Adaptive Bayesian Nonparametric Smoothing with Markov Random Fields and Shrinkage Priors

James Robert Faulkner

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2019

Reading Committee:

Vladimir N. Minin, Chair

Peter Guttorp

Trevor Branch

Program Authorized to Offer Degree:
Quantitative Ecology and Resource Management

University of Washington

Abstract

Adaptive Bayesian Nonparametric Smoothing with Markov Random Fields and Shrinkage Priors

James Robert Faulkner

Chair of the Supervisory Committee:
Professor Vladimir N. Minin
Department of Statistics

The need to estimate unknown functions or surfaces arises in many disciplines in science and there are many statistical methods available to do this. Our interest lies in using Bayesian nonparametric approaches to estimate unknown functions. One such approach to nonparametric estimation is based on the Gaussian Markov random field priors. This class of computationally efficient and flexible methods is widely used in applications. There is frequently the need to estimate functions with change points, discontinuities, or abrupt changes, or functions with varying levels of smoothness. Gaussian Markov random fields have limited ability to accurately capture such features. We develop a locally adaptive version of Markov random fields that uses shrinkage priors on the order-k increments of the discretized function and has the flexibility to accommodate a large class of functional behaviors. We show that the horseshoe prior results in superior performance in comparison to other shrinkage priors. The horseshoe prior induces sparsity in the increments, which provides good smoothing properties, and at the same time the heavy tails of the prior allow for jumps and discontinuities in the field. We first apply the method to some standard settings where we use simulated data to compare to other methods and then apply the models to two benchmark data examples frequently used to test nonparametric methods. We use Hamiltonian Monte Carlo to approximate the posterior distribution of model parameters because this method provides superior performance in the presence of the high dimensionality and strong parameter correlations exhibited by our models. We then extend the method to the estimation of effective population sizes

using the coalescent process and genetic sequence data. For that application, we develop a custom Markov chain Monte Carlo sampler based on a combination of elliptical slice sampling and Gibbs sampling. We test the method using simulated data and then use it to reconstruct past changes in genetic diversity of human hepatitis C virus in Egypt and to estimate population size changes of ancient and modern steppe bison. Finally, we extend the method for use in the spatial setting, where we apply the method to disease mapping and to the estimation of the intensity of an inhomogeneous spatial point process. Overall, we find that this method is flexible enough to accommodate a variety of data generating models and offers the adaptive properties and computational tractability that make it a useful addition to the Bayesian nonparametric toolbox.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to start by thanking the National Oceanic and Atmospheric Administration and my superiors at the Northwest Fisheries Science Center. NOAA provided funding for me to attend classes in my first two years of study through the NOAA Advanced Studies Program, and allowed me to continue to work at my position during my participation in the program. Director John Ferguson opened the door for me to go back to school and my supervisor at the time, Rich Zabel, repeatedly encouraged me to pursue the opportunity (until I finally did), and Mark Scheuerell, my current supervisor, supported me through the process once I finally took it on. I am very thankful to them and others at NOAA who helped make this dream a reality for me.

Next I want to express my deep gratitude to Vladimir Minin for agreeing to be my adviser and for his tireless support over the years. Not only is he an excellent statistician, mathematician, and scientist, but he is also an outstanding teacher, a resourceful, patient, and supporting mentor, a great conversationalist, and has a wicked sense of humor. He provides a high-quality research environment for his students and sets a standard of world-class excellence while being a gentleman and a statesman. I could not have asked for a better adviser and I am proud to be able to say that I was part of his research group. I look forward to our future collaborations.

I want to thank the QERM program for accepting me as a student and for providing a top quality academic experience. Loveday Conquest and Tim Essington provided thoughtful guidance to me and excellent leadership for the program. Joanne Besch and Erica Owens worked hard behind the scenes to make the program work and I thank them for all of the advice they provided to me. I thank Peter Guttorp, Jim Anderson, and Mark Kot for some great discussions and positive encouragement. I especially want to thank all of the great QERM students that I had the opportunity to interact with during my time in the program. In particular, I want to recognize my fellow cohort:

Brooke Davis, Austin Phillips, and Kai Ross. They are one tough and bright group of people and I am thankful to have shared the experience of the first year with them.

I thank my Committee members Peter Guttorp, Trevor Branch, Mathias Drton, and Soumik Pal for agreeing to be on my committee and for the valuable feedback and discussion they provided.

I also want to thank my fellow students in the Minin research group that I got to interact with over the years. Amrit Dhar, Jon Fintzi, Jan Irvahn, Micheal Karcher, Amanda Koepcke, Jane Lange, Andy Magee, Mingwei Tang, and Jason Xu are all incredibly smart and talented people and they all held a high standard in their research and in their contributions to the group. I will always be thankful for the experiences we shared. Among them, I especially want to thank Andy Magee for all the programming and testing he did to get our models to work in `RevBayes`.

I also want to thank the members of the Space-Time Reading Group at UW for a shared enthusiasm for spatial statistics and for going out of their way to participate and keep the group going. I thank Peter Guttorp and Paul Sampson for their early efforts to create the group and for their involvement over the years. I especially want to thank the dedicated student members like Serge Aleshin-Guendel, John Best, Hannah Director, Peter Gao, Johnny Paige, Max Schneider, and Aaron Zimmerman for all of their great presentations and insightful discussions.

Most importantly of all, I would like to thank my wife Yuri and our two wonderful children, Nikolas and Maya, for their patience, love, and understanding over my long years of study. They are my motivation. Without them, I could not have completed this degree.

DEDICATION

To my dear family: Yuri, Nikolas, and Maya

Chapter 1

INTRODUCTION

There are many situations in science when a researcher has data that were generated from a process that they do not fully understand but they need to estimate the shape of the underlying function, trend, or surface representing that process. This is often a necessary step in initial discovery that will lead to the construction of more complex models that provide a better understanding of the process and the forces that drive it. Estimation can also be the goal in itself, for a primary process or for a secondary process that needs to be estimated so that higher-level processes can be understood. Frequently the processes of interest exhibit smooth or gradual changes over space or time or some set of explanatory variables and do not exhibit break points or rapid changes. Simple linear associations between variables are commonly assumed, or nonlinear parametric models are used to describe the true underlying processes. Many scientific disciplines assume the processes of interest can be described by differential equations, and the resulting solutions are nonlinear and continuous.

However, there are also many situations where the underlying processes of interest exhibit rapid changes, breaks, discontinuities, or complex surfaces that are not well described by simple parametric functions. In ecological or biological systems, gradual changes in an underlying process can cause disproportionate, nonlinear responses in the overlying process when threshold levels of the underlying process are reached (Scheffer et al., 2001). As an example, coral bleaching can occur when water temperatures exceed thresholds that are stressful or lethal to the symbiotic algae that live with the corals (Brown, 1997). Coral bleaching can lead to widespread mortality in corals and the organisms that depend on them, resulting in abrupt changes in population sizes. In the study of natural populations, other sudden changes in population size could be brought on by rapid extinction events, outbreaks of deadly diseases, abrupt changes in habitat, or rapid expansion due to colonization. In medical image analysis, one goal is to find specific areas of unusual activity, such as areas of heightened brain activity in neural imaging (Brezger et al., 2007). These areas can be distinctly defined and arise as abrupt changes relative to adjacent areas in the image. In epidemiology, there is interest in mapping the incidence of particular diseases within a country or across political boundaries. There can be stark differences in disease incidence across some boundaries and very little difference across other boundaries (Knorr-Held and Raßer, 2000). In

infectious disease dynamics, rapid changes in the number of infected individuals could be caused by sudden changes in contact rates due to behavioral changes or quarantine, or sudden changes in the infection rate due to introduction of treatment or vaccine.

Flexible, nonparametric statistical methods like smoothing splines, kernel smoothing, wavelets, or Gaussian processes are frequently used to estimate functions of unknown form. Despite the misleading implication of their name, nonparametric methods are typically high dimensional or highly parameterized. This increased dimensionality is what allows for the flexibility to fit to a wide range of functional forms. In practice, methods like smoothing splines and wavelets are types of convolution methods where a set of basis functions are convolved with set of random weights to produce a smoothed curve or surface. The shapes of the basis function and the distribution of the weights will determine the properties of the resulting curves. Kernel smoothers use probability density kernels centered at each data point location to get local values of weighted averages of all the data points. These are types of convolution methods where the kernels are convolved with the data. Gaussian processes are distributions of random functions, where the functions are generated by high-dimensional (infinite-dimensional in theory) multivariate normal distributions and their shapes are dependent on the covariance function. Gaussian processes can also be represented as convolutions (Higdon, 1998).

Many of these standard methods have been generalized to allow for more flexibility to adapt to rapid changes in a function or varying levels of smoothness. Some examples include adaptive smoothing splines (Mammen et al., 1997), adaptive kernel methods (Terrell et al., 1992), trend filtering (Kim et al., 2009; Tibshirani, 2014), adaptive Gaussian Markov random fields (Brezger et al., 2007; Yue and Speckman, 2010) and nonstationary Gaussian processes (Sampson and Guttorp, 1992; Higdon, 1998; Paciorek and Schervish, 2004). The local adaptivity of many of these methods is achieved by allowing some component of the model that was otherwise constant to vary over the domain of the function of interest. For example, variance or shape of kernels, shape or coverage of basis functions, or distributions of random weights may be allowed to vary. In some types of adaptive Gaussian Markov random fields, local variances are introduced. In nonstationary Gaussian processes, properties of the covariance function are allowed to vary over the domain.

Trend filtering is different in that it is a type of penalized spline that has a penalty that induces sparsity and allows it to capture abrupt changes. Although these approaches can adapt to a wide variety of functional forms, most do not have the ability to simultaneously estimate sharp break points and smooth over the remaining function. Two exceptions are trend filtering and some adaptive Gaussian Markov random fields. We will draw on the properties of both of these methods in what follows.

This work was motivated by the need for a curve fitting method that could adapt to local changes in smoothness of a function, including abrupt changes or discontinuities, and could be used with a variety of types of observational data and associated likelihoods. One of our main goals was to have a method that could be incorporated into multi-level Bayesian models with complex data generating processes. Specifically, we needed the method to work with models that jointly estimate effective population size changes and the nucleotide substitution process from genetic sequence data. We also wanted the method to be fully Bayesian so that we could characterize the uncertainty in the estimated parameters. Finally, we wanted the method to be relatively simple to implement and computationally efficient. This work describes the development of such a method, the development of methods for posterior inference for the method, assessment of the performance of the method and comparison to that of other state-of-the-art methods in current use, and the application of the method to several empirical examples from different disciplines.

In Chapter 2 we provide some background information on some technical topics that are referenced frequently throughout the remaining document. There we describe Gaussian Markov random fields and how they are applied to data. We also provide some background on Gaussian processes, which have a relationship to Gaussian Markov random fields and are very popular in Bayesian nonparametrics. Finally, we provide an introduction to Hamiltonian Monte Carlo. This is an efficient Markov chain Monte Carlo technique for posterior inference that we draw upon frequently in the remaining chapters.

Chapter 3 is a core chapter that describes the main research contributions in this document. Here we describe a new Bayesian nonparametric smoothing method that allows for estimation of unknown functions with sharp features such as change points or with varying levels of smoothness.

This is achieved by using shrinkage priors as the distributions on the independent increments of Gaussian Markov random field priors. The type of shrinkage prior used will affect the amount of smoothing and the ability to capture sharp features. We find that the combination of high probability mass near zero and long tails offered by the horseshoe distribution provides the best performance in terms of bias and precision when compared to other shrinkage priors using simulated data. We use Hamiltonian Monte Carlo for posterior inference for these models and develop an R package for fitting the models. We also apply the models to two benchmark data examples frequently used to test nonparametric methods.

Chapter 4 builds upon the methods introduced in Chapter 3 by extending the methods for use in estimating effective population size trajectories from genetic sequence data and gene genealogies. Nonparametric methods are important for estimating effective population size trajectories and there is a need for more flexible methods that can capture features like bottlenecks and rapid changes in population size. We use the horseshoe Markov random field method developed in Chapter 3 to fill this need. We also develop a novel method for posterior inference based on a combination of elliptical slice sampling and Gibbs sampling. We test our model using simulated data and then use it to reconstruct past changes in genetic diversity of human hepatitis C virus in Egypt and to estimate population size changes of ancient and modern steppe bison.

In Chapter 5, we further extend the model from Chapter 3 to the spatial setting. Various nonparametric methods have been developed for fitting spatial surfaces that have discontinuities or abrupt changes. We develop a fully Bayesian nonparametric method for spatial smoothing based on the horseshoe Markov random field that allows for adaptation to abrupt changes while maintaining good smoothing properties. We propose an efficient method for posterior inference that takes advantage of the sparsity in the precision matrices and uses Hamiltonian Monte Carlo. We use simulations to compare performance of the model to other methods and then we apply our method to disease mapping and to estimating the intensity of an inhomogeneous spatial point process.

Chapter 6 contains some concluding discussion about the work presented and offers possible avenues of future research. There were many discoveries made during the course of conducting the research presented here that could not be fully investigated or pursued. These include model exten-

sions, possible customized methods for marginal likelihood estimation, and alternative computing methods.

Chapter 2

TECHNICAL BACKGROUND OF SELECTED TOPICS

This chapter provides an introduction to a few topics that are important to understanding ideas that are developed in later chapters. Gaussian Markov random fields are used extensively in my research as a foundation for the nonparametric smoothing techniques that are discussed in Chapters 3, 4, and 5. In general, Gaussian Markov random fields are approximations to Gaussian processes, which are commonly used as prior distributions to estimate unknown functions non-parametrically in a Bayesian framework. I introduce Gaussian processes and cover some topics on approximation and on computational issues. Finally, I cover some of the background on Hamiltonian Monte Carlo, which is used for posterior inference in Chapters 3, 4, and 5.

2.1 Gaussian Markov Random Fields

Gaussian Markov random fields (GMRFs) are used in many disciplines as a nonparametric smoothing method. They are finite dimensional approximations to Gaussian processes (GPs), and this dimension reduction often results in computational benefits. Here I touch on some of the main topics related to GMRFs. Note that the information presented in this section was drawn from Brémaud (1999), Rue and Held (2005), and the lecture notes of Vladimir Minin.

2.1.1 Conditional Independence and Markov Random Fields

Two random variables x and y are *independent* if and only if their joint distribution is the product of their marginal distributions: $p(x, y) = p(x)p(y)$. Another way to express the independence of x and y is: $x \perp y$. Two random variables x and y are *conditionally independent* given random variable z if and only if $p(x, y \mid z) = p(x \mid z)p(y \mid z)$. Alternatively, this can be written: $x \perp y \mid z$. Conditional independence of x and y given z means that given a known fixed value of z, knowledge of y provides no new information about x.

For a discrete-time process $\{X_k\}_{k=0}^n$, the Markov property can be defined as

$$\Pr(X_i \mid X_{i-1}, X_{i-2}, \ldots, X_0) = \Pr(X_i \mid X_{i-1}).$$

This says that the probability distribution of the current state only depends on the previous state

and not the states before. An alternative way to write this is:

$$\Pr(X_i \mid X_0, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n) = \Pr(X_i \mid X_{i-1}, X_{i+1}).$$

With this representation in mind, if we think of i as a site and sites $i-1$ and $i+1$ as neighbors of i, then we can say that given the values of X_{i-1} and X_{i+1}, the value of X_i is independent of the values at the non-neighbors of i.

We can generalize this property to apply to a graph in order to define a Markov random field. A process $\{X_s\}_{s\in S}$ with $X_s \in \Lambda$ is a *random field* over Λ, where S is a finite set of sites or locations, and Λ is a finite set of phases. Suppose the sites s are connected with a labeled and undirected graph $\mathcal{G} = (S, \mathcal{E})$, where $\mathcal{E}$ is the set of edges and S is the set of sites (vertices or nodes). Let $(s_1, s_2) \in \mathcal{E}$ mean that sites s_1 and s_2 are connected by an edge. For a subset $A \in S$, let $\mathbf{X}_A$ be the set of all X_s where $s \in A$, and let $\mathbf{X}_{-A}$ be the set of all X_s where $s \notin A$. Now we can define N_s to be a *neighborhood* of s if $s \notin N_s$ and for $t \neq s$, N_s is the set of all t for which $(t, s) \in \mathcal{E}$. Now we are ready to define a Markov random field. A random field X is a *Markov random field* with respect to the graph $\mathcal{G}$ if

$$\Pr(X_s = x_s \mid \mathbf{X}_{-s} = x_{-s}) = \Pr(X_s = x_s \mid \mathbf{X}_{N_s} = \mathbf{x}_{N_s}) \tag{2.1}$$

for all $s \in S$ and $\mathbf{x} \in \Lambda^{|S|}$. This says that the distribution of X_s only depends on its neighbors, or alternatively, conditional on its neighbors, X_s is independent of the values of the field at all sites that are not its neighbors.

2.1.2 Multivariate Normal Distribution

The multivariate normal or multivariate Gaussian distribution is central to the concept of a GMRF. Consider the random variable $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$, where $n < \infty$ and $\boldsymbol{x} \in \mathbb{R}^n$. We say that $\boldsymbol{x}$ has a multivariate normal distribution with $n \times 1$ mean vector $\boldsymbol{\mu}$ and $n \times n$ symmetric positive definite covariance matrix $\boldsymbol{\Sigma}$ if it has the following probability density function:

$$p(\boldsymbol{x}) = (2\pi)^{-n/2}|\boldsymbol{\Sigma}|^{-1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right). \tag{2.2}$$

Here $|\mathbf{\Sigma}|$ is the matrix determinant of $\mathbf{\Sigma}$. The marginal distributions of the x_i are also normal, with $x_i \sim \mathcal{N}(\mu_i, \Sigma_{ii})$. That is $\mathrm{E}[x_i] = \mu_i$ and $\mathrm{Var}[x_i] = \Sigma_{ii}$. The covariance between x_i and x_j is $\mathrm{Cov}[x_i, x_j] = \Sigma_{ij}$. Sometimes, as we will see in the next section, it is more convenient to work the precision matrix, $\mathbf{Q} = \mathbf{\Sigma}^{-1}$, rather than with the covariance matrix.

2.1.3 GMRF Definition

In brief, a GMRF is a multivariate normal distribution with a precision matrix which is defined by the neighborhood structure and conditional independencies among sites. More specifically, for $i \neq j$, $x_i \perp x_j \mid \boldsymbol{x}_{-ij} \iff Q_{ij} = 0$, which says the precision is zero among conditionally independent sites. This means non-zero values in $\mathbf{Q}$ indicate sites that are adjacent or share an edge in $\mathcal{G}$ and zero values indicate sites that are not neighbors. Therefore, parameterization of $\mathbf{Q}$ can be constructed from knowledge of $\mathcal{G}$ as long as $\mathbf{Q}$ is restricted to be positive definite. It follows that $\mathbf{Q}$ is completely dense (all values non-zero) when the elements of $\mathcal{G}$ are completely connected.

Now we formally define a GMRF. A random vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T \in \mathbb{R}^n$ is a GMRF with respect to graph $\mathcal{G} = (S, \mathcal{E})$ with mean $\boldsymbol{\mu}$ and precision matrix $\mathbf{Q}$ if and only if $Q_{ij} \neq 0 \iff i, j \in \mathcal{E}$ for all $i \neq j$ and the density for $\boldsymbol{x}$ has the form

$$p(\boldsymbol{x}) = (2\pi)^{-n/2} |\mathbf{Q}|^{1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \mathbf{Q} (\boldsymbol{x} - \boldsymbol{\mu})\right). \tag{2.3}$$

If $\boldsymbol{x}$ is a GMRF, then the following conditional properties hold:

$$\begin{aligned}
\mathrm{E}[x_i \mid \boldsymbol{x}_i] &= \mu_i - \frac{1}{Q_{ii}} \sum_{j:j\ i} Q_{ij}(x_j - \mu_j), \\
\mathrm{Prec}[x_i \mid \boldsymbol{x}_i] &= Q_{ii}, \\
\mathrm{Corr}[x_i, x_j \mid \boldsymbol{x}_{ij}] &= -\frac{Q_{ij}}{\sqrt{Q_{ii} Q_{jj}}}, \quad i \neq j.
\end{aligned}$$

Expressing a GMRF in terms of its conditional distributions can be useful for implementing Gibbs samplers for posterior inference.

2.1.4 Intrinsic GMRFs

Intrinsic GMRFs (IGMRF) are considered *improper* because their precision matrix is less than full rank. They are invariant to the addition of a trend that is a polynomial of the node locations where the of order of the polynomial is dependent on the rank of the precision matrix. This is useful when a underlying mean of the field is measured as a function of a set of fixed covariates, and so the GMRF essentially models the temporal residuals of the process.

To define an improper GMRF, first we let $\mathbf{Q}$ be an $n \times n$ symmetric and positive semi-definite precision matrix with rank $n - k > 0$. The random field $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$ is an improper GMRF if it has density function

$$p(\boldsymbol{x}) = 2\pi^{\frac{-(n-k)}{2}}(|\mathbf{Q}|^*)^{1/2} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\mathbf{Q}(\boldsymbol{x}-\boldsymbol{\mu})\right), \tag{2.4}$$

where $|\cdot|^*$ denotes the generalized determinant, calculated as the product of the non-zero eigenvalues of $\mathbf{Q}$, and parameters $\boldsymbol{\mu}$ and $\mathbf{Q}$ are not the mean and precision of the field, since they formally do not exist, although we will refer to them as the mean and precision.

IGMRFs of order k on a line or directional graph are typically built on forward differences of order k. We define the first-order difference of a function $f(\cdot)$ as

$$\Delta f(x) = f(x+1) - f(x)$$

and the second-order forward difference as

$$\Delta^2 f(x) = f(x+2) - 2f(x+1) + f(x).$$

Higher-order differences are defined recursively: $\Delta^k f(x) = \Delta\Delta^{k-1} f(x)$. The k-order forward difference is discrete approximation to the kth order derivative of $f(x)$ is

$$f'(x) = \lim_{h\to 0} \frac{f(x+h) - f(x)}{h},$$

which is equal to the first-order forward difference when $h = 1$.

Here we consider first-order and second-order random walks on a line. Random walks of higher order can be constructed in a similar manner to those below but are rarely used in practice. We

assume the locations of the nodes i are at positive integers, $i = 1, 2, \ldots, n$, such as the case when i are located at equally-spaced, discrete times. A first-order random walk, or IGMRF of order 1, is constructed by assuming the increments Δx_i are independent and identically distributed

$$\Delta x_i \sim \mathcal{N}(0, \kappa^{-1}), \tag{2.5}$$

where κ is the precision parameter. Given the summation properties of normal random variables, this implies that for $i < j$, $x_j - x_i \sim \mathcal{N}(0, (j-i)\kappa^{-1})$. Therefore, the marginal distributions of the x_i are normally distributed with variance increasing linearly with i. Assuming $\boldsymbol{\mu} = \mathbf{0}$ for notational convenience, the density for $\boldsymbol{x}$ is

$$\begin{aligned} p(\boldsymbol{x} \mid \kappa) &\propto \kappa^{(n-1)/2} \exp\left(-\frac{\kappa}{2} \sum_{i=1}^{n-1} (\Delta x_i)^2\right) \\ &= \kappa^{(n-1)/2} \exp\left(-\frac{\kappa}{2} \sum_{i=1}^{n-1} (x_{i+1} - x_i)^2\right). \end{aligned} \tag{2.6}$$

We define a $(n-1) \times n$ first-order differencing operator matrix $\mathbf{D}_1$, where

$$\mathbf{D}_1 = \begin{pmatrix} -1 & 1 & & & & \\ & -1 & 1 & & & \\ & & \ddots & \ddots & & \\ & & & -1 & 1 & \\ & & & & -1 & 1 \end{pmatrix}. \tag{2.7}$$

Note that $\mathbf{D}_1$ is sparse and the zeros are represented as blanks here. It follows that $\mathbf{D}_1\boldsymbol{x} = (\Delta x_1, \Delta x_2, \ldots \Delta x_{n-1})^T$ and so

$$\sum_{i=1}^{n-1} (\Delta x_i)^2 = (\mathbf{D}_1\boldsymbol{x})^T(\mathbf{D}_1\boldsymbol{x}) = \boldsymbol{x}^T\mathbf{D}_1^T\mathbf{D}_1\boldsymbol{x} = \boldsymbol{x}^T\mathbf{R}_1\boldsymbol{x},$$

where the $n \times n$ matrix $\mathbf{R}_1$ is

$$\mathbf{R}_1 = \begin{pmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & \ddots & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{pmatrix}. \tag{2.8}$$

We can then see that

$$p(\boldsymbol{x} \mid \kappa) \propto \kappa^{(n-1)/2} \exp\left(-\frac{1}{2}\boldsymbol{x}^T \mathbf{Q}_1 \boldsymbol{x}\right), \tag{2.9}$$

where $\mathbf{Q}_1 = \kappa \mathbf{R}_1$. The matrix $\mathbf{R}_1$ is sometimes referred to as the *structure* matrix since it defines the structure of the precision matrix $\mathbf{Q}_1$. Note that the proportionality sign in equation 2.9 arises because $(2\pi)^{\frac{-(n-1)}{2}}$ is a constant and $(|\mathbf{Q}_1|^*)^{1/2} = (|\kappa \mathbf{R}_1|^*)^{1/2} = \kappa^{(n-1)/2}(|\mathbf{R}_1|^*)^{1/2} \propto \kappa^{(n-1)/2}$, since $(|\mathbf{R}_1|^*)^{1/2}$ is a constant.

In a similar fashion, we can construct a second-order random walk on a line by using the order-2 increments and assuming for all $i = 1, 2 \ldots, n-2$,

$$\Delta^2 x_i \sim \mathcal{N}(0, \kappa^{-1}).$$

Again assuming $\boldsymbol{\mu} = \mathbf{0}$, the density for $\boldsymbol{x}$ given κ is then

$$p(\boldsymbol{x} \mid \kappa) \propto \kappa^{(n-2)/2} \exp\left(\frac{\kappa}{2} \sum_{i=1}^{n-2} (x_i - 2x_{i+1} + x_{i+2})^2\right). \tag{2.10}$$

We construct the $(n-2) \times n$ second-order differencing operator $\mathbf{D}_2$ recursively using the $(n-1) \times n$ matrix $\mathbf{D}_1$ from equation 2.7 and a $(n-2) \times (n-1)$ version of $\mathbf{D}_1$, call it $\mathbf{D}_{1,-1}$, that removes the rightmost column and bottom row of $\mathbf{D}_1$. That is $\mathbf{D}_2 = \mathbf{D}_{1,-1}\mathbf{D}_1$, or

$$\mathbf{D}_2 = \begin{pmatrix} 1 & -2 & 1 & & & & \\ & 1 & -2 & 1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & 1 & -2 & 1 & \\ & & & & 1 & -2 & 1 \end{pmatrix}. \tag{2.11}$$

It follows that $\mathbf{D}_2 \boldsymbol{x} = (\Delta^2 x_1, \Delta^2 x_2, \ldots \Delta^2 x_{n-2})^T$ and so

$$\sum_{i=1}^{n-2} (\Delta^2 x_i)^2 = (\mathbf{D}_2 \boldsymbol{x})^T (\mathbf{D}_2 \boldsymbol{x}) = \boldsymbol{x}^T \mathbf{D}_2^T \mathbf{D}_2 \boldsymbol{x} = \boldsymbol{x}^T \mathbf{R}_2 \boldsymbol{x},$$

where

$$\mathbf{R}_2 = \begin{pmatrix} 1 & -2 & 1 & & & & \\ -2 & 5 & -4 & 1 & & & \\ 1 & -4 & 6 & -4 & 1 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & 1 & -4 & 6 & -4 & 1 \\ & & & 1 & -4 & 5 & -2 \\ & & & & 1 & -2 & 1 \end{pmatrix}. \tag{2.12}$$

Then we can write the density of $\boldsymbol{x}$ for the second-order random walk as:

$$p(\boldsymbol{x} \mid \kappa) \propto \kappa^{(n-2)/2} \exp\left(-\frac{1}{2}\boldsymbol{x}^T \mathbf{Q}_2 \boldsymbol{x}\right), \tag{2.13}$$

where $\mathbf{Q}_2 = \kappa \mathbf{R}_2$.

2.1.5 Proper GMRF Random Walks

We can make proper versions of the intrinsic random walk models by putting a prior distribution on the first value of the field, x_1, and, for $k > 2$, by putting prior distributions on the first $n - k$ differences of order less than k. We will also introduce a constant overall mean μ. We can construct a proper first-order GMRF by placing a distribution on x_1, where $x_1 \sim \mathcal{N}(\mu, \omega^{-1})$, μ is the mean of x_1 and the overall mean of the field, and ω is the precision parameter for x_1. Using the Markov property, we have

$$\begin{aligned} p(\boldsymbol{x} \mid \mu, \omega, \kappa) &= p(x_1 \mid \mu, \omega) p(x_2 \mid x_1, \kappa) \cdots p(x_n \mid x_{n-1}, \kappa) \\ &= p(x_1 \mid \mu, \omega) p(\Delta x_1 \mid \kappa) \cdots p(\Delta x_{n-1} \mid \kappa). \end{aligned} \tag{2.14}$$

The second line of equation 2.14 above follows from the fact that the density for $x_i \sim \mathcal{N}(x_{i-1}, \kappa^{-1})$ is the same as $\Delta x_{i-1} \sim \mathcal{N}(0, \kappa^{-1})$, since $\exp\left(-\frac{\kappa}{2}(x_i - x_{i-1})^2\right) = \exp\left(-\frac{\kappa}{2}(\Delta x_{i-1})^2\right)$. It follows that the density for $\boldsymbol{x}$ is

$$\begin{aligned} p(\boldsymbol{x} \mid \mu, \omega, \kappa) &= (2\pi)^{-1/2} \omega^{1/2} \exp\left(-\frac{\omega}{2}(x_1 - \mu)^2\right) \prod_{i=2}^{n} (2\pi)^{-1/2} \kappa^{1/2} \exp\left(-\frac{\kappa}{2}(x_i - x_{i-1})^2\right) \\ &= (2\pi)^{-n/2} \omega^{1/2} \kappa^{(n-1)/2} \exp\left(-\frac{\omega}{2}(x_1 - \mu)^2\right) \prod_{i=2}^{n} \exp\left(-\frac{\kappa}{2}(x_i - x_{i-1})^2\right). \end{aligned} \tag{2.15}$$

For given values of the precision parameters ω and κ, we define the $n \times n$ precision matrix for a proper first-order random walk as:

$$\mathbf{Q}_1 = \kappa \begin{pmatrix} \frac{\omega}{\kappa} + 1 & -1 & & & & & \\ -1 & 2 & -1 & & & & \\ & -1 & 2 & -1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & -1 & 2 & -1 & \\ & & & & -1 & 1 \end{pmatrix}. \tag{2.16}$$

It can then be shown that

$$\exp\left(-\frac{\omega}{2}(x_1 - \mu)^2\right) \prod_{i=2}^{n} \exp\left(-\frac{\kappa}{2}(x_i - x_{i-1})^2\right) = \exp\left(-\frac{1}{2}(\boldsymbol{x} - \mu)^T \mathbf{Q}_1 (\boldsymbol{x} - \mu)\right).$$

It also turns out that $|\mathbf{Q}_1|^{1/2} = (\kappa^n \omega / \kappa)^{1/2} = \omega^{1/2} \kappa^{(n-1)/2}$. It follows that the density in equation 2.15 is equivalent to that in equation 2.3 and is therefore a proper GMRF.

Constructing a proper second-order random walk proceeds in a similar manner except that we need to set a prior on $\Delta x_1 = x_2 - x_1$ and account for this increment in the precision matrix. We assume $\Delta x_1 \sim \mathcal{N}(0, \kappa^{-1})$ for simplicity here, but could allow this increment to have a different precision value. As with the IGMRF version, we also assume $\Delta^2 x_i \sim \mathcal{N}(0, \kappa^{-1})$. The density for $\boldsymbol{x}$ for the proper second-order random walk is then

$$\begin{aligned} p(\boldsymbol{x} \mid \mu, \omega, \kappa) &= (2\pi)^{-1} \omega^{1/2} \exp\left(-\frac{\omega}{2}(x_1 - \mu)^2\right) \kappa^{1/2} \exp\left(-\frac{\kappa}{2}(x_2 - x_1)^2\right) \\ &\quad \prod_{i=1}^{n-2} (2\pi)^{-1/2} \kappa^{1/2} \exp\left(-\frac{\kappa}{2}(x_i - 2x_{i+1} + x_{i+2})^2\right) \\ &= (2\pi)^{-n/2} \omega^{1/2} \kappa^{(n-1)/2} \exp\left(-\frac{\kappa}{2} \boldsymbol{x}^T \mathbf{Q}_2 \boldsymbol{x}\right), \end{aligned} \tag{2.17}$$

where

$$\mathbf{Q}_2 = \kappa \begin{pmatrix} \omega/\kappa + 2 & -3 & 1 & & & & \\ -3 & 6 & -4 & 1 & & & \\ 1 & -4 & 6 & -4 & 1 & & \\ & \ddots & \ddots & \ddots & \ddots & \ddots & \\ & & 1 & -4 & 6 & -4 & 1 \\ & & & 1 & -4 & 5 & -2 \\ & & & & 1 & -2 & 1 \end{pmatrix}.$$

2.1.6 Sparsity and Computation

One of the major benefits of using GMRFs is the computational efficiency that can be gained by exploiting the sparsity of the precision matrix. The covariance matrix $\mathbf{\Sigma}$ of a multivariate normal distribution is typically dense. This is also true of the covariance matrix of a GMRF. Direct inversion of large matrices can be computationally costly and is rarely done in practice. Typically one uses the Cholesky factorization, $\mathbf{\Sigma} = \tilde{\mathbf{L}}\tilde{\mathbf{L}}^T$, where $\tilde{\mathbf{L}}$ is a lower triangular matrix. For a multivariate normal distribution defined in terms of the covariance matrix, the log density can be calculated as

$$\log p(\boldsymbol{x}) = \frac{n}{2}\log(2\pi) - \sum_{i=1}^{n} \log \tilde{L}_{ii} - \frac{1}{2}\boldsymbol{u}^T\boldsymbol{u},$$

where $\boldsymbol{u}$ is the solution to $\tilde{\mathbf{L}}\boldsymbol{u} = \boldsymbol{x} - \boldsymbol{\mu}$, and $\frac{1}{2}\log|\mathbf{\Sigma}| = \sum_{i=1}^{n} \log \tilde{L}_{ii}$. For a GMRF, the Cholesky factorization of the precision matrix is $\mathbf{Q} = \mathbf{L}\mathbf{L}^T$. The log density of a GMRF can be calculated as

$$\log p(\boldsymbol{x}) = \frac{n}{2}\log(2\pi) + \sum_{i=1}^{n} \log L_{ii} - \frac{1}{2}q,$$

where $q = (\boldsymbol{x}-\boldsymbol{\mu})^T\mathbf{Q}(\boldsymbol{x}-\boldsymbol{\mu})$ and $\frac{1}{2}\log|\mathbf{Q}| = \sum_{i=1}^{n} \log L_{ii}$. Notice the precision matrix approach does not require finding the solution for $\boldsymbol{u}$.

Further benefits of using the precision matrix can be had if methods for sparse matrices are used in calculating the Cholesky factorization and in matrix multiplication. Computing the Cholesky factorization of the full $\mathbf{Q}$ requires $O(n^3)$ floating point operations per second (flops), while using a sparse Cholesky factorization will reduce the flops to $O(n)$ for one-dimensional GMRFs and

$O(n^{3/2})$ for spatial fields (Rue and Held, 2005). In many cases, using optimal permutations of the precision matrix to reduce the bandwidth will also increase computational efficiency.

The Cholesky factorization and sparse matrix operations can also be used for sampling from a GMRF. Suppose we want to sample $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{Q}^{-1})$, where $\mathbf{Q} = \mathbf{L}\mathbf{L}^T$. We can take advantage of the fact that if $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, the solution of $\mathbf{L}^T\boldsymbol{x} = \boldsymbol{z}$ has covariance matrix $\text{Cov}(\boldsymbol{x}) = \text{Cov}(\mathbf{L}^{-T}\boldsymbol{z}) = (\mathbf{L}\mathbf{L}^T)^{-1} = \mathbf{Q}^{-1}$. This suggests the following approach to sample $\boldsymbol{x}$: 1) compute $\mathbf{Q} = \mathbf{L}\mathbf{L}^T$, 2) sample $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, 3) solve $\mathbf{L}^T\boldsymbol{v} = \boldsymbol{z}$ for $\boldsymbol{v}$ using back substitution (not direct inversion), and 4) compute $\boldsymbol{x} = \boldsymbol{\mu} + \boldsymbol{v}$.

2.2 Gaussian Processes

2.2.1 Definitions

A Gaussian process is a stochastic process typically defined over space and/or time, but can also be defined over any continuously indexed measure. A Gaussian process can be viewed as a distribution over continuously indexed functions, which is a useful viewpoint when one is trying to estimate an unknown function nonparametrically.

Formally, following the definition in Rue and Held (2005), let $\boldsymbol{Z}(\boldsymbol{s})$ be a stochastic process, where $\boldsymbol{s} \in D$ is a location in the domain $D \in \mathbb{R}^d$, with d typically equal to 1, 2, or 3. The process $\boldsymbol{Z}(\boldsymbol{s})$ is a Gaussian process or Gaussian field if for any $k \geq 1$ and any locations $\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_k \in D$, vector $(\boldsymbol{Z}(\boldsymbol{s}_1), \boldsymbol{Z}(\boldsymbol{s}_2), \ldots, \boldsymbol{Z}(\boldsymbol{s}_k))^T$ is normally distributed. A Gaussian process is completely determined by its mean and covariance function, where $\text{E}[\boldsymbol{Z}(\boldsymbol{s})] = \boldsymbol{\mu}(\boldsymbol{s})$ is the mean and $\text{C}[\boldsymbol{s}, \boldsymbol{t}] = \text{Cov}[\boldsymbol{Z}(\boldsymbol{s}), \boldsymbol{Z}(\boldsymbol{t})]$ is the covariance function, which are both assumed to exist for all $\boldsymbol{s}$ and $\boldsymbol{t}$.

A Gaussian process is *stationary* if the covariance function only depends on $\boldsymbol{s} - \boldsymbol{t}$ and $\boldsymbol{\mu}(\boldsymbol{s}) = \boldsymbol{\mu}$ for all $\boldsymbol{s} \in D$. A Gaussian process is *isotropic* if the covariance function only depends on the Euclidean distance $h = \sqrt{\| \boldsymbol{s} - \boldsymbol{t} \|}$ between $\boldsymbol{s}$ and $\boldsymbol{t}$, that is, $\text{C}[\boldsymbol{s}, \boldsymbol{t}] = \text{C}(h)$. Some examples of draws from Gaussian processes with different parameter values for covariance functions are shown in Figure 2.1.

A very popular covariance function used with Gaussian processes is the Matérn covariance

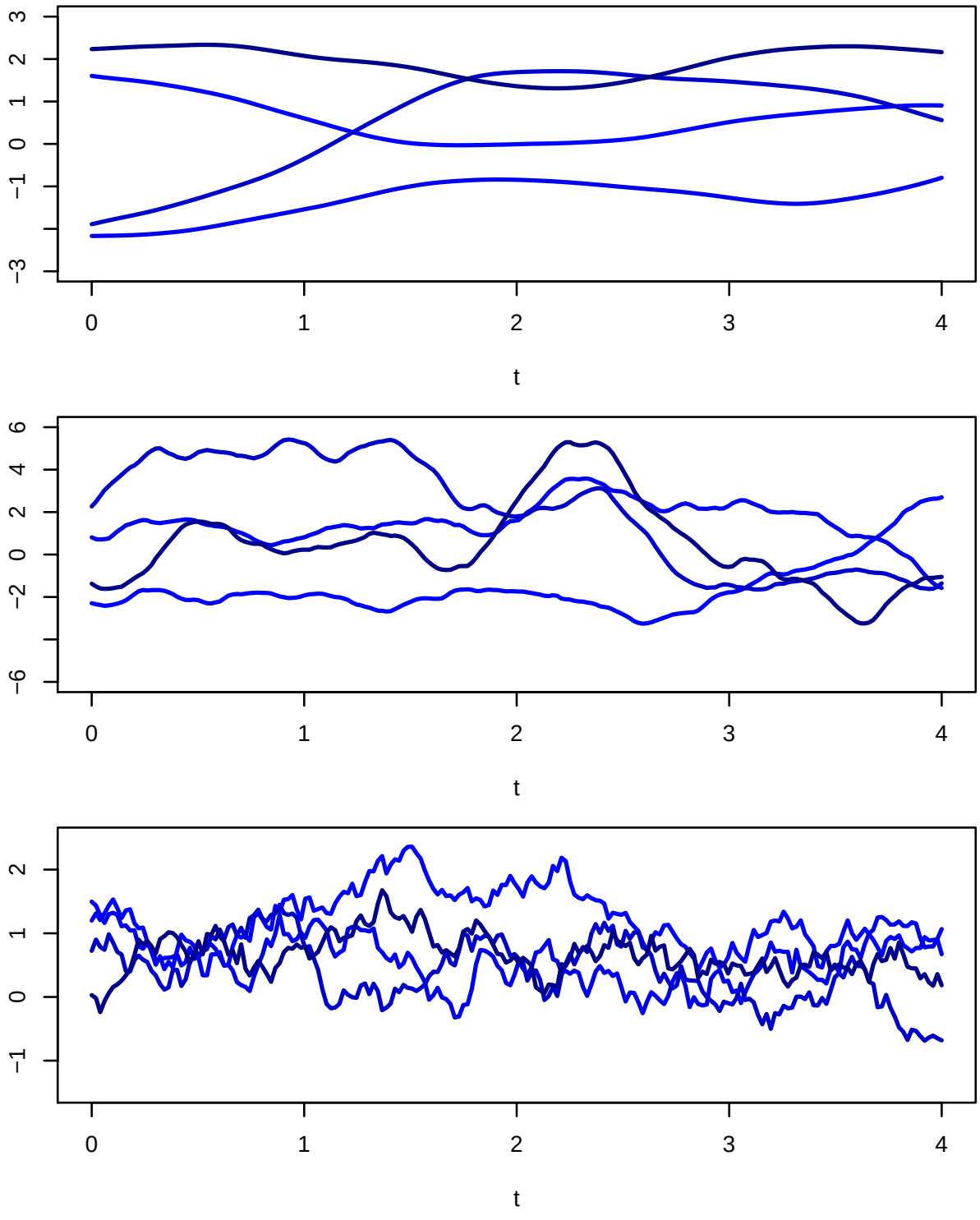

Figure 2.1: Draws from Gaussian processes with Matérn covariance functions with different parameter values.

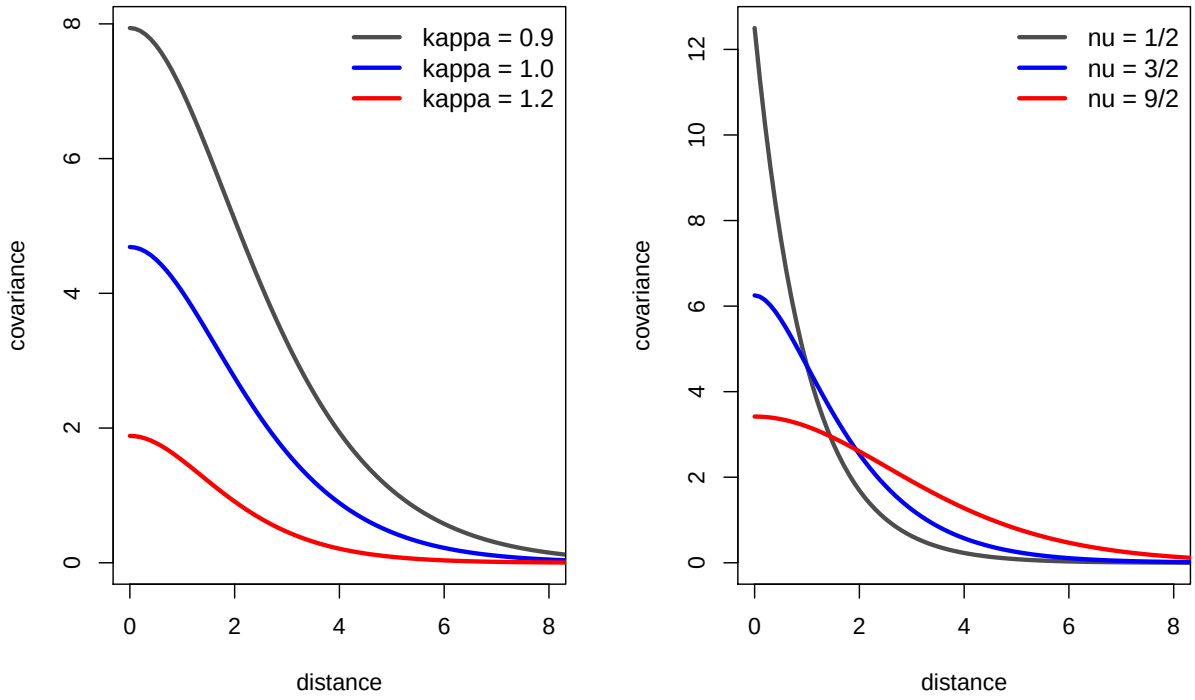

Figure 2.2: Matérn covariance functions with different parameter values.

function:

$$C(\mathbf{h}) = \frac{2^{1-\nu}\phi^2}{(4\pi)^{\frac{d}{2}}\Gamma(\nu + \frac{d}{2})\kappa^{2\nu}}(\kappa\|\mathbf{h}\|)^{\nu}K_{\nu}(\kappa\|\mathbf{h}\|).$$

Here $\|\mathbf{h}\|$ is a Euclidean norm, d is the dimension of the domain, ν is a shape parameter, κ is a scale parameter, ϕ^2 is a variance parameter, and K is a modified Bessel function of the second kind. This covariance function results in a stationary and isotropic process. The Matérn covariance function is flexible and widely used, and certain parameter values result in special cases of other well-known covariance functions, such as the exponential and Gaussian covariance functions. Figure 2.2 shows the effect of changing different parameters on the shape of the Matérn covariance function.

2.2.2 *Approximations*

Inference for Gaussian processes in high dimension can be difficult due to computational issues. In particular, calculating the density function requires inverting the covariance matrix (typically using the Cholesky factorization), which can be costly in high dimensions. Many methods have be developed to approximate Gaussian processes while reducing the dimensionality. One method we already discussed is the GMRF. Other approximations include process convolutions (Higdon, 1998), basis function representations (Katzfuss, 2017), and stochastic partial differential equations (Lindgren et al., 2011). Some other approximation techniques reduce computations by manipulating the covariance matrix via covariance tapering (Furrer et al., 2006), or by defining neighborhoods over which to apply the covariance function via nearest-neighbor models (Datta et al., 2016). There are many more examples, but all have the goal of reducing the dimensionality of the problem. Next, we show brief examples of a couple of these approximation techniques.

A process convolution is one way to express a Gaussian process on $\mathbb{R}^d$:

$$X(\mathbf{s}) = \int_{\mathbb{R}^d} k(\mathbf{s}, \mathbf{u}) \mathcal{B}(\mathrm{d}\mathbf{u}),$$

where k is a deterministic kernel function and $\mathcal{B}$ is a Brownian sheet (Higdon, 2002). The covariance function for X is $C(\mathbf{h}) = \int k(\mathbf{u} - \mathbf{h}) k(\mathbf{u}) \mathrm{d}\mathbf{u}$, which is a function of the kernel. Many different covariance functions, including nonstationary ones, can be induced using different kernels or by allowing the kernel parameterization to vary over space. The integrals are approximated numerically, so this method can still be computationally costly.

Lindgren et al. (2011) showed that a Gaussian process with a Matérn covariance function can be approximated using the solution to a set of stochastic partial differential equations. A Gaussian Matérn field $X(\mathbf{s})$ is the solution to the following stochastic partial differential equation:

$$\left(\kappa^2 - \Delta\right)^{\frac{\alpha}{2}} X(\mathbf{s}) = \phi \mathcal{W}(\mathbf{s}),$$

where $\mathcal{W}(\mathbf{s})$ is a spatial white noise process with unit variance, ϕ is a variance parameter, $\alpha = \nu + d/2$, and $\Delta = \sum_{i=1}^{d} \frac{\partial^2}{\partial s_i^2}$ is the Laplacian operator. The solution can be approximated using a compact

and sparse set of basis functions in combination with a GMRF, which results in a computationally efficient method.

2.3 Applying GMRFs to Data

So far we have only considered the descriptions of the probabilistic models for the underlying data generating processes, but have not considered how these models are applied to actual data. GMRF models have been used in a wide variety of data applications.

In time series analysis, autoregressive models are frequently used and are types of GMRFs on a linear graph. State-space models are also popular in time series analysis and these models partition an unobserved state process from the observation process (Brockwell and Davis, 2002). The state process is typically modeled as a random walk or an autoregressive process, which make state-space models a variation on GMRF models. These models have been extended to the multivariate time series setting (Holmes et al., 2012). GMRFs have been used in semiparametric regression and splines. Fahrmeir and Lang (2001) used GMRF priors in combination with covariates to create semiparametric generalized additive models. Lang and Brezger (2004) put random walk priors on spline coefficients to create Bayesian penalized splines. GMRFs have seen frequent use in image analysis. A small number of the many examples include restoring ultrasound images (Husby and Rue, 2004), object identification (Rue and Hurn, 1999), smoothing fMRI images of brain activity (Brezger et al., 2007; Yue et al., 2010), estimating densities of crowds (Lamba and Nain, 2019), and segmentation of remote sensing images (Zheng and Yao, 2019). GMRF models have also been used extensively in spatial statistics. Often the intrinsic forms of either conditional autoregressive models or GMRF models are used. A few of the many spatial applications include analysis of agricultural field experiments (Besag and Higdon, 1999), disease mapping (Besag et al., 1991; Knorr-Held and Rue, 2002), and modeling species abundance (Chakraborty et al., 2010), species diversity (Gelfand et al., 2005), static species occupancy (Broms et al., 2016), and dynamic species occupancy for colonization (Broms et al., 2016). The very popular SPDE approach for approximating Gaussian processes uses GMRF priors for generating the weights that provide the values of the spatial field at basis function locations (Lindgren et al., 2011). GMRFs are also used

in the INLA computing package as nonparametric smoothing functions over time, space, or other explanatory variables (Rue et al., 2009, 2017).

Suppose we have a set of observations y_i, $i = 1, 2, \ldots, n$, which are collected at regular points in time or space. We assume the y_i values follow a probability distribution which can be described by the probability density or likelihood function $p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\xi})$, where $\boldsymbol{x}$ is assumed to be a GMRF, and $\boldsymbol{\xi}$ are other parameters related to the likelihood. We assume that the y_is are conditionally independent given the value of the underlying field parameters x_i. The posterior distribution for the parameters given the data can be written as

$$p(\boldsymbol{x}, \mathbf{Q}, \boldsymbol{\xi}, \mid \mathbf{y}) \propto p(\boldsymbol{y} \mid \boldsymbol{x}, \boldsymbol{\xi}) p(\boldsymbol{x} \mid \boldsymbol{\mu}, \mathbf{Q}) p(\boldsymbol{\xi}) p(\boldsymbol{\mu}) p(\mathbf{Q}), \tag{2.18}$$

where $p(\boldsymbol{x} \mid \boldsymbol{\mu}, \mathbf{Q})$ is the GMRF prior and $p(\boldsymbol{\xi})$, $p(\boldsymbol{\mu})$, $p(\mathbf{Q})$ are priors for the remaining parameters. Note that the precision matrix $\mathbf{Q}$ is typically dependent on a single precision parameter.

In a Bayesian setting, typically Markov chain Monte Carlo (MCMC) is used to approximate the posterior distribution of the parameters. An exception to this would be to approximate the posterior using something like integrated nested Laplace approximations (INLA; Rue et al., 2009). MCMC for GMRFs can be problematic due to the large number of parameters and the correlations among the parameters. One approach that has been shown to improve mixing and convergence is block updating groups of parameters (Rue, 2001; Knorr-Held and Rue, 2002). In this approach joint proposals are made for blocks of parameters and acceptance/rejection of the parameters is done jointly. Another method that has proven useful for latent Gaussian models is elliptical slice sampling (Murray et al., 2010), which also jointly proposes the field parameters. A more general MCMC method that jointly proposes parameters and is very efficient is Hamiltonian Monte Carlo (Neal, 2011), which we discuss in detail in the next section.

2.4 Hamiltonian Monte Carlo

Hamiltonian dynamics describes the state of a d-dimensional frictionless dynamic physical system where total energy is conserved. For illustration, consider a frictionless system where an object of mass m glides over a two-dimensional surface of varying height with velocity $\boldsymbol{v}$. The state space

of the system consists of the position of the object, denoted by the coordinate vector $\boldsymbol{q}$ and its momentum (equal to mass times velocity), denoted by vector $\boldsymbol{p}$. There are two forms of energy present in the system. The first is potential energy, $U(\boldsymbol{p})$, which is proportional to the height of the surface at position $\boldsymbol{q}$. The second is kinetic energy, $K(\boldsymbol{p})$, which is equal to $m|\boldsymbol{v}|^2/2$ or $K(\boldsymbol{p}) = |\boldsymbol{p}|^2/(2m)$. On the level part of the surface, the object moves at a constant velocity, equal to $\boldsymbol{p}/m$. As the object moves upward in slope, the physics of the system dictate that potential energy increases and kinetic energy decreases, until kinetic energy reaches zero at a peak. At this point the object moves downslope while potential energy decreases and kinetic energy increases. This system can be represented by the Hamiltonian function, which describes the total energy in the system:

$$H(\boldsymbol{q}, \boldsymbol{p}) = U(\boldsymbol{q}) + K(\boldsymbol{p}).$$

The partial derivatives of $H(\boldsymbol{q}, \boldsymbol{p})$ determine how $\boldsymbol{q}$ and $\boldsymbol{p}$ change over time. These are known as Hamilton's equations:

$$\begin{aligned} \frac{dq_i}{dt} &= \frac{\partial H}{\partial p_i} = \frac{\partial K}{\partial p_i}, \\ \frac{dp_i}{dt} &= -\frac{\partial H}{\partial q_i} = -\frac{\partial U}{\partial q_i}, \end{aligned} \tag{2.19}$$

for $i = 1, \ldots, d$. These equations define a mapping T_s from the state at some time t to the state at some time $t + s$.

Say we have a set of variables with unknown probability distribution of dimension d which we want to draw samples from. If we define the potential energy function to be minus the log of the density function of this distribution (plus any constant), and therefore let the position variables $\boldsymbol{q}$ represent the variables of interest, then we can use Hamiltonian dynamics to sample from this distribution. We also need to introduce a set of auxiliary variables to represent the momentum variables, $\boldsymbol{p}$, which are also of dimension d and have a distribution defined by the kinetic energy function. The variables $\boldsymbol{q}$ and $\boldsymbol{p}$ are independent and their joint density is defined as

$$P(\boldsymbol{q}, \boldsymbol{p}) = \frac{1}{Z} \exp\left[-H(\boldsymbol{q}, \boldsymbol{p})\right] = \frac{1}{Z} \exp\left[-U(\boldsymbol{q})\right] \exp\left[-K(\boldsymbol{p})\right].$$

The kinetic energy function is usually defined as $K(\boldsymbol{p}) = \boldsymbol{p}^T \boldsymbol{M}^{-1} \boldsymbol{p}/2$, where the mass matrix $\boldsymbol{M}$ is typically diagonal with diagonal elements $m_1, \ldots, m_d$. This gives $K(\boldsymbol{p}) = \sum_i p_i^2/(2m_i)$, which means the momentum variables, p_is, are independent and each p_i follows a Gaussian distribution with mean zero and variance m_i.

For application to Bayesian statistics where we are interested in sampling from an unknown posterior distribution, we let the position variables, $\boldsymbol{q}$, take the role of the unknown model parameters. We can then express the posterior distribution using a potential energy function as

$$U(\boldsymbol{q}) = -\log\left[P(\boldsymbol{q})L(\boldsymbol{q} \mid D)\right],$$

where $P(\boldsymbol{q})$ is the prior density and $L(\boldsymbol{q} \mid D)$ is the likelihood of the parameters given the data D.

Hamiltonian dynamics is reversible, leaves H invariant, and preserves volume (Neal, 2011), which are properties that allow it to be used for an MCMC proposal. In theory we could now use Hamilton's equations applied for a specific time period to propose a new state within the Metropolis MCMC algorithm. This proposal would always be accepted under the ideal conditions where we have continuous time and can solve Hamilton's equations exactly. However, in practice this is not possible, and we need to approximate the equations using discrete time. A discretization method known as the leapfrog method is typically used for this and works as follows:

$$\begin{aligned}
p_j(t + \epsilon/2) &= p_j(t) - (\epsilon/2)\frac{\partial U}{\partial q_j}(q(t)), \\
q_j(t + \epsilon) &= q_j(t) + \epsilon\frac{\partial K}{\partial p_j}(p(t + \epsilon/2)), \\
p_j(t + \epsilon) &= p_j(t + \epsilon/2) - (\epsilon/2)\frac{\partial U}{\partial q_j}(q(t + \epsilon)).
\end{aligned}$$

We can now use these leapfrog steps to make a MCMC proposal. We start at state $(\boldsymbol{q}, \boldsymbol{p})$ and fictitious time $t = 0$, and move the system ahead some number of steps L with step size ϵ for a time $t = L\epsilon$ to the proposed state $(\boldsymbol{q}^*, \boldsymbol{p}^*)$. The proposal is then accepted as the next state in the Markov chain with probability

$$\min[1, \exp(-H(\boldsymbol{q}^*, \boldsymbol{p}^*) + H(\boldsymbol{q}, \boldsymbol{p}))] = \min\left[1, \frac{\exp(U(\boldsymbol{q}))\exp(K(\boldsymbol{p}))}{\exp(U(\boldsymbol{q}^*))\exp(K(\boldsymbol{p}^*))}\right],$$

and set to the current state otherwise. These proposals will not always be accepted since H does not remain perfectly constant in practice. However, these Metropolis updates leave H approximately constant, and therefore do not explore the entire joint distribution of $\boldsymbol{q}$ and $\boldsymbol{p}$. This makes it necessary for the HMC algorithm to alternate Metropolis updates with updates in which the momentum variables are sampled from their Gaussian distribution. This HMC algorithm results in proposals far from the current state which are accepted with high probability.

One potential drawback of HMC is the need to calculate gradients of the log posterior. Gradients can be difficult or impossible to derive analytically in some cases. Solutions to this problem can be provided by numerical differentiation methods or automatic differentiation. Another potential drawback of HMC is that performance can be sensitive to the step length ϵ and number of steps L, which can potentially result in the need for extensive manual tuning to achieve optimal performance. To overcome this issue, Hoffman and Gelman (2014) proposed a modification of HMC which adaptively updates the step length and step size. The open source package `rstan` (Stan Development Team, 2015a) provides a platform for fitting models using HMC in the `R` computing environment (R Core Team, 2017). This package incorporates automatic differentiation for calculating gradients and uses the adaptive HMC algorithm of Hoffman and Gelman (2014). We used `rstan` for posterior inference for the models investigated in this paper.

We note that the performance of HMC applied to hierarchical models can be greatly improved by using the non-centered parameterization methods described by Papaspiliopoulos et al. (2003, 2007). Hierarchical models have layers of dependent parameters due to the conditional nature of the distributions. Small changes in a few global (lower level) hyperparameters can affect many upper level parameters, resulting in large changes in the distribution. Non-centered parameterizations break the dependencies among parameters by introducing deterministic transformations of the parameters. The MCMC algorithm then operates directly on the *a priori* independent parameters. Consider the following simple example, which is a centered parameterization of a normal means

model with known error variance σ^2:

$$\begin{aligned} y_i \,|\, \theta_i &\sim \mathrm{N}(\theta_i, \sigma^2), \\ \theta_i \,|\, \tau &\sim \mathrm{N}(0, \tau^2), \\ \tau &\sim \mathrm{C}^+(0, 1). \end{aligned} \tag{2.20}$$

We can create a non-centered parameterization of (2.20) by making the transformation $\theta_i = \tau z_i$, where $z_i \sim \mathrm{N}(0, 1)$. The new variable z_i is *a priori* independent of τ. The distribution for τ remains unchanged and the model becomes

$$\begin{aligned} y_i \,|\, z_i, \tau &\sim \mathrm{N}(\tau z_i, \sigma^2), \\ z_i &\sim \mathrm{N}(0, 1), \\ \tau &\sim \mathrm{C}^+(0, 1). \end{aligned} \tag{2.21}$$

The origin of HMC goes back to Alder and Wainwright (1959) who used Hamiltonian dynamics to deterministically simulate the motion of molecules. Duane et al. (1987) combined Hamiltonian dynamics with Markov chain Monte Carlo for lattice field theory simulations. They called the method hybrid Monte Carlo. Neal (1996) was one of the first to use hybrid Monte Carlo for statistical inference. The method found limited use in statistics until pivotal papers by Girolami and Calderhead (2011) and Neal (2011) brought more attention to the method. As previously mentioned, the automatic tuning methods introduced by Hoffman and Gelman (2014) made HMC more accessible to the general user. Monnahan et al. (2017) provide a user-friendly description of HMC and find with a simulation study that HMC is more efficient than Gibbs sampling.

Many of the topics presented in this chapter appear as foundational elements in the research developments presented in later chapters. We modify the standard Gaussian Markov random field to create a new nonparametric smoothing technique in Chapter 3, and extend the method further in Chapters 4 and 5. The method of using stochastic partial differential equations to approximate Gaussian processes is drawn on in Chapter 3 when extending our second-order models to irregular grids, and is referred to in Chapter 6 as the proposed method for extending our models to continuous space. Finally, Hamiltonian Monte Carlo is used for posterior inference in Chapters 3, 4, and 5.

Chapter 3

LOCALLY-ADAPTIVE SMOOTHING WITH MARKOV RANDOM FIELDS AND SHRINKAGE PRIORS

3.1 Introduction

Nonparametric curve fitting methods find extensive use in many aspects of statistical modeling such as nonparametric regression, spatial statistics, and survival models, to name a few. Although these methods form a mature area of statistics, many computational and statistical challenges remain when such curve fitting needs to be incorporated into multi-level Bayesian models with complex data generating processes. This work is motivated by the need for a curve fitting method that could adapt to local changes in smoothness of a function, including abrupt changes or jumps, and would not be restricted by the nature of observations and/or their associated likelihood. Our desired method should offer measures of uncertainty for use in inference, should be relatively simple to implement and computationally efficient. There are many methods available for nonparametric curve fitting, but few which meet all of these criteria.

Gaussian process (GP) regression (Neal, 1998; Rasmussen and Williams, 2006) is a popular Bayesian nonparametric approach for functional estimation that places a GP prior on the function of interest. The covariance function must be specified for the GP prior, and the isotropic covariance functions typically used are not locally adaptive. Nonstationary covariance functions have been investigated to make GP regression locally adaptive (Brahim-Belhouari and Bermak, 2004; Paciorek and Schervish, 2004, 2006). Any finite dimensional representation of GPs involves manipulations of, typically high dimensional, Gaussian vectors with mean vector and covariance matrix induced by the GP. Many GPs, including the ones with nonstationary covariance functions, suffer from high computational cost imposed by manipulations (e.g., Cholesky factorization) of the dense covariance matrix in the finite dimensional representation.

Sparsity can be imposed in the precision matrix (inverse covariance matrix) by constraining a finite dimensional representation of a GP to be a Gaussian Markov random field (GMRF), and then computational methods for sparse matrices can be employed to speed computations (Rue, 2001; Rue and Held, 2005). Fitting smooth functions with GMRFs has been practiced widely. These methods use difference equations as approximations to continuous function derivatives to induce smoothing, and have a direct relationship to smoothing splines (Speckman and Sun, 2003).

GMRFs have also been used to develop Bayesian adaptive smoothing splines (Lang et al., 2002; Yue et al., 2012, 2014). A similar approach is the nested GP (Zhu and Dunson, 2013), which puts a GP prior on the order-k function derivative, which is in turn centered on another GP. This approach has good adaptive properties but has not been developed for non-Gaussian data.

Differencing has commonly been used as an approach to smoothing and trend estimation in time series analysis, signal processing, and spatial statistics. Its origins go back at least to Whittaker (1922), who suggested a need for a trade off between fidelity to the data and smoothness of the estimated function. This idea is the basis of some frequentist curve-fitting methods based on penalized least squares, such as the smoothing spline (Reinsch, 1967; Wahba, 1975) and the trend filter (Kim et al., 2009; Tibshirani, 2014). These penalized least-squares methods are closely related to regularization methods for high-dimensional regression such as ridge regression (Hoerl and Kennard, 1970) and the lasso (Tibshirani, 1996) due to the form of the penalties imposed.

Bayesian versions of methods like the lasso (Park and Casella, 2008) utilize shrinkage priors in place of penalties. Therefore, it is interesting to investigate how these shrinkage priors (Polson and Scott, 2010; Griffin et al., 2013; Bhattacharya et al., 2015) perform when applied to differencing-based time series smoothing. Although shrinkage priors have been used explicitly in the Bayesian nonparametric regression setting for regularization of wavelet coefficients (Abramovich et al., 1998; Johnstone and Silverman, 2005; Reményi and Vidakovic, 2015) and for shrinkage of order-k differences of basis spline coefficients in adaptive Bayesian P-splines (Scheipl and Kneib, 2009), a Bayesian version of the trend filter and Markov random field (MRF) smoothing with shrinkage priors has not been thoroughly investigated. To our knowledge, only Roualdes (2015), independently from our work, looked at Laplace prior-based Bayesian version of the trend filter in the context of a normal response model. In this paper, we conduct a thorough investigation of smoothing with shrinkage priors applied to MRFs for Gaussian and non-Gaussian data. We call the resulting models shrinkage prior Markov random fields (SPMRFs).

We borrow the idea of shrinkage priors from the sparse regression setting and apply it to the problem of function estimation. We take the perspective that nonparametric curve fitting is essentially a regularization problem where estimation of an unknown function can be achieved by

inducing sparsity in its order-k derivatives. We propose a few fully Bayesian variations of the trend filter (Kim et al., 2009; Tibshirani, 2014) which utilize shrinkage priors on the kth-order differences in values of the unknown target function. The shrinkage imposed by the priors induces a locally adaptive smoothing of the trend. The fully Bayesian implementation allows representation of parameter uncertainty through posterior distributions and eliminates the need to specify a single global smoothing parameter by placing a prior distribution on the smoothing parameter, although complete automation is not possible so we offer ways to parameterize the global smoothing prior. In Section 3.2 we provide a derivation of the models starting from penalized frequentist methods and we show the relationship to GMRF models. In Section 3.2 we also describe our method of sampling from the posterior distribution of the parameters using Hamiltonian Monte Carlo (HMC), which is efficient and straightforward to implement. In Section 3.3 we use simulations to investigate performance properties of the SPMRF models under two different prior formulations and we compare results to those for a GMRF with constant precision. We show that the choice of shrinkage prior will affect the smoothness and local adaptive properties. In Section 3.4 we apply the method to two example data sets which are well known in the nonparametric regression setting.

3.2 Methods

3.2.1 Preliminaries

We start by reviewing a locally adaptive penalized least squares approach to nonparametric regression known as the trend filter (Kim et al., 2009; Tibshirani and Taylor, 2011; Tibshirani, 2014) and use that as a basis to motivate a general Bayesian approach that utilizes shrinkage priors in place of roughness penalties. We first consider the standard nonparametric regression problem to estimate the unknown function f. We let $\boldsymbol{\theta}$ represent a vector of values of f on a discrete uniform grid $t \in \{1, 2, \ldots, n\}$, and we assume $\boldsymbol{y} = \boldsymbol{\theta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \mathbf{I}\sigma^2)$, and $\boldsymbol{y}$ and $\boldsymbol{\epsilon}$ are vectors of length n. Here all vectors are column vectors. Following Tibshirani (2014) with slight modification, the least squares estimator of the kth order trend filtering estimate $\hat{\boldsymbol{\theta}}$ is

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \ \|\mathbf{y} - \boldsymbol{\theta}\|_2^2 + \lambda \|\mathbf{D}^{(k)}\boldsymbol{\theta}\|_1 \,, \tag{3.1}$$

where $\|\cdot\|_q$ represents the L_q vector norm, and $\mathbf{D}^{(k)}$ is an $(n-k)\times n$ forward difference operator matrix of order k, such that the ith element of the vector $\Delta^k\boldsymbol{\theta} = \mathbf{D}^{(k)}\boldsymbol{\theta}$ is the forward difference $\Delta^k\theta_i = (-1)^k \sum_{j=0}^{k}(-1)^j\binom{k}{j}\theta_{i+j}$. Note that $\mathbf{D}^{(k)}$ has recursive properties such that $\mathbf{D}_n^{(k)} = \mathbf{D}_{n-k+1}^{(1)}\mathbf{D}_n^{(k-1)}$, where $\mathbf{D}_m^{(h)}$ has dimensions $(m-h)\times m$. The objective function in equation (3.1) balances the trade-off between minimizing the squared deviations from the data (the first term in the sum on the right) with minimizing the discretized roughness penalty of the function f (the second term in the sum on the right). The smoothing parameter $\lambda \geq 0$ controls the relative influence of the roughness penalty. Setting λ to 0 we get least squares estimation. As λ gets large, the roughness penalty dominates, resulting in a function with k-th order differences approaching 0 for all t. The trend filter produces a piecewise polynomial function of $t_1, \ldots, t_n$ with degree $k-1$ as an estimator of the unknown function f. Increasing the order of the difference operator will enforce a smoother function.

The L_1 penalty in equation (3.1) results in the trend filter having locally adaptive smoothing properties. Tibshirani (2014) shows that the trend filter is very similar in form and performance to smoothing splines and locally adaptive regression splines, but the trend filter has a finer level of local adaptivity than smoothing splines. A main difference between the trend filter and smoothing splines is that the latter uses a squared L_2 penalty, which is the same penalty used in ridge regression (Hoerl and Kennard, 1970). Note that the L_1 penalty used by the trend filter is also used by the lasso regression (Tibshirani, 1996), and the trend filter is a form of generalized lasso (Tibshirani and Taylor, 2011; Tibshirani, 2014). In the linear regression setting with regression coefficients β_js, the L_1 and L_2 penalties can be represented by the generalized ridge penalty $\lambda \sum_j |\beta_j|^q$ (Frank and Friedman, 1993), where $q = 2$ corresponds to the ridge regression penalty, $q = 1$ to the lasso penalty, and sending q to zero results in all subsets selection regression (Tibshirani, 2011). Based on what we know about lasso regression, subset selection regression, and ridge regression, we expect a penalty closer to subset selection to do better for fitting functions with a small number of large jumps, a trend filter penalty (L_1) to do better for fitting functions with small to moderate deviations from polynomials of degree $k-1$, and a smoothing spline (squared L_2) penalty to do better for smooth polynomial-like functions with no jumps. This distinction will become important later when we assess the performance of different Bayesian formulations of the trend filter.

One can translate the penalized least squares formulation in equation (3.1) into either a penalized likelihood formulation or a Bayesian formulation. Penalized least squares can be interpreted as minimizing the penalized negative log-likelihood $-l_p(\boldsymbol{\theta} \mid \boldsymbol{y}) = -l(\boldsymbol{\theta} \mid \boldsymbol{y}) + p(\boldsymbol{\theta} \mid \lambda)$, where $l(\boldsymbol{\theta} \mid \boldsymbol{y})$ is the unpenalized log-likelihood and $p(\boldsymbol{\theta} \mid \lambda)$ is the penalty. It follows that maximization of the penalized log-likelihood is directly comparable to finding the mode of the log-posterior in the Bayesian formulation, where the penalty is represented as a prior. This implies independent Laplace (double-exponential) priors on the $\Delta^k\theta_j$, where $j = 1, \ldots, n-k$, for the trend filter formulation in equation (3.1). That is, $p(\Delta^k\theta_j \mid \lambda) = \frac{\lambda}{2}\exp\left(-\lambda\,|\,\Delta^k\theta_j\,|\right)$. This is a well-known result that has been used in deriving a Bayesian form of the lasso (Tibshirani, 1996; Figueiredo, 2003; Park and Casella, 2008). Note that putting independent priors on the kth order differences results in improper joint prior $p(\boldsymbol{\theta} \mid \lambda)$, which can be made proper by including a proper prior on the first k θs.

The Laplace prior falls into a class of priors commonly known as shrinkage priors. An effective shrinkage prior has the ability to shrink noise to zero yet retain and accurately estimate signals (Polson and Scott, 2010). These properties translate into a prior density function that has a combination of high mass near zero and heavy tails. The high density near zero acts to shrink small values close to zero, while the heavy tails allow large signals to be maintained. A simple prior developed for subset selection in Bayesian setting is the spike-and-slab prior, which is a mixture distribution between a point mass at zero and a continuous distribution (Mitchell and Beauchamp, 1988). This prior works well for model selection, but some drawbacks are that it forces small signals to be exactly zero, and computational issues can make it difficult to use (Polson and Scott, 2010). There has been much interest in developing priors with continuous distributions (one group) that retain variable selection properties of the spike-and-slab (two-group) yet do so by introducing sparsity through shrinkage (Polson and Scott, 2010). This approach allows all of the coefficients to be nonzero, but most are small and only some are large. Many such shrinkage priors have been proposed, including the normal-gamma (Griffin et al., 2010), generalized double-Pareto (Armagan et al., 2013), horseshoe (Carvalho et al., 2010), horseshoe+ (Bhadra et al., 2017), and Dirichlet-Laplace (Bhattacharya et al., 2015). The Laplace prior lies somewhere between the normal prior

and the spike-and-slab in its shrinkage abilities, yet most shrinkage priors of current research interest have sparsity inducing properties closer to those of the spike-and-slab. Our main interest is in comparing the Laplace prior to other shrinkage priors in the context of nonparametric smoothing.

3.2.2 Model Formulation

It is clear that shrinkage priors other than the lasso could represent different smoothing penalties and therefore could lead to more desirable smoothing properties. There is a large and growing number of shrinkage priors in the literature. It is not our goal to compare and characterize properties of Bayesian nonparametric function estimation under all of these priors. Instead, we wish to investigate a few well known shrinkage priors and demonstrate as proof of concept that adaptive functional estimation can be achieved with shrinkage priors. Further research can focus on improvements to these methods. What follows is a general description of our modeling approach and the specific prior formulations that will be investigated through the remainder of the paper.

We assume the n observations y_i, where $i = 1, \ldots, n$, are independent and follow some distribution dependent on the unknown function values θ_i and possibly other parameters $\boldsymbol{\xi}$ at discrete points t. We further assume that the order-k forward differences in the function parameters, $\Delta^k\theta_j$, where $j = 1, \ldots, n-k$, are independent and identically distributed conditional on a global scale parameter which is a function of the smoothing parameter λ. These assumptions result in the following general hierarchical form:

$$y_i \mid \theta_i, \boldsymbol{\xi} \sim p(y_i \mid \theta_i, \boldsymbol{\xi}), \quad \Delta^k\theta_j \mid \lambda \sim p(\Delta^k\theta_j \mid \lambda), \quad \lambda \sim p(\lambda), \quad \boldsymbol{\xi} \sim p(\boldsymbol{\xi}). \tag{3.2}$$

One convenient trait of many shrinkage priors, including the Laplace, the logistic, and the t-distribution, is that they can be represented as scale mixtures of normal distributions (Andrews and Mallows, 1974; West, 1987; Polson and Scott, 2010). The conditional form of scale mixture densities leads naturally to hierarchical representations. This can allow some otherwise intractable density functions to be represented hierarchically with standard distributions and can ease computation. To take advantage of this hierarchical structure, we restrict densities $p(\Delta^k\theta_j \mid \lambda)$ to be scale mixtures of normals, which allows us to induce a hierarchical form to our model formulation by

introducing latent local scale parameters, τ_j. Here the order-k differences in the function parameters, $\Delta^k\theta_j$, are conditionally normally distributed with mean zero and variance τ_j^2, and the τ_j are independent and identically distributed with a global scale parameter which is a function of the smoothing parameter λ. The distribution statement for $\Delta^k\theta_j$ in Equation (3.2) can then be replaced with the following hierarchical representation:

$$\Delta^k\theta_j \mid \tau_j \sim \mathrm{N}(0, \tau_j^2), \quad \tau_j \mid \lambda \sim p(\tau_j \mid \lambda). \tag{3.3}$$

To complete the model specification, we place proper priors on $\theta_1, \ldots, \theta_k$. This maintains propriety and can improve computational performance for some Markov chain Monte Carlo (MCMC) samplers. We start by setting $\theta_1 \sim \mathrm{N}(\mu, \omega^2)$, where μ and ω can be constants or allowed to follow their own distributions. Then for $k \geq 2$ and $h = 1, \ldots, k-1$, we let $\Delta^h\theta_1 \,|\, \alpha_h \sim \mathrm{N}(0, \alpha_h^2)$ and $\alpha_h \,|\, \lambda \sim p(\alpha_h \,|\, \lambda)$, where $p(\alpha \,|\, \lambda)$ is the same form as $p(\tau \,|\, \lambda)$. That is, we assume the order-h differences are independent with scale parameters that follow the same distribution as the order-k differences. For most situations, the order of k will be less than 4, so issues of scale introduced by assuming the same distribution on the scale parameters for the lower and higher order differences will be minimal. One could alternatively adjust the scale parameter of each $p(\alpha_h \,|\, \lambda)$ to impose smaller variance for lower order differences.

For the remainder of the paper we investigate two specific forms of shrinkage priors: the Laplace and the horseshoe. We later compare the performance of these two priors to the case where the order-k differences follow identical normal distributions. The following provides specific descriptions of our shrinkage prior formulations.

Laplace. As we showed previously, this prior arises naturally from an L_1 penalty, making it the default prior for Bayesian versions of the lasso (Park and Casella, 2008) and trend filter. The Laplace distribution is leptokurtic and features high mass near zero and exponential tails (Figure 3.1). Various authors have investigated its shrinkage properties (Griffin et al., 2010; Kyung et al., 2010; Armagan et al., 2013). We allow the order-k differences $\Delta^k\theta_j$ to follow a Laplace distribution conditional on a global scale parameter $\gamma = 1/\lambda$, and we allow γ to follow a half-Cauchy

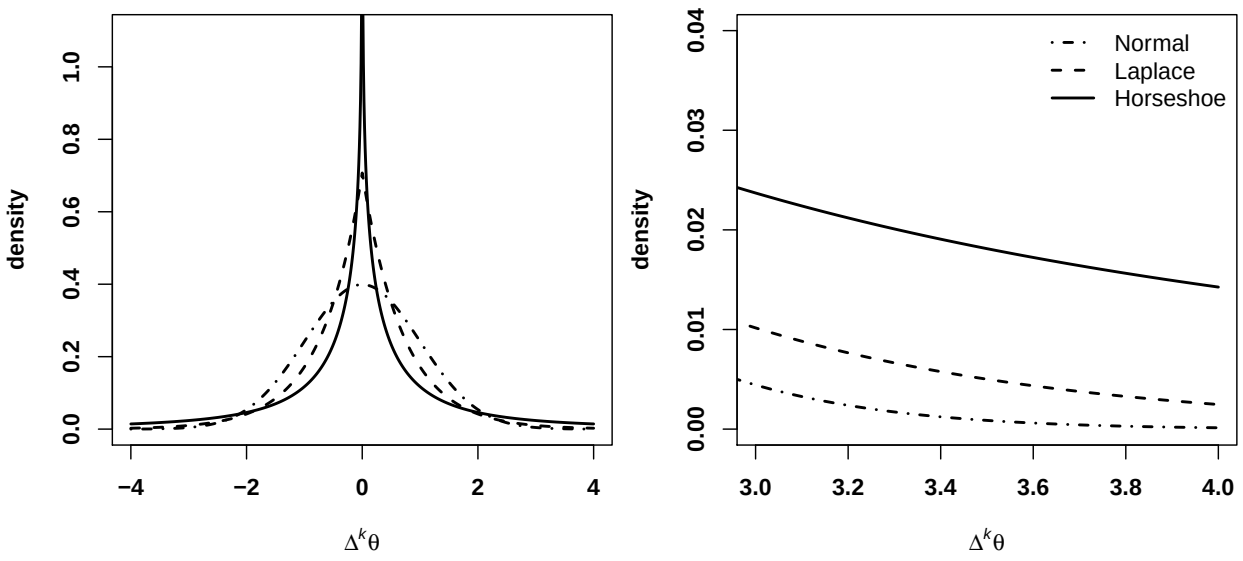

Figure 3.1: Shapes of prior distributions (left) and associated tail behavior (right) for priors used for $p(\Delta^k\theta \mid \lambda)$.

distribution with scale parameter ζ. That is,

$$\Delta^k\theta_j \mid \gamma \sim \text{Laplace}(\gamma), \qquad \gamma \sim \text{C}^+(0, \zeta). \tag{3.4}$$

The use of a half-Cauchy prior on γ is a departure from Park and Casella (2008), who make λ^2 follow a gamma distribution to induce conjugacy in the Bayesian lasso. We chose to use the half-Cauchy prior on γ because its single parameter simplifies implementation, it has desirable properties as a prior on a scale parameter (Gelman et al., 2006; Polson and Scott, 2012a), and it allowed us to be consistent across methods (see horseshoe specification below). The hierarchical form of the Laplace prior arises when the mixing distribution on the square of the local scale parameter τ_j is an exponential distribution. Specifically, we specify $\tau_j^2 \mid \lambda \sim \text{Exp}(\lambda^2/2)$ and $\Delta^k\theta_j \mid \tau_j \sim \text{N}(0, \tau_j^2)$ in the hierarchical representation.

Horseshoe. The horseshoe prior (Carvalho et al., 2010) has an infinite spike in density at zero but also exhibits heavy tails (Figure 3.1). This combination results in excellent performance as a shrinkage prior (Polson and Scott, 2010), and gives the horseshoe shrinkage properties more

similar to the spike-and-slab variable selection prior than those of the Laplace prior. We allow the order-k differences $\Delta^k\theta_j$ to follow a horseshoe distribution conditional on global scale parameter $\gamma = 1/\lambda$, and allow γ to follow a half-Cauchy distribution with scale parameter ζ. That is,

$$\Delta^k\theta_j \mid \gamma \sim \text{HS}(\gamma), \qquad \gamma \sim \text{C}^+(0, \zeta). \tag{3.5}$$

The horseshoe density function does not exist in closed form, but we have derived an approximate closed-form solution using the known function bounds (see Appendix A.1), which could be useful for application in some settings. Carvalho et al. (2010) represent the horseshoe density hierarchically as a scale mixture of normals where the local scale parameters τ_j are distributed half-Cauchy. In our hierarchical version, the latent scale parameter $\tau_j \mid \gamma \sim \text{C}^+(0, \gamma)$ and then conditional on τ_j the distribution on the order-k differences is $\Delta^k\theta_j \mid \tau_j \sim \text{N}(0, \tau_j^2)$.

The horseshoe prior arises when the mixing distribution on the local scale parameter τ_j is half-Cauchy, which is a special case of a half-t-distribution where degrees of freedom (df) equal 1. Setting $df > 1$ would result in a prior with lighter tails than the horseshoe, and setting $0 < df < 1$ would result in heavier tails. We tested half-t formulations with df between 1 and 5 in test scenarios, but did not find an appreciable difference in performance relative to the horseshoe. We also attempted to place a prior distribution on the df parameter, but found the data to be insufficient to gain information in the posterior for df in our test scenarios, so we did not pursue this further.

Normal. The normal distribution arises as a prior on the order-k differences when the penalty in the penalized likelihood formulation is a squared L_2 penalty. The normal prior is also the form of prior used in Bayesian smoothing splines. The normal is not considered a shrinkage prior and does not have the flexibility to allow locally adaptive smoothing behavior. We use it for comparison to demonstrate the local adaptivity allowed by the shrinkage priors. For our investigations, the distribution on the order-k differences and associated scale parameter is:

$$\Delta^k\theta_j \mid \gamma \sim \text{N}(0, \gamma^2), \qquad \gamma \sim \text{C}^+(0, \zeta). \tag{3.6}$$

3.2.3 Connections to Markov Random Fields

Here we briefly show the models represented by (3.2) can be expressed with GMRF priors for $\boldsymbol{\theta}$ conditional on the local scale parameters $\boldsymbol{\tau}$. It is instructive to start with the normal increments model (3.6), which belongs to a class of time series models known as autoregressive models of order k. Rue and Held (2005) call this model a k-th order random walk and show that it is a GMRF with respect to a k-th order chain graph — a graph with nodes $\{1, 2, \ldots, n\}$, where the nodes $i \neq j$ are connected by an edge if and only if $|i - j| \leq k$. Since the normal model (3.6) does not fully specify the joint distribution of $\boldsymbol{\theta}$, it is an intrinsic (improper) GMRF. We make it a proper GMRF by specifying a prior density of the first k components of $\boldsymbol{\theta}$, $p(\theta_1, \ldots, \theta_k)$. The Markov property of the model manifests itself in the following factorization:

$$p(\boldsymbol{\theta}) = p(\theta_1, \ldots, \theta_k)p(\theta_{k+1} \mid \theta_1, \ldots, \theta_k) \cdots p(\theta_n \mid \theta_{n-1}, \ldots, \theta_{n-k}).$$

Equipped with initial distribution $p(\theta_1, \ldots, \theta_k)$, models (3.5) and (3.4) also admit this factorization, so they are k-th order Markov, albeit not Gaussian models. However, if we condition on the latent scale parameters $\boldsymbol{\tau}$, both the Laplace and horseshoe models become GMRFs, or more specifically k-th order normal random walks. One important feature of these random walks is that each step in the walk has its own precision. To recap, under prior specifications (3.5) and (3.4) $p(\boldsymbol{\theta} \mid \gamma)$ is a non-Gaussian Markov field, while $p(\boldsymbol{\theta} \mid \boldsymbol{\tau}, \gamma) = p(\boldsymbol{\theta} \mid \boldsymbol{\tau})$ is a GMRF.

Our GMRF point of view is useful in at least three respects. First, GMRFs with constant precision have been used for nonparametric smoothing in many settings (see Rue and Held (2005) for examples). GMRFs with nonconstant precision have been used much less frequently, but one important application is to the development of adaptive smoothing splines by allowing order-k increments to have nonconstant variances (Lang et al., 2002; Yue et al., 2012). The approach of these authors is very similar to our own but differs in at least two important ways. First, we specify the prior distribution on the latent local scale parameters τ_j with the resulting marginal distribution of $\Delta^k\theta_j$ in mind, such as the Laplace or horseshoe distributions which arise as scale mixtures of normals. This allows a better understanding of the adaptive properties of the resulting marginal prior in advance of implementation. In contrast, Lang et al. (2002) and Yue et al. (2012) appear

to choose the distribution on local scale parameters based on conjugacy and do not consider the effect on the marginal distribution of $\Delta^k \theta_j$. Second, we allow the local scale parameters τ_j to be independent, whereas Lang et al. (2002) and Yue et al. (2012) impose dependence among the scale (precision) parameters by forcing them to follow another GMRF. Allowing the local scale parameters to be independent allows the model to be more flexible and able to adapt to jumps and sharp local features. We should also note that Rue and Held (2005) in section 4.3 show that the idea of scale mixtures of normal distributions can be used with GMRFs to generate order-k differences which marginally follow a t-distribution by introducing latent local scale parameters. Although they do not pursue this further, we mention it because it bears similarity to our approach.

Second, viewing the SPMRF models as conditional GMRFs allows us to utilize some of the theoretical results and computational methods developed for GMRFs. In particular, one can take advantage of more complex forms of precision matrices such as circulant or seasonal trend matrices (see Rue and Held (2005) for examples). One can also employ the computational methods developed for sparse matrices, which speed computation times (Rue, 2001; Rue and Held, 2005). We note that simple model formulations such as the kth-order random walk models can be coded with state-space formulations based on forward differences, which speed computation times by eliminating the operations on covariance matrices necessary with multivariate Gaussian formulations.

A third advantage of connecting our models to GMRFs is that the GMRF representation allows us to connect our first-order Markov models to subordinated Brownian motion (Bochner, 1955; Clark, 1973), a type of Lévy process recently studied in the context of scale mixture of normal distributions (Polson and Scott, 2012b). Polson and Scott (2012b) use the theory of Lévy processes to develop shrinkage priors and penalty functions. Let us briefly consider a simple example of subordinated Brownian motion. Let W be a Weiner process, so that $W(t+s) - W(t) \sim \mathrm{N}(0, s\sigma^2)$, and W has independent increments. Let T be a subordinator, which is a Lévy process that is non-decreasing with probability 1, has independent increments, and is independent of W. The subordinated process Z results from observing W at locations $T(t)$. That is, $Z(t) = W[T(t)]$. The subordinator essentially generates a random set of irregular locations over which the Brownian

motion is observed, which results in a new process. In our hierarchical representation of Laplace and horseshoe priors *for the first order differences*, we can define a subordinator process $T_j = \sum_{i=1}^{j} \tau_i^2$, so that the GMRF $p(\boldsymbol{\theta} \mid \boldsymbol{\tau})$ can be thought of as a subordinated Brownian motion or as a realization of a Brownian motion with unit variance on the random latent irregular grid $T_1, \ldots, T_n$. The subordinated Brownian motion interpretation is not so straight forward when applied to higher-order increments, but we think this interpretation will be fruitful for extending our SPMRF models in the future. One example where this interpretation is useful is when observations occur on an irregularly spaced grid, which we explore in the following section.

3.2.4 Extension to Irregular Grids

So far we have restricted our model formulation to the case where data are observed at equally-spaced locations. Here we generalize the model formulation to allow for data observed at locations with irregular spacing. This situation arises with continuous measurements over time, space, or some covariate, or when gaps are left by missing observations.

For a GMRF with constant precision (normally distributed kth-order differences), we can use integrated Wiener processes to obtain the precision matrix (see Rue and Held (2005) and Lindgren and Rue (2008) for details). However, properly accounting for irregular spacing in our models with Laplace or horseshoe kth-order differences is more difficult. To use tools similar to those for integrated Wiener processes we would need to show that the processes built on Laplace and horseshoe increments maintain their distributional properties over any subinterval of a continuous measure. Polson and Scott (2012b) show that processes with Laplace or horseshoe first-order increments can be represented as subordinated Brownian motion. However, to meet the necessary condition of an infinitely divisible subordinator, the subordinator for the Laplace process needs to be on the precision scale and the subordinator for the horseshoe process needs to be on the log-variance scale. Both resulting processes are Lévy process, which means they have independent and stationary increments, but the increments are no longer over the continuous measure of interest. This makes representation of these processes over continuous time difficult and development of the necessary theory is out of the scope of this paper.

Absent theory to properly address this problem, we instead start with our hierarchical model formulations and assume that conditional on a set of local variance parameters $\boldsymbol{\tau}$, we can use methods based on integrated Wiener processes to obtain the precision matrices for the latent GMRFs. This requires the assumption that local variances are constant within respective intervals between observations. Let $s_1 < s_2 < ... < s_n$ be a set of locations of observations, and let $\delta_j = s_{j+1} - s_j$ be the distance between adjacent locations. We assume we have a discretely observed continuous process and denote by $\theta(s_j)$ the value of the process at location s_j. For the first-order model and some interval $[s_j, s_{j+1}]$, we assume that conditional on local variance τ_j, $\theta(s)$ follows a Wiener process where $\theta(s_j + h) - \theta(s_j) \mid \tau_j \sim \mathrm{N}(0, h\tau_j^2)$ for all $0 \leq h \leq \delta_j$. If we let $\Delta\theta_j = \theta(s_{j+1}) - \theta(s_j)$, the resulting variance of $\Delta\theta_j$ is

$$\mathrm{Var}(\Delta\theta_j) = \delta_j\tau_j^2.$$

Note that the resulting marginal distribution of $\theta(s_j + h) - \theta(s_j)$ after integrating over τ_j is therefore assumed to be Laplace or horseshoe for all h, with the form of the marginal distribution dependent on the distribution of τ_j. We know this cannot be true in general given the properties of these distributions, but we assume it approximately holds for $h \leq \delta$.

The situation becomes more complex for higher order models. We restrict our investigations to the second-order model and follow the methods of Lindgren and Rue (2008), who use a Galerkin approximation to the stochastic differential equation representing the continuous process. The resulting formula for a second-order increment becomes

$$\Delta^2\theta_j = \theta(s_{j+2}) - \left(1 + \frac{\delta_{j+1}}{\delta_j}\right)\theta(s_{j+1}) + \frac{\delta_{j+1}}{\delta_j}\theta(s_j),$$

and the variance of a second-order increment conditional on τ_j is

$$\mathrm{Var}(\Delta^2\theta_j) = \frac{\delta_{j+1}^2(\delta_j + \delta_{j+1})}{2}\tau_j^2.$$

This adjustment of the variance results in good consistency properties for GMRFs with constant precision (Lindgren and Rue, 2008), so should also perform well over intervals with locally constant precision. We show in Appendices A.1 and A.2 that integrating over the local scale parameter τ_j maintains the distance correction as a multiplicative factor on the scale of the resulting marginal

distribution. We also provide a data example involving a continuous covariate in Appendix A.3 where we apply the methods above for irregular grids.

3.2.5 Posterior Computation

Since we have two general model formulations, marginal and hierarchical, we could use MCMC to approximate the posterior distribution of heights of our piecewise step functions, $\boldsymbol{\theta}$, by working with either one of the two corresponding posterior distributions. The first one corresponds to the marginal model formulation:

$$p(\boldsymbol{\theta}, \gamma, \boldsymbol{\xi} \mid \mathbf{y}) \propto \prod_{i=1}^{n} p(y_i \mid \theta_i, \boldsymbol{\xi}) p(\boldsymbol{\theta} \mid \gamma) p(\boldsymbol{\xi}) p(\gamma), \tag{3.7}$$

where $p(\boldsymbol{\theta} \mid \gamma)$ is a Markov field induced by the normal, Laplace, or horseshoe densities, and p(γ) is a half-Cauchy density. Note that a closed-form approximation to the density function for the horseshoe prior (see Appendix A.1) is needed for the marginal formulation using the horseshoe. The second posterior corresponds to the hierarchical model with latent scale parameters $\boldsymbol{\tau}$:

$$p(\boldsymbol{\theta}, \boldsymbol{\tau}, \gamma, \boldsymbol{\xi} \mid \mathbf{y}) \propto \prod_{i=1}^{n} p(y_i \mid \theta_i, \boldsymbol{\xi}) p(\boldsymbol{\theta} \mid \boldsymbol{\tau}) \prod_{j=1}^{n-k} p(\tau_j \mid \gamma) p(\boldsymbol{\xi}) p(\gamma), \tag{3.8}$$

where $p(\boldsymbol{\theta} \mid \boldsymbol{\tau})$ is a GMRF and the choice of $p(\tau_j \mid \gamma)$ makes the marginal prior specification for $\boldsymbol{\theta}$ correspond either to a Laplace or to a horseshoe Markov random field. Notice that the unconditional GMRF (normal prior) has only the marginal specification.

Both of the above model classes are highly parameterized with dependencies among parameters induced by differencing and the model hierarchy. It is well known that high-dimensional, hierarchical models with strong correlations among parameters can create challenges for standard MCMC samplers, such as component-wise random walk Metropolis or Gibbs updates. When faced with these challenges, random walk behavior can result in inefficient exploration of the parameter space, which can lead to poor mixing and prohibitively long convergence times. Many approaches have been proposed to deal with these issues, including block updating (Knorr-Held and Rue, 2002), elliptical slice sampling (Murray et al., 2010; Murray and Adams, 2010), the Metropolis adjusted

Langevin algorithm (MALA) (Roberts and Stramer, 2002), and Hamiltonian Monte Carlo (HMC) (?Neal, 1993, 2011). All of these approaches jointly update some or all of the parameters at each MCMC iteration, which usually improves mixing and speeds up convergence of MCMC. Among these methods, HMC offered the most practical choice due to its ability to handle a wide variety of models and its relative ease in implementation via readily available software such as `stan` (Carpenter et al., 2016). We used a modification of HMC proposed by Hoffman and Gelman (2014) which automatically adjusts HMC tuning parameters. We used the open source package `rstan` (Stan Development Team, 2015a), which provides a platform for fitting models using HMC in the `R` computing environment (R Core Team, 2014).

Even with HMC, slow mixing can still arise with hierarchical models and heavy-tailed distributions due to the inability of a single set of HMC tuning parameter values to be effective across the entire model parameter space. Fortunately this problem can often be remedied by model reparameterizations that change the geometry of the sampled parameter space. For hierarchical models, the non-centered parameterization methods described by Papaspiliopoulos et al. (2003, 2007) and Betancourt and Girolami (2015) can be useful. Non-centered parameterizations break the dependencies among parameters by introducing deterministic transformations of the parameters. The MCMC algorithm then operates directly on the independent parameters. Betancourt and Girolami (2015) discuss non-centered parameterizations in the context of HMC, and further examples of these and other reparameterization methods that target heavy-tailed distributions are provided in the documentation for `stan` (Stan Development Team, 2015b).

We note that after employing reparameterizations, HMC with stationary distribution equal to the hierarchical model posterior (3.8) had good convergence and mixing properties for each of our models and in nearly all of our numerical experiments. HMC that targeted the marginal model posterior (3.7) had fast run times and good mixing for the normal and Laplace formulations, but we could not effectively reparameterize the (approximate) marginal horseshoe distribution to remove the effects of its heavy tails, which resulted in severe mixing problems for the marginal horseshoe-based model. Therefore, in the rest of the manuscript we work with the hierarchical model posterior distribution (3.8) for all models.

For SPMRF and GMRF models, the computation time needed to evaluate the log-posterior and its gradient scales as $O(n)$, where n is the grid size. However, the hierarchical SPMRF models have approximately twice as many parameters as the GMRF or marginal SPMRF models. These hierarchical SPMRF methods are therefore slower than their GMRF counterparts. Since the computational cost of evaluating the log-posterior is only one factor determining the MCMC speed, we compared run times of the SPMRF and GMRF models on simulated and real data (see Appendix A.7). Our results show that SPMRF models are slower than GMRFs, but not prohibitively so.

We developed an `R` package titled `spmrf` which allows for easy implementation of our models via a wrapper to the `rstan` tools. The package code is publicly available at `https://github.com/jrfaulkner/spmrf`.

3.3 Simulation Study

3.3.1 Simulation Protocol

We use simulations to investigate the performance of two SPMRF formulations using the Laplace and horseshoe shrinkage priors described in Section 3.2.2 and compare results to those using a normal distribution on the order-k differences. We refer to the shrinkage prior methods as adaptive due to the local scale parameters, and the method with normal prior as non-adaptive due to the use of a single scale parameter. We constructed underlying trends with a variety of characteristics following approaches similar to those of other authors (Scheipl and Kneib, 2009; Yue et al., 2012; Zhu and Dunson, 2013). We investigated four different types of underlying trend (constant, piecewise constant, smooth function, and function with varying smoothness). The first row of Figure 3.2 shows examples of the trend functions, each illustrated with simulated normal observations centered at the function values over a regular grid. We used three observation types for each trend type where the observations were conditionally independent given the trend function values θ_i, where $i = 1, \ldots, n$. The observation distributions investigated were 1) normal: $y_i \,|\, \theta_i \sim \mathrm{N}(\theta_i, \sigma^2)$, where $\sigma = 1.5$ or $\sigma = 4.5$; 2) Poisson: $y_i \,|\, \theta_i \sim \mathrm{Pois}(\exp(\theta_i))$; and 3) binomial: $y_i \,|\, \theta_i \sim \mathrm{Binom}(m, (1 + \exp(-\theta_i))^{-1})$, where $m = 20$ for all scenarios.

Note that we constructed the function values for the scenarios with normally distributed observations so that each function would have approximately the same mean and variance, where the mean and variance were calculated across the function values realized at the discrete time points. This allowed us to specify observation variances which resulted in the same signal-to-noise ratio for each function, where signal-to-noise ratio is defined as the standard deviation of function values divided by the standard deviation of observations. The signal-to-noise ratios for our scenarios with normal observations were 6 for $\sigma = 1.5$ and 2 for $\sigma = 4.5$. We chose the mean sizes for the Poisson scenarios and sample sizes for the binomial scenarios so that the resulting signal-to-noise ratios would be similar to those for the normal scenarios with $\sigma = 4.5$. These levels allowed us to assess the ability of the models to adapt to local features when the signal is not overwhelmed by noise. We describe the trend functions further in what follows.

Constant. This scenario uses a constant mean across all points. We use this scenario to investigate the ability of each method to find a straight horizontal line in the presence of noisy data. The values used for the constant mean were 20 for normal and Poisson observations, and 0.5 for binomial observations.

Piecewise constant. This type of function has been used by Tibshirani (2014) and others such as Scheipl and Kneib (2009) and Zhu and Dunson (2013). The horizontal trends combined with sharp breaks offer a difficult challenge for all methods. For the scenarios with normal or Poisson observations, the function values were 25, 10, 35, and 15 with break points at $t \in \{20, 40, 60\}$. For the binomial observations the function values on the probability scale were 0.65, 0.25, 0.85, and 0.45 with the same break points as the other observation types.

Smooth trend. We use this as an example to test the ability of the adaptive methods to handle a smoothly varying function. We generated the function $\boldsymbol{f}$ as a GP with squared exponential covariance function. That is, $\boldsymbol{f} \sim \text{GP}(\mu, \boldsymbol{\Sigma}), \Sigma_{i,j} = \sigma_f^2 \exp\left[-(t_j - t_i)^2/(2\rho^2)\right]$, where $\Sigma_{i,j}$ is the covariance between points i and j, $\sigma_f^2 > 0$ is the signal variance and $\rho > 0$ is the length scale. We set $\mu = 10$, $\sigma_f^2 = 430$, and $\rho = 10$ for the scenarios with normal or Poisson observations. For binomial observations, $\boldsymbol{f}$ was generated in logit space with $\mu = -0.5$, $\sigma_f^2 = 3$, and $\rho = 10$ and then back-transformed to probability space. For all scenarios the function was generated with the same

random number seed.

Varying smoothness. This function with varying smoothness was initially presented by DiMatteo et al. (2001) and later used by others, including Yue et al. (2012). We adapted the function to a uniform grid, $t \in [1, n]$, where $n = 100$ in our case, resulting in the function

$$g(t) = \sin\left(\frac{4t}{n} - 2\right) + 2\exp\left(-30\left(\frac{4t}{n} - 2\right)^2\right).$$

For the normal and Poisson observations we made the transformation $f(t) = 20 + 10g(t)$. For binomial observations we used $f(t) = 1.25g(t)$ on the logit scale.

We generated 100 datasets for each combination of trend and observation type. This number of simulations was sufficient to identify meaningful differences between models without excessive computation time. Each dataset had 100 equally-spaced sample points over the interval $[1, 100]$. For each dataset we fit models representing three different prior formulations for the order-k differences, which were 1) normal, 2) Laplace, and 3) horseshoe. We used the hierarchical prior representations for these models given in Section 3.2.2. We selected the degree of k-th order differences for each model based on knowledge of the shape of the underlying function. We fit first-order models for the constant and piecewise constant functions, and we fit second-order models for the smooth and varying smooth functions. For the scenarios with normal observations, we set $\sigma \sim \mathrm{C}^+(0, 5)$. In all cases, $\theta_1 \sim \mathrm{N}(\mu, \omega^2)$, where μ is set to the sample mean and ω is two times the sample standard deviation of the observed data transformed to match the scale of θ. We also set $\gamma \sim \mathrm{C}^+(0, 0.01)$ for all models.

We used HMC to approximate the posterior distributions. For each model we ran four independent chains with different randomly generated starting parameter values and initial burn-in of 500 iterations. For all scenarios except for normal observations with $\sigma = 1.5$, each chain had 2,500 posterior draws post-burn-in that were thinned to keep every 5th draw. For scenarios with normal observations with $\sigma = 1.5$, chains with 10,000 iterations post-burn-in were necessary, with additional thinning to every 20th draw. In all cases, these settings resulted in 2,000 posterior draws retained per model. We found that these settings consistently resulted in good convergence properties, where convergence and mixing were assessed with a combination of trace plots, au-

tocorrelation values, effective sample sizes, and potential scale reduction statistics (Gelman and Rubin, 1992).

We assessed the relative performance of each model using three different summary statistics. We compared the posterior medians of the trend parameters ($\hat{\theta}_i$) to the true trend values (θ_i) using the mean absolute deviation (MAD):

$$\text{MAD} = \frac{1}{n}\sum_{i=1}^{n}|\hat{\theta}_i - \theta_i|. \tag{3.9}$$

We assessed the width of the 95% Bayesian credible intervals (BCIs) using the mean credible interval width (MCIW):

$$\text{MCIW} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{97.5,i} - \hat{\theta}_{2.5,i}, \tag{3.10}$$

where $\hat{\theta}_{97.5,i}$ and $\hat{\theta}_{2.5,i}$ are the 97.5% and 2.5% quantiles of the posterior distribution for θ_i. We also computed the mean absolute sequential variation (MASV) of $\hat{\boldsymbol{\theta}}$ as

$$\text{MASV} = \frac{1}{n-1}\sum_{i=1}^{n-1}|\hat{\theta}_{i+1} - \hat{\theta}_i|. \tag{3.11}$$

We compared the observed MASV to the true MASV (TMASV) in the underlying trend function, which is calculated by substituting true θ's into equation for MASV.

3.3.2 Simulation Results

In the interest of space, we emphasize results for the scenarios with normally distributed observations with $\sigma = 4.5$ here. This level of observation variance was similar to that for Poisson and binomial observations and therefore offered results similar to those scenarios. We follow these results with a brief summary of results for the other observation types, and we provide further summary of other results in Appendix A.4.

Constant. The three models performed similarly in terms of absolute value of all the metrics (Table 3.1 and Figure 3.2), but the Laplace and normal models were slightly better at fitting straight lines than the horseshoe. This is evidenced by the fact that the horseshoe had larger MCIW and

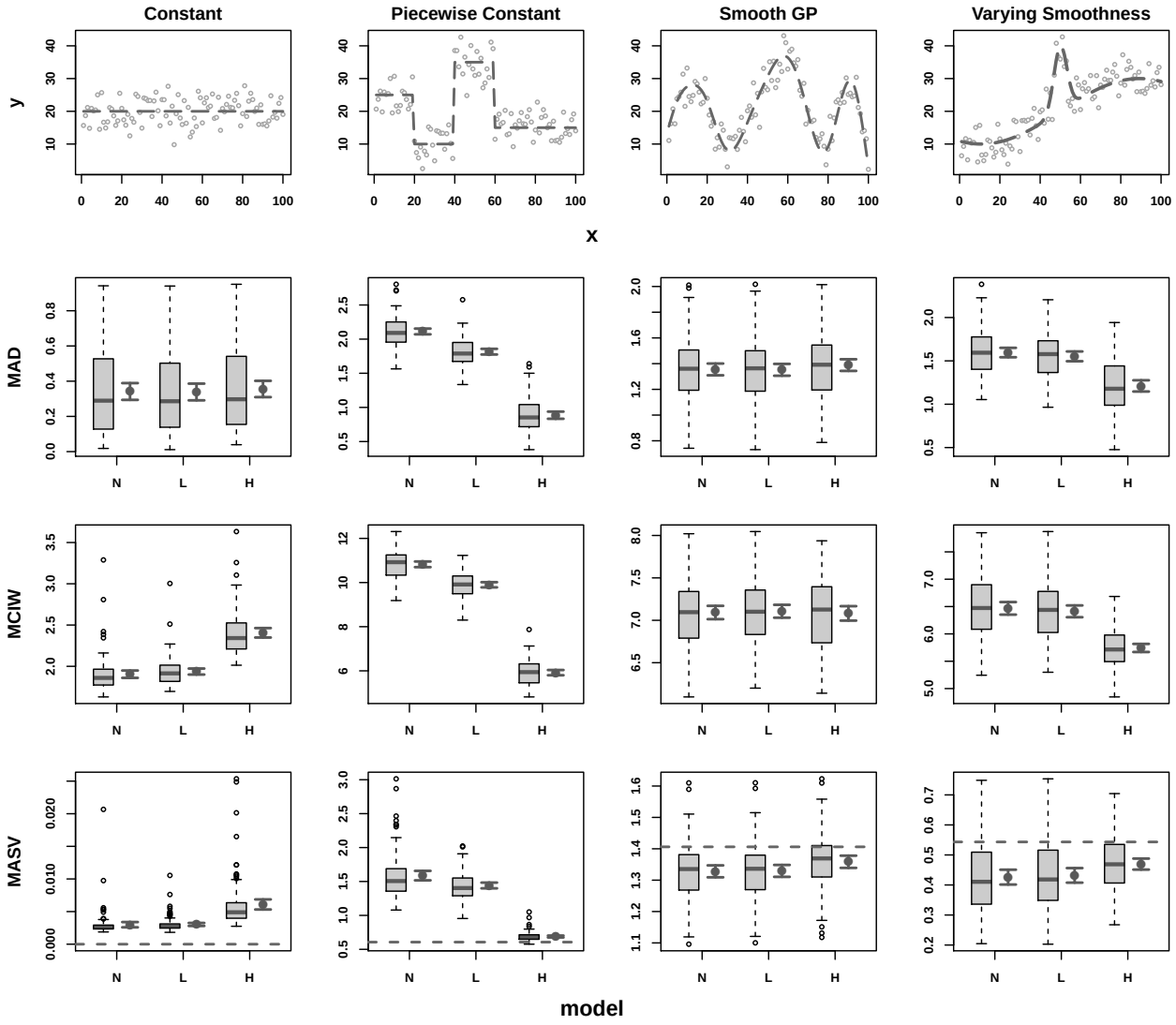

Figure 3.2: Functions used in simulations and simulation results by model (N=normal, L=Laplace, H=horseshoe) and function type for normally distributed data with $\sigma = 4.5$. Top row shows true functions (dashed lines) with example simulated data. Remaining rows show mean absolute deviation (MAD), mean credible interval width (MCIW), and mean absolute sequential variation (MASV). Horizontal dashed line in plots on bottom row is the true mean absolute sequential variation (TMASV). Shown for each model are standard boxplots of simulation results (left) and mean values with 95% frequentist confidence intervals (right).

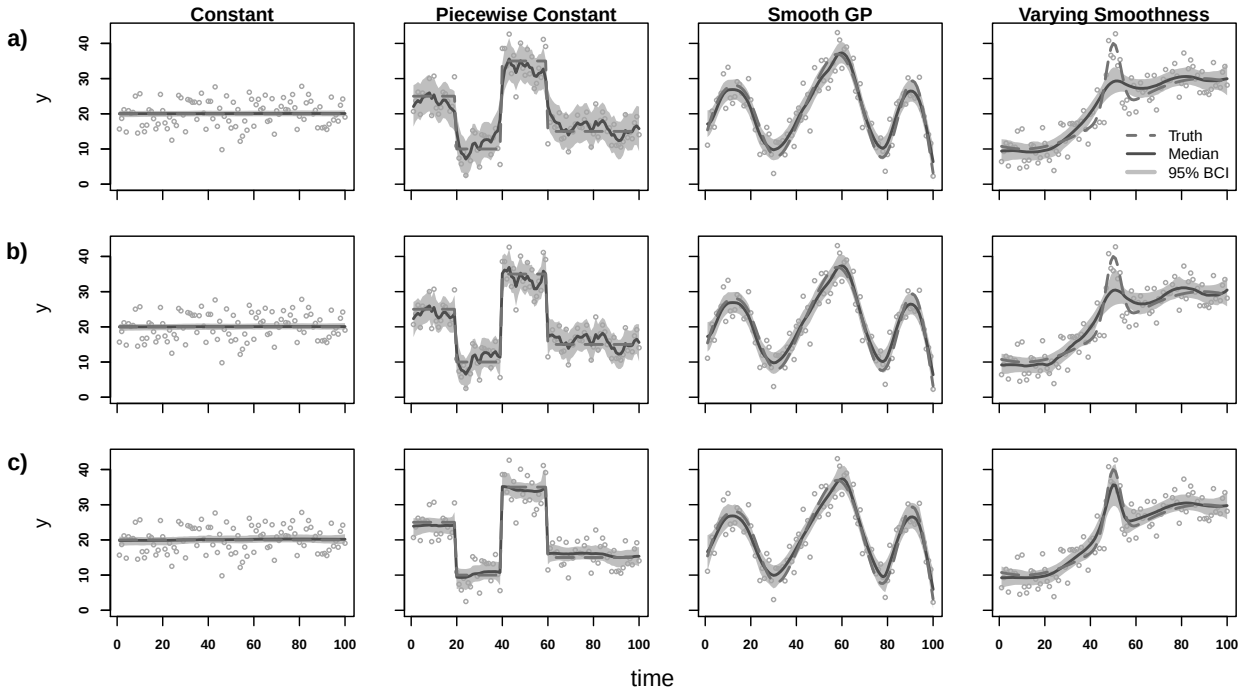

Figure 3.3: Example fits for models using a) normal, b) Laplace, and c) horseshoe priors where observations are drawn from normal distributions with SD = 4.5. Plots show true functions (dashed gray lines), posterior medians (solid dark gray lines), and associated 95% Bayesian credible intervals (BCI; gray bands) for each θ. Values between observed locations are interpolated for plotting.

larger MASV than the other methods. The first column of plots in Figure 3.3 provides a visual example of the extra variation exhibited by the horseshoe.

Piecewise constant. The horseshoe model performed the best in all categories for this scenario and the normal model performed the worst (Table 3.1 and Figure 3.2). The Laplace model was closer to the normal model in performance. The horseshoe was flexible enough to account for the large function breaks yet still able to limit variation in the constant segments. Example fits for the piecewise constant function are shown in the second column of plots in Figure 3.3.

Smooth trend. The different models were all close in value of the performance metrics for the smooth trend scenario (Table 3.1 and Figure 3.2). The normal and Laplace models had smallest MAD, but the horseshoe had MSAV closer to the true MSAV. The fact that the values of the metrics

Table 3.1: Mean values of performance measures across 100 simulations for normal observations (σ = 4.5) for each model and trend function type.

Function	**Model**	**MAD**	**MCIW**	**MASV**	**TMASV**
Constant	Normal	0.341	1.904	0.003	0.000
	Laplace	0.339	1.937	0.003	0.000
	Horseshoe	0.356	2.406	0.006	0.000
Piecewise Const.	Normal	2.112	10.826	1.587	0.606
	Laplace	1.816	9.899	1.441	0.606
	Horseshoe	0.886	5.919	0.689	0.606
Smooth	Normal	1.355	7.092	1.328	1.406
	Laplace	1.352	7.106	1.329	1.406
	Horseshoe	1.389	7.081	1.359	1.406
Varying Smooth	Normal	1.596	6.467	0.426	0.543
	Laplace	1.552	6.413	0.432	0.543
	Horseshoe	1.211	5.743	0.470	0.543

were similar for all models suggests that not much performance is lost in fitting a smooth trend with the adaptive methods in comparison to non-adaptive.

Varying Smoothness. Again the models all performed similarly in terms of absolute value of the metrics, but there was a clear ordering among models in relative performance (Table 3.1 and Figure 3.2). The horseshoe model performed the best relative to the other models on all metrics. This function forces a compromise between having large enough local variance to capture the spike and small enough local variance to remain smooth through the rest of the function. The horseshoe was more adaptive than the other two methods and therefore better able to meet the compromise. The plots in the last column of Figure 3.3 provide example fits for this function.

The results for the scenarios with normal observations with σ = 1.5 and Poisson and binomial

observations (see Appendix A.4) showed similar patterns to those with normal observations and $\sigma = 4.5$. For the constant function, the normal prior performed the best and the horseshoe prior the worst, although differences in terms of absolute values of the performance metrics were small. The relative differences were more pronounced with the scenarios with normal observations with $\sigma = 1.5$. For the piecewise constant function, the horseshoe prior performed the best for all scenarios and the normal prior the worst. All methods performed similarly for the smooth function, with the normal and Laplace generally performing a little better than the horseshoe. For the function with varying smoothness, the horseshoe performed the best and the normal the worst for all scenarios.

3.4 Data Examples

Here we provide two examples of fitting SPMRF models to real data. Each example uses a different probability distribution for the observations. The first example exhibits a change point, which makes it amenable to adaptive smoothing methods. The second example has a more uniformly smooth trend but also shows a period of rapid change, so represents a test for all methods. First we address the issue of setting the hyperparameter for the global smoothing parameter.

3.4.1 Parameterizing the Global Smoothing Prior

The value of the global smoothing parameter λ determines the precision of the marginal distributions of the order-k differences, which influences the smoothness of the estimated trend. Selection of the global smoothing parameter in penalized regression models is typically done via cross-validation in the frequentist setting (Tibshirani, 1996) or marginal maximum likelihood in the empirical Bayes setting (Park and Casella, 2008). Our fully Bayesian formulation eliminates the need for these additional steps, but in turn requires selection of the hyperparameter controlling the scale of the prior on the smoothing parameter. The value of this hyperparameter will depend on the order of the model, the grid resolution, and the variability in the latent trend parameters. Therefore, a single hyperparameter value cannot be used in all situations. Some recent studies have focused on methods for more careful and principled specification of priors for complex hier-

archical models (Fong et al., 2010; Simpson et al., 2014; Sørbye and Rue, 2014). The method of Sørbye and Rue (2014) was developed for intrinsic GMRF priors and we adapt their approach to our specific models in what follows.

We wish to specify values of the hyperparameter ζ for various situations, where the global scale parameter $\gamma \sim \mathrm{C}^{+}(0, \zeta)$. Let $\boldsymbol{Q}$ be the precision matrix for the Markov random field corresponding to the model of interest (see Appendix A.5 for examples), and $\boldsymbol{\Sigma} = \boldsymbol{Q}^{-1}$ be the covariance matrix with diagonal elements Σ_{ii}. The marginal standard deviation of all components of $\boldsymbol{\theta}$ for a fixed value of γ is $\sigma_{\gamma}(\theta_i) = \gamma\sigma_{\text{ref}}(\boldsymbol{\theta})$, where $\sigma_{\text{ref}}(\boldsymbol{\theta})$ is the geometric mean of the individual marginal standard deviations when $\gamma = 1$ (Sørbye and Rue, 2014). We want to set an upper bound U on the average marginal standard deviation of θ_i, such that $\Pr(\sigma_{\gamma}(\theta_i) > U) = \alpha$, where α is some small probability. Using the cumulative probability function for a half-Cauchy distribution, we can find a value of ζ for a given value of $\sigma_{\text{ref}}(\boldsymbol{\theta})$ specific to a model of interest and given common values of U and α by:

$$\zeta = \frac{U}{\sigma_{\text{ref}}(\boldsymbol{\theta}) \tan\left(\frac{\pi}{2}(1-\alpha)\right)}. \tag{3.12}$$

By standardizing calculations to be relative to the average marginal standard deviation, the methods of Sørbye and Rue (2014) allow us to easily calculate ζ for a model of different order or a model with a different density of grid points. For practical purposes we apply the same method to the normal and SPMRF models. This is not ideal in terms of theory, however, since the horseshoe distribution has infinite variance and the corresponding SPMRF will clearly not have the same marginal variance as a GMRF. This is not necessarily problematic since GMRF approximation will result in an estimate of ζ under the horseshoe SPMRF which is less informative than would result under similar methods derived specifically for the horseshoe SPMRF, and could therefore be seen as more conservative in terms of guarding against over smoothing. In contrast, the Laplace SPMRF has finite marginal variance that is well approximated by the GMRF methods. We apply these methods in the data examples that follow.

3.4.2 Coal Mining Disasters

This is an example of estimating the time-varying intensity of an inhomogeneous Poisson process that exhibits a relatively rapid period of change. The data are on the time intervals between successive coal-mining disasters, and were originally presented by Maguire et al. (1952), with later corrections given by Jarrett (1979) and Raftery and Akman (1986). We use the data format presented by Raftery and Akman (1986). A disaster is defined as an accident involving 10 or more deaths. The first disaster was recorded in March of 1851 and the last in March of 1962, with 191 total event times during the period 1 January, 1851 through 31 December, 1962. Visual inspection of the data suggests a decrease in rate of disasters over time, but it is unclear by eye alone whether this change is abrupt or gradual. The decrease in disasters is associated with a few changes in the coal industry at the time. A sharp decline in labor productivity at the end of the 1880's is thought to have decreased the opportunity for disasters, and the formation of the Miner's Federation, a labor union, in late 1889 brought added safety and protection to the workers (Raftery and Akman, 1986).

This data set has been of interest to various authors due to uncertainty in the timing and rate of decline in disasters and the computational challenge presented by the discrete nature of the observations. Some authors have fit smooth curves exhibiting gradual change (Adams et al., 2009; Teh and Rao, 2011) and others have fit change-point models with abrupt, instantaneous change (Raftery and Akman, 1986; Carlin et al., 1992; Green, 1995). An ideal model would provide the flexibility to automatically adapt to either scenario.

We assumed an inhomogeneous Poisson process for the disaster events and binned the event counts by year. We fit first-order models using the normal, Laplace, and horseshoe prior formulations. We assumed the event counts, y_i, were distributed Poisson conditional on the θ_i: $y_i \,|\, \theta_i \sim \text{Pois}\left(\exp(\theta_i)\right)$. The marginal prior distributions for the first-order increments were $\Delta\theta_j \sim \text{N}(0, \gamma^2)$ for the Normal, $\Delta\theta_j \sim \text{Laplace}(\gamma)$ for the Laplace, and $\Delta\theta_j \sim \text{HS}(\gamma)$ for the horseshoe. We used the same prior specifications as those used in the simulations for the remaining parameters, except we used the guidelines in Section 3.4.1 to set the hyperparameter on the global scale prior. Using calculations outlined in Appendix A.5, we set $\sigma_{\text{ref}}(\boldsymbol{\theta}) = 6.47$ and $U = 0.860$. Setting $\alpha = 0.05$ and

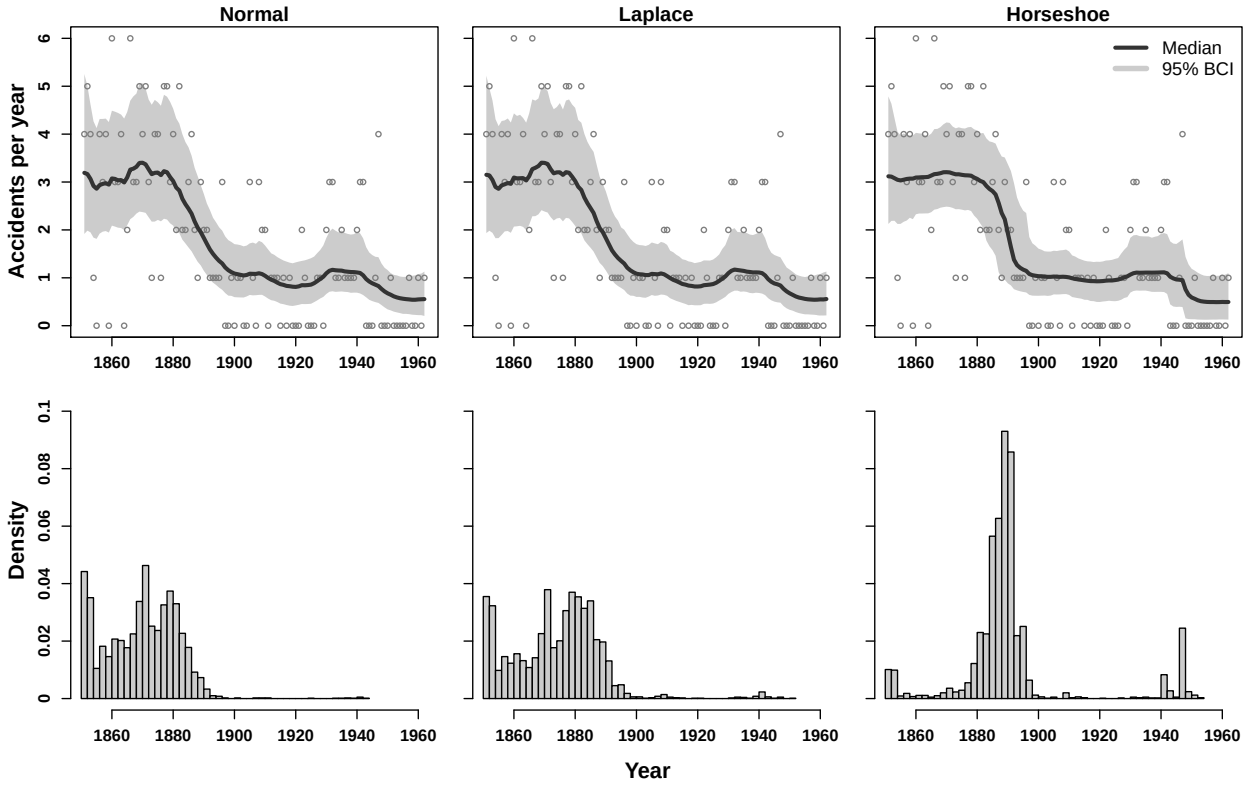

Figure 3.4: Top row: fits to coal mining disaster data for different prior distributions. Posterior medians (lines), 95% credible intervals (shaded regions), and data points are shown. Bottom row: associated posterior distributions for change points.

substituting into Equation (3.12) results in $\zeta = 0.0105$, so $\gamma \sim \mathrm{C}^{+}(0, 0.0105)$ for each model. We used HMC for approximating the posterior distributions. For each model we ran four independent chains, each with a burn-in of 500 followed by 6,250 iterations thinned at every 5. This resulted in a total of 5,000 posterior samples for each model. We were interested in finding the best representation of the process over time as well as finding the most likely set of years associated with the apparent change point. For this exercise we arbitrarily defined a change point as the maximum drop in rate between two consecutive time points.

Plots of the fitted trends (Figure 3.4) indicate that the horseshoe model picked up a sharper change in trend and had narrower BCIs than the other models. The normal and Laplace models

did not have sufficient flexibility to allow large jumps and produced a gradual decline in accidents rate, which is less plausible than a sharp decline in light of the additional information about change in coal mining industry safety regulations. The relative qualitative performance of the normal, Laplace, and horseshoe densities is similar to that for the piecewise constant scenario from our simulation study. The posterior distributions of the change point times are shown in Figure 3.4. The horseshoe model clearly shows a more concentrated posterior for the break points, and that distribution is centered near the late 1880's, which corresponds to the period of change in the coal industry. Therefore, we think the Bayesian trend filter with the horseshoe prior is a better default model in cases where sharp change points are expected.

It is important to point out that we tried other values for the scale parameter (ζ) in the prior distribution for γ and found that the models were somewhat sensitive to that hyperparameter for this data set. In particular, the horseshoe results for $\zeta = 1$ looked more like those for the other two models in Figure 3.4, but when $\zeta = 0.0001$, the horseshoe produced more defined break points and straighter lines with narrower BCIs compared to the results with $\zeta = 0.01$ (see Appendix A.6).

3.4.3 Tokyo Rainfall

This problem concerns the estimation of the time-varying mean of an inhomogeneous binomial process. We are interested in estimating the seasonal trend in daily probability of rainfall. The data are binary indicators of when daily rainfall exceeded 1 mm in Tokyo, Japan, over the course of 39 consecutive years (1951-1989). The indicators were combined by day of year across years, resulting in a sample size of $m = 39$ for each of 365 out of 366 possible days, and a size of $m = 10$ for the additional day that occurred in each of the 10 leap years. The observation variable y is therefore a count, where $y \in \{0, 1, \ldots, 39\}$. Data were obtained from the NOAA's National Center for Climate Information (https://www.ncdc.noaa.gov). A smaller subset of these data (1983-1984) was initially analyzed by Kitagawa (1987) and later by several others, including Rue and Held (2005).

We fit SPMRF models with Laplace and horseshoe priors and a GMRF model (normal prior).

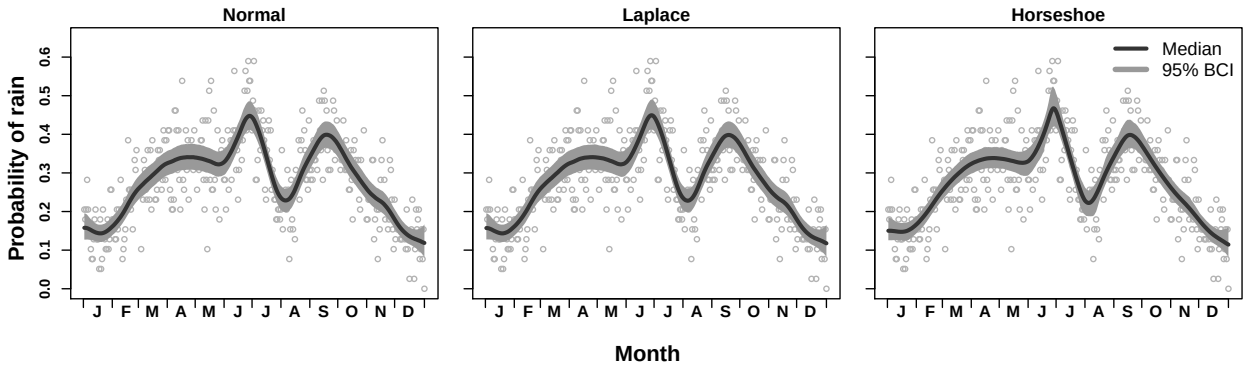

Figure 3.5: Fits to Tokyo rainfall data for different prior distributions. Posterior medians (lines), 95% credible intervals (shaded regions), and estimated probabilities (y_i/n_i) are shown.

All models were based on second-order differences. The observation model was

$$y_i \mid \theta_i \sim \text{Bin}\left(m_i, \frac{1}{1+\exp(-\theta_i)}\right),$$

and the marginal prior distributions for the second-order differences were $\Delta^2\theta_j \sim \text{N}(0, \gamma^2)$ for the normal prior, $\Delta^2\theta_j \sim \text{Laplace}(\gamma)$ for the Laplace, and $\Delta^2\theta_j \sim \text{HS}(\gamma)$ for the horseshoe. We used the same prior specifications as those used in the simulations for the remaining parameters, except we used the guidelines in Section 3.4.1 to set the hyperparameter on the global scale prior. Using calculations outlined in Appendix A.5, we set $\sigma_{\text{ref}}(\boldsymbol{\theta}) = 906.7$ and $U = 0.679$. Setting $\alpha = 0.05$ and substituting into Equation (3.12) results in $\zeta = 5.89 \times 10^{-5}$, so $\gamma \sim \text{C}^+(0, 5.89 \times 10^{-5})$ for each model. We ran four independent chains for each model, each with a burn-in of 500 followed by 6,250 draws thinned at every 5. This resulted in a total of 5,000 MCMC samples retained for each model.

The resulting function estimates for all models reveal a sharp increase in probability of rain in June followed by a sharp decrease through July and early August and a subsequent sharp increase in late August and September (Figure 3.5). Changes through the rest of the months were relatively smooth. The estimated function displays some variations in smoothness similar to the function with varying smoothness used in our simulations. All methods resulted in a similar estimated function, but the horseshoe prior resulted in a smoother function that displayed sharper features at

transition points in late June and early August, yet also had narrower credible intervals over most of the function. The normal and Laplace models resulted in a little more variability in the trend in January-April and in November. In their analysis of a subset of these data, Rue and Held (2005) used a circular constraint to tie together the endpoints of the function at the beginning and end of the year. We did not use such a constraint here, although it is possible with the SPMRF models. Even so, it is evident that the horseshoe model resulted in more similar function estimates at the endpoints than did the other two models.

3.5 Discussion

We presented a method for curve fitting in a Bayesian context that achieves locally adaptive smoothing by exploiting the sparsity-inducing properties of shrinkage priors and the smoothing properties of GMRFs. We compared the performance of the Laplace prior, which simply reformulates the frequentist trend filter to a Bayesian analog, to a more aggressive horseshoe shrinkage prior by using simulations and found that the horseshoe provided the best balance between bias and precision. The horseshoe prior has the greatest concentration of density near zero and the heaviest tails among the priors we investigated. This combination allows smooth functions to be fit in regions with weak signals or noisy data while still allowing for recovery of sharp functional changes when supported by informative data. The Laplace prior allowed more functional changes of moderate value to be retained and could not accommodate large changes without compromising the ability to shrink the noisy and smaller functional changes. This resulted in greater variability in the estimated functions and wider associated credible intervals for the models with the Laplace prior in comparison to those with the horseshoe prior when the underlying true functions had jumps or varying smoothness. The Laplace prior did have adaptive ability not possessed by the normal prior, but the horseshoe prior clearly had the best adaptive properties among the priors we investigated.

The Laplace prior performed better than the horseshoe for the constant and smooth functions in our simulations, with results closer to those of the normal prior, although the differences in performance among the three methods were relatively small. These functions do not have large deviations in order-k differences, and so there are many small or medium sized values for the estimated $\Delta^k\theta$.

This situation is reflective of cases described by Tibshirani (1996) where the lasso and ridge regression perform best, which helps explain why the analogous SPMRF models with Laplace or normal prior distributions do better here. We expect that non-adaptive or mildly adaptive methods will perform better when used on functions which do not exhibit jumps or varying smoothness. However, it is reassuring that an adaptive method does nearly as well as a non-adaptive method for these functions. This allows an adaptive model such as that using the horseshoe to be applied to a variety of functions with minimal risk of performance loss.

Our fully Bayesian implementation of the SPMRF models eliminates the need to explicitly select the global smoothing parameter λ , either directly or through selection methods such as cross-validation (e.g., Tibshirani (1996)) or marginal maximum likelihood (e.g., Park and Casella (2008)). However, the fully Bayesian approach does still require attention to the selection of the hyperparameter that controls the prior distribution on the smoothing parameter. We found the methods of Sørbye and Rue (2014) to offer practical guidelines for selecting this hyperparameter, and we successfully applied a modification of those methods in our data examples. A highly informative prior on the global smoothing parameter can result in over-smoothing if the prior overwhelms the information in the data, while a diffuse prior may result in a rougher function with insufficient smoothing. Noisier data are therefore more sensitive to choice of parameterization of the prior on the global smoothing parameter. We tested prior sensitivity in the coal mining example and found that the horseshoe prior was more responsive to changes in hyperparameter values than the normal and Laplace priors (see Appendix A.6). However, the results for the simulations and for the Tokyo rainfall example were much more robust to the value of the hyperparameter on the global scale due to the information in the data. As a precaution, we recommend first applying methods such as those by Sørbye and Rue (2014) to set the hyperparameter, but then also paying attention to prior sensitivity when analyzing noisy data with the SPMRF models.

We only addressed one-dimensional problems here, but we think the GMRF representation of these models can allow extension to higher dimensions such as the spatial setting by incorporating methods used by Rue and Held (2005) and others. We also plan to extend these methods to semi-parametric models that allow additional covariates.

Chapter 4

HORSESHOE-BASED BAYESIAN NONPARAMETRIC ESTIMATION OF EFFECTIVE POPULATION SIZE TRAJECTORIES

4.1 Introduction

Estimation of population sizes and population dynamics over time is an important task in ecology and epidemiology. Census population sizes can be difficult to estimate due to infeasible sampling requirements or study costs. Genetic sequences are a growing source of information that can be used to infer past population sizes from the signatures of genetic diversity. Phylodynamics is a discipline that uses genetic sequence data to estimate past population dynamics. Many phylodynamic models draw on coalescent theory (Kingman, 1982; Griffiths and Tavaré, 1994), which provides a probabilistic framework that connects the branching times of a genealogical tree with the effective population size and other demographic variables, such as migration rates, of the population from which the genealogy was drawn. Effective population size can be interpreted as a measure of genetic diversity in a population and is proportional to census population size if coalescent model assumptions are met. When genetic diversity is high, the effective population size approaches the census population size, given random mating and no inbreeding or genetic drift, but is otherwise smaller than the census size. In our work we concentrate on estimation of effective population sizes over evolutionary time, which can be short for rapidly evolving virus populations and longer (but still estimable with preserved ancient molecular sequence samples) for more slowly-evolving organisms. Some examples of successful application of phylodynamics include describing seasonal trends of influenza virus spread around the world (Rambaut et al., 2008), quantifying dynamics of outbreaks like hepatitis C (Pybus et al., 2003) and Ebola viruses (Alizon et al., 2014; Volz and Pond, 2014), and assessing the effects of climate change on populations of large mammals during the ice ages using ancient DNA (Shapiro et al., 2004; Lorenzen et al., 2011).

Some approaches to phylodynamics use parametric functional relationships to describe effective population size trajectories (*e.g.,* Pybus et al., 2003; Rasmussen et al., 2014), but nonparametric methods offer a flexible alternative when an accurate estimate of a complex population size trajectory is needed and knowledge of the mechanisms driving population size changes is incomplete. Nonparametric models have a long history of use in inferring effective population size trajectories. Pybus et al. (2000) introduced a nonparametric method, called the skyline plot, that

produced point-wise estimates of population size, where the number of estimates was equal to the number of sampled genetic sequences minus one. The estimates from this method were highly variable, so a modification, referred to as the generalized skyline plot, created a set of discrete time interval groups that shared a single effective population size (Strimmer and Pybus, 2001). These likelihood-based approaches were adapted to a Bayesian framework with the Bayesian skyline plot (Drummond et al., 2005) and the variable-knot spline approach of Opgen-Rhein et al. (2005). Minin et al. (2008) provided an alternative to these change-point methods by introducing a Gaussian Markov random field (GMRF) smoothing prior that connected the piecewise-constant population size estimates between coalescent events without needing to specify or estimate knot locations. Palacios and Minin (2012) and Gill et al. (2013) extended the GMRF approach of Minin et al. (2008) by constructing a GMRF prior on a discrete uniform grid. A grid-free approach, introduced by Palacios and Minin (2013), allowed the population size trajectories to vary continuously by using a Gaussian process (GP) prior.

Modern nonparametric Bayesian methods offer the state-of-the-art for recovering effective population size trajectories of unknown form. However, current methods cannot accurately recover trajectories that exhibit challenging features such as abrupt changes or varying levels of smoothness. Such features may arise in populations in the form of bottlenecks, rapid population changes, or aperiodic fluctuations with varying amplitudes. Accurate estimation of features like these can be important for understanding the demographic history of a population. Outside of phylodynamics, various nonparametric statistical methods have been developed to deal with such nonstationary or locally-varying behavior under more standard likelihoods. These methods include, but are not limited to, GPs with nonstationary covariance functions (Paciorek and Schervish, 2006), nonstationary process convolutions (Higdon, 1998; Fuentes, 2002), non-Gaussian Matérn fields (Wallin and Bolin, 2015), and adaptive smoothing splines (Yue et al., 2012, 2014). Each of these methods has good qualities and could potentially be adapted for inferring effective population sizes, but methods based on continuous random fields or process convolutions can be computationally challenging for large data sets, and some spline methods require selection or modeling of the number and location of knots.

A recent method by Faulkner and Minin (2018) uses shrinkage priors in combination with Markov random fields to perform nonparametric smoothing with locally-adaptive properties. This is a fully Bayesian method that does not require the use of knots and avoids the costly computations of inverting dense covariance matrices. Computations instead take advantage of the sparsity in the precision matrix of the Markov random field to avoid matrix inversion. Faulkner and Minin (2018) compared different specifications of their shrinkage prior Markov random field (SPMRF) models and found that putting a horseshoe prior on the kth order differences between successive function values had superior performance when applied to underlying functions with sharp breaks or varying levels of smoothness. We refer to the model with the horseshoe prior as a horseshoe Markov random field (HSMRF).

In this paper, we propose an adaptation of the HSMRF approach of Faulkner and Minin (2018) for use in phylodynamic inference with coalescent priors. We devise a new MCMC scheme for the model that uses efficient, tuning-parameter-free, high-dimensional block updates. We provide an implementation of this MCMC in the program `RevBayes`, which allows us to target the joint distribution of genealogy, evolutionary model parameters, and effective population size parameters. We also develop a method for setting the hyperparameter on the prior for the global shrinkage parameter for coalescent data. We use simulations to compare the performance of the HSMRF model to that of a GMRF model and show that our model has lower bias and higher precision across a set of population trajectories that are difficult to estimate. We then apply our model to two real data examples that are well-known in the phylodynamics literature and compare its performance to other popular nonparametric methods. The first example reanalyzes epidemiological dynamics of hepatitis C virus in Egypt and the second looks at estimation of ancient bison population size changes from DNA data.

4.2 Methods

4.2.1 Sequence Data and Substitution Model

Suppose we have a set of n aligned RNA or DNA sequences for a set of L sites within a gene. We assume the sequences come from a random sample of n individuals from a well-mixed population, where samples were collected potentially at different times. Let $\mathbf{Y}$ be the $n \times L$ sequence alignment matrix. We assume the sites are fully linked with no recombination possible between the sequences. This allows us to assume the existence of a genealogy $\boldsymbol{g}$, which is a rooted bifurcating tree that describes the ancestral relationships among the sampled individuals.

We assume that $\mathbf{Y}$ is generated by a continuous time Markov chain (CTMC) substitution model that models the evolution of the discrete states (*e.g.*, A,C,T,G for DNA) along the genealogy $\boldsymbol{g}$ for each alignment site. A variety of substitution models are available and are typically differentiated by the form of the transition matrix $M(\boldsymbol{\Omega})$, which controls the substitution rates in the CTMC for the nucleotide bases with a set of parameters $\boldsymbol{\Omega}$ (see Yang (2014) for examples). Let the likelihood of the sequence data given the genealogy and substitution parameters be denoted by $p(\mathbf{Y} \mid \boldsymbol{g}, \boldsymbol{\Omega})$. This is often referred to as the Felsenstein likelihood and can be efficiently computed using Felsenstein's pruning algorithm (Felsenstein, 1981).

4.2.2 Coalescent

Suppose that we now have a genealogy $\boldsymbol{g}$, where branch lengths of the genealogical tree are measured in units of clock time (*e.g.*, years). To build a Bayesian hierarchical model, we need a prior density for $\boldsymbol{g}$. The times at which two lineages merge into a common ancestor on the tree are called coalescent times. The coalescent model provides a probabilistic framework for relating the coalescent times in the sample to the effective size of the population. Kingman (1982) developed the coalescent model for a constant effective population size and Griffiths and Tavaré (1994) extended it for varying effective population sizes.

Let the $n-1$ coalescent times arising from genealogy $\boldsymbol{g}$ be denoted by $0 < t_{n-1} < \cdots < t_1$, where 0 is the present and time is measured backward from there. We will assume the general

case where sampling of the genetic sequences occurs at different times (*heterochronous sampling*), which will include the special case where all sampling occurs at time 0 (*isochronous sampling*). We denote the set of unique sampling times as $s_m = 0 < s_{m-1} < \cdots < s_1 < t_1$ for samples of size $n_m, \ldots, n_1$, respectively, where $n = \sum_{j=1}^{m} n_j$ and we assume no sample times are equal to coalescent times (Figure 3.4). We let $\boldsymbol{s}$ denote the vector of sampling times. Further, we let the intervals that end with a coalescent event be denoted $I_{0,k} = (\max\{t_{k+1}, s_j\}, t_k]$, for $s_j < t_k$and $k = 1, \ldots, n-1$, and let the intervals that end with a sampling event be denoted $I_{i,k} = (\max\{t_{k+1}, s_{j+i}\}, s_{j+i-1}]$, for $s_{j+i-1} > t_{k+1}$ and $s_{j+i} < t_k$, $k = 1, \ldots, n-1$. For $k = n-1$, we substitute $t_{k+1} = 0$. We let $n_{i,k}$ be the number of lineages present in interval $I_{i,k}$ and let the vector of number of lineages be denoted $\boldsymbol{n}$. Further, we denote the number of unique sampling times in interval $(t_{k+1}, t_k]$ as m_k, where $m = 1 + \sum_{k=1}^{n-1} m_k$. The joint density of the coalescent times given $\boldsymbol{s}$ and the effective population size trajectory $N_e(t)$ can then be written as

$$\begin{aligned} p(t_1, t_2, \ldots, t_{n-1} \mid \boldsymbol{s}, \boldsymbol{n}, N_e(t)) &= \prod_{k=1}^{n-1} p(t_k \mid t_{k+1}, \boldsymbol{s}, \boldsymbol{n}, N_e(t)) \\ &= \prod_{k=1}^{n-1} \frac{C_{0,k}}{N_e(t_k)} \exp\left\{ - \sum_{i=0}^{m_k} \int_{I_{i,k}} \frac{C_{i,k}}{N_e(t)} dt \right\}, \end{aligned} \tag{4.1}$$

where $C_{i,k} = \binom{n_{i,k}}{2}$ is the coalescent factor (Felsenstein and Rodrigo, 1999). This model can be seen as an inhomogeneous Markov point process where the conditional intensity is $C_{i,k}[N_e(t)]^{-1}$ (Palacios and Minin, 2013).

Here we assume $N_e(t)$ is an unknown continuous function, so the integrals in equation (4.1) must be computed with numerical approximation techniques. We follow Palacios and Minin (2012), Gill et al. (2013), and Lan et al. (2015) and use discrete approximations of the integrals over a finite grid. We construct a regular grid, $\boldsymbol{x} = \{x_h\}_{h=1}^{H+1}$, and set the end points of the grid $\boldsymbol{x}$ such that $x_1 = 0$ and $x_{H+1} = t_1$ (Figure 4.1). This results in H grid cells and $H+1$ cell boundaries. Now for $t \in (x_h, x_{h+1}]$, we have $N_e(t) \approx \exp[\theta_h]$, where θ_h is an unknown model parameter. This implies that $\boldsymbol{\theta} = \{\theta_h\}_{h=1}^{H}$ is a piecewise-constant approximation to $f(t) = \ln[N_e(t)]$ for $t \in [s_m, t_1]$. The piecewise constant population size can be integrated analytically, leading to a discrete approximation to the likelihood in Equation (4.1). The details of this approximation are provided in Appendix B.1.

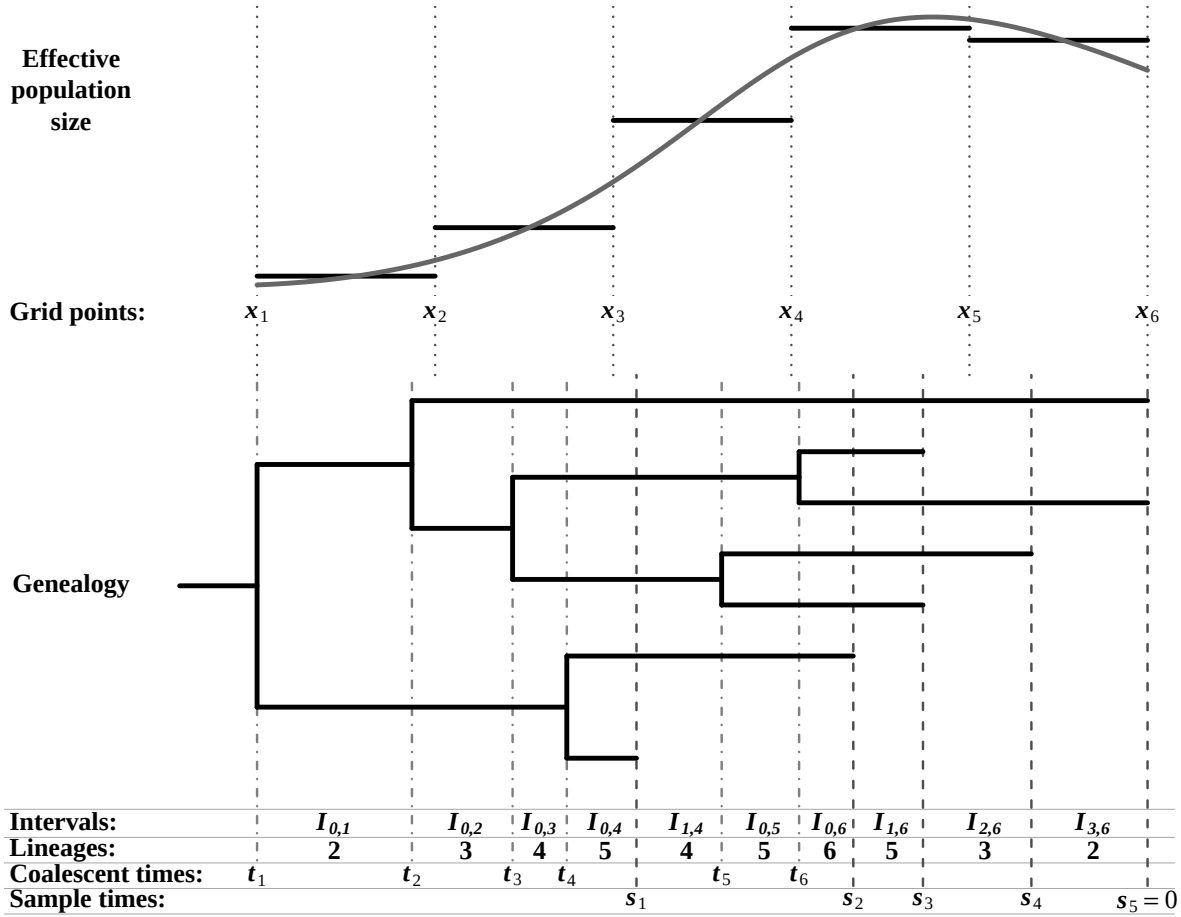

Figure 4.1: Effective population size trajectory and associated genealogical tree under heterochronous sampling. The top panel shows a continuous effective population size trajectory (gray) and an associated piecewise constant approximation to it. Also shown are the relationships between the genealogy and sampling times s_i, coalescent times t_i, intervals $I_{i,k}$, number of lineages $n_{i,k}$, and the uniform grid points, x_h, used for approximating coalescent densities.

4.2.3 *Prior for Effective Population Size Trajectory*

Next we develop a prior for the unknown function $N_e(t)$ that describes the effective population size trajectory over time. Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_H)$ be a vector of parameters that govern the effective population size trajectory $N_e(t)$. We propose using a SPMRF model (Faulkner and Minin, 2018) for $\boldsymbol{\theta}$, which is a type of Markov model where the pth-order differences in the forward-time evolution of the sequence $\{\theta_h\}_{h=1}^{H}$ are independent and follow a shrinkage prior distribution. We define the pth-order forward difference as $\Delta^p\theta_l \equiv (-1)^p \sum_{j=0}^{p}(-1)^j\binom{p}{j}\theta_{l+j-p+1}$, for $l = p, \ldots, H-1$, which is a discrete approximation to the pth derivative of $f(t)$ evaluated at t. If we assume a horseshoe distribution (Carvalho et al., 2010) as our shrinkage prior on the order-p differences in $\boldsymbol{\theta}$, then

$$\Delta^p\theta_l \mid \gamma \sim \mathcal{HS}(\gamma), \tag{4.2}$$

where the location parameter of the horseshoe distribution is zero and γ is the scale parameter and controls how much $f(t)$ is allowed to vary *a priori*. Following Carvalho et al. (2010), we put a half-Cauchy prior on γ with scale hyperparameter ζ, so that $\gamma \sim C^+(0, \zeta)$. We chose the half-Cauchy here because it has desirable properties as a prior on a scale parameter (Gelman et al., 2006; Polson and Scott, 2012a) and its single parameter simplifies implementation. Depending on the order p of the model, we also place proper priors on $\theta_1, \ldots, \theta_p$. To do this, we start by setting $\theta_1 \sim \mathcal{N}(\mu, \sigma^2)$, where μ and σ are hyperparameters typically set to create a diffuse prior. Then for $p \geq 2$ and $q = 1, \ldots, p-1$, we let $\Delta^q\theta_q \mid \gamma \sim \mathcal{HS}(a_q\gamma)$, where $a_q = 2^{-(p-q)/2}$, which follows from the recursive property and independence of the order-p differences. For example, for $p = 2$, $a_1 = 2^{-1/2}$, and for $p = 3$, $a_2 = 2^{-1/2}$ and $a_1 = 4^{-1/2}$. We will refer to this specific model formulation as a state-space formulation of a horseshoe Markov random field (HSMRF).

The horseshoe distribution is leptokurtic with an infinite spike in density at zero and Cauchy-like tails. In our setting, this combination results in small θ differences being shrunk toward zero and larger differences being maintained, which corresponds to smoothing over smaller noisy signals while retaining the ability to adapt to rapid functional changes. This is in contrast to the normal distribution, which has higher density around medium-sized values and normal tails. These attributes result in noisier estimates and reduced ability to capture abrupt functional changes. Dif-

ferent shrinkage priors will result in different levels of shrinkage and therefore different smoothing behavior. Faulkner and Minin (2018) found that the horseshoe prior performed better than the Laplace prior in terms of bias and precision for nonparametric smoothing with SPMRFs, but we do not investigate the effect of different shrinkage priors here.

The horseshoe density does not have a closed form (although see Faulkner and Minin (2018) for an approximation in closed form). However, a horseshoe distribution can be represented hierarchically as a scale mixture of normal distributions by introducing a latent scale parameter that follows a half-Cauchy distribution (Carvalho et al., 2010). That is, if $\tau_l \sim C^+(0, \gamma)$ and $\Delta^p\theta_l \mid \tau_l \sim \mathcal{N}(0, \tau_l^2)$, then integrating over τ_l results in the marginal relationship in equation (4.2).

The hierarchical HSMRF models are a type of pth-order normal random walk with separate variance parameters for each increment. The inherent Markov properties and properties of the normal distribution allow the joint distribution of $\boldsymbol{\theta}$ conditional on the vector of scale parameters $\boldsymbol{\tau}$ to be expressed $p(\boldsymbol{\theta} \mid \boldsymbol{\tau}, \mu, \sigma^2) = p(\theta_1 \mid \mu, \sigma^2)p(\Delta^1\theta_1, \ldots, \Delta^p\theta_p, \Delta^p\theta_{p+1}, \ldots \Delta^p\theta_{H-1} \mid \boldsymbol{\tau})$, which results in a multivariate normal distribution with mean $\boldsymbol{\mu}$ and precision matrix $\mathbf{Q}(\boldsymbol{\tau})$. Specifically, $\boldsymbol{\theta}$ follows a Gaussian Markov random field (GMRF; Rue and Held, 2005) conditional on $\boldsymbol{\tau}$, where the order p of the differencing in θ determines the structure of the sparse $\mathbf{Q}(\boldsymbol{\tau})$. For the models presented here, $\boldsymbol{\mu} = \mu\mathbf{1}$, where μ is a constant and $\mathbf{1}$ is a vector of ones. We specify $p(\boldsymbol{\tau})$ by assuming that the τ's are independent $C^+(0, \gamma)$-distributed random variables, where $\tau_l \sim C^+(0, \gamma)$ for $l = p, \ldots, H-1$ and $\tau_l \sim C^+(0, a_l\gamma)$ for $l = 1, \ldots, p-1$ and $p \geq 2$. The marginal joint distribution of $\boldsymbol{\theta}$ that results from integrating over $\boldsymbol{\tau}$ is a HSMRF. Note that a GMRF model results when a single scale parameter τ is used for all order-p differences in $\boldsymbol{\theta}$. For our GMRF models, we use $\tau \sim C^+(0, \zeta)$, where ζ is a fixed hyperparameter. The order of the HSMRF will determine the amount of smoothing, with higher orders resulting in more smoothing. We only consider first-order and second-order models here. In practice, we use the state-space formulation described previously but with the independent hierarchical representations of the horseshoe distributions for the individual order-p differences, which improves computational efficiency over the conditional multivariate normal representation.

4.2.4 Posterior Inference

For the case where we have a fixed genealogical tree, $\boldsymbol{g}$, which consists of sampling times $\boldsymbol{s}$ and coalescent times $\boldsymbol{t}$, the posterior distribution of the parameters $\{\boldsymbol{\theta}, \boldsymbol{\tau}, \gamma\}$ can be written as

$$p(\boldsymbol{\theta}, \boldsymbol{\tau}, \gamma \mid \boldsymbol{g}) \propto p(\boldsymbol{g} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{\tau}) p(\boldsymbol{\tau} \mid \gamma) p(\gamma). \tag{4.3}$$

Here $\boldsymbol{g}$ is considered data and we assume the coalescent times are known. Then $p(\boldsymbol{g} \mid \boldsymbol{\theta})$ is the coalescent likelihood and $p(\boldsymbol{\theta} \mid \boldsymbol{\tau}) p(\boldsymbol{\tau} \mid \gamma) p(\gamma)$ is the HSMRF prior described in Section 4.2.3. For our GMRF models, the righthand side of equation 4.3 becomes $p(\boldsymbol{g} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \tau) p(\tau)$.

For our analyses with fixed genealogical trees, we follow Faulkner and Minin (2018) and Lan et al. (2015) and use Hamiltonian Monte Carlo (HMC; Neal, 2011) for posterior inference. HMC performs joint proposals for the parameters that are typically far from the current parameter state and have high acceptance rates, resulting in efficient posterior sampling. We used the `Stan` computing environment (Carpenter et al., 2016) for implementing HMC. Specifically, we used the open source package `rstan` (Stan Development Team, 2017), which provides a platform for fitting models using HMC in the `R` computing environment (R Core Team, 2017). Our `R` package titled `spmrf` allows for easy implementation of our models for use on fixed genealogical trees via a wrapper to the `rstan` tools. The package code is publicly available at `https://github.com/jrfaulkner/spmrf`. We present a method for objectively setting the scale hyperparameter ζ of the prior distribution of the global smoothing parameter γ in Appendix B.2.

When there are genetic sequence data available and we want to jointly estimate evolutionary parameters, coalescent times, and population size trajectories, our posterior can be written as

$$p(\boldsymbol{g}, \boldsymbol{\Omega}, \boldsymbol{\theta}, \boldsymbol{\tau}, \gamma \mid \mathbf{Y}) \propto p(\mathbf{Y} \mid \boldsymbol{g}, \boldsymbol{\Omega}) p(\boldsymbol{g} \mid \boldsymbol{\theta}) p(\boldsymbol{\Omega}) p(\boldsymbol{\theta} \mid \boldsymbol{\tau}) p(\boldsymbol{\tau} \mid \gamma) p(\gamma), \tag{4.4}$$

where $\mathbf{Y}$ are the sequence data and $\boldsymbol{\Omega}$ are the parameters related to the DNA substitution model. The likelihood of the sequence data given the parameters is $p(\mathbf{Y} \mid \boldsymbol{g}, \boldsymbol{\Omega})$, and now $p(\boldsymbol{g} \mid \boldsymbol{\theta})$ is a prior for the genealogy given the population sizes and is proportional to $p(\boldsymbol{g} \mid \boldsymbol{\theta})$ in equation (4.3). The remaining components are the prior for the evolution parameters $p(\boldsymbol{\Omega})$ and the HSMRF prior as

in equation (4.3). HMC requires the calculation of gradients over continuous parameter space and therefore cannot be used for inference on discrete parameters. To the best of our knowledge, HMC has not been fully developed for inference on phylogenetic trees due to the presence of discrete parameters related to the tree configuration, although some recent progress has been made by Dinh et al. (2017). Therefore, we developed a custom MCMC algorithm that uses a combination of Gibbs sampling, elliptical slice sampling, and the Metropolis-Hastings (MH) algorithm to sample from the joint posterior of the evolution parameters and the effective population size parameters. In particular, elliptical slice sampling (Murray et al., 2010) was used to sample from the joint field of log effective population sizes conditional on the latent scale parameters, a Gibbs sampler based on an approach developed by (Makalic and Schmidt, 2016) for horseshoe random variables was used to sample the latent scale parameters conditional on the field parameters, and standard phylogenetic MH steps were used to update the genealogy and substitution model parameters. We implemented our custom MCMC in `RevBayes` — a statistical computing environment geared primarily for phylogenetic inference (Höhna et al., 2016). The standard phylogenetic MH updates mentioned above were already implemented in `RevBayes`, so we contributed a heterochronous coalescent likelihood calculator, elliptical slice sampling, and Gibbs updates of our model parameters to the `RevBayes` source code. The details of the sampling scheme are provided in the Appendix B.3 and code for implementing our methods for analyzing sequence data is available at `https://github.com/jrfaulkner/phylocode`.

4.3 Results

4.3.1 Simulated Data

We used simulated data to assess the performance of the HSMRF model relative to the GMRF model. We investigated four scenarios with different trajectories for $N_e(t)$: (1) Bottleneck (BN), (2) Boom-Bust (BB), (3) Broken Exponential (BE), and (4) Nonstationary Gaussian Process (NGP) realization. The trajectory shapes are shown at the top of Figure 4.2. For each scenario, we generated 100 data sets of coalescent times and fit GMRF and HSMRF models of first and second

order using the fixed-tree approach. The scenario descriptions and further methodological details of the simulations are provided in Appendix B.4.

We assessed the relative performance of the models using a set of summary statistics. As a measure of bias, we used the mean absolute deviation (MAD) to compare the posterior medians of the trend parameters ($\hat{\theta}_i$) to the true trend values (θ_i): MAD = $\frac{1}{H}\sum_{i=1}^{H}|\hat{\theta}_i - \theta_i|$. Smaller value of MAD indicate lower bias. We assessed the width of the 95% Bayesian credible intervals (BCIs) using the mean credible interval width (MCIW): MCIW = $\frac{1}{H}\sum_{i=1}^{H}\left(\hat{\theta}_{97.5,i} - \hat{\theta}_{2.5,i}\right)$, where $\hat{\theta}_{97.5,i}$ and $\hat{\theta}_{2.5,i}$ are the 97.5% and 2.5% quantiles of the posterior distribution for θ_i. Smaller values of MCIW indicate less uncertainty in posterior estimates of θ. We assessed the coverage of BCIs using Envelope = $\frac{1}{H}\sum_{i=1}^{H} I(\theta_i \in [\hat{\theta}_{97.5,i}, \hat{\theta}_{2.5,i}])$, where $I(\cdot)$ is the indicator function. Average Envelope values closer to the nominal coverage of 95% are considered better. To measure local variability in the estimated population trend, we used the mean absolute sequential variation (MASV) of $\hat{\boldsymbol{\theta}}$, which was computed as MASV = $\frac{1}{H-1}\sum_{i=1}^{H-1}|\hat{\theta}_{i+1} - \hat{\theta}_i|$. We compared the observed MASV to the true MASV (TMASV) in the underlying trend function, which is calculated by substituting true θ's into the equation for MASV. For a measure of model complexity, we estimated the effective number of parameters p_{eff} using an approach suggested by Raftery et al. (2006): $p_{eff} = \frac{2}{R-1}\sum_{r=1}^{R}(\mathcal{L}_r - \bar{\mathcal{L}})^2$, where $\mathcal{L}_r$ is the log-likelihood evaluated at the parameter values for the rth of R samples from the posterior, and $\bar{\mathcal{L}}$ is the mean value of $\mathcal{L}$ across the R samples. We used the Watanabe-Akaike information criterion (WAIC; Watanabe, 2010) to calculate model weights and rank model performance. The weight for model m was calculated as $w_m = \exp(-0.5\Delta\mathrm{W}_m) \,/\, \sum_{j=1}^{M}\exp\left(-0.5\Delta\mathrm{W}_j\right)$ for a set of M models, where $\Delta\mathrm{W}_m = \mathrm{WAIC}_m - \min_{j\in M}\mathrm{WAIC}_j$. We utilized the `loo` package (Vehtari et al., 2017) to calculate WAIC. For a measure of computational efficiency, we calculated the mean effective sample size (ESS) of the posterior samples across parameters for each model and simulated data set and used those with the total sampling times to calculate the mean ESS per second of sampling time.

For the BN scenario, the HSMRF model clearly had better performance than the GMRF model for the main performance metrics for both model orders (Figure 4.2 , Tables 4.1 and 4.2). Example model fits from each scenario provide some intuition for the simulation results (Figure 4.3).

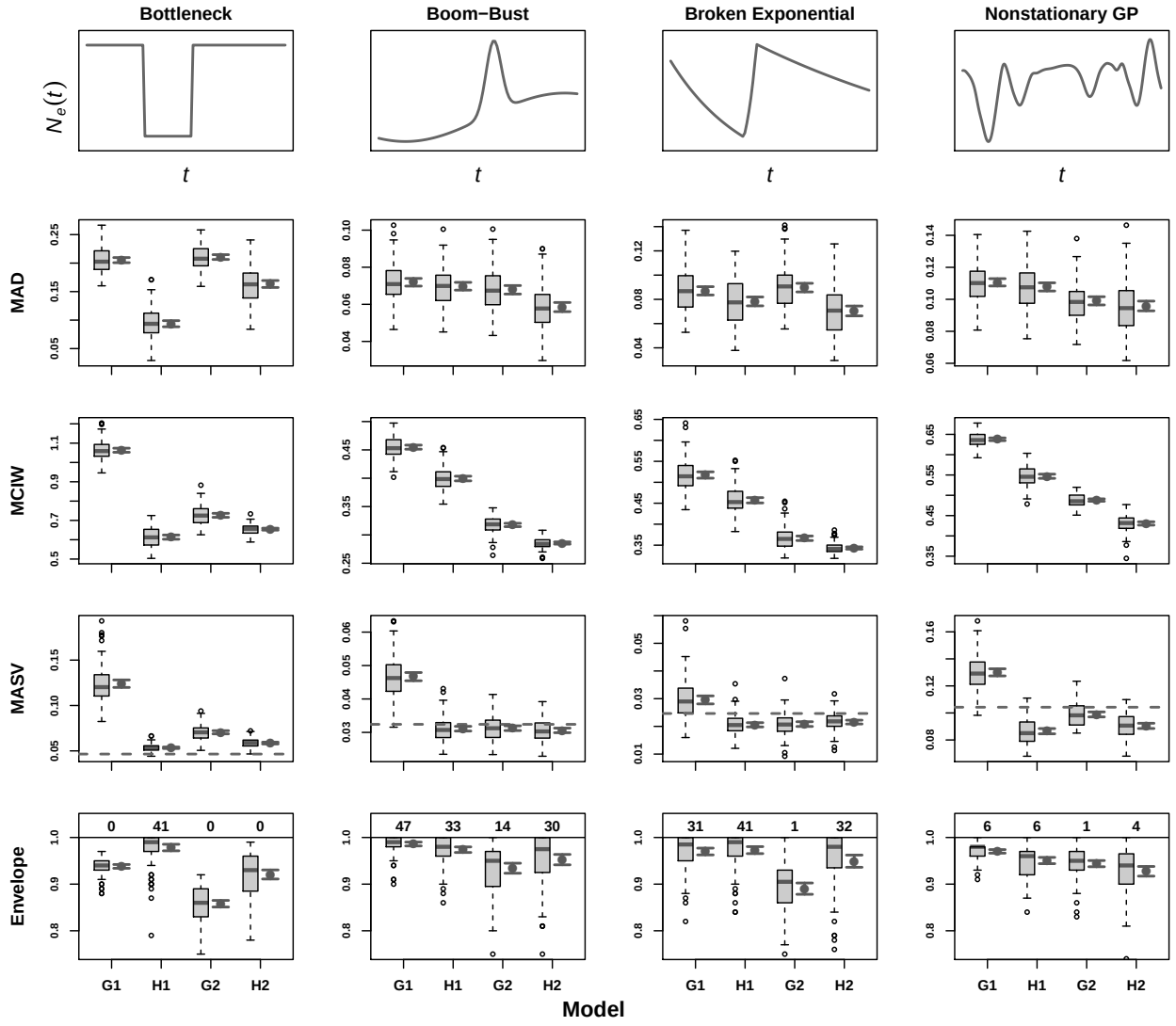

Figure 4.2: Effective population size trajectories used in simulations and simulation results by model and scenario. Models are GMRF of order 1 (G1) and order 2 (G2) and HSMRF of order 1 (H1) and order 2 (H2). Top row shows true effective population size trajectories used to simulate coalescent data. Remaining rows show mean absolute deviation (MAD), mean credible interval width (MCIW), mean absolute sequential variation (MASV), and credible interval Envelope. Horizontal dashed lines in the third row plots indicate the true mean absolute sequential variation (TMASV) values. Shown for each model are standard boxplots of the performance metrics (left) and mean values with 95% frequentist confidence intervals (right). Also shown for Envelope are the number of simulations with Envelope equal to 1.0.

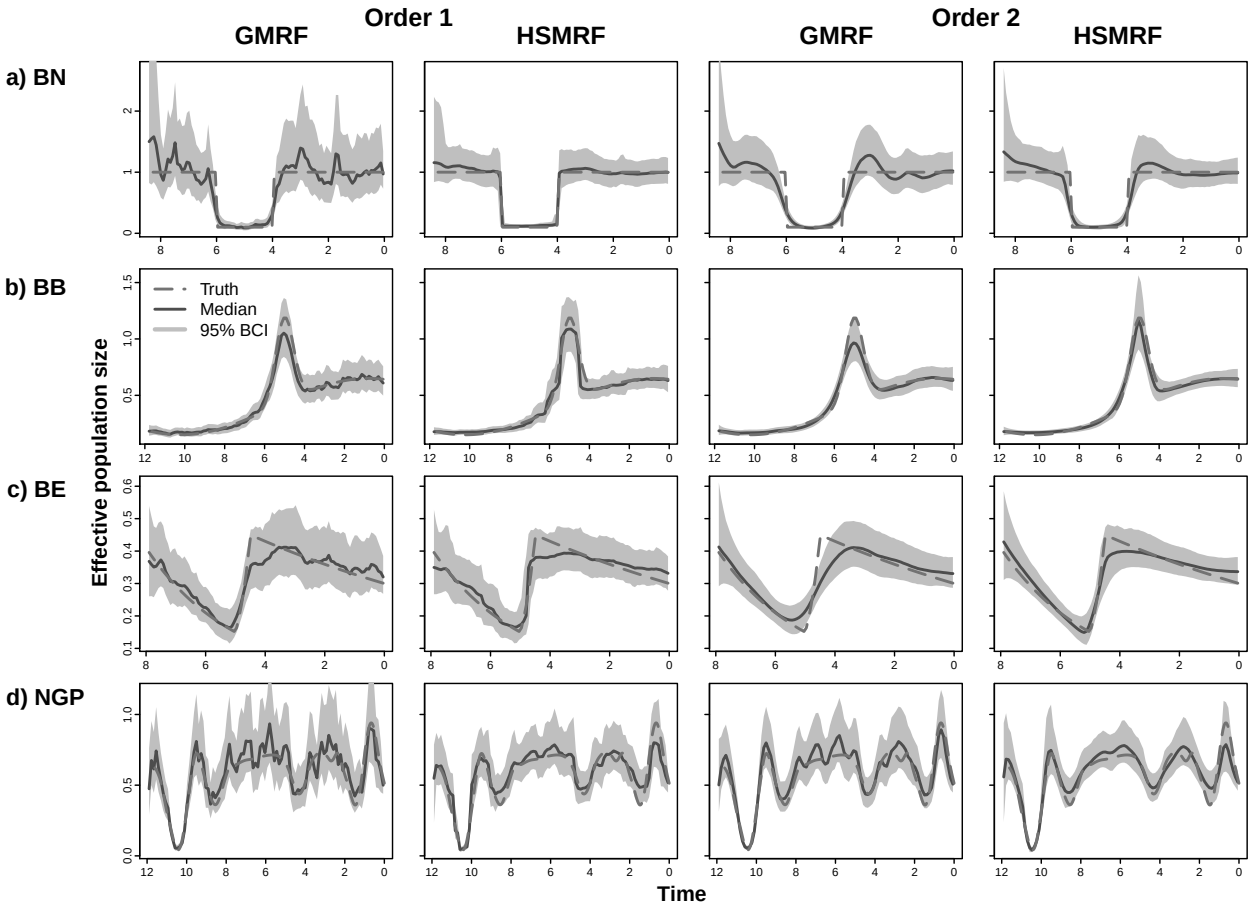

Figure 4.3: Example fits of first- and second-order Gaussian Markov random field (GMRF) and horseshoe Markov random field (HSMRF) models for four different simulation scenarios. Scenarios are a) Bottleneck (BN), b) Boom-Bust (BB), c) Broken Exponential (BE), and d) Nonstationary Gaussian Process (NGP). Results for all models within a particular scenario are for the same set of simulated data. Shown are the true effective population size trajectories that generated the data (dashed line), posterior medians of estimated trajectories (solid line) and associated 95% Bayesian credible intervals (shaded band).

Table 4.1: Mean values of performance measures across 100 simulations for each model and scenario. Scenarios are Bottleneck (BN), Boom-Bust (BB), Broken Exponential (BE), and Non-stationary Gaussian Process (NGP). Models are GMRF of order 1 (G1) and order 2 (G2) and HSMRF of order 1 (H1) and order 2 (H2). Abbreviations for measures are mean absolute deviation (MAD), mean credible interval width (MCIW), envelope (Env), mean absolute sequential variation (MASV), true MASV (TMASV), effective number of parameters (p_{eff}), and mean effective sample size (ESS). Values of MAD, MCIW, Env, MASV, and TMASV are multiplied by 100 for readability.

Scenario	**Model**	**MAD**	**MCIW**	**Env**	**MASV**	**TMASV**	p_{eff}	**ESS/s**
BN	G1	20.52	106.3	**93.8**	12.40	4.65	55.1	**8.95**
	H1	**9.37**	**61.4**	97.8	**5.34**	4.65	**18.9**	2.47
	G2	21.06	72.6	85.8	7.05	4.65	39.4	0.78
	H2	16.34	65.5	92.1	5.86	4.65	34.4	0.29
BB	G1	7.19	45.5	98.6	4.67	3.24	39.4	**9.10**
	H1	6.98	40.0	97.4	3.11	3.24	30.6	1.25
	G2	6.78	31.8	93.4	**3.13**	3.24	25.5	0.66
	H2	**5.85**	**28.6**	**95.3**	3.06	3.24	**17.8**	0.13
BE	G1	8.70	51.7	97.0	2.96	2.47	31.0	**10.22**
	H1	7.83	45.7	97.3	2.06	2.47	21.1	2.47
	G2	8.97	36.6	89.0	2.08	2.47	19.7	1.52
	H2	**7.04**	**34.3**	**94.9**	**2.16**	2.47	**15.2**	0.53
NGP	G1	11.06	63.8	97.0	13.00	10.42	64.7	**1.72**
	H1	10.78	54.7	**95.1**	8.65	10.42	61.4	0.54
	G2	9.90	48.8	94.4	**9.89**	10.42	**46.9**	0.07
	H2	**9.58**	**43.1**	92.8	9.04	10.42	50.4	0.05

First order models did better than second order models within model types for the BN scenario. Differences between model types were not as strong for the other scenarios. The second-order

Table 4.2: Summary of model selection criteria across 100 simulations by scenario and model set. WAIC weights were calculated and the best model (greatest WAIC weight) was determined for each simulated data set within each scenario and model set. Metrics shown are the percentage of simulations each model was determined best and the mean model weight across simulations. Values for each metric are compared among models within each scenario and model set. Highest percentage of best models is in bold within each scenario and model set. Scenarios are Bottleneck (BN), Boom-Bust (BB), Broken Exponential (BE), and Nonstationary Gaussian Process (NGP). Models are GMRF of order 1 (G1) and order 2 (G2) and HSMRF of order 1 (H1) and order 2 (H2).

Metric	**Model Set**	**Model**	**BN**	**BB**	**BE**	**NGP**
Best Model (%)	All Models	G1	1	9	13	1
		H1	**93**	14	34	9
		G2	0	3	1	24
		H2	6	**74**	**52**	**66**
	Order 1	G1	1	**51**	29	50
		H1	**99**	49	**71**	50
	Order 2	G2	9	9	5	27
		H2	**91**	**91**	**95**	**73**
Mean Weight	All Models	G1	0.03	0.11	0.14	0.04
		H1	0.89	0.15	0.35	0.09
		G2	0.01	0.10	0.07	0.26
		H2	0.08	0.63	0.44	0.61
	Order 1	G1	0.03	0.48	0.24	0.46
		H1	0.97	0.52	0.76	0.54
	Order 2	G2	0.12	0.11	0.11	0.43
		H2	0.88	0.89	0.89	0.57

HSMRF performed the best in terms of MAD, MCIW, and WAIC for the remaining scenarios. Among second-order models, the HSMRF was clearly favored over the GMRF in terms of WAIC across all scenarios. However, the HSMRF models were not noticeably different from the second-order GMRF in terms of MASV for the BB and BE scenarios. The second-order GMRF had mean MASV closer to TMASV than did the second-order HSMRF for the NGP scenario. Although the GMRF tended to estimate excess variation in the middle section of the trend for the NGP scenario, it did capture the peaks and troughs a little better than the HSMRF in other parts of the trend (see Figure 4.3 for an example). In all scenarios, the HSMRF had lower p_{eff} compared to the GMRF of the same order. The GMRF was consistently more computationally efficient than the HSMRF, with mean ESS/sec approximately 1.5 to 6 times higher for models of the same order. These differences are due to the additional parameters in the HSMRF models. The second-order models were relatively slow for both model types, but the HSMRF was always slower. As we show in the following data examples, however, the differences in computational speed between the HSMRF and GMRF models is negligible when genealogies and effective population size trajectories are jointly estimated.

4.3.2 Egyptian Hepatitis C Virus

The hepatitis C virus (HCV) is a blood-borne RNA virus that exclusively infects humans. HCV infection is often asymptomatic, but can lead to liver disease and liver failure. HCV infections have historically had high prevalence in Egypt (Miller and Abu-Raddad, 2010). This is thought to be due to past widespread use of unsanitary medical practices in the region. Of particular interest is a treatment for the parasite disease schistosomiasis known as parenteral antischistosomal therapy (PAT), which uses intravenous injections. PAT was practiced from the 1920's to 1980's in Egypt and is thought to have contributed to the spread of HCV during that period due to unsterilized injection equipment (Frank et al., 2000; Medhat et al., 2002).

We analyze 63 RNA sequences of type 4 with 411 base pairs from the E1 region of the HCV genome that were collected in 1993 in Egypt (Ray et al., 2000). Pybus et al. (2003) used a piecewise demographic model for effective population size with a period of exponential growth between two

periods of constant population size and concluded that the HCV population grew exponentially during the period of PAT treatment. Other authors have applied nonparametric methods to estimate the effective population size trajectory for these data (*e.g.*, Drummond et al., 2005; Minin et al., 2008; Palacios and Minin, 2013). Different nonparametric methods lead to different estimated trajectories and different levels of uncertainty. This example is interesting for the nature of the change in population size during the epidemic.

We fit six different models to these data: 1) Bayesian Skyline — a piecewise constant/linear model with estimable locations of change-points (SkyLine; Drummond et al., 2005), 2) Bayesian Skyride (SkyRide; Minin et al., 2008) 3) GMRF-1 (similar to Bayesian Skygrid; Gill et al. (2013)), 4) GMRF-2, 5) HSMRF-1, and 6) HSMRF-2. We note that the SkyRide model is also a type of GMRF model where the non-uniform grid cell boundaries are determined by coalescent events. For all six models we jointly estimated the evolutionary model parameters, genealogies, and effective population size parameters. We used the program `BEAST` implementation of the SkyLine and SkyRide models (Drummond et al., 2012), and used our own `RevBayes` implementation of the GMRF and HSMRF models. Although the Skygrid implementation of the GMRF-1 model is available in `BEAST`, the GMRF-2 and the HSMRF models are not, so we decided to use common software for the GMRF and HSMRF models. We used the same distributional forms and parameterizations for priors representing model components in common across the models. Where possible, we also attempted to use the same proposal distributions and maintain the same relative weighting of different MCMC moves in common to the models across the two software packages. We fixed the mean mutation rate to 7.9×10^{-4} substitutions/site/year, which is a value estimated by Pybus et al. (2001) and used by others for these data. We used the HKY nucleotide substitution model (Hasegawa et al., 1985) with gamma distributed rate heterogeneity and invariant sites (Yang, 1994). For the GMRF and HSMRF models we used 100 equally-spaced grid cells where the first 99 ended at a fixed boundary of 227 years before 1993, and the final cell captured any coalescent events beyond the boundary (see Appendix B.5 for discussion on setting grids). The SkyLine model requires specification of the number of discrete population intervals, where each interval describes a piecewise constant population size between two coalescent events. We used 20

population intervals to allow fair flexibility to capture sharp features in the population trajectory. Further details about the MCMC implementation and computation times are provided in Appendix B.6. For model comparison, we calculated posterior model probabilities using marginal likelihood estimates calculated with steppingstone sampling (see Appendix B.7 for details; Xie et al., 2011).

The six models differed in both the estimated trajectories and the uncertainty about those trajectories (Figure 4.4). The SkyLine and HSMRF-1 models accounted for the majority of the posterior model probability mass, with the SkyLine favored a little over the HSMRF-1 (Figure 4.4). The shape of the median trajectory from the HSMRF-1 model was similar to that of the SkyLine model, yet the HSMRF-1 model showed a very rapid increase in population between 1925 and 1945, while the SkyLine and other models showed more gradual increases that started earlier and ended later. The increase estimated by the SkyRide model lasted the longest, starting near 1900 and ending near 1970. The HSMRF and the SkyLine also showed relatively constant population size following the increase in the mid 20th century, while the SkyRide and GMRF-1 models showed a decrease after 1970. The posterior mean densities of frequencies of coalescent times provide an indication of when the HCV epidemic started (Figure 4.4). The results of the HSMRF-1 support the idea that HCV epidemic started after PAT was introduced and suggest that early PAT campaigns may have used less sanitary practices and contributed more to the spread of HCV than the major PAT campaigns started in the 1950's. Plots of the effective population trajectories covering the entire span of the coalescent times are provided with further discussion in Appendix B.8.

4.3.3 Beringian Steppe Bison

Modern molecular methods have allowed the recovery of DNA samples from specimens that lived hundreds to hundreds of thousands of years ago (Pääbo et al., 2004; Shapiro and Hofreiter, 2014). Large mammals that lived in the Northern Hemisphere during the Pleistocene and Holocene epochs have been a valuable source of this ancient DNA due to conditions favorable for specimen preservation in the northern latitudes (*e.g.,* Shapiro et al., 2004; Lorenzen et al., 2011). We focus on bison (*Bison* spp.) that lived on the steppe-tundra of Northern Asia and Europe and crossed into North America over the Bering land bridge during the middle Pleistocene (Shapiro et al., 2004). Interest

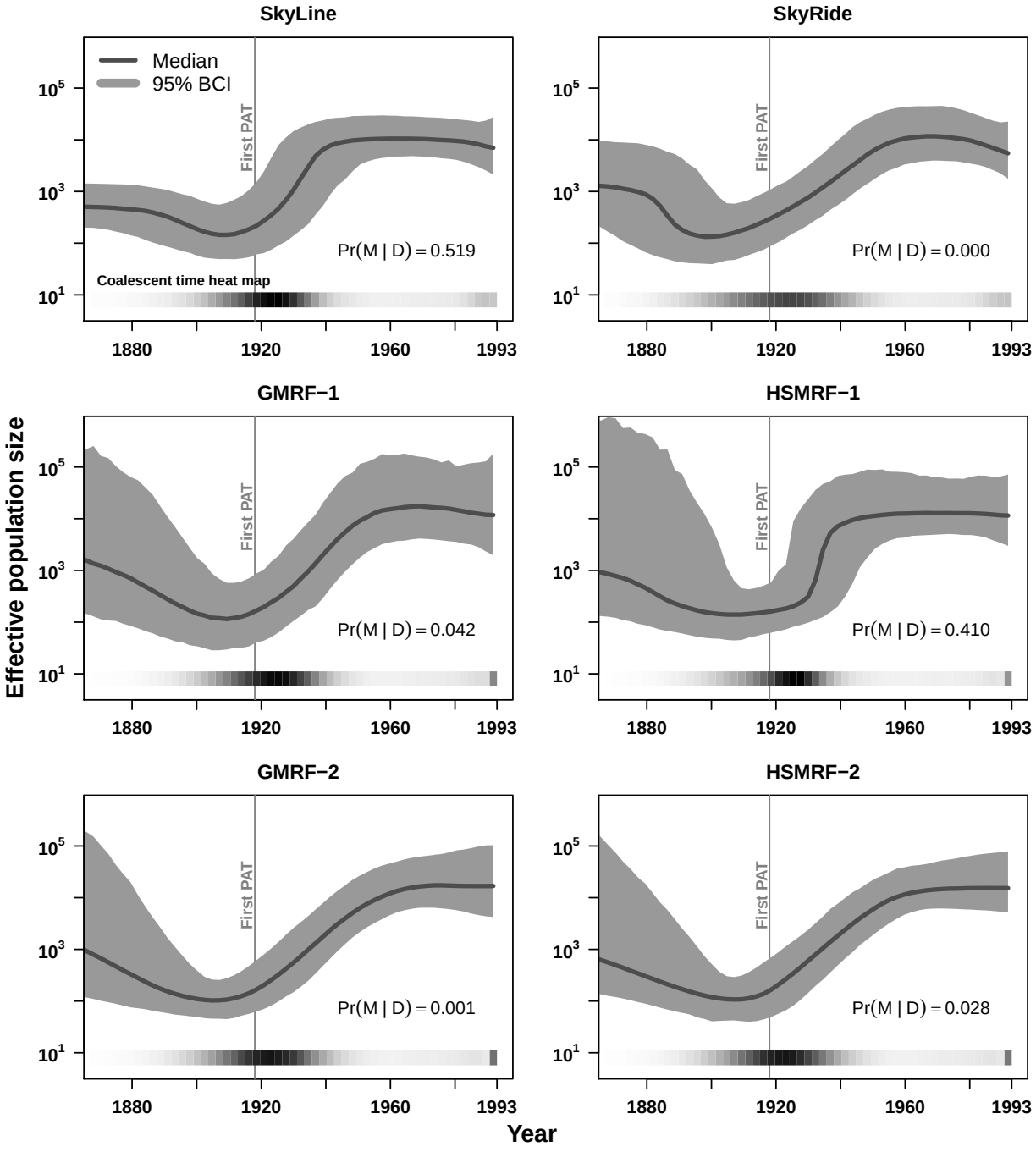

Figure 4.4: Posterior medians (solid black lines) of effective population sizes and associated 95% credible intervals (grey shaded areas) for the HCV data for the Bayesian Skyline (SkyLine), Bayesian Skyride (SkyRide), Gaussian Markov random field of order 1 (GMRF-1) and order 2 (GMRF-2), and horseshoe Markov random field of order 1 (HSMRF-1) and order 2 (HSMRF-2). Also shown for each model are posterior model probabilities (Pr(M | D)) and heat maps of mean posterior frequencies of coalescent times. A vertical reference line is shown at year 1918, which is the year PAT was introduced.

has been in determining whether human impact or climate and related habitat change were behind the decline of bison across their range during the late Pleistocene. Shapiro et al. (2004) used a parametric piecewise-exponential model for the bison effective population size and estimated that the time of transition from population growth to decline was 37 thousand years ago (kya). Drummond et al. (2005) used the more flexible SkyLine model, which indicated a more rounded and prolonged peak in population size followed by a rapid decline and bottleneck around 10 kya. Here we use a modified version of the bison data described by Shapiro et al. (2004) and fit coalescent models directly to the sequence data as with the HCV data. We make qualitative comparisons among the resulting estimated population trajectories and in relation to some benchmark times describing the arrival of humans and the period of the Last Glacial Maximum (LGM).

We analyze 152 sequences (135 ancient and 17 modern) of mitrochondrial DNA with 602 base pairs from the mitochondrial control region. DNA was extracted from bison fossils from Alaska (68), Canada (46), Siberia (13), the lower 48 United States (6), and China (2). Sample dates were estimated for the ancient samples using radiocarbon dating, with dates ranging up to 59k years. We treat the calibrated radiocarbon dates as known in the following analyses. These data are the same as those used by Gill et al. (2013), and are slightly modified from the data first described by Shapiro et al. (2004) to remove sequences identified as potentially contaminated with young radiocarbon (Shapiro et al., 2010) and include additional sequences generated since generation of the initial data set. In this data set, radiocarbon dates are calibrated to calendar time using the IntCal09 calibration curve (Reimer et al., 2009).

The LGM in the Northern Hemisphere is estimated to have occurred between 26.5 to 19 kya (Clark et al., 2009). A small, isolated population of humans existed in central Beringia, including, potentially, the land bridge that connected the continents during the LGM (Llamas et al., 2016). Humans may have ventured into eastern Beringia (Alaska and Yukon) as early as 26 kya (Bourgeon et al., 2017), but there is as yet no evidence of continuous occupation until 14 kya (Easton et al., 2011; Holmes, 2011). Humans probably first entered continental North America via a western coastal route that became available close to 16 kya (Llamas et al., 2016; Heintzman et al., 2016), where they would have encountered the population of steppe bison that were isolated in the south

with the coalescence of the Laurentide and Cordilleran glaciers (Shapiro et al., 2004; Heintzman et al., 2016). Because the majority of our bison samples were collected in North America, we used 16-14 kya as the time of first human occupation.

As with the HCV data, we used `BEAST` to fit the SkyLine and SkyRide models and used `RevBayes` to fit the GMRF and HSMRF models. We used 15 groups for the SkyLine model to match the approach used by Drummond et al. (2005), which allowed for sufficient flexibility to fit important change points without introducing computational challenges associated with more groups. To improve mixing and reduce computation time, we used a strict molecular clock with mutation rate set to 5.38×10^{-7} substitutions per year, which was based on initial runs in `BEAST` where the clock rate was estimated under a Skygrid model for the effective population size trajectory. We used the HKY nucleotide substitution model with gamma distributed rate heterogeneity. Further details on MCMC implementation and on computation times are provided in Appendix B.6. We used steppingstone sampling (Xie et al., 2011) to estimate marginal likelihoods for calculating posterior model probabilities (see Appendix B.7 for details). We also calculated posterior distributions for the time of the peak in population size.

Results indicated quite different population trajectories from the different models, but the HSMRF-1 model had the highest posterior model probability (Figure 4.5). The posterior median trajectory from the HSMRF-1 model was most similar to the SkyLine model, but the credible intervals for the HSMRF-1 model were most similar to the GMRF-1 model. The second-order models both produced strongly piecewise-linear trajectories with relatively narrow credible intervals, but had low posterior probability and were smoothing over some of the local features displayed by other models. The HSMRF-1 model displayed a more complex descent from the peak size to the present in comparison to the other models, and the areas of rapid descent are coincident with the arrival of humans in eastern Beringia and ice-free North America and the initial retreat of the glaciers, both of which are coincident with changes in habitat. All models suggested that the overall decline in population size started before the LGM, and all had median time of population peak between 41.6 and 47.3 kya, but uncertainty in the time of peak population size varied widely across the models.

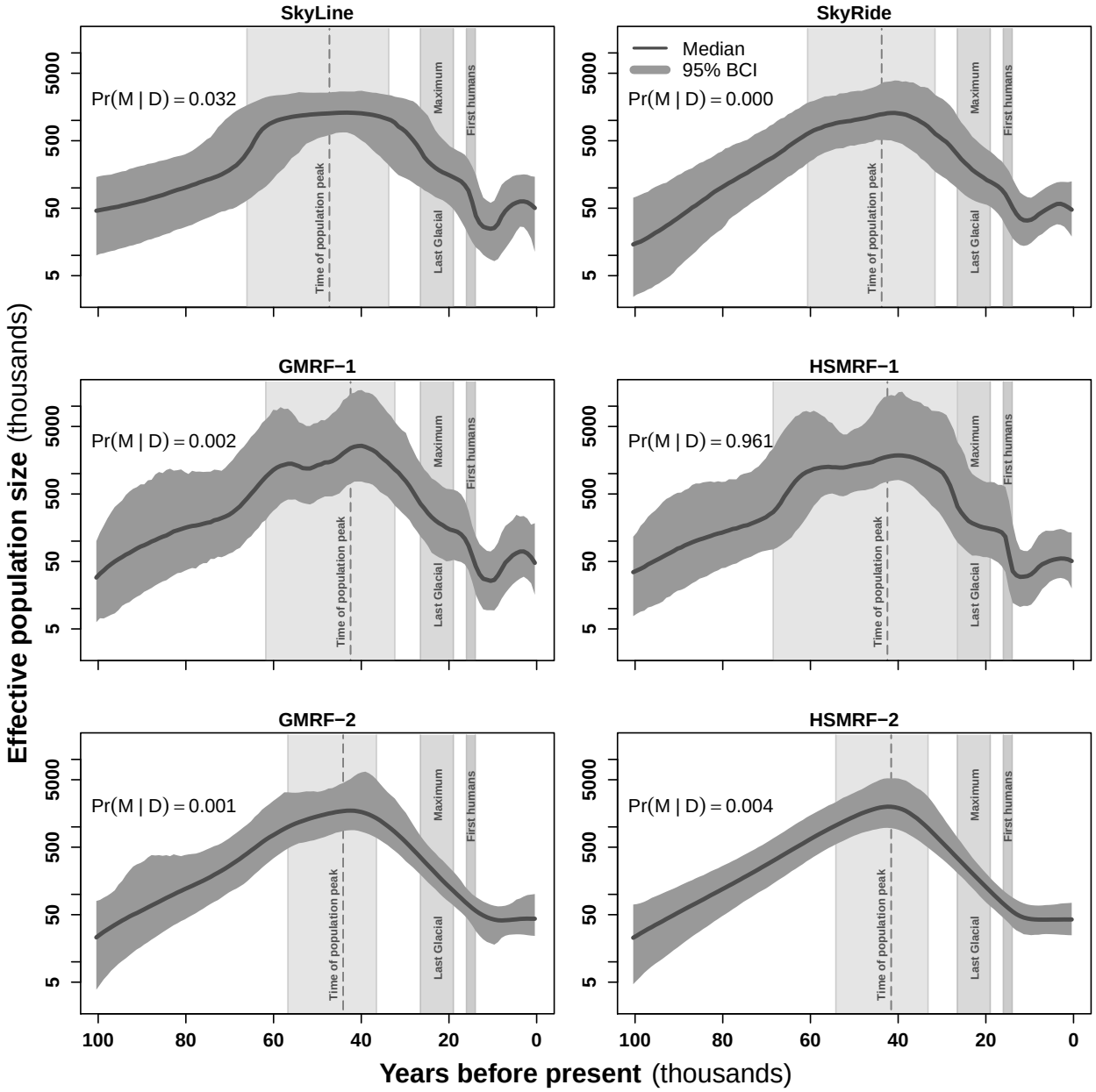

Figure 4.5: Posterior medians of effective population sizes and associated 95% credible intervals obtained from the bison DNA sequence data using the Bayesian Skyline (SkyLine), Bayesian Skyride (SkyRide), and GMRF and HSMRF models of order 1 and order 2. Also shown for each model are posterior model probabilities (Pr(M | D)) and posterior median and 95% credible intervals for the time of peak effective population size. The period of the Last Glacial Maximum and timing of first human settlement in North America are shown for reference.

4.4 Discussion

We introduced a novel and fully Bayesian method for nonparametric inference of changes in effective population size that we call the HSMRF. This method utilizes a shrinkage prior known as the horseshoe distribution, which allows more flexibility to respond to rapid changes in effective population size trajectories, yet also generates smoother trajectories in comparison to standard GMRF methods. Our simulations demonstrated that the HSMRF had lower bias and higher precision than the GMRF and was able to recover the underlying true trajectories better in most cases.

There are many situations where the local adaptivity of the HSMRF models would provide advantages over the GMRF and other models. Examples in infectious disease dynamics that could lead to rapid changes in effective population sizes include sudden changes in contact rates due to behavioral changes or quarantine, or sudden changes in the infection rate due to introduction of treatment or vaccine. At a macro-evolutionary scale, sudden changes in the effective population size could be brought on by rapid extinction events or rapid expansion due to colonization. As we have demonstrated, in situations like these the GMRF and other models tend to smooth over the sharp changes that the HSMRF can capture.

In our HCV example, the HSMRF-1 showed that HCV may have undergone a sharper increase in effective population size than previous estimates had shown. In addition, the HSMRF-estimated timing of this increase was later than the estimated timing from other methods. The later increase is more plausible in light of other epidemiological information that links increased prevalence of HCV in Egypt with mass administration of PAT (Frank et al., 2000; Medhat et al., 2002). In the bison example, the HSMRF-1 captured some features in the population trajectory that the other models did not, such as a sharp decline near the time the first humans arrival. Our results from both data examples indicated that the properties of the population size trajectories estimated by the HSMRF-1 model were somewhere between those from the GMRF-1 model and the SkyLine model. The SkyLine model is a type of change-point model, which suggests the HSMRF-1 can produce behavior of change-point models without explicitly needing to specify number or location of change points.

We demonstrated in our simulations that second-order models for either the HSMRF or GMRF formulations can perform better than first-order models in many cases. Although the second-order models did not perform as well as the first-order models in our particular data examples, they would likely do well in other examples with smoother trajectories. Among the second-order models, the HSMRF did as well or better than the GMRF for the simulated examples and had higher posterior model probabilities for both of the data examples.

Second-order models have not been used much for estimating effective population sizes previously. Palacios and Minin (2013), whose method assumes a fixed and known genealogy, tested an integrated Brownian motion (IBM) prior for their GP model for the purpose of testing prior sensitivity but did not use the prior beyond that. The IBM prior is equivalent to the second-order GMRF in continuous time. Our use of second-order GMRF model for jointly estimating genealogy and effective population size trajectory is the first we are aware of in the literature. The second order GMRF and HSMRF can have similar performance in many cases, but HSMRF has the advantage of added flexibility when needed, so it is a reasonable default choice over the GMRF. We suggest that researchers fit both orders and use a metric such as Bayes factors to select the best order model for the data.

Chapter 5

HORSESHOE MARKOV RANDOM FIELDS FOR SPATIAL PROCESSES

5.1 Introduction

Spatial processes play important roles in many disciplines including ecology, epidemiology, image analysis, meteorology, oceanography, geophysics, and astronomy. Spatial analysis focuses on extracting information about spatial processes from spatially referenced data. An important aspect of spatial analysis is the estimation of an underlying spatial field that generates a set of observed data. There are many cases of interest where the underlying field could contain abrupt or rapid changes in combination with areas with minimal variation. Some examples include distributions of individuals of a sessile species on a landscape (Yue and Loh, 2011), areas of brain activity on neural images (Brezger et al., 2007), spatial clusters of high disease incidence (Knorr-Held and Raßer, 2000), and areas of extreme precipitation events (Wallin and Bolin, 2015).

We focus here on Bayesian nonparametric methods for estimating spatial fields. In particular, we restrict our focus to methods developed for discrete space where the spatial domain of interest is partitioned into a set of areal units of regular or irregular shapes. Examples of such discrete units of irregular shape could be counties or states in a country, and regular units could be formed by imposing a regular grid over the spatial domain.

The two most common methods for estimating spatial fields over areal units are conditionally autoregressive models (CAR; Besag, 1974), and Gaussian Markov random fields (GMRF; Rue and Held, 2005). Both methods use the Markov property and conditional independence to define correlation structures based on neighborhoods of adjoining units. CAR models are built on pairwise differences among neighboring units and are directly comparable to first-order GMRFs (Rue and Held, 2005). Both methods assume that the pairwise differences follow Gaussian distributions.

Besag (1986) termed the group of CAR models that placed functions or distributions on the pairwise differences as pairwise interaction models. Some versions of pairwise interaction models were constructed to allow for abrupt jumps or discontinuities among adjacent pixels in image analysis were introduced by Geman and McClure (1985) and Geman and Reynolds (1992). Besag et al. (1991) used a Laplace distribution for the pairwise differences to account for discontinuities, and Green (1990) introduced a distribution for the differences that was between a Laplace distri-

bution and a normal distribution. As a way to account for outliers and jumps, Besag and Higdon (1999) introduced a model that placed t-distributions on the the pairwise differences by putting independent gamma distributions on the the precisions of the individual differences.

Extensions to the distributions of increments of GMRFs of first or second order have also been proposed by various authors to make locally-adaptive fields that can account for jumps and discontinuities. In the one-dimensional setting, Lang et al. (2002) suggested using independent local precision parameters for the increments. In particular, they suggested using gamma distributions to induce marginal t-distributions for the increments. Methods for using hierarchical scale mixture of normal distributions to induce non-normal distributions on increments of GMRFs was discussed by Rue and Held (2005), where Laplace, logistic, and t-distributions with ν degrees of freedom were described. Brezger et al. (2007) also used independent gamma distributions on local precision parameters to induce t-distributions on the increments and recommended using degrees of freedom equal to one to induce a Cauchy distribution. Yue and Speckman (2010) placed another GMRF on the log of local precision parameters to allow for correlation among the local precisions.

Another approach to allow for local adaptivity or sharp break points is the spatial partition model, which uses CAR models in combination with models for adjacency or group membership. Knorr-Held and Raßer (2000) used a model for group membership to cluster areal units within contiguous regions, resulting in piecewise-constant fields where the areal units in an assigned region shared the same value of the field. In both the CAR and GMRF models, there needs to be some kind of adjacency matrix or way to keep track of which units are adjacent. Typically this is done with a set of variables w_{ij}, where $w_{ij} = 1$ if units i and j are neighbors and $w_{ij} = 0$ otherwise. Lu et al. (2007) proposed treating the w_{ij} of adjacent units as binary parameters to be estimated. This flexibility allows units close in space to be uncorrelated and therefore allows for the possibility of abrupt changes in the field. This approach leads to a large number of additional parameters and is computationally expensive. Rushworth et al. (2017) proposed modeling adjacency on the unit interval rather than as a binary variable, which allows a continuum of strength of adjacency. This was done by modeling a probability surface as a GMRF on the logit scale.

Note that Gaussian processes (GP) are commonly used for spatial problems where the data

are indexed in continuous space. A GP can be approximated in discretized space by using some measure of distance between areal units, such as the distance between midpoints of the units, which allows the covariances among locations to be calculated with the covariance function. There are various methods for creating nonstationary GPs that would potentially allow for modeling discontinuities in the field or varying levels of smoothness. Some methods include specifying nonstationary covariance functions (Paciorek and Schervish, 2006), using nonstationary process convolutions (Higdon, 1998; Fuentes, 2002), or adaptive smoothing splines (Yue et al., 2014). We chose to focus on GMRF methods since these are specifically tailored for use in discrete space, and lower-order fields, *e.g.*, first-order spatial random walks, are naturally able to accommodate discontinuities in the field.

Here we propose an extension of the HSMRF models described in previous chapters to the spatial setting. Our approach places a horseshoe distribution on the pairwise differences of the field parameters at neighboring locations, which allows for jumps and discontinuities in the field while also imposing smoothing over regions with little variation. We present computationally efficient methods for inference using sparse matrix operations in combination with Hamiltonian Monte Carlo (HMC). We demonstrate our methods on simulated data and apply the models to cancer incidence data and to a spatial point process.

5.2 Methods

5.2.1 HSMRF Prior for Discrete Space

We extend our HSMRF model presented in Chapter 2 to two-dimensional space. Assume there is a process in continuous space that follows an unknown function $f(\boldsymbol{s})$, where $\boldsymbol{s} \in \mathbb{R}^2$. Let $\theta_i = \tilde{f}(\boldsymbol{a}_i)$ be the expected value of the surface over a discrete areal unit $\boldsymbol{a}_i$, where $\boldsymbol{a}_i = \int_{D_i} \boldsymbol{s} d\boldsymbol{s}$ and $D_i \subset D$, where D is the spatial domain over which the process is defined and $i = 1, \ldots, N$. The areal units can include irregularly shaped spatial units (e.g., county or state boundaries) or cells of a discrete (typically uniform) grid defined over the domain of interest. Let the first-order difference between

neighboring units be defined as

$$\Delta\theta_{ij} = \theta_i - \theta_j,$$

where $i = 1, \ldots, N$ and $j = 1, \ldots, N$ but $i \neq j$ and $\Delta\theta_{ij}$ is only defined for unique pairs of neighboring units. We assume the $\Delta\theta_{ij}$'s follow independent and identical horseshoe distributions with scale parameter γ. That is

$$\Delta\theta_{ij} \mid \gamma \sim \mathcal{HS}(\gamma).$$

We set the prior for γ to be $C^+(0, \zeta)$ as in previous chapters. These marginal distributions for the $\Delta\theta_{ij}$ are represented in practice using the following hierarchical form:

$$\begin{aligned} \Delta\theta_{ij} \mid \tau_{ij} &\sim \mathcal{N}(0, \tau_{ij}^2) \\ \tau_{ij} \mid \gamma &\sim C^+(0, \gamma). \end{aligned} \tag{5.1}$$

This hierarchical version formulates the otherwise intractable horseshoe distribution as a scale mixture of normal distributions (Carvalho et al., 2010).

It follows that $\boldsymbol{\theta}$ follows a Gaussian Markov random field (GMRF; Rue and Held, 2005) conditional on mean $\boldsymbol{\mu}$ and precision matrix $\mathbf{Q}(\boldsymbol{\tau})$. The conditional probability density function for $\boldsymbol{\theta}$ is therefore:

$$p(\boldsymbol{\theta} \mid \boldsymbol{\tau}, \boldsymbol{\mu}) = (2\pi)^{-n/2} |\mathbf{Q}(\boldsymbol{\tau})|^{1/2} \exp\left(-\frac{1}{2} (\boldsymbol{\theta} - \boldsymbol{\mu})^T \mathbf{Q}(\boldsymbol{\tau}) (\boldsymbol{\theta} - \boldsymbol{\mu}) \right). \tag{5.2}$$

Here $\boldsymbol{\tau}$ is a vector of scale parameters and the structure of the sparse $\mathbf{Q}(\boldsymbol{\tau})$ is determined by the differencing in $\boldsymbol{\theta}$ and the neighborhood structure of the set of areal units $\boldsymbol{a}$ (see below). We assume $\boldsymbol{\mu} = \mathbf{1}\mu$ where $\mathbf{1}$ is a vector of ones and μ is a scalar for the models we present here. These models could be generalized to allow $\boldsymbol{\mu} = \mathbf{X}\boldsymbol{\beta}$, where $\mathbf{X}$ is an $n \times m$ matrix of m covariates and $\boldsymbol{\beta}$ is a $m \times 1$ vector of regression parameters. This model extension is a topic of future research.

We only consider first-order fields, although higher-order fields could be constructed. The elements of the precision matrix $\mathbf{Q}(\boldsymbol{\tau})$ for a first-order HSMRF is given by:

$$Q_{ij} = \begin{cases} w_{i+} & j = i \\ -w_{ij} & j \neq i \\ 0 & \text{otherwise,} \end{cases} \tag{5.3}$$

where $w_{ij} = 1/(d_{ij}\tau_{ij}^2)$ and d_{ij} can be specified as the distance between the centroids of units $\boldsymbol{a}_i$ and $\boldsymbol{a}_j$. We set $d_{ij} = 1$ for all i and j for analyses presented here.

5.2.2 Posterior Inference

We avoid the necessity to specify a precision matrix and calculate its determinant by assuming the $\Delta\theta_{ij}$ are independent. This assumption allows for faster computation times and easier model implementation. However, the assumption of independence is only correct if we account for the set of hidden constraints on the $\Delta\theta_{ij}$ imposed by the complex geometry of the neighborhood structures (Rue and Held, 2005). These constraints are unknown or difficult to recover at best, but they are implicit within the structure of the precision matrix in equation 5.2.1. It turns out that the independence assumption is actually not of consequence since any correlations among the $\Delta\theta_{ij}$ will be recovered in the posterior. We therefore assume independence in the $\Delta\theta_{ij}$ in what follows.

Let the $(K + 1) \times N$ matrix $\mathbf{B}$ consist of differencing operators for the set of K unique pairwise differences between adjacent map units, such that

$$\mathbf{B}\boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \Delta\theta_1 \\ \vdots \\ \Delta\theta_K \end{bmatrix} = \mathbf{z}. \tag{5.4}$$

Note that the first row of $\mathbf{B}$ has a 1 in the first column and 0's in the remaining columns. The remaining rows have a 1 in column i, a -1 in column j, and 0's in the remaining columns, where i and j are in the set of K unique pairs of adjacent units. For computational efficiency, we want to sample $\mathbf{z}$ instead of $\boldsymbol{\theta}$, but we still need $\boldsymbol{\theta}$ for likelihood computations. To recover $\boldsymbol{\theta}$, we need to solve equation 5.4 for $\boldsymbol{\theta}$ using $\mathbf{B}$ and $\mathbf{z}$.

Note that $\mathbf{B}$ is rectangular but can be decomposed with a QR decomposition. The QR decomposition will provide an exact solution for $\boldsymbol{\theta}$ here because it acts as a deterministic transformation. Let $\mathbf{B} = \mathbf{Q}_B\mathbf{R}_B$, where $\mathbf{Q}_B$ is a $(K + 1) \times (K + 1)$ orthogonal matrix and $\mathbf{R}_B$ is a $(K + 1) \times N$

upper-triangular matrix. Then the solution for $\boldsymbol{\theta}$ is found by

$$\boldsymbol{\theta} = \mathbf{R}_B^{-1}\mathbf{Q}_B^T\mathbf{z}.$$

Note $\mathbf{R}_B^{-1}$ and $\mathbf{Q}_B^T$ are partially sparse and only need to be computed once and saved. In practice, we compute $\mathbf{R}_B^{-1}$ and $\mathbf{Q}_B^T$ using sparse matrix tools in the `R` package `Matrix`. We then convert the resulting matrices to compressed sparse row format for faster computation in `stan`.

Our posterior distribution is

$$p(\boldsymbol{\theta}, \boldsymbol{\tau}, \gamma, \boldsymbol{\xi} \mid \mathbf{y}) \propto \prod_{i=1}^{N} p(y_i \mid \theta_i, \boldsymbol{\xi}) p(z_1 \mid \mu, \omega^2) \prod_{k=2}^{K+1} p(z_k \mid \tau_{k-1}) p(\tau_{k-1} \mid \gamma) p(\boldsymbol{\xi}) p(\gamma), \tag{5.5}$$

where $p(y_i \mid \theta_i, \boldsymbol{\xi})$ is the likelihood and $\boldsymbol{\xi}$ are additional parameters associated with the likelihood, $p(z_1 \mid \mu, \omega^2)$ is the prior distribution on $z_1 = \theta_1$, where $z_1 \sim \mathcal{N}(\mu, \omega^2)$, $p(z_k \mid \tau_{k-1})$ are independent conditionally normal priors for the z_k, $p(\tau_{k-1} \mid \gamma)$ and $p(\gamma)$ are previously described priors, and $p(\boldsymbol{\xi})$ is the prior distribution for $\boldsymbol{\xi}$.

We used HMC (Neal, 2011) for posterior inference with the `Stan` computing environment (Carpenter et al., 2016). Specifically, we used the open source package `rstan` (Stan Development Team, 2017), which provides a platform for fitting models using HMC in the `R` computing environment (R Core Team, 2017).

5.3 Results

5.3.1 Simulated Data

We used simulated data to investigate properties of the spatial HSMRF. We compared performance to two other nonparametric methods based on Markov random field priors. One method is a first-order GMRF. This method is considered non-adaptive. For this method, we have $\Delta\theta_{ij} \mid \gamma \sim \mathcal{N}(0, \gamma^2)$, where $\gamma \sim C{+}(0, \zeta)$. The other method is based on a locally-adaptive approach introduced by Brezger et al. (2007) and adapted by Yue and Speckman (2010) and Yue and Loh (2011). This method uses a hierarchical scale mixture of normal distributions with results marginally in a Cauchy distribution on the first-order increments. The hierarchical formulation is $\Delta\theta_{ij} \mid \lambda_{ij}, \delta \sim$

$\mathcal{N}(0, \delta^2\lambda_{ij}^2)$, $\lambda_{ij}^2 \sim \mathcal{IG}(\nu/2, \nu/2)$, where $\mathcal{IG}()$ is an inverse gamma distribution and $\nu = 1$, and $\delta^2 \sim \mathcal{IG}(0.001, 0.001)$. After integration over the λ_{ij}, the marginal formulation is $\Delta\theta_{ij} \mid \delta \sim C(0, \delta)$. We will refer to this model as a Cauchy Markov random field (CMRF).

We investigated four different scenarios: 1) Halves, 2) Blocks, 3) Smooth, and 4) Hot spots (Figure 5.1). For each scenario, we generated a mean surface on a 30×30, uniformly spaced grid.

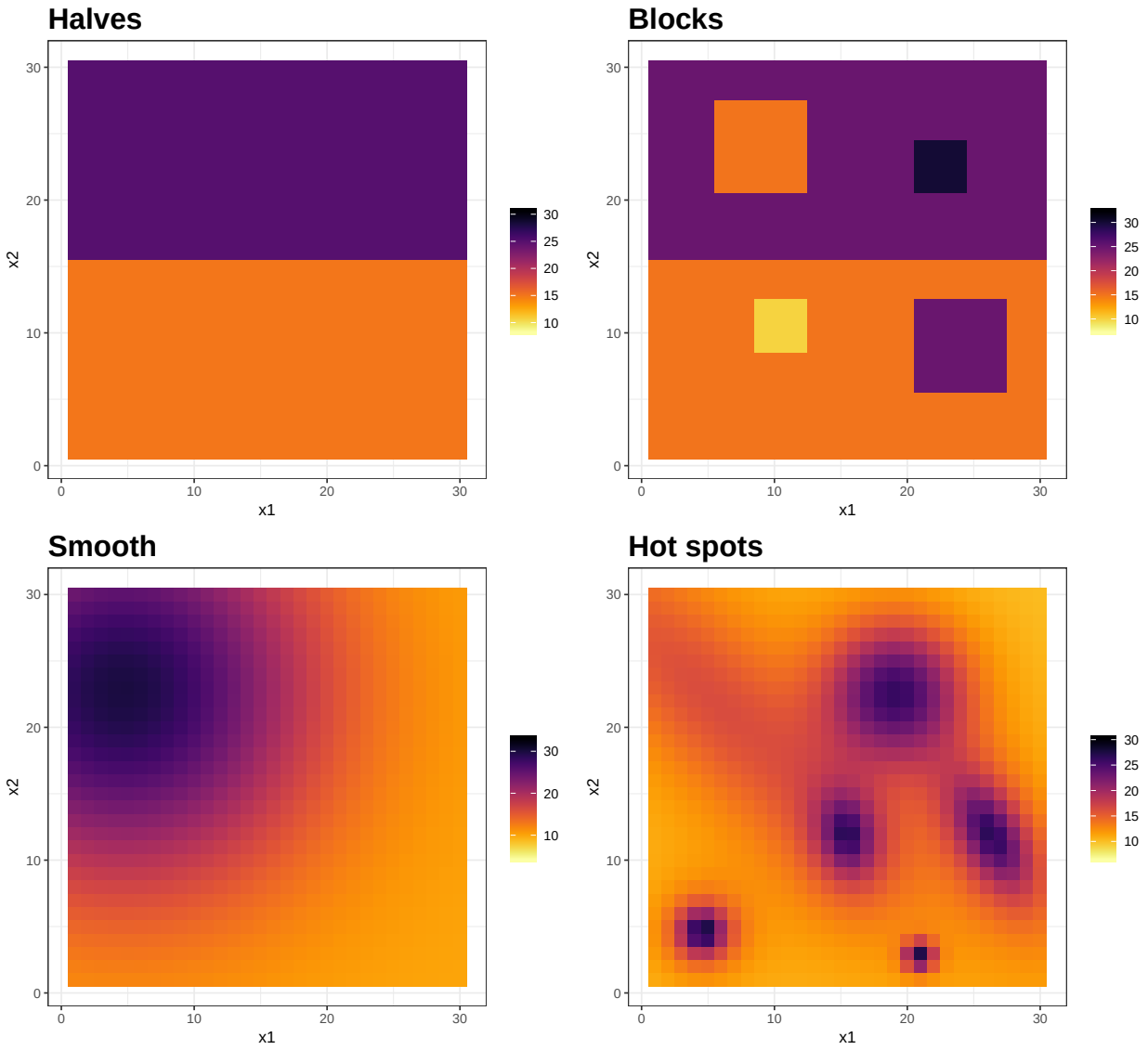

Figure 5.1: Scenarios for generating simulated data. Mean surfaces are shown.

We then generated data representing observational noise from a $\mathcal{N}(0, \sigma^2)$ distribution, where each of the 900 grid cells had a single observation. For each scenario we have currently only generated a single data realization (but will generate many more simulated data sets in the near future).

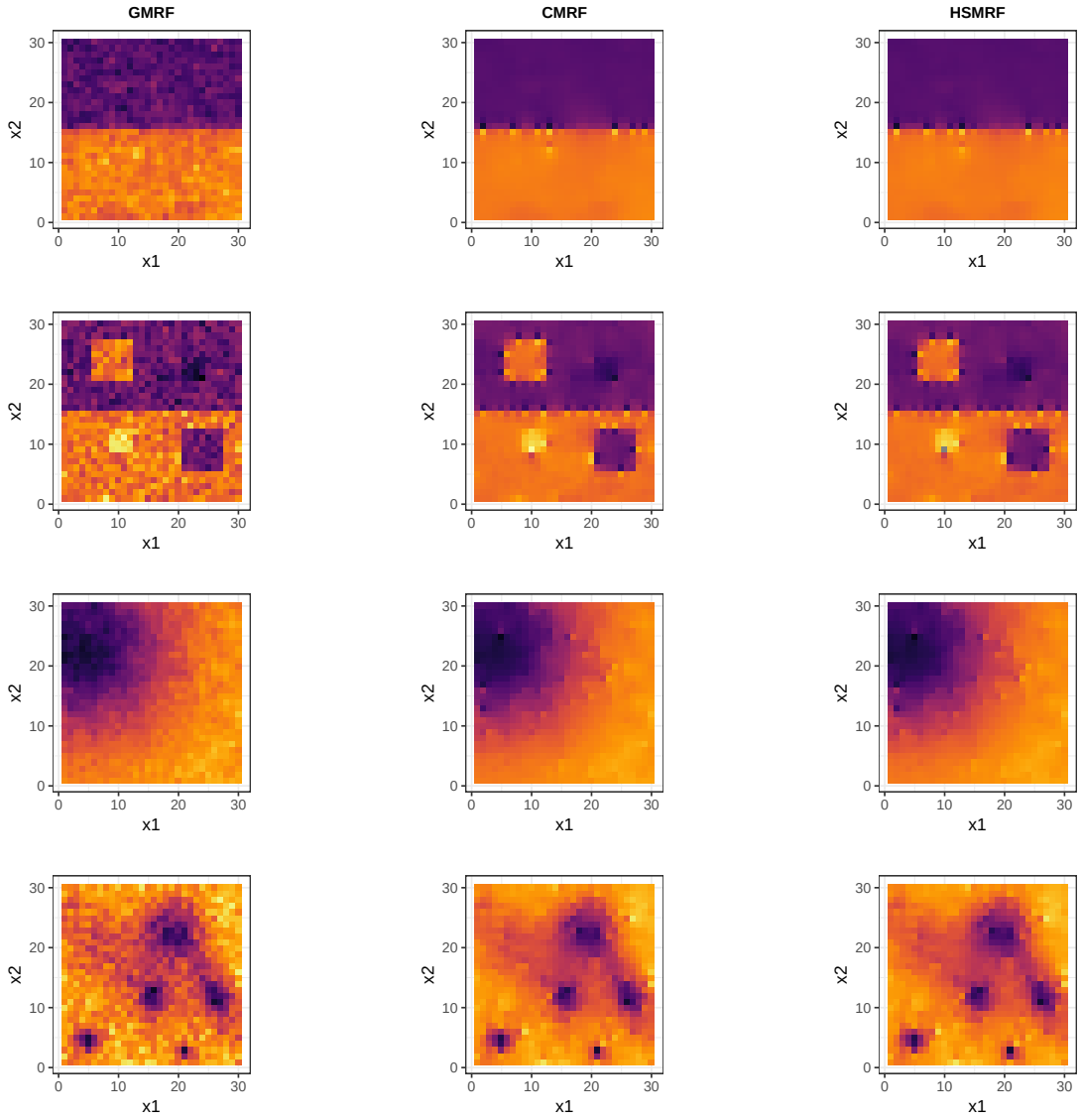

Figure 5.2: Posterior median values of the estimated surfaces fit to the simulated data for each of the scenarios and each of the three models.

For each data set we calculated some performance metrics for comparison among models. We

used the mean absolute deviation (MAD) as a measure of bias, where MAD $= \frac{1}{N}\sum_{i=1}^{N}|\hat{\theta}_i - \theta_i|$, and θ_i are the true field values and $\hat{\theta}_i$ are the posterior medians. We calculated the mean squared error as a measure of bias and precision, where MSE $= \frac{1}{N}\sum_{i=1}^{N}\left(\hat{\theta}_i - \theta_i\right)^2$. As a measure of bias in the individual increments, we used the mean absolute deviation in the differences (MADD), where MADD $= \frac{1}{K}\sum_{i=1}^{K}|\hat{\Delta}\theta_k - \Delta\theta_k|$, and $\Delta\theta_k$ are the true increment values and $\hat{\Delta}\theta_k$ are the posterior medians. Smaller values of MAD, MSE, and MADD indicate lower bias in estimated posterior medians. We used the mean width of 95% credible intervals as a measure of precision: MCIW $= \frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_{97.5,i} - \hat{\theta}_{2.5,i})$. Here $\hat{\theta}_{97.5,i}$ and $\hat{\theta}_{2.5,i}$ are the 97.5% and 2.5% quantiles of the posterior distribution for θ_i. Smaller values of MCIW indicate less uncertainty in the posterior estimates of θ. We assessed the coverage of credible intervals using Envelope $= \frac{1}{N}\sum_{i=1}^{N} I(\theta_i \in [\hat{\theta}_{97.5,i}, \hat{\theta}_{2.5,i}])$, where $I(\cdot)$ is the indicator function. Average Envelope values closer to the nominal coverage of 95% are considered better. Finally, we calculated the posterior median and 95% credible interval for σ^2, with the idea that an over-fitted surface will result in a negatively biased estimate of the noise variance. Estimated values of σ closer to 2.0 are better for these simulations. For the GMRF and HSMRF models we used $\sigma \sim C^+(0, 5)$ as a prior for σ and used $\sigma^2 \sim \mathcal{IG}(0.001, 0.001)$ for the CMRF.

The results indicate that the locally adaptive models (CMRF and HSMRF) performed better than the GMRF (Table 5.1 and Figure 5.2). The CMRF and HSMRF had nearly identical performance measures, although the HSMRF performed slightly better for many of the metrics and scenarios.

5.3.2 Disease Mapping

Here we investigate the risk of oral cavity cancer in Germany. The observation data are number of deaths from oral cavity cancer in males in 544 districts from 1986-1990. The data also includes expected number of deaths in each district based on age and population size. There were a total of 15,466 deaths, with median per district of 19 and range 1–501. Our interest is in estimating relative risk of oral cancer spatially for the time period represented by the data. Among others, these data were used by Knorr-Held and Raßer (2000) for detecting clusters and discontinuities in

Table 5.1: Mean values of performance measures and posterior median and 95% credible intervals for σ for simulated data for each scenario and model. Best values for each performance measure within each scenario are shown in bold.

Scenario	**Model**	**MAD**	**MSE**	**MCIW**	**Env**	**MADD**	σ
Halves	GMRF	0.999	1.586	4.449	0.921	1.176	1.53 (1.30, 1.72)
	CMRF	0.381	0.430	**2.541**	0.968	0.300	1.97 (1.87, 2.07)
	HSMRF	**0.378**	**0.425**	2.557	**0.968**	**0.299**	**1.97** (1.87, 2.09)
Blocks	GMRF	1.515	3.524	2.414	0.474	2.121	0.59 (0.40, 0.95)
	CMRF	0.766	1.193	3.983	**0.954**	0.860	1.87 (1.75, 2.00)
	HSMRF	**0.762**	**1.192**	**3.946**	**0.954**	**0.854**	**1.89** (1.76, 2.02)
Smooth	GMRF	0.747	0.901	4.206	**0.970**	0.841	1.64 (1.48, 1.79)
	CMRF	**0.567**	**0.641**	4.015	0.980	0.509	1.82 (1.70, 1.94)
	HSMRF	0.567	0.642	**4.008**	0.982	**0.505**	**1.82** (1.71, 1.94)
Hot spots	GMRF	1.303	2.625	3.518	0.727	1.766	0.97 (0.74, 1.28)
	CMRF	0.767	1.093	4.495	0.966	0.841	1.67 (1.52, 1.82)
	HSMRF	**0.759**	**1.085**	**4.491**	**0.964**	**0.826**	**1.69** (1.54, 1.83)

spatial patterns of disease and by Held et al. (2005) for joint mapping of multiple diseases. We obtained the data from the `INLA` package for `R` (Rue et al., 2017).

We base the general form of the models on an approach introduced by Besag et al. (1991). Let y_i be the observed number of cancer deaths and E_i be the expected number in district i over the time period of interest. We assume the counts are distributed Poisson and we use the following expression for the likelihood:

$$y_i \mid \boldsymbol{\theta}, \boldsymbol{\omega} \sim \mathcal{P}\left(\exp(\theta_i + \omega_i)E_i\right), \tag{5.6}$$

where the θ_is represent the spatial relative risks and ω_is are independent and distributed as $\omega_i \mid \rho \sim \mathcal{N}(0, \rho^2)$, with $\rho \sim C^+(0, 5)$.

We fit three models, each with different priors for $\boldsymbol{\theta}$. The priors were the GMRF, CMRF, and HSMRF described in section 5.3.1. Note that the field for $\boldsymbol{\theta}$ is estimated on the natural logarithm scale.

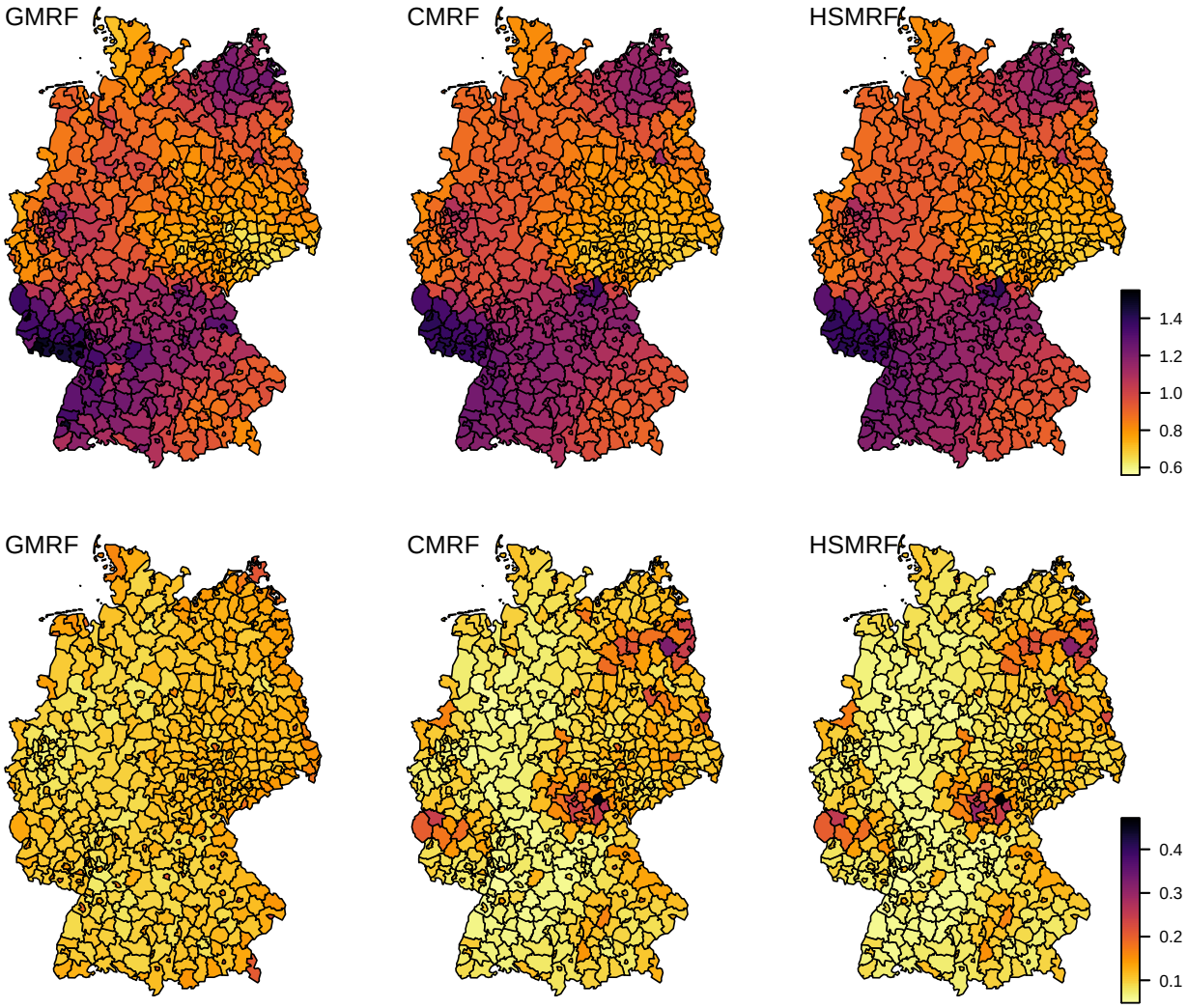

Figure 5.3: Posterior median relative risk (top row) and standard deviation of logarithm of relative risk (bottom row) for each of the three models.

The posterior median relative risks showed the biggest differences between the non-adaptive

and adaptive methods (Figures 5.3 and 5.4). The adaptive models displayed higher levels of smoothing of the relative risks yet also allowed for some abrupt discontinuities. There was little difference between the CMRF and HSMRF in terms of posterior median relative risk. The biggest differences between the HSMRF and the other two models occurred in two adjacent districts near the middle of the country. Interestingly, the boundary between these districts is part of the old boundary between East and West Germany (see Figure 5.4 and Figure C.1 in Appendix C). The standard deviations in the posterior estimates for the fields covered a wider range for the CMRF and HSMRF compared to the CMRF.

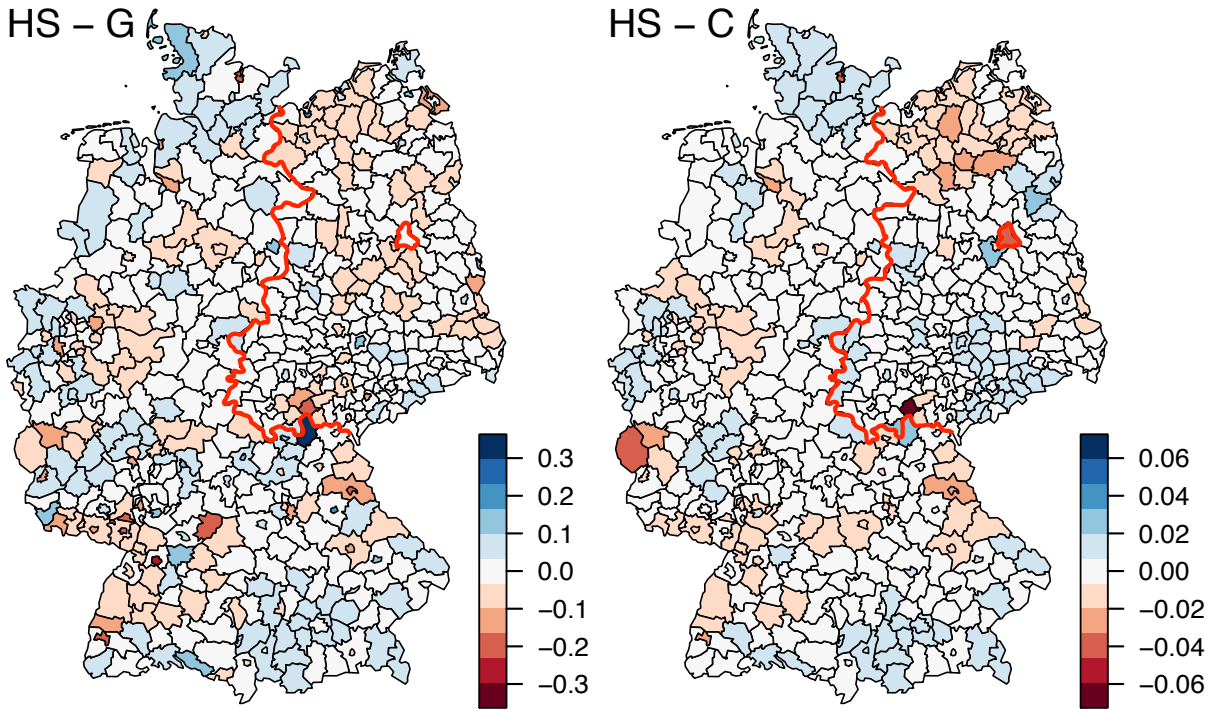

Figure 5.4: Difference in posterior median relative risk for HSMRF minus GMRF (HS - G) and HSMRF minus CMRF (HS - C). Note the differences in scale between the two plots. Old boundary between East and West Germany is shown in red.

5.3.3 *Inhomogeneous Point Process*

A Cox process is a type of point process used to model aggregated spatial point patterns. A Cox process is composed of an inhomogeneous Poisson point process driven by a random intensity measure. A log-Gaussian Cox process models the logarithm of the intensity measure with a Gaussian process (Møller et al., 1998).

We use an example of an inhomogeneous point process to estimate the underlying point process intensity using a log-Gaussian Cox process. We use data for the locations of 3,413 individual trees of the species *Beilschmiedia pendula* from a census of a $1,000 \times 500$ m (50 ha) plot on Barro Colorado Island in Panama in 1995 (see Hubbell, 1983; Condit et al., 1996; Condit, 1998). This is a subset of a larger data set used by others for point process methods (*e.g.*, Rue et al., 2009; Waagepetersen, 2007). We divided the study area into a regular grid of $25 \times 50 = 1,250$ cells, each of 400m^2 and counted the number of trees in each grid cell. The individual tree locations and grid counts are shown in Figure (5.5).

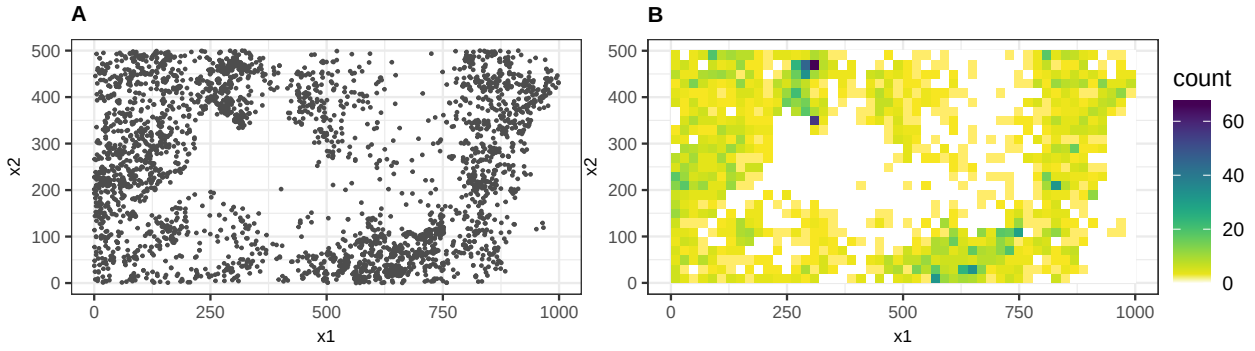

Figure 5.5: Locations of *B. pendula* trees (A) and binned counts of trees (B) on a 25×50 grid.

The likelihood is similar to that used in the disease mapping example:

$$y_i \mid \boldsymbol{\theta}, \boldsymbol{\omega} \sim \mathcal{P}\left(\exp(\theta_i + \omega_i)|A_i|\right), \tag{5.7}$$

where $|A_i|$ is the area (m^2) of grid cell i, θ_i represent the spatial component of the log intensity and

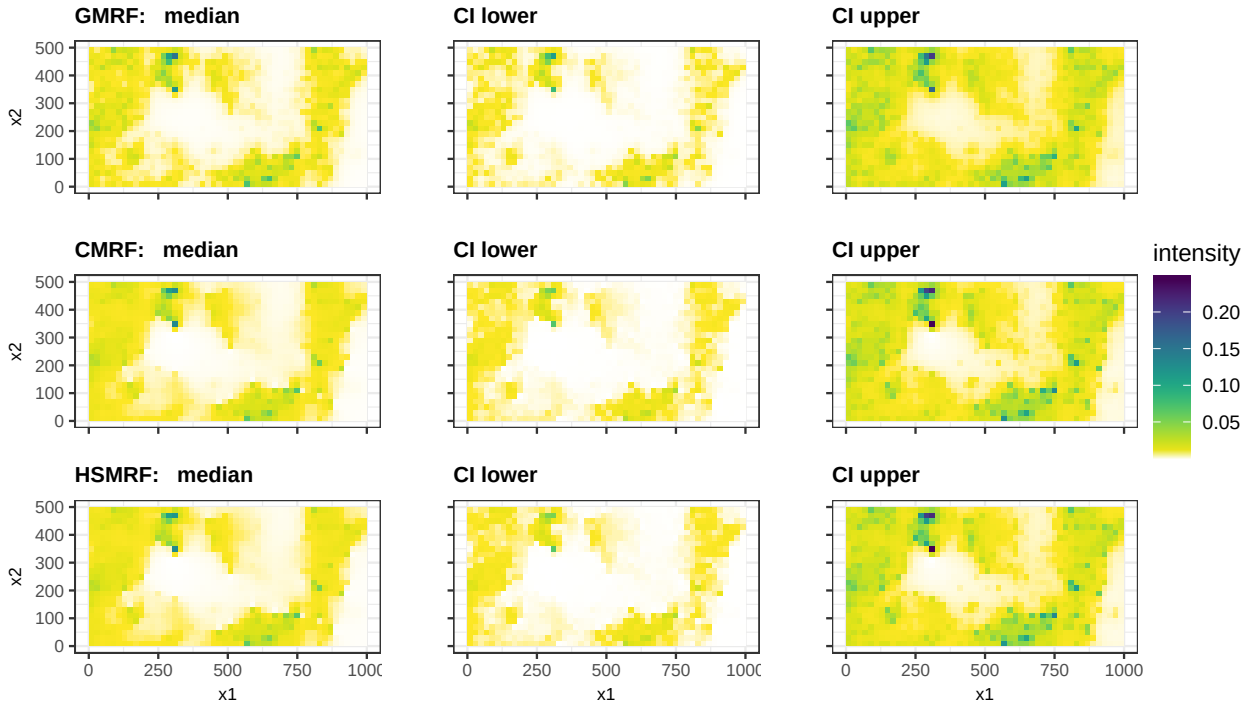

Figure 5.6: Posterior medians and 95% credible intervals for the estimated point process intensities for *B. pendula* locations by model.

ω_i are independent and distributed as $\omega_i \mid \rho \sim \mathcal{N}(0, \rho^2)$, with $\rho \sim C^+(0, 5)$. Here $|A_i| = 400\text{m}^2$ for all i.

Results indicate that the CMRF and the HSMRF were able to capture abrupt changes in the point process intensity yet still were able to smooth the surface better than the GMRF (Figure 5.6). Note the surfaces are much easier to compare if you zoom in on the plots. The differences between the estimated intensities for the HSMRF and CMRF were much more subtle, but the HSMRF did a little better at smoothing the intensity.

5.4 Discussion

We successfully extended the HSMRF to spatial processes defined on a discrete grid or set of areal units. As with the one-dimensional case, we demonstrated that placing independent horseshoe pri-

ors on the individual increments of the spatial field resulted in a combined ability to capture abrupt changes as well as smooth over unvarying portions of the field. As outlined in the introduction to this chapter, the idea of allowing individual increments to have non-normal distributions is not new. However, we are the first to use shrinkage priors for the increments and we have demonstrated this can result in better performance than other distributions.

That being said, the results for the CMRF and HSMRF were very similar in the examples we investigated. This is probably because the scale of the Cauchy distribution can be made small, resulting in high probability mass near zero, yet the distribution can still maintain long tails. As we have seen for the horseshoe distribution, this combination results in the combination of smoothing and ability to capture abrupt changes. The differences between the two distributions are in the higher mass at zero for the horseshoe and less mass in the middle values for the horseshoe due to the leptokurtic shape of the distribution. These differences should result in an advantage in smoothing due to the added shrinkage of the horseshoe.

Two things we are lacking in our HSMRF models are the ability to include covariates and the ability to fit continuous fields without discretization with a grid. These are both topics of future research and are discussed in the following chapter.

Chapter 6

FUTURE RESEARCH

6.1 Future Research

6.1.1 HSMRF Model for Continuous Space

The next step in our work on spatial HSMRFs is to extend the method to continuous space. This would allow for application to spatial data that are sampled at random locations. One example of such an application is the spatial point process of tree locations described in Chapter 5. In that example we had to create a discrete grid and bin the counts of trees in each grid cell. A model for continuous space would allow us to predict the point process intensity at any location in the spatial domain. Our approach is inspired by the SPDE approach for non-Gaussian fields developed by Bolin (2014) and Wallin and Bolin (2015), which is an extension of the work of Lindgren et al. (2011) on SPDEs for Gaussian Matérn fields. Most of the information presented in this subsection was drawn from their work.

The SPDE for a non-Gaussian Matérn field $X(\mathbf{s})$ is

$$\left(\kappa^2 - \Delta\right)^{\frac{\alpha}{2}} X(\mathbf{s}) = \dot{M}(\mathbf{s}),$$

where κ and α are parameters of the Matérn covariance function, $\Delta = \sum_{i=1}^{d} \frac{\partial^2}{\partial s_i^2}$ is the Laplacian operator, d is the dimension of the domain, and $\dot{M}(\mathbf{s})$ is non-Gaussian noise process. A Gaussian noise process is produced if $\dot{M}(\mathbf{s}) = \phi\mathcal{W}(\mathbf{s})$, where ϕ is a scalar scale parameter and $\mathcal{W}(\mathbf{s})$ a white noise process (differentiated Brownian sheet). Wallin and Bolin (2015) restricted their focus to the case where $M(\mathbf{s})$ is a type G Lévy process. Note that for the models we discuss, $\dot{M}(\mathbf{s}) = \phi(\mathbf{s})\mathcal{W}(\mathbf{s})$, where $\phi(\mathbf{s})$ is a random scale process.

A type G Lévy process has increments that can be written as a scale mixture of Gaussian variables $V^{1/2}Z$. Here V is a non-negative and infinitely divisible random variable, and Z is a standard Gaussian variable. For infinitely divisible V, there exists a non-decreasing Lévy process $V(\mathbf{s})$ with increments distributed the same as V. The generalized hyperbolic distributions are a subtype of type G Lévy process. Wallin and Bolin (2015) restricted the class of distributions they investigated to those closed under convolution so that V_i has known distribution for any increment area. This restricted them to the normal inverse Gaussian (NIG) and generalized asymmetric Laplace (GAL)

distributions.

The infinite divisibility property of V is important because the SPDE solutions use a set of basis functions that are defined over discrete space, and each basis can cover different amounts of the domain. Direct extension of this method to a continuous HSMRF is not possible because the half-Cauchy scale process is not infinitely divisible (self similar). This can be seen by the fact that the sum of half-Cauchy random variables does not follow a half-Cauchy distribution. One heuristic solution to this problem would be to treat the half-Cauchy distribution as if it is infinitely divisible. This is similar to our approach in Chapter 3, Section 3.2.4 for extending the one-dimensional HSMRF to an irregular grid. The resulting scale process would preserve the shrinkage properties of the horseshoe distribution.

6.1.2 HSMRF Models with Covariates

A clear limitation of the models presented in this dissertation is that they did not allow the underlying spatial or temporal fields to be modeled with covariates. It is standard practice in semi-parametric models or spatial generalized linear models to allow fixed covariates (*e.g.*, Rue et al., 2009; Fong et al., 2010; Yue and Speckman, 2010). These approaches typically use intrinsic GMRFs or intrinsic CAR models so that the fields are invariant to the addition of a fixed trend. We would need to make adjustments to our model formulations to add covariate effects.

One potential concern with adding covariates to models that have random effects correlated in space or time is the potential confounding that can occur between the random effects and any covariate effects that are also correlated with space or time (Reich et al., 2006; Hughes and Haran, 2013). We would need to investigate the effects of such spatial confounding on the estimation of the parameters related to the covariate effects.

6.1.3 Effects of HSMRF Posterior Geometry on Marginal Likelihood Estimation

In Chapter 4, we used marginal likelihood estimates to calculate Bayes factors and posterior model probabilities. We used a stepping stone approach (Xie et al., 2011) for the HCV and bison analy-

ses, both of which were based on sequence data. This method appeared to work well despite some variation in estimates due to Monte Carlo error. The stepping stone method is quite computationally intensive and I was not able to get it to work in `stan` for analyses on coalescent times from fixed genealogies.

As part of the simulation analysis based on fixed genealogies, we investigated the use of bridge sampling (Gelman and Meng, 1998) for estimating marginal likelihoods. This method can be implemented with `stan` but it does require 10-20 times more posterior samples than would be needed for standard estimation of posterior distributions. However, we discovered that the bridge sampling estimates for the HSMRF models were unstable and could result in potentially biased estimates with large variance. We believe this was due to the geometry of the posterior distributions for the HSMRF models. In particular, there are non-linear, funnel-like correlations between the local scale parameters (τ_j) and their associated $\Delta\theta_j$'s. This becomes most apparent when the local scale parameters are displayed on the log scale. The result is that the proposal distributions used in the bridge sampling algorithm do not match the geometry of the posteriors from the HSMRF models. This is because the bridge sampling algorithms used in the `bridgesampling` package use either multivariate normal distributions or "warped" (skewed) multivariate distributions. These distributions will place posterior mass in large areas where the HSMRF posteriors have low mass. The result will be poor estimates of marginal likelihoods.

An interesting topic of future research would be to investigate the effect of custom proposal distributions for bridge sampling that account for the complex geometry of the posteriors from the HSMRF models. These methods could be applied to the temporal and spatial models with standard likelihoods described in Chapters 3 and 5. We could also add functionality to our `spmrf` package that would use the modified bridge sampling algorithms. We are unsure if the stepping stone method could suffer from similar problems to the bridge sampling algorithm. We could also investigate effects of difficult posterior geometries generated by the HSMRF models on stepping stone or path sampling algorithms.

6.1.4 Custom HMC Code with Sparse Matrix Operations

We predominantly used the `stan` programming language to implement our models using Hamiltonian Monte Carlo, with the exception of the models for sequence data presented in Chapter 4 that utilized our custom elliptical slice and Gibbs sampler. We ran into some limitations with `stan` when implementing our spatial models because `stan` did not have the structure in place to take advantage of sparse matrix operations. The GMRF and HSMRF models take advantage of conditional independence properties which allow for sparsity in the precision matrix. This sparsity can provide substantial benefits in terms of computation times if one can take advantage of it. Unfortunately, in `stan` for some models we needed to use standard methods for matrix inversion that did not use sparse operations.

In the absence of an upgrade to `stan` to allow for sparse matrix operations in the near future, an avenue of future research for me would be to code a custom HMC sampler in the `C++` language that takes advantage of available libraries for sparse matrix operations. This would allow for faster computation times and for more complex structure in the precision matrix. The main challenge of this project would be to derive a set of formulas for gradient calculations for a fixed set of model forms. HMC requires the gradient to generate parameter proposals. A major strength of the `stan` program is that it uses a set of fast and accurate automatic differentiation methods to calculate gradients.

BIBLIOGRAPHY

F. Abramovich, T. Sapatinas, and B. W. Silverman. Wavelet thresholding via a Bayesian approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 60(4):725–749, 1998.

R. P. Adams, I. Murray, and D. J. MacKay. Tractable nonparametric Bayesian inference in Poisson processes with Gaussian process intensities. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 9–16. ACM, 2009.

B. J. Alder and T. E. Wainwright. Studies in molecular dynamics. i. general method. *The Journal of Chemical Physics*, 31(2):459–466, 1959.

S. Alizon, S. Lion, C. L. Murall, and J. L. Abbate. Quantifying the epidemic spread of Ebola virus (EBOV) in Sierra Leone using phylodynamics. *Virulence*, 5(8):825–827, 2014.

D. F. Andrews and C. L. Mallows. Scale mixtures of normal distributions. *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(1):99–102, 1974.

A. Armagan, D. B. Dunson, and J. Lee. Generalized double Pareto shrinkage. *Statistica Sinica*, 23 (1):119–143, 2013.

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

J. Besag. On the statistical analysis of dirty pictures. *Journal of the Royal Statistical Society: Series B (Methodological)*, 48(3):259–279, 1986.

J. Besag and D. Higdon. Bayesian analysis of agricultural field experiments. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(4):691–746, 1999.

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

M. Betancourt and M. Girolami. Hamiltonian Monte Carlo for hierarchical models. In S. K. Upadhyay, U. Singh, D. K. Dey, and A. Loganathan, editors, *Current Trends in Bayesian Methodology with Applications*, pages 79–97. CRC Press, 2015.

A. Bhadra, J. Datta, N. G. Polson, and B. Willard. The horseshoe+ estimator of ultra-sparse signals. *Bayesian Analysis*, 12(4):1105–1131, 2017.

A. Bhattacharya, D. Pati, N. S. Pillai, and D. B. Dunson. Dirichlet-Laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, 110(512):1479–1490, 2015.

S. Bochner. *Harmonic Analysis and the Theory of Probability*. University of California Press, 1955.

D. Bolin. Spatial Matérn fields driven by non-Gaussian noise. *Scandinavian Journal of Statistics*, 41(3):557–579, 2014.

L. Bourgeon, A. Burke, and T. Higham. Earliest human presence in North America dated to the last glacial maximum: new radiocarbon dates from Bluefish Caves, Canada. *PloS ONE*, 12(1): e0169486, 2017.

S. Brahim-Belhouari and A. Bermak. Gaussian process for nonstationary time series prediction. *Computational Statistics & Data Analysis*, 47(4):705–712, 2004.

P. Brémaud. *Markov hains: Gibbs fields, Monte Carlo simulation, and queues*, volume 31 of *Texts in Applied Mathematics*. Springer Science & Business Media, 1999.

A. Brezger, L. Fahrmeir, and A. Hennerfeind. Adaptive Gaussian Markov random fields with applications in human brain mapping. *Journal of the Royal Statistical Society, Series C*, 56(3): 327–345, 2007.

P. J. Brockwell and R. A. Davis. *Inroduction to time series and forecasting*. Springer-Verlag, 2nd edition, 2002.

K. M. Broms, M. B. Hooten, D. S. Johnson, R. Altwegg, and L. L. Conquest. Dynamic occupancy models for explicit colonization processes. *Ecology*, 97(1):194–204, 2016.

B. E. Brown. Coral bleaching: causes and consequences. *Coral reefs*, 16(1):S129–S138, 1997.

B. P. Carlin, A. E. Gelfand, and A. F. Smith. Hierarchical Bayesian analysis of changepoint problems. *Applied Statistics*, 41(2):389–405, 1992.

B. Carpenter, A. Gelman, M. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. A. Brubaker, J. Guo, P. Li, and A. Riddell. Stan: A probabilistic programming language. *Journal of Statistical Software*, 20:1–37, 2016.

C. M. Carvalho, N. G. Polson, and J. G. Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.

A. Chakraborty, A. E. Gelfand, A. M. Wilson, A. M. Latimer, J. A. Silander Jr, et al. Modeling large scale species abundance with latent spatial processes. *The Annals of Applied Statistics*, 4 (3):1403–1429, 2010.

P. K. Clark. A subordinated stochastic process model with finite variance for speculative prices. *Econometrica*, 41(1):135–155, 1973.

P. U. Clark, A. S. Dyke, J. D. Shakun, A. E. Carlson, J. Clark, B. Wohlfarth, J. X. Mitrovica, S. W. Hostetler, and A. M. McCabe. The last glacial maximum. *Science*, 325(5941):710–714, 2009.

R. Condit. *Tropical forest census plots: methods and results from Barro Colorado Island, Panama and a comparison with other plots*. Springer Science & Business Media, 1998.

R. Condit, S. P. Hubbell, and R. B. Foster. Changes in tree species abundance in a neotropical forest: impact of climate change. *Journal of Tropical Ecology*, 12(2):231–256, 1996.

A. Datta, S. Banerjee, A. O. Finley, and A. E. Gelfand. On nearest-neighbor Gaussian process models for massive spatial data. *Wiley Interdisciplinary Reviews: Computational Statistics*, 8 (5):162–171, 2016.

I. DiMatteo, C. R. Genovese, and R. E. Kass. Bayesian curve-fitting with free-knot splines. *Biometrika*, 88(4):1055–1071, 2001.

V. Dinh, A. Bilge, C. Zhang, and F. A. Matsen, IV. Probabilistic path Hamiltonian Monte Carlo. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 1009–1018, 2017.

A. J. Drummond, A. Rambaut, B. Shapiro, and O. G. Pybus. Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular Biology and Evolution*, 22(5): 1185–1192, 2005.

A. J. Drummond, M. A. Suchard, D. Xie, and A. Rambaut. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular Biology and Evolution*, 29(8):1969–1973, 2012.

S. Duane, A. D. Kennedy, B. J. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Physics Letters B*, 195(2):216–222, 1987.

N. A. Easton, G. R. Mackay, P. B. Young, P. Schnurr, and D. R. Yesner. Chindadn in Canada? emergent evidence of the Pleistocene transition in southeast Beringia as revealed by the Little John Site, Yukon. In *From the Yenisei to the Yukon: Interpreting Lithic Assemblage Variability in Late Pleistocene/Early Holocene Beringia*, pages 289–307. Texas A&M University Press, 2011.

L. Fahrmeir and S. Lang. Bayesian inference for generalized additive mixed models based on markov random field priors. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 50(2):201–220, 2001.

J. R. Faulkner and V. N. Minin. Locally adaptive smoothing with Markov random fields and shrinkage priors. *Bayesian Analysis*, 13:225–252, 2018.

J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17(6):368–376, 1981.

J. Felsenstein and A. G. Rodrigo. Coalescent approaches to HIV population genetics. In K. A. Crandall, editor, *The Evolution of HIV*, pages 233–272. Johns Hopkins University Press, 1999.

M. A. Figueiredo. Adaptive sparseness for supervised learning. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(9):1150–1159, 2003.

Y. Fong, H. Rue, and J. Wakefield. Bayesian inference for generalized linear mixed models. *Biostatistics*, 11(3):397–412, 2010.

C. Frank, M. K. Mohamed, G. T. Strickland, D. Lavanchy, R. R. Arthur, L. S. Magder, T. El Khoby, Y. Abdel-Wahab, W. Anwar, and I. Sallam. The role of parenteral antischistosomal therapy in the spread of hepatitis C virus in Egypt. *The Lancet*, 355(9207):887–891, 2000.

I. E. Frank and J. H. Friedman. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.

N. Friel and A. N. Pettitt. Marginal likelihood estimation via power posteriors. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(3):589–607, 2008.

M. Fuentes. Spectral methods for nonstationary spatial processes. *Biometrika*, 89(1):197–210, 2002.

R. Furrer, M. G. Genton, and D. Nychka. Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, 15(3):502–523, 2006.

A. E. Gelfand, A. M. Schmidt, S. Wu, J. A. Silander Jr, A. Latimer, and A. G. Rebelo. Modelling species diversity through species level hierarchical modelling. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(1):1–20, 2005.

A. Gelman and X.-L. Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical Science*, 13:163–185, 1998.

A. Gelman and D. B. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, 7(4):457–472, 1992.

A. Gelman et al. Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, 1(3):515–534, 2006.

D. Geman and G. Reynolds. Constrained restoration and the recovery of discontinuities. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 14(3):367–383, 1992.

S. Geman and D. E. McClure. Bayesian image analysis: an application to single photon emission tomography. *Proceedings of the American Statistical Association, Statistical Computing Section*, pages 12–18, 1985.

M. S. Gill, P. Lemey, N. R. Faria, A. Rambaut, B. Shapiro, and M. A. Suchard. Improving Bayesian population dynamics inference: a coalescent-based model for multiple loci. *Molecular Biology and Evolution*, 30(3):713–724, 2013.

M. S. Gill, P. Lemey, S. N. Bennett, R. Biek, and M. A. Suchard. Understanding past population dynamics: Bayesian coalescent-based modeling with covariates. *Systematic Biology*, 65(6): 1041–1056, 2016.

M. Girolami and B. Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

P. J. Green. Bayesian reconstructions from emission tomography data using a modified EM algorithm. *IEEE Transactions on Medical Imaging*, 9(1):84–93, 1990.

P. J. Green. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82(4):711–732, 1995.

J. E. Griffin, P. J. Brown, et al. Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis*, 5(1):171–188, 2010.

J. E. Griffin, P. J. Brown, et al. Some priors for sparse regression modelling. *Bayesian Analysis*, 8 (3):691–702, 2013.

R. C. Griffiths and S. Tavaré. Sampling theory for neutral alleles in a varying environment. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 344(1310):403–410, 1994.

M. Hasegawa, H. Kishino, and T. Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution*, 22(2):160–174, 1985.

P. D. Heintzman, D. Froese, J. W. Ives, A. E. Soares, G. D. Zazula, B. Letts, T. D. Andrews, J. C. Driver, E. Hall, P. G. Hare, et al. Bison phylogeography constrains dispersal and viability of the Ice Free Corridor in western Canada. *Proceedings of the National Academy of Sciences*, 113 (29):8057–8063, 2016.

L. Held, I. Natário, S. E. Fenton, H. Rue, and N. Becker. Towards joint disease mapping. *Statistical Methods in Medical Research*, 14(1):61–82, 2005.

D. Higdon. A process-convolution approach to modelling temperatures in the North Atlantic Ocean. *Environmental and Ecological Statistics*, 5(2):173–190, 1998.

A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

M. D. Hoffman and A. Gelman. The no-u-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *The Journal of Machine Learning Research*, 15(1):1593–1623, 2014.

S. Höhna, M. J. Landis, T. A. Heath, B. Boussau, N. Lartillot, B. R. Moore, J. P. Huelsenbeck, and F. Ronquist. RevBayes: Bayesian phylogenetic inference using graphical models and an interactive model-specification language. *Systematic Biology*, 65(4):726–736, 2016.

C. E. Holmes. The Beringian and Transitional periods in Alaska. In *From the Yenisei to the Yukon:*

Interpreting Lithic Assemblage Variability in Late Pleistocene/Early Holocene Beringia, pages 179–191. Texas A&M University Press, 2011.

E. E. Holmes, E. J. Ward, and K. Wills. MARSS: Multivariate autoregressive state-space models for analyzing time-series data. *R journal*, 4(1), 2012.

S. P. Hubbell. *Diversity of canopy trees in a neotropical forest and implications for conservation*, pages 25–41. Oxford: Blackwell Scientific Publications, 1983.

J. Hughes and M. Haran. Dimension reduction and alleviation of confounding for spatial generalized linear mixed models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(1):139–159, 2013.

O. Husby and H. Rue. Estimating blood vessel areas in ultrasound images using a deformable template model. *Statistical Modelling*, 4(3):211–226, 2004.

R. G. Jarrett. A note on the intervals between coal-mining disasters. *Biometrika*, 66(1):191–193, 1979.

I. M. Johnstone and B. W. Silverman. Empirical Bayes selection of wavelet thresholds. *Annals of Statistics*, pages 1700–1752, 2005.

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

M. Katzfuss. A multi-resolution approximation for massive spatial datasets. *Journal of the American Statistical Association*, 112(517):201–214, 2017.

S.-J. Kim, K. Koh, S. Boyd, and D. Gorinevsky. ℓ_1 trend filtering. *Siam Review*, 51(2):339–360, 2009.

J. F. C. Kingman. The coalescent. *Stochastic Processes and Their Applications*, 13(3):235–248, 1982.

G. Kitagawa. Non-Gaussian state-space modeling of nonstationary time series. *Journal of the American Statistical Association*, 82(400):1032–1041, 1987.

L. Knorr-Held and G. Raßer. Bayesian detection of clusters and discontinuities in disease maps. *Biometrics*, 56(1):13–21, 2000.

L. Knorr-Held and H. Rue. On block updating in Markov random field models for disease mapping. *Scandinavian Journal of Statistics*, 29(4):597–614, 2002.

M. Kyung, J. Gill, M. Ghosh, and G. Casella. Penalized regression, standard errors, and Bayesian lassos. *Bayesian Analysis*, 5(2):369–411, 2010.

S. Lamba and N. Nain. A texture based mani-fold approach for crowd density estimation using Gaussian Markov random field. *Multimedia Tools and Applications*, 78(5):5645–5664, 2019.

S. Lan, J. A. Palacios, M. Karcher, V. N. Minin, and B. Shahbaba. An efficient Bayesian inference framework for coalescent-based nonparametric phylodynamics. *Bioinformatics*, 31(20):3282–3289, 2015.

S. Lang and A. Brezger. Bayesian P-splines. *Journal of Computational and Graphical Statistics*, 13(1):183–212, 2004.

S. Lang, E.-M. Fronk, and L. Fahrmeir. Function estimation with locally adaptive dynamic models. *Computational Statistics*, 17(4):479–499, 2002.

N. Lartillot and H. Philippe. Computing Bayes factors using thermodynamic integration. *Systematic Biology*, 55(2):195–207, 2006.

F. Lindgren and H. Rue. On the second-order random walk model for irregular locations. *Scandinavian Journal of Statistics*, 35(4):691–700, 2008.

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

B. Llamas, L. Fehren-Schmitz, G. Valverde, J. Soubrier, S. Mallick, N. Rohland, S. Nordenfelt, C. Valdiosera, S. M. Richards, A. Rohrlach, et al. Ancient mitochondrial DNA provides high-resolution time scale of the peopling of the Americas. *Science Advances*, 2(4):e1501385, 2016.

E. D. Lorenzen, D. Nogués-Bravo, L. Orlando, J. Weinstock, J. Binladen, K. A. Marske, A. Ugan, M. K. Borregaard, M. T. P. Gilbert, R. Nielsen, et al. Species-specific responses of Late Quaternary megafauna to climate and humans. *Nature*, 479(7373):359–364, 2011.

H. Lu, C. S. Reilly, S. Banerjee, and B. P. Carlin. Bayesian areal wombling via adjacency modeling. *Environmental and Ecological Statistics*, 14(4):433–452, 2007.

B. A. Maguire, E. S. Pearson, and A. H. A. Wynn. The time intervals between industrial accidents. *Biometrika*, 39(1/2):168–180, 1952.

E. Makalic and D. F. Schmidt. A simple sampler for the horseshoe estimator. *IEEE Signal Processing Letters*, 23(1):179–182, 2016.

E. Mammen, S. van de Geer, et al. Locally adaptive regression splines. *The Annals of Statistics*, 25(1):387–413, 1997.

A. Medhat, M. Shehata, L. S. Magder, N. Mikhail, L. Abdel-Baki, M. Nafeh, M. Abdel-Hamid, G. T. Strickland, and A. D. Fix. Hepatitis C in a community in upper Egypt: risk factors for infection. *The American Journal of Tropical Medicine and Hygiene*, 66(5):633–638, 2002.

F. D. Miller and L. J. Abu-Raddad. Evidence of intense ongoing endemic transmission of Hepatitis C virus in Egypt. *Proceedings of the National Academy of Sciences*, 107(33):14757–14762, 2010.

V. N. Minin, E. W. Bloomquist, and M. A. Suchard. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Molecular Biology and Evolution*, 25(7):1459–1471, 2008.

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

J. Møller, A. R. Syversveen, and R. P. Waagepetersen. Log Gaussian Cox processes. *Scandinavian Journal of Statistics*, 25(3):451–482, 1998.

C. C. Monnahan, J. T. Thorson, and T. A. Branch. Faster estimation of Bayesian models in ecology using Hamiltonian Monte Carlo. *Methods in Ecology and Evolution*, 8(3):339–348, 2017.

I. Murray and R. P. Adams. Slice sampling covariance hyperparameters of latent Gaussian models. In *Advances in Neural Information Processing Systems*, pages 1732–1740, 2010.

I. Murray, R. P. Adams, and D. Mackay. Elliptical slice sampling. *Journal of Machine Learning Research*, 9:541–548, 2010.

R. Neal. MCMC using Hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2, 2011.

R. M. Neal. Probabilistic inference using Markov chain Monte Carlo methods. Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto, 1993.

R. M. Neal. Bayesian learning for neural networks. *Lecture Notes in Statistics*, 118, 1996.

R. M. Neal. Regression and classification using Gaussian process priors. *Bayesian statistics*, 6: 475–501, 1998.

R. Opgen-Rhein, L. Fahrmeir, and K. Strimmer. Inference of demographic history from genealogical trees using reversible jump Markov chain Monte Carlo. *BMC Evolutionary Biology*, 5(1):6, 2005.

S. Pääbo, H. Poinar, D. Serre, V. Jaenicke-Després, J. Hebler, N. Rohland, M. Kuch, J. Krause, L. Vigilant, and M. Hofreiter. Genetic analyses from ancient DNA. *Annual Review of Genetics*, 38:645–679, 2004.

C. Paciorek and M. Schervish. Nonstationary covariance functions for Gaussian process regression. *Advances in Neural Information Processing Systems*, 16:273–280, 2004.

C. J. Paciorek and M. J. Schervish. Spatial modelling using a new class of nonstationary covariance functions. *Environmetrics*, 17(5):483–506, 2006.

J. A. Palacios and V. N. Minin. Integrated nested Laplace approximation for Bayesian nonparametric phylodynamics. *Proceedings of the Twenty-Eighth Conference on Uncertainty in Artificial Intelligence*, pages 726–735, 2012.

J. A. Palacios and V. N. Minin. Gaussian process-based Bayesian nonparametric inference of population size trajectories from gene genealogies. *Biometrics*, 69(1):8–18, 2013.

O. Papaspiliopoulos, G. O. Roberts, and M. Sköld. Non-centered parameterisations for hierarchical models and data augmentation. In J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, and M. West, editors, *Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting*, pages 307–326. Oxford University Press, USA, 2003.

O. Papaspiliopoulos, G. O. Roberts, and M. Sköld. A general framework for the parametrization of hierarchical models. *Statistical Science*, 22(1):59–73, 2007.

T. Park and G. Casella. The Bayesian lasso. *Journal of the American Statistical Association*, 103 (482):681–686, 2008.

N. G. Polson and J. G. Scott. Shrink globally, act locally: sparse Bayesian regularization and prediction. *Bayesian Statistics*, 9:501–538, 2010.

N. G. Polson and J. G. Scott. On the half-Cauchy prior for a global scale parameter. *Bayesian Analysis*, 7(4):887–902, 2012a.

N. G. Polson and J. G. Scott. Local shrinkage rules, Lévy processes and regularized regression. *Journal of the Royal Statistical Society: Series B (Methodological)*, 74(2):287–311, 2012b.

O. G. Pybus, A. Rambaut, and P. H. Harvey. An integrated framework for the inference of viral population history from reconstructed genealogies. *Genetics*, 155(3):1429–1437, 2000.

O. G. Pybus, M. A. Charleston, S. Gupta, A. Rambaut, E. C. Holmes, and P. H. Harvey. The epidemic behavior of the Hepatitis C virus. *Science*, 292(5525):2323–2325, 2001.

O. G. Pybus, A. J. Drummond, T. Nakano, B. H. Robertson, and A. Rambaut. The epidemiology and iatrogenic transmission of Hepatitis C virus in Egypt: a Bayesian coalescent approach. *Molecular Biology and Evolution*, 20(3):381–387, 2003.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL `http://www.R-project.org`.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL `http://www.R-project.org`.

A. E. Raftery and V. E. Akman. Bayesian analysis of a Poisson process with a change-point. *Biometrika*, 73(1):85–89, 1986.

A. E. Raftery, M. A. Newton, J. M. Satagopan, and P. N. Krivitsky. Estimating the integrated likelihood via posterior simulation using the harmonic mean identity. Technical Report 499, University of Washington, 2006.

A. Rambaut, O. G. Pybus, M. I. Nelson, C. Viboud, J. K. Taubenberger, and E. C. Holmes. The genomic and epidemiological dynamics of human Influenza A virus. *Nature*, 453(7195):615, 2008.

C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006.

D. A. Rasmussen, E. M. Volz, and K. Koelle. Phylodynamic inference for structured epidemiological models. *PLoS Computational Biology*, 10(4):e1003570, 2014.

S. C. Ray, R. R. Arthur, A. Carella, J. Bukh, and D. L. Thomas. Genetic epidemiology of Hepatitis C virus throughout Egypt. *The Journal of Infectious Diseases*, 182(3):698–707, 2000.

B. J. Reich, J. S. Hodges, and V. Zadnik. Effects of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics*, 62(4):1197–1206, 2006.

P. J. Reimer, M. G. Baillie, E. Bard, A. Bayliss, J. W. Beck, P. G. Blackwell, C. B. Ramsey, C. E. Buck, G. S. Burr, R. L. Edwards, et al. IntCal09 and Marine09 radiocarbon age calibration curves, 0–50,000 years cal BP. *Radiocarbon*, 51(4):1111–1150, 2009.

C. H. Reinsch. Smoothing by spline functions. *Numerische Mathematik*, 10(3):177–183, 1967.

N. Reményi and B. Vidakovic. Wavelet shrinkage with double Weibull prior. *Communications in Statistics-Simulation and Computation*, 44(1):88–104, 2015.

G. O. Roberts and O. Stramer. Langevin diffusions and Metropolis-Hastings algorithms. *Methodology and Computing in Applied Probability*, 4(4):337–357, 2002.

E. A. Roualdes. Bayesian trend filtering. *arXiv preprint arXiv:1505.07710*, 2015.

H. Rue. Fast sampling of Gaussian Markov random fields. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):325–338, 2001.

H. Rue and L. Held. *Gaussian Markov Random Fields: Theory and Applications*. CRC Press, 2005.

H. Rue and M. A. Hurn. Bayesian object identification. *Biometrika*, 86(3):649–660, 1999.

H. Rue, S. Martino, and N. Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392, 2009.

H. Rue, A. Riebler, S. H. Sørbye, J. B. Illian, D. P. Simpson, and F. K. Lindgren. Bayesian computing with INLA: a review. *Annual Review of Statistics and Its Application*, 4:395–421, 2017.

A. Rushworth, D. Lee, and C. Sarran. An adaptive spatiotemporal smoothing model for estimating trends and step changes in disease risk. *Journal of the Royal Statistical Society, Series C*, 66(1): 141–157, 2017.

P. Sagulenko, V. Puller, and R. A. Neher. TreeTime: maximum-likelihood phylodynamic analysis. *Virus evolution*, 4(1):vex042, 2018.

P. D. Sampson and P. Guttorp. Nonparametric estimation of nonstationary spatial covariance structure. *Journal of the American Statistical Association*, 87(417):108–119, 1992.

M. Scheffer, S. Carpenter, J. A. Foley, C. Folke, and B. Walker. Catastrophic shifts in ecosystems. *Nature*, 413(6856):591, 2001.

F. Scheipl and T. Kneib. Locally adaptive Bayesian p-splines with a normal-exponential-gamma prior. *Computational Statistics & Data Analysis*, 53(10):3533–3552, 2009.

B. Shapiro and M. Hofreiter. A paleogenomic perspective on evolution and gene function: new insights from ancient DNA. *Science*, 343(6169):1236573, 2014.

B. Shapiro, A. J. Drummond, A. Rambaut, M. C. Wilson, P. E. Matheus, A. V. Sher, O. G. Pybus, M. T. P. Gilbert, I. Barnes, J. Binladen, et al. Rise and fall of the Beringian steppe bison. *Science*, 306(5701):1561–1565, 2004.

B. Shapiro, S. Y. Ho, A. J. Drummond, M. A. Suchard, O. G. Pybus, and A. Rambaut. A Bayesian phylogenetic method to estimate unknown sequence ages. *Molecular Biology and Evolution*, 28 (2):879–887, 2010.

D. P. Simpson, T. G. Martins, A. Riebler, G.-A. Fuglstad, H. Rue, and S. H. Sørbye. Penalising model component complexity: A principled, practical approach to constructing priors. *arXiv preprint arXiv:1403.4630*, 2014.

S. H. Sørbye and H. Rue. Scaling intrinsic Gaussian Markov random field priors in spatial modelling. *Spatial Statistics*, 8:39–51, 2014.

P. L. Speckman and D. Sun. Fully Bayesian spline smoothing and intrinsic autoregressive priors. *Biometrika*, 90(2):289–302, 2003.

Stan Development Team. RStan: the R interface to Stan, Version 2.6.2, 2015a. URL `http://mc-stan.org/rstan.html`.

Stan Development Team. *Stan Modeling Language Users Guide and Reference Manual, Version 2.6.2*, 2015b. URL `http://mc-stan.org/`.

Stan Development Team. RStan: the R interface to Stan, Version 2.14.2, 2017. URL `http://mc-stan.org/rstan.html`.

K. Strimmer and O. G. Pybus. Exploring the demographic history of DNA sequences using the generalized skyline plot. *Molecular Biology and Evolution*, 18(12):2298–2305, 2001.

Y. W. Teh and V. Rao. Gaussian process modulated renewal processes. In *Advances in Neural Information Processing Systems*, pages 2474–2482, 2011.

G. R. Terrell, D. W. Scott, et al. Variable kernel density estimation. *The Annals of Statistics*, 20 (3):1236–1265, 1992.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

R. Tibshirani. Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(3):273–282, 2011.

R. J. Tibshirani. Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics*, 42(1):285–323, 2014.

R. J. Tibshirani and J. Taylor. The solution path of the generalized lasso. *The Annals of Statistics*, 39(3):1335–1371, 2011.

T.-H. To, M. Jung, S. Lycett, and O. Gascuel. Fast dating using least-squares criteria and algorithms. *Systematic Biology*, 65(1):82–97, 2015.

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

E. Volz and S. Pond. Phylodynamic analysis of Ebola virus in the 2014 Sierra Leone epidemic. *PLoS Currents*, 6, 2014.

R. P. Waagepetersen. An estimating function approach to inference for inhomogeneous Neyman–Scott processes. *Biometrics*, 63(1):252–258, 2007.

G. Wahba. Smoothing noisy data with spline functions. *Numerische Mathematik*, 24(5):383–393, 1975.

J. Wallin and D. Bolin. Geostatistical modelling using non-Gaussian Matérn fields. *Scandinavian Journal of Statistics*, 42(3):872–890, 2015.

S. Watanabe. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11:3571–3594, 2010.

M. West. On scale mixtures of normal distributions. *Biometrika*, 74(3):646–648, 1987.

E. T. Whittaker. On a new method of graduation. *Proceedings of the Edinburgh Mathematical Society*, 41:63–75, 1922.

W. Xie, P. O. Lewis, Y. Fan, L. Kuo, and M.-H. Chen. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Systematic Biology*, 60(2):150–160, 2011.

Z. Yang. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *Journal of Molecular Evolution*, 39:306–314, 1994.

Z. Yang. *Molecular evolution: a statistical approach*. Oxford University Press, 2014.

Y. Yue and P. L. Speckman. Nonstationary spatial Gaussian Markov random fields. *Journal of Computational and Graphical Statistics*, 19(1):96–116, 2010.

Y. R. Yue and J. M. Loh. Bayesian semiparametric intensity estimation for inhomogeneous spatial point processes. *Biometrics*, 67(3):937–946, 2011.

Y. R. Yue, J. M. Loh, and M. A. Lindquist. Adaptive spatial smoothing of fMRI images. *Statistics and Its Interface*, 3(1):3–13, 2010.

Y. R. Yue, P. L. Speckman, and D. Sun. Priors for Bayesian adaptive spline smoothing. *Annals of the Institute of Statistical Mathematics*, 64(3):577–613, 2012.

Y. R. Yue, D. Simpson, F. Lindgren, and H. Rue. Bayesian adaptive smoothing splines using stochastic differential equations. *Bayesian Analysis*, 9(2):397–424, 2014.

C. Zheng and H. Yao. Segmentation for remote-sensing imagery using the object-based Gaussian-Markov random field model with region coefficients. *International Journal of Remote Sensing*, 40(11):1–32, 2019.

B. Zhu and D. B. Dunson. Locally adaptive Bayes nonparametric regression via nested Gaussian processes. *Journal of the American Statistical Association*, 108(504):1445–1456, 2013.

Appendix A

APPENDIX FOR CHAPTER 3

A.1 Approximation to the Horseshoe Density

There is no exact closed-form expression available for the horseshoe density function. We present an approximation to the horseshoe density that can be used without the need for explicit specification of the nuisance local scale parameters. Following Carvalho et al. (2010), the marginal distribution of u given global scale parameter γ is found by integrating over possible values of the local scale parameter τ, where $u \,|\, \tau \sim \mathrm{N}(0, \delta\tau^2)$ and $\tau \,|\, \gamma \sim \mathrm{C}^+(0, \gamma)$. Here δ is a constant representing a scale factor for the distance between adjacent points when this distribution is used for the increments of a kth-order smoothing model. This leads to

$$\begin{aligned} p(u \,|\, \delta, \gamma) &= \int_0^\infty p(u \,|\, \delta, \tau, \gamma) p(\tau \,|\, \gamma) d\tau \\ &= \int_0^\infty \frac{1}{\sqrt{2\pi\delta\tau^2}} \exp\left(-\frac{u^2}{2\delta\tau^2}\right) \frac{2\gamma}{\pi(\tau^2 + \gamma^2)} d\tau \end{aligned}$$

We let $B = 2\gamma/(\sqrt{2\pi^3\delta})$ and introduce the substitution $\omega = \tau^{-2}$, which gives $d\tau = -1/(2\omega^{3/2})$, resulting in

$$\begin{aligned} p(u \,|\, \delta, \gamma) &= B \int_0^\infty \frac{\omega^{1/2}}{2\omega^{3/2}} \exp\left(-\frac{u^2\omega}{2\delta}\right) \frac{1}{\omega^{-1} + \gamma^2} d\omega \\ &= \frac{B}{2} \int_0^\infty \frac{1}{\omega} \exp\left(-\frac{u^2\omega}{2\delta}\right) \frac{\omega}{1 + \omega\gamma^2} d\omega \\ &= \frac{B}{2} \int_0^\infty \frac{1}{1 + \omega\gamma^2} \exp\left(-\frac{u^2\omega}{2\delta}\right) d\omega. \end{aligned}$$

Now we introduce the substitution $z = 1 + \omega\gamma^2$, which gives $d\omega = \gamma^{-2} dz$, and results in

$$\begin{aligned} p(u \,|\, \delta, \gamma) &= \frac{B}{2} \int_1^\infty \frac{1}{z\gamma^2} \exp\left\{-\left(\frac{u^2\omega}{2\delta}\right)\frac{z}{\gamma^2} + \left(\frac{u^2}{2\delta}\right)\frac{1}{\gamma^2}\right\} dz \\ &= \frac{B}{2\gamma^2} \exp\left(\frac{u^2}{2\delta\gamma^2}\right) \int_1^\infty \frac{1}{z} \exp\left(-\frac{u^2 z}{2\delta\gamma^2}\right) dz \\ &= \left(\frac{1}{2\pi^3\delta\gamma^2}\right)^{1/2} \exp\left(\frac{u^2}{2\delta\gamma^2}\right) \mathrm{E}_1\left(\frac{u^2}{2\delta\gamma^2}\right) dz, \end{aligned}$$

where E_1 is the exponential integral function. Note that $\lim_{x\to 0^+} \mathrm{E}_1(x) = \infty$, but for $x > 0$, the function $\mathrm{E}_1(x)$ is bounded as follows:

$$\frac{1}{2}e^{-x}\ln\left(1+\frac{2}{x}\right) < \mathrm{E}_1(x) < e^{-x}\ln\left(1+\frac{1}{x}\right).$$

Then for $u \in \{\mathbb{R} : u \neq 0\}$ we have

$$\frac{1}{2}\exp\left(\frac{-u^2}{2\delta\gamma^2}\right)\ln\left(1+\frac{4\delta\gamma^2}{u^2}\right) < \mathrm{E}_1\left(\frac{u^2}{2\delta\gamma^2}\right) < \exp\left(\frac{-u^2}{2\delta\gamma^2}\right)\ln\left(1+\frac{2\delta\gamma^2}{u^2}\right).$$

It follows that the target density is bounded by

$$\frac{1}{2}\left(\frac{1}{2\pi^3\delta\gamma^2}\right)^{1/2}\ln\left(1+\frac{4\delta\gamma^2}{u^2}\right) < p(u\,|\,\gamma) < \left(\frac{1}{2\pi^3\delta\gamma^2}\right)^{1/2}\ln\left(1+\frac{2\delta\gamma^2}{u^2}\right). \tag{A.1}$$

Let the left bound in equation (A.1) be denoted $B_1(u)$ and the right bound $B_2(u)$. Note that as $u \to 0$, each of $B_1(u)$, $p(u\,|\,\gamma)$ and $B_2(u)$ approach ∞. It can be shown that $\int_{-\infty}^{\infty} B_1(u)du = \sqrt{2/\pi}$ and $\int_{-\infty}^{\infty} B_2(u)du = 2/\sqrt{\pi}$. Since $\sqrt{2/\pi} < 1 < 2/\sqrt{\pi}$, these bounds can be used to find an approximate expression for $p(u\,|\,\gamma)$ that integrates to 1 and still satisfies equation (A.1). We set

$$\tilde{p}(u\,|\,\gamma) = wB_1(u) + (1-w)B_2(u) \tag{A.2}$$

with constraints $0 < w < 1$ and $\int_{-\infty}^{\infty} wB_1(u)+(1-w)B_2(u)du = 1$. Using the values for the integrated bounds and solving gives $w = (\sqrt{\pi}-2)/(\sqrt{2}-2)$. Substituting this value for w into equation (A.2) and simplifying gives the following closed-form approximation to the horseshoe density function:

$$\tilde{p}(u\,|\,\gamma) = \left(\frac{1}{2\pi^3\delta\gamma^2}\right)^{1/2}\left[\frac{\sqrt{\pi}-2}{2\sqrt{2}-4}\ln\left(1+\frac{4\delta\gamma^2}{u^2}\right) + \frac{\sqrt{2}-\sqrt{\pi}}{\sqrt{2}-2}\ln\left(1+\frac{2\delta\gamma^2}{u^2}\right)\right]. \tag{A.3}$$

A.2 Marginal Laplace Distribution with Irregular Grid Spacing

The following is a derivation of the marginal prior distribution for the order-k differences when grid spacing is unequal. These derivations are based on the scale-mixture representation of the Laplace distribution. These results are known to apply to the first-order and second-order models, but higher orders.

Let $u_j = \Delta^k \theta_j$ and let δ_j be a constant representing a scale factor for the distance between adjacent points when this distribution is used for the increments of a kth-order smoothing model. For convenience, subscripts on u and δ are dropped from here forward. We assume $u|\tau, \delta \sim \mathrm{N}(0, \delta\tau^2)$ and $\tau^2|\lambda^2 \sim \mathrm{Exp}(\lambda^2/2)$. Here $\lambda = 1/\gamma$ is the global shrinkage parameter. It follows that

$$\begin{aligned} p(u|\delta, \lambda) &= \int_0^\infty \frac{\lambda^2}{2} \exp\left(-\frac{\tau^2\lambda^2}{2}\right) \frac{1}{\sqrt{2\pi\delta\tau^2}} \exp\left(-\frac{u^2}{2\delta\tau^2}\right) d\tau^2 \\ &= A \int_0^\infty \frac{1}{\tau} \exp\left(-\frac{\tau^2\lambda^2}{2} - \frac{u^2}{2\delta\tau^2}\right) d\tau^2, \end{aligned}$$

where $A = \frac{\lambda^2}{2\sqrt{2\pi\delta\tau^2}}$. Now we make the substitution $\omega = 1/\tau^2$, which gives $d\tau^2 = -\omega^{-2} d\omega$, and the marginal density for u becomes

$$\begin{aligned} p(u|\delta, \lambda) &= A \int_0^\infty \omega^{-3/2} \exp\left\{-\frac{\lambda^2}{2\omega} - \frac{u^2\omega}{2\delta}\right\} d\omega \\ &= A \int_0^\infty \omega^{-3/2} \exp\left\{-\frac{u^2\omega}{2\delta} - \frac{\lambda^2}{2\omega} + \frac{\lambda|u|}{\delta^{1/2}} - \frac{\lambda|u|}{\delta^{1/2}}\right\} d\omega \\ &= A \int_0^\infty \omega^{-3/2} \exp\left\{-\frac{|u|^2}{2\delta\omega}\left(\omega^2 - \frac{2\delta^{1/2}\omega\lambda}{|u|} + \frac{\delta\lambda^2}{|u|^2}\right) - \frac{\lambda|u|}{\delta^{1/2}}\right\} d\omega \\ &= \frac{\lambda}{2\sqrt{\delta}} \exp\left\{-\frac{\lambda|u|}{\sqrt{\delta}}\right\} \int_0^\infty \frac{\lambda}{\sqrt{2\pi}\omega^{3/2}} \exp\left\{-\frac{\lambda^2}{2\delta\omega(\lambda^2/|u|^2)}\left(\omega - \frac{\sqrt{\delta}\lambda}{|u|}\right)^2\right\} d\omega \\ &= \frac{\lambda}{2\sqrt{\delta}} \exp\left\{-\frac{\lambda|u|}{\sqrt{\delta}}\right\}, \end{aligned}$$

where the last line follows from the fact that the integrand in the second-to-last line is the pdf of an inverse-Gaussian distribution with mean parameter $\mu = \sqrt{\delta}\lambda/|u|$ and shape parameter $\lambda = \lambda^2$. The result is the pdf of a Laplace distribution with mean zero and scale parameter $\lambda/\sqrt{\delta}$. Note that

the variance of the Laplace distribution is $2\delta/\lambda^2$, which implies that the grid spacing δ scales the variance of the increments u.

A.3 Data Example with Irregular Grid

We apply the SPMRF models to a data set with a continuous covariate. The response data are rent per square meter of floor space in Munich, Germany, and the covariate is the floor space in square meters. These data were analysed by Rue and Held (2005) using a second-order GMRF with irregular spacing. Here we apply a second-order GMRF and SPMRF models using the methods described in Section 3.2.4 of the main text.

Let x represent the floor space measurements, and let $x_1 < x_2 < ... < x_n$ be the ordered set of unique floor measurement values. Further, let $\delta_j = x_{j+1} - x_j$ be the distance between adjacent floor space measurements. The marginal prior distributions for the second-order differences were $\Delta^2\theta_j \sim \mathrm{N}(0, d_j\gamma^2)$ for the normal prior, $\Delta^2\theta_j \sim \mathrm{Laplace}(d_j^{1/2}\gamma)$ for the Laplace, and $\Delta^2\theta_j \sim \mathrm{HS}(d_j^{1/2}\gamma)$ for the horseshoe, where

$$\Delta^2\theta_j = \theta(x_{j+2}) - \left(1 + \frac{\delta_{j+1}}{\delta_j}\right)\theta(x_{j+1}) + \frac{\delta_{j+1}}{\delta_j}\theta(x_j),$$

and

$$d_j = \frac{\delta_{j+1}^2(\delta_j + \delta_{j+1})}{2}.$$

Using methods described in Section 3.4.1 of the main text and Appendix A.5, we calculated the value of the hyperparameter for the global scale parameter to be $\zeta = 0.00094$, so $\gamma \sim \mathrm{C}^+(0, 0.00094)$ for all models. The results are shown in Figure A.1.

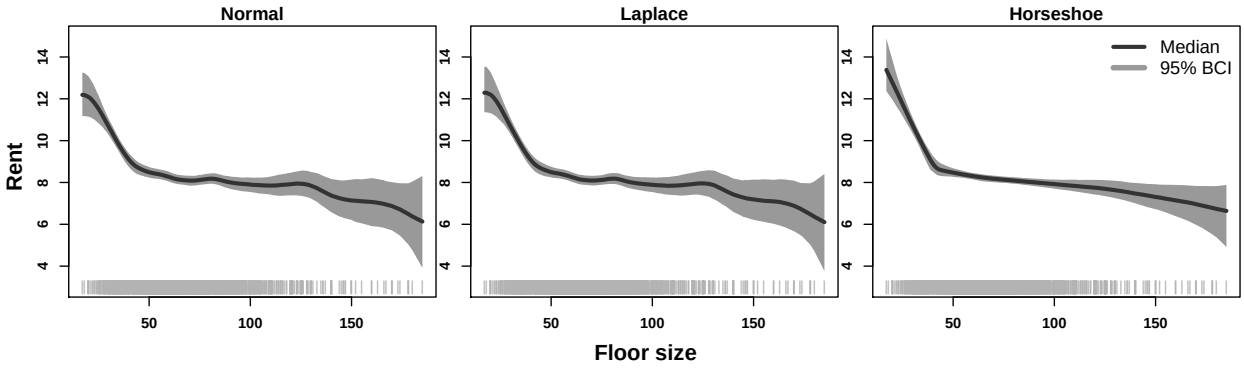

Figure A.1: Results for models using irregular grids for Munich rent data. Posterior medians (dark line) are shown with 95% Bayesian credible intervals (BCIs). Locations of data are shown with vertical bars at the bottom of plots.

A.4 Additional Simulation Results

Here we display plots with simulation results for normal data with $\sigma = 1.5$ (Figure A.2), Poisson data (Figure A.3), and binomial data (Figure A.4). Summary measures for all data types show similar patterns to each other and to those for normal data with $\sigma = 4.5$ (Figure 2 in main article).

Table A.1: Mean values of performance measures across 100 simulations for normal observations ($\sigma = 1.5$) for each model and trend function type.

Function	**Model**	**MAD**	**MCIW**	**MASV**	**TMASV**
Constant	Normal	0.115	0.698	0.002	0.000
	Laplace	0.116	0.731	0.002	0.000
	Horseshoe	0.127	0.921	0.004	0.000
Piecewise Const.	Normal	1.040	5.479	1.647	0.606
	Laplace	0.899	3.282	1.557	0.606
	Horseshoe	0.281	1.918	0.638	0.606
Smooth	Normal	0.565	2.985	1.391	1.406
	Laplace	0.561	2.985	1.393	1.406
	Horseshoe	0.565	2.946	1.414	1.406
Varying Smooth	Normal	0.586	3.036	0.613	0.543
	Laplace	0.550	2.898	0.592	0.543
	Horseshoe	0.438	2.228	0.558	0.543

Table A.2: Mean values of performance measures across 100 simulations for Poisson observations for each model and trend function type.

Function	**Model**	**MAD**	**MCIW**	**MASV**	**TMASV**
Constant	Normal	0.022	0.142	0.001	0.000
	Laplace	0.023	0.149	0.001	0.000
	Horseshoe	0.025	0.167	0.001	0.000
Piecewise Const.	Normal	0.109	0.557	0.077	0.030
	Laplace	0.092	0.529	0.064	0.030
	Horseshoe	0.051	0.334	0.036	0.030
Smooth	Normal	0.078	0.379	0.072	0.079
	Laplace	0.078	0.380	0.072	0.079
	Horseshoe	0.079	0.382	0.073	0.079
Varying Smooth	Normal	0.067	0.296	0.020	0.023
	Laplace	0.066	0.295	0.020	0.023
	Horseshoe	0.058	0.277	0.020	0.023

Table A.3: Mean values of performance measures across 100 simulations for binomial observations for each model and trend function type.

Function	**Model**	**MAD**	**MCIW**	**MASV**	**TMASV**
Constant	Normal	0.042	0.249	0.001	0.000
	Laplace	0.043	0.262	0.001	0.000
	Horseshoe	0.047	0.311	0.002	0.000
Piecewise Const.	Normal	0.229	1.191	0.166	0.066
	Laplace	0.193	1.126	0.137	0.066
	Horseshoe	0.108	0.690	0.076	0.066
Smooth	Normal	0.139	0.733	0.110	0.117
	Laplace	0.139	0.735	0.111	0.117
	Horseshoe	0.143	0.740	0.113	0.117
Varying Smooth	Normal	0.188	0.730	0.056	0.068
	Laplace	0.183	0.726	0.056	0.068
	Horseshoe	0.149	0.676	0.058	0.068

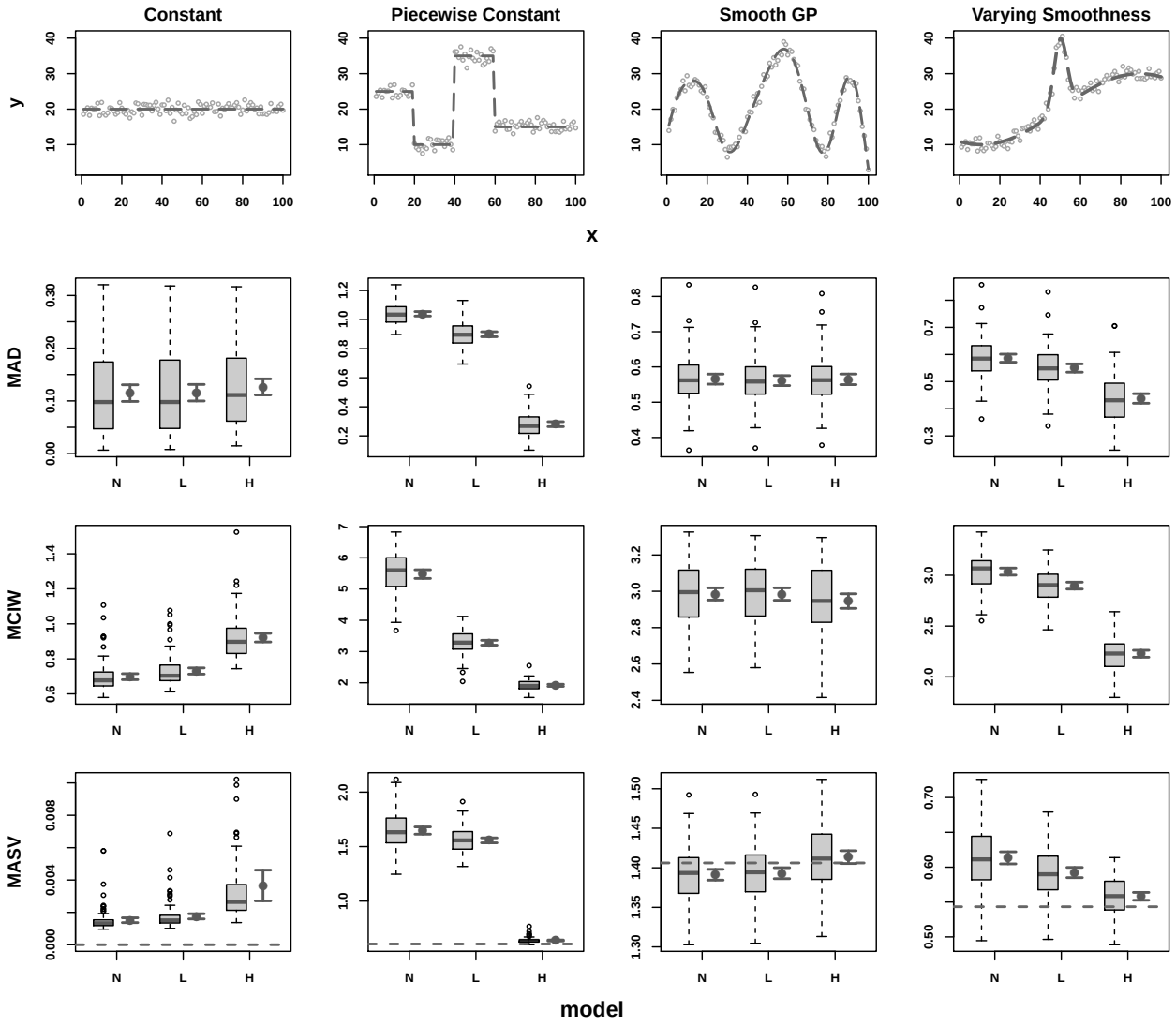

Figure A.2: Functions used in simulations and simulation results by model (N=normal, L=Laplace, H=horseshoe) and function type for normally distributed data with $\sigma = 1.5$. Top row shows true functions (dashed lines) with example simulated data. Remaining rows show mean absolute deviation (MAD), mean credible interval width (MCIW), and mean absolute sequential variation (MASV). Horizontal dashed line in plots on bottom row is the true mean absolute sequential variation (TMASV). Shown for each model are standard boxplots of simulation results (left) and mean values with 95% frequentist confidence intervals (right).

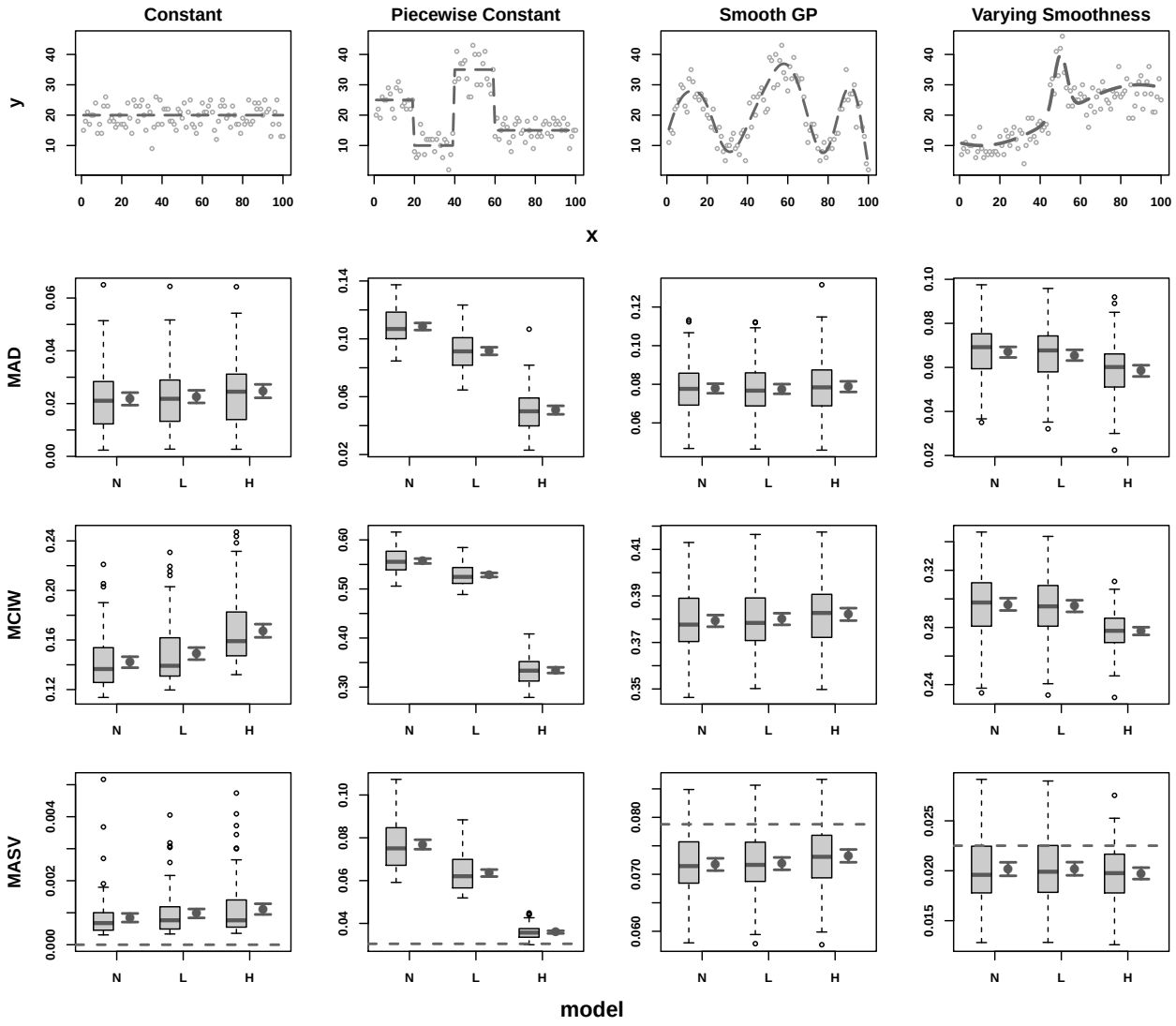

Figure A.3: Functions used in simulations and simulation results by model (N=normal, L=Laplace, H=horseshoe) and function type for Poisson distributed data. Top row shows true functions (dashed lines) with example simulated data. Remaining rows show mean absolute deviation (MAD), mean credible interval width (MCIW), and mean absolute sequential variation (MASV). Horizontal dashed line in plots on bottom row is the true mean absolute sequential variation (TMASV). Shown for each model are standard boxplots of simulation results (left) and mean values with 95 % frequentist confidence intervals (right).

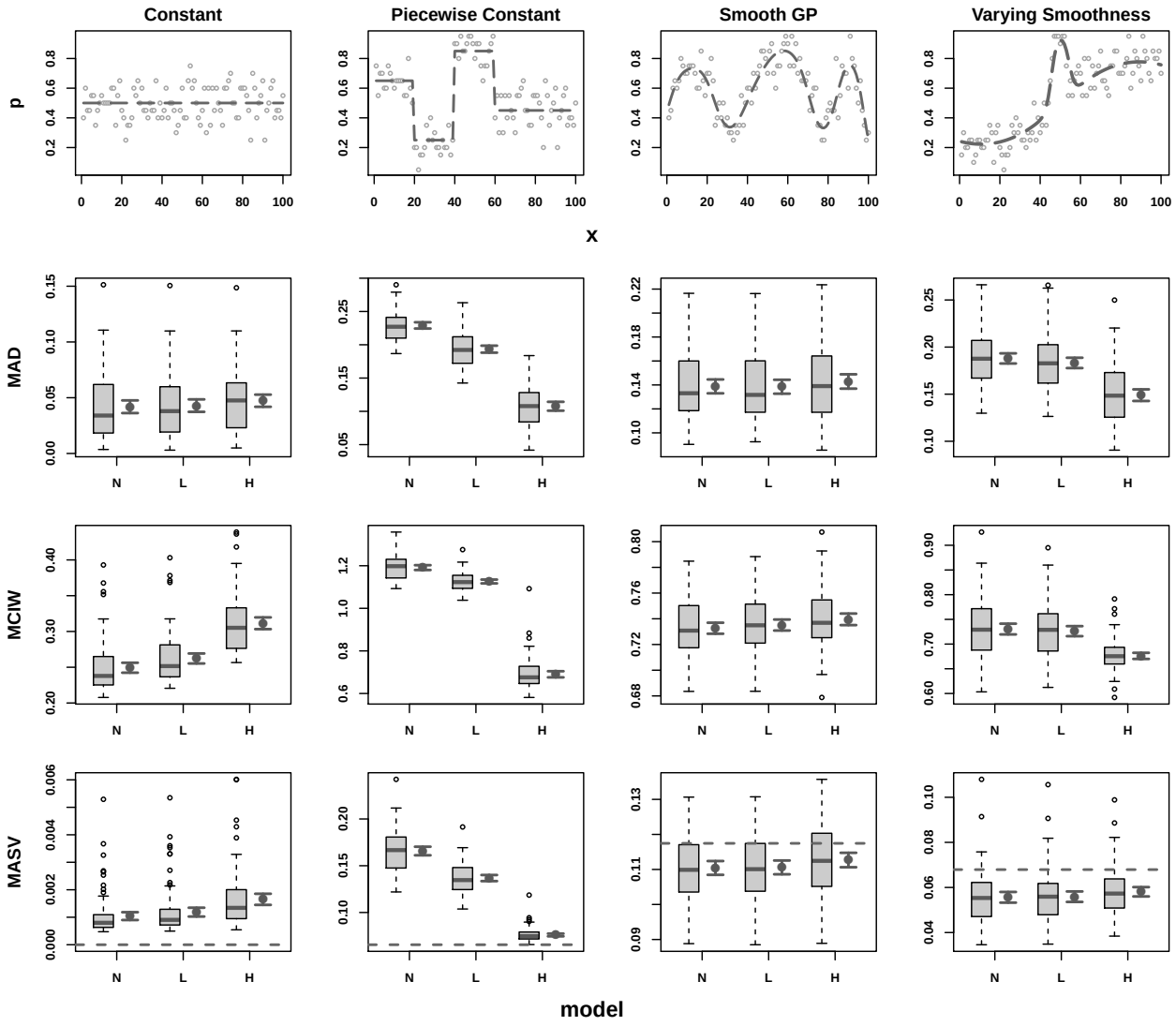

Figure A.4: Functions used in simulations and simulation results by model (N=normal, L=Laplace, H=horseshoe) and function type for binomial distributed data. Top row shows true functions (dashed lines) with empirical probability estimates from example simulated data. Remaining rows show mean absolute deviation (MAD), mean credible interval width (MCIW), and mean absolute sequential variation (MASV). Horizontal dashed line in plots on bottom row is the true mean absolute sequential variation (TMASV). Shown for each model are standard boxplots of simulation results (left) and mean values with 95% frequentist confidence intervals (right).

A.5 Parameterizing the Global Smoothing Prior

Here we provide additional details for calculating the hyperparameter ζ for the prior on the global scale parameter γ, where $\gamma \sim \mathrm{C}^+(0, \zeta)$. First let $\boldsymbol{Q}$ be the precision matrix for the Markov random field corresponding to the model of interest (see examples below), and $\boldsymbol{\Sigma} = \boldsymbol{Q}^{-1}$ be the covariance matrix with diagonal elements Σ_{ii}. Following Sørbye and Rue (2014), the marginal standard deviation of all components of $\boldsymbol{\theta}$ for a fixed value of γ is $\sigma_\gamma(\theta_i) = \gamma\sigma_{\text{ref}}(\boldsymbol{\theta})$, where

$$\sigma_{\text{ref}}(\boldsymbol{\theta}) = \exp\left(\frac{1}{n}\sum_{i=1}^{n}\log\sigma_{\{\gamma=1\}}(\theta_i)\right) = \exp\left(\frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\log\left(\Sigma_{ii}\right)\right). \tag{A.4}$$

That is, $\sigma_{\text{ref}}(\boldsymbol{\theta})$ is the geometric mean of the individual marginal standard deviations when $\gamma = 1$.

Before going further, let us describe two precision matrices used in the accompanying paper and their associated covariance matrices. Sørbye and Rue (2014) use intrinsic formulations of a Gaussian Markov random field (GRMF), which is also possible with our models, but we chose to use proper GMRF models. This requires specification of the variance of the first θ, which we denote $\omega^2 = \mathrm{Var}(\theta_1)$. For a given value of γ, the $n \times n$ precision matrix for a first-order model is:

$$Q_1 = 1/\gamma^2\begin{pmatrix} \gamma^2/\omega^2+1 & -1 & & & & & \\ -1 & 2 & -1 & & & & \\ & -1 & 2 & -1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & -1 & 2 & -1 & \\ & & & & -1 & 1 \end{pmatrix}, \tag{A.5}$$

and the corresponding covariance matrix is:

$$\Sigma_1 = \begin{pmatrix} \omega^2 & \omega^2 & \cdots & & & \omega^2 \\ \omega^2 & \omega^2+\gamma^2 & \omega^2+\gamma^2 & \cdots & & \omega^2+\gamma^2 \\ \vdots & \omega^2+\gamma^2 & \omega^2+2\gamma^2 & \omega^2+2\gamma^2 & \cdots & \omega^2+2\gamma^2 \\ & \vdots & \omega^2+2\gamma^2 & \ddots & & \vdots \\ & & \vdots & & \omega^2+(n-2)\gamma^2 & \omega^2+(n-2)\gamma^2 \\ \omega^2 & \omega^2+\gamma^2 & \omega^2+2\gamma^2 & \cdots & \omega^2+(n-2)\gamma^2 & \omega^2+(n-1)\gamma^2 \end{pmatrix}. \tag{A.6}$$

Therefore, the marginal variances for the first-order model are $\Sigma_{1,ii} = \omega^2 + (i-1)\gamma^2$. For the second-order model, the $n \times n$ precision matrix is:

$$Q_2 = 1/\gamma^2 \begin{pmatrix} \gamma^2/\omega^2 + 2 & -3 & 1 & & & & & & \\ -3 & 6 & -4 & 1 & & & & & \\ 1 & -4 & 6 & -4 & 1 & & & & \\ & 1 & -4 & 6 & -4 & 1 & & & \\ & & \ddots & \ddots & \ddots & \ddots & \ddots & & \\ & & & 1 & -4 & 6 & -4 & 1 & \\ & & & & 1 & -4 & 6 & -4 & 1 \\ & & & & & 1 & -4 & 5 & -2 \\ & & & & & & 1 & -2 & 1 \end{pmatrix}. \tag{A.7}$$

There is an analytical form for the covariance matrix for the second-order model, but it suffices here to know that the form of the marginal variances is:

$$\Sigma_{2,ii} = \omega^2 + \frac{i(i-1)(2i-1)}{6}\gamma^2. \tag{A.8}$$

Note that if we were using an intrinsic GMRF model, we would assume that ω^2 is infinite, which would result in a covariance matrix of rank $n-k$. Following Sørbye and Rue (2014) we would then use the generalized inverse of the precision matrix to calculate the marginal variances.

In practice we use the variance of the data (transformed data if the θ parameters are on a transformed scale) as an estimate of ω^2. Although this is using the data twice, this offers a reasonable constraint on the marginal variances of the θs.

We want to set an upper bound U on the average marginal standard deviation of θ_i, such that $\Pr(\sigma_\gamma(\theta_i) > U) = \alpha$, where α is some small probability. Using the cumulative probability function for a half-Cauchy distribution, we can find a value of ζ for a given value of $\sigma_{\text{ref}}(\boldsymbol{\theta})$ specific to a model of interest and given common values of U and α by:

$$\zeta = \frac{U}{\sigma_{\text{ref}}(\boldsymbol{\theta}) \tan\left(\frac{\pi}{2}(1-\alpha)\right)}. \tag{A.9}$$

It may be useful to note here that the median of a half-Cauchy distribution is equal to its scale parameter, since the median may be a more intuitive measure of the effect of different values of ζ.

For our data examples in the main text, we let U be the estimated standard deviation of the data on the appropriate scale. We know that the marginal variances of the θs should not exceed the variance in the observed data, on average. We set $\alpha = 0.05$ as the probability of the average marginal standard deviation exceeding U. For the coal mining example in the main text, we found an estimate of the variance of the data on the log scale by $\sum_{i=1}^{n} \ln(y_i + 0.5)/(n-1)$, where y_i is the observed count at time $i = 1, \ldots, n$. For the Tokyo rain example, we estimated the variance of the data on the logit scale as $\sum_{i=1}^{n} \text{logit}((y_i + q_i)/m_i))/(n-1)$, where y_i is the number of years with rain on day i out of m_i possible years, and $q_i = 0.005 I_{y_i=0} - 0.005 I_{y_i=1} + 0 I_{y_i \notin \{0,1\}}$, where I is an indicator function.

Suppose we have calculated ζ_{o1}, the hyperparameter for a first-order model given the corresponding average marginal standard deviation $\sigma_{\text{ref}}(\boldsymbol{\theta}_{o1})$ using Equation (A.9). If we wish to calculate the value of ζ_{o2} for a second-order model we can simply use

$$\zeta_{o2} = \zeta_{o1} \frac{\sigma_{\text{ref}}(\boldsymbol{\theta}_{o1})}{\sigma_{\text{ref}}(\boldsymbol{\theta}_{o2})}.$$

Now suppose we have a model with n equidistant nodes and want to increase the density of the grid to kn nodes without changing the range of the grid. For a first-order model, $\text{Var}(\Delta\theta_{\text{new}}) = \frac{1}{k}\text{Var}(\Delta\theta)$, and for a second-order model $\text{Var}(\Delta^2\theta_{\text{new}}) = \frac{1}{k^3}\text{Var}(\Delta^2\theta)$ (Lindgren and Rue, 2008; Sørbye and Rue, 2014). In terms of the hyperparameter for the global smoothing prior, for the first-order model $\zeta_{o1,\text{new}} = k^{-1/2}\zeta_{o1}$, and for the second-order model $\zeta_{o2,\text{new}} = k^{-3/2}\zeta_{o2}$.

A.6 Prior Sensitivity

We tested the sensitivity of the three prior formulations (normal, Laplace, and horseshoe) to the value of the hyperparameter (ζ) which controls the scale of the distribution on the smoothing parameter γ, where $\gamma \sim \mathrm{C}^{+}(0, \zeta)$ A smaller value of ζ constricts γ to be closer to zero, which in turn constricts the scales of the priors on the order-k differences. We tested three levels for the hyperparameter: a) $\zeta = 1$, b) $\zeta = 0.01$, and c) $\zeta = 0.0001$. In general, we expect noisier data sets should be more sensitive to prior settings. The coal mine disaster data offered a good test set because the observations are relatively noisy.

Clearly the horseshoe prior was the most sensitive to the level of ζ (Figure A.5). In particular, the horseshoe results for $\zeta = 1$ looked more like those for the other two models in Figure 3.4, but when $\zeta = 0.0001$, the horseshoe produced more defined break points and straighter lines with narrower BCIs compared to the results with $\zeta = 0.01$.

A.7 Computational Efficiency

To compare SPMRF and GMRF models' computational efficiency, we calculated the effective number of posterior samples per second of computation time (ESSps) for different model formulations and data configurations. We used the scenario with a piecewise-constant expected value from our main simulations (Section 3) to test the effect of model type, model order, and number of grid cells (n) on the ESS per second of sampling time. Here sampling time is defined as the total run time minus the time spent in the adaptation (warm-up or burn-in) phase, where time is measured in seconds of CPU time. We also calculated the ESSps for the coal mining and Tokyo rainfall examples.

There were three simulated scenarios with piecewise-constant trend: 1) order-1 with $n = 100$ observations and grid points (one observation per grid point), 2) order-1 with $n = 200$, and 3) order-2 with $n = 100$. The observations in these scenarios were normally distributed with standard deviation $\sigma = 4.5$. For each of these scenarios we ran 4 independent chains each with 1,000 iterations of burn-in and 2,500 iterations post burn-in thinned at every 5 for a total of 2,000 retained

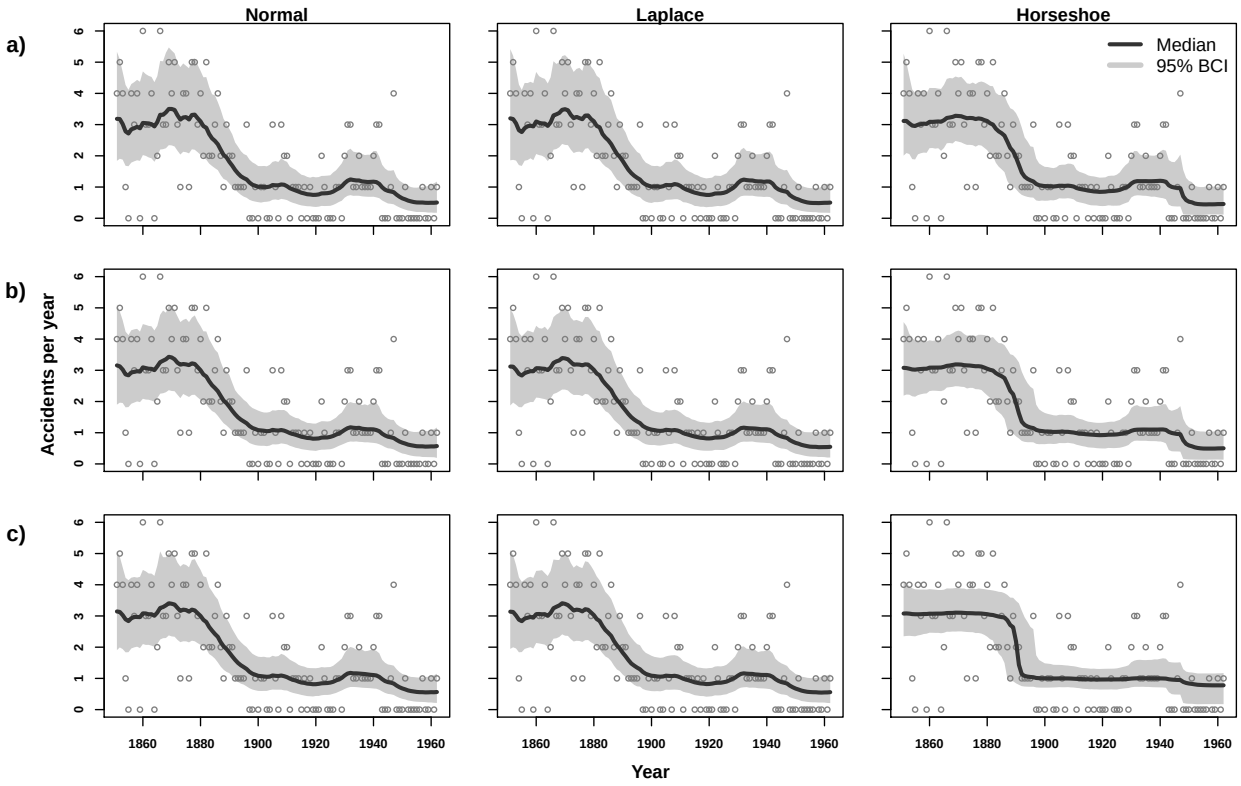

Figure A.5: Models fits to coal mining accidents data by model type and value of hyperparameter for global scale: a) $\zeta = 1$, b) $\zeta = 0.01$, and c) $\zeta = 0.0001$. Posterior medians and associated 95% Bayesian credible intervals are shown along with observed data.

posterior samples combined across chains. The maximum ESS would therefore be 2,000 for these scenarios. The chains were run in sequence so that the total time (TCPU) sampling times (SCPU) times are the respective cumulative times across 4 chains. For the coal mining and Tokyo rainfall examples we used the same settings for number of iterations and thinning as was used in the main text (Section 4). We calculated effective sample size using the methods described in the documentation for `stan` (Stan Development Team, 2015b).

Tests indicated that doubling the number of grid points (with a single observation per grid point) resulted in approximately 60% fewer ESSps for each model formulation (63% fewer for both the Normal and Laplace and 53% for the Horseshoe), and changing from a first- to second-order model resulted in approximately 90% fewer ESSps (91% fewer for the Normal, 90% for the Laplace, and 83% for the Horseshoe). On average across the five scenarios investigated, the Laplace formulation resulted in 77% fewer ESSps compared to the Normal formulation, and the Horseshoe resulted in 90% fewer ESSps. In terms of sampling times, the Laplace formulations on average took 4.8 times longer to achieve the same number of effective samples as the Normal formulations (range: 2.5 to 5.7 times longer), and the Horseshoe formulations took an average of 12.2 times longer than the Normal (range: 4.6 to 17.4).

Table A.4: Measures of computational efficiency for each model type (Normal (N), Laplace (L), or Horseshoe (H)) for three simulated data scenarios and two real data examples. Model order and number of parameters (p) are shown. The total CPU time (TCPU: includes adaptive phase) and sampling CPU (SCPU) time are in seconds. The minimum and mean effective sample sizes per second (ESSps) of SCPU are also shown.

Scenario	**Model**	**Order**	***p***	**TCPU**	**SCPU**	**Min. ESSps**	**Mean ESSps**
Piece. Const. (n = 100)	N	1	102	74	52	25.14	36.79
	L	1	201	422	290	5.13	6.58
	H	1	201	1,228	897	0.80	2.12
Piece. Const. (n = 200)	N	1	202	198	141	10.52	13.82
	L	1	401	1,195	794	1.86	2.44
	H	1	401	2,709	1,898	0.19	1.00
Piece. Const. (n = 100)	N	2	102	797	592	2.61	3.26
	L	2	200	3,916	2,770	0.52	0.68
	H	2	200	4,822	3,473	0.12	0.37
Coal Mining	N	1	113	42	37	121.00	133.48
	L	1	224	228	200	20.07	24.51
	H	1	224	639	580	5.57	8.20
Tokyo Rainfall	N	2	367	20,991	18,206	0.24	0.27
	L	2	731	53,304	45,629	0.09	0.11
	H	2	731	94,128	81,891	0.03	0.06

Appendix B

APPENDIX FOR CHAPTER 4

B.1 Discrete Approximation to Coalescent Likelihood

Here we assume $N_e(t)$ is an unknown continuous function, so the integrals in equation (1) must be computed with numerical approximation techniques. We follow Palacios and Minin (2012), Gill et al. (2013), and Lan et al. (2015) and use discrete approximations of the integrals over a finite grid. We assume $N_e(t) = \exp\left[f(t)\right]$, where $f(t)$ is a function of continuous time. We approximate $f(t)$ by estimating it at discrete locations on a fixed grid with uniform spacing. We construct a regular grid, $\boldsymbol{x} = \{x_h\}_{h=1}^{H+1}$, and set the end points of the grid $\boldsymbol{x}$ such that $x_1 = 0$ and $x_{H+1} = t_1$ (see Figure 1 of main text). This results in H grid cells and $H+1$ cell boundaries. Now for $t \in (x_h, x_{h+1}]$, we have $N_e(t) \approx \exp[\theta_h]$, where θ_h is an unknown model parameter. This implies that $\boldsymbol{\theta} = \{\theta_h\}_{h=1}^{H}$ is a piecewise-constant approximation to $f(t) = \ln[N_e(t)]$ for $t \in [s_m, t_1]$.

Calculating the likelihood in equation (1) of the main text requires first sorting the combined set of time points $\{\boldsymbol{t}, \boldsymbol{s}, \boldsymbol{x}\}$ and creating a new set of $D = n+m+H-3$ half-open subintervals $\{I'_d\}_{d=1}^{D}$, such that for each $d = 1, \ldots, D$ there exists an i, k, and h that satisfy $I'_d = I_{i,k} \cap (x_h, x_{h+1}]$. Now the integrals in equation (1) can be approximated by

$$\int_{I_{i,k}} \frac{C_{i,k}}{N_e(t)} dt \approx \sum_{I'_d \subset I_{i,k}} \frac{C_{i,k}}{\exp[\theta_h]} \Delta_d, \tag{B.1}$$

where Δ_d is the length of the subinterval I'_d. If we introduce an auxiliary variable z_d that takes the value 1 if interval I_d ends with a coalescent event ($I'_d \subseteq I_{0,k}$) and 0 otherwise, then we can use equation (B.1) to write an approximation to the component of the density in equation (1) of the main text associated with interval $(x_h, x_{h+1}]$ as

$$p(\boldsymbol{z}_h \mid \boldsymbol{s}, \boldsymbol{n}, N_e(t)) = \prod_{I'_d \subset (x_h, x_{h+1}]} \left\{ \frac{C_{i,k}}{\exp[\theta_h]} \right\}^{z_d} \exp\left\{ -\frac{C_{i,k}}{\exp[\theta_h]} \Delta_d \right\}, \tag{B.2}$$

where $\boldsymbol{z}_h$ is the vector of z_d values such that $I'_d \subset (x_h, x_{h+1}]$. An approximation to the complete density in equation (1) is then the product of the components in equation (B.2):

$$p(t_1, \ldots, t_{n-1} \mid \boldsymbol{s}, \boldsymbol{n}, N_e(t)) \approx \prod_{h=1}^{H} p(\boldsymbol{z}_h \mid \boldsymbol{s}, \boldsymbol{n}, N_e(t)). \tag{B.3}$$

B.2 Setting the Global Smoothing Hyperparameter

The global smoothing parameter γ controls the variation in the estimated effective population size trajectory. It is therefore important to have a way to select the scale hyperparameter ζ of the prior distribution of the global smoothing parameter that reduces subjectivity. We follow a method suggested Sørbye and Rue (2014) for intrinsic GMRF models and modified by Faulkner and Minin (2018) for SPMRF models for selecting this hyperparameter. Let $\boldsymbol{Q}$ be the precision matrix for the Markov random field corresponding to the model of interest (see Faulkner and Minin (2018) for examples), and $\boldsymbol{\Sigma} = \boldsymbol{Q}^{-1}$ be the covariance matrix with diagonal elements Σ_{ii}. The marginal standard deviation of all components of $\boldsymbol{\theta}$ for a fixed value of γ is $\sigma_\gamma(\theta_i) = \gamma\sigma_{\text{ref}}(\boldsymbol{\theta})$, where $\sigma_{\text{ref}}(\boldsymbol{\theta})$ is the geometric mean of the individual marginal standard deviations when $\gamma = 1$ (Sørbye and Rue, 2014). We want to set an upper bound U on the average marginal standard deviation of θ_i, such that $\Pr(\sigma_\gamma(\theta_i) > U) = \alpha$, where α is some small probability (typically 0.01 to 0.05). Using the cumulative probability function for a half-Cauchy distribution, we can find a value of ζ for a given value of $\sigma_{\text{ref}}(\boldsymbol{\theta})$ specific to a model of interest and given common values of U and α by:

$$\zeta = \frac{U}{\sigma_{\text{ref}}(\boldsymbol{\theta}) \tan\left(\frac{\pi}{2}(1-\alpha)\right)}. \tag{B.4}$$

For phylodynamic inference, we set U equal the estimated standard deviation of the log-transformed values the Skyline estimates of population size (Pybus et al., 2000) based on a fixed genealogy and set of sample times. We choose this value of U since we know that the marginal variances of the θs should not exceed the variance in the log-Skyline estimates, on average. For the examples in this paper, we set $\alpha = 0.05$ as the probability of the average marginal standard deviation exceeding U.

B.3 Elliptical Slice within Gibbs Sampler

For models based on sequence data, we used a combination of elliptical slice sampling (Murray et al., 2010) for the latent effective population size parameters and Gibbs sampling for the latent local and global scale parameters. The Gibbs sampler was based on a modification of the approach derived by Makalic and Schmidt (2016) for Gibbs sampling of horseshoe random variables.

B.3.1 Model Specifications

HSMRF-1

Using a state-space representation of the HSMRF where μ is the fixed overall mean and σ^2 is a fixed variance for θ_1 and ζ is the fixed hyperparameter on the global scale, following Makalic and Schmidt (2016) the first-order HSMRF model conditional on a set of auxiliary variables can be written:

$$
\begin{aligned}
\boldsymbol{y} \mid \boldsymbol{\theta} &\sim \mathcal{L}(\boldsymbol{y} \mid \boldsymbol{\theta}) \\
\Delta\theta_j &\sim \mathcal{N}(0, \lambda_j^2 \eta^2 \zeta^2) \quad j = 1, \ldots, H-1 \\
\theta_1 &\sim \mathcal{N}(\mu, \sigma^2) \\
\theta_i &= \theta_1 + \sum_{j=1}^{i-1} \Delta\theta_j \quad i = 2, \ldots, H \\
\lambda_j^2 \mid \psi_j &\sim \mathcal{IG}(1/2, 1/\psi_j) \\
\eta^2 \mid \xi &\sim \mathcal{IG}(1/2, 1/\xi) \\
\psi_1, \ldots, \psi_{H-1}, \xi &\sim \mathcal{IG}(1/2, 1),
\end{aligned}
$$

where $\boldsymbol{y}$ is the coalescent data, $\mathcal{L}$ is the coalescent density, and $\mathcal{IG}$ is an inverse-gamma distribution. This formulation implies that $\lambda_j \sim C^+(0, 1)$ and $\eta \sim C^+(0, 1)$. We translate this to our original model formulation by allowing the global scale parameter $\gamma \sim C^+(0, \zeta)$, where $\gamma = \eta\zeta$, and the local scale parameters $\tau_j \sim C^+(0, \gamma)$, where $\tau_j = \lambda_j\gamma = \lambda_j\eta\zeta$. This implies that $\Delta\theta_j \sim \mathcal{N}(0, \tau_j^2)$, which is our original way of formulating the model.

HSMRF-2

The second-order HSMRF model conditional on a set of auxiliary variables can be written:

$$
\begin{aligned}
\boldsymbol{y} \mid \boldsymbol{\theta} &\sim \mathcal{L}(\boldsymbol{y} \mid \boldsymbol{\theta}) \\
\Delta\theta_1 &\sim \mathcal{N}(0, \frac{1}{2}\lambda_1^2\eta^2\zeta^2) \\
\Delta^2\theta_j &\sim \mathcal{N}(0, \lambda_j^2\eta^2\zeta^2) \quad j = 2, \ldots, H-1 \\
\theta_1 &\sim \mathcal{N}(\mu, \sigma^2) \\
\theta_2 &= \theta_1 + \Delta\theta_1 \\
\theta_j &= \Delta^2\theta_{j-1} + 2\theta_{j-1} - \theta_{j-2} \quad j = 3, \ldots, H \\
\lambda_j^2 \mid \psi_j &\sim \mathcal{IG}(1/2, 1/\psi_j) \quad j = 1, \ldots, H-1 \\
\eta^2 \mid \xi &\sim \mathcal{IG}(1/2, 1/\xi) \\
\psi_1, \ldots, \psi_{H-1}, \xi &\sim \mathcal{IG}(1/2, 1),
\end{aligned}
$$

where $\Delta\theta_1 = \theta_2 - \theta_1$, and $\Delta^2\theta_j = \theta_{j+1} - 2\theta_j + \theta_{j-1}$ for $j = 2, \ldots, H-1$.

GMRF-1

Similar to the HSMRF-1 model above but absent the local scale parameters, the first-order GMRF model can be written:

$$
\begin{aligned}
\boldsymbol{y} \mid \boldsymbol{\theta} &\sim \mathcal{L}(\boldsymbol{y} \mid \boldsymbol{\theta}) \\
\Delta\theta_j &\sim \mathcal{N}(0, \eta^2\zeta^2) \quad j = 1, \ldots, H-1 \\
\theta_1 &\sim \mathcal{N}(\mu, \sigma^2) \\
\theta_i &= \theta_1 + \sum_{j=1}^{i-1} \Delta\theta_j \quad i = 2, \ldots, H \\
\eta^2 \mid \xi &\sim \mathcal{IG}(1/2, 1/\xi) \\
\xi &\sim \mathcal{IG}(1/2, 1).
\end{aligned}
$$

GMRF-2

The second-order GMRF model can be written:

$$
\begin{aligned}
\boldsymbol{y} \mid \boldsymbol{\theta} &\sim \mathcal{L}(\boldsymbol{y} \mid \boldsymbol{\theta}) \\
\Delta\theta_1 &\sim \mathcal{N}(0, \frac{1}{2}\eta^2\zeta^2) \\
\Delta^2\theta_j &\sim \mathcal{N}(0, \eta^2\zeta^2) \quad j = 2, \ldots, H-1 \\
\theta_1 &\sim \mathcal{N}(\mu, \sigma^2) \\
\theta_2 &= \theta_1 + \Delta\theta_1 \\
\theta_j &= \Delta^2\theta_{j-1} + 2\theta_{j-1} - \theta_{j-2} \quad j = 3, \ldots, H \\
\eta^2 \mid \xi &\sim \mathcal{IG}(1/2, 1/\xi) \\
\xi &\sim \mathcal{IG}(1/2, 1).
\end{aligned}
$$

where $\Delta\theta_1 = \theta_2 - \theta_1$, and $\Delta^2\theta_j = \theta_{j+1} - 2\theta_j + \theta_{j-1}$ for $j = 2, \ldots, H-1$.

B.3.2 Full Conditional Distributions

HSMRF-1

First we describe the full conditional distributions of the latent scale and auxiliary variables used in the Gibbs sampler for the first-order HSMRF. It can be shown that for $j = 1, \ldots, H-1$, the full conditional distributions are:

$$
\begin{aligned}
p(\lambda_j^2 \mid \cdot) &\propto \mathcal{IG}\left(1, \ \frac{1}{\psi_j} + \frac{\Delta\theta_j^2}{2\eta^2\zeta^2}\right) \\
p(\eta^2 \mid \cdot) &\propto \mathcal{IG}\left(\frac{H}{2}, \ \frac{1}{\xi} + \frac{1}{2\zeta^2}\sum_{j=1}^{H-1}\frac{\Delta\theta_j^2}{\lambda_j^2}\right) \\
p(\psi_j \mid \cdot) &\propto \mathcal{IG}\left(1, \ 1 + \frac{1}{\lambda_j^2}\right) \\
p(\xi \mid \cdot) &\propto \mathcal{IG}\left(1, \ 1 + \frac{1}{\eta^2}\right)
\end{aligned}
$$

HSMRF-2

The full conditional distributions of the latent scale and auxiliary variables for the second-order HSMRF are:

$$
\begin{aligned}
p(\lambda_1^2 \mid \cdot) &\propto \mathcal{IG}\left(1,\ \frac{1}{\psi_1} + \frac{\Delta\theta_1^2}{\eta^2\zeta^2}\right) \\
p(\lambda_j^2 \mid \cdot) &\propto \mathcal{IG}\left(1,\ \frac{1}{\psi_j} + \frac{\Delta^2\theta_j^2}{2\eta^2\zeta^2}\right) \quad j = 2, \ldots, H-1 \\
p(\eta^2 \mid \cdot) &\propto \mathcal{IG}\left(\frac{H}{2},\ \frac{1}{\xi} + \frac{\Delta\theta_1^2}{\lambda_1^2\zeta^2} + \frac{1}{2\zeta^2}\sum_{j=2}^{H-1}\frac{\Delta^2\theta_j^2}{\lambda_j^2}\right) \\
p(\psi_j \mid \cdot) &\propto \mathcal{IG}\left(1,\ 1 + \frac{1}{\lambda_j^2}\right) \\
p(\xi \mid \cdot) &\propto \mathcal{IG}\left(1,\ 1 + \frac{1}{\eta^2}\right)
\end{aligned}
$$

GMRF-1

Similarly, the full conditional distributions for the scale and auxiliary variables for the first-order GMRF are:

$$
\begin{aligned}
p(\eta^2 \mid \cdot) &\propto \mathcal{IG}\left(\frac{H}{2},\ \frac{1}{\xi} + \frac{1}{2\zeta^2}\sum_{j=1}^{H-1}\Delta\theta_j^2\right) \\
p(\xi \mid \cdot) &\propto \mathcal{IG}\left(1,\ 1 + \frac{1}{\eta^2}\right)
\end{aligned}
$$

GMRF-2

The full conditional distributions for the scale and auxiliary variables for the second-order GMRF are:

$$
\begin{aligned}
p(\eta^2 \mid \cdot) &\propto \mathcal{IG}\left(\frac{H}{2},\ \frac{1}{\xi} + \frac{\Delta\theta_1^2}{\zeta^2} + \frac{1}{2\zeta^2}\sum_{j=2}^{H-1}\Delta^2\theta_j^2\right) \\
p(\xi \mid \cdot) &\propto \mathcal{IG}\left(1,\ 1 + \frac{1}{\eta^2}\right)
\end{aligned}
$$

B.3.3 Elliptical Slice and Gibbs Sampling

We follow the algorithm in Figure 2 (pg 543) of Murray et al. (2010) for elliptical slice sampling, but with a few modifications. Suppose the elements of our observation variable $\boldsymbol{y}$ are conditionally independent given a function of underlying latent Gaussian variables $\boldsymbol{\theta} = \boldsymbol{f} + \mu$, where $\boldsymbol{f} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ and μ is a fixed constant. We denote the likelihood of $\boldsymbol{y}$ conditional on $\boldsymbol{\theta}$ as $\mathcal{L}(\boldsymbol{y} \mid \boldsymbol{\theta})$. Following Murray et al. (2010), let $\boldsymbol{f}$ be the current state of the zero-centered field parameters on the natural log scale. The algorithm proceeds by first selecting $\boldsymbol{\nu} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$ and drawing $u \sim \mathcal{U}(0, 1)$. We set the slice value $s = \ln u + \ln \mathcal{L}(\boldsymbol{y} \mid \boldsymbol{f} + \mu)$. We then draw a proposed angle $\alpha \sim \mathcal{U}(0, 2\pi)$, and define a bracket $[\alpha_{min}, \alpha_{max}] = [\alpha - 2\pi, \alpha]$. The current proposal is $\boldsymbol{f}' = \boldsymbol{f} \cos\alpha + \boldsymbol{\nu} \sin\alpha$. If $\ln \mathcal{L}(\boldsymbol{y} \mid \boldsymbol{f}' + \mu) > s$, then we accept and set $\boldsymbol{f} = \boldsymbol{f}'$. Otherwise, we shrink the bracket by setting $\alpha_{min} = \alpha$ if $\alpha < 0$ or setting $\alpha_{max} = \alpha$ if $\alpha \geq 0$, and draw a new $\alpha \sim \mathcal{U}(\alpha_{min}, \alpha_{max})$. We then calculate a new proposal and keep shrinking the bracket in this manner until the proposal is accepted.

One modification we make to this process is in drawing the initial $\boldsymbol{f}$ and subsequent $\boldsymbol{\nu}$ vectors. Instead of using the multivariate normal specification, we use the state-space formulation. To do this, we first draw $\nu_1 \sim \mathcal{N}(0, \sigma^2)$ and then draw $\Delta\nu_j \sim \mathcal{N}(0, \lambda_j^2 \eta^2 \zeta^2)$ for $j = 1, \ldots, n-1$ and calculate $\nu_i = \nu_1 + \sum_{j=1}^{i-1} \Delta\nu_j$ for $i = 2, \ldots, n$. Then, prior to evaluating the likelihood, we need to calculate $\boldsymbol{\theta} = \boldsymbol{f} + \mu$. The likelihood is then $\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{y})$. This approach allows us to sample the variables as multivariate normal with mean zero without needing to use the multivariate normal distribution and costly computations that come with it.

We can specify μ and σ^2 using the natural log of the maximum likelihood estimates (MLE) of $N_e(t)$ on a grid, where the coalescent times are obtained from a fixed maximum clade credibility tree, where μ is the log of the mean of the MLE estimates and σ^2 is 4 times their variance. These should provide reasonable hyperparameters that will not result in too diffuse of a sampling distribution.

We use elliptical slice sampling to sample from the field parameters $\boldsymbol{\theta}$ conditional on the latent scale parameters and use Gibbs sampling to update the latent scale parameters conditional on the

field and other parameters. We alternate between these updates until convergence and the desired number of posterior samples are obtained.

B.3.4 Checking Validity of Algorithms

We performed two checks of our implementation of the random field models in `RevBayes`. We simulated coalescent times from the four trajectories that were used in our simulations and generated genealogical trees from those times. Our first-pass check of our elliptical-slice-within-Gibbs sampler in `RevBayes` was to feed these trees directly into `RevBayes` as fixed (as in Section 3.1 of the main text) and compare the results to those obtained with our `spmrf` package using Hamiltonion Monte Carlo (HMC). Trace plots for a few parameters from the RevBayes implementation indicate decent mixing (Figure S1), and plots of trends from RevBayes implementations do not show appreciable differences from those estimated using HMC with the spmrf package (Figure S2).

We then tested our joint inference procedure in `RevBayes`, estimating the tree topology, coalescent times, and coalescent trajectory. To reduce computation times, we first down-sampled each tree to 100 tips, ensuring that the retained tips spanned the entire range of non-contemporaneous tips. We then simulated alignments of 500 sites using mutation rates that produced alignments with an expectation of ≈ 0.93 substitutions per site. Thus the simulated alignments were approximately the size of the empirical alignments and contained approximately the same amount of information (number of substitutions). For simplicity, we employed the Jukes-Cantor substitution model (with no free parameters) with no rate heterogeneity across sites. When performing the full joint analyses on these datasets, we assumed the clock rate was known (as with the Bison analysis). All tests indicated that our MCMC sampler was working correctly. The code we used for conducting these tests is available at `https://github.com/jrfaulkner/phylocode`

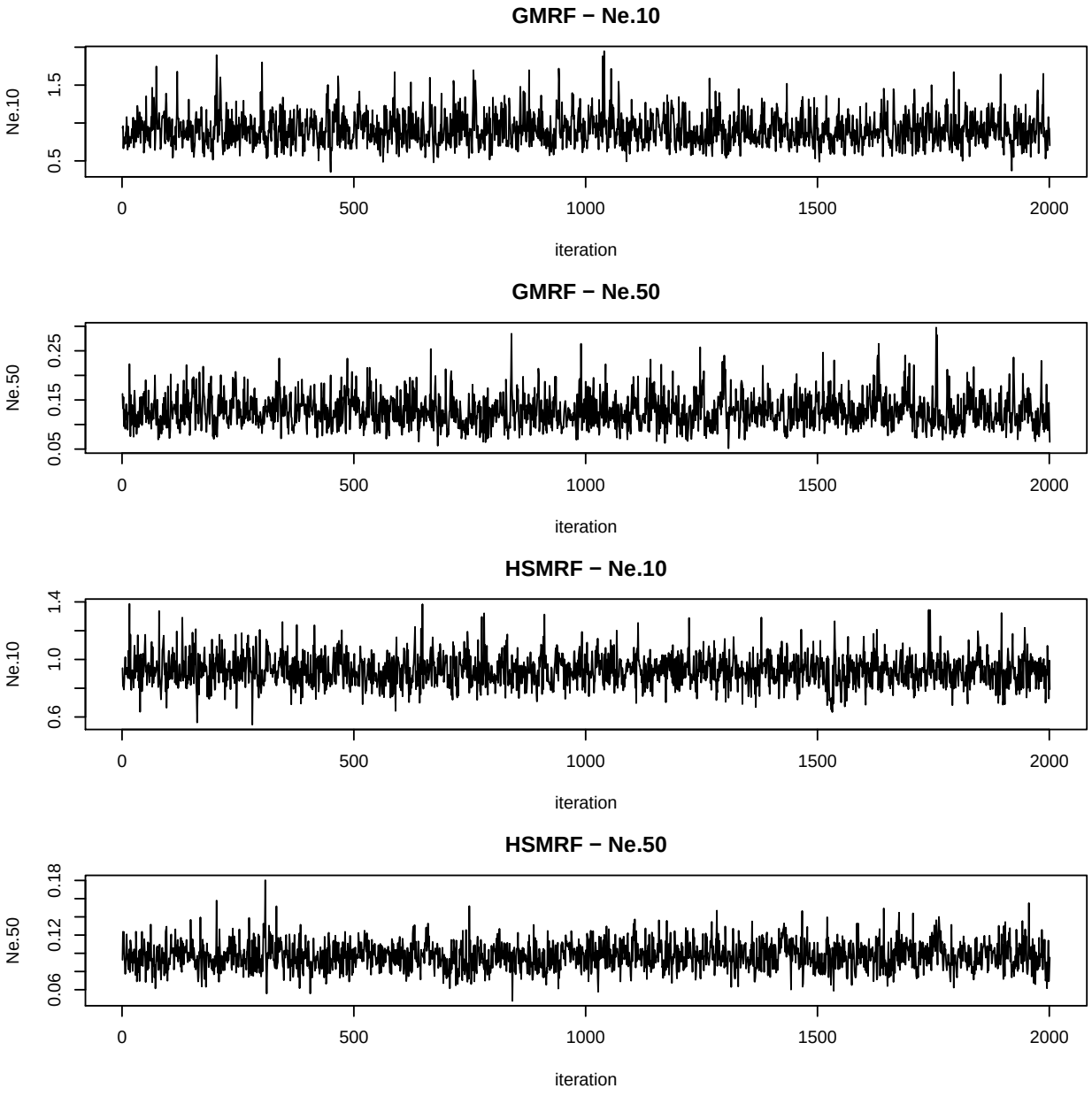

Figure B.1: Trace plots for posterior samples from two N_e parameters from models fit using our elliptical-slice-within-Gibbs sampler in `RevBayes`. Examples are for fixed tree coalescent data generated from the Bottleneck scenario used in the main simulations.

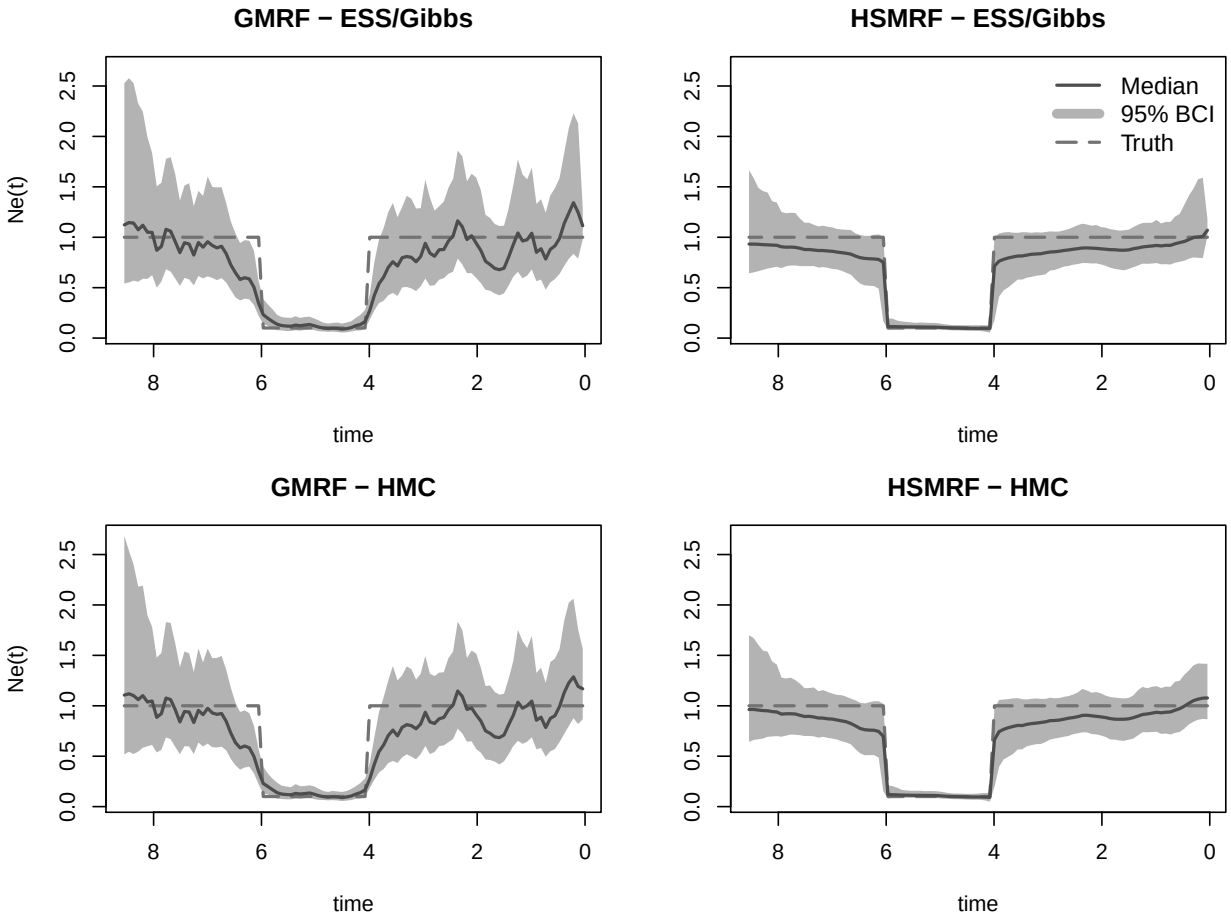

Figure B.2: Posterior medians and 95% credible intervals for N_e trajectories for two different MCMC samplers. The top row shows results from the elliptical-slice-within-Gibbs sampler in `RevBayes`, and the bottom shows results from the HMC sampler in `Stan` interfaced from the `spmrf` package. Examples are for fixed tree coalescent data generated from the Bottleneck scenario used in the main simulations.

B.4 Simulation Details

We used simulated data to assess the performance of the HSMRF model relative to the GMRF model. We investigated four scenarios with different trajectories for $N_e(t)$: (1) Bottleneck (BN), (2) Boom-Bust (BB), (3) Broken Exponential (BE), and (4) Nonstationary Gaussian Process (NGP). The BN scenario had true $N_e(t) = 0.1$ for $4 \leq t \leq 6$ and $N_e(t) = 1.0$ elsewhere. The BB scenario had $N_e(t) = 0.4 + 0.25[\sin((5.5 - t)/3) + 0.75\exp(-2.5(t - 5)^2)]$. The BE trajectory was $N_e(t) = \exp(-1.20 + 0.09t)$ for $0 \leq t < 4.5$, $N_e(t) = \exp(9.09 - 2.20t)$ for $4.5 \leq t < 5$, and $N_e(t) = \exp(-3.57 + 0.33t)$ for $t \geq 5$. The trajectory for the NGP scenario, was generated from a Gaussian process with mean 0.55 and a nonstationary Matérn covariance function (Paciorek and Schervish, 2006). The covariance function was constructed so that the length scale increased rapidly in the center of the domain, resulting in a smoother $N_e(t)$ trajectory in the center. The trajectories used for each scenario are shown at the top of Figure 2 of the main text. These effective population sizes were set to be small so the the coalescent times would be quick and would mostly fall within a time window specified for each scenario (see below).

For each scenario we generated 100 simulated data sets by first generating a random set of sampling times over a fixed interval and then generating a corresponding random set of coalescent times using the thinning algorithm proposed by Palacios and Minin (2013) and the true deterministic $N_e(t)$ trajectories defined for each scenario. For each simulated data set, this is equivalent to assuming we know the fixed genealogical tree for a sample of DNA sequences. We found that 100 simulations per scenario was sufficient to identify meaningful differences between models without excessive computation time. We used heterochronous sampling and set the sample sizes based on the complexity of each scenario. The sample sizes were $n = 500$, 2,000, 1,000, and 2,000 and the number of lineages sampled at time zero were $n_m = 50$, 50, 100, and 200 for the BN, BB, BE, and NGP scenarios, respectively. The remaining sample times followed a uniform distribution on $[0, S]$, where the values of S were 8.0, 11.8, 7.8, and 11.8 for the BN, BB, BE, and NGP scenarios, respectively. We used a fixed grid of 100 cells where the boundary of the 99th cell was T and the final cell collected any coalescent times greater than T. The values of T were 8.37, 11.73, 7.86, and

11.84 for the BN, BB, BE, and NGP scenarios, respectively. These values were chosen such that the final grid cell contained at least one coalescent time for all of the simulated data sets in a scenario. For a single fixed tree analysis, we would typically use the final coalescent time as the end of the grid. However, for the simulations we wanted to keep the grid cells of uniform width across the data sets in a scenario for comparability of results, so we used the same fixed grid for each data set in a scenario. We chose to use 100 grid cells because that allowed for sufficient resolution to capture features in the underlying trends, and more cells would have increased computation times.

We used HMC to approximate the posterior distribution of model parameters. For each simulated data set we ran four independent chains, where each chain had 1,000 iterations of adaptation followed by 500 sampling iterations. This resulted in a total of 2,000 posterior samples. The hyperparameter on the global scale parameter was selected using the method described in Web Appendix B based on the order of the model and the observations from a single data set generated for a scenario.

B.5 Guidelines for Constructing Grids

The length of a grid and the number of cells will affect resolution of the estimated effective population size trajectory and its uncertainty, and will also affect computation times. Here we set some general guidelines that will help with setting up a grid.

The number of grid cells will determine the resolution of detail in the estimated effective population size trajectory. There have to be enough cells to capture important features in the trajectory, but there should also be enough data to support the number of cells. As a general rule, we suggest selecting the total number of grid cells, H, such that $H = \min\{0.8(n-1), 500\}$, where n is the number of sequences in the data set ($n-1$ is the number of coalescent times). This is not a hard and fast rule, but estimates of effective population size tend to behave better when there is at least one coalescent event in a cell or in an adjacent cell. The upper bound of 500 is arbitrary, but the resolution provided by grid densities greater than 500 is typically not worth the cost of additional computing time. We should point out that we broke this rule with the HCV example (100 grid cells for 62 coalescent times) because the majority of the coalescent times occurred within the first

50% of the time domain and we wanted better grid resolution in that first half of the grid to capture details of the population trend. However, we did follow the rule for the bison example which had 151 coalescent times and we used 120 grid cells, which is approximately 0.8 times 151.

Another important factor in setting up a grid is where to place the boundaries on the final cell. When performing analyses using coalescent times from a single fixed genealogical tree, the end of the grid is typically set equal to the final coalescent time and all grid cells are equally spaced between time zero and the end of the grid. However, when analyzing sequence data, the genealogical tree is being estimated, which results in a different set of coalescent times for each MCMC iteration. Since the width of the grid cells will affect the value of the local and global scale parameters of the random fields used to estimate the effective population sizes, it is necessary to fix the location of the grid boundaries across MCMC iterations. The location of the boundary T between grid cells $H-1$ and H is important because the uncertainty in the effective population size will be inflated in the final cell if it rarely contains coalescent times across MCMC iterations. Gill et al. (2016) suggested setting T for the Skygrid model so that the final grid cell spans the interval between T and infinity but the remaining grid cells are equally spaced. Note that the final cell is effectively only as wide as the oldest coalescent time, so the difference in width compared to the other cells should not be that great. We expand on this idea by formulating a general probability rule for setting T based on data. We want to find T such that

$$\Pr(\text{TMRCA} > T) = 1 - \alpha_T, \tag{B.5}$$

where TMRCA is the time to most recent common ancestor, and also the last coalescent time. We want α_T to be small, so that there is a high probability that the final grid cell contains the TMRCA in each MCMC iteration.

If a posterior distribution for the TMRCA is available from a previous analysis of the data, then one could use that to find T based on a given α_T. Alternatively, if point estimates and measures of uncertainty for the TMRCA have been published for the data of interest from another study, then those could be used to derive a value for T by a assuming a distribution on the TMRCA, such as a log-normal distribution (see examples below). In the absence of published estimates,

one could use quick frequentist methods such as those proposed by To et al. (2015) or Sagulenko et al. (2018) to get initial estimates of the TMRCA based on the sequence data without needing to perform a complete Bayesian analysis. Note that in some instances after the initial grid setup it may be necessary to adjust the grid after running models and assessing the results to make sure the final cells are adequately capturing the final coalescent times.

For the HCV example, we used the estimated TMRCA of 283 years and associated 95% credible interval of 246 to 320 years reported by Pybus et al. (2003). Assuming a log-normal distribution for TMRCA, we estimated the standard deviation of the distribution on the log scale to be $\hat{\sigma} = (\ln(283) - \ln(246))/1.96 = 0.071$ and the mean on the log scale to be $\hat{\mu} = \ln(283) = 5.64$. Setting $\alpha_T = 0.001$, the quantile of a log-normal $\mathcal{LN}(\hat{\mu}, \hat{\sigma}^2)$ distribution that satisfies equation (B.5) is $T = 227$ years. For the bison example, we used the estimated TMRCA of 136 kya and 95% credible interval of 111 to 164 kya reported by Shapiro et al. (2004). Following the same procedure as with the HCV example and setting $\alpha_T = 0.001$, the quantile of a $\mathcal{LN}(\hat{\mu}, \hat{\sigma}^2)$ distribution that satisfies equation (B.5) is $T = 98.7$ kya. We rounded up and set $T = 100$ kya.

B.6 Implementation Details for Data Examples

For each of the models run in `RevBayes` for the HCV example, we ran four chains each with 1 million iterations of burn-in followed by 25 million iterations of sampling thinned at intervals of 20,000 iterations. For the SkyLine, SkyRide, and SkyGrid models, each had 1 million burn-in followed by 50 million thinned at every 10,000 iterations. This resulted in 5,000 posterior samples for each model. For each of the models run in `RevBayes` for the bison example, we ran four chains each with 1 million iterations of burn-in followed by 80 million iterations of sampling thinned at intervals of 64,000 iterations. Each of the models run in `BEAST` for the bison example had 1 million iterations of burn-in. The SkyLine, SkyRide, and SkyGrid had 250 million, 100 million, and 100 million iterations of sampling, respectively, each were thinned at 50,000, 20,000, and 20,000, respectively. These settings for the models in the bison example resulted in a total of 5,000 posterior samples per model.

Note that where possible, we attempted to maintain the same relative weighting of different

MCMC moves in common to the models across the two software packages. This corresponds to a somewhat non-standard MCMC scheme in RevBayes where one iteration is equal to one generation. In RevBayes, typically move weights are generally chosen such that each MCMC iteration contains a number of generations.

All model runs were performed on a cluster running the Centos 7 Linux operating system with Intel® Xeon® X7550 2.0GHz 64-bit processors and 15.3 GB of RAM. We used `BEAST` version 1.10.4 with the `BEAGLE` library activated. We used `RevBayes` version 1.0.11, which includes the modifications made to implement the coalescent models described in the main text.

We calculated the total run times for each of the models by including the time for burn-in and the total time run for the sampling iterations. For the examples run with multiple chains, we calculated the total time by summing the sampling times of each chain and then adding the average burn-in time across chains. This combined time is an estimate of time it would have taken to run as single chain with a single burn-in period that would result in the same total number of samples from the combined chains. The combined times from `RevBayes` are reported with the times from the single long chains run in `BEAST` (Appendix Table B.1). The run times and ESS/hr differed substantially between the models run in `BEAST` and the models run in `RevBayes`, but the run times were almost identical for models of the same order run with `RevBayes`. `RevBayes` was constructed to be a flexible modeling platform, but some of the programming approaches that allowed that flexibility have resulted in long run times for some models. This problem is known and the `RevBayes` developers are working to address it. With the differences between `BEAST` and `RevBayes` aside, it is reassuring that the run times for the HSMRF models are not practically different from those of the GMRF models. Sampling the additional number of parameters in the HSMRF does not add much if any computation time because the Gibbs samplers and elliptical slice samplers are opterated on vectors of parameters. Most of the compution time in `RevBayes` appears to be dominated by sampling the genealogical trees and evolution parameters. If run times cannot be improved in `RevBayes`, we will look to implement our HSMRF models in `BEAST`. Either way, whether implemented in `BEAST` or in an optimzed version of `RevBayes`, we expect that the difference in run times between the GMRF and HSMRF models will remain minimal.

Table B.1: Run times (hrs) , mean effective sample sizes (ESS) for the log effective population size parameters, and mean number of effective samples per hour (ESS/hr) for the HCV and bison data examples.

Example	Software	Model	Time	ESS	ES/hr
HCV	BEAST	SkyLine	2.7	1,853.2	681.3
		SkyRide	3.4	4,469.2	1,308.7
		SkyGrid	3.6	2,441.1	668.8
	RevBayes	GMRF-1	522.9	1,307.5	2.5
		HSMRF-1	524.1	1,073.5	2.0
		GMRF-2	521.3	468.6	0.9
		HSMRF-2	521.0	457.2	0.9
Bison	BEAST	SkyLine	16.8	1,596.9	94.9
		SkyRide	9.1	3,983.5	436.3
		SkyGrid	9.4	3,398.8	353.3
	RevBayes	GMRF-1	339.0	2,833.5	8.4
		HSMRF-1	341.5	2,517.4	7.4
		GMRF-2	338.0	2,282.3	6.8
		HSMRF-2	337.9	2,758.1	8.2

B.7 Calculating Posterior Model Probabilities

Calculation of posterior model probabilities requires estimates of marginal (or integrated) likelihood values. For both the HCV and bison data examples, we used steppingstone sampling (Xie et al., 2011) to estimate marginal likelihoods. Steppingstone sampling is a type of path sampling algorithm (Gelman and Meng, 1998) which is similar to the thermodynamic integration methods developed by Lartillot and Philippe (2006) and Friel and Pettitt (2008) but incorporates importance sampling to improve computational efficiency.

The stepping stone and thermodynamic methods evaluate the posterior where the likelihood is raised to a power over a sequence of power values between 0.0 and 1.0, which spans the set of densities between the posterior and the prior. For each model in each data example, we used 50 stones, where the stones represent a set of quantiles of a Beta(α, 1.0) distribution which are used as the sequence of power values. We set $\alpha = 0.2$ in the beta distribution. We used the same total number of iterations used in the main analyses but spread equally among the stones.

Once we have marginal likelihood estimates, we can calculate Bayes factors (Kass and Raftery, 1995) and posterior model probabilities to compare evidence for different models. The posterior odds of Model 1 ($\mathcal{M}_1$) relative to Model 2 ($\mathcal{M}_2$) conditional on the data ($\mathcal{D}$) is calculated as

$$\frac{\Pr(\mathcal{M}_1 \mid \mathcal{D})}{\Pr(\mathcal{M}_2 \mid \mathcal{D})} = \frac{\Pr(\mathcal{D} \mid \mathcal{M}_1)}{\Pr(\mathcal{D} \mid \mathcal{M}_2)} \frac{\Pr(\mathcal{M}_1)}{\Pr(\mathcal{M}_2)},$$

where $\Pr(\mathcal{D} \mid \cdot)$ is the marginal likelihood of the data given a particular model, and the Bayes factor is the ratio of marginal likelihoods: $B_{12} = \Pr(\mathcal{D} \mid \mathcal{M}_1)/\Pr(\mathcal{D} \mid \mathcal{M}_2)$. For our set of six models $\mathcal{M}_1, \ldots, \mathcal{M}_6$, we calculated the posterior probability of $\mathcal{M}_k$ as

$$\Pr(\mathcal{M}_k \mid \mathcal{D}) = \alpha_k B_{k1} / \sum_{r=1}^{6} \alpha_r B_{r1}$$

where $\alpha_k = \Pr(\mathcal{M}_k)/\Pr(\mathcal{M}_1)$ is the prior odds of $\mathcal{M}_k$ relative to $\mathcal{M}_1$ (HSMRF-1), and $B_{11}, \ldots, B_{61}$ are the Bayes factors calculated relative to $\mathcal{M}_1$. We assumed equal prior model probabilities, so $B_{11} = \alpha_1 = 1$. Appendix Table B.2 shows results for marginal likelihoods, Bayes factors, and posterior model probabilities from the HCV and bison data examples used in the main text.

B.8 Additional Results for HCV Example

In the main text for the HCV example, we focused on the period of rapid increase in the effective population size trajectory that generally occured after the year 1900. Here we report results for the entire time domain. The heatmaps of the posterior frequencies of coalescent event times show that coalescent events were very unlikely between approximately 1770 and 1870 (Appendix Figure B.3) . The GMRF and HSMRF models had similar levels of uncertainty in the population size trajectories during that gap in coalescent events, but the posterior distributions for effective population

Table B.2: Summary of model selection results for data examples. Shown are natural log of marginal likelihood values (logML), natural log of Bayes factors (logBF), Bayes factors (BF), and posterior model probabilities (Pr(M|D)) by model for the HCV and bison data examples. Bayes factors and posterior model probabilities are calculated relative to the HSMRF-1 model.

Example	**Model**	**logML**	**logBF**	**BF**	**Pr(M|D)**
HCV	SkyLine	-6,396.02	0.23	1.2644	0.5188
	SkyRide	-6,424.75	-28.49	0.0000	0.0000
	GMRF-1	-6,398.54	-2.29	0.1017	0.0417
	HSMRF-1	-63,96.26	0.00	1.0000	0.4103
	GMRF-2	-6,402.83	-6.57	0.0014	0.0006
	HSMRF-2	-6,398.93	-2.67	0.0695	0.0285
Bison	SkyLine	-3,717.53	-3.39	0.0336	0.0322
	SkyRide	-3,731.68	-17.54	0.0000	0.0000
	GMRF-1	-3,720.27	-6.13	0.0022	0.0021
	HSMRF-1	-3,714.14	0.00	1.0000	0.9611
	GMRF-2	-3,721.26	-7.12	0.0008	0.0008
	HSMRF-2	-3,719.66	-5.52	0.0040	0.0038

sizes were more skewed for the HSMRF models than for the GMRF models (Appendix Figure 3). The result of this is that the 95% credible intervals were wider for the HSMRF models compared to the GMRF models, but the corresponding 90% and 80% credible intervals were narrower for the HSMRF models compared their GMRF counterparts. Both the SkyLine and SkyRide models had narrow credible intervals during the gap in coalescent times, with the SkyRide model showing more uncertainty than the SkyLine. Both of these models use piecewise constant trajectories between coalescent events, which means they are restricted to be nearly flat over the long period without coalescent events. In contrast, the GMRF and HSMRF models had a large number of grid cells covering the gap in coalescent events and the effective population size was allowed to be dif-

ferent in each grid cell. The uncertainty in effective population sizes during the gap in coalescent times was likely grossly underestimated for the SkyLine and SkyRide due to the constrained nature of those models.

The GMRF and HSMRF models show a small change the population size trajectory and associated credible intervals in the final grid cell. This is because the final grid cell extends from year 1766 to infinity and likely capture all of the final four coalescent events for many of the posterior draws. The change from zero to four coalescent events in the final grid cell is the likely explanation for the abrupt change in estimated effective population size seen on the plots.

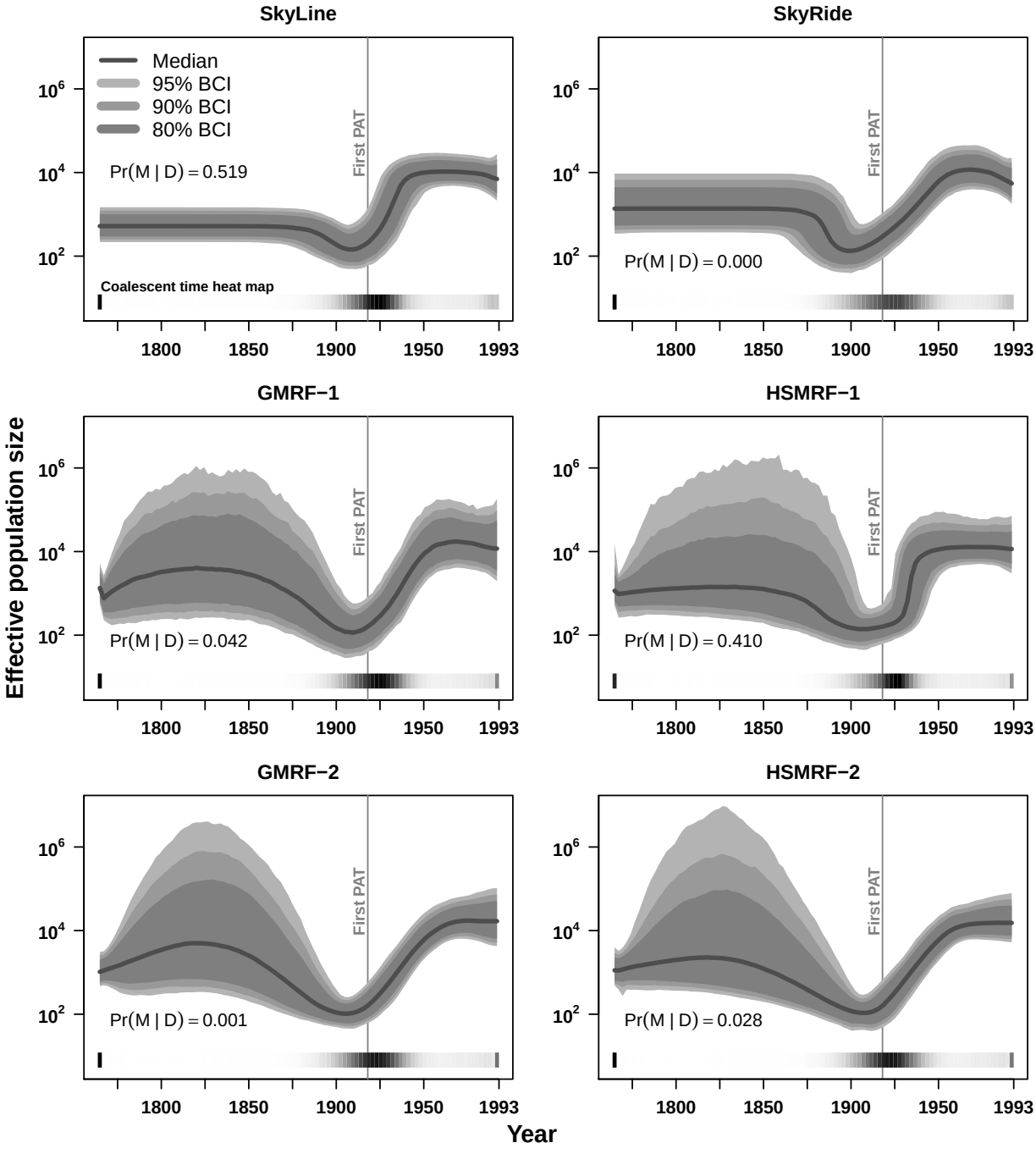

Figure B.3: Posterior medians (solid black lines) of effective population sizes and associated 95%, 90%, and 80% credible intervals (layered grey shaded areas) for the HCV data for the complete time domain for the Bayesian Skyline (SkyLine), Bayesian Skyride (SkyRide), Gaussian Markov random field of order 1 (GMRF-1) and order 2 (GMRF-2), and horseshoe Markov random field of order 1 (HSMRF-1) and order 2 (HSMRF-2). Also shown for each model are posterior model probabilities (Pr(M | D)) and heat maps of mean posterior frequencies of coalescent times. A vertical reference line is shown at year 1918, which is the year PAT was introduced.

Appendix C

APPENDIX FOR CHAPTER 5

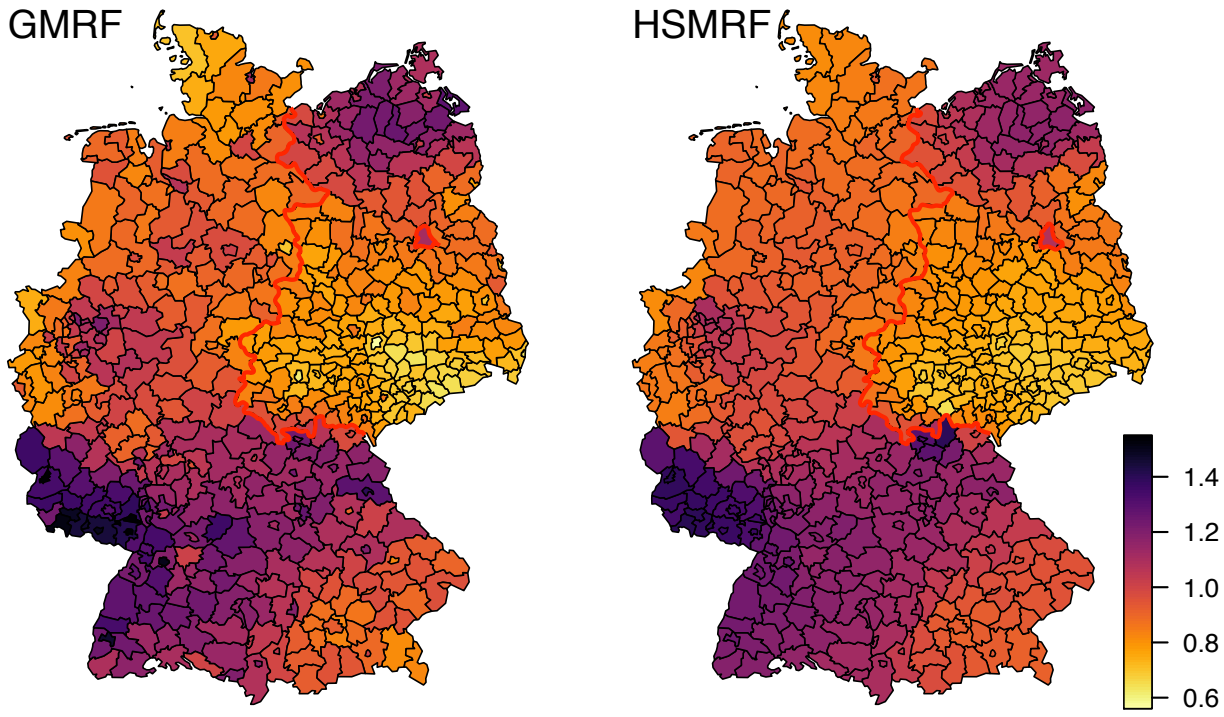

Figure C.1: Posterior median relative risk for the GMRF and HSMRF with old boundary between East and West Germany (in red).