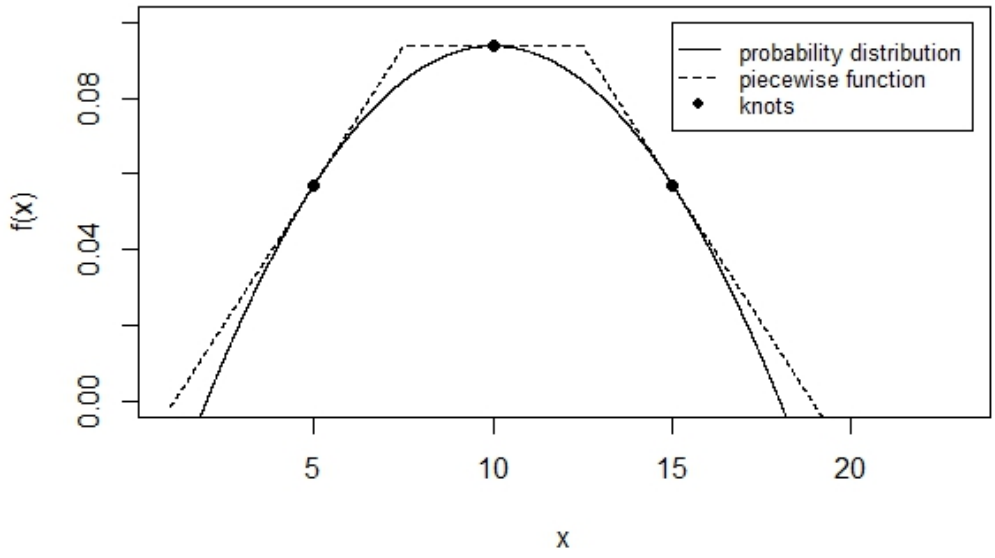


Figure 4.5: Example to illustrate how the adaptive rejection Metropolis sampling works: For a concave distribution with solid line, and knots specified at 5, 10 and 15, the rejection sampling procedure can be performed using a piecewise function with dash line as the rejection envelop.

where u and v are two coordinate parameters (measured in meters), and were generated by Lambert conformal conic projection for the State of Massachusetts. A thin-plate, equally spaced by 6.6 km, as well as the contours of spatial effect in (4.12) are shown in Figure 5.2. We did not include an intercept in the data generating model because the nonlinear spatial effect inherently accounts for this. The grid size of 495 and a sample size of $N = 1000$ was used. We used prior parameter values of $a = 100, b = 1, c = 0.001, d = 8$ for the hyperpriors of δ and η. For sampling from the posterior distribution of z, the vector z was divided into 44 blocks, and each block contained 9 14 elements, as shown in Figure 4.8a. To sample γ, two sets of blocks were needed, and the second set of blocks (42 in total) is shown in Figure 4.8b. For this simulated example, $15,000$ MCMC samples were drawn, and we chose a burn-in of $10,000$ to ensure convergence. The trace plots of the posterior draws are shown in Figure 4.15

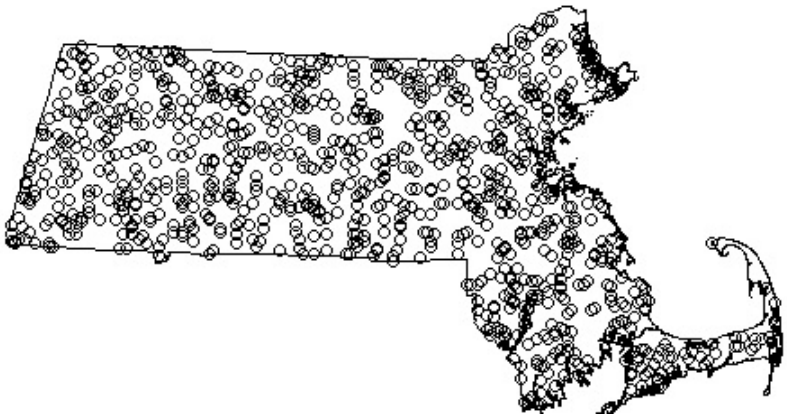


Figure 4.6: Geographical distribution of the simulated data from Massachusetts State.

We compare the performance of the Bayesian adaptive method with the performance of the GAM, which is commonly used to estimate the adjusted spatial effects for discrete disease outcomes. The GAM method is implemented using the `GAM` package in `R` with LOESS as the smoothing method. Two spans are chosen for comparison for the GAM method. First, a choice of 0.2 (suggested by AIC as the optimal span size) and a choice of 0.1 to illustrate the effect of potential under-smoothing in the GAM.

The scatterplot of the posterior mean of z versus the true adjusted spatial effects of all estimates are plotted in Figure 4.9. When the span size for the GAM is large (Figure 4.9a), the overall variation of the estimates is relatively small, but for the area where the spatial

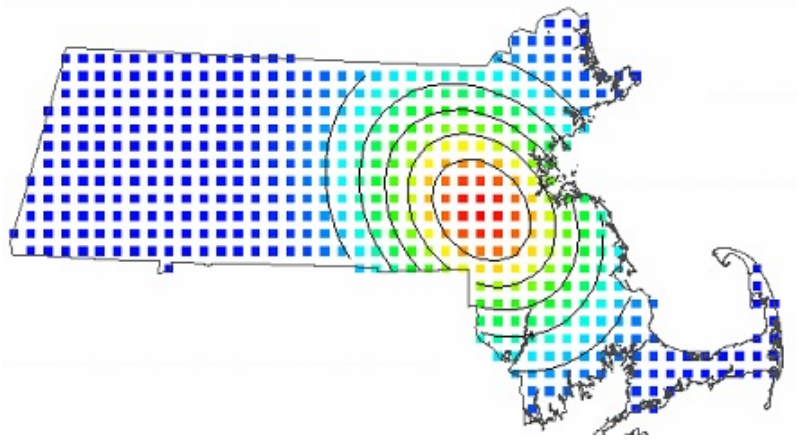
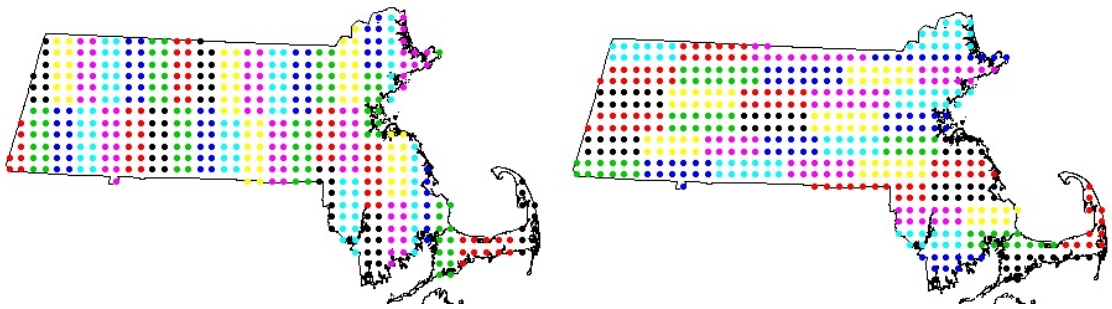


Figure 4.7: The colormap with contours of the spatial effect function.

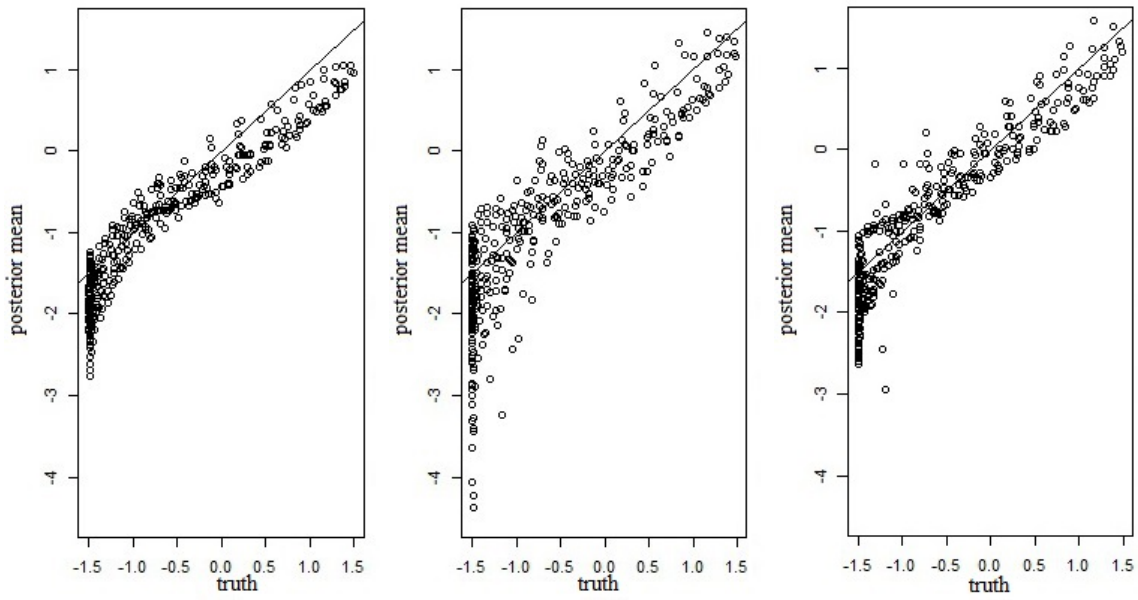


(a) First set of blocks. (b) Second set of blocks.

Figure 4.8: Two sets of blocks used in the simulation study when the block sampling is performed to sample from the posterior distribution of z and γ. (a) has 44 blocks and was used when sampling z. (b) has 42 blocks and was used when sampling γ.

effect changes quickly, the bias model fits is pronounced. Conversely, when the span size for the GAM is small (Figure 4.9b), the resulting bias is relatively small, but near the area where the spatial effect is flat the variation of the estimates is large. However, with the proposed adaptive method the results indicate both approximately unbiased estimates and along with relatively smaller overall variation (Figure 4.9c).

To further assess the performance of the proposed method, we divide the map into six regions as shown in Figure 4.10, and compare the MSE of the three methods region by region in Table 4.1. The MSE for the proposed adaptive smoothing method is 0.158, which is smaller than the GAM methods with either span size (0.173 for the span size of 0.2 and 0.393 for the span size of 0.1). Not surprisingly, the GAM method with a span size suggested by AIC (span size of 0.2) attained smaller MSEs than the GAM method using a smaller span size in all the six regions. However, the adaptive smoothing procedure performs better than the GAM with a span size of 0.2 in all other regions other than region B, where spatial effects are consistently homogeneous but near the high spatial effects of Region C.



(a) GAM (span=0.2). (b) GAM (span=0.1). (c) Adaptive smoothing.

Figure 4.9: Performance comparison (estimated adjusted spatial effect versus the true adjusted spatial effect) of the GAM with LOESS as the smoothing method and the proposed adaptive smoothing method. (a) GAM model using span size of 0.2 (as selected by AIC); (b) GAM model using span size of 0.1; (c) The proposed adaptive smoothing method.

4.4 Application to Massachusetts Preterm Birth Study

In this section we apply our proposed adaptive smoothing method to a study aimed at evaluating the spatial effect on preterm birth (< 37 weeks) risk (see Chapter 1 for further details). The data consist of birth outcomes data from the state of Massachusetts and are comprised of data from $N = 61,942$ live births. Geolocations of each mother's residence are used as spatial parameters (measured in meters). Potential confounding factors also available for analysis include maternal race, and indicator of maternal smoking during pregnancy, maternal age, median income within the census block of the residence of the mother, insurance type, an indicator of maternal alcohol consumption during pregnancy, father's education and parity. Given prior evidence on the associations between each of the above covariates and the risk of preterm birth, all variables are adjusted for as potential confounding factors. The spatial distribution of all mothers as well as pteterm birth mothers are shown in Figure 4.11.

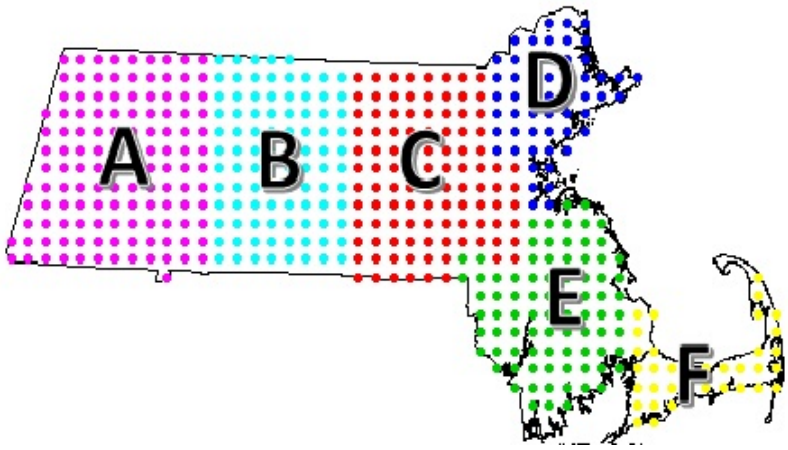


Figure 4.10: The map is divided into six regions to compare MSE of the competing methods region by region (results shown in Table).

Table 4.1: MSE of the three methods (GAM with span of 0.2, GAM with span of 0.1 and adaptive smoothing) in the six regions as shown in Figure 4.10. Results are based upon a single simulated dataset of 1000 observations on MA state.

MSE	A	B	C	D	E	F	Overall
GAM (span=0.2)	0.099	0.089	0.108	0.147	0.281	0.606	0.173
GAM (span=0.1)	0.357	0.179	0.202	0.334	0.637	1.147	0.393
Adaptive smoothing	0.095	0.105	0.082	0.104	0.232	0.575	0.158
%change	4%	-6%	18.5%	25%	17.8%	5%	11.5%

We used the same grid and set of blocks as in the simulation study presented in Section 4.3 to sample z and γ. The acceptance rates for the actual data were between 0.2 and 0.3. We used prior parameter values of $a = 80, b = 1, c = 0.001, d = 10$ for the hyperpriors of δ and η. Posterior means were calculated as summary measures of the posterior distriutions of model parameters. The MCMC procedure for the adaptive model was run for 10,000 iterations with a burn-in of 5,000. Trace plots resulting from the MCMC algorithm are shown in Figure 4.16. Because the sample size of the actual data is approximately 60 times larger than that used in the simulation study, the MCMC sampling converged faster than the simulation study. For comparison, we also fit a (nonadaptive) GAM model with span size of 0.3 (chosen by minimizing AIC). Figure 4.12b shows the estimated odds ratio of adjusted spatial effects

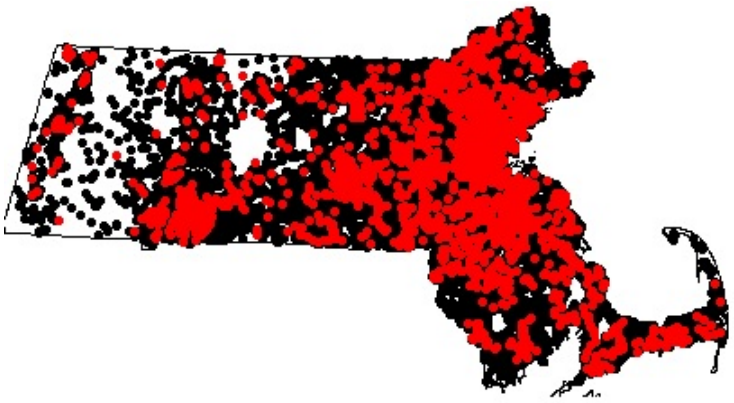
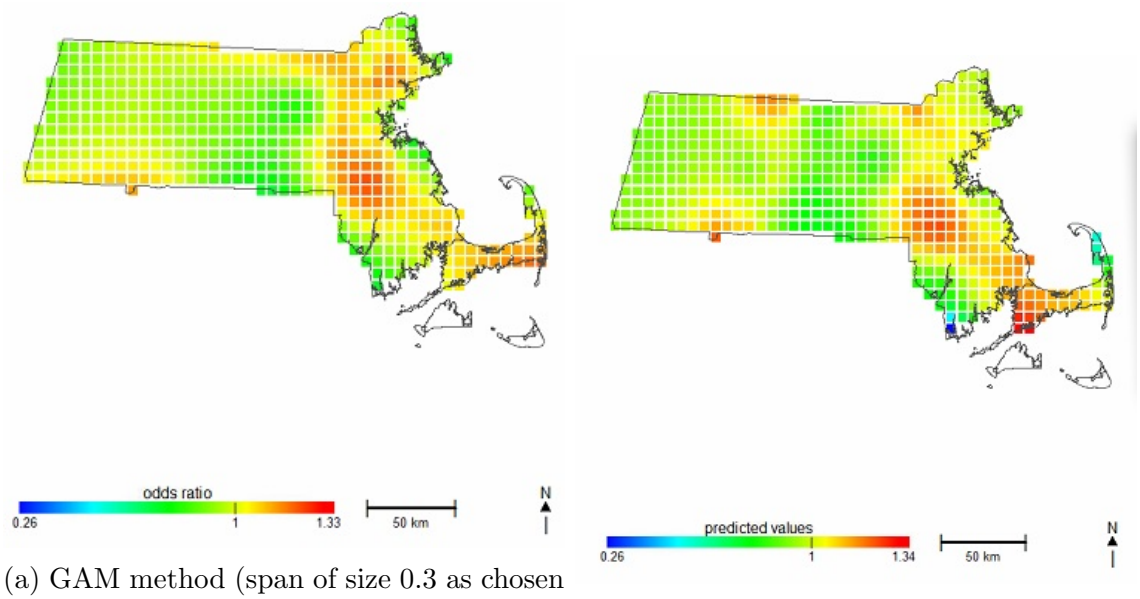


Figure 4.11: Spatial distribution of the mothers in the Massachusetts data set.

compared to median odds (marginalized over the state) using the adaptive method. Figure 4.12a displays the resulting odds ratio estimates using the GAM method.



(a) GAM method (span of size 0.3 as chosen by AIC).

(b) Adaptive method.

Figure 4.12: Colormap of the estimated odds ratio of adjusted spatial effects compared to the median odds over the state using (a): GAM method (span size of 0.3 as chosen by AIC) and (b): proposed adaptive smoothing method.

In Figure 4.13 we circle four areas where the estimates using the two methods differ. These same areas are consistent with those areas in the simulation study that indicated better performance of the adaptive smoothing procedure relative to the GAM (data in the simulation

were generated using a similar population density). Areas of significant spatial effects as suggested by 95% probability (point-wise 95%confidence) intervals based on the two methods are presented in Figure 4.14. Notably, the cluster in the southwest area of the state that is estimated to be significantly different by the adaptive smoothing approach is not highlighted by the GAM method.

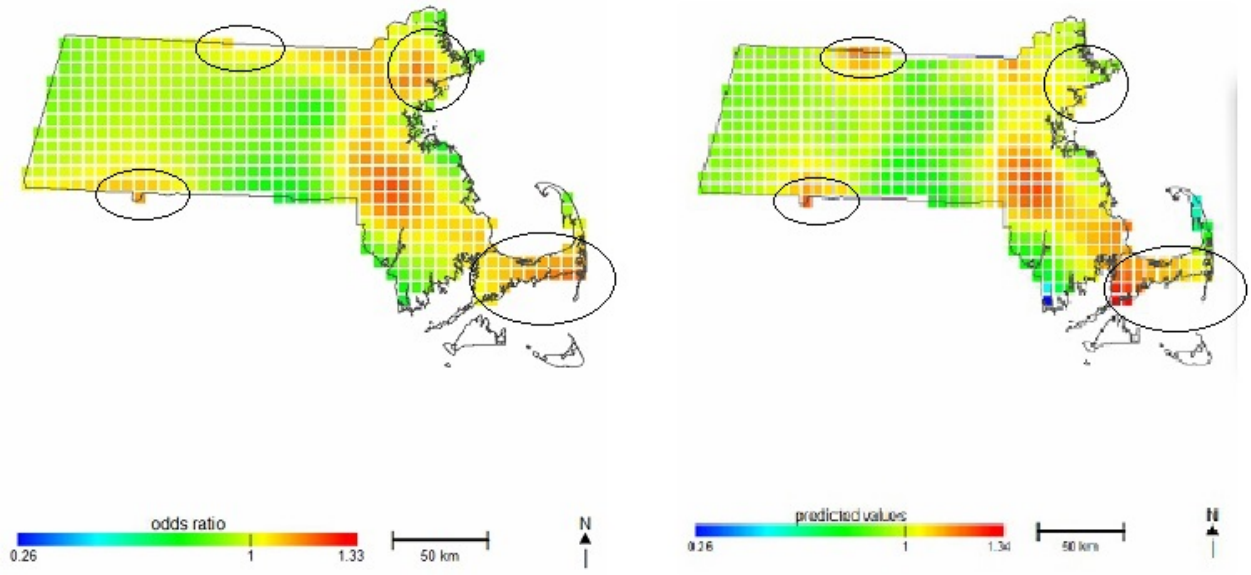
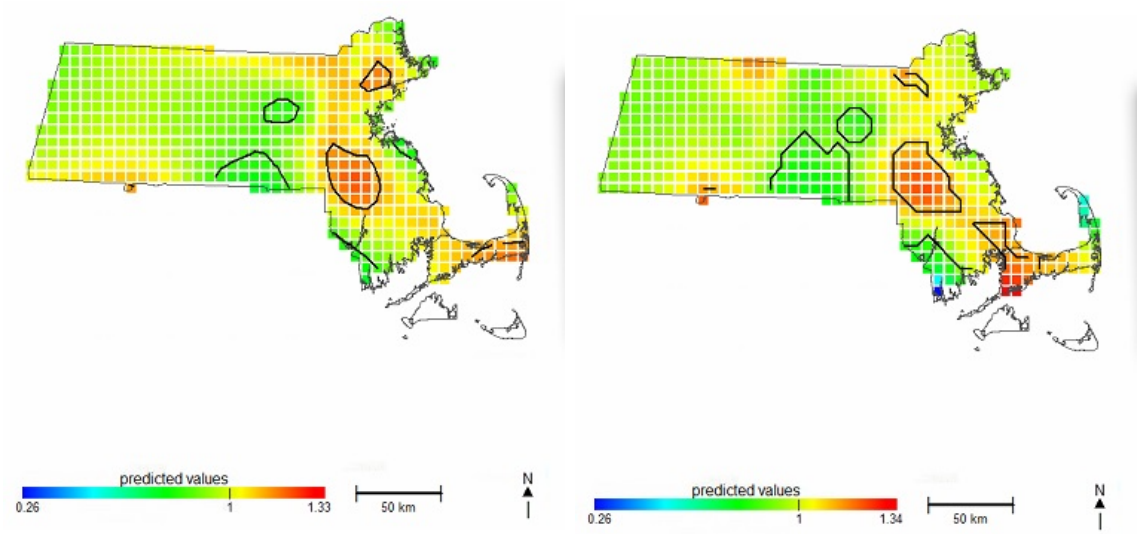


Figure 4.13: Four areas are circled where the GAM method and adaptive smoothing method show different patterns.

4.5 Discussion

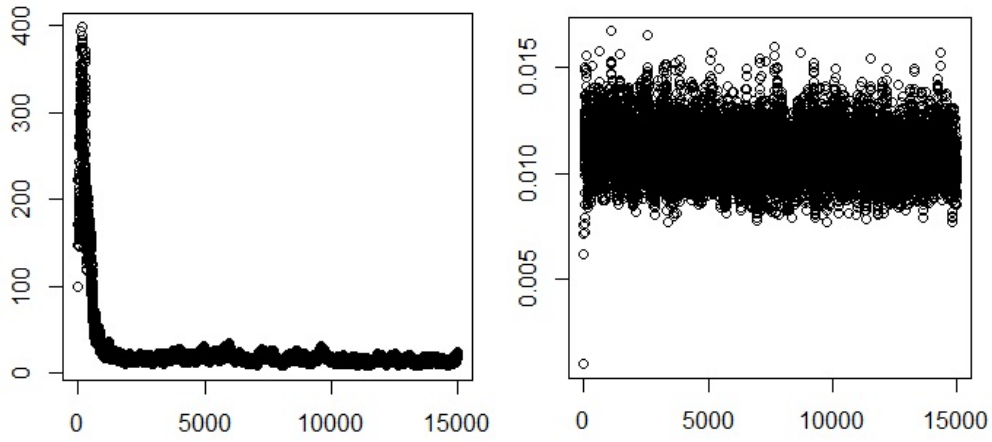
In this chapter, we introduced a novel adaptive GMRF by using a spatially varying variance component with its own GMRF prior to estimate the adjusted spatial effects for disease outcomes while accounting for confounding factors. The result is a Bayesian hierarchical model corresponding to a discretized adaptive thin-plate spline suitable for non-stationary spatial data. Bayesian computation is based on an efficient Gibbs sampler, in which two block-sampling Metropolis-Hastings algorithms are used. The performance of our non-stationary spatial model has been demonstrated by a simulated example informed by real spatial epidemiologic data. Further, the method was applied to preterm birth data from the state of



(a) GAM (span size of 0.2).

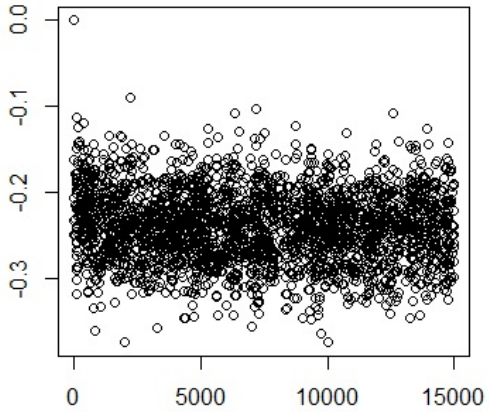
Figure 4.14: Significant areas for the preterm birth risk in Massachusetts state suggested using (a) the GAM method and (b)the proposed adaptive smoothing method.

Massachusetts to evaluate the role of geospatial location on preterm birth risk. When compared to a more standard GAM, the applied results indicate that the adaptive smoothing approach was able to identify significantly different geographic clusters indicating differential risk of preterm birth. Future work may include more efficient approaches to specifying prior parameter values for the hyperpriors of the model, and extending the work to estimate the possible spatio-temporal effects in epidemiology studies.

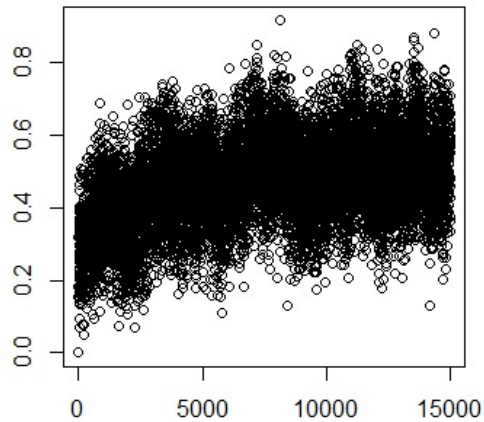


(a) δ.

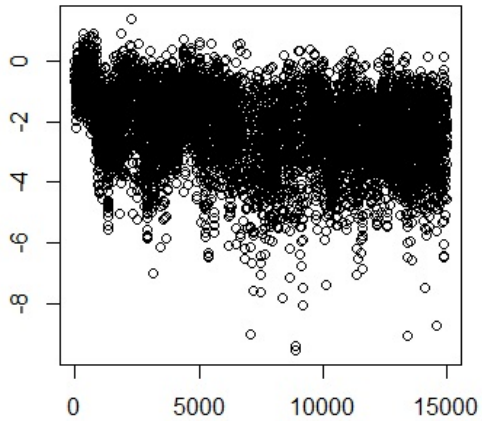
(b) η.



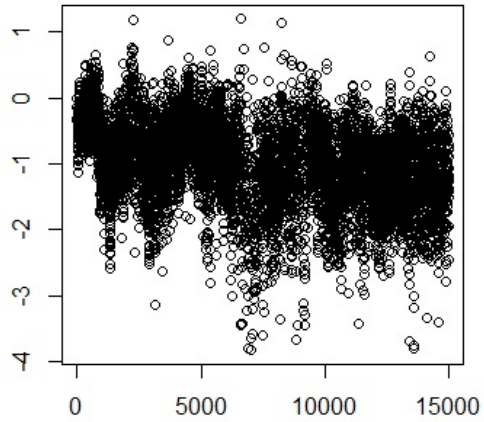
(c) β_1.



(d) β_2.

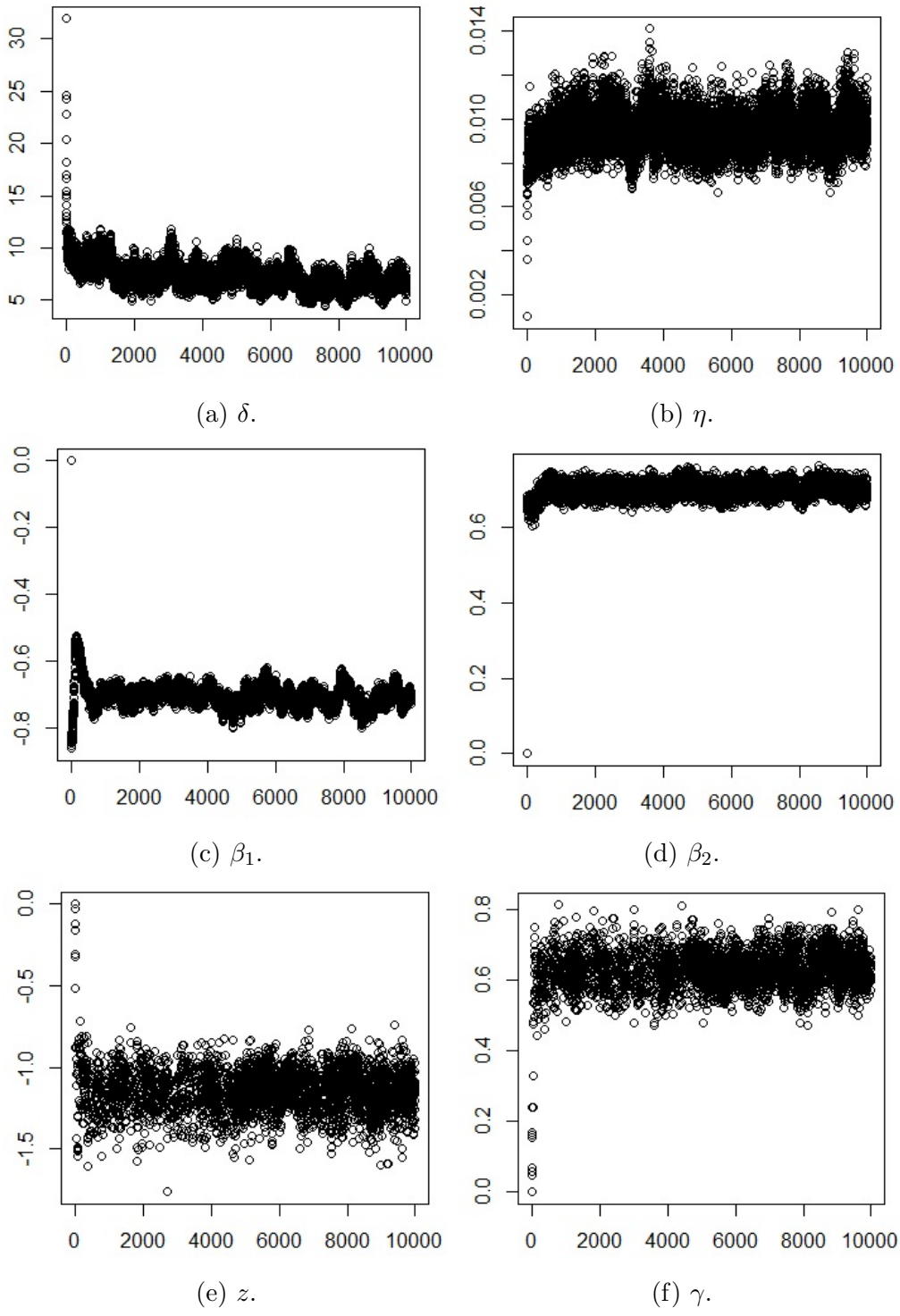


(e) z.



(f) γ.

Figure 4.15: Traces of the MCMC samples in the simulation study.



(a) δ. (b) η.

(c) β_1. (d) β_2.

(e) z. (f) γ.

Figure 4.16: Trace plots of the MCMC samples when estimating the adjusted spatial effect on preterm birth risk based on the Massachusetts data.

Chapter 5

Spatial Analysis for Censored Survial Data via Adaptive smoothing

Spatial effect estimation in the context of survival outcomes is of great interest in epidemiology studies (Henderson et al., 2002; Bristow et al., 2014). In this chapter, we extend the adaptive smoothing technique proposed in Chapter 4 to the analysis of right-censored survival outcomes. A Weibull distribution is adopted to model the time-to-event response, and the Bayesian hierarchical model for adaptive smoothing of adjusted survival rates over geolocations is introduced in Section 5.1. Performance of the method is evaluated by a simulation study in Section 5.2, and an application to a population-based epidemiologic study investigating spatial heterogeneity in survival times among California advanced-stage ovarian cancer patients is presented in Section 5.3. We conclude the chapter with a discussion of the proposed methods and of future work in Section 5.4.

5.1 methods

5.1.1 Data Model

We consider a Weibull distribution for the survival time T, so that the survival function is given by

$$S_T(t) = Pr(t > T) \exp\{-\lambda t^\alpha\}. \tag{5.1}$$

and the probability distribution for survival time is given by

$$f_T(t) = -\alpha\lambda t^{\alpha-1} \exp\{-\lambda t^\alpha\}. \tag{5.2}$$

Let (u, v) represent the longitude and latitude of the location of a given observation, and let $X \in \mathbb{R}^p$ denote a vector of covariates to be adjusted in the model. Then we assume that

$$\log(\lambda) = X\beta + f(u, v),$$

where β reflects the association between the adjustment covariates and the survival hazard, and $f(u, v)$ represents the spatial effect of location on the survival hazard. As noted in Chapter 2, for right-censored survival data with an observed time t and the corresponding censoring status δ, the likelihood L can be written as

$$\begin{aligned} L &= f(t)^\delta S(t)^{1-\delta} \\ &= (\alpha\lambda t^{\alpha-1})^\delta \exp\{\lambda t^\alpha\}. \end{aligned}$$

Assuming there are N independent observations, let (u_i, v_i) represent the longitude and latitude of the i^{th} observation's location on a map, and let t_i, δ_i, and X_i represent the observed time, censoring status and adjustment covariates values of subject i, respectively.

We consider a thin-plate spline estimator where f is the solution to

$$\hat{f} = \min_f \left\{ -\sum_i \log(L(t_i, \delta_i, u_i, v_i, X_i)) + \phi J_2(f) \right\},$$

with $J_2(f)$ representing the roughness penalty for f on $\mathbb{R}^2$ such that

$$\begin{aligned} J_2(f) &= \iint_{\mathbb{R}^2} \left[\left(\frac{\partial^2 f(u,v)}{\partial u^2} \right)^2 + \left(\frac{\partial^2 f(u,v)}{\partial u \partial v} \right)^2 + \left(\frac{\partial^2 f(u,v)}{\partial v^2} \right)^2 \right] du dv \\ &= \iint_{\mathbb{R}^2} \left[\left(\frac{\partial^2}{\partial u^2} + \frac{\partial^2}{\partial v^2} \right) f(u,v) \right] du dv, \end{aligned} \tag{5.3}$$

and ϕ is a smoothing parameter to control the trade-off between the likelihood of the data and the smoothness from the penalty term $J_2(f)$.

To introduce adaptivity, a spatially varying function $\phi(u, v)$ is used instead of a global ϕ. Therefore, $\phi J_2(f)$ in (**??**) is replaced by $J_2^*(f)$ given by

$$J_2^*(f) = \iint_{\mathbb{R}^2} \phi(u,v) \left[\left(\frac{\partial^2}{\partial u^2} + \frac{\partial^2}{\partial v^2} \right) f(u,v) \right] du dv.$$

5.1.2 Spatially Varying Thin-Plate Priors

We start by generating an evenly-space grid with n grid points across the map, and vectorized longitudes and latitudes of the grid points denoted by $\tilde{u} = [\tilde{u}_1, \cdots, \tilde{u}_n]^T$ and $\tilde{v} = [\tilde{v}_1, \cdots, \tilde{v}_n]^T$, respectively. Let the vector z denote the spatial effects on the vectorized locations, and D denote the incidence matrix (see Chapter 2). Then we can vectorize the data model as

$$\log(\lambda) = X\beta + Dz.$$

As in Chapter 4, the penalty term $J_2^*(f)$ can be approximated by

$$J_2^*(f) \approx ||(\exp\{\gamma/2\})^T Bz||^2 = z^T A_\gamma z,$$

where $\exp\{\gamma/2\} = [e^{\gamma_1/2}, e^{\gamma_2/2}, \cdots, e^{\gamma_{n-1}/2}]^T$, and $A_\gamma = B' \text{diag}\{e^{\gamma_1}, e^{\gamma_2}, \cdots, e^{\gamma_{n-1}}\} B$. The matrix B is obtained by deleting the first row of the second order difference structure matrix,

$\tilde{B}$, defined as

$$[\tilde{B}z]_j = (z_{jR}I_{jR} + z_{jL}I_{jL} + z_{jD}I_{jD} + z_{jU}I_{jU}) - z_j(I_{jR} + I_{jL} + I_{jD} + I_{jU}), \tag{5.4}$$

where I_{jL} is an indicator that there exists a grid point at the left of the point ($\tilde{u}_j$,$\tilde{v}_j$), and I_{jR}, I_{jU} and I_{jD} are indicators at other three positions: right, upper and lower respectively. Let z_{jL} denote the smooth function of the grid point at the left of the point ($\tilde{u}_j$,$\tilde{v}_j$), and z_{jR}, z_{jU} and z_{jD} denote corresponding smooth functions at the other three positions: right, upper and lower, respectively. It is the vector γ that introduces spatially varying weights based on the local extent of the smooth function when calculating the penalty term $J_2^*(f)$ and hence allows for spatial smoothing adaptivity.

From the above, the optimization problem becomes

$$\hat{z} = \min_z \left\{ -\sum_i \log(L(t_i, \delta_i, z_i, X_i)) + z^T A_\gamma z \right\}, \tag{5.5}$$

which suggests an IGMRF of the first order prior for z given by

$$z \propto \exp\left\{ -\frac{\zeta_1}{2} z^T A_\gamma z \right\}.$$

We then specify a simpler IGMRF of the first order prior for γ with

$$\gamma \propto \exp\left\{ -\frac{\zeta_1\zeta_2}{2} \gamma^T M \gamma \right\} I_{\mathbf{1}^T\gamma=0},$$

where $M = C^TC$, and C is the first order difference structure matrix such that the jth element of $C\gamma$ yields

$$[C\gamma]_j = \gamma_{jR}I_{jR} + \gamma_{jD}I_{jD} - \gamma_j(I_{jR} + I_{jD}). \tag{5.6}$$

5.1.3 Removing Links Between Points Without Spatial Effect Correlation

In section 5.1.2, IGMRF priors are specified for z and γ. Because of the local Markov property of IGMRFs, the spatial effect at one grid point is more correlated with the spatial effects within it's neighborhood. However, in epidemiology studies, spatial effects on outcome are not always highly correlated between two geographically close points. As one example, consider in the California ovarian cancer study presented in Chapter 1, where the spatial effect is hypothesized to be driven by healthcare resource utilization disparities. However, linear spatial distance on a map may not be a good indicator for healthcare access. For example, If there is a mountain or forest between two seemingly neighboring geolocations, then residents within the two locations may not share any common healthcare resources due to the travel times required to move from one location to the other. Because of this, the spatial effects at the two locations would not be correlated one appearing to be geographically in the neighborhood of the other.

One benefit of using an IGMRF prior in the context of a thin-plate spline is that it is convenient to remove on grid point from the neighborhood of another point when specifying IGMRF priors. For example, in the grid shown in Figure 5.1, if it is assumed that the spatial effects between the two black points are not correlated, then the link between the two points needs to be removed. When specifying the matrices $\tilde{B}$ (5.4) and C (5.6), the link between the two points can easily be removed by setting the indicator $I_{jR} = 0$ for the left grid point and the indicator $I_{jL} = 0$ for the right grid point. The end result is that the smoothing process will deterministically remove the influence of the neighboring points as may be desired in cases where natural obstructions exist between two neighboring locations.

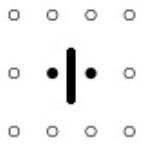


Figure 5.1: An example of removing links between two points.

5.1.4 Hyperpriors and MCMC Sampling Scheme

We specify a gamma(a_0, b_0) distribution as the prior for α in the data model (5.1) and (5.2), and a diffuse Gaussian prior $\mathcal{N}(\mu_0, \Sigma_0)$ for β. The hyperpriors for ζ_1 and ζ_2 are taken to be a Pareto(c, d) prior (5.7) and an inverse-gamma(a, b) prior (5.8), respectively, so that

$$\zeta_1 \propto \frac{c^d}{(c+\zeta_1)^{d+1}}; \tag{5.7}$$

$$\zeta_2 \propto \zeta_2^{-a-1} \exp(-b/\zeta_2). \tag{5.8}$$

As noted in Chapter 4, the Pareto distribution can be written as a scale mixture of exponentials by introducing a latent variable θ: $\zeta_1|\theta \sim \text{Exp}(\theta); \theta \sim \text{Gamma}(d, c)$, and we do this to facilitate sampling from the posterior distribution.

Given the above specifications, the conditional posterior distributions for the model parameters are

$$\begin{aligned}
&z|t, \delta, X, u, v, \gamma, \beta, \alpha, \zeta_1, \propto \exp\left\{\textstyle\sum_i L(t_i, \delta_i, X_i, u_i, v_i) - \frac{\zeta_1}{2} z^T A_\gamma z\right\} \\
&\beta|t, \delta, X, u, v, z \propto \exp\left\{\textstyle\sum_i L(t_i, \delta_i, X_i, u_i, v_i) - (\beta-\mu_0)^T \Sigma^{-1}(\beta-\mu_0)/2\right\} \\
&\alpha|t, \delta, X, u, v, z, \beta \propto \alpha^{a_0-1} \exp\left\{\textstyle\sum_i L(t_i, \delta_i, X_i, u_i, v_i) - b_0\alpha\right\} \\
&\gamma|z, \zeta_1, \zeta_2, \theta \propto \exp\left\{-\zeta_1 z^T A_\gamma z/2 - \zeta_1\zeta_2\gamma^T M\gamma/2\right\} I_{\mathbf{1}^T\gamma=0} \\
&\zeta_1|z, \zeta_2, \theta \sim \text{Gamma}(n-1/2, z^T A_\gamma z/2 + \zeta_2\gamma^T M\gamma/2 + \theta) \\
&\zeta_2|\gamma \propto \zeta_2^{(n-2)/2-a-1} \exp\{-\zeta_1\zeta_2\gamma^T M\gamma/2 - b/\zeta_2\} \\
&\theta|\zeta_1 \sim \text{Gamma}(d+1, \zeta_1 + c).
\end{aligned} \tag{5.9}$$

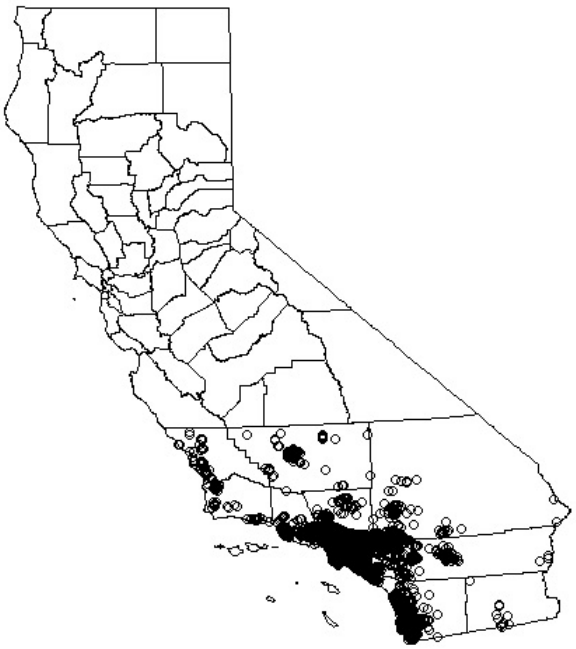


Figure 5.2: Spatial distribution of the patients in the southern California ovarian cancer dataset.

A block-sampling procedure can be used to sample from the posterior distribution of z by dividing the vector into several blocks. The block-sampling procedure is also used to sample γ, but two sets of the blocks must to used alternately to deal with the following identifitability constraint for: γ of $\mathbf{1}^T\gamma = 0$. The adaptive rejection Metropolis sampling method can be used to sample α and ζ_2. Finally, the posterior distribution of ζ_1 can be identified as a Gamma distribution, and hence is trivial to sample from.

5.2 Simulation study

We performed a simulation study using data generated on the map of Southern California. The geographical distribution of the data is shown in Figure 5.2. For the simulation study we assumed a sample size is $N = 6,878$, as is consistent with the California ovarian cancer data presented in Chapter 1.

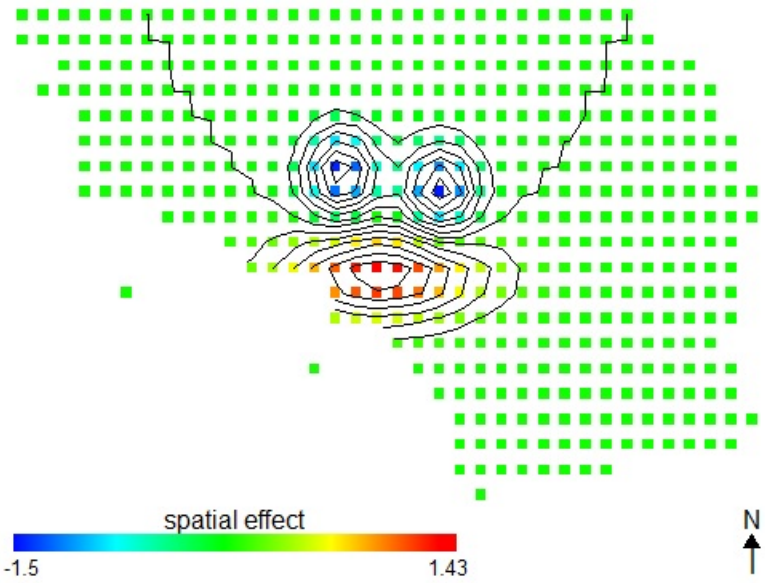


Figure 5.3: Contours of the adjusted spatial effect on Southern California state as well as the thin-plate used in the Bayesian methods.

Simulated survival times were generated from a Weibull distribution with $\alpha = 0.8$ and

$$\log(\lambda) = X\beta + f(u, v),$$

where $\beta = 0.3$ and the nonlinear spatial effect f were specified as

$$\begin{aligned} f(u,v) = \; & -1.5\exp\left(-0.1\left((u-2355000)/10000\right)^2 - 0.1\left((v+13000)/10000\right)^2\right) \\ & -1.5\exp\left(-0.1\left((u-2440000)/10000\right)^2 - 0.1\left((v+25000)/10000\right)^2\right) \\ & +1.5\exp\left(-1/70\left((u-2390000)/10000\right)^2 - 0.1\left((v+85000)/10000\right)^2\right), \end{aligned} \tag{5.10}$$

with u and v denoting coordinate parameters (measured in meters), and generated by Lambert conformal conic projection for the State of California. The contours of the data-generating spatial effect function (5.10) and the thin-plate grid used in the simulation are shown in Figure 5.3. The generated survival times had a median of 12.99 years, and independent censoring times were sampled from a uniform$(0, 25)$ distribution. Then the observed time in the simulation was then calculated as the minimum of the true survival time and the censoring time. An event status indicator of whether the survival time occurred at or before the respective time of censoring was then generated for each subject.

The thin-plate grid used in the simulation study contained 458 grid points. We considered two types of GMRFs. One is based on the regular grid (denoted GMRF1). For the other grid (denoted GMRF2), because of the mountains lying to the north of Los Angeles, links between the grid points seperated by the red line shown in Figure 5.4 were deleted when calculating the matrix B for prior of z and the matrix M for prior of γ. Given the size of of the mountain range and the travel time from one side of the mountains to the other, this removal is scientifically justified if it is believe that healthcare access is the primary driver of spatial impacts on the survival times of patients.

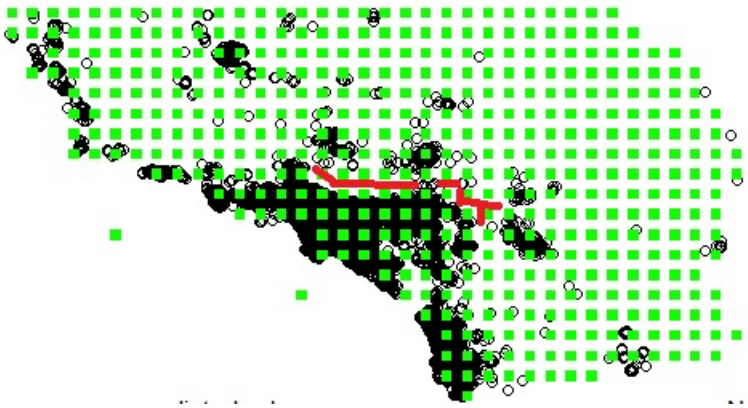


Figure 5.4: Links where the red line crosses are deleted when calculating the IGMRF structure matrices on the southern California map.

We used parameter values of $a_0 = 2, b_0 = 2, a = 50, b = 1, c = 0.01, d = 4$ for the priors of α,ζ_1 and ζ_2. $10,000$ MCMC iterations are preform with a burn-in of $5,000$. Examples of the MCMC sample trace plots are shown in Figure 5.10. The performance of the Bayesian adaptive method was compared with the performance of the Cox proportional hazards additive model with a LOESS smoothing method (for details see Chapter 3). The span size of the LOESS smoother in the Cox additive model was chosen to minimize AIC, resulting in a span size value of 0.3.

A scatterplot of posterior mean of z versus the true adjusted spatial effect function using GMRF1 and GMRF2 are plotted in Figure 5.5b and Figure 5.5c, respectively. The estimates

from GMRF1 show some bias. Specifically, the positive areas (increased hazard) tend to be under-estimated and the negative areas (decreased hazard) tend to be over-estimated. This is because the two areas are linear close to one another, and hence have influenced the estimate of each other as expected with the smoothing procedure. The estimates of GMRF2, where the links between points separated by a mountain range have been removed, are generally more consistent with the truth function. The variation around zero is large because the sample size in those areas is small and the geolocation distribution is sparse. Figure 5.5a shows the scatter plot of the estimates using the Cox additive model with LOESS smoothing versus the true adjusted spatial effect function. The observed bias of the LOESS fit is between 1.5 and -1.5 and is more pronounced the two adaptive methods when the spatial effect function varies more dramatically. This is expected as the adaptive approaches are designed to identify these local changes in spatial effect.

We divide the map into four regions as shown in Figure 5.6, and summarize the overall MSE and the MSE in the each of the for regions in Table 5.1. The overall MSE from the adaptive smoothing model with GMRF2 was calculated to be 0.0347 which is 20% smaller than the MSE from the LOESS method. In region A and region B, the adaptive smoothing yielded MSEs that are around 20% smaller than the LOESS method. In region C and region D, the performances of the two approaches were similar. The GMRF1 model also out performs the LOESS method, indicating that adaptive smoothing is effective for estimating spatially-varying survival effects, even when links between natural barriers are not removed in the estimation procedure.

GMRF1 and GMRF2 also showed similar performances in region C and region D. However, in region A and region B, the MSE of GMRF2 is around 5% smaller than that of GMRF1. Thus, even though the two areas are spatially close, if we believe that the outcomes of the two areas are not correlated based on scientific grounds, deletion of the link between the two regions in the GMRF can provided improved fitting performance.

Table 5.1: MSE of the three methods (LOESS with span of 0.3 and adaptive smoothing with GMRF1 and adaptive smoothing with GMRF2) in the four regions as shown in Figure 5.6.

MSE	A	B	C	D	Overall
GAM (span=0.3)	0.0782	0.0769	0.0523	0.0423	0.0445
GMRF1	0.0618	0.0695f	0.0548	0.0427	0.0357
GMRF2	0.0541	0.0634	0.0543	0.0437	0.0347
%change	27%	26%	-0.4%	-0.4%	22%

5.3 Application to Ovarian Cancer Study

In this section we apply the proposed adaptive approach to the true California ovarian cancer data (see Chapter 1 for further details of the study). Figure 5.2 shows the geographical distribution of the data. The data consist of survival times and locations of $N = 6,878$ patients in total. All patients were diagnosed with ovarian cancer at stage III/IV. Observed survival times and event indicators were obtained from hospital medical records. Age, race, insurance type, socioeconomic status score, FIGO stage of the tumor, tumor grade, tumor size, tumor histology, whether receiving treatment in a high-volume hospital (treating more than 20 cases pre year in average), and whether the treatment adhered to the NCCN guidelines were available on all subjects and are adjusted for in all analytic results.

The same grid and sets of blocks that were used in the simulation study in Section 5.2 to sample z and γ were also used for the actual data. We specified prior parameters of $a_0 = 2, b_0 = 2, a = 55, b = 1, c = 0.01, d = 7$ for the hyperpriors of α, ζ_1 and ζ_2. The MCMC was run for 10,000 iterations with a burn-in of 5,000 samples. The posterior mean of z was calculated as the estimate for the adjusted spatial effect. The colormap of the hazards ratio of the adjusted spatial effects compared to the median hazard marginalized over the full area is shown in Figure 5.7b. For comparison, we also fit a Cox proportional hazards additive model with a LOESS smoother. The span size was taken to be 0.25, and was chosen to minimize AIC. Figure 5.7a shows the colormap and the estimated hazards ratio of adjusted spatial effects compared to the median hazard over the are. The obvious difference between

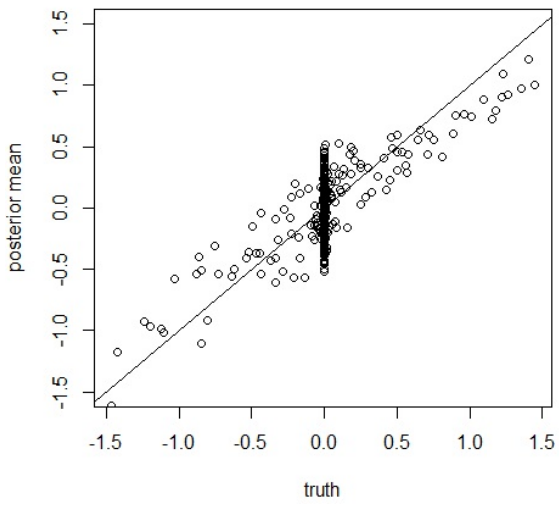
the estimated hazard ratio estimates using the two methods is the circled area in Figure 5.8, where the adaptive smoothing procedure results in higher estimates when compared to the LOESS smoother.

The significant areas suggested by the 95% probability intervals of the adaptive smoothing procedure and by the pointwise 95% confidence intervals of the Cox additive model are shown in Figure 5.9. The cluster in the middle area of southern California (where the estimates are also different) was highlighted as significantly different based on the adaptive smoothing method. This areas was not determined as significantly different when estimation was performed with the GAM.

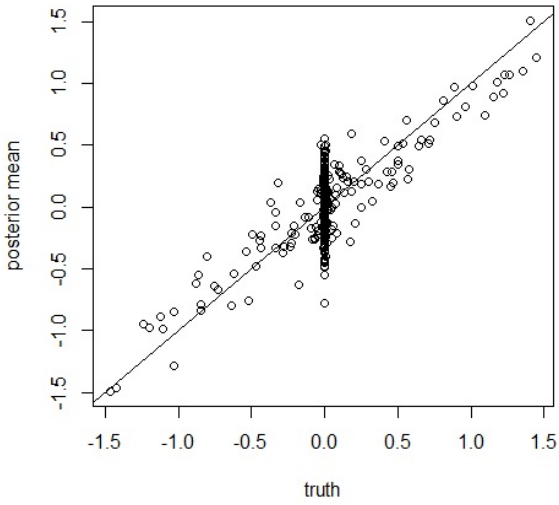
5.4 Discussion

In this chapter, we extended the adaptive smoothing methods presented in Chapter 4 to the setting of right-censored survival outcomes. The approach considered a parametric survival model, namely the Weibull model, where the scale parameter of the Weibull distribution is modeled using a linear predictor of adjustment covariates and a nonlinear function of spatial effect. A Bayesian hierarchical model was constructed corresponding to a thin-plate spline where the spatial effect is modeled via an IGMRF in the first hierarchy, and the spatially-varying variance component is modeled using an IGMRF in the second hierarchy. The same Bayesian computation method as discussed in Chapter refChapter4 proves to be efficient in the survival setting. The performance of the adaptive spatial smoothing model was evaluated by a simulated example. We also illustrated a promising property of the GMRF prior in epidemiologic studies, by noting that it is easy to remove spatially close points from the neighborhood one another if it is hypothesized that the spatial effect on the outcome is not correlated between the regions. Finally, we applied the method to data on ovarian cancer survival times in southern California. Comparisons with the more commonly

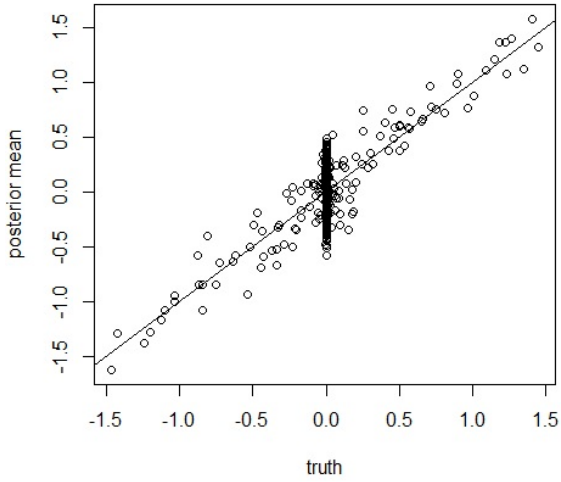
used LOESS smoother indicate that the proposed adaptive approach may be more sensitivity in identifying high risk geographic areas. Future work may include modeling the survival data using a mixtures of Weibull distributions for robust to model misspecification, and developing a tri-variate smoothing procedure to allow for estimation spatial effects that vary with time.



(a) GAM (span=0.3 as chosen by AIC).



(b) Adaptive smoothing with GMRF1.



(c) Adaptive smoothing with GMRF2.

Figure 5.5: Performance comparison (estimated adjusted spatial effect versus the true adjusted spatial effect) of the Cox additive model with LOESS as the smoothing method and the adaptive method proposed in this paper. (a) LOESS using a span size of 0.3 (chosen to minimize AIC); (b) Adaptive smoothing with an GMRF1; (c) Adaptive smoothing with GMRF2 (deleting links).

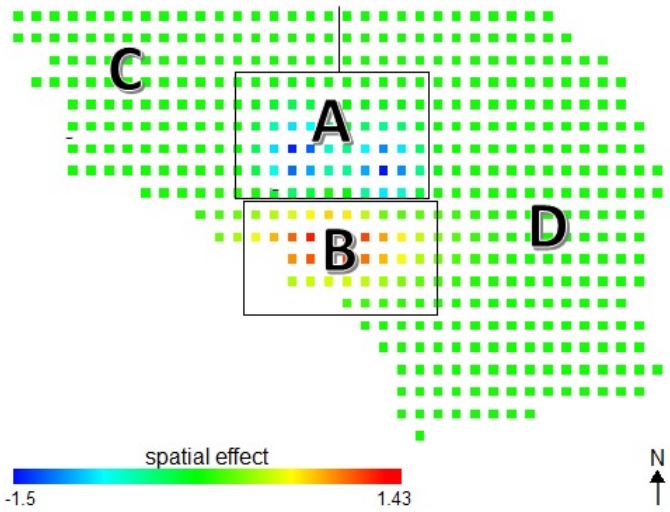
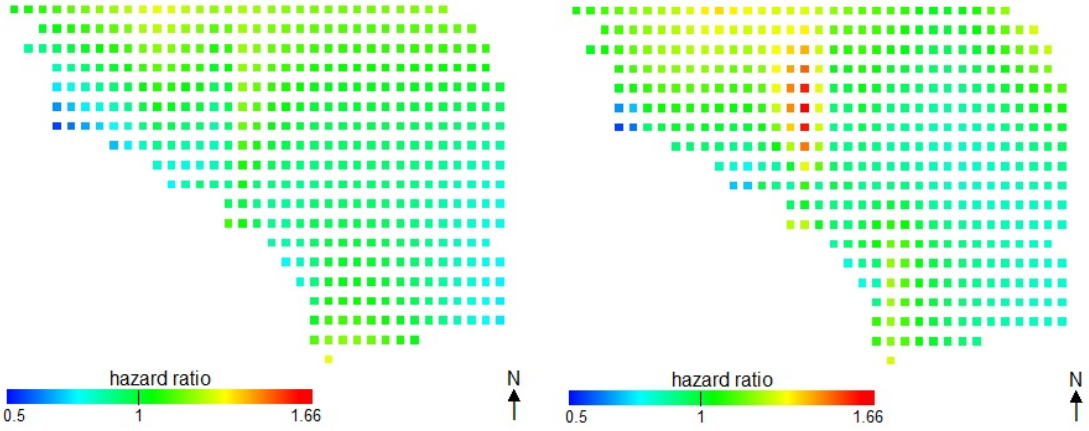


Figure 5.6: Depiction of the four regions for which the MSE from the proposed fitting approaches are compared.



(a) LOESS smoothing.

(b) Adaptive smoothing.

Figure 5.7: Estimated hazards ratio of adjusted spatial effects compared to median hazard using (a) a LOESS smoother and (b): the proposed adaptive smoothing method.

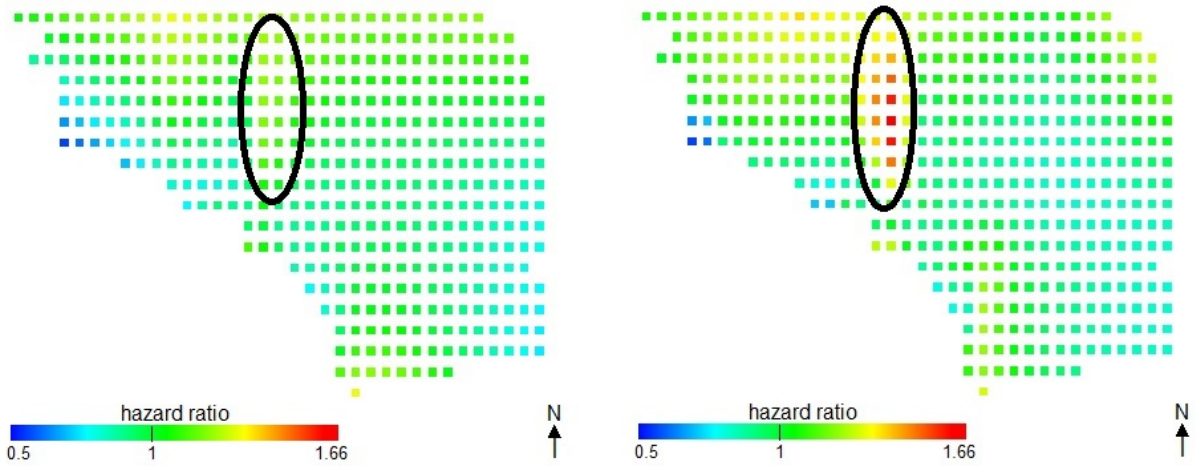


Figure 5.8: Circled areas indicate where the LOESS smoothing and adaptive smoothing methods show discrepant estimates.

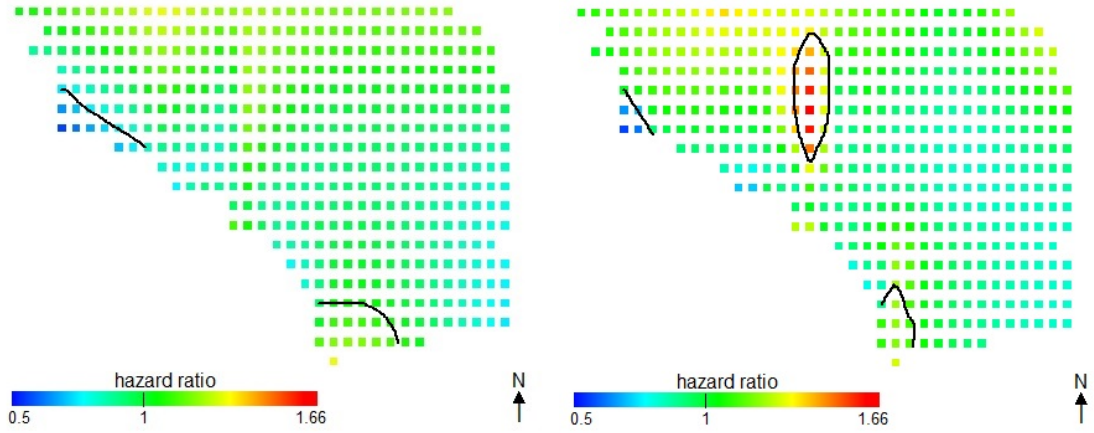
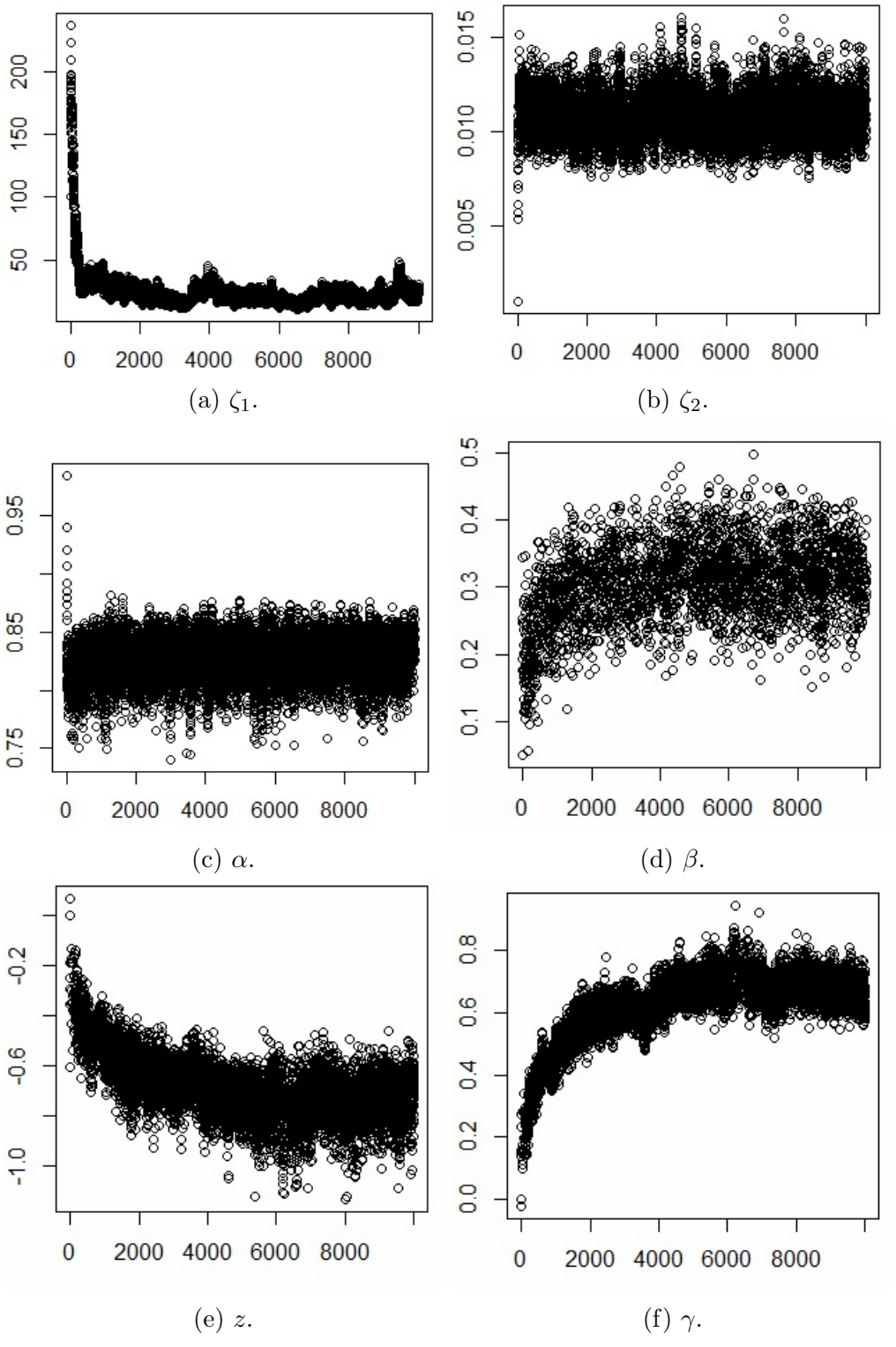


Figure 5.9: Estimated hazard ratio of adjusted spatial effects compared to the median hazard (marginalized over the full area) using the proposed adaptive method.



(a) ζ_1. (b) ζ_2.

(c) α. (d) β.

(e) z. (f) γ.

Figure 5.10: Example trace plots for the MCMC samples for the simulation study (based on $N = 6,878$ samples) to compare the performance of adaptive smoothing to that of the LOESS smoother.

Chapter 6

Summary and Future Work

This dissertation has focused on the development of statistical methodology for quantifying spatial effects on disease incidence using individual-level data. In Chapter 3 we implemented a proportional hazards additive model for right-censored survival data by including a bivariate LOESS smoother into Cox's proportional hazards model to estimate adjusted spatial effects on survival outcomes. An R package was developed to estimate and predict covariate-adjusted spatial effects using individual-level data. Generalized additive models for analyzing continuous, binary, count and censored survival outcomes are implemented in the package along with a full suite of visualization tools for mapping and interpreting spatial effects.

The bivariate LOESS smoother is a computational efficient analytic tool for population-based epidemiologic spatial analyses in which the population density varies over geolocations. However, it uses the same amount of smoothing across all geolocations, as is done with other commonly used smoothing methods. In Chapter 4, we developed an Bayesian thin-plate spline method to smooth continuous and discrete disease outcomes adaptively as the size and shape of the spatial effect changes over locations. The approach also accommodates adjustment for potential confounding factors. Based on a fine enough thin-plate grid, modeling the second order differences of spatial effect on each grid point with a Gaussian distribution results in an IGMRF prior for the spatial effect. We introduced adaptivity by allowing for

the variance component of the Gaussian distributions of the second order differences to vary across geolocations. The end result is that the amount of smoothing can be altered adaptively based on the shape of the spatial effect. We then specified the IGMRF prior for the variance component in a second hierarchy. We further extended our work to the analysis of censored survival outcomes in Chapter 5. One promising property of the IGMRF prior that was highlighted in Chapter 5 is that it the method makes it easy to systematically define the neighborhood for each point on a spatial map. For example, most smoothing methods define the neighborhood based on spatial distances between two points. However, if there is a natural barrier between two locations so that the disease outcome in one location does not have close correlation with a seemingly neighboring location (as measured by linear distance), then the two locations are not in the same "scientific" neighborhood of each other. It was shown that the IGMRF can easily account for this by deleting the links between areas that lie along such a natural barrier. In Chapter 5, illustrated the utility of link deletion using a simulation study inspired by the California ovarian cancer data presented in Chapter 1 of the thesis.

One limitation of our current work is that we build our model on a regular (equally spaced) thin-plate grid. If a researcher would like to increase the resolution and want more detail in particular geographic areas, then the current approach would require increasing the resolution of the whole map. This approach may be computationally infeasible in some settings. In future work we plan to construct the model on an irregular grid. Doing so would allow analysts to create different resolutions for different areas without the computational cost associated with increasing the resolution of the whole map.

Another potential limitation of the methods developed here is the reliance of a parametric (Weibull) survival distribution in Chapter 5. ,The area of Bayesian non-parametrics has grown rapidly following the work of Ferguson (1973) on the Dirichlet process (DP), a random probability measure on spaces of distribution functions. Such methods are well-suited for

survival spatial data analysis, and would enable flexible modeling for the unknown survival function, cumulative hazard function or hazard function. The use of a DP prior survival model would also allow one to incorporation prior information into parameter estimation. In the future, we plan to extend our approach in Chapter 5 to allow for non-parametric survival regression, leading to potential greater robustness against mode mis-specification.

Finally, we plan to extend our proposed method to the setting of adaptive spatial-temporal analyses by creating a grid in three dimensions so that disease incidence can be smoothed spatially and temporally simultaneously. The end result would allow for assessing potential time-varying effects of geospatial location on disease incidence.

Bibliography

Aha, D. (1997). Special ai review issue on lazy learning. *Artificial Intelligence Review*, 11.

Akullian, A., Kohler, P., Kinuthia, J., Laserson, K., Mills, L. A., Okanda, J., Olilo, G., Ombok, M., Odhiambo, F., Rao, D., Wakefield, J., and John-Stewart, G. (2014). Geographic distribution of hiv stigma among women of childbearing age in rural kenya. *AIDS*, 28:1665–1672.

Andersen, P. and Gill, R. (1982). Cox's regression model for counting process: a large sample study. *The Annals of Statistics*, 10:1100–1120.

Assael, B. M., Castellani, C., Ocampo, M. B., Iansa, P., Callegaro, A., and Valssecchi, M. G. (2002). Epideomology and survival analysis of cystic fibrosis in an area of intense neonatal screening over 30 years. *American Journal of Epidemiology*, 156:397–401.

Assuncao, R., Potter, J., and Cavenaghi, S. (2002). A baysian space varying parameter model applied to estimating fertility schedules. *Statsitics in Medicine*, 21:2057–2075.

Baker, S., Holt, K. E., Clements, A. C., Karkey, A., Arjyal, A., Boni, M. F., Dongol, S., Hammond, N., Koirala, S., Duy, P. T., Nga, T. V. T., Campbell, J. I., Dolecek, C., Basnyat, B., Dougan, G., and Farrar, J. J. (2011). Combined high-resolution genotyping and geospatial analysis reveals modes of endemic urban typhoid fever transmission. *Open biology*, 1(2):110008.

Baladandayuthapani, V., M. B. K. and Carroll, R. J. (2005). Spatially adaptive bayesian penalized regression splines (p-splines). *Statistics in Medicine*, 26:5047–5080.

Banerjee, S., Carlin, B., and Gelfand, A. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman and Hall, Florida, U.S.A.

Besag, J., York, J., and Mollie, A. (1991). Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43:1–21.

Bramball, S., Allum, W., Jones, A., Allwood, A., Cummins, C., and Neoptolemos, J. (1995). Treatment and survival in 13560 patients with pancreatic cancer, and incidence of the disease, in the west midland: A epidemiology study. *British Journal of Surgery*, 82:111–115.

Breiman, L. and Freidman, J. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of American Statistical Association*, 80:580–597.

Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics*, 30:89–99.

Breslow, N. E. and Clayton, D. G. (1993). Approximate inference in generalized linear mixed models.

Brezger, A., F. L. and Hennerfeind, A. (2007). Adaptive gaussian markov random fields with applications in human brain mapping. *Drug Information Journal*, 35:1095–1112.

Bristow, R. E., Chang, J., Ziogas, A., Anton-Culver, H., and Vieira, Veronica, M. (2014). Spatial analysis of adherence to treatment guidelines for advanced-stage ovarian cancer and the impact of race and socioeconomic status. *Gynecologic Oncology*, 134:60–67.

Cater, C. and Kohn, R. (1996). Markov chain monte carlo in conditionally gaussian state space models. *Biometrika*, 83:589–601.

Chambers, J. and Hastie, T. J. (1992). *Statistical models in S*. Chapman and Hall/CRC.

Clark, R. (1977). Non-parametric estimation of a smooth regression function. *Journal of Royal Statistical Society. Series B*, 39:107–113.

Clayton, D. and Kaldor, J. (1987). Empirical bayes estimates of age-standardized relative risks for use in disease mapping. *Biometrics*, 43:671–681.

Cleveland, W., Devlin, S., and Grosse, E. (1988). Regression by local fitting: methods, properties and computational algorithms. *Journal of Econometrics*, 37:87–114.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74:829–836.

Cleveland, W. S. (1981). Lowess: A program for smoothing scatterplots by robust locally weighted regression. *The American Statistician*, 35:54.

Cleveland, W. S. and Devlin, S. J. (1988). Locally-weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83:596610.

Cox, D. (1972). Regression models and life-tables. *Journal of Royal Statistical Society. Series B (Methodological)*, 34(2):187–220.

Cox, D. (1975). Partial likelihood. *Biometrika*, 62(2):269–276.

Crainiceanu, C., R. D. C. R. A. J. and Goodner, B. (2007). Spatially adaptive penalized splines with heteroscedastic errors. *Communications in Statistics, Part A- Theory and Methods*, 13:2315–2338.

Cressie, N. (1993). *Statistics for spatial data*. Wiley, New York.

De Boor, C. (1978). *A practical guide to splines*. Springer-Verlag.

Devlin, S. (1988). Locally-weighted multiple regression: statistical properties and a test of linearity. *Bell Communications Research technical memorandum*.

Diggle, P., Moyeed, R., and Tawn, J. (1998). Model-based geostatistics. *Applied Statistics*.

Duchon, J. (1977). *Splines minimizing rotation-invariant semi-norms in Solobev spaces in Solobev spaces in construction theory of functions of several variabels*. Springer, Berlin.

Efron, B. (1977). The efficiency of cox's likelihood function for censored data. *Journal of the Americian Statistical Association*, 72:557–565.

Elliott, P. and Wartenberg, D. (2004). Spatial epidemioloty: Current approaches and future challenges. *Environmental Health Perspectives*, 112:998–1106.

Ezekiel, M. (1941). *Methods of correlation analysis*. J. Wiley and Sons, New York.

Friedman, J. and Stuetzle, W. (1984). Smoothing of scatterplots. *Tech. Rept. Orion 3, Department of Statistics, Stanford University*.

Gamerman, D., Moreira, A., and Rue, H. (2003). Space-varying regression models: specifications and simulations. *Computational Statsitics and Data Analysis*, 42:513–533.

Gilks, W. and Wild, P. (1992). Adaptive rejection sampling for gibbs sampling. *Applied Statistics*, 41:337–348.

Green, P. and Silverman, B. (1994). *Nonparametric regression and generalized linear models*. CPC Press.

Greenland, S. (1992). Divergent biases in ecologic and individual-level studies. *Statistics in Medicine*.

Hardle, W. and Scott, D. (1992). Smoothing by weighted averaging of rounded points. *Computational Statistics*, 7:97–128.

Hastie, T. (2004). *gam: Generalized Additive Models*. R package version 1.12.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, 1:297–310.

Henderson, R., Shimakura, S., and Gorst, D. (2002). Modeling spaital variation in leukemia survival data. *Journal of the American Statistical Association*, 97:965 – 975.

Ho, K. K., Pinsky, J. L., Kannel, W. B., Levy, D., and Pitt, B. (1993). The epidemiology of heart failure: the framingham study. *Jounal of American Coll Cardiol*, 22:A6–A13.

Hoffman, K., Aschengrau, A., Webster, T. F., Bartell, S. M., and Vieira, V. M. (2015). Associations between residence at birth and mental health disorders: A spatial analysis of retrospective cohort data. *BMC Public Health*, 15(688).

Hurvich, C., Simonoff, J., and Tsai, C. (1998). Smoothing parameter selction in nonparametric regression using an improved akaike informaiton criterion. *Journal of the Royal Statistical Society B*, 60:9–31.

Hutchinson, M. (1995). Interpolating mean rainfall using thin plate smoothing splines. *International Journal of Geographical Information Systems*, 9.

Ickovics, J. R., Hamburger, M. E., Vlahov, D., Schoenbaum, E. E., Schuman, P., and Moore, J. (2001). Mortality, cd4 cell count decline, and the depressive symptoms among hiv-seropositive women longitudinal analysis from the hiv epidemiology research study. *JAMA*, 285:1466–1474.

Kalbfleisch, J. and Prentice, R. (1973). Marginal likelihoods based on cox's regression and life model. *Biometrika*, 60:267–278.

Kelsall, J. E. and Diggle, P. J. (1998). Spatial variation in risk of disease: A nonparametric binary regression approach. *Applied Statistics*, pages 559–573.

Kneib, T. (2006). *Mixed model based inference in structured additive regression*. Ph.D. thesis, LMU Munchen, Faculty of Mathematics, Computer Science and Statistics.

Koch, T. (2005). *Cartographies of disease*. ESRI Press.

Krivobokova, T., C. C. M. and Kauermann, G. (2008). fast adaptive penalized splines. *Biometrics*, 59:770–777.

Lancaster, P. and Salkauskas, K. (1986). *Curve and surface fitting*. Elsevier Science and Technology Books.

Lanczos, C. (1950). An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *Journal of Research of the National Institute of Standards and Technology*, 45:255–282.

Lang, S., Fronk, E., and Fahrmeir, L. (2002). Function estimation with locally adaptive dynamic models. *Computational Statistics*, 17:479–499.

Lu, H., Reilly, C. S., Banerjee, S., and Carlin, B. P. (2007). Bayesian areal wombling via adjacency modeling. *Environmental Ecology Statistics*.

McCullagh, P. and Nelder, J. (1983). *Generalized linear models*. Chapman and Hall, London, New York.

Nadaraya, E. (1964). On estimating regression. *Theor. Prob. Appl.*, 9:141–142.

Nelder, J. and Wedderburn, R. (1972). Generalized linear models. *Journal of the Royal statistical Society. Series A (General)*, 135:370–384.

Pardo-Crespo, M. R., Narla, N. P., R., W. A., Beebe, T. J., Sloan, J., Yawn, B. P., Wheeler, P. H., and Juhn, Y. J. (2013). Comparison of individual-level versus area-level socioeconomic measures in assessing health outcomes of children in olmsted county, minnesota. *Journal of Epidemiology Community Health*.

Peto, R. (1972). Contribution to the discussion of 'regression modes and life-tables' by dr cox. *Journal of the Royal Statistical Society, Series B*, 34:205–207.

Polissar, L. (1980). The effect of migration on comparison of disease rates in geographic studies in the united states. *American Journal of Epidemiology*.

Priestley, M. and Chao, M. (1972). Non-parametric function fitting. *Journal of Royal Statistical Society. Series B*, 34:385–392.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Reinsch, C. (1967). Smoothing by spline functions. *Numeraical Mathematics*, 10:177–183.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape,(with discussion). *Applied Statistics*, 54:507–554.

Rodriguez, P. P. and Komarek, A. (2014). *Adaptive Rejection Sampling*. R package version 3.1.2.

Rue, H. and Held, L. (2005). *Gaussian markov random fields: theory and applications*. Chapman and Hall/CRC.

Ruppert, D. and Carroll, R. J. (2000). Spatially-adaptive penalties for spline fitting. Technical report, Department of Biostatistics, Harvard School of Public Health.

Scott, D. (2003). *Multivariate Density Estimation: Theory, Practice, and Visualization*. Wiley.

Siegel, R., Ma, J., Zou, Z., and Jemal, A. (2014). Cancer statistics 2014. *CA Cancer Journal for Clinicians*, 64(1):9–29.

Smith, G., Neaton, J., Wentworth, D., Stamler, R., and Stamler, J. (1996). Socioeconomic differentials in mortality risk among men screened for the multiple risk factor intervention trial. *American Journal of Public Health*, 86:486–496.

Speed, T. and Kiiveri, H. (1986). Gaussian markov distributions over finite graphs. *The Annals of Statistics*, 14:138–150.

Stasinopoulos, D. M. and Rigby, R. A. (2007). Generalized additive models for location scale and shape (gamlss) in r. *Journal of Statistical Software*.

Stasinopoulos, M., Rigby, B., and Mortan, N. (2015). *gamlss.cens: Fitting an interval response variable using gamlss.family distributions*. R package version 4.3.1.

Stein, M. (1999). *Interpolation of spatial data: some theory for kriging*. Springer.

Steliarova-Foucher, E., Stiller, C., Kaatsch, P., Berrino, F., Coebergh, J.-W., Lacour, B., and Perkin, M. (2004). Geographical patterns and time trends of cancer incidence and survival among children and adolescents in europe since the 1970s: an epidemiological study. *The Lancet*, 364:11–17.

Stevens, K. B. and Pfeiffer, D. U. (2011). Spatial modelling of disease using data- and knowledge-driven approaches. *Spatial and Spatio-temporal Epidemiology*, 2:125–133.

Tibshirani, R. and Hastie, T. (1987). Local likelihood estimation. *Journal of American Statistical Association*, 82:559–568.

Van Der Linde, A. (2003). Pca-based dimension reduction for splines. *Journal of Nonparametric Statistics*, 15:77–92.

Vieira, V. M., Webster, T. F., Weinberg, J. M., and Aschengrau, A. (2008). Spatial-temporal analysis of breast cancer in upper cape cod, massachusetts. *International Journal of the Health Geographics*, 7(46).

W., Y. T. (2007). *VGAM: Vector Generalized Linear and Additive Models*. VGAM: Vector Generalized Linear and Additive Models.

Wahba, G. (1990). *Spline models for observational data*. SIAM, Philadelphia.

Waller, L. A. and Gotway, C. A. (2004). *Applied spatial statistics for public health data*. John Wiley and Sons.

Webster, T., Vieira, V., Weinberg, J., and Aschengrau, A. (2006). Method for mapping population-based case-control studies: An application using generalized additive models. *International Journal of Health Geographics*, 5(26).

Williams, C. (1998). Prediction with gaussian processes: from linear regression to linear prediction and beyond. *Learning in Graphical Models*, 89:599–621.

Wood, S. (2009). *Mixed Gam Computation Vehicle with GCV/AIC/REML Smoothness Estimation*.

Wood, S. N. (2006). *Generalized additive models*. Chapman and Hall: London.

Yue, Y. and L., S. P. (2010). Nonstationary spatial gaussian markov random fields. *Journal of Computational and Graphical Statistics*, 19:96–116.

Yue, Y., Loh, J. M., and Lindquist, M. A. (2010). Adaptive spatial smoothing of fmri images. *Statistics And Its Interface*, 3:3–13.