UNIVERSITY OF CALIFORNIA
SANTA CRUZ

DISTRIBUTED INFERENCE AND DATA SKETCHING FOR HIGH DIMENSIONAL SPATIAL REGRESSION MODELS

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

DOCTOR OF PHILOSOPHY

in

STATISTICAL SCIENCE

by

Laura N. Baracaldo

June 2022

The Dissertation of Laura N. Baracaldo
is approved:

Professor Rajarshi Guhaniyogi, Chair

Professor Herbert Lee

Professor John Kornak

Professor Robert Lund

Peter Biehl
Vice Provost and Dean of Graduate Studies

Table of Contents

List of Figures

List of Tables

Abstract

Distributed Inference and Data Sketching for High Dimensional Spatial Regression Models

by

Laura N. Baracaldo

Modeling spatial data with flexible statistical models has become an enormously active area of research over the last decade in many disciplines. This work focuses on scaling MC computations for large-scale Bayesian inference in complex spatial models with adequate point estimation and uncertainty in inference and prediction. We first derive a three-step distributed Bayesian inferential framework for multivariate spatial generalized linear mixed effect models (MVspGLMMs) for big data. The proposed approach delivers fully model-based Bayesian parameter inference based on the construction of the “meta posterior” as the Wasserstein Barycenter of pseudo posterior distributions obtained from the partition of the data into independent subsets.

We introduce Bayesian data sketching for spatially varying coefficient regression models (SVCM) to obviate computational challenges presented by large numbers of spatial locations. To address the challenges of analyzing very large spatial data, we compress spatially oriented data by a random linear transformation to achieve dimension reduction and conduct inference on the compressed data. We establish posterior contraction rates for estimating the spatially varying coefficients and predicting the outcome at new locations under the randomly compressed data model.

Finally, we present a novel idea that employs data sketching for distributed Bayesian inference. The proposed model takes advantage of parallel computation by performing Bayesian inference built on the aggregation of “sketched subset posteriors". This approach addresses spatial variable selection in SVCMs with big data without developing fundamentally new models or algorithms or making use of any specialized computational hardware. The models are empirically illustrated by simulation experiments and by conducting a spatial analysis of remote sensed vegetation data.

Acknowledgments

I would like to express my deepest gratitude to the countless people who have held me up during my doctorate. First and foremost, I want to acknowledge the invaluable impact of my advisor Prof. Raj Guhaniyogi on my research process and my career. He not only devoted his time, experience and knowledge into guiding me throughout this work, but also motivated and supported me in my academic aspirations. I will always be grateful for his generosity and thoughtful advise. I am indebted also to Prof. Sudipto Banerjee, whose vision was fundamental to great part of this dissertation.

I am thankful to all Faculty members at the Statistics department, who contributed in many ways to my professional growth and statistical learning through the entirety of my studies. I want to specially thank Professors Robert Lund, Herbert Lee and John Kornak for serving on my defense reading committee.

I feel fortunate to having had the opportunity to share this journey with my cohort fellows, specially Hyotae and Wenjie, who made the process more enriching and enjoyable.

I am endlessly grateful to my family, Andrés and Luffy, who held my spirits up at every moment with their joy, patience and unconditional company. I am thankful to my parents, and my little sisters, who are part of every step I accomplish.

Chapter 1

Introduction

The growing capabilities of Geographical Information Systems (GIS) and associated software have effectuated unprecedented access to spatial data. Statisticians today routinely encounter geographically referenced datasets containing a large number of irregularly located observations on multiple variables. This has, in turn, fueled considerable interest in statistical modeling for location-referenced spatial data; see, for example, the books by Schabenberger and Gotway (2004), Gelfand *et al.* (2010a), Cressie and Wikle (2015) and Banerjee *et al.* (2014a) for a variety of methods and applications. The need to model spatially-referenced outcomes, perhaps vector-valued, across large domains is ripe in many fields, such as environmental and health sciences.

Gaussian processes and their variants are often used to model the complex spatial dependence between variables. For example, regression between different spatially geocoded variables are often performed using spatial varying coefficient models, where the coefficients corresponding to every spatially varying predictor is also assumed to be spatially varying, and a stochastic process prior, most often a Gaussian process prior, is assigned on these spatially varying coefficients. Details on this topic is available in chapter 2. While posterior Monte Carlo (MC; Markov chain or sequential Monte Carlo) computations in high-dimensional linear models have received significant attention, computations in even state-of-the-art spatial models with stochastic process priors on spatially varying functions are limited to sample sizes that are much smaller than the realistic ones. This prohibits applications of hierarchical nonparametric spatial Bayesian models in massive data settings unless restrictive assumptions are imposed or the accuracy of inference is compromised. Indeed, computational bottlenecks are a major barrier in the application of nonparametric Bayesian models to massive spatial databases in the environmental and health sciences, respectively.

Developing computationally tractable Bayesian spatial models is an active area of research. Structured approximations are the most common tools adopted to resolve the intractability of Bayesian methods. Optimization-based methods are efficient in obtaining an analytic approximation of the true posterior distribution, with Laplace approximation, variational Bayes, and expectation propagation being the most common techniques. Optimization, however, can be nontrivial for complex likelihoods frequently used in nonparametric Bayesian models based on stochastic processes. It is also known that variational Bayes and expectation propagation can often be highly biased in the estimation of posterior uncertainty and dependence. The literature on a sampling-based approach to posterior inference with large data has exploded in the last two decades, with a majority of methods largely falling into a few categories. First, there are model-based solutions which replace Gaussian process or any other computationally prohibitive stochastic process by their computationally efficient variants. Second, there are likelihood approximation techniques which replace the computationally cumbersome likelihood for the correlated spatial data with their computationally efficient versions, mainly by ignoring dependence between random variables at some level.

There are also algorithmic approaches to solve to computational bottleneck in MC computation of large data. For example, there are subsampling-based methods which obtain posterior samples conditioned on a small fraction of the data. Frequently, sampling is coupled with modified Hamiltonian or Langevin Dynamics for improved posterior exploration. Some MC methods replace the exact Markov transition kernel by an approximation that significantly reduces the time required to finish an iteration of the sampler. Another approach proposes dividing the data into a large number of subsets, draw inference on data subsets to construct subset posteriors, followed by combining subset posteriors using a notion of geometric center in the space of distributions. These methods are referred to as the divide-and-conquer strategy which has gained attention in the last few years. On the divide-and-conquer front, there is recent literature on exploiting this strategy to draw scalable inference in parametric models or non-parametric spatial process models with Gaussian errors. However, no effort has been undertaken as of yet to develop divide-and-conquer methods for scalable inference with non-Gaussian spatial data.

This thesis has proposed divide-and-conquer methodology for non-Gaussian multivariate spatial data. The core to the divide-and-conquer algorithm is combining subset posteriors through their geometric center in the distribution space. To this end, we will employ Wasserstein Barycenter of order 2 for subset posteriors which is conceptual similarity with an ordinary mean of scalar quantities. We provide a brief overview of Wasserstein Barycenter in the next section.

1.1 Wasserstein Barycenter of Distributions

This section provides a brief description of Wasserstein barycenter. Let $(\mathcal{X}, \rho)$ be a complete separable metric space and $\mathcal{F}(\mathcal{X})$ be the space of all probability measures on $\mathcal{X}$. The Wasserstein space of order 2 is a set of probability distributions defined as $\mathcal{F}_2(\mathcal{X}) = \{F \in \mathcal{F}(\mathcal{X}) : \int_{\mathcal{X}} \rho^2(\alpha, \alpha_0)F(d\alpha) < \infty\}$, where $\alpha_0 \in \mathcal{X}$ is arbitrary and $\mathcal{F}_2(\mathcal{X})$ does not depend on the choice of α_0. The Wasserstein distance of order 2, denoted as W_2, is a metric on $\mathcal{F}_2(\mathcal{X})$. Let F_1, F_2 be two probability measures in $\mathcal{F}_2(\mathcal{X})$ and $\Pi(F_1, F_2)$ be the set of all probability measures on $\mathcal{X} \times \mathcal{X}$ with marginals F_1 and F_2, then W_2 distance between F_1 and F_2 is defined as $W_2(F_1, F_2) = \{\inf\limits_{\pi \in \Pi(F_1,F_2)} \int_{\mathcal{X}\times\mathcal{X}} \rho^2(x, y)\, d\pi(x, y)\}^{1/2}$. Let $\nu_1, \dots, \nu_K \in \mathcal{F}_2(\mathcal{X})$, then the Wasserstein barycenter of $\nu_1, \dots, \nu_K$ is defined as

$$\overline{\nu} = \underset{\nu \in \mathcal{F}_2(\mathcal{X})}{\arg\min} \frac{1}{K} \sum_{k=1}^{K} W_2^2(\nu, \nu_k), \tag{1.1}$$

which can be viewed as the geometric center of the probability measures $\nu_1, \dots, \nu_K$. It is known that $\overline{\nu}$ exists and is unique (Agueh and Carlier, 2011a).

In the one-dimensional case, the Wasserstein barycenter has an explicit relation with the K probability measures. Let $F^{-1}(u) = inf\ \{x : F(x) \geq u\}$ be the quantile function of a generic univariate distribution function $F(x)$. Let F_1 and F_2 be two univariate distributions in $\mathcal{F}_2(\mathcal{X})$, with quantile functions $F_1^{-1}(u)$ and $F_2^{-1}(u)$, for any $u \in (0, 1)$, respectively. Then the W_2 distance between F_1 and F_2 has an explicit expression by Lemma 8.2 of Bickel and Freedman (1981)

$$W_2(F_1, F_2) = \left[\int_0^1 \left\{F_1^{-1}(u) - F_2^{-1}(u)\right\}^2 du\right]^{1/2}. \tag{1.2}$$

(1.2) provides an explicit expression for the Wasserstein barycenter as given by the following lemma.

Lemma 1.1.1. *For one dimensional case, Wasserstein barycenter $\overline{\nu}$ for $\nu_1, ..., \nu_K$ satisfies $\overline{\nu}^{-1}(u) = \frac{1}{K}\sum_{k=1}^{K} \nu_k^{-1}(u)$, for all $u \in (0, 1)$.*

Proof. We will show that $\alpha^{-1}(u) = \frac{1}{K}\sum_{k=1}^{K} \nu_k^{-1}(u)$ minimizes $\sum_{k=1}^{K} W_2^2(\nu, \nu_k)$ over all $\nu \in \mathcal{F}_2(\mathcal{X})$. For any $\nu \in \mathcal{F}_2(\mathcal{X})$,

$$\begin{aligned}
&\sum_{k=1}^{K} W_2^2(\nu, \nu_k) = \sum_{k=1}^{K} \int_0^1 \left(\nu^{-1}(u) - \nu_k^{-1}(u)\right)^2 du \\
=&\sum_{k=1}^{K} \int_0^1 \left(\left(\nu^{-1}(u) - \alpha^{-1}(u)\right) - \left(\nu_k^{-1}(u) - \alpha^{-1}(u)\right)\right)^2 du \\
=&\sum_{k=1}^{K} W_2^2(\nu, \alpha) + \sum_{k=1}^{K} W_2^2(\alpha, \nu_k) - \sum_{k=1}^{K} \int_0^1 2\left(\alpha^{-1}(u) - \nu^{-1}(u)\right)\left(\alpha^{-1}(u) - \nu_k^{-1}(u)\right) du
\end{aligned}$$

Note that, for every $u \in (0,1)$, $\sum_{k=1}^{K}(\alpha^{-1}(u) - \nu_k^{-1}(u)) = 0$, by the definition of $\alpha^{-1}(u)$. Thus,

$$\begin{aligned}
&\sum_{k=1}^{K} \int_0^1 2\left(\alpha^{-1}(u) - \nu^{-1}(u)\right)\left(\alpha^{-1}(u) - \nu_k^{-1}(u)\right) du \\
&= \int_0^1 2\left(\alpha^{-1}(u) - \nu^{-1}(u)\right) \sum_{k=1}^{K}\left(\alpha^{-1}(u) - \nu_k^{-1}(u)\right) du = 0,
\end{aligned}$$

since both $\alpha^{-1}(u)$ and $\nu^{-1}(u)$ can be taken outside of the summation. Therefore, $\sum_{k=1}^{K} W_2^2(\nu, \nu_k) = \sum_{k=1}^{K} W_2^2(\nu, \alpha) + \sum_{k=1}^{K} W_2^2(\alpha, \nu_k)$. As $W_2^2(\nu, \alpha) \geq 0$, this implies that $\overline{\nu}^{-1}(u) = \alpha^{-1}(u)$, which proves the theorem.

Therefore $\overline{\nu}$ in (1.1) is related to $\nu_1, \ldots, \nu_K$ by

$$\overline{\nu}^{-1}(u) = \frac{1}{K} \sum_{k=1}^{K} \nu_k^{-1}(u), \tag{1.3}$$

where $\nu_k^{-1}(u)$ and $\overline{\nu}^{-1}(u)$ are the quantile functions of ν_k and $\overline{\nu}$, respectively. This expression for the one-dimensional W_2 barycenter has been derived in Agueh and Carlier (2011b) from an optimal transport perspective. The relation indicates that for a scalar functional, the average of subset posterior quantiles produces another quantile function that corresponds exactly to the one-dimensional Wasserstein posterior.

1.2 Data Sketching Algorithms

This thesis also proposes a new idea of data sketching to address the computational issue in spatial models with massive data. Sketching is a probabilistic data compression technique that has been largely explored and developed in Computer Science,(Cormode *et al.* (2011)) to address the challenges of storing and processing massive data sets. The advantages of data sketching approaches include less memory consumption, faster implementation of algorithms, and reduced bandwidth requirements in distributed computing environments.

The domain of sketching algorithms has been extended to a wide range, including processing massive data streams with low memory footprint, "compressed sensing" (Candes and Tao (2005); Donoho (2006)) for irreversible compression of signals with few linear measurements, and dimensionality reduction or "random projection" methods for speedups in large-scale linear algebra algorithms (Braverman *et al.* (2010); Dasgupta *et al.* (2010); Kane and Nelson (2010)) , and high-dimensional computational geometry (Clarkson and Woodruff (2017); Meng and Mahoney (2013); Nelson and Nguyên (2013)).

A *sketch* (Broder (1997)) is defined as a small-space data structure which acts as a compact representation of a much larger data set and that allows the execution of a set of pre-specified tasks. More specifically, we consider the case where the data can be expressed as a high-dimensional vector $y \in \mathbb{R}^N$ and the *linear sketch* of y is obtained by some randomized linear mapping $y \mapsto \Phi y$, where Φ denotes a $M \times N$ compression matrix, with $M << N$. The construction of sketching algorithms is based on the idea that the sketching matrix Φ is an ϵ-subspace embedding (Woodruff (2014), Meng and Mahoney (2013), Yang *et al.* (2015)) for the source data-set. Broadly speaking, an ϵ-subspace embedding preserves the linear structure of the source up to some multiplicative $(1 \pm \epsilon)$ factor. When ϵ

is small, the linear mapping Φ preserves the covariance structure of the original data set, therefore, the compressed data serves as a surrogate of the full data. Formally, for a given $N \times P$ matrix X, we call a $M \times N$ matrix Φ and ϵ-subspace embedding for X, if for all vectors $z \in \mathbb{R}^P$

$$(1-\epsilon)\|Xz\|_2^2 \leq \|\Phi Xz\|_2^2 \leq (1+\epsilon)\|Xz\|_2^2 \tag{1.4}$$

There are two broad classes of distributions for the random matrix Φ, data aware random projections and data oblivious random projections. While data aware random projections use information from the source data y, data oblivious random projections can be obtained without any knowledge on y. The former category is linked to finite population sampling methods, and the later category is related to dimension reduction techniques such as multidimensional scaling as these matrices are designed to offer a ϵ-subspace embedding for an arbitrary data collection with high probability.

The idea of data sketching has been prevalent in computer science and machine learning literature, mainly in unsupervised models, ordinary linear regressions and high dimensional penalized linear regressions. However, the usage of data compression to address challenges in high dimensional Bayesian inference, especially in spatial regression models with big data has not been explored. An important contribution of this thesis is to introduce these concepts in the realm of high dimensional spatial models under the Bayesian framework. Please refer to Chapter 3 and 4 for more details.

1.3 Roadmap

In Chapter 2, we investigate the three step divide-and-conquer Bayesian inferential framework to enhance scalability of multivariate spatial generalized linear mixed models (spGLMMs) for multivariate binary spatial data, where multivariate spatial Gaussian processes employ linear model co-regionalization (LMC) to account for correlation between multivariate spatial observations. The approach is conceptually simple, yields parametric inference with uncertainty and seamlessly scales with big data due to the distributed framework which avoids storage and computation of subsets of the data in different processors, and minimizes communication between different processors. We provide empirical illustration that reveals inferential accuracy of our approach on multivariate binary observations.

In chapter 3, we introduce data sketching within the framework of Bayesian spatially varying coefficient regression models to obviate computational challenges that arise from the analysis of large numbers of spatial locations. We start by reviewing current developments in the implementation of data sketching in the context of ordinary linear regression and penalized linear regression models. Then, the model based on data sketching is proposed. We adapt results from random matrix theory and develop new theoretical results to establish posterior contraction rates for estimating the spatially varying coefficients and predicting the outcome at new locations under the randomly compressed data model. Finally, we present some simulation experiments and conduct a spatial analysis of remote sensed vegetation data to empirically illustrate the inferential and computational efficiency of our approach.

In chapter 4, we propose a three-stage strategy built on the idea of Bayesian data sketching to simultaneously perform variable selection and coefficient estimation in non-parametric spatially varying coefficient models (SVCMs). This scheme, which incorporates data compression within the divide and conquer architecture, takes advantage of parallel computation by independently fitting the model across data sketches to obtain “sketched subset posteriors” for varying coefficients, which are combined through an aggregation algorithm with the purpose of delivering fully model-based Bayesian inference and prediction. The proposed approach solves the issue of sensitivity due to subset selection in the divide-and-conquer Bayesian inference for spatial data. Simulation studies and geo-statistical analysis of a remote sensing data validate our approach. Finally, Chapter 5 concludes the thesis with a brief discussion on future work. Proofs of the theorems are available in Appendix A.

Chapter 2

Distributed Inference for Multivariate Spatial Generalized Linear Models

2.1 Introduction

With the rapid development of Geographical Information Systems (GIS), along with related software, statisticians today routinely encounter large spatial datasets containing multiple spatially correlated variables observed across thousands of locations. This chapter is motivated by applications where spatially correlated variables are binary and model fitting becomes necessary for binary non-Gaussian spatially correlated datasets. To this end, multivariate spatial generalized linear mixed effect models (spGLMMs) offer a remarkably flexible class of models that uses multivariate spatial random effects to capture space-varying association between responses, and enables interpolating observations across a continuous spatial domain. We focus on scenarios where spGLMMs are applied for point-referenced/geo-spatial data where the correlation between random effects are captured using multivariate Gaussian process priors or their variants (Banerjee *et al.*, 2014b; Cressie and Wikle, 2015). However, such models implemented with the Markov Chain Monte Carlo (MCMC) algorithm involves matrix decompositions whose complexity increases as $O(N^3Q^3)$ in the number of observations, N, and the outcome dimensions Q, at every iteration of the MCMC algorithm. These computationally intensive matrix calculations limit inference with multivariate binary spatial Gaussian process models to even moderately large datasets (Cressie and Wikle, 2015; Banerjee *et al.*, 2014b).

There is a burgeoning literature on the analysis of large spatial datasets mainly in the context of univariate spatial data. Briefly, these methods are based on approximating Gaussian process models with a low-rank model or with a sparse model. Low-rank processes are usually derived from expressing the Gaussian process using basis functions (Cressie and Johannesson, 2008; Banerjee *et al.*, 2008; Finley *et al.*, 2009; Guhaniyogi *et al.*, 2011). Scalability for low-rank models are further enhanced by fitting them into multiple resolutions, where the resolutions at finer scales capture local variation of the response (Guhaniyogi and Sansó, 2018; Katzfuss and Guinness, 2021). In contrast, sparse models assume that spatial correlation between two distantly located observations is nearly zero, so little information is lost by assuming conditional independence given the intermediate locations. A few important classes of sparse models have emerged recently, such as models based on covariance tapering (e.g., Furrer *et al.* (2006), Kaufman *et al.* (2008), Du *et al.* (2009), Shaby and Ruppert (2012)), composite likelihoods (e.g., Eidsvik *et al.* (2014)) or nearest neighbor models (e.g., Vecchia (1988a); Rue *et al.* (2009); Stein *et al.* (2004); Datta *et al.* (2016a)). Another class of approaches divide the domain into subdomains and fit stochastic process models in different subdomains while stiching inferences in subdomains together with another layer of Markov random fields (Gramacy and Lee, 2008; Peruzzi *et al.*, 2020). Recent articles extend low-rank or sparse processes to enhance scalability of multivariate spatial data (Banerjee *et al.*, 2010; Guhaniyogi *et al.*, 2013; Guhaniyogi, 2017). While multivariate spatial modeling of big data with Gaussian errors have received attention, considerably less attention has been given in the literature to boost scalability of multivariate spGLMMs for big data, except for a few recent articles such as Guan and Haran (2018) which discuss scalability for big data but struggle to scale beyond a few thousand observations. While the authors discuss the strategy to bring this approach under the INLA (Rue *et al.*, 2009; Lindgren *et al.*, 2011) framework to accrue additional computational gain, it is likely to face significant challenges in terms of accurately estimating spatial surface.

This chapter proposes a three step distributed Bayesian inferential approach for multivariate spGLMMs with multivariate binary spatial data to address the computational issue. The first step of the algorithm constructs K subsets by randomly sampling M data points without replacement from the full data of size N, where K is large and posterior computations with M data points is tractable. The second step fits the spGLMM for the binary spatial data in K subsets in parallel to obtain MCMC based approximation of the full data posterior from each subset. We only retain the posterior samples from each subset which are saved in a CPU dedicated to itself. The posteriors from various subsets (also known as "subset posteriors") are then combined optimally in the third step to yield a single posterior distribution (the "meta-posterior") for the model parameters that substitutes the full posterior distribution for all inferential tasks. To combine subset posteriors, we adapt the notion of Wasserstein barycenter of subset posteriors (see, e.g., Guhaniyogi *et al.* (2020b); Guhaniyogi *et al.* (2020a)). Though the idea has been proposed in the context of predictive models for independent data (Srivastava *et al.*, 2015), for scaling univariate (Guhaniyogi *et al.*, 2020b) and multivariate spatial models (Guhaniyogi *et al.*, 2020a) with continuous response, this chapter proposes to employ this technique to enhance scalability for spGLMMs with binary multivariate spatial data. The approach simply demands obtaining posterior samples for the process parameters from a multivariate spatial model fitted on multiple subsets in parallel, followed by deriving the meta-posterior. We emphasize that the third step of combining subset posteriors is agnostic to the specific spGLMM used in the second step of the algorithm; hence the algorithm offers a general framework for distributed inference with spGLMMs.

The remainder of the chapter evolves as follows. Section 2.2 discusses the divide-and-conquer framework for spGLMMs with binary data, while Section 2.3 demonstrates its empirical performance with a brief simulation study. Finally, Section 2.5 concludes the chapter with some discussion and general conclusions.

2.2 Distributed Bayesian Inference with Multivariate Spatial Generalized Linear Mixed Effect Models with Binary Data

Let $\mathcal{D} \subset \Re^2$ be the spatial domain of interest and let s be a generic point in $\mathcal{D}$. Let $y(s) = (y_1(s), ..., y_Q(s))^T$ denote the response and $X(s)$ denote a $Q \times P$ predictor matrix at location $s \in \mathcal{D} \subseteq \mathbb{R}^2$ with its l-th row given by $x_l(s)^T$. The relationship between $y(s)$ and $X(s)$ is expressed using a spatial generalized linear mixed effect model (spGLMM) given by,

$$g(E[y(s)|\beta, w(s)]) = X(s)\beta + w(s), \tag{2.1}$$

where $g(\cdot)$ is an appropriate link function, $\beta \in \mathcal{R}^P$ is the $P \times 1$ coefficient vector corresponding to the predictor matrix $X(s)$ and $w(s) = (w_1(s), ..., w_Q(s))^T$ is a $Q \times 1$ vector of spatial process, modeling local patterns with structured dependence to the mean. The model specification suggests that conditional on the unknown function $w(\cdot)$ at two locations s and t, $y(s)$ and $y(t)$ are independent of each other. Here on, we will focus on a popular example for spGLMMs on spatial binary data, the binary response being modeled by a Bernoulli distribution with the logit link. Therefore, when $y_l(s) \in \{0, 1\}$, (2.1) with $g(\cdot)$ as the logit link function becomes

$$p(y_l(s)|\beta, w_l(s)) = \frac{\exp(y_l(s)(x_l(s)^T\beta + w_l(s)))}{1 + \exp(x_l(s)^T\beta + w_l(s))}, \ \ l = 1, ..., Q. \tag{2.2}$$

Although our approach is illustrated based on (2.2) from the class of spGLMMs, the approach presented in this article generalizes to other link functions and observation models, and to the cases where an additional nugget term is present in (2.1) (Berrett and Calder, 2016).

The customary process specification for $w(s)$ is a zero-centered Q-variate Gaussian process. The process $w(s)$ is completely specified by its *cross-covariance* function $C_w(s, t; \theta)$, which, for any pair of locations s and t, is a $Q \times Q$ matrix with $\text{cov}\{w_{l_1}(s), w_{l_2}(t)\}$ as its (l_1, l_2)-th element and θ is a collection of process parameters. Therefore, $C_w(s, s; \theta)$ is precisely the variance-covariance matrix for the elements of $w(s)$ within site s. While this chapter considers one popular way to construct the cross-covariance matrix by linear model co-regionalization (LMC) as described below, we refer to Gelfand *et al.* (2010b); Gneiting *et al.* (2010); Guhaniyogi *et al.* (2013) for versatile constructions of valid cross-covariance matrix functions.

Assume, $v(s) = (v_1(s), ..., v_Q(s))^T$ is a Q-variate vector with its components $v_l(s)$'s are following independent univariate Gaussian processes with mean 0 and covariance function $\alpha(s,t;\Delta_l)$. The LMC approach (Gelfand *et al.*, 2004) models correlation between components of $w(s)$ by assuming $w(s) = \Gamma v(s)$, which yields a structured cross-covariance function given by,

$$C_w(s,t;\theta) = \Gamma C_v(s,t;\Delta)\Gamma^T = \sum_{l=1}^{Q} \gamma_l \gamma_l^T \alpha(s,t;\Delta_l) \ , \tag{2.3}$$

where γ_l is the l-th column of Γ, $\Delta = \{\Delta_1, ..., \Delta_Q\}$ (in univariate case $\Delta = \Delta_1$) and θ is the collection of all $\Delta_1, .., \Delta_Q$ and Γ. One natural choice of $\alpha(\cdot,\cdot)$ comes from the class of Matérn correlation functions,

$$\alpha(s,t;\Delta_l) = \frac{1}{2^{\xi_{2,l}-1}\Gamma(\xi_{2,l})}(||s-t||\xi_{1,l})^{\xi_{2,l}} K_{\xi_{2,l}}(||s-t||;\xi_{1,l});\ \xi_{1,l} > 0, \xi_{2,l} > 0, \tag{2.4}$$

Where K_ν is a modified Bessel function of the second kind of order ν, and $\xi_{1,l}$, $\xi_{2,l}$ are the decay and smoothness parameters respectively; $\Delta_l = (\xi_{1,l}, \xi_{2,l})$. This implies that for any finite set of n locations, say $S = \{s_1, s_2, \ldots, s_N\}$, the $NQ \times 1$ vector of realizations, $w = (w(s_1)^T, w(s_2)^T, \ldots, w(s_N)^T)^T$ follows a multivariate normal distribution with zero mean and a $NQ \times NQ$ blocked covariance matrix $C_w(\theta)$ whose (i,j)-th block is given by the $Q \times Q$ matrix $C_w(s_i, s_j; \theta)$.

2.2.1 Challenges in Posterior Computation in spGLMMs with Big Data

With $y(s)$ and $X(s)$ observed at a set of locations $S = \{s_i : i = 1, 2, \ldots, N\}$, (2.2) does not allow full conditional posterior distributions of parameters in closed forms. However, the posterior computation is enormously simplified using the data augmentation approach based on the result, $\frac{e^{a\psi}}{(1+e^{\psi})^b} = 2^{-b} e^{\kappa\psi} \int_0^\infty e^{-\omega\psi^2/2} p_\omega(\omega) d\omega$ (Theorem 1 in Polson *et al.* (2013)), where $p_\omega(\omega)$ stands for a Polya-Gamma distribution with parameters 0 and b, denoted as $PG(b, 0)$, $\kappa = a - b/2$. Introducing a latent variable $\omega_{i,l}$ corresponding to the lth component of the ith observation, (2.2) can be equivalently written as

$$p(y_l(s_i)|\omega_{i,l}, \beta, w_l(s_i)) \propto \exp\{-\omega_{i,l}(x_l(\boldsymbol{s}_i)^T\beta - \kappa_{i,l}/\omega_{i,l})^2/2\},\ \ \omega_{i,l} \sim PG(1,0),$$
$$\kappa_{i,l} = y_l(\boldsymbol{s}_i) - 1/2. \tag{2.5}$$

Assume that $z = (z_1^T, ..., z_N^T)^T$ and $\Omega = diag(\omega_1, ..., \omega_N)$, where $z_i = (\kappa_{i,1}/\omega_{i,1}, ..., \kappa_{i,Q}/\omega_{i,Q})^T$ and $\omega_i = (\omega_{i,1}, ..., \omega_{i,Q})^T$. Then (2.2) leads to the hierarchical mixed model framework

$$z = X\beta + w + \epsilon,\ \ \epsilon \sim N(0, \Omega^{-1}),\ \ \omega_{i,l} \sim PG(1,0), \tag{2.6}$$

where X is the $NQ \times P$ matrix of regressors ($P < N$) with $X(s_i)$ as its i-th block row of dimension $Q \times P$. We specify $\beta \sim N(\mu_\beta, \Sigma_\beta)$ as the prior distribution for the slope vector, where μ_β and Σ_β are assumed fixed, and θ is assigned a proper prior distribution $p(\theta)$. Bayesian inference proceeds, customarily, by sampling $\Theta = \{\beta, \theta\}$ from (2.6) using Markov chain Monte Carlo (MCMC) methods. Irrespective of the specific parametrization or estimation algorithm, model fitting usually involves matrix decompositions for $C_w(\theta)$ requiring $\sim (NQ)^3$ floating point operations (flops) and $\sim (NQ)^2$ memory units in storage. These become prohibitive for large N since $C_w(\theta)$, in general, has no exploitable structure.

2.2.2 Distributed Inference with the Meta Kriging Approach

Let $\mathcal{S}$ be partitioned into K exhaustive and mutually exclusive subsets $\mathcal{S}_1, ..., \mathcal{S}_K$. Let $\{y_h, X_h\}$ be the corresponding data partitions, for $h = 1, 2, \dots, K$, where y_h is an $N_hQ \times 1$ vector and X_h is an $N_hQ \times P$ matrix. This chapter assumes all subsets are equal sized and each of them contains $M = N/K$ data points. Assume that we are able to obtain posterior samples for Θ from (2.6) applied independently to each of K subsets of the data. Let $\Pi(\Theta|y_h, X_h)$ be the posterior distribution of Θ from the hth subset, referred to as the hth subset posterior. Assume that $\Theta_{h,1}, ..., \Theta_{h,F}$ are the F post burn-in posterior samples from $\Pi(\Theta|y_h, X_h)$. Following Guhaniyogi *et al.* (2020a), we propose to combine $\Pi(\Theta|y_h, X_h)$'s to arrive at a legitimate probability density $\Pi_M(\Theta|y, X)$, referred to as the "meta-posterior".

For notational simplicity, we denote $\Pi(\Theta \mid y_h, X_h)$ by Π_h. We define the "meta posterior" as the Wasserstein barycenter Π_M as in (1.1), which provides a general notion of obtaining the mean of K possibly dependent subset posterior distributions in the space of distributions. While the posterior distribution $\Pi(\Theta|y, X)$ obtained from the full data are analytically intractable and computationally prohibitive, it can be well approximated by the meta posterior. Given the MCMC samples of Θ from subset posteriors are available, one can conveniently estimate the empirical version of the meta posterior. Since the primary interest often lies in the Bayesian inference of one dimensional functionals of the model parameters, we adopt the algorithm outlined in Li *et al.* (2017) and Guhaniyogi *et al.* (2020b) following the general discussion in Section 1.1 of Chapter 1 to compute Wasserstein barycenter of the posterior distributions of one-dimensional parameters. More specifically, if Θ is one-dimensional and $\Theta_{h,(q^*)}$ denotes the q^*-th empirical quantile of Θ obtained from Π_h, $h = 1, ..., K$, then $\overline{\Theta}_{q^*} = (1/K)\sum_{h=1}^{K} \Theta_{h,(q^*)}$, $q^* = 1, ..., Q^*$, denotes the q^*th empirical quantile of Π_M. We choose $Q^* = 10^4$ and compute quantiles on an equi-spaced grid of size Q^* on $(0, 1)$.

While, for empirical illustration, we employ the algorithm leading to computation of Wasserstein barycenter for one-dimensional parameters, it is possible to utilize the existing literature focusing on combination of distributions for multivariate parameters. For example, one can efficiently solve a sparse linear program as described in Cuturi and Doucet (2014); Srivastava *et al.* (2018) to compute Wasserstein barycenter when Θ is multivariate. Alternatively, we can follow combination algorithm outlined in Guhaniyogi *et al.* (2020a) to compute an empirical approximation to the meta posterior. It has been shown that for independent data, the Wasserstein barycenter is a preferable choice to several other combination methods; for example, directly averaging over many subset posterior densities with different means can usually result in an undesirable multimodal meta posterior distribution, but the Wasserstein barycenter does not have this problem and can recover a unimodal posterior (Srivastava *et al.*, 2018). Besides, it does not rely on the asymptotic normality of the subset posterior distributions as in other approaches, such as consensus Monte Carlo (Scott *et al.*, 2016).

Choice of K: One important ingredient in the distributed inference is the choice of subsets K. While choosing K small will not be useful for computational efficiency, choice of K to be large will yield less accurate inference. Thus, K needs to be chosen by striking a balance between inferential accuracy and computational efficiency. There is existing literature that studies theoretically the "optimal choice of K" depending on N and the smoothness of the spatial surface for spatial Gaussian process model (Guhaniyogi *et al.*, 2020b) and spatio-temporal varying coefficient models (Guhaniyogi *et al.*, 2020a) with continuous outcomes. In absence of an analogous theoretical results for spGLMMs we choose K to be moderately large and ensure that subset posteriors are not very different from each other by calculating the KL-divergence between the subset posteriors empirically. We plan to explore this topic theoretically in more detail elsewhere.

2.3 Empirical Study

This section presents empirical performance of the proposed distributed framework for spGLMMs with a simulated multivariate spatial binary data. In our simulation example the data is randomly divided into K exhaustive and mutually exclusive subsets. Our approach is implemented in `R`, with `foreach` and `doParallel` packages are used for multicore parallelization. A detailed implementation with open source codes can be found in the GitHub page `https://github.com/LauraBaracaldo/Spatial-Meta-Kriging-for-Distributed-Inference-for-Binary-Response`.

To illustrate the performance, we set $Q = 2$ and simulate $N = 10,000$ bivariate observations within a unit square domain from model (2.2). The covariate matrix $X(s)$ at any location is taken to be a 2×4 matrix with the first row $x_1(s)^T = (1, u_1(s), 0, 0)$, and the second row $x_2(s)^T = (0, 0, 1, u_2(s))$ where $u_1(s), u_2(s)$ are drawn i.i.d from N(0,1), and the corresponding coefficient $\beta = (\beta_{01}, \beta_{11}, \beta_{02}, \beta_{12})^T$. An exponential spatial correlation function was assumed for all spatial processes, i.e., $\xi_{2,l}$ was fixed at 0.5 in (2.4) for $l = 1, 2$. For the sake of identifiability of each element of Γ, we assume $\Gamma = \begin{pmatrix} \gamma_{11} & 0 \\ \gamma_{21} & \gamma_{22} \end{pmatrix}$, with $\gamma_{11} > 0$. The column labeled *True* in Table 2.1 depicts the true parameter values used to generate the data. We simulate 10 datasets with the true parameter values.

We fit our distributed approach for $K = 10, 20$ to assess how the inference improves as K decreases. For estimating the LMC model in each subset, we assign a flat prior to each component of the intercept β and γ_{21}, and Jeffrey's prior on γ_{11} and γ_{22}. The decay parameters $\xi_{1,1}$ and $\xi_{1,2}$ in the exponential correlation functions are assigned $U(2, 12)$ prior that gives fairly wide support for range parameters given that the maximum inter-location distance in the generated data is 1.34.

Table 2.1: The median and 95% Bayesian credible intervals of parameters for distributed multivariate spGLMM (dmv-spGLMM) for $K = 10, 20$. We also present True Positive Rate (TPR) and True Negative Rate (TNR) for out of sample classification performance. Computation time in minutes are also presented.

	True	dmv-spGLMM ($K = 20$)	dmv-spGLMM ($K = 10$)
β_{01}	1	1.24 (0.83, 1.48)	1.09 (0.75 , 1.43)
β_{11}	-1	-0.94 (-1.22, 0.76)	-0.91 (-1.16, -0.67)
β_{02}	-1	-1.42 (-1.75, -0.89)	-1.26 (-1.62 , -0.92)
β_{12}	1	1.05 (0.77, 1.39)	0.97 (0.73,1.23)
γ_{11}	1	1.14 (0.71, 1.72)	0.81 (0.34, 1.54)
γ_{12}	-0.9	-1.10 (-1.58, -0.79)	-1.11 (-1.77, -0.55)
γ_{22}	1	1.19 (0.84, 2.01)	0.75 (0.41, 1.22)
$\xi_{1,1}$	5	7.25 (4.49, 11.35)	6.97 (4.32, 10.96)
$\xi_{1,2}$	6	7.79 (4.01, 11.56)	7.99 (4.29, 11.79)
TPR	–	0.72	0.74
TNR	–	0.71	0.76
Time(in min)	–	22.46	84.21

Given that γ's and $\xi_{1,l}$'s, for $l = 1, 2$, are not strongly identifiable together (Zhang, 2004), it is a common practice to propose a uniform prior with bounded support on $\xi_{1,l}$'s. Further, by moderately perturbing the range of the uniform distribution, we do not observe any significant change in the performance.

Table 2.1 presents posterior medians along with 95% credible intervals for all the parameters for a reprersentative simulated data. All parameters are estimated accurately by our distributed approach, both for $K = 10, 20$. For both $K = 10, 20$, the point estimates are close to the truth and the 95% credible intervals of all parameters have covered the truth. However, the point estimates seem to be little more accurate corresponding to $K = 10$ than $K = 20$. On a similar note, the 95% credible intervals corresponding to all parameters are narrower for $K = 10$, suggesting improved uncertainty quantification for the parameters as the number of subsets decreases. Overall, we find pretty robust parameter estimation when the number of subsets varies within a certain range. Also, the inference turns out to be similar across all ten simulated datasets.

We also compute True Positive Rate (TPR) and True Negative Rate (TNR) for predictive classification with our meta-approach at 100 out of sample observations. Table 2.1 presents the TPR and TNR values averaged over all ten simulated datasets. Since TPR and TNR vary between 0 and 1 with values close to 1 indicate excellent classification performance, the results show reasonably good predictive classification performance with our approach. As expected, the classification performance is marginally improved when the number of subsets in reduced from $K = 20$ to $K = 10$.

Finally, we note that the computation complexity of our approach is dominated by $(MQ)^3$ floating point operations which leads to manageable run times even with a non-optimized implementation in R. In fact, full Bayesian computation of the model requires less than two hours for both $K = 10, 20$ with $10,000$ data points. We emphasize that the computation can be further enhanced by replacing multivariate Gaussian processes on $w(s)$ with their computationally efficient variants. However, we plan to explore it elsewhere.

2.3.1 Sensitivity Analysis

In order to asses the impact of the prior choice in the posterior inference and prediction, we carry out a sensitivity analysis for four different scenarios. The first scenario contemplates a non informative choice for the priors, so that, $p(\beta)$ has a flat prior $p(\beta) \propto 1$, Γ is Inverse-Wishart with degrees of freedom $df = q = 2$, which leads to high uncertainty about the information in the scale matrix Ψ, $\Gamma \sim IW(df = q, \Psi = 0.1I)$ and $\xi_{1,1}, \xi_{1,2}$ are set to be uniform over a wide positive interval; $\xi_{1,1}, \xi_{1,2} \sim Unif(3, 15)$. Scenario 2, keeps the same priors as in scenario 1, except for β which is multivariate normal $\beta \sim N(0, I)$. On the other hand, scenario 3, contemplates the same priors as in scenario 1 except for Γ, which follows and Inverse-Wishart but with $df = q + 3$, which makes it more informative over the scale parameter Ψ. Lastly, scenario 4 specifies a flat distribution for β, a non-informative prior on Γ but with different scale matrix, $IW(df = q+1, \Psi = I)$, and $\xi_{1,1}, \xi_{1,2}$ are chosen to be uniform over a narrower interval $\xi_{1,1}, \xi_{1,2} \sim Unif(4, 7)$.

Table 2.2 presents the results in terms of point-wise and interval estimation for all parameters, as well as predictive classification performance. As we observe, although results are fairly similar overall, there exist some variations throughout different scenarios, which indicates that the model fitting is subject to the prior choice. We conclude that in general, scenario 1 provides the most robust results.

	$n = 10,000$, $(k = 10)$			$n = 10,000$, $(k = 20)$	
	True	*Case 1*	*Case 2*	*Case 3*	*Case 4*
β_{01}	1	1.08 (0.75, 1.43)	0.44 (-0.27, 0.96)	0.38 (0.19, 0.76)	0.60 (0.16, 1.11)
β_{11}	-1	-0.91 (-1.16,-0.67)	-1.01 (-1.39,-0.66)	-0.52 (-0.87,-0.23)	-0.43(-0.79, 0.01)
β_{02}	-1	-1.26 (-1.62,-0.92)	-0.32 (-0.73,0.07)	-0.57 (-0.83, -0.29)	-0.72 (-1.18,-0.25)
β_{12}	1	0.97 (0.73, 1.23)	0.93 (0.59,1.30)	0.51 (0.25,0.95)	0.22 (0.18 ,0.46)
γ_{11}	1	0.81 (0.34, 1.54)	0.98 (0.54,1.75)	0.28 (0.04,0.49)	0.79(0.30,2.24)
γ_{21}	-1	-1.11 (-1.77,-0.55)	-0.88 (-1.51,-0.47)	-0.36 (-0.72,-0.02)	-0.49 (-1.29,0.01)
γ_{22}	1	0.79 (0.41, 1.22)	1.21 (0.68, 2.22)	0.49 (0.14,0.78)	0.82(0.33,2.07)
$\xi_{1,1}$	5	6.54 (4.32, 10.96)	6.32 (3.51,9.56)	6.22 (3.36,9.68)	5.46 (4.57,6.44)
$\xi_{1,2}$	6	6.99 (4.29, 11.79)	6.69 (3.34,9.80)	6.24 (3.29,9.74)	5.49 (4.61,6.74)
TPR	-	0.74	0.69	0.67	0.68
TNR	-	0.76	0.73	0.65	0.66

Table 2.2: Sensitivity analysis for four prior set-ups: Case 1. Non informative priors: $p(\beta) \propto 1$, $\Gamma \sim IW(df = q, \Psi = 0.1I)$, $\xi_{1,1}, \xi_{1,2} \sim Unif(3, 15)$; Case 2. $\beta \sim N(0, I)$; Case 3. Informative for $\Gamma \sim IW(df = q + 3, \Psi = 0.1I)$; Case 4. $\xi_{1,1}, \xi_{1,2} \sim Unif(4, 7)$, $\Gamma \sim IW(df = q + 1, \Psi = I)$

2.4 Real Data Application

This section is motivated by the study of Biogeochemical cycles (BGC) and its association to the biome, which corresponds to a distinctive classification of large areas based on its dominant plant and vegetation formations. These cycles describe the natural pathways by which essential chemical substances such as carbon, phosphorus and nitrogen circulate through the biosphere, and demonstrate the way in which the energy is used.

In this application we seek to predict the presence or absence of certain terrestrial vegetation. Among the different modeling techniques available, the most relevant prediction methods are the dynamic global vegetation models (Prentice *et al.* (1992); Kucharik *et al.* (2000)) and the ecological niche or bioclimatic/environmental envelope models (Guisan and Zimmermann (2000); Peterson *et al.* (2011); Araújo and Peterson (2012)). The former considers physiological characteristics of biomes or plant functional types, that respond to the environment in terms of phenology, and carbon stocks, among other predictors (Harrison *et al.* (2010)). Other techniques involve the use of Random forest regression and neural networks, which do not consider a structure of spatial correlation (Wessels *et al.* (2011), Sato and Ise (2022)). The purpose of our work was to model the annual grass vegetation $y_1(s) = \mathbb{1}_{AGV}(s)$ and evergreen needleleaf vegetation $y_2(s) = \mathbb{1}_{ENV}(s)$, based on energy storage measurements such as red light reflectance and latent heat flux. Annual grass vegetation is dominated by herbaceous annuals including cereal croplands, whereas evergreen needleleaf vegetation is characterized by evergreen conifer trees and shrubs, with a percentage of woody vegetation cover of over 10%.

We model the bi-variate response $y(s) = (y_1(s), y_2(s))^T$ at locations $s = 1, \ldots, N$ which are projected on a sinusoidal grid, located on the western coast of the United States, between $30°N$ to $40°N$ latitude and $104°W$ to $130°W$ longitude. The covariance matrix $X(s)$ at any location is set to be a 2×4 matrix with the first row $x_1(s)^T = (1, u_1(s), 0, 0)$, and the second row $x_2(s)^T = (0, 0, 1, u_2(s))$, where $u_1(s)$ and $u_2(s)$ are defined in table 2.3. The corresponding vector of coefficients is defined as $\beta = (\beta_{01}, \beta_{11}, \beta_{02}, \beta_{12})^T$, so that β_{01}, β_{02} represent intercept effects, and β_{11}, β_{12} represent the effect of variables $u_1(s)$ and $u_2(s)$ respectively.

Our analysis was focused on a data set with 51000 locations where variables $y_1(s)$, $y_2(s)$, $u_1(s)$ and $u_2(s)$ were observed. For the model fitting process we kept $N = 50000$ locations randomly chosen, whereas the remaining $N^* = 1000$ observations were held out for prediction assessment. Further, we ran the distributed approach based on K exhaustive and mutually exclusive subsets, for $K = 20, 50$ with the purpose of studying differences in estimation and prediction performance based on distinct number of cores.

As in section 2.3, we model the spatial processes through a exponential correlation kernel. The prior specification of all parameters is set to be non informative, so that we assigned a flat prior for coefficients β, a uniform distribution over the interval $(3, 15)$ for decay parameters $\xi_{1,1}$ and $\xi_{1,2}$, and $\Gamma \sim IW(df = 2, \Phi = 0.1I)$.

Table 2.4 offers posterior medians for all parameters along with their corresponding 95% credible intervals. Point-wise estimation resulted similar for both values of $K = 20, 50$, however, we observe that $K = 20$ yields to narrower credible intervals for all parameters, in comparison to $K = 50$, which results in an improvement in terms of estimation precision. Additionally, we present (TPR) and (TNR) which measure the sensitivity and specificity of the predictive classification respectively, based on $N^* = 1000$ out of the sample observations. The results show fairly good classification performance, with a marginal improvement when the number of subsets is lowered to $K = 20$.

	Variable	Definition
AGV	$y_1(s)$	Presence of Annual grass Vegetation.
ENV	$y_2(s)$	Presence of Evergreen Needleleaf Vegetation.
RED	$u_1(s)$	% of red light reflectance.
LE	$u_2(s)$	Latent heat flux.

Table 2.3: Data description.

Table 2.4: The median and 95% Bayesian credible intervals of parameters for distributed multivariate spGLMM (dmv-spGLMM) for $K = 20, 50$. We also present True Positive Rate (TPR) and True Negative Rate (TNR) for out of sample classification performance. Computation time in minutes are also presented.

	dmv-spGLMM ($K = 50$)	dmv-spGLMM ($K = 20$)
$Intercept_1$	0.075(-0.236, 0.389)	0.021(-0.288, 0.291)
RED	-8.555(-10.270, -6.817)	-8.481(-10.261,-6.769)
$Intercept_2$	2.059(1.658, 2.475)	2.112(1.706, 2.491)
LE	-0.375(-0.446, -0.308)	-0.390(-0.461, -0.309)
γ_{11}	4.721(3.976, 5.643)	6.081(4.835, 7.696)
γ_{12}	-5.327(-6.202, -4.537)	-8.241(-9.791, -6.823)
γ_{22}	6.433(5.390, 7.835)	9.235(8.304, 10.401)
$\xi_{1,1}$	8.386(7.065, 8.963)	8.438(7.031, 8.971)
$\xi_{1,2}$	4.399(2.162,7.636)	4.614(2.487, 7.501)
TPR	0.712	0.739
TNR	0.817	0.819
Runtime(in min)	248.9	445.8

2.5 Summary

Many scientific applications encounter large spatial data with multivariate binary observations. Fitting spGLMMs to such data is computationally expensive. We propose a three-step divide-and-conquer approach wherein we first partition the data, fit multivariate spGLMMs with spatial random effects modeled through multivariate Gaussian processes to each subset, and finally combine inferences from subsets. The proposed idea offers a theoretically justifiable framework to scale multivariate spGLMMs with binary responses to large number of observations in manageable computation time. Our empirical investigation shows satisfactory estimation of model parameters by the proposed framework.

Chapter 3

Bayesian Data Sketching for Spatial Regression Models

In this chapter, we develop an inferential framework for spatial data analysis using Bayesian data sketching to achieve scalable inference for massive spatial data sets. “Data sketching” (Vempala, 2005; Halko *et al.*, 2011; Mahoney, 2011; Woodruff, 2014; Guhaniyogi and Dunson, 2015, 2016) is a method of compression that is being increasingly employed for analysing massive amounts of data. The entire data set is compressed before being analysed for computational efficiency. Data sketching proceeds by transforming the original data through a random linear transformation to produce a much smaller number of data samples and we conduct the analysis on the compressed data thereby achieving dimension reduction. Furthermore, the original data is neither accessed nor exactly recoverable from the compressed data, which preserves data confidentiality.

While such developments have primarily focused on ordinary linear regression and penalised linear regression (Zhang *et al.*, 2013; Chen *et al.*, 2015; Dobriban and Liu, 2018; Drineas *et al.*, 2011; Ahfock *et al.*, 2017; Huang, 2018), our innovation lies in developing such methods for spatial regression models. The primary challenge distinguishing the current chapter from existing data sketching methods is our pursuit of inference for the underlying spatial effects in the context of spatially-varying regression models. While bearing some similarities, our current contribution differs from compressed sensing (Donoho, 2006; Ji *et al.*, 2008; Candes and Tao, 2006; Eldar and Kutyniok, 2012; Yuan *et al.*, 2014) in the inferential objectives. Specifically, compressed sensing solves an inverse problem by "nearly" recovering a sparse vector of responses from a smaller set of random linear transformations. In contrast, our spatially referenced response vector is not necessarily sparse. Also, we do not seek to (approximately) recover the response vector, so our method is applicable to situations where preserving confidentiality of the response (and predictors) is important.

We consider a spatially-varying regression model with response $y(s) \in \mathcal{Y} \subseteq \mathbb{R}$ and P predictors $x_1(s), ..., x_P(s) \in \mathcal{X} \subseteq \mathbb{R}$, $s \in \mathcal{D} \subseteq \mathbb{R}^2$ related according to the model

$$y(s) = \sum_{j=1}^{P} x_j(s)\beta_j + \sum_{j=1}^{\tilde{P}} \tilde{x}_j(s)w_j(s) + \epsilon(s) = x(s)^{\mathrm{T}}\beta + \tilde{x}(s)^{\mathrm{T}}w(s) + \epsilon(s) \ , \quad (3.1)$$

where $\beta = (\beta_1, \beta_2, \ldots, \beta_P)^{\mathrm{T}}$ is a $P \times 1$ vector of spatially static coefficients, $\tilde{x}(s) = (\tilde{x}_1(s), \tilde{x}_2(s), \ldots, \tilde{x}_{\tilde{P}}(s))^{\mathrm{T}}$ is a $\tilde{P} \times 1$ vector comprising a subset of predictors from $x(s)$ (so $\tilde{P} \leq P$), $w(s) = (w_1(s), w_2(s), \ldots, w_{\tilde{P}}(s))^{\mathrm{T}}$ is the $\tilde{P} \times 1$ vector of spatially varying regression slopes, and $\epsilon(s) \overset{iid}{\sim} N(0, \sigma^2)$ captures measurement error variation at location s. Such spatially-varying regression coefficient models are effective tools for estimating the spatially varying impact of predictors on the response over space (see, e.g., Gelfand *et al.*, 2003; Wheeler and Calder, 2007; Finley *et al.*, 2011; Guhaniyogi *et al.*, 2013; Kim and Wang, 2021, and references therein). Customary geostatistical regression models with only a spatially-varying intercept emerge if the first column of $x(s)$ is the intercept and $\tilde{P} = 1$ with $\tilde{x}_1(s) = 1$. Spatially-varying coefficient models also offer a process-based alternative to widely used geographically weighted regression (see, e.g., Brunsdon *et al.*, 1996) for modelling nonstationary behaviour in the mean. Finley (2011) offers a comparative analysis and highlights the richness of (3.1) in ecological applications.

Bayesian inference for (3.1) is computationally expensive for large spatial data sets, as are commonplace today, due to the presence of the high-dimensional spatial covariance matrix introduced by $w(s)$ in (3.1). High-dimensional spatial modelling has been attracting significant interest and the burgeoning literature on diverse aspects of scalable methods is too vast to be comprehensively reviewed here (see, e.g., Banerjee, 2017; Heaton *et al.*, 2019, for reviews). Briefly, model-based dimension reduction in spatial models have proceeded from low-rank or fixed rank representations (e.g., Cressie and Johannesson, 2008; Banerjee *et al.*, 2008; Wikle, 2010), multi-resolution approaches (e.g., Nychka *et al.*, 2015; Katzfuss, 2017; Guhaniyogi and Sansó, 2018), sparsity-inducing processes (e.g., Vecchia, 1988b; Datta *et al.*, 2016b; Katzfuss and Guinness, 2021; Peruzzi *et al.*, 2020) and divide-and-conquer approaches such as meta-kriging (Guhaniyogi and Banerjee, 2018; Guhaniyogi *et al.*, 2020b). While most of the aforementioned methods entail new classes of models and approximations, or very specialised high-performance computing architectures, Bayesian data sketching has the advantage that customary exploratory data analysis tools, well-established methods and well-tested available algorithms for implementing (3.1) can be applied to the sketched data set without recourse to new algorithmic or software development.

We pursue fully model-based Bayesian data sketching, where inference proceeds from a hierarchical model (Cressie and Wikle, 2015; Banerjee *et al.*, 2014a). The hierarchical approach to spatial data analysis is widely employed for inferring on model parameters that may be weakly identified from the likelihood alone and, more relevantly for substantive inference, for estimating the latent spatial process over the domain of interest. For analytic tractability we model the varying coefficients using basis expansions (Wikle, 2010; Wang *et al.*, 2008; Wang and Xia, 2009; Bai *et al.*, 2019) rather than Gaussian processes. We exploit and adapt some recent developments in theory of random matrices to relate the inference from the compressed data with the full scale spatial model. We establish consistency of the posterior distributions of the spatially varying coefficients and analyse the predictive efficiency of our models based upon the compressed data. Posterior contraction of varying coefficient (VC) models have been investigated by a few recent articles. For example, Guhaniyogi *et al.* (2020a) derive minimax-optimal posterior contraction rates for Bayesian VC models under GP priors when the number of predictors P is fixed. Deshpande *et al.* (2020) also derived near-optimal posterior contraction rates under BART priors, and Bai *et al.* (2019) showed asymptotically optimal rate of estimation for varying coefficients with a variable selection prior on varying coefficients. We address these questions in the context of data compression, which has largely remained unexplored.

3.1 Bayesian Compressed Spatially Varying Coefficient Models

We model each spatially varying coefficient $w_j(s)$ in (3.1) as

$$w_j(s) = \sum_{h=1}^{H} B_{jh}(s)\gamma_{jh} \ , \quad j = 1, ..., \tilde{P} \ , \tag{3.2}$$

where each $B_{jh}(s)$ is a basis function evaluated at location s for $h = 1, ..., H$, and γ_{jh}'s are the corresponding basis coefficients. The distribution of these γ_{jh}'s yields a multivariate process with $\text{cov}(w_i(s), w_j(s')) = B_i(s)^{\mathrm{T}}\text{cov}(\gamma_i, \gamma_j)B_j(s)$, where $B_i(s)$ and γ_i are $H \times 1$ with elements $B_{ih}(s)$ and γ_{ih}, respectively, for $h = 1, \ldots, H$.

Appropriate choices for basis functions can produce appropriate classes of multivariate spatial processes. A number of choices are available. For example, Biller and Fahrmeir (2001) and Huang *et al.* (2015) use splines to model the $B_{jh}(s)$'s and place Gaussian priors on the basis coefficients γ_{jh}. Li *et al.* (2015) propose a scale-mixture of multivariate normal distributions to shrink groups of basis coefficients towards zero. More recently, Bai *et al.* (2019) proposed using B-spline basis functions and multivariate spike-and-slab discrete mixture prior distributions on basis coefficients to aid functional variable selection. Other popular choices for basis functions include the wavelet basis (Vidakovic, 2009; Cressie and Wikle, 2015), radial basis (Bliznyuk *et al.*, 2008) and locally bi-square (Cressie and Johannesson, 2008) or elliptical basis functions (Lemos and Sansó, 2009). Alternatively, a basis representation of $w_j(s)$ can be constructed by envisioning $w_j(s)$ as the projection of a Gaussian process $w_j(s)$ onto a set of reference locations, or “knots”, which yields predictive processes and other variants (Banerjee *et al.*, 2008; Guhaniyogi *et al.*, 2013). More generally, each $w_j(s)$ can also be modelled using multi-resolution analogues to the aforesaid models to carefully capture global variations at the lower resolution and local variations at the higher resolutions (Katzfuss, 2017; Guhaniyogi and Sansó, 2018).

Let $\{y(s_i), x(s_i)\}$ be observations at N spatial locations $S = \{s_1, s_2, \ldots, s_N\}$. Using (3.2) in (3.1) yields the Gaussian linear mixed model

$$y = X\beta + \tilde{X}B\gamma + \epsilon\ , \quad \epsilon \sim N(0, \sigma^2 I_N)\ . \tag{3.3}$$

where $y = (y(s_1), y(s_2), \ldots, y(s_N))^{\mathrm{T}}$ and $\epsilon = (\epsilon(s_1), \epsilon(s_2), \ldots, \epsilon(s_N))^{\mathrm{T}}$ are $N \times 1$ vectors of responses and errors, respectively, X is $N \times P$ with n-th row $x(s_n)^{\mathrm{T}}$, $\tilde{X}$ is the $N \times N\tilde{P}$ block-diagonal matrix with (n,n)-th block $\tilde{x}(s_n)^{\mathrm{T}}$, $B = (B(s_1)^{\mathrm{T}}, \ldots, B(s_N)^{\mathrm{T}})^{\mathrm{T}}$ is $N\tilde{P} \times H\tilde{P}$ with $B(s_n)$ a block-diagonal $\tilde{P} \times H\tilde{P}$ matrix whose j-th diagonal block is $(B_{j1}(s_n), \ldots, B_{jH}(s_n))$. The coefficient $\gamma = (\gamma_1^{\mathrm{T}}, ..., \gamma_{\tilde{P}}^{\mathrm{T}})^{\mathrm{T}}$ is $H\tilde{P} \times 1$ with each $\gamma_j = (\gamma_{j1}, \ldots, \gamma_{jH})^{\mathrm{T}}$ being $H \times 1$. Bayesian methods for estimating (3.3) typically employ a multivariate normal prior (Biller and Fahrmeir, 2001; Huang *et al.*, 2015) or its scale-mixture (discrete as well as continuous) variants (Li *et al.*, 2015; Bai *et al.*, 2019) on γ.

Working with (3.3) will be expensive for large N. Instead, we consider data compression or sketching using a random linear mapping to reduce the size of the dataset from N to M observations. For this, we use M one-dimensional linear mappings of the data encoded by an $M \times N$ compression matrix Φ with $M << N$. This compression matrix is applied to y, X and $\tilde{X}$ to construct the $M \times 1$ compressed response vector $y_\Phi = \Phi y$ and the matrices $X_\Phi = \Phi X$ and $\tilde{X}_\Phi = \Phi \tilde{X}$. We will return to the specification of Φ, which, of course, will be crucial for relating the inference from the compressed data with the full model. For now assuming that we have fixed Φ, we construct a Bayesian hierarchical model for the compressed data

$$p(\psi, \beta, \gamma, \sigma^2 \,|\, y_\Phi, \Phi) \propto p(\psi, \sigma^2, \beta, \gamma) \times N(y_\Phi \,|\, X_\Phi \beta + \tilde{X}_\Phi B \gamma, \sigma^2 I_M) \;, \tag{3.4}$$

where ψ denotes additional parameters specifying the prior distributions on either γ or β. For example, a customary specification is

$$p(\psi, \sigma^2, \beta, \gamma) = \prod_{i=1}^{\tilde{P}} IG(\tau_i^2 \,|\, a_\tau, b_\tau) \times IG(\sigma^2 \,|\, a_\sigma, b_\sigma) \times N(\beta \,|\, \mu_\beta, V_\beta) \times N(\gamma \,|\, 0, \Delta_\psi) \;, \tag{3.5}$$

where $\psi = \{\tau_1^2, ..., \tau_{\tilde{P}}^2\}$ and Δ is $H\tilde{P} \times H\tilde{P}$ block-diagonal with j-th block given by $\tau_j^2 I_H$, for $j = 1, ..., \tilde{P}$. While (4.8) is a convenient choice for empirical investigations due to conjugate full conditional distributions, our method applies broadly to any basis function and any discrete or continuous mixture of Gaussian priors on the basis coefficients. In applications where the associations among the latent regression slopes is of importance, one could, for instance, adopt $p(\psi, \gamma) = IW(\psi \,|\, r, \Omega) \times N(\gamma \,|\, 0, \psi)$ with ψ as the $H\tilde{P} \times H\tilde{P}$ covariance matrix for γ. Our current focus is not, however, on such multivariate models, so we do not discuss them further except to note that (3.4) accommodates such extensions.

The likelihood in (3.4) is different from that by applying Φ to (3.3) because the error distribution in (3.4) is retained as the usual noise distribution without any effect of Φ. Hence, the model in (3.4) is a model analogous to (3.3) but applied to the *new* compressed data set $\{y_\Phi, X_\Phi, \tilde{X}_\Phi\}$. Working with a Φ-transformed model (3.3), where the distribution of the noise will be transformed according $\Phi\epsilon$, will not deliver the computational benefits, and is somewhat detrimental to the cause of data confidentiality (as in that case, the analyst need to know Φ) that are provided by (3.4).

For specifying Φ we pursue the idea of data oblivious Gaussian sketching (Sarlos, 2006), where we draw the elements of $\Phi = (\Phi_{ij})$ independently from $\mathrm{N}(0, 1/N)$ and fix them. The dominant computational operations for obtaining the sketched data using Gaussian sketches is $O(MN^2\tilde{P})$. While alternative computationally efficient data oblivious options such as the Hadamard sketch (Ailon and Chazelle, 2009) and the Clarkson-Woodruff sketch (Clarkson and Woodruff, 2017) are available for Φ, it is less pertinent in Bayesian settings since computation time of (3.4) far exceeds that for the sketching matrix. The compressed data serves as a surrogate for the Bayesian regression analysis with spatially varying coefficients. Since the number of compressed records is much smaller than the number of records in the uncompressed data matrix, spatial model fitting becomes computationally efficient and economical in terms of storage as well as the number of floating point operations (flops). Importantly, original data are not recoverable from the compressed data, and the compressed data effectively reveal no more information than would be revealed by a completely new sample (Zhou *et al.*, 2008). In fact, the original uncompressed data does not need to be stored or accessed at any stage in the course of the analysis.

3.1.1 Efficient Posterior Computation & Approximate Predictive Inference

In what follows, we discuss efficient computation offered by the data sketching framework. With prior distributions on parameters specified as in (3.5), posterior computation requires drawing Markov chain Monte Carlo (MCMC) samples sequentially from the full conditional posterior distributions of $\gamma|-$, $\beta|-$, $\sigma^2|-$ and $\tau_j^2|-$, $j = 1, \dots, \tilde{P}$. To this end, $\sigma^2|- \sim IG(a_\sigma + M/2, b_\sigma + ||y_\Phi - X_\Phi\beta - \tilde{X}_\Phi B\gamma||^2/2)$, $\beta|- \sim N\left((X_\Phi^{\mathrm{T}} X_\Phi/\sigma^2 + I)^{-1} X_\Phi^{\mathrm{T}}(y_\Phi - \tilde{X}_\Phi B\gamma)/\sigma^2, (X_\Phi^{\mathrm{T}} X_\Phi/\sigma^2 + I)^{-1}\right)$ and $\tau_j^2|- \sim IG(a_\tau + H/2, b_\tau + ||\gamma_j||^2/2)$ do not present any computational obstacles. The main computational bottleneck lies with $\gamma|-$,

$$N\left(\left(\frac{B^{\mathrm{T}}\tilde{X}_\Phi^{\mathrm{T}}\tilde{X}_\Phi B}{\sigma^2} + \Delta^{-1}\right)^{-1} B^{\mathrm{T}}\tilde{X}_\Phi^{\mathrm{T}}\frac{(y_\Phi - X_\Phi\beta)}{\sigma^2}, (B^{\mathrm{T}}\tilde{X}_\Phi^{\mathrm{T}}\tilde{X}_\Phi B/\sigma^2 + \Delta^{-1})^{-1}\right). \tag{3.6}$$

Efficient sampling of γ relies upon the Cholesky decomposition of the matrix $\left(B^{\mathrm{T}}\tilde{X}_\Phi^{\mathrm{T}}\tilde{X}_\Phi B/\sigma^2 + \Delta^{-1}\right)$ and solves triangular linear systems to draw a sample from (3.6). While numerically robust for small to moderately large H, computing and storing the Cholesky factor of this matrix involves $O((H\tilde{P})^3)$ and $O((H\tilde{P})^2)$ floating point operations, respectively (Golub and Van Loan, 2012). This results in computational and memory bottlenecks for a large number of basis functions, which may be required to estimate the spatial surface with sufficient local variation.

To achieve computational efficiency, we adapt a recent algorithm proposed in Bhattacharya *et al.* (2016) (in the context of ordinary linear regression with uncompressed data and small sample size) to our setting: (i) draw $\tilde{\gamma}_1 \sim N(0, \Delta)$ and $\tilde{\gamma}_2 \sim N(0, I_M)$; (ii) set $\tilde{\gamma}_3 = \tilde{X}_\Phi B \tilde{\gamma}_1/\sigma + \tilde{\gamma}_2$; (iii) solve $(\tilde{X}_\Phi B \Delta B^{\mathrm{T}} \tilde{X}_\Phi^{\mathrm{T}}/\sigma^2 + I_M)\tilde{\gamma}_4 = ((y_\Phi - X_\Phi \beta)/\sigma - \tilde{\gamma}_3)$; and (iv) set $\tilde{\gamma}_5 = \tilde{\gamma}_1 + \Delta B^{\mathrm{T}} \tilde{X}_\Phi^{\mathrm{T}} \tilde{\gamma}_4/\sigma$. The resulting $\tilde{\gamma}_5$ is a draw from the full conditional posterior distribution of γ. The computation is dominated by step (iii), which comprises $O(M^3 + M^2 H \tilde{P})$. Finally, note that when basis functions involve parameters, they are updated using Metropolis-Hastings steps since no closed form full conditionals are generally available for them.

Predictive inference on $y(s_0)$ will proceed from the posterior predictive distribution

$$\mathbb{E}[p(y(s_0) \mid y_\Phi, \beta, \gamma, \sigma^2)] = \int p(y(s_0) \mid y_\Phi, \beta, \gamma, \sigma^2) p(\beta, \gamma, \sigma^2 \mid y_\Phi, \Phi) d\beta d\gamma d\sigma^2 \ , \tag{3.7}$$

where $\mathbb{E}[\cdot]$ is the expectation with respect to the posterior distribution in (3.4). This is easily achieved by drawing $y(s_0)^{(l)} \sim N(\sum_{p=1}^{P} x_p(s_0)\beta_p^{(l)} + \sum_{j=1}^{\tilde{P}} \tilde{x}_j(s_0) w_j(s_0)^{(l)}, \sigma^{2(l)})$ for each posterior sample $\{\beta^{(l)}, \gamma^{(l)}, \sigma^{2(l)}\}$ drawn from (3.4), where $w_j(s_0)^{(l)}$ is obtained from $\gamma^{(l)}$ using (3.2) and $l = 1, 2, \ldots, L$ indexes the L (post-convergence) posterior samples. The next section offers theoretical results related to the large sample consistency of the posterior distribution from the compressed varying coefficients model (3.4) and the posterior predictive distribution in (3.7) with respect to the probability law for the uncompressed oracle model in (3.1).

3.2 Posterior contraction from data sketching

3.2.1 Definitions and Notations

This section proves the posterior contraction properties of varying coefficients under the proposed framework. In what follows, we add a subscript N to the compressed response vector $y_{\Phi,N}$, compressed predictor matrix $\tilde{X}_{\Phi,N}$, dimension of the compression matrix M_N and the number of basis functions H_N to indicate that all of them increase with the sample size N. Naturally, the dimension of the basis coefficient vector γ and the compression matrix Φ are also functions of N, though we keep this dependence implicit. Since we do not assume a functional variable selection framework, we keep P fixed throughout, and not a function of N. We assume that $s_1, ..., s_N$ follow i.i.d. distribution G on $\mathcal{D}$ with G having a Lebesgue density g, which is bounded away from zero and infinity uniformly over $\mathcal{D}$. The true regression function is also given by (3.1), with the true varying coefficients $w_1^*(s), ..., w_{\tilde{P}}^*(s)$ belonging to the class of functions

$$\mathcal{F}_\xi(\mathcal{D}) = \{f : f \in L_2(\mathcal{D}) \cap \mathcal{C}^\xi(\mathcal{D}), E_{\mathcal{S}}[|f|] < \infty\}, \tag{3.8}$$

where $L_2(\mathcal{D})$ is the set of all square integrable functions on $\mathcal{D}$, $\mathcal{C}^\xi(\mathcal{D})$ is the class of at least ξ-times continuously differentiable functions in $\mathcal{D}$ and $E_{\mathcal{S}}$ denotes the expectation under the density of g. The probability and expectation under the true data generating model are denoted by P^* and E^*, respectively. For algebraic simplicity, we make a few simplifying assumptions in the model. To be more specific, we assume that $\beta = 0$ and $\sigma^2 = \sigma^{*2}$ is known and fixed at 1. The first assumption is mild since P does not vary with N and we do not consider variable selection. The second assumption is also customary in asymptotic studies (Vaart and Zanten, 2011). Furthermore, the theoretical results obtained by assuming σ^2

as a fixed value is equivalent to those obtained by assigning a prior with a bounded support on σ^2 (Van der Vaart *et al.*, 2009).

For a vector $v = (v_1, ..., v_N)^{\mathrm{T}}$, we let $|| \cdot ||_1, || \cdot ||_2$ and $|| \cdot ||_\infty$ denote the L_1, L_2 and L_∞ norms, respectively, defined as $||v||_2 = (\sum_{n=1}^{N} v_n^2)^{1/2}$, $||v||_1 = \sum_{n=1}^{N} |v_n|$ and $||v||_\infty = \max_{n=1,..,N} |v_n|$, respectively. The number of nonzero elements in a vector is given by $|| \cdot ||_0$. In the case of a square integrable function $f(s)$ on $\mathcal{D}$, we denote the integrated L_2-norm of f by $||f||_2 = \left(\int_{\mathcal{D}} f(s)^2 g(s) ds\right)^{1/2}$ and the sup-norm of f by $||f||_\infty = \sup_{s \in \mathcal{D}} |f(s)|$. Thus $|| \cdot ||_\infty$ and $|| \cdot ||_2$ are used both for vectors and functions, and they should be interpreted based on the context. Finally, $e_{\min}(A)$ and $e_{\max}(A)$, respectively, represent the minimum and maximum eigenvalues of the square matrix A. The Frobenius norm of the matrix A is given by $||A||_F = \sqrt{\mathrm{tr}(A^{\mathrm{T}} A)}$. For two nonnegative sequences $\{a_N\}$ and $\{b_N\}$, we write $a_N \asymp b_N$ to denote $0 < \liminf_{N \to \infty} a_N / b_N \leq \limsup_{N \to \infty} a_N / b_N < \infty$. If $\lim_{N \to \infty} a_N / b_N = 0$, we write $a_N = o(b_N)$ or $a_N \prec b_N$. We use $a_N \lesssim b_N$ or $a_N = O(b_N)$ to denote that for sufficiently large N, there exists a constant $C > 0$ independent of N such that $a_N \leq C b_N$.

3.2.2 Assumption, Framework and Main Results

For simplicity, we assume $\Delta = I$ and that the random covariates $x_p(s)$, $p = 1, ..., P$ follow distributions which are independent of the distribution of the idiosyncratic error ϵ. We now state the following assumptions on the basis functions, H_N, M_N, covariates and the sketching or compression matrix.

(A) For any $w_j^*(s) \in \mathcal{F}_\xi(\mathcal{D})$, there exists γ_j^* such that $||w_j^* - B_j^{\mathrm{T}} \gamma_j^*||_\infty = \sup_{s \in \mathcal{D}} |w_j^*(s) - \sum_{h=1}^{H_N} B_{jh}(s) \gamma_{jh}^*| = O(H_N^{-\xi})$, for $j = 1, ..., \tilde{P}$, and $||\gamma^*||_2^2 \prec M_N^{1/(1+\xi)}$.

(B) N, M_N, H_N satisfy $M_N = o(N)$ and $H_N \asymp M_N^{1/(2\xi+2)}$.

(C) $||\Phi\Phi^{\mathrm{T}} - I_{M_N}||_F \leq C'\sqrt{M_N/N}$, for some constant $C' > 0$, for all large N.

(D) The random covariate $x_p(s)$ are uniformly bounded for all $s \in \mathcal{D}$, and w.l.g., $|x_p(s)| \leq 1$, for all $p = 1, ..., P$ and for all $s \in \mathcal{D}$.

(E) There exists a sequence κ_N such that $||\tilde{X}_{\Phi,N}\alpha||^2 \asymp \kappa_N||\tilde{X}_N\alpha||^2$, such that $1 \prec N\kappa_N \prec M_N$ for any vector $\alpha \in \mathbb{R}^{N\tilde{P}}$.

Assumption (A) holds for orthogonal Legendre polynomials, Fourier series, B-splines and wavelets (Shen and Ghosal, 2015). Assumption (B) provides an upper bound on the growth of M_N and H_N as a function of N. Assumption (C) is a mild assumption based on the theory of random matrices and occurs with probability at least $1 - e^{-C''M_N}$ when Φ is constructed using the Gaussian sketching for a constant $C'' > 0$ (see Lemma 5.36 and Remark 5.40 of Vershynin (2010)). Assumption (D) is a technical condition customarily used in functional regression analysis (Bai *et al.*, 2019). Finally, Assumption (E) characterises the class of feasible compression matrices, roughly explaining how the linear structure of the columns of the original predictor matrix is related to that of the compressed predictor matrix. Such an assumption is reasonable for the set of random compression matrices for a sequence κ_N depending on N, M_N and $\tilde{P}$ (Ahfock *et al.*, 2017).

Let $w(s) = (w_1(s), ..., w_{\tilde{P}}(s))^{\mathrm{T}}$ and $w^*(s) = (w_1^*(s), ..., w_{\tilde{P}}^*(s))^{\mathrm{T}}$ be the $\tilde{P}$-dimensional fitted and true varying coefficients. Let $||w - w^*||_2 = \sum_{j=1}^{\tilde{P}} ||w_j - w_j^*||_2$ denote the sum of integrated L_2 distances between the true and the fitted varying coefficients. Define the set $\mathcal{C}_N = \left\{w : ||w - w^*||_2 > \tilde{C}\theta_N\right\}$, for some constant $\tilde{C}$ and some sequence $\theta_N \to 0$ and $M_N\theta_N^2 \to \infty$. Further suppose $\pi_N(\cdot)$ and $\Pi_N(\cdot)$ are the prior and posterior densities of w with N observations, respectively. From equation (3.2), the prior distribution on w is governed by the prior distribution on γ, so that the posterior probability of $\mathcal{C}_N$ can be expressed as,

$$\Pi_N(\mathcal{C}_N|y_{\Phi,N},\tilde{X}_{\Phi,N}) = \frac{\int_{\mathcal{C}_N} f(y_{\Phi,N}|\tilde{X}_{\Phi,N},\gamma)\pi_N(\gamma)}{\int f(y_{\Phi,N}|\tilde{X}_{\Phi,N},\gamma)\pi_N(\gamma)},$$

where $f(y_{\Phi,N}|\tilde{X}_{\Phi,N},\gamma)$ is the joint density of $y_{\Phi,N}$ under model (3.4). We begin with the following important result from the random matrix theory.

Lemma 3.2.1. *Consider the $M_N \times N$ compression matrix Φ with each entry being drawn independently from $N(0,1/N)$. Then, almost surely*

$$(\sqrt{N} - \sqrt{M_N} - o(\sqrt{N}))^2/N \le e_{\min}(\Phi\Phi^{\mathrm{T}}) \le e_{\max}(\Phi\Phi^{\mathrm{T}}) \le (\sqrt{N} + \sqrt{M_N} + o(\sqrt{N}))^2/N, \tag{3.9}$$

when both $M_N, N \to \infty$.

Proof. This is a consequence of Theorem 5.31 and Corollary 5.35 of Vershynin (2010).

The inequalities in (3.9) is used to derive the following two results, which we present as Lemma 3.2.2 and 3.2.3.

Lemma 3.2.2. *Let P^* denote the true probability distribution of y_N and $f^*(y_{\Phi,N}|\gamma^*)$ denotes the density of $y_{\Phi,N}$ (omitting explicit dependence on $\tilde{X}_{\Phi,N}$) under the true data generating model. Define*

$$\mathcal{A}_N = \left\{ y : \int \{f(y_{\Phi,N}|\gamma)/f^*(y_{\Phi,N}|\gamma^*)\}\, \pi_N(\gamma)d\gamma \le \exp(-CM_N\theta_N^2) \right\}. \tag{3.10}$$

Then $P^*(\mathcal{A}_N) \to 0$ as $M_N, N \to \infty$ for any constant $C > 0$.

Proof. See Appendix A.

Lemma 3.2.3. *Let γ^* be any fixed vector in the support of γ and let $\mathcal{B}_N = \{\gamma : ||\gamma - \gamma^*||_2 \leq C_{2w}\theta_N H_N^{1/2}\}$ for some constant $C_{2w} > 0$. Then there exists a sequence ζ_N of random variables depending on $\{y_{\Phi,N}, X_{\Phi,N}\}$ and taking values in $(0,1)$ such that*

$$\mathbb{E}^*(\zeta_N) \lesssim \exp(-M_N\theta_N^2) \text{ and } \sup_{\gamma \in \mathcal{B}_N^c} \mathbb{E}_\gamma(1-\zeta_N) \lesssim \exp(-M_N\theta_N^2), \tag{3.11}$$

where $\mathbb{E}_\gamma$ and $\mathbb{E}^*$ denote the expectations under the distributions $f(\cdot \mid \gamma)$ and $f^*(\cdot \mid \gamma^*)$, respectively.

Proof. See Appendix A.

We use the above results to establish the posterior contraction result for the proposed model.

Theorem 3.2.4. *Under Assumptions (A)-(E), our proposed model (3.4) satisfies $\max_{j=1,\dots,\tilde{P}} \sup_{w_j^* \in \mathcal{F}_\xi(\mathcal{D})} \mathbb{E}^*\Pi_N(\mathcal{C}_N \mid y_{\Phi,N}, \tilde{X}_{\Phi,N}) \to 0$, as $N, M_N \to \infty$ and with the posterior contraction rate $\theta_N \asymp M_N^{-\xi/(2\xi+2)}$.*

Proof. See Appendix A.

Since $\theta_N \to 0$ as $N \to \infty$, the model consistently estimates the true varying coefficients under the integrated L_2-norm. Further, data compression decreases the effective sample size from N to M_N, hence, the contraction rate θ_N obtained in Theorem 3.2.4 is optimal and adaptive to the smoothness of the true varying coefficients. Our next theorem justifies the two-stage prediction strategy described in Section 3.1.1.

Theorem 3.2.5. *For any location s_0 drawn randomly with the density g and corresponding predictors $\tilde{x}_1(s_0), \dots, \tilde{x}_{\tilde{P}}(s_0)$, let f_u be the predictive density $p(y(s_0) \mid \tilde{x}_1(s_0), \dots, \tilde{x}_{\tilde{P}}(s_0), w(s_0))$ derived from (3.1) without data compression. Let f^* be the* *true data generating model (i.e., (3.1) with $w(s_0)$ fixed at $w^*(s_0)$). Given s_0 and $\tilde{x}_1(s_0), \ldots, \tilde{x}_{\tilde{P}}(s_0)$, define $h(f_u, f^*) = \int(\sqrt{f_u} - \sqrt{f^*})^2$ as the Hellinger distance between the densities f_u and f^*. Then*

$$\mathbb{E}^*\mathbb{E}\mathbb{E}_{\mathcal{S}}[h(f_u, f^*) \mid \tilde{X}_{\Phi,N}, y_{\Phi,N}] \to 0, \text{ as } N, M_N \to \infty, \tag{3.12}$$

where $\mathbb{E}_{\mathcal{S}}$, $\mathbb{E}$ and $\mathbb{E}^*$ stand for expectations with respect to the density g, the posterior density $\Pi_N(\cdot|\tilde{X}_{\Phi,N}, y_{\Phi,N})$ and the true data generating distribution, respectively.

Proof. See Appendix A.

The theorem states that the predictive density of the VCM model in (3.1) is arbitrarily close to the true predictive density even when we plug-in inference on parameters from (3.4).

3.3 Simulation Results

3.3.1 Inferential performance

We empirically validate our proposed approach using (3.4), henceforth abbreviated as *geoS*, by comparing its inferential performance and computational efficiency with the uncompressed model (3.3) on some simulated data. We simulate data by using a fixed set of spatial locations $s_1, \ldots, s_N$ that were drawn uniformly over the domain $\mathcal{D} = [0,1] \times [0,1]$. We set $\tilde{P} = P = 3$ and assume $\beta = 0$, i.e., all predictors have purely space-varying coefficients. We set $\tilde{x}_1(s_i) = 1$, for all $i = 1, \ldots, N$, while the values of $\tilde{x}_j(s_1), \ldots, \tilde{x}_j(s_N)$ for $j = 2, 3$ were set to independently values from $N(0,1)$. For each $n = 1, \ldots, N$, the response $y(s_n)$ is drawn independently from $N(w_1^*(s_n) + w_2^*(s_n)\tilde{x}_2(s_n) + w_3^*(s_n)\tilde{x}_3(s_n), \sigma^{*2})$ following

(3.3), where σ^{*2} is set to be 0.1. The true space-varying coefficients ($w_j^*(s)$s) are simulated from a Gaussian process with mean 0 and covariance kernel $C(\cdot, \cdot; \theta_j)$, i.e., $(w_j^*(s_1), ..., w_j^*(s_N))^{\mathrm{T}}$ is drawn from $N(0, C^*(\theta_j))$, for each $j = 1, \ldots, \tilde{P}$, where $C^*(\theta_j)$ is an $N \times N$ matrix with the (n, n')th element $C(s_n, s_{n'}; \theta_j)$. We set the covariance kernel $C(\cdot, \cdot; \theta_j)$ to be the exponential covariance function given by

$$C(s, s'; \theta_j) = \delta_j^2 \exp\left\{-\frac{1}{2}\left(\frac{||s - s'||}{\phi_j}\right)\right\}, \;\; j = 1, 2, 3, \tag{3.13}$$

with the true values of $\delta_1^2, \delta_2^2, \delta_3^2$ set to $1, 0.8, 1.1$, respectively. We fix the true values of ϕ_1, ϕ_2, ϕ_3 at $1, 1.25, 2$, respectively.

While fitting *geoS* and its uncompressed analogue (3.3), the varying coefficients are modelled through the linear combination of H basis functions as in (3.2), where these basis functions are chosen as the tensor-product of B-spline bases of order $q = 4$ (Shen and Ghosal, 2015). More specifically, for $s = (s^{(1)}, s^{(2)})$, the j-th varying coefficient is modelled as

$$w_j(s) = \sum_{h_1=1}^{H_1} \sum_{h_2=1}^{H_2} B_{jh_1}^{(1)}(s^{(1)}) B_{jh_2}^{(2)}(s^{(2)}) \gamma_{jh_1h_2} \,, \tag{3.14}$$

where the marginal B-splines $B_{jh_1}^{(1)}$, $B_{jh_2}^{(2)}$ are defined on sets of H_1 and H_2 knots, respectively. The knots are chosen to be equally-spaced so the entire set of $H = H_1H_2$ knots is uniformly spaced over the domain $\mathcal{D}$. We complete the hierarchical specification by assigning independent $IG(2, 0.1)$ priors (mean 0.1 with infinite variance) for σ^2 and τ_j^2 for each $j = 1, \ldots, P$.

We implemented our models in the `R` statistical computing environment on a Dell XPS 13 PC with Intel Core i7-8550U CPU @ 4.00GHz processors at 16 GB of RAM. For each of our simulation datasets we ran a single-threaded MCMC chain for 5000 iterations. Posterior inference was based upon 2000 samples retained after adequate convergence was diagnosed using Monte Carlo standard errors and effective sample sizes (ESS) using the `mcmcse` package in `R`. All source codes for these experiments are available from `https://github.com/LauraBaracaldo/Bayesian-Data-Sketching-in-Spatial-Regression-Models`.

Table 3.1 summarises the estimates of varying coefficients and the predictive performance for *geoS* in comparison to the uncompressed model. We applied these models to data generated with $N = 5000$ (case 1) and $N = 10000$ (case 2). For both cases the compressed dimension is taken to be $M \approx 10\sqrt{N}$ which seems to be effective from empirical considerations in our simulations. We provide further empirical justification for this choice in Section 3.3.2. Our *geoS* approach compresses the sample sizes to $M = 700$ and $M = 1000$ in cases 1 and 2, respectively. The number of fitted basis functions in cases 1 & 2 are $H = 225, 256$, respectively.

Figures 3.1 and 3.2 present the estimated varying coefficients by *geoS* and the uncompressed data model for cases 1 and 2, respectively. These figures reveal similar point estimation offered by *geoS* and the uncompressed model. The mean squared error of estimating varying coefficients, defined as $\sum_{j=1}^{3}\sum_{n=1}^{N}(\widehat{w}_j(s_n) - w_j^*(s_n))^2/(3N)$ (where $\widehat{w}_j(s_n)$ is the posterior median of $w_j(s_n)$), also confirms very similar point estimates offered by the compressed and uncompressed models (see Table 3.1). Further, *geoS* offers close to nominal coverage for 95% credible intervals for varying coefficients, with little wider credible intervals compared to uncompressed data model. This can be explained by the smaller sample size for the *geoS* model, though the difference turns out to be minimal. We also carry out predictive inference using *geoS* (Section 3.1.1). Table 3.1 presents mean squared predictive error (MSPE), average length and coverage for the 95% predictive intervals, based on $N^* = 500$ out of the sample observations. We find *geoS* delivers posterior predictive estimates and predictive coverage that are very consistent with the uncompressed model, perhaps with marginally wider predictive intervals than those without compression.

Finally, the computational efficiency of both models are computed based on the metric $\log_2(ESS/\text{Computation Time})$, where ESS denotes the effective sample size averaged over the MCMC samples of all parameters. We find *geoS* is almost 270% and 223% more efficient than the uncompressed model for $N = 5,000$ and $N = 10,000$, respectively, while delivering almost indistinguishable substantive inference on the spatial effects.

	$N = 5000$, $H = 225$		$N = 10000$, $H = 256$	
	(geoS) $M = 700$	*Uncompressed*	*(geoS)* $M = 1000$	*Uncompressed*
MSE (SVC)	0.0474	0.0168	0.0429	0.0178
95% CI length	0.8368	0.6182	0.7222	0.5531
95% CI Coverage	0.9448	0.9322	0.9153	0.9026
MSPE	0.2574	0.1833	0.2283	0.1605
95% PI length	1.9717	1.5168	1.8613	1.5148
95% PI coverage	0.936	0.925	0.954	0.930
Computation efficiency	2.2050	0.8079	0.9755	0.4356

Table 3.1: Results for simulation cases 1 & 2 for the compressed *geoS* and uncompressed models. Mean Squared Error (MSE), length and coverage of 95% CI for the spatially varying coefficients. We also present mean squared prediction error (MSPE), coverage and length of 95% predictive intervals for the competing models. Computation efficiency for the *geoS* with the uncompressed data model is also recorded.

3.3.2 Choice of the dimension of the compression matrix

We present investigations into the choice of the appropriate compression matrix size M. For simulated data with sample size $N = 10000$, we ran our model for different values of $M = k\sqrt{N}$, $k = 1, \ldots, 20$. Figure 3.3 shows the variations in point-wise and interval prediction reflected in the $MSPE$ and 95% predicted interval coverage and length, respectively. Unsurprisingly, as M increases the MSPE drops with a diminished rate of decline until the $k \sim 10$. In terms of in-

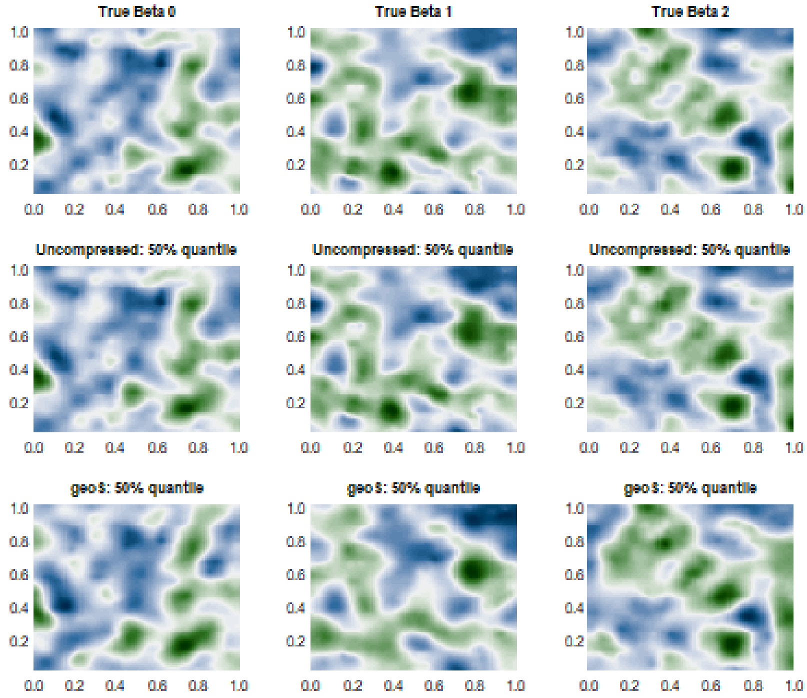


Figure 3.1: Simulation case 1: $(N, H) = (5000, 225)$. Two-dimensional true and predicted surfaces over the unit square $\mathcal{D} = [0,1] \times [0,1]$. First row corresponds to the surfaces of true space-varying coefficients $\beta_p^*(s)$, $p = 1, 2, 3$. Rows 2 and 3 correspond to the predicted 50% quantile surfaces for the uncompressed and compressed *geoS* models respectively.

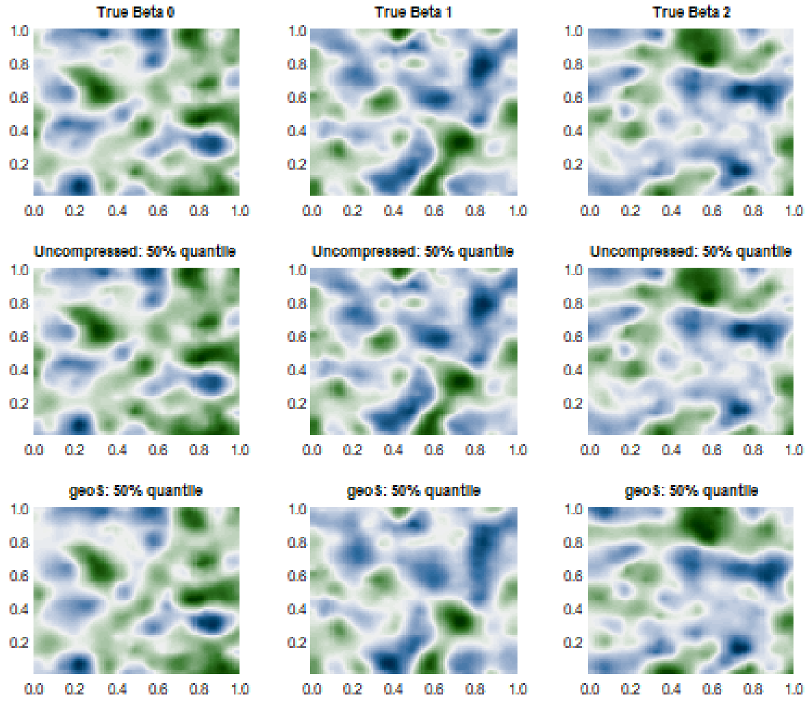
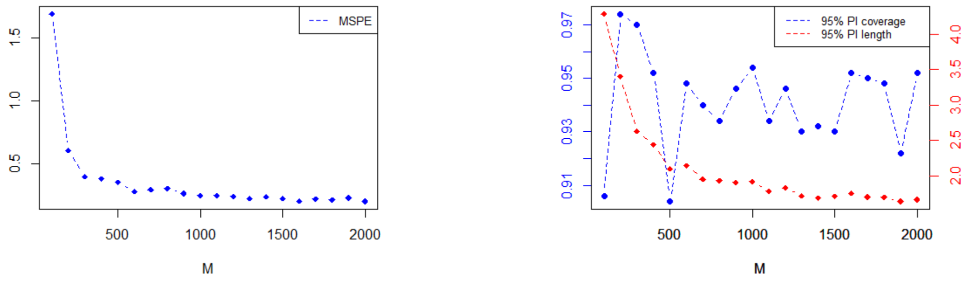


Figure 3.2: Simulation case 2: $(N, H) = (10000, 256)$. Two-dimensional true and predicted surfaces over the unit square $\mathcal{D} = [0, 1] \times [0, 1]$. First row corresponds to the surfaces of true space-varying coefficients $\beta_p^*(s)$, $p = 1, 2, 3$. Rows 2 and 3 correspond to the predicted 50% quantile surfaces for the uncompressed and compressed *geoS* models respectively.

terval prediction, predictive coverage seems to oscillate within the narrow interval $(0.9, 0.97)$ for all values of M, but the length of the predictive interval improves as M increases and starts to stabilise at around $k \sim 10$. We observe that the choice of $M \sim 10\sqrt{N}$ leads to good performance across various simulations and real data analysis.



(a) **(b)**

Figure 3.3: (a) MSPE, (b) 95% predictive interval coverage and length for different choices of M

3.4 Vegetation Data Analysis

We implement *geoS* to analyse vegetation data gathered through the Moderate Resolution Imaging Spectroradiometer (MODIS), which resides aboard the Terra and Aqua platforms on NASA spacecrafts. MODIS vegetation indices, produced on 16-day intervals and at multiple spatial resolutions, provide consistent information on the spatial distribution of vegetation canopy greenness, a composite property of leaf area, chlorophyll and canopy structure. The variable of interest will be the Normalised Difference Vegetation Index (NDVI), which quantifies the relative vegetation density for each pixel in a satellite image, by measuring the difference between the reflection in the near-infrared spectrum (NIR) and the red light reflection (RED): $NDVI = \frac{NIR-RED}{NIR+RED}$. High NDVI values, ranging between 0.6 and 0.9 indicate high density of green leaves and healthy vegetation, whereas low values, 0.1 or below, correspond to low or absence of vegetation as in the case of urbanised areas. When analysed over different locations, NDVI can reveal changes in vegetation due to human activities such as deforestation and natural phenomena such as wild fires and floods.

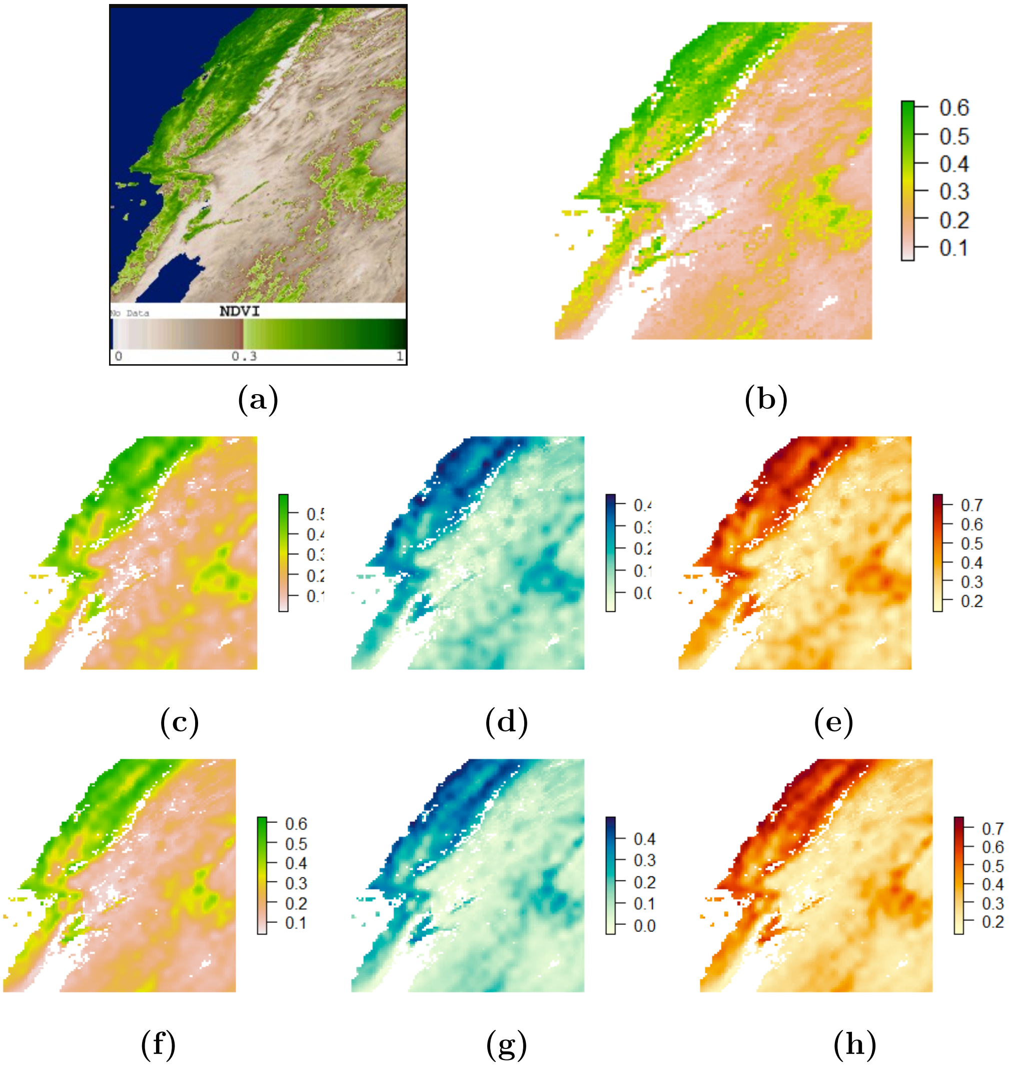
Our analysis will be focused on geographical data that was mapped on a sinusoidal (SIN) projected grid, located on the western coast of the United States, more precisely zone *h08v05*, between $30°N$ to $40°N$ latitude and $104°W$ to $130°W$ longitude (see Figure 3.4(a)). The data set, which was downloaded using the `R` package MODIS, comprises $133,000$ observed locations where the response was measured through the MODIS tool over a 16-day period in April, 2016. We retained $N = 113,000$ observations (randomly chosen) for model fitting and held out the rest for prediction. In order to fit (3.1), we set $y(s_n)$ to be the transformed NDVI ($\log(NDVI) + 1$), $P = \tilde{P} = 2$ and consider the $P \times 1$ vector of predictors that includes an intercept and a binary index of urban area, both with fixed effects and spatially varying coefficients, i.e., $x(s_n) = \tilde{x}(s_n) = (1,\ x_2(s_n))^{\mathrm{T}}$, with $x_2(s_n) = \mathbb{1}_U(s_n)$, where U denotes an urban area.

As in Section 3.3, we fit *geoS* with $M \sim 10\sqrt{N} = 2300$ and its uncompressed counterpart (4.3), by modelling the varying coefficients through a linear combination of basis functions constructed using the tensor-product of B-splines of order $q = 4$ as in (4.12). We set the number of knots $H = H_1 H_2 = 39^2 = 1521$ to be uniformly distributed over the domain $\mathcal{D}$, which results in $HP = 3042$ basis coefficients γ_{jh} that are estimated. Specification of priors are identical to the simulation studies for σ^2, and τ_j^2, $j = 1, ..., P$; for β_j, $j = 1, ..., P$ we set a flat prior.

We ran an MCMC chain for 5000 iterations and retained 2000 samples for posterior inference after adequate convergence was diagnosed. The posterior mean of β_1 and β_2, along with their estimated 95% credible intervals corresponding to *geoS* and the uncompressed model are presented in Table 3.2. Additionally, Table 3.2 offers predictive inference from both competitors based on $N^* = 20,000$ test observations. According to both models there is a global pattern of relatively low vegetation density for areas with positive urban index as the estimated slope coefficient β_2 is negative in the compressed *geoS* and in the uncompressed models. In terms of point prediction and quantification of predictive uncertainty, the two competitors offer practically indistinguishable results, as revealed by Table 3.2. Further, Figure 3.4 shows that the 2.5%, 50% and 97.5% quantiles for the posterior predictive distribution are almost identical for the two competitors across the spatial domain, with the exception of neighbourhoods around locations having lower NDVI values. Notably, *geoS* offers nominal coverage for 95% prediction intervals, even with a significant reduction in the sample size from $N = 113,000$ to $M = 2300$. Data sketching to such a scale considerably reduces the computation time, leading to a much higher computation efficiency of *geoS* in comparison with its uncompressed analogue.

	(geoS) $M = 2300$	*Uncompressed*
β_1	0.222 (0.212, 0.230)	0.229 (0.219, 0.237)
β_2	-0.060 (-0.074, -0.047)	-0.071 (-0.082, -0.059)
$MSPE$	0.00327	0.00276
95% PI length	0.23445	0.22136
95% PI coverage	0.95250	0.95411
Computation efficiency	3.5424	0.46901

Table 3.2: Median and 95% credible interval of β_1, β_2 for geoS and its uncompressed analogue are presented for the Vegetation data analysis. We also present MSPE, coverage and length of 95% predictive intervals for the competing models. Computational efficiency for the two competing models are also provided.



(a) **(b)**

(c) **(d)** **(e)**

(f) **(g)** **(h)**

Figure 3.4: Coloured NDVI images of western United States (zone h08v05). (a) Satellite image: MODIS/Terra Vegetation Indices 16-Day L3 Global 1 km SIN Grid - 2016.04.06 to 2016.04.21; (b) True NDVI surface (raw data). Figures (c), (d) & (e) present NVDI predicted 50%, 2.5% and 97.5% quantiles for the *geoS* model. Figures (f), (g) & (h) present NVDI Predicted 50%, 2.5% and 97.5% quantiles for the uncompressed model.

3.5 Summary

We have developed Bayesian sketching for spatially oriented data using spatial regression models. The method achieves dimension reduction by compressing the data using a random linear transformation. The approach is different to the prevalent methods for large spatial data in that no new models or algorithms need to be developed since those available for existing spatially varying regression models can be directly applied to the compressed data. We establish attractive concentration properties of the posterior and posterior predictive distributions and empirically demonstrate the effectiveness of this method for analysing large spatial data sets. Access to the values of the response and predictors in the full data are not required at stage of inference, which preserves data confidentiality should that be of concern in the application.

Chapter 4

Distributed Bayesian Inference with Sketched Data

This chapter extends the data sketching idea proposed in Chapter 3 to develop a distributed inferential framework using Bayesian data sketching for simultaneous variable selection and estimation of varying coefficients in a spatially varying coefficient model with large spatial data. As discussed in Chapter 3, the core to data sketching is the usage of random compression matrices, which are employed to compute random linear transformations of the original data to produce a much smaller number of transformed data samples. Model fitting and inference is performed with the transformed data. While data sketching has an extensive literature in machine learning (please see the introduction of Chapter 3), to the best of our knowledge, this chapter is the first to study data sketching for efficient Bayesian inference in functional estimation and variable selection for varying coefficient models with large spatial data.

For this chapter, we consider (3.1) to describe relationship between response and predictors with all predictors are assumed to have spatially varying coefficients ($\tilde{P} = P$). This assumption is consistent with the recent literature on spatial variable selection (Bai *et al.*, 2019). Thus the model can be written as

$$y(s) = \sum_{j=1}^{P} x_j(s)\beta_j + \sum_{j=1}^{P} x_j(s)w_j(s) + \epsilon(s) = x(s)^{\mathrm{T}}\beta + x(s)^{\mathrm{T}}w(s) + \epsilon(s) \ , \quad (4.1)$$

where $y(s) \in \mathcal{Y} \subseteq \mathbb{R}$ ($s \in \mathcal{D} \subseteq \mathbb{R}^2$) is the spatially varying response function and $x_1(s), ..., x_P(s) \in \mathcal{X} \subseteq \mathbb{R}$ are possibly spatially varying predictors, with the $P \times 1$ vector $\beta = (\beta_1, ..., \beta_P)^{\mathrm{T}}$ representing the vector of spatially static coefficients corresponding to the P predictors, $w(s) = (w_1(s), w_2(s), \ldots, w_P(s))^{\mathrm{T}}$ corresponds to the $P \times 1$ vector of spatially varying regression coefficients which capture nonlinear dependence of the response function on the covariates, and $\epsilon(s) \overset{iid}{\sim} N(0, \sigma^2)$ captures measurement error variation at location s. The SVCM setup in (4.1) has advantages over its peers in the literature. For example, the varying coefficients $w(s)$ provide a more flexible and realistic modeling of responses with spatial indices (see, e.g., Gelfand *et al.*, 2003; Wheeler and Calder, 2007; Finley *et al.*, 2011; Guhaniyogi *et al.*, 2013; Kim and Wang, 2021, and references therein), so they perform better in practice than fitting deterministic trends in covariates, such as polynomial regression (Gelfand *et al.*, 2003). Customary geostatistical regression models with only a spatially-varying intercept emerge if the first column of $x(s)$ is the intercept and $P = 1$. Spatially-varying coefficient models also offer a process-based alternative to widely used geographically weighted regression (see, e.g., Brunsdon *et al.*, 1996) for modelling nonstationary behaviour in the mean. Chapter 3 provides a detailed review of varying coefficient models.

This chapter offers a novel inferential framework to address spatial variable selection with massive data. Our approach does not entail development of new class of models and approximations, rather we propose a three-stage framework that can be applied on well-established and well-tested models to enhance their scalability by multiple folds. The outline of our framework is as follows. First, we construct a number of (say, K) random matrices each of dimensions $M \times N$, where N is the sample size and $M << N$, and construct K compressed response vectors and predictor matrices by pre-multiplying the original response vector and predictor matrix by each random matrix. Second, posterior inference is drawn on varying coefficients after fitting a Bayesian SVCM, aided with a variable selection architecture, to the K compressed data in parallel. The K posterior distributions of model parameters computed in this manner are referred to as the "sketched posteriors." To reduce sensitivity on inference due to the choice of a random matrix in the second stage, the third stage computes Wasserstein barycenter of sketched posteriors for model parameters to derive "sketched pseudo posterior", that replaces the computationally expensive full data posterior distribution. For our exposition, we model the varying coefficients using basis expansions (Wikle, 2010; Wang *et al.*, 2008; Wang and Xia, 2009; Bai *et al.*, 2019) while computing the sketched posterior in the second stage, as it offers the most popularly used technique for variable selection in functional regressions.

Our proposal bears connection with the growing literature on divide-and-conquer Bayesian inference with large spatial datasets (Guhaniyogi and Banerjee, 2018; Guhaniyogi *et al.*, 2020b,a). This literature advocates dividing the data into a large number of subsets, fits a spatial model, e.g., an SVCM with each data subset, followed by aggregating subset posteriors through a measure of centrality on the space of distributions, such as their Wasserstein barycenter. Chapter 2 offers detailed description of the recently emerging divide-and-conquer methodology in Bayesian inference. Construction of data subsets is an important ingredient to this framework as it has been demonstrated that the resulting inference is somewhat sensitive to the choice of data subsetting strategy, see Guhaniyogi *et al.* (2020b) for a discussion on this topic with an empirical illustration of the issue. While our proposal follows a similar three-step strategy, it bypasses the construction of data subsets through the usage of random compression matrices and addresses the sensitivity to the choice of random matrices by computing the Wasserstein barycenter of sketched posteriors. The proposed framework is significantly different from a few recent articles (Maillard and Munos, 2009; Fard *et al.*, 2012; Guhaniyogi and Dunson, 2015, 2016) where the main objective is to facilitate efficient computation using random compression matrices in ordinary high-dimensional regression with small sample and large number of predictors.

Our novel contributions to the literature on spatially varying coefficient models are twofold. Methodologically, the main innovations are developing a three-stage inferential framework to accomplish variable selection and functional estimation for big spatial data. No restrictive data- or model-specific assumptions (e.g., the independence between data subsets or independence between blocks of parameters) and new algorithmic or software development are adopted and the framework still allows principled Bayesian inference on varying coefficients. The proposal also has an attractive feature of preserving confidentiality of response and predictor vectors, as the analysts can only be supplied with the compressed data which are much lower-dimensional and does not allow recovering the full uncompressed data.

The rest of the chapter proceeds as following. Section 4.1 discusses the framework for spatial variable selection in spatially varying coefficient models. Section 4.2 outlines the three stage distributed Bayesian framework for Bayesian implementation of SVCMs endowed with a functional variable selection architecture. Section 4.3 and Section 4.4 demonstrate performance of the proposed approach with simulation examples and a remote sensing data analysis, respectively. Finally, Section 4.5 summarizes our contribution in this chapter.

4.1 Bayesian Compressed Spatially Varying Coefficient Models

To illustrate our approach, we adopt the basis representation for the spatially varying coefficients $w_j(s)$ in (4.1). More specifically, each $w_j(s)$ assumes a basis representation with respect to H basis functions, given by,

$$w_j(s) = \sum_{h=1}^{H} B_{jh}(s)\gamma_{jh} \ , \quad j = 1, ..., P. \tag{4.2}$$

Each $B_{jh}(s)$ is a basis function evaluated at location s for $h = 1, ..., H$, and γ_{jh}'s are the corresponding basis coefficients. Detailed discussion on the choice of basis functions is available in Section 3.1 of Chapter 3.

Suppose $y(s_i)$ and $x(s_i)$ represent the data observed at the ith location s_i, $i = 1, ..., N$. Let y be the N dimensional response vector with its ith entry as $y(s_i)$, X be an $N \times P$ dimensional matrix with its ith row as $x(s_i)^T$ and $\tilde{X}$ be an $N \times NP$ dimensional block-diagonal matrix with its (i, i)th diagonal block as $x(s_i)^T$. Using (4.2) in (4.1) yields the Gaussian linear mixed model

$$y = X\beta + \tilde{X}B\gamma + \epsilon \ , \quad \epsilon \sim N(0, \sigma^2 I_N) \ . \tag{4.3}$$

where $B = (B(s_1)^{\mathrm{T}}, \ldots, B(s_N)^{\mathrm{T}})^{\mathrm{T}}$ is $NP \times HP$ with $B(s_n)$ a block-diagonal $P \times HP$ matrix whose j-th diagonal block is $(B_{j1}(s_n), \ldots, B_{jH}(s_n))$. The coefficient $\gamma = (\gamma_1^{\mathrm{T}}, ..., \gamma_P^{\mathrm{T}})^{\mathrm{T}}$ is $HP \times 1$ with each $\gamma_j = (\gamma_{j1}, \ldots, \gamma_{jH})^{\mathrm{T}}$ being $H \times 1$.

To determine which predictors are influential in explaining the response, we assign a block spike-and-slab mixture prior (Ishwaran and Rao, 2005; Li *et al.*, 2015; Bai *et al.*, 2019) to the basis coefficients corresponding to each predictor,

$$\begin{aligned}
&(\beta_j, \gamma_j^{\mathrm{T}})^{\mathrm{T}} | \tau_j^2 \overset{ind.}{\sim} (1-\pi_0) N(0, \sigma^2 \tau_j^2 I_{H+1}) + \pi_0 \delta_0, \\
&\tau_j^2 \overset{i.i.d.}{\sim} Gamma\left(\frac{H+1}{2}, \frac{\theta^2}{2}\right), \;\; \pi_0 \sim Beta(a_{\pi_0}, b_{\pi_0}), \\
&\qquad \theta^2 \sim Gamma(a_\theta, b_\theta), \qquad (4.4)
\end{aligned}$$

where δ_0 is the Dirac-delta function at 0, and the parameter π_0 corresponds to the probability of the zero mixture component. Note that if the jth predictor is not influential in predicting the response then, a-posteriori, $(\beta_j, \gamma_j^{\mathrm{T}})^{\mathrm{T}}$ should have a high probability being 0. Thus, based on the posterior probability of the event $\{(\beta_j, \gamma_j^{\mathrm{T}})^{\mathrm{T}} = 0\}$, it will be possible to identify unimportant spatial predictors in the regression. The beta prior distribution on π_0 ensures multiplicity correction in the variable selection framework (Scott and Berger, 2010). With prior distribution on γ set as a Gaussian scale-mixture distribution from the class of distributions given by (4.4), posterior computation using a blocked Metropolis-within-Gibbs algorithm cycles through updating the full conditional distributions: (a) $(\beta^{\mathrm{T}}, \gamma^{\mathrm{T}})^{\mathrm{T}} | \lambda, \sigma, \tau, \pi_0, \theta$, (b) $\lambda | \beta, \gamma, \sigma, \tau, \pi_0, \theta$, (c) $\sigma | \lambda, \beta, \gamma, \tau, \pi_0, \theta$, (d) $\tau | \lambda, \beta, \gamma, \sigma, \pi_0, \theta$, (e) $\pi_0 | \lambda, \beta, \gamma, \sigma, \tau, \theta$ and (f) $\theta | \lambda, \beta, \gamma, \sigma, \tau, \pi_0$. While updating (b), (c) (d),(e) and (f) do not face any computational challenge due to big N or P, full conditional posterior updating of $(\beta^{\mathrm{T}}, \gamma^{\mathrm{T}})^{\mathrm{T}} | \lambda, \sigma, \tau$ has the form given by

$$(1-\tilde{\pi}_0) N\left(\left(\tilde{X}_B^{\mathrm{T}} \tilde{X}_B + \Delta^{-1}\right)^{-1} \tilde{X}_B^{\mathrm{T}} y, \sigma^2 (\tilde{X}_B^{\mathrm{T}} \tilde{X}_B + \Delta^{-1})^{-1}\right) + \tilde{\pi}_0 \delta_0, \qquad (4.5)$$

where $\tilde{X}_B = [X : \tilde{X}B]$ is an $N \times (H+1)P$ matrix, $\Delta = \sigma^2 \mathrm{diag}(\tau_1^2, ..., \tau_P^2, \tau_1^2 I_H, .., \tau_P^2 I_H)$, and $\tilde{\pi}_0 = \pi_0 N(y|0, \sigma^2 I_N) / \{\pi_0 N(y|0, \sigma^2 I_N) + (1-\pi_0) N(y|0, \sigma^2 (I_N + \tilde{X}_B \Delta \tilde{X}_B^T))\}$ ($N(y|\mu, \Sigma)$ represents a normal density with mean μ and covariance matrix Σ).

The most efficient algorithm to sample from $(\beta^{\mathrm{T}},\gamma^{\mathrm{T}})^{\mathrm{T}}$ (Rue, 2001) computes Cholesky decomposition of $\left(\tilde{X}_B^T\tilde{X}_B+\Delta^{-1}\right)$ and employs the Cholesky factor to solve a series of linear systems to draw a sample from (4.5). In absence of any easily exploitable structure, computing and storing the Cholesky factor of this matrix involves $O((H+1)^3P^3)$ and $O((H+1)^2P^2)$ floating point operations, respectively (Golub and Van Loan, 2012), which leads to computational and storage bottlenecks with a large P, N and H. The next section proposes a three-step framework for efficient inference on varying coefficients.

4.2 Efficient Three-Step Bayesian Inference with Data Sketching

4.2.1 First step: construction of multiple sketched datasets

We construct K matrices $\Phi_1,...,\Phi_K$, each of dimensions $M\times N$, where $M\leq N$ and each entry of Φ_k is drawn randomly from a distribution $G(\cdot)$. We adopt the similar choice of $G(\cdot)$ as in Chapter 3 wherein each entry of Φ_k is drawn from $\mathrm{N}(0,1/N)$. A desirable choice of M depends on the smoothness of the spatially varying coefficients and the number of unimportant predictors. K is chosen moderately large, typically of the order of ~ 20. Each compression matrix, say Φ_k, $k=1,...,K$, is applied to y and $\tilde{X}_B$ to construct the compressed response vector $y_{\Phi_k}=\Phi_k y$ of dimensions $M\times 1$ and the compressed predictor matrix $\tilde{X}_{B,\Phi_k}=\Phi_k\tilde{X}_B$ of dimensions $M\times(H+1)P$. The compressed response vectors and predictor matrices will be used as surrogates to the uncompressed data for efficient model fitting of (4.3).

4.2.2 Second step: construction of sketched posteriors

With the kth compressed data $\{y_{\Phi_k}, \tilde{X}_{B,\Phi_k}\}$, we propose to fit the varying coefficient model (4.3) as in Chapter 3,

$$y_{\Phi_k} = \tilde{X}_{B,\Phi_k}(\beta^{\mathrm{T}}, \gamma^{\mathrm{T}})^{\mathrm{T}} + \tilde{\epsilon}\ , \quad \tilde{\epsilon} \sim N(0, \sigma^2 I_M),\ \ k = 1, ..., K. \tag{4.6}$$

In order to fit model (4.6), the analyst only needs to be supplied with the compressed data. The original predictor matrix $\tilde{X}_B$ is not possible to recover from $\tilde{X}_{B,\Phi_k} = \Phi_k \tilde{X}_B$, since the system of equations are grossly under-determined (even if Φ_K is known), as $M << N$.

The hierarchical Bayesian model constructed from (4.6) takes the form

$$\begin{aligned} p(\tau, \beta, \gamma, \sigma^2, \lambda, \pi_0, \theta \,|\, y_{\Phi_k}, \tilde{X}_{B,\Phi_k}) \propto\ & p(\tau, \sigma^2, \beta, \gamma, \lambda, \pi_0, \theta) \\ & \times N(y_{\Phi_k} \,|\, \tilde{X}_{B,\Phi_k}(\beta^{\mathrm{T}}, \gamma^{\mathrm{T}})^{\mathrm{T}}, \sigma^2 I_M)\ , \end{aligned} \tag{4.7}$$

where λ has already been defined earlier as the parameters used to specify the basis functions, if any. In our context, we have

$$\begin{aligned} p(\tau, \sigma^2, \beta, \gamma, \lambda, \pi_0) = & \prod_{i=1}^{P} \{(1-\pi_0) N((\beta_i^{\mathrm{T}}, \gamma_i^{\mathrm{T}})^{\mathrm{T}} \,|\, 0, \sigma^2 \tau^2 I_{H+1}) I[(\beta_i^{\mathrm{T}}, \gamma_i^{\mathrm{T}})^{\mathrm{T}} \neq 0] + \pi_0 \delta_0\} \\ & \times \prod_{i=1}^{P} Gamma\left(\tau_i^2 \,|\, \frac{H+1}{2}, \frac{\theta^2}{2}\right) \times Beta(\pi_0 | a_{\pi_0}, b_{\pi_0}) \\ & \times IG(\sigma^2 \,|\, a_\sigma, b_\sigma) \times Gamma(\theta^2 | a_\theta, b_\theta) \times p(\lambda)\ , \end{aligned} \tag{4.8}$$

The quantity $p(\tau, \sigma^2, \beta, \gamma, \lambda, \pi_0, \theta \,|\, y_{\Phi_k}, \tilde{X}_{B,\Phi_k})$ represents posterior distribution of the parameters given a specific compressed/sketched data, and is referred to as the "sketched posterior" distribution of the parameters.

Sketched posterior specified as in (4.8) leads to computational benefits over the full data posterior of the parameters. To see this, note that the posterior computation requires drawing Markov chain Monte Carlo (MCMC) samples sequentially from the full conditional posterior distributions of (a) $(\beta^{\mathrm{T}},\gamma^{\mathrm{T}})^{\mathrm{T}}|\lambda,\tau,\sigma,\pi_0,\theta$, (b) $\sigma^2|\beta,\gamma,\lambda,\tau,\pi_0,\theta$, (c) $\lambda|\tau,\sigma,\beta,\gamma,\pi_0,\theta$, (d) $\pi_0|\tau,\sigma,\beta,\gamma,\lambda,\theta$, (e) $\tau_j^2|\lambda,\beta,\gamma,\sigma,\pi_0,\theta$ and $\theta^2|\tau,\lambda,\beta,\gamma,\sigma,\pi_0$, $j=1,\ldots,P$. To this end, we have the following form of the full conditionals,

$$\begin{aligned}
\sigma^2|- &\sim IG(a_\sigma + M/2 + (H+1)P^*/2,\\
&\qquad b_\sigma + ||y_{\Phi_k} - \tilde{X}_{B,\Phi_k}(\beta^{\mathrm{T}},\gamma^{\mathrm{T}})^{\mathrm{T}}||^2/2 + \sum_{j=1}^{P}||(\beta_j,\gamma_j^{\mathrm{T}})||^2/(2\tau_j^2))\\
\tau_j^2|- &\sim Gamma((H+1)/2,\theta^2/2), \qquad\qquad \text{if} \quad (\beta_j,\gamma_j^T)^T = 0\\
1/\tau_j^2|- &\sim Inv-Gaussian(\theta\sigma/||(\beta_j,\gamma_j^{\mathrm{T}})||,\theta^2) \qquad \text{if} \quad (\beta_j,\gamma_j^T)^T \neq 0\\
\theta^2|- &\sim Gamma(P(H+1)/2 + a_\theta, \sum_{j=1}^{P}\tau_j^2/2 + b_\theta)\\
\pi_0|- &\sim Beta(a_{\pi_0} + P_0, b_{\pi_0} + P - P_0)
\end{aligned}$$

Where P^* is the number of nonzero $(\beta_j,\gamma_j^{\mathrm{T}}) \neq 0$, and $P_0 = P - P^*$. The computational challenges are presented by the full conditional distribution of $(\beta^{\mathrm{T}},\gamma^{\mathrm{T}})^{\mathrm{T}}$, which is given by

$$\begin{cases} N\left(\left(\tilde{X}_{B,\Phi_k}^{\mathrm{T}}\tilde{X}_{B,\Phi_k} + \Delta^{-1}\right)^{-1}\tilde{X}_{B,\Phi_k}^{\mathrm{T}}y_{\Phi_k}, \sigma^2(\tilde{X}_{B,\Phi_k}^{\mathrm{T}}\tilde{X}_{B,\Phi_k} + \Delta^{-1})^{-1}\right) & \text{w.p. } (1-\tilde{\pi}_{\Phi_k,0})\\ 0 & \text{w.p. } \tilde{\pi}_{\Phi_k,0}, \end{cases} \tag{4.9}$$

where $\tilde{\pi}_{\Phi_k,0} = \pi_0 N(y_{\Phi_k}|0,\sigma^2 I_M)/\{\pi_0 N(y_{\Phi_k}|0,\sigma^2 I_M) + (1-\pi_0)N(y_{\Phi_k}|0,\sigma^2(I_M + \tilde{X}_{B,\Phi_k}\Delta\tilde{X}_{B,\Phi_k}^{\mathrm{T}}))\}$. Computing $\tilde{\pi}_{\Phi_k,0}$ only requires Cholesky decomposition of the matrix $(I_M + \tilde{X}_{B,\Phi_k}\Delta\tilde{X}_{B,\Phi_k}^{\mathrm{T}})$, incurring $\sim M^3$ floating point operations and is efficient as $M << min(N, P)$. To achieve computational efficiency in drawing $(\beta^{\mathrm{T}}, \gamma^{\mathrm{T}})^{\mathrm{T}}$ jointly from the distribution $N\left(\left(\tilde{X}_{B,\Phi_k}^{\mathrm{T}}\tilde{X}_{B,\Phi_k} + \Delta^{-1}\right)^{-1}\tilde{X}_{B,\Phi_k}^{\mathrm{T}}y_{\Phi_k}, \sigma^2(\tilde{X}_{B,\Phi_k}^{\mathrm{T}}\tilde{X}_{B,\Phi_k} + \Delta^{-1})^{-1}\right)$, we adapt a recent algorithm proposed in Bhattacharya *et al.* (2016) (in the context of ordinary linear regression with uncompressed data and small sample size) to our setting. More specifically, we follow the steps given as here: (i) draw $\tilde{\gamma}_1 \sim N(0, \Delta)$ and $\tilde{\gamma}_2 \sim N(0, I_M)$; (ii) set $\tilde{\gamma}_3 = \tilde{X}_{B,\Phi_k}B\tilde{\gamma}_1/\sigma + \tilde{\gamma}_2$; (iii) solve $(\tilde{X}_{B,\Phi_k}\Delta\tilde{X}_{B,\Phi_k}^{\mathrm{T}}/\sigma^2 + I_M)\tilde{\gamma}_4 = (y_{\Phi_k}/\sigma - \tilde{\gamma}_3)$; and (iv) set $\tilde{\gamma}_5 = \tilde{\gamma}_1 + \Delta\tilde{X}_{B,\Phi}^{\mathrm{T}}\tilde{\gamma}_4/\sigma$. The resulting $\tilde{\gamma}_5$ is a draw from the full conditional posterior distribution of γ. The computation is dominated by step (iii), which incurs $\sim (M^3 + M^2(H+1)P)$ floating point operations. Finally, note that when basis functions involve parameters λ, they are updated using Metropolis-Hastings steps since no closed form full conditionals are generally available for them.

Predictive inference on $y(s_0)$, where s_0 is a new location, will proceed from the sketched posterior predictive distribution induced from the sketched posterior of parameters, given by

$$p(y(s_0)\,|\,y_{\Phi_k}, \tilde{X}_{B,\Phi_k}) = \int p(y(s_0)\,|\,y_{\Phi_k}, \tilde{X}_{B,\Phi_k}, \beta, \gamma, \sigma^2, \lambda, \pi_0)p(\beta, \gamma, \sigma^2, \lambda, \pi_0\,|\,y_{\Phi_k}, \tilde{X}_{B,\Phi_k}) \, d(\beta, \gamma, \sigma^2, \lambda, \pi_0). \tag{4.10}$$

Samples are drawn from the sketched posterior predictive distribution using composition sampling. To elaborate on it, for each post burn-in posterior samples $\{\beta^{(l)}, \gamma^{(l)}, \sigma^{2(l)}, \lambda^{(l)}, \pi_0^{(l)}\}$, we draw a sample $y(s_0)^{(l)} \sim N(\sum_{j=1}^{P} x_j(s_0)\beta_j^{(l)} + \sum_{j=1}^{P} x_j(s_0)w_j(s_0)^{(l)}, \sigma^{2(l)})$ from the SVCM model (4.1), where $w_j(s_0)^{(l)}$ is obtained from $\gamma^{(l)}$ and $\lambda^{(l)}$ using (4.2) and $l = 1, 2, \ldots, L$ indexes the L post burn-in posterior samples. The samples $\{y(s_0)^{(1)}, ..., y(s_0)^{(L)}\}$ constitute post-convergence MCMC samples from the sketched posterior predictive distribution.

4.2.3 Third step: construction of sketched pseudo posterior

The sketched posteriors are combined following the notion of Wasserstein barycenter, as defined and discussed in Section 1.1 of Chapter 1. In the context of our framework, we consider $\nu_1, ..., \nu_K$ as the sketched posteriors corresponding to the random matrices $\Phi_1, ..., \Phi_K$, respectively, i.e., $\nu_k = p(\theta | y_{\Phi_k}, X_{B,\Phi_k})$, where θ is a function of model parameters. We define the "sketched pseudo posterior" as the Wasserstein barycenter $\overline{\nu}$ as in (1.1), which provides a general notion of obtaining the mean of K possibly dependent sketched posterior distributions in the space of distributions. While the posterior $p(\theta|y, X)$ obtained from the uncompressed data are analytically intractable and computationally prohibitive, it is well approximated by the sketched pseudo posterior. Given that the MCMC samples of θ from sketched posteriors are available from the second step, one can conveniently estimate the empirical version of the sketched pseudo posterior following the algorithm outlined in Section 1.1 to compute Wasserstein barycenter of the posterior distributions of one-dimensional parameters. More specifically, if θ is one-dimensional and $\theta_{k,q}$ denotes the q-th empirical quantile of θ obtained from ν_k, $k = 1, ..., K$, then $\overline{\theta}_q = (1/k) \sum_{q=1}^{Q} \theta_{k,q}$, $q = 1, ..., Q$, denotes the qth empirical quantile of $\overline{\nu} = \overline{p}(\theta)$. We choose $Q = 10^4$ and compute quantiles on an equi-spaced grid of size Q on $(0, 1)$. To draw predictive inference at a new location s_0, we consider $\nu_k = p(y(s_0) \,|\, y_{\Phi_k}, \tilde{X}_{B,\Phi_k})$ as the sketched posterior predictive distribution (4.7) and draw samples from the Wasserstein barycenter of the sketched posterior predictive distributions, following the same strategy as outlined above.

The three-step approach to construct the sketched pseudo posterior closely resembles the construction of "meta posterior" in the recent literature on divide-and-conquer Bayesian inference with large spatial data (Guhaniyogi and Banerjee, 2018; Guhaniyogi *et al.*, 2020b,a) with an important difference. The current literature on divide-and-conquer inference in spatial models allows users to construct the data subsets. Empirical investigation in this literature reveals that the inference is somewhat sensitive to the choice of data subsets (Guhaniyogi *et al.*, 2020b). In contrast, our approach constructs multiple randomly linear transformation of the original data and reduces sensitivity of inference to the choice of the random linear transformations by computing Wasserstein barycenter of sketched posteriors. The remaining sections will empirically justify our proposed approach with simulation studies and a real data analysis.

4.3 Simulation Results

4.3.1 Inferential performance

This section empirically illustrates sketched pseudo posterior, by comparing its inferential and predictive performance, along with computational efficiency with a number of competitors on simulated data. To simulate the data, a fixed set of spatial locations $s_1, \dots, s_N$ are drawn uniformly over the domain $\mathcal{D} = [0,1] \times [0,1]$ and the number of spatially varying predictors is fixed at $P = 20$. For all $i = 1, ..., N$, we set $x_1(s_i) = 1$, and simulate $x_j(s_i)$ for $j = 2, ..., P$ independently from $N(0,1)$. To facilitate performance of our approach as a tool to spatial variable selection, we introduce sparsity in the data generation scheme. Specifically, out of $P = 20$ predictors, only the first $P^* = 3$ spatially varying predictors are assumed to be related to the response. Since our main focus is on the estimation of purely spatially varying predictor coefficients, our data generation scheme assumes that all nonzero predictor coefficients are purely spatially varying, i.e., $\beta = 0$ in the truth. Hence the response $y(s_i)$ is drawn independently from $N(\sum_{j=1}^{P^*} x_j(s_i) w_j^*(s_i), \sigma^{*2})$

following (4.1), where σ^{*2} is set to be 0.1. The true spatially varying coefficients ($w_j^*(s)$, $j = 1, \ldots, P^*$) are simulated from a Gaussian process with mean 0 and covariance kernel $C(\cdot, \cdot; \theta_j)$, i.e., $(w_j^*(s_1), ..., w_j^*(s_N))^{\mathrm{T}}$ is drawn from $N(0, C^*(\theta_j))$, for each $j = 1, \ldots, P^*$, where $C^*(\theta_j)$ is an $N \times N$ matrix with the (i, i')th element $C(s_i, s_{i'}; \theta_j)$. We set the covariance kernel $C(\cdot, \cdot; \theta_j)$ to be the exponential covariance function given by

$$C(s, s'; \theta_j) = \kappa_j^2 \exp\left\{-\frac{1}{2}\left(\frac{||s - s'||}{\phi_j}\right)\right\}, \ j = 1, .., P^*, \tag{4.11}$$

with the true values of $\kappa_1^2, \ldots, \kappa_{P^*}^2$ set to 1, 1.2, 0.8, respectively. We fix the true values of $\phi_1, \ldots, \phi_{P^*}$ at 1, 2, 1.25, for simulation case 1; and at 0.2, 0.15 and 0.1 for the simulation case 2 respectively.

While fitting our proposed approach, the varying coefficients are modelled through the linear combination of H basis functions as in (4.2), where these basis functions are chosen as the tensor-product of B-spline bases of order $\zeta = 4$ (Shen and Ghosal, 2015). More specifically, for $s = (s^{(1)}, s^{(2)})$, the j-th varying coefficient is modelled as

$$w_j(s) = \sum_{h_1=1}^{H_1} \sum_{h_2=1}^{H_2} B_{jh_1}^{(1)}(s^{(1)}) B_{jh_2}^{(2)}(s^{(2)}) \gamma_{jh_1h_2} \ , \tag{4.12}$$

where the marginal B-splines $B_{jh_1}^{(1)}$, $B_{jh_2}^{(2)}$ are defined on sets of H_1 and H_2 knots, respectively. The knots are chosen to be equally-spaced so the entire set of $H = H_1H_2$ knots is uniformly spaced over the domain $\mathcal{D}$. We complete the hierarchical specification by assigning independent $IG(a_\sigma = 2, b_\sigma = 0.1)$ priors (mean 0.1 with infinite variance) for σ^2, $IG(a_\tau = 2, b_\tau = 0.1)$ τ_j^2 for each $j = 1, \ldots, P$, $Beta(a_{\pi_0} = 1, b_{\pi_0} = 1)$ for π_0.

Two sets of closely related competitors are compared with our approach to assess its inferential and predictive performance. The first set of competitors is a three step approach where the data is divided randomly into K exhaustive subsets in the first step with each subset having exactly M data points. The second step fits a spatially varying coefficient model (spSVC) using the R package `spBayes` on each data subset (without any variable selection) and the third step combines subset posteriors using the Wasserstein barycenter of subset posteriors. This competitor is constructed following the recent literature on divide-and-conquer (d-&-c) inference on spatial models (Guhaniyogi *et al.*, 2020b), which bears close connection to our approach as discussed in Section 4.2. We implement d-&-c spSVC with two different subset construction schemes. in the first scheme, subsets are constructed randomly, and, in the second scheme the spatial domain is divided into sub-domains and each subset contains representative samples from each sub-domain. We refer to the first competitor as the d-&-c spSVC-random and the second competitor as the d-&-c spSVC-designed. Further, we implement another competitor that is identical to our approach in fitting (4.6) with different Φ_k's, except for the fact that a standard multivariate normal prior on $\boldsymbol{\gamma}$ are assigned without any variable selection framework in each subset, as we do in Chapter 3. We refer to the second competitor as geostatistical sketching (geoS), with a slight abuse of nomenclature, which only differs from our proposed approach in terms of not taking into account variable selection in fitting the sketched posteriors. We refer to our approach as geostatistical sketching with variable selection (geoS-VS). Since none of the competitors of geoS-VS addresses spatial variable selection, comparison with these competitors will show the relative inferential advantage of adding the variable selection architecture within our framework when the response variable in true SVCM is only influenced by a subset of predictors. Additionally, inference from d-&-c spSVC-random and d-&-c spSVC-designed will highlight the sensitivity to the choice of subsets in d-&-c Bayesian inference with spatial data, thereby offering the rationale to employ sketched pseudo posterior which addresses this issue. All four distributed methods are implemented with different choices of $K = 10, 15, 20$.

We applied these competitors to data generated with $N = 5000$ (case 1) and $N = 10000$ (case 2). For both cases the compressed dimension is taken to be $M \approx 10\sqrt{N}$ which seems to be effective from empirical considerations in our simulations. We provide further empirical justification for this choice in Section 3.3.2. Our approach compresses the sample sizes to $M = 700$ and $M = 1000$ in cases 1 and 2, respectively. For a fair comparison, M is kept the same for the competitors of our approach. The number of fitted basis functions in cases 1 & 2 are $H = 225, 256$, respectively.

All competing methods run in the `R` statistical computing environment on a Dell XPS 13 PC with Intel Core i7-8550U CPU @ 4.00GHz processors at 16 GB of RAM. 5000 MCMC iterations run for Bayesian inference with data subsets in d-&-c spSVC methods and for Bayesian inference with sketched posteriors. Posterior inference is based upon 2000 samples retained after adequate convergence is diagnosed using Monte Carlo standard errors and effective sample sizes (ESS) using the `mcmcse` package in `R`.

Figures 3.1 and 3.2 present the estimated truly nonzero varying coefficients by the competitors for cases 1 and 2, respectively. Since d-&-c spSVC-random and d-&-c spSVC-designed yield similar estimation of nonzero varying coefficients, we just present the estimated map for one of them. The figures show satisfactory estimation of nonzero varying coefficients by our approach with the estimated map closely capturing features of the true coefficients. Also, the estimated varying coefficients from spSVC undergo more smoothing than our approach. To investigate it further, we present mean squared error of estimating truly nonzero varying coefficients, defined as $MSE_0 = \sum_{j=1}^{P*} \sum_{i=1}^{N} (\widehat{w}_j(s_i) - w_j^*(s_i))^2/(P^*N)$ and the mean squared error of estimating all varying coefficients, defined as $MSE = \sum_{j=1}^{P} \sum_{i=1}^{N} (\widehat{w}_j(s_i) - w_j^*(s_i))^2/(PN)$ (where $\widehat{w}_j(s_n)$ is the posterior median of $w_j(s_n)$) in Table 4.1. The coverage of 95% credible intervals for varying coefficients are also presented in Table 4.1. The results demonstrate significantly better performance of our approach geoS-VS over d-&-c-spSVC methods in terms of offering smallest MSE and MSE$_0$ of estimating the varying coefficients. Our approach also enjoys superior performance over geoS. Both of this can be attributed to the variable selection architecture embedded in the geoS-VS method. The results also indicate discrepancy in the performances of d-&-c-spSVC-random and d-&-c-spSVC-designed in some cases, though the differences are not stark.

To assess the performance of our approach geoS-VS in terms of spatial variable selection, we present F_1 score, defined as $2 * \text{precision} * \text{recall}/(\text{precision} + \text{recall})$, for the sketched pseudo posterior. The highest possible value of an F$_1$-score is 1, indicating perfect precision and recall, and the lowest possible value is 0, if either the precision or the recall is zero. This score is only presented for our approach since our approach is the only one among competitors which is endowed with a variable selection framework. Table 4.1 shows perfect F$_1$-score under both simulation settings and for all choices of $K = 10, 15, 20$, confirming our approach being an efficient tool for spatial variable selection.

Table presents mean squared predictive error (MSPE), and coverage for the 95% predictive intervals, based on $N^* = 500$ out of the sample observations. Similar to the estimation of coefficients, we observe geoS-VS performs significantly better than d&-c-spSVC methods, and somewhat better over geoS without variable selection. The predictive coverage of all competitors are found to be close to nominal.

Finally, the computational efficiency of all competitors are computed based on the metric $\log_2(ESS/\text{Computation Time})$, where ESS denotes the effective sample size averaged over the MCMC samples of all parameters. The data sketching based approaches geoS and geoS-VS emerge as the two most computationally efficient methods among the competitors, whereas d&-c-spSVC methods yield much reduced metric for computational efficiency.

Method	F_1 score	MSE	MSE_0	95%CI CVG	95%CI CVG_0	95%PI CVG	$MSPE$	$C.Eff$
$N = 5000$, $H = 225$, $M = 710$								
spSVC-rdm ($K = 10$)	-	0.0161	0.0451	0.974	0.924	0.963	0.2037	4.26
spSVC-rdm ($K = 15$)	-	0.0205	0.0628	0.971	0.912	0.960	0.2676	4.68
spSVC-rdm ($K = 20$)	-	0.0216	0.0876	0.983	0.891	0.954	0.3269	4.96
spSVC-dsg ($K = 10$)	-	0.0227	0.0462	0.997	0.938	0.960	0.2123	4.32
spSVC-dsg ($K = 15$)	-	0.0221	0.0655	0.993	0.932	0.964	0.2763	4.89
spSVC-dsg ($K = 20$)	-	0.0216	0.0836	0.996	0.912	0.952	0.3266	4.95
geoS ($K = 10$)	-	0.0032	0.0198	0.965	0.911	0.946	0.1659	7.92
geoS ($K = 15$)	-	0.0026	0.0161	0.971	0.925	0.942	0.1452	7.91
geoS ($K = 20$)	-	0.0025	0.0159	0.963	0.934	0.938	0.1456	7.80
geoS-VS ($K = 10$)	1	0.0020	0.0138	0.969	0.923	0.976	0.1403	7.95
geoS-VS ($K = 15$)	1	**0.0019**	**0.0132**	0.979	0.939	0.978	**0.1393**	7.88
geoS-VS ($K = 20$)	1	0.0020	0.0133	0.962	0.918	0.980	0.1405	7.79
Method	F_1 score	MSE	MSE_0	95%CI CVG	95%CI CVG_0	95%PI CVG	$MSPE$	$C.Eff$
$N = 10000$, $H = 256$, $M = 1000$								
spSVC-rdm ($K = 10$)	-	0.0811	0.1351	0.981	0.910	0.958	0.1591	2.35
spSVC-rdm ($K = 15$)	-	0.0973	0.1596	0.990	0.901	0.961	0.1687	3.45
spSVC-rdm ($K = 20$)	-	0.1062	0.1831	0.976	0.903	0.967	0.1793	4.19
spSVC-dsg ($K = 10$)	-	0.0926	0.1478	0.976	0.927	0.959	0.1577	2.32
spSVC-dsg ($K = 15$)	-	0.1329	0.1712	0.971	0.913	0.959	0.1718	3.09
spSVC-dsg ($K = 20$)	-	0.0827	0.1565	0.983	0.918	0.964	0.1789	4.02
geoS ($K = 10$)	-	0.0027	0.0101	0.996	0.932	0.962	0.1523	8.89
geoS ($K = 15$)	-	0.0023	0.0089	0.998	0.943	0.967	0.1522	8.15
geoS ($K = 20$)	-	0.0027	0.0102	0.999	0.936	0.969	0.1547	8.19
geoS-VS ($K = 10$)	1	0.0021	0.0085	0.996	0.928	0.983	0.1485	8.96
geoS-VS ($K = 15$)	1	0.0019	0.0076	0.998	0.926	0.983	0.1438	7.93
geoS-VS ($K = 20$)	1	**0.0018**	**0.0072**	0.982	0.941	0.985	**0.1370**	7.74

Table 4.1: Mean Square Error for estimation of all spatially-varying coefficients (MSE) and truly nonzero spatially-varying coefficients (MSE_0), MSPE, coverage of 95% predictive intervals for the competing model. Computational efficiency for all models is also provided.

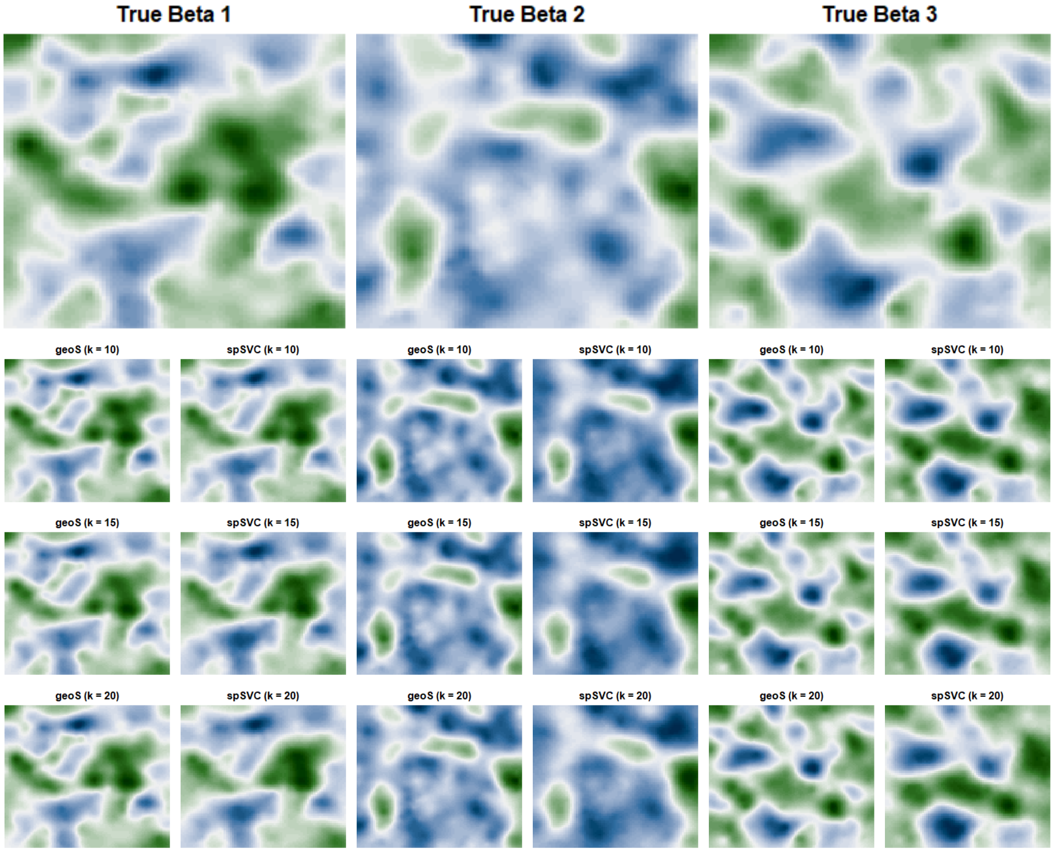


Figure 4.1: Simulation case 2: $(N, H) = (5000, 225)$. First row corresponds to the true surfaces of nonzero space-varying coefficients, second, third and fourth row present the predicted 50% quantile surfaces for the *geoS-VS* and *spSVC-designed* for different values of $K = 10, 15, 20$.

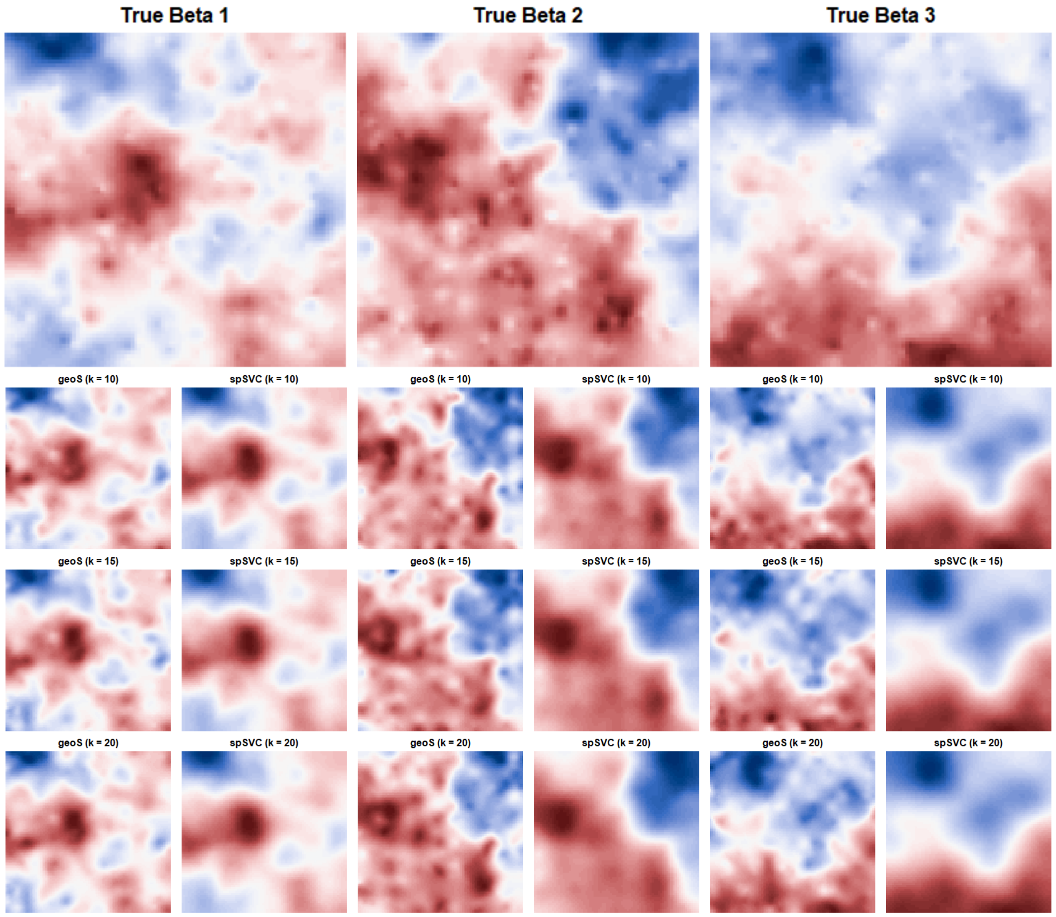
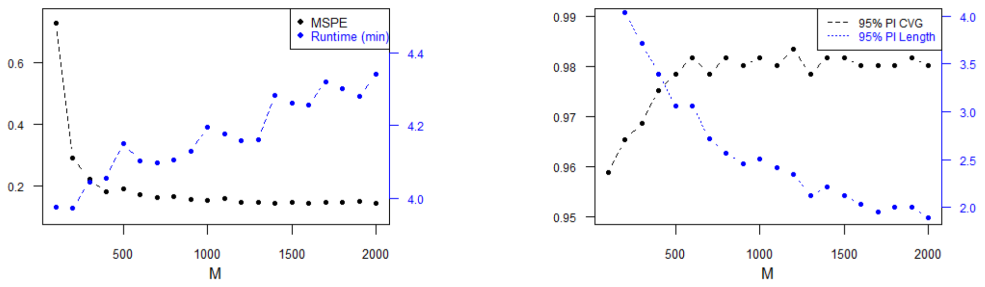


Figure 4.2: Simulation case 2: $(N, H) = (10000, 256)$. First row corresponds to the true surfaces of nonzero space-varying coefficients, second, third and fourth row present the predicted 50% quantile surfaces for the *geoS-VS* and *spSVC-random* for different values of $K = 10, 15, 20$.

4.3.2 Choice of the dimension of the compression matrix

In this section, we conduct empirical examinations around the choice of the pertinent compression matrix size M. For simulated data with sample size $N = 10000$ and number of servers $K = 20$, we ran our model for different values of $M = k\sqrt{N}$, $k = 1, \ldots, 20$. Figure 4.3 shows the variations in point-wise and interval prediction reflected in the $MSPE$ and 95% predicted interval coverage and length, respectively. Unsurprisingly, as M increases the MSPE drops with a diminished rate of decline until the $k \sim 10$. In terms of interval prediction, predictive coverage seems to fluctuate within the narrow interval $(0.97, 0.99)$ for $k > 5$, whereas the length of the predictive interval improves as M increases. Throughout several simulations and real data analysis we observe that the choice of $M \sim 10\sqrt{N}$ leads to good performance.



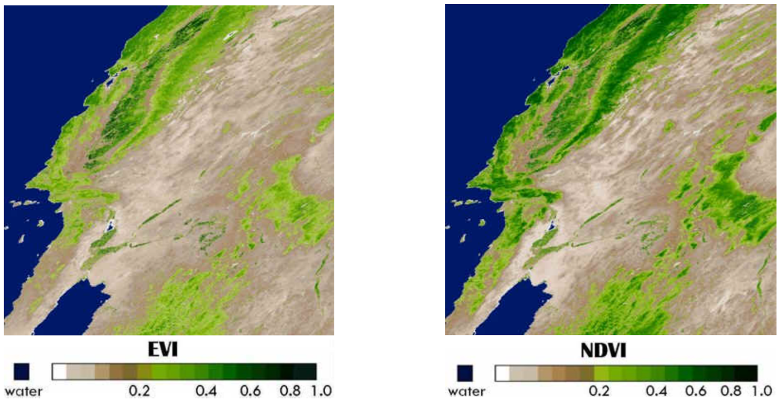
(a) **(b)**

Figure 4.3: (a) MSPE, (b) 95% predictive interval coverage and length for different choices of M

4.4 Vegetation Data Analysis

In this section, we present a real data application of our model for the prediction of the Enhanced Vegetation Index (EVI). As in section 3.4, we consider vegetation indices retrieved from the MODIS sensors aboard Terra and Aqua satellites which are broadly deployed in all ecosystem, climate, and natural resources management studies and operational research. EVI, which characterizes the global range of vegetation states, offers an alternative vegetation quantification to address some of the shortcomings of the normalized difference vegetation index (NDVI). Although both indices are derived from atmospherically-corrected reflection factors in the red, near-infrared, and blue wavebands; EVI is constructed by removing saturation and canopy background signals, and reducing atmospheric influences (Chen *et al.* (2005), Deng *et al.* (2007), Fensholt (2004), Huete *et al.* (2002), Huete *et al.* (1997), Miura *et al.* (2001), Potithep *et al.* (2013)); this ultimately results in a reduction of canopy-soil variations and in an improvement of sensitivity to changes in canopy structure, including leaf area index (LAI), and plant phenology and stress over regions of dense vegetation conditions and high biomass. Due to its robustness, EVI has become an effective index in tracking phenological events of crop growth, assessing and monitoring seasonal variations of crops and evergreen vegetation and quantifying evapotranspiration or water-use efficiency, which are influential in multiple applications including global biogeochemical and hydrologic modeling, agricultural monitoring and forecasting and land-use planning (Gurung *et al.* (2009), Potgieter *et al.* (2007), Wardlow *et al.* (2007)). The enhanced Vegetation index is calcualted as: $EVI = G\frac{NIR-RED}{NIR+C_1*RED-C_2*BLUE+L}$, where (NIR), (RED) and (BLUE) are atmospherically-corrected surface reflectances, C_1 and C_2 are the atmospheric correction coefficients of aerosol resistance, L is the canopy background adjustment that addresses non-linear, differential NIR and red radiant transfer through a canopy and G is a gain factor. The coefficients adopted in the MODIS-EVI algorithm are $L = 1$, $C_1 = 6$, $C_2 = 7.5$, and $G = 2.5$. Figure 4.4 shows the comparison between NDVI and EVI. The NDVI image shows a greater area in dark green because NDVI loses sensitivity to changes in vegetation in areas of higher biomass. The EVI image keeps a more consistent sensitivity to changes in vegetation and, in this example, has a more even distribution of vegetation greenness values.

We centered our analysis on geostatistical data that was projected on a sinusoidal (SIN) grid, located on the western coast of the United States, more precisely zone *h08v05*, between $30°N$ to $40°N$ latitude and $104°W$ to $130°W$ longitude. The database, which



(a) (b)

Figure 4.4: (a) MODIS EVI compared to NDVI (b) for western United States (*zone h08v05*) on August 12, 2001.

is distributed by the United States Geological Survey (USGS) as a standard MODIS product, is accessible through the `R` package, `MODIStsp`. We process a total number of 130000 observed locations, where all variables were derived over a 16-day period in April, 2016. For model fitting we kept $N = 80000$ observations randomly chosen, and held out the remaining $N^* = 50000$ for prediction and model assessment. The variable of interest $y(s_n)$ is the transformed EVI ($log(EVI + 1)$), which is to be predicted based on a set of $\tilde{P} = 8$ predictors, including variables $x_1(s_n), \ldots x_7(s_n)$ (see table 4.2) and an intercept. For the sake of identifying relevant spatially varying effects exclusively, we set $\beta = 0$, so that the response is fitted through the model $E(y(s_n)) = \tilde{x}(s_n)^T w(s_n)$, where $\tilde{x}(s_n) = (1, x_1(s_n), \ldots, x_7(s_n))^T$.

The large size of the data and limited computational resources at our disposal preclude fitting d-&-c-spSVC methods. Therefore, we only fit geoS-VS with geoS in order to detect any differences in estimation and predictive performance due to the introduction of the variable selection scheme. In concordance with 4.3.2, we set the dimension of the compression matrix to be $M \sim 10\sqrt{N} = 2800$ for both models. We deploy a linear combination of basis functions across a set of $H = H_1 H_2 = 15^2 = 225$ knots, distributed over the entire domain $\mathcal{D}$, in order to model the spatially-varying coefficients $\{\gamma_{jh}, j = 1, \ldots, 8; h = 1, \ldots, 225\}$, which results in $HP = 1800$ coefficients to be estimated. The specification for the basis functions is set to be the tensor-product of uni-dimensional B-splines of order $q = 4$.

We ran an MCMC chain for 5000 iterations and retained 2000 samples for posterior inference after proper convergence was determined. In Tables 4.3 and 4.4, we present numerical results on variable selection for model geoS-VS and prediction performance for both models respectively. According to model geoS-VS, only the intercept and variable **ET** are deemed to be relevant in explaining the response, as their posterior probabilities of inclusion are greater that 0.5. While there is some uncertainty regarding **URB** whose posterior probability of inclusion is between $10 - 30\%$, the model shows no uncertainty about not including the other variables in the set of influential predictors. Further, Table 4.4 reveals better performance of model geoS-VS over model geoS in terms of predictive accuracy which is reflected in the mean squared predictive error (MSPE), and significantly better performance in terms of precision as length of 95% PI from geoS-VS is notably narrower for all values of $K = 5, 10, 15$. Further, predictive coverage for both models are found to be close to nominal. This finding in further corroborated in Figure 4.5. The figure shows that both models yield excellent point prediction for the Enhanced Vegetation Index, as both satisfactorily capture global and local variations in the the response, however, model geoS-VS presents superior precision, which is exhibited in a lower standard deviation in the prediction (please refer to second and third rows in the first column). This can be attributed to the variable selection architecture, which disregards unimportant information leading to a more precise prediction of the response variable over the domain. Similar to simulation studies, no significant difference is found in terms of computation efficiency of the two competing methods. Overall, the data analysis serves as a demonstration for excellent performance of our proposed approach if offering efficient computation and accurate inference for spatial variable selection with large data.

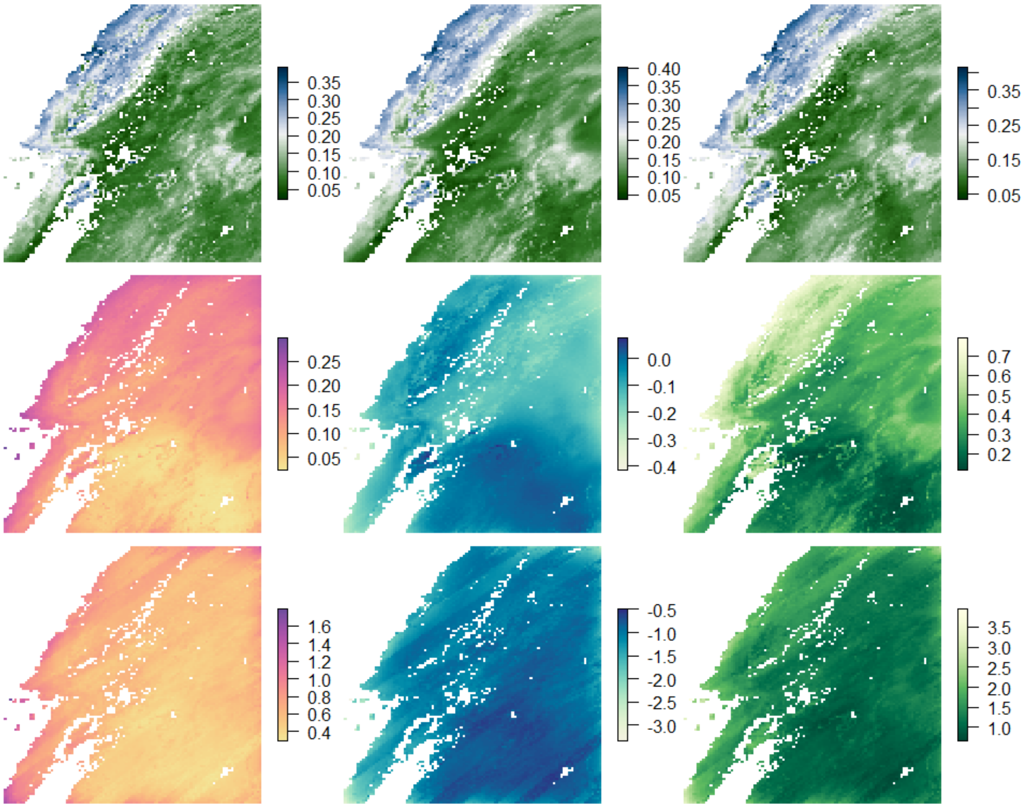


Figure 4.5: Coloured EVI images of western United States (zone h08v05). First row shows to the true EVI surface, and 50% quantile estimates for *geoS-VS* and *geoS* models respectively. Second row corresponds to the Std Dev, 2.5% and 97.5% quantiles for the *geoS-VS* model respectively. Third row presents the Std Dev, 2.5% and 97.5% quantiles for the *geoS* model respectively.

	Variable		Definition
GPP	$x_1(s)$	Gross primary productivity	Vegetation photosynthesis at the ecosystem scale.
LCType4	$x_2(s)$	Land Cover Type 4	Categorical. Annual BIOME-Biogeochemical Cycles - BGC classification. $x_2(s) \in \{0, 1, 2, 3, 4, 5, 6, 7, 8\}$
ET	$x_3(s)$	Evapotranspiration	Water loss occurring by the processes of evaporation and transpiration.
URB	$x_4(s)$	Urban area Index	Binary. $x_2(s) = \mathbb{1}_U(s)$, U denotes urban area.
EV	$x_5(s)$	Evergreen Vegetation	Binary. $x_4(s) = \mathbb{1}_E(s)$, E denotes presence of evergreen vegetation.
SAA	$x_6(s)$	Sun azimuth angle	Sun's relative angle along the local horizon.
SZA	$x_7(s)$	Sun zenith angle	Angle between the sun's rays and the vertical direction or Sun's apparent altitude.

Table 4.2: Vegetation data analysis set of predictors.

Predictor	$K = 5$	$K = 10$	$K = 15$
Intercept	1	1	1
GPP	0	0	0
LCType4	0	0	0
ET	1	1	1
URB	0.13	0.30	0.22
EV	0	0	0
SAA	0	0	0
SZA	0	0	0

Table 4.3: Posterior probabilities of inclusion for all predictors in geoS-VS with $K = 5, 10, 15$.

4.5 Summary

We have developed three-step distributed Bayesian inference for spatially varying coefficient models equipped with variable selection. The proposed approach constructs many sketches of the original data, fits spatially varying coefficient model with variables selection with each data sketch, followed by combining inferences of parameters obtained from applying the model with different data sketches. The proposed approach successfully marries two powerful ideas, divide-and-conquer inference and data sketching using random matrices, for scalable inference in spatial variable selection problem with big data. While the recently popular approaches in divide-and-conquer Bayesian inference can be sensitive to the choice of data subsets, our approach bypasses the sensitivity of distributed inference due to the choice of data subsets. Further, access to the values of the

		geoS-VS $M = 2800$	*geoS* $M = 2800$
$K = 5$	$MSPE$	**0.00045**	0.00053
	95% PI length	**0.4411**	2.2162
	95% PI coverage	0.990	0.991
	Runtime (sec)	444.59	367.198
	C.Eff	7.9807	8.3501
$K = 10$	$MSPE$	**0.00048**	0.00088
	95% PI length	**0.4264**	2.1111
	95% PI coverage	0.990	0.995
	Runtime (sec)	394.86	367.19
	C.Eff	7.9801	8.0283
$K = 15$	$MSPE$	**0.00050**	0.00065
	95% PI length	**0.4195**	2.311
	95% PI coverage	0.988	0.993
	Runtime (sec)	531.24	369.22
	C.Eff	6.9995	5.4675

Table 4.4: MSPE, coverage and length of 95% predictive intervals for geoS and geoS-VS are presented. Computational efficiency and computation time for both models are also provided. The divide and conquer spSVC methods could not be added due to excess computational burden and limited computational resources at our disposal.

response and predictors in the full data are not required at stage of inference, which preserves data confidentiality should that be of concern in the application.

Chapter 5

Conclusion and Future Work

This dissertation develops novel scalable Bayesian framework for large spatial data. In Chapter 2, we develop a divide-and-conquer Bayesian inferential tool for multivariate spatial generalized linear mixed effect models (spGLMMs) with large data. Scalable computation in spGLMMs is a methodologically challenging problem with the state-of-the-art approaches struggle to scale with more than ~ 50000 observations. To this end, we employ a three step approach that divides computation of the model into multiple processors with smaller subset of data in each processor. With abundant computational resources, the proposed method can scale computation of these models at an unprecedented level. Empirical analysis shows desirable inference with model parameters with the distributed Bayesian approach.

Chapter 3 discusses another novel concept of achieving scale and accuracy simultaneously in spatially varying coefficient (SVC) models with big data. Instead of fitting SVC models with big data that incurs computational issues, this chapter proposes fitting SVC models with a few random linear transformations of the original data. Theoretical results developed in this chapter shows close to asymptotically optimal rate of estimation of varying coefficients using the proposed approach. One of the significant novelties of this chapter is that it develops the theoretical and computational framework for the usage of random compression matrices in high dimensional Bayesian computation. Furthermore, the proposed approach requires only a few random linear combination of the original data be revealed to the analyst, thus protecting privacy of data samples. Empirical analysis shows state-of-the-art performance of the proposed approach.

Chapter 4 borrows strength from ideas developed in both Chapter 2 and 3 to offer a novel divide-and-conquer Bayesian inference with spatially varying coefficient models equipped with a functional variable selection framework. SVCs with functional variable selection architecture is remarkably inefficient in terms of computation with large spatial data. Extending ideas from Chapter 2 and 3, we develop a three stage divide-and-conquer framework to solve this problem. Specifically, we construct multiple random matrices to yield multiple randomly compressed data, fit SVC with the functional variable selection architecture with each compressed data, followed by combining inferences from them. Apart from providing a method for computationally efficient functional variable selection, this approach also eliminates the sensitivity of inference due to the choice of data subsets for the divide-and-conquer inferential framework described in Chapter 2. Empirical analysis shows excellent performance in terms of identifying important functional variables and the spatial variation of the corresponding coefficients.

A number of future directions emerge from here. While Chapter 2 empirically validates performance of the divide-and-conquer strategy with spGLMMs, we feel that it is important to derive theoretical results to assess the number of subsets as a function of sample size and smoothness of the spatial surface. As an immediate future work, we plan to build on the existing theoretical framework for divide-and-conquer strategy on continuous outcome model (Guhaniyogi and Banerjee, 2017;

Guhaniyogi *et al.*, 2020b,a) to the binary spGLMM framework. As another future work, we propose to investigate the framework in Chapter 2 when the response variables are observed over both space and time, and inferential objective lies in assessing both spatial correlation and the temporal trend of the underlying physical process. There is also a possibility of enhancing scalability in each subset by employing computationally efficient variants of multivariate Gaussian processes.

Chapter 3 and 4 of the thesis focuses on application of random compression idea to facilitate scaling of varying coefficient models in large sample size. Notably, the use of random compression matrices as demonstrated in these chapters is sufficiently general to be applied to other models. For example, high dimensional multivariate reduced-rank models face severe computational challenges when number of predictors and sample size are both large. The Bayesian data sketching idea, as proposed in Chapters 3 and 4, may offer a solution to this problem. The concept of data sketching also finds application in drawing scalable inference for high dimensional spatial generalized linear mixed effect models (spGLMMs). Another future project can evolve from extending the approaches in Chapter 3 and 4 for large spatio-temporal data. Finally, as discussed in the earlier chapters, the idea of random compression masks the original data to the analysts leading to privacy protected Bayesian inference. Notably, there is an extensive and ever growing literature on differential privacy in computer science which received very little attention by the Bayesian practitioners. As part of our future exploration, we seek to rigorously develop the connection between the data compression idea and differential privacy.

Appendix A

Theoretical Results

This section contains theoretical results building up to the proofs of Theorems 3.2.4 and 3.2.5. Lemma 3.2.1 states an important result from random matrix theory that is easily obtained from Theorem 5.31 and Corollary 5.35 of Vershynin (2010). We prove Lemmas 3.2.2 and 3.2.3. The results in Lemma 3.2.1-3.2.3 are further used to prove Theorems 3.2.4 and 3.2.5.

Proof of Lemma 3.2.2

Proof. Define

$$\mathcal{A}_{1N} = \left\{ K(f^*, f) \leq M_N \theta_N^2, \ V(f^*, f) \leq M_N \theta_N^2 \right\}. \tag{A.1}$$

By Lemma 10 in Ghosal *et al.* (2007), to show (3.10) it is enough to show that $\Pi(\mathcal{A}_{1N}) \gtrsim \exp(-C_2 M_N \theta_N^2)$,

for some constant $C_2 > 0$. Let e_k, $1 \leq k \leq M_N$ be the ordered eigenvalues of

$(\Phi\Phi^{\mathrm{T}})^{-1}$. After some calculations, we derive the following expressions,

$$K(f^*, f) = \frac{1}{2}\left\{\sum_{k=1}^{M_N}(e_k - 1 - \log(e_k)) + \mathbb{E}_{\mathcal{S}}\mathbb{E}_{\mathcal{X}}\left[||\tilde{X}_{\Phi,N}B(\gamma-\gamma^*) - \tilde{X}_{\Phi,N}\eta^*||_2^2\right]\right\} \text{ and}$$
$$V(f^*, f) = \sum_{k=1}^{M_N}\frac{(1-e_k)^2}{2} + \mathbb{E}_{\mathcal{S}}E_{\mathcal{X}}\left[||(\Phi\Phi^{\mathrm{T}})^{-1}(\tilde{X}_{\Phi,N}B(\gamma-\gamma^*) - \tilde{X}_{\Phi,N}\eta^*)||_2^2\right], \tag{A.2}$$

where $\eta^* = (\eta^*(s_1)^{\mathrm{T}}, ..., \eta^*(s_N)^{\mathrm{T}})^{\mathrm{T}}$, $\eta^*(s) = (\eta_1^*(s), \ldots, \eta_{\tilde{P}}^*(s))^{\mathrm{T}}$, $\eta_j^*(s) = w_j^*(s) - \sum_{h=1}^{H_N} B_{jh}(s)\gamma_{jh}^*$. Expanding $\log(e_k)$ in the powers of $(1-e_k)$ and using Lemma 1 in Jeong and Ghosal (2020) we find $(e_k - 1 - \log(e_k)) \sim (1-e_k)^2/2$. Another use of Lemma 1 in Jeong and Ghosal (2020) yields $\sum_{k=1}^{M_N}(1-e_k)^2 \lesssim ||I - \Phi\Phi^{\mathrm{T}}||_F^2 \lesssim M_N/N \leq M_N\theta_N^2$. Using Lemma 3.2.1, $e_k \asymp 1$ for all $k = 1, ..., M_N$. Hence, from (A.2)

$$\begin{aligned}\Pi(\mathcal{A}_{1N}) &\gtrsim \Pi\left(\left\{\gamma : \mathbb{E}_{\mathcal{S}}\mathbb{E}_{\mathcal{X}}\left[||\tilde{X}_{\Phi,N}B(\gamma-\gamma^*) - \tilde{X}_{\Phi,N}\eta^*||_2^2\right] \lesssim M_N\theta_N^2\right\}\right)\\ &\geq \Pi\left(\left\{\gamma : \mathbb{E}_{\mathcal{S}}\mathbb{E}_{\mathcal{X}}\left[||\tilde{X}_{\Phi,N}B(\gamma-\gamma^*)||_2^2\right] + \mathbb{E}_{\mathcal{S}}\mathbb{E}_{\mathcal{X}}\left[||\tilde{X}_{\Phi,N}\eta^*||_2^2\right] \lesssim M_N\theta_N^2/2\right\}\right),\end{aligned} \tag{A.3}$$

where we use $||a-b||_2^2 \leq 2(||a||_2^2 + ||b||_2^2)$, for all $a, b \in \mathbb{R}$. Let $B_j(s_n) = (B_{j1}(s_n), ..., B_{jH_N}(s_n))^{\mathrm{T}}$, for $n = 1, ..., N$ and $j = 1, ..., \tilde{P}$. By Assumption (E),

$$\begin{aligned}\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\mathcal{X}}\left[||\tilde{X}_{\Phi,N}B(\gamma-\gamma^*)||_2^2\right] &\asymp \kappa_N E_{\mathcal{S}}\mathbb{E}_{\mathcal{X}}\left[||\tilde{X}_N B(\gamma-\gamma^*)||_2^2\right]\\ &= \kappa_N(\gamma-\gamma^*)^{\mathrm{T}}\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\mathcal{X}}\left[B^{\mathrm{T}}\tilde{X}_N^{\mathrm{T}}\tilde{X}_N B\right](\gamma-\gamma^*).\end{aligned}$$

Recalling that $B^{\mathrm{T}}\tilde{X}_N^{\mathrm{T}}\tilde{X}_N B$ is a $H_N\tilde{P}\times H_N\tilde{P}$ matrix with the (j,j')-th block given by $\sum_{n=1}^N \tilde{x}_j(s_n)B_j(s_n)B_{j'}(s_n)^{\mathrm{T}}\tilde{x}_{j'}(s_n)$, we obtain

$$\begin{aligned}\mathbb{E}_{\mathcal{S}}E_{\mathcal{X}}\left[\sum_{n=1}^N \tilde{x}_j(s_n)B_j(s_n)B_{j'}(s_n)^{\mathrm{T}}\tilde{x}_{j'}(s_n)\right] &\asymp \mathbb{E}_{\mathcal{S}}\left[\sum_{n=1}^N B_j(s_n)B_{j'}(s_n)^{\mathrm{T}}\right]\\ &= N\mathbb{E}_{\mathcal{S}}\left[B_j(s_1)B_{j'}(s_1)^{\mathrm{T}}\right],\end{aligned}$$

where the last equation follows since $s_1,...,s_N$ are i.i.d.. Hence,

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\mathcal{X}}\left[||\tilde{X}_{\Phi,N}B(\gamma-\gamma^*)||_2^2\right] \asymp N\kappa_N\mathbb{E}_{\mathcal{S}}\left[||B(s_1)(\gamma-\gamma^*)||_2^2\right] \asymp N\kappa_N||\gamma-\gamma^*||_2^2/H_N, \tag{A.4}$$

where $B(s)=[B_1(s):\cdots:B_{\tilde{P}}(s)]^{\mathrm{T}}$. The last expression follows from Lemma A.1 of Huang *et al.* (2004). From Assumption (E) again,

$$\begin{aligned}\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\mathcal{X}}\left[||\tilde{X}_{\Phi,N}\eta^*||_2^2\right] &\asymp \kappa_N\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\mathcal{X}}\left[||\tilde{X}_N\eta^*||_2^2\right] = \kappa_N\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\mathcal{X}}\left[\sum_{n=1}^N\sum_{j=1}^{\tilde{P}}\tilde{x}_j(s_n)^2\eta_j^*(s_n)^2\right]\\ &\asymp \kappa_N\mathbb{E}_{\mathcal{S}}\left[\sum_{n=1}^N\sum_{j=1}^{\tilde{P}}\eta_j^*(s_n)^2\right] \lesssim N\kappa_N H_N^{-2\xi}, \end{aligned}\tag{A.5}$$

where the last inequality follows from Assumption (A). From (A.3),

$$\begin{aligned}\Pi(\mathcal{A}_{1N}) &\gtrsim \Pi\left(\gamma : N\kappa_N||\gamma-\gamma^*||_2^2/H_N + N\kappa_N H_N^{-2\xi} \lesssim M_N\theta_N^2/2\right)\\ &\gtrsim \Pi\left(\gamma : N\kappa_N||\gamma-\gamma^*||_2^2 \le M_N H_N\theta_N^2\right),\end{aligned}$$

where the last step follows from Assumptions (B) and (E). Using the fact that $\int_a^b \exp(-x^2/2)dx \geq \exp(-(a^2+b^2)/2)(b-a)$, we obtain

$$\begin{aligned}\Pi\left(\gamma : N\kappa_N||\gamma-\gamma^*||_2^2 \leq M_N H_N \theta_N^2\right) &\geq \prod_{h,j=1}^{H_N,\tilde{P}} \Pi(|\gamma_{jh}-\gamma^*_{jh}| \leq \theta_N/\sqrt{\tilde{P}}) \\ &\geq \exp(-||\gamma^*||_2^2 - \theta_N^2 H_N)(2\theta_N/\sqrt{\tilde{P}})^{H_N\tilde{P}} \gtrsim \exp(-M_N\theta_N^2 C_2),\end{aligned}$$

for any $C_2 > 0$, where the first inequality follows from Assumption (E) and the last inequality follows from $H_N P \log(\sqrt{\tilde{P}}/2\theta_N) \prec M_N\theta_N^2$ (since $M_N\theta_N^2 \asymp M_N^{1/(1+\xi)}$ while $H_N \prec M_N^{1/(1+\xi)}$).

Proof of Lemma 3.2.3

Proof. Denote $\tilde{X}_{\Phi,B,N} = \tilde{X}_{\Phi,N}B$, $\widehat{\gamma} = (\tilde{X}^{\mathrm{T}}_{\Phi,B,N}\tilde{X}_{\Phi,B,N})^{-1}\tilde{X}^{\mathrm{T}}_{\Phi,B,N}y_{\Phi,N}$ and a sequence of random variables $\zeta_N = I(||\tilde{X}_{\Phi,B,N}\widehat{\gamma} - \tilde{X}_{\Phi,B,N}\gamma^*||_2 \gtrsim \theta_N M_N^{1/2})$. Then,

$$\begin{aligned}\mathbb{E}^*(\zeta_N) &= P^*(||\tilde{X}_{\Phi,B,N}\widehat{\gamma} - \tilde{X}_{\Phi,B,N}\gamma^*||_2 \gtrsim \theta_N M_N^{1/2}) \\ &= P^*(||P_{\tilde{X}_{\Phi,B,N}}\tilde{X}_{\Phi,N}\eta^* + P_{\tilde{X}_{\Phi,B,N}}\epsilon||_2^2 \gtrsim \theta_N^2 M_N) \\ &\leq P^*(||P_{\tilde{X}_{\Phi,B,N}}\tilde{X}_{\Phi,N}\eta^*||_2^2 + ||P_{\tilde{X}_{\Phi,B,N}}\epsilon||_2^2 \gtrsim \theta_N^2 M_N),\end{aligned}$$

where $P_{\tilde{X}_{\Phi,B,N}}$ denotes the projection matrix corresponding to the matrix $\tilde{X}_{\Phi,B,N}$. Note that

$$||P_{\tilde{X}_{\Phi,B,N}}\tilde{X}_{\Phi,N}\eta^*||_2^2 \leq \eta^{*\mathrm{T}}\tilde{X}^{\mathrm{T}}_{\Phi,N}P_{\tilde{X}_{\Phi,B,N}}\tilde{X}_{\Phi,N}\eta^* \leq ||\tilde{X}_{\Phi,N}\eta^*||_2^2.$$

We then refer to equation (A.5) to see that $E_{\mathcal{S}}E_X||\tilde{X}_{\Phi,N}\eta^*||_2^2 \lesssim N\kappa_N M_N^{-\xi/(\xi+1)} \prec M_N\theta_N^2$. The above two facts together conclude that

$$\mathbb{E}_{\mathcal{S}}\mathbb{E}_X[||P_{\tilde{X}_{\Phi,B,N}}\tilde{X}_{\Phi,N}\eta^*||_2^2] \lesssim N\kappa_N M_N^{-\xi/(\xi+1)} \prec M_N\theta_N^2.$$

$\mathbb{E}^*(\zeta_N) \lesssim P^*(||P_{\tilde{X}_{\Phi,B,N}}\epsilon||_2^2 \gtrsim \theta_N^2 M_N) = P^*(\epsilon^{\mathrm{T}}P_{\tilde{X}_{\Phi,B,N}}\epsilon \gtrsim \theta_N^2 M_N)$.

Note that under P^*, $\epsilon \sim N(0,\Phi\Phi^{\mathrm{T}})$, and, $e_{max}(\Phi\Phi^{\mathrm{T}}) \asymp 1$ (by Lemma 3.2.1). Also note that Lemma 1 of Laurent and Massart (2000) can be simplified to write $P^*(\chi^2_{p^*} > x) \leq \exp(-x/4)$, for $x \geq 8p^*$. Further, $\epsilon^{\mathrm{T}}P_{\tilde{X}_{\Phi,B,N}}\epsilon$ follows a χ^2 distribution with degree of freedom less than equal to $H_N\tilde{P} \prec M_N\theta_N^2 = M_N^{1/(1+\xi)}$. Using all the above facts, we conclude that $E^*(\zeta_N) \lesssim \exp(-M_N\theta_N^2)$.

Next, for $\gamma \in \mathcal{B}_N^c$, we show that $\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\mathcal{X}}||\tilde{X}_{\Phi,B,N}\gamma - \tilde{X}_{\Phi,B,N}\gamma^*||_2^2 \gtrsim M_N\theta_N^2$. To see this, note that

$$\begin{aligned}\mathbb{E}_{\mathcal{S}}E_{\mathcal{X}}||\tilde{X}_{\Phi,B,N}\gamma - \tilde{X}_{\Phi,B,N}\gamma^*||_2^2 &= \mathbb{E}_{\mathcal{S}}\mathbb{E}_{\mathcal{X}}\left[(\gamma-\gamma^*)^{\mathrm{T}}\tilde{X}_{\Phi,B,N}^{\mathrm{T}}\tilde{X}_{\Phi,B,N}(\gamma-\gamma^*)\right] \\ &\asymp \kappa_N\mathbb{E}_{\mathcal{S}}\mathbb{E}_{\mathcal{X}}\left[(\gamma-\gamma^*)^{\mathrm{T}}B^{\mathrm{T}}\tilde{X}_N^{\mathrm{T}}\tilde{X}_N B(\gamma-\gamma^*)\right] \asymp N\kappa_N||\gamma-\gamma^*||_2^2/H_N \gtrsim M_N\theta_N^2,\end{aligned}$$

where the second line follows using similar calculations leading to equation (A.4).

Now, using the fact that $||\tilde{X}_{\Phi,B,N}\widehat{\gamma} - \tilde{X}_{\Phi,B,N}\gamma||_2 \geq -||\tilde{X}_{\Phi,B,N}\widehat{\gamma} - \tilde{X}_{\Phi,B,N}\gamma^*||_2 + ||\tilde{X}_{\Phi,B,N}\gamma - \tilde{X}_{\Phi,B,N}\gamma^*||_2$, we obtain

$$\begin{aligned}\mathbb{E}_\gamma(1-\zeta_N) &= P_\gamma(||\tilde{X}_{\Phi,B,N}\widehat{\gamma} - \tilde{X}_{\Phi,B,N}\gamma^*||_2 \lesssim \theta_N M_N^{1/2}) \\ &= P_\gamma(||\tilde{X}_{\Phi,B,N}\widehat{\gamma} - \tilde{X}_{\Phi,B,N}\gamma||_2 \gtrsim \theta_N M_N^{1/2}) \\ &\leq P_\gamma(||P_{\tilde{X}_{\Phi,B,N}}\epsilon||_2^2 \gtrsim \theta_N^2 M_N) \lesssim \exp(-M_N\theta_N^2),\end{aligned}$$

where the last inequality follows from simplifying the conclusion for Lemma 1 of Laurent and Massart (2000) (as is done before) and the fact that under P_γ,

$\epsilon \sim N(0, I)$.

A.0.1 Proof of Lemma 3.2.4

Proof. Note that,

$$\begin{aligned}||w - w^*||_2 &\leq ||w - \tilde{w}^* + \tilde{w}^* - w^*||_2 \leq ||w - \tilde{w}^*||_2 + ||\tilde{w}^* - w^*||_2 = ||w - \tilde{w}^*||_2 + ||\eta^*||_2 \\ &\lesssim ||w - \tilde{w}^*||_2 + P^{1/2} H_N^{-\xi} \asymp ||\gamma - \gamma^*||_2 H_N^{-1/2} + P^{1/2} H_N^{-\xi} \\ &\asymp ||\gamma - \gamma^*||_2 H_N^{-1/2} + P^{1/2} M_N^{-\xi/(2\xi+2)},\end{aligned}$$

where $\tilde{w}^*(s) = (\sum_{h=1}^{H_N} B_{1h}(s)\gamma_{1h}^*, \ldots, \sum_{h=1}^{H_N} B_{\tilde{P}h}(s)\gamma_{\tilde{P}h}^*)^{\mathrm{T}}$, and the first inequality in the second line follows from the property of B-splines (Huang *et al.*, 2004). The second expression in the second line follows from Lemma A.1 of Huang *et al.* (2004). Using the fact that $\tilde{P}^{1/2} M_N^{-\xi/(2\xi+2)} = O(\theta_N)$, we have $\Big\{w : ||w - w^*||_2 \geq \tilde{C}\theta_N\Big\} \subset \Big\{\gamma : ||\gamma - \gamma^*||_2 H_N^{-1/2} \geq C_{2w}\theta_N\Big\}$, for some constant $C_{2w} > 0$.

Denote $\mathcal{B}_N = \Big\{\gamma : ||\gamma - \gamma^*||_2 H_N^{-1/2} \leq C_{2w}\theta_N\Big\}$. To prove the theorem, it is enough to establish

$$\mathbb{E}^* \Pi(||\gamma - \gamma^*||_2 H_N^{-1/2} \geq C_{2w}\theta_N | y_{\Phi,N}, \tilde{X}_{\Phi,N}) \to 0, \text{ as } N \to \infty, \tag{A.6}$$

Note that,

$$\begin{aligned}\mathbb{E}^*[\Pi(\mathcal{B}_N^c \,|\, y_{\Phi,N}, \tilde{X}_{\Phi,N})] &\leq \mathbb{E}^*\zeta_N + \mathbb{E}^*[\Pi(\mathcal{B}_N^c \,|\, y_{\Phi,N}, \tilde{X}_{\Phi,N})(1 - \zeta_N) 1_{y_N \in \mathcal{A}_N^c}] + P^*(\mathcal{A}_N) \\ &= \mathbb{E}^*[\zeta_N] + \mathbb{E}^*\left[1_{y_N \in \mathcal{A}_N^c} \frac{\Big\{(1 - \zeta_N) \int_{\mathcal{B}_N^c} \{f(y_{\Phi,N}|\gamma)/f^*(y_{\Phi,N}|\gamma^*)\}\pi_N(\gamma) d\gamma\Big\}}{\{\int \{f(y_{\Phi,N}|\gamma)/f^*(y_{\Phi,N}|\gamma^*)\}\pi_N(\gamma) d\gamma\}}\right] + P^*(\mathcal{A}_N),\end{aligned} \tag{A.7}$$

where $\mathcal{A}_N$ is a set defined in the statement of Lemma 3.2.2 and ζ_N can be regarded as a sequence of random variables as defined in Lemma 3.2.3. By Lemma 3.2.2, $P^*(\mathcal{A}_N) \to 0$, as $N, M_N \to \infty$. Also, by Lemma 3.2.3, $\mathbb{E}^*\zeta_N \to 0$, as $N, M_N \to \infty$. To show (A.6), it remains to prove that

$$\frac{\mathbb{E}^*\left[1_{y_N \in \mathcal{A}_N^c} \int_{\mathcal{B}_N^c} \{f(y_{\Phi,N}|\gamma)/f^*(y_{\Phi,N}|\gamma^*)\}\pi_N(\gamma)d\gamma\right]}{[\int\{f(y_{\Phi,N}|\gamma)/f^*(y_{\Phi,N}|\gamma^*)\}\pi_N(\gamma)d\gamma]} \to 0 \text{ as } N, M_N \to \infty.$$

To this end, we have

$$\begin{aligned}\mathbb{E}^*\left[1_{y_N \in \mathcal{A}_N^c} \int_{\mathcal{B}_N^c} \{f(y_{\Phi,N}|\gamma)/f^*(y_{\Phi,N}|\gamma^*)\}\pi_N(\gamma)d\gamma\right] &\leq \sup_{\gamma \in \mathcal{B}_n^c} \mathbb{E}_\gamma(1-\zeta_N)\Pi(\mathcal{B}_N^c) \\ &\leq \exp(-C_{2w}M_N\theta_N^2),\end{aligned}$$

where $\Pi(\mathcal{B}_N^c)$ is the prior probability of the set $\mathcal{B}_N^C$. The denominator $\int\{f(y_{\Phi,N}|\gamma)/f^*(y_{\Phi,N}|\gamma^*)\}\pi(\gamma)d\gamma \geq \exp(-C_1M_N\theta_N^2)$ on $\mathcal{A}_N$, where C_1 is chosen so that $C_1 < C_{2w}$. Thus, $\mathbb{E}^*\Pi(\mathcal{B}_N^c \mid y_{\Phi,N}, \tilde{X}_{\Phi,N})1_{y_N \in \mathcal{A}_N^c} \leq \exp(-(C_{2w}-C_1)M_N\theta_N^2) \to 0$, as $N, M_N \to \infty$.

A.0.2 Proof of Theorem 3.2.5

Proof. For densities f_u and f^*, we have

$$\begin{aligned}h(f_u, f^*) &= 1 - \exp\left\{-\left(\sum_{j=1}^{\tilde{P}} \tilde{x}_j(s_0)w_j(s_0) - \sum_{j=1}^{\tilde{P}} \tilde{x}_j(s_0)w_j^*(s_0)\right)^2/8\right\} \\ &\leq 1 - \exp\left\{-\tilde{P}\sum_{j=1}^{\tilde{P}}\left(w_j(s_0) - w_j^*(s_0)\right)^2/8\right\} \\ &\leq 1 - \exp\left\{-\tilde{P}||w(s_0) - w^*(s_0)||_2^2/8\right\}\end{aligned}$$

Then, $\mathbb{E}_{\mathcal{S}}[h(f_u, f^*)] \leq 1 - \exp\left(-\tilde{P}||w - w^*||_2^2/8\right)$, by Jensen's inequality. Further,

$\mathbb{E}^*\mathbb{E}\mathbb{E}_{\mathcal{S}}[h(f_u, f^*)|\tilde{X}_{\Phi,N}, y_{\Phi,N}] = \left\{1 - \exp\left(-\tilde{P}\tilde{C}^2\theta_N^2/8\right)\right\} + 2\Pi_N(||w - w^*||_2 \geq \tilde{C}\theta_N)$,

which implies

$$\mathbb{E}^*\mathbb{E}\mathbb{E}_{\mathcal{S}}[h(f_u, f^*)] \leq \left\{1 - \exp\left(-\tilde{P}\tilde{C}^2\theta_N^2/8\right)\right\} + 2\mathbb{E}^*\Pi_N(||w - w^*||_2 \geq \tilde{C}\theta_N) \to 0$$

as $N, M_N \to \infty$, where the last expression followed by the conclusion of Theorem 3.2.4 and the fact that $\theta_N \to 0$ as $N, M_N \to \infty$.

Bibliography

Agueh, M. and Carlier, G. (2011a). Barycenters in the wasserstein space. *SIAM Journal on Mathematical Analysis*, **43**(2), 904–924.

Agueh, M. and Carlier, G. (2011b). Barycenters in the Wasserstein space. *SIAM Journal on Mathematical Analysis*, **43**(2), 904–924.

Ahfock, D., Astle, W. J., and Richardson, S. (2017). Statistical properties of sketching algorithms. *arXiv preprint arXiv:1706.03665*.

Ailon, N. and Chazelle, B. (2009). The fast johnson–lindenstrauss transform and approximate nearest neighbors. *SIAM J. Comput.*, **39**, 302–322.

Araújo, M. B. and Peterson, A. T. (2012). Uses and misuses of bioclimatic envelope modeling. *Ecology*, **93**(7), 1527–1539.

Bai, R., Boland, M. R., and Chen, Y. (2019). Fast algorithms and theory for high-dimensional bayesian varying coefficient models. *arXiv preprint arXiv:1907.06477*.

Banerjee, S. (2017). High-dimensional Bayesian geostatistics. *Bayesian Analysis*, **12**, 583–614.

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

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

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014a). *Hierarchical Modeling and Analysis for Spatial Data*. Chapman & Hall/CRC, Boca Raton, FL, second edition.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2014b). *Hierarchical modeling and analysis for spatial data*. Crc Press.

Berrett, C. and Calder, C. A. (2016). Bayesian spatial binary classification. *Spatial Statistics*, **16**, 72–102.

Bhattacharya, A., Chakraborty, A., and Mallick, B. K. (2016). Fast sampling with gaussian scale mixture priors in high-dimensional regression. *Biometrika*, page asw042.

Bickel, P. J. and Freedman, D. A. (1981). Some asymptotic theory for the bootstrap. *The annals of statistics*, **9**(6), 1196–1217.

Biller, C. and Fahrmeir, L. (2001). Bayesian varying-coefficient models using adaptive regression splines. *Statistical Modelling*, **1**(3), 195–211.

Bliznyuk, N., Ruppert, D., Shoemaker, C., Regis, R., Wild, S., and Mugunthan, P. (2008). Bayesian calibration and uncertainty analysis for computationally expensive models using optimization and radial basis function approximation. *Journal of Computational and Graphical Statistics*, **17**(2), 270–294.

Braverman, V., Ostrovsky, R., and Rabani, Y. (2010). Rademacher chaos, random eulerian graphs and the sparse johnson-lindenstrauss transform. *arXiv preprint arXiv:1011.2590*.

Broder, A. (1997). On the resemblance and containment of documents, in "compression and complexity of sequences (sequences'97)".

Brunsdon, C., Fotheringham, A. S., and Charlton, M. E. (1996). Geographically weighted regression: A method for exploring spatial nonstationarity. *Geographical Analysis*, **28**(4), 281–298.

Candes, E. J. and Tao, T. (2005). Decoding by linear programming. *IEEE transactions on information theory*, **51**(12), 4203–4215.

Candes, E. J. and Tao, T. (2006). Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE transactions on information theory*, **52**(12), 5406–5425.

Chen, S., Liu, Y., Lyu, M. R., King, I., and Zhang, S. (2015). Fast relative-error approximation algorithm for ridge regression. In *UAI*, pages 201–210.

Chen, X., Vierling, L., and Deering, D. (2005). A simple and effective radiometric correction method to improve landscape change detection across sensors and across time. *Remote Sensing of Environment*, **98**(1), 63–79.

Clarkson, K. and Woodruff, D. (2017). Low-rank approximation and regression in input sparsity time. *Journal of the ACM*, **63**, 1–45.

Cormode, G., Garofalakis, M., Haas, P. J., Jermaine, C., *et al.* (2011). Synopses for massive data: Samples, histograms, wavelets, sketches. *Foundations and Trends® in Databases*, **4**(1–3), 1–294.

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

Cressie, N. and Wikle, C. K. (2015). *Statistics for spatio-temporal data*. John Wiley & Sons.

Cuturi, M. and Doucet, A. (2014). Fast computation of wasserstein barycenters. In *International conference on machine learning*, pages 685–693. PMLR.

Dasgupta, A., Kumar, R., and Sarlós, T. (2010). A sparse johnson: Lindenstrauss transform. In *Proceedings of the forty-second ACM symposium on Theory of computing*, pages 341–350.

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016a). Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, **111**(514), 800–812.

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016b). Hierarchical nearest-neighbor gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, **111**, 800–812.

Deng, F., Su, G., and Liu, C. (2007). Seasonal variation of modis vegetation indexes and their statistical relationship with climate over the subtropic evergreen forest in zhejiang, china. *IEEE Geoscience and Remote Sensing Letters*, **4**(2), 236–240.

Deshpande, S. K., Bai, R., Balocchi, C., Starling, J. E., and Weiss, J. (2020). Vcbart: Bayesian trees for varying coefficients. *arXiv preprint arXiv:2003.06416*.

Dobriban, E. and Liu, S. (2018). A new theory for sketching in linear regression. *arXiv preprint arXiv:1810.06089*.

Donoho, D. L. (2006). Compressed sensing. *IEEE Transactions on information theory*, **52**(4), 1289–1306.

Drineas, P., Mahoney, M. W., Muthukrishnan, S., and Sarlós, T. (2011). Faster least squares approximation. *Numerische mathematik*, **117**(2), 219–249.

Du, J., Zhang, H., and Mandrekar, V. (2009). Fixed-domain asymptotic properties of tapered maximum likelihood estimators. *The Annals of Statistics*, **37**(6A), 3330–3361.

Eidsvik, J., Shaby, B. A., Reich, B. J., Wheeler, M., and Niemi, J. (2014). Estimation and prediction in spatial models with block composite likelihoods. *Journal of Computational and Graphical Statistics*, **23**(2), 295–315.

Eldar, Y. C. and Kutyniok, G. (2012). *Compressed sensing: theory and applications*. Cambridge university press.

Fard, M. M., Grinberg, Y., Pineau, J., and Precup, D. (2012). Compressed least-squares regression on sparse spaces. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*.

Fensholt, R. (2004). Earth observation of vegetation status in the sahelian and sudanian west africa: comparison of terra modis and noaa avhrr satellite data. *International Journal of Remote Sensing*, **25**(9), 1641–1659.

Finley, A. O. (2011). Comparing spatially-varying coefficients models for analysis of ecological data with non-stationary and anisotropic residual dependence. *Methods in Ecology and Evolution*, **2**(2), 143–154.

Finley, A. O., Banerjee, S., Waldmann, P., and Ericsson, T. (2009). Hierarchical spatial modeling of additive and dominance genetic variance for large spatial trial datasets. *Biometrics*, **65**(2), 441–451.

Finley, A. O., Banerjee, S., and MacFarlane, D. W. (2011). A hierarchical model for quantifying forest variables over large heterogeneous landscapes with uncertain forest areas. *Journal of the American Statistical Association*, **106**(493), 31–48.

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

Gelfand, A. E., Kim, H.-J., Sirmans, C., and Banerjee, S. (2003). Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association*, **98**(462), 387–396.

Gelfand, A. E., Schmidt, A. M., Banerjee, S., and Sirmans, C. (2004). Nonstationary multivariate process modeling through spatially varying coregionalization. *Test*, **13**(2), 263–312.

Gelfand, A. E., Diggle, P., Guttorp, P., and Fuentes, M. (2010a). *Handbook of spatial statistics*. CRC press.

Gelfand, A. E., Diggle, P., Guttorp, P., and Fuentes, M., editors (2010b). *Handbook of Spatial Statistics*. CRC Press, Boca Raton, FL.

Ghosal, S., Van Der Vaart, A., *et al.* (2007). Convergence rates of posterior distributions for noniid observations. *Annals of Statistics*, **35**(1), 192–223.

Gneiting, T., Kleiber, W., and Schlather, M. (2010). Matérn cross-covariance functions for multivariate random fields. *Journal of the American Statistical Association*, **105**(491), 1167–1177.

Golub, G. H. and Van Loan, C. F. (2012). *Matrix computations*, volume 3. JHU press.

Gramacy, R. B. and Lee, H. K. H. (2008). Bayesian treed gaussian process models with an application to computer modeling. *Journal of the American Statistical Association*, **103**(483), 1119–1130.

Guan, Y. and Haran, M. (2018). A computationally efficient projection-based approach for spatial generalized linear mixed models. *Journal of Computational and Graphical Statistics*, **27**(4), 701–714.

Guhaniyogi, R. (2017). Multivariate bias adjusted tapered predictive process models. *Spatial Statistics*.

Guhaniyogi, R. and Banerjee, S. (2017). Meta-kriging: Scalable Bayesian modeling and inference for massive spatial datasets. *https://www.soe.ucsc.edu/sites/default/files/technical-reports/UCSC-SOE-17-07.pdf*.

Guhaniyogi, R. and Banerjee, S. (2018). Meta-kriging: Scalable bayesian modeling and inference for massive spatial datasets. *Technometrics*, **60**(4), 430–444.

Guhaniyogi, R. and Dunson, D. B. (2015). Bayesian compressed regression. *Journal of the American Statistical Association*, **110**(512), 1500–1514.

Guhaniyogi, R. and Dunson, D. B. (2016). Compressed gaussian process for manifold regression. *The Journal of Machine Learning Research*, **17**(1), 2472–2497.

Guhaniyogi, R. and Sansó, B. (2018). Large multi-scale spatial kriging using tree shrinkage priors. *arXiv preprint arXiv:1803.11331*.

Guhaniyogi, R., Finley, A. O., Banerjee, S., and Gelfand, A. E. (2011). Adaptive gaussian predictive process models for large spatial datasets. *Environmetrics*, **22**(8), 997–1007.

Guhaniyogi, R., Finley, A. O., Banerjee, S., and Kobe, R. K. (2013). Modeling complex spatial dependencies: Low-rank spatially varying cross-covariances with application to soil nutrient data. *Journal of agricultural, biological, and environmental statistics*, **18**(3), 274–298.

Guhaniyogi, R., Li, C., Savitsky, T. D., and Srivastava, S. (2020a). Distributed bayesian varying coefficient modeling using a gaussian process prior. *arXiv preprint arXiv:2006.00783*.

Guhaniyogi, R., Li, C., Savitsky, T. D., and Srivastava, S. (2020b). A divide-and-conquer bayesian approach to large-scale kriging. *arXiv preprint arXiv:1712.09767*.

Guisan, A. and Zimmermann, N. E. (2000). Predictive habitat distribution models in ecology. *Ecological modelling*, **135**(2-3), 147–186.

Gurung, R. B., Breidt, F. J., Dutin, A., and Ogle, S. M. (2009). Predicting enhanced vegetation index (evi) curves for ecosystem modeling applications. *Remote Sensing of Environment*, **113**(10), 2186–2193.

Halko, N., Martinsson, P.-G., and Tropp, J. A. (2011). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, **53**(2), 217–288.

Harrison, S. P., Prentice, I. C., Barboni, D., Kohfeld, K. E., Ni, J., and Sutra, J.-P. (2010). Ecophysiological and bioclimatic foundations for a global plant functional classification. *Journal of vegetation Science*, **21**(2), 300–317.

Heaton, M. J., Datta, A., Finley, A. O., Furrer, R., Guinness, J., Guhaniyogi, R., Gerber, F., Gramacy, R. B., Hammerling, D., Katzfuss, M., *et al.* (2019). A case study competition among methods for analyzing large spatial data. *Journal of Agricultural, Biological and Environmental Statistics*, **24**(3), 398–425.

Huang, J. Z., Wu, C. O., and Zhou, L. (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica*, pages 763–788.

Huang, Z. (2018). Near optimal frequent directions for sketching dense and sparse matrices. In *International Conference on Machine Learning*, pages 2048–2057. PMLR.

Huang, Z., Li, J., Nott, D., Feng, L., Ng, T.-P., and Wong, T.-Y. (2015). Bayesian estimation of varying-coefficient models with missing data, with application to the singapore longitudinal aging study. *Journal of Statistical Computation and Simulation*, **85**(12), 2364–2377.

Huete, A., Liu, H., Batchily, K., and Van Leeuwen, W. (1997). A comparison of vegetation indices over a global set of tm images for eos-modis. *Remote sensing of environment*, **59**(3), 440–451.

Huete, A., Didan, K., Miura, T., Rodriguez, E. P., Gao, X., and Ferreira, L. G. (2002). Overview of the radiometric and biophysical performance of the modis vegetation indices. *Remote sensing of environment*, **83**(1-2), 195–213.

Ishwaran, H. and Rao, J. S. (2005). Spike and slab variable selection: Frequentist and Bayesian strategies. *Annals of Statistics*, **33**(2), 730–773.

Jeong, S. and Ghosal, S. (2020). Unified bayesian asymptotic theory for sparse linear regression. *arXiv preprint arXiv:2008.10230*.

Ji, S., Xue, Y., and Carin, L. (2008). Bayesian compressive sensing. *IEEE Transactions on signal processing*, **56**(6), 2346–2356.

Kane, D. M. and Nelson, J. (2010). A derandomized sparse johnson-lindenstrauss transform. *arXiv preprint arXiv:1006.3585*.

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

Katzfuss, M. and Guinness, J. (2021). A General Framework for Vecchia Approximations of Gaussian Processes. *Statistical Science*, **36**(1), 124 – 141.

Kaufman, C. G., Schervish, M. J., and Nychka, D. W. (2008). Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association*, **103**(484), 1545–1555.

Kim, M. and Wang, L. (2021). Generalized spatially varying coefficient models. *Journal of Computational and Graphical Statistics*, **30**(1), 1–10.

Kucharik, C. J., Foley, J. A., Delire, C., Fisher, V. A., Coe, M. T., Lenters, J. D., Young-Molling, C., Ramankutty, N., Norman, J. M., and Gower, S. T. (2000). Testing the performance of a dynamic global ecosystem model: water balance, carbon balance, and vegetation structure. *Global Biogeochemical Cycles*, **14**(3), 795–825.

Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338.

Lemos, R. T. and Sansó, B. (2009). A spatio-temporal model for mean, anomaly, and trend fields of north atlantic sea surface temperature. *Journal of the American Statistical Association*, **104**(485), 5–18.

Li, C., Srivastava, S., and Dunson, D. B. (2017). Simple, scalable and accurate posterior interval estimation. *Biometrika*, **104**(3), 665–680.

Li, J., Wang, Z., Li, R., and Wu, R. (2015). Bayesian group lasso for nonparametric varying-coefficient models with application to functional genome-wide association studies. *The annals of applied statistics*, **9**(2), 640.

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

Mahoney, M. W. (2011). Randomized algorithms for matrices and data. *arXiv preprint arXiv:1104.5557*.

Maillard, O. and Munos, R. (2009). Compressed least-squares regression. *Advances in neural information processing systems*, **22**.

Meng, X. and Mahoney, M. W. (2013). Low-distortion subspace embeddings in input-sparsity time and applications to robust linear regression. In *Proceedings of the forty-fifth annual ACM symposium on Theory of computing*, pages 91–100.

Miura, T., Huete, A. R., Yoshioka, H., and Holben, B. N. (2001). An error and sensitivity analysis of atmospheric resistant vegetation indices derived from dark target-based atmospheric correction. *Remote sensing of Environment*, **78**(3), 284–298.

Nelson, J. and Nguyên, H. L. (2013). Osnap: Faster numerical linear algebra algorithms via sparser subspace embeddings. In *2013 ieee 54th annual symposium on foundations of computer science*, pages 117–126. IEEE.

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

Peruzzi, M., Banerjee, S., and Finley, A. O. (2020). Highly scalable bayesian geostatistical modeling via meshed gaussian processes on partitioned domains. *Journal of the American Statistical Association (in press)*.

Peterson, A. T., Soberón, J., Pearson, R. G., Anderson, R. P., Martínez-Meyer, E., Nakamura, M., and Araújo, M. B. (2011). Ecological niches and geographic distributions (mpb-49). In *Ecological Niches and Geographic Distributions (MPB-49)*. Princeton University Press.

Polson, N. G., Scott, J. G., and Windle, J. (2013). Bayesian inference for logistic models using pólya–gamma latent variables. *Journal of the American statistical Association*, **108**(504), 1339–1349.

Potgieter, A. B., Apan, A., Dunn, P., and Hammer, G. (2007). Estimating crop area using seasonal time series of enhanced vegetation index from modis satellite imagery. *Australian Journal of Agricultural Research*, **58**(4), 316–325.

Potithep, S., Nagai, S., Nasahara, K. N., Muraoka, H., and Suzuki, R. (2013). Two separate periods of the lai–vis relationships using in situ measurements in a deciduous broadleaf forest. *Agricultural and forest meteorology*, **169**, 148–155.

Prentice, I. C., Cramer, W., Harrison, S. P., Leemans, R., Monserud, R. A., and Solomon, A. M. (1992). Special paper: a global biome model based on plant physiology and dominance, soil properties and climate. *Journal of biogeography*, pages 117–134.

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

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

Sarlos, T. (2006). Improved approximation algorithms for large matrices via random projections. In *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '06, page 143–152, USA. IEEE Computer Society.

Sato, H. and Ise, T. (2022). Predicting global terrestrial biomes with the lenet convolutional neural network. *Geoscientific Model Development*, **15**(7), 3121–3132.

Schabenberger, O. and Gotway, C. A. (2004). *Statistical methods for spatial data analysis*. CRC press.

Scott, J. G. and Berger, J. O. (2010). Bayes and empirical-bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, pages 2587–2619.

Scott, S. L., Blocker, A. W., Bonassi, F. V., Chipman, H. A., George, E. I., and McCulloch, R. E. (2016). Bayes and big data: The consensus monte carlo algorithm. *International Journal of Management Science and Engineering Management*, **11**(2), 78–88.

Shaby, B. and Ruppert, D. (2012). Tapered covariance: Bayesian estimation and asymptotics. *Journal of Computational and Graphical Statistics*, **21**(2), 433–452.

Shen, W. and Ghosal, S. (2015). Adaptive bayesian procedures using random series priors. *Scandinavian Journal of Statistics*, **42**(4), 1194–1213.

Srivastava, S., Cevher, V., Dinh, Q., and Dunson, D. (2015). WASP: Scalable Bayes via Barycenters of subset posteriors. In *Proceedings of the 18th International Conference on Artificial Intelligence and Statistics*, pages 912–920.

Srivastava, S., Li, C., and Dunson, D. B. (2018). Scalable bayes via barycenter in wasserstein space. *The Journal of Machine Learning Research*, **19**(1), 312–346.

Stein, M. L., Chi, Z., and Welty, L. J. (2004). Approximating likelihoods for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **66**(2), 275–296.

Vaart, A. v. d. and Zanten, H. v. (2011). Information rates of nonparametric gaussian process methods. *Journal of Machine Learning Research*, **12**(Jun), 2095–2119.

Van der Vaart, A. W., van Zanten, J. H., *et al.* (2009). Adaptive bayesian estimation using a gaussian random field with inverse gamma bandwidth. *The Annals of Statistics*, **37**(5B), 2655–2675.

Vecchia, A. V. (1988a). Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, **50**(2), 297–312.

Vecchia, A. V. (1988b). Estimation and model identification for continuous spatial processes. *Journal of the Royal Statistical Society: Series B*, **50**, 297–312.

Vempala, S. S. (2005). *The random projection method*, volume 65. American Mathematical Soc.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

Vidakovic, B. (2009). *Statistical modeling by wavelets*, volume 503. John Wiley & Sons.

Wang, H. and Xia, Y. (2009). Shrinkage estimation of the varying coefficient model. *Journal of the American Statistical Association*, **104**(486), 747–757.

Wang, L., Li, H., and Huang, J. Z. (2008). Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements. *Journal of the American Statistical Association*, **103**(484), 1556–1569.

Wardlow, B. D., Egbert, S. L., and Kastens, J. H. (2007). Analysis of time-series modis 250 m vegetation index data for crop classification in the us central great plains. *Remote sensing of environment*, **108**(3), 290–310.

Wessels, K., Steenkamp, K., Von Maltitz, G., and Archibald, S. (2011). Remotely sensed vegetation phenology for describing and predicting the biomes of south africa. *Applied Vegetation Science*, **14**(1), 49–66.

Wheeler, D. C. and Calder, C. A. (2007). An assessment of coefficient accuracy in linear regression models with spatially varying coefficients. *Journal of Geographical Systems*, **9**(2), 145–166.

Wikle, C. K. (2010). Low-rank representations for spatial processes. *Handbook of Spatial Statistics*, pages 107–118. Gelfand, A. E., Diggle, P., Fuentes, M. and Guttorp, P., editors, Chapman and Hall/CRC, pp. 107-118.

Woodruff, D. P. (2014). Sketching as a tool for numerical linear algebra. *arXiv preprint arXiv:1411.4357*.

Yang, J., Meng, X., and Mahoney, M. W. (2015). Implementing randomized matrix algorithms in parallel and distributed environments. *Proceedings of the IEEE*, **104**(1), 58–92.

Yuan, X., Llull, P., Brady, D. J., and Carin, L. (2014). Tree-structure bayesian compressive sensing for video. *arXiv preprint arXiv:1410.3080*.

Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, **99**(465), 250–261.

Zhang, L., Mahdavi, M., Jin, R., Yang, T., and Zhu, S. (2013). Recovering the optimal solution by dual random projection. In *Conference on Learning Theory*, pages 135–157.

Zhou, S., Wasserman, L., and Lafferty, J. D. (2008). Compressed regression. In *Advances in Neural Information Processing Systems*, pages 1713–1720.