BAYESIAN NONPARAMETRIC MODELING USING LÉVY PROCESS PRIORS WITH APPLICATIONS FOR FUNCTION ESTIMATION, TIME SERIES MODELING AND SPATIO-TEMPORAL MODELING

by

Chong Tu

Institute of Statistics and Decision Sciences
Duke University

Date: ____________________________
Approved:

____
Dr. Robert L. Wolpert, Supervisor

____
Dr. Merlise Clyde, Supervisor

____
Dr. David Banks

____
Dr. Prasad Kasibhatla

Dissertation submitted in partial fulfillment of the
requirements for the degree of Doctor of Philosophy
in the Institute of Statistics and Decision Sciences
in the Graduate School of
Duke University

2006

ABSTRACT

(Statistics)

BAYESIAN NONPARAMETRIC MODELING USING LÉVY PROCESS PRIORS WITH APPLICATIONS FOR FUNCTION ESTIMATION, TIME SERIES MODELING AND SPATIO-TEMPORAL MODELING

by

Chong Tu

Institute of Statistics and Decision Sciences
Duke University

Date: ______________________________
Approved:

____
Dr. Robert L. Wolpert, Supervisor

____
Dr. Merlise Clyde, Supervisor

____
Dr. David Banks

____
Dr. Prasad Kasibhatla

An abstract of a dissertation submitted in partial
fulfillment of the requirements for the degree
of Doctor of Philosophy in the
Institute of Statistics and Decision Sciences in the Graduate School of
Duke University

2006

Abstract

In this dissertation, we propose a new class of Bayesian method for nonparametric function estimation. We denote the new model as *Lévy adaptive regression kernel* or “LARK”. The LARK model is based on a stochastic expansion of functions in an overcomplete dictionary, which can be formulated as a stochastic integration problem with a random measure.

The unknown function is represented as a weighted sum of kernel or generator functions with arbitrary location parameters. Scaling parameters of the kernels are also taken as location specific and thus are adaptive, as with wavelets bases and dictionaries. Lévy random fields are introduced to construct prior distributions on the unknown functions, which lead to the specification of a joint prior distribution for the number of kernels, kernel regression coefficients and kernel associated parameters. Under Gaussian errors, the problem may be formulated as a sparse regression problem, with regularization induced through the Lévy random field prior. To make posterior inference on the unknown functions, a reversible jump MCMC algorithm is developed.

The LARK framework developed in this dissertation can be used to model both Gaussian and nonstationary Non-Gaussian data. The adaptability of the kernels is especially useful for modeling spatially inhomogeneous functions. Unlike many wavelet based methods, there is no requirement that the data are equally spaced. The RJ-MCMC algorithm developed for fitting the LARK model provides an automatic search mechanism for finding sparse representations of a function. Fitting a LARK model does not involve matrix calculation, thus the model is amenable to large data set.

We start with reviews on some basic properties and theories of Lévy processes in Chapter 1, which serve as the theoretical foundations for this dissertation. Chapter 2 develops LARK model in the context of nonparametric regression problems. Both simulated and real examples are used to illustrate the method. Chapter 3 applies LARK model for multivariate air pollutant time series modeling. Based on LARK framework, we develop a new class of spatio-temporal models in Chapter 4. A simulated data set and SO_2 monitoring data from the Environmental Protection Agency are used to demonstrate the model. We conclude the dissertation in Chapter 5 by summarizing the LARK framework and pointing out directions for future research.

Contents

List of Tables

List of Figures

Acknowledgements

I would like to take this opportunity to acknowledge with gratitude those people at ISDS who have given me help and support over the past five years. First of all, I deeply thank my advisers Dr. Robert Wolpert and Dr. Merlise Clyde for their encouragement and guidance throughout the development of this research. Their broad knowledge and enthusiasm for sciences have always been an inspiration for me. I am very thankful for having them along on the journey.

I also wish to thank my committee members, Dr. David Banks Dr. Prasad Kasibhatla for their interesting and illuminating comments. I sincerely thank Dr. Edwin Iversen for his invaluable guidance during my first two years at ISDS. I am also very thankful to Dr. Alan Gelfand, Dr. Feng Liang and Dr. Dalene Stangl for their encouragements during difficult times.

Special thanks is reserved for all my friends at ISDS, especially Laura Gunn, Carlos Carvalho, Jun Duan, Eric Laber and Scotland Leman for their wonderful friendship and for making my time at Duke University a great memorable personal experience.

I am also grateful to Pat Johnson, Krista Moyle, Elizabeth Clark, Susan Gillispie for their indispensable help I received at ISDS.

Most of all I would like to thank my parents to whom this work is dedicated. Without their unconditional love and support this work would not have been possible.

Chapter 1

Introduction

In this dissertation, we explore a new class of Bayesian nonparametric regression models based on Lévy process priors. We denote it as *Lévy adaptive regression kernel* or "LARK". The LARK framework falls into the general category of stochastic expansions of functions in an overcomplete dictionary (Abramovich *et al.*, 2000). The unknown function is represented as a weighted sum of kernel functions at arbitrary locations, with the number of kernels treated as a free parameter. Scaling parameters of the kernels are taken as location specific and thus are adaptive. The join prior distribution for the number of kernels, kernel regression coefficients and kernel associated parameters is constructed through Lévy random fields. We highlight both the modeling and computational advantage of LARK models. In addition, variety of simulated and real data sets are analysed to illustrate LARK models. We start by reviewing theory of Lévy processes on which LARK models are built. We then introduce the LARK model in the context of nonparametric regression problem followed by discussion of its applications for constructing flexible time series models and spatio-temporal models.

1.1 Lévy Processes

A Lévy process is a continuous-time analogue of a random walk. Theory of general Lévy processes is receiving increasing attention due to its flexibility in modeling jumps, extremes and other anomalous behavior of phenomena. Lévy processes have been successfully applied in many fields such as statistics, finance and physics. In this section we first recall the definition of a Lévy process and some of its basic properties. We then discuss gamma process in detail since it is the prior process we used in this dissertation. The connection between gamma processes and Bayesian adaptive kernel methods is introduced together with a couple of illustrative examples.

Definition 1.1. *(Lévy process) Let* $\mathbb{P}$ *be a probability measure on* $(\Omega, \mathcal{F})$. $X = \{X_t\}$ *for* $t \geq 0$ *is said to be a Lévy process for* $(\Omega, \mathcal{F}, \mathbb{P})$ *if*

1. *X has independent increments.*
2. *$X_0 = 0$ almost surely.*
3. *X is stochastically continuous, i.e., for any $s, t \geq 0$, $X_{t+s} - X_s \xrightarrow{\mathbb{P}} 0$ as $t \to 0$*
4. *X is time homogeneous, i.e., for $t, s \geq 0$, the law of $X_{t+s} - X_s$ does not depend on s.*
5. *X is right continuous with left limits almost surely.*

Without 5, X is said to be a Lévy process in law. It can be proved that each Lévy process in law has a *modification* that is a Lévy process, where modifications are defined as:

Definition 1.2. *Two stochastic processes $\{X_t\}$ and $\{Y_t\}$ are modifications of each other if $\mathbb{P}(X_t = Y_t) = 1$ for all t.*

So, we consider two processes identical in law as the same process. Let $\phi_X(u) \equiv \mathbb{E}[e^{iuX}]$ be the characteristic function for a random variable X. It can be shown that, for a Lévy process X_t, $\phi_{X_t}(u) = (\phi_{X_1}(u))^t$. So, the marginal distribution of a Lévy process X_t is determined by X_1.

There is a strong relationship between Lévy processes and infinitely divisible distributions, which we shall define now:

Definition 1.3. *A random variable X is infinitely divisible (ID) if, $\forall n \in \mathbb{N}$, there is an i.i.d. sequence $X_1, X_2, \cdots, X_n$, such that $X \overset{d}{=} X_1 + X_2 + \cdots + X_n$.*

The following theorem links infinitely divisible distributions to Lévy processes:

Theorem 1.1. *For any random variable Z, the following conditions are equivalent:*

1. *Z is infinitely divisible.*
2. *$Z \overset{d}{=} X_1$ for some Lévy process X_t.*

Actually, there is a one-to-one correspondence between the collection of all infinitely divisible distributions and the collection of all Lévy processes.(Sato, 1999). Next, we state a result that is one of the most fundamental one in probability theory:

Theorem 1.2. *(Lévy-Khinchine Formula) (Jacod and Shiryaev, 1987, p.75). X is a random variable which is infinitely divisible, then for some $a \in \mathbb{R}$, some $\sigma^2 \geq 0$ and some σ-finite measure ν on $\mathbb{R}/0$, s.t. $\int_{\mathbb{R}}(1 \wedge u^2)\nu(du) \leq \infty$. X has a characteristic function of the form:*

$$\mathbb{E}[e^{i\theta X}] = \exp\left\{ i\theta a - \theta^2\sigma^2/2 + \int_{\mathbb{R}} (e^{i\theta u} - 1 - i\theta h(u))\nu(du) \right\}$$

where $h(u) = u + \mathcal{O}(u^2)$ *near* $u = 0$*; if* $\int_{\mathbb{R}}(1 \wedge |u|)\nu(du) \le \infty$*, then we may take* $h(u) = 0$*.*

If the above representation exists, we call the measure ν Lévy measure. Here are some examples of the characteristic functions and their corresponding Lévy measures of some common infinitely divisible random variables:

Table 1.1: Several common infinitely divisible random variables and their corresponding Lévy measures.

Distribution	Log Chf	Lévy measure
$\mathsf{Po}(\lambda)$	$\lambda(e^{i\omega} - 1)$	$\lambda\delta_1(du)$
$\mathsf{Ga}(\alpha, \lambda)$	$-\alpha\log(1 - i\lambda\omega)$	$\alpha e^{-\lambda u}u^{-1}\,du$
$\mathsf{C}(\gamma, 0)$	$-\gamma\|\omega\|$	$\gamma\, u^{-2}\,du$
$\mathrm{S}\alpha\mathrm{S}(\alpha, \beta, \gamma)$	$-\gamma\|\omega\|^{\alpha}[1 - i\beta\tan\frac{\pi\alpha}{2}sgn(\omega)]$	$c_\alpha\gamma u^{-1-\alpha}\,du$

The main results stated above are for one dimensional case, but all the results can be extended to higher dimensions.

Common Lévy processes include Brownian motion, compound Poisson process, gamma process etc. A direct consequence of Lévy Khinchine theorem states that any Lévy process can be decomposed into sum of a Brownian motion with a drift and a pure jump process. Actually, Brownian motion with a drift term is the only Lévy process that has continuous path. In this dissertation, we only consider pure jump Lévy processes. In particular, we consider a class of increasing Lévy processes X_t, which is usually called subordinators. For a subordinator, its associated Lévy measure ν has the following interpretation. For a Borel measurable set $A \subset \mathbb{R}_+$, the number of jumps of sizes $(X_t - X_{t^-}) \in A$ follows a Poisson distribution with mean $\nu(A)$. Implicit in the Lévy-Khinchine theorem, a subordinator X_t can be represented as a stochastic integral of a Poisson random measure on $\mathbb{R}_+^2$. But first, we shall define a Poisson random field on $\mathbb{R}^d$.

Definition 1.4. *Let Π be a random subset of $\mathbb{R}^d$ and let $N(A)$ denote the number of points in $\{\Pi \cap A\}$. A Poisson random field on $\mathbb{R}^d$ is a random countable subset Π of $\mathbb{R}^d$, such that:*

1. *for any disjoint measurable subsets $A_1, A_2, \cdots, A_n$ of $\mathbb{R}^d$, the random variables $N(A_1), N(A_2), \cdots, N(A_n)$ are independent.*
2. *$N(A)$ has Poisson distribution $\mathsf{Po}(\mu(A))$, where μ is a positive measure on $\mathbb{R}^d$.*

If X_t is an increasing Lévy process, it can be shown that

$$X_t = \iint_{\mathbb{R}_+ \times [0,t]} u N(du\, ds)$$

where $N(du\, ds)$ defines a Poisson random field on $\mathbb{R}_+^2$ with mean measure $\mathbb{E}[N(du\, ds)]$ $= \nu(du\, ds)$, and $\nu(du\, ds)$ is the Lévy measure. More generally, for a Borel measurable function $\psi : \mathbb{T} \to \mathbb{R}$, where $\mathbb{T}$ is a bounded Borel subset of $\mathbb{R}_+$, we can construct a stochastic integral with respect a Lévy random measure $X(dt)$ as follows:

$$X[\psi] \equiv \int_{\mathbb{T}} \psi(t) X(dt) \equiv \iint_{\mathbb{R}_+ \times \mathbb{T}} u\psi(t) N(du\, dt) \tag{1.1}$$

If the integrability condition $\int_{\mathbb{R}} (|u| \wedge 1)\nu(du) < \infty$ satisfies, the characteristic functional of $X[\psi]$ is:

$$\mathbb{E}\left[e^{iX[\psi]}\right] = \exp\left\{\iint_{\mathbb{R}_+ \times \mathbb{T}} (e^{iu\psi(t)} - 1)\nu(du\, dt)\right\}$$

1.2 Adaptive Kernel Methods and Gamma Process Priors

In this section, we consider a particular type of Lévy process, i.e., a gamma process, in detail. We then explain how a gamma process can be used to construct the prior distribution for a wide class of models based on adaptive kernel smoothing. We start with the conventional definition of a gamma process.

Definition 1.5. *A gamma process $\mathcal{GP}(\alpha, \beta)$ denotes a stochastic process X_t ($t \geq 0$) with the properties:*

1. $X_0 = 0$ *a.s.*
2. X_t *has independent increments on disjoint intervals.*
3. *for any* $0 \leq s < t$, $X_t - X_s \sim \mathsf{Ga}(\alpha(t-s),\, \beta)$.

where $\alpha()$ is the mean measure and β is the scale parameter. If a Lebesgue measure is used for α, a gamma process X_t can be represented as a stochastic integral:

$$X_t = \iint_{\mathbb{R}_+ \times [0,t]} uN(du\, ds)$$

where $N(du\, ds)$ denotes a Poisson random measure defined on $\mathbb{R}_+ \times [0, T]$ for some $T > 0$, with mean measure:

$$\mathbb{E}[N(du\, ds)] = \nu(du)ds = \alpha u^{-1} e^{-\beta u} ds$$

where $\nu(du)$ denotes the Lévy measure associated with a gamma random variable. So, for any $\epsilon > 0$, a gamma process defined on $[0, T]$ can be viewed as jumps of size $u \geq \epsilon$ arriving as a Poisson process with rate $T\nu([\epsilon, \infty)) = T \int_\epsilon^\infty \alpha u^{-1} e^{-\beta u}\, du$.

Notice that $\nu([0,\infty)) = \int_0^\infty \alpha e^{-\beta u} u^{-1} du = \infty$. This means that there are infinitely many jumps in a unit interval with probability one. But for any $\epsilon > 0$, $\nu([\epsilon,\infty)) < \infty$. So we can choose a threshold $\epsilon > 0$ and the number of jumps of sizes bigger than ϵ is finite with probability one. There are still infinitely many of jumps of sizes smaller than ϵ, but they add up to a finite number, since for a gamma process, $\int_0^\infty (1 \wedge u)\nu(du) < \infty$. We can thus approximate the Lévy measure $\nu(du)$ by $\nu_\epsilon(du)$:

$$\nu_\epsilon(du) \equiv \alpha e^{-\beta u} u^{-1} I_{\{u \geq \epsilon\}} du\,.$$

Let Δ_ϵ denotes the sum of jumps of sizes smaller than ϵ for a gamma process defined on a unit interval, we have

$$\begin{aligned} \mathbb{E}[\Delta_\epsilon] &= \int_0^\epsilon u\alpha e^{-\beta u} u^{-1} du = \alpha(1 - e^{-\beta u})/\beta \ \leq \alpha\epsilon \\ \mathbb{V}[\Delta_\epsilon] &= \int_0^\epsilon u^2 \alpha e^{-\beta u} u^{-1} du = \frac{\alpha}{\beta^2}(1 - e^{-\beta\epsilon} - \epsilon\beta e^{-\beta\epsilon}) \ \leq \alpha\epsilon^2/2, \end{aligned}$$

thus we can approximate a gamma random measure arbitrarily well with arbitrarily small ϵ.

Conventional gamma processes only allow positive jumps. But the Poisson representation offers a nice way to extend the conventional gamma process to allow negative jumps:

Definition 1.6. *X_t $(t \geq 0)$ is said to be a symmetric gamma process $\mathcal{GP}_s(\alpha,\beta)$ if:*

$$X_t \equiv \iint_{\mathbb{R}\times[0,t]} |u| N(du\,ds)$$

where $N(du\,ds)$ denotes a Poisson random field on $\mathbb{R} \times [0,T]$ with mean measure $\mathbb{E}[N(du\,ds)] = \int_{-\infty}^\infty \alpha |u|^{-1} e^{-\beta|u|}\,du$ for some $\alpha > 0$ and $\beta > 0$.

It is useful to allow negative jumps in certain applications, as we will show a few examples in Chapter 2.

With Poisson representation, we can easily extend one dimensional gamma processes to d dimensional gamma random fields.

Definition 1.7. *A gamma random field* Γ *is a continuous (in probability) linear operator that assigns random variables* $\Gamma[\psi]$ *for bounded Borel measurable functions* ψ *defined on some subset* $\mathbb{S} \subset \mathbb{R}^d$*, such that* $\Gamma[I_A] \sim \mathsf{Ga}(\alpha|A|, \beta)$ *(*A *being a Borel set in* $\mathbb{S}$ *and* $|A|$ *being the size of* A*.)*

We can construct a gamma random field Γ through a Poisson random field:

$$\Gamma[\psi] \equiv \iint_{\mathbb{R}_+ \times \mathbb{S}} \psi(s) u N(du\, ds) \tag{1.2}$$

where $N(du\, ds)$ defines a Poisson random field on $\mathbb{R}_+ \times \mathbb{S}$ with mean measure $\mathbb{E}[N(du\, ds)] = \alpha u^{-1} e^{-\beta u} ds$

The Poisson representation suggests an elegant way to sample from a gamma random field, as was discovered by Wolpert and Ickstadt (1998), and is named as *Inverse Lévy Measure* (ILM). We state below a simplified version of the ILM procedure which suffices for the purpose of this dissertation. More general version can be found in the cited paper.

1. Generate J independent identically distributed random variables $\{\tau_1, \cdots, \tau_J\}$, where τ_j is uniformly distributed on $\mathbb{S}$, a bounded subset of $\mathbb{R}^d$.

2. Generate J successive jump times of a standard Poisson process $\{\xi_1, \cdots, \xi_J\}$.

3. Set $u_j \equiv E_1^{-1}(\xi_j/\alpha)/\beta$, for $j = 1, \cdots, J$.

The random field $\Gamma[I_A] \equiv \sum_{j=1}^{J} u_j I_{\{\tau_j \in A\}}(\tau_j)$ has the gamma random field distribution $\Gamma[I_A] \sim \mathsf{Ga}(\alpha\,|A|, \beta)$.

For a kernel function $k(s;\sigma,\theta)$ defined on $\mathbb{S} \times \Theta$, where $\mathbb{S}$ is a bounded subset of $\mathbb{R}^d$, $s \in \mathbb{S}$, $\sigma \in \mathbb{S}$ and $\theta \in \Theta$, set $f(s)$ by:

$$f(s) \equiv \Gamma[k(s;\sigma,\theta)] \equiv \iint_{\mathbb{R}_+ \times \Theta} u k(s;\sigma,\theta) N(du\, d\sigma\, d\theta) \tag{1.3}$$

where $N(du\, d\sigma\, d\theta)$ defines a Poisson random field on $\mathbb{R}_+ \times \mathbb{S} \times \Theta$, with mean measure $\mathbb{E}[du\, d\sigma\, d\theta] = \nu(du) d\sigma \pi(d\theta) = \alpha u^{-1} e^{-\beta u} d\sigma \pi(d\theta)$, where $\pi(d\theta)$ defines a finite measure on Θ. The discrete nature of a Poisson random field allows the following representation of Eqn. (1.3):

$$f(s) = \sum_{j=1}^{J} u_j k(s;\sigma_j,\theta_j) \tag{1.4}$$

In nonparametric regression context, Eqn. (1.4) can be viewed as a stochastic expansion of a d-dimensional function f defined on $\mathbb{S}$, i.e., $f(s)$ is the sum of J weighted kernel elements, with $(u_1, \cdots, u_J)$ being the weights (coefficients) and $(\sigma_1, \cdots, \sigma_J)$ being the locations of the J kernels. The kernels are adaptive since at each location σ_j, the kernel has its own parameter θ_j. The number of kernels J follows a Poisson distribution. Eqn. (1.4) is essentially an adaptive kernel model since it can also be interpreted as J smoothing kernels smooth out J jumps with magnitudes $(u_1, \cdots, u_J)$. The gamma random field Γ defined on $\mathbb{S} \times \Theta$ governs the joint probability distribution of $(J, u_1, \cdots, u_J, \sigma_1, \cdots, \sigma_J, \theta_1, \cdots, \theta_J)$. In Bayesian framework, $\Gamma[k]$ serves as a natural prior distribution for functions defined on $\mathbb{S}$. It can generate wide classes of functions. We conclude this chapter by two demonstration examples shown in Fig. (1.1) and Fig. (1.2). In the next three chapters, we discuss in detail the applications of the model defined in Eqn. (1.4) in the areas of nonparametric regression, semiparametric time series modeling and semiparametric spatio-temporal modeling.

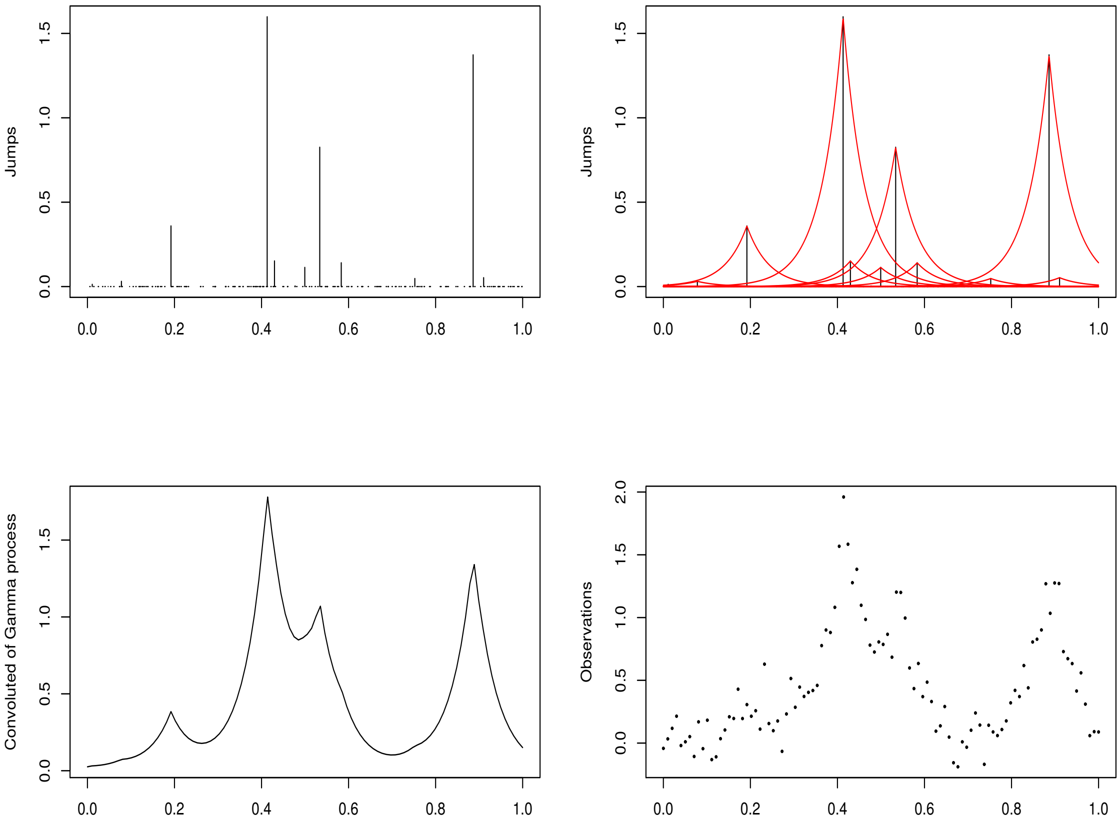

Figure 1.1: Gamma process and adaptive kernel smoothing in one dimension. The upper left figure shows a realization of a $\mathcal{GP}(2.0, 1.0)$ defined on $[0, 1]$; the upper right figure shows the kernel $k(t; \tau, \lambda) = e^{-\lambda|t-\tau|}$ putting at each jump, we choose $\lambda = 20.0$ for all jumps; the lower left figure shows the kernel smoothing of $\mathcal{GP}(2.0, 1.0)$; the lower right figure shows the observations generated by adding white noise $\mathsf{N}(0, 0.15^2)$ to the mean curve at 100 equally spaced points.

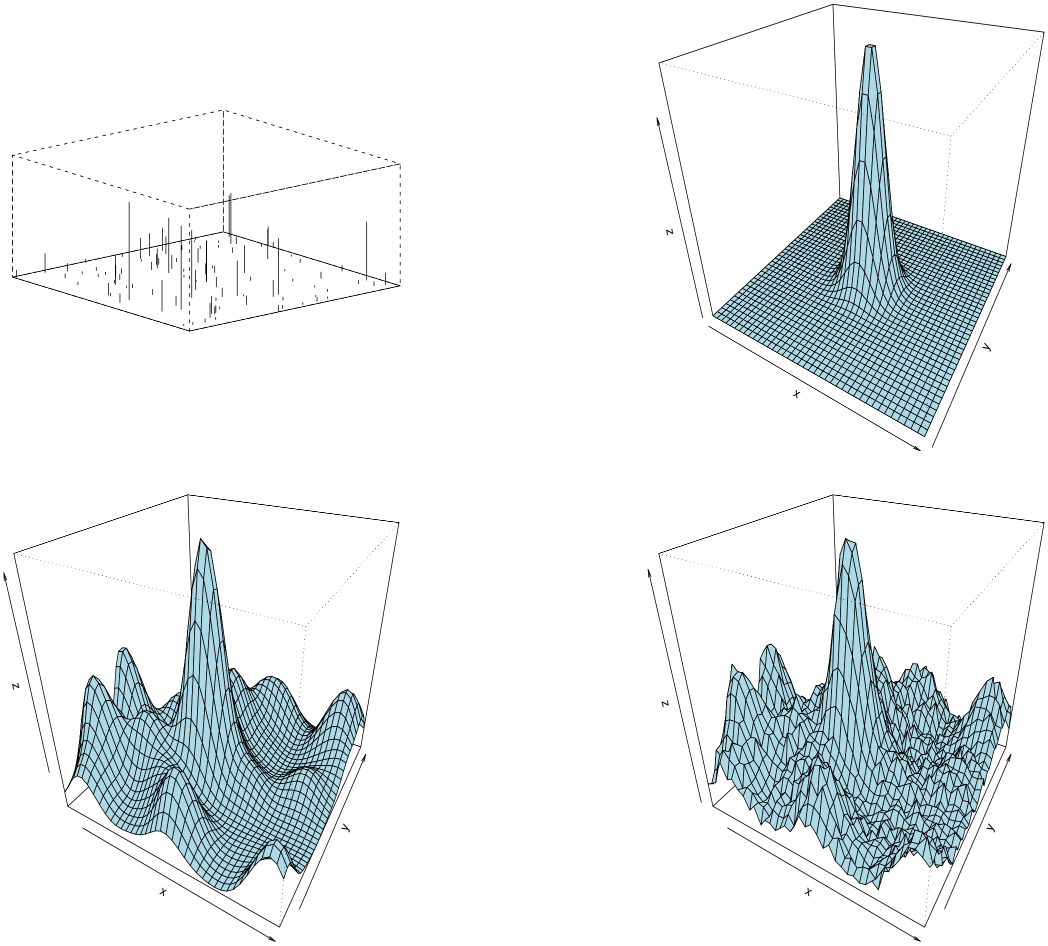

Figure 1.2: Gamma random field and adaptive kernel smoothing in two dimensions. The upper left figure shows a realization of a gamma random field $\Gamma(2.0, 0.5)$ defined on $[0,1] \times [0,1]$; the upper right figure shows the spatial kernels $k(s; \sigma, \lambda) = e^{-\lambda ||s-\sigma||^2}$, we choose $\lambda = 125.0$ for all jumps; the lower left figure shows the spatial surface obtained by kernel smoothing of $\Gamma(2.0, 1.0)$; the lower right figure shows the observations generated by adding white noise $\mathsf{N}(0, 0.25^2)$ to the mean curve at 40×40 equally spaced grid points.

Chapter 2

Lévy Adaptive Regression Kernels

2.1 Introduction

Suppose we have n noisy measurements $Y_1, \ldots, Y_n$ of an unknown real-valued function $f : \mathcal{X} \to \mathbb{R}$ on some compact space $\mathcal{X}$,

$$Y_i = f(x_i) + e_i \qquad e_i \overset{iid}{\sim} \mathsf{N}(0, \sigma^2) \tag{2.1}$$

observed at points $x_i \in \mathcal{X}$. In nonparametric regression models, the mean function $f(\cdot)$ is often regarded as an element of some Hilbert space $\mathcal{H}$ of real-valued functions on $\mathcal{X}$, and is expressed as a linear combination of basis functions $\{g_j\} \subset \mathcal{H}$:

$$f(x_i) = \sum_{0 \le j < J} g_j(x_i)\beta_j \tag{2.2}$$

with unknown coefficients in the expansion $\{\beta_j\}_{0 \le j < J}$. There is a vast literature on classical and Bayesian approaches for estimating f from noisy data using such methods as regression splines, Fourier expansions, wavelets expansions, and kernel methods, including kernel regression and support (or relevance) vector machines (see Chu and Marron, 1991; Cristianini and Shawe-Taylor, 2000; Denison *et al.*, 2002; Vidakovic, 1999; Wahba, 1990, 1992, for background and references). Solutions using smoothing splines and support vector machines (among others) generally use as many basis elements, J, as there are data points, n, but employ regularization to avoid over-fitting. Sparser solutions (using fewer basis elements, $J < n$) may be obtained through other choices of penalty in the regularization problem, as in the LASSO (Tibshirani, 1996), or (often equivalently) through choice of prior distributions, as in relevance vector machines (Tipping, 2001). Sparse solutions may also be achieved by using variable selection techniques to choose a few well-placed basis functions, perhaps in conjunction with regularization (Chen *et al.*, 1998; Denison *et al.*, 1998; DiMatteo *et al.*, 2001; Mallat and Zhang, 1993; Johnstone and Silverman, 2005; Smith and Kohn, 1996; Wolfe *et al.*, 2004).

In most signal processing and other applications where functions exhibit nonstationarity, no single (especially orthonormal) basis will lead to a sparse representation (Donoho and Elad, 2003; Wolfe *et al.*, 2004). Overcomplete dictionaries and frames (Daubechies, 1992; Mallat and Zhang, 1993) provide a larger collection of generating elements $\{g_\omega\}_{\omega \in \Omega}$ than with a single basis for $\mathcal{H}$, potentially allowing for more effective signal extraction and data compression for functions. Examples of overcomplete dictionaries include unions of bases, Gabor frames, nondecimated or translational invariant wavelets, wavelet packets, or more general kernel functions or generating functions $g(x, \omega)$ where $\omega \in \Omega$ controls local or global features of the generating function, such as translations, dilations, modulations, shape parameters or other features. Because of the redundancy inherent in the overcomplete representation, coefficients for expansions using the complete dictionary are no longer unique. This lack of uniqueness is advantageous, as it is possible to find a more parsimonious representation from the dictionary than those obtained using any single basis.

In this paper, we develop a fully Bayesian method for the sparse regression problem using overcomplete dictionaries. We begin in Section 2.2 by introducing Lévy random fields, which are used to induce prior distributions for $f \in \mathcal{H}$ via a kernel convolution with an infinitely divisible random prior measure. We denote the new class of kernel models as *Lévy adaptive regression kernel* or "LARK" models. The LARK framework allows both the number of kernels and kernel-specific parameters to adapt to the unknown degree of sparsity in representing f. For many Lévy measures, this results in a stochastic expansion with an infinite number of terms. In Section 2.3 we provide an approximation to the Lévy random field that permits tractable computation via a sequence of compound Poisson random fields. Exploiting the construction of Lévy random fields through Poisson random fields, we develop a hierarchical representation of the LARK model in Section 2.4 and describe posterior inference for the LARK model using reversible jump Markov chain Monte Carlo (RJ-MCMC) algorithms. We then compare our LARK method to other procedures in a simulation study and on a real example in Section 2.5. We conclude in Section 2.6 by contrasting the LARK formulation with other Bayesian and regularization methods for nonparametric regression and discussing possible extensions of the LARK model.

2.2 Prior Distributions

To make Bayesian inference about the unknown function $f \in \mathcal{H}$ of Eqn. (2.1), we must propose a prior distribution on $\mathcal{H}$ for f, with f represented as an expansion of dictionary elements g. Let Ω to be a complete separable metric space and choose a Borel measurable function $g : \mathcal{X} \times \Omega \to \mathbb{R}$, and set $g_j(x_i) = g(x_i, \omega_j)$. Possible choices for $g(x, \omega)$ include kernel functions, such as a Gaussian kernel

$$g_G(x, \omega) = \exp\{-\frac{1}{2}\lambda(x - \chi)^2\} \tag{2.3}$$

with $\omega \equiv (\chi, \lambda) \in \mathcal{X} \times \mathbb{R}^+ \equiv \Omega$, or a Laplace kernel

$$g_L(x, \omega) = \exp\{-\lambda|x - \chi|\} \tag{2.4}$$

with ω and Ω defined as in the Gaussian case. However, there is no need to restrict attention to symmetric kernels; asymmetric kernels, such as a one-sided exponential

$$g_E(x, \omega) = \exp\{-\lambda|x - \chi|\}\mathbf{1}_{\{x>\chi\}} \tag{2.5}$$

are also of interest, for example, in modeling pollution dissipation over time. Generating functions such as step functions (as in the Haar wavelet)

$$g_H(x, \omega) = I_{0<(x-\chi)<\lambda} \tag{2.6}$$

with $\Omega \equiv [0, 1] \times \Lambda$, or continuous rescaling and shifting of other wavelet functions $\phi(\cdot)$,

$$g_\phi(x, \omega) = \lambda^{1/2}\phi(\lambda(x - \chi)) \tag{2.7}$$

are other possibilities. In each of the examples above, the parameter λ controls scaling of the function and χ is viewed as a location parameter. While we will focus on the case $\mathcal{X} \subset \mathbb{R}$, generating functions $g(\cdot, \omega)$ may be defined similarly for higher dimensional spaces $\mathcal{X}$.

As a slight extension of the basis expansion of Eqn. (2.2), set

$$f(x) \equiv \sum_{0\le j<J} g(x, \omega_j)\,\beta_j \tag{2.8}$$

for a random number $J \leq \infty$ of pairs $\beta_j \in \mathbb{R}$, $\omega_j \in \Omega$. Note this is equivalent to specifying a random (signed) Borel measure $\mathcal{L}(d\omega) = \sum \beta_j \delta_{\omega_j}(d\omega)$ on Ω, giving the equivalent representation:

$$f(x) = \int_\Omega g(x,\omega)\mathcal{L}(d\omega). \tag{2.9}$$

Thus to assign prior distributions on functions of the form (2.8), we need to specify a prior distribution for the random measure $\mathcal{L}(d\omega)$. It is convenient to view a random measure as a stochastic process or random field, $\mathcal{L}(\cdot)$, indexed by sets A_i, where $(\mathcal{L}(A_1), \ldots, \mathcal{L}(A_k))$ is a random vector. Lévy random fields are ideal for this purpose as they facilitate construction of non-negative functions, as well as real valued functions, and, as we will see in section 2.4.1, are amenable to posterior simulation.

2.2.1 Lévy Random Fields

For the random measure $\mathcal{L}(d\omega)$ and disjoint Borel sets $A_i \subset \Omega$, the random variables $\mathcal{L}(A_i) = \int_{A_i} \mathcal{L}(d\omega)$ will be independent, if we choose any positive number $\nu_+ > 0$, any probability distribution π on $\mathbb{R} \times \Omega$, and give J a Poisson distribution, $J \sim \mathsf{Po}(\nu_+)$, and, conditional on J, accord the $(\beta_j,\, \omega_j) \in \mathbb{R} \times \Omega$ independent, identical distributions, $(\beta_j, \omega_j) \overset{iid}{\sim} \pi(d\beta,\, d\omega)$. In that case, $\mathcal{L}$ will assign independent infinitely-divisible (or ID) random variables $\mathcal{L}(A_i)$ to disjoint Borel sets $A_i \subset \Omega$, with characteristic functions

$$\mathsf{E}\left[\exp\left\{it\mathcal{L}(A_i)\right\}\right] = \exp\left\{\iint_{\mathbb{R}\times A_i} \left(e^{it\beta} - 1\right)\nu(d\beta, d\omega)\right\} \tag{2.10}$$

where $\nu(d\beta, d\omega) = \nu_+ \pi(d\beta, d\omega)$ is the product of the Poisson rate ν_+ for J and the distribution $\pi(d\beta, d\omega)$ for $\{(\beta_j, \omega_j)\}$. When they exist, each $\mathcal{L}(A_i)$ has mean

$\mathsf{E}[\mathcal{L}(A_i)] = \iint_{\mathbb{R}\times A_i} \beta\, \nu(d\beta, d\omega)$ and variance $\mathsf{V}ar[\mathcal{L}(A_i)] = \iint_{\mathbb{R}\times A_i} \beta^2\, \nu(d\beta, d\omega)$.

Such a random measure $\mathcal{L}$ induces a continuous (in probability) linear mapping $g \mapsto \mathcal{L}[g]$; the collection of $\{\mathcal{L}[g] : x \in \mathcal{X}, g \in \mathcal{G}\}$ is called a Lévy random field. The random measure assigns random variables,

$$\mathcal{L}[g] \equiv \int_\Omega g(x,\omega)\, \mathcal{L}(d\omega) = \sum_{0\le j<J} g(x,\omega_j)\, \beta_j \tag{2.11}$$

to continuous compactly-supported functions g on $\mathcal{X} \times \Omega$, with the characteristic functional for $\mathcal{L}[g]$ given by

$$\mathsf{E}[\exp\left(it\mathcal{L}[g]\right)] = \exp\left\{\iint_{\mathbb{R}\times\Omega} \left(e^{itg(x,\omega)\beta} - 1\right) \nu(d\beta, d\omega)\right\}. \tag{2.12}$$

When $\nu(\mathbb{R}\times\Omega)$ is finite then $\mathcal{L}[g]$ is called a *compound Poisson* random field, and J in the representation (2.11) is almost surely finite.

More generally, the "Lévy measure" $\nu(d\beta, d\omega)$ in (2.10) need not be finite, for the random measure $\mathcal{L}$ to exist and, as we will see, for the random field $\mathcal{L}[g]$ to be well defined. In constructing a random measure, for any countable partition of a Borel set A into disjoint sets A_i, $\mathcal{L}(A) = \sum_i L(A_i)$. In constructing prior distributions for $\mathcal{L}$, it is convenient to use a family of distributions such that the random variables $L(A_i)$ are closed under summation for arbitrary partitions; indeed this is a defining characteristic of infinitely divisible (ID) distributions. By the Lévy-Khinchine representation theorem (Rajput and Rosiński, 1989, Proposition2.1), an infinitely divisible (ID) random variable $\mathcal{L}(A)$ has characteristic function

$$\mathsf{E}[\exp\left(it\mathcal{L}(A)\right)] = \exp\left\{it\delta(A) - \frac{1}{2}t^2\Sigma(A) + \iint_{\mathbb{R}\times A} \left(e^{it\beta} - 1 - it\beta\mathbf{1}_{[-1,1]}(\beta)\right) \nu(d\beta, d\omega)\right\} \tag{2.13}$$

where the characteristic triplet of measures (Σ, ν, δ) uniquely characterizes the random measure $\mathcal{L}$: $\delta(d\omega)$ is a signed measure on Ω, $\Sigma(d\omega)$ is a positive sigma-finite measure on Ω, and $\nu(d\beta, d\omega)$ is a positive sigma-finite measure on $\mathbb{R} \times \Omega$, satisfying

$$\iint_{\mathbb{R}\times A} (1 \wedge \beta^2)\nu(d\beta, d\omega) < \infty \tag{2.14}$$

and $\nu(\{0\}, A) = 0$ (for more details on the non-stationary version of the classic Lévy-Khinchine (Khinchine and Lévy, 1936) representation see Jacod and Shiryaev (1987, *p.* 75), Sato (1999, §9), Cont and Tankov (2004, *pp.* 457–459) or Wolpert and Taqqu (2005)). This representation implies that the ID random measure $\mathcal{L}$ may be decomposed into a deterministic "drift" term based on δ, a continuous Gaussian component with covariance determined by the measure Σ, and a discontinuous pure jump component given by the last expression with Lévy measure ν, which controls the rate and size of jumps. When the Lévy measure ν satisfies (2.14), but $\nu([-1, 1], A)$ is not finite, the term $\beta\mathbf{1}_{[-1,1]}(\beta)$ in the characteristic function is required to "compensate" for the infinite number of small jumps which are not absolutely summable. Other compensating functions may be used, with an adjustment to δ. For every bounded measurable function h which satisfies

$$h(\beta) = \beta + O(\beta^2) \tag{2.15}$$

for β in a neighborhood around 0, an equivalent version of the Lévy-Khinchine representation may be obtained as

$$\mathsf{E}[\exp(it\mathcal{L}[A_i])] = \exp\left\{it\delta_h(A_i) - \frac{1}{2}t^2\Sigma(A_i) + \iint_{\mathbb{R}\times A_i}\left(e^{it\beta} - 1 - it\,h(\beta)\right)\nu(d\beta, d\omega)\right\} \tag{2.16}$$

where

$$\delta_h(d\omega) \equiv \delta(d\omega) + \int_{\mathbb{R}} \left(h(\beta) - \beta\mathbf{1}_{[-1,1]}(\beta)\right)\nu(d\beta, d\omega) \tag{2.17}$$

(see Jacod and Shiryaev (1987, Chapter II) or Cont and Tankov (2004, Chapter 3)). While ν and Σ are unaffected by the choice of compensator or truncation function h, the so-called "drift" term δ_h is dependent on the choice of h, thus the characteristic triplet (Σ, ν, δ_h) of the Lévy random field is given with respect to the choice of compensator function h.

This more general construction of $\mathcal{L}$ has both continuous and discrete components. Because we are interested in stochastic expansions of f with a countable basis, from now on we will consider random measures without a Gaussian component and take $\Sigma \equiv 0$ and set $\delta = 0$. As suggested by the connection between the compound Poisson random field and Lévy random field in section 2.2.1, a random measure $\mathcal{L}(d\omega)$ may be formally constructed in terms of a (compensated) Poisson random measure plus a deterministic drift component. Besides providing a more intuitive understanding of the random measures, this representation is key to the development of tractable posterior inference using stochastic computation.

2.2.2 Poisson Construction of Lévy Random Fields

As a generalization of the compound Poisson random field, we begin with a Lévy measure $\nu(d\beta, d\omega)$ satisfying a more restrictive condition

$$\iint_{\mathbb{R}\times A} (1 \wedge |\beta|)\, \nu(d\beta, d\omega) < \infty \tag{2.18}$$

for every $A \subset \Omega$. One can always construct a Poisson random measure $N(d\beta, d\omega)$ on the space $\mathbb{R} \times \Omega$ with intensity (sometimes called *control*) measure $\nu(d\beta, d\omega)$ assigning independent Poisson-distributed random variables $N(B_i) \sim \mathsf{Po}\big(\mu(B_i)\big)$ to disjoint sets $B_i \subset \mathbb{R} \times \Omega$ of finite μ-measure. Then the random measure $\mathcal{L}$ on Ω with characteristic triplet $(0, \nu, 0)$ may be constructed using a Poisson random measure on the larger space $\mathbb{R} \times \Omega$ as

$$\mathcal{L}(A_i) = \int_{A_i} L(d\omega) \overset{d}{=} \iint_{\mathbb{R}\times A_i} \beta N(d\beta, d\omega) = \sum_{0 \le j < J} \beta_j \mathbf{1}_{(\beta_j, \omega_j)}(d\beta, d\omega) \tag{2.19}$$

(Wolpert and Taqqu, 2005). Intuitively, the Poisson measure $N(d\beta, d\omega)$ describes the number of points or jumps in $d\beta \times d\omega$ (or A_i), which occur with intensity $\mathsf{E}[N(d\beta, d\omega)] = \nu(d\beta, d\omega)$. With β interpreted as jump height, then $\mathcal{L}(A_i)$ is the sum of all jumps in $\mathbb{R}\times A_i$. Of course, when $J \equiv \nu(\mathbb{R}, \Omega) = \infty$, the Poisson measure will have an infinite number of jumps. A discrete signed random measure $\mathcal{L}(d\omega)$ will still make sense with infinitely many support points $\{\omega_j\}$ with associated jumps of size $\{\beta_j\}$, as long as the jumps are absolutely summable. For $J = \infty$, we must have only finitely many large (in absolute value) jumps, so for every $\epsilon > 0$, we require

$$\nu((\epsilon, \epsilon)^c, \Omega) < \infty. \tag{2.20}$$

Thus while $\nu((-\epsilon, \epsilon), \Omega)$ is infinite, leading to an infinite number of "small" jumps, it will suffice to have

$$\iint_{(-\epsilon,\epsilon)\times A_i} |\beta| \nu(d\beta, d\omega) < \infty. \tag{2.21}$$

to ensure the absolute summability of small jumps. For ν satisfying the more restrictive integrability condition (2.18), the Poisson sums are all well defined, so that no compensation is required ($h = 0$). For construction of a strictly positive random measure $\mathcal{L}(d\omega)$ consisting solely of pure jumps, the condition in (2.18) is actually necessary to guarantee absolute summability.

The situation is a little more delicate in the case where the Lévy measure does not satisfy the condition in (2.18), but the more general condition in (2.14) holds. For construction of signed measures, while absolute summability is sufficient for the existence of the characteristic function of $\mathcal{L}(A_i)$, it is not necessary as long as there is a suitable "cancellation" of the infinite number of small positive and negative jumps, which is achieved through the compensator h. The compensated Poisson representation of (2.19) is

$$\begin{aligned}
\mathcal{L}[A_i] &= \int_{A_i} L(d\omega) \\
&\overset{d}{=} \int_{A_i} \delta_h(d\omega) + \iint_{\mathbb{R}\times A_i} (\beta - h(\beta))\ N(d\beta, d\omega) + \iint_{\mathbb{R}\times A_i} h(\beta)\,\tilde{N}(d\beta, d\omega) \qquad (2.22) \\
&= \iint_{[-1,1]^c\times A_i} \beta N(d\beta, d\omega) + \iint_{[-1,1]\times A_i} \beta \tilde{N}(d\beta, d\omega) \qquad (2.23)
\end{aligned}$$

where $\tilde{N}(d\beta, d\omega) \equiv N(d\beta, d\omega) - \nu(d\beta, d\omega)$ is the compensated or centered Poisson measure, with mean 0 (Sato, 1999, page 38). If $\nu((-\epsilon, \epsilon)^c \times A_i) < \infty$, then $\mathcal{L}[A_i]$ may be equivalently represented as

$$\mathcal{L}[A_i] \overset{d}{=} \iint_{(-\epsilon,\epsilon)^c\times A_i} \beta N(d\beta, d\omega) - \iint_{[-1,-\epsilon)\cup(\epsilon,1]\times A_i} \beta \nu(d\beta, d\omega) + \iint_{(-\epsilon,\epsilon)\times A_i} \beta \tilde{N}(d\beta, d\omega). \tag{2.24}$$

For jumps larger than ϵ in absolute value, there are only finitely many points, thus the first integral is in fact a finite sum. So while $\iint_{(-\epsilon,\epsilon)\times A_i} |\beta| \nu(d\beta, d\omega)$ is infinite, the integral $\iint_{(-\epsilon,\epsilon)\times A_i} \beta^2 \nu(d\beta, d\omega)$ is finite, as is the integral with respect to the compensated Poisson measure $\tilde{N}$ (Samorodnitsky and Taqqu, 1994, page 158).

2.2.3 Construction of Lévy Random Fields $\mathcal{L}[g]$

We start by considering initially $\mathcal{G}$ defined by continuous compactly-supported functions $g(x, \omega)$ defined on $\mathcal{X} \times \Omega$. This includes the Gaussian and Laplace kernels defined over a compact Ω. Let $\mathcal{L}$ be an ID random measure with characteristic triplet $(0, \nu(d\beta, d\omega), 0)$, where the Lévy measure $\nu(d\beta, d\omega)$ satisfies the integrability condition of (2.18). Then

$$\begin{aligned} \mathcal{L}[g] &= \int g(x,\omega)\mathcal{L}(d\omega) = \iint_{\mathbb{R}\times\Omega} g(x,\omega)\,\beta\,N(d\beta, d\omega) \\ &= \sum_{0\le j<J} g(x,\omega)\,\beta_j, \end{aligned} \tag{2.25}$$

where $\{(\beta_j, \omega_j)\}$ are the (at most countable) support-points of the Poisson measure $N(d\beta, d\omega)$, thus justifying the expansion in (2.11) for the case with infinite Lévy measure ν. If the Lévy measure satisfies the more general integrability condition(2.14), then the random field is given by

$$\mathcal{L}[g] = \int g(x,\omega)\mathcal{L}(d\omega) = \iint_{(-1,1)^c\times\Omega} g(x,\omega)\,\beta\,N(d\beta, d\omega) + \iint_{[-1,1]\times\Omega} g(x,\omega)\,\beta\,\tilde{N}(d\beta, d\omega) \tag{2.26}$$

The Poisson integrals in (2.25)and (2.26) are well-defined because $g(x,\omega)\beta$ is ν-integrable on $[-1,1]\times\Omega$ and bounded on $[-1,1]^c\times\Omega$. For the general case, when the Lévy measure satisfies (2.14), the characteristic functional may be written as

$$\begin{aligned} \mathsf{E}[\exp(it\mathcal{L}[g])] &= \exp\Bigg\{ it\int_\Omega g(x,\omega)\delta_h(d\omega) + \\ &\iint_{\mathbb{R}\times\Omega} \left(e^{itg(x,\omega)\beta} - 1 - it\,g(x,\omega)h(\beta)\right)\nu(d\beta, d\omega)\Bigg\} \end{aligned} \tag{2.27}$$

where $\delta_h(d\omega)$ and h may be taken as 0 when condition (2.18) holds.

2.2.4 Extending $\mathcal{L}[k]$

We will need to define integrals $\int_\Omega g(x,\omega)\mathcal{L}(d\omega)$ for measurable functions $g : \mathbb{R}\times\Omega \to \mathbb{R}$ that may *not* be continuous or compactly-supported in the variable

ω. While $\mathcal{L}[g] = \int_\Omega g(\omega)\mathcal{L}(d\omega)$ is defined initially only for continuous compactly-supported functions g on Ω, it can be extended uniquely to the linear space g of limits of sequences $\{g_n\}$ for which $\mathcal{L}[g_n]$ converges in probability. The necessary and sufficient condition for a measurable function $g : \Omega \to \mathbb{R}$ to be in $\mathcal{G}$, hence for $\mathcal{L}[g]$ to be well defined, is that the real and imaginary parts of (2.27) exist and be finite, *i.e.*, that $g \in L^1(\Omega, d\delta_h)$ and

$$\begin{aligned} \iint_{\mathbb{R}\times\Omega} \left|\cos\big(g(\omega)\beta\big) - 1\right| \,\nu(d\beta, d\omega) \;&<\; \infty \\ \iint_{\mathbb{R}\times\Omega} \left|\sin\big(g(\omega)\beta\big) - g(\omega)h(\beta)\right| \,\nu(d\beta, d\omega) \;&<\; \infty. \end{aligned}$$

If ν satisfies equation (2.18), then it is enough that

$$\iint_{\mathbb{R}\times\Omega} \big(1 \wedge |\beta|\big)|g(\omega)|\nu(d\beta, d\omega) < \infty; \tag{2.28}$$

thus $\mathcal{G}$ includes all bounded measurable functions. When compensation is necessary, it is always sufficient for g that

$$\iint_{\mathbb{R}\times\Omega} \big(1 \wedge \beta^2\big)\big[g(x,\omega)^2\big]\nu(d\beta, d\omega) < \infty, \tag{2.29}$$

(Samorodnitsky and Taqqu, 1994, page 158). More generally, Rajput and Rosiński (1989) show that the space of functions that are integrable with respect to an ID random measure $\mathcal{L}(d\omega)$ are certain *Musielak-Orlicz* modular spaces.

The integral $\mathcal{L}[g]$ has a finite expectation whenever it exists. Moments of $f(x) = \mathcal{L}[g]$, when they exist, are easy to compute from (2.22) or using the characteristic function (2.27):

$$\mathsf{E}\big\{f(x)\big\} = \int_{\Omega} g(x,\omega)\delta_h(d\omega) + \iint_{\mathbb{R}\times\Omega} g(x,\omega)\,\big[\beta - h(\beta)\big]\nu(d\beta,d\omega) \tag{2.30}$$

$$\mathsf{C}ov\big\{f(x_1),f(x_2)\big\} = \iint_{\mathbb{R}\times\Omega} g(x_1,\omega)\,g(x_2,\omega)\,\beta^2\,\nu(d\beta,d\omega). \tag{2.31}$$

2.2.5 Choice of Lévy Measure

We now consider specific examples of Lévy measures and the corresponding kernel integrals. Well known examples of Lévy random fields include the Poisson random field, Gamma random field, Cauchy and Stable random fields. The Gamma random field is an example of a Lévy random field with infinite Lévy measure that satisfies the integrability condition (2.18), where (in the stationary case)

$$\nu(d\beta,d\omega) = \alpha(d\omega)\beta^{-1}e^{-\beta/\tau}\mathbf{1}_{\{\beta>0\}}\,d\beta \tag{2.32}$$

for some σ-finite measure $\alpha(d\omega)$ on Ω, giving $\mathcal{L}(A) \sim \mathsf{Ga}(\alpha(A),1/\tau)$ (with mean $\alpha(A)\tau$) for Borel measurable $A \subset \Omega$. Because the Gamma random measure has only positive jumps, coefficients in any expansion will be non-negative. Combined with generating functions that are always non-negative, this provides a direct way to model non-negative functions without having to transform the response Y as with Gaussian random field priors.

The measure in (2.32) may be generalized to allow construction of signed measures using a symmetric version of the Lévy measure,

$$\nu(d\beta,d\omega) = \alpha(d\omega)|\beta|^{-1}e^{-|\beta|/\tau}\,d\beta \tag{2.33}$$

for $\beta \in \mathbb{R}$, leading to symmetric Gamma random measures. Both the standard positive and symmetric Gamma random measures satisfy the bound given by (2.18), thus no compensation is required and $h \equiv 0$ and $\delta_h \equiv 0$.

The Gaussian kernels, Laplace, and exponential, given by equations (2.3), (2.4), and (2.5), (2.3), respectively, with $\omega \equiv (\chi, \lambda)$, are each in $\mathcal{G}$ for the stationary Gamma (positive and symmetric) random field where $\alpha(d\omega) = \alpha \, d\chi \, \pi(\lambda) \, d\lambda$, $\Omega = \mathbb{R} \times \mathbb{R}^+$ and $\pi(\lambda)$ is a density with respect to Lebesgue measure on $\mathbb{R}^+$. Means may be obtained directly from (2.30). In the case of the symmetric stationary Gamma random field, $\mathsf{E}[f(x)] = 0$ for any kernel that is in $L_1(\Omega, \alpha)$ and that is symmetric in x and χ, i.e. $g(x, \omega) = k(\lambda r(|x - \chi|))$ for some function $r(\cdot)$. This holds more generally for any Lévy measure $\nu(d\beta, d\omega)$ that is symmetric in β, as long as the expectation exists. Specific examples of covariance functions may be obtained from (2.31). For the Gaussian case, we obtain

$$\mathsf{C}ov\big[f(x_1), f(x_2)\big] = 2\alpha\tau^2 \int_{\mathbb{R}^+} \sqrt{\frac{\pi}{\lambda}} e^{-\frac{\lambda(x_1 - x_2)^2}{4}} \pi(\lambda) \, d\lambda$$

which, in the case that $\lambda \sim \mathsf{Ga}(a, b)$, simplifies to

$$\mathsf{C}ov\big[f(x_1), f(x_2)\big] = 2\alpha\tau^2 \sqrt{\pi} b \frac{\Gamma(a - 1/2)}{\Gamma(a)} \left[1 + \frac{(x_1 - x_2)^2}{4b}\right]^{1/2 - a}$$

for $a > 1/2$ and $b > 0$. Using the Gamma prior on the scale parameter, λ, with $a > 1$ and $b > 0$ the covariance function using the exponential kernel is

$$\begin{aligned} \mathsf{C}ov\big[f(x_1), f(x_2)\big] &= \alpha\tau^2 \int_{\mathbb{R}^+} \frac{1}{\lambda} e^{-\lambda|x_1 - x_2|} \pi(\lambda) d\lambda \\ &= \frac{\alpha\tau^2 b}{a - 1} \left[1 + \frac{|x_1 - x_2|}{b}\right]^{1 - a} \end{aligned}$$

and, similarly, the covariance function is

$$\begin{aligned}\mathsf{C}ov\big[f(x_1), f(x_2)\big] &= 2\alpha\tau^2 \int_{\mathbb{R}^+} \frac{1}{\lambda} e^{-\lambda|x_1-x_2|}\big(1+\lambda|x_1-x_2|\big)\pi(\lambda)d\lambda \\ &= 2\alpha\tau^2 \frac{a|x_1-x_2|+b}{(a-1)\left[1+\frac{|x_1-x_2|}{b}\right]^a}\end{aligned}$$

for the Laplace kernel.

The symmetric α-stable (SαS) for $0 < \alpha < 1$ is another example of a Lévy random field with infinite Lévy measure

$$\nu(d\beta, d\omega) = c_\alpha \gamma(d\omega)|\beta|^{-1-\alpha}\, d\beta$$

for some constant $c_\alpha > 0$ and sigma finite measure γ, giving $\mathcal{L}[A] \sim \mathsf{St}(\alpha, 0, \gamma(A), 0)$. The construction of S$\alpha$S random fields with $1 \leq \alpha < 2$ (including the Cauchy, with $\alpha = 1$) requires a compensation function. For the choice $\mathbf{1}_{|\beta|<1}\beta$, (or any h that is an odd function satisfying the condition (2.15)), however, the term δ_h is identically zero, so that no "drift" term is required in the measure $\mathcal{L}$. In the symmetric stable case, where $0 < \alpha < 2$, it is enough that $g \in L^\alpha(\Omega, d\gamma)$ so that the function $f(x)$ is well defined.

2.3 Approximating Kernel Integrals

If $\nu(\mathbb{R} \times \Omega)$ is finite the number $J = N(\mathbb{R} \times K)$ of support points in the random measure $\mathcal{L}$ will almost surely be finite and the ID random measure $\mathcal{L}$ may be constructed by generating the number of support point J, and given J, generate the random support points and jump heights of the distribution as follows:

$$\begin{aligned} J &\sim \mathsf{Po}\big(\nu(\mathbb{R}\times\Omega)\big) \\ (\beta_j,\omega_j)\mid J &\overset{iid}{\sim} \pi(\beta_j,\omega_j) \equiv \frac{\nu(d\beta_j,d\omega_j)}{\nu(\mathbb{R},\Omega)} \qquad \text{for } j=1,\ldots,J. \end{aligned}$$

In the case of infinite Lévy measure, J will be infinite almost surely, so that while the integrals $\mathcal{L}[g]$ will be well behaved, stochastic expansions with an infinite number of terms are not practical for simulation or posterior inference for f. However, when the integrability condition (2.18) holds, we can always approximate $\mathcal{L}$ and $\mathcal{L}[g]$ by choosing some small $\epsilon > 0$, and replacing $\mathbb{R}$ in Eqn. (2.25) by $[-\epsilon,\epsilon]^c$. Thus,

$$\mathcal{L}_\epsilon[g] \equiv \iint_{[-\epsilon,\epsilon]^c\times\Omega} g(\omega)\beta\, N(d\beta,d\omega) \overset{d}{=} \iint_{\mathbb{R}\times\Omega} g(\omega)\beta\, N_\epsilon(d\beta,d\omega) \tag{2.34}$$

where N_ϵ is a Poisson measure on $\mathbb{R}\times\Omega$ with intensity measure $\nu_\epsilon(d\beta,d\omega) \equiv \nu(d\beta,d\omega)\mathbf{1}_{\{|\beta|>\epsilon\}}$. Thus the Lévy random field $\mathcal{L}[g]$ may be approximated by a sequence of compound Poisson random fields, $\mathcal{L}_\epsilon[g]$ where $\mathcal{L}_\epsilon[g]$ converges in distribution to $\mathcal{L}[g]$ as $\epsilon \to 0$. This in fact gives us a simple way to specify the prior and simulate from $\mathcal{L}[g]$ (approximately), as for any fixed $\epsilon > 0$, $J \sim \mathsf{Po}\big(\nu_\epsilon(\mathbb{R}\times\Omega)\big)$ will be finite almost surely. As $\mathcal{L}_\epsilon$ has a finite Lévy measure, we can generate $\mathcal{L}_\epsilon$ as in the compound Poisson case,

$$J \mid \epsilon \;\sim\; \mathsf{Po}\big(\nu(-[\epsilon,\epsilon]^c\times\Omega)\big) \tag{2.35}$$

$$(\beta_j,\omega_j)\mid J,\,\epsilon \;\overset{iid}{\sim}\; \pi(\beta_j,\omega_j\mid\epsilon) \equiv \frac{\nu(d\beta_j,d\omega_j)\mathbf{1}_{\beta_j\in[-\epsilon,\epsilon]^c}}{\nu([-\epsilon,\epsilon]^c,\Omega)} \tag{2.36}$$

When ν satisfies the more general integrability condition, we may proceed similarly, but may be required to adjust for compensation. As $\nu([-\epsilon,\epsilon]^c,\Omega) < \infty$

for $\epsilon > 0$, $\mathcal{L}$ may be approximated based on the first two terms of Eqn. (2.24),

$$\mathcal{L}_\epsilon[g] = \iint_{[-\epsilon,\epsilon]^c \times \Omega} g(x,\omega)\beta \, N(d\beta, d\omega) - \iint_{([-1,-\epsilon] \cup [\epsilon,1]) \times \Omega} g(x,\omega)\beta \, \nu(d\beta, d\omega). \tag{2.37}$$

The Poisson integral may be simulated using Eqns. (2.35, 2.36), while the second term is a deterministic integral that may be evaluated either analytically or via simulation. In the case of the Cauchy random field or other SαS random fields, the second term is actually zero for $\epsilon > 0$, due to the symmetry in the Lévy measure, thus $\mathcal{L}[g]$ may be approximated directly by Eqns. (2.35, 2.36).

Rather than "truncating" the support of the Lévy measure, we note that any other sequence of finite measures $\nu_a(d\beta, d\omega)$ defined on $\mathbb{R} \times \Omega$ where $\lim_a \nu_a(d\beta, d\omega) \to \nu(d\beta, d\omega)$ could be used instead, and $\mathcal{L}_a[g]$ will also converge in distribution to $\mathcal{L}[g]$ due to the convergence of the characteristic functionals. In both cases, $\mathcal{L}[g]$ is approximated by a sequence of compound Poisson random fields. Unlike the approximation ν_ϵ, however, these may place positive probability on β_j in an ϵ neighborhood around zero. The approximation based on the truncated Lévy measure maintains the same relative prior density for $\beta_j \epsilon$, and sets to zero only the smallest coefficients. This focus on the large magnitude coefficients is desirable, as we are interested in potentially sparse expansions.

2.4 Sparse Lévy Adaptive Regression Kernel Models

Exploiting the Poisson construction of the Lévy random field $\mathcal{L}[g]$ for Lévy measures satisfying Eqn. (2.18), we may express the LARK model given by Eqn. (2.1) and Eqns. (2.35)–(2.36) in a hierarchical fashion. For the remainder of the paper, we consider the specific case of the symmetric stationary Gamma model; other Lévy measures may be implemented similarly using Eqns. (2.35)–(2.37). When compensation is required, we may use the same hierarchical model, but will need to include a deterministic offset to $f(x)$ corresponding to the second term in Eqn. (2.37). However, as noted in the previous section, for the Cauchy and other SαS random fields, this offset will be zero, so that we may proceed directly with the hierarchical specification given below with the appropriate changes in measures.

In the symmetric stationary Gamma model in Eqn. (2.32), the Lévy measure is decomposed into a product measure, so that in the approximate compound Poisson formulation of the problem, β_j and the components of ω_j are independent. The normalizing constant for the distribution of β and ω as well as the mean of the Poisson distribution for J depend on $\nu([-\epsilon,\epsilon]^c,\Omega) = \alpha(|\mathcal{X}| \times \Omega)E_1(\epsilon/\tau)$, where $E_1(x) \equiv \int_x^\infty u^{-1}e^{-u}du$ denotes the exponential integral function. In this parameterization, τ controls the size of the jumps β, but because of the truncation, also affects the number of jumps. Because a priori the appropriate scaling of the jumps is unknown, this makes selection of an objective value of ϵ more difficult. We also would prefer to have the number of jumps be independent from their size in the approximate model, as they are in the limiting case. To resolve these issues, we work instead with the equivalent random field (in that $\mathcal{L}[g]$ has the same characteristic functional) expressed as

$$\mathcal{L}[g] = \tau \iint_{\mathbb{R}\times\Omega} g(x,\omega)\beta N(d\beta, d\omega)$$

where now $\mathsf{E}[N(d\beta,d\omega)] = \nu(d\beta,d\omega) = \alpha(d\omega)|\beta|^{-1}e^{-|\beta|}$. In this version, the jumps β are scale free, and taking $\epsilon = 0.01$ (for example) makes sense across problems of different scales. To complete the choice of Lévy measure, we take $\alpha(d\omega)$ to be a product measure on $\mathcal{X} \times \mathbb{R}^+$ where $\alpha(d\omega) \equiv \alpha d\chi\, \pi_\lambda(d\lambda)$, where χ is uniform on $\mathcal{X}$ and λ has a Gamma distribution on $\mathbb{R}^+$. With this Lévy measure, the hierarchical model is expressed as

$$\begin{aligned}
Y_i \mid \boldsymbol{\omega}, \boldsymbol{\beta}, J, \epsilon &\overset{iid}{\sim} \mathsf{N}(f(x_i), \sigma^2) \quad \text{where} && (2.38)\\
f(x_i) &\equiv \beta_0 + \tau \sum_{j=1}^{J} g(x_i, \omega_j)\beta_j && (2.39)\\
(\beta_j, \omega_j) \mid J, \epsilon &\overset{iid}{\sim} \pi(\beta_j, \omega_j \mid \epsilon) \equiv \frac{\nu_\epsilon(d\beta_j, d\omega_j)}{\nu_\epsilon(\mathbb{R}, \Omega)} && (2.40)\\
J \mid \epsilon &\sim \mathsf{Po}(\nu_\epsilon(\mathbb{R}, \Omega)) \quad \text{where } \nu_\epsilon(\mathbb{R}, \Omega) = 2\alpha|\mathcal{X}|E_1(\epsilon) && (2.41)
\end{aligned}$$

where $\beta_j, \chi_j, \lambda_j$ are independent and identically distributed from the following distributions:

$$\begin{aligned}
\pi_\beta(\beta_j|\epsilon) &= \frac{|\beta_j|^{-1}e^{-|\beta_j|}}{2E_1(\epsilon)}, \qquad \beta_j \in [-\epsilon, \epsilon]^c && (2.42)\\
\pi_\chi(\chi_j) &= 1/|\mathcal{X}|, \qquad \chi_j \in \mathcal{X} && (2.43)\\
\lambda_j &\sim \mathsf{Ga}(a_\lambda, b_\lambda), \qquad \lambda_j \in \mathbb{R}^+. && (2.44)
\end{aligned}$$

For the scalar parameters $(\alpha, \tau, \beta_0, \sigma^2)$, we use the following distributions

$$\begin{aligned}
\alpha &\sim \mathsf{Ga}(a_\alpha, b_\alpha), \qquad \alpha \in \mathbb{R}^+\\
\tau &\sim \mathsf{Ga}(a_\tau, b_\tau), \qquad \tau \in \mathbb{R}^+\\
\pi(\beta_0, \sigma^2) &\propto 1/\sigma^2
\end{aligned}$$

We place a Gamma prior on the hyperparameter α in the Gamma Lévy measure, leading to Negative-Binomial distribution for J given ϵ. This corresponds to a mixture of Lévy processes and provides more robustness to the choice of hyperparameters for the distribution of J for a given ϵ. We select the values a_α and

b_α based on quantiles of J, for example by fixing a prior probability that $J = 0$ and fixing the 95th percentile of J. The Gamma prior on τ controls the overall jump size. Finally, in the absence of prior information, we adopt the independent Jefferey's prior for (β_0, σ^2),

$$\pi(\beta_0, \sigma^2) \propto \frac{1}{\sigma^2} \tag{2.45}$$

as both parameters are present in all "models".

2.4.1 Posterior Inference

Given observations $\mathbf{Y}$, the joint posterior distribution of all unknowns under the LARK model is

$$\begin{aligned} p(\boldsymbol{\beta}, \boldsymbol{\omega}, J, \alpha, \tau, \beta_0, \sigma^2 \mid \mathbf{Y}) \quad \propto \quad & \left(\frac{1}{\sigma^2}\right)^{-\frac{n}{2}-1} \prod_{i=1}^{n} \exp\left\{ -\frac{1}{2\sigma^2} \left(Y_i - \beta_0 - \tau \sum_j g(x_i, \omega_j)\beta_j \right)^2 \right\} \\ & \frac{\exp(-\nu_\epsilon(\mathbb{R} \times \Omega))}{J!} \left\{ \prod_{j=1}^{J} \nu_\epsilon(d\beta_j, d\omega_j) \right\} \pi_\alpha(\alpha)\pi_\tau(\tau). \end{aligned} \tag{2.46}$$

While both β_0 and σ^2 may be integrated out analytically, the posterior distribution or full conditional distributions of the other components do not exist in closed form. As we have fixed dimensional parameters $(\beta_0, \alpha, \sigma^2)$ and varying dimensional parameters (β_j, ω_j), $j = 1, \cdots, J$, the dimension of the parameter space is not fixed and trans-dimensional Markov chain Monte Carlo (MCMC), such as a reversible jump MCMC algorithm (Green, 1995; Wolpert *et al.*, 2003; DiMatteo *et al.*, 2001) must be used to provide samples from Eqn. (2.46) for posterior inference.

The RJ-MCMC procedure for sampling varying dimensional parameters typically involves three types of moves: Birth, Death and Update. A birth step involves adding a new point (β^*, ω^*) to $((\beta_1, \omega_1), \cdots, (\beta_J, \omega_J))$ and increasing J by one; a death step involves selecting an index $j \in \{1, \cdots, J\}$ and removing (β_j, ω_j) from $((\beta_1, \omega_1), \cdots, (\beta_J, \omega_J))$ and decreasing J by one; an update step involves selecting a point (β_j, ω_j) and updating its values (β_j^*, ω_j^*). A Metropolis-Hastings algorithm is used to sample the fixed dimensional parameters. Details of the MCMC algorithm may be found in the Appendix.

We now turn our attention to simulated and real examples to illustrate the performance of the LARK models in practice.

2.5 Examples and Illustrations

In this section, we conducted a simulated study to compare the performance of the LARK model to other nonparametric methods. An application to a motorcycle crash experiment data is then presented to illustrate the methodology with unequally spaced data.

2.5.1 Simulation Study

We carried out a simulation study for the model on the four test functions commonly used in the wavelet literature: Blocks, Bumps, Doppler and Heavysine (Donoho and Johnstone, 1994). For each function, data were generated by adding independent Gaussian random noise $\mathsf{N}(0, \sigma^2)$ to the true target function f at 1024 equally spaced points on $[0, 10]$. The value of σ was chosen such that the root signal-to-noise ratio was seven, where (RSNR $\equiv \sqrt{\int (f(x) - f\bar(x))\, dx}/\sigma$ and $\bar{f} \equiv \int f(x)\, dx$, as in Abramovich *et al.* (1998). For each function we generate 100 replicate data sets to evaluate the performance of LARK and other methods.

According to the features of the test functions, we use a different kernel for each test function, as indicated in Table 2.1. In Section 2.4, we described the prior distributions for the LARK model in the symmetric stationary Gamma random field. In order to implement the model, we need to specify several hyperparameters. Sensitivity analysis shows that within a wide range, the results are insensitive to the choice of ϵ; for this analysis we have used $\epsilon = 0.5$. We use a $\mathsf{Ga}(a_\lambda, b_\lambda)$ distribution (with mean a_λ/b_λ) as the prior distribution for λ_j. Because each kernel has its own scale parameter, the choice of a_λ and b_λ is important; this is similar to the issue of bandwidth selection problem in other kernel smoothing methods. A large λ is necessary to fit a very "spiky" part of a curve, while small values of λ are needed to fit smoother part of a curve, thus the prior distribution of λ needs to support an adequate range of values in order to fit a spatially inhomogeneous curve. For the model to be well defined with a finite covariance function (for the Gaussian kernel), we need $a_\lambda > 1$. By specifying 25% and 75% quantiles (for example) of $\mathsf{Ga}(a_\lambda, b_\lambda)$ based on a range of widths or support of the kernel, we can solve for the corresponding a_λ and b_λ. In this example, a $\mathsf{Ga}(2,\, 0.1)$ (with mean being 20) is used for λ_j. Lastly, we need to specify the prior distributions for α and τ. Fixing α a priori leads to a Poisson prior $\mathsf{Po}(2\alpha E_1(\epsilon)\mathcal{X})$ for J, which can be too concentrated because of the mean/variance relationship. Taking a Gamma prior distribution for α and integrating over α, leads to a Negative Binomial prior distribution for J, which is more robust to prior misspecification J, as the Negative Binomial has a larger variance than Poisson distribution. A Gamma distribution, $\mathsf{Ga}(a_\alpha, b_\alpha)$, is used as the prior for α, which controls the number of kernels in a unit interval. A priori, the expected number of kernels per unit interval is $2\alpha E_1(\epsilon)$. The smaller α is, the larger models are penalized. One way to

Test Function	Kernel $g(x_i;\chi_j,\lambda_j)$
blocks	$I_{\{0<(x_i-\chi_j)<\lambda_j\}}$
bumps	$e^{-\lambda_j\|x_i-\chi_j\|}$
doppler	$e^{-0.5\lambda_j^2(x_i-\chi_j)^2}$
heavysine	$e^{-0.5\lambda_j^2(x_i-\chi_j)^2} I_{\{\|x_i-\chi_j\|<2.0\}}$

Table 2.1: Kernel functions used for four test functions

choose an appropriate a_α and b_α is through specifying 25% and 75% quantiles of J and solving for the corresponding a_α and b_α. We choose $a_\alpha = 2$ and $b_\alpha = 0.5$, which in in this example corresponds to 6 and 20 for 25% and 75% percentiles, respectively, of J a priori. The height of any function at $x = \chi_j$ is $\tau\beta_j$. Recall that β_j is scale free such that scale parameter τ controls the overall size of jump. To complete our specification, a $\mathsf{Ga}(2.0, 0.2)$ is used as the prior distribution for τ to cover the observed range of the data.

We compare LARK with a number of wavelet-based methods. The Translational-Invariant Marginal Maximum Likelihood (TI-MML) approach of Johnstone and Silverman (2005) is one of the best wavelet methods currently available for inhomogeneous function estimation using an overcomplete representation. In addition to TI-MML using the Laplace prior, we compare LARK with a number of other methods previously used for these functions, including BayesThresh (Abramovich *et al.*, 1998), GlobalSure, Cross-validation (Nason, 1996) and False Discovery Rate (Abramovich and Benjamini, 1996).

The performance of each method was measured by its average mean square error (AMSE), which is defined as the average of MSE $= n^{-1}\sum_{i=1}^{n}(\widehat{f(x_i)} - f(x_i))^2$, over the 100 replicated simulations. Overall, the performance of the LARK model is excellent (Table 2.2). In terms of AMSE, LARK outperformed all methods. In

Method	Blocks	Bumps	HeavySine	Doppler
LARK	0.027	0.092	0.041	0.117
Laplace TI-MML	0.096	0.307	0.118	0.202
BayesThresh	0.38	0.45	0.10	0.16
Cross-validation	0.41	0.46	0.10	0.21
GlobalSure	0.42	0.48	0.12	0.21
False Discovery Rate	0.96	1.23	0.12	0.39

Table 2.2: Average over 100 replications of mean square errors of the four test functions for different models and methods. Laplace TI-MML the Translational Invariant Maximum Marginal Likelihood approach using a Laplace prior from Johnstone and Silverman (2005). The results for BayesThresh, Cross-validation, GlobalSure and False Discovery Rate are from Abramovich *et al.* (1998).

Fig. (2.1), we compare the reconstruction between TI-MML and LARK. The left and the right columns are the constructions using Laplace TI-MML and LARK models, respectively. The figures show that LARK generally gives a better visual reconstruction in the sense that the fit is smoother and less noisy than the Laplace. The adaptive smoothing provided by LARK preserves local features such as peaks, while eliminating variation due to noise in other regions. Another important advantage the LARK model offers is the sparsity. On average, it takes the LARK model 14.0, 15.0, 20.9 and 32.4 kernels to represent blocks, bumps, heavysine and doppler functions, while TI-MML uses 284.8, 570.4, 0.6 and 147.9 non-zero coefficients to reconstruct the same functions (this is in addition to the 1024 non-zero coefficients for the scaling function, which are not thresholded in the TI-MML approach). Overall, the LARK model provides excellent MSE performance while identifying sparse representations.

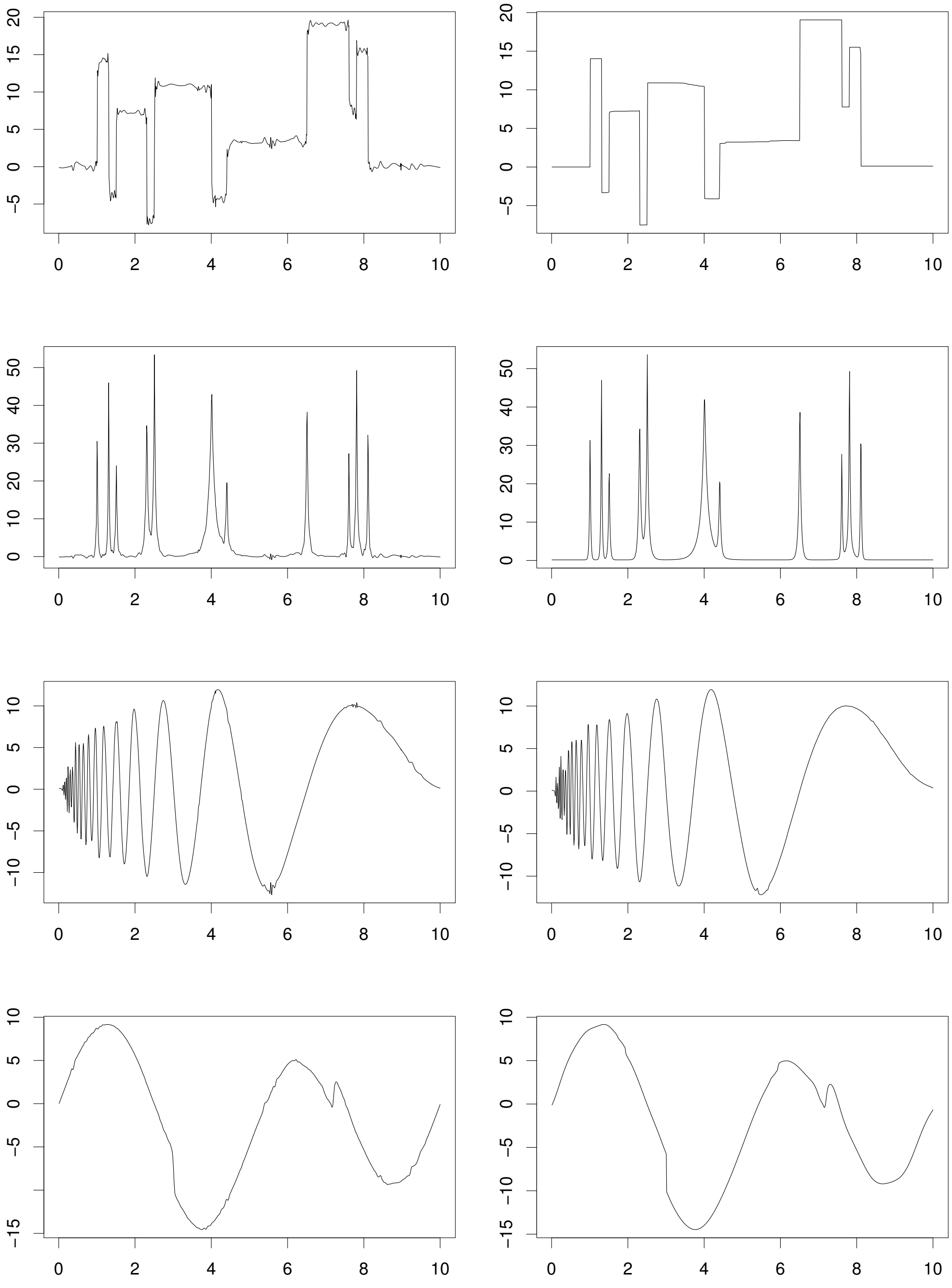

Figure 2.1: Comparison of fitted functions using TI-MML Laplace (Johnstone and Silverman, 2005) (left column) and Lévy Adaptive Regression Kernels (LARK) (right column) for the four test functions. From row one to row four, the test functions are Blocks, Bumps, Doppler and Heavysine, respectively.

2.5.2 Examples: Motorcycle Crash Data

To further illustrate the method, we consider the motorcycle crash experiment used by Silverman(1985) (Fig. (2.2)). There are 133 observations, however, the time points of the observations are not equally spaced and there are repeated observations at some time points. While, it is clear from the figure that the variance of accelerations at different time points is different, we are not going to address this issue here and for demonstrative purposes, assume an independent normal error model as the focus of this paper is to discern the general shape of the curve and concentrate on modeling inhomogeneous features of the curve. An interesting extension of the LARK model, but beyond the scope of this paper, is to simultaneously model the error process, in addition to the mean function.

In the simulation study, the power ρ in the kernel $g(x; \chi, \lambda, \rho) = \exp\{-\lambda|x - \chi|^{\rho}\}$ is fixed and chosen based on the characteristics of the test functions. Rather than making an ad hoc choice of ρ, the Bayesian paradigm permits treating ρ as an unknown parameter and making inference regarding it from the data. For the motorcycle example, we assume a common, but unknown ρ for all kernels. A relatively concentrated Gamma prior $\mathsf{Ga}(2.0, 0.75)$ is assumed for ρ, which includes the Laplace and Gaussian kernels. We summarize the results in Fig. (2.2). The solid line is the fitted mean and the dotted lines are 5% and 95% pointwise credible intervals for the fitted mean respectively. We see clearly from the figure that the fitted mean captures the general pattern of the data very well, with minimal boundary effects. The model is parsimonious in the sense we only need 4 kernels on average to fit the data. In Fig. (2.3), we show the histogram of posterior draws of ρ from the MCMC output, and overlay the histogram with the prior density of ρ. The posterior mean of ρ is around 3, with most of the mass of the posterior distribution above 2 (the Gaussian kernel).

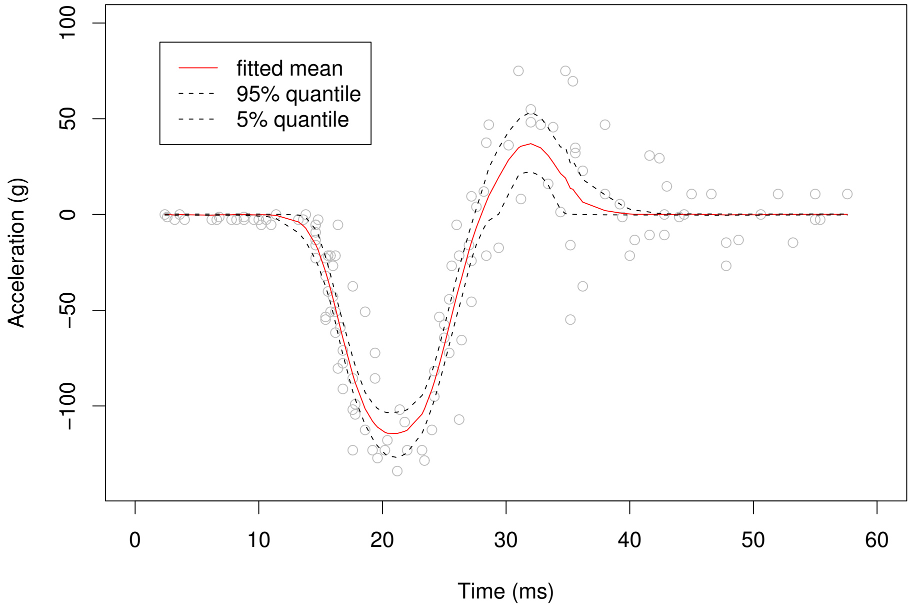

Figure 2.2: Results of the LARK model for the motorcycle crash data: the circles represent the observations; the solid line is the posterior mean and the dotted lines are 5% and 95% Bayesian credible intervals for the mean function.

2.6 Discussion

In this paper, we have developed a fully Bayesian adaptive kernel method, LARK, for nonparametric function estimation. The LARK model is based on a stochastic expansion of functions in an overcomplete dictionary, which may be formulated as a stochastic integration problem with a random measure. The unknown function may be approximated as a finite sum of kernel functions at arbitrary locations where the number of kernels is a free parameter. The kernel parameters are location-specific and thus are adaptively updated given the data. The adapt-

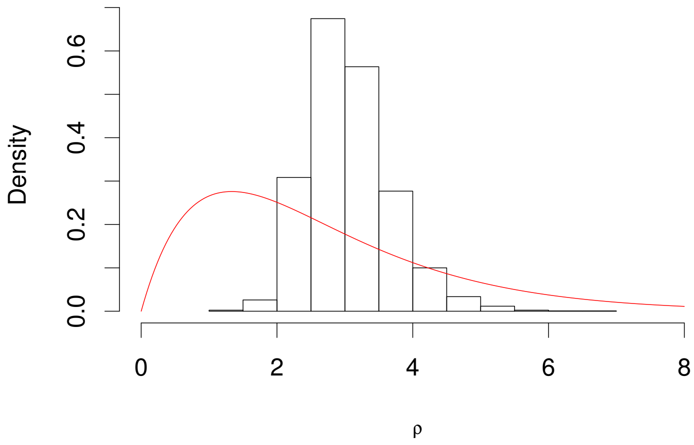

Figure 2.3: Histogram of posterior samples of the kernel power parameter ρ: the solid line is the prior density function for ρ.

ability of the kernels is especially useful for modeling "spatially" inhomogeneous functions. Unlike many wavelet based methods, there is no requirement that the data are equally spaced. As with wavelets, the adaptive smoothing using LARK preserves local features such as high peaks and jumps. The RJ-MCMC algorithm developed for fitting LARK models provides an automatic search mechanism for finding sparse representations of a function.

2.6.1 Relation to Other Sparse Regression Methods

Bayesian shrinkage and variable selection have been used successfully to achieve sparsity in standard finite dimensional regression formulations, and provide a canonical framework for sparse representations in nonparametric models (see Clyde and George (2004) for an overview). One of the first examples, Smith and Kohn (1996) adopted the conjugate Stochastic Search Variable Selection (SSVS) framework of George and McCulloch (1997) to nonparametric cubic spline regression. In the case of a cubic spline model in one dimension,

$$f(x) = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \sum_{j=1}^{J} (x - \tilde{x}_j)_+^3 \beta_{j+3}$$

where $(z)_+ = \max(0, z)$ and $\tilde{x}_1, \ldots, \tilde{x}_J$ are potential knot locations, typically chosen at some subset of observed quantiles of the data. The model with all knots can be re-expressed in matrix form as a linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathsf{N}_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$$

with fixed design matrix $\mathbf{X}$ derived from the polynomial and spline. Bayesian variable selection techniques are then used to identify the basis vectors or knot locations with non-zero coefficients. This is achieved by introducing a J dimensional vector $\boldsymbol{\gamma} = (\gamma_1, \ldots \gamma_J)^T$ of binary random variables that indicate which of the spline coefficients are non-zero. Under a particular $\boldsymbol{\gamma}$, the model for the data is

$$\mathbf{Y} = \mathbf{X}_{\boldsymbol{\gamma}} \boldsymbol{\beta}_{\boldsymbol{\gamma}} + \boldsymbol{\epsilon}$$

where $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ are the nonzero elements of $\boldsymbol{\beta}$ and $\mathbf{X}_{\boldsymbol{\gamma}}$ is the design matrix with columns of $\mathbf{X}$ corresponding to those elements of $\boldsymbol{\gamma}$ that are equal to one. This framework provides the canonical model for soft and hard thresholding of wavelet coefficients Abramovich *et al.* (1998); Clyde *et al.* (1998). In overcomplete Gabor frames, Wolfe *et al.* (2004) use the SSVS framework, but allow the number of frame of vectors J to exceed n.

Many examples in the literature use independent Bernoulli priors on the inclusion indicators, $\gamma_j \overset{iid}{\sim} \mathsf{Ber}(\pi)$, in which case, the induced prior on the model dimension $J = \sum_j \gamma_j$ is $\mathsf{Bin}(J, \pi)$. Letting J go to infinity, such that $J\pi$ converges to a constant μ, we arrive at a Poisson distribution for the number of knots, with expected number of knots μ, as in the LARK framework. DiMatteo *et al.* (2001) adopt Poisson (or truncated Poisson) prior distributions on the number of knots and allow random knot locations by using a uniform prior on knot locations. The free-knot framework allows more flexibility over models with knots restricted to certain quantiles, as both the number of knots and their locations may adapt based on the observed data, and is the closest to the LARK framework.

The main difference in the LARK formulation and the above approaches concerns prior specifications on the coefficients in the expansion f and choice of dictionary elements. Letting $\mathbf{X}_{\boldsymbol{\gamma}}$ denote the $k_{\boldsymbol{\gamma}} \times n$ design matrix conditional on the knot locations with non-zero coefficients $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$, Smith and Kohn (1996) and DiMatteo *et al.* (2001) use Zellner's (1986) g-prior

$$\boldsymbol{\beta}_{\boldsymbol{\gamma}} \mid \sigma^2 \sim \mathsf{N}_{k_{\boldsymbol{\gamma}}}(\mathbf{0}, g\sigma^2(\mathbf{X}_{\boldsymbol{\gamma}}^T \mathbf{X}_{\boldsymbol{\gamma}})^{-1})$$

with hyperparameter g. This prior has been widely adapted for variable selection because of its computational tractability and that it requires specification of only one hyperparameter g. DiMatteo *et al.* (2001) take $g = n$ as in the unit-information prior of Kass and Wasserman (1995), while Smith and Kohn (1996) use g in the range of 10 to 1000 after standardizing $\mathbf{X}_{\boldsymbol{\gamma}}$. While the g-prior formulation permits efficient MCMC computation, a drawback is that model comparisons based on g-priors have several undesirable inconsistency properties as discussed in Berger and Pericchi (2001); Berger *et al.* (2003); Liang *et al.* (2005). Paciorek (2006) shows that the use of the unit information prior, a special case of the g-prior with $g = n$, may lead to serious distortion of the posterior distributions in complex, high dimensional problems, like the free-knot spline model.

An alternative framework is based on independent prior distributions. Clyde and George (2000); Tipping (2001); Johnstone and Silverman (2004); Wolfe *et al.* (2004), for example, achieve sparsity using independent heavy-tailed priors derived as scale mixtures of normals. The LARK framework, which uses independent heavy tailed priors on the coefficients in the expansion, however, does not a priori restrict support for ω to a lattice as in wavelets (decimated and non-decimated) and Gabor frames, and thus, as in free-knot splines, allows greater adaptation. Abramovich *et al.* (2000) also propose stochastic expansions for overcomplete wavelet dictionaries allowing arbitrary locations and scales. They use normal priors on the coefficients in the expansion, but with restrictions on the variances to ensure that the random functions are in a pre-specified Besov space with probability one. While this model is also a special case of the LARK model, in practice, a priori determination of the parameters of the Besov space is difficult. Secondly, the heavy-tailed priors induced by the two-sided Gamma process or Cauchy process provide additional robustness over normal priors. Viewing the log of Eqn. (2.46) as a penalized log-likelihood, the approximate Lévy measure may be seen as inducing a sparsity generating penalty on the addition of terms to the function $f(x)$, similar (or stronger in fact) to the L_1 penalties of the LASSO (Tibshirani, 1996).

The Gamma and Cauchy processes are both examples of non-finite Lévy measures, and in order to restrict the expansion to a finite number of terms, we must restrict $|\beta| > \epsilon$. This may be related to the idea of practical significance in the non-conjugate version of the SSVS algorithm George and McCulloch (1993); Chipman *et al.* (1997) where the prior distribution on β was a mixture of two normal distributions; one fairly dispersed and the other concentrated around zero. The variance of the concentrated distribution was chosen to reflect values of the coefficient that for all practical purposes indicated that the variable could be dropped from the model. If the coefficient had a high posterior probability of coming from the concentrated component of the mixture, then that variable would be dropped. Elicitation of the choice of ϵ in the LARK framework can be based on practical significance for estimating f in the presence of noise.

2.6.2 Extensions

Besides the model presented in Section 2.4, there are a number of possible extensions. Although we have introduced the LARK model in the context of functions on a one dimensional space, the method may be readily adapted to model higher dimensional data. Current applications include extending the LARK framework for multiple time series modeling and spatio-temporal modeling and non-Gaussian likelihoods. While we have focused on normal error models, the LARK model provides a flexible way to model non-Gaussian data as well, with a nominal change in the RJ-MCMC algorithm.

In much higher dimensional problems, computational complexity increases and methods to improve the MCMC algorithms will be of importance. In the current computational algorithm, we start the MCMC chain at the initial values which are sampled from priors. Choosing good starting values, for example, starting with estimates based on kernels at observed quantiles of the data, deserves further investigation, since it can drastically speed up the convergence. Other improvements include proposals for the addition of a new kernel. In the current BIRTH step, we propose a new kernel location at a random location generated from the prior. Because of the local nature of the kernels, one possibility is to bias proposed locations based on residuals of the current model fit, thus making it more likely that we add a component in any area where the model exhibits more lack of fit.

The kernels used in this paper are from symmetric exponential family distributions. It will be interesting to consider a wider class of kernels such as wavelets and splines. Although we have used symmetrical kernel functions in the examples, this is not a restriction for the LARK model. In Section 2.5, we used different kernel functions for different test functions. Automatic selection of kernel functions may be handled by adding another level of adaptability in the model, such as a "mark". For example, we can introduce categorical variables $(\gamma_1, \cdots, \gamma_J)$, where $\gamma_j \in \{1, \cdots, K\}$, indicating kernels from K different families. We can combine all K kernel functions this way and a fully Bayesian approach allows the computational algorithm to choose the optimal kernel at each location based on the data. Finally, we have used a symmetric Gamma random field to construct a joint prior distribution for the model parameters. Other prior processes such as Cauchy process and Stable process are also good candidates, thus it will be of interest to compare the performance (empirical and theoretical) of different ID random measures.

Chapter 3

An Adaptive Kernel Smoothing Approach to Modeling Multivariate Time Series of Air Pollutants

3.1 Introduction

PM stands for particulate matter and is also known as particle pollution. PM is a complex mixture of extremely small particles and liquid droplets. It is one of the six criteria pollutants, which consist of Ozone (O_3), Particulate Matter (PM), Carbon Monoxide (CO), Sulfur Dioxide (SO_2), Nitrogen Oxides (NO_x) and Lead (P_b). PM has primary and secondary sources depending on their origins. Primary PM is emitted directly into the air whereas secondary PM is formed in the atmosphere through chemical and physical conversion of gaseous precursors. In typical urban areas, two broad sets of source categories exist: one is combustion sources, such as automobile exhaust and emissions from power plants; the other is mechanical forces such as prevailing wind and vehicle traffic. The Environmental Protection Agency (EPA) categorizes PM by its size: one is PM_{10} (coarse particles that can usually be found near roadways and dusty industries) that have diameters range from 2.5 to 10 micrometers; the other is $\mathrm{PM}_{2.5}$ (fine particles that can usually be found in smoke and haze) that have diameters smaller than 2.5 micrometers.

According to EPA, particles that are 10 micrometers in diameter or smaller are particularly harmful because once inhaled, these particles can have serious adverse effects on the health of heart and lungs. PM exposure can affect breathing, aggravate existing respiratory and cardiovascular disease and damage the body's immune system. A comprehensive report of the particulate matter can be found by EPA (2004). For the 20 largest U.S. cities, Dominici *et al.* (2000) found that a $10\mu g/m^3$ increase in PM_{10} on the current day was associated with a 0.05 to 0.92 percent increasing elderly mortality. Review articles on the epidemiological studies of the health effects of PM can be found by Thurston (1996), Dockery and Pope (1994) and Bell *et al.* (2004).

Developing sensible time series models for PM is essential for studying its health effects. It is well known that meteorological conditions have strong influence on ambient pollutant concentrations. For example, in a dessert area, strong prevailing wind can bring sand storm and raise PM level significantly. Cold winter temperatures usually result in higher PM concentrations due to increased energy use. PM and other major air pollutants in urban area have common sources and their concentrations may be correlated. For example, combustion of sulfur-containing fuels is common source for both PM and sulfur dioxide. Studies have associated other criteria pollutants such as CO and O_3 with adverse health effects. Thus, developing multivariate time series models for ambient pollutants that incorporate meteorological covariates is essential.

In this paper, we develop a class of time series models that are based on adaptive kernel smoothing. The adaptive kernel approach we will describe serves two purposes: the first is to offer a satisfying time series model that can capture the stylized features of air pollutants data; the second is to build a model that has good predictive power. We start with a univariate time series model for hourly PM_{10} data and then extend it for a bivariate time series model for hourly PM_{10} and CO data.

3.2 Univariate Time Series Modeling for Hourly PM_{10} Data

In this section a time series model based on adaptive kernel smoothing method is developed and applied to model PM_{10} time series data.

3.2.1 The Model

The data we use are from Maricopa county in Arizona. Particulate matter is a year round problem in Maricopa county and in 1997 the EPA reclassified the county's non-attainment area as "serious". For illustrative purpose, 30 consecutive days of hourly PM_{10} data are selected from one monitor starting on April 8, 1998. The data are displayed in Fig. (3.1). The 30 day hourly PM_{10} shows both slowly varying time trend and sharply falling local peaks. This stylized feature makes it difficult to apply classic ARMA type of time series models. Since PM_{10} concentration is non-negative, a log transformation of the data is needed to apply a Gaussian time series model. However, a log transformation is not a particularly good choice since it removes the high peaks which are one of the most important features to model. The non-stationary, non-Gaussian features make the adaptive kernel method (*LARK*) developed by Tu *et al.* (2005) a suitable building block for this

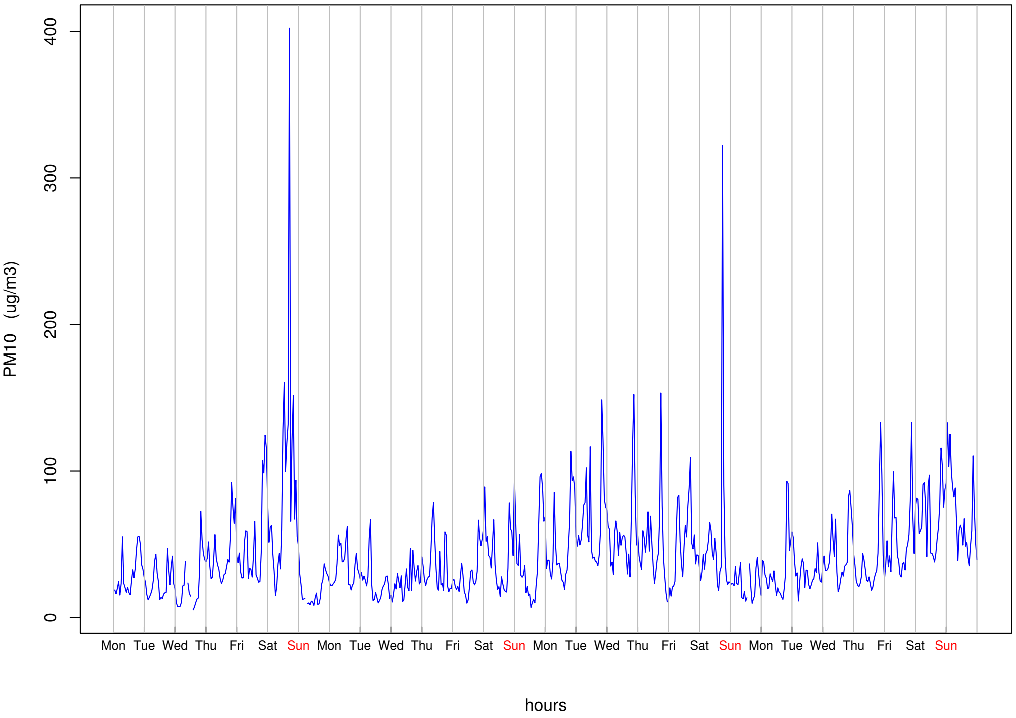

Figure 3.1: Time series plot of 30 days of hourly PM_{10} concentration measurements collected at one monitor in Maricopa county, Arizona. The starting day is April 8, 1998.

type of data sets. We assume a time series of the form:

$$Y(t) = \mu(t) + \epsilon(t), \quad \mu(t) > 0\,, \tag{3.1}$$

where $Y(t)$ denotes the observation at time t, $\mu(t)$ is the mean process and $\epsilon(t)$ is the measurement error process. In this paper, $\epsilon(t)$ is considered to be a Gaussian white noise process and the focus is placed on modeling time trend $\mu(t)$. Exploratory analysis shows daily pattern embedded in the data, thus the mean process $\mu(t)$ is modeled by decomposing it into aperiodic and daily periodic parts:

$$\mu(t) = b_0 + b_1 \int_{\Theta_1 \cup \Theta_2} k(t;\theta)\ L(d\theta)\,, \tag{3.2}$$

where the underlying non-stationary Lévy random field $L(d\theta)$ defined on the disjoint subspaces Θ_1 (the aperiodic part) and Θ_2 (the daily periodic part) are independent. Baseline level is modeled via b_0, and b_1 is used to make the latent process $L(d\theta)$ unit free. Weekly, monthly or any other seasonal behavior can be modeled using this approach. Let $\Theta \equiv \Theta_1 \cup \Theta_2 = \{[0, T_1] \times \mathbb{R}^+ \times \{1\}\} \cup \{[0, T_2] \times \mathbb{R}^+ \times \{2\}\}$ for $T_2 = 24$ and $T_1 > 24$. Each point $\theta \in \Theta$ has three components, $\theta = (\tau, \lambda, a)$ with $\tau \in [0, T_1] \cup [0, T_2]$, $\lambda \in \mathcal{R}^+$ and $a \in \{1, 2\}$. With Poisson construction of Lévy random field (For more details, please refer to Jacod and Shiryaev (1987), Sato (1999), or Wolpert and Taqqu (2005).), we can represent Eqn. (3.2) by a Poisson stochastic integral:

$$\mu(t) = b_0 + b_1 \iint_{\mathbb{R}^+ \times \Theta} u \, k(t; \theta) N(du, d\theta) \,, \tag{3.3}$$

where $N(du, d\theta)$ defines a Poisson random measure on the space $\mathbb{R}^+ \times \Theta$ with mean measure:

$$\mathbb{E}[N(du, d\theta)] \equiv \nu_\epsilon(du, d\theta) \equiv (\alpha_1 I_{\theta \in \Theta_1}(\theta) + \alpha_2 I_{\theta \in \Theta_2}(\theta)) u^{-1} e^{-u} I_{[\epsilon, \infty)}(u) \, du \pi(d\theta) \,.$$

The positive measure $\nu_\epsilon(du, d\theta)$ is usually called Lévy measure, and $\pi(d\theta)$ is a finite measure defined on Θ. For any $\epsilon > 0$, it can be shown that $\iint_{\mathbb{R}^+ \times \Theta} \nu_\epsilon(du, d\theta) < \infty$. An equivalent representation of Eqn. (3.3) is:

$$\mu(t) = b_0 + b_1 \sum_{j=1}^{J} u_j k(t; \tau_j, \lambda_j, a_j) \,, \tag{3.4}$$

where the smoothing kernel is defined as:

$$k(t; \tau_j, \lambda_j, a_j) = \begin{cases} e^{-\lambda_j |t - \tau_j|} & a_j = 1 \\ \frac{e^{-\lambda_j \left[(t-\tau_j) \mod 24\right]} \left(1 + e^{-2\lambda_j \left[12 - (t-\tau_j) \mod 24\right]}\right)}{1 - e^{-24\lambda_j}} & a_j = 2 \,. \end{cases} \tag{3.5}$$

The derivation of the kernel for the periodic jumps ($a_j = 2$) is detailed in the appendix B.1. The representation of $\mu(t)$ specified by Eqn. (3.4) and Eqn. (3.5) has a natural interpretation: the mean process $\mu(t)$ is driven by J latent jumps with magnitudes $u_1, \cdots, u_J$ at time $\tau_1, \cdots, \tau_J$, where J is a free parameter. Particulate matter includes dust, dirt, soot, smoke and liquid droplets directly emitted into the air by sources such as factories, power plants, cars, construction activity, fires and natural windblown dust. A latent jump j can be interpreted as a point source emitted and peaks with magnitude u_j at time τ_j and dissipates exponentially with rate λ_j. A categorical variable a_j is introduced to label each jump j. A jump j with $a_j = 1$ is an aperiodic jump, and it only occurs once on $[0, T_1]$; a jump j with $a_j = 2$ is a daily periodic jump, and it recurs at the same time τ_j everyday. Periodic jumps are used to model cyclic pattern of the data.

Daily meteorological data are available in the same area where the monitors are located. Incorporating meteorological covariates into the model helps build more sensible model with better predictive power. We introduce daily intensity factor $d_1, \cdots, d_{\lceil T/24 \rceil}$, which are modeled by meteorological variables $\mathbf{X}$ as follows:

$$d_l \sim \mathsf{LN}(\mathbf{X}_l' \boldsymbol{\gamma}, \sigma_d^2), \quad l = 1, 2, \cdots, \lceil T/24 \rceil \,.$$

The meteorological covariates matrix $\mathbf{X}$ are orthogonalized daily temperature, wind speed, their interaction term and their the quadratic terms. Precipitation is another important meteorological factor that influences PM_{10} level. We did not include it in the model due to the fact that precipitation is negligible in April and May in Maricopa, Arizona. The daily factors $\mathbf{d}$ serve as scale parameters for daily periodic jumps, which can inflate or deflate the magnitude of a daily jump, reflecting the belief that the meteorological factors are one of the driving forces that influence air pollutants level.

The Bayesian paradigm is chosen for inference and the final model can be represented in the following hierarchical fashion:

$$\begin{aligned}
Y(t_i) &= \mu(t_i) + \epsilon(t_i), \quad \mu(t_i) > 0, \quad \text{for } i = 1, \cdots, n \\
\epsilon(t_i) &\overset{iid}{\sim} \mathsf{N}(0, \sigma^2) \\
\mu(t_i) &= b_0 + b_1 \sum_{j=1}^{J} \left[u_j k(t_i; \theta_j) I_{\{a_j=1\}}(a_j) + d_{\lceil t_i/24 \rceil} u_j k(t_i; \theta_j) I_{\{a_j=2\}}(a_j) \right] \\
(u_j, \theta_j) \mid J, \epsilon &\overset{iid}{\sim} \pi(u_j, \theta_j \mid \epsilon) \equiv \frac{\nu_\epsilon(du, d\theta)}{\nu_\epsilon(\mathbb{R}_+ \times \Theta)} \\
J &\sim \mathsf{Po}\left(\sum_{k=1}^{K} \alpha_k T_k E_1(\epsilon) \right)
\end{aligned}$$

where u_j and components of θ_j are independently and identically distributed from the following distributions:

$$\begin{aligned}
\pi(u_j \mid \epsilon) &= \frac{u_j^{-1} e^{-u_j}}{E_1(\epsilon)} I_{[\epsilon,\infty)}(u_j), \quad u_j \in [\epsilon, \infty) \\
\pi(\lambda_j) &\sim \mathsf{Ga}(\alpha_\lambda, \beta_\lambda) \\
\pi(\tau_j | a_j) &\sim \mathsf{Un}[0, T_{a_j}] \\
\pi(a_j) &= \prod_{k=1}^{K} p_k^{I_{\{a_j=k\}}(a_j)}, \; a_j \in \{1, \cdots, K\}, \;\; p_k = \frac{\alpha_k T_k}{\sum_{m=1}^{K} \alpha_m T_m}, \quad k = 1, \cdots, K\,.
\end{aligned}$$

The distributions of the rest of the parameters are as follows:

$$\begin{aligned}\log(d_l) &\sim \mathsf{N}(\mathbf{x}_l'\boldsymbol{\gamma},\ \sigma_d^2), \quad l = 1, \cdots, \lceil t_n/24 \rceil \\ \pi(b_0) &\sim \mathsf{Ga}(\alpha_{b_0}, \beta_{b_0}) \\ \pi(b_1) &\sim \mathsf{Ga}(\alpha_{b_1}, \beta_{b_1}) \\ \pi(\boldsymbol{\gamma}, \sigma_d^2) &\sim \frac{1}{\sigma_d^2} \\ \pi(\sigma^2) &\sim \frac{1}{\sigma^2}\,,\end{aligned}$$

where $E_1(\epsilon) \equiv \int_\epsilon^\infty\ u^{-1}e^{-u}\ du$, which is called the exponential integral function, vector $\mathbf{x}_l$ is the meteorological data for day l. We justify the choice of $p_1, \cdots, p_K$ (in the univariate daily model, $K = 2$) by introducing K auxiliary variables

$$J_k = \sum_{j=1}^{J} I_{\{a_j=k\}}(a_j), \quad \text{for } k = 1, 2, \cdots, K\,,$$

where J_k represents the number of type k jumps. A priori, the mean of J_k is set to be $\alpha_k T_k E_1(\epsilon)$. Note that the expectation of J_k can be calculated through an alternative way:

$$\mathbb{E}[J_k] \;=\; \mathbb{E}[\mathbb{E}[J_k\ |J]] = \mathbb{E}\left[\mathbb{E}\left[\sum_{j=1}^{J}\ I_{a_j=k}(a_j)\ |J\right]\right] = \mathbb{E}[Jp_k] = p_k\mathbb{E}[J]\,,$$

such that,

$$p_k = \frac{\mathbb{E}[J_k]}{\mathbb{E}[J]} = \frac{\alpha_k T_k}{\sum_{m=1}^{K}\ \alpha_m T_m} \quad k = 1, 2, \cdots, K\,.$$

The above specification of p_k satisfies the conditions that $p_1, \cdots, p_K \geq 0$ and $\sum_{k=1}^{K}\ p_k = 1$.

To finish the model specification we need to choose hyperpriors and hyperparameters. Cut-off value ϵ controls the minimum size of latent jump. The scale parameter b_1 makes the size of latent jump unit free. Simulation studies show that the result is normally insensitive to ϵ within a wide range of values. In this paper, we have used $\epsilon = 0.5$. The expected number of type k jumps in a unit time interval is $\alpha_k E_1(\epsilon)$. If we fix $\alpha_k,\ k = 1, \cdots, K$, the prior distribution for J_k is Poisson:

$$J_k \mid \alpha_k \;\sim\; \mathsf{Po}(\alpha_k E_1(\epsilon) T_k), \quad k = 1, \cdots, K\,,$$

which is a relatively concentrated prior distribution. Instead, a Gamma distribution is used as the prior for α_k:

$$\alpha_k \;\sim\; \mathsf{Ga}(a_{\alpha_k}, b_{\alpha_k}), \quad k = 1, \cdots, K\,.$$

We can show that the marginal distribution of J_k is:

$$\pi(J_k) = \frac{\Gamma(J_k + a_{\alpha_k})}{\Gamma(a_{\alpha_k})\Gamma(J_k+1)} \left[\frac{b_{\alpha_k}}{b_{\alpha_k} + E_1(\epsilon)T_k}\right]^{a_{\alpha_k}} \left[1 - \frac{b_{\alpha_k}}{b_{\alpha_k} + E_1(\epsilon)T_k}\right]^{J_k}\,.$$

If a_{α_k} is chosen to be a positive integer, J_k follows a Negative-Binomial distribution. Negative-Binomial distribution is a more dispersed distribution compared to its Poisson counterpart. The prior mean and variance of J_k is:

$$\mathbb{E}[J_k] = \frac{a_{\alpha_k} E_1(\epsilon) T_k}{b_{\alpha_k}} \quad , \quad \mathbb{V}[J_k] = \mathbb{E}[J_k]\left(1 + \frac{\mathbb{E}[J_k]}{a_{\alpha_k}}\right)\,.$$

By choosing prior mean and variance for J_k, we can solve the corresponding a_{α_k} and b_{α_k}. The kernel parameter λ_j controls the decay rate of a latent jump. We fix α_λ to be a constant and assumes a log Normal prior for β_λ:

$$\log \beta_\lambda \sim \mathsf{N}(m_b, v_b)\,.$$

With experts' opinion, we can choose the 10%, 50% and 90% quantiles of λ_j, and solve the corresponding values of α_λ, m_b and v_b. Numerical procedures are used since explicit solutions are not available. A Gamma distribution $\mathsf{Ga}(\alpha_{b_0}, \beta_{b_0})$ (with mean α_{b_0}/β_{b_0}) is chosen as the prior for b_0, which serves as the baseline level. We set $\mathbb{E}[b_0] = \bar{\mathbf{y}}_{\text{obs}}$ and $\mathbb{V}[b_0]$ to be some large multiple of $\mathbb{E}[b_0]$, and $(\alpha_{b_0}, \beta_{b_0})$ can be solved correspondingly. The values of $(\alpha_{b_1}, \beta_{b_1})$ are chosen such that prior distribution of $\mu(t)$ covers the range of the observations.

3.2.2 Posterior Simulation

The posterior inference is investigated through MCMC procedure. Since J, the number of latent jumps, is treated as a free parameter, the dimension of the parameter space varies. A birth-death type of reversible-jump MCMC algorithm (Green, 1995) is implemented to sample the varying dimensional parameters. The detailed algorithm to update varying dimensional parameters can be found in the paper by Tu *et al.* (2005). In the appendix B.2, we detailed the MCMC sampling scheme for fixed dimensional parameters.

3.2.3 Forecasting

In this section, the forecasting method for the time series model is discussed. Let $\mathbf{y} \equiv (y_1, y_2, \cdots, y_n)$ denote the data observed on the time interval $[0, T]$. In a Bayesian framework, predicting $Y(t)$ for $t \in [T, T + \Delta T]$ amounts to finding the predictive distribution $[Y(t) \,|\, \mathbf{y}]$. Let $\omega \equiv (\{u_j, \theta_j\}_{j \le J}, b_0, b_1, d_{\lceil t/24 \rceil}, J)$ denotes the parameter vector on which $\mu(t)$ depends. Note that:

$$[Y(t) \mid \mathbf{y}] \;=\; \int [Y(t) \mid \mu(t), \sigma^2]\, [\mu(t) \mid \omega]\, [\omega \mid \mathbf{y}]\, [\sigma^2|\mathbf{y}] \; d\mu(t) d\omega d\sigma^2 . \quad (3.6)$$

Dependent samples from $[Y(t)\,|\,\mathbf{y}]$ can be obtained using a Monte Carlo approach. Posterior samples of $(b_0, b_1, \sigma^2, \boldsymbol{\gamma}, \sigma_d^2)$, and posterior samples of $(\{u_j, \theta_j\}_{j\leq J}, J)$ that are associated with daily periodic jumps and with aperiodic jumps that occur in the interval $[0, T]$ on which we have observations, are available through fitting the model, and we denote them by: $(\tilde{b_0}^{(m)}, \tilde{b_1}^{(m)}, \tilde{\sigma}^{2^{(m)}}, \{\tilde{u_j}^{(m)}, \tilde{\theta_j}^{(m)}\}_{j\leq \tilde{J}^{(m)}}, \tilde{J}^{(m)}, \tilde{\boldsymbol{\gamma}}^{(m)}, \tilde{\sigma_d}^{2^{(m)}})$, for $m = 1, 2, \cdots, M$, where m denotes the m^{th} thinned posterior samples. Posterior samples of $(\{u_j, \theta_j\}_{j\leq J}, J)$ that are associated with aperiodic jumps in the interval $(T, T + \Delta T]$ are obtained from their prior distributions (conditioning on the hyperparameters), since there are no observations in $(T, T{+}\Delta T]$. We denote them by $(\{\breve{u_j}^{(m)}, \breve{\theta_j}^{(m)}\}_{j\leq \breve{J}^{(m)}}, \breve{J}^{(m)})$, which can be obtained by: first draw $\breve{J}^{(m)} \sim \mathsf{Po}(\breve{\alpha}_1^{(m)} \Delta T E_1(\epsilon))$; fix $a_j = 1$ since it is for the aperiodic part; conditioning on $\breve{J}^{(m)}$, we can sample $\{u_j^{(m)}, \tau_j^{(m)}, \lambda_j^{(m)}\}_{j\leq \breve{J}^{(m)}}$ from the prior distribution as we specified in Section 3.2.1. The hyperparameters $(\breve{\alpha}_1^{(m)}, \breve{\beta}_\lambda^{(m)})$ are available from model fitting. We can use a composition sampling scheme to draw samples from Eqn. (3.6) through a hierarchical fashion:

$$
\begin{aligned}
d_l^{(m)} &\sim \mathsf{N}(\mathbf{x}_l' \tilde{\boldsymbol{\gamma}}^{(m)}, \tilde{\sigma_d}^{2^{(m)}}), \quad l = \lceil t/24 \rceil \\
\mu^{(m)}(t) &= \tilde{b_0}^{(m)} + \tilde{b_1}^{(m)} \sum_{j=1}^{\breve{J}^{(m)}} \breve{u}_j^{(m)} k(t; \breve{\theta}_j^{(m)}) I_{\{\breve{a}_j^{(m)}=1\}}(\tilde{a}_j^{(m)}) \\
&\quad + \tilde{b_1}^{(m)} \sum_{j=1}^{\tilde{J}^{(m)}} \left[\tilde{u}_j^{(m)} k(t; \tilde{\theta}_j^{(m)}) I_{\{\tilde{a}_j^{(m)}=1\}}(\tilde{a}_j^{(m)}) + d_l^{(m)} \tilde{u}_j^{(m)} k(t; \tilde{\theta}_j^{(m)}) I_{\{\tilde{a}_j^{(m)}=2\}}(\tilde{a}_j^{(m)}) \right] \\
y^{(m)}(t) &\sim \mathsf{N}(\mu^{(m)}(t), \tilde{\sigma}^{2^{(m)}}),
\end{aligned}
$$

for $m = 1, 2, \cdots, M$, where weather forecast including wind-speed and temperature for day l is used to calculate $\mathbf{x}_l$. The collection $\left\{y^{(1)}(t), y^{(2)}(t), \cdots, y^{(M)}(t)\right\}$ are samples from the posterior predictive density $[Y(t)|\mathbf{y}]$, which can be used to obtain a point estimate and credible interval for $Y(t)$, among others.

3.2.4 Results

We fit the model specified in Section 3.2.1 to a real data set consisting of 30 consecutive days of hourly PM_{10} data. There are seven missing values in the 720 observations. We first discuss the hyperparameters used in this example. The cutoff value ϵ is chosen to be 0.15 in this example. Let $\lambda_{0.1}$, $\lambda_{0.5}$ and $\lambda_{0.9}$ denote 10%, 50% and 90% quantiles of λ_j, respectively. We set:

$$\begin{aligned} \exp(-24\lambda_{0.1}) &= 0.1 \implies \lambda_{0.1} = 0.01 \\ \exp(-2.0\lambda_{0.5}) &= 0.1 \implies \lambda_{0.5} = 1.15 \\ \exp(-0.5\lambda_{0.9}) &= 0.1 \implies \lambda_{0.9} = 4.6\,, \end{aligned}$$

$\lambda_{0.1}, \lambda_{0.5}$ and $\lambda_{0.9}$ are chosen such that the time for a latent point source decaying to 10% of its maximum intensity is $24, 2$ and 0.15 hours, respectively. Analytical solutions for the corresponding α_λ, m_b and v_b do not exist, so a numerical procedure based on a grid search is used. For each grid point (α_λ, m_b, v_b), we can simulate N realizations of λ based on the specified priors and calculate the 10%, 50% and 90% quantiles of λ. Then we calculate l^2-distance between the vector $(\lambda_{0.1}, \lambda_{0.5}, \lambda_{0.9})$ and $(0.01, 1.15, 4.6)$. The values of (α_λ, m_b, v_b) are chosen such that this l^2-distance is minimized. Following this procedure, α_λ, m_b and v_b are chosen to be 2.3, 0.53, and 0.76, respectively. We choose $a_{\alpha_1} = a_{\alpha_2} = 1.0$ and $b_{\alpha_1} = b_{\alpha_2} = 2.0$, which corresponds to about one jump per hour in expectation for both periodic and aperiodic components, a priori. We run the MCMC chain for $100,000$ iterations, with the first $25,000$ iterations discarded as burn-in. The remaining samples are used for posterior inference.

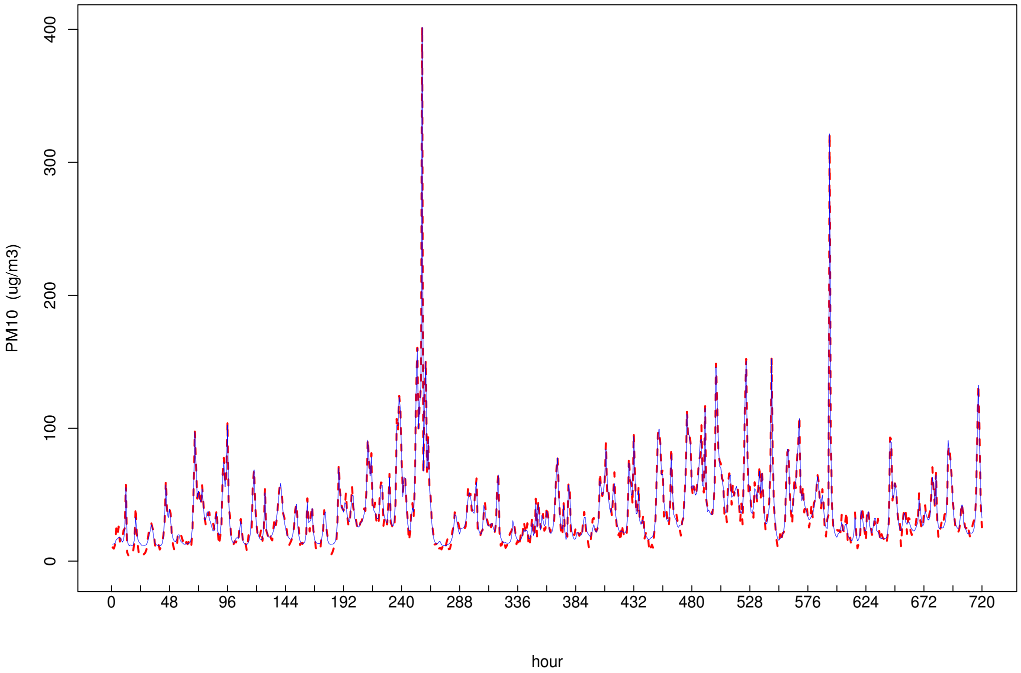

Figure 3.2: Model fit of 30 days of hourly PM_{10} data. The solid line represents the posterior mean and the dotted line represents the true observations. The fitted RMSE is 4.8. The starting day is April 8, 1998.

The fitted mean process is shown in Fig. (3.2). The result is satisfying in the sense that model captures both local peaks and global trends. Fig. (3.3) shows the residuals and the autocorrelation plot. The autocorrelation plot strongly indicates that the model assumption of iid measurement errors is valid. Fig. (3.4) shows the decomposition of the mean process into its aperiodic and daily periodic components. The posterior distribution of the latent process itself is also of interest. We show in Fig. (3.5) and Fig. (3.6) the time and magnitudes of the latent jumps from one posterior sample. Posterior distribution of the latent process for the periodic part shows two strong peaks at around 7 am and 8 pm everyday, which may be

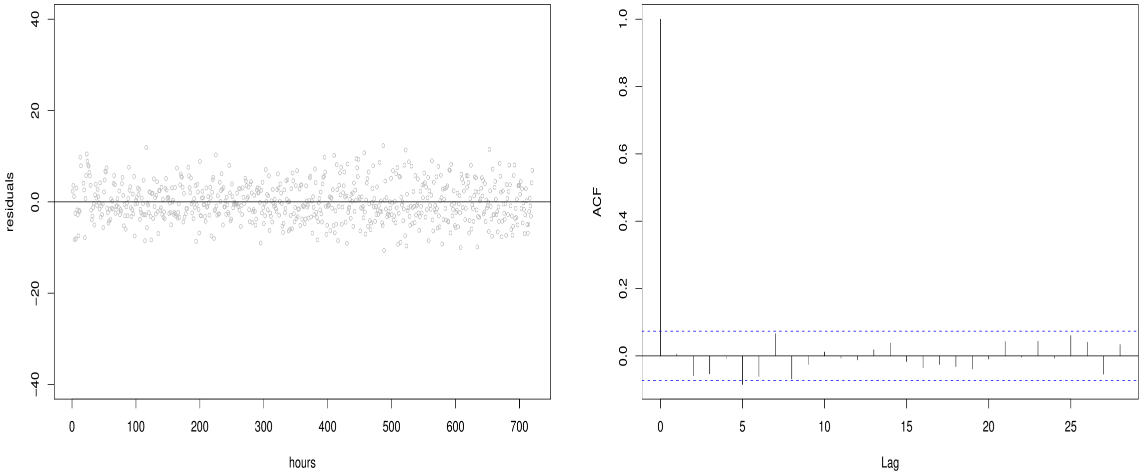

Figure 3.3: Residuals (resid $\equiv y_{\text{fit}} - y_{\text{obs}}$). The left diagram shows the residuals. The right plot shows the autocorrelation plot of the residuals.

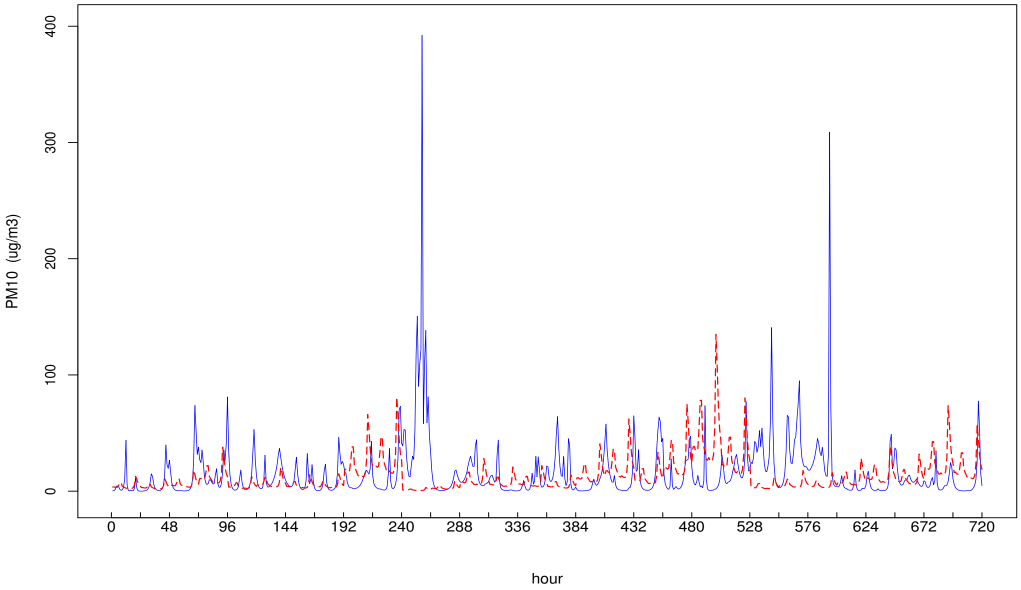

Figure 3.4: Aperiodic and periodic decomposition. The solid line represents the posterior mean from the aperiodic part and the dashed line represents the posterior mean from the daily periodic part.

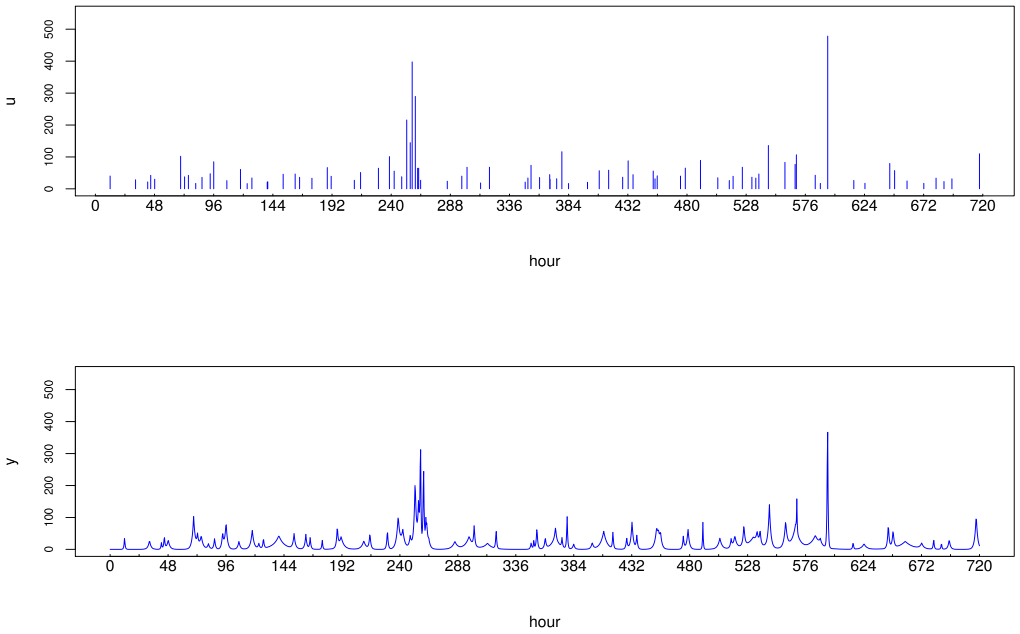

Figure 3.5: Aperiodic jumps. The upper diagram shows the time and magnitudes of the aperiodic latent jumps from one posterior sample. The bottom plot shows the mean process generated by the kernel convolution of the aperiodic jumps.

due to daily traffic. We plot the daily factors d_l for $l = 1, \cdots, 30$ with daily wind speed and temperature in Fig. (3.7), which indicates a positive correlation between daily factor and wind speed. A five day out of sample prediction was done to test the predictive performance of the model. The forecasting method described in Section 3.2.3 is implemented. Instead of using weather forecast, real meteorological data are used. In reality, uncertainty due to meteorological forecast needs to be addressed in the forecasting model. The result is shown in Fig. (3.8). The 90% pointwise Bayesian credible interval is constructed and the coverage is 91.5%. The root mean square error (RMSE $\equiv \sqrt{(\sum(y_{\text{fit}} - y_{\text{obs}})^2)/n}$) for the prediction is

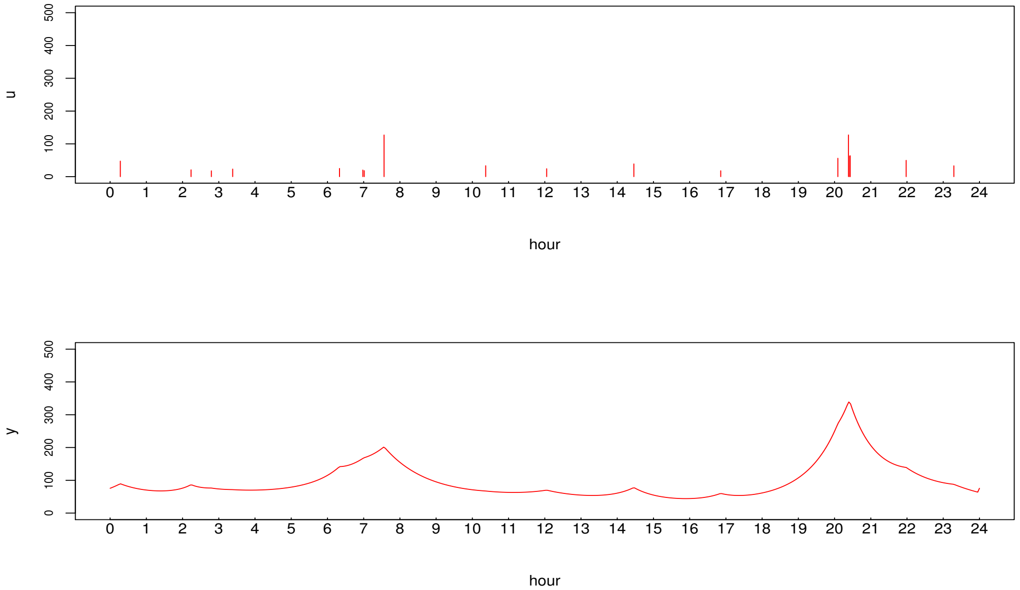

Figure 3.6: Daily periodic jumps. The upper diagram shows the time and magnitudes of the daily latent jumps from one posterior sample. The lower plot shows the mean process generated by kernel convolution of the daily jumps.

28.1. For comparison, we fit a Bayesian AR(2) model that includes covariates:

$$Y(t_i) = b_0 + \alpha_1 Y(t_{i-1}) + \alpha_2 Y(t_{i-2}) + \mathbf{X}(t_i)'\boldsymbol{\beta} + \epsilon(t_i)\,,$$

where the covariates vector $\mathbf{X}(t_i)$ includes day, hour and hourly meteorological data including wind speed, temperature and their quadratic terms. The forecasting based on the AR model is compared with the adaptive kernel model. The adaptive kernel model performs better: for the adaptive kernel method, the coverage of the 90% credible interval is 91.5%, and the RMSE is 28.1; while for the AR model, the coverage of the 90% credible interval is only 67.3% and the RMSE is only 31.2.

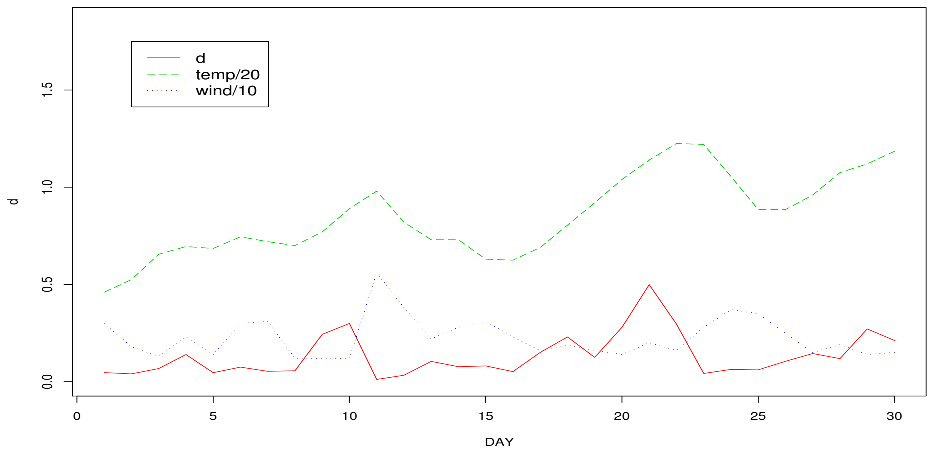

Figure 3.7: Daily Factor. The solid line is the posterior mean of the daily factor. The dotted and dashed lines are rescaled daily wind speed and temperature, respectively.

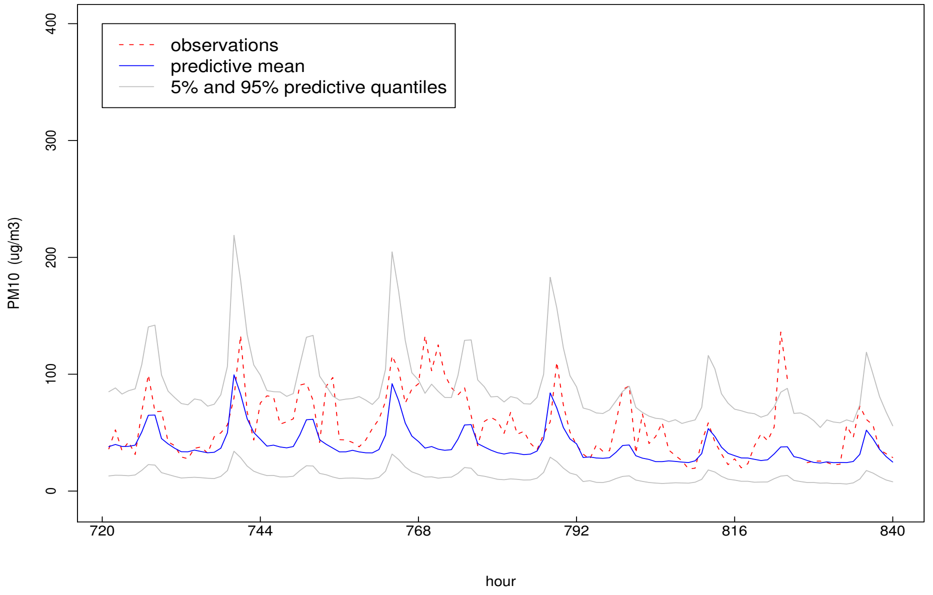

Figure 3.8: Five day out of sample predictions. The dashed line represents the true observations; the solid line is the posterior predictive mean; the two grey solid lines are pointwise 5% and 95% quantile estimates. The coverage of the 90% Bayesian credible interval is 91.5%, and the RMSE is 28.1.

3.3 Multiple Time Series Modeling

In the previous section, we developed a Bayesian semiparametric model for univariate air pollutant time series data through an adaptive kernel convolution method. In this section, the method is extended to model multivariate time series data for air pollutants. The scheme we are about to discuss is general for modeling K-variate processes. A bivariate process example is used to illustrate the method.

3.3.1 The Model

We assume a bivariate time series model of the form:

$$\begin{aligned} Y_1(t_i) &= \mu_1(t_i) + \epsilon_1(t_i) \\ Y_2(t_i) &= \mu_2(t_i) + \epsilon_2(t_i) \end{aligned}$$

$$\epsilon_1(t_i) \overset{iid}{\sim} \mathsf{N}(0, \sigma_1^2), \quad \epsilon_2(t_i) \overset{iid}{\sim} \mathsf{N}(0, \sigma_2^2), \quad \epsilon_1(t_i) \perp\!\!\!\perp \epsilon_2(t_i), \quad \text{for } i = 1, 2, \cdots, n\,.$$

We assume independent Gaussian white noise error processes. The association between two time series is modeled through their mean processes $\mu_1(t)$ and $\mu_2(t)$. As in Section 3.2.1, the mean processes are modeled by a latent marked point process. Each latent jump j is assigned with a categorical variable a_j to indicate its type. We show in Table 3.1 the assignment of a_j.

Table 3.1: Assignment of categorical variable a_j.

a_j	polutant-1	pollutant-2	shared jumps
periodic	1	2	3
aperiodic	4	5	6

Six types of jumps are needed in this bivariate time series model, i.e., the aperiodic jump for pollutant 1, the aperiodic jump for pollutant 2, the daily jump for pollutant 1, the daily jump for pollutant 2, the aperiodic jump shared by both pollutants and the daily jump shared by both pollutants. The latent jumps shared by both pollutants introduce correlation between two processes. Non-linear association between two time series can be modeled through this way, which cannot be accomplished by classical multivariate time series models that use cross-covariance functions to model correlations.

Each of $\mu_1(t)$ and $\mu_2(t)$ is decomposed into four components: their own aperiodic and daily periodic components, and their shared aperiodic and daily periodic components:

$$\mu_1(t) = b_{01} + b_{11}\left\{\sum_{j=1}^{J} u_j k(t;\theta_j) I(a_j)_{\{4,6\}} + d_{1\lceil t/24\rceil}\sum_{j=1}^{J} u_j k(t;\theta_j) I(a_j)_{\{1,3\}}\right\} \quad (3.7)$$

$$\mu_2(t) = b_{02} + b_{12}\left\{\sum_{j=1}^{J} u_j k(t;\theta_j) I(a_j)_{\{5,6\}} + d_{2\lceil t/24\rceil}\sum_{j=1}^{J} u_j k(t;\theta_j) I(a_j)_{\{2,3\}}\right\}. \quad (3.8)$$

In the above equations, b_{0k} is the baseline level for pollutant k, and b_{1k} is the overall scale factor which standardizes the measurements for pollutant k. Since different pollutants are measured on heterogeneous scales, b_{1k} makes the model robust with respect to the choice of scales. d_{kl} is daily intensity factor on day l for pollutant k, and is modeled through meteorological covariates as we did in the univariate case. Kernel parameter $\theta_j \equiv (\tau_j, \lambda_{1j}, \lambda_{2j}, a_j)$. There are two decay parameters λ_{1j} and λ_{2j} assigned to each jump j, which represent decay rates for pollutant 1 and pollutant 2, respectively. Different pollutants may dissipate with different rates, as a result, for a jump that is shared by both pollutants, we assign each pollutant its own decay parameter. The kernel $k(t;\theta_j)$ is defined as:

$$k(t;\theta_j) = \begin{cases} \frac{e^{-\lambda_{1j}\left[(t-\tau_j) \mod 24\right]}\left(1+e^{-2\lambda_{1j}\left[12-(t-\tau_j) \mod 24\right]}\right)}{1-e^{-24\lambda_{1j}}} & \text{if } a_j = 1 \text{ or } a_j = 3 \\ e^{-\lambda_{1j}|t-\tau_j|} & \text{if } a_j = 4 \text{ or } a_j = 6 \\ \frac{e^{-\lambda_{2j}\left[(t-\tau_j) \mod 24\right]}\left(1+e^{-2\lambda_{2j}\left[12-(t-\tau_j) \mod 24\right]}\right)}{1-e^{-24\lambda_{2j}}} & \text{if } a_j = 2 \text{ or } a_j = 3 \\ e^{-\lambda_{2j}|t-\tau_j|} & \text{if } a_j = 5 \text{ or } a_j = 6, \end{cases}$$

The prior specifications and the computational algorithms are similar to the univariate case, which we do not detail here.

3.3.2 Example: A Joint Model of PM_{10} and CO

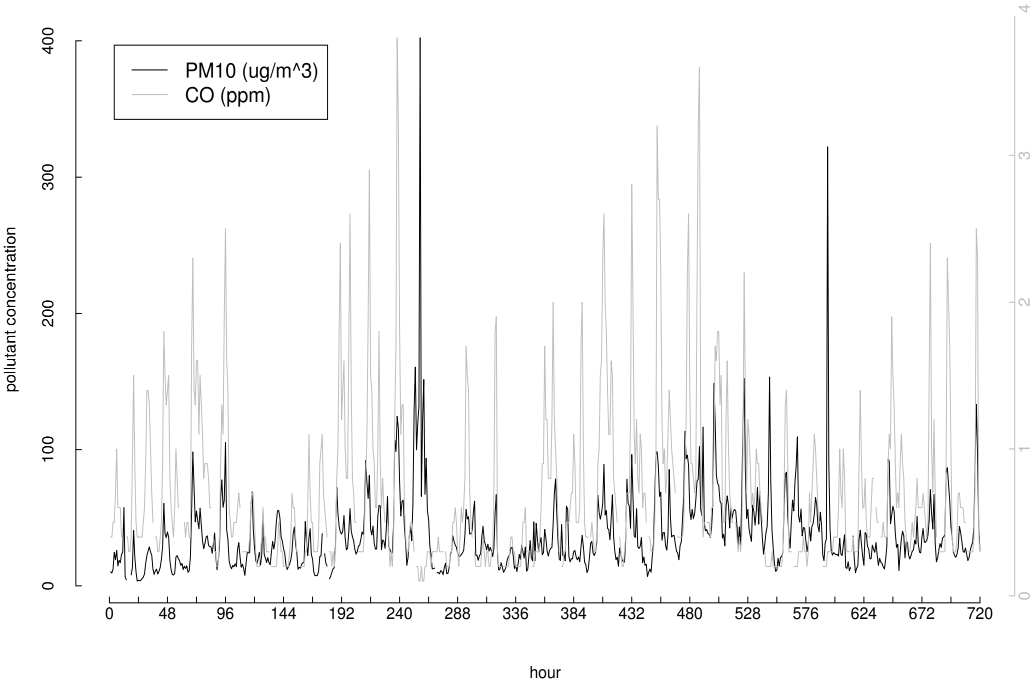

Figure 3.9: Thirty days of hourly PM_{10} and CO concentration data in Maricopa county, Arizona. The starting day is April 6, 1998. The Grey line is the time series for CO and the black line is the time series for PM_{10}.

In this section, we fit the model developed in the previous section to 30 consecutive days of hourly PM_{10} and CO data measured at the same location in Maricopa county of Arizona. The data are displayed in Fig. (3.9). Note that PM_{10} is measured in ug/m^3 and CO is measured in PPM. There are 7 missing observations for PM_{10} and 34 for CO. Missing observations are excluded from the likelihood evaluation. The specifications for hyperpriors and hyperparameters are similar to the univariate case, which we do not repeat here.

The fitted bivariate times series model is shown in Fig. (3.10) and Fig. (3.11).

The root mean square errors (RMSE) for the fitting are 4.42 and 0.068, for PM_{10} and CO respectively. In Fig. (3.12) and Fig. (3.13) we show the decomposition for the posterior mean process. The results show that there are significant number of shared aperiodic jumps between PM_{10} and CO, but not so much for the the shared daily periodic jumps. The five day out of sample prediction for the joint model is shown in Fig. (3.14) and Fig. (3.15). We compare the results of modeling PM_{10} and CO separately to the results of modeling PM_{10} and CO jointly in Table 3.2. The results are mixed: fitting of PM_{10} is improved in the joint model, but the prediction performance for PM_{10} is slightly worse in the joint model; fitting of CO is worse in the joint model but the prediction performance is better in the joint model.

Table 3.2: Comparisons of the results from univariate time series models of PM_{10} and CO to the results from the bivariate time series model.

Model	RMSE (fit)	RMSE (pred)	Coverage (pred)
PM_{10}	4.8	28.1	91.5%
PM_{10} (joint model)	4.42	33.07	80.5%
CO	0.046	0.48	87.8%
CO (joint model)	0.068	0.42	91.3%

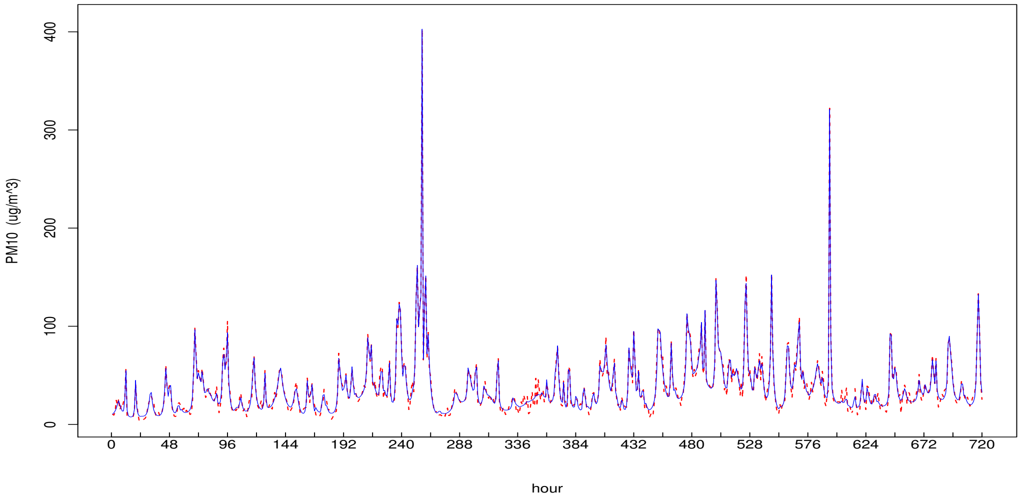

Figure 3.10: Fitted model of 30 days of hourly PM_{10} data for the joint model of PM_{10} and CO. The solid line represents the posterior mean and the dotted line represents the observations. The fitted RMSE is 4.42.

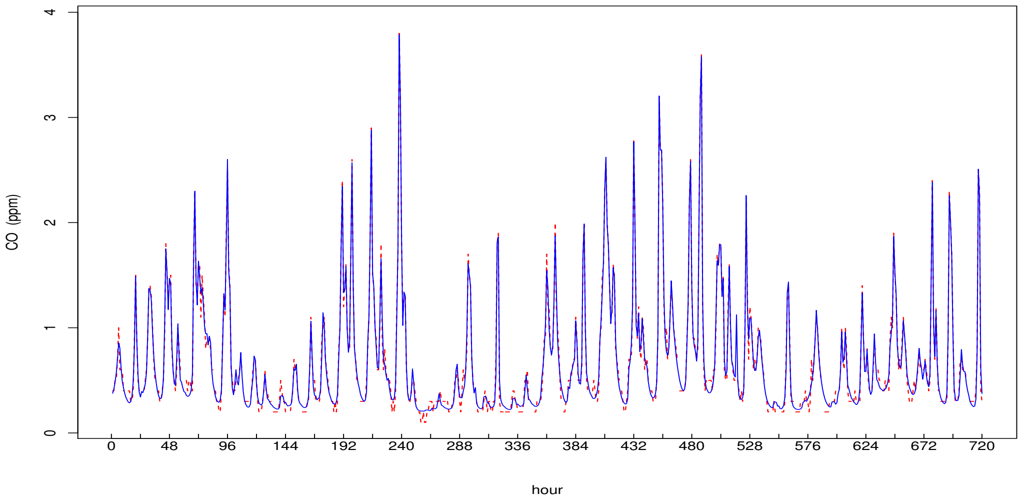

Figure 3.11: Fitted model of 30 days of hourly CO data for the joint model of PM_{10} and CO. The solid line represents the posterior mean and the dotted line represents the observations. The fitted RMSE is 0.068.

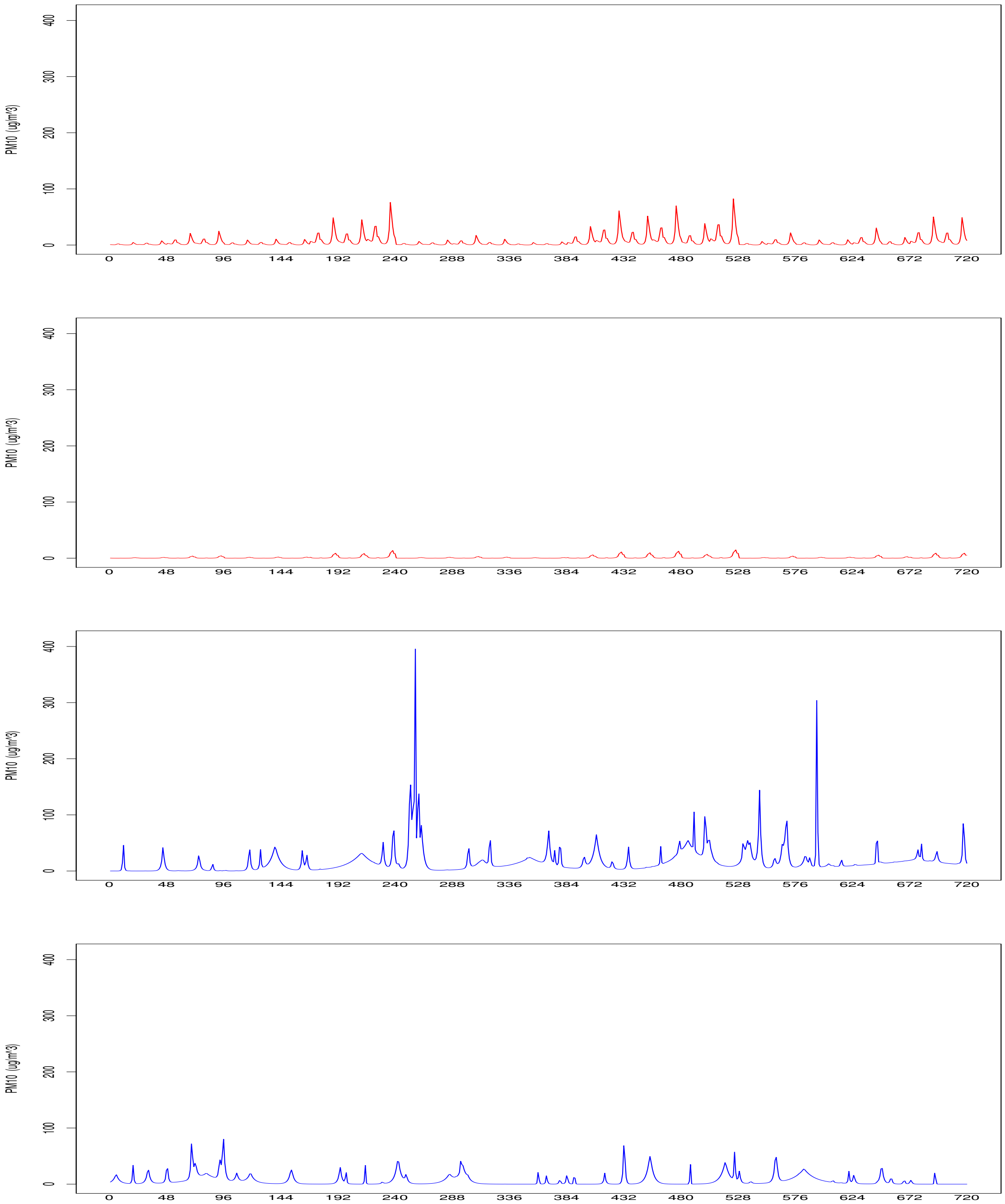

Figure 3.12: Decomposition of posterior mean process for PM_{10}. From the top to the bottom, the four plots represent the mean processes generated from daily jumps of PM_{10}, daily shared jumps between PM_{10} and CO, aperiodic jumps of PM_{10} and aperiodic shared jumps between PM_{10} and CO, respectively.

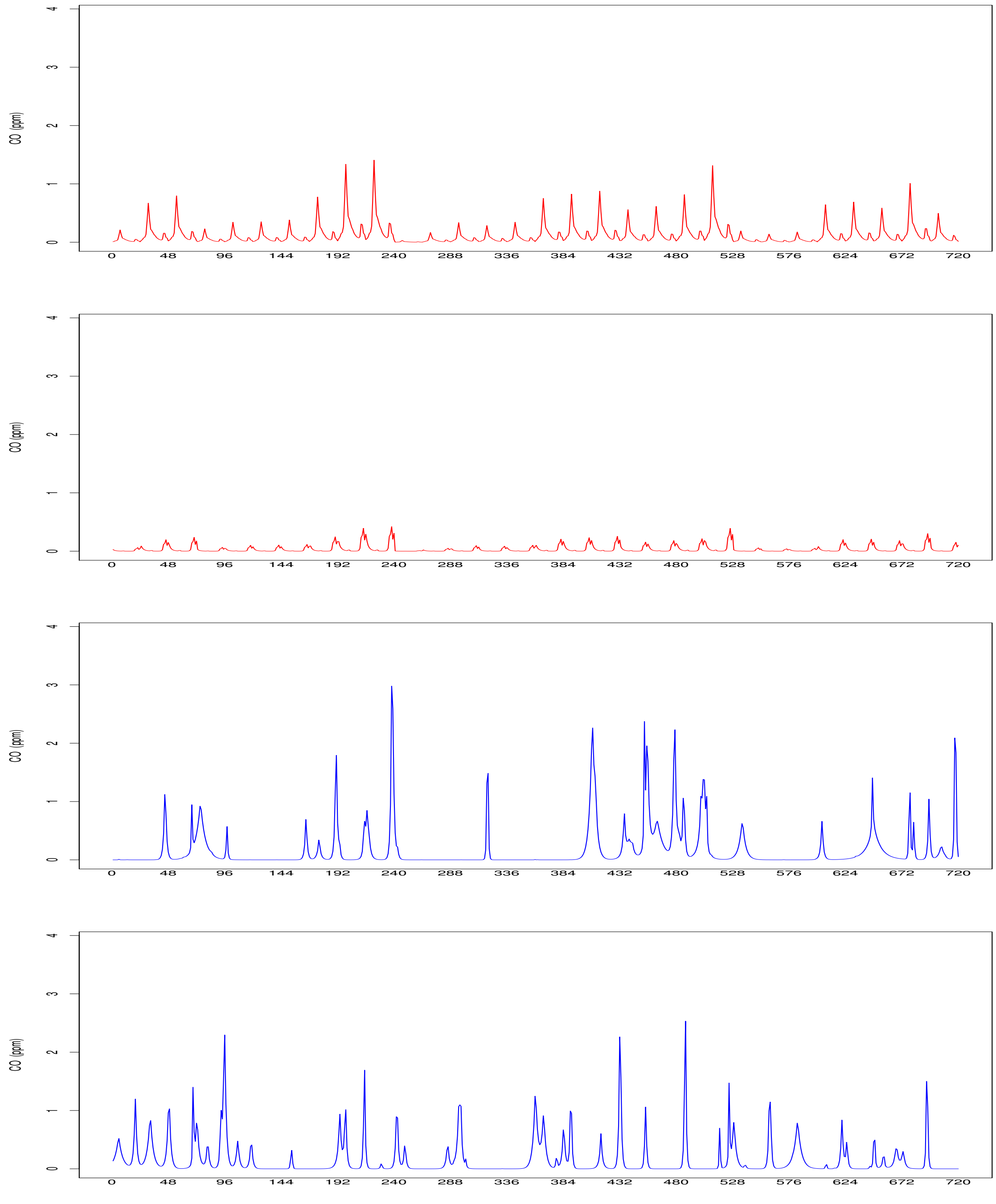

Figure 3.13: Decomposition of posterior mean process for CO. From the top to the bottom, the four plots represent the mean processes generated from daily jumps of CO, daily shared jumps between PM_{10} and CO, aperiodic jumps of CO and aperiodic shared jumps between PM_{10} and CO, respectively.

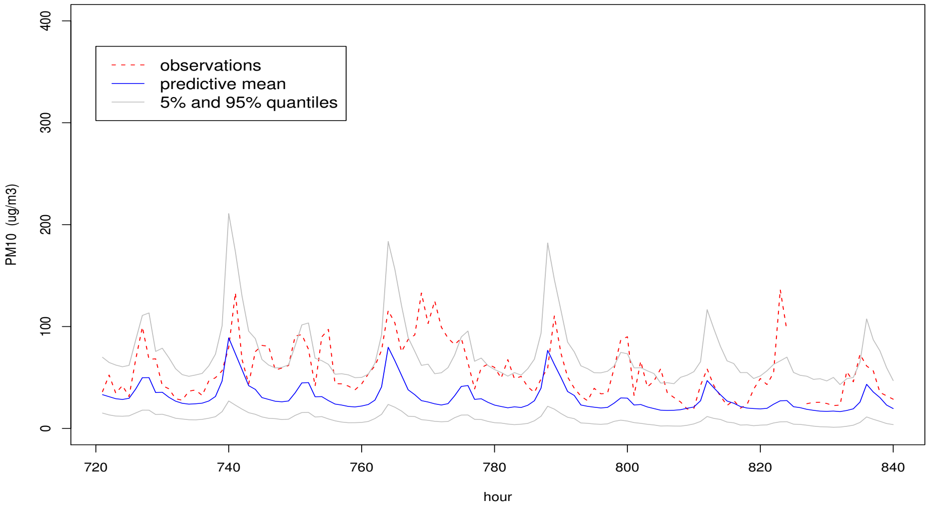

Figure 3.14: Five day out of sample prediction for PM_{10} in the joint model of PM_{10} and CO. The coverage of the Bayesian 90% credible interval is 80.5%. The RMSE is 33.07.

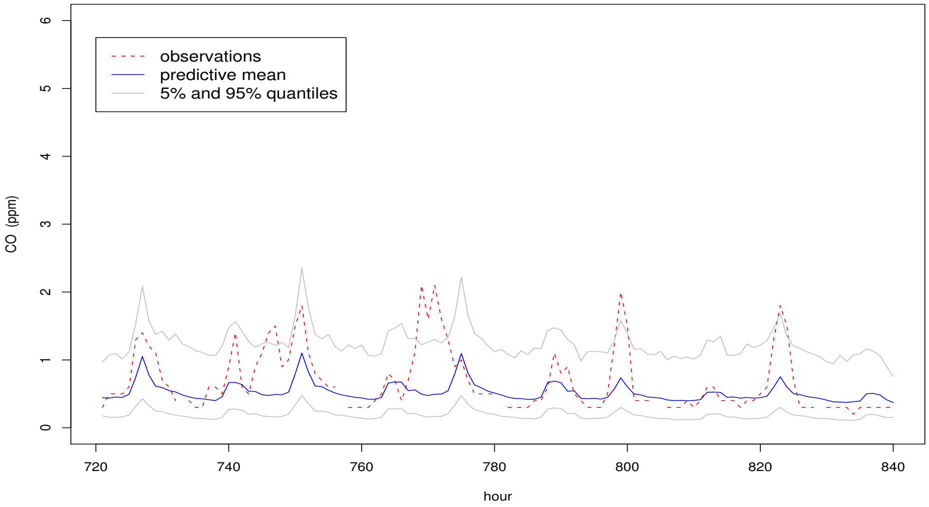

Figure 3.15: Five day out of sample prediction for CO in the joint model of PM_{10} and CO. The coverage of the Bayesian 90% credible interval is 91.3%. The RMSE is 0.42.

3.4 Discussion

In this paper, a class of Bayesian semiparametric time series models is developed for modeling multivariate ambient pollutants. The models are built on adaptive kernel smoothing of latent Gamma processes. In stead of focusing on modeling residual process, as most time series models do, we focus on modeling mean process nonparametrically.

The mean processes is represented as a weighted sum of kernels with arbitrary location parameters. The kernels are location specific and thus are adaptive. The adaptive kernels allow us to model both slowly varying long-term trend and fast decaying local peaks. The Gamma process based approach allows to model non-negative process directly without the need for data transformations. The method developed provides a framework to model non-Gaussian and non-stationary time series data. Many environmental time series data are known to be non-Gaussian and non-stationary, which are extremely hard for classical ARMA type of time series model to handle. The adaptive kernel methods developed in this paper will serve as a strong candidate to model this type of data. Classical multivariate time series models resort to cross-covariance functions to model inter-dependence structure. It can only capture linear association between two processes. The latent process approach we developed can model non-linear association among time series, thus are potentially more general. Classical time series models usually require the data being collected at regular grid points. Since the latent process is defined on continuous time, the time series models we developed is continuous in time, and can be used to analyze data that are collected at irregularly spaced time points.

To conclude, we present here a promising applications of the time series models we developed in this article. As we discussed in the beginning, air pollution has strong adverse effect on human health. Studies that try to connect air pollutant concentration and mortality are drawing ever increasing attention. The frequently used health effect models that relate mortality rate to pollutants concentration are regression type of models which include pollutants concentration measured at discrete time points as regressors. With our latent process approach, we can build a health effect model that links health outcomes to latent pollutants exposure field using evolutionary covariates that integrate over past exposures in time. This avoids the sometimes ad hoc choice of lag structures. It will be interesting to compare the results of health effects model based on those different time series models. Maybe more exciting fact is that the latent process approach we developed in this paper to analyze multivariate time series data can be easily extended to build multivariate space-time model, which can be used for spatial epidemiology study. This is part of our ongoing research.

Chapter 4

Adaptive Kernel Methods in Spatio-Temporal Modeling

4.1 Introduction

Spatio-temporal modeling has gained increased attention in both applied and theoretical work in the last several years. To analyze point-referenced space-time data, the common thread of modeling is through Gaussian processes. Modeling through Gaussian processes requires specification of a valid spatio-temporal covariance function: that for any set of locations and time points the covariance function for the set of random variables must be positive definite. A frequently used space-time covariance function takes a separable form which is simply a valid two-dimensional spatial covariance function multiplied by a valid one-dimensional autocovariance function. Non-separable space-time covariance functions that allow more flexible space-time interaction have been studied by Cressie and Huang (1999) and Stein (2005). Another class of models for space-time data is dynamic (West, 1997) spatio-temporal models, which includes classical approaches (Huang and Cressie, 1996; Wikle and Cressie, 1999; Mardia *et al.*, 1998) and Bayesian approaches. (Gelfand *et al.*, 2005; Stroud *et al.*, 2001; Huerta *et al.*, 2004).

An alternative constructive approach to generating a Gaussian process in $\mathbb{R}^d$ is to place latent independent identically distributed Gaussian random variables on a lattice in $\mathbb{R}^d$ and convolve them with smoothing kernels. The process convolution method (Higdon, 1998) is used successfully for spatial and spatio-temporal modeling. Calder (2003) extended the method by combining the process convolution approach with space-time Kalman filter. Defining a Gaussian process through process convolution is equivalent to defining a Gaussian process through specification of a covariance function. The advantage of using the process convolution approach, however, reaches beyond Gaussian process modeling. It can be readily extended to model non-stationary non-Gaussian spatio-temporal processes, which are the primary focus of this paper.

In the previous papers pertaining to process convolution, the latent discrete white noise process is defined on a coarse lattice in the studying region $\mathcal{S}$. An important modeling decision when applying the process convolution method is to appropriately choose the number and locations of the grid points. An ad hoc method of choosing such points is discussed by Calder (2003) if both the latent process and the convolution kernel are Gaussian. However it is difficult to find a consistent and appropriate technique for specifying the location of the underlying process. To remedy this problem, we generate the latent process through first distributing random variables over space $\mathcal{S}$ and time $\mathcal{T}$ according to a marked point process. We then associate each random variable with a kernel defined over space $\mathcal{S}$ and time $\mathcal{T}$. The continuous spatio-temporal process is obtained by smoothing the latent process according to its kernels. Through this procedure, we develop a fully Bayesian adaptive kernel convolution approach for spatio-temporal modeling. In this paper, Section 4.2 describes the Bayesian adaptive kernel convolution approach for spatio-temporal modeling. A Gamma random field is also discussed, which serves as prior distribution for the latent process. We then discuss the prior distributions and choice of hyperparameters. In Section 4.3, the prediction method is discussed. We illustrate the method in a simulation study followed by a real example in Section 4.4. Conclusions are given in Section 4.5.

4.2 Space-Time Modeling using Adaptive Kernel Convolutions

We consider point-referenced location and continuous time. Let $Z_i = Z(s_i, t_i)$ denote the observation at location s_i and time t_i for $i = 1, \cdots, n$. We write the general form for Z_i as follows:

$$Z_i = \mu(s_i, t_i) + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} \mathsf{N}(0, \eta)$$

where $\mu(s_i, t_i)$ denotes the mean structure and ϵ_i denotes the error process. For nonparametric modeling, there are two categories of approaches from which we can choose. The first approach is to model the mean structure nonparametrically. Conditional on the mean process, a simple parametric iid model for the error process is assumed. The second approach is to choose a parametric model for the mean function but model the error process nonparametrically. In this paper we adopt the first approach and focus on building the mean spatio-temporal surface nonparametrically.

4.2.1 Space-Time Models Based on Kernel Convolutions

As discussed in Section 4.1, we take a constructive approach for creating a spatio-temporal process $\mu(s,t)$ over $\mathcal{S} \times \mathcal{T}$, where $\mathcal{S} \subset \mathbb{R}^2$ and $\mathcal{T} \subset \mathbb{R}$. Let Θ be a Borel measurable subset of $\mathbb{R}^d$, choose a Borel measurable function $k : \mathcal{S} \times \mathcal{T} \times \Theta \to \mathbb{R}$ and set:

$$\mu(s,t) = b_0 + b_1 \sum_{j=1}^{J} u_j k(s,t;\theta_j) \tag{4.1}$$

for a random number $J \leq \infty$ of pairs (u_j, θ_j), where $u_j \in \mathbb{R}$, $\theta_j \in \Theta$. We use b_0 to model the baseline level, b_1 serves as a scale factor which makes the coefficients $\{u_j\}_{j \leq J}$ independent of the measurement unit of the observations. Each point $\theta \in \Theta$ contains the following elements: $\theta = (\sigma, \tau, \lambda_\tau, \lambda_1, \lambda_2, \phi)$, where $\sigma \in \mathcal{S}$ and $\tau \in \mathcal{T}$ denote the location and time of the underlying process respectively, with the remaining elements as the kernel associated parameters. Notice that if we fix J as well as the locations and time of the latent process $\{\sigma_j, \tau_j\}$, set underlying variables $\{u_j\}$ to be iid Normal random variables, and choose a non-adaptive kernel $k(s,t;\theta)$, Model (4.1) becomes a discrete process convolution model that has been studied extensively by Higdon (1998) and Calder (2003), among others.

Model (4.1) contains several important new features. First, the number of underlying variables as well as the location and time of the underlying variables are random. Under a Bayesian framework, posterior distributions are used to determine how many underlying variables are needed in addition to where and when to place the underlying variables. This relieves the researcher from using an ad hoc method for choosing J and $\{\sigma_j, \tau_j\}_{j \leq J}$. Allowing the latent variables to evolve spatially and temporally may help uncover the data generating mechanisms.

Assigning each latent variable, u_j, its own kernel $k(\cdot;\theta_j)$ facilitates modeling of non-stationary spatio-temporal processes. Assigning u_j a distribution other than a Normal offers us an alternative for modeling non-Gaussian data.

To fully specify Model (4.1), we must specify a joint probability distribution for J and $\{(u_j,\theta_j)\}_{j=1}^{J}$. A Lévy random field is a natural choice for this purpose. With a Poisson construction of a Lévy random field, we can represent Eqn. (4.1) through a stochastic integral

$$\mu(s,t) = b_0 + b_1 \int_{\Theta} k(s,t;\theta)\, L(d\theta)$$

where $L(d\theta)$ defines a Lévy random field on Θ. The mean surface $z(s,t)$ is driven by the underlying latent Lévy random field $L(d\theta)$, which distributes J (where J is random) point sources of magnitude u_j at time τ_j and location σ_j for $j = 1, \cdots, J$. The point sources dissipate in time and space with decay rates controlled by $k(s,t;\theta)$. The mean process $\mu(s,t)$ is the weighted average of J point sources at time $\{\tau_j\}_{j=1}^{J}$ and location $\{\sigma_j\}_{j=1}^{J}$, with weights determined by the smoothing spatio-temporal kernels. A detailed discussion of Lévy random field priors is in Section 4.2.5.

4.2.2 Spatio-Temporal Kernel

A flexible spatio-temporal kernel is essential in our approach to characterize space-time dependence. In order to allow non-stationarity, we incorporate location-and-time-specific parameters into the three-dimensional kernel. Recall that (s,t) and (σ,τ) are used to denote location and time of observations and the latent process, respectively. Under Cartesian coordinates x-y, let $s \equiv (x,y)$ and $\sigma_j \equiv (\sigma_{x_j}, \sigma_{y_j})$.

A separable form of spatio-temporal kernel is defined as follows:

$$k(s,t;\sigma_j,\tau_j,\lambda_{\tau_j},\lambda_{1_j},\lambda_{2_j},\phi_j) = k_t(t;\tau_j,\lambda_{\tau_j})k_s(s;\sigma_j,\lambda_{1_j},\lambda_{2_j},\phi_j) \qquad (4.2)$$

For the temporal kernel, the kernel of a double exponential probability density function is used, i.e.,

$$k_t(t;\tau_j,\lambda_{\tau_j}) \equiv e^{-\lambda_{\tau_j}|t-\tau_j|}.$$

For the spatial kernel, the kernel of a bivariate Normal probability density function is used. A standard parametrization of the bivariate Normal distribution is as follows:

$$\text{Let} \quad \Sigma_j \equiv \begin{pmatrix} \psi_{1_j}^2 & \kappa_j\psi_{1_j}\psi_{2_j} \\ \kappa_j\psi_{1_j}\psi_{2_j} & \psi_{2_j}^2 \end{pmatrix}, \ \Delta s_j \equiv s - \sigma_j \equiv \begin{pmatrix} x - \sigma_{x_j} \\ y - \sigma_{y_j} \end{pmatrix}$$

$$k_s(s;\sigma_j,\psi_{1_j},\psi_{2_j},\kappa_j) = \exp\left\{-\frac{1}{2}\Delta s_j^T \Sigma_j^{-1} \Delta s_j\right\}.$$

This parametrization, however, does not have a clear spatial interpretation. Furthermore, it may cause mixing problems in posterior simulation. To solve these problems, we reparametrize the bivariate Normal distribution by decomposing the covariance function as follows:

$$\Sigma_j = R(\phi_j)\begin{pmatrix} \lambda_{1_j} & 0 \\ 0 & \lambda_{2_j} \end{pmatrix} R(-\phi_j), \qquad \text{where } R(\phi_j) \equiv \begin{pmatrix} \cos\phi_j & -\sin\phi_j \\ \sin\phi_j & \cos\phi_j \end{pmatrix}.$$

Notice that $R(\phi_j)$ and $R(-\phi_j)$ are orthonormal matrices satisfying $R(\phi_j)R(-\phi_j) = I$, where I denotes the identity matrix. $R(\phi_j)$ corresponds to a rotation matrix that rotates the x-y coordinates clockwise by an angle ϕ_j. Solving this equation yields:

$$\begin{aligned}
\psi_{1_j}^2 &= \lambda_{1_j}\cos^2\phi_j + \lambda_{2_j}\sin^2\phi_j, \quad \psi_{2_j}^2 = \lambda_{1_j}\sin^2\phi_j + \lambda_{2_j}\cos^2\phi_j \\
\kappa_j &= \frac{(\lambda_{1_j} - \lambda_{2_j})\sin\phi_j\cos\phi_j}{\sqrt{(\lambda_{1_j}\cos^2\phi_j + \lambda_{2_j}\sin^2\phi_j)(\lambda_{1_j}\sin^2\phi_j + \lambda_{2_j}\cos^2\phi_j)}}.
\end{aligned}$$

The spatial kernel k_s can now be reparametrized using parameters $\lambda_{1_j}, \lambda_{2_j}$, and ϕ_j defined in the above equations. If a constraint is included, such that $\lambda_{1_j} \geq \lambda_{2_j}$, then a very nice geometric interpretation of the parameters λ_{1_j}, λ_{2_j}, and ϕ_j in the bivariate Normal distribution results: $\sqrt{\lambda_{1_j}}$ and $\sqrt{\lambda_{2_j}}$ correspond to the major axis and minor axis of one of the elliptical contours and ϕ_j is the angle between the major axis and the x-coordinate. Swall (1999) suggests additional parameterizations of spatial kernels.

4.2.3 Modeling Periodic Patterns

Many time series data exhibit periodic patterns. To better understand the data and efficiently predict future values, it is imperative to capture periodic features. This section extends the spatio-temporal model to allow cyclic pattern modeling. Following the classic paradigm, the process is decomposed into its periodic and aperiodic components. In the context of the model, this can be done by introducing an indicator variable. For illustrative purposes, we build a model that can capture daily patterns hidden in the data. Our method is general and can be extended to model other cyclic features. To model the daily features embedded in the observed time series data, each jump j is associated with a binary random variable a_j ($a_j \in \{1, 2\}$). Set $\Pr(a_j = 1) = p$ and $\Pr(a_j = 2) = 1 - p$ for some $0 \leq p \leq 1$. Jump j for which $a_j = 1$ appears only once at time τ_j and is considered an aperiodic jump. Jump j for which $a_j = 2$ recurs at the same time of every day and is called a daily periodic jump. The mean spatio-temporal process is modeled by decomposing it into periodic and aperiodic parts:

$$\mu(s,t) \quad = \quad b_0 + b_1 \int_{\Theta_1 \cup \Theta_2} k(s,t,\theta)\, L(d\theta)\,,$$

where $L(d\theta)$ defines a non-stationary Lévy random field on $\Theta = \Theta_1 \cup \Theta_2 = (\mathcal{S} \times \mathcal{T}_1 \times \mathcal{K} \times \{1\}) \cup (\mathcal{S} \times \mathcal{T}_2 \times \mathcal{K} \times \{2\})$. In this example we set $\mathcal{T}_2 \equiv [0, 24)$, since daily patterns are modeled. Without loss of generality, we set $\mathcal{T}_1 \equiv [0, T]$, for $T > 24$. Each point $\theta \in \Theta$ contains the following elements: $\theta = (\sigma, \tau, \lambda_\tau, \lambda_1, \lambda_2, \phi, a)$, where $\sigma \in \mathcal{S}$ denotes jump location, $\tau \in \mathcal{T}_1 \cup \mathcal{T}_2$ denotes jump time, $(\lambda_\tau, \lambda_1, \lambda_2, \phi) \in \mathbb{R}^3_+ \times [0, 2\pi)$ are kernel associated parameters, and $a \in \{1, 2\}$ denotes the type of jump, i.e., periodic or aperiodic jump. Let $\delta \equiv (t - \tau_j) \mod 24$. The contribution of a daily periodic jump j at time t is:

$$\sum_{n=-\infty}^{\infty} e^{-\lambda_{\tau_j}|t-\tau_j-24n|} = \sum_{n=0}^{\infty} e^{-\lambda_{\tau_j}\delta - 24\lambda_{\tau_j} n} + \sum_{n=0}^{\infty} e^{-\lambda_{\tau_j}|\delta-24| - 24\lambda_{\tau_j} n}$$

$$= \quad \frac{e^{-\lambda_{\tau_j}\delta}}{1 - e^{-24\lambda_{\tau_j}}} + \frac{e^{-\lambda_{\tau_j}(24-\delta)}}{1 - e - 24\lambda_{\tau_j}} = \frac{e^{-\lambda_{\tau_j}\delta}\left(1 + e^{-2\lambda_{\tau_j}(12-\delta)}\right)}{1 - e^{-24\lambda_{\tau_j}}}\,.$$

Thus, the temporal kernel in this seasonal model is defined as follows:

$$k_t(t; \tau_j, \lambda_{\tau_j}, a_j) = \begin{cases} e^{-\lambda_{\tau_j}|t-\tau_j|} & a_j = 1 \\ \frac{e^{-\lambda_{\tau_j}\left[(t-\tau_j) \mod 24\right]}\left(1 + e^{-2\lambda_{\tau_j}\left[12 - (t-\tau_j) \mod 24\right]}\right)}{1 - e^{-24\lambda_{\tau_j}}} & a_j = 2\,. \end{cases} \tag{4.3}$$

4.2.4 Multiple Processes Modeling

The adaptive process convolution model introduced in the previous sections can be extended to model multivariate spatio-temporal processes. Since various spatio-temporal processes may be correlated with each other, development of models to capture their correlation is needed. We introduce a bivariate spatio-temporal model and the extension to higher dimensional problems follows similar techniques.

Suppose there are two spatio-temporal processes $\Xi_1(s,t)$ and $\Xi_2(s,t)$. The correlation between the two processes is most naturally introduced by allowing them to share some common Lévy random field in the construction. For example, $\Xi_1(s,t)$ and $\Xi_2(s,t)$ can be constructed as:

$$\begin{aligned}
\Xi_1(s,t) &= b_{01} + b_{11} \int_{\Theta_1 \cup \Theta_3} k_1(s,t;\theta) L(d\theta) \\
\Xi_2(s,t) &= b_{02} + b_{12} \int_{\Theta_2 \cup \Theta_3} k_2(s,t;\theta) L(d\theta)\,,
\end{aligned}$$

where the underlying Lévy random field $L(d\theta)$ on the disjoint subspaces Θ_1, Θ_2 and Θ_3 are all independent. The correlation between $\Xi_1(s,t)$ and $\Xi_2(s,t)$ arises from their shared dependence on $L(d\theta)$ on Θ_3.

This idea clearly extends to multiple space-time fields. To model multiple spatio-temporal processes, define a latent Lévy random field on a union of subspaces $\Theta = \cup \Theta_l$, with kernels $k_l(s,t;\theta)$ supported on the appropriate subspaces. For k processes, a maximum of $2^k - 1$ subspaces are needed, although very often in practice, a much smaller number of subspaces is needed to account for their dependence structure. The implementation of the extension can be conveniently done by assigning each jump with an indicator variable to determine the type of jump.

4.2.5 Gamma Random Field

As discussed in Section 4.1, the prior distribution for $\mu(s,t)$ is constructed from a kernel convolution of Lévy random fields. A Gamma random field is used as an example to illustrate the construction. For discussions of general Lévy process priors, see Tu *et al.* (2005). The motivating example used in Section 4.4 is spatio-temporal data of hourly measurements of SO_2 concentration. A Gamma random field ensures a non-negative $\mu(s,t)$ which is a natural choice for this type of data. A seasonal model that allows modeling daily patterns is used here. Following the model specified in Section 4.2.3, the mean spatio-temporal process is modeled through:

$$\mu(s,t) = b_0 + b_1 \int_{\Theta_1 \cup \Theta_2} k(s,t;\theta) L(d\theta) \ .$$

We start from defining a positive measure $\nu_\epsilon(du, d\theta)$ defined on $\mathbb{R}^+ \times \Theta$ by setting:

$$\nu_\epsilon(du, d\theta) = \left(\alpha_1 I_{\theta\in\Theta_1}(\theta) + \alpha_2 I_{\theta\in\Theta_2}(\theta)\right) u^{-1} e^{-u} I_{(\epsilon,\infty)}(u) du \pi(d\theta) \ ,$$

where $\pi(d\theta)$ defines a finite positive measure on $\Theta = \Theta_1 \cup \Theta_2$. Let N be a Poisson random measure defined on $\mathbb{R}^+ \times \Theta$ with mean measure $\nu_\epsilon(du, d\theta)$. For any bounded Borel measurable function $\psi\colon \Theta \to \mathbb{R}$, define $L[\psi] \equiv \int_\Theta \psi(\theta) L(d\theta)$ as follows:

$$L[\psi] \equiv \int_\Theta \psi(\theta) L(d\theta) \equiv \iint_{\mathbb{R}^+ \times \Theta} u\psi(\theta) N(du, d\theta) \ .$$

Where $N(du, d\theta)$ defines a Poisson random field on $\mathbb{R}_+ \times \Theta$ with intensity measure $\mathbb{E}[N(du, d\theta)] = \nu_epsilon(du, d\theta)$, and we can calculate the characteristic functional of $L[\psi]$ through the Lévy-Khinchine formula:

$$\mathbb{E}\left[e^{iL[\psi]}\right] = \exp\left\{ \iint_{\mathbb{R}_+ \times \Theta} (e^{iu\psi(\theta)} - 1) \nu_\epsilon(du, d\theta) \right\} \ .$$

In particular, for an indicator function $\psi(\theta) = I_B(\theta)$ for any Borel measurable set $B \subset \Theta$,

$$\mathbb{E}\left[e^{iL[\psi]}\right] = \exp\left\{C\int_{\epsilon}^{\infty}(e^{iu}-1)u^{-1}e^{-u}du\right\} ,$$

where $C = (\alpha_1\pi(B\cap\Theta_1) + \alpha_2\pi(B\cap\Theta_2))$. When $\epsilon = 0$, this is the characteristic function for a Gamma random variable, and $L(d\theta)$ defines a Gamma random field on the space Θ.

4.2.6 Prior Distributions

With the Poisson construction for a Lévy random field, $\mu(s,t)$ can be represented using discrete summation:

$$\begin{aligned}
\mu(s,t) &= b_0 + b_1\int_{\Theta} k(s,t;\theta)\, L(d\theta) \\
&= b_0 + b_1\iint_{\mathbb{R}_+\times\Theta} u\ k(s,t;\theta)N(du,d\theta) \\
&= b_0 + b_1\sum_{j=1}^{J} u_j k(s,t;\theta_j)\ ,
\end{aligned}$$

where $J \sim \mathsf{Po}\left(\nu_\epsilon(\mathbb{R}_+\times\Theta)\right)$. Let $E_1(\epsilon) = \int_\epsilon^\infty u^{-1}e^{-u}du$ and let the area of $\mathcal{S}$ be S. Since $\pi(d\theta)$ defines a finite measure on Θ,

$$\iint_{\mathbb{R}_+\times\Theta}\nu_\epsilon(du,d\theta) = (T\alpha_1 + 24\alpha_2)S\int_\epsilon^\infty u^{-1}e^{-u}du = (T\alpha_1+24\alpha_2)SE_1(\epsilon).$$

For any $\epsilon > 0$, it can be shown that $E_1(\epsilon) < \infty$, but for $\epsilon = 0$, $E_1(0) = \infty$. In practice, it is only feasible to implement a model that has finitely many parameters. To ensure a finite J (in probability) in the summation, a positive number ϵ

is chosen. With Poisson construction of a Lévy random field, our spatio-temporal model can be represented in the following hierarchical form:

$$
\begin{aligned}
Z(s_i,t_i) \mid \mathbf{u},\boldsymbol{\theta},J,\epsilon &\overset{iid}{\sim} \mathsf{N}(\mu(s_i,t_i),\eta) \qquad \text{for } i=1,\cdots,n \\
\mu(s_i,t_i) &= b_0 + b_1\sum_{j=1}^{J} k(s_i,t_i;\theta_j)u_j \qquad \mu(s_i,t_i)\geq 0 \\
(u_j,\theta_j)\mid J,\epsilon &\overset{iid}{\sim} \pi(u_j\mid\epsilon)\pi(\theta_j) \qquad \text{for } j=1,\cdots,J \\
\pi(u_j\mid\epsilon) &= \frac{u_j^{-1}e^{-u_j}}{E_1(\epsilon)} I_{[\epsilon,\infty)}(u_j)\ ,\ u_j\in[\epsilon,\infty) \\
\pi(\theta_j) &= \mathsf{Un}(\sigma_j;\mathcal{S})\left(\mathsf{Un}(\tau_j;\mathcal{T}_1)I_{\{a_j=1\}}+\mathsf{Un}(\tau_j;\mathcal{T}_2)I_{\{a_j=2\}}\right)\mathsf{Ga}(\lambda_{\tau_j};\alpha_\tau,\beta_\tau) \\
&\quad \mathsf{Ga}(\lambda_{1_j};\alpha_{\lambda_1},\beta_{\lambda_1})\mathsf{Ga}(\lambda_{2_j};\alpha_{\lambda_2},\beta_{\lambda_2})\mathsf{Un}(\phi_j;0,2\pi)p^{I_{\{a_j=1\}}}(1-p)^{I_{\{a_j=2\}}} \\
J\mid\boldsymbol{\alpha},\epsilon &\sim \mathsf{Po}\left((T\alpha_1+24\alpha_2)E_1(\epsilon)S\right)\ ,\ J\in\mathbb{N}\cup\{0\} \\
\pi(\eta) &\propto \frac{1}{\eta}\ , \quad \eta>0 \\
\pi(b_0) &= \mathsf{Ga}(\alpha_{b_0},\beta_{b_0}) \\
\pi(b_1) &= \mathsf{Ga}(\alpha_{b_1},\beta_{b_1})\ .
\end{aligned}
$$

The baseline level is modeled by b_0. Hyperparameters $(\alpha_{b_0},\beta_{b_0})$ are chosen such that $\mathbb{E}[b_0]$ is close to the sample mean of the observations $\mathbf{z}$, with relatively large variance of b_0. The overall size of latent jump is controlled by b_1. Hyperparameters $(\alpha_{b_1},\beta_{b_1})$ are chosen such that $\mu(s,t)$ covers the range of the observations.

Hyperparameters $\zeta=(\alpha_\tau,\beta_\tau,\alpha_{\lambda_1},\beta_{\lambda_1},\alpha_{\lambda_2},\beta_{\lambda_2},\alpha_1,\alpha_2,p,\epsilon)$ are associated with θ. The minimum size of latent jump is controlled by ϵ. Recall that the sizes of latent jumps $\{u_j\}$ are unit free. We choose $\epsilon=0.15$ in this example, but we varied ϵ over a wide range and found no apparent sensitivity to the choice of ϵ. For a given $\epsilon>0$, α_1 and α_2 control the average number of jumps of size $u_j>\epsilon$.

Gamma distributions are used as priors for α_1 and α_2 such that the posterior distribution is in closed form. In order to calculate posterior distributions of α_1 and α_2, two auxiliary variables J_1 and J_2 are introduced. Let $J_1 \equiv \sum_{j=1}^{J} I_{\{a_j=1\}}$ and $J_2 \equiv \sum_{j=1}^{J} I_{\{a_j=2\}}$, i.e., J_1 is the number of jumps in $[\epsilon, \infty) \times \Theta_1$ and J_2 is the number of jumps in $[\epsilon, \infty) \times \Theta_2$. Assume $\mathbb{E}[J_1|\alpha_1] = \nu(\mathbb{R}_+ \times \Theta_1) = TS\alpha_1 E_1(\epsilon)$ and $\mathbb{E}[J_2|\alpha_2] = \nu(\mathbb{R}_+ \times \Theta_2) = 24S\alpha_2 E_1(\epsilon)$. Conditional on $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$, p can be solved through:

$$\mathbb{E}[J_1] = \mathbb{E}\left[\mathbb{E}\left[\sum_{j=1}^{J} I_{a_j=1}|J\right]\right] = \mathbb{E}\left[\sum_{j=1}^{J} Pr(a_j = l)\right] = \mathbb{E}[Jp] = p\mathbb{E}[J] \ ,$$

such that,

$$p = \frac{\mathbb{E}[J_1]}{\mathbb{E}[J]} = \frac{T\alpha_1}{T\alpha_1 + 24\alpha_2} \ .$$

By the definition of J_1 and J_2, $J_1|J, \boldsymbol{\alpha} \sim \mathsf{Bin}(J, p)$ and $J_2|J, \boldsymbol{\alpha} \sim \mathsf{Bin}(J, 1-p)$. Since $J \sim \mathsf{Po}((T\alpha_1 + 24\alpha_2)SE_1(\epsilon))$, it can be derived that,

$$\pi(J_1|\boldsymbol{\alpha}) \sim \mathsf{Po}(\alpha_1 T E_1(\epsilon)) \quad \text{and} \quad \pi(J_2|\boldsymbol{\alpha}) \sim \mathsf{Po}(\alpha_2 24 E_1(\epsilon)).$$

If a Gamma distribution is specified as the prior for α_1 and α_2, i.e.,

$$\pi(\alpha_1) \sim \mathsf{Ga}(a_{\alpha_1}, b_{\alpha_1}) \quad \text{and} \quad \pi(\alpha_2) \sim \mathsf{Ga}(a_{\alpha_2}, b_{\alpha_2}),$$

then,

$$\alpha_1|J_1 \sim \mathsf{Ga}(a_{\alpha_1} + J_1, b_{\alpha_1} + TSE_1(\epsilon)) \quad \text{and} \quad \alpha_2|J_2 \sim \mathsf{Ga}(a_{\alpha_2} + J_2, b_{\alpha_2} + 24SE_1(\epsilon)).$$

Here we give some guidance for the choice of the values for a_{α_1}, b_{α_1}, a_{α_1} and b_{α_1}. Let $p_1 \equiv \frac{b_{\alpha_1}}{b_{\alpha_1} + E_1(\epsilon)TS}$, and $p_2 \equiv \frac{b_{\alpha_2}}{b_{\alpha_2} + 24E_1(\epsilon)S}$, and a priori, the marginal distribution for J_1 and J_2 is Negative-Binomial (See the appendix C.1 for further details):

$$J_1 \sim \mathsf{NB}(a_{\alpha_1}, p_1) \quad \text{and} \quad J_2 \sim \mathsf{NB}(a_{\alpha_2}, p_2).$$

In order to favor parsimonious models, we control the size of J a priori. Specifically, a_{α_1}, a_{α_2}, b_{α_1} and b_{α_2} are chosen so that the mean and variance of J_1 and J_2 are relatively small. This usually requires a very large b_{α_1} and b_{α_2} and relatively small a_{α_1} and a_{α_2}. By choosing prior mean and variance for J_1 and J_2 we can solve the corresponding a_{α_1}, a_{α_2}, b_{α_1} and b_{α_2}. There is trade-off between the periodic and the aperiodic parts. Based on the model formulation, the decomposition of $\mu(s,t)$ into periodic and aperiodic parts is not unique. An extreme case is that the periodic part can be dropped and the model still fits the data very well using only the aperiodic part. The purpose of introducing the periodic part is to offer insights to the data generation mechanism and to facilitate prediction, which we discuss in Section 4.3. Simulation experiments show that in order to capture the periodic pattern embedded in the data effectively, large value of J_1 needs to be penalized more than large J_2 through appropriate prior specification for J_1 and J_2, i.e., make $a_{\alpha_1}/b_{\alpha_1} \gg a_{\alpha_2}/b_{\alpha_2}$. (Notice that $E_1(\epsilon)a_{\alpha_1}/b_{\alpha_1}$ is the prior mean of the number of jumps in a unit interval in space $\mathbb{R}_+ \times \Theta_1$, and $E_1(\epsilon)a_{\alpha_2}/b_{\alpha_2}$ is the prior mean of the number of jumps in a unit interval in space $\mathbb{R}_+ \times \Theta_2$.)

We now discuss the choice of hyperpriors for hyperparameters $(\alpha_\tau, \beta_\tau, \alpha_{\lambda_1}, \beta_{\lambda_1}, \alpha_{\lambda_2}, \beta_{\lambda_2})$. One important feature of the spatio-temporal model is the adaptiveness of the spatio-temporal kernel. Instead of having a universal kernel for all time and locations, we allow the kernel to be time and location specific. This is particularly useful to model data that show strong space and time heterogeneity. Placing appropriate prior distributions on the kernel associated parameters $\{\lambda_{\tau_j}, \lambda_{1_j}, \lambda_{2_j}\}_{j \le J}$ is an important part of the model specification. We parametrize the prior distributions in a hierarchical fashion, i.e., instead of choosing specific values for $(\alpha_\tau, \beta_\tau, \alpha_{\lambda_1}, \beta_{\lambda_1}, \alpha_{\lambda_2}, \beta_{\lambda_2})$, prior distributions are specified for each. There are several advantages to hyperprior specification. First, the hierarchical construction can alleviate the over-fitting associated with many nonparametric methods. Second, the hierarchical priors introduce dependency structure among parameters which is more realistic than using iid priors for multiple parameters. Finally, using hierarchical priors for kernel parameters facilitates predictions. Because the kernel is adaptive, posterior distribution of kernel parameters at location and time points where no observations exist coincides with their prior distribution. One remedy is to use hierarchical priors for kernel parameters. The information associated with kernel parameters at location and time points where no observations occur is gathered indirectly through hyperparameters $(\alpha_\tau, \beta_\tau, \alpha_{\lambda_1}, \beta_{\lambda_1}, \alpha_{\lambda_2}, \beta_{\lambda_2})$. We discuss the choice of hyperpriors for $(\alpha_\tau, \beta_\tau)$, and other hyperpriors can be chosen in a similar fashion. Simulation study shows that it is difficult to identify both α_τ and β_τ. As a result, we choose to fix α_τ and assign a log Normal prior for β_τ. The hierarchical prior for the temporal kernel parameter λ_τ is:

$$\begin{aligned} \lambda_{\tau_j} | J, \alpha_\tau, \beta_\tau &\overset{iid}{\sim} \mathsf{Ga}(\alpha_\tau, \beta_\tau), \quad \text{for } j = 1, 2, \cdots, J \\ \log(\beta_\tau) &\sim \mathsf{N}(m_b, v_b)\,. \end{aligned}$$

Let $\lambda_{0.1}$, $\lambda_{0.5}$ and $\lambda_{0.9}$ denote the 10%, 50% and 90% quantiles of λ_{τ_j}, respectively. Since λ_{τ_j} controls the decay rate in the temporal domain for the underlying point source, it is natural to obtain an estimate of $\lambda_{0.1}$, $\lambda_{0.5}$ and $\lambda_{0.9}$ a priori from experts' opinion or historical data. The 10%, 50% and 90% quantiles of λ_{τ_j} can be matched with the desired values by choosing the appropriate α_τ, m_b and v_b. Solving this analytically is not feasible, so a numerical procedure is used.

4.2.7 Posterior Distributions

Given observations $\{z_i\}_{i=1}^n$, the joint posterior distribution of all unknown parameters is:

$$
\begin{aligned}
p(\mathbf{u}, \boldsymbol{\theta}, J, \boldsymbol{\alpha}, b_0, b_1, \eta^2 \mid \mathbf{Z}) \quad \propto \quad & \left(\frac{1}{\eta}\right)^{-\frac{n}{2}-1} \prod_{i=1}^{n} \exp\left\{ -\frac{1}{2\eta} \left(Z_i - b_0 - b_1 \sum_{j=1}^{J} k(s_i, t_i; \theta_j) u_j \right)^2 \right\} \\
& \frac{e^{-\mu(\nu,\alpha)}}{J!} \left\{ \prod_{j=1}^{J} \nu_\epsilon(du_j, d\theta_j) \right\} \pi(\boldsymbol{\alpha}) \pi(b_1) \pi(b_0) \qquad (4.4)
\end{aligned}
$$

The posterior distributions for all the parameters, except for η, are not in closed form. Because J is random, the dimension of the parameter space is not fixed. A reversible jump MCMC scheme is therefore used to draw samples from the posterior distributions of the parameters.

4.3 Predictions

A primary goal for building this spatio-temporal model is prediction. Our interest focuses on two types of predictive distributions. One is the predictive distribution at a location that has no observation but at a time which observations at other locations occur: this predictive distribution is called *spatio-temporal interpolation.* The other is the predictive distribution at a future time for any location, which is referred to as *spatio-temporal extrapolation.* In this section, we propose a scheme to handle both types of predictions.

We begin discussion with the spatio-temporal interpolation. Let $D \equiv (z_1, z_2, \cdots, z_n)$ denote the n observations where $z_i = Z(s_i, t_i)$, $s_i \in \mathcal{S}$ and $t_i \in \mathcal{T}_1$ for $i = 1, 2, \cdots, n$. We are interested in the posterior predictive distribution of $Z_0 \equiv Z(s_0, t_0)$ at a new point (s_0, t_0), $s_0 \in \mathcal{S}$ and $t_0 \in \mathcal{T}_1$. Based on the spatio-temporal model, $Z_0 = \mu_0 + \epsilon_0$, where $\mu_0 \equiv \mu(s_0, t_0) = b_0 + b_1 \sum_{j=1}^{J} k(s_0, t_0; \theta_j) u_j$ and $\epsilon_0 \sim \mathsf{N}(0, \eta)$. We use $[Z \mid X]$ to denote the conditional density function of Z given X. Let $\omega \equiv (b_0, b_1, \{u_j, \theta_j\}_{j \leq J}, J)$ denotes all the parameters on which μ_0 depends, the posterior predictive distribution can be represented by:

$$[Z_0 \mid D] = \int [Z_0 \mid \mu_0, \eta, D][\mu_0 \mid \omega, D][\omega \mid D][\eta \mid D] \; d\mu_0 d\omega d\eta \tag{4.5}$$

Dependent samples from $[Z_0 \mid D]$ can be drawn using a Monte Carlo approach. From the MCMC simulation, we obtain $N_p \equiv \lfloor (N_r - N_b)/\Delta \rfloor$ samples from the joint posterior distribution $[\omega, \eta \mid D]$, and we denote them by $\{(\tilde{\omega}_i, \tilde{\eta}_i)\}_{i=1}^{N_p}$, with which, we then obtain samples from the posterior predictive distribution, $[z_0|D]$.

We first take a posterior sample $\tilde{\omega}_i$, from which we calculate $\tilde{\mu}_{0_i}$ by setting:

$$\tilde{\mu}_{0_i} = \tilde{b}_{0_i} + \tilde{b}_{1_i} \sum_{j=1}^{\tilde{J}_i} \tilde{u}_{j_i} k(s_0, t_0; \tilde{\theta}_{j_i}) \, ,$$

where $\tilde{u}_{j_i}$ and $\tilde{\theta}_{j_i}$ denote posterior samples of u_j and θ_j from the i^{th} thinned iteration. The spatio-temporal kernel $k(s, t; \theta)$ takes the separable form $k(s, t; \theta) = k_t(t; ...) k_s(s; ...)$ as in Eqn. (4.2), and the temporal part takes the form as in Eqn. (4.3). We then draw $\tilde{z}_{0_i}$ from a Normal distribution with mean $\tilde{\mu}_{0_i}$ and variance $\tilde{\eta}_i$. Repeat this process N_p times to obtain N_p dependent samples from the posterior predictive distribution $[z_0|D]$, which we can use to estimate the predictive mean and predictive intervals, among other results.

The spatio-temporal extrapolation is slightly more complicated. In order to proceed, an efficient algorithm to sample from a Lévy random field is needed. The Inverse Lévy Measure (ILM) algorithm developed by Wolpert and Ickstadt (1998) offers a very efficient scheme to sample from a wide class of Lévy random fields.

Appendix C.3 describes the procedure we use here and more examples can be found in the work cited. For spatio-temporal extrapolation, our interest focuses on the predictive distribution of $[z_0|D]$ at a point $z_0 = z(s_0, t_0)$, $s_0 \in \mathcal{S}$ but $t_0 \notin \mathcal{T}_1$. Once again, we take $\mathcal{T}_1 = [0, T]$ and $\mathcal{T}_2 = [0, 24]$ with $T > 24$, and choose some $\Delta_T > 0$ such that $t_0 \in (T, T + \Delta_T]$. In order to obtain samples from the posterior predictive distribution $[z_0 \mid D]$, samples must be drawn from $[\mu(s_0, t_0) \mid D]$. The predictive mean $\mu(s_0, t_0)$ can be represented as the following stochastic integral:

$$\begin{aligned}
\mu(s_0, t_0) &= b_0 + b_1 \int_{\mathcal{S}\times[0,T+\Delta_T]\times\mathcal{K}\times\{1,2\}} k(s_0, t_0; \theta) L(d\theta) \\
&= b_0 + b_1 \int_{\mathcal{S}\times[0,T]\times\mathcal{K}\times\{1,2\}} k(s_0, t_0; \theta) L(d\theta) \\
&\qquad + b_1 \int_{\mathcal{S}\times(T,T+\Delta_T]\times\mathcal{K}\times\{1\}} k(s_0, t_0; \theta) L(d\theta) \quad (*)
\end{aligned}$$

Using posterior samples of $\theta \in \mathcal{S} \times [0, T] \times \mathcal{K} \times \{1, 2\}$, samples can be drawn from the first stochastic integral in equation $(*)$ as we did for spatio-temporal interpolation. Since we have no observations in $\mathcal{S} \times (T, T + \Delta_T]$, the posterior distribution of $\theta \in \mathcal{S} \times (T, T + \Delta_T] \times \mathcal{K} \times \{1, 2\}$ is just the prior distribution (more precisely, it is the prior distribution of θ conditional on the hyperparameters for which posterior samples are collected). So, we propose the following way to sample from $[z_0 \mid D]$:

$$\tilde{\mu}_{0_i} = \tilde{b}_{0_i} + \tilde{b}_{1_i} \left\{ \sum_{j=1}^{\tilde{J}_i} \tilde{u}_{j_i} k(s_0, t_0; \tilde{\theta}_{j_i}) + \sum_{m=1}^{\breve{M}_i} \breve{u}_{m_i} k(s_0, t_0; \breve{\theta}_{m_i}) \right\} . \tag{4.6}$$

In the first summation of Eqn. (4.6), as in the spatio-temporal interpolation, $\tilde{b}_{0_i}$, $\tilde{b}_{1_i}$, $\tilde{J}_i$, $\tilde{u}_{j_i}$ and $\tilde{\theta}_{j_i}$ are posterior samples from the i^{th} thinned iteration. For the second summation, first draw $\breve{M}_i$ from a Poisson distribution $\mathsf{Po}(\breve{\alpha}_{1_i} \Delta_T S E_1(\epsilon))$ where

$\breve{\alpha}_{1_i}$ is a posterior sample from the i^{th} thinned iteration. Given $\breve{M}_i$, $\{(\breve{u}_{m_i}, \breve{\theta}_{m_i})\}_{m=1}^{\breve{M}_i}$ are generated from the Lévy process, $L(d\theta)$, defined on $\theta \in \mathcal{S} \times (T, T + \Delta_T] \times \mathcal{K}$ with hyperparameters $(\breve{\beta}_{\tau_i}, \breve{\beta}_{\lambda_{1_i}}, \breve{\beta}_{\lambda_{2_i}})$ as posterior samples from the i^{th} thinned iteration, i.e.: $\breve{u}_{m_i} \sim \pi(u)$, where $\pi(u) = \frac{u^{-1}e^{-u}}{E_1(\epsilon)} I_{(\epsilon,\infty)}(u)$. (See the appendix C.3 for details on how to sample from $\pi(u)$). Furthermore, $\breve{\tau}_{m_i} \sim \mathsf{Un}(T, T + \Delta_T)$, $\breve{\sigma}_{m_i} \sim \mathsf{Un}(\mathcal{S})$, $\breve{\lambda}_{\tau_{m_i}} \sim \mathsf{Ga}(\alpha_\tau, \breve{\beta}_{\tau_i})$, $\breve{\lambda}_{1_{m_i}} \sim \mathsf{Ga}(\alpha_{\lambda_1}, \breve{\beta}_{\lambda_{1_i}})$, $\breve{\lambda}_{2_{m_i}} \sim \mathsf{Ga}(\alpha_{\lambda_2}, \breve{\beta}_{\lambda_{2_i}})$, $\breve{\phi}_{j_i} \sim \mathsf{Un}(0, 2\pi)$ and $\breve{a}_{m_i} = 1$ (since this is for the aperiodic part), for $m = 1, \cdots, \breve{M}_i$. After obtaining a sample, $\tilde{\mu}_{0_i}$, draw $\tilde{z}_{0_i}$ from a Normal distribution, $\mathsf{N}(\tilde{\mu}_{0_i}, \tilde{\eta}_i)$, where $\tilde{\eta}_i$ is a posterior sample from the i^{th} thinned iteration. Repeat the above steps for $i = 1, 2, \cdots, N_p$, and obtain N_p samples from the posterior predictive distribution $[z_0 \mid D]$.

4.4 Illustrative Examples

In this section, we consider two examples to illustrate the proposed Bayesian spatio-temporal model. The posterior distributions of the parameters are investigated using a reversible-jump Markov chain Monte Carlo algorithm(Green, 1995), since the dimension of the parameter space varies. Tu *et al.* (2005) provides a detailed computational algorithm.

4.4.1 A Simulated Example

We first illustrate the spatio-temporal model with a simulation study. Observations at 33 locations (labeled from 1 to 33 in Fig. (4.1)) are generated on a $[0, 31] \times [0, 31]$ square. The sample frequency is one hour. Four days of hourly data (96 in total) are collected at each location. At time t_i and location $s_i = (x_i, y_i)$

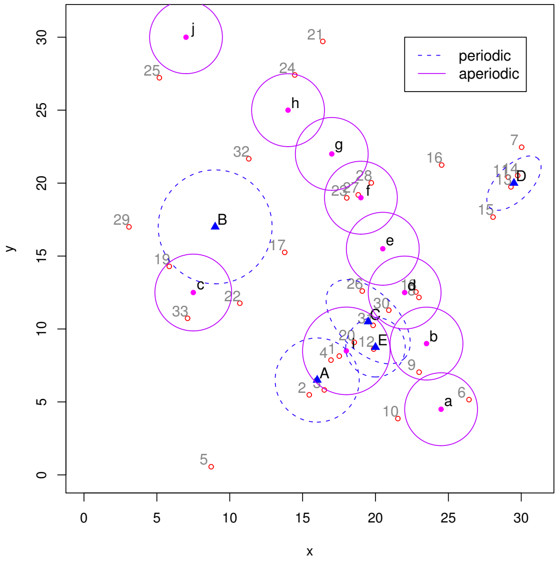

Figure 4.1: Locations of latent point sources that generate observations. Open circles 1-33 denote locations where observations are collected. Solid circles *a*-*j* denote locations of aperiodic latent point sources and solid triangles *A*-*E* denote locations of daily periodic latent point sources. The ellipsis associated with each point source is used to delineate the contours of the bivariate Normal spatial kernel where the signal weakens to 50% of its maximum intensity at the boundary.

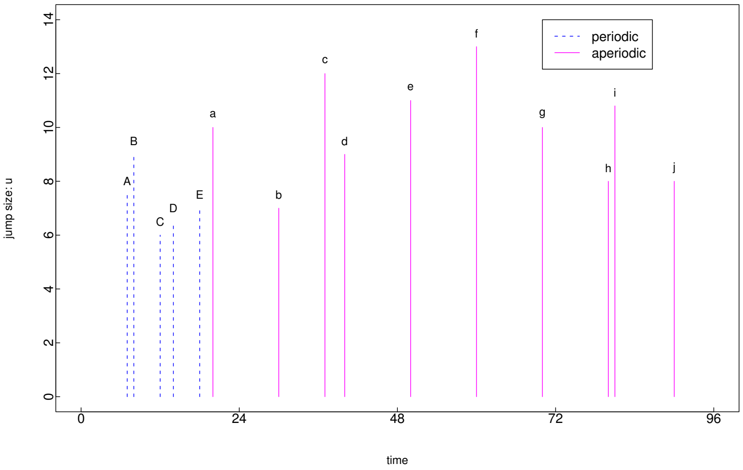

Figure 4.2: Time and magnitudes of point sources that generate observations. The solid lines are aperiodic point sources and the dashed lines are periodic point sources.

the observation $z_i = z(s_i, t_i)$ is generated by:

$$z_i = \sum_{j=1}^{15} u_j \; k_t(t_i;\; \tau_j, a_j, \lambda_{\tau_j}) k_s(s_i; \sigma_j, \lambda_{1_j}, \lambda_{2_j}, \phi_j) + \epsilon_i \,, \tag{4.7}$$

where $\epsilon_i \overset{iid}{\sim} \mathsf{N}(0, 0.2^2)$. The form of spatial kernel k_s is discussed in Section 4.2.2 and the temporal kernel k_t is specified in Eqn. (4.3). The locations and contours of the spatial kernels of the latent point sources are shown in Fig. (4.1). The time of the point sources are shown in Fig. (4.2). The values we used for all the parameters in Eqn. (4.7) are listed in Table C.1 in the appendix C.2. The data are driven by 15 underlying point sources. Five are daily periodic point sources which occur at a certain time every day and are labeled from A to E. Ten of the point sources are aperiodic that occur only once in the entire study period and are labeled from a to j. Fig. (4.3) shows the time series generated by the Eqn. (4.7) on four of the 33 sites. Note that the time series generated at different sites have very different features.

We fit the model in Section 4.2.6 to the simulated data set. The performance is examined through the following criteria. We first check model fit in temporal dimension through plotting the true mean function at each location and superimposing the posterior mean function. The results are shown at four locations in Fig. (4.3). We then check the fit in spatial dimension by comparing the true spatial surface and the posterior mean spatial surface at different time points. The result at $t = 36$ is shown in Fig. (4.6). Both results portray a good fit for the model.

A more challenging task is to recover the latent process which can be extremely hard for non-adaptive kernel methods. The posterior distribution of the latent

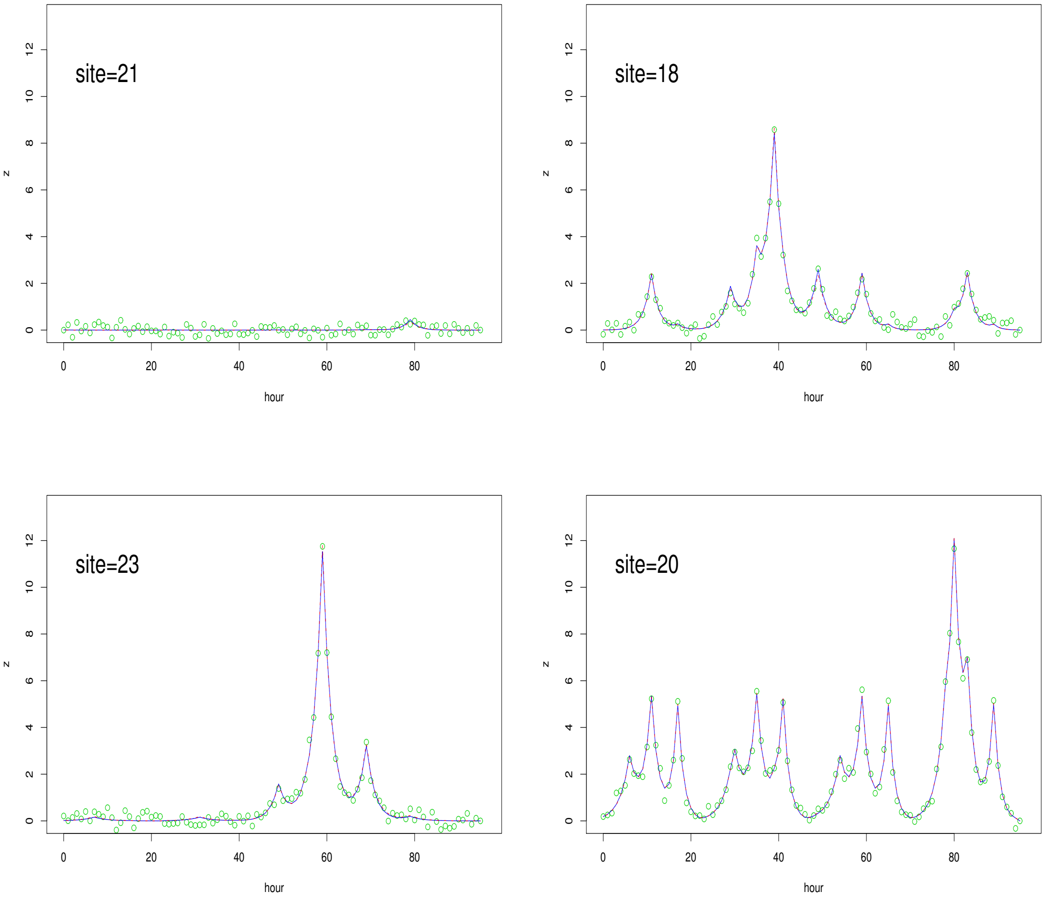

Figure 4.3: Fitted results at four locations: the open circles are observations, the dashed line is true mean function and the solid line is the fitted mean function.

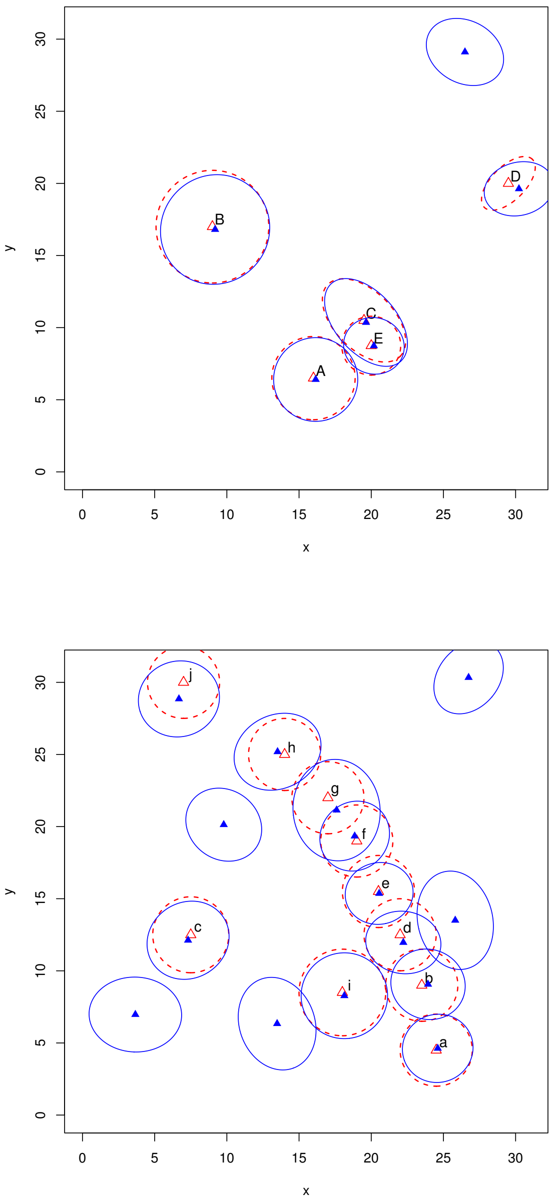

Figure 4.4: Locations and spatial kernels of latent point sources. The left diagram is for aperiodic point sources and the right plot is for daily periodic point sources. The solid triangles and solid ellipses are locations and spatial kernels respectively, drawn from their posterior distributions; and the open triangle and open ellipses are true locations and spatial kernels of the latent point sources respectively.

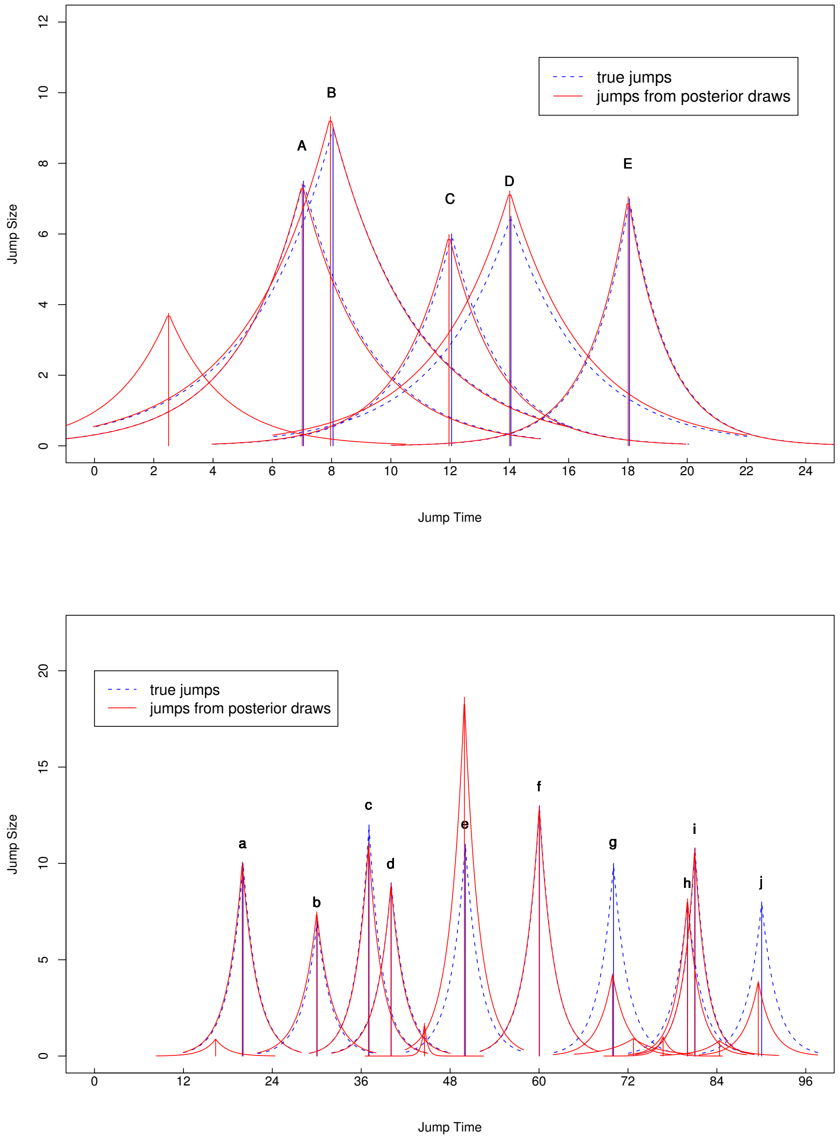

Figure 4.5: Times and temporal kernels of latent point sources. The left diagram is for aperiodic point sources and the right plot is for daily periodic point sources. The x axis represents jump time and y axis represents jump magnitude. The solid lines and solid curves are times and temporal kernels respectively, drawn from their posterior distributions, and the dashed lines and dashed curves are true times and temporal kernels of the latent point sources respectively.

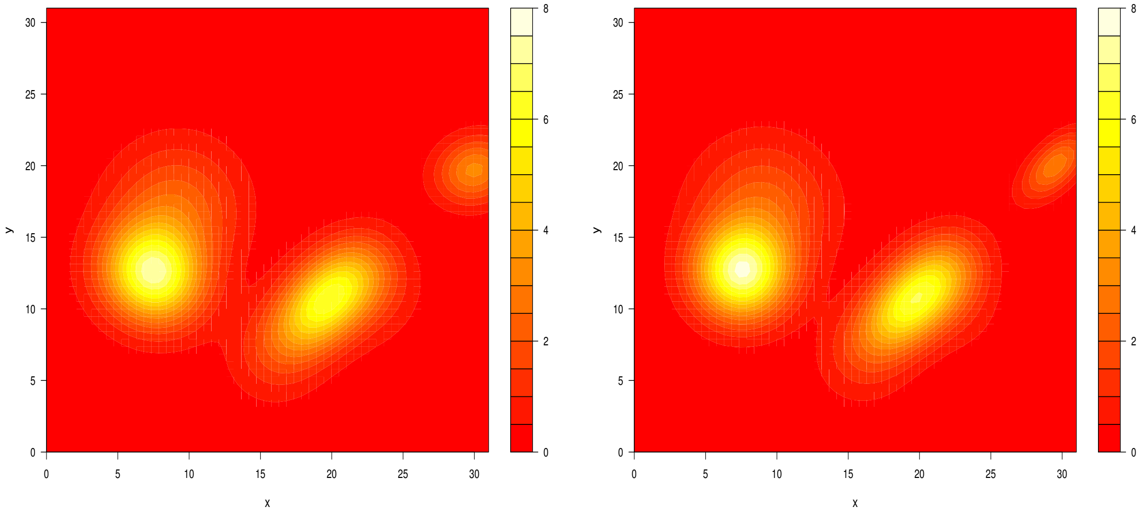

Figure 4.6: A snapshot of the spatial surface at $t = 36$. The left diagram is the posterior mean spatial surface and the right plot is the true spatial surface.

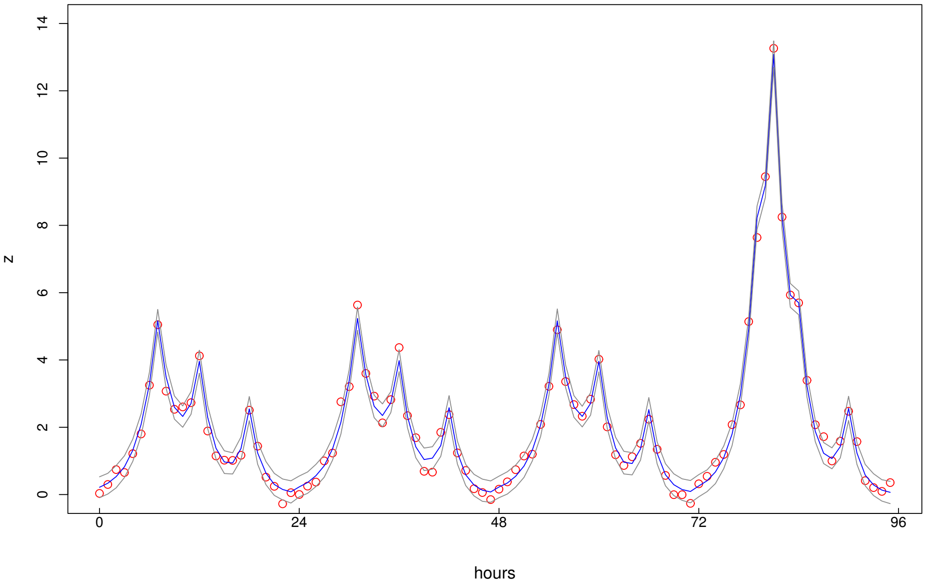

Figure 4.7: Leave-one-out prediction. The open circles represent observations. The solid line is the predictive mean curve and the two grey lines are 5% and 95% predictive quantiles respectively. The coverage of the 90% Bayesian credible interval is 86.5% and the RMSE is 0.224.

process is compared with the true one. In Fig. (4.4), the true locations and spatial kernels of the latent point sources are plotted and superimposed with one sample from the posterior distribution. In Fig. (4.5) the true time and temporal kernels of the latent point sources are plotted and superimposed with one sample from the posterior distribution. Plots like Fig. (4.4) and Fig. (4.5) can be reproduced using additional posterior samples. Animations created from these plots are available from the author's website. The results indicate that the true latent process can be recovered very well. We found all 15 point sources around the right time and locations. Spatial and temporal kernels were also identified very well. Not surprisingly, the latent periodic signals were identified more accurately than the latent aperiodic signals. Notice that several false positive signals exist, however the magnitudes of the false positive signals are usually small compared to the true signals. We finish the simulation example with a leave-one-out prediction study to test the performance of the model. The model is fitted without the data at location 31 and a prediction of the observations at location 31 for the entire study period (96 hours in total) using the prediction method we described in Section 4.3 is conducted. Fig. (4.7) shows the results.

4.4.2 SO_2 Monitoring Data

In this section we test the performance of the spatio-temporal model on a real data set collected by the Environmental Protection Agency (EPA). The data consist of hourly sulfur dioxide (SO_2) concentration levels (ppm) taken at 33 locations across Pennsylvania, New Jersey, Delaware and Maryland. The locations of the monitoring stations are shown in Fig. (4.8). A Lambert projection method is used to reduce the distortion caused by the curvature of the earth's surface. The

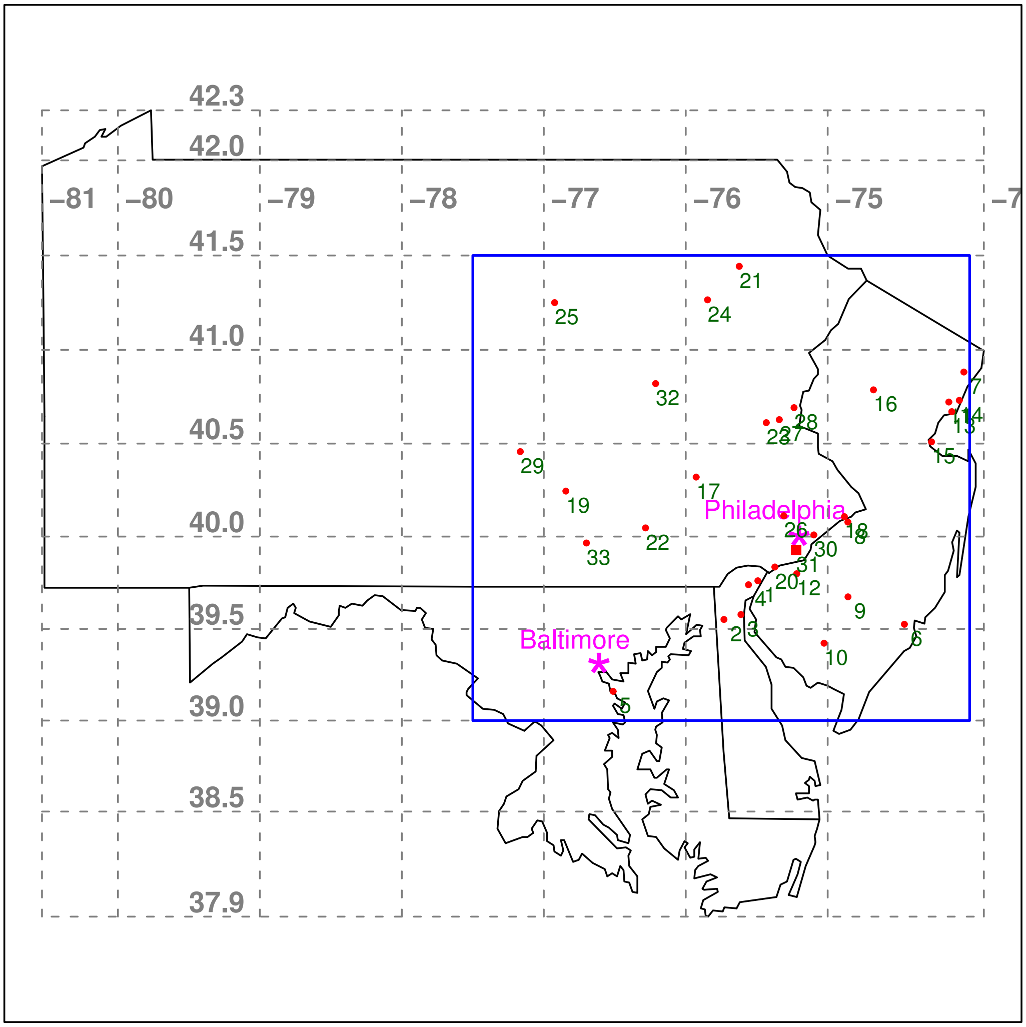

Figure 4.8: The 33 circles represent 33 monitors used by the EPA to measure hourly SO_2 concentration in the year 2002. Monitors used in this study are in the area delineated by the rectangle, which includes part of the following four states: PA, MD, NJ and DE. The grey grid is the actual longitude and latitude, and the map is drawn after *Lambert projection* is implemented. The solid square denotes site 31.

coordinates used in the analysis are rescaled Lambert coordinates. The study region delineated by a rectangle in Fig. (4.8) covers a 310 km × 310 km area. For demonstration purposes, we restrict our analysis to 144 consecutive hours from September of 2002. About 5% of SO_2 readings are missing from the data set. This is not a problem in the setting of our model; the missing values can be either omitted from the likelihood calculation or treated as unknown parameters. The first approach is taken in this paper. We select eight sites and plot the observed time series at each site in Fig. (4.9). Notice that the time series at different sites have very different features. This indicates that the target spatio-temporal field can be highly non-stationary. Another feature of the data is the spikes embedded in the time series. The spikes represent a high concentration level of SO_2. Modeling the spikes is one of the primary goals of the study, since an important question of future consideration is whether high level of SO_2 is associated with a high incidence of human respiratory disease. One common approach to modeling air pollutants data is conducted using Gaussian processes. Since the mean function to model is strictly positive, a log transformation of the data is usually necessary. But a log transformation will eliminate the spiky feature from the data. As a result, a Gaussian process approach may not be an ideal choice. On the other hand, the adaptive kernel approach developed in this paper can be used to model non-negative process directly and is especially good for modeling this type of non-stationary non-Gaussian data.

We fit the SO_2 data using the spatio-temporal model we developed in Section 4.2.6. The MCMC algorithm was run for 75000 iterations. The first 50000 iterations were discarded as burn-in with the remaining used for posterior inference. Convergence was diagnosed by examining trace plots of the log likelihood.

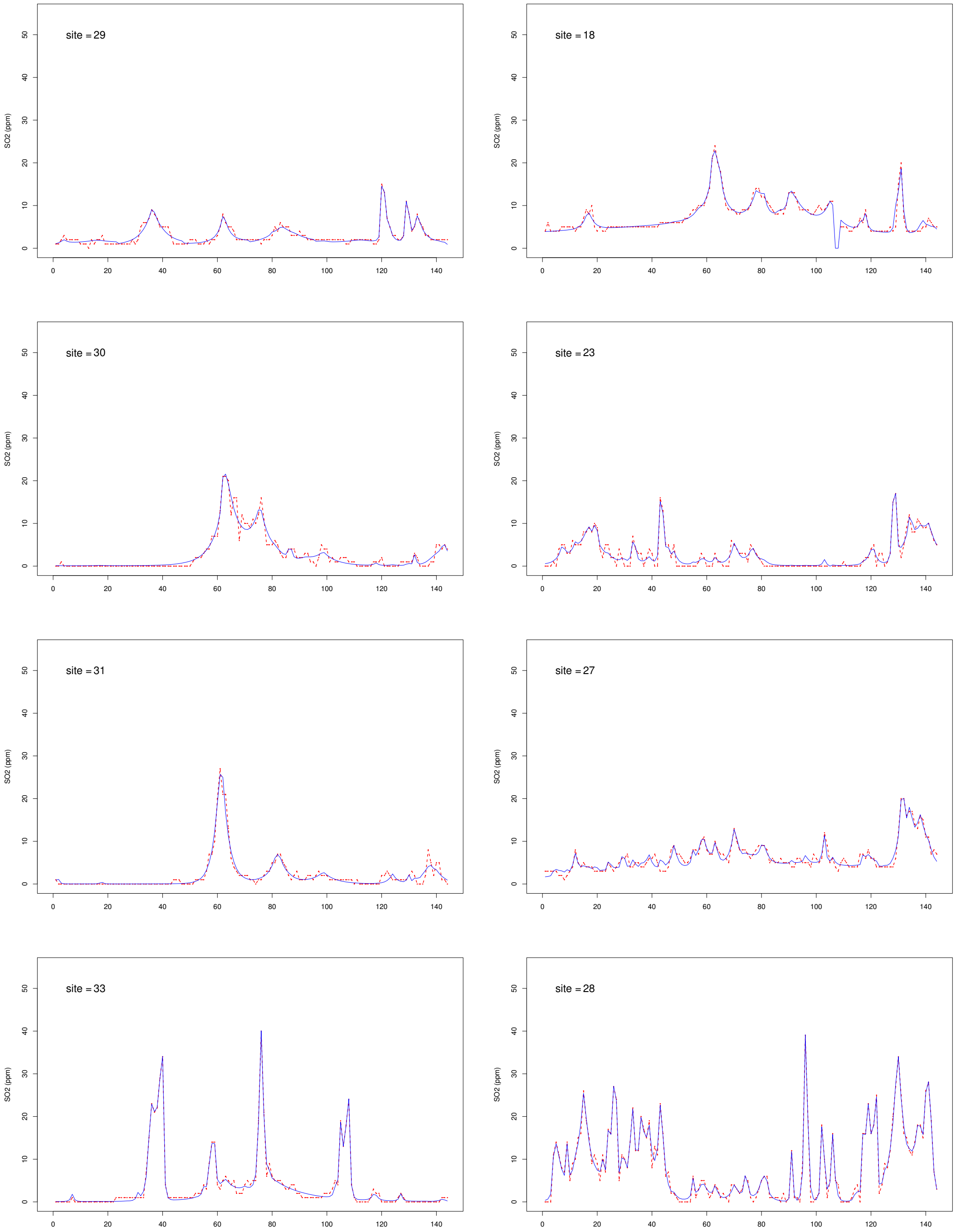

Figure 4.9: Model fitting of SO_2 at 8 out of 33 monitors. The dashed line represents the observations, and the solid line represents the posterior mean.

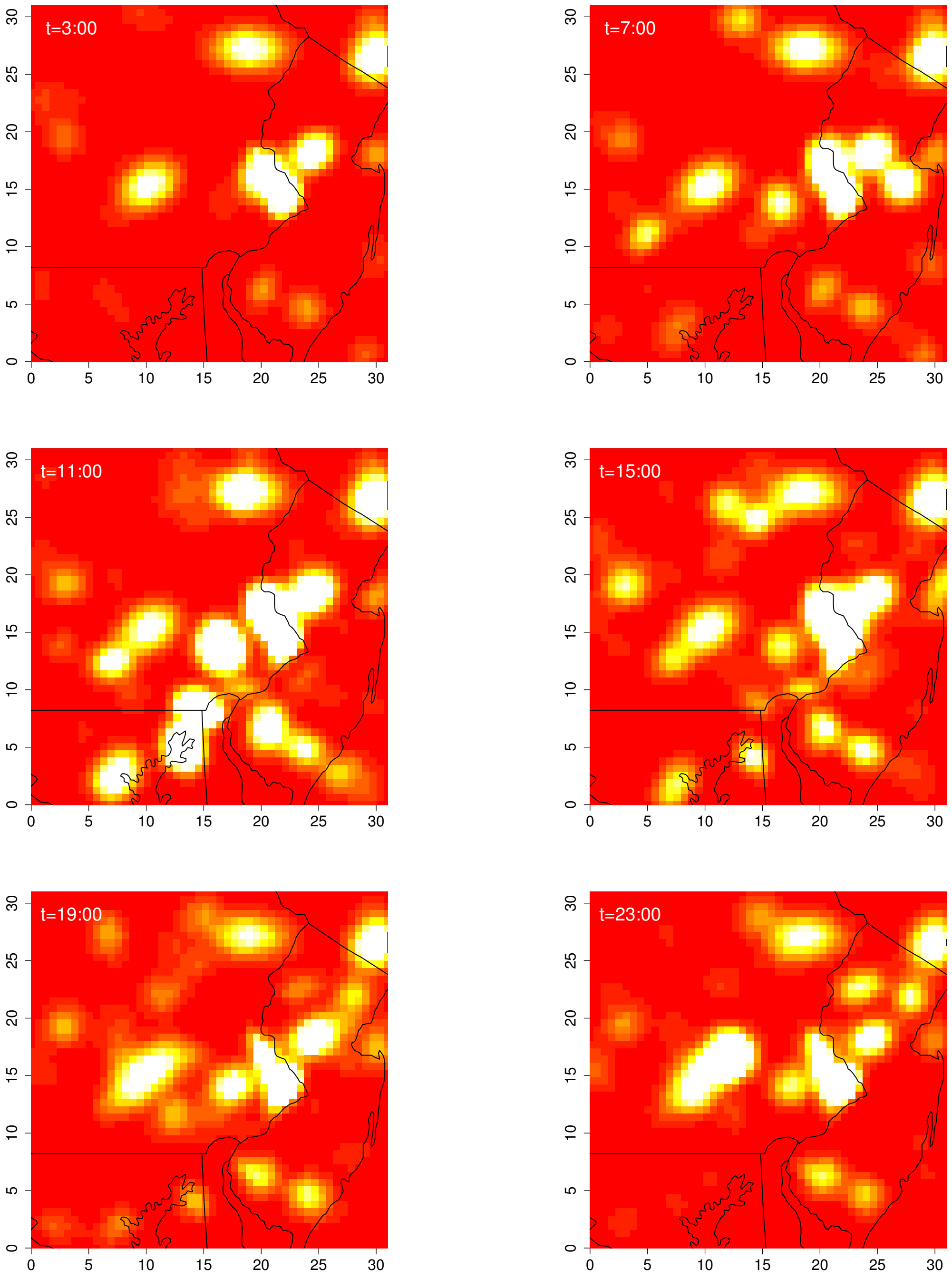

Figure 4.10: Posterior mean spatio-temporal surface. From top to bottom and left to right, each plot shows the fitted spatio-temporal surface at time $3, 7, 11, 15, 19, 23$ o'clock on Sep 3^{rd}, 2002 respectively. Lighter region indicates higher level of SO_2.

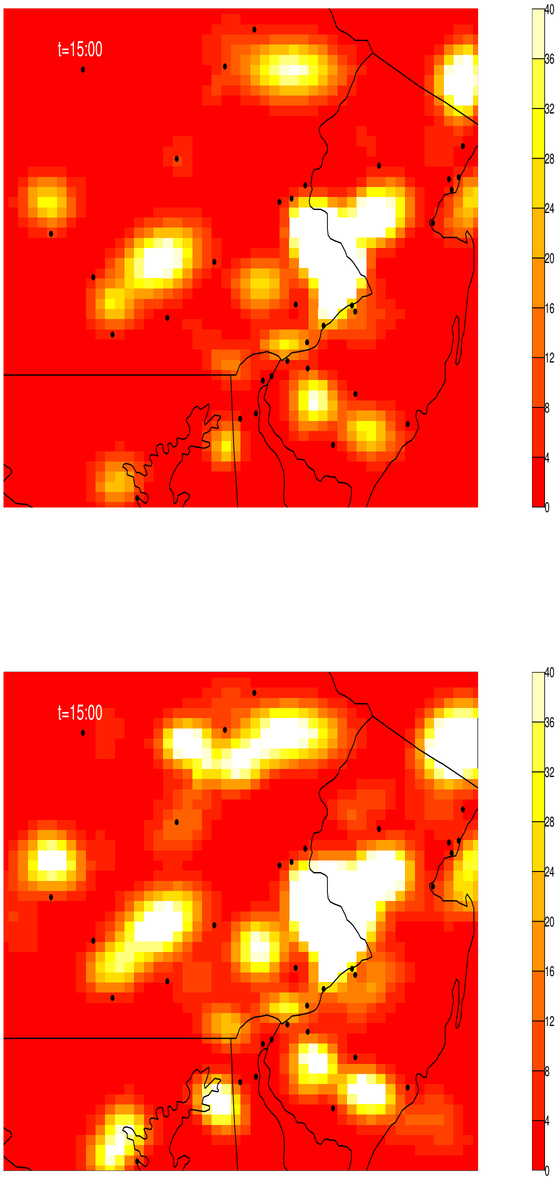

Figure 4.11: Quantiles of the spatio-temporal surface. The top diagram shows the 25% quantile of the posterior spatio-temporal surface $\mu(s, t)$ for $t = 15.00$ on Sep 3$^{\text{rd}}$, 2002. The bottom plot shows the 75% quantile.

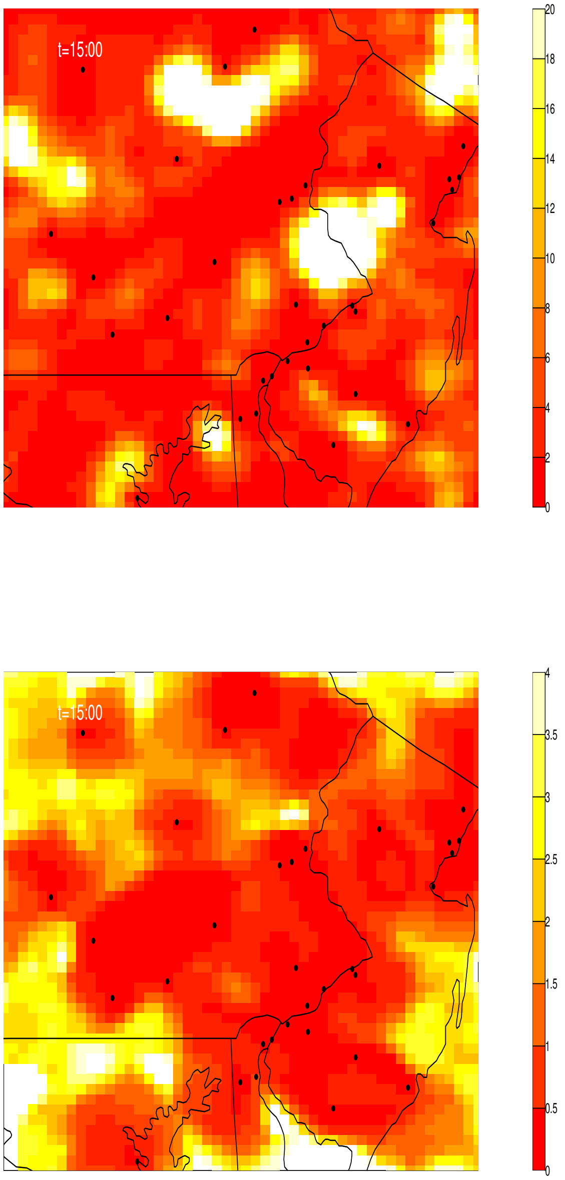

Figure 4.12: Uncertainty map. The top digram shows the standard deviation map. The bottom plot shows that coeffecient of variation (CV) map. The CV is defined as: $\mathrm{CV}(s,t) \equiv \frac{\sqrt{\mathbb{V}[\mu(s,t)|\mathbf{z}]}}{\mathbb{E}[\mu(s,t)|\mathbf{z}]}$. The 33 dots denote the 33 SO_2 monitors.

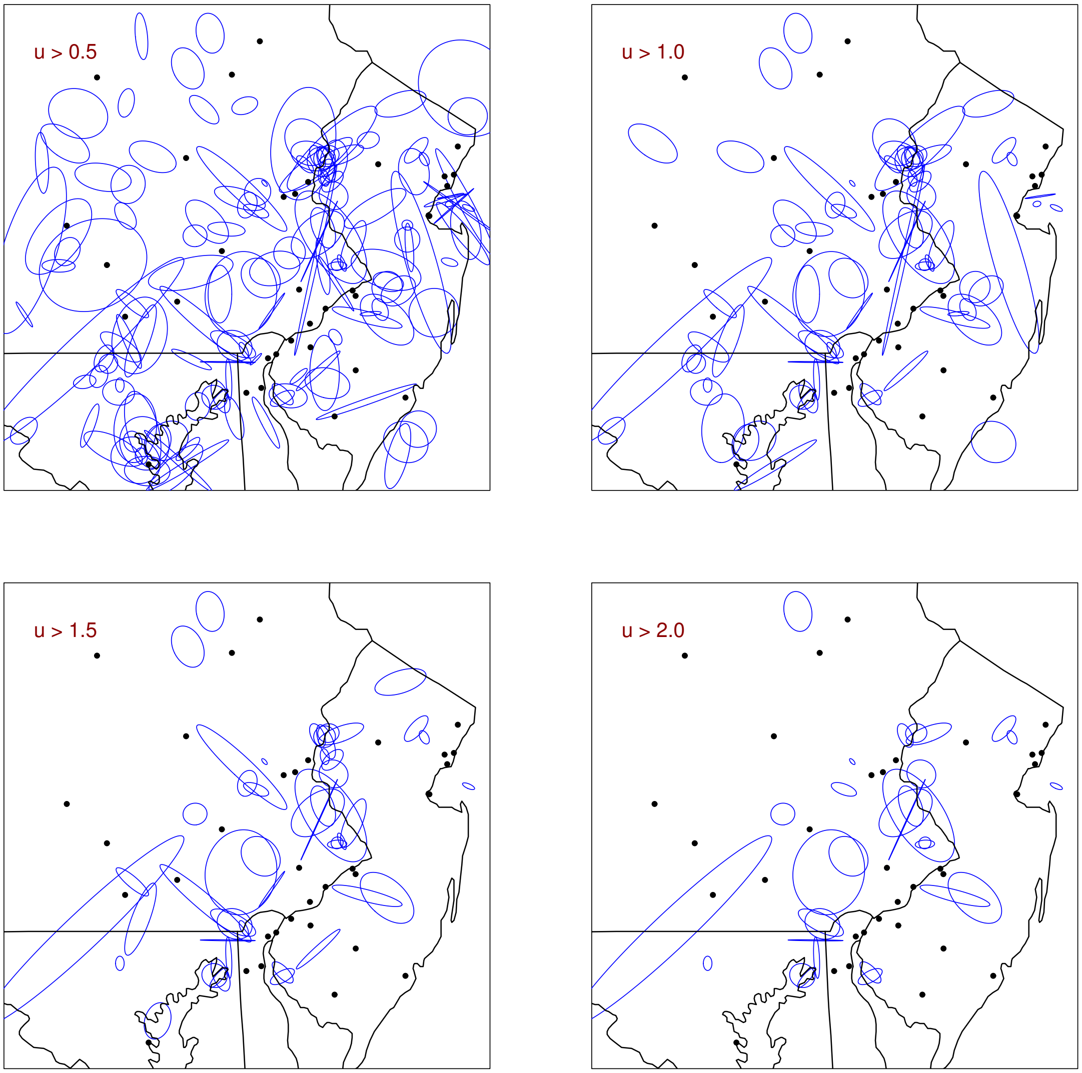

Figure 4.13: Spatio kernels from one posterior sample. The ellipses represent the contours of all the spatial kernels whose corresponding jumps have magnitudes bigger than some threshold. From top to bottom and left to right the threshold is 0.5, 1.0, 1.5 and 2.0 respectively. The dots denote the monitors.

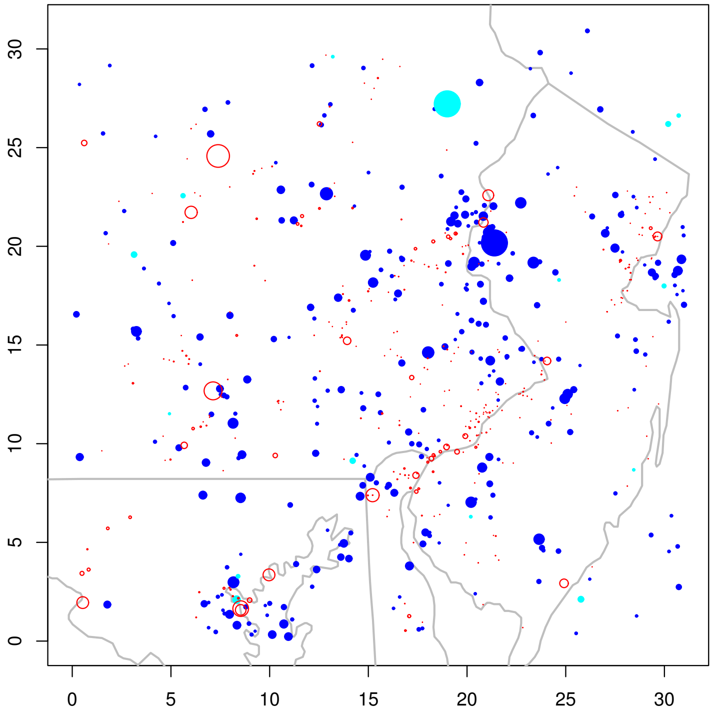

Figure 4.14: Locations of latent point sources from one sample of the posterior distribution. The dark solid circles represent the aperiodic point sources, and the grey solid circles represent the daily periodic point sources. The open circles denotes the real point sources published by EPA. The area of a circle is proportional to the magnitude of the point source it represents. The unit for the x and y axis is 10 km.

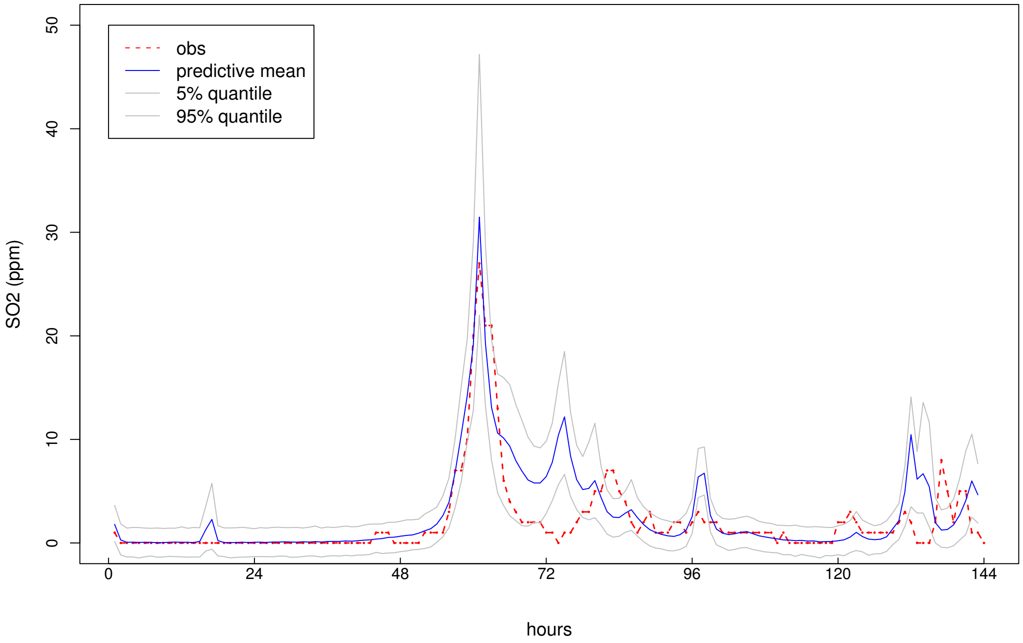

Figure 4.15: Leave-one-out prediction. The dashed line represents observed time series at location 31. The solid line is the predictive mean curve, and the two grey lines are 5% and 95% posterior predictive quantiles respectively. The coverage of the 90% Bayesian credible interval is 80.3% and the RMSE is 2.94.

The posterior means of $\mu(s, t)$ are summarized in Fig. (4.9). At each location, we plot the fitted mean overlaid with the observed time series. It is evident that the model not only has a smoothing effect shared by most kernel methods that is useful for estimating slowly-varying time trends but also can fit the local peaks of the data very well. We can also construct the spatio-temporal surface using the spatio-temporal interpolation method we described in Section 4.3. The posterior mean spatio-surfaces at time $3, 7, 11, 15, 19, 23$ o'clock on September 3 are shown In Fig. (4.10). We show 25% and 75% posterior quantiles of the spatio-surface for $t = 15$ in Fig. (4.11). To evaluate the uncertainty of the estimated spatio-temporal surface, we set up a 50×50 regular grid points in the study region and apply the spatio-temporal extrapolation method to obtain samples from the posterior predictive distribution at each grid point. The standard deviation and coefficient of variation (CV(s,t) $\equiv \frac{\sqrt{\mathbb{V}[\mu(s,t)|\mathbf{z}]}}{\mathbb{E}[\mu(s,t)|bz]}$) of the posterior samples at each grid point are calculated and plotted in Fig. (4.12). The coefficient variation map shows higher CV values for region where no monitor exists. The spatial kernels from one posterior sample are displayed in Fig. (4.13). Spatial kernels associated with large jumps are mostly local, resulting in relatively spiky spatio-temporal surface as shown in Fig. (4.10) . We plot in Fig. (4.14) the locations of latent point sources from one posterior sample. Spatial patterns associated with the latent process can be seen in Fig. (4.14). In particular, we found clusters of point sources with relatively large magnitudes around the Baltimore metropolitan area and the boarder of New Jersey and Pennsylvania. The latent point process in our model is more than just a modeling instrument. It has an attractive interpretation as we discussed in Section 4.2. The point source map such as Fig. (4.14) can help us to identify underlying sources of pollution. It can also help to make future decisions on where to place new monitors and which existing monitors can be removed. Finally we test the predictive power of the model through a leave-one-out prediction experiment. We fit the model without the data at site 31 and use the output to predict the observations for the entire 144 hours at site 31. The result shown in Fig. (4.15) is promising. We were able to predict the major peak with good accuracy. The 90% Bayesian credible intervals cover about 80% of the true observations.

4.5 Discussions and Conclusions

The proposed spatio-temporal model provides a flexible framework for modeling spatio-temporal data. The model is constructed through adaptive kernel convolution, i.e, we define a discrete process over space and time and then smooth it with time-and-location-specific kernels. This constructive approach is appealing since it allows for non-stationary, non-Gaussian models and non-separable space-time dependence structures. In addition, the MCMC algorithm developed for fitting the model is computationally tractable even for large data sets. The computing time is mainly spent on likelihood calculation. But the likelihood calculation in our model formulation does not involve matrix computation, and the parallel computing techniques can be applied to speed up the computation. A Normal additive error model is often used to facilitate computation, although it does not serve such purpose in our model. The spatio-temporal model we introduced in this paper can be easily modified to use other error models. A Gamma multiplicative error model is used for the SO_2 monitoring data, and the result is satisfying. (Detailed results are not shown here). The model developed in this paper is valid for any dataset that is continuous in time and continuous in space, which include data collected on irregular time and locations. This is an important advantage over many other existing methods that require lattice-based data.

In addition to the model presented, we are interested in a number of extensions. Exploring non-separable space-time kernels is of interest because it helps build more flexible and realistic space-time dependence structures. We are also working on introducing covariates into our model. In the SO_2 concentration modeling example, introducing meteorological information can help build a more sensible model. For example, the direction and magnitude of the prevailing winds can be used to model the directions and decay rates of the spatial kernels. The EPA has publicly available information on point pollutant sources such as large chemical plants, power plants, among others. This information can be incorporated into our model by adding latent point sources with known fixed locations. The spatio-temporal model we developed in this paper is very flexible in the sense that it can be used to model both moving and immobile latent point sources simultaneously. Additional important ongoing work is to model multivariate spatio-temporal data. This can be accomplished by applying the idea we discussed in Section 4.2.3. We use Gamma processes to construct the prior distributions for the model, but other Lévy processes, such as Stable processes, are also good candidates. It will be interesting to compare the performance of different Lévy processes.

Chapter 5

Conclusions and Future Work

The LARK model developed in this dissertation provides an attractive solution for sparse nonparametric regression problems in a Bayesian framework. The possible applications of LARK models are broad. We demonstrated its applications in the areas of nonparametric curve fitting, time series modeling and spatio-temporal modeling.

The Lark framework offers a number of advantages. In the area of nonparametric curve fitting, it provides a method to represent inhomogeneous functions sparsely. Unlike traditional basis expansion methods which artificially constrain the locations and scales of basis functions, the LARK framework chooses the locations and scales of kernels (basis elements) optimally based on the data. In addition, the choice of finite Lévy measure $\nu_\epsilon(du\,d\theta)$ can have $\mathbb{E}[J]$, the prior expectation of the number of kernels, as small as we desire, leading to remarkably parsimonious representation of an unknown function. In the area of multivariate time series modeling, the LARK models provides a flexible framework for modeling temporal trends, seasonality and exogenous predictor variables. They also offer a unique way to model association among processes. Classical time series models use cross covariance function which can only capture pairwise linear association between processes. The LARK framework models association through shared latent jumps among processes, which goes far beyond linear association. In the area of spatio-temporal modeling, a common tactic is to use separable covariance structure. The LARK framework allows us to build more general space-time dependence models. In addition, nonstationary processes can be modeled through the LARK framework by allowing the smoothing kernels to evolve over time and space. Since the space-time field is constructed by smoothing out the latent process, non-Gaussian fields can be constructed using non-Gaussian latent processes. The LARK models are valid for data set that is continuous in time and space, which includes data observed at irregular grids that change over time. This is a major advantage over most existing models that require lattice-based data. To conclude this dissertation, we discuss a few possible directions for future work.

Theoretical Properties

In chapter 2, we introduced the LARK models for nonparametric regression problems. We represent an unknown function $f(x)$ as:

$$f(x) = \int_{\Theta} k(x;\theta)L(d\theta) \tag{5.1}$$

for some kernel function $k(x;\theta)$ and Lévy process $L(d\theta)$ defined on Θ. The choice of kernel function, the Lévy process and the prior distribution for θ determines the class of functions Eqn. (5.1) can represent. Relevant work was done by Abramovich *et al.* (2000). In the paper, they consider a class of random functions expanded in an overcomplete wavelet dictionary and prove that under certain regularity conditions of the prior distributions, the expanded functions fall within certain

Besov spaces. Similar work needs to be done for the LARK model. Simulation studies indicate that the LARK model can represent a wide class of functions. Theoretical work to investigate the regularities of the random functions generated by the proposed LARK model is left as future research.

Another challenging future theoretical work concerns the Bayesian consistency properties of the LARK model. Eqn. (5.1) can be viewed as an integral equation (Wolpert *et al.*, 2003). The goal is to impute the unknown measure $L(d\theta)$ from finitely many observations $\{y_i\}$, $i = 1, 2, \cdots, n$, where $y_i = f(x_i) + \epsilon_i$ for measurement error ϵ_i. In the Bayesian framework, the Lévy random field can be used as a prior distribution for unknown measure $L(d\theta)$. With the MCMC based inference procedure we developed, we can obtain the posterior samples of the unknown measure. It will be interesting to know whether or not $\pi_n(\cdot|y_1, \cdots, y_n)$, the posterior distribution of L, is consistent, i.e. for any function $g \in C_b(\Theta)$, $\int_\Theta g(\theta)\pi_n(d\theta) \to \int_\Theta g(\theta)L(d\theta)$.

Computational Issues

The algorithm developed for posterior inference is tractable even for large data set. Here we discuss a few possible ways to speed up the computation. The current MCMC chain starts at initial values drawn from prior distributions. More intelligent choice of good starting values can speed up convergence significantly. One possibility is to put one kernel at each data point and use existing fast algorithms to estimate the regression coefficients, then select the kernels with coefficients bigger than the threshold (ϵ) and start the chain at those kernels with their corresponding coefficient estimates. The most involved part of the computation to fit the LARK models is likelihood evaluation. When the sample size becomes really large (in the order of thousands), fitting LARK models can be slow. Two possible ways can relieve the computation of likelihood calculation. One approach is to develop likelihood approximation that can be calculated relatively fast. The other one is to resort to parallel computing technique which is receiving increasing attention in statistical community. The algorithm we developed to fit a LARK model is parallelizable since the likelihood is evaluated at one point a time. A naive way to parallelize the algorithm is to distribute the likelihood evaluations to multiple computers.

Modeling Extensions

We focused in this dissertation on usage of Gamma random field in construction of LARK models. Other Lévy random fields are readily available as candidates for LARK models. It is worth studying a range of different Lévy random fields and seeing how they affect posterior inference. In chapter two, we mainly considered kernels from exponential families. It will be interesting to explore other well established kernel functions such as wavelet and smoothing spline. On the spatio-temporal modeling for ambient pollutants, we would like to extend the univariate model to multivariate and introduce meteorological covariates as we did in the time series modeling. Finally, we would like to apply the proposed space-time model for epidemiological studies of air pollution, which is one of our primary motivation for this dissertation. A predominant approach to study the effect of ambient pollutants on mortality rate is through Poisson regression:

$$\begin{aligned} Y_t &\sim \mathsf{Po}(\mu_t) \\ \log(\mu_t) &= \mathbf{X}_t'\beta \end{aligned}$$

where Y_t denotes the daily number of deaths and $\mathbf{X}_t$ represents the vector of covariates for predicting mortality on day t, including concentrations of pollutants, important confounding variables, lags of variables and semi-parametric functions of time to account for a baseline risk. There are a number of concerns for the above heath effect model. Monitoring air pollutants are expensive, thus the monitors are sparse. In general, the mortality data and the air pollutants data are often spatially misaligned. The aggregated group level exposure is often approximated by measurements from monitoring sites which can be quite inaccurate. As the timing and duration of exposures are unknown, daily lags of pollutants concentrations are used as predictors. The choice of lags can be artificial and ad hoc. Every now and then, monitors may not be in operation and the simple area-wide average may be biased due to the missing data. As we stated earlier, the space-time model we built is continuous in space and time. With observations at discrete time points and sparse locations, we can obtain the posterior distribution of spatio-temporal surface of the air pollutants. Thus we can build models for individual lifetimes using evolutionary covariates that integrate over past exposures in space and time. This avoids the problem of spatially misaligned health outcome data and the need for choosing lag structures.

Appendix A

The Appendix for Chapter 2

A.1 MCMC Algorithm

MCMC Algorithms: We use $\pi()$ to denote prior distribution, $q()$ to denote Metropolis Hastings proposal distribution and $l()$ to denote likelihood function. Let p_+, p_- and $p_=$ be three positive numbers that satisfy $p_+ + p_- + p_= = 1$.

We use superscript t on the parameters to denote the posterior samples at iteration t and let N_b to denote the burn-in period, N_r to denote the run length, and Δ to denote the thinning rate (to reduce correlations among samples). The algorithm we use to draw posterior samples can be summarized in the following manner:

1. Start the chain at $t = 0$ by initializing the parameters $(\beta_0^0, \alpha^0, \sigma^{2^0}, J^0, \boldsymbol{\beta}^0, \boldsymbol{\omega}^0)$.

2. With probability p_f we update the fixed dimensional parameters and with probability $1 - p_f$ we implement one of the three moves (BIRTH, DEATH, UPDATE) to update varing dimensional parameters.

3. If $t \geq N_r - N_b$ and $(t - N_b) \mod \Delta = 0$, save state for later analysis.

4. Increase t by one and return to step 2 above.

Following the above procedures, we obtain $\lfloor (N_r - N_b)/\Delta \rfloor$ posterior samples of the parameters which can be used to draw inference on the parameters or the functions of the parameters.

The trans-dimensional steps may be summarized as follows:

- BIRTH Step

  With probability p_+, we set $J^* = J^{t-1}+1$ and generate a random index r uniformly from $1, \cdots, J^{t-1}+1$ and sample a new point (β^*, ω^*) from proposal distribution $q^b(\beta, \omega)$, where $q^b()$ ensures that $\beta^* > \epsilon$. Set $(\boldsymbol{\beta}^*, \boldsymbol{\omega}^*)$ by letting: $(\beta^*_j, \omega^*_j) = (\beta^{t-1}_j, \omega^{t-1}_j)$, for $j = 1, \cdots, r-1$; $(\beta^*_r, \omega^*_r) = (\beta^*, \omega^*)$; $(\beta^*_j, \omega^*_j) = (\beta^{t-1}_{j-1}, \omega^{t-1}_{j-1})$, for $j = r+1, \cdots, J^*$. Let $\boldsymbol{\Theta}^* = (\boldsymbol{\beta}^*, \boldsymbol{\omega}^*, J^*, \beta^{t-1}_0, \sigma^{2^{t-1}}, \alpha^{t-1})$, $\boldsymbol{\Theta}^{t-1} = (\boldsymbol{\beta}^{t-1}, \boldsymbol{\omega}^{t-1}, J^{t-1}, \beta^{t-1}_0, \sigma^{2^{t-1}}, \alpha^{t-1})$ and $\boldsymbol{\theta}^t = (\boldsymbol{\beta}^t, \boldsymbol{\omega}^t, J^t, \beta^t_0, \sigma^{2^t}, \alpha^t)$. With probability $\min(1, H)$, we accept the proposal and set $\boldsymbol{\Theta}^t \equiv \boldsymbol{\Theta}^*$; with probability $1 - \min(1, H)$, we reject the proposal and set $\boldsymbol{\Theta}^t \equiv \boldsymbol{\Theta}^{t-1}$. The Hastings ratio H for this move is:

$$\begin{aligned} H &= \frac{l(\mathbf{y}|\boldsymbol{\Theta}^*)}{l(\mathbf{y}|\boldsymbol{\Theta}^{t-1})} \times \frac{\pi(\beta^*, \omega^*|J^*)\pi(J^*)}{\pi(\beta^{t-1}, \omega^{t-1}|J^{t-1})\pi(J^{t-1})} \\ &\times \frac{\left(p_- + p_= \int_{-\infty}^{\epsilon} q^d(\beta|\beta^*)d\beta\right)/J^*}{p_+/J^*} \times \frac{1}{q^b(\beta^*, \omega^*)} \end{aligned}$$

- UPDATE Step

  With probability $1-p_+$, generate a random index r uniformly from $1, \cdots, J_{t-1}$. With probability $p_=$, we propose a new point β^*_r from proposal distribution $q^d(\beta|\beta^{t-1}_r)$. If $\beta^*_r \geq \epsilon$, we implement UPDATE step; otherwise we implement DEATH step which we specify below. We first update β_r. Let $\boldsymbol{\Theta}_{(-\beta_r)}$ denote the rest of the parameters in the model. With probability $\min(1, H)$, we accept the proposal and set $\beta_r^t \equiv \beta_r^*$; with probability $1 - \min(1, H)$, we reject it and set $\beta_r^t \equiv \beta_r^{t-1}$. The Hastings ratio H is:

$$H = \frac{l\left(\mathbf{y}|\beta_r^*, \mathbf{\Theta}_{(-\beta_r)}^{t-1}\right)\pi(\beta_r^*)q^d(\beta_r^{t-1}|\beta_r^*)}{l\left(\mathbf{y}|\beta_r^{t-1}, \mathbf{\Theta}_{(-\beta_r)}^{t-1}\right)\pi(\beta_r^{t-1})q^d(\beta_r^*|\beta_r^{t-1})}$$

We then update every component of ω_r in a similar fashion.

- DEATH Step

  Using the random index r and β_r^* generated in the UPDATE step, with probability $p_- + p_= \times \Pr(\beta_r^* < \epsilon)$, set $J^* \equiv J^{t-1} - 1$ and generate $(\mathbf{u}^*, \boldsymbol{\theta}^*)$ by deleting the r-th component from $(\boldsymbol{\beta}, \boldsymbol{\omega})$, i.e. let $(\beta_j^*, \omega_j^*) = (\beta_j^{t-1}, \omega_j^{t-1})$, for $j = 1, 2, \cdots, r-1$ and $(\beta_j^*, \omega_j^*) = (\beta_{j+1}^{t-1}, \omega_{j+1}^{t-1})$ for $j = r, r+1, \cdots, J^*$. Let $\mathbf{\Theta}^* = (\boldsymbol{\beta}^*, \boldsymbol{\omega}^*, J^*, \beta_0^{t-1}, {\sigma^2}^{t-1}, \alpha^{t-1})$, $\mathbf{\Theta}^{t-1} = (\boldsymbol{\beta}^{t-1}, \boldsymbol{\omega}^{t-1}, J^{t-1}, \beta_0^{t-1}, {\sigma^2}^{t-1}, \alpha^{t-1})$ and $\mathbf{\Theta}^t = (\boldsymbol{\beta}^t, \boldsymbol{\omega}^t, J^t, \beta_0^t, {\sigma^2}^t, \alpha^t)$. With probability $\min(1, H)$, we accept the DEATH move and set $\mathbf{\Theta}^t = \mathbf{\Theta}^*$; with probability $1 - \min(1, H)$, we reject the DEATH move and set $\mathbf{\Theta}^t = \mathbf{\Theta}^{t-1}$. The Hastings ratio for this move is:

$$\begin{aligned} H &= \frac{l(\mathbf{y}|\mathbf{\Theta}^*)}{l(\mathbf{y}|\mathbf{\Theta}^{t-1})} \times \frac{\pi(\boldsymbol{\beta}^*, \boldsymbol{\omega}^*|J^*)\pi(J^*)}{\pi(\boldsymbol{\beta}^{t-1}, \boldsymbol{\omega}^{t-1}|J^{t-1})\pi(J^{t-1})} \\ &\times \frac{p_+/J^*}{\left(p_- + p_= \int_{-\infty}^{\epsilon} q^d(\beta|\beta_r^{t-1})d\beta\right)/J^*} \times q^b(\beta_r^{t-1}, \omega_r^{t-1}) \end{aligned}$$

- Update $(\beta_0, \sigma^2, \alpha)$ We update $(\beta_0, \sigma^2, \alpha)$ element by element.

  Sample a candidate point β_0^* from $q(\beta_0^*|\beta_0^{t-1})$. With probability $\min(1, H)$, we accept the proposal and set $\beta_0^t = \beta_0^*$; with probability $1 - \min(1, H)$, we reject the proposal and set $\beta_0^t = \beta_0^{t-1}$. The Hastings ratio for this move is:

$$H = \frac{l(\mathbf{y}|\beta_0^*)\pi(\beta_0^*)q(\beta_0^{t-1}|\beta_0^*)}{l(\mathbf{y}|\beta_0^{t-1})\pi(\beta_0^{t-1})q(\beta_0^*|\beta_0^{t-1})}$$

Updating parameter α depends on the choice of prior Lévy process. For certain Lévy process, there exists conjugate prior for α and we can use a Gibbs step to update α but for certain Lévy process there is no conjugate prior for *alpha* and under this circumstance, we update α in a similar fashion to the update of β_0.

Assuming an independent normal error model and a prior $\pi(\sigma^2) \propto \frac{1}{\sigma^2}$, the conditional distribution of σ^2 is InverseGamma, and easily updated using a Gibbs step. Note that if $x \sim$ Inv-Gamma(α, β), with density function $\frac{\beta^\alpha}{\Gamma(\alpha)} x^{-(\alpha+1)} e^{-\beta/x}$, then the full conditional posterior distribution for σ^2 is

$$\sigma^2|\mathbf{y}, \mathbf{\Theta} \quad \sim \quad \text{Inv-Gamma}\left(\frac{n}{2}, \; \frac{1}{2}\sum_{i=1}^{n}\left[y_i - \sum_{j=1}^{J}\beta_j k(x_i;\omega_j)\right]^2\right)$$

Appendix B

The Appendix for Chapter 3

B.1 Kernel Function for Periodic Jumps

Let $\delta \equiv (t - \tau_j) \mod 24$, The contribution of a daily periodic jump j at time t can be calculated in the following fashion:

$$\sum_{n=-\infty}^{\infty} e^{-\lambda_j |t-\tau_j-24n|} = \sum_{n=0}^{\infty} e^{-\lambda_j \delta - 24\lambda_j n} + \sum_{n=0}^{\infty} e^{-\lambda_j |\delta - 24| - 24\lambda_j n}$$

$$= \frac{e^{-\lambda_j \delta}}{1 - e^{-24\lambda_j}} + \frac{e^{-\lambda_j (24-\delta)}}{1 - e^{-24\lambda_j}} = \frac{e^{-\lambda_j \delta}\left(1 + e^{-2\lambda_j (12-\delta)}\right)}{1 - e^{-24\lambda_j}}$$

So, the kernel function $k(t; \tau_j, \lambda_j, a_j = 2) = \frac{e^{-\lambda_j\left[(t-\tau_j) \mod 24\right]}\left(1+e^{-2\lambda_j\left[12-(t-\tau_j) \mod 24\right]}\right)}{1-e^{-24\lambda_j}}$

B.2 Updating Fixed Dimensional Parameters

1. Update α_k for $k = 1, 2, \cdots, K$

Let $J_k = \sum_{j=1}^{J} I_{\{a_j=k\}}(a_j)$, if $\alpha_k \sim \mathsf{Ga}(a_{\alpha_k}, b_{\alpha_k})$, we can prove that

$$\begin{aligned} J_k \mid J, \alpha_k &\sim \mathsf{Bi}(J, p_k) \\ J_k \mid \alpha_k &\sim \mathsf{Po}\left(\alpha_k T_k E_1(\epsilon)\right) \\ \alpha_k \mid J_k &\sim \mathsf{Ga}\left(a_{\alpha_k} + J_k \ , \ b_{\alpha_k} + T_k E_1(\epsilon)\right) \end{aligned}$$

2. Metropolis Step to Update b_λ

   The full conditional distribution for b_λ

$$\begin{aligned} (b_\lambda \mid \boldsymbol{\lambda}, J) &\propto \left\{ \prod_{j=1}^{J} \frac{b_\lambda^{a_\lambda}}{\Gamma(a_\lambda)} \lambda_j^{a_\lambda - 1} e^{-b_\lambda \lambda_j} \right\} \frac{1}{b_\lambda} \exp\left\{ -\frac{(\log b_\lambda - m_b)^2}{2v_b} \right\} \\ &\propto b_\lambda^{J a_\lambda - 1} \exp\left\{ -b_\lambda \sum_{j=1}^{J} \lambda_j - \frac{(\log b_\lambda - m_b)^2}{2v_b} \right\} \end{aligned}$$

   is not in closed form. A Metropolis algorithm is used to update b_λ. We sample a new point b_λ^* from the proposal distribution $q(b_\lambda^* | b_\lambda) \sim \mathsf{LN}(\log(b_\lambda), \sigma_{b_\lambda}^2)$. With probability $\min(1, H)$, we accept the proposal and with probability $1 - \min(1, H)$ we reject the proposal. The Hastings ratio H is:

$$H = \left\{ \frac{\pi(b_\lambda^* \mid \boldsymbol{\lambda}, J) q(b_\lambda | b_\lambda^*)}{\pi(b_\lambda \mid \boldsymbol{\lambda}, J) q(b_\lambda^* | b_\lambda)} \right\}$$

3. Update $\mathbf{d}$, $\boldsymbol{\gamma}$ and σ_d^2

   The full conditional distribution for d_k is not in closed form and we choose Metropolis step to update it. A log normal proposal distribution $q(d_k^* | d_k) \sim \mathsf{LN}(\log(d_k), \sigma_d^2)$ is used. The Hastings ratio H is:

$$\alpha = \frac{f(\mathbf{y} | d_k^*, \ \_) \pi(d_k^* | \boldsymbol{\gamma}, \sigma_d^2) q(d_k | d_k^*)}{f(\mathbf{y} | d_k, \ \_) \pi(d_k | \boldsymbol{\gamma}, \sigma_d^2) q(d_k^* | d_k)}$$

Since $\log(\mathbf{d}) \sim \mathsf{MN}(\mathbf{X}\boldsymbol{\gamma},\ \sigma_d^2 I)$, under the priors specified in Section **??** section, we can use Gibbs step to sample them $\boldsymbol{\gamma}$ and σ_d^2.

$$\boldsymbol{\gamma} \mid \mathbf{d}, \mathbf{X}, \sigma_d^2 \ \sim\ \mathsf{MN}([\mathbf{X}^{'}\mathbf{X}]^{-1}\mathbf{X}^{'}\log(\mathbf{d})\ ,\ \sigma_d^2[\mathbf{X}^{'}\mathbf{X}]^{-1})$$

Let D denote the number of days:

$$\sigma_d^2 \mid \mathbf{d}, \mathbf{X}, \boldsymbol{\gamma} \ \sim\ Inv - \mathsf{Ga}\ \left(\frac{D+1}{2}\ ,\ \frac{\sum_{k=1}^{D}(\log d_k - \mathbf{X}_k^{'}\boldsymbol{\gamma})^2}{2}\right)$$

Appendix C

The Appendix for Chapter 4

C.1 Derivation of $\pi(J_1 \mid \boldsymbol{\alpha})$, $\pi(J_1)$ and $\pi(J_2)$

- Derivation of $\pi(J_1|\boldsymbol{\alpha})$:

  Let $\boldsymbol{\alpha} \equiv (\alpha_1, \alpha_2)$, $p = \frac{T\alpha_1}{T\alpha_1+24\alpha_2}$ and $\gamma = (T\alpha_1 + 24\alpha_2)E_1(\epsilon)S$. Notice that $J_1 \mid J, \boldsymbol{\alpha} \sim \mathsf{Bin}(J, p)$ and $J \mid \boldsymbol{\alpha} \sim \mathsf{Po}(\gamma)$, for any nonnegative integer k, we have:

$$\begin{aligned}
\Pr(J_1 = k \mid \boldsymbol{\alpha}) &= \sum_{j=0}^{\infty} \Pr(J_1 = k, J = j \mid \boldsymbol{\alpha}) \\
&= \sum_{j=0}^{\infty} \Pr(J_1 = k \mid J = j, \boldsymbol{\alpha}) \Pr(J = j \mid \boldsymbol{\alpha}) \\
&= \sum_{j=k}^{\infty} \left(\begin{array}{c} j \\ k \end{array} \right) p^k (1-p)^{j-k} \frac{e^{-\gamma}\gamma^j}{j!} = \sum_{j=k}^{\infty} (\gamma p)^k \frac{[\gamma(1-p)]^{j-k} e^{-\gamma}}{k!(j-k)!} \\
&= \frac{(\gamma p)^k}{k!} \frac{e^{-\gamma}}{e^{-\gamma(1-p)}} \sum_{j=0}^{\infty} \frac{[\gamma(1-p)]^j e^{-\gamma(1-p)}}{j!} = \frac{(\gamma p)^k e^{-\gamma p}}{k!} \\
&= \frac{(\alpha_1 T E_1(\epsilon))^k \, e^{-\alpha_1 T E_1(\epsilon)}}{k!}
\end{aligned}$$

So, $\pi(J_1 \mid \boldsymbol{\alpha}) \sim \mathsf{Po}(\alpha_1 T E_1(\epsilon))$. Similarly, we can derive that $\pi(J_2 \mid \boldsymbol{\alpha}) \sim \mathsf{Po}(\alpha_2 24 E_1(\epsilon))$.

- Derivation of $\pi(J_1)$ and $\pi(J_2)$:

  Since $\pi(J_1 \mid \boldsymbol{\alpha}) \sim \mathsf{Po}(\alpha_1 T E_1(\epsilon))$ and $\pi(\alpha_1) \sim \mathsf{Ga}(a_{\alpha_1}, b_{\alpha_1})$, for any nonnegative integer J_1, we have:

$$\begin{aligned}
\pi(J_1) &= \int_0^\infty \pi(J_1|\alpha_1)\pi(\alpha_1)\, d\alpha_1 \\
&= \int_0^\infty \frac{\exp\{-\alpha_1 E_1(\epsilon)TS\}[\alpha_1 E_1(\epsilon)TS]^{J_1}}{J_1!} \frac{b_{\alpha_1}^{a_{\alpha_1}}}{\Gamma(a_{\alpha_1})} \alpha_1^{a_{\alpha_1}-1} e^{-b_{\alpha_1}\alpha_1}\, d\alpha_1 \\
&= \frac{\Gamma(J_1 + a_{\alpha_1})}{\Gamma(a_{\alpha_1})\Gamma(J_1+1)} \left[\frac{b_{\alpha_1}}{b_{\alpha_1} + E_1(\epsilon)TS}\right]^{a_{\alpha_1}} \left[1 - \frac{b_{\alpha_1}}{b_{\alpha_1} + E_1(\epsilon)TS}\right]^{J_1}
\end{aligned}$$

  If we choose a_{α_1} to be a positive integer, then $J_1 \sim \mathsf{NB}(a_{\alpha_1}, p_1)$ for $p_1 \equiv \frac{b_{\alpha_1}}{b_{\alpha_1}+E_1(\epsilon)TS}$. Similarly, we can show that $J_2 \sim \mathsf{NB}(a_{\alpha_2}, p_2)$, for $p_2 \equiv \frac{b_{\alpha_2}}{b_{\alpha_2}+24E_1(\epsilon)S}$.

C.2 The Complete Information of Point Sources Used in the Simulation Study

Table C.1: Information of parameters used for the simulation study.

	u	$(\sigma_x,$	$\sigma_y)$	τ	λ_t	$(\lambda_1,$	$\lambda_2)$	ϕ	a
a	10.0	(24.5,	4.50)	20	0.50	(4.5,	4.5)	1.571	1
b	7.0	(23.5,	9.00)	30	0.50	(4.5,	4.5)	1.571	1
c	12.0	(7.5,	12.50)	37	0.55	(5.0,	5.0)	1.571	1
d	9.0	(22.0,	12.50)	40	0.50	(4.5,	4.5)	1.571	1
e	11.0	(20.5,	15.50)	50	0.50	(4.5,	4.5)	1.571	1
f	13.0	(19.0,	19.00)	60	0.50	(4.5,	4.5)	1.571	1
g	10.0	(17.0,	22.00)	70	0.50	(4.5,	4.5)	1.571	1
h	8.0	(14.0,	25.00)	80	0.50	(4.5,	4.5)	1.571	1
i	10.8	(18.0,	8.50)	81	0.60	(6.5,	6.5)	1.571	1
j	8.0	(7.0,	30.00)	90	0.50	(4.5,	4.5)	1.571	1
A	7.5	(16.0,	6.50)	7*	0.45	(6.0,	6.0)	1.571	2
B	9.0	(9.0,	17.00)	8*	0.35	(11.0,	11.0)	1.571	2
C	6.0	(19.5,	10.50)	12*	0.60	(9.0,	3.0)	0.785	2
D	6.5	(29.5,	20.00)	14*	0.40	(4.0,	1.0)	−0.785	2
E	7.0	(20.0,	8.75)	18*	0.75	(3.0,	3.0)	1.571	2

C.3 Algorithms for Sampling from $[u|\beta = 1]$

From Section **??**, we find out the prior distribution of $[u|\beta = 1]$ is as follows:

$$f(u) = \frac{u^{-1}e^{-u}}{E_1(\epsilon)} I_{(\epsilon,\infty)}(u) \ , \ u \in (\epsilon, \infty)$$

where $E_1(x) = \int_x^\infty u^{-1}e^{-\beta u}du$ is the so called exponential integral function. A simple calculation gives us the CDF $F(u)$, where $F(u) \equiv \Pr(U \leq u)$:

$$F(u) = \left(1 - \frac{E_1(u)}{E_1(\epsilon)}\right) \ , \ u \in (\epsilon, \infty)$$

$E_1(x)$ is a strictly increasing function, thus the inverse of $F(u)$ exists:

$$F^{-1}(u) = E_1^{-1}\left(E_1(\epsilon)(1-u)\right) \ , \ u \in (0,1)$$

So, as long as we can invert the function $E_1(x)$, u can be sampled efficiently using probability integral transform, i.e., sample x from a uniform distribution on $[0,1]$, generate u by setting $u = E_1^{-1}\left(E_1(\epsilon)(1-u)\right)$. To invert $E_1(x)$, we refer to Wolpert and Ickstadt (1998), let χ_d^2 be a random variable that has chi-square distribution with degrees of freedom d, we have:

$$E_1(x) = \lim_{d \to 0}\left(\frac{2}{d}\Pr(\chi_d^2 > 2x)\right)$$

Thus $E_1^{-1}(x)$ can be approximated by:

$$E_1^{-1}(x) \approx \frac{1}{2}q_{\chi_d^2}\left(1 - \frac{xd}{2}\right)$$

where $a_{\chi_d^2}$ is the quantile function for chi-square distribution with d degrees of freedom and d is a small positive number, usually choosing $d = 1e-9$ gives an approximation precise enough for most applications.

Bibliography

Abramovich, F. and Benjamini, Y. (1996). Adaptive thresholding of wavelet coefficients. *Computational Statistics and Data Analysis* **22**, 351–361.

Abramovich, F., Sapatinas, T., and Silverman, B. W. (1998). Wavelet thresholding via a Bayesian approach. *J. Roy. Stat. Soc. B* **60**, 1, 1–52.

Abramovich, F., Sapatinas, T., and Silverman, B. W. (2000). Stochastic expansions in an overcomplete wavelet dictionary. *Probab. Theory Rel.* **117**, 1, 133–144.

Bell, M. L., Samet, J. M., and Dominici, F. (2004). Time-series studies of particulate matter. *Annual Review of Public Health* **25**, 247–280.

Berger, J. O., Ghosh, J. K., and Mukhopadhyay, N. (2003). Approximations and consistency of Bayes factors as model dimension grows. *J. Stat. Plan. Infer.* **112**, 1-2, 241–258.

Berger, J. O. and Pericchi, L. R. (2001). Objective Bayesian methods for model selection: Introduction and comparison. In P. Lahiri, ed., *Model Selection*, vol. 38 of *Lecture Notes in Statistics*, 135–193. Inst. Math. Statist., Hayward, CA.

Calder, K. (2003). *Exploring Latent Structure in Spatial Temporal Processes Using Process Convolutions*. Ph.D. dissertation, Duke University ISDS, Durham, NC.

Chen, S. S., Donoho, D. L., and Saunders, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM J. Sci. Comp.* **20**, 1, 33–61.

Chipman, H. A., Kolaczyk, E. D., and McCulloch, R. E. (1997). Adaptive Bayesian wavelet shrinkage. *J. Am. Stat. Assoc.* **92**, 1413–1421.

Chu, C.-k. and Marron, J. S. (1991). Choosing a kernel regression estimator (Disc: *p.* 419–436). *Stat. Sci.* **6**, 4, 404–419.

Clyde, M. and George, E. I. (2000). Flexible empirical Bayes estimation for wavelets. *J. Roy. Stat. Soc. B* **62**, 4, 681–698.

Clyde, M. and George, E. I. (2004). Model uncertainty. *Stat. Sci.* **19**, 2, 81–94.

Clyde, M., Parmigiani, G., and Vidakovic, B. (1998). Multiple shrinkage and subset selection in wavelets. *Biometrika* **85**, 391–401.

Cont, R. and Tankov, P. (2004). *Financial modelling with jump processes.* Chapman & Hall/CRC, London, UK.

Cressie, N. and Huang, H. (1999). Classes of nonseperable, spatio-temporal stationary covariance functions. *Journal of the American Statistical Association* **94**, 1330–1340.

Cristianini, N. and Shawe-Taylor, J. (2000). *An introduction to support vector machines: and other kernel-based learning methods.* Cambridge University Press, Cambridge, UK.

Daubechies, I. (1992). *Ten Lectures on Wavelets*, vol. 61 of *CBMS-NSF Regional Conference Series in Applied Mathematics.* SIAM, Philadelphia, PA.

Denison, D. G. T., Holmes, C. C., Mallick, B. K., and Smith, A. F. M. (2002). *Bayesian Methods for Nonlinear Classification and Regression.* John Wiley & Sons, New York, NY.

Denison, D. G. T., Mallick, B. K., and Smith, A. F. M. (1998). Automatic Bayesian curve fitting. *J. Roy. Stat. Soc. B* **60**, 333–350.

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

Dockery, D. and Pope, C. A. (1994). Acute respiratory effects of particulate air pollution. *Annual Review of Public Health* **15**, 107–132.

Dominici, F., Samet, J. M., and Zeger, S. L. (2000). Combining evidence on air pollution and daily mortality from the twenty largest U.S. cities: A hierarchical modeling strategy. *Journal of the Royal Statistical Society, Series A* **163**, 3, 263–284.

Donoho, D. L. and Elad, M. (2003). Maximal sparsity representation via l_1 minimization. *Proc. Nat. Aca. Sci.* **100**, 2197–2202.

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425–455.

EPA (2004). Air quality criteria for particulate matter. Tech. rep., U.S. Environmental Protection Agency.

Gelfand, A., S.Banerjee, and D.Gamerman (2005). Spatial process modelling for univariate and multivariate dynamic spatial data. *Environmetrics* **16**, 1–15.

George, E. I. and McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *J. Am. Stat. Assoc.* **88**, 881–889.

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Stat. Sinica* **7**, 339–374.

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

Higdon, D. M. (1998). A process-convolution approach to modeling temperatures in the north atlantic ocean. *Journal of Environmental and Ecological Statistics* **5**, 173–190.

Huang, H. and Cressie, N. (1996). Spatio-temporal prediction of snow water equivalent using the kalman filter. *Computational Statistics and Data Analysis* **22**, 159–175.

Huerta, G., Sansó, B., and Stroud, J. R. (2004). A spatio-temporal model for mexico city ozone levels. *Journal of the Royal Statistical Society, Series C* **53**, 231–248.

Jacod, J. and Shiryaev, A. N. (1987). *Limit Theorems for Stochastic Processes*, vol. 288 of *Grundlehren der mathematischen Wissenschaften*. Springer-Verlag, Berlin, DE.

Johnstone, I. M. and Silverman, B. W. (2004). Needles and hay in haystacks: Empirical Bayes estimates of possibly sparse sequences. *Ann. Stat.* **32**, 1, 4.

Johnstone, I. M. and Silverman, B. W. (2005). Empirical Bayes selection of wavelet thresholds. *Ann. Stat.* **33**, 1700–1752.

Kass, R. E. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. Am. Stat. Assoc.* **90**, 928–934.

Khinchine, A. Y. and Lévy, P. (1936). Sur les lois stables. *C. R. Acad. Sci. Paris* **202**, 374–376.

Liang, F., Paulo, R., Molina, G., Clyde, M., and Berger, J. O. (2005). Mixtures of g-priors for Bayesian variable selection. Tech. Rep. 05-12, Duke University ISDS, USA.

Mallat, S. G. and Zhang, Z. (1993). Matching pursuit with time-frequency dictionaries. *IEEE T. Signal Proces.* **41**, 3397–3415.

Mardia, K., Goodall, C.R., Redfern, E., and Alonso, F. (1998). The kriged kalman filter. *TEST* **7**, 217–252.

Nason, G. (1996). Wavelet shrinkage using cross-validation. *J. Roy. Stat. Soc. B* **58**, 463–479.

Paciorek, C. J. (2006). Misinformation in the conjugate prior for the linear model with implications for free-knot spline modelling. *Bayesian Analysis* .

Rajput, B. S. and Rosiński, J. (1989). Spectral representations of infinitely divisible processes. *Probab. Theory Rel.* **82**, 3, 451–487.

Samorodnitsky, G. and Taqqu, M. S. (1994). *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*, vol. 1 of *Stochastic Modeling Series*. Chapman & Hall, New York, NY.

Sato, K.-i. (1999). *Lévy Processes and Infinitely Divisible Distributions.* Cambridge University Press, Cambridge, UK.

Smith, M. and Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *J. Econometrics* **75**, 317–343.

Stein, M. L. (2005). Space-time covariance function. *Journal of the American Statistical Association* **100**, 1, 310–321.

Stroud, J. R., Müller, P., and Sansó, B. (2001). Dynamic models for spatio-temporal data. *Journal of the Royal Statistical Society, Series B* **63**, 4, 673–689.

Swall, J. (1999). *Non-stationary Spatial Modeling Using A Process Convolution Approach*. Ph.D. dissertation, Duke University ISDS, Durham, NC.

Thurston, G. D. (1996). A critical review of PM10-mortality time-series studies. *Journal of Exposure Analysis and Environmental Epidemiology* **6**, 3–21.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Stat. Soc. B* **58**, 1, 267–288.

Tipping, M. E. (2001). Sparse Bayesian learning and the Relevence Vector Machine. *J. Mach. Learn. Res.* **1**, 211–244.

Tu, C., Wolpert, R. L., and Clyde, M. A. (2005). Lévy adaptive regression kernels. Tech. rep., Duke University ISDS.

Vidakovic, B. (1999). *Statistical Modeling by Wavelets*. Computational & Graphical Statistics. John Wiley & Sons, New York, NY.

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

Wahba, G. (1992). Multivariate function and operator estimation, based on smoothing splines and reproducing kernels. In M. Casdagli and S. G. Eubank, eds., *Nonlinear Modeling and Forecasting: Proceedings of the Workshop on Nonlinear Modeling and Forecasting held September, 1990, in Santa Fe, New Mexico*, vol. XII of *SFI Studies in the Sciences of Complexity*, 95–112, Redwood, CA. Addison-Wesley.

West, M. (1997). *Bayesian Forecasting and Dynamic Models(Second Edition)*. Springer, New York, NY.

Wikle, C. and Cressie, N. (1999). Dimension-reduced approach to space-time kalman filtering. *Biometrika* **86**, 815–829.

Wolfe, P. J., Godsill, S. J., and Ng, W.-J. (2004). Bayesian variable selection and regularisation for time-frequency surface estimation. *J. Roy. Stat. Soc. B* **66**, 575–589.

Wolpert, R. L. and Ickstadt, K. (1998). Simulation of Lévy random fields. In D. K. Dey, P. Müller, and D. Sinha, eds., *Practical Nonparametric and*

Semiparametric Bayesian Statistics, vol. 133 of *Lecture Notes in Statistics*, 227–242. Springer-Verlag, New York, NY.

Wolpert, R. L., Ickstadt, K., and Hansen, M. B. (2003). A nonparametric Bayesian approach to inverse problems (with discussion). In J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith, and M. West, eds., *Bayesian Statistics 7*, 403–418, Oxford, UK. Oxford Univ. Press.

Wolpert, R. L. and Taqqu, M. S. (2005). Fractional Ornstein-Uhlenbeck Lévy processes and the Telecom process: Upstairs and downstairs. *Signal Processing* **85**, 8, 1523–1545.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In P. K. Goel and A. Zellner, eds., *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, 233–243, Amsterdam, NL. Elsevier.

Biography

I was born December 10, 1976 in Wuhan, China. I received a B.S. in Biology with honors in 1999 from Zhejiang University in Hangzhou, China. In 2003, I earned an M.S. in statistics from the Institute of Statistics and Decision Sciences, Duke University. I have co-authored the following articles:

1. Clyde, M., House, L., Tu, C., Wolpert, R. "Bayesian Nonparametric Function Estimation Using Overcomplete Representations and Lévy Random Field Priors." *Statistische und Probabilistische Methoden der Modellwahl. Oberwolfach Report 47* (2005).

2. Tu, C., Clyde, M.A., Wolpert, R.L. "Lévy Adaptive Regression Kernels.", *Technical Report.* Institute of Statistics and Decision Sciences, Duke University (2006).

3. Tu, C., Clyde, M.A., Wolpert, R.L. "An Adaptive Kernel Smoothing Approach to Modeling Multivariate Time Series of Air Pollutants", *Technical Report.* Institute of Statistics and Decision Sciences, Duke University (2006).

4. Tu, C., Clyde, M.A., Wolpert, R.L. "Adaptive Kernel Methods in Spatio-Temporal Modeling.", *Technical Report.* Institute of Statistics and Decision Sciences, Duke University (2006).