UNIVERSITY OF CALIFORNIA

Los Angeles

On Simplified Bayesian Modeling for Massive Geostatistical Datasets: Conjugacy and Beyond

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Biostatistics

by

Lu Zhang

2020

ABSTRACT OF THE DISSERTATION

On Simplified Bayesian Modeling for Massive Geostatistical Datasets:
Conjugacy and Beyond

by

Lu Zhang
Doctor of Philosophy in Biostatistics
University of California, Los Angeles, 2020
Professor Sudipto Banerjee, Chair

With continued advances in Geographic Information Systems and related computational technologies, researchers in diverse fields like forestry, environmental health, climate sciences etc. have growing interests in analyzing large scale data sets measured at a substantial number of geographic locations. Geostatistical models used to capture the space varying relationships in such data are often accompanied by onerous computations which prohibit the analysis of large scale spatial data sets. Less burdensome alternatives proposed recently for analyzing massive spatial datasets often lead to inaccurate inference or require slow sampling process. Bayesian inference, while attractive for accommodating uncertainties through their hierarchical structures, can become computationally onerous for modeling massive spatial data sets because of their reliance on iterative estimation algorithms. My dissertation research aims at developing computationally scalable Bayesian geostatistical models that provide valid inference through highly accelerated sampling process. We also study the asymptotic properties of estimators in spatial analysis.

In Chapter 2 and 3, we develop conjugate Bayesian frameworks for analyzing univariate and multivariate spatial data. We propose a conjugate latent Nearest-Neighbor Gaussian Process (NNGP) model in Chapter 2, which uses analytically tractable posterior distributions to obtain posterior inferences, including the large dimensional latent process. In

Chapter 3, we focus on building conjugate Bayesian frameworks for analyzing multivariate spatial data. We utilize Matrix-Normal Inverse-Wishart(MNIW) prior to propose conjugate Bayesian frameworks and algorithms that can incorporate a family of scalable spatial modeling methodologies.

In Chapter 4, we pursue general Bayesian modeling methodologies beyond a conjugate Bayesian hierarchical modeling. We build scalable versions of a hierarchical linear model of coregionalization (LMC) and spatial factor models, and propose a highly accelerated block update MCMC algorithm. Using the proposed Bayesian LMC model, we extend scalable modeling strategies for a single process into multivariate process cases.

All proposed frameworks are tested on simulated data and fit to real data sets with observed locations numbering in the millions. Our contribution is to offer practicing scientists and spatial analysts practical and flexible scalable hierarchical models for analyzing massive spatial data sets.

In Chapter 5, we investigate the asymptotic properties of the estimators in spatial analysis. We formally establish results on the identifiability and consistency of the nugget in spatial models based upon the Gaussian process within the framework of in-fill asymptotics, i.e. the sample size increases within a sampling domain that is bounded. We establish the identifiability of parameters in the Matérn covariance function and the consistency of their maximum likelihood estimators in the presence of discontinuities due to the nugget.

The dissertation of Lu Zhang is approved.

Yifang Zhu

Robert Erin Weiss

Hua Zhou

Sudipto Banerjee, Committee Chair

University of California, Los Angeles

2020

To whom it may concern.

Table of Contents

List of Figures

List of Tables

Acknowledgments

First of all, I thank my advisor Sudipto Banerjee for his patient guidance and kind support throughout my Ph.D. years. Not only does he provide me with guidance on concrete problems, but he also leads by example to teach me how to become a professional researcher. He always encourages me to be positive and confident in my research. He teaches things that make me being better prepared for my future academic career. I look back upon the time under the instruction of Sudipto as so important in the development of the person that I am today. I feel so lucky to be his student.

I am very honored that Prof. Hua Zhou, Prof. Robort Weiss and Prof. Yifang Zhu agreed to serve on my thesis committee. I appreciate the continuous help from Prof. Zhou throughout my graduate study. I would like to thank him for the inspiring discussions, the support and writing the several recommendation letters for me. I thank Prof. Weiss for the instructions of course arrangements, writing and presentation skills. I thank Prof. Zhu for broadening my horizon in the real data applications.

I would like to thank Prof. Dorota M Dabrowska for spending a lot of time with me and teaching me about probability theory and stochastic processes individually. I appreciate her great suggestions in my career plan. I am grateful to Prof. Andrew Gelman who offered me the opportunity to work with Stan team, and I thank Prof. Bob Carpenter, Prof. Michael Betancourt and Prof. Aki Vehtari for their warmly helps in writing study case. I thank Prof. Ramirez for providing me the chance of collaborating on COVID-19 related researches. I thank Prof. Aki Nishimura, Prof. Bao Wang, Prof. Chenchen Mou and Dr. Jiajun Tong for the several fruitful discussions.

I am sincerely grateful to my collaborators Prof. Wenpin Tang, Prof. Andrew O. Finley and Prof. Abhirup Datta. I really appreciate their excellent work and help in theory development, computing, writing and proof reading. I thank them for their patience, encouragement and friendship. I learned a lot of great ideas and skills while working together

with them.

In biostatistics department of UCLA, I met many wonderful people. I had a colorful life in the biostatistics department because of all the friends: Alec Chan-Golston, Nada Abdalla, Justin Williams, Jason Wang, Leiwen Gao, Eric Kawaguchi, Juhyun Kim, Jason Clague etc. I really enjoy being with them as well as discussing different statistics with them.

I would like to express my gratitude to my family, especially my dad. Without their support and understanding, I would not have been able to pursue my study and career abroad. Their whole-hearted love sustains my life. Their faith and confidence in me, without any limitations, shaped me to be who I am.

Last but not least, I thank my boyfriend Yuming Zhang, whose accompany has been one of my biggest source of happiness. I am grateful for his care and trust and I enjoy learning some mathematics from him. Wish him all the best in his postdoc researches and in his future endeavors.

Vita

2010-2014	B.S. (Mathematics), Fudan University.
2014–present	Teaching Assistant, Biostatistics Department, UCLA.
2014–present	Research Assistant, Biostatistics Department, UCLA.

Publications

Lu Zhang, Sudipto Banerjee and Andrew O. Finley (In review). High-dimensional Multivariate Geostatistics: A Conjugate Bayesian Matrix-Normal Approach. *arXiv:2003.10051.*

Wenpin Tang, Lu Zhang and Sudipto Banerjee (In review). On identifiability and consistency of the nugget in Gaussian spatial process models. *arXiv:1908.05726.*

Lu Zhang, Abhirup Datta and Sudipto Banerjee (2019). Practical Bayesian modeling and inference for massive spatial datasets on modest computing environments. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 12(3):197-209.

Lu Zhang and Suditpo Banerjee (2018). A Note on the comparison of Nearest Neighbor Gaussian Process (NNGP) based models. *arXiv:1811.03735.*

Lu Zhang, LiZhen Nie, Sudipto Banerjee, *JALAJni: A JAVA package providing a java interface for lapack and blas library Github:*`https://github.com/JaLAJni/JaLAJni`

Xiang Chen, Lu Zhang, Sudipto Banerjee, *JAMAJniLite: A JAVA package providing a java interface for lapack and blas libraries and using the classes defined by JAMA Package Github:*`https://github.com/JAMAJni/JAMAJniLite`

Lu Zhang and Jun Yin (2018). *phase1PRMD: Personalized Repeated Measurement Design for Phase I Clinical Trials. R package version 0.1.0. CRAN:*`https://cran.r-project.org/web/packages/phase1PRMD/index.html`

CHAPTER 1

Introduction

Technical advances in sensing, transferring and storing have enabled the collection of massive-scale data in this era. Human beings create data in explosively growing scale across diverse disciplines. The information explosion has invoked growing interests in analyzing very large data sets. In fields like forestry, environmental health and climate sciences, variables are measured at a large number of locations. Statistical models used to capture the space varying relationships in such data are often accompanied by onerous computations that is prohibitive for analyzing large-scale data sets. This has generated substantial interest over the last decade in scalable methodologies for analyzing large spatial datasets. Scalable spatial process models using the Bayesian paradigm have been found especially attractive due to their richness and flexibility in hierarchical model settings. However, inference in Bayesian hierarchical modeling usually relies on iterative methods like MCMC algorithms. These algorithms can become computationally expensive and have a prohibitively slow sampling process. Moreover, when the full Bayesian inference in a spatial regression model includes latent processes, which depict spatially-varying randomness over the study domain, the dimension of the parameter space becomes linear in the number of observed locations. This high-dimensional parameter space usually results in a very slow convergence rate of MCMC algorithms. The vast majority of research articles [see, e.g., Banerjee, 2017, Heaton et al., 2017, and references therein] present scalable spatial modeling and have been geared toward innovative theory or more complex model development, yet very limited attention has been accorded to approaches for easily implementable scalable hierarchical models for the practicing scientist or spatial analyst.

This dissertation research aims at developing computationally scalable Bayesian geostatistical models that provide valid inference with highly accelerated sampling process. Our methodologies can be incorporated with a family of existing scalable spatial models. Meanwhile, the proposed models feature efficient posterior sampling over a high-dimensional parameter space including latent processes. In general, we make Bayesian inference more feasible through two strategies, constructing conjugate Bayesian framework to avoid slow iterative sampling algorithm or accelerating the MCMC algorithm. We also study the asymptotic properties of estimators in spatial analysis.

The remainder of this chapter is organized as follows. Section 1.1 provides a brief background introduction to hierarchical spatial modeling and Gaussian process. Section 1.2 gives a road-map of this dissertation and a brief summary of the contribution to the field.

1.1 Background to Hierarchical Spatial Modeling

Much of spatial modeling is carried out within the familiar hierarchical modeling paradigm,

$$[\text{data} \,|\, \text{process}] \times [\text{process} \,|\, \text{parameters}] \times [\text{parameters}] \;. \tag{1.1.1}$$

For point-referenced data sets, where spatial locations are indexed by coordinates on a map, the "process" is modeled as a spatial random field over the domain of interest and the observations are treated as a finite realization of this random field. The Gaussian process (GP) is, perhaps, the most prominent of process specifications and offers flexibility and richness in modeling. It is denoted as $\{w(\mathbf{s}) \sim \mathrm{GP}(m_\theta(\cdot), C_\theta(\cdot,\cdot)), \mathbf{s} \in \mathcal{D}\}$, where $\{w(\mathbf{s}),\ \mathbf{s} \in \mathcal{D}\}$ is the random field defined on domain $\mathcal{D}$, θ is a set of unknown parameters, $m_\theta(\cdot)$ is a mean function defining the trend, and $C_\theta(\cdot,\cdot)$ is a positive definite covariance function. The GP's popularity as a modeling tool is enhanced due to their extensibility to multivariate and spatial-temporal geostatistical settings, although we do not pursue such generalizations in this Chapter. They also provide comparatively greater theoretical tractability among spatial processes [Stein, 1999].

Fitting GPs incur onerous computational costs that severely hinders their implementation for large datasets [see, e.g., Gelfand et al., 2010, Cressie and Wikle, 2011, Banerjee et al., 2014]. The key bottleneck stems from the massive spatial covariance matrix present in the multivariate normal density for the finite realizations of the GP. For irregularly situated spatial locations, as is common in geostatistics, these matrices are typically dense and carry no exploitable structure to facilitate computations. To be precise, if $\mathbf{w}(\mathcal{S})$ is the $n \times 1$ vector of $w(\mathbf{s}_i)$'s over a set $\mathcal{S} = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n\}$ of n locations. Then $\mathbf{w}(\mathcal{S}) \sim \mathrm{N}(\mathbf{m}_\theta(\mathcal{S}), \mathbf{C}_\theta(\mathcal{S}, \mathcal{S}))$, where $\mathbf{m}_\theta(\mathcal{S})$ is the corresponding $n \times 1$ mean vector and $\mathbf{C}_\theta(\mathcal{S}, \mathcal{S})$ is an $n \times n$ covariance matrix whose entries are given by the covariance function $C_\theta(\mathbf{s}_i, \mathbf{s}_j)$. Computing this density requires the Cholesky decomposition for $\mathbf{C}_\theta(\mathcal{S}, \mathcal{S})$ involving storage in the order of $\sim n^2$ and floating point operations in the order of $\sim n^3$, usually performed in each iteration of the fitting algorithm. Even for a modestly large number of points ($\approx 50,000$ or greater), the computational demands become prohibitive for a modern computer and preclude inference from GP models.

1.2 Dissertation Outline and Contributions

In the next two chapters, we develop conjugate Bayesian frameworks for analyzing univariate and multivariate spatial data. In Chapter 2, we propose a conjugate latent Nearest-Neighbor Gaussian Process (NNGP) model, which uses analytically tractable posterior distributions to obtain posterior inferences, including the large dimensional latent process. NNGP is a well-defined spatial process providing finite-dimensional densities with sparse precision matrices. We implement a conjugate gradient algorithm to accelerate the process of posterior sampling in a constructed conjugate NNGP based model. Our algorithm has an unparalleled speed in the implementation for large-scale spatial data. A key emphasis is on implementation within very standard (modest) computing environments (e.g., a standard desktop or laptop) using easily available statistical software packages. We provide the detailed algorithm and illustrate the model with a data analysis of sea surface temperature collected over 2.5 million observed locations on a standard laptop.

Joint modeling of spatially-oriented dependent variables are commonplace in the environmental sciences,where scientists seek to estimate the relationships among a set of environmental outcomes accounting for dependence among these outcomes and the spatial dependence for each outcome. In Chapter 3, we extend the work in Chapter 2 from univariate modeling to multivariate modeling. We utilize Matrix-Normal Inverse-Wishart(MNIW) prior to construct conjugate Bayesian framework for modeling large-scale multivariate spatial data sets. We develop algorithms for obtaining posterior inference from the proposed Bayesian framework, and show that the proposed model can be incorporated with a family of scalable spatial modeling methodologies. Moreover, we discuss differences between modeling the multivariate response itself as a spatial process and that of modeling a latent process. We illustrate the computational and inferential benefits of these models using simulation studies and real data analyses for a vegetation indices dataset with observed locations numbering over three million.

There is, by now, a burgeoning literature on scalable models for an individual process, but methods for scalable multivariate spatial process are limited in comparison. In Chapter 4, we propose frameworks to extend scalable modeling strategies for a single process into multivariate process cases. We pursue modeling methodologies that are more flexible and versatile than a conjugate Bayesian framework as established in Chapter 2 and 3. We propose a Bayesian linear model of coregionalization (BLMC) and a block update MCMC algorithm. The modeling approach we develop in Chapter 4 enriches the popular linear models of coregionalization [Bourgault and Marcotte, 1991, Wackernagel, 2003, Gelfand et al., 2004, Chiles and Delfiner, 2009, Genton and Kleiber, 2015] using a Matrix-Normal prior to model the linear transformation on latent spatial processes. A key contribution here is that we provide a fully model-based enhancement for misaligned data, where not all responses are recorded over the same set of locations. We further expand this contribution by using the Matrix-Normal family to model the loading matrix in spatial factor models. In the latter context, our current contribution can be seen as enhancements to earlier contributions by Lopes et al. [2008], Ren and Banerjee [2013] and Taylor-Rodriguez et al. [2019]. We illustrate the computational and inferential benefits of our algorithms over competing methods using simulation studies and real data analyses for a vegetation index dataset observed in over a million locations.

The MCMC chains for Bayesian spatial models are often observed to be unstable for some of the hyper-parameters in Bayesian spatial model fitting. This invokes the research interests in the theoretical studies on the asymptotic properties of the parameter estimators in spatial modeling. Chapter 5 is devoted to investigating the asymptotic properties of the estimators in spatial analysis. We formally establish results on the identifiability and consistency of the nugget in spatial models based upon the Gaussian process within the framework of in-fill asymptotics, i.e. the sample size increases within a sampling domain that is bounded. Our main contribution in this chapter is that we extend results in fixed domain asymptotics for spatial models without the nugget effect, the presence of noise process, into the scenario with the nugget effect. More specifically, we establish the identifiability of parameters in the Matérn covariance function and the consistency of their maximum likelihood estimators in the presence of discontinuities due to the nugget. We present simulation studies to demonstrate the role of the identifiable quantities in spatial interpolation.

CHAPTER 2

Massive Spatial Data Analysis On Modest Computing Environments

2.1 Introduction

Rapidly increasing usage and growing capabilities of Geographic Information Systems (GIS) have spawned considerable research in modeling and analyzing spatial datasets in diverse disciplines including, but not limited to, environmental sciences, economics, biometry and so on [see, e.g., Gelfand et al., 2010, Cressie and Wikle, 2011, Banerjee et al., 2014]. A substantial literature exists on methodologies for massive spatial datasets [see, e.g., Banerjee, 2017, Heaton et al., 2017, and references therein]. Some are more amenable than others to the hierarchical setup in (1.1.1). Even within the hierarchical paradigm, there is already a burgeoning literature on massively scalable spatial process models. There are two pressing issues facing the practicing spatial analyst. The first is to analyze massive amounts of spatial data on "modest" computing environments such as standard desktop or laptop architectures. The second pressing issue is that of *full inference* that subsumes parameter estimation, spatial prediction of the outcome, and estimation of the underlying latent process. Yet the size of the datasets easily exceed the CPU memory available for computing, which means that we need to rely upon statistical models that will enable analysis with the available memory.

Some scalable processes such as the multi-resolution predictive process models proposed by Katzfuss [2017] or the nearest-neighbor Gaussian process (NNGP) models by Datta et al. [2016a] can be programmed in modest computing environments to estimate parameters and predict outcomes, but not necessarily infer on the latent process efficiently. Katzfuss [2017] does not address this, while Datta et al. [2016a] and Datta et al. [2016b] implement high-dimensional Gibbs sampling algorithms that had to be run for several iterations on a high-performance computing environment to yield adequate convergence due to high autocorrelations. Other approaches such as Gaussian Markov random field (GMRF) approximations to spatial processes [Rue et al., 2009, Lindgren et al., 2011] use Integrated Nested Laplace Approximations (INLA) for computing the marginal distribution of the process at given locations. These approximations can be implemented on standard environments for a variety of spatial models using the `R-INLA` software (`www.r-inla.org`). This is computationally more promising than MCMC, but is still an iterative procedure requiring convergence assessment. Its performance is yet to be demonstrated for analyzing massive spatial data with millions of spatial locations on modest computing environments.

This Chapter outlines strategies for achieving fully model-based Bayesian inference including parameter estimation, response surface predictions and interpolation of the latent spatial process for massive spatial datasets on modest computing environments. To achieve this goal, we need a massively scalable spatial process that will be able to estimate (1.1.1) by obviating the memory obstacles. Here, there are a few choices that are well-suited for (1.1.1) all of whom seem to be competitive based upon the recent "contest" paper by Heaton et al. [2017], but we opt for the sparsity-inducing Nearest-neighbor Gaussian process (NNGP) primarily because of its ease of use and also because of its easier accessibility through the `spNNGP` package available from `cran.r-project.org/web/packages/spNNGP` (see Section 2.2).

In fact, Finley et al. [2019] outlines several strategies for estimating NNGP models, including a conjugate response NNGP model and a collapsed NNGP model. The conjugate response NNGP model can provide exact inference without requiring MCMC and has been demonstrated to effectively fit a dataset with approximately 5 million locations in a matter of seconds on a Linux workstation. However, the response model does not accommodate the latent process and, hence, is restrictive in its inferential capabilities compared to (1.1.1). The collapsed NNGP model, on the other hand, is embedded within MCMC algorithms and is able to provide the posterior inference of the latent process. It can exploit permutation-based sparse Cholesky methods, but the approach requires specialized libraries and can still be too expensive for massive datasets in the order of 10^6 locations for standard computing environments. We briefly introduce the conjugate response NNGP model in section 2.3.2, and the discussion of the collapsed NNGP model can be found in section 2.2.2. Our contribution lies in casting the latent process models of Datta et al. [2016a] within a conjugate Bayesian framework for exact inference so as to avoid MCMC while being able to achieve full Bayesian inference including estimation of the latent process. We propose a conjugate latent NNGP model that exploits conjugacy in conjunction with cross-validatory estimation of a small set of process parameters, and the model formulation and computations will not require loading large data objects into memory at any point, allowing fitting for massive datasets in the order of 10^6 on computer environments like standard desktop or laptop architecture. The details of this Bayesian formulation and the algorithms for their effective implementation constitute the novelty of this Chapter.

The remainder of the Chapter evolves as follows. Section 2.2 provides a brief review of nearest-neighbor Gaussian process and NNGP based models. Section 2.3 develops the conjugate NNGP based models, emphasizing the conjugate latent NNGP model, and devises algorithms for practical implementation. A simulation study is presented in Section 2.4 for discussing the performance of the proposed models, while an analysis on sea surface temperature with over 2.5 million locations is conducted in Section 2.5. Finally, we conclude with some discussion in Section 2.6.

2.2 The nearest-neighbor Gaussian process

The computational burden in GP models arises from the $n \times n$ covariance matrix $\mathbf{C}_\theta(\mathcal{S}, \mathcal{S})$, where $\mathcal{S} = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n\}$ is the set of observed locations. The (i, j)-th element of $\mathbf{C}_\theta(\mathcal{S}, \mathcal{S})$ is the value of a spatial covariance function evaluated at locations $\mathbf{s}_i$ and $\mathbf{s}_j$. Spatial covariance functions in general do not produce exploitable structures in the resulting matrix. One effective approach to achieve efficient computations is to replace $\mathbf{C}_\theta(\mathcal{S}, \mathcal{S})$ with an approximate $\tilde{\mathbf{C}}_\theta(\mathcal{S},\mathcal{S})$ such that the inverse of $\tilde{\mathbf{C}}_\theta(\mathcal{S},\mathcal{S})$ is sparse. There are multiple options, but notable among them are approximations based upon Gaussian Markov random fields or GMRFs [see, e.g., Rue and Held, 2005, Rue et al., 2009] that yield computationally efficient sparse representations. An alternative approach exploits an idea familiar in graphical models or Bayesian networks [see, e.g., Lauritzen, 1996, Bishop, 2006, Murphy, 2012] that has also been exploited by Vecchia [1988], Stein et al. [2004] and Stroud et al. [2017] to construct composite likelihoods for inference. Datta et al. [2016a,b] extended this idea to construct a Nearest Neighbor Gaussian Process (NNGP) for modeling large spatial data. NNGP is a well defined Gaussian Process that yields finite dimensional Gaussian densities with sparse precision matrices. It delivers massive scalability both in terms of parameter estimation and spatial prediction or "kriging".

2.2.1 Response NNGP model

Consider modeling a point-referenced outcome as a partial realization of a Gaussian process, $\{y(\mathbf{s}) : \mathbf{s} \in \mathcal{D}\} \sim \mathrm{GP}(m_\theta(\mathbf{s}), C_\theta(\cdot,\cdot))$ on a spatial domain $\mathcal{D} \in \mathbb{R}^d$. The mean and covariance functions are assumed to be determined by one or more parameters in a set θ. The finite-dimensional distribution for the $n \times 1$ vector $\mathbf{y}(\mathcal{S})$ with elements $y(\mathbf{s}_i)$ is multivariate normal with mean $\mathbf{m}_\theta(\mathcal{S})$ and covariance matrix $\mathbf{C}_\theta(\mathcal{S},\mathcal{S})$. As a directed acyclic graph (DAG) [Bishop, 2006], the joint density is $p(\mathbf{y}(\mathbf{S})) = \displaystyle\prod_{i=1}^{n} p(y(\mathbf{s}_i) \,|\, \mathbf{y}(\mathrm{Pa}[\mathbf{s}_i])$, where $\mathrm{Pa}[\mathbf{s}_1]$ is the empty set and $\mathrm{Pa}[\mathbf{s}_i] = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_{i-1}\}$ for $i = 2, 3, \ldots, n-1$ is the set of parent nodes with directed edges to $\mathbf{s}_i$. Vecchia [1988] suggested approximating the multivariate normal likelihood by shrinking $\mathrm{Pa}[\mathbf{s}_i]$ from the set of all nodes preceding $\mathbf{s}_i$ to a much smaller subset of locations preceding $\mathbf{s}_i$ that are among the m (a fixed small number) nearest neighbors of $\mathbf{s}_i$ based upon their Euclidean distance. Datta et al. [2016a] extended that notion to arbitrary points in the domain by defining

$$\mathrm{Pa}[\mathbf{s}] = \begin{cases} \text{empty set} & \text{if } \mathbf{s} = \mathbf{s}_1 \ , \\ \{\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_{i-1}\} & \text{if } \mathbf{s} \in \mathcal{S} \text{ and } i = 1, 2, \dots, m \ , \\ m \text{ closest points to } \mathbf{s} \text{ among } \{\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_{i-1}\} & \text{if } \mathbf{s} \in \mathcal{S} \text{ and } i > m \ , \\ m \text{ closest points to } \mathbf{s} \text{ among } \mathcal{S} & \text{if } \mathbf{s} \notin \mathcal{S} \ . \end{cases}$$

for any arbitrary point $\mathbf{s}$ in the domain, where m is the fixed number of nearest neighbors. This results in another multivariate Gaussian density

$$p(\mathbf{y}(\mathcal{S})) = \mathrm{N}(\mathbf{y}(\mathcal{S}) \,|\, \mathbf{m}_\theta(\mathcal{S}), \mathbf{C}_\theta(\mathcal{S}, \mathcal{S})) \approx \mathrm{N}(\mathbf{y}(\mathcal{S}) \,|\, \mathbf{m}_\theta(\mathcal{S}), \tilde{\mathbf{C}}_\theta(\mathcal{S}, \mathcal{S})) \ , \tag{2.2.1}$$

where $\tilde{\mathbf{C}}_\theta(\mathcal{S}, \mathcal{S})^{-1} = (\mathbf{I} - \mathbf{A}(\mathcal{S}))^\top \mathbf{D}(\mathcal{S})^{-1}(\mathbf{I} - \mathbf{A}(\mathcal{S}))$ is sparse, $\mathbf{A}(\mathcal{S})$ is sparse and strictly lower triangular with $\mathbf{A}(\mathcal{S})(i,i) = 0$ for $i = 1, 2, \dots, n$ and at most m non-zero entries in each row, and $\mathbf{D}(\mathcal{S})$ is diagonal whose elements are the conditional variances $\mathrm{var}\{y(\mathbf{s}_i) \,|\, \mathbf{y}(\mathrm{Pa}[\mathbf{s}_i])\}$ based upon the full GP model, i.e., $\mathbf{D}(\mathcal{S})(1,1) = C_\theta(\mathbf{s}_1, \mathbf{s}_1)$ and $\mathbf{D}(\mathcal{S})(i,i) = C_\theta(\mathbf{s}_i, \mathbf{s}_i) - \mathbf{C}_\theta(\mathbf{s}_i, \mathrm{Pa}[\mathbf{s}_i])\mathbf{C}_\theta(\mathrm{Pa}[\mathbf{s}_i], \mathrm{Pa}[\mathbf{s}_i])^{-1}\mathbf{C}_\theta(\mathrm{Pa}[\mathbf{s}_i], \mathbf{s}_i)$ for $i = 2, \dots, n$. Turning to the structure of $\mathbf{A}(\mathcal{S})$, all its elements are completely determined from $\mathbf{C}_\theta(\mathcal{S}, \mathcal{S})$. Its first row, i.e., $\mathbf{A}(\mathcal{S})(1,)$ has all zeroes. For the $i+1$-th row, the nonzero entries appear in the positions indexed by $\mathrm{Pa}[\mathbf{s}_{i+1}]$ and are obtained as row vectors,

$$\mathbf{A}(\mathcal{S})(i+1, \mathrm{Pa}[\mathbf{s}_{i+1}]) = \mathbf{C}_\theta(\mathbf{s}_{i+1}, \mathrm{Pa}[\mathbf{s}_{i+1}])\mathbf{C}_\theta(\mathrm{Pa}[\mathbf{s}_{i+1}], \mathrm{Pa}[\mathbf{s}_{i+1}])^{-1} \ .$$

The nonzero entries in each row of $\mathbf{A}(\mathcal{S})$ are precisely the "kriging" weights of $y(\mathbf{s}_i)$ based upon the values of $y(\mathbf{s})$ at neighboring locations, i.e., $\mathrm{Pa}[\mathbf{s}_i]$ [Chiles and Delfiner, 2009]. The $\tilde{\mathbf{C}}_\theta(\mathcal{S}, \mathcal{S})$, constructed as above, is called an NNGP approximation to $\mathbf{C}_\theta(\mathcal{S}, \mathcal{S})$.

With the above definition of $\mathrm{Pa}[\mathbf{s}]$, we can express the partial realizations of an NNGP as a linear model. Let $\mathcal{S}$ be the set of the n observed locations as defined earlier (and n is assumed to be large) and let $\mathcal{U} = \{\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_{n'}\}$ be a set of n' arbitrary locations where we wish to predict $y(\mathbf{s})$. Then,

$$\underbrace{\begin{bmatrix} \mathbf{y}(\mathcal{S}) \\ \mathbf{y}(\mathcal{U}) \end{bmatrix}}_{\mathbf{y}} = \underbrace{\begin{bmatrix} \mathbf{m}_\theta(\mathcal{S}) \\ \mathbf{m}_\theta(\mathcal{U}) \end{bmatrix}}_{\mathbf{m}_\theta} + \underbrace{\begin{bmatrix} \mathbf{A}(\mathcal{S}) \\ \mathbf{A}(\mathcal{U}) \end{bmatrix}}_{\mathbf{A}} (\mathbf{y}(\mathcal{S}) - \mathbf{m}_\theta(\mathcal{S})) + \boldsymbol{\eta} \ , \tag{2.2.2}$$

where $\boldsymbol{\eta} \sim \mathrm{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \mathbf{D}(\mathcal{S}) & \mathbf{0} \\ \mathbf{0} & \mathbf{D}(\mathcal{U}) \end{bmatrix}\right)$, $\mathbf{D}(\mathcal{U})$ is $n' \times n'$ diagonal and $\mathbf{A}(\mathcal{U})$ is sparse $n' \times n$ formed by extending the definitions of $\mathbf{D}(\mathcal{S})$ and $\mathbf{A}(\mathcal{S})$ as

$$\begin{aligned} \mathbf{D}(\mathcal{U})(i,i) &= C_\theta(\mathbf{u}_i, \mathbf{u}_i) - \mathbf{C}_\theta(\mathbf{u}_i, \mathrm{Pa}[\mathbf{u}_i])\mathbf{C}_\theta(\mathrm{Pa}[\mathbf{u}_i], \mathrm{Pa}[\mathbf{u}_i])^{-1}\mathbf{C}_\theta(\mathrm{Pa}[\mathbf{u}_i], \mathbf{u}_i)\ , \\ \mathbf{A}(\mathcal{U})(i, \mathrm{Pa}[\mathbf{u}_i]) &= \mathbf{C}_\theta(\mathbf{u}_i, \mathrm{Pa}[\mathbf{u}_i])\mathbf{C}_\theta(\mathrm{Pa}[\mathbf{u}_i], \mathrm{Pa}[\mathbf{u}_i])^{-1}\ . \end{aligned} \tag{2.2.3}$$

Each row of $\mathbf{A}(\mathcal{U})$ has exactly m nonzero entries corresponding to the column indices in $\mathrm{Pa}[\mathbf{u}_i]$. The above structure implies that $y(\mathbf{s})$ and $y(\mathbf{s}')$ are conditionally independent for any two points $\mathbf{s}$ and $\mathbf{s}'$ outside of $\mathcal{S}$, given $\mathbf{y}(\mathcal{S})$. The parameters θ will be estimated from the data $\mathbf{y}(\mathcal{S})$ and predictions will be carried out using the conditional distribution of $\mathbf{y}(\mathcal{U})$ given $\mathbf{y}(\mathcal{S})$. In a Bayesian setting, θ will be sampled from its posterior distribution $p(\theta \,|\, \mathbf{y}(\mathcal{S}))$,

$$p(\theta) \times \left(\prod_{i=1}^{n} \frac{1}{\sqrt{\mathbf{D}(\mathcal{S})(i,i)}}\right) \times \exp\left\{-\frac{1}{2}\mathbf{z}_\theta(\mathcal{S})^\top(\mathbf{I} - \mathbf{A}(\mathcal{S})^\top)\mathbf{D}(\mathcal{S})^{-1}(\mathbf{I} - \mathbf{A}(\mathcal{S}))\mathbf{z}_\theta(\mathcal{S})\right\}\ , \tag{2.2.4}$$

where $\mathbf{z}_\theta(\mathcal{S}) = \mathbf{y}(\mathcal{S}) - \mathbf{m}_\theta(\mathcal{S})$ and $p(\theta)$ is the prior distribution for θ.

Consider a specific example with the covariance function $C_\theta(\mathbf{s}, \mathbf{s}') = \sigma^2 \exp(-\phi\|\mathbf{s} - \mathbf{s}'\|) + \tau^2\delta_{\mathbf{s}=\mathbf{s}'}$, where $\delta_{\mathbf{s}=\mathbf{s}'}$ is equal to one if $\mathbf{s} = \mathbf{s}'$ and 0 otherwise, and $m_\theta(\mathbf{s}) = \mathbf{x}(\mathbf{s})^\top\boldsymbol{\beta}$ is a linear regression with spatial predictors $\mathbf{x}(\mathbf{s})$ and corresponding slope vector $\boldsymbol{\beta}$. Then $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \sigma^2, \phi, \tau^2\}$ and one choice of priors could be

$$p(\boldsymbol{\theta}) \propto \mathrm{U}(\phi \,|\, a_\phi, b_\phi) \times \mathrm{IG}(\sigma^2 \,|\, a_\sigma, b_\sigma) \times \mathrm{IG}(\tau^2 \,|\, a_\tau, b_\tau) \times \mathrm{N}(\boldsymbol{\beta} \,|\, \boldsymbol{\mu}_\beta, \mathbf{V}_\beta)\ ,$$

where we are using standard notations for the above distributions as, e.g., in Gelman et al. [2013]. The parameter space for this model is not high-dimensional and MCMC algorithms such as Gibbs sampling in conjunction with random-walk Metropolis (RWM) or Hamiltonian Monte Carlo (HMC) can be easily implemented. Other approximate algorithms such as Variational Bayes or INLA can also be used.

Once the parameter estimates (i.e., posterior samples) are obtained from (2.2.4) we can carry out predictive inference for $\mathbf{y}(\mathcal{U})$ from the posterior predictive distribution

$$p(\mathbf{y}(\mathcal{U}) \,|\, \mathbf{y}(\mathcal{S})) = \int p(\mathbf{y}(\mathcal{U}) \,|\, \mathbf{y}(\mathcal{S}), \boldsymbol{\theta}) p(\boldsymbol{\theta} \,|\, \mathbf{y}(\mathcal{S})) d\boldsymbol{\theta} = \mathrm{E}_{\boldsymbol{\theta} \,|\, \mathbf{y}(\mathcal{S})} \left[\mathrm{N}(\mathbf{y}(\mathcal{U}) \,|\, \boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathcal{U}|\cdot), \mathbf{D}_{\mathcal{U}}) \right] \ , \tag{2.2.5}$$

where $p(\mathbf{y}(\mathcal{U}) \,|\, \mathbf{y}(\mathcal{S}), \boldsymbol{\theta})$ is an n'-dimensional multivariate normal distribution with mean $\boldsymbol{\mu}_{\boldsymbol{\theta}}(\mathcal{U} \,|\, \cdot) = \mathbf{m}_{\boldsymbol{\theta}}(\mathcal{U}) + \mathbf{A}(\mathcal{U})(\mathbf{y}(\mathcal{S}) - \mathbf{m}_{\boldsymbol{\theta}}(\mathcal{S}))$ and conditional covariance matrix $\mathbf{D}_{\mathcal{U}}$. Since $\mathbf{D}(\mathcal{U})$ is diagonal, it is easy to sample from $p(\mathbf{y}(\mathcal{U}) \,|\, \mathbf{y}(\mathcal{S}), \boldsymbol{\theta})$. For each $\boldsymbol{\theta}$ sampled from (2.2.4), we sample an n'-dimensional vector $\mathbf{y}(\mathcal{U})$ from $p(\mathbf{y}(\mathcal{U}) \,|\, \mathbf{y}(\mathcal{S}), \boldsymbol{\theta})$. The resulting $\mathbf{y}(\mathcal{U})$'s are samples from (2.2.5). The NNGP exploits the conditional independence between the elements of $\mathbf{y}(\mathcal{U})$, given $\mathbf{y}(\mathcal{S})$ and $\boldsymbol{\theta}$, to achieve efficient posterior predictive sampling for $\mathbf{y}(\mathcal{U})$. This assumption of conditional independence is not restrictive as the samples from (2.2.5) are not independent. In fact, the marginal covariance matrix of $\mathbf{y}(\mathcal{U})$, given $\boldsymbol{\theta}$ only, is $\mathbf{A}(\mathcal{U})\tilde{\mathbf{C}}_{\boldsymbol{\theta}}(\mathcal{S}, \mathcal{S})\mathbf{A}(\mathcal{U})^{\top} + \mathbf{D}(\mathcal{U})$, which is clearly not diagonal.

2.2.2 Latent NNGP model

Rather than model the outcome as an NNGP, as was done for the response model in the preceding subsection, one could use the NNGP as a prior for the latent process [Datta et al., 2016a]. In fact, as discussed in Section 4 of [Datta et al., 2016a], the response model does not strictly follow the paradigm in (1.1.1) and it is not necessarily possible to carry out inference on a latent or residual spatial process after accounting for the mean.

A more general setting envisions a spatial regression model at any location s

$$y(\mathbf{s}) = m_{\boldsymbol{\theta}}(\mathbf{s}) + w(\mathbf{s}) + \epsilon(\mathbf{s}) \ , \quad \epsilon(\mathbf{s}) \overset{iid}{\sim} \mathrm{N}(0, \tau^2) \ , \tag{2.2.6}$$

where, usually, $m_{\boldsymbol{\theta}}(\mathbf{s}) = \mathbf{x}(\mathbf{s})^{\top}\boldsymbol{\beta}$ and $w(\mathbf{s})$ is a latent spatial process capturing spatial dependence. Using definitions analogous to Section 2.2.1, we assume $\{w(\mathbf{s}) : \mathbf{s} \in \mathcal{D}\} \sim \mathrm{NNGP}(0, \tilde{C}_{\boldsymbol{\theta}}(\cdot, \cdot))$, which means that for any $\mathcal{S}$ and $\mathcal{U}$, as constructed in (2.2.2), $\mathbf{w} \equiv \mathbf{w}(\mathcal{S} \cup \mathcal{U})$ will have a zero-centered multivariate normal law with covariance matrix $(\mathbf{I} - \mathbf{A})^{-1}\mathbf{D}(\mathbf{I} -$

$\mathbf{A})^{-\top}$. The posterior distribution to be sampled from is now given by

$$p(\boldsymbol{\theta}) \times \mathrm{N}(\mathbf{w} \,|\, \mathbf{0}, \tilde{\mathbf{C}}_{\boldsymbol{\theta}}(\mathcal{S}, \mathcal{S})) \times \prod_{i=1}^{n} \mathrm{N}(y(\mathbf{s}_i) \,|\, m_{\boldsymbol{\theta}}(\mathbf{s}_i) + w(\mathbf{s}_i), \tau^2) \; . \tag{2.2.7}$$

It is easier to sample from (2.2.4) than from (2.2.7) since the parameter space in the latter includes the high-dimensional random vector $\mathbf{w}$ in addition to $\boldsymbol{\theta}$. One option is to integrate out $\mathbf{w}$ from (2.2.7) which yields the posterior

$$p(\boldsymbol{\theta}) \times (\det(\tilde{\mathbf{C}}_{\boldsymbol{\theta}}(\mathcal{S}, \mathcal{S}) + \tau^2 \mathbf{I}_n))^{-\frac{1}{2}} \times \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} \mathbf{z}_{\boldsymbol{\theta}}(\mathcal{S})^\top \left(\tilde{\mathbf{C}}_{\boldsymbol{\theta}}(\mathcal{S}, \mathcal{S}) + \tau^2 \mathbf{I}_n \right)^{-1} \mathbf{z}_{\boldsymbol{\theta}}(\mathcal{S}) \right\} \; , \tag{2.2.8}$$

where $\det(\mathbf{A})$ is the determinant of matrix $\mathbf{A}$, $p(\boldsymbol{\theta})$ and $\mathbf{z}_{\boldsymbol{\theta}}(\mathcal{S})$ are as defined for (2.2.4). The parameter space has collapsed from $\{\boldsymbol{\theta}, \mathbf{w}\}$ to $\boldsymbol{\theta}$, so (2.2.8) is called the collapsed version of (2.2.7). Efficient computations for obtaining (2.2.8) requires a sparse-Cholesky decomposition for the large matrix $\left(\tilde{\mathbf{C}}_{\boldsymbol{\theta}}(\mathcal{S}, \mathcal{S})^{-1} + \tau^{-2} \mathbf{I} \right)$. This step can be complicated and expensive. To exacerbate the matter further, full Bayesian inference requires calculating the likelihood (2.2.8) in each MCMC iteration as described in the algorithm of the "collapsed" model in Section 2.1 of Finley et al. [2019]. To avoid such expenses, we turn to conjugate models in the next section.

2.3 Conjugate Bayesian model

The response NNGP and latent NNGP models outlined in Sections 2.1 and 2.2, respectively, will still require iterative simulation methods such as MCMC for full Bayesian inference. Conjugate models, i.e., using conjugate priors, can provide exact Bayesian inference by exploiting analytic forms for the posterior distributions. While some specific assumptions are needed, these models are much faster to implement even for massive datasets. Here we develop conjugate NNGP models using the tractable Normal Inverse-Gamma (NIG) family of conjugate priors. We formulate a *conjugate response model* (also formulated in Finley et al. [2019] and is available in the `spNNGP` package from `cran.r-project.org/web/packages/spNNGP`) and a new conjugate latent NNGP model. These are conjugate versions of the models described in Sections 2.1 and 2.2. We especially focus on the conjugate latent NNGP model and show how it can exploit sparsity by sampling from latent spatial processes over massive numbers of locations efficiently using a conjugate gradient algorithm for solving large sparse systems.

2.3.1 The NIG conjugate prior family

Let the spatial linear regression model be specified as

$$\mathbf{y}(\mathcal{S}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{w}(\mathcal{S}) + \boldsymbol{\epsilon}(\mathcal{S}) \tag{2.3.1}$$

where $\mathbf{y}(\mathcal{S})$, $\mathbf{w}(\mathcal{S})$ and $\boldsymbol{\epsilon}(\mathcal{S})$ are the realization of the corresponding processes defined in (2.2.6) over the n observed locations $\mathcal{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$, $\mathbf{X}$ is the $n \times p$ matrix of regressors with i-th row being a $1 \times p$ vector of regressors, $x(\mathbf{s}_i)^\top$ at location $\mathbf{s}_i \in \mathcal{S}$. Henceforth, we suppress the dependence of $\mathbf{y}$, $\mathbf{w}$, $\boldsymbol{\epsilon}$ and their covariance matrix on $\mathcal{S}$ when this will not lead to confusion. Assume that $\mathbf{w} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{C})$, $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \delta^2\sigma^2\mathbf{I}_n)$, where $\mathbf{C}$ and $\delta^2 = \frac{\tau^2}{\sigma^2}$ are known. Let $\boldsymbol{\gamma}^\top = [\boldsymbol{\beta}^\top, \mathbf{w}^\top]$, $\boldsymbol{\mu}_\gamma^\top = [\boldsymbol{\mu}_{\boldsymbol{\beta}}^\top, \mathbf{O}]$ and $\mathbf{V}_\gamma = \begin{bmatrix} \mathbf{V}_{\boldsymbol{\beta}} & \mathbf{O} \\ \mathbf{O} & \mathbf{C} \end{bmatrix}$. The Normal-Inverse-Gamma (NIG) density yields a convenient conjugate prior,

$$p(\boldsymbol{\gamma}, \sigma^2) = \mathrm{NIG}(\boldsymbol{\gamma}, \sigma^2 \,|\, \boldsymbol{\mu}_\gamma, \mathbf{V}_\gamma, a, b) = \mathrm{N}(\boldsymbol{\gamma} \,|\, \boldsymbol{\mu}_\gamma, \sigma^2\mathbf{V}_\gamma) \times \mathrm{IG}(\sigma^2 \,|\, a, b) \; . \tag{2.3.2}$$

The posterior distribution of the parameters, up to proportionality, is

$$p(\boldsymbol{\gamma}, \sigma^2 \,|\, \mathbf{y}) \propto \mathrm{NIG}(\boldsymbol{\gamma},\sigma^2 \,|\, \boldsymbol{\mu}_\gamma, \mathbf{V}_\gamma, a_\sigma, b_\sigma) \times \mathrm{N}(\mathbf{y} \,|\, [\mathbf{X} : \mathbf{I}_n]\boldsymbol{\gamma}, \delta^2\sigma^2\mathbf{I}_n) \; . \tag{2.3.3}$$

The joint posterior distribution is of the form $\mathrm{NIG}(\boldsymbol{\mu}^*, \mathbf{V}^*, a^*, b^*)$, where

$$\begin{aligned}
\mathbf{y}^* \quad &= \tfrac{1}{\delta}\mathbf{y} \; , \quad \mathbf{X}^* = \left[\tfrac{1}{\delta}\mathbf{X}, \tfrac{1}{\delta}\mathbf{I}_n\right] \; , \\
\boldsymbol{\mu}^* \quad &= [\mathbf{V}_\gamma^{-1} + \mathbf{X}^{*\top}\mathbf{X}^*]^{-1}(\mathbf{V}_\gamma^{-1}\boldsymbol{\mu}_\gamma + \mathbf{X}^{*\top}\mathbf{y}^*) \; , \\
\mathbf{V}^* \quad &= [\mathbf{V}_\gamma^{-1} + \mathbf{X}^{*\top}\mathbf{X}^*]^{-1} \; , \\
a^* \quad &= a_\sigma + \tfrac{n}{2} \; , \\
b^* \quad &= b_\sigma + \tfrac{1}{2}[\boldsymbol{\mu}_\gamma^\top\mathbf{V}_\gamma\boldsymbol{\mu}_\gamma + \mathbf{y}^{*\top}\mathbf{y}^* - \boldsymbol{\mu}^{*\top}\mathbf{V}^{*-1}\boldsymbol{\mu}^*] \; .
\end{aligned} \tag{2.3.4}$$

The prior of the regression coefficients $\boldsymbol{\beta}$ is formulated as $\mathrm{N}(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \mathbf{V}_{\boldsymbol{\beta}})$. The above model, however, also allows improper priors for $\boldsymbol{\beta}$. When assigning improper priors for $\boldsymbol{\beta}$, the precision matrix of the prior of $\boldsymbol{\gamma}$ in (2.3.4) becomes $\mathbf{V}_{\gamma}^{-1} = \begin{bmatrix} \mathbf{O} & \mathbf{O} \\ \mathbf{O} & \mathbf{C}^{-1} \end{bmatrix}$, showing that no information from $\boldsymbol{\beta}$'s prior contributes to the posterior distribution, and we can assume $\boldsymbol{\mu}_{\gamma}^{\top} = [\mathbf{O}, \mathbf{O}]$ in (2.3.4). The marginal posterior distribution of σ^2 follows an $\mathrm{IG}(a^*, b^*)$ and the marginal posterior distribution of $\boldsymbol{\gamma}$ can be identified as a multivariate t-distribution with mean $\boldsymbol{\mu}^*$, variance $\frac{b^*}{a^*}\mathbf{V}^*$ and degree of freedom $2a^*$(i.e. MVS-$t_{2a^*}(\boldsymbol{\mu}^*, \frac{b^*}{a^*}\mathbf{V}^*)$). Exact Bayesian inference is carried out by sampling directly from the joint posterior density: we sample σ^2 from $\mathrm{IG}(a^*, b^*)$ and then, for each sampled σ^2, we draw $\boldsymbol{\gamma}$ from its conditional posterior density $\mathrm{N}(\boldsymbol{\mu}^*, \sigma^2\mathbf{V}^*)$. This yields posterior samples from (2.3.3). Furthermore, note that once the posterior samples of σ^2 are obtained, we can obtain samples from $p(\tau^2 \,|\, \mathbf{y})$ by simply multiplying the sampled σ^2s with δ^2. Thus, posterior samples are obtained without recourse to MCMC or other iterative algorithms.

2.3.2 Conjugate response NNGP model

Finley et al. [2019] formulated a conjugate NNGP model for the response model described in Section 2.1. This is formed by integrating out $\mathbf{w}(\mathcal{S})$ from (2.3.1) and applying an NNGP approximation to the marginal covariance matrix of $\mathbf{y}(\mathcal{S})$. The model can be cast as a conjugate Bayesian linear regression model

$$p(\boldsymbol{\beta}, \sigma^2 \,|\, \mathbf{y}) \propto \mathrm{NIG}(\boldsymbol{\beta}, \sigma^2 \,|\, \boldsymbol{\mu}_{\boldsymbol{\beta}}, \mathbf{V}_{\boldsymbol{\beta}}, a_\sigma, b_\sigma) \times \mathrm{N}(\mathbf{y} \,|\, \mathbf{X}\boldsymbol{\beta}, \sigma^2\tilde{\mathbf{K}}) \;, \tag{2.3.5}$$

where $\tilde{\mathbf{K}}$ is the NNGP approximation of $\mathbf{K} = \mathbf{C} + \delta^2\mathbf{I}$, $\mathbf{C}$ and δ^2 are as described in Section 2.3.1. Also, $\tilde{\mathbf{K}}^{-1} = \sigma^2(\mathbf{I} - \mathbf{A}(\mathcal{S})^{\top})\mathbf{D}(\mathcal{S})^{-1}(\mathbf{I} - \mathbf{A}(\mathcal{S}))$ with $\mathbf{A}(\mathcal{S})$ and $\mathbf{D}(\mathcal{S})$ as described in Section 2.2.1. We will refer to (2.3.5) as the conjugate response NNGP model. Note that this model can estimate $\{\boldsymbol{\beta}, \sigma^2\}$ and also impute the outcome at unknown locations, but does not permit inference on the latent process $w(\cdot)$. The reason why a conjugate response NNGP model cannot provide inference on the latent process is that the construction of the response NNGP will not guarantee the existence of a well-defined latent process. It is pointed out in Section 4 of [Datta et al., 2016a] that the eigenvalue of $\tilde{\mathbf{K}}$ may be less than δ^2, consequently the covariance matrix of the posterior distribution of $\mathbf{w}$ need not be positive definite for every proper δ^2, $\boldsymbol{\mu}_{\boldsymbol{\beta}}$ and $\mathbf{V}_{\boldsymbol{\beta}}$. We address this shortcoming with a new conjugate latent NNGP model in the next section.

2.3.3 Conjugate latent NNGP model

The conjugate models in Section 2.3.1 works for any covariance matrix $\mathbf{C}$. Here, we derive a conjugate latent NNGP model that will subsume inference on $w(\cdot)$. We rewrite the covariance matrix $\tilde{\mathbf{C}}_{\boldsymbol{\theta}}(\mathcal{S},\mathcal{S})$ in section 2.2.2 for $\mathbf{w}(\mathcal{S})$ as $\sigma^2\tilde{\mathbf{M}}_\phi$ with fixed parameter ϕ. Note that $\tilde{\mathbf{M}}_\phi$ is the NNGP approximation of the dense matrix $\mathbf{M}$, where $\mathbf{C} = \sigma^2\mathbf{M}$. Specifically, $\tilde{\mathbf{M}}_\phi^{-1} = (\mathbf{I}-\mathbf{A}_{\mathbf{M}})^\top\mathbf{D}_{\mathbf{M}}^{-1}(\mathbf{I}-\mathbf{A}_{\mathbf{M}})$, where $\mathbf{A}_{\mathbf{M}}$ and $\mathbf{D}_{\mathbf{M}}$ depend only on ϕ. We recast the model as

$$\underbrace{\begin{bmatrix} \frac{1}{\delta}\mathbf{y} \\ \mathbf{L}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}} \\ \mathbf{0} \end{bmatrix}}_{\mathbf{y}_*} = \underbrace{\begin{bmatrix} \frac{1}{\delta}\mathbf{X} & \frac{1}{\delta}\mathbf{I}_n \\ \mathbf{L}_{\boldsymbol{\beta}}^{-1} & \mathbf{O} \\ \mathbf{O} & \mathbf{D}_{\mathbf{M}}^{-\frac{1}{2}}(\mathbf{I}-\mathbf{A}_{\mathbf{M}}) \end{bmatrix}}_{\mathbf{X}_*} \underbrace{\begin{bmatrix}\boldsymbol{\beta} \\ \mathbf{w}\end{bmatrix}}_{\gamma} + \underbrace{\begin{bmatrix}\boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 \\ \boldsymbol{\eta}_3\end{bmatrix}}_{\boldsymbol{\eta}} \tag{2.3.6}$$

where $\mathbf{L}_{\boldsymbol{\beta}}$ is the Cholesky decomposition of the $p \times p$ matrix $\mathbf{V}_{\boldsymbol{\beta}}$, and $\boldsymbol{\eta} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I}_{2n+p})$. The joint posterior distribution of $\boldsymbol{\gamma}$ and σ^2 follows an NIG distribution

$$p(\boldsymbol{\gamma}, \sigma^2 \,|\, \mathbf{y}) = \mathrm{NIG}(\boldsymbol{\gamma}, \sigma^2 \,|\, \hat{\boldsymbol{\gamma}}, (\mathbf{X}_*^\top\mathbf{X}_*)^{-1}, a_*, b_*) \tag{2.3.7}$$

where $\hat{\boldsymbol{\gamma}} = (\mathbf{X}_*^\top\mathbf{X}_*)^{-1}\mathbf{X}_*^\top\mathbf{y}_*$, $a_* = a_\sigma + \dfrac{n}{2}$ and $b_* = b_\sigma + \dfrac{1}{2}(\mathbf{y}_* - \mathbf{X}_*\hat{\boldsymbol{\gamma}})^\top(\mathbf{y}_* - \mathbf{X}_*\hat{\boldsymbol{\gamma}})$. Evaluating the posterior mean of $\boldsymbol{\gamma}$ involves solving $\mathbf{X}_*^\top\mathbf{X}_*\hat{\boldsymbol{\gamma}} = \mathbf{X}_*^\top\mathbf{y}_*$, which requires $\mathcal{O}(\frac{1}{3}(n+p)^3)$ flops. However, when $p \ll n$, the structure of $\mathbf{X}_*$ ensures low storage complexity. Also, $\mathbf{X}_*^\top\mathbf{X}_* =$

$$\begin{bmatrix} \frac{1}{\delta^2}\mathbf{X}^\top\mathbf{X} + \mathbf{L}_{\boldsymbol{\beta}}^{-\top}\mathbf{L}_{\boldsymbol{\beta}}^{-1} & \frac{1}{\delta^2}\mathbf{X}^\top \\ \frac{1}{\delta^2}\mathbf{X} & \frac{1}{\delta^2}\mathbf{I}_n + (\mathbf{I}_n - \mathbf{A}_{\mathbf{M}})^\top\mathbf{D}_{\mathbf{M}}^{-1}(\mathbf{I}_n - \mathbf{A}_{\mathbf{M}}) \end{bmatrix} \tag{2.3.8}$$

Since $(\mathbf{I}_n - \mathbf{A}_{\mathbf{M}})$ has less than $n(m+1)$ nonzero elements and each of its row has at most $m+1$ nonzero elements, the storage of the $n \times n$ matrix $(\mathbf{I}_n - \mathbf{A}_{\mathbf{M}})^\top\mathbf{D}_{\mathbf{M}}^{-1}(\mathbf{I}_n - \mathbf{A}_{\mathbf{M}})$ is less than $n(m+1)^2$, and the computational complexity is less than $nm + n(m+1)^2$.

This sparsity in $\mathbf{X}_*^\top \mathbf{X}_*$ can be exploited by a conjugate gradient (CG) method [see, e.g., Golub and Van Loan, 2012]. CG is an iterative method for solving $\mathbf{Ax} = \mathbf{b}$ when $\mathbf{A}$ is a symmetric positive definite matrix. The underlying idea is to recognize that a solution of the linear system $\mathbf{Ax} = \mathbf{b}$ minimizes the quadratic function $\phi(\mathbf{x}) = \frac{1}{2}\mathbf{x}^\top \mathbf{Ax} - \mathbf{x}^\top \mathbf{b}$. CG is an iterative procedure that generates a sequence of approximate solutions $\{\mathbf{x}_k\}_{k=1,2,\ldots}$ that converges to $\mathbf{x} = \mathbf{A}^{-1}\mathbf{b}$ in at most n iterations. Briefly, the procedure starts with an initial value $\mathbf{x}_0$ and setting $\mathbf{r}_0 = \mathbf{b} - \mathbf{Ax}_0$ and $\mathbf{q}_0 = \mathbf{r}_0$. Then, at the $k+1$-th iteration we compute the following three quantities for each $k = 0, 1, 2, \ldots$: (i) $\mathbf{x}_{k+1} = \mathbf{x}_k + \frac{\|\mathbf{r}_k\|^2}{\mathbf{q}_{k+1}^\top \mathbf{A} \mathbf{q}_{k+1}} \mathbf{q}_{k+1}$; (ii) $\mathbf{r}_{k+1} = \mathbf{b} - \mathbf{Ax}_{k+1}$; and (iii) $\mathbf{q}_{k+1} = \mathbf{r}_{k+1} + \left(\frac{\|\mathbf{r}_{k+1}\|}{\|\mathbf{r}_k\|}\right)^2 \mathbf{q}_k$. The matrix $\mathbf{A}$ is involved only in matrix-vector multiplications. Due to the sparsity of $\mathbf{A}$, the computational cost per iteration is $\mathcal{O}(n)$ flops. The sparsity in $\mathbf{A}$ also implies that CG is more memory efficient than direct methods such as the Cholesky decomposition. A sufficiently good approximation is often obtained in iterations much less than n [Banerjee and Roy, 2014], hence the performance of the conjugate gradient algorithm will be competitive when n is large. This enables posterior sampling of the latent process $\mathbf{w}(\mathcal{S})$ in high-dimensional settings. The algorithm for sampling $\{\gamma, \sigma^2\}$ from (2.3.7) using the conjugate gradient method is given below.

Algorithm 2.1: Sample $\{\boldsymbol{\gamma}, \sigma^2\}$ from conjugate latent NNGP model

1. Fixing ϕ and δ^2, obtain $\mathbf{L}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}}$ and $\mathbf{L}_{\boldsymbol{\beta}}^{-1}$:
   - Compute a Cholesky decomposition of $\mathbf{V}_{\boldsymbol{\beta}}$ to get $\mathbf{L}_{\boldsymbol{\beta}}$ $\quad \mathcal{O}(p^3)$
   - Compute $\mathbf{L}_{\boldsymbol{\beta}}^{-1}$ and $\mathbf{L}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}}$ $\quad \mathcal{O}(p^2)$
2. Obtain the posterior mean for $\boldsymbol{\gamma}$:
   - Construct $\mathbf{A}_\mathbf{M}$ and $\mathbf{D}_\mathbf{M}$ as described, for example, in Finley et al. [2019] $\quad \mathcal{O}(nm^3)$
   - Construct $\mathbf{X}_*$ and $\mathbf{Y}_*$ from (2.3.6) $\quad \mathcal{O}(nm)$
   - Calculate $\mathbf{X}_*^\top \mathbf{X}_*$ and $\mathbf{X}_*^\top \mathbf{y}_*$ $\quad \mathcal{O}(n(m+1)^2)$
   - Use conjugate gradient to solve $\mathbf{X}_*^\top \mathbf{X}_* \hat{\boldsymbol{\gamma}} = \mathbf{X}_*^\top \mathbf{y}_*$
3. Obtain posterior samples of σ^2
   - Calculate a_* and b_* as given below (2.3.7) $\quad \mathcal{O}(n(m+4+p))$
   - Sample σ^2 from $\mathrm{IG}(a_*, b_*)$
4. Obtain posterior samples of γ

- Generate $\mathbf{u} \sim \mathrm{N}(\mathbf{0}, \sigma^2 \mathbf{I}_{2n+p})$
- Calculate $\mathbf{v}$ by solving $\mathbf{X}_*^\top \mathbf{X}_* \mathbf{v} = \mathbf{X}_*^\top \mathbf{u}$ using conjugate gradient
- Obtain $\boldsymbol{\gamma} = \hat{\boldsymbol{\gamma}} + \mathbf{v}$ $\mathcal{O}(n)$

It is readily seen that the $\mathbf{v}$ in step 4 follows a Gaussian distribution with variance $\sigma^2(\mathbf{X}_*^\top \mathbf{X}_*)^{-1}$. Note that Algorithm 2.1 implements the conjugate gradient method for an $n+p$-dimensional linear system in steps 2 and 4. Since $\mathbf{X}_*$ and y_* depend only on $\{\phi, \delta^2\}$, the linear equation in step 2 only need to be solved once for each choice of $\{\phi, \delta^2\}$.

The main contribution of the conjugate gradient method lies in obtaining the posterior estimator $\hat{\boldsymbol{\gamma}}$ (step 2) and generating samples from a high dimensional Gaussian distribution (step 4). It is worth pointing out that the conjugate gradient method does not easily produce the determinant of a large matrix. Hence, a sparse Cholesky decomposition is still unavoidable for the collapsed NNGP model formulated in equation (2.2.8), where det $\Big(\tilde{\mathbf{C}}_{\boldsymbol{\theta}}(\mathcal{S}, \mathcal{S}) + \tau^2 \mathbf{I}\Big)$ changes with the hyper-parameters $\boldsymbol{\theta}$.

2.3.4 Posterior predictive inference for conjugate latent NNGP

We extend the predictive inference for the response NNGP model in Section 2.1 to the conjugate latent NNGP model. Assume $\mathbf{w}(\mathcal{U})$ and $\mathbf{y}(\mathcal{U})$ are the realization of the latent process and the response process over the n' locations $\mathcal{U} = \{\mathbf{u}_1, \ldots, \mathbf{u}_{n'}\}$ where we wish to predict. let $\mathbf{C}_{\boldsymbol{\theta}}(\cdot, \cdot)$ be the covariance function for the latent process $w(\mathbf{s})$ in (2.2.6), $\mathrm{Pa}[\mathbf{u}_i]$ be the nearest neighbors of ith location in $\mathcal{U}$ as defined in section 2.1. Define $\mathbf{A}_{\mathcal{U}} = \mathbf{A}(\mathcal{U})$ and $\mathbf{D}_{\mathcal{U}} = \frac{1}{\sigma^2}\mathbf{D}(\mathcal{U})$ where $\mathbf{A}(\mathcal{U})$ and $\mathbf{D}(\mathcal{U})$ are constructed by (2.2.3). Here, σ^2 refers to the variance of the latent process $w(\mathbf{s})$, and $\mathbf{A}_{\mathcal{U}}$ and $\mathbf{D}_{\mathcal{U}}$ are defined in the way that they only depend on fixed parameter ϕ. According to the definition of NNGP process over the whole domain given in section 2, the joint distribution of $\mathbf{w}(\mathcal{U})$ and $\boldsymbol{\gamma}, \sigma^2$ given $\mathbf{y}(\mathcal{S})$ follows:

$$p(\mathbf{w}(\mathcal{U}), \boldsymbol{\gamma}, \sigma^2 \,|\, \mathbf{y}(\mathcal{S})) = \mathrm{N}(\mathbf{w}(\mathcal{U}) \,|\, [\mathbf{O} : \mathbf{A}_{\mathcal{U}}]\boldsymbol{\gamma}, \sigma^2 \mathbf{D}_{\mathcal{U}}) \times \mathrm{NIG}(\boldsymbol{\gamma}, \sigma^2 \,|\, \hat{\boldsymbol{\gamma}}, (\mathbf{X}_*^\top \mathbf{X}_*)^{-1}, a_*, b_*) \tag{2.3.9}$$

Marginalizing the joint distribution (2.3.9) over $\boldsymbol{\gamma}$ and σ^2, the posterior distribution of $\mathbf{w}(\mathcal{U})$ can be identified as a multivariate t-distribution:

$$\mathbf{w}(\mathcal{U}) \,|\, \mathbf{y}(\mathcal{S}) \sim \text{MVS-}t_{2a_*}\left(\boldsymbol{\mu}_{wu}, \frac{b_*}{a_*}\mathbf{V}_{wu}\right) \tag{2.3.10}$$

where

$$\boldsymbol{\mu}_{wu} = [\mathbf{O} : \mathbf{A}_{\mathcal{U}}]\hat{\gamma} \ , \ \mathbf{V}_{wu} = [\mathbf{O} : \mathbf{A}_{\mathcal{U}}](\mathbf{X}_*^\top \mathbf{X}_*)^{-1} \begin{bmatrix} \mathbf{O} \\ \mathbf{A}_{\mathcal{U}}^\top \end{bmatrix} + \mathbf{D}_{\mathcal{U}} \ .$$

It is straightforward to see that the joint posterior distribution of $\{\mathbf{y}(\mathcal{U}), \mathbf{w}(\mathcal{U}), \boldsymbol{\gamma}, \sigma^2\}$ is

$$p(\mathbf{y}(\mathcal{U}), \mathbf{w}(\mathcal{U}), \boldsymbol{\gamma}, \sigma^2 \,|\, \mathbf{y}(\mathcal{S})) = \text{N}(\mathbf{y}(\mathcal{U}) \,|\, \mathbf{X}(\mathcal{U})\boldsymbol{\beta} + \mathbf{w}(\mathcal{U}), \sigma^2\delta^2\mathbf{I}_{n'}) \times p(\mathbf{w}(\mathcal{U}), \boldsymbol{\gamma}, \sigma^2 \,|\, \mathbf{y}(\mathcal{S})) \ , \tag{2.3.11}$$

which is the product of the conditional distribution of $\mathbf{y}(\mathcal{U})$ from the spatial linear regression model (2.2.6) and the posterior distribution (2.3.9). It can be shown that the posterior distribution of the predictive process $\mathbf{y}(\mathcal{U})$ and σ^2 follows an NIG after marginalizing out $\boldsymbol{\gamma}$ and $\mathbf{w}(\mathcal{U})$, and the posterior distribution of $\mathbf{y}(\mathcal{U})$ follows a multivariate t-distribution:

$$\mathbf{y}(\mathcal{U}) \,|\, \mathbf{y}(\mathcal{S}) \sim \text{MVS-}t_{2a_*}\left(\boldsymbol{\mu}_{yu}, \frac{b_*}{a_*}\mathbf{V}_{yu}\right) \tag{2.3.12}$$

where

$$\boldsymbol{\mu}_{yu} = [\mathbf{X}(\mathcal{U}) : \mathbf{A}_{\mathcal{U}}]\hat{\gamma} \ \text{ and } \ \mathbf{V}_{yu} = [\mathbf{X}(\mathcal{U}) : \mathbf{A}_{\mathcal{U}}](\mathbf{X}_*^\top \mathbf{X}_*)^{-1} \begin{bmatrix} \mathbf{X}(\mathcal{U})^\top \\ \mathbf{A}_{\mathcal{U}}^\top \end{bmatrix} + \delta^2\mathbf{I}_{n'} + \mathbf{D}_{\mathcal{U}} \ .$$

Sampling $\mathbf{w}(\mathcal{U})$ $\mathbf{y}(\mathcal{U})$ from their posterior distribution requires taking Cholesky decomposition of matrix $\mathbf{V}_{wu}$ and $\mathbf{V}_{yu}$. Since the matrix $(\mathbf{X}_*^\top \mathbf{X}_*)^{-1}$ is involved in the calculation, the required computation power is expensive and the calculation quickly become forbidden when the number of locations to predict is large. Rather than direct sampling, we recommend using a two stage sampling method based on the joint distribution (2.3.9) and (2.3.11) in this subsection. First, obtain the posterior samples $\{\boldsymbol{\gamma}^{(l)}, \sigma^{2(l)}\}_{l=1}^L$. Then generate the posterior samples of $\mathbf{w}(\mathcal{U})$ through $\mathbf{w}(\mathcal{U})^{(l)} \sim \text{N}([\mathbf{O} : \mathbf{A}_{\mathcal{U}}]\boldsymbol{\gamma}^{(l)}, \sigma^{2(l)}\mathbf{D}_{\mathcal{U}})$ for $l = 1, \ldots, L$. Finally use $\mathbf{y}(\mathcal{U})^{(l)} \sim \text{N}(\mathbf{X}(\mathcal{U})\boldsymbol{\beta}^{(l)} + \mathbf{w}(\mathcal{U})^{(l)}, \delta^2\sigma^{2(l)})$ to generate the posterior samples of $\mathbf{y}(\mathcal{U})$.

2.3.5 Inference of hyper-parameters

Algorithm 2.1 provides the exact posterior sampling of the process parameters after specifying ϕ and δ^2. This motivates us to estimate all the process parameters by first obtaining the inference of a small set of parameters ϕ and δ^2, then implementing Algorithm 2.1 to sample $\{\boldsymbol{\gamma}, \sigma^2\}$. When we fix ϕ and δ^2 at a point estimator (i.e. $\arg\max\{p(\phi, \delta^2 \,|\, \mathbf{y})\}$), the conjugate latent NNGP model becomes a special case of fitting latent NNGP model with Empirical Bayes method.

Here we propose a K-folder cross-validation algorithm for picking a point estimate of $\{\phi, \delta^2\}$ of the conjugate Latent NNGP. We first split the data randomly into K folds and denote the k-th folder of the observed locations $\mathcal{S}[k]$, whereas $\mathcal{S}[-k]$ denotes the observed locations without $\mathcal{S}[k]$. Then we fit the predictive mean $E[\mathbf{y}(\mathcal{S}[k]) \,|\, \mathbf{y}(\mathcal{S}[-k])]$ by the posterior distribution given in (2.3.12). We use the Root Mean Square Predictive Error (RMSPE)(Yeniay and Goktas [2002]) to select ϕ and δ^2 from a gird of candidate values. The initial candidates for $\{\phi, \delta^2\}$ comes from a coarse grid. The range of the grid is decided based on interpretation of the hyper-parameters. Specifically, the spatial decay ϕ describes how the spatial correlation decreases as the distance between two locations increases. Define $\text{maxdist}(\mathcal{S}) := \max_{\mathbf{s},\mathbf{t}\in\mathcal{S}}\{d(\mathbf{s}, \mathbf{t})\}$ where $d(\mathbf{s}, \mathbf{t})$ is the distance between location $\mathbf{s}$ and $\mathbf{t}$. The lower bound of the candidate value of ϕ is set at $3/\text{maxdist}(\mathcal{S})$, which indicates that the spatial correlation drops below 0.05 when the distance reaches $\text{maxdist}(S)$. The upper bound can be initially set as 100 times of the lower bound $300/\text{maxdist}(S)$. For δ^2, we need to use reasonable assumptions on the variance components. A suggested wide range for δ^2 can be $[0.001, 1000]$, which accommodates one variance component substantially dominating the other in either direction. The prior information from the related studies of the data as well as the estimators from the variogram also provide the candidate value of $\{\phi, \delta^2\}$. Functions like `variofit` in the R package `geoR` [Ribeiro Jr and Diggle, 2012] can provide empirical estimates for $\{\phi, \delta^2\}$ from an empirical variogram. After initial fitting, we can shrink the range and refine the grid of the candidate values for more precise estimators. Algorithm 2.2 describes K-fold cross-validation for choosing ϕ, δ^2 in the conjugate latent NNGP model.

Algorithm 2.2: Cross-validation of tuning ϕ, δ^2 for conjugate latent NNGP model

1. Split the data into K folds, and build neighbor index.
    - Build nearest neighbors for $\mathcal{S}[-k]$
    - Find the collection of nearest neighbor set for $\mathcal{S}[k]$ among $\mathcal{S}[-k]$.
2. Fix ϕ and δ^2, Obtain the posterior mean for $\boldsymbol{\gamma}_k = \{\boldsymbol{\beta}, \mathbf{w}(\mathcal{S}[-k])\}$ after removing the k^{th} fold of the data:
    - Use step 1-2 in Algorithm 2.1 to obtain $\hat{\boldsymbol{\gamma}}_k$
3. Predicting posterior means of $\mathbf{y}(\mathcal{S}[k])$
    - Construct matrix $\mathbf{A}_{\mathcal{U}}$ for $\mathcal{S}[k]$
    - According to (2.3.12), the predicted posterior mean follows $\hat{\mathbf{y}}(\mathcal{S}[k]) = E[\mathbf{y}(\mathcal{S}[k]) \,|\, \mathbf{y}(\mathcal{S}[-k])] = [\mathbf{X}(\mathcal{U}) : \mathbf{A}_{\mathcal{U}}]\hat{\boldsymbol{\gamma}}$
4. Root Mean Square Predictive Error (RMSPE) over K folds
    - Initialize $e = 0$

```
for ($k$ in $1 : K$)
    for ($\mathbf{s}_i$ in $\mathcal{S}[k]$)
        $e = e + (y(\mathbf{s}_i) - \hat{y}(\mathbf{s}_i))^2$
```

5. Cross validation for choosing ϕ and δ^2
    - Repeat steps (2) - (4) for all candidate values of ϕ and δ^2
    - Choose ϕ_0 and δ_0 as the value that minimizes the average RMSPE

The main computational burden lies in step 1 in Algorithm 2.2. However, step 1 serves as a pre-calculation for the whole cross-validation since it only need to be calculated for once. We recommend using a KD-tree algorithm provided in R package `spNNGP` [Finley et al., 2017] to build the nearest neighbor matrics. Step 2 dominates the computational requirement in Algorithm 2.2 after the pre-calculation, which calls Algorithm 1 for k times for each choice of $\{\phi, \delta^2\}$.

An alternative approach for choosing point estimates of $\{\phi, \delta^2\}$ is to carry out the cross-validation with the conjugate response NNGP model in (2.3.5). The practical advantage here is that the function `spConjNNGP` within the `spNNGP` package in `R` can be used to carry out the cross-validation. The algorithm behind `spConjNNGP` is exactly linear in n and highly efficient in its implementation. Empirical studies reveal that the response NNGP model and the latent NNGP model provide similar optimal choices for $\{\phi, \delta^2\}$ when using the K- folder cross-validation.

2.4 Simulation Study

We use a simulation study in this section to discuss the performance of the aforementioned models in Sections 2.2 and 2.3. Algorithm 2.1 were programmed in R which calls the `Rstan` environment [Stan Development Team, 2016] for building matrix $\mathbf{A_M}$ and $\mathbf{D_M}$. The conjugate gradient solver for sparse linear systems was implemented through `RcppEigen` [Bates and Eddelbuettel, 2013], which calls a Jacobi preconditioner [see, e.g., page 653 in Golub and Van Loan, 2012] by default. We provide a brief discussion on preconditioned conjugate gradient algorithms in Section 2.6. The nearest-neighbor sets were built using the `spConjNNGP` function in the `spNNGP` package. All simulations were conducted on a OS High sierra system (version 10.13.4) with 16GB RAM and one 3.1 GHz Intel-Core i7 processors.

2.4.1 Univariate simulation study

We generated data using the spatial regression model in (2.2.6) over a set of $n = 1200$ spatial locations within a unit square. The true values of the parameters generating the data are supplied in Table 2.1. The size of the data set was kept moderate to permit comparisons with the expensive full GP models. The model had an intercept and a single predictor $x(\mathbf{s})$ generated from a standard normal distribution. An exponential covariance function was used to generate the data.

Candidate models for fitting the data included full Gaussian process based model (labeled as full GP in Table 2.1), a latent NNGP model with $m = 10$ neighbors and a conjugate latent NNGP model with $m = 10$ neighbors. These models were trained using $n = 1000$ of the 1200 observed locations. And the remaining 200 observations were withheld to assess predictive performance. The full Gaussian process based model was implemented with

Table 2.1: Simulation study summary table: posterior mean (2.5%, 97.5%) percentiles

	True	Full GP	NNGP	Conj LNNGP
β_0	1	1.07(0.72, 1.42)	1.10 (0.74, 1.43)	1.06 (0.76, 1.46)
β_1	-5	-4.97 (-5.02, -4.91)	-4.97 (-5.02, -4.91)	-4.97 (-5.02, -4.91)
σ^2	2	1.94 (1.63, 2.42)	1.95 (1.63, 2.41)	1.94 (1.77, 2.12)
τ^2	0.2	0.14 (0.07, 0.23)	0.15 (0.06, 0.24)	0.17 (0.16, 0.19)
ϕ	16	19.00 (13.92, 23.66)	18.53 (14.12, 24.17)	17.65
KL-D	–	4.45(1.16, 9.95)	5.13(1.66, 11.39)	3.58(1.27, 8.56)
MSE($\mathbf{w}$)	–	297.45(231.62, 444.79)	303.38(228.18, 429.54)	313.28 (258.96, 483.75)
RMSPE	–	0.94	0.94	0.94
time(s)	–	2499 + 23147	109.5	12 + 0.6

function *spLM* in R package *spBayes* [Finley et al., 2007]. The latent NNGP model was conducted with function *spNNGP* in R package *spNNGP*. The fixed parameters$\{\phi, \delta^2\}$ for the conjugate latent NNGP model were picked through the k-th folder cross-validation algorithm (Algorithm 2.2). And the choice from `spConjNNGP` coincide with the cross-validation for the conjugate latent NNGP model.

The intercept and slope parameters $\boldsymbol{\beta}$ were assigned improper flat priors. The spatial decay ϕ was modeled using a fairly wide uniform prior $\mathrm{U}(2.2, 220)$. We use Inverse-Gamma priors $\mathrm{IG}(2, b)$ (mean b) for the nugget (τ^2) and the partial sill (σ^2) in order to compare the conjugate Bayesian models with other models. The shape parameter was fixed at 2 and the scale parameter was set from the empirical estimate provided by the variogram using the `geoR` package [Ribeiro Jr and Diggle, 2012]. The parameter estimates and performance metrics are provided in Table 2.1.

The summaries for the full Gaussian process based model and the latent NNGP model were based on 1 MCMC chain with $20,000$ iterations. The number of iterations was taken to be large enough to guarantee the convergence of the MCMC chains. We took the first half of the MCMC chains as burn-in. The inference from the conjugate latent NNGP model were based on 300 samples. 300 samples is sufficient for the conjugate latent NNGP model since the conjugate model provides independent samples from the exact posterior distribution. We don't need extra memory for burn-in, and the samples from the conjugate model are more efficient than that from MCMC algorithms.

All models were assessed by the Kullback-Leibler divergence (labeled KL-D; Gneiting and Raftery [2007]) and the out-of-sample root mean squared prediction error (RMSPE) (Yeniay and Goktas [2002]). The KL-D between true distribution Q and fitted distribution $P_{\boldsymbol{\theta}}$ is measured by:

$$d(P_{\boldsymbol{\theta}}, Q) = \frac{1}{2}\{tr(\boldsymbol{\Sigma}_P^{-1}\boldsymbol{\Sigma}_Q) - \log\det(\boldsymbol{\Sigma}_P^{-1}\boldsymbol{\Sigma}_Q) + (\boldsymbol{\mu}_P - \boldsymbol{\mu}_Q)'\boldsymbol{\Sigma}_P^{-1}(\boldsymbol{\mu}_P - \boldsymbol{\mu}_Q) - n\} \tag{2.4.1}$$

where P_θ and Q define Gaussian distributions on $\mathbb{R}^n$ with mean vectors μ_P and μ_Q, respectively, and covariance matrices Σ_P and Σ_Q, respectively. The KL-D in Table 2.1 are on the collapsed space $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \sigma^2, \tau^2, \phi\}$. We estimated the KL-D by the empirical estimator:

$$E_{\boldsymbol{\theta}\,|\,\mathbf{y}(\mathcal{S})}(d(P_{\boldsymbol{\theta}}, Q)) \approx \frac{1}{L}\sum_{i=1}^{L} d(P_{\boldsymbol{\theta}_{(i)}}, Q)\ , \tag{2.4.2}$$

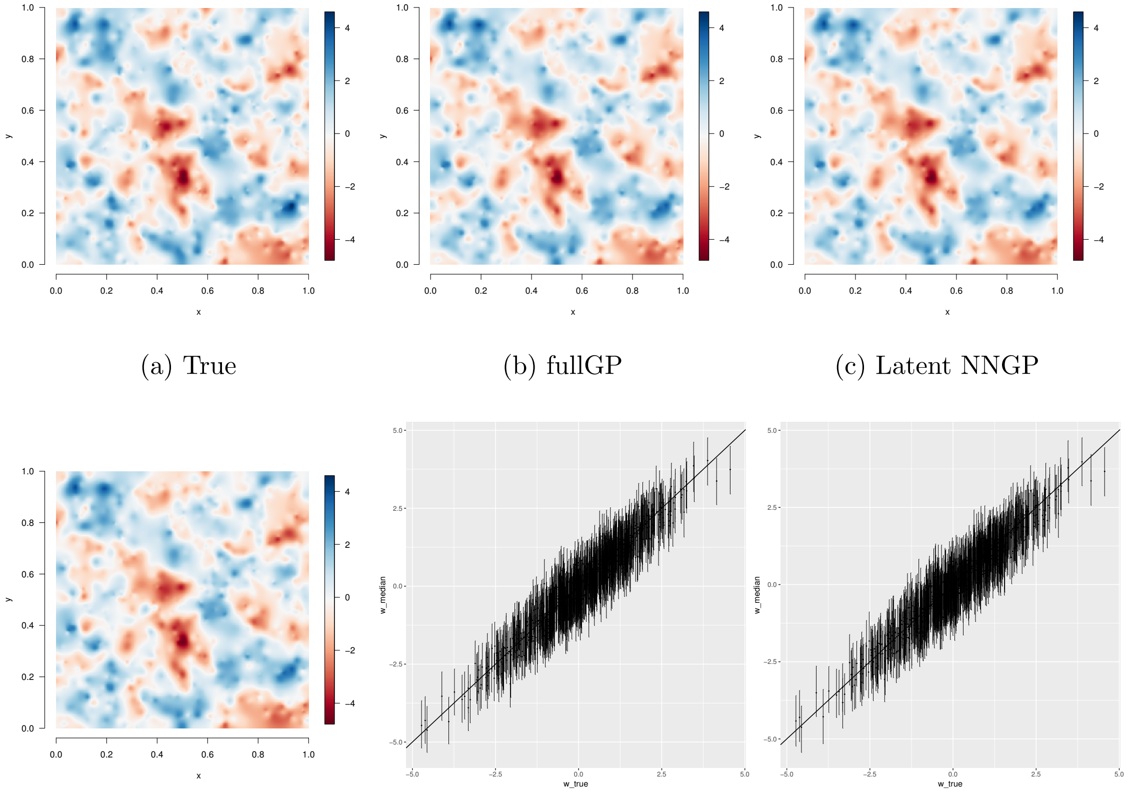
where $\boldsymbol{\theta}_{(i)}, i = 1, \ldots, L$ are L samples from the posterior distribution of $\boldsymbol{\theta}$. We also present the 95% credible intervals for $d(P_{\boldsymbol{\theta}}, Q)$ in Table 2.1. The predicted outcome at any withheld location s_0 was estimated as

$$\hat{y}(\mathbf{s}_0) = E[\tilde{y}(\mathbf{s}_0)\,|\,\mathbf{y}(\mathcal{S})] \approx \frac{1}{L}\sum_{i=1}^{L} \tilde{y}_{\boldsymbol{\theta}_{(i)}}(\mathbf{s}_0)\ , \tag{2.4.3}$$

where $\tilde{y}_{\boldsymbol{\theta}_{(i)}}(\mathbf{s}_0) \sim p(y(\mathbf{s}_0)\,|\,\mathbf{y}(\mathcal{S}), \boldsymbol{\theta}_{(i)})$ and $p(\cdot\,|\,\mathbf{y}(\mathcal{S}), \boldsymbol{\theta}_{(i)})$ is the likelihood for the respective model. These were used to calculate the RMSPE using the 200 hold-out values. We randomly picked 300 out of the 10000 samples from the post burn-in MCMC chains for calculating the KL-D and RMSPE. The $y(\mathbf{s}_0)$ for full Gaussian process based and the latent NNGP model are sampled by function *spPredict*. For the purpose of assessing the performance of recovering spatial latent process, we also report the Mean Squared Error (MSE) with respect to the true values of the spatial latent process (MSE($\mathbf{w}$)) over the observed locations in the simulation. The KL-D, MSE($\mathbf{w}$) and RMSPE metrics reveal that the NNGP provides a highly competitive alternative to the full Gaussian process based model.

Table 2.1 lists the parameter estimates and performance metrics for the candidate models. The posterior inference of the regression coefficients $\boldsymbol{\beta}$ are close for all three models. While the posterior estimates of $\{\sigma^2.\tau^2, \phi\}$ are similar for full Gaussian process based model and latent NNGP model but, somewhat expectedly, different from the conjugate latent NNGP model. The 95% confidence interval for σ^2 and τ^2 are narrower since we fix the parameter ϕ, δ^2. The KL-Ds on the parameter space $\{\mathbf{w}, \boldsymbol{\beta}, \tau^2\}$ show that the conjugate latent NNGP provides reliable inference for the latent process and the regression coefficients. The same RMSPE across all three models also support that conjugate latent NNGP is comparable with full Gaussian process based model in prediction. The latent NNGP model is 200 times faster than the full Gaussian process based model, while the conjugate latent NNGP model use one tenth of the time required for the latent NNGP model to obtain similar inference on the regression coefficients and latent process. Notice that the time for the sampling of the 300 samples after fixing the parameter ϕ and δ^2 in the conjugate latent NNGP model is less than one second. And the conjugate latent NNGP spare the effect of testing the tuning parameters in MCMC algorithm. Based on KL-D and RMSPE, the conjugate latent NNGP models emerge as highly competitive alternatives to latent NNGP models for prediction and inference on the latent process.

Figure 2.1 shows interpolated surfaces from the simulation example: 2.1(a) shows an interpolated map of the "true" spatial latent process w, 2.1(b)–(d) are maps of the posterior means of the latent process using a full GP model, a latent NNGP model and a conjugate latent NNGP model, respectively. Figure 2.1(e)–(f) present the 95% confidence intervals for bw from a full GP model and a conjugate latent NNGP model. The recovered spatial residual surfaces are almost indistinguishable, and are comparable to the true interpolated surface of $w(\mathbf{s})$. Notice that the posterior mean of $\mathbf{w}$ of the conjugate latent NNGP model can be theoretically calculated by the $\hat{\boldsymbol{\gamma}}$ in (2.3.7). Thus the posterior samples of the latent process $\mathbf{w}$ is only required for measuring uncertainty. Figure 2.1f provides the 95% confidence interval for all latent process $\mathbf{w}$ from the conjugate latent NNGP model. There are 955 out of 1000 95% confidence intervals successfully include the true value. This is comparable to



(a) True

(b) fullGP

(c) Latent NNGP

(d) Conjugate Latent NNGP

(e) CIs of w from fullGP

(f) CIs of w from Conjugate Latent NNGP

Figure 2.1: Interpolated maps of (a) the true generated surface, the posterior means of the spatial latent process $w(\mathbf{s})$ for (b) the full Gaussian Process (Full GP), (c) the latent NNGP and (d) the conjugate latent NNGP. The 95% confidence intervals for w from (e) the full GP and (f) the conjugate latent NNGP. The models in (c), and (d) were all fit using $m = 10$ nearest neighbors.

the full Gaussian process based model (fig 2.1e) which has 946 out of 1000 95% confidence intervals covering the true value.

2.5 Sea surface temperature analysis

Global warming continues to be an ongoing concern among scientists. In order to develop conceptual and predictive global models, NASA monitors temperature and other atmospheric properties of the Earth regularly by two Moderate Resolution Imaging Spectroradiometer (MODIS) instruments in Aqua and Terra platforms. There is an extensive global satellite-based database processed and maintained by NASA. Details of the data can be found in `http://modis-atmos.gsfc.nasa.gov/index.html`. In particular, inferring on processes generating sea surface temperatures (SST) are of interest to atmospheric scientists studying exchange of heat, momentum, and water vapor between the atmosphere and ocean. Our aforementioned development will enable scientists to analyze large spatially-indexed datasets using a Bayesian geostatistical model easily implementable on modest computing platforms.

Model-based inference is obtained rapidly using the conjugate latent NNGP model and, based on simulation studies, will be practically indistinguishable from MCMC-based output from more general NNGP specification. The dataset we analyze here consists of 2,827,252 spatially indexed observations of sea surface temperature (SST) collected between June 18-26, 2017, the data covers the ocean from longitude -140° to °0 and from latitude 0° to 60°. Among the 2,827,252 observations, $n = 2,544,527$ (90%) were used for model fitting and the rest were withheld to assess predictive performance of the candidate models. Figure 2.3a depicts an interpolated map of the observed SST records over training locations. The temperatures are color-coded from shades of blue indicating lower temperatures, primarily seen in the higher latitudes, to shades of red indicating high temperatures. The missing data are colored by yellow and the gray part refers to land. To understand trends across the coordinates, we used sinusoidally projected coordinates (scaled to 1000km units) as explanatory variables. The sinusoidal projection is a popular equal-area projection [see, e.g., Banerjee

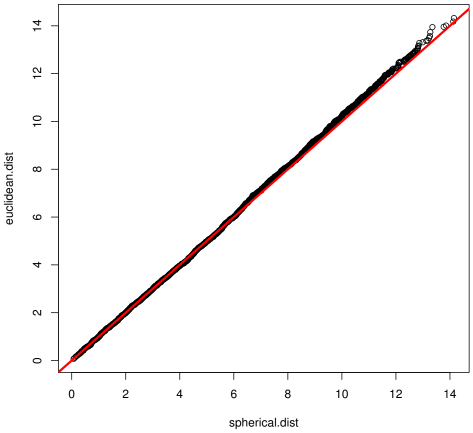
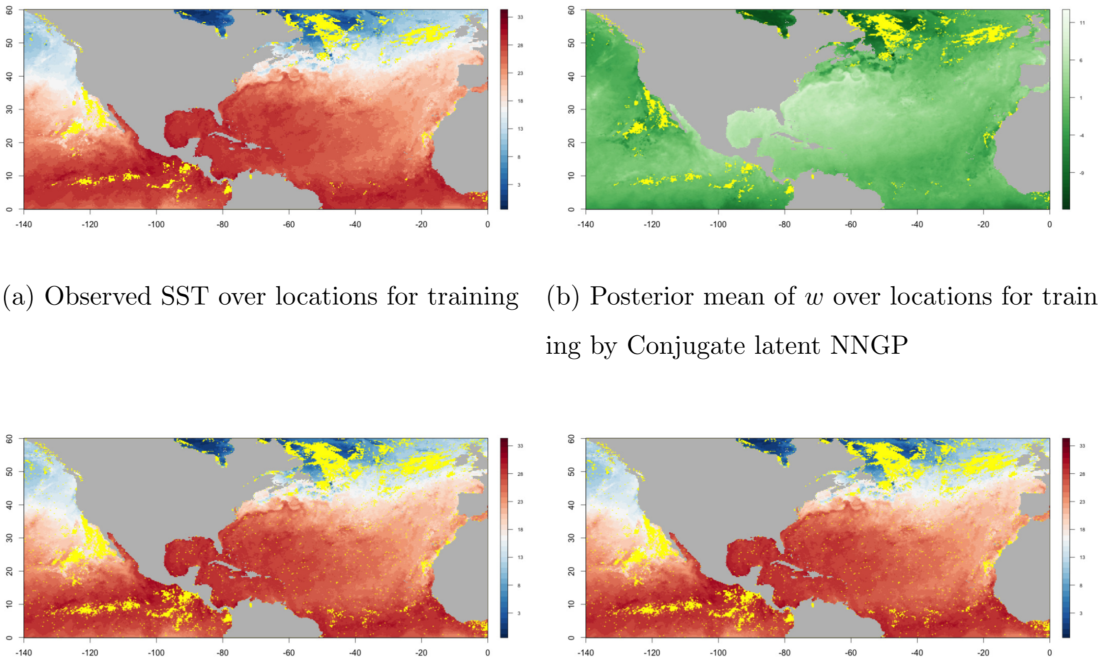


Figure 2.2: The Q-Q plot of the euclidean distance v.s. the spherical distance of 4000 pairs of observed locations over the study domain of the SST analysis. The red line is the 45 degree line

[2005] or page 10 in Banerjee et al. [2014]]. We compare the Euclidean distances computed from a sinusoidal projection and the spherical or geodesic distance over the study domain by checking the two distances for 4000 pairs of locations randomly selected from the observed location set. The Q-Q plot (figure 2.2) shows that the Euclidean distance based on sinusoidal projects serves as a good measure of distance over the study domain. An exponential spatial covariance function with sinusoidally projected distance was used for the model. Further model specifications included non-informative flat priors for the intercept and regression coefficients, inverse-gamma priors for τ^2 and σ^2 with shape parameter 2 and scale parameter equaling the respective estimates from an empirical variogram.

We fit the conjugate Bayesian model with fixed ϕ and δ^2 using the algorithm 2.1 in Section 2.3.3 with $m = 10$ nearest neighbors. We implement Algorithm 2.2 to choose the values of $\{\phi, \delta^2\}$ at $\phi = 7$, $\delta^2 = 0.001$. Figures 2.3b shows the posterior means for the latent process of the conjugate latent NNGP model. The temperatures are color-coded from light green indicating high temperatures to dark of green indicating low temperatures. The map of the latent process w indicates lower temperature on the east coast and higher



(a) Observed SST over locations for training

(b) Posterior mean of w over locations for training by Conjugate latent NNGP

(c) Observed SST over locations for testing

(d) Posterior mean of SST over locations for testing

Figure 2.3: Notes: (a) Observed SST over locations for training (b) Posterior mean of $\mathbf{w}$ over locations for training by Conjugate latent NNGP (c) Posterior mean of SST over locations for testing (d) Posterior mean of SST over locations for testing The land is colored by gray, locations in the ocean without observations are colored by yellow.

Table 2.2: Real data analysis summary table. Parameter Posterior summary mean (2.5, 97.5) percentiles

	Non-spatial	Conjugate latent NNGP [a]
Intercept	31.92(31.91, 31.92)	31.43 (31.28, 31.59)
x-coordinate (10^3km)	0.12 (0.12, 0.12)	0.07 (0.05, 0.09)
y-coordinate (10^3km)	-3.07 (-3.07, -3.07)	-3.03 (-3.08, -2.99)
σ^2	–	3.95 (3.94, 3.95)
ϕ	–	7.00
τ^2	11.44 (11.43, 11.46)	$3.95e^{-3}$ ($3.94e^{-3}$, $3.95e^{-3}$)
RMSPE	3.39	0.31

temperature on the west coast. At the same time, we observed high temperture at center of the map. These features coincide with the ocean current, suggesting that the ocean current plays an important role in the sea surface temperature.

Parameter estimates along with their estimated 95% credible intervals and performance metrics for candidate models are shown in Table 2.2.

The RMSPE for a non-spatial linear regression model, conjugate latent NNGP model were 1.13, 0.31, respectively. Compared to the spatial models, the non-spatial models have substantially higher values of RMSPE, which suggest that coordinates alone does not adequately capture the spatial structure of SST. The fitted SST map over the withheld locations (Fig 2.3d) using conjugate latent NNGP model is almost indistinguishable from the real SST map (Fig 2.3c). All the inference from the conjugate latent NNGP model are based on 300 samples. The sampling process took 2367 seconds. In average, the posterior mean of the latent process w can be obtained within 20 seconds.

2.6 Discussion

This Chapter has attempted to address some practical issues encountered by scientists and statisticians in the hierarchical modeling and analysis for very large geospatial datasets. Building upon some recent work on nearest-neighbor Gaussian processes for massive spatial data, we build conjugate Bayesian spatial regression models and propose strategies for rapidly deliverable inference on modest computing environments equipped with user-friendly and readily available software packages. In particular, we have demonstrated how judicious use of a conjugate latent NNGP model can be effective for estimation and uncertainty quantification of latent (underlying) spatial processes. This provides an easily implementable practical alternative to computationally onerous Bayesian computing approaches. All the computations done in this Chapter were implemented on a standard desktop using `R` and `Stan`. This Chapter intends to contribute toward innovations in statistical practice rather than novel methodologies.

The subsequent research of speeding up Algorithm 2.1 will include the following two aspects. Firstly, the speed of convergence of the regular CG algorithm to the solution of a symmetric positive definite linear system $Ax = b$ depends on the condition number of the matrix A. In practice, a *preconditioned* CG is much more beneficial. Preconditioning of the CG method in Algorithm 2.1 is achieved by using a symmetric positive definite preconditioner matrix, say $\mathbf{M} = \mathbf{L}\mathbf{L}^\top$, to solve $\tilde{\mathbf{A}}\tilde{\mathbf{x}} = \tilde{\mathbf{b}}$, where $\tilde{\mathbf{A}} = \mathbf{L}^{-1}\mathbf{A}\mathbf{L}^{-\top}$ and $\tilde{\mathbf{b}} = \mathbf{L}^{-1}\mathbf{b}$. The solution for $\mathbf{A}\mathbf{x} = \mathbf{b}$ is then obtained as $\mathbf{x} = \mathbf{L}^{-\top}\tilde{\mathbf{x}}$. The preconditioner should be chosen carefully. It should enjoy high memory efficiency and also ensure that $\kappa(\tilde{\mathbf{A}})$ is close to 1, where $\kappa(\cdot)$ denotes the condition number of a matrix. Without these conditions, the benefits of preconditioning will not be evident and further investigations are needed to specify efficient preconditioners for modifying Algorithm 2.1. The second aspect is parallel computing. The posterior samples generated by Algorithm 2.1 are independent, allowing the possibility of generating them simultaneously. One could explore the use of different parallel programming paradigms such as message parsing interfaces and GPUs to dramatically reduce the sampling times in Algorithm 1.

It is important to recognize that the conjugate Bayesian models outlined here are not restricted to the NNGP. Any spatial covariance structure that leads to efficient computations can, in principle, be used. There are a number of recently proposed approaches that can be adopted here. These include, but are not limited to, multi-resolution approaches [e.g., Nychka et al., 2002, 2015, Katzfuss, 2017], covariance tapering and its use in full-scale approximations [e.g., Furrer et al., 2006, Sang and Huang, 2012, Katzfuss, 2013], and stochastic partial differential equation approximations [Lindgren et al., 2011], among several others [see, e.g., Banerjee, 2017, and references therein].

With regard to the NNGP specifically, our choice was partially dictated by its easy implementation in `R` using the `spNNGP` package and in `Stan` as described in `http://mc-stan.org/users/documentation/case-studies/nngp.html`. The NNGP is built upon a very effective likelihood approximation [Vecchia, 1988, Stein et al., 2004], which has also been explored recently by several authors in a variety of contexts [Stroud et al., 2017, Guinness, 2016]. Guinness [2016] provides empirical evidence about Vecchia's approximation outperforming other alternate methods, but also points out some optimal methods for permuting the order of the spatial locations before constructing the model. His methods for choosing the order of the locations can certainly be executed prior to implementing the models proposed in this Chapter. Finally, an even more recent article by Katzfuss and Guinness [2017] proposes further extensions of the Vecchia approximation, but its practicability for massive datasets on modest computing environments with easily available software packages is yet to be ascertained.

Supplementary Material

All computer programs implementing the examples in this Chapter can be found in the public domain and downloaded from `https://github.com/LuZhangstat/ConjugateNNGP`.

CHAPTER 3

High-dimensional Multivariate Geostatistics: A Conjugate Bayesian Matrix-Normal Approach

3.1 Introduction

Analysis for environmental data sets often require joint modeling of multiple spatially dependent variables accounting for dependence among the variables and the spatial association for each variable. For point-referenced variables, multivariate Gaussian processes (GPs) serve as versatile tools for joint modeling of spatial variables [see, e.g., Schabenberger and Gotway, 2004, Cressie and Wikle, 2011, Banerjee et al., 2014, and references therein]. However, for a dataset with n observed locations, fitting a GP based spatial model typically requires floating point operations (flops) and memory requirements of the order $\sim n^3$ and $\sim n^2$, respectively. This is challenging when n is large. This "Big Data" problem has received much attention in the literature and a comprehensive review is beyond the scope of this Chapter; see, e.g., Banerjee [2017], Heaton et al. [2019], Sun et al. [2011] for a review and comparison of scalable modeling methods. Much of the aforementioned literature for scalable models focused on univariate spatial processes, i.e., assuming only one response for each location.

Multivariate processes [see, e.g., Genton and Kleiber, 2015, Salvaña and Genton, 2020, Le and Zidek, 2006, and references therein], has received relatively limited developments in the context of massive data. Bayesian models are attractive for inference on multivariate spatial processes because they can accommodate uncertainties in the process parameters more flexibly through their hierarchical structure. Multivariate spatial interpolation using conjugate Bayesian modeling can be found in Brown et al. [1994], Le et al. [1997], Sun et al. [1998], Le et al. [2001], Gamerman and Moreira [2004], but these methods do not address the challenges encountered in massive data sets. More flexible methods for joint modeling, including spatial factor models, have been investigated in Bayesian contexts [see, e.g. Schmidt and Gelfand, 2003, Ren and Banerjee, 2013, Taylor-Rodriguez et al., 2019], but these methods have focused upon delivering full Bayesian inference through iterative algorithms such as Markov chain Monte Carlo (MCMC).

In this Chapter, we extend the work in Chapter 2 to address the "Big Data" problem in multivariate spatial data modeling. we propose an augmented Bayesian multivariate linear model framework that accommodates conjugate distribution theory, similar to Gamerman and Moreira [2004], but that can scale up to massive data sets with locations numbering in the millions. More specifically, we embed the Nearest-Neighbor Gaussian process (NNGP) [Datta et al., 2016a] within our conjugate Bayesian framework. We will consider two classes of models. The first is obtained by modeling the spatially dependent variables jointly as a multivariate spatial process, while the second models a latent multivariate spatial process in a hierarchical setup. We refer to the former as the "response" model and the latter as the "latent" model and we explore some properties of these models.

The remainder of this Chapter is arranged as follows. Section 3.2 develops a conjugate Bayesian multivariate spatial regression framework using Matrix-Normal and Inverse-Wishart prior distributions. We first develop two classes of models, response models and latent models using Gaussian spatial processes, in Section 3.2.1. Subsequently, in Section 3.2.2 we develop scalable versions of these models using the Nearest Neighbor Gaussian process (NNGP). We develop NNGP response models and NNGP latent models in this conjugate Bayesian framework. A cross-validation algorithm to fix certain hyperparameters in these models is presented in Section 3.2.3 and some theoretical attributes of these models are presented in Section 3.2.4. Section 3.3 present some simulation experiments, while Section 3.4 analyzes a massive Normalized Difference Vegetation Index data with a few million locations. Finally, Section 3.5 concludes the Chapter with some discussion.

3.2 Bayesian Multivariate Geostatistical Modeling

3.2.1 Conjugate Multivariate Spatial Models

Conjugate Multivariate Response Model Let $\mathbf{y}(\mathbf{s}) = (y_1(\mathbf{s}), \ldots, y_q(\mathbf{s}))^\top \in \mathbb{R}^q$ be a $q \times 1$ vector of outcomes at location $\mathbf{s} \in \mathcal{D} \subset \mathbb{R}^d$ and $\mathbf{x}(\mathbf{s}) = (x_1(\mathbf{s}), \ldots, x_p(\mathbf{s}))^\top \in \mathbb{R}^p$ be a corresponding $p \times 1$ vector of explanatory variables. Conditional on these explanatory variables, the response is assumed to follow a multivariate Gaussian process,

$$\mathbf{y}(\mathbf{s}) \sim \mathrm{GP}(\boldsymbol{\beta}^\top \mathbf{x}(\mathbf{s}), \mathbf{C}(\cdot,\cdot)) \; ; \quad \mathbf{C}(\mathbf{s}, \mathbf{s}') = [\rho_\psi(\mathbf{s}, \mathbf{s}') + (\alpha^{-1} - 1)\delta_{\mathbf{s}=\mathbf{s}'}]\boldsymbol{\Sigma} \; , \tag{3.2.1}$$

where the mean of $\mathbf{y}(\mathbf{s})$ is $\boldsymbol{\beta}^\top \mathbf{x}(\mathbf{s})$, $\boldsymbol{\beta}$ is a $p \times q$ matrix of regression coefficients and $\mathbf{C}(\mathbf{s}, \mathbf{s}') = \{\mathrm{cov}\{y_i(\mathbf{s}), y_j(\mathbf{s}')\}\}$ is a $q \times q$ cross-covariance matrix [Genton and Kleiber, 2015] whose (i, j)-th element is the covariance between $y_i(\mathbf{s})$ and $y_j(\mathbf{s}')$. The cross-covariance matrix is defined for each pair of locations and is further specified as a multiple of a nonspatial positive definite matrix $\boldsymbol{\Sigma}$. The multiplication factor is a function of the two locations and is composed of two components: a spatial correlation function, $\rho_\psi(\mathbf{s}, \mathbf{s}')$, which introduces spatial dependence between the outcomes through hyperparameters ψ, and a micro-scale adjustment $(1/\alpha - 1)\delta_{\mathbf{s}=\mathbf{s}'}$, where $\delta_{\mathbf{s}=\mathbf{s}'} = 1$ if $\mathbf{s} = \mathbf{s}'$ and is 0 if $\mathbf{s} \neq \mathbf{s}'$, and $\alpha \in (0, 1]$ is a scalar parameter representing the overall strength of the spatial variability as a proportion of the total variation.

The covariance among the elements of $\mathbf{y}(\mathbf{s})$ within a location $\mathbf{s}$ is given by the elements of $\mathbf{C}(\mathbf{s}, \mathbf{s}) = (1/\alpha)\boldsymbol{\Sigma}$ suggesting the interpretation of $\boldsymbol{\Sigma}$ as the within-location (nonspatial) dependence among the outcomes adjusted by a scale of $1/\alpha$ to accommodate additional variation at local scales. The interpretation of α is equivalent to the ratio of the "partial sill" to the "sill" in classical geostatistics. For example, in the special case that $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_q$, $\mathrm{cov}\{y_i(\mathbf{s}), y_j(\mathbf{s}')\} = \sigma^2 \rho(\mathbf{s}, \mathbf{s}') + \sigma^2(1/\alpha - 1)\delta_{\mathbf{s}=\mathbf{s}'}$, which shows that $\sigma^2(1/\alpha - 1) = \tau^2$ is the variance of micro-scale processes (or the "nugget"), so that $\alpha = \sigma^2/(\sigma^2 + \tau^2)$ is the ratio of the spatial variance (partial sill) to the total variance (sill). A similar interpretation for α results in the univariate setting with $q = 1$.

Let $\mathcal{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\} \subset \mathcal{D}$ be a set of n locations yielding observations on $\mathbf{y}(\mathbf{s})$. Then $\mathbf{Y} = \mathbf{y}(\mathcal{S}) = [\mathbf{y}(\mathbf{s}_1) : \cdots : \mathbf{y}(\mathbf{s}_n)]^\top$ is $n \times q$ and $\mathbf{X} = \mathbf{x}(\mathcal{S}) = [\mathbf{x}(\mathbf{s}_1) : \cdots : \mathbf{x}(\mathbf{s}_n)]^\top$ is the corresponding $n \times p$ matrix of explanatory variables observed over $\mathcal{S}$. We will assume that $\mathbf{X}$ has full column rank ($= p < n$). The likelihood emerging from (3.2.1) is $\mathbf{Y} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma} \sim \mathrm{MN}_{n,q}(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\mathcal{K}}, \boldsymbol{\Sigma})$, where MN denotes the Matrix-Normal distribution defined in Ding and Cook [2014], i.e.,

$$\mathrm{MN}_{n,q}(\mathbf{Y} \mid \mathbf{X}\boldsymbol{\beta}, \boldsymbol{\mathcal{K}}, \boldsymbol{\Sigma}) = \frac{\exp\left(-\frac{1}{2}\operatorname{tr}\left[\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^\top \boldsymbol{\mathcal{K}}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\right]\right)}{(2\pi)^{np/2}|\boldsymbol{\Sigma}|^{n/2}|\boldsymbol{\mathcal{K}}|^{p/2}} , \tag{3.2.2}$$

where tr denotes trace, $\boldsymbol{\mathcal{K}} = \boldsymbol{\rho}_\psi + (\alpha^{-1} - 1)\mathbf{I}_n$ and $\boldsymbol{\rho}_\psi = \{\rho_\psi(\mathbf{s}_i, \mathbf{s}_j)\}$ is the $n \times n$ spatial correlation matrix. A conjugate Bayesian model is obtained by a Matrix-Normal-Inverse-Wishart (MNIW) prior on $\{\boldsymbol{\beta}, \boldsymbol{\Sigma}\}$, which we denote as

$$\mathrm{MNIW}(\boldsymbol{\beta}, \boldsymbol{\Sigma} \mid \boldsymbol{\mu}_\beta, \mathbf{V}_r, \boldsymbol{\Psi}, \nu) = \mathrm{IW}(\boldsymbol{\Sigma} \mid \boldsymbol{\Psi}, \nu) \times \mathrm{MN}_{p,q}(\boldsymbol{\beta} \mid \boldsymbol{\mu}_{\boldsymbol{\beta}}, \mathbf{V}_r, \boldsymbol{\Sigma}) , \tag{3.2.3}$$

where $\mathrm{IW}(\boldsymbol{\Sigma} \mid \boldsymbol{\Psi}, \nu)$ is the Inverse-Wishart distribution with parameters ν and $\boldsymbol{\Psi}$ describing the degrees of freedom and the scale matrix [see section 3.6 in Gelman et al., 2013]. The MNIW family is a conjugate prior with respect to the likelihood (3.2.2) and, for any fixed values of α, ψ and the hyperparameters in the prior density, we obtain the posterior density

$$\begin{aligned} p(\boldsymbol{\beta}, \boldsymbol{\Sigma} \mid \mathbf{Y}) &\propto \mathrm{MNIW}(\boldsymbol{\beta}, \boldsymbol{\Sigma} \mid \boldsymbol{\mu}_\beta, \mathbf{V}_r, \boldsymbol{\Psi}, \nu) \times \mathrm{MN}_{n,q}(\mathbf{Y} \mid \mathbf{X}\boldsymbol{\beta}, \boldsymbol{\mathcal{K}}, \boldsymbol{\Sigma}) \\ &\propto \mathrm{MNIW}(\boldsymbol{\mu}^*, \mathbf{V}^*, \boldsymbol{\Psi}^*, \nu^*) , \end{aligned} \tag{3.2.4}$$

where

$$\begin{aligned} \mathbf{V}^* &= (\mathbf{X}^\top \boldsymbol{\mathcal{K}}^{-1}\mathbf{X} + \mathbf{V}_r^{-1})^{-1} , \\ \boldsymbol{\mu}^* &= \mathbf{V}^*(\mathbf{X}^\top \boldsymbol{\mathcal{K}}^{-1}\mathbf{Y} + \mathbf{V}_r^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}}) , \\ \boldsymbol{\Psi}^* &= \boldsymbol{\Psi} + \mathbf{Y}^\top \boldsymbol{\mathcal{K}}^{-1}\mathbf{Y} + \boldsymbol{\mu}_{\boldsymbol{\beta}}^\top \mathbf{V}_r^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}} - \boldsymbol{\mu}^{*\top}\mathbf{V}^{*-1}\boldsymbol{\mu}^* \quad \text{and} \\ \nu^* &= \nu + n . \end{aligned} \tag{3.2.5}$$

Direct sampling from the MNIW posterior distribution in (3.2.4) is achieved by first sampling $\boldsymbol{\Sigma} \sim \mathrm{IW}(\boldsymbol{\Psi}^*, \nu^*)$ and then sampling one draw of $\boldsymbol{\beta} \sim \mathrm{MN}_{p,q}(\boldsymbol{\mu}^*, \mathbf{V}^*, \boldsymbol{\Sigma})$ for each draw of $\boldsymbol{\Sigma}$. The resulting pairs $\{\boldsymbol{\beta}, \boldsymbol{\Sigma}\}$ will be samples from (3.2.4). Since this scheme draws directly from the posterior distribution, the sample is exact and does not require burn-in or convergence.

Turning to predictions, let $\mathcal{U} = \{\mathbf{u}_1, \ldots, \mathbf{u}_{n'}\}$ be a finite set of locations where we intend to predict or impute the value of $\mathbf{y}(\mathbf{s})$ based upon an observed $n' \times p$ design matrix $\mathbf{X}_{\mathcal{U}} = [\mathbf{x}(\mathbf{u}_1) : \cdots : \mathbf{x}(\mathbf{u}_{n'})]^\top$ for $\mathcal{U}$. If $\mathbf{Y}_{\mathcal{U}} = [\mathbf{y}(\mathbf{u}_1) : \cdots : \mathbf{y}(\mathbf{u}_{n'})]^\top$ is the $n' \times q$ matrix of predictive random variables, then the conditional predictive distribution is

$$\begin{aligned} p(\mathbf{Y}_{\mathcal{U}} \,|\, \mathbf{Y}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) =& \mathrm{MN}_{n',q}(\mathbf{Y}_{\mathcal{U}} \,|\, \mathbf{X}_{\mathcal{U}}\boldsymbol{\beta} + \boldsymbol{\rho}_\psi(\mathcal{U},\mathcal{S})\boldsymbol{\mathcal{K}}^{-1}[\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}], \\ & \boldsymbol{\rho}_\psi(\mathcal{U},\mathcal{U}) + (\alpha^{-1} - 1)\mathbf{I}_{n'} - \boldsymbol{\rho}_\psi(\mathcal{U},\mathcal{S})\boldsymbol{\mathcal{K}}^{-1}\boldsymbol{\rho}_\psi(\mathcal{S},\mathcal{U}),\ \boldsymbol{\Sigma}) \ , \end{aligned} \tag{3.2.6}$$

where $\boldsymbol{\rho}_\psi(\mathcal{U},\mathcal{S}) = \{\rho_\psi(\mathbf{u}_i, \mathbf{s}_j)\}$ is $n' \times n$ and $\boldsymbol{\rho}_\psi(\mathcal{S},\mathcal{U}) = \{\rho_\psi(\mathbf{s}_i, \mathbf{u}_j)\} = \boldsymbol{\rho}_\psi(\mathcal{U},\mathcal{S})^\top$. Predictions can also be directly carried out in posterior predictive fashion, where we sample from

$$\begin{aligned} p(\mathbf{Y}_{\mathcal{U}} \,|\, \mathbf{Y}) = \int & \mathrm{MN}_{n',q}(\mathbf{X}_{\mathcal{U}}\boldsymbol{\beta} + \boldsymbol{\rho}_\psi(\mathcal{U},\mathcal{S})\boldsymbol{\mathcal{K}}^{-1}[\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}], \\ & \boldsymbol{\rho}_\psi(\mathcal{U},\mathcal{U}) + (\alpha^{-1} - 1)\mathbf{I}_{n'} - \boldsymbol{\rho}_\psi(\mathcal{U},\mathcal{S})\boldsymbol{\mathcal{K}}^{-1}\boldsymbol{\rho}_\psi(\mathcal{S},\mathcal{U}),\ \boldsymbol{\Sigma}) \\ & \times \mathrm{MNIW}(\boldsymbol{\mu}^*, \mathbf{V}^*, \boldsymbol{\Psi}^*, \nu^*)\, d\boldsymbol{\beta}\, d\boldsymbol{\Sigma} \ . \end{aligned} \tag{3.2.7}$$

Sampling from (3.2.7) is achieved by drawing one $\mathbf{Y}_{\mathcal{U}}$ from (3.2.6) for each posterior draw of $\{\boldsymbol{\beta}, \boldsymbol{\Sigma}\}$.

Conjugate Multivariate Latent Model We now discuss a conjugate Bayesian model for a latent process. Consider the spatial regression model

$$\mathbf{y}(\mathbf{s}) = \boldsymbol{\beta}^\top\mathbf{x}(\mathbf{s}) + \boldsymbol{\omega}(\mathbf{s}) + \boldsymbol{\epsilon}(\mathbf{s}) \ , \ \mathbf{s} \in \mathcal{D} \ , \tag{3.2.8}$$

where $\boldsymbol{\omega}(\mathbf{s}) \sim \mathrm{GP}(\mathbf{0}_{q\times 1}, \rho_\psi(\cdot,\cdot)\boldsymbol{\Sigma})$ is a $q \times 1$ multivariate latent process with cross-covariance matrix $\rho_\psi(\mathbf{s},\mathbf{s}')\boldsymbol{\Sigma}$ and $\boldsymbol{\epsilon}(\mathbf{s}) \overset{iid}{\sim} \mathrm{N}(\mathbf{0}_{q\times 1}, (\alpha^{-1} - 1)\boldsymbol{\Sigma})$ captures micro-scale variation. The "proportionality" assumption for the variance of $\boldsymbol{\epsilon}(\mathbf{s})$ will allow us to derive analytic posterior distributions using conjugate priors.

The latent process $\boldsymbol{\omega}(\mathbf{s})$ captures the underlying spatial pattern and holds specific interest in many applications. Let $\boldsymbol{\omega} = \boldsymbol{\omega}(\mathcal{S}) = [\boldsymbol{\omega}(\mathbf{s}_1) : \cdots : \boldsymbol{\omega}(\mathbf{s}_n)]^\top$ be $n \times q$. The parameter space with the latent process is $\{\boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\Sigma}\}$. Letting $\boldsymbol{\gamma}^\top = [\boldsymbol{\beta}^\top, \boldsymbol{\omega}^\top]$ be $q \times (p + n)$, we assume that

$\{\boldsymbol{\gamma}, \boldsymbol{\Sigma}\} \sim \text{MNIW}(\boldsymbol{\mu}_{\boldsymbol{\gamma}}, \mathbf{V}_{\boldsymbol{\gamma}}, \boldsymbol{\Psi}, \nu)$, where $\boldsymbol{\mu}_{\boldsymbol{\gamma}}^{\top} = [\boldsymbol{\mu}_{\boldsymbol{\beta}}^{\top}, \mathbf{0}_{q\times n}]$ and $\mathbf{V}_{\boldsymbol{\gamma}} = \text{blockdiag}\{\mathbf{V}_r, \boldsymbol{\rho}_{\psi}(\mathcal{S}, \mathcal{S})\}$. The posterior density is

$$\begin{aligned} p(\boldsymbol{\gamma}, \boldsymbol{\Sigma} \,|\, \mathbf{Y}) &\propto \text{MNIW}(\boldsymbol{\gamma}, \boldsymbol{\Sigma} \,|\, \boldsymbol{\mu}_{\boldsymbol{\gamma}}, \mathbf{V}_{\boldsymbol{\gamma}}, \boldsymbol{\Psi}, \nu) \times \text{MN}_{n,q}(\mathbf{Y}_{n\times q} \,|\, [\mathbf{X} : \mathbf{I}_n]\boldsymbol{\gamma}, (\alpha^{-1} - 1)\mathbf{I}_n, \boldsymbol{\Sigma}) \\ &\propto \text{MNIW}(\boldsymbol{\gamma}, \boldsymbol{\Sigma} \,|\, \boldsymbol{\mu}_{\boldsymbol{\gamma}}^{*}, \mathbf{V}^{*}, \boldsymbol{\Psi}^{*}, \nu^{*}) \;, \end{aligned} \tag{3.2.9}$$

where

$$\begin{aligned} &\mathbf{V}^* = \begin{bmatrix} \frac{\alpha}{1-\alpha}\mathbf{X}^{\top}\mathbf{X} + \mathbf{V}_r^{-1} & \frac{\alpha}{1-\alpha}\mathbf{X}^{\top} \\ \frac{\alpha}{1-\alpha}\mathbf{X} & \boldsymbol{\rho}_{\psi}^{-1}(\mathcal{S}, \mathcal{S}) + \frac{\alpha}{1-\alpha}\mathbf{I}_n \end{bmatrix}^{-1} \;, \quad \boldsymbol{\mu}_{\boldsymbol{\gamma}}^{*} = \mathbf{V}^* \begin{bmatrix} \frac{\alpha}{1-\alpha}\mathbf{X}^{\top}\mathbf{Y} + \mathbf{V}_r^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}} \\ \frac{\alpha}{1-\alpha}\mathbf{Y} \end{bmatrix} , \\ &\boldsymbol{\Psi}^* = \boldsymbol{\Psi} + \frac{\alpha}{1-\alpha}\mathbf{Y}^{\top}\mathbf{Y} + \boldsymbol{\mu}_{\boldsymbol{\beta}}^{\top}\mathbf{V}_r^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}} - \boldsymbol{\mu}_{\boldsymbol{\gamma}}^{*\top}\mathbf{V}^{*-1}\boldsymbol{\mu}_{\boldsymbol{\gamma}}^{*} \;\text{ and } \nu^* = \nu + n \;. \end{aligned} \tag{3.2.10}$$

For prediction on a set of location $\mathcal{U}$, we can estimate the unobserved latent process $\boldsymbol{\omega}_{\mathcal{U}} = \boldsymbol{\omega}(\mathcal{U}) = [\boldsymbol{\omega}(\mathbf{u}_1) : \cdots : \boldsymbol{\omega}(\mathbf{u}_{n'})]^{\top}$ and the response $\mathbf{Y}_{\mathcal{U}}$ through

$$\begin{aligned} p(\mathbf{Y}_{\mathcal{U}}, \boldsymbol{\omega}_{\mathcal{U}} \,|\, \mathbf{Y}) = \int & \text{MN}_{n',q}(\mathbf{Y}_{\mathcal{U}} \,|\, \mathbf{X}_{\mathcal{U}}\boldsymbol{\beta} + \boldsymbol{\omega}_{\mathcal{U}},\ (\alpha^{-1} - 1)\mathbf{I}_{n'},\ \boldsymbol{\Sigma}) \\ & \times \text{MN}_{n',q}(\boldsymbol{\omega}_{\mathcal{U}} \,|\, \mathbf{M}_{\mathcal{U}}\boldsymbol{\omega}, \mathbf{V}_{\omega_{\mathcal{U}}}, \boldsymbol{\Sigma}) \times \text{MNIW}(\boldsymbol{\gamma}, \boldsymbol{\Sigma} \,|\, \boldsymbol{\mu}_{\boldsymbol{\gamma}}^{*}, \mathbf{V}^*, \boldsymbol{\Psi}^*, \nu^*)\; d\boldsymbol{\gamma} d\boldsymbol{\Sigma} \;, \end{aligned} \tag{3.2.11}$$

where $\mathbf{M}_{\mathcal{U}} = \boldsymbol{\rho}_{\psi}(\mathcal{U}, \mathcal{S})\boldsymbol{\rho}_{\psi}^{-1}(\mathcal{S}, \mathcal{S})$ and $\mathbf{V}_{\omega_{\mathcal{U}}} = \boldsymbol{\rho}_{\psi}(\mathcal{U},\mathcal{U}) - \boldsymbol{\rho}_{\psi}(\mathcal{U}, \mathcal{S})\boldsymbol{\rho}_{\psi}^{-1}(\mathcal{S}, \mathcal{S})\boldsymbol{\rho}_{\psi}(\mathcal{S},\mathcal{U})$. Posterior predictive inference proceeds by sampling one draw of $\boldsymbol{\omega}_{\mathcal{U}} \sim \text{MN}_{n',q}(\boldsymbol{\omega}_{\mathcal{U}} \,|\, \mathbf{M}_{\mathcal{U}}\boldsymbol{\omega}, \mathbf{V}_{\omega_{\mathcal{U}}}, \boldsymbol{\Sigma})$ for each posterior draw of $\{\boldsymbol{\gamma}, \boldsymbol{\Sigma}\}$ and then one draw of $\mathbf{Y}_{\mathcal{U}} \sim \text{MN}(\mathbf{X}_{\mathcal{U}}\boldsymbol{\beta}+\boldsymbol{\omega}_{\mathcal{U}}, (\alpha^{-1}-1)\mathbf{I}_{n'},\ \boldsymbol{\Sigma})$ for each drawn $\{\boldsymbol{\omega}_{\mathcal{U}}, \boldsymbol{\gamma}, \boldsymbol{\Sigma}\}$.

3.2.2 Scalable Conjugate Bayesian Multivariate Models

Conjugate multivariate response NNGP model A conjugate Bayesian modeling framework is appealing for massive spatial data sets because the posterior distribution of the parameters are available in closed form circumventing the need for MCMC algorithms. The key computational bottleneck for Bayesian estimation of spatial process models concerns the computation and storage involving $\boldsymbol{\mathcal{K}}^{-1}$ in (3.2.5). The required matrix computations require $\mathcal{O}(n^3)$ flops and $\mathcal{O}(n^2)$ storage when $\boldsymbol{\mathcal{K}}$ is $n \times n$ and dense. While conjugate models reduce computational expenses by enabling direct sampling from closed-form posterior and posterior predictive distributions, the computation and storage of $\boldsymbol{\mathcal{K}}$ is still substantial for massive datasets.

One approach to obviate the overwhelming computations is to develop a sparse alternative for $\boldsymbol{\mathcal{K}}^{-1}$ in (3.2.5). One such approximation that has generated substantial recent attention in the spatial literature is an approximation due to Vecchia [Vecchia, 1988]. Consider the spatial covariance matrix $\boldsymbol{\mathcal{K}} = \boldsymbol{\rho}_{\psi} + \delta_{\mathbf{s}=\mathbf{s}'}\mathbf{I}_n$ in (3.2.2). This is a dense $n \times n$ matrix with no apparent exploitable structure. Instead, we specify a sparse Cholesky representation

$$\boldsymbol{\mathcal{K}}^{-1} = (\mathbf{I} - \mathbf{A}_{\boldsymbol{\mathcal{K}}})^{\top}\mathbf{D}_{\boldsymbol{\mathcal{K}}}^{-1}(\mathbf{I} - \mathbf{A}_{\boldsymbol{\mathcal{K}}}) \ , \tag{3.2.12}$$

where $\mathbf{D}_{\boldsymbol{\mathcal{K}}}$ is a diagonal matrix and $\mathbf{A}_{\boldsymbol{\mathcal{K}}}$ is a sparse lower-triangular matrix with 0 along the diagonal and with no more than a fixed small number m of nonzero entries in each row of $\mathbf{A}_{\boldsymbol{\mathcal{K}}}$. The diagonal entries of $\mathbf{D}_{\boldsymbol{\mathcal{K}}}^{-1}$ and the nonzero entries of $\mathbf{A}_{\boldsymbol{\mathcal{K}}}$ are obtained from the conditional variance and conditional expectations for a Gaussian process with covariance function $\rho_{\psi}(\mathbf{s}, \mathbf{s}')$. To be precise, we consider a fixed order of locations in $\mathcal{S}$ and define $N_m(\mathbf{s}_i)$ to be the set comprising at most m neighbors of $\mathbf{s}_i$ among locations $\mathbf{s}_j \in \mathcal{S}$ such that $j < i$. The (i, j)-th entry of $\mathbf{A}_{\boldsymbol{\mathcal{K}}}$ is 0 whenever $\mathbf{s}_j \notin N_m(\mathbf{s}_i)$. If $j_1 < j_2 < \cdots < j_m$ are the m column indices indicating the nonzero entries in the i-th row of $\mathbf{A}_{\boldsymbol{\mathcal{K}}}$, then the (i, j_k)-th element of $\mathbf{A}_{\boldsymbol{\mathcal{K}}}$ is set equal to the k-th element of the $1 \times m$ vector $\mathbf{a}_i^{\top} = \boldsymbol{\rho}_{\psi}(\mathbf{s}_i, N_m(\mathbf{s}_i))\boldsymbol{\rho}_{\psi}(N_m(\mathbf{s}_i), N_m(\mathbf{s}_i))^{-1}$. The (i, i)-th diagonal element of $\mathbf{D}_{\boldsymbol{\mathcal{K}}}$ is given by $\rho_{\psi}(\mathbf{s}_i, \mathbf{s}_i) - \mathbf{a}_i^{\top}\boldsymbol{\rho}_{\psi}(N_m(\mathbf{s}_i), \mathbf{s}_i)$. Repeating these calculations for each row completes the construction of $\mathbf{A}_{\boldsymbol{\mathcal{K}}}$ and $\mathbf{D}_{\kappa}$ and yields a sparse $\boldsymbol{\mathcal{K}}^{-1}$ in (3.2.12). This construction can be performed in parallel and requires storage or computation of at most $m \times m$ matrices, where $m << n$, costing $\mathcal{O}(n)$ flops and storage. Further algorithmic details about this construction can be found in Finley et al. [2019].

Based on Section 3.2.1, the posterior distribution $\boldsymbol{\beta}, \boldsymbol{\Sigma} \,|\, \mathbf{Y}$ follows $\text{MNIW}(\boldsymbol{\mu}^*, \mathbf{V}^*, \boldsymbol{\Psi}^*, \nu^*)$ where $\{\boldsymbol{\mu}^*, \mathbf{V}^*, \boldsymbol{\Psi}^*, \nu^*\}$ are given in (3.2.5). With the sparse representation of $\boldsymbol{\mathcal{K}}^{-1}$ in (3.2.12), the process of obtaining posterior inference for $\{\boldsymbol{\beta}, \boldsymbol{\Sigma}\}$ only involves steps with storage and computational requirement in $\mathcal{O}(n)$.

The predictions on the unobserved locations $\mathcal{U} = \{\mathbf{u}_1, \ldots, \mathbf{u}_{n'}\}$ is also simplified as follows.

We extend the definition of $N_m(\mathbf{s}_i)$'s to arbitrary locations by defining $N_m(\mathbf{u}_i)$ to be the set of m nearest neighbors of $\mathbf{u}_i$ from $\mathcal{S}$. Furthermore, we assume that $\mathbf{y}(\mathbf{u})$ and $\mathbf{y}(\mathbf{u}')$ are conditionally independent of each other given $\mathbf{Y} = \mathbf{y}(\mathcal{S})$ and the other model parameters. Thus, for any $\mathbf{u}_i \in \mathcal{U}$, we have

$$\mathbf{y}(\mathbf{u}_i) \,|\, \mathbf{Y}, \boldsymbol{\beta}, \boldsymbol{\Sigma} \sim \mathrm{N}(\boldsymbol{\beta}^\top \mathbf{x}(\mathbf{u}_i) + [\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}]^\top \tilde{\mathbf{a}}_i,\, \tilde{d}_i \boldsymbol{\Sigma}),\ i = 1, \ldots, n' \ , \tag{3.2.13}$$

where $\tilde{\mathbf{a}}_i$ is an $n \times 1$ vector with m non-zero elements. If $N_m(\mathbf{u}_i) = \{\mathbf{s}_{j_k}\}_{k=1}^m$, then

$$\begin{aligned} (\{\tilde{\mathbf{a}}_i\}_{j_1}, \ldots, \{\tilde{\mathbf{a}}_i\}_{j_m}) &= \boldsymbol{\rho}_\psi(\mathbf{u}_i, N_m(\mathbf{u}_i))\{\boldsymbol{\rho}_\psi(N_m(\mathbf{u}_i), N_m(\mathbf{u}_i)) + (\alpha^{-1} - 1)\mathbf{I}_m\}^{-1} \ , \\ \tilde{d}_i &= \alpha^{-1} - \boldsymbol{\rho}_\psi(\mathbf{u}_i, N_m(\mathbf{u}_i))[\boldsymbol{\rho}_\psi(N_m(\mathbf{u}_i), N_m(\mathbf{u}_i)) + (\alpha^{-1} - 1)\mathbf{I}_m]^{-1}\boldsymbol{\rho}_\psi(N_m(\mathbf{u}_i), \mathbf{u}_i) \ . \end{aligned} \tag{3.2.14}$$

If $\tilde{\mathbf{A}} = [\tilde{\mathbf{a}}_1 : \cdots : \tilde{\mathbf{a}}_{n'}]^\top$ and $\tilde{\mathbf{D}} = \mathrm{diag}(\{\tilde{d}_i\}_{i=1}^n$, then the conditional predictive density for $\mathbf{Y}_\mathcal{U}$ is

$$\mathbf{Y}_\mathcal{U} \,|\, \mathbf{Y}, \boldsymbol{\beta}, \boldsymbol{\Sigma} \sim \mathrm{MN}(\mathbf{X}_\mathcal{U}\boldsymbol{\beta} + \tilde{\mathbf{A}}[\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}], \tilde{\mathbf{D}}, \boldsymbol{\Sigma}) \ . \tag{3.2.15}$$

Since the posterior distribution of $\{\boldsymbol{\beta}, \boldsymbol{\Sigma}\}$ and the conditional predictive distribution of $\mathbf{Y}_\mathcal{U}$ are both available in closed form, direct sampling from the posterior predictive distribution is straightforward. A detailed algorithm for obtaining the posterior inference on parameter set $\{\boldsymbol{\beta}, \boldsymbol{\Sigma}\}$ and the posterior prediction over a new set of location $\mathcal{U}$ is given as below.

Algorithm 3.1: Obtaining posterior inference of $\{\boldsymbol{\beta}, \boldsymbol{\Sigma}\}$ and predictions on $\mathcal{U}$ for conjugate multivariate response NNGP

```
1. Construct $\mathbf{V}^*$, $\boldsymbol{\mu}^*$, $\boldsymbol{\Psi}^*$ and $\nu^*$:
    (a) Compute $\mathbf{L}_r$ the Cholesky decomposition of $\mathbf{V}_r$
    (b) Compute $\mathbf{DIAX} = \mathbf{D}_k^{-\frac{1}{2}}(\mathbf{I} - \mathbf{A}_k)\mathbf{X}$ and $\mathbf{DIAY} = \mathbf{D}_k^{-\frac{1}{2}}(\mathbf{I} - \mathbf{A}_k)\mathbf{Y}$
        • Construct $\mathbf{A}_K$ and $\mathbf{D}_K$ as described, for example, in Finley et al. [2019]    $\mathcal{O}(nm^3)$
        • Compute $\mathbf{DIAX} = \mathbf{D}_k^{-\frac{1}{2}}(\mathbf{I} - \mathbf{A}_k)\mathbf{X}$ and $\mathbf{DIAY} = \mathbf{D}_k^{-\frac{1}{2}}(\mathbf{I} - \mathbf{A}_k)\mathbf{Y}$    $\mathcal{O}(n(m+1)(p+q+2))$
    (c) Obtain $\mathbf{V}^*$, $\boldsymbol{\mu}^*$ and $\boldsymbol{\Psi}^*$
        • Compute $\mathbf{V}^* = (\mathbf{DIAX}^\top\mathbf{DIAX} + \mathbf{V}_r^{-1})^{-1}$ and its Cholesky decomposition $\mathbf{L}_{v*}$    $\mathcal{O}(np^2)$
        • Compute $\boldsymbol{\mu}^* = \mathbf{V}^*(\mathbf{DIAX}^\top\mathbf{DIAY} + \mathbf{V}_r^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}})$    $\mathcal{O}(npq)$
        • Compute $\boldsymbol{\Psi}^* = \boldsymbol{\Psi} + \mathbf{DIAY}^\top\mathbf{DIAY} + (\mathbf{L}_r^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}})^\top(\mathbf{L}_r^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}}) - (\mathbf{L}_{v*}^{-1}\boldsymbol{\mu}^*)^\top(\mathbf{L}_{v*}^{-1}\boldsymbol{\mu}^*)$    $\mathcal{O}(nq^2)$
        • Compute $\nu^* = \nu + n$
```

```
2. Generate posterior samples $\{\mathbf{Y}_{\mathcal{U}}^{(l)}\}_{l=1}^{L}$ on a new set $\mathcal{U}$ given $\mathbf{X}_{\mathcal{U}}$
    (a) Construct $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{D}}$ as described in (3.2.14)    $\mathcal{O}(n'm^3)$
    (b) For $l$ in $1:L$
        i. Sample $\mathbf{\Sigma}^{(l)} \sim \mathrm{IW}(\mathbf{\Psi}^*, \nu^*)$
        ii. Sample $\boldsymbol{\beta}^{(l)} \sim \mathrm{MN}(\boldsymbol{\mu}^*, \mathbf{V}^*, \mathbf{\Sigma}^{(l)})$
            • Calculate Cholesky decomposition of $\mathbf{\Sigma}^{(l)}$, $\mathbf{\Sigma}^{(l)} = \mathbf{L}_{\Sigma^{(l)}} \mathbf{L}_{\Sigma^{(l)}}^{\top}$
            • Sample $\mathbf{u} \sim \mathrm{MN}(\mathbf{0}, \mathbf{I}_p, \mathbf{I}_q)$ (i.e. $\mathrm{vec}(\mathbf{u}) \sim \mathrm{MVN}(0, \mathbf{I}_{pq})$)
            • Generate $\boldsymbol{\beta}^{(l)} = \boldsymbol{\mu}^* + \mathbf{L}_{v*}\mathbf{u}\mathbf{L}_{\mathbf{\Sigma}^{(l)}}^{\top}$
        iii. Sample $\mathbf{Y}_{\mathcal{U}}^{(l)} \sim \mathrm{MN}(\mathbf{X}_{\mathcal{U}}\boldsymbol{\beta}^{(l)} + \tilde{\mathbf{A}}[\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^{(l)}], \tilde{\mathbf{D}}, \mathbf{\Sigma}^{(l)})$
            • Sample $\mathbf{u} \sim \mathrm{MN}(\mathbf{0}, \mathbf{I}_{n'}, \mathbf{I}_q)$.
            • Generate $\mathbf{Y}_{\mathcal{U}}^{(l)} = \mathbf{X}_{\mathcal{U}}\boldsymbol{\beta}^{(l)} + \tilde{\mathbf{A}}[\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}^{(l)}] + \tilde{\mathbf{D}}^{\frac{1}{2}}\mathbf{u}\mathbf{L}_{\mathbf{\Sigma}^{(l)}}^{\top}$    $\mathcal{O}((n'+n)pq + n'(q^2+mq))$
```

Conjugate multivariate latent NNGP model Bayesian estimation for the conjugate multivariate latent model is more challenging because inference is usually sought on the (high-dimensional) latent process itself. In particular, the calculations involved in $\mathbf{V}^*$ in (3.2.9) are often too expensive for large data sets even when the precision matrix $\boldsymbol{\rho}_{\psi}^{-1}(\mathcal{S}, \mathcal{S})$ is sparse. Here, the latent process $\boldsymbol{\omega}(\mathbf{s})$ in (3.2.8) follows a multivariate Gaussian process so that its realizations over $\mathcal{S}$ follows $\boldsymbol{\omega} \sim \mathrm{MN}(\mathbf{O}_{n\times q}, \tilde{\boldsymbol{\rho}}, \mathbf{\Sigma})$, where $\tilde{\boldsymbol{\rho}}$ is the Vecchia approximation of $\boldsymbol{\rho}_{\psi}(\mathcal{S}, \mathcal{S})$. Hence, $\tilde{\boldsymbol{\rho}}^{-1} = (\mathbf{I} - \mathbf{A}_{\boldsymbol{\rho}})^{\top}\mathbf{D}_{\boldsymbol{\rho}}^{-1}(\mathbf{I} - \mathbf{A}_{\boldsymbol{\rho}})$, where $\mathbf{A}_{\boldsymbol{\rho}}$ and $\mathbf{D}_{\boldsymbol{\rho}}$ are constructed analogous to $\mathbf{A}_{\boldsymbol{\mathcal{K}}}$ and $\mathbf{D}_{\boldsymbol{\mathcal{K}}}$ in (3.2.12) with $\boldsymbol{\mathcal{K}}$ replaced by $\boldsymbol{\rho}_{\psi}(\mathcal{S}, \mathcal{S})$. This corresponds to modeling $\boldsymbol{\omega}(\mathbf{s})$ with a Nearest-Neighbor Gaussian Process (NNGP) [see, e.g., Datta et al., 2016a,b, Banerjee, 2017, for details].

The posterior distribution of $\{\boldsymbol{\gamma}, \mathbf{\Sigma}\}$ follows a Matrix-Normal distribution similar to (3.2.9), but with $\boldsymbol{\rho}_{\psi}(\mathcal{S}, \mathcal{S})^{-1}$ in (3.2.10) replaced by its Vecchia approximation $\tilde{\boldsymbol{\rho}}_{\psi}(\mathcal{S}, \mathcal{S})$. We will solve the linear system $\mathbf{X}^{*\top}\mathbf{X}^*\boldsymbol{\mu}^* = \mathbf{X}^{*\top}\mathbf{Y}^*$ for $\boldsymbol{\mu}^*$, compute $\{\mathbf{\Psi}^*, \nu^*\}$ and generate posterior samples of $\mathbf{\Sigma}$ from $\mathrm{IW}(\mathbf{\Psi}^*, \nu^*)$. Posterior samples of $\boldsymbol{\gamma}$ are obtained by generating $\boldsymbol{\eta} \sim \mathrm{MN}(\mathbf{O}, \mathbf{I}_{2n+p}, \mathbf{\Sigma})$, solving $\mathbf{X}^{*\top}\mathbf{X}^*\mathbf{v} = \mathbf{X}^{*\top}\boldsymbol{\eta}$ for $\mathbf{v}$ and then obtaining posterior samples of $\boldsymbol{\gamma}$ from $\boldsymbol{\gamma} = \boldsymbol{\mu}^* + \mathbf{v}$.

However, sampling $\{\boldsymbol{\gamma}, \boldsymbol{\Sigma}\}$ is still challenging for massive data sets, where we seek to minimize storage and operations with large matrices. Here we introduce a useful representation. Let $\mathbf{V}_{\boldsymbol{\rho}}$ be a non-singular square matrix such that $\boldsymbol{\rho}_{\psi}^{-1}(\mathcal{S}, \mathcal{S}) = \mathbf{V}_{\boldsymbol{\rho}}^{\top}\mathbf{V}_{\boldsymbol{\rho}}$ where we write $\mathbf{V}_{\boldsymbol{\rho}} = \mathbf{D}_{\boldsymbol{\rho}}^{-1/2}(\mathbf{I} - \mathbf{A}_{\boldsymbol{\rho}})$. We treat the prior of $\boldsymbol{\gamma}$ as additional "observations" and recast $p(\mathbf{Y}, \boldsymbol{\gamma} \,|\, \boldsymbol{\Sigma}) = p(\mathbf{Y} \,|\, \boldsymbol{\gamma}, \boldsymbol{\Sigma}) \times p(\boldsymbol{\gamma} \,|\, \boldsymbol{\Sigma})$ into an augmented linear model

$$\underbrace{\begin{bmatrix} \sqrt{\frac{\alpha}{1-\alpha}}\mathbf{Y} \\ \mathbf{L}_r^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}} \\ \mathbf{0} \end{bmatrix}}_{\mathbf{Y}^*} = \underbrace{\begin{bmatrix} \sqrt{\frac{\alpha}{1-\alpha}}\mathbf{X} & \sqrt{\frac{\alpha}{1-\alpha}}\mathbf{I}_n \\ \mathbf{L}_r^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_{\boldsymbol{\rho}} \end{bmatrix}}_{\mathbf{X}^*} \underbrace{\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\omega} \end{bmatrix}}_{\boldsymbol{\gamma}} + \underbrace{\begin{bmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 \\ \boldsymbol{\eta}_3 \end{bmatrix}}_{\boldsymbol{\eta}} , \qquad (3.2.16)$$

where $\mathbf{L}_r$ is the Cholesky decomposition of $\mathbf{V}_r$, and $\boldsymbol{\eta} \sim \text{MN}(\mathbf{0}, \mathbf{I}_{2n+p}, \boldsymbol{\Sigma})$. With a flat prior for $\boldsymbol{\beta}$, $\mathbf{L}_r^{-1}$ degenerates to $\mathbf{O}$ and does not contribute to the linear system. The expression in (3.2.10) can now be simplified as follows

$$\begin{aligned} &\mathbf{V}^* = (\mathbf{X}^{*\top}\mathbf{X}^*)^{-1} \;,\; \boldsymbol{\mu}^* = (\mathbf{X}^{*\top}\mathbf{X}^*)^{-1}\mathbf{X}^{*\top}\mathbf{Y}^* \;, \\ &\boldsymbol{\Psi}^* = \boldsymbol{\Psi} + (\mathbf{Y}^* - \mathbf{X}^*\boldsymbol{\mu}^*)^{\top}(\mathbf{Y}^* - \mathbf{X}^*\boldsymbol{\mu}^*) \;,\; \nu^* = \nu + n \;. \end{aligned} \qquad (3.2.17)$$

Following developments in Zhang et al. [2019] for the univariate case, one can efficiently generate posterior samples through a conjugate gradient algorithm exploiting the sparsity of $\mathbf{V}_{\boldsymbol{\rho}}$. The sampling process for $\boldsymbol{\gamma}$ will be scalable when there is a sparse precision matrix $\boldsymbol{\rho}_{\psi}^{-1}(\mathcal{S}, \mathcal{S})$. It is also possible to construct $\mathbf{V}^*$ and $\boldsymbol{\mu}^*$ in (3.2.17) using $\boldsymbol{\rho}_{\psi}^{-1}(\mathcal{S}, \mathcal{S})$ instead of $\mathbf{V}_{\boldsymbol{\rho}}$. We refer to Zhang et al. [2019] for further details of this construction. We provide a detailed algorithm of the conjugate multivariate latent NNGP model below, where we implement a "Sparse Equations and Least Squares" (LSMR) algorithm [Fong and Saunders, 2011] to solve the linear system $\mathbf{X}^{*\top}\mathbf{X}^*\boldsymbol{\mu}^* = \mathbf{X}^{*\top}\mathbf{Y}^*$ and $\mathbf{X}^{*\top}\mathbf{X}^*\mathbf{v} = \mathbf{X}^{*\top}\boldsymbol{\eta}$ needed to generate $\boldsymbol{\gamma}$. LSMR is a conjugate-gradient type algorithm for solving sparse linear equations $\mathbf{A}\mathbf{x} = \mathbf{b}$ where the matrix $\mathbf{A}$ may be square or rectangular. The matrix $\mathbf{A} := \mathbf{X}^*$ is a sparse tall matrix. LSMR only requires storing $\mathbf{X}^*$, $\mathbf{Y}^*$ and $\boldsymbol{\eta}^*$ and, unlike the conjugate gradient algorithm, avoids $\mathbf{X}^{*\top}\mathbf{X}^*$, $\mathbf{X}^{*\top}\mathbf{Y}$ and $\mathbf{X}^{*\top}\boldsymbol{\eta}$. LSMR also tends to produce more stable estimates than conjugate gradient. We have also tested a variety of conjugate gradient methods and preconditioning methods, where we have observed that their performances varied across different data sets.

The LSMR without conditioning showed a relatively good performance for the latent models. Therefore, we choose LSMR without preconditioning for our current illustrations.

Posterior predictive inference will adapt from (3.2.11) for scalable models. After sampling $\{\boldsymbol{\gamma}, \boldsymbol{\Sigma}\}$, we sample one draw of $\boldsymbol{\omega}_{\mathcal{U}} \sim \boldsymbol{\omega}_{\mathcal{U}} \,|\, \boldsymbol{\gamma}, \boldsymbol{\Sigma} \sim \mathrm{MN}([\mathbf{O}_{n' \times p}, \tilde{\mathbf{A}}]\boldsymbol{\gamma}, \tilde{\mathbf{D}}, \boldsymbol{\Sigma})$ for each sampled $\{\boldsymbol{\gamma}, \boldsymbol{\Sigma}\}$, where $\tilde{\mathbf{A}} = [\tilde{\mathbf{a}}_1 : \cdots : \tilde{\mathbf{a}}_n]^\top$, $\tilde{\mathbf{D}} = \mathrm{diag}(\{\tilde{d}_i\}_{i=1}^n)$ with

$$\begin{aligned} (\{\tilde{\mathbf{a}}_i\}_{j_1}, \ldots, \{\tilde{\mathbf{a}}_i\}_{j_m}) &= \boldsymbol{\rho}_\psi(\mathbf{u}_i, N_m(\mathbf{u}_i))\boldsymbol{\rho}_\psi^{-1}(N_m(\mathbf{u}_i), N_m(\mathbf{u}_i)) \;, \\ \tilde{d}_i &= 1 - \boldsymbol{\rho}_\psi(\mathbf{u}_i, N_m(\mathbf{u}_i))\boldsymbol{\rho}_\psi^{-1}(N_m(\mathbf{u}_i), N_m(\mathbf{u}_i))\boldsymbol{\rho}_\psi(N_m(\mathbf{u}_i), \mathbf{u}_i) \;. \end{aligned} \tag{3.2.18}$$

Finally, for each sampled $\{\boldsymbol{\beta}, \boldsymbol{\omega}_{\mathcal{U}}, \boldsymbol{\Sigma}\}$ we make one draw of $\mathbf{Y}_{\mathcal{U}} \sim \mathrm{MN}(\mathbf{X}_{\mathcal{U}}\boldsymbol{\beta} + \boldsymbol{\omega}_{\mathcal{U}}, (\alpha^{-1} - 1)\mathbf{I}_{n'}, \boldsymbol{\Sigma})$. The following provides details of the algorithm for predictive inference.

Algorithm 3.2: Obtaining posterior inference of $\{\boldsymbol{\gamma}, \boldsymbol{\Sigma}\}$ and predictions on set $\mathcal{U}$ for conjugate multivariate latent NNGP

```
1. Construct $\mathbf{X}^*$ and $\mathbf{Y}^*$ in (3.2.16)
   (a) $\mathbf{L}_r^{-1}$ and $\mathbf{L}_r^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}}$
       • Compute the Cholesky decomposition of $\mathbf{V}_r$, $\mathbf{L}_r$
       • Compute $\mathbf{L}_r^{-1}$ and $\mathbf{L}_r^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}}$
   (b) $\mathbf{V}_{\boldsymbol{\rho}}$
       • Construct $\mathbf{A}_{\boldsymbol{\rho}}$ and $\mathbf{D}_{\boldsymbol{\rho}}$ as described, for example, in Finley et al. [2019]    $\mathcal{O}(nm^3)$
       • Compute $\mathbf{V}_{\boldsymbol{\rho}} = \mathbf{D}_{\boldsymbol{\rho}}^{-\frac{1}{2}}(\mathbf{I} - \mathbf{A}_{\boldsymbol{\rho}})$    $\mathcal{O}(n(m+1))$
   (c) Construct $\mathbf{X}^*$ and $\mathbf{Y}^*$
2. Obtain $\boldsymbol{\mu}^*$, $\boldsymbol{\Psi}^*$ and $\nu^*$.
   (a) Obtain $\boldsymbol{\mu}^* = [\boldsymbol{\mu}_1^* : \cdots : \boldsymbol{\mu}_q^*]$
       • Solve $\boldsymbol{\mu}_i^*$ from $\mathbf{X}^*\boldsymbol{\mu}_i^* = \mathbf{Y}_i^*$ by LSMR for $i = 1, \ldots, q$.
   (b) Obtain $\boldsymbol{\Psi}^*$ and $\nu^*$
       • Generate $\mathbf{u} = \mathbf{Y}^* - \mathbf{X}^*\boldsymbol{\mu}^*$    $\mathcal{O}(n(1 + (p + m + 1)q))$
       • Compute $\boldsymbol{\Psi}^* = \boldsymbol{\Psi} + \mathbf{u}^\top\mathbf{u}$    $\mathcal{O}(nq^2)$
       • Compute $\nu^* = \nu + n$
3. Generate posterior samples of $\{\boldsymbol{\gamma}^{(l)}, \boldsymbol{\Sigma}^{(l)}\}_{l=1}^L$. For $l$ in $1 : L$
   (a) Sample $\boldsymbol{\Sigma}^{(l)} \sim \mathrm{IW}(\boldsymbol{\Psi}^*, \nu^*)$
   (b) Sample $\boldsymbol{\gamma}^{(l)} \sim \mathrm{MN}(\boldsymbol{\mu}^*, \mathbf{V}^*, \boldsymbol{\Sigma}^{(l)})$
       • Sample $\mathbf{u} \sim \mathrm{MN}(0, \mathbf{I}_{2n+p}, \mathbf{I}_q)$    $\mathcal{O}(2nq)$
       • Calculate Cholesky decomposition of $\boldsymbol{\Sigma}^{(l)}$, $\boldsymbol{\Sigma}^{(l)} = \mathbf{L}_{\boldsymbol{\Sigma}^{(l)}}\mathbf{L}_{\boldsymbol{\Sigma}^{(l)}}^\top$
       • Generate $\boldsymbol{\eta} = \mathbf{u}\mathbf{L}^{(l)\top} = [\boldsymbol{\eta}_1 : \cdots : \boldsymbol{\eta}_q]$    $\mathcal{O}(2nq^2)$
       • Solve $\mathbf{v}_i$ from $\mathbf{X}^*\mathbf{v}_i = \boldsymbol{\eta}_i$ by LSMR for $i = 1, \ldots, q$.
```

```
        • Generate $\boldsymbol{\gamma}^{(l)} = \boldsymbol{\mu}^* + \mathbf{v}$ with $\mathbf{v} = [\mathbf{v}_1 : \cdots : \mathbf{v}_q]$        $\mathcal{O}(nq)$

4. Generate posterior samples of $\{\mathbf{Y}_{\mathcal{U}}^{(l)}\}$ on a new set $\mathcal{U}$ given $\mathbf{X}_{\mathcal{U}}$.
   (a) Construct $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{D}}$ using (3.2.18)        $\mathcal{O}(n'm^3)$
   (b) For $l$ in $1 : L$
       i. Sample $\boldsymbol{\omega}_{\mathcal{U}}^{(l)} \sim \mathrm{MN}([\mathbf{0}_{n'\times p}, \tilde{\mathbf{A}}]\boldsymbol{\gamma}^{(l)}, \tilde{\mathbf{D}}, \boldsymbol{\Sigma}^{(l)})$
          • Sample $\mathbf{u} \sim \mathrm{MN}(\mathbf{0}, \mathbf{I}_{n'}, \mathbf{I}_q)$        $\mathcal{O}(n'q)$
          • Generate $\boldsymbol{\omega}_{\mathcal{U}}^{(l)} = [\mathbf{0}_{n'\times p}, \tilde{\mathbf{A}}]\boldsymbol{\gamma}^{(l)} + \tilde{\mathbf{D}}^{\frac{1}{2}}\mathbf{u}\mathbf{L}_{\boldsymbol{\Sigma}^{(l)}}^{\top}$        $\mathcal{O}(n'mq + n'q^2)$
       ii. Sample $\mathbf{Y}_{\mathcal{U}}^{(l)} \mid \boldsymbol{\omega}_{\mathcal{U}}^{(l)}, \boldsymbol{\gamma}^{(l)}, \boldsymbol{\Sigma}^{(l)} \sim \mathrm{MN}(\mathbf{X}_{\mathcal{U}}\boldsymbol{\beta} + \boldsymbol{\omega}_{\mathcal{U}}, (\alpha^{-1} - 1)\mathbf{I}_{n'}, \boldsymbol{\Sigma})$
          • Sample $\mathbf{u} \sim \mathrm{MN}(\mathbf{0}, I_{n'}, I_q)$        $\mathcal{O}(n'q)$
          • Generate $\mathbf{Y}_{\mathcal{U}}^{(l)} = \mathbf{X}_{\mathcal{U}}\boldsymbol{\beta} + \boldsymbol{\omega}_{\mathcal{U}}^{(l)} + (\alpha^{-1} - 1)\mathbf{u}\mathbf{L}_{\boldsymbol{\Sigma}^{(l)}}^{\top}$        $\mathcal{O}(n'pq + n'q^2)$
```

3.2.3 Cross-validation for Conjugate Multivariate NNGP Models

Conjugate Bayesian multivariate regression models will depend upon fixing hyperparameters in the model. Here, we apply a K-fold cross-validation algorithm for choosing $\{\psi, \alpha\}$. This algorithm is a straightforward generalization of the univariate algorithm in [Finley et al., 2019]. We run the conjugate models for each point $\{\psi, \alpha\}$ on a grid and choose the value that produces the least magnitude of root mean square prediction error. The inference on that point is then presented. This is appealing for scalable Gaussian process models that, for any fixed $\{\psi, \alpha\}$, can deliver posterior inference at new locations requiring storage and flops in $\mathcal{O}(n)$.

Algorithm 3.3: Cross-validation of tuning ψ, α for conjugate multivariate response or latent NNGP model

```
1. Split $\mathcal{S}$ into $K$ folds, and build neighbor index.
   • Split $\mathcal{S}$ into $K$ folds $\{\mathcal{S}_k\}_{k=1}^{K}$. We use $\mathcal{S}_{-k}$ to denote $\mathcal{S}$ without the locations in $\mathcal{S}_k$.
   • Build nearest neighbors for $\{\mathcal{S}_{-k}\}_{k=1}^{K}$
   • Find the collection of nearest neighbor set for $\mathcal{S}_k$ among $\mathcal{S}_{-k}$ for $k = 1, \ldots, K$.

2. (For response NNGP) Fix $\psi$ and $\alpha$, obtain posterior mean of $\boldsymbol{\beta}$ after removing the $k^{th}$ fold of the data:
   • Use step 1 in Algorithm 3.1 to obtain $\hat{\boldsymbol{\beta}}_k$ by taking $\mathcal{S}$ to be $\mathcal{S}_{-k}$ and $\boldsymbol{\mu}^*$ to be $\hat{\boldsymbol{\beta}}_k$.
   (For latent NNGP) Fix $\psi$ and $\alpha$, obtain posterior mean of $\boldsymbol{\gamma}_k = \{\boldsymbol{\beta}, \boldsymbol{\omega}(\mathcal{S}_{-k})\}$ after removing the $k^{th}$ fold of the data:
   • Use step 1-2 in Algorithm 3.2 to obtain $\hat{\boldsymbol{\gamma}}_k$ by taking $\mathcal{S}$ to be $\mathcal{S}_{-k}$ and $\boldsymbol{\mu}^*$ to be $\hat{\boldsymbol{\gamma}}_k$.

3. (For response NNGP) Predict posterior means of $\mathbf{y}(\mathcal{S}_k)$
   • Construct matrix $\tilde{\mathbf{A}}$ through (3.2.14) by taking $\mathcal{S}$ to be $\mathcal{S}_{-k}$ and $\mathcal{U}$ to be $\mathcal{S}_k$.
   • According to (3.2.15), the predicted posterior mean of $\mathbf{y}(\mathcal{S}_k)$ follows
     $\hat{\mathbf{y}}(\mathcal{S}_k) = \mathrm{E}[\mathbf{y}(\mathcal{S}_k) \mid \mathbf{y}(\mathcal{S}_{-k})] = \mathbf{x}(\mathcal{S}_k)\hat{\boldsymbol{\beta}}_k + \tilde{\mathbf{A}}[\mathbf{y}(\mathcal{S}_{-k}) - \mathbf{x}(\mathcal{S}_{-k})\hat{\boldsymbol{\beta}}_k]$
```

(For latent NNGP) Predict posterior means of $\mathbf{y}(\mathcal{S}_k)$

- Construct matrix $\tilde{\mathbf{A}}$ by taking $\mathcal{S}$ to be $\mathcal{S}_{-k}$ and $\mathcal{U}$ to be $\mathcal{S}_k$.
- The predicted posterior mean of $\mathbf{y}(\mathcal{S}_k)$ follows
  $\hat{\mathbf{y}}(\mathcal{S}_k) = \mathrm{E}[\mathbf{y}(\mathcal{S}_k) \,|\, \mathbf{y}(\mathcal{S}_{-k})] = \mathrm{E}_{\boldsymbol{\omega}}[\mathrm{E}_{\mathbf{y}}[\mathbf{y}(\mathcal{S}_k) \,|\, \boldsymbol{\omega}(\mathcal{S}_{-k}), \mathbf{y}(\mathcal{S}_{-k})]] = [\mathbf{x}(\mathcal{S}_k), \tilde{\mathbf{A}}]\hat{\boldsymbol{\gamma}}_k$

4. Root Mean Square Predictive Error (RMSPE) over K folds

```
• Initialize $e = 0$
    for ($k$ in $1 : K$)
        for ($\mathbf{s}_i$ in $\mathcal{S}_k$)
            $e = e + \|\mathbf{y}(\mathbf{s}_i) - \hat{\mathbf{y}}(\mathbf{s}_i)\|^2$
```

5. Cross validation for choosing ψ and α

   - Repeat steps (2) - (4) for all candidate values of ψ and α
   - Choose ψ_0 and α_0 as the value that minimizes the average RMSPE

3.2.4 Comparison of Response and Latent Models

Modeling the response as an NNGP produces a different model from modeling the latent process as an NNGP. In the former, Vecchia's approximation to the joint density of the response yields a sparse precision matrix for the response. In the latter, it is the precision matrix of the realizations of the latent process that is sparse. This has been discussed in Datta et al. [2016a] and also explored in greater generality by Katzfuss and Guinness [2017]. Comparisons based on the Kullback-Leibler divergence (KL-D) between the NNGP based models and their parent full GP models reveal that the latent NNGP model tends to be closer to the full GP than the response NNGP. A proof of such a result is provided by Katzfuss and Guinness [2017], but this result holds only in the context of an augmented directed acyclical graphical model with nodes comprising the response and the latent variables. However, if we compute the KL-D between the NNGP models and their full GP counterparts in terms of the collapsed or marginal distribution for $\mathbf{Y}$, then it is theoretically possible for the response model to be closer to the full GP.

Here we provide a simple example where a response NNGP model outperforms a latent NNGP model on a collapsed space. Assume the observed location set is $\mathcal{S} = \{\mathbf{s}_1, \mathbf{s}_2, \mathbf{s}_3\}$, $\omega(\mathcal{S})$ has covariance matrix $\sigma^2\mathbf{R}$ with correlation matrix

$$\mathbf{R} = \begin{bmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{bmatrix} . \tag{3.2.19}$$

Let us suppress the connection between knots $\mathbf{s}_1$ and $\mathbf{s}_3$ in the directed acyclic graph corresponding to the finite realization of the NNGP on $\mathcal{S}$. Then the covariance matrix of of the response NNGP model $\mathbf{\Sigma}_R$ and that of the latent NNGP model $\mathbf{\Sigma}_l$ have the following forms:

$$\mathbf{\Sigma}_R = \sigma^2 \begin{bmatrix} 1+\delta^2 & \rho_{12} & \frac{\rho_{12}\rho_{23}}{1+\delta^2} \\ \rho_{12} & 1+\delta^2 & \rho_{23} \\ \frac{\rho_{12}\rho_{23}}{1+\delta^2} & \rho_{23} & 1+\delta^2 \end{bmatrix} , \; \mathbf{\Sigma}_l = \sigma^2 \begin{bmatrix} 1+\delta^2 & \rho_{12} & \rho_{12}\rho_{23} \\ \rho_{12} & 1+\delta^2 & \rho_{23} \\ \rho_{12}\rho_{23} & \rho_{23} & 1+\delta^2 \end{bmatrix} , \tag{3.2.20}$$

where $\delta^2 = \frac{\tau^2}{\sigma^2}$ is the noise-to-signal ratio with τ^2 as the variance of the noise process $\epsilon(s)$. Since R is positive-definite, we must have

$$1 - (\rho_{12}^2 + \rho_{13}^2 + \rho_{23}^2) + 2\rho_{12}\rho_{13}\rho_{23} > 0 \; , \; 1 - \rho_{12}^2 > 0 \; . \tag{3.2.21}$$

It is easy to show that $\mathbf{\Sigma}_R$ and $\mathbf{\Sigma}_l$ are also positive-definite. If $\rho_{13} = \frac{\rho_{12}\rho_{23}}{1+\delta^2}$, then the KL-D from the response NNGP model to the true model always equals zero, which is no more than the KL-D from the latent NNGP model to the true model. If $\rho_{13} = \rho_{12}\rho_{23}$, then the KL-D of the latent NNGP model to the true model always equals zero, which reverses the relationship. Numerical examples can be found in `https://luzhangstat.github.io/notes/KL-D_com.html`

Still, our simulations indicate that the latent NNGP model tends to outperform the response NNGP model in approximating their parent GP based models. This is consistent with the theoretical result of [Katzfuss and Guinness, 2017] and also with our intuition: the presence of the latent process should certainly improve the goodness of fit of the model. Without loss of generality, our discussion here considers the univariate case, but the argument applies to the multivariate setting as well. Let $\{y(\mathbf{s}) : \mathbf{s} \in \mathcal{D}\}$ be the process of interest over $\mathcal{D} \subset \mathbb{R}^d, d \in N^+$, and let $y(\mathbf{s}) = \omega(\mathbf{s}) + \epsilon(\mathbf{s})$ for some latent spatial GP $\omega(\mathbf{s})$ and white noise process $\epsilon(\mathbf{s})$. A response NNGP model specifies the NNGP on $y(\mathbf{s})$, while a latent NNGP model assumes that $\omega(\mathbf{s})$ follows the NNGP.

Let the covariance matrix of $\mathbf{y} = y(\mathcal{S})$ of the parent GP based models be $\mathbf{C} + \tau^2\mathbf{I}$, where $\mathbf{C}$ is the covariance matrix of the latent process $\omega(\mathcal{S})$. Consider the Vecchia approximation of the precision matrices $\mathbf{C}^{-1}$ and $\mathbf{K}^{-1} = \{\mathbf{C} + \tau^2\mathbf{I}\}^{-1}$:

$$\text{Vecchia}(\mathbf{C}^{-1}) = \tilde{\mathbf{C}}^{-1}\ ,\ \text{Vecchia}(\mathbf{K}^{-1}) = \tilde{\mathbf{K}}^{-1}\ . \tag{3.2.22}$$

The covariance matrix of $y(\mathcal{S})$ from the latent NNGP model is $\tilde{\mathbf{C}} + \tau^2\mathbf{I}$, while the precision matrix of $y(\mathcal{S})$ from the response NNGP model is $\tilde{\mathbf{K}}^{-1}$. We denote the error matrix of the Vecchia approximation of $\mathbf{C}^{-1}$ by $\mathbf{E}$. We assume that $\mathbf{E}$ is small so that $\tilde{\mathbf{C}}^{-1}$ approximates $\mathbf{C}^{-1}$ well. With the same observed location $\mathcal{S}$ and the fixed number of nearest neighbors, the error matrix of the Vecchia approximation of $\mathbf{K}^{-1}$ is believed to be close to $\mathbf{E}$, i.e.,

$$\mathbf{C}^{-1} = \tilde{\mathbf{C}}^{-1} + \mathbf{E}\ ;\ \mathbf{K}^{-1} = \tilde{\mathbf{K}}^{-1} + \mathcal{O}(\mathbf{E}). \tag{3.2.23}$$

Representing the precision matrices of $y(\mathcal{S})$ of the parent GP based model and the latent NNGP model by

$$\begin{aligned}(\mathbf{C} + \tau^2\mathbf{I})^{-1} &= \mathbf{C}^{-1} - \mathbf{C}^{-1}\mathbf{M}^{-1}\mathbf{C}^{-1}\ , \mathbf{M} = \mathbf{C}^{-1} + \tau^{-2}\mathbf{I}\ ,\\ (\tilde{\mathbf{C}} + \tau^2\mathbf{I})^{-1} &= \tilde{\mathbf{C}}^{-1} - \tilde{\mathbf{C}}^{-1}\mathbf{M}^{*-1}\tilde{\mathbf{C}}^{-1}\ , \mathbf{M}^* = \tilde{\mathbf{C}}^{-1} + \tau^{-2}\mathbf{I}\ ,\end{aligned} \tag{3.2.24}$$

we find that the difference between the precision metrics over the collapsed space for the parent NNGP and for the latent NNGP model is

$$\begin{aligned}(\mathbf{C} + \tau^2\mathbf{I})^{-1} - (\tilde{\mathbf{C}} + \tau^2\mathbf{I})^{-1} &= \mathbf{C}^{-1} - \mathbf{C}^{-1}\mathbf{M}^{-1}\mathbf{C}^{-1} - \tilde{\mathbf{C}}^{-1} + \tilde{\mathbf{C}}^{-1}\mathbf{M}^{*-1}\tilde{\mathbf{C}}^{-1}\\ &= \underbrace{\mathbf{E} - \mathbf{E}\mathbf{M}^{-1}\tilde{\mathbf{C}}^{-1} - \tilde{\mathbf{C}}^{-1}\mathbf{M}^{-1}\mathbf{E} - \tilde{\mathbf{C}}^{-1}(\mathbf{M}^{-1} - \mathbf{M}^{*-1})\tilde{\mathbf{C}}^{-1}}_{\mathbf{B}} - \underbrace{\mathbf{E}\mathbf{M}^{-1}\mathbf{E}}_{\mathcal{O}(\mathbf{E}^2)}\end{aligned}$$

Representing $\mathbf{B}$ in terms of $\tilde{\mathbf{C}}^{-1}$, $\mathbf{M}^*$ and $\mathbf{E}$, where $\mathbf{E}$ is assumed to be nonsingular, we find

$$\begin{aligned}\mathbf{B} = {} & \mathbf{E} - \mathbf{E}\mathbf{M}^{*-1}\tilde{\mathbf{C}}^{-1} + \mathbf{E}\mathbf{M}^{*-1}(\mathbf{E}^{-1} + \mathbf{M}^{*-1})^{-1}\mathbf{M}^{*-1}\tilde{\mathbf{C}}^{-1} - \tilde{\mathbf{C}}^{-1}\mathbf{M}^{*-1}\mathbf{E}\\ & + \tilde{\mathbf{C}}^{-1}\mathbf{M}^{*-1}(\mathbf{E}^{-1} + \mathbf{M}^{*-1})^{-1}\mathbf{M}^{*-1}\mathbf{E} + \tilde{\mathbf{C}}^{-1}\mathbf{M}^{*-1}(\mathbf{E}^{-1} + \mathbf{M}^{*-1})^{-1}\mathbf{M}^{*-1}\tilde{\mathbf{C}}^{-1}\ .\end{aligned} \tag{3.2.25}$$

Using the familiar Woodbury matrix identity and the expansion $(\mathbf{I}+\mathbf{X})^{-1} = \sum_{n=0}^{\infty}\{-\mathbf{X}\}^n$, we find

$$\begin{aligned}(\mathbf{E}^{-1}+\mathbf{M}^{*-1})^{-1}\mathbf{M}^{*-1} &= \{\mathbf{M}^*(\mathbf{E}^{-1}+\mathbf{M}^{*-1})\}^{-1} = \{\mathbf{M}^*\mathbf{E}^{-1}+\mathbf{I}\}^{-1}\\ &= \mathbf{I}-\{\mathbf{I}+\mathbf{E}\mathbf{M}^{*-1}\}^{-1} = \mathbf{I}-\{\mathbf{I}-\mathbf{E}\mathbf{M}^{*-1}+\mathcal{O}(\mathbf{E}^2)\}\\ &= \mathbf{E}\mathbf{M}^{*-1}+\mathcal{O}(\mathbf{E}^2)\ .\end{aligned}$$

Using the above equations and excluding the terms of order $\mathcal{O}(\mathbf{E}^2)$ in the expression of $\mathbf{B}$, the leading term in the difference is

$$\mathbf{B} = (\mathbf{I}-\tilde{\mathbf{C}}^{-1}\mathbf{M}^{*-1})\mathbf{E}(I-\mathbf{M}^{*-1}\tilde{\mathbf{C}}^{-1}) = (\mathbf{I}+\tau^2\tilde{\mathbf{C}}^{-1})^{-1}\mathbf{E}(\mathbf{I}+\tau^2\tilde{\mathbf{C}}^{-1})^{-1}\ . \tag{3.2.26}$$

Using the spectral decomposition $(\mathbf{I}+\tau^2\tilde{\mathbf{C}}^{-1}) = \mathbf{P}^\top(\mathbf{I}+\tau^2\mathbf{D})\mathbf{P}$, where $\mathbf{P}$ is orthogonal and $\mathbf{D}$ is diagonal with positive elements on the diagonal, we obtain

$$\begin{aligned}\|\mathbf{B}\|_F &= \|\mathbf{P}^\top(\mathbf{I}+\tau^2\mathbf{D})^{-1}\mathbf{P}\mathbf{E}\mathbf{P}^\top(\mathbf{I}+\tau^2\mathbf{D})^{-1}\mathbf{P}\|_F = \|(\mathbf{I}+\tau^2\mathbf{D})^{-1}\mathbf{P}\mathbf{E}\mathbf{P}^\top(\mathbf{I}+\tau^2\mathbf{D})^{-1}\|_F\\ &\leq \|\mathbf{P}\mathbf{E}\mathbf{P}^\top\|_F = \|\mathbf{E}\|_F\ ,\end{aligned} \tag{3.2.27}$$

where $\|\cdot\|_F$ denotes the Frobenius matrix norm. The inequality also holds for the absolute value of the determinant and p norms. And the equality holds if and only if $\tau^2 = 0$ when the difference is the same as the error matrix for response NNGP model. Thus, the latent model tends to shrink the error from the Vecchia approximation, which explains the expected superior performance of the latent NNGP model over the response NNGP model based on KL-Ds.

3.3 Simulation

Our proposed models were implemented in `Julia` 1.2.0 [Bezanson et al., 2017]. All models were run on a Linux environment (Ubuntu 18.04.2 LTS), with 32 Gbytes of random-access memory and 1 Intel Core i7-7700K CPU @ 4.20GHz processor with 4 cores each and 2 threads per core - totaling 8 possible threads for use in parallel. Model diagnostics and other posterior summaries were implemented within the Julia statistical environment and R 3.6.1. Each model was compared in terms of the posterior inference of parameters (posterior mean and 95% confidence interval), root mean squared predict error (RMSPE= $n^{-1}\sum_{i=1}^{n}((y_j(\mathbf{s}_i)-\hat{y}_j(\mathbf{s}_i))^2)^{\frac{1}{2}}, j = 1,\ldots,q$), mean squared error of intercept-centered latent processes (MSEL = $n^{-1}\sum_{i=1}^{n}((\omega_j(\mathbf{s}_i)+\boldsymbol{\beta}_{1j}-\hat{\omega_j}(\mathbf{s}_i)-\hat{\boldsymbol{\beta}}_{1j})^2)^{\frac{1}{2}}, j = 1,\ldots,q$), prediction interval coverage (CVG; the percent of intervals containing the true value), interval coverage for intercept-centered latent process of observed response (CVGL), mean continuous rank probability score (MCRPS $= n'^{-1}\sum_{i=1}^{n'}\mathrm{CRPS}_j(u_i), j = 1,\ldots,q,$ where $\mathrm{CPRS}_j(\mathbf{u}_i)$ is the CRPS of j-th response on held location $\mathbf{u}_i$ see Gneiting and Raftery [2007]), and run time. To calculate $\mathrm{CRPS}_j(\mathbf{u}_i)$, we assumed the associated predictive distribution was well approximated by a Gaussian distribution with mean centered at the predicted value $\hat{y}_j(\mathbf{u}_i)$ and standard deviation equal to the predictive standard error $\hat{\sigma}_j(\mathbf{u}_i)$, $\mathrm{CPRS}_j(\mathbf{u}_i) = \hat{\sigma}_j(\mathbf{u}_i)[1/\sqrt{\pi}-2\varphi(z_{ij})-z_{ij}(2\Phi(z_{ij})-1)]$ with $z_{ij} = (y_j(\mathbf{u}_i) - \hat{y}_j(\mathbf{u}_i))/\hat{\sigma}_j(\mathbf{u}_i)$, φ and Φ denoting the probability density function and the cumulative distribution function of a standard Gaussian variable. All NNGP models in this section specified at most $m = 10$ nearest neighbors.

We simulated $\mathbf{y}(\mathbf{s})$ using model (3.2.8) with $q = 2, p = 2$ over 1200 randomly generated locations inside a unit square. The design matrix $\mathbf{X}$ consisted of a column of 1's and a single predictor generated from a standard normal distribution. An exponential covariance function with decay ϕ was used to model $\rho_\psi(\cdot,\cdot)$ in (3.2.8), i.e., $\rho_\psi(\mathbf{s}',\mathbf{s}'') = \exp\left(-\phi\|\mathbf{s}'-\mathbf{s}''\|\right)$, for $\mathbf{s}',\mathbf{s}'' \in \mathcal{D}$, with $\|\mathbf{s}'-\mathbf{s}''\|$ be the L^2 norm of $\mathbf{s}'-\mathbf{s}''$ and $\psi = \phi$. The value of parameters are listed in table 3.1. We withheld 200 locations to assess predictive performance for conjugate models and benchmark models. NNGP based BSLMC model was also tested here for a comparison.

We assigned a flat prior for $\boldsymbol{\beta}$, the prior of $\boldsymbol{\Sigma}$ was set to follow $\mathrm{IW}(\boldsymbol{\Psi},\nu)$ with $\boldsymbol{\Psi} = \mathrm{diag}([1.0, 1.0])$ and $\nu = 3$. The candidate values for $\{\phi,\alpha\}$ used in cross-validation algorithm were a 25 by 25 grid over $[2.12, 26.52] \times [0.8, 0.99]$. The posterior inference of conjugate response and latent NNGP models were based on 500 samples. The run times for conjugate models include the time for choosing hyper-parameters through cross-validation and the time for the sampling process. We summarize posterior inference for regression coefficients $\boldsymbol{\beta} = \{\boldsymbol{\beta}_{ij}\}_{i=1,j=1}^{p,q}$, covariance of measurement error (labeled as $\text{cov}(\boldsymbol{\epsilon})$ in summary table), covariance across different latent processes (labeled as $\text{cov}(\boldsymbol{\omega})$ in summary table) and hyperparameters $\{\phi, \alpha\}$ in Table 3.1.

Table 3.1 lists the parameter estimates and performance metrics of the candidate models. The posterior inference of regression slopes $\{\boldsymbol{\beta}_{21}, \boldsymbol{\beta}_{22}\}$ are close among two models. The 95% confidence intervals of the intercepts $\{\boldsymbol{\beta}_{11}, \boldsymbol{\beta}_{12}\}$ all include the actual value.

The interpolated map of the recovered posterior mean of latent processes (figure 3.1) capture the patterns of the underlying latent processes. The conjugate NNGP models all yielded close RMSPEs and MCRPSs. The CVG and CVGL are close to 0.95, supporting reliable inference from conjugate NNGP models. The two models finished within a minute. The simulation example shows that fitting a conjugate model is a pragmatic method for quick inference in multivariate spatial data analysis.

3.4 Vegetation Indices Data Analysis

We implement all proposed models on a real dataset to test their performances in a realistic analysis scenario. Our dataset comprises Vegetation Indices data and landcover data [see Ramon Solano et al., 2010, Sulla-Menashe and Friedl, 2018, for further details]. The Vegetation Indices data records the standard Normalized Difference Vegetation Index (NDVI) and Enhanced Vegetation Index (EVI). These two indices are robust, empirical measures of vegetation activity at the land surface, that are studied for an understanding of the global distribution of vegetation types as well as their biophysical and structural properties and spatial/temporal variations [Ramon Solano et al., 2010]. Provided along with the two vegetation indexes are red reflectance, near-infrared (NIR) reflectance, blue reflectance mid-infrared (MIR) reflectance, view zenith angle, sun zenith angle and relative azimuth angle. All data were mapped to Euclidean coordinates using the Sinusoidal (SIN) grid projection. We chose zone *h08v05* which covers 11,119,505 to 10,007,555 meters south of the prime meridian and

Table 3.1: Simulation 1 study summary table: posterior mean (2.5%, 97.5%) percentiles

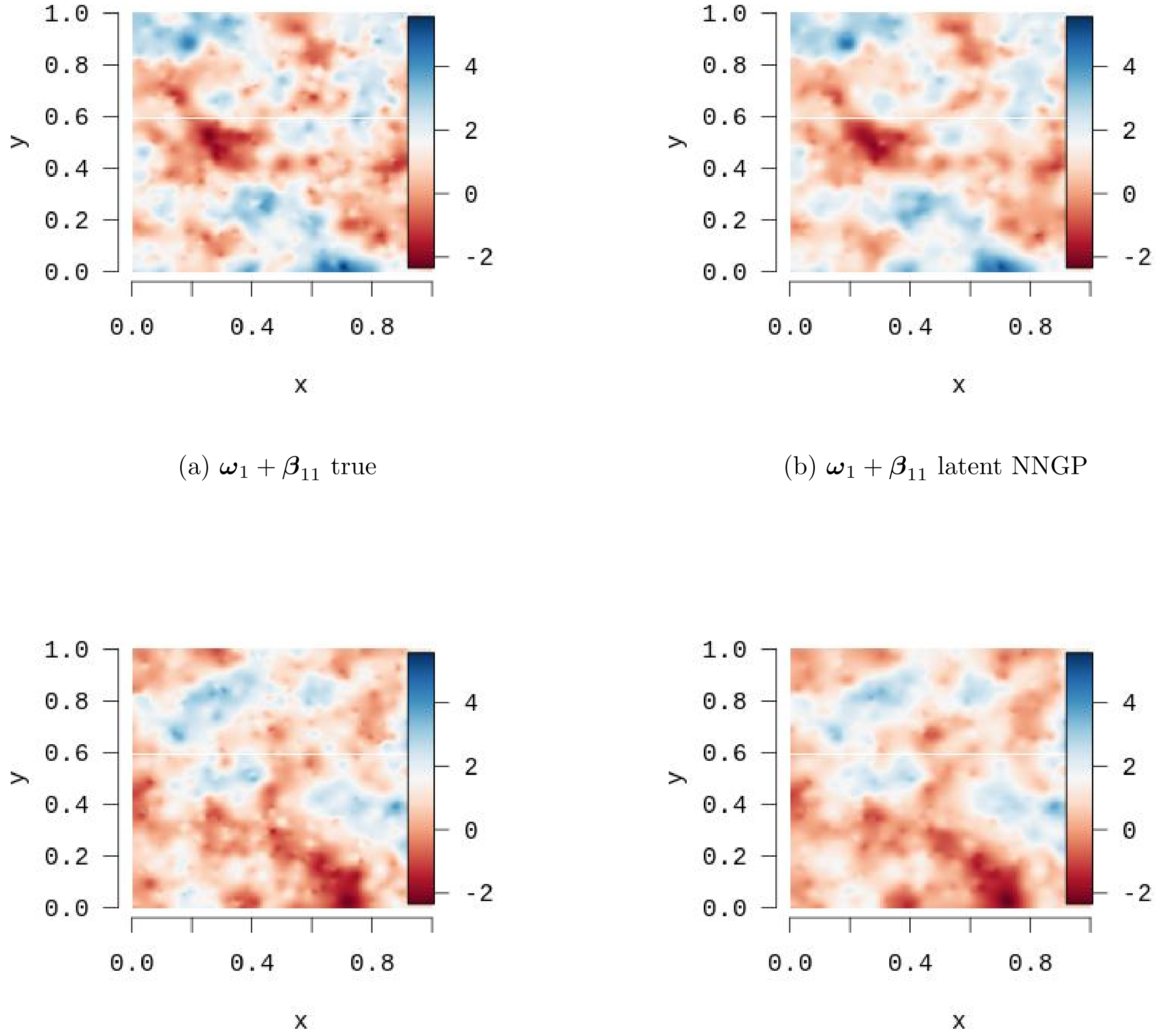
	True	Conj resp	Conj latent
$\boldsymbol{\beta}_{11}$	1.0	1.391 (0.814, 1.902)	1.459 (0.865, 2.057)
$\boldsymbol{\beta}_{12}$	-1.0	0.813 (0.344, 1.286)	0.734(0.201, 1.276)
$\boldsymbol{\beta}_{21}$	-2.0	-1.978 (-2.114, -1.841)	-1.979 (-2.121, -1.842)
$\boldsymbol{\beta}_{22}$	2.0	2.076 (1.952, 2.21)	2.082 (1.961, 2.208)
$\text{cov}(\boldsymbol{\epsilon})_{11}$	0.222	0.226 (0.205, 0.248)	0.231 (0.212, 0.252)
$\text{cov}(\boldsymbol{\epsilon})_{12}$	-0.111	-0.113(-0.129, -0.099)	-0.115(-0.128, -0.103)
$\text{cov}(\boldsymbol{\epsilon})_{22}$	0.167	0.172 (0.158, 0.188)	0.175 (0.16, 0.189)
$\text{cov}(\boldsymbol{\omega})_{11}$	1.234	–	1.208 (1.148, 1.268)
$\text{cov}(\boldsymbol{\omega})_{12}$	-0.701	–	-0.705(-0.75, -0.658)
$\text{cov}(\boldsymbol{\omega})_{22}$	1.077	–	1.077 (1.023, 1.131)
ϕ	6.0	8.220	7.204
α	0.9	0.863	0.871
RMSPE[a]	–	[0.727, 0.602, 0.668]	[0.723, 0.6, 0.664]
MSEL	–	–	[0.112, 0.112, 0.103]
CVG[a]	–	[0.935, 0.955, 0.945]	[0.925, 0.95, 0.9375]
CVGL[a]	–	–	[0.957, 0.945, 0.951]
MCRPS[a]	–	[-0.408, -0.336, -0.372]	[-0.405, -0.334, -0.37]
time(s)	–	[12, 1][b]	[17, 1][b]

[a] [response 1, response 2, all responses]

[b] [time for cross-validation, time for sampling]

[c] [time for MCMC sampling, time for recovering $\boldsymbol{\beta}$ and predictions]

3,335,852 to 4,447,802 meters north of the equator. The land region in zone *h08v05* is the western United States. We generated a dummy variable for no vegetation or urban area through the 2016 landcover data, and took it along with the intercept as the explanatory variables in the analysis. All other data were measured through MODIS satellite over a 16-days period from 2016.04.06 to 2016.04.21. Some variables were rescaled and transformed



(a) $\boldsymbol{\omega}_1 + \boldsymbol{\beta}_{11}$ true

(b) $\boldsymbol{\omega}_1 + \boldsymbol{\beta}_{11}$ latent NNGP

(c) $\boldsymbol{\omega}_2 + \boldsymbol{\beta}_{12}$ true

(d) $\boldsymbol{\omega}_2 + \boldsymbol{\beta}_{12}$ latent NNGP

Figure 3.1: Interpolated maps of (a) & (c) the true generated latent processes and the posterior means of the spatial latent process $\boldsymbol{\omega}$ from the (b) & (d) conjugate latent NNGP model. The NNGP based models were all fit using $m = 10$ nearest neighbors.

in exploratory data analysis for the sake of better model fitting. The datasets were downloaded using the `R` package *MODIS*, and the code for exploratory data analysis is provided on `https://github.com/LuZhangstat/Conj_Multi_NNGP`.

There are 3,115,934 observed locations. We chose transformed NDVI ($\log(\text{NDVI} + 1)$ labeled as NDVI) and red reflectance (red refl) as responses. Bayesian linear models were fitted for comparison. All NNGP based models specified at most $m = 10$ nearest neighbors. We randomly held out 1% of observed locations and then held all responses over region 10,400,000 to 10,300,000 meters south of the prime meridian and 3,800,000 to 3,900,000 meters north of the equator to examine the predictive performance of models on randomly missing locations and a missing region. There were in total 67,132 locations held for prediction. Figure 3.2a illustrates the map of the transformed NDVI data. The white square region within the Continent is the region held out for prediction.

Posterior inference from our conjugate models were based on 500 independent samples from the posterior distribution. Recall that the samples are directly drawn from the conjugate posterior distribution and, hence, there is no need to monitor convergence of these samples. The priors for all parameters, except the decay, follow those in the simulation section. We recursively shrink the domain and grid of candidate values $\{\phi, \alpha\}$ through repeatedly using cross-validation algorithms for fixing parameters. The recorded run time for running the cross-validation algorithms, therefore, varied a lot across different models. The number of threads used in the cross-validation algorithm for conjugate models and response NNGP models with misalignment were equal to the number of folders. The remaining part of all the code were run with single thread.

The results for the conjugate models are listed in Table 3.2. Consistent with the related background, the regression coefficients of the index of no vegetation or urban area show relatively low biomass (low NDVI) and high red reflectance over no vegetation or urban area. The inference of the covariance of the noise and non-spatial covariance of the latent process shows a negative association between the residuals and latent processes of transformed NDVI and red reflectance, which satisfies the underlying relationship between two responses. The maps of the latent processes recovered by conjugate latent NNGP shown in Figure 3.2 also support this relationship.

Model performances were compared in terms of RMSPE, CVG, MCRPS and run time. The spatial models, unsurprisingly, greatly improved predictive accuracy. Conjugate Bayesian spatial models effected 35% shrinkage over the (non-spatial) Bayesian linear model in the magnitude of RMSPE. The performance in terms of the CVG is similar among all the models, but all the spatial models provided more accurate predictions than the Bayesian linear models based on MCRPS. Visual inspections of the predictive surfaces based on conjugate response NNGP model are shown in Figure 3.2. Notably, the proposed methods smooth out the predictions in the held-out region.

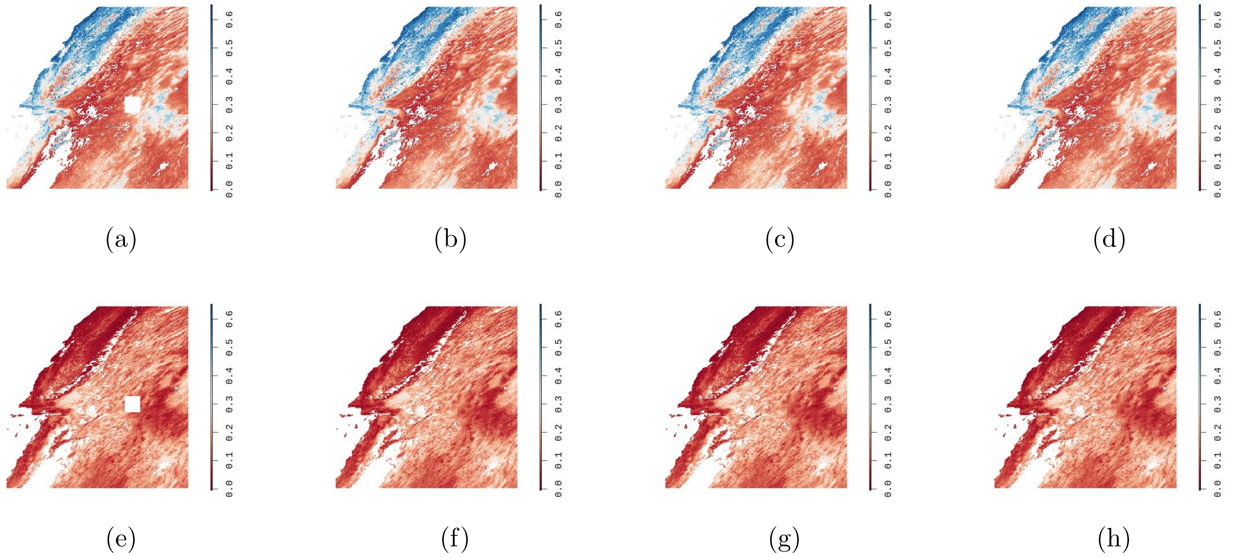
Posterior sampling for the conjugate response and latent models cost 1.8 and 18.88 minutes, respectively, which is impressive given our sample sizes of around 3 million locations. The run time for both the cross-validation algorithm and sampling for conjugate models is appealing for such massive datasets.

Table 3.2: Real data analysis summary table 1: posterior mean (2.5%, 97.5%) percentiles

	Bayesian linear model	conj response	conj latent
intercept$_1$	0.25144 (0.25131, 0.25158)	0.1023(0.0822, 0.1223)	0.240729 (0.240723, 0.240736)
intercept$_2$	0.13951 (0.13944, 0.13958)	0.2218(0.2094, 0.2338)	0.144277 (0.144273, 0.144281)
no vege or urban area$_1$	-0.13375(-0.13425, -0.13329)	-8.010e-3(-8.233e-3, -7.796e-3)	-8.025e-3 (-8.050e-3, -8.001e-3)
no vege or urban area$_2$	6.026e-2(6.002e-2, 6.052e-2)	4.381e-3 (4.261e-3, 4.514e-3)	4.390e-3 (4.376e-3, 4.402e-3)
$\text{cov}(\boldsymbol{\epsilon})_{11}$	1.599e-2 (1.596e-2, 1.602e-2)	3.493e-5 (3.487e-5, 3.499e-5)	3.125e-5 (3.120e-5, 3.130e-5)
$\text{cov}(\boldsymbol{\epsilon})_{12}$	-6.494e-3 (-6.505e-3, -6.483e-3)	-1.214e-5 (-1.217e-5, -1.212e-5)	-1.086e-5 (-1.089e-5, -1.085e-5)
$\text{cov}(\boldsymbol{\epsilon})_{22}$	3.656e-3 (3.651e-3, 3.662e-3)	1.090e-5(1.089e-5, 1.092e-5)	9.760e-6 (9.745e-6, 9.776e-6)
$\text{cov}(\boldsymbol{\omega})_{11}$	–	7.776e-2 (7.764e-2, 7.789e-2)	1.7192e-2 (1.7190e-2, 1.7193e-2)
$\text{cov}(\boldsymbol{\omega})_{12}$	–	-2.703e-2 (-2.709e-2, -2.697e-2)	-7.0307e-3(-7.0314e-3, -7.03e-3)
$\text{cov}(\boldsymbol{\omega})_{22}$	–	2.428e-2(2.424e-2, 2.432e-2)	3.8897e-3 (3.8893e-3, 3.8901e-3)
ϕ	–	17.919 ($\alpha = 0.999551$)	20.1755 ($\alpha = 0.999551$)
RMSPE^a	[0.09899 0.04932 0.07821]	[0.05707 0.03187 0.04622]	[0.0503 0.02572 0.03995]
MCRPS^a	[-0.05588 -0.02818 -0.04203]	[-0.03301 -0.0188 -0.02591]	[-0.0314 -0.01748 -0.02444]
CVG^a	[0.9664 0.9847 0.9755]	[0.9756 0.9707 0.9732]	[0.9764 0.9715 0.974]
time(mins)^b	–	[1012.18, 1.8]	[270.28, 18.88]

[a] [1st response transformed NDVI, 2nd response red reflectance, all responses]

[b] [time for cross-validation, time for generating 500 samples]



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 3.2: Colored NDVI and red reflectance images (first and second row respectively) of western United States (zone h08v05). Maps of raw data (a) & (e), raw data with predictions fitted by NNGP based conjugate response model (b) & (f), raw data with predictions fitted by NNGP based conjugate latent model (c) & (g) and the posterior mean of the intercept-centered latent process recovered from NNGP based conjugate latent model (d) & (h).

3.5 Discussion

We have presented a conjugate Bayesian multivariate spatial regression model using Matrix-Normal and Inverse-Wishart distributions in this Chapter. A specific contribution is to embed the latent spatial process within an augmented Bayesian multivariate regression to obtain posterior inference for the high-dimensional latent process with stochastic uncertainty quantification. For scalability to massive spatial datasets—our examples here comprise locations in the millions—we adopt the increasingly popular Vecchia approximation and, more specifically, the NNGP models that render savings in terms of storage and floating point operations. We present elaborate simulation experiments to test the performance of different models using datasets exhibiting different behaviors. Our conjugate modeling framework fixes hyperparameters using a K-fold cross-validation approach. While our analysis is based upon fixing these hyperparameters, the subsequent inference obtained is seen to be effective in capturing the features of the generating latent process (in our simulation experiments) and is orders of magnitude faster than iterative alternatives at such massive scales as ours. We also applied our models, and compared them, in our analysis of an NDVI dataset. The scalability of our approach is guaranteed when univariate scalable model can exploit a tractable precision or covariance matrix. Our approach can, thereofore, incorporate other methods such as multiresolution approximation (MRA) and more general Vecchia-type of approximations [see, e.g. Katzfuss and Guinness, 2017].

Future work can extend and adapt this framework to univariate and multivariate spatiotemporal modeling. A modification is to use a dynamic nearest-neighbor Gaussian process (DNNGP) [Datta et al., 2016b] instead of the NNGP in our models, which dynamically learns about space-time neighbors rather than fixing them. We can also develop conjugate Bayesian modeling frameworks for spatially-varying coefficient models, where the regression coefficients $\boldsymbol{\beta}$ are themselves random fields capturing the spatially-varying impact of predictors on the vector of outcomes. While conceptually straighforward, their actual implementation at massive scales will require substantial development.

Developments in scalable statistical models must be accompanied by explorations in high performance computing. While the algorithms presented here are efficient in terms of storage and flops, they have been implemented on modest hardware. Implementations exploiting Graphical Processing Units (GPUs) and parallel CPUs can be further explored. For the latent NNGP models, the algorithms relied upon sparse solvers such as conjugate gradients and LSMR matrix algorithms. Adapting such libraries to GPUs and other high performance computing hardware will need to be explored and tested further in the context of our spatial Gaussian process models.

Supplementary Material

All computer programs implementing the examples in this Chapter can be found in the public domain and downloaded from `https://github.com/LuZhangstat/Conj_Multi_NNGP`.

CHAPTER 4

Spatial Factor Modeling: A Bayesian Matrix-Normal Approach for Misaligned data

4.1 Introduction

This Chapter develops and investigates a new class of hierarchical models for analyzing multiple spatially oriented variables in high-dimensional settings. We address multivariate spatial modeling in high-dimensional settings, where we have measured a potentially large number of dependent variables over a massive number of locations. While statistical methods for analyzing massive spatial and spatial-temporal databases have received much attention [see, e.g., Sun et al., 2011, Banerjee, 2017, Heaton et al., 2019, and references therein for an account of the expanding literature in this domain], the bulk of these methods have focused on one or very few (two or three) spatially dependent variables. For even moderately large number of dependent variables (e.g. in tens or hundreds), modeling the cross-covariance matrix becomes challenging as it needs to capture the associations among all of the dependent variables over each pair of locations. Even for stationary cross-covariance functions, where we assume that the associations among the variables do not change over space and the spatial association for each variable depends only on the translation vector connecting two locations, matters become computationally challenging.

In this Chapter we devise scalable modeling strategies for multivariate spatial models. The modeling approach we develop in this Chapter enriches the popular linear models of coregionalization [Bourgault and Marcotte, 1991, Wackernagel, 2003, Gelfand et al., 2004, Chiles and Delfiner, 2009, Genton and Kleiber, 2015] using a Matrix-Normal prior to model the linear transformation on latent spatial processes. A key contribution here is that we provide a fully model-based enhancements for misaligned data, where not all responses are recorded over the same set of locations. We further expand this contribution by using the Matrix-Normal family to model the loading matrix in spatial factor models. In the latter context, our current contribution can be seen as enhancements to earlier contributions by Lopes et al. [2008], Ren and Banerjee [2013] and Taylor-Rodriguez et al. [2019]. Lopes et al. [2008] extend earlier work by Lopes et al. [2008] to formulate spatial dynamic factor models for large number of outcomes, but they did not explore dimension reduction in the number of locations. Ren and Banerjee [2013] proposed low-rank specifications for spatially-varying factors to achieve dimension reduction in number of locations and number of variables, but it has since been demonstrated that such low-rank specifications will likely over-smooth when estimating the latent process from massive data sets containing millions of locations. More recently, Taylor-Rodriguez et al. [2019] consider Nearest-Neighbor Gaussian process [Datta et al., 2016a] for spatial factors with the usual constrained loading matrices in non-spatial factor models, which are less general than are strictly required for spatially correlated factors [see, e.g. Ren and Banerjee, 2013].

This Chapter develops as follows. In the next section, we collect some important definitions and results in multivariate geostatistical modeling. We propose our models in Section 4.2. A detailed algorithms for implementing models in Section 4.2 with Nearest Neighbor Gaussian Process (NNGP) [Datta et al., 2016a] are given in Section 4.2.3. In Section 4.3, we state some theoretical results about posterior consistency for the proposed models. Simulation studies for exploring the performance of proposed models are summarized in Section 4.4. An analysis illustrating our methods is presented in Section 4.5.

4.2 Multivariate spatial processes

Let $\mathbf{z}(\mathbf{s}) = (z_1(\mathbf{s}), \ldots, z_q(\mathbf{s}))^\top$ be a $q \times 1$ stochastic process, where each $z_i(\mathbf{s})$ is a real-valued random variable at location $\mathbf{s} \in \mathcal{D} \subseteq \Re^d$. The process is specified by its mean $\mathrm{E}[z_i(\mathbf{s})] = \mu_i(\mathbf{s})$

and, customarily, second-order stationary covariances $C_{ij}(\mathbf{h}) = \text{Cov}\{z_i(\mathbf{s}), z_j(\mathbf{s}+\mathbf{h})\}$ for $i, j = 1, 2, \ldots, q$. These covariances define the matrix-valued $q \times q$ cross-covariance function $\mathbf{C}(\mathbf{h}) = \{C_{ij}(\mathbf{h})\}$ with (i, j)-th entry $C_{ij}(\mathbf{h})$.

While there is no loss of generality in assuming the process mean to be zero by absorbing the mean into a separate regression component in the model, as we will do here, modeling the cross-covariance function requires care. From its definition, $\mathbf{C}(\mathbf{h})$ need not be symmetric, but must satisfy $\mathbf{C}(\mathbf{h})^\top = \mathbf{C}(-\mathbf{h})$. Also, since $\text{var}\{\sum_i^n \mathbf{a}_i^\top \mathbf{z}(\mathbf{s}_i)\} \geq 0$ for any set of finite locations $\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n \in \mathcal{D}$ and any set of constant vectors $\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n \in \Re^q$, we have $\sum_{i,j=1}^n \mathbf{a}_i^\top \mathbf{C}(\mathbf{s}_i - \mathbf{s}_j)\mathbf{a}_i \geq 0$. An excellent review of cross-covariance functions, including several theoretical characterizations, can be found in Genton and Kleiber [2015] and other references on multivariate spatial statistics provided in Section 4.1.

While theoretical characterizations rely upon spectral theory and are useful in understanding the local behavior of random fields, perhaps the most widely used approach for constructing multivariate random fields is the linear model of coregionalization (LMC). The underlying idea is that invertible linear maps of independent spatial processes will yield valid spatial processes. If $\mathbf{f}(\mathbf{s}) = (f_1(\mathbf{s}), f_2(\mathbf{s}), \ldots, f_K(\mathbf{s}))^\top$ is a $K \times 1$ vector of spatial processes, independent of each other so that $\text{cov}\{f_i(\mathbf{s}), f_j(\mathbf{s}')\} = 0$ for all $i \neq j$ and any two locations $\mathbf{s}$ and $\mathbf{s}'$ (same or distinct), then the any new process $\mathbf{z}(\mathbf{s}) = \boldsymbol{\Lambda}^\top \mathbf{f}(\mathbf{s})$, where $\boldsymbol{\Lambda}$ is $K \times q$ will have $q \times q$ cross-covariance matrix $\mathbf{C}_z(\mathbf{h}) = \boldsymbol{\Lambda}^\top \mathbf{C}_f(\mathbf{h})\boldsymbol{\Lambda}$. This cross-covariance will yield non-degenerate process-realizations whenever $K = q$ and $\boldsymbol{\Lambda}$ is nonsingular. The key question, then, becomes how to model $\boldsymbol{\Lambda}$, whose rows determine the subspace where the set of independent factors is mapped.

We follow the developments in Bourgault and Marcotte [1991], which we call the simplified LMC (or SLMC). The multivariate random field $\mathbf{z}(\mathbf{s})$ is a linear combination of K independent univariate random fields

$$\mathbf{z}(\mathbf{s}) = \sum_{k=1}^{K} \boldsymbol{\lambda}_k f_k(\mathbf{s}) = \boldsymbol{\Lambda}^\top \mathbf{f}(\mathbf{s}) \, , \tag{4.2.1}$$

where $\boldsymbol{\lambda}_k$ is the k-th row of $\boldsymbol{\Lambda}$ and each $f_k(\mathbf{s})$ is an independent Gaussian process with covariance function $\rho_{\psi_k}(\cdot,\cdot)$. The corresponding cross-covariance function for $\mathbf{z}(\mathbf{s})$ is $\mathbf{C}(\mathbf{s},\mathbf{s}') = \sum_{k=1}^{K}\rho_{\psi_k}(\mathbf{s},\mathbf{s}')\boldsymbol{\lambda}_k\boldsymbol{\lambda}_k^\top$. We call model (4.2.1) a simplified linear model of coregionalization (SLMC). Other versions include Schmidt and Gelfand [2003], who model the multivariate spatial process through a hierarchical spatial conditional modeling approach, whereupon $\boldsymbol{\Lambda}^\top$ in (4.2.1) is a $q \times q$ lower triangular matrix. The SLMC (4.2.1) connects univariate spatial models to multivariate spatial models using linear transformations of univariate processes. The flexibility offered in modeling the linear transformation is appealing and, in particular, can be used to accrue computational benefits in high-dimensional settings. Other approaches for building cross-covariance functions such as convolutions, latent dimensions, Matérn cross-covariances and other methods reviewed in Genton and Kleiber [2015] do not always provide the flexibility and scalability we seek. Hence, in the remainder of this Chapter we focus on the SLMC model and some its special cases.

4.2.1 A Bayesian SLMC factor model

Let $\mathbf{y}(\mathbf{s}) = (y_1(\mathbf{s}),\ldots,y_q(\mathbf{s}))^\top \in \mathbb{R}^q$ denote the $q \times 1$ vector of dependent outcomes in location $\mathbf{s} \in \mathcal{D} \subset \mathbb{R}^d$, $\mathbf{x}(\mathbf{s}) = (x_1(\mathbf{s}),\ldots,x_p(\mathbf{s}))^\top \in \mathbb{R}^p$ be the corresponding explanatory variables, and $\boldsymbol{\beta}$ be a $p \times q$ regression coefficient matrix that are related as below in a multivariate spatial model

$$\mathbf{y}(\mathbf{s}) = \boldsymbol{\beta}^\top\mathbf{x}(\mathbf{s}) + \boldsymbol{\Lambda}^\top\mathbf{f}(\mathbf{s}) + \boldsymbol{\epsilon}(\mathbf{s})\ ,\ \mathbf{s} \in \mathcal{D}\ , \tag{4.2.2}$$

where the latent process $\boldsymbol{\Lambda}^\top\mathbf{f}(\mathbf{s})$ is an SLMC as described above. Elements in $\mathbf{f}(\mathbf{s})$ are as described in (4.2.1), while the noise process $\boldsymbol{\epsilon}(\mathbf{s}) \overset{iid}{\sim} \mathrm{N}(\mathbf{0},\boldsymbol{\Sigma})$ with covariance matrix $\boldsymbol{\Sigma}$. We model $\{\boldsymbol{\beta},\boldsymbol{\Lambda},\boldsymbol{\Sigma}\}$ using a Matrix-Normal-Inverse-Wishart family. To be precise,

$$\boldsymbol{\beta}\,|\,\boldsymbol{\Sigma} \sim \mathrm{MN}(\boldsymbol{\mu}_{\boldsymbol{\beta}},\mathbf{V}_{\boldsymbol{\beta}},\boldsymbol{\Sigma})\ ;\ \boldsymbol{\Lambda}\,|\,\boldsymbol{\Sigma} \sim \mathrm{MN}(\boldsymbol{\mu}_{\boldsymbol{\Lambda}},\mathbf{V}_{\boldsymbol{\Lambda}},\boldsymbol{\Sigma})\ ;\ \boldsymbol{\Sigma} \sim \mathrm{IW}(\boldsymbol{\Psi},\nu) \quad , \tag{4.2.3}$$

where $\boldsymbol{\mu}_{\boldsymbol{\Lambda}}$ a $q \times K$ matrix and $\mathbf{V}_{\boldsymbol{\Lambda}}$ a $K \times K$ positive definite matrix. A random matrix $\mathbf{Z}_{n\times p} \sim \mathrm{MN}_{n,p}(\mathbf{M},\mathbf{U},\mathbf{V})$ has the probability density function

$$p(\mathbf{Z} \mid \mathbf{M},\mathbf{U},\mathbf{V}) = \frac{\exp\left(-\frac{1}{2}\,\mathrm{tr}\left[\mathbf{V}^{-1}(\mathbf{Z}-\mathbf{M})^T\mathbf{U}^{-1}(\mathbf{Z}-\mathbf{M})\right]\right)}{(2\pi)^{np/2}|\mathbf{V}|^{n/2}|\mathbf{U}|^{p/2}}\ , \tag{4.2.4}$$

where tr denotes trace, $\mathbf{M}$ is the mean matrix, $\mathbf{U}$ is the first scale matrix with dimension $n \times n$ and $\mathbf{V}$ is the second scale matrix with dimension $p \times p$. This distribution is equivalent to $\text{vec}(\mathbf{Z}) \sim \text{N}_{np}(\text{vec}(\mathbf{M}), \mathbf{V} \otimes \mathbf{U})$, where $\otimes$ is the Kronecker product and $\text{vec}(\mathbf{Z}) = \left[\mathbf{z}_1^\top, \ldots, \mathbf{z}_p^\top\right]^\top$ is the vectorization of the $n \times p$ random matrix $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_p]$.

The assigned priors in (4.2.3) yield conditional posterior distributions in a closed form for all the parameters, except $\{\psi_k\}_{k=1}^K$. This supports a block update MCMC algorithm for posterior sampling. Assume $\mathcal{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ is the set of locations with at least one observed response, $\{\mathcal{S}_i\}$ is the observed location set for i-th response, $\cup_{i=1}^q \mathcal{S}_i = \mathcal{S}$. $\mathcal{M}_i = \mathcal{S} \setminus \mathcal{S}_i$ is the set of locations where at least one response, but not the ith response, is observed, $\cup_{i=1}^q \mathcal{M}_i = \mathcal{M}$. Without misalignment, the observation model can be cast as

$$\mathbf{Y}_{n\times q} = \mathbf{X}_{n\times p}\boldsymbol{\beta}_{p\times q} + \mathbf{F}_{n\times K}\boldsymbol{\Lambda}_{K\times q} + \boldsymbol{\epsilon}_{n\times q} \ , \tag{4.2.5}$$

where $\mathbf{Y} = \mathbf{y}(\mathcal{S}) = [\mathbf{y}(\mathbf{s}_1) : \cdots : \mathbf{y}(\mathbf{s}_n)]^\top$ is the $n \times q$ response matrix, $\mathbf{X} = \mathbf{x}(\mathcal{S}) = [\mathbf{x}(\mathbf{s}_1) : \cdots : \mathbf{x}(\mathbf{s}_n)]^\top$ is the corresponding design matrix with full rank ($n > p$), and $\mathbf{F}$ is the $n \times K$ matrix with j-th column being the $n \times 1$ vector comprising $\mathbf{f}_j(\mathbf{s}_i)$'s for $i = 1, 2, \ldots, n$.

We derive the posterior distribution of $\mathbf{F}$ and the unobserved response $\{y_i(\mathcal{M}_i)\}_{i=1}^q$ conditional on $\{\boldsymbol{\beta}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}, \{\psi_k\}_{k=1}^K\}$. Let $\mathbf{P}$ be the $nq \times nq$ permutation matrix such that $\mathbf{P}\text{vec}(\mathbf{Y}) = \{\mathbf{y}(\mathbf{s}_i)_{os_i}\}_{i=1}^n$, where we use the suffix os to denote the index of the observed responses for $\mathbf{s} \in \mathcal{S}$. Therefore, $\mathbf{P}$ reorders the observed responses from $\text{vec}(\mathbf{Y})$ in locations $\{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$. Then, the joint distribution of $\text{vec}(\mathbf{F})$ and $\{\mathbf{y}(\mathbf{s}_i)_{os_i}\}_{i=1}^n$, given $\{\boldsymbol{\beta}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}, \{\psi_k\}_{k=1}^K\}$, can be represented through the augmented linear system,

$$\begin{bmatrix} \{(\mathbf{y}(\mathbf{s}_i) - \mathbf{x}(\mathbf{s}_i)^\top\boldsymbol{\beta})_{os_i}\}_{i=1}^n \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} \mathbf{P}(\boldsymbol{\Lambda}^\top \otimes \mathbf{I}_n) \\ \mathbf{I}_K \otimes \mathbf{I}_n \end{bmatrix} \text{vec}(\mathbf{F}) + \begin{bmatrix} \boldsymbol{\epsilon}_1 \\ \boldsymbol{\epsilon}_2 \end{bmatrix} , \tag{4.2.6}$$

where $\boldsymbol{\epsilon}_1 \sim \text{N}(\mathbf{0}, \oplus_{i=1}^n\{\boldsymbol{\Sigma}_{os_i}\})$, $\boldsymbol{\epsilon}_2 \sim \text{N}(\mathbf{0}, \oplus_{k=1}^K\{\boldsymbol{\rho}_{\psi_k}(\mathcal{S}, \mathcal{S})\})$, $\boldsymbol{\rho}_{\psi_k}(\mathcal{S}, \mathcal{S})$ is the $n \times n$ spatial correlation matrix corresponding to $\mathbf{f}_k = (f_k(\mathbf{s}_1), f_k(\mathbf{s}_2), \ldots, f_k(\mathbf{s}_n))^\top$, and $\oplus_{i=1}^n$ represents the block diagonal operator stacking matrices along the diagonal. Letting $\mathbf{D}_{\boldsymbol{\Sigma}_o}^{-\frac{1}{2}} = \oplus_{i=1}^n\{\boldsymbol{\Sigma}_{os_i}^{-\frac{1}{2}}\}$

and $\mathbf{V}_{\mathbf{F}} = \oplus_{k=1}^{K}\{\mathbf{V}_k\}$, where $\boldsymbol{\rho}_{\psi_k}^{-1}(\mathcal{S}, \mathcal{S}) = \mathbf{V}_k^\top \mathbf{V}_k$, we obtain

$$\underbrace{\begin{bmatrix} \mathbf{D}_{\mathbf{\Sigma}_o}^{-\frac{1}{2}}\{(\mathbf{y}(\mathbf{s}_i) - \mathbf{x}(\mathbf{s}_i)^\top \boldsymbol{\beta})_{os_i}\}_{i=1}^{n} \\ \mathbf{0} \end{bmatrix}}_{\tilde{\mathbf{Y}}} = \underbrace{\begin{bmatrix} \mathbf{D}_{\mathbf{\Sigma}_o}^{-\frac{1}{2}}\mathbf{P}\mathbf{\Lambda}^\top \otimes \mathbf{I}_n \\ \mathbf{V}_{\mathbf{F}} \end{bmatrix}}_{\tilde{\mathbf{X}}} \mathrm{vec}(\mathbf{F}) + \underbrace{\begin{bmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 \end{bmatrix}}_{\tilde{\boldsymbol{\eta}}} . \tag{4.2.7}$$

The elements of $\tilde{\boldsymbol{\eta}}$ are independent error terms, each with unit variance. The full conditional distribution $\mathrm{vec}(\mathbf{F}) \mid \{\mathbf{y}(\mathbf{s}_i)_{os_i}\}_{i=1}^{n}, \boldsymbol{\beta}, \mathbf{\Lambda}, \mathbf{\Sigma}, \{\psi_k\}_{k=1}^{K}$ for the SLMC model in (4.2.2) then follows

$$\mathrm{vec}(\mathbf{F}) \mid \{\mathbf{y}(\mathbf{s}_i)_{os_i}\}_{i=1}^{n}, \boldsymbol{\beta}, \mathbf{\Lambda}, \mathbf{\Sigma}, \{\psi_k\}_{k=1}^{K} \sim \mathrm{N}((\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^\top \tilde{\mathbf{Y}},\, (\tilde{\mathbf{X}}^\top \tilde{\mathbf{X}})^{-1}). \tag{4.2.8}$$

Turning to the unobserved variables, let ms be the suffix for $\mathbf{s} \in \mathcal{M}$. Then, the conditional distribution of $\mathbf{y}(\mathbf{s})_{ms}$ given the parameters $\{\mathbf{F}, \{\mathbf{y}(\mathbf{s}_i)_{os_i}\}_{i=1}^{n}, \boldsymbol{\beta}, \mathbf{\Lambda}, \mathbf{\Sigma}\}$ is

$$\mathrm{N}([\boldsymbol{\mu}_{\mathbf{s}}]_{ms} + \mathbf{\Sigma}_{[ms,os]}\mathbf{\Sigma}_{[os,os]}^{-1}(\mathbf{y}(\mathbf{s})_{os} - [\boldsymbol{\mu}_{\mathbf{s}}]_{os}), \mathbf{\Sigma}_{[ms,ms]} - \mathbf{\Sigma}_{[ms,os]}\mathbf{\Sigma}_{[os,os]}^{-1}\mathbf{\Sigma}_{[os,ms]}) \, , \tag{4.2.9}$$

where $\boldsymbol{\mu}_{\mathbf{s}} = \boldsymbol{\beta}^\top \mathbf{x}(\mathbf{s}) + \mathbf{\Lambda}\mathbf{f}(\mathbf{s})$, $\mathbf{\Sigma}_{[ms,os]}$ is a sub-matrix of $\mathbf{\Sigma}$ extracted with row index ms and column index os, and $\mathbf{\Sigma}_{[os,os]}^{-1}$ is the inverse of matrix $\mathbf{\Sigma}_{[os,os]}$. With the priors given in (4.2.3), we let $\mathbf{V}_{\Lambda} = \mathbf{L}_{\Lambda}\mathbf{L}_{\Lambda}^\top$ and define $\boldsymbol{\gamma} = [\boldsymbol{\beta}^\top, \mathbf{\Lambda}^\top]^\top$. The conditional posterior distribution $\boldsymbol{\gamma} \mid \mathbf{\Sigma}, \mathbf{F}, \mathbf{Y}$ can be found through the augmented linear system,

$$\underbrace{\begin{bmatrix} \mathbf{Y} \\ \mathbf{L}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}} \\ \mathbf{L}_{\mathbf{\Lambda}}^{-1}\boldsymbol{\mu}_{\mathbf{\Lambda}} \end{bmatrix}}_{\mathbf{Y}^*} = \underbrace{\begin{bmatrix} \mathbf{X} & \mathbf{F} \\ \mathbf{L}_{\boldsymbol{\beta}}^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{L}_{\mathbf{\Lambda}}^{-1} \end{bmatrix}}_{\mathbf{X}^*} \underbrace{\begin{bmatrix} \boldsymbol{\beta} \\ \mathbf{\Lambda} \end{bmatrix}}_{\gamma} + \underbrace{\begin{bmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 \\ \boldsymbol{\eta}_3 \end{bmatrix}}_{\boldsymbol{\eta}^*} , \tag{4.2.10}$$

where $\boldsymbol{\eta}^* \sim \mathrm{MN}(\mathbf{0}_{(n+p+K)\times q}, \mathbf{I}_{n+p+K}, \mathbf{\Sigma})$. Using standard distribution theory, we can show that $\boldsymbol{\gamma}, \mathbf{\Sigma} \mid \mathbf{F}, \mathbf{Y}$ follows $\mathrm{MNIW}(\boldsymbol{\mu}^*, \mathbf{V}^*, \mathbf{\Psi}^*, \nu^*)$, where

$$\begin{aligned} &\mathbf{V}^* = [\mathbf{X}^{*\top}\mathbf{X}^*]^{-1} \, , \; \boldsymbol{\mu}^* = \mathbf{V}^*[\mathbf{X}^{*\top}\mathbf{Y}^*] \, , \\ &\mathbf{\Psi}^* = \mathbf{\Psi} + (\mathbf{Y}^* - \mathbf{X}^*\boldsymbol{\mu}^*)^\top(\mathbf{Y}^* - \mathbf{X}^*\boldsymbol{\mu}^*) \, , \; \nu^* = \nu \, . \end{aligned} \tag{4.2.11}$$

In particular, if $\mathbf{\Sigma} = \oplus_{i=1}^{q}\{\sigma_i^2\}$, then we specify $\sigma_i^2 \sim \mathrm{IG}(a_i, b_i)$ for $i = 1, \dots q$ where $a_1 = a_2 = \ldots = a_q = a$. We can show that the marginal posterior distribution of σ_i^2 given $\mathbf{Y}, \mathbf{F}$

follows $\mathrm{IG}(a^*, b_i^*)$ with

$$a^* = a + \frac{n}{2}\ ,\ b_i^* = b_i + \frac{1}{2}(\mathbf{Y}^* - \mathbf{X}^*\boldsymbol{\mu}^*)_i^\top(\mathbf{Y}^* - \mathbf{X}^*\boldsymbol{\mu}^*)_i\ ,\ i = 1, \ldots, q. \tag{4.2.12}$$

Here $(\mathbf{Y}^* - \mathbf{X}^*\boldsymbol{\mu}^*)_i$ is the i-th column of $\mathbf{Y}^* - \mathbf{X}^*\boldsymbol{\mu}^*$. Through the linear system (4.2.10), the conditional distribution $\boldsymbol{\gamma} \mid \boldsymbol{\Sigma}, \mathbf{F}, \mathbf{Y}$ follows $\mathrm{MN}(\boldsymbol{\mu}^*, \mathbf{V}^*, \boldsymbol{\Sigma})$.

The full conditional distributions for $\{\psi_k\}_{k=1}^K$ are not available in closed form. However, since $\{\psi_k\}_{k=1}^K$ and $\mathbf{Y}$ are conditionally independent given $\{\mathbf{F}, \boldsymbol{\gamma}, \boldsymbol{\Sigma}\}$, and $\mathbf{f}_k$ are independent for $k = 1, 2, \ldots, K$, we obtain $p(\psi_k \mid \mathbf{F}, \mathbf{Y}, \boldsymbol{\gamma}, \boldsymbol{\Sigma}, \{\psi_j\}_{j \neq k})$ up to a proportionality constant as

$$p(\mathbf{Y} \mid \mathbf{F}, \boldsymbol{\gamma}, \boldsymbol{\Sigma}) \times p(\boldsymbol{\gamma}, \boldsymbol{\Sigma}) \times \prod_{k=1}^K p(\mathbf{f}_k \mid \psi_k) \times p(\psi_k) \propto p(\mathbf{f}_k \mid \psi_k) \times p(\psi_k)\ , \tag{4.2.13}$$

for each $k = 1, \ldots, K$, where $p(\psi_k)$ is the prior for ψ_k. Often, the right hand side of (4.2.13) is much easier to calculate than a direct formulation of the posterior distribution of ψ_k, especially when $\mathbf{f}_k$'s are modeled using scalable spatial models.

Next, consider the posterior predictive distribution on a new set of locations $\mathcal{U} = \{\mathbf{u}_1, \ldots, \mathbf{u}_{n'}\}$. Through the definition of SLMC (4.2.2), the prediction $\mathbf{Y}_\mathcal{U}$ on $\mathcal{U}$ is conditionally independent to $\{\mathbf{y}(\mathbf{s}_i)_{os_i}\}_{i=1}^n$ given $\{\boldsymbol{\beta}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}\}$ and $\mathbf{f}(\mathbf{s})$ over $\mathcal{U}$. We denote $f_k(\mathcal{U})$ as the realization of the kth element in $\mathbf{f}(\mathbf{s})$ on $\mathcal{U}$, then

$$f_k(\mathcal{U}) \mid f_k(\mathcal{S}), \psi_k \sim \mathrm{N}(\boldsymbol{\rho}_{\psi_k}(\mathcal{U}, \mathcal{S})\boldsymbol{\rho}_{\psi_k}^{-1}(\mathcal{S}, \mathcal{S})f_k(\mathcal{S}),\ \boldsymbol{\rho}_{\psi_k}(\mathcal{U}, \mathcal{S})\boldsymbol{\rho}_{\psi_k}^{-1}(\mathcal{S}, \mathcal{S})\boldsymbol{\rho}_{\psi_k}(\mathcal{S}, \mathcal{U}))\ . \tag{4.2.14}$$

Since $\mathbf{F}_\mathcal{U} = [f_1(\mathcal{U}) : \cdots : f_k(\mathcal{U})]^\top$ is conditionally independent to $\{\mathbf{y}(\mathbf{s}_i)_{os_i}\}_{i=1}^n$ given $\mathbf{F}$ and $\{\psi_k\}_{k=1}^K$, we have $p(\mathbf{Y}_\mathcal{U}, \mathbf{F}_\mathcal{U} \mid \{\mathbf{y}(\mathbf{s}_i)_{oi}\}_{i=1}^n) \propto$

$$p(\mathbf{Y}_\mathcal{U} \mid \mathbf{F}_\mathcal{U}, \boldsymbol{\beta}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}) \times p(\mathbf{F}_\mathcal{U} \mid \mathbf{F}, \{\psi_k\}_{k=1}^K) \times p(\boldsymbol{\beta}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}, \mathbf{F}, \{\psi_k\}_{k=1}^K \mid \{\mathbf{y}(\mathbf{s}_i)_{os_i}\}_{i=1}^n)\ . \tag{4.2.15}$$

There is no closed form of the posterior predictive distribution, but we can use Monte Carlo methods and equation (4.2.15) and (4.2.14) to obtain the posterior predictive samples over new locations after obtaining posterior samples of $\boldsymbol{\beta}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}, \mathbf{F}, \{\psi_k\}_{k=1}^K$.

4.2.2 The block update MCMC algorithm

We formulate an efficient MCMC algorithm for obtaining full Bayesian inference as follows. From l th iteration with $\{\boldsymbol{\beta}^{(l)}, \boldsymbol{\Lambda}^{(l)}, \boldsymbol{\Sigma}^{(l)}, \{\psi_k^{(l)}\}_{k=1}^K\}$, we first generate $\mathbf{F}^{(l+1)}$ from its full conditional distribution (4.2.8). Next, we impute the missing response $\{\mathbf{y}(\mathbf{s}_i)_{mi}^{(l+1)}\}_{\mathbf{s}_i \in \mathcal{M}}$ on $\mathcal{M}$ through (4.2.9) and then update $\{\boldsymbol{\beta}^{(l+1)}, \boldsymbol{\Lambda}^{(l+1)}, \boldsymbol{\Sigma}^{(l+1)}\}$ using (4.2.11). We complete the iteration by drawing $\{\psi_k^{(l+1)}\}_{k=1}^K$ using a Metropolis-Hasting (M-H) random walk step using (4.2.13). Repeating these iterations will eventually, after a suitably diagnosed burn-in period, will generate samples from the desired joint posterior distribution.

For each iteration after burn-in, we sample $\mathbf{F}_{\mathcal{U}}$ from (4.2.14), given the posterior samples of $\mathbf{F}$ and $\{\psi_k\}_{k=1}^K$, then generate posterior predictions of $\mathbf{Y}_{\mathcal{U}}$ given posterior samples of $\{\boldsymbol{\beta}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}, \mathbf{F}_{\mathcal{U}}\}$. Applying the SCAM algorithm introduced in Haario et al. [2005], one can avoid tuning parameters in M-H algorithm by warming up each MCMC chain of $\{\psi_k\}_{k=1}^K$ with an adaptive proposal distribution. In our implementation, we use the proposal distribution defined by equation (2.1) in Roberts and Rosenthal [2009] with an empirical estimate of the covariance of the target distribution based on half of the chain's history. We ran the adaptive algorithm for the first quarter of the MCMC chains and then fixed the proposal distribution for the rest of the MCMC chains in the simulation studies and Vegetation Indices data analysis in Section 4.4 & 4.5.

The parameters $\boldsymbol{\Lambda}$ and $\mathbf{F}$ are not jointly identified, but we can transform back to $\boldsymbol{\omega} = \mathbf{F}\boldsymbol{\Lambda}$ and obtain inference for the latent process. This parametrization has the advantage of conditional conjugacy, which brings more efficient computation in posterior sampling. Since all elements in $\mathbf{F}$ are sampled simultaneously from a Gaussian distribution, the sample of $\mathbf{F}$ can be generated through a linear transformation of $n \times K$ independent parameters as shown in (4.2.7). The sampling of $\{\boldsymbol{\beta}, \boldsymbol{\Lambda}\}$ follows the same trick. Hence, we can dramatically improve the convergence of the Markov chain by reducing the posterior dependence among the parameter in this Gibbs with M-H algorithm [Gelman et al., 2013]. Since $\mathbf{F}$ is sensitive to the value of the intercept, we recommend using an intercept-centered latent process to obtain inference for the latent spatial pattern and the non-spatial covariance of the latent process. The initial value of $\mathbf{\Lambda}$ should never be a zero matrix. Otherwise, $\mathbf{F}$ may get an extreme initial value, slowing down the convergence of the MCMC chains.

4.2.3 Scalable Modeling for Block-update MCMC

Analogous to the conjugate multivariate latent model, we use a conjugate gradient method to facilitate the sampling of $\mathbf{F}$ when there exists a sparse precision matrix $\boldsymbol{\rho}_{\psi_k}^{-1}(\mathcal{S}, \mathcal{S})$ for $k = 1, \ldots, K$. The idea of accelerating MCMC sampling through a conjugate gradient method has an excellent implementation in Nishimura and Suchard [2018]. We develop a Bayesian framework to implement this sampling scheme in massive multivariate spatial data modeling. Here, we illustrate a detailed algorithm for a BSLMC model, where each element of the factor process $\mathbf{f}(\mathbf{s})$ is modeled as a Nearest-Neighbor Gaussian Process (NNGP).

Let each $f_k(\mathbf{s}), \mathbf{s} \in \mathcal{D}$ be an NNGP$(0, \rho_{\psi_k}(\cdot, \cdot))$, which implies that $\mathbf{f}_k \sim N(\mathbf{0}, \tilde{\boldsymbol{\rho}}_k)$ for each $k = 1, 2, \ldots, K$, where $\tilde{\boldsymbol{\rho}}_k = (\mathbf{I} - \mathbf{A}_{\rho_k})^{-1}\mathbf{D}_{\rho_k}(\mathbf{I} - \mathbf{A}_{\rho_k})^{-\top}$, $\mathbf{A}_{\rho_k}$ is a sparse-lower triangular matrix with no more than a specified small number, m, of nonzero entries and $\mathbf{D}_{\rho_k}$ is a diagonal matrix. The diagonal entries of $\mathbf{D}_{\rho_k}$ and the nonzero entries of $\mathbf{A}_{\rho_k}$ are obtained from the conditional variance and conditional expectations for a Gaussian process with covariance function $\rho_{\psi_k}(\mathbf{s}, \mathbf{s}')$. To be precise, we consider a fixed order of locations in $\mathcal{S}$ and define $N_m(\mathbf{s}_i)$ to be the set comprising at most m neighbors of $\mathbf{s}_i$ among locations $\mathbf{s}_j \in \mathcal{S}$ such that $j < i$. The (i, j)-th entry of $\mathbf{A}_{\rho_k}$ is 0 whenever $\mathbf{s}_j \notin N_m(\mathbf{s}_i)$. If $j_1 < j_2 < \cdots < j_m$ are the m column indices for the nonzero entries in the i-th row of $\mathbf{A}_{\rho_k}$, then the (i, j_k)-th element of $\mathbf{A}_{\rho_k}$ is the k-th element of the $1 \times m$ vector $\mathbf{a}_i^\top = \boldsymbol{\rho}_{\psi_k}(\mathbf{s}_i, N_m(\mathbf{s}_i))\boldsymbol{\rho}_{\psi}(N_m(\mathbf{s}_i), N_m(\mathbf{s}_i))^{-1}$. The (i, i)-th diagonal element of $\mathbf{D}_{\rho_k}$ is given by $\rho_{\psi_k}(\mathbf{s}_i, \mathbf{s}_i) - \mathbf{a}_i^\top \boldsymbol{\rho}_{\psi_k}(N_m(\mathbf{s}_i), \mathbf{s}_i)$. Repeating these calculations for each row completes the construction of $\mathbf{A}_{\rho_k}$ and $\mathbf{D}_{\rho_k}$ and yields a sparse $\tilde{\boldsymbol{\rho}}_k^{-1}$. This construction can be performed in parallel and requires storage or computation of at most $m \times m$ matrices, where $m << n$, costing $\mathcal{O}(n)$ flops and storage.

For posterior predictions, we use $N_m(\mathbf{u}_i)$ to denote the m neighbors of $\mathbf{u}_i \in \mathcal{U}$ among $\mathcal{S}$.

The posterior prediction for $f_k(\mathcal{U})$ given in (4.2.14) follows

$$f_k(\mathcal{U}) \,|\, f_k(\mathcal{S}), \psi_k \sim \mathrm{N}(\tilde{\mathbf{A}} f_k(\mathcal{S}), \tilde{\mathbf{D}}) \;, \tag{4.2.16}$$

where the (i,j)-th entry of $\tilde{\mathbf{A}}$ is 0 when $\mathbf{s}_j \notin N_m(\mathbf{u}_i)$, and, similar to $\mathbf{A}_{\rho_k}$, the m nonzero entries in the i-th row of $\tilde{\mathbf{A}}$ corresponds to the elements of the $1 \times m$ vector $\tilde{\mathbf{a}}_i^\top = \rho_{\psi_k}(\mathbf{u}_i, N_m(\mathbf{u}_i))\rho_{\psi_k}(N_m(\mathbf{u}_i), N_m(\mathbf{u}_i))^{-1}$. The (i,i)-th diagonal element of $\tilde{\mathbf{D}}$ equals $\rho_{\psi_k}(\mathbf{u}_i, \mathbf{u}_i) - \tilde{\mathbf{a}}_i^\top \boldsymbol{\rho}_{\psi_k}(N_m(\mathbf{u}_i), \mathbf{u}_i)$. And the posterior sample of $\mathbf{Y}_\mathcal{U}$ after giving posterior sample of $\boldsymbol{\beta}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}$ and $\mathbf{F}_\mathcal{U}$ can be sampled through

$$\mathrm{MN}(\mathbf{X}_\mathcal{U}\boldsymbol{\beta} + \mathbf{F}_\mathcal{U}\boldsymbol{\Lambda}, \mathbf{I}_{n'}, \boldsymbol{\Sigma}) \tag{4.2.17}$$

A detailed algorithm is presented below.

Algorithm 4.1: Obtaining posterior inference of $\{\boldsymbol{\gamma}, \boldsymbol{\Sigma}, \boldsymbol{\omega}\}$ and predictions on a new set $\mathcal{U}$ for NNGP based BSLMC model

```
1. Precalculation and preallocation for the MCMC algorithm
    (a) Find location sets $\mathcal{S}$, $\mathcal{M}$ and the index of the observed and missing response $\{os_i\}_{i=1}^n$ and $\{ms_i\}_{i=1}^n$.
    (b) Build the nearest neighbor for $\mathcal{S}$
    (c) Calculate Cholesky decompositions $\mathbf{V}_{\boldsymbol{\Lambda}} = \mathbf{L}_{\boldsymbol{\Lambda}}\mathbf{L}_{\boldsymbol{\Lambda}}^\top$ and $\mathbf{V}_{\boldsymbol{\beta}} = \mathbf{L}_{\boldsymbol{\beta}}\mathbf{L}_{\boldsymbol{\beta}}^\top$
    (d) Preallocate MCMC samples and initalize MCMC chain with $\boldsymbol{\beta}^{(0)}$, $\boldsymbol{\Lambda}^{(0)}$, $\boldsymbol{\Sigma}^{(0)}$ and $\{\psi_k^{(0)}\}_{k=1}^K$
2. Block update MCMC alogrithm. For $l = 1 : L$
    (a) Update $\mathbf{F}^{(l)}$ and impute missing response $\{\mathbf{y}(\mathbf{s}_i)_{mi}^{(l)}\}_{\mathbf{s}_i \in \mathcal{M}}$
        • Construct $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$ in (4.2.7)
            – Build the matrix $\mathbf{D}_{\boldsymbol{\Sigma}_o}^{\frac{1}{2}} = \mathrm{diag}(\{\boldsymbol{\Sigma}_{os_i}^{-\frac{1}{2}}\}_{i=1}^n\})$ in (4.2.7)    $\mathcal{O}(n)$
            – Construct $\{\mathbf{A}_{\rho_k}\}_{k=1}^K$ and $\{\mathbf{D}_{\rho_k}\}_{k=1}^K$ as described, for example, in Finley et al. [2019]    $\mathcal{O}(Knm^3)$
            – Construct $\tilde{\mathbf{X}}$ and $\tilde{\mathbf{Y}}$ in (4.2.7) with $\mathbf{V}_k = \mathbf{D}_{\rho_k}^{-\frac{1}{2}}(\mathbf{I} - \mathbf{A}_{\rho_k})$    $\mathcal{O}(nK(m+1+q)+npq)$
        • Use LSMR to generate sample of $\mathbf{F}^{(l)}$
            – Sample $\mathbf{u} \sim N(\mathbf{0}, I_{Kn})$    $\mathcal{O}(nK)$
            – Solve $\mathrm{vec}(\mathbf{F})^{(l)}$ from $\tilde{\mathbf{X}}\mathrm{vec}(\mathbf{F})^{(l)} = \tilde{\mathbf{Y}} + \mathbf{u}$ by LSMR
        • Impute missing response $\{\mathbf{y}(\mathbf{s}_i)_{ms_i}^{(l)}\}_{\mathbf{s}_i \in \mathcal{M}}$ over $\mathcal{M}$ through (4.2.9)
            – Calculate $\boldsymbol{\mu}_\mathbf{s} = \boldsymbol{\beta}^{(l)\top}\mathbf{x}(\mathbf{s}) + \boldsymbol{\Lambda}^{(l-1)}\mathbf{f}(\mathbf{s})$ for $\mathbf{s} \in \mathcal{M}$
            – Sample $\mathbf{y}(\mathbf{s})_{ms}^{(l)}$ by (4.2.9) for $\mathbf{s} \in \mathcal{M}$
    (b) Use MNIW to update $\{\boldsymbol{\beta}^{(l)}, \boldsymbol{\Lambda}^{(l)}, \boldsymbol{\Sigma}^{(l)}\}$
        • Construct $\mathbf{X}^*$ and $\mathbf{Y}^*$ in (4.2.10)
        • Generate $\boldsymbol{\Sigma}^{(l)}$
        – (When $\boldsymbol{\Sigma}$ is a positive symmetric matrix)
            * Calculate $\boldsymbol{\mu}^*$, $\mathbf{V}^{*-1}$, $\boldsymbol{\Psi}^*$ and $\nu^*$ by (4.2.11)    $\mathcal{O}(n(p+K)(p+K+q))$
            * Sample $\boldsymbol{\Sigma}^{(l)}$ from $\mathrm{IW}(\boldsymbol{\Psi}, \nu^*)$
        – (When $\boldsymbol{\Sigma}$ is diagonal)
            * Calculate $\boldsymbol{\mu}^*$ by (4.2.11)    $\mathcal{O}(n(p+K)(p+K+q))$
            * Sample elements of $\boldsymbol{\Sigma}^{(l)}$ from Inverse-Gamma with parameters provided in (4.2.12)
      • Sample $\boldsymbol{\gamma}^{(l)} = [\boldsymbol{\beta}^{(l)\top}, \boldsymbol{\Lambda}^{(l)^\top}]^\top$ from $\mathrm{MN}(\boldsymbol{\mu}^*, \mathbf{V}^*, \boldsymbol{\Sigma}^{(l)})$
          i. Sample $\mathbf{u} \sim \mathrm{MN}(\mathbf{0}, \mathbf{I}_{p+K}, \mathbf{I}_q)$
         ii. Calculate Cholesky decomposition $\mathbf{V}^{*-1} = \mathbf{L}_{\mathbf{V}}\mathbf{L}_{\mathbf{V}}^\top$ and $\boldsymbol{\Sigma}^{(l)} = \mathbf{L}_{\boldsymbol{\Sigma}^{(l)}}\mathbf{L}_{\boldsymbol{\Sigma}^{(l)}}^\top$
        iii. Generate $\boldsymbol{\gamma}^{(l)} = \boldsymbol{\mu}^* + \mathbf{L}_{\mathbf{V}}^{-\top}\mathbf{u}\mathbf{L}_{\boldsymbol{\Sigma}^{(l)}}^\top$
  (c) (Optional) Use Metropolis-Hasting to update $\{\Psi_k^{(l)}\}_{k=1}^K$
       i. Propose new $\{\Psi_k^*\}_{k=1}^K$ based on $\{\Psi_k^{(l-1)}\}_{k=1}^K$
      ii. Calculate the likelihood of the new proposed $\{\Psi_k^*\}_{k=1}^K$ and $\{\Psi_k^{(l-1)}\}_{k=1}^K$ given $\mathbf{F}^{(l)}$ using (4.2.13) $\mathcal{O}(Knm^3)$
     iii. Accept the new $\{\Psi_k^*\}_{k=1}^K$ as $\{\Psi_k^{(l)}\}_{k=1}^K$ with the probability of the ratio of the likelihood of $\{\Psi_k^*\}_{k=1}^K$ and
          $\{\Psi_k^{(l-1)}\}_{k=1}^K$. Let $\{\Psi_k^{(l)}\}_{k=1}^K = \{\Psi_k^{(l-1)}\}_{k=1}^K$ when the new proposal is rejected.
3. Generate posterior samples of $\{\mathbf{F}_{\mathcal{U}}^{(l)}, \mathbf{Y}_{\mathcal{U}}^{(l)}\}$ on a new set $\mathcal{U}$
  (a) Construct $\tilde{\mathbf{A}}$ and $\tilde{\mathbf{D}}$ in (4.2.16)    $\mathcal{O}(n'm^3K)$
  (b) Generate $f_k(\mathcal{U})^{(l)} \sim \mathrm{N}(\tilde{\mathbf{A}}f_k(\mathcal{S}), \tilde{\mathbf{D}})$ for $k = 1, \ldots, K$    $\mathcal{O}(n'Km)$
  (c) Sample $\mathbf{Y}_{\mathcal{U}}^{(l)} \,|\, \boldsymbol{\omega}_{\mathcal{U}}^{(l)}, \boldsymbol{\gamma}^{(l)}, \boldsymbol{\Sigma}^{(l)}, \mathbf{F}_{\mathcal{U}}^{(l)} \sim \mathrm{MN}(\mathbf{X}_{\mathcal{U}}\boldsymbol{\beta} + \mathbf{F}_{\mathcal{U}}\boldsymbol{\Lambda}, \mathbf{I}_{n'}, \boldsymbol{\Sigma}^{(l)})$

      • Sample $\mathbf{u} \sim \mathrm{MN}(\mathbf{0}, \mathbf{I}_{n'}, \mathbf{I}_q)$    $\mathcal{O}(n'q)$
      • Generate $\mathbf{Y}_{\mathcal{U}}^{(l)} = \mathbf{X}_{\mathcal{U}}\boldsymbol{\beta} + \mathbf{F}_{\mathcal{U}}\boldsymbol{\Lambda} + \mathbf{u}\mathbf{L}_{\boldsymbol{\Sigma}^\top_{(l)}}$ with $\mathbf{F}_{\mathcal{U}}^{(l)} = [f_1(\mathcal{U})^{(l)} : \cdots : f_K(\mathcal{U})^{(l)}]$    $\mathcal{O}(n'(pq + Kq + q^2)$
```

We conclude this section with a remark on the BSLMC model with diagonal $\boldsymbol{\Sigma}$. This specification is desirable for data sets with a massive number of responses q. Compared to BSLMC, BSLMC with diagonal $\boldsymbol{\Sigma}$ can avoid the quadratic growth of the number of parameters in $\boldsymbol{\Sigma}$ as q increases. When $K < q$, it becomes a factor model that can fit the latent process with a low-rank structure. We provide an example in next section to illustrate an NNGP based factor BSLMC with diagonal $\boldsymbol{\Sigma}$.

4.3 On posterior consistency

We present some theoretical results on the Matrix-Normal models constructed in the previous section. Specifically, we investigate the behavior of the posterior distribution as the sample size increases. Here, for establishing the results, we will assume conjugate MNIW models with no misalignment.

First consider modeling $\mathbf{y}(\mathbf{s})$ as a spatial process without explicitly introducing a latent process. Let

$$\mathbf{y}(\mathbf{s}) \sim \mathrm{GP}(\boldsymbol{\beta}^\top \mathbf{x}(\mathbf{s}), \mathbf{C}(\cdot,\cdot)) \;,\; \mathbf{C}(\mathbf{s},\mathbf{s}') = [\rho_\psi(\mathbf{s},\mathbf{s}') + (\alpha^{-1}-1)\delta_{\mathbf{s}=\mathbf{s}'}]\boldsymbol{\Sigma} \;, \tag{4.3.1}$$

where $\rho_\psi(\cdot,\cdot)$ is a spatial correlation function defined through hyperparameter ψ, δ denotes Dirac's delta function, and $\alpha^{-1}\boldsymbol{\Sigma}$ is the non-spatial covariance matrix of $\mathbf{y}(\mathbf{s})$. The scalar α is assumed fixed represents the proportion of total variability allocated to the spatial process. This implies that $\mathbf{Y} \,|\, \boldsymbol{\beta}, \boldsymbol{\Sigma} \sim \mathrm{MN}_{n,q}(\mathbf{X}\boldsymbol{\beta}, \boldsymbol{\mathcal{K}}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mathcal{K}} = \boldsymbol{\rho}_\psi(\mathcal{S},\mathcal{S}) + (\alpha^{-1}-1)\mathbf{I}_n$. We model $\{\boldsymbol{\beta}, \boldsymbol{\Sigma}\}$ using the conjugate MNIW prior

$$\boldsymbol{\beta} \,|\, \boldsymbol{\Sigma} \sim \mathrm{MN}_{p,q}(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \mathbf{V}_r, \boldsymbol{\Sigma}) \;,\; \boldsymbol{\Sigma} \sim \mathrm{IW}(\boldsymbol{\Psi}, \nu) \;, \tag{4.3.2}$$

with prefixed $\{\boldsymbol{\mu}_{\boldsymbol{\beta}}, \mathbf{V}_r, \boldsymbol{\Psi}, \nu\}$. Closely following the developments in Gamerman and Moreira [2004], we obtain the posterior distribution of $\{\boldsymbol{\beta}, \boldsymbol{\Sigma}\}$ as $\mathrm{MNIW}(\boldsymbol{\mu}^*, \mathbf{V}^*, \boldsymbol{\Psi}^*, \nu^*)$, where

$$\begin{aligned} &\mathbf{V}^* = (\mathbf{X}^\top \boldsymbol{\mathcal{K}}^{-1}\mathbf{X} + \mathbf{V}_r^{-1})^{-1} \;,\; \boldsymbol{\mu}^* = \mathbf{V}^*(\mathbf{X}^\top \boldsymbol{\mathcal{K}}^{-1}\mathbf{Y} + \mathbf{V}_r^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}}) \;, \\ &\boldsymbol{\Psi}^* = \boldsymbol{\Psi} + \mathbf{Y}^\top \boldsymbol{\mathcal{K}}^{-1}\mathbf{Y} + \boldsymbol{\mu}_{\boldsymbol{\beta}}^\top \mathbf{V}_r^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}} - \boldsymbol{\mu}^{*\top}\mathbf{V}^{*-1}\boldsymbol{\mu}^* \;,\; \text{and } \nu^* = \nu + n \;. \end{aligned} \tag{4.3.3}$$

Consider the spatial regression model

$$\mathbf{y}(\mathbf{s}) = \boldsymbol{\beta}^\top \mathbf{x}(\mathbf{s}) + \boldsymbol{\omega}(\mathbf{s}) + \boldsymbol{\epsilon}(\mathbf{s}) \;,\; \mathbf{s} \in \mathcal{D} \tag{4.3.4}$$

where $\boldsymbol{\omega}(\mathbf{s}) \sim \mathrm{GP}(\mathbf{0}_{q\times 1}, \rho_\psi(\cdot,\cdot)\boldsymbol{\Sigma})$ is a latent process and $\boldsymbol{\epsilon}(\mathbf{s}) \sim \mathrm{N}(\mathbf{0}_{q\times 1}, (\alpha^{-1}-1)\boldsymbol{\Sigma})$ is measurement error. Define $\boldsymbol{\omega} = \boldsymbol{\omega}(\mathcal{S}) = [\boldsymbol{\omega}(\mathbf{s}_1) : \cdots : \boldsymbol{\omega}(\mathbf{s}_n)]^\top$. For theoretical tractability, we restrict posterior inference on $\{\boldsymbol{\beta}, \boldsymbol{\omega}, \boldsymbol{\Sigma}\}$, assuming that the scalar α is fixed. Assuming that the joint distribution of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ are given in (4.2.3) and that $\boldsymbol{\omega} \,|\, \boldsymbol{\Sigma} \sim \mathrm{MN}_{n\times q}(\mathbf{0}, \boldsymbol{\rho}_\psi(\mathcal{S},\mathcal{S}), \boldsymbol{\Sigma})$, the posterior distribution of $\boldsymbol{\gamma}^\top = [\boldsymbol{\beta}^\top, \boldsymbol{\omega}^\top]$ is $p(\boldsymbol{\gamma}, \boldsymbol{\Sigma} \,|\, \mathbf{Y}) = \mathrm{MNIW}(\boldsymbol{\mu}^*_{\boldsymbol{\gamma}}, \mathbf{V}^*, \boldsymbol{\Psi}^*, \nu^*)$, where

$$\begin{aligned}
\mathbf{V}^* &= \begin{bmatrix} \frac{\alpha}{1-\alpha}\mathbf{X}^\top\mathbf{X} + \mathbf{V}_r^{-1} & \frac{\alpha}{1-\alpha}\mathbf{X}^\top \\ \frac{\alpha}{1-\alpha}\mathbf{X} & \boldsymbol{\rho}_\psi^{-1}(\mathcal{S}, \mathcal{S}) + \frac{\alpha}{1-\alpha}\mathbf{I}_n \end{bmatrix}^{-1} , \\
\boldsymbol{\mu}^*_{\boldsymbol{\gamma}} &= \mathbf{V}^* \begin{bmatrix} \frac{\alpha}{1-\alpha}\mathbf{X}^\top\mathbf{Y} + \mathbf{V}_r^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}} \\ \frac{\alpha}{1-\alpha}\mathbf{Y} \end{bmatrix} , \\
\boldsymbol{\Psi}^* &= \boldsymbol{\Psi} + \frac{\alpha}{1-\alpha}\mathbf{Y}^\top\mathbf{Y} + \boldsymbol{\mu}_{\boldsymbol{\beta}}^\top \mathbf{V}_r^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}} - \boldsymbol{\mu}^{*\top}_{\boldsymbol{\gamma}}\mathbf{V}^{*-1}\boldsymbol{\mu}^*_{\boldsymbol{\gamma}} \ , \text{ and} \\
\nu^* &= \nu + n \ ,
\end{aligned} \tag{4.3.5}$$

Let $\mathbf{V}_{\boldsymbol{\rho}}$ be a non-singular square matrix such that $\boldsymbol{\rho}_\psi^{-1}(\mathcal{S}, \mathcal{S}) = \mathbf{V}_{\boldsymbol{\rho}}^\top \mathbf{V}_{\boldsymbol{\rho}}$. Treat the prior of $\boldsymbol{\gamma}$ as additional observations and recast $p(\mathbf{Y}, \boldsymbol{\gamma} \,|\, \boldsymbol{\Sigma}) = p(\mathbf{Y} \,|\, \boldsymbol{\gamma}, \boldsymbol{\Sigma}) \times p(\boldsymbol{\gamma} \,|\, \boldsymbol{\Sigma})$ into an augmented linear model

$$\underbrace{\begin{bmatrix} \sqrt{\frac{\alpha}{1-\alpha}}\mathbf{Y} \\ \mathbf{L}_r^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}} \\ \mathbf{0} \end{bmatrix}}_{\mathbf{Y}^*} = \underbrace{\begin{bmatrix} \sqrt{\frac{\alpha}{1-\alpha}}\mathbf{X} & \sqrt{\frac{\alpha}{1-\alpha}}\mathbf{I}_n \\ \mathbf{L}_r^{-1} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_{\boldsymbol{\rho}} \end{bmatrix}}_{\mathbf{X}^*} \underbrace{\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\omega} \end{bmatrix}}_{\boldsymbol{\gamma}} + \underbrace{\begin{bmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 \\ \boldsymbol{\eta}_3 \end{bmatrix}}_{\boldsymbol{\eta}} , \tag{4.3.6}$$

where $\mathbf{L}_r$ is the Cholesky decomposition of $\mathbf{V}_r$, and $\boldsymbol{\eta} \sim \mathrm{MN}(\mathbf{0}, \mathbf{I}_{2n+p}, \boldsymbol{\Sigma})$. When having a flat prior for $\boldsymbol{\beta}$, $\mathbf{L}_r^{-1}$ degenerates to a zero matrix, showing no information from $\boldsymbol{\beta}$'s prior contributes to the linear system. The expression in (4.3.5) can be simplified as

$$\begin{aligned}
&\mathbf{V}^* = (\mathbf{X}^{*\top}\mathbf{X}^*)^{-1} \ , \ \boldsymbol{\mu}^* = (\mathbf{X}^{*\top}\mathbf{X}^*)^{-1}\mathbf{X}^{*\top}\mathbf{Y}^* \ , \\
&\boldsymbol{\Psi}^* = \boldsymbol{\Psi} + (\mathbf{Y}^* - \mathbf{X}^*\boldsymbol{\mu}^*)^\top(\mathbf{Y}^* - \mathbf{X}^*\boldsymbol{\mu}^*) \ , \ \nu^* = \nu + n \ .
\end{aligned} \tag{4.3.7}$$

We explore the behavior of the above posterior density as the number of observations becomes large under a true data generating distribution. Assume that the true distribution of the dependent variables is included in the parametric family $f(\mathbf{Y}) = p(\mathbf{Y} \,|\, \boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0)$ for some $\boldsymbol{\Sigma}_0$ and $\boldsymbol{\beta}_0$. For distinguishing the variables based on the number of observations, we make the dependence upon n explicit. Denote $\mathbf{X}(n)_{n\times p} = [\mathbf{x}(\mathbf{s}_1) : \cdots : \mathbf{x}(\mathbf{s}_n)]^\top$, $\mathbf{Y}(n)_{n\times q} = [\mathbf{y}(\mathbf{s}_1) : \cdots : \mathbf{y}(\mathbf{s}_n)]^\top$, $\mathcal{S}(n) = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$, $\boldsymbol{\mathcal{K}}(n) = \mathbf{C}(\mathcal{S}(n), \mathcal{S}(n)) + (\alpha^{-1} - 1)\mathbf{I}_n$. $\mathbf{X}^*(n)$ and $\mathbf{Y}^*(n)$ are $\mathbf{X}^*$ and $\mathbf{Y}^*$ in (4.3.6) using $\mathbf{X}(n)$ and $\mathbf{Y}(n)$ instead of $\mathbf{X}$ and $\mathbf{Y}$.

We will establish the posterior consistency of $\{\boldsymbol{\beta}, \boldsymbol{\Sigma}\}$ for the model in (4.3.1) and for $\{\boldsymbol{\gamma}, \boldsymbol{\Sigma}\}$ for (4.3.4). In the following results, we denote $\mathbf{P}(n) = \mathbf{X}(n)^\top \boldsymbol{\mathcal{K}}(n)^{-1}\mathbf{X}(n)$, $\mathbf{A} \geq \mathbf{B}$ to mean that $\mathbf{A} - \mathbf{B}$ is a positive semi-definite matrix, and $\mathbf{A}_{ij}$ to be the (i, j)-th element of $\mathbf{A}$. Since the marginal distribution of $\boldsymbol{\Sigma} \,|\, \mathbf{Y}$ and $\boldsymbol{\beta} \,|\, \boldsymbol{\Sigma}, \mathbf{Y}$ are essentially the same for conjugate response model and conjugate latent model, proof of one model can be adapted for the other.

Lemma 4.3.1. *The matrix* $\boldsymbol{\Sigma}$ *in the conjugate multivariate models is posterior consistent if and only if* $\boldsymbol{\Psi}^*(n)_{ij}/n \to \{\boldsymbol{\Sigma}_0\}_{ij}$ *a.s. for* $1 \leq i, j \leq q$ *with* $\boldsymbol{\Psi}^*(n)$ *defined by* (4.3.3) *&* (4.3.5)

Proof. Conjugate multivariate response model yields $\boldsymbol{\Sigma} \,|\, \mathbf{Y}(n) \sim \text{IW}(\boldsymbol{\Psi}^*(n), \nu^*(n))$ with

$$\mathbf{M}_{ij} = \mathbb{E}(\boldsymbol{\Sigma}_{ij} \,|\, \mathbf{Y}(n)) = \frac{\boldsymbol{\Psi}^*(n)_{ij}}{c-1} \ , \ \text{Var}(\boldsymbol{\Sigma}_{ij} \,|\, \mathbf{Y}(n)) = \frac{(c+1)\boldsymbol{\Psi}^*(n)^2_{ij} + (c-1)\boldsymbol{\Psi}^*(n)_{ii}\boldsymbol{\Psi}^*(n)_{jj}}{c(c-1)^2(c-3)} \ ,$$

where $c = \nu^*(n) - p$, $\boldsymbol{\Psi}^*(n)$ and $\nu^*(n)$ are defined in (4.3.3).

(1)Necessity: If $\boldsymbol{\Sigma}$ is posterior consistent, i.e., for any $\epsilon > 0$

$$\text{lim}_{n\to\infty}\text{Pr}(|\boldsymbol{\Sigma}_{ij} - \boldsymbol{\Sigma}_{0ij}| > \epsilon \,|\, \mathbf{Y}(n)) = 0 \text{ for } 1 \leq i, j \leq q \ .$$

Then $\text{lim}_{n\to\infty}\mathbb{E}(\boldsymbol{\Sigma}_{ij} - \boldsymbol{\Sigma}_{0ij} \,|\, \mathbf{Y}(n)) \leq \text{lim}_{n\to\infty}\mathbb{E}(|\boldsymbol{\Sigma}_{ij} - \boldsymbol{\Sigma}_{0ij}| \,|\, \mathbf{Y}(n)) < \epsilon$ for any $\epsilon > 0$, i.e, $\text{lim}_{n\to\infty}\mathbb{E}(\boldsymbol{\Sigma}_{ij} \,|\, \mathbf{Y}(n)) = \boldsymbol{\Sigma}_{0ij}$ a.s. Therefore we have $\boldsymbol{\Psi}^*(n)_{ij}/n \to \boldsymbol{\Sigma}_{0ij}$ *a.s.* for $1 \leq i, j \leq q$.

(2)Sufficiency: When $\boldsymbol{\Psi}^*(n)_{ij}/n \to \boldsymbol{\Sigma}_{0ij}$ *a.s.* for $1 \leq i, j \leq q$, by the posterior distribution of $\boldsymbol{\Sigma}$ we have $\text{lim}_{n\to\infty}\mathbb{E}(\boldsymbol{\Sigma}_{ij} \,|\, \mathbf{Y}(n)) = \boldsymbol{\Sigma}_{0ij}$ and the variance of each element converges to 0 at the rate of $1/n$. Use triangle inequality and Chebyshev's inequality, we have for any $\epsilon > 0$

$$\begin{aligned}\text{Pr}(|\boldsymbol{\Sigma}_{ij} - \boldsymbol{\Sigma}_{0ij}| > \epsilon \,|\, \mathbf{Y}(n)) &\leq \text{Pr}(|\boldsymbol{\Sigma}_{ij} - \mathbf{M}_{ij}| > \epsilon/2 \,|\, \mathbf{Y}(n)) + \text{Pr}(|\mathbf{M}_{ij} - \boldsymbol{\Sigma}_{0ij}| > \epsilon/2 \,|\, \mathbf{Y}(n)) \\ &\leq 4\text{Var}(\boldsymbol{\Sigma}_{ij} \,|\, \mathbf{Y}(n))/\epsilon^2 + \text{Pr}(|\mathbf{M}_{ij} - \boldsymbol{\Sigma}_{0ij}| > \epsilon/2 \,|\, \mathbf{Y}(n)) \to 0 \ a.s.\end{aligned}$$

Since lemma 4.3.1 holds for conjugate multivariate response model, it also holds for conjugate multivariate latent model with $\boldsymbol{\Psi}^*(n)$ defined by (4.3.5). □

Theorem 4.3.2. *The matrix* $\boldsymbol{\Sigma}$ *in conjugate multivariate models is posterior consistent.*

Proof. Consider the augmented linear model (4.3.6) built upon conjugate multivariate latent model. Let $\mathbf{u}(n) = \mathbf{Y}^*(n) - \mathbf{X}^*(n)\boldsymbol{\gamma}$, $\mathbf{u}(n) \mid \boldsymbol{\Sigma}_0 \sim \mathrm{MN}_{(2n+p)\times q}(\mathbf{0}, \mathbf{I}_{2n+p}, \boldsymbol{\Sigma}_0) = [\mathbf{u}_1, \ldots, \mathbf{u}_{2n+p}]^\top$,

$$\boldsymbol{\Psi}^*(n)/n = \boldsymbol{\Psi}/n + \frac{1}{n}\mathbf{u}(n)^\top \mathbf{u}(n) - \frac{1}{n}\mathbf{u}(n)^\top \mathbf{X}^*(n)(\mathbf{X}^*(n)^\top \mathbf{X}^*(n))^{-1}\mathbf{X}^*(n)^\top \mathbf{u}(n) \tag{4.3.8}$$

Since $\mathbf{X}^*(n)(\mathbf{X}^*(n)^\top \mathbf{X}^*(n))^{-1}\mathbf{X}^*(n)^\top$ is idempotent with rank $p+n$, there exist an orthogonal matrix $\mathbf{Q}(n)$ such that $\mathbf{X}^*(n)(\mathbf{X}^*(n)^\top \mathbf{X}^*(n))^{-1}\mathbf{X}^*(n)^\top = \mathbf{Q}(n)^\top \begin{bmatrix} \mathbf{I}_{p+n} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \mathbf{Q}(n)$. Let $\mathbf{v}(n) = \mathbf{Q}(n)\mathbf{u} = [\mathbf{v}_1, \ldots, \mathbf{v}_{2n+p}]^\top$, then $\mathbf{v}(n) \sim \mathbf{MN}_{(2n+p)\times q}(\mathbf{0}, \mathbf{I}_{2n+p}, \boldsymbol{\Sigma}_0)$ then by Khimchine-Kolmogorov strong law of large number, we have $\lim_{n\to\infty}\frac{2}{2n}\{\mathbf{u}(n)^\top \mathbf{u}(n)\}_{ij} = 2\boldsymbol{\Sigma}_{0ij}$ a.s. for $1 \le i, j, \le q$ and

$$\lim_{n\to\infty}\frac{1}{n}\{\mathbf{u}(n)^\top \mathbf{X}^*(n)(\mathbf{X}^*(n)^\top \mathbf{X}^*(n))^{-1}\mathbf{X}^*(n)^\top \mathbf{u}(n)\}_{ij} = \lim_{n\to\infty}\frac{1}{n}\sum_{l=1}^{p}\{\mathbf{v}_l^\top \mathbf{v}_l\}_{ij} = \boldsymbol{\Sigma}_{0ij}\ a.s.$$

where $\{A\}_{ij}$ is the (i,j)-th element of matrix A. Hence, $\lim_{n\to\infty}\boldsymbol{\Psi}^*(n)_{ij}/n = \boldsymbol{\Sigma}_{0ij}$ a.s. By lemma 4.3.1, we prove the posterior consistency of $\boldsymbol{\Sigma}$ □

Theorem 4.3.3. *$\boldsymbol{\beta}$ is posterior consistent for both conjugate models if and only if $lim_{n\to\infty}\lambda_{\min}(\mathbf{P}(n)) = \infty$, where $\lambda_{\min}(\mathbf{P}(n))$ is the smallest eigenvalue of matrix $\mathbf{P}(n)$.*

Proof. Through the augmented linear system (4.3.6) we can see that the marginal posterior mean of $\boldsymbol{\beta}$ based on $\boldsymbol{\beta} \mid \boldsymbol{\Sigma}, \mathbf{Y}(n)$ is an unbiased estimator of $\boldsymbol{\beta}_0$. When $\boldsymbol{\beta}$ is posterior consistent, i.e. $\lim_{n\to\infty}\Pr(|\boldsymbol{\beta}_{ij} - \boldsymbol{\beta}_{0ij}| > \epsilon \mid \mathbf{Y}(n)) = 0$ for any $\epsilon > 0$ for $1 \le i \le p, 1 \le j \le q$ then $\lim_{n\to\infty}\mathrm{Var}(\boldsymbol{\beta}_{ij} \mid \mathbf{Y}(n)) = 0$ a.s. Moreover, we can show that $\lim_{n\to\infty}\mathrm{Var}(\boldsymbol{\beta}_{ij} \mid \mathbf{Y}(n)) = 0$ a.s. is a sufficient condition for the posterior consistency of $\boldsymbol{\beta}$ through Chebyshev's inequality. Consider the posterior marginal distribution of $\boldsymbol{\beta}$ of conjugate multivariate response model, which follows a matrix-t distribution $\boldsymbol{\beta} \mid \mathbf{Y}(n) \sim \mathrm{T}_{p,q}(\nu^*(n) - q + 1, \boldsymbol{\mu}^*(n), \mathbf{V}^*(n), \boldsymbol{\Psi}^*(n))$ with parameters given in (4.3.3). As proved in Theorem 4.3.2, $\lim_{n\to\infty}\boldsymbol{\Psi}^*(n)_{ij}/n = \boldsymbol{\Sigma}_{0ij}$ a.s. for $1 \le i, j, \le q$, therefore we have $\lim_{n\to\infty}\mathrm{Var}(\boldsymbol{\beta}_{ij} \mid \mathbf{Y}(n)) = 0$ a.s. if and only if $\lim_{n\to\infty}\{\mathbf{V}^*(n)\}_{ii} = 0$ for all $i = 1, \ldots, q$, Then follows Eicker [1963]'s proof of Theorem 1, the sufficient and necessary condition is $\lim_{n\to\infty}\lambda_{\min}(\mathbf{V}^{*-1}(n)) = \infty$. Since

$$\lambda_{\min}(\mathbf{P}(n)) + \lambda_{\max}(\mathbf{V_r}^{-1}) \ge \lambda_{\min}(\mathbf{V}^{*-1}(n)) = \lambda_{\min}(\mathbf{P}(n) + \mathbf{V_r}^{-1}) \ge \lambda_{\min}(\mathbf{P}(n))\ , \tag{4.3.9}$$

the condition can be simplified into $\lim_{n\to\infty}\lambda_{\min}(\mathbf{P}(n))=\infty$. □

Remark 4.3.4. *$\lambda_{\min}(\mathbf{P}(n))$ is non-decreasing, and when $\boldsymbol{\beta}$ is posterior consistent, $\lim_{n\to\infty}\mathbf{P}(n)_{ii}=\infty$ since $\mathbf{P}(n)_{ii}\geq\lambda_{\min}(\mathbf{P}(n))$*

Proof. Let $\mathbf{X}(n+1)=[\mathbf{X}(n)^\top,x_{n+1}^\top]^\top$, $\boldsymbol{\mathcal{K}}(n+1)=\begin{bmatrix}\boldsymbol{\mathcal{K}}(n) & \boldsymbol{\mathcal{K}}_{(n),n+1}\\ \boldsymbol{\mathcal{K}}_{n+1,(n)} & \alpha^{-1}\end{bmatrix}$. Then

$$\begin{aligned}\mathbf{P}(n+1)&=[\mathbf{X}(n)^\top,x_{n+1}^\top]\begin{bmatrix}\boldsymbol{\mathcal{K}}(n) & \boldsymbol{\mathcal{K}}_{(n),n+1}\\ \boldsymbol{\mathcal{K}}_{n+1,(n)} & \alpha^{-1}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{X}(n)\\ x_{n+1}\end{bmatrix}\\ &=\mathbf{P}(n)+(\mathbf{X}(n)^\top\boldsymbol{\mathcal{K}}(n)^\top\boldsymbol{\mathcal{K}}_{(n),n+1}-x_{n+1}^\top)d(\boldsymbol{\mathcal{K}}_{n+1,(n)}\boldsymbol{\mathcal{K}}(n)\mathbf{X}(n)-x_{n+1})\\ &=\mathbf{P}(n)+\mathbf{A}(n)\end{aligned}\tag{4.3.10}$$

where $d=(\alpha^{-1}-\boldsymbol{\mathcal{K}}_{n+1,(n)}\boldsymbol{\mathcal{K}}(n)^{-1}\boldsymbol{\mathcal{K}}_{(n),n+1})>0$ and $\mathbf{A}(n)$ is positive semi-definite symmetric matrix. Thus

$$\lambda_{\min}(\mathbf{P}(n+1))=\lambda_{\min}(\mathbf{P}(n)+\mathbf{A}(n))\geq\lambda_{\min}(\mathbf{P}(n))$$

□

Remark 4.3.5. *When $\mathbf{X}(n)\sim MN(\mathbf{0},\boldsymbol{\mathcal{K}}(n),\boldsymbol{\Sigma}^*)$ for some $\boldsymbol{\Sigma}^*$, $\boldsymbol{\beta}$ is posterior consistent.*

Proof. Let $\boldsymbol{\mathcal{K}}(n)^{-\frac{1}{2}}$ be the square root of $\boldsymbol{\mathcal{K}}(n)^{-1}$. Then we have $\boldsymbol{\mathcal{K}}(n)^{-\frac{1}{2}}\mathbf{X}(n)\sim\mathrm{MN}(\mathbf{0},\mathbf{I}_n,\boldsymbol{\Sigma}^*)$. By Khimchine-Kolmogorov strong law of large number, we have

$$\lim_{n\to\infty}\{\frac{1}{n}\mathbf{P}(n)\}_{ij}=\lim_{n\to\infty}\{\frac{1}{n}\mathbf{X}^\top(n)\boldsymbol{\mathcal{K}}(n)^{-\frac{1}{2}}\boldsymbol{\mathcal{K}}(n)^{-\frac{1}{2}}\mathbf{X}(n)\}_{ij}=\boldsymbol{\Sigma}^*_{ij}\ a.s.\text{ for }1\leq i,j\leq p\ .$$

Hence $\lambda_{\min}(\mathbf{P}(n))\to\infty$. □

This Remark shows that when the explanatory variables shares the same spatial correlation with the responses, the sufficient and necessary condition in Theorem 4.3.3 holds.

Remark 4.3.6. *When $\mathbf{X}(n)\sim MN(\mathbf{0},\mathbf{I}_n,\boldsymbol{\Sigma}^*)$ for some $\boldsymbol{\Sigma}^*$, $\boldsymbol{\beta}$ is posterior consistent*

Proof. For any n, there exist an orthogonal matrix $\mathbf{Q}(n)$ and a diagonal matrix $\mathbf{D}(n) = \text{diag}(\{d_i\}_{i=1}^n)$ such that $\mathbf{C}(\mathcal{S}(n), \mathcal{S}(n)) = \mathbf{Q}(n)^\top \mathbf{D}(n)\mathbf{Q}(n)$

$$\begin{aligned}\mathbf{P}(n) &= \mathbf{X}(n)^\top \mathbf{Q}(n)^\top (\mathbf{D}(n) + (\alpha^{-1} - 1)\mathbf{I}_n)^{-1}\mathbf{Q}(n)\mathbf{X}(n) \\ &= \mathbf{Z}(n)^\top \text{diag}(\{\frac{1}{d_i + (\alpha^{-1} - 1)}\}_{i=1}^n)\mathbf{Z}(n) = \sum_{i=1}^n \frac{1}{d_i + (\alpha^{-1} - 1)}\mathbf{z}_i\mathbf{z}_i^\top\end{aligned} \tag{4.3.11}$$

where $\mathbf{Z}(n) = [\mathbf{z}_i : \cdots : \mathbf{z}_n]^\top \sim \text{MN}(\mathbf{0}, \mathbf{I}_n, \boldsymbol{\Sigma}^*)$ and $\sum_{i=1}^n d_i = n$, $d_i \geq 0, i = 1, \ldots, n$. Define $\mathbf{V}_i = \mathbf{z}_i\mathbf{z}_i^\top, i = 1, \ldots, n$, by matrix version of Cauchy-Schwarz inequality [Marshall and Olkin, 1990, equation 4]

$$\sum_{i=1}^n (d_i + (\alpha^{-1} - 1)) \sum_{i=1}^n \frac{1}{d_i + (\alpha^{-1} - 1)}\mathbf{V}_i \geq \left\{\sum_{i=1}^n \sqrt{d_i + (\alpha^{-1} - 1)}\frac{\mathbf{V}_i^{\frac{1}{2}}}{\sqrt{d_i + (\alpha^{-1} - 1)}}\right\}^2,$$

where $\mathbf{V}_i^{\frac{1}{2}}\mathbf{V}_i^{\frac{1}{2}} = \mathbf{V}_i$, we have

$$\mathbf{P}(n) \geq \alpha\left\{\frac{\sum_{i=1}^n \mathbf{V}_i^{\frac{1}{2}}}{\sqrt{n}}\right\}^2 \tag{4.3.12}$$

Let $\mathbf{V}_i = \mathbf{z}_i\mathbf{z}_i^\top = \lambda_i\mathbf{u}_i\mathbf{u}_i^\top$ where $\lambda_i = \mathbf{z}_i^\top\mathbf{z}_i = \|\mathbf{z}_i\|^2$ and $\mathbf{u}_i = \frac{\mathbf{z}_i}{\|\mathbf{z}_i\|}$. Then we have $\mathbf{V}_i^{\frac{1}{2}} = \sqrt{\lambda_i}\mathbf{u}_i\mathbf{u}_i^\top$. Now change n into np and rewrite $\sum_{i=1}^n \mathbf{V}_i^{\frac{1}{2}}$ into $\sum_{i=1}^n\sum_{k=1}^p \mathbf{V}_{ik}^{\frac{1}{2}}$, where $\{\sum_{k=1}^p \mathbf{V}_{ik}^{\frac{1}{2}}\}$ for each i is a full rank $p \times p$ matrix with probability 1, we recast (4.3.12) into

$$\mathbf{P}(np) \geq \frac{\alpha}{p}\left\{\frac{\sum_{i=1}^n\sum_{k=1}^p \sqrt{\lambda_{ik}}\mathbf{u}_{ik}\mathbf{u}_{ik}^\top}{\sqrt{n}}\right\}^2 \tag{4.3.13}$$

And we are going to show that the smallest eigenvalue of the matrix on the right side goes to infinity as $n \to \infty$, which implies that $\lambda_{\min}(\mathbf{P}(np)) \to \infty$. Since $\sum_{k=1}^p \mathbf{V}_{ik}^{\frac{1}{2}} = \sum_{k=1}^p \sqrt{\lambda_{ik}}\mathbf{u}_{ik}\mathbf{u}_{ik}^\top \sim \mathbf{W}_p(\boldsymbol{\Sigma}^{*\frac{1}{2}}, p)$, where $\mathbf{W}_p$ is Wishart distribution, $\{\mathbf{u}_{i1}, \ldots, \mathbf{u}_{ip}\}$ compose the bases of the space $\mathbb{R}^p$ with probability 1. For any $\mathbf{u} \in \mathcal{R}^p$, $\|\mathbf{u}\| = 1$, we have

$$\mathbf{u}^\top \sum_{k=1}^p \left(\sqrt{\lambda_{ik}}\mathbf{u}_{ik}\mathbf{u}_{ik}^\top\right)\mathbf{u} \geq \min_{k=1,\ldots,p}\{\sqrt{\lambda_{ik}}\}$$

Hence, $\lambda_{\min}(\sum_{i=1}^n\sum_{k=1}^p \sqrt{\lambda_{ik}}\mathbf{u}_{ik}\mathbf{u}_{ik}^\top) \geq \sum_{i=1}^n \min_{k=1,\ldots,p}\{\sqrt{\lambda_{ik}}\}$. Since $\lambda_{ik} = \|\mathbf{z}_{ik}\|^2$ where $\mathbf{z}_{ik} \sim \text{N}(\mathbf{0}, \boldsymbol{\Sigma}^*)$, $\min_{k=1,\ldots,p}\{\sqrt{\lambda_{ik}}\}$ are independent and identically distributed with a positive mean $\mathbb{E}(\min_{k=1,\ldots,p}\{\sqrt{\lambda_{ik}}\}) = c^* > 0$ and a finite variance σ^{2*}. By law of large number, we have

$\lim_{n\to\infty} \sum_{i=1}^{n} \min\limits_{k=1,\dots,p}\{\sqrt{\lambda_{ik}}\}/n = c^*$ a.s.. Therefore,

$$\lambda_{\min}\left(\left\{\frac{\sum_{i=1}^{n}\sum_{k=1}^{p}\sqrt{\lambda_{ik}}\mathbf{u}_{ik}\mathbf{u}_{ik}^{\top}}{\sqrt{n}}\right\}^2\right) \geq \frac{1}{n}\left\{\sum_{i=1}^{n}\min_{k=1,\dots,p}\{\sqrt{\lambda_{ik}}\}\right\}^2 \to \infty$$

By (4.3.13), $\lim_{n\to\infty}\lambda_{\min}(\mathbf{P}(n)) = \infty$. □

This Remark shows that when the explanatory variables can be viewed as independent observations, the sufficient and necessary condition in Theorem 4.3.3 holds. Remark 2.2 and 2.3 discuss two common situations for the design matrix, which support that the condition in Theorem 2 is a general condition.

Now let us consider spatial regression model

$$\mathbf{y}(\mathbf{s}) = \boldsymbol{\beta}^{\top}\mathbf{x}(\mathbf{s}) + \boldsymbol{\omega}(\mathbf{s}) + \boldsymbol{\epsilon}(\mathbf{s}) \;\; \text{with } \boldsymbol{\omega}(\mathbf{s}) \sim \text{GP}(\mathbf{0}_{q\times 1}, \rho_{\psi}(\cdot,\cdot)\boldsymbol{\Sigma}_{\boldsymbol{\omega}})\ ,\ \boldsymbol{\epsilon}(\mathbf{s}) \sim \text{N}(\mathbf{0}_{q\times 1}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})\ , \tag{4.3.14}$$

where the covariance matrix of $\boldsymbol{\epsilon}(\mathbf{s})$ is $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ and non-spatial covariance matrix of $\boldsymbol{\omega}(\mathbf{s})$ is $\boldsymbol{\Sigma}_{\boldsymbol{\omega}}$.

Corollary 4.3.7. *With any covariance matrix of noise process $\boldsymbol{\epsilon}(\mathbf{s})$, $\boldsymbol{\beta}\,|\,\boldsymbol{\Sigma}, \mathbf{Y}(n)$ given in form of matrix normal with parameters defined in* (4.3.3) *&* (4.3.5) *are consistent if and only if* $\lim_{n\to\infty}\lambda_{\min}(\mathbf{P}(n)) = \infty$.

Proof. Without loss of generalization, assign a flat prior for $\boldsymbol{\beta}$. Reformulate the augmented linear system (4.3.6) as

$$\underbrace{\begin{bmatrix} \sqrt{\frac{\alpha}{1-\alpha}}\mathbf{Y}(n) \\ \mathbf{0} \end{bmatrix}}_{\mathbf{Y}^*(n)} = \underbrace{\begin{bmatrix} \sqrt{\frac{\alpha}{1-\alpha}}\mathbf{X}(n) & \sqrt{\frac{\alpha}{1-\alpha}}\mathbf{I}_n \\ \mathbf{0} & \mathbf{V}_{\boldsymbol{\rho}} \end{bmatrix}}_{\mathbf{X}^*(n)} \underbrace{\begin{bmatrix} \boldsymbol{\beta} \\ \boldsymbol{\omega}(\sqrt{\frac{\alpha}{1-\alpha}}\boldsymbol{\Sigma}_{\boldsymbol{\omega}}^{-\frac{1}{2}}\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}^{\frac{1}{2}}) \end{bmatrix}}_{\tilde{\gamma}(n)} + \underbrace{\begin{bmatrix} \tilde{\boldsymbol{\eta}}_1 \\ \tilde{\boldsymbol{\eta}}_2 \end{bmatrix}}_{\tilde{\boldsymbol{\eta}}(n)}\ ,$$

then $\tilde{\boldsymbol{\eta}}(n) \sim \text{MN}(\mathbf{0}, \mathbf{I}_{2n}, \frac{\alpha}{1-\alpha}\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$, $\tilde{\gamma}(n)\,|\,\boldsymbol{\Sigma}_{\boldsymbol{\omega}}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}, \mathbf{Y}(n) \sim \text{MN}(\boldsymbol{\mu}^*(n), \mathbf{V}^*(n), \frac{\alpha}{1-\alpha}\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$ where $\boldsymbol{\mu}^*(n)$ and $\mathbf{V}^*(n)$ are in form of (4.3.7). Through the above linear system, we can see than the first p rows of $\boldsymbol{\mu}^*(n)$ provides an unbiased estimator of $\boldsymbol{\beta}$ with any $\alpha \in (0, 1)$. If elements in $\lim_{n\to\infty}\boldsymbol{\Psi}^*(n)/n$ are bounded, then we can finish the proof by following the proof of Theorem 4.3.3.

Next, check elements in $\lim_{n\to\infty} \mathbf{\Psi}^*(n)/n$. Let $\mathbf{u}(n) = \mathbf{Y}^*(n) - \mathbf{X}^*(n)\boldsymbol{\gamma} = [\mathbf{u}_1 : \cdots : \mathbf{u}_n : \mathbf{u}_{n+1} \cdots \mathbf{u}_{2n}]^\top$, follow the definition of model (4.3.14) we have $\mathbf{u}_i \sim \mathrm{N}(\mathbf{0}, \frac{\alpha}{1-\alpha}\mathbf{\Sigma}_{\boldsymbol{\epsilon}})$ for $i = 1, \ldots, n$ and $\mathbf{u}_i \sim \mathrm{N}(\mathbf{0}, \mathbf{\Sigma}_{\boldsymbol{\omega}})$ for $i = n+1, \ldots, 2n$, Through (4.3.8) we have

$$\mathbf{\Psi}^*(n)/n \leq \mathbf{\Psi}/n + \frac{1}{n}\mathbf{u}(n)^\top \mathbf{u}(n) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{u}_i \mathbf{u}_i^\top + \frac{1}{n}\sum_{i=n+1}^{2n} \mathbf{u}_i \mathbf{u}_i^\top$$

Since $\sum_{i=1}^{n} \mathbf{u}_i \mathbf{u}_i^\top \sim \mathrm{W}_q(\frac{\alpha}{1-\alpha}\mathbf{\Sigma}_{\boldsymbol{\epsilon}}, n)$ and $\sum_{i=n+1}^{2n} \mathbf{u}_i \mathbf{u}_i^\top \sim \mathrm{W}_q(\mathbf{\Sigma}_{\boldsymbol{\omega}}, n)$ we can see that

$$\lim_{n\to\infty} \left\{ \frac{1}{n}\sum_{i=1}^{n} \mathbf{u}_i \mathbf{u}_i^\top + \frac{1}{n}\sum_{i=n+1}^{2n} \mathbf{u}_i \mathbf{u}_i^\top \right\}_{ij} = \left\{ \frac{\alpha}{1-\alpha}\mathbf{\Sigma}_{\boldsymbol{\epsilon}} \right\}_{ij} + \mathbf{\Sigma}_{\boldsymbol{\omega} ij} \ a.s. \ 1 \leq i, j \leq q$$

So elements in $\lim_{n\to\infty} \mathbf{\Psi}^*(n)/n$ are bounded. □

4.4 Simulation

We present two simulation examples here. The first compares our proposed BSLMC model with other multivariate Bayesian spatial models. The second assesses our factor BSLMC model when K is not excessively large. Our proposed models were implemented in Julia 1.2.0 [Bezanson et al., 2017]. We modeled the univariate processes in the proposed BSLMC by NNGP. We took the Bayesian LMC model proposed by Schmidt and Gelfand [2003] as a benchmark in the first simulation example. The benchmark model was implemented in R 3.4.4 through function *spMisalignLM* in the R package *spBayes*[Finley et al., 2007]. We also fitted a response NNGP model with misalignment in Julia 1.2.0 in the first example. The detailed algorithms for the response NNGP model with misalignment is in the Section A.1 in Appendix. The most demanding model took approximately 21 hours to deliver its entire inferential output involving 20,000 MCMC iterations on a single 8 Intel Core i7-7700K CPU @ 4.20GHz processor with 32 Gbytes of random-access memory running Ubuntu 18.04.2 LTS. Convergence diagnostics and other posterior summaries were implemented within the Julia statistical environment. Each model was compared in terms of the posterior inference of parameters (posterior mean and 95% confidence interval), root mean squared predict error (RMSPE), mean squared error of intercept-centered latent processes (MSEL), prediction interval coverage (CVG; the percent of intervals containing the true value), interval coverage for intercept-centered latent process of observed response (CVGL), average continuous rank probability score (CRPS; see Gneiting and Raftery [2007]) for responses, and the average interval score (INT; see Gneiting and Raftery [2007]) for responses and run time. To calculate the CRPS and INT, we assumed that the associated predictive distribution was well approximated by a Gaussian distribution with mean centered at the predicted value and standard deviation equal to the predictive standard error. All NNGP models were specified with at most $m = 10$ nearest neighbors.

4.4.1 Simulation Example 1

We simulated the response $\mathbf{y}(\mathbf{s})$ from the SLMC model in (4.2.2) with $q = 2, p = 2, K = 2$ over 1200 randomly generated locations over a unit square. The size of the data set was kept moderate to enable comparisons with the expensive full GP based LMC models for experiments conducted on the computing setup described earlier. The explanatory variable $\mathbf{x}(\mathbf{s})$ consists of an intercept and a single predictor generated from a standard normal distribution. An exponential correlation function was used to model $\{\rho_{\psi_k}(\cdot,\cdot)\}_{k=1}^{K}$, i.e., $\rho_{\psi_k}(\mathbf{s},\mathbf{s}') = \exp\left(-\phi_k\|\mathbf{s}-\mathbf{s}'\|\right)$, for $\mathbf{s},\mathbf{s}' \in \mathcal{D}$, where $\|\mathbf{s}-\mathbf{s}'\|$ is the Euclidean distance between $\mathbf{s}$ and $\mathbf{s}'$, and $\psi_k = \phi_k$ is the decay for each k. We took $\boldsymbol{\Sigma} = \text{diag}([0.3, 0.2])$ and let $\boldsymbol{\Lambda}$ in (4.2.2) be an upper triangular matrix. We randomly picked 200 locations for predicting each response to examine the predictive performance. Since the data set has misalignment, we present inference from a response NNGP model with misalignment (resp NNGP), BSLMC, and Benchmark LMC model. The values of covariance of measurement error (labeled as $\text{cov}(\boldsymbol{\epsilon})$) and non-spatial covariance of latent process (labeled as $\text{cov}(\boldsymbol{\omega})$) as well as other parameters are listed in table 4.1.

We assigned flat priors for $\{\boldsymbol{\beta}, \boldsymbol{\Lambda}\}$ for response NNGP model with misalignment and BSLMC, respectively. The prior for $\boldsymbol{\Sigma}$ for two models was set to follow $\text{IW}(\boldsymbol{\Psi}, \nu)$ with $\boldsymbol{\Psi} = \text{diag}([1.0, 1.0])$ and $\nu = 3$. For the benchmark LMC, we assigned a flat prior for $\boldsymbol{\beta}$, $\text{IW}(\boldsymbol{\Psi}, \nu)$ with $\boldsymbol{\Psi} = \text{diag}([1.0, 1.0])$ and $\nu = 3$ for the cross-covariance matrix $\boldsymbol{\Lambda}^\top\boldsymbol{\Lambda}$, and

IG$(2, 0.5)$ for each diagonal element of $\mathbf{\Sigma}$. The candidate values for $\{\phi, \alpha\}$ used in cross-validation algorithm for response NNGP model with misalignment over a 25 by 25 grid over $[2.12, 26.52] \times [0.8, 0.99]$. We gave unif$(2.12, 212)$ as priors of decays for BSLMC and benchmark LMC model. The posterior inference from the response NNGP with misalignment, BSLMC as well as the benchmark LMC model were based on an MCMC chain with 20,000 iterations, and we took the first $15,000$ samples as burn-in. The number of iterations of all MCMC chains was taken to be large enough to guarantee their convergence.

All three models provided close posterior inferences for $\{\boldsymbol{\beta}_{21}, \boldsymbol{\beta}_{21}\}$. The 95% confidence intervals of the intercepts $\{\boldsymbol{\beta}_{11}, \boldsymbol{\beta}_{12}\}$ all include the true value used to generate the data. With a mismatch of data generating schemes and model assumptions, the response NNGP model with misalignment provided incorrect inference for cov$(\boldsymbol{\epsilon})$ when compared to the other two candidate models. The RMSPEs and CVGs, however, are close to BSLMC and benchmark LMC. Compared to benchmark LMC which cost around 21 hours, the response NNGP model spent less that 0.5 minute, suggesting that fitting the response NNGP model with misalignment is a pragmatic way to have reliable interpolation and predictions. The NNGP based BSLMC model costs 4.5 minutes, while the Benchmark LMC model costs around 21 hours. Yet, despite the shorter running time, we observed superior performance of the NNGP based BSLMC models than the benchmark LMC for inferring on the latent process using CVGL. Moreover, the interpolated map of the recovered intercept-centered latent processes (figure 4.1) by BSLMC and benchmark LMC are almost indistinguishable from each other. BSLMC and benchmark LMC produce very similar MSELs, RMSPEs, CRPS and INT. Benchmark LMC yields better estimates for the spatial decays but poorer inference for cov$(\boldsymbol{\omega})$. The differences in estimates between the two models is likely emerging from the different prior settings and sampling schemes. Benchmark LMC restricts the loading matrix $\mathbf{\Lambda}$ to be upper triangular, while BSLMC does not, resulting in greater flexibility in fitting latent process. On the other hand, the unidentifiable parameter setting of BSLMC may cause less somewhat less stable inference for the hyperparameters $\{\phi_1, \phi_2\}$.

Table 4.1: Simulation study summary table: posterior mean (2.5%, 97.5%) percentiles

	True	resp NNGP	BSLMC	benchmark LMC
$\boldsymbol{\beta}_{11}$	1.0	0.761(0.13, 1.376)	0.877(0.399, 1.378)	0.79 (0.344, 1.229)
$\boldsymbol{\beta}_{12}$	-1.0	-1.048(-1.971, -0.09)	-1.605(-2.078, -0.977)	-0.795(-2.069, 0.74)
$\boldsymbol{\beta}_{21}$	-5.0	-4.958(-5.068, -4.847)	-4.968(-5.113, -4.819)	-4.968(-5.115, -4.822)
$\boldsymbol{\beta}_{22}$	2.0	1.925(1.763, 2.087)	1.93(1.719, 2.124)	1.933 (1.731, 2.134)
$\mathrm{cov}(\boldsymbol{\epsilon})_{11}$	0.3	**0.17 (0.156, 0.185)**	0.277 (0.231, 0.324)	0.275 (0.233, 0.326)
$\mathrm{cov}(\boldsymbol{\epsilon})_{12}$	0.0	**-0.052(-0.071, -0.036)**	0.023 (-0.031, 0.073)	0.0
$\mathrm{cov}(\boldsymbol{\epsilon})_{22}$	0.2	**0.376(0.344, 0.411)**	0.221 (0.145, 0.307)	0.244 (0.165, 0.322)
$\mathrm{cov}(\boldsymbol{\omega})_{11}$	0.683	**1.58(1.451, 1.719)**	0.707 (0.636, 0.778)	0.706 (0.639, 0.773)
$\mathrm{cov}(\boldsymbol{\omega})_{12}$	-0.616	**-0.488 (-0.656, -0.33)**	-0.596(-0.685, -0.504)	**-0.07(-0.115, -0.024)**
$\mathrm{cov}(\boldsymbol{\omega})_{22}$	4.517	**3.5(3.203, 3.826)**	4.372 (4.2, 4.536)	**4.311(4.15, 4.455)**
ϕ_1	6.0	7.204($\alpha = 0.903$)	**2.926(2.213, 3.941)**	8.63 (5.251, 12.711)
ϕ_2	6.0	7.204($\alpha = 0.903$)	7.771(3.963, 12.226)	6.045(3.731, 8.526)
RMSPE[a]	–	[0.643, 0.948, 0.81]	[0.633, 0.917, 0.788]	[0.633, 0.918, 0.788]
MSEL [b]	–	–	[0.111, 0.139, 0.125]	[0.111, 0.14, 0.126]
CRPS[a]	–	[-0.366, -0.535, -0.45]	[-0.359, -0.515, -0.437]	[-0.359, -0.515, -0.437]
CRPSL[b]	–	–	[-0.031, -0.036, -0.033]	[-0.189, -0.212, -0.2]
CVG[a]	–	[0.945, 0.955, 0.95]	[0.965, 0.945, 0.955]	[0.965, 0.945, 0.955]
CVGL [b]	–	–	[0.941, 0.971, 0.956]	[0.791, 0.816, 0.803]
INT[a]	–	[3.031, 4.324, 3.678]	[2.929, 4.327, 3.628]	[2.927, 4.315, 3.621]
INTL[b]	–	–	[0.253, 0.278, 0.265]	[1.535, 1.728, 1.631]
time(s)	–	[14, 2, 12][c]	270	[51456, 23973][d]

[a] [response 1, response 2, all responses]

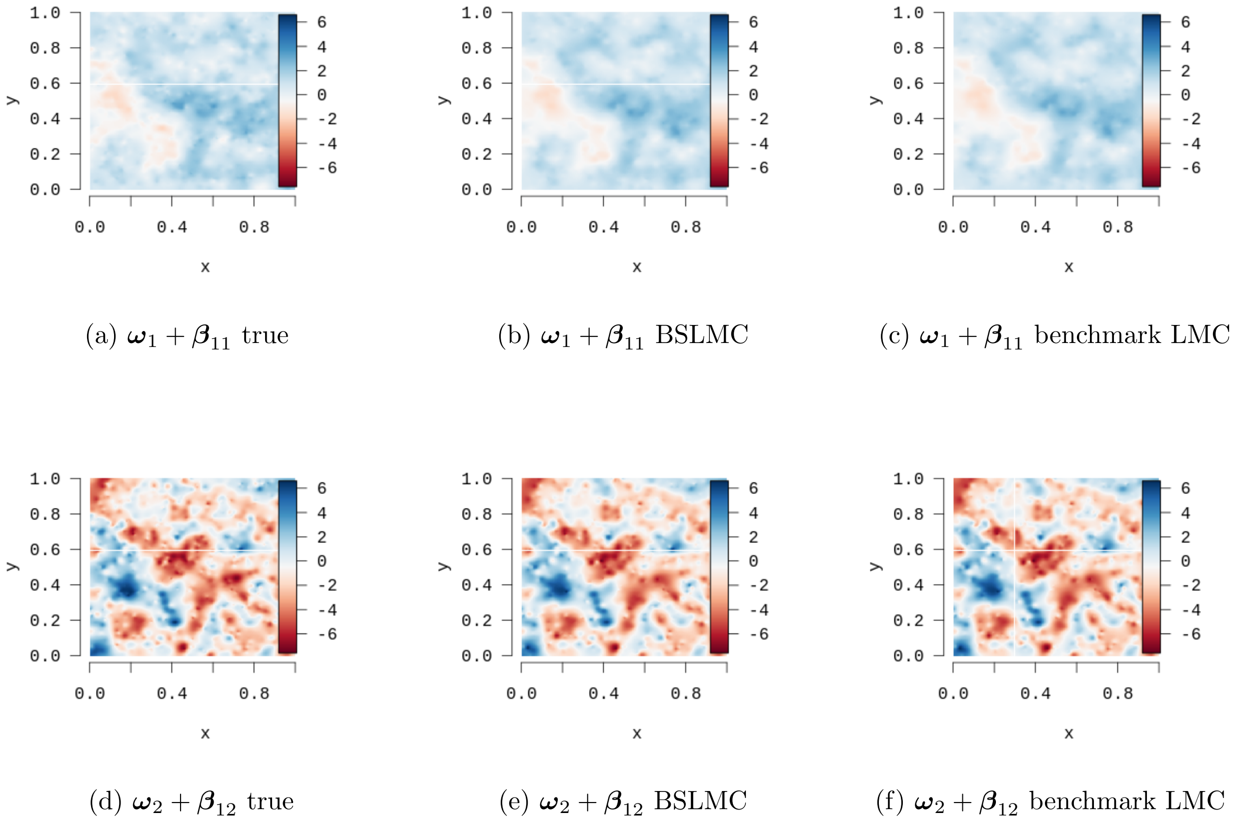
[b] intercept + latent process on 1000 observed locations for [response 1, response 2, all responses]

[c] [time for cross-validation, time for MCMC sampling, time for recovering $\boldsymbol{\beta}$ and predictions]

[d] [time for MCMC sampling, time for recovering predictions]

4.4.2 Simulation Example 2

We generated the second dataset by SLMC model (4.2.2) with a diagonal $\boldsymbol{\Sigma}$ and $q = 10, p = 3, K = 50$ over 1200 randomly generated locations over a unit square. The explanatory variable $\mathbf{x}(\mathbf{s})$ was composed of an intercept and two predictors generated independently



(a) $\boldsymbol{\omega}_1 + \boldsymbol{\beta}_{11}$ true

(b) $\boldsymbol{\omega}_1 + \boldsymbol{\beta}_{11}$ BSLMC

(c) $\boldsymbol{\omega}_1 + \boldsymbol{\beta}_{11}$ benchmark LMC

(d) $\boldsymbol{\omega}_2 + \boldsymbol{\beta}_{12}$ true

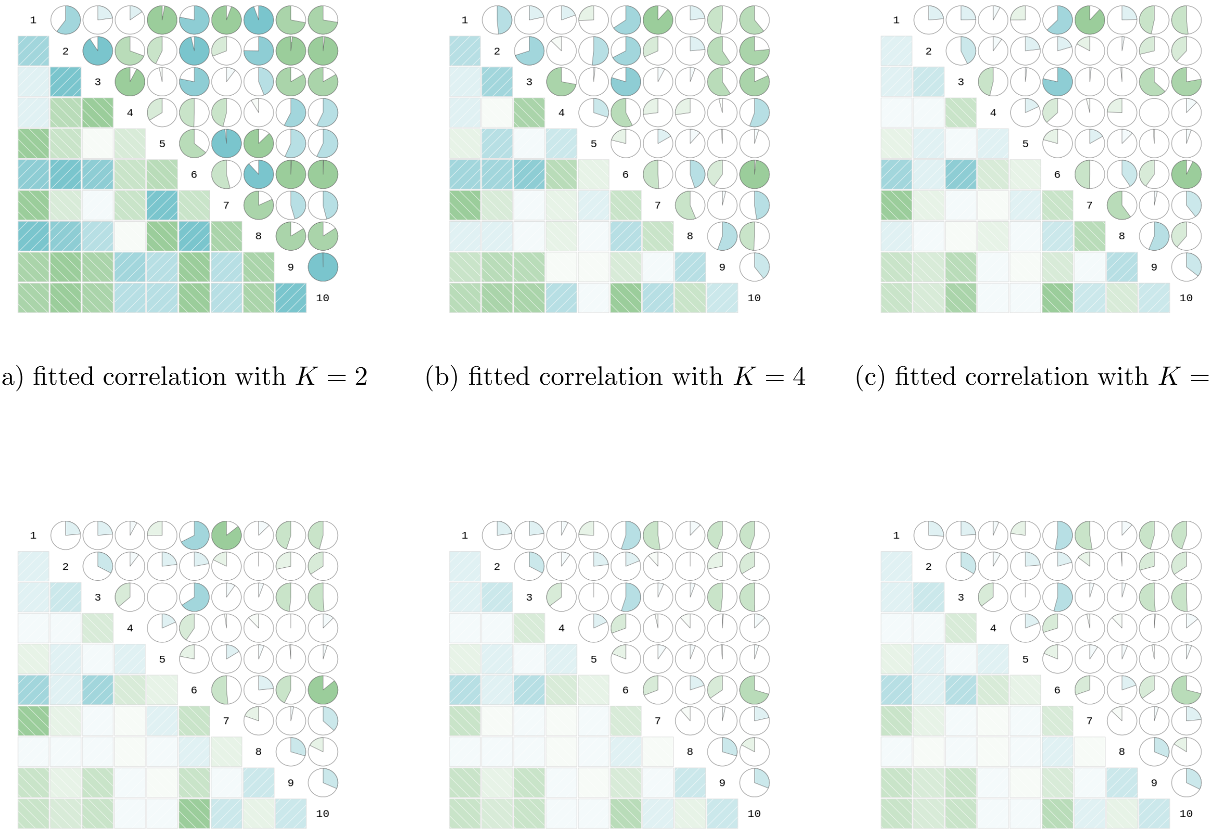
(e) $\boldsymbol{\omega}_2 + \boldsymbol{\beta}_{12}$ BSLMC

(f) $\boldsymbol{\omega}_2 + \boldsymbol{\beta}_{12}$ benchmark LMC

Figure 4.1: Interpolated maps of (a) & (d) the true generated intercept-centered latent processes, the posterior means of the intercept-centered latent process $\boldsymbol{\omega}$ from the (b) & (e) NNGP based BSLMC model and the (c) & (f) benchmark LMC model.

from a standard normal distribution. We used an exponential covariance function to model $\{\rho_{\psi_k}(\cdot,\cdot)\}_{k=1}^{K}$, where $\psi_k = \phi_k$ denotes the decay for $k = 1, \ldots, K$. This data set features a relatively large number of responses ($q = 10$) and a complicated pattern in latent processes ($K = 50$). We randomly selected 200 locations for prediction for each response.

We fitted a factor BSLMC model with diagonal $\boldsymbol{\Sigma}$ with K from 1 to 10. The goal of this simulation example is to check the performance of a factor BSLMC model, especially in recovering latent processes, when K is not sufficiently large. We assigned a $\Gamma(2, 11.67)$ prior for all $\{\phi_k\}_{k=1}^{K}$ and we set flat priors for $\boldsymbol{\beta}$ and $\boldsymbol{\Lambda}$. All diagonal elements of $\boldsymbol{\Sigma}$ were assigned an $\text{IG}(2, 1.0)$ prior. The setting for the MCMC sampling scheme follows that of BSLMC in



(a) fitted correlation with $K = 2$

(b) fitted correlation with $K = 4$

(c) fitted correlation with $K = 6$

(d) fitted correlation with $K = 8$

(e) fitted correlation $K = 10$

(f) correlation of the raw data

Figure 4.2: Heat-maps of the (l) actual and (g)-(k) fitted non-spatial correlation of $\boldsymbol{\omega}(\mathbf{s})$

the first example.

The running time for executing the models along with CVGL, CVG, and RMSPE are listed in table 4.2. We also added CVG-slope in table 4.2, which counts the number of 95% CIs of regression slopes that include the true value. Inference for the regression slopes was found to be robust to the choice of K. The CVG for each K was close to 0.95, while RMSPE decreased rapidly as K increased. We also found that factor BSLMC would fit the latent processes better for some of the responses that the others when K was small. The performance metrics quickly improved as K increased from 1 to 10, RMSPE decreased by 45.7% and CVGL hit 75% for $K \geq 7$.

Table 4.2: Simulation study summary table 2:

K =	1	2	3	4	5	6	7	8	9	10
CVG-slope	19/20	18/20	18/20	18/20	19/20	19/20	19/20	20/20	20/20	20/20
CVGL	0.3068	0.4175	0.5103	0.5999	0.6642	0.7236	0.7864	0.7934	0.8462	0.8681
CVG	0.9445	0.942	0.9375	0.943	0.9435	0.9435	0.941	0.938	0.9385	0.94
RMSPE	4.6531	4.3852	4.0105	3.7076	3.5578	3.2946	3.0944	2.9314	2.716	2.527
time(s)	235	422	836	1268	1891	2417	3214	3635	4880	5248

We compare the correlation across different latent processes (referred as non-spatial correlation) to check the performance of different models in estimating the latent processes. Figures 4.2a through 4.2f illustrate the heat-maps of the non-spatial correlation of the fitted and the true latent processes. As K increases from 2 to 10, the estimated heat-maps approach the true correlation matrix. It can be seen that the heat-map for the fitted correlation with $K = 10$ shared a similar pattern with that of the actual correlation. Given that our data set follows an SLMC model with $K = 50$, we can conclude that the factor BSLMC is efficient in obtaining inference for the latent processes even when K is not adequately large. The test also shows that the choice of K is important for obtaining reliable inference when using BSLMC with a diagonal $\mathbf{\Sigma}$ as a factor model. We recommend choosing K based on scientific considerations for the problem at hand and exploratory data analyses. One can also check the RMSPE value for different K and use an elbow rule [Thorndike, 1953] to choose K .

4.5 Real Data Analysis

We apply our proposed models to analyze Normalized Difference Vegetation Indices (NDVI) and Enhanced Vegetation Indices (EVI) measuring vegetation activity on the land surface, which can help us understand the global distribution of vegetation types as well as their biophysical and structural properties and spatial variations. Apart from the NDVI, we consider Gross Primary Productivity data, Global Terrestrial Evapotranspiration (ET) Product, and landcover data [see Ramon Solano et al., 2010, Mu et al., 2013, Sulla-Menashe and Friedl, 2018, for further details]. The geographic coordinates of our variables were mapped on a Sinusoidal (SIN) projection grid. We chose zone *h08v05*, which covers 11,119,505 to 10,007,555 meters south of the prime meridian and 3,335,852 to 4,447,802 meters north of the equator. The land region in zone *h08v05* is situated in the western United States. Our explanatory variables included an intercept and a binary indicator for no vegetation or urban area through the 2016 landcover data. All other variables were measured through MODIS satellite over a 16-days period from 2016.04.06 to 2016.04.21. Some variables were rescaled and transformed in exploratory data analysis for the sake of better model fitting. The data sets were downloaded using the R package *MODIS* and the code for the exploratory data analysis is provided as supplementary material to this paper.

Our data set comprises 1,020,000 randomly selected observed locations to illustrate BSLMC, response NNGP with misalignment and a factor BSLMC model with diagonal $\mathbf{\Sigma}$. Our spatially dependent outcomes were the transformed NDVI ($\log(\text{NDVI} + 1)$ labeled as NDVI) and red reflectance (red refl). A Bayesian linear model on the two data sets were also fitted for comparison. All NNGP based models specified at most $m = 10$ nearest neighbors. We randomly held 10% of each response and then held all responses over region 10,400,000 to 10,300,000 meters south of the prime meridian and 3,800,000 to 3,900,000 meters north of the equator to examine the predictive performance of models over a missing region and randomly missing locations. Figure 4.3a illustrates the map of the transformed NDVI data. The white square region within the Continent is the region held out for prediction.

The posterior inference for BSLMC and response NNGP with misalignment were based on an MCMC chain with 10,000 iterations. The priors for all parameters except decays follow those in the simulation section. We assigned $\Gamma(200, 0.02)$ and $\Gamma(200, 0.04)$ for ϕ_1 and ϕ_2 for BSLMC based on variograms fitted in exploratory data analysis. We recursively shrink the domain and the grid of candidate values $\{\phi, \alpha\}$ through repeatedly using cross-validation algorithms for fixing parameters for the response NNGP model with misalignment. The number of threads used in the cross-validation algorithms for response NNGP models with misalignment were equal to the number of folders. The remaining part of all the code were run with single thread.

Table 4.3 gives the results of BSLMC and response NNGP with misalignment. Consistent with the related background, the regression coefficients of the index of no vegetation or urban area show relatively low biomass (low NDVI) and high red reflectance over no vegetation or urban area. The inference of the covariance of the noise and non-spatial covariance of the latent process shows a negative association between the residuals and latent processes of transformed NDVI and red reflectance, which satisfies the underlying relationship between two responses. BSLMC captured a high negative correlation (≈ -0.87) between the latent processes of two responses, indicating that the spatial pattern of the latent processes of NDVI and red-reflectance are almost the reverse of each other. The maps of the latent processes recovered by BSLMC, presented in Figure 3.2, also support this relationship.

Each model was compared in terms of RMSPE, CVG, CRPS, INT and run time. It is clear that the spatial models greatly improved predictive accuracy. BSLMC and the response NNGP with misalignment reduced at least 50% RMSPE compared to the Bayesian linear model. CVG is similar among all models, while all spatial models provided a more accurate prediction than the Bayesian linear models based on INT and CRPS. Visual inspections of the recovered latent processes based on BSLMC are shown in figure 3.2. Notably, the proposed methods smooth out the predictions in the held-out region. The BSLMC model took around 44.7 hours. Regarding the scale of the multivariate spatial data set, the run time for BSLMC model is still appealing.

We also fitted a factor BSLMC with diagonal $\boldsymbol{\Sigma}$ to explore the underlying latent processes of ten (transformed) responses: (i) NDVI, (ii) EVI, (iii) Gross Primary Productivity (GPP), (iv) Net Photosynthesis (PsnNet), (v) red reflectance (red refl), (vi) blue reflectance (blue refl), (vii) average daily global evapotranspiration (ET), (viii) latent heat flux (LE), (ix) potential ET (PET) and (x) potential LE (PLE). We held all responses over the region picked in the previous example and randomly held 10% of each response to examine the predictive performance. There are in total $12,057$ locations with no responses and $656,366$

observed locations with misaligned data (at least one but not all responses), which covers 65.12% of observed locations. We provide a heat-map (figure 4.4k) to present the status of misalignment over the study domain.

Based on the exploratory analysis, we observed two groups of responses that have high within-group correlations but relatively low between-group correlations (see figure 4.3g). Hence we picked $K = 2$ for the factor BSLMC with diagonal $\mathbf{\Sigma}$. The results of the factor BSLMC model is presented in table 4.4. As shown in the summary table, no vegetation or urban area tend to have lower vegetation indexes (lower NDVI and EVI) and lower production of chemical energy in organic compounds by living organisms (lower GPP and PsnNet). We observe a trend of higher blue reflectance, red reflectance, evapotranspiration (higher ET LE) and lower potential evapotranspiration (lower PET PLE) in urban area and area with no vegetation. We present maps of the posterior prediction for all 10 variables in figure 4.4

The latent processes corresponding to transformed NDVI and red reflectance fitted through BSLMC and the factor BSLMC with diagonal $\mathbf{\Sigma}$ in figure 4.3 share a similar pattern. Finally, the heat map of the nonspatial correlation of the latent processes fitted by the factor BSLMC with diagonal $\mathbf{\Sigma}$, presented in figure 4.3h, reveals a high underlying correlation among NDVI, EVI, GPP, PsnNet, red and blue reflectance, and that LE and ET are slightly more correlated with NDVI and EVI than PLE and PET. The total run time for factor BSLMC with diagonal $\mathbf{\Sigma}$ was around 75 hours (4518.67 minutes).

4.6 Discussion

This Chapter aims at extending existing methodologies of scalable univariate spatial modeling to the multivariate cases. In this Chapter, we developed a variety of models that can implement scalable modeling methodologies for univariate spatial data in Bayesian multivariate spatial data analyses. One of the features of our models is that we can obtain a posterior inference of the high-dimensional latent process when it is well defined. The other main

Table 4.3: Real data analysis summary table 1: posterior mean (2.5%, 97.5%) percentiles

	Bayesian linear model	Response NNGP with misalign	BSLMC
intercept$_1$	0.25146(0.25117, 0.25176)	0.1662(0.1579, 0.1742)	–
intercept$_2$	0.13951(0.13937, 0.13965)	0.178(0.1733 , 0.1827)	–
no vege or urban area$_1$	-0.1338(-0.1349, -0.1327)	-1.066e-2 (-1.085e-2, -1.047e-2)	-1.385e-2 (-1.430e-2, -1.342e-2)
no vege or urban area$_2$	6.039e-2 (5.989e-2, 6.09e-2)	5.4625e-3 (5.3478e-3, 5.5733e-3)	7.831e-3 (7.584e-3, 8.097e-3)
$\mathrm{cov}(\boldsymbol{\epsilon})_{11}$	1.599e-2 (1.595e-2, 1.604e-2)	2.4628e-4 (2.4566e-4, 2.4702e-4)	3.51e-4 (3.48e-4, 3.55e-4)
$\mathrm{cov}(\boldsymbol{\epsilon})_{12}$	-6.494e-3(-6.515e-3, -6.474e-3)	-8.617e-5 (-8.649e-5, -8.585e-5)	-1.08e-4 (-1.10e-4, -1.07e-4)
$\mathrm{cov}(\boldsymbol{\epsilon})_{22}$	3.657e-3(3.647e-3, 3.668e-3)	7.672e-5 (7.652e-5, 7.692e-5)	1.07e-4 (1.06e-4, 1.08e-4)
$\mathrm{cov}(\boldsymbol{\omega})_{11}$	–	3.334e-2(3.325e-2, 3.344e-2)	1.675e-2(1.674e-2, 1.676e-2)
$\mathrm{cov}(\boldsymbol{\omega})_{12}$	–	-1.166e-2(-1.171e-2, -1.162e-2)	-6.873e-3(-6.879e-3, -6.867e-3)
$\mathrm{cov}(\boldsymbol{\omega})_{22}$	–	1.039e-2 (1.036e-2, 1.041e-2)	3.764e-3 (3.760e-3, 3.768e-3)
ϕ_1	–	26.414 ($\alpha = 0.99267$)	3.942 (3.857, 4.013)
ϕ_2	–	26.414($\alpha = 0.99267$)	12.358 (11.601, 13.162)
RMSPE[a]	[0.074, 0.0359, 0.0582]	[0.03172, 0.01743, 0.02559]	[0.0326, 0.0171, 0.0260]
CRPS[a]	[-0.0414, -0.01052, -0.02596]	[-0.01523, -0.00875, -0.01199]	[-0.01561, -0.00879, -0.0122]
CVG[a]	[0.9526, 0.9547, 0.9537]	[0.9515, 0.9427, 0.9471]	[0.954, 0.947, 0.950]
INT[a]	[0.34868, 0.17283, 0.26077]	[0.1909, 0.10172, 0.14631]	[0.1965, 0.09952, 0.14802]
time(mins)	–	[169.19, 65.52, 51.13][b]	2684.75

[a] [response 1, response 2, all responses]

[b] [time for cross-validation, time for MCMC sampling, time for recovering $\boldsymbol{\beta}$ and predictions]

feature of the proposed models is the introduction of conjugacy and conditional conjugacy through MNIW that facilitates the posterior sampling process. We devoted a significant part of this Chapter to formulate and illustrate our models in conjunction with NNGP. Meanwhile, we took an elaborate design in our simulation studies to test the performance of all models using datasets with different behaviors. We also showed our models are capable of conducting various analyses for massive multivariate spatial data through implementations on a real dataset with observed locations in millions.

Here, we emphasizes the modeling for purely spatial datasets, whereas the topics on spatio-temporal modeling and spatially-varying coefficient modeling are essential in Multivariate Geostatistics. We omitted these topics in the main body of the Chapter for the sake of conciseness, and we briefly discuss here as a supplement. For spatio-temporal modeling, it can

Table 4.4: Real data analysis summary table 2: posterior mean (2.5%, 97.5%)

response	intercept	slope	nugget ($\mathbf{\Sigma}_{ii}$)
NDVI	-0.176(-0.179, -0.172)	-0.0121 (-0.0125, -0.0116)	7.45e-4 (7.43e-4, 7.48e-4)
EVI	-0.076(-0.077, -0.074)	-4.4e-3(-4.7e-3, -4.1e-3)	8.68e-4(8.65e-4, 8.7e-4)
GPP	-6.939(-6.957, -6.919)	-0.197(-0.199, -0.194)	0.0244(0.0243, 0.0245)
PsnNet	-4.282 (-4.289, -4.275)	-4.5e-3(-5.5e-3, -3.6e-3)	5.34e-3(5.32e-3, 5.36e-3)
red refl	0.358 (0.356, 0.359)	4.5e-3 (4.2e-3, 4.8e-3)	9.84e-4(9.81e-4, 9.87e-4)
blue refl	0.186(0.185, 0.187)	0.0123 (0.0121, 0.0124)	2e-4(2.59e-4, 2.61e-4)
LE	4.601(4.586 , 4.616)	0.091(0.088, 0.093)	0.0531 (0.0529, 0.0533)
ET	1.154(1.139, 1.169)	0.092 (0.089, 0.094)	0.0531(0.053, 0.0533)
PLE	0.7126 (0.7118, 0.7132)	-3.6e-3 (-3.8e-3, -3.3e-3)	2.1e-5 (2.09e-5, 2.11e-5)
PET	2.255 (2.252, 2.257)	-4.6e-3(-5.5e-3, -3.8e-3)	6.4e-5 (6.3e-5, 6.4e-5)

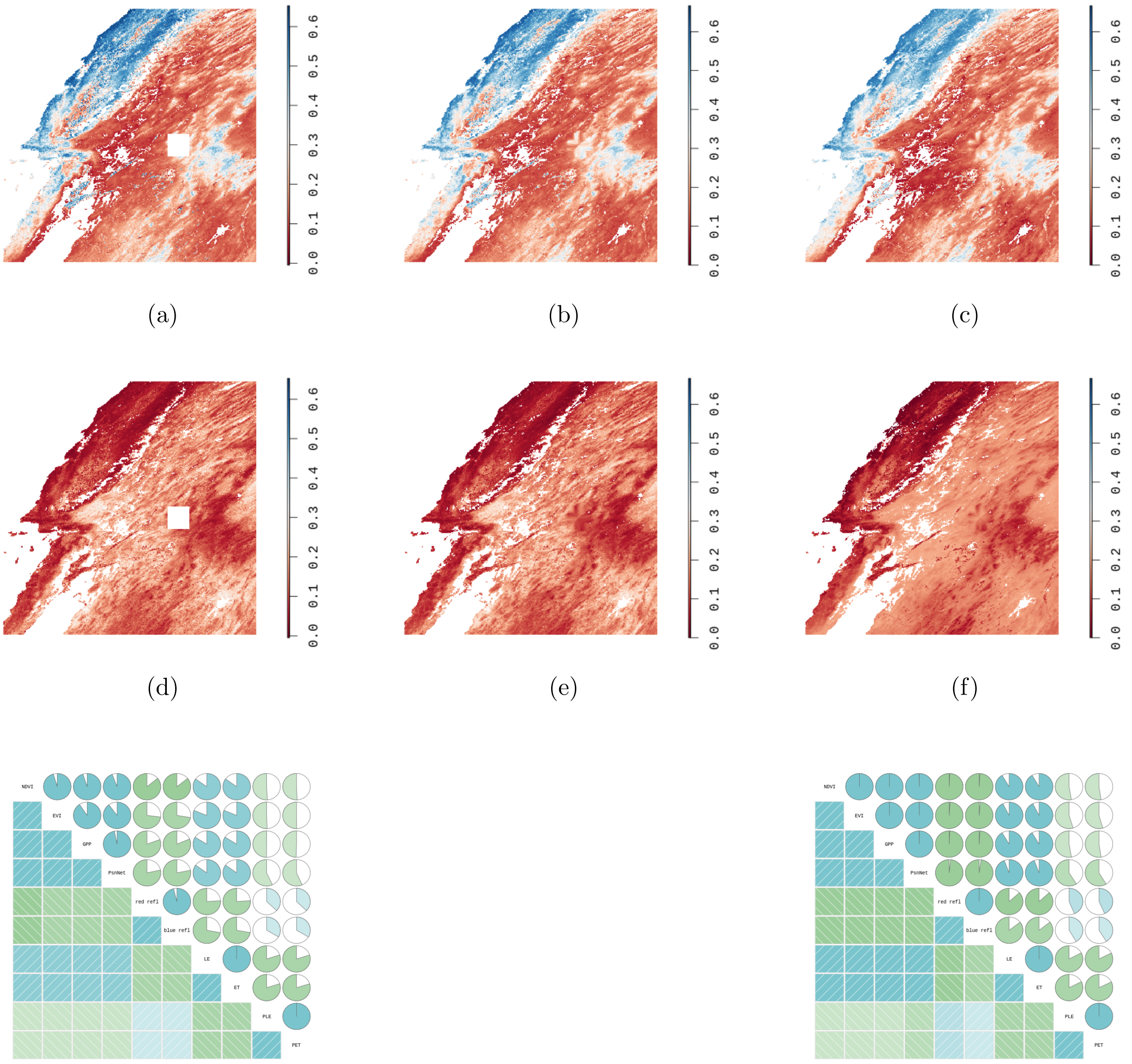
be seen that the models proposed in this Chapter can readily be modified for modeling multivariate spatio-temporal random fields. A modification is to use a dynamic nearest-neighbor Gaussian process (DNNGP) [Datta et al., 2016] instead of the NNGP in our models. When considering spatially-varying coefficient modeling, i.e., we model the regression coefficients $\boldsymbol{\beta}$ by a random field to allow the analysis of the spatial pattern of the regression coefficients. We can assign the prior of the regression coefficients $\boldsymbol{\beta}$ by a multivariate Gaussian random field with a proportional cross-covariance function. Then the prior of $\boldsymbol{\beta}$ over observed locations follows a Matrix-Normal distribution, which is the prior we designed for $\boldsymbol{\beta}$ in all proposed models in this Chapter. While the modification seems to be easy, the actual implementation requires a more detailed exploration, and we leave these topics for further studies.

The scalability of proposed algorithms is guaranteed when the utilized univariate scalable modeling method can yield a sparse precision matrix. Hence our approach can adapt to other methods such as multiresolution approximation (MRA), NNGP, spatial partitioning, etc. [see, e.g. Katzfuss and Guinness, 2017]. For brevity, we elaborated on the algorithms for NNGP based models in the supplementary material, but provided a limited discussion on alternative models. Further exploration with various univariate scalable modeling methods still require careful discussions. Also, scalable modeling methods that introduce sparsity in covariance matrics such as covariance tapering are beyond the scope of our discussion.

An important future research topic is to explore high-performance programming. The programs provided in this Chapter are for illustration and have limited usage in GPU computing and parallel CPU computing. Since the algorithms of Conjugate models are parallelizable, GPU parallel computing can dramatically reduce the run time. A parallel CPU computing for the BSLMC model can simultaneously sample multiple MCMC chains, improving the performance of the actual implementations. Implementations with modeling methods such as MRA [Katzfuss, 2017] also requires dedicated programming with GPU. Hence, efficient programming for developing related packages is demanded in the follow-up researches.

Supplementary Material

All computer programs implementing the examples in this Chapter can be found in the public domain and downloaded from `https://github.com/LuZhangstat/Multi_NNGP`.

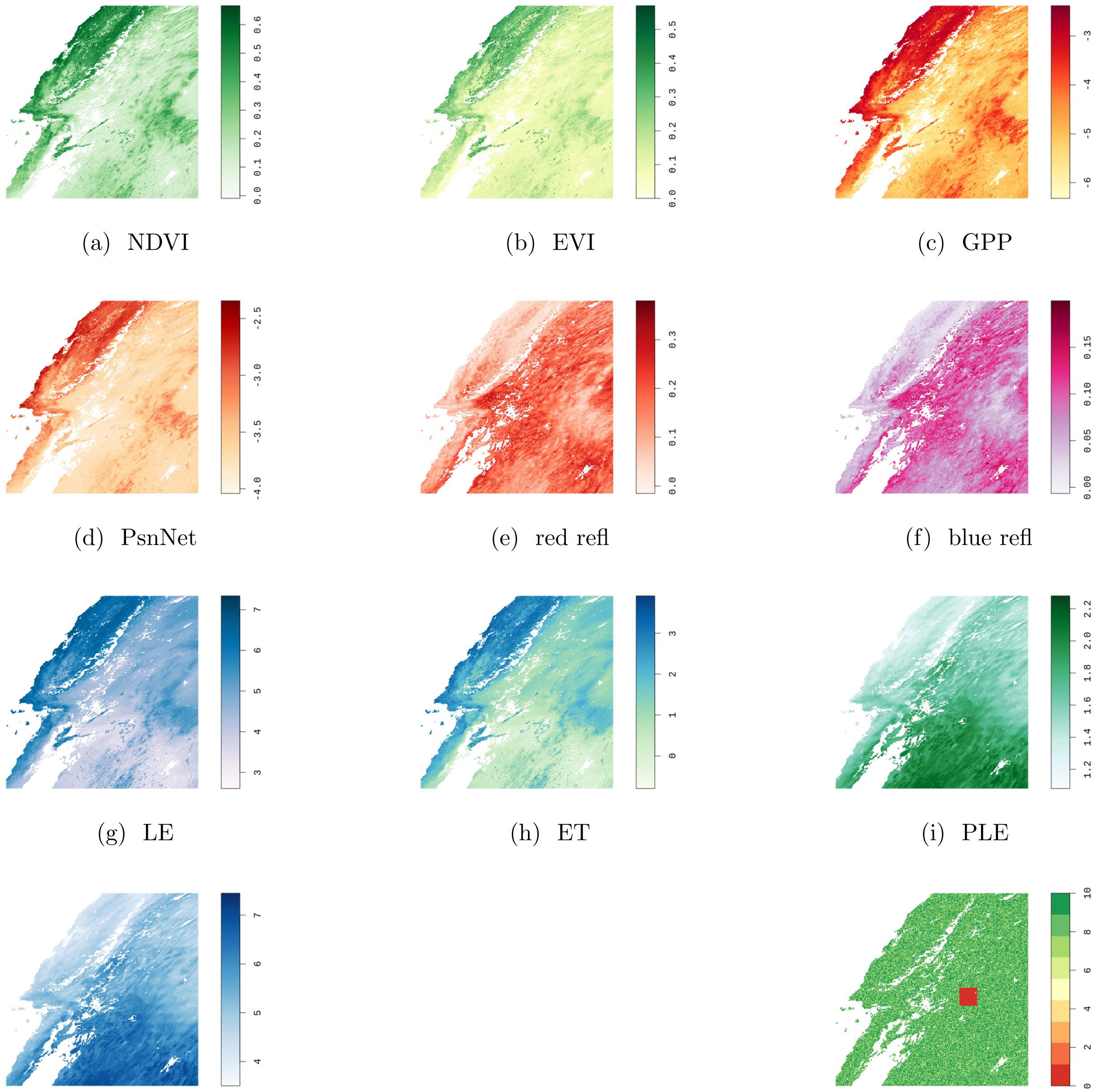


(a) (b) (c)

(d) (e) (f)

(g) (h)

Figure 4.3: Colored NDVI and red reflectance images of western United States (zone h08v05). Maps of raw data (a) & (d) and the posterior mean of the intercept-centered latent process recovered from (b) & (e) BSLMC and (c) & (f) factor BSLMC with diagonal $\mathbf{\Sigma}$. Correlation of responses (g) and nonspatial correlation of latent process fitted by the factor BSLMC model with diagonal $\mathbf{\Sigma}$ (h).



(a) NDVI

(b) EVI

(c) GPP

(d) PsnNet

(e) red refl

(f) blue refl

(g) LE

(h) ET

(i) PLE

(j) PET

(k) Misalignment

Figure 4.4: Maps (a)-(j) of predicted value on $1,020,000$ observed locations for 10 variables in Section 3.4. The deeper the color, the higher the value. Some variables are transformed for better model fitting. Each map has its own color scale. Heat-map (k) of counts of observed response, the greener the color, the higher the value.

CHAPTER 5

On identifiability and consistency of the nugget in Gaussian spatial process models

5.1 Introduction

In the process of obtaining posterior inferences for Bayesian spatial models as shown in Chapter 2 3 and 4, MCMC chains or posterior samples of some hyper-parameters are often observed to be unstable even with a large sample size. This phenomenon invokes the research interests in the theoretical studies on the asymptotic properties of the parameter estimators in spatial modeling. The analysis of point-referenced spatial data relies heavily on stationary Gaussian processes for modeling spatial dependence. Let $y(\mathbf{s})$ be the outcome measured at a location $\mathbf{s} \in S \subset \mathbb{R}^d$, where S is a bounded region within $\mathbb{R}^d$. The outcome is customarily modeled as

$$y(\mathbf{s}) = \mu(\mathbf{s}) + w(\mathbf{s}) + \epsilon(\mathbf{s}), \quad \mathbf{s} \in S \subset \mathbb{R}^d \;, \tag{5.1.1}$$

where $\mu(\mathbf{s})$ models the trend, $w(\mathbf{s})$ is a Gaussian process capturing spatial dependence, and $\epsilon(\mathbf{s})$ is a white noise process modeling measurement error or micro-scale variation. Matérn [1986] introduced a flexible class of covariance functions for modeling $w(\mathbf{s})$ that has been widely used in spatial modeling ever since it was recommended in Stein [1999]. The finite dimensional realizations of $\epsilon(\mathbf{s})$ are modeled independently and identically as $\mathrm{N}(0, \tau^2)$ over any finite collection of locations. The variance parameter τ^2 is called the “nugget”.

Our intended contribution in this Chapter is to formally establish the identifiability and consistency of the process parameters in (5.1.1) in the presence of an unknown nugget under infill or fixed domain asymptotics, where the sample size increases with increasing numbers of locations within a domain that is fixed and does not expand. This distinguishes the Chapter from existing results on inference for process parameters in Matérn models that have, almost exclusively, been studied without the presence of an unknown nugget. Zhang and Zimmerman [2005] compared infill and expanding domain asymptotic paradigms and elucidate a preference for the former for analyzing the limiting distributions of parameters in the Matérn family. Zhang [2004] showed that not all parameters in the Matérn family can be consistently estimated under infill asymptotics, but certain *microergodic* parameters, which play a crucial role in the identifiability of Gaussian processes with the Matérn covariogram (see Section 5.2.1 for further details), are consistently estimable. Du et al. [2009] derived the asymptotic normality of the maximum likelihood estimator for such microergodic parameters. Kaufman and Shaby [2013] extended these asymptotic results to the case of jointly estimating the spatial range and the variance parameters in the Matérn family, and explore the effect of a prefixed range verses a joint estimated range on inference when having relatively small sample size.

These studies have focused upon settings without the presence of a nugget. In practice, modeling the measurement error, or nugget effect, in (5.1.1) is prevalent in geostatistical modeling. Zhang and Zimmerman [2005] offered some heuristic arguments for the consistency and asymptotic normality of the maximum likelihood estimators of microergodic parameters in (5.1.1). Chen et al. [2000] demonstrated that the presence of measurement error can have a big impact on the parameter estimates for Ornstein-Uhlenbeck processes over bounded intervals. Their proof exploits the Markovian property as well as the explicit formula of the maximum likelihood estimator of the one-dimensional Ornstein-Uhlenbeck process that are not available in the case of Matérn model over $\mathbb{R}^d$ with $d \geq 2$.

Returning to (5.1.1), it will be sufficient for our subsequent development to assume that $\mu(\mathbf{s}) = 0$, i.e., the data have been de-trended. We specify $\{w(\mathbf{s}) : \mathbf{s} \in \mathcal{S} \subset \mathbb{R}^d\}$ as a zero-centered stationary Gaussian process with isotropic Matérn covariogram,

$$K_w(\mathbf{x}; \sigma^2, \phi, \nu) := \frac{\sigma^2(\phi\|\mathbf{x}\|)^\nu}{\Gamma(\nu)2^{\nu-1}} K_\nu(\phi\|\mathbf{x}\|), \quad \|\mathbf{x}\| \geq 0 , \tag{5.1.2}$$

where $\sigma^2 > 0$ is called the *partial sill* or spatial variance, $\phi > 0$ is the scale or decay parameter, $\nu > 0$ is a smoothness parameter, $\Gamma(\cdot)$ is the Gamma function, and $K_\nu(\cdot)$ is the modified Bessel function of order ν [Abramowitz and Stegun, 1965, Section 10]. The corresponding spectral density is

$$f(\mathbf{u}) = C\frac{\sigma^2\phi^{2\nu}}{(\phi^2+u^2)^{\nu+d/2}} \quad \text{for some } C>0. \tag{5.1.3}$$

When $\nu = 1/2$, the covariogram (5.1.2) simplifies to the exponential (Ornstein-Uhlenbeck in one-dimension) kernel $K_w(\mathbf{x};\,\sigma^2,\phi) := \sigma^2\exp(-\phi\|\mathbf{x}\|)$. For the measurement error, we assume $\{\epsilon(\mathbf{s}) : \mathbf{s}\in\mathcal{S}\subset\mathbb{R}^d\}$ is Gaussian white noise with covariogram $K_\epsilon(\mathbf{y};\,\tau^2) := \tau^2\delta_0$, where δ_0 is the Dirac mass at 0 and τ^2 is the nugget. The processes $\{w(\mathbf{s}),\,\mathbf{s}\in\mathcal{D}\subset\mathbb{R}^d\}$ and $\{\epsilon(\mathbf{s}),\,\mathbf{s}\in\mathcal{D}\subset\mathbb{R}^d\}$ are independent. Hence, a Matérn model with measurement error is a stationary Gaussian process with covariogram

$$K(\mathbf{x};\,\tau^2,\sigma^2,\phi,\nu) := K_w(\mathbf{x};\,\sigma^2,\phi,\nu) + K_\epsilon(\mathbf{x};\,\tau^2). \tag{5.1.4}$$

Our approach will depend upon identifying microergodic parameters in the above model. The remainder of the Chapter evolves as follows. We review the discussion in Zhang [2004] for the Matérn model with measurement error, claiming that only $\boldsymbol{\theta} = \{\sigma^2\phi^{2\nu},\tau^2\}$ can have infill consistent estimators when $d\le 3$. Subsequently, we establish that the maximum likelihood estimates for $\boldsymbol{\theta}$ are consistent and are asymptotically normal. This extends the main results in Chen et al. [2000] to the case with dimension $d \le 3$. The asymptotic properties of interpolation are explored mainly through simulations, and we demonstrate the role of $\boldsymbol{\theta}$ in interpolation. We conclude with some insights and directions for future work.

5.2 Asymptotic theory for estimation and prediction

5.2.1 Identifiability

Zhang [2004] showed that for the Matérn model without measurement error, when fixing the smoothness parameter $\nu > 0$ and $d \le 3$, there are no (weakly) infill consistent estimators for either the partial sill σ^2 or the scale parameter ϕ. Such results rely upon the equivalence and orthogonality of Gaussian measures. Two probability measures P_1 and P_2 on a measurable space $(\Omega, \mathcal{F})$ are said to be *equivalent*, denoted $P_1 \equiv P_2$, if they are absolutely continuous with respect to each other. Thus, $P_1 \equiv P_2$ implies that for all $A \in \mathcal{F}$, $P_1(A) = 0$ if and only if $P_2(A) = 0$. On the other hand, P_1 and P_2 are orthogonal, denoted $P_1 \perp P_2$, if there exists $A \in \mathcal{F}$ for which $P_1(A) = 1$ and $P_2(A) = 0$. While measures may be neither equivalent nor orthogonal, Gaussian measures are in general one or the other. For a Gaussian probability measure $P_{\boldsymbol{\theta}}$ indexed by a set of parameters $\boldsymbol{\theta}$, we say that $\boldsymbol{\theta}$ is *microergodic* if $P_{\boldsymbol{\theta}_1} \equiv P_{\boldsymbol{\theta}_2}$ if and only if $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$. For further background, see Chapter 6 in Stein [1999] and Zhang [2012]. Furthermore, two Gaussian probability measures defined by Matérn covariograms $K_w(\cdot;\, \sigma_1^2, \phi_1, \nu)$ and $K_w(\cdot;\, \sigma_2^2, \phi_2, \nu)$ are equivalent if and only if $\sigma_1^2\phi_1^{2\nu} = \sigma_2^2\phi_2^{2\nu}$ [Theorem 2 in Zhang, 2004] and, consequently, one cannot consistently estimate σ^2 or ϕ in the Matérn model (5.1.2) [Corollary 1 in Zhang, 2004].

We first characterize identifiability for the Matérn model with measurement error, i.e., with covariogram given by (5.1.4). Over a closed set $S \subset \mathbb{R}^d$, let $G_S(m, K)$ denote the Gaussian measure of the random field on S with mean function m and covariance function K. Consider two different specifications for $w(\mathbf{s})$ in (5.1.1) corresponding to mean m_i and covariogram K_i for $i = 1, 2$. The respective measures on the realizations of $w(\mathbf{s})$ over S will be denoted by $G_S(m_i, K_i)$ for $i = 1, 2$. If $\chi = \{\mathbf{s}_1, \mathbf{s}_2, \ldots\}$ is a sequence of points in S, then the probability measure for the sequence of outcomes over χ, i.e., $\{y(\mathbf{s}_j) : \mathbf{s}_j \in \chi\}$, is denoted $G_\chi(m_i, K_i, \tau_i^2)$ under model i. The following lemma is familiar.

Lemma 5.2.1. *Let S be a closed set, $w(\mathbf{s})$ be a mean square continuous process on S under $G_S(m_1, K_1)$, and χ be a dense sequence of points in S. Then, (i) if $\tau_1^2 \neq \tau_2^2$, then $G_\chi(m_1, K_1, \tau_1^2) \perp G_\chi(m_2, K_2, \tau_2^2)$; and (ii) if $\tau_1^2 = \tau_2^2$, then $G_\chi(m_1, K_1, \tau_1^2) \equiv G_\chi(m_2, K_2, \tau_2^2)$ if and only if $G_S(m_1, K_1) \equiv G_S(m_2, K_2)$.*

Proof. See Theorem 6 in Chapter 4 of Stein [1999]. □

We adapt Lemma 5.2.1 to the Matérn model with measurement error. The following result summarizes the identifiability issue for Matérn model with measurement error.

Theorem 5.2.2. *Let $S \subset \mathbb{R}^d$ be a compact set. For $i = 1, 2$, let P_i be the probability measure of the Gaussian process on S with mean zero and covariance $K(\cdot; \tau_i^2, \sigma_i^2, \phi_i, \nu)$ defined by* (5.1.4)*. Then, (i) if $\tau_1^2 \neq \tau_2^2$, then $P_1 \perp P_2$; and (ii) if $\tau_1^2 = \tau_2^2$, then for $d \leq 3$, $P_1 \equiv P_2$ if and only if $\sigma_1^2\phi_1^{2\nu} = \sigma_2^2\phi_2^{2\nu}$, and for $d \geq 5$, $P_1 \equiv P_2$ if and only if $(\sigma_1^2, \phi_1) = (\sigma_2^2, \phi_2)$.*

Proof. Denote K_i for $K_w(\cdot;\, \sigma_i^2, \phi_i, \nu)$. It is easy to see that $w(\mathbf{s})$ is mean square continuous on S under $G_S(0, K_i)$. From Lemma 5.2.1, we know that if $\tau_1^2 \neq \tau_2^2$, for any dense sequence χ, $G_\chi(0, K_1, \tau_1^2) \perp G_\chi(0, K_2, \tau_2^2)$. Therefore, $P_1 \perp P_2$. This proves (i).

Next, suppose $\tau_1^2 = \tau_2^2$. From Theorem 2 in Zhang [2004], we know that for $d \leq 3$ $G_S(0, K_1) \equiv G_S(0, K_2)$ if and only if $\sigma_1^2\phi_1^{2\nu} = \sigma_2^2\phi_2^{2\nu}$. Corollary 3 in Anderes [2010] shows that, for $d \geq 5$, $G_S(0, K_1) \perp G_S(0, K_2)$ if $\{\sigma_1^2, \phi_1\} \neq \{\sigma_2^2, \phi_2\}$. A straightforward application of Lemma 5.2.1 proves (ii). □

Combining Theorem 5.2.2 with the argument in Corollary 1 given in Zhang [2004], we obtain that σ^2 and ϕ are not consistently estimable in the following sense.

Corollary 5.2.3. *Let $y(\mathbf{s})$, $\mathbf{s} \in S \subset \mathbb{R}^d, d \leq 3$ be a Gaussian process with covariogram as in* (5.1.4)*, and S_n, $n \geq 1$ be an increasing sequence of subsets of S. Given observations of $y(\mathbf{s})$, $\mathbf{s} \in S_n$, there do not exist estimates $\widehat{\sigma}_n^2$ and $\widehat{\phi}_n$ that are consistent.*

Theorem 5.2.2 characterizes equivalence and orthogonality of Matérn based Gaussian measures according to the values of their parameters. The results in Zhang [2004] emerge as special cases when $\tau_1^2 = \tau_2^2 = 0$ and $\sigma_1^2\phi_1^{2\nu} = \sigma_2^2\phi_2^{2\nu}$ for $d \leq 3$. The characterization of equivalence and orthogonality of P_1 and P_2 remains open when $d = 4$. In the rest of this Chapter, we focus on the asymptotic properties of parameter estimates and predictions for the case with $d \leq 3$.

5.2.2 Parameter estimation

Theorem 5.2.2 implies that if ν is fixed in the specification of $w(\mathbf{s})$ in (5.1.1), then $\sigma^2\phi^{2\nu}$ and the nugget τ^2 will be identifiable. In view of this, we consider the estimation of the microergodic parameter $\kappa := \sigma^2\phi^{2\nu}$ and the nugget τ^2 with fixed decay ϕ. Our main results concern the consistency and the asymptotic normality of the maximum likelihood estimators of κ and τ^2 when the observations are taken from $y(\cdot)$ modeled by (5.1.1).

To proceed further, we need some notations. Let $\chi_n = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ be the sampled points in S, $y_i := y(\mathbf{s}_i)$, $i = 1, \ldots, n$ be the corresponding observations, and let $\mathbf{K}_n := \left\{K_w(\mathbf{s}_i - \mathbf{s}_j;\, \sigma^2, \phi, \nu)\right\}_{1\le i,j\le n}$ denote the $n \times n$ Matérn covariance matrix over locations χ_n. Let $\{\lambda_i^{(n)},\ i = 1, \ldots, n\}$ be the eigenvalues of $\mathbf{K}_n$ in decreasing order. The covariance matrix of the observations $\mathbf{y} = (y_1, \ldots, y_n)^\top$ is $\mathbf{V}_n = \tau^2\mathbf{I}_n + \mathbf{K}_n$, and the (rescaled) negative log-likelihood is

$$\ell(\tau^2, \sigma^2, \phi) := \log\det \mathbf{V}_n + \mathbf{y}^\top \mathbf{V}_n^{-1}\mathbf{y}\ . \tag{5.2.1}$$

Let $\{\sigma_0^2, \phi_0, \tau_0^2\}$ be the true generating values of $\{\sigma^2, \phi, \tau^2\}$, $\kappa_0 = \sigma_0^2\phi_0^{2\nu}$. Assume that the smoothness parameter $\nu > 0$ is known. For any fixed $\phi_1 > 0$, let $(\widehat{\tau}_n^2(\phi_1), \widehat{\sigma}_n^2(\phi_1))$ be the maximum likelihood estimators of $\mathcal{L}(\tau^2, \sigma^2, \phi_1)$. That is,

$$(\widehat{\tau}_n^2(\phi_1), \widehat{\sigma}_n^2(\phi_1)) : = \underset{(\tau^2,\sigma^2)\in D}{\arg\max}\, \mathcal{L}(\tau^2, \sigma^2, \phi_1) = \underset{(\tau^2,\sigma^2)\in D}{\operatorname{argmin}}\ \ell(\tau^2, \sigma^2, \phi_1) \tag{5.2.2}$$

where $D = [a, b] \times [c, d]$ with $0 < a < b < \infty$ and $0 < c < d < \infty$. To simplify notations, write $\widehat{\tau}_n^2$, $\widehat{\sigma}_n^2$ for $\widehat{\tau}_n^2(\phi_1)$, $\widehat{\sigma}_n^2(\phi_1)$. Unlike the Matérn model (5.1.2), there is no explicit formula for $\widehat{\tau}_n^2$ and $\widehat{\sigma}_n^2$ in the Matérn model with measurement error. Another difficulty of the analysis is that $\mathcal{L}$ is not concave, so the (rescaled) negative log-likelihood $\ell(\tau^2, \sigma^2, \phi_1)$ may have local minima and stationary points. Nevertheless, we are able to establish the theorems regarding the consistency and asymptotic normality at these stationary points under some assumptions of the eigenvalue asymptotics.

We first give an upper bound for the eigenvalues $\lambda_i^{(n)}$, which is of independent interest. The argument we provide below works for a large class of covariograms, including the Matérn model. In the sequel, the symbol $\asymp$ indicates asymptotically bounded from below and above.

We follow closely the presentation of Belkin [2018]. Let Ω be a domain of $\mathbb{R}^d$, and $K(\cdot)$ be a positive definite radial basis kernel on $\mathbb{R}^d$. Denote $\mathcal{H}$ to be the Reproducing Kernel Hilbert Space corresponding to the kernel K, which is also the native space associated to Kernel K. Given a probability measure μ on Ω, define the integral operator $\mathcal{K}_\mu : L^2_\mu \to L^2_\mu$ by

$$\mathcal{K}_\mu f(\mathbf{x}) := \int_\Omega K(\mathbf{x}-\mathbf{z})f(\mathbf{z})\mu(d\mathbf{z}).$$

In particular, if $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{\mathbf{s}_i}$, $\mathcal{K}_\mu$ corresponds to the kernel matrix $\{\frac{1}{n}K(\mathbf{s}_i - \mathbf{s}_j)\}_{1\le i,j\le n}$. It is well known that $\mathcal{K}_\mu f \in \mathcal{H}$ for $f \in L^2_\mu$ [see Belkin, 2018, Section 2], and any function in $\mathcal{H}$ induces a function in L^2_μ by restricting it to the support of μ. Call $\mathcal{R}_\mu : \mathcal{H} \to L^2_\mu$ the restriction operator. The key idea of Belkin [2018] is to get a measure-independent upper bound for the eigenvalues of $\mathcal{K}_\mu$ for infinitely smooth kernels. While the argument can carry over to kernels with limited smoothness; that is, the spectral density of K satisfies $f(u) \asymp u^{-\beta-d}$ (β-smooth). By (5.1.3), the Matérn covariogram is 2ν-smooth. Given $\chi = \{\mathbf{s}_1, \dots, \mathbf{s}_n\} \subset \Omega$, let $S_\chi : \mathcal{H} \to \mathcal{H}$ be the interpolation operator defined by

$$S_\chi f(\mathbf{x}) = \sum_{i=1}^n \alpha_i K(\mathbf{x}_i - \mathbf{x}),$$

where $(\alpha_1, \dots, \alpha_n)^\top = \mathbf{K}_n^{-1}(f(\mathbf{x}_1), \dots, f(\mathbf{x}_n))^\top$ with $\mathbf{K}_n = \{K(\mathbf{s}_i - \mathbf{s}_j)\}_{1\le i,j\le n}$. By letting $h = \max_{\mathbf{s}\in S} \min_{1\le i\le n} \|\mathbf{s} - \mathbf{s}_i\|$, Santin and Schaback [2016, p985] proved that there exists $C > 0$ (independent of n) such that

$$\|\mathcal{R}_\mu - S_\chi\|_{\mathcal{H}\to L^2_\mu} \le Ch^{(\beta+d)/2}. \tag{5.2.3}$$

Here $\|\cdot\|_{\mathcal{H}\to L^2_\mu}$ denotes the operator norm. So (5.2.3) is a limited smoothness version of Belkin [2018, Theorem A p5]. The following result is adapted from Theorem 1 in Belkin [2018] to the β-smooth kernel.

Theorem 5.2.4. *Suppose $\mathcal{T} : V \to \mathcal{H}$ is a map from a Banach space V to a Reproducing Kernel Hilbert Space of functions on $\mathbb{R}^d$, $\mathcal{H}$ corresponding to a β-smooth radial basis kernel. Then there exists a map $\mathcal{T}_n$ from V to an n-dimensional linear subspace $\mathcal{H}_n \subset \mathcal{H}$, such that*

$$\|\mathcal{T} - \mathcal{T}_n\|_{V\to L^2_\mu} \le C\|\mathcal{T}\|_{V\to\mathcal{H}} n^{-\frac{\beta+d}{d}}$$

for $C > 0$ indepedent of $\mathcal{T}$ and μ. Moreover, (1) the subspace $\mathcal{H}_n$ is independent of $\mathcal{T}$; (2) if $\mathcal{T}$ is linear operator, $\mathcal{T}_n$ is also a linear operator.

Proof. The proof is exactly as in Theorem 1 in Belkin [2018]. □

The following Theorem is adapted from Theorem 2 in Belkin [2018] for β-smooth kernels.

Theorem 5.2.5. *Let K be a β-smooth radial basis kernel, and $\lambda_i(\mathcal{K}_\mu)$ be the i^{th} largest eigenvalue of $\mathcal{K}_\mu$. Then there exists $C > 0$ such that*

$$\lambda_i(\mathcal{K}_\mu) \leq C i^{-\frac{\beta+d}{d}}.$$

Proof. The proof follows by combining Theorem 5.2.4 above with Lemma 1 in Belkin [2018]. □

Corollary 5.2.6. *Assume that $\max_{\mathbf{s}\in S} \min_{1\leq i\leq n} \|\mathbf{s}-\mathbf{s}_i\| \asymp n^{-1/d}$. There exists $C > 0$ such that*

$$\lambda_i^{(n)} \leq Cni^{-2\nu/d-1} \quad \text{for all } i = 1, \ldots, n. \tag{5.2.4}$$

Proof. This follows by applying Theorem 5.2.5 with $\mu = \frac{1}{n}\sum_{i=1}^n \delta_{s_i}$ and $\beta = 2\nu$. □

It is natural to expect a matching lower bound for the eigenvalues $\lambda_i^{(n)}$ under a suitable condition on the sampled point locations.

Assumption 5.2.7. *Assume that $\min_{1\leq i\neq j\leq n} \|\mathbf{s}_i-\mathbf{s}_j\| \asymp n^{-1/d}$. There exists $c > 0$ such that*

$$\lambda_i^{(n)} \geq cni^{-2\nu/d-1} \quad \text{for all } i = 1, \ldots, n. \tag{5.2.5}$$

The lower bound (5.2.5) holds for the largest eigenvalues. A lesser known result of Schaback [1995] shows that (5.2.5) also holds for $i \asymp n$. Particularly interesting cases are $i \asymp n^\alpha$ for $0 < \alpha < 1$, which leave the lower bound (5.2.5) open. Furthermore, for the regular grid $\chi_n = [0,1)^d \cap n^{-1/d}\mathbb{Z}^d$, the scaled covariance matrix $\frac{1}{n}\mathbf{K}_n$ is viewed as the discretization of the integral operator

$$\mathcal{K}f(\mathbf{x}) := \int_{[0,1]^d} K_w(\mathbf{s}-\mathbf{t};\sigma^2,\phi,\nu) f(\mathbf{t}) d\mathbf{t},$$

where f is a test function. The integral operator $\mathcal{K}$ has eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots > 0$. Intuitively, $\lambda_i^{(n)}/n \approx \lambda_i$ which is at least true for fixed i. Santin and Schaback [2016] observed that $\lambda_i = d_{i-1}^2$, with d_i the i-width of the unit Sobolev ball in the L^2 space. Using a differential operator approach, Jerome [1972] showed that $\lim_{i\to\infty} i^{\frac{2\nu+d}{2d}} d_i = C'$. The above two results imply that $\lim_{i\to\infty} i^{2\nu/d+1}\lambda_i = C'^2$. Though the error $|\lambda_i^{(n)}/n - \lambda_i|$ is not easy to estimate, the following asymptotics is expected.

Assumption 5.2.8. *Let $\chi_n = [0,1)^d \cap n^{-1/d}\mathbb{Z}^d$. There exists $A > 0$ such that*

$$\lambda_i^{(n)}/(ni^{-2\nu/d-1}) \to A \quad \text{as } n, i \to \infty \tag{5.2.6}$$

To proceed further, we need the following lemma which is proved by elementary calculus.

Lemma 5.2.9. *Assume that $\max_{\mathbf{s}\in S}\min_{1\leq i\leq n}\|\mathbf{s}-\mathbf{s}_i\| \asymp n^{-1/d}$ and $\min_{1\leq i\neq j\leq n}\|\mathbf{s}_i-\mathbf{s}_j\| \asymp n^{-1/d}$. Let $a_{ni} = 1/(\widehat{\tau}_n^2 + \widehat{\sigma}_n^2\lambda_i^{(n)})$, $b_{ni} = \lambda_i^{(n)}/(\tau_0^2 + \widehat{\sigma}_n^2\lambda_i^{(n)})$, $a_{ni}^0 = 1/(\tau_0^2 + \sigma^2\lambda_i^{(n)})$, and $b_{ni}^0 = \lambda_i^{(n)}a_{ni}^0$.*

(1) There exists $C > 0$ such that

$$\sum_{i=1}^n a_{ni}^2 \asymp n \ , \ \sum_{i=1}^n \lambda_i^{(n)}a_{ni}^2 \leq Cn^{\frac{1}{2\nu/d+1}} \ , \ \sum_{i=1}^n b_{ni} \leq Cn^{\frac{1}{2\nu/d+1}} \ , \ \sum_{i=1}^n b_{ni}^2 \leq Cn^{\frac{1}{2\nu/d+1}} \ .$$

(2) Under Assumption 5.2.7,

$$\sum_{i=1}^n a_{ni}^2 \asymp n \ , \ \sum_{i=1}^n \lambda_i^{(n)}a_{ni}^2 \asymp n^{\frac{1}{2\nu/d+1}} \ , \ \sum_{i=1}^n b_{ni} \asymp n^{\frac{1}{2\nu/d+1}} \ , \ \sum_{i=1}^n b_{ni}^2 \asymp n^{\frac{1}{2\nu/d+1}} \ .$$

(3) Under Assumption 5.2.8, there exist $c_1(\sigma), c_2(\sigma), c_3(\sigma) > 0$ such that as $n \to \infty$,

$$\frac{1}{n}\sum_{i=1}^n (a_{ni}^0)^2 \to c_1(\sigma), \quad \frac{1}{n}\sum_{i=1}^n (a_{ni}^0)^4 \to c_2(\sigma), \quad \frac{1}{n^{1/(1+2\nu/d)}}\sum_{i=1}^n (b_{ni}^0)^2 \to c_3(\sigma).$$

Theorem 5.2.10. *Assume that $(\tau_0^2, \sigma_0^2) \in D$, $\chi_n := \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ satisfy*

$$\max_{\mathbf{s}\in S}\min_{1\leq i\leq n}\|\mathbf{s}-\mathbf{s}_i\| \asymp n^{-1/d} \quad \text{and} \quad \min_{1\leq i\neq j\leq n}\|\mathbf{s}_i-\mathbf{s}_j\| \asymp n^{-1/d} \ ,$$

and the conditions in Assumption 5.2.7 hold. Then $\widehat{\tau}_n^2 \to \tau_0^2$ almost surely and $\widehat{\sigma}_n^2\phi_1^{2\nu} \to \kappa_0$ almost surely under the Matérn model with covariogram $K(\cdot;\tau_0^2,\sigma_0^2,\phi_0,\nu)$.

Proof. Let P_0 be the probability measure for y corresponding to the Matérn covariogram $K(\cdot;\, \tau_0^2, \sigma_0^2, \phi_0, \nu)$, and P_1 be that for $K(\cdot;\, \tau_0^2, \sigma_1^2, \phi_1, \nu)$ where $\sigma_1^2 := \kappa_0/\phi_1^{2\nu}$. We first prove that $\widehat{\tau}_n^2 \to \tau_0^2$ almost surely under P_0. By Theorem 5.2.2, $P_0 \equiv P_1$. Thus, it suffices to prove that $\widehat{\tau}_n^2 \to \tau_0^2$ almost surely under P_1. Under P_1, we can rewrite (5.2.1) as

$$\ell(\tau^2, \widehat{\sigma}_n^2, \phi_1) = \sum_{i=1}^n \frac{\tau_0^2 + \sigma_1^2\lambda_i^{(n)}}{\tau^2 + \widehat{\sigma}_n^2\lambda_i^{(n)}} W_i^2 + \sum_{i=1}^n \log(\tau^2 + \widehat{\sigma}_n^2\lambda_i^{(n)}), \tag{5.2.7}$$

where $W_i \overset{iid}{\sim} \mathcal{N}(0,1)$. The maximum likelihood estimator $\widehat{\tau}_n^2$ of τ^2 satisfies

$$(\tau_0^2 - \widehat{\tau}_n^2) \cdot \sum_{i=1}^n W_i^2 a_{ni}^2 = \sum_{i=1}^n \widehat{\tau}_n^2(1 - W_i^2)a_{ni}^2 + \sum_{i=1}^n (\widehat{\sigma}_n^2 - \sigma_1^2 W_i^2)\lambda_i^{(n)} a_{ni}^2. \tag{5.2.8}$$

where $a_{ni} = 1/(\widehat{\tau}_n^2 + \widehat{\sigma}_n^2\lambda_i^{(n)})$. By Lemma 5.2.9 (1), we have $\sum_{i=1}^n a_{ni}^2 \asymp n$ and $\sum_{i=1}^n \lambda_i^{(n)} a_{ni}^2 \le Cn^{1/(2\nu/d+1)}$ for some $C > 0$. Using Etemadi [2006, Theorem 1], we obtain

$$\frac{\sum_{i=1}^n W_i^2 a_{ni}^2}{\sum_{i=1}^n a_{ni}^2} \to 1, \quad \frac{\sum_{i=1}^n \widehat{\tau}_n^2(1 - W_i^2)a_{ni}^2}{\sum_{i=1}^n a_{ni}^2} \to 0 \quad \text{and} \quad \frac{\sum_{i=1}^n (\widehat{\sigma}_n^2 - \sigma_1^2 W_i^2)\lambda_i^{(n)} a_{ni}^2}{\sum_{i=1}^n a_{ni}^2} \to 0.$$

Combining the above with (5.2.8), we have $\widehat{\tau}_n^2 \to \tau_0^2$ almost surely under P_1.

Next, we prove that $\widehat{\sigma}_n^2\phi_1^{2\nu} \to \kappa_0$ almost surely under P_0. Since $\widehat{\tau}_n^2 \to \tau_0^2$ almost surely under P_0 and $\sigma_1^2 = \kappa_0/\phi_1^{2\nu}$, it suffices to prove that $\widehat{\sigma}_n'^2 := \operatorname{argmin}_{\sigma^2\in[c,d]} \ell(\tau_0^2, \sigma^2, \phi_1)$ converges almost surely to σ_1^2 under P_0. Again, since $P_0 \equiv P_1$, it suffices to prove that $\widehat{\sigma}_n'^2 \to \sigma_1^2$ almost surely under P_1. Under P_1,

$$\ell(\tau_0^2, \sigma^2, \phi_1) = \sum_{i=1}^n \frac{\tau_0^2 + \sigma_1^2\lambda_i^{(n)}}{\tau_0^2 + \sigma^2\lambda_i^{(n)}} W_i^2 + \sum_{i=1}^n \log(\tau_0^2 + \sigma^2\lambda_i^{(n)}). \tag{5.2.9}$$

Taking the derivative of (5.2.9) with respect to σ^2 and equating to zero, we obtain

$$\sum_{i=1}^n b_{ni}(W_i^2 - 1) = (\widehat{\sigma}_n'^2 - \sigma_1^2)\sum_{i=1}^n b_{ni}^2 W_i^2. \tag{5.2.10}$$

with $b_{ni} = \lambda_i^{(n)}/(\tau_0^2 + \widehat{\sigma}_n'^2\lambda_i^{(n)})$. It suffices to prove that $\sum_{i=1}^n b_{ni}(W_i^2 - 1)/\sum_{i=1}^n b_{ni}^2 W_i^2$ converges almost surely to 0. Since

$$\frac{\sum_{i=1}^n b_{ni}(W_i^2 - 1)}{\sum_{i=1}^n b_{ni}^2 W_i^2} = \frac{\sum_{i=1}^n b_{ni}(W_i^2 - 1)}{\sum_{i=1}^n b_{ni}} \cdot \frac{\sum_{i=1}^n b_{ni}}{\sum_{i=1}^n b_{ni}^2} \cdot \frac{\sum_{i=1}^n b_{ni}^2}{\sum_{i=1}^n b_{ni}^2 W_i^2},$$

and $\sum_{i=1}^n b_{ni} \asymp n^{1/(2\nu/d+1)}$, $\sum_{i=1}^n b_{ni}^2 \asymp n^{1/(2\nu/d+1)}$ by Lemma 5.2.9 (2), we get

$$\frac{\sum_{i=1}^n b_{ni}(W_i^2-1)}{\sum_{i=1}^n b_{ni}} \longrightarrow 0 \quad \text{and} \quad \frac{\sum_{i=1}^n b_{ni}^2}{\sum_{i=1}^n b_{ni}^2 W_i^2} \longrightarrow 1 \quad a.s.$$

Combining the above estimates with (5.2.10), we have $\widehat{\sigma}_n'^2 \to \sigma_1^2$ almost surely under P_1. □

Remark 5.2.11. *From the proof of the consistency of $\widehat{\tau}_n^2$ provided above, (1) $\widehat{\tau}_n^2 \to \tau_0^2$ is true without Assumption 5.2.7 on the lower bound for eigenvalues; (2) $\widehat{\tau}_n^2$ remains consistent even when σ^2 and ϕ are misspecified.*

It is difficult to establish the consistency of the joint maximum likelihood estimates of $\{\kappa, \tau^2, \phi\}$ (i.e., ϕ is not fixed). A related result can be found in Theorem 2 of Kaufman and Shaby [2013] without a nugget effect. In the presence of a nugget effect, constructing such a proof becomes difficult due to the analytic intractability of the maximum likelihood estimators for $\{\kappa, \tau^2, \phi\}$. Nevertheless, our simulation studies in Section 5.3.3 seem to support consistent estimation of $\{\kappa, \tau^2\}$ even when ϕ is not fixed.

Given the consistency of the maximum likelihood estimators, we turn to their asymptotic distributions. For simplicity of presentation, we let $S = [0,1]^d$ in the following theorem. The asymptotic normality described below holds for any compact set $S \subset \mathbb{R}^d$.

Theorem 5.2.12. *Assume that n is the d^{th} power of some positive integer, $\chi_n = [0,1)^d \cap n^{-1/d}\mathbb{Z}^d$, and the conditions in Assumption 5.2.8 hold. Let*

$$a_{ni}^0 := 1/(\tau_0^2 + \sigma_1^2 \lambda_i^{(n)}) \quad \text{and} \quad b_{ni}^0 := \lambda_i^{(n)} a_{ni}^0 \quad \text{for } 1 \le i \le n.$$

There exist constants $c_1, c_2, c_3 > 0$ such that as $n \to \infty$,

$$\frac{1}{n}\sum_{i=1}^n (a_{ni}^0)^2 \to c_1, \quad \frac{1}{n}\sum_{i=1}^n (a_{ni}^0)^4 \to c_2, \quad \frac{1}{n^{1/(1+2\nu/d)}}\sum_{i=1}^n (b_{ni}^0)^2 \to c_3. \tag{5.2.11}$$

We have

$$\sqrt{n}(\widehat{\tau}_n^2 - \tau_0^2) \xrightarrow{(d)} \mathcal{N}(0, 2\tau_0^4 c_2/c_1^2), \tag{5.2.12}$$

and

$$n^{1/(2+4\nu/d)}(\widehat{\sigma}_n^2 \phi_1^{2\nu} - \kappa_0) \xrightarrow{(d)} \mathcal{N}(0, 2\phi_1^{4\nu}/c_3), \tag{5.2.13}$$

Proof. With Assumption 5.2.8, the limits (5.2.11) follows from Lemma 5.2.9 (3). By (5.2.8) and Theorem 5.2.10, we have

$$\sqrt{n}(\tau_0^2 \quad - \quad \widehat{\tau}_n^2) = \\ (\quad 1 + o(1)) \, \frac{\tau_0^2 \sqrt{n} \sum_{i=1}^n (1 - W_i^2)(a_{ni}^0)^2 + \sigma_1^2 \sqrt{n} \sum_{i=1}^n (1 - W_i^2) \lambda_i^n (a_{ni}^0)^2}{\sum_{i=1}^n W_i^2 (a_{ni}^0)^2} \tag{5.2.14}$$

We know that $\sum_{i=1}^n W_i^2 (a_{ni}^0)^2 / \sum_{i=1}^n (a_{ni}^0)^2 \longrightarrow 1$. In addition,

$$\begin{aligned} \frac{\tau_0^2 \sqrt{n} \sum_{i=1}^n (1 - W_i^2)(a_{ni}^0)^2}{\sum_{i=1}^n (a_{ni}^0)^2} &= \frac{\sum_{i=1}^n (1 - W_i^2)(a_{ni}^0)^2}{\sqrt{2 \sum_{i=1}^n (a_{ni}^0)^4}} \cdot \frac{\tau_0^2 \sqrt{2n \sum_{i=1}^n (a_{ni}^0)^4}}{\sum_{i=1}^n (a_{ni}^0)^2} \\ &\xrightarrow{(d)} \mathcal{N}(0, 2\tau_0^4 c_2 / c_1^2), \end{aligned} \tag{5.2.15}$$

where the first term on the right hand side converges to $\mathcal{N}(0,1)$ by Lindeberg's central limit theorem, and the second term converges to $\tau_0^2 \sqrt{2c_2}/c_1$. Similarly,

$$\frac{\sigma_1^2 \sqrt{n} \sum_{i=1}^n (1 - W_i^2) \lambda_i^n (a_{ni}^0)^2}{\sum_{i=1}^n (a_{ni}^0)^2} = \frac{\sum_{i=1}^n (1 - W_i^2) \lambda_i^n (a_{ni}^0)^2}{\sqrt{2 \sum_{i=1}^n (\lambda_i^n)^2 (a_{ni}^0)^4}} \cdot \frac{\sigma_1^2 \sqrt{2n \sum_{i=1}^n (\lambda_i^n)^2 (a_{ni}^0)^4}}{\sum_{i=1}^n (a_{ni}^0)^2} \longrightarrow 0, \tag{5.2.16}$$

where the first term on the right hand side converges to $\mathcal{N}(0,1)$, and the second term converges to 0 since $\sum_{i=1}^n (\lambda_i^n)^2 (a_{ni}^0)^4 \asymp n^{1/(1+2\nu/d)}$. Combining (5.2.14), (5.2.15) and (5.2.16) leads to (5.2.12).

By (5.2.10) and Theorem 5.2.10, we get

$$n^{1/(2+4\nu/d)} (\widehat{\sigma}_n^2 - \sigma_1^2) = (1 + o(1)) \, \frac{n^{1/(2+4\nu/d)} \sum_{i=1}^n b_{ni}^0 (W_i^2 - 1)}{\sum_{i=1}^n (b_{ni}^0)^2 W_i^2}. \tag{5.2.17}$$

Moreover,

$$\begin{aligned} \frac{n^{1/(2+4\nu/d)} \sum_{i=1}^n b_{ni}^0 (W_i^2 - 1)}{\sum_{i=1}^n (b_{ni}^0)^2 W_i^2} &= \frac{\sum_{i=1}^n b_{ni}^0 (W_i^2 - 1)}{\sqrt{2 \sum_{i=1}^n (b_{ni}^0)^2}} \cdot \frac{\sqrt{2} n^{1/(2+4\nu/d)}}{\sqrt{\sum_{i=1}^n (b_{ni}^0)^2}} \cdot \frac{\sum_{i=1}^n (b_{ni}^0)^2}{\sum_{i=1}^n (b_{ni}^0)^2 W_i^2} \\ &\xrightarrow{(d)} \mathcal{N}(0, 2/c_3), \end{aligned} \tag{5.2.18}$$

where the first term on the right hand side converges to $\mathcal{N}(0,1)$, the second term converges to $\sqrt{2/c_3}$, and the third term converges to 1. Combining (5.2.17) and (5.2.18) yields (5.2.13).

□

Du et al. [2009] showed that for the Matérn model without measurement error, the maximum likelihood estimator $\widehat{\sigma}_n^2$ converges to σ_1^2 at a $\sqrt{n}$-rate. Theorem 5.2.12 shows that in the presence of measurement error, the maximum likelihood estimator $\widehat{\tau}_n^2$ has a $\sqrt{n}$-rate while $\widehat{\sigma}_n^2$ has a slower $n^{1/(2+4\nu/d)}$-rate. By taking $\nu = \frac{1}{2}$ and $d = 1$, we recover the results of Ying [1991], Chen et al. [2000] for the Ornstein-Uhlenbeck process, where the maximum likelihood estimator $\widehat{\sigma}_n^2$ converges at a $\sqrt{n}$-rate without measurement error, but at a $\sqrt[4]{n}$-rate in the presence of measurement error.

5.2.3 Interpolation at new locations

We now turn to predicting the value of the process at unobserved locations. Without the nugget (i.e., $\tau = 0$ in (5.1.1)), Stein [1988, 1993, 1999] establish that predictions under different measures tend to agree as sample size $n \to \infty$. However, in the presence of a nugget effect, the predictive variance of $y(s)$ at an unobserved location may not decrease to zero with increasing sample size. In fact, the squared prediction error for any linear predictor is expected to be at least τ^2. For example, let $\widehat{y}_0 = \mathbf{v}^\top \mathbf{y}$ be a linear predictor of $y_0 = y(\mathbf{s}_0)$ at the unobserved location $\mathbf{s}_0, \mathbf{s}_0 \notin \chi_n$. Let $\mathbf{w} = \{w(\mathbf{s}_1), \ldots, w(\mathbf{s}_n)\}$, $\boldsymbol{\epsilon} = \{\epsilon(\mathbf{s}_1), \ldots, \epsilon(\mathbf{s}_n)\}$, $w_0 = w(\mathbf{s}_0)$ and $\epsilon_0 = \epsilon(\mathbf{s}_0)$. The expected squared prediction error satisfies

$$\mathbb{E}[(\widehat{y}_0 - y_0)^2] = \mathbb{E}[\{(\mathbf{v}^\top \mathbf{w} - w_0) + (\mathbf{v}^\top \boldsymbol{\epsilon} - \epsilon_0)\}^2] = \mathbb{E}[(\mathbf{v}^\top \mathbf{w} - w_0)^2] + \mathbb{E}[(\mathbf{v}^\top \boldsymbol{\epsilon} - \epsilon_0)^2] \geq \tau^2.$$

To see whether there can be a consistent linear (unbiased) estimate of the underlying process $w(\cdot)$ at unobserved locations, consider the universal kriging estimator at an unobserved location $\mathbf{s}_0$ given by

$$\widehat{Z}_n(\tau^2, \sigma^2, \phi) := \boldsymbol{\gamma}_n(\sigma^2, \phi)^\top \boldsymbol{\Gamma}_n(\tau^2, \sigma^2, \phi)^{-1} \mathbf{y} \ , \tag{5.2.19}$$

where $\{\boldsymbol{\gamma}_n(\sigma^2, \phi)\}_i := K_w(\mathbf{s}_0 - \mathbf{s}_i; \, \sigma^2, \phi, \nu)$, and $\{\boldsymbol{\Gamma}_n(\tau^2, \sigma^2, \phi)\}_{ij} := K_w(\mathbf{s}_i - \mathbf{s}_j; \, \sigma^2, \phi, \nu) + \tau^2 \delta_0(i - j)$ for $i, j = 1, \ldots, n$. The interpolant $\widehat{Z}_n(\tau^2, \sigma^2, \phi)$ provides a best linear unbiased estimate of w_0 under the Matérn model with measurement error (5.1.4). By letting $\{\mathbf{K}_n(\phi)\}_{ij} := K_w(\mathbf{s}_0 - \mathbf{s}_i; \, 1, \phi, \nu)$, we have the mean squared error of the estimator (5.2.19) follows

$$\begin{aligned}\mathrm{Var}_{\tau_0^2,\sigma_0^2,\phi_0}\{\widehat{Z}_n(\tau^2,\sigma^2,\phi)-w_0\} &= \sigma_0^2\{1-2\boldsymbol{\gamma}_n(\sigma^2,\phi)^\top\boldsymbol{\Gamma}_n(\tau^2,\sigma^2,\phi)^{-1}\boldsymbol{\gamma}_n(\sigma_0^2,\phi_0)\\ &+\boldsymbol{\gamma}_n(\sigma^2,\phi)^\top\boldsymbol{\Gamma}_n(\tau^2,\sigma^2,\phi)^{-1}\mathbf{K}_n(\phi_0)\boldsymbol{\Gamma}_n(\tau^2,\sigma^2,\phi)^{-1}\boldsymbol{\gamma}_n(\sigma^2,\phi)\} \quad (5.2.20)\\ &+\tau_0^2\boldsymbol{\gamma}_n(\sigma^2,\phi)^\top\boldsymbol{\Gamma}_n(\tau^2,\sigma^2,\phi)^{-2}\boldsymbol{\gamma}_n(\sigma^2,\phi)\ ,\end{aligned}$$

where $\{\tau_0^2,\sigma_0^2,\phi_0\}$ are the true generating values of $\{\sigma^2,\phi,\tau^2\}$. Setting $(\tau^2,\sigma^2,\phi)=(\tau_0^2,\sigma_0^2,\phi_0)$ in (5.2.20) yields

$$\begin{aligned}&\mathrm{Var}_{\tau_0^2,\sigma_0^2,\phi_0}\{\widehat{Z}_n(\tau_0^2,\sigma_0^2,\phi_0)-w_0\}=\\ &\qquad\sigma_0^2\{1-\boldsymbol{\gamma}_n(\sigma_0^2,\phi_0)^\top\boldsymbol{\Gamma}_n(\tau_0^2,\sigma_0^2,\phi_0)^{-1}\boldsymbol{\gamma}_n(\sigma_0^2,\phi_0)\} \qquad (5.2.21)\end{aligned}$$

Theorem 8 in Chapter 3 of Stein [1999] characterizes the mean squared error of the best linear unbiased estimate at location 0 as $\dfrac{(2\pi c)^{1/\alpha}}{\alpha\sin(\pi/\alpha)}\left(\delta\tau^2\right)^{1-1/\alpha}$ with observations at δj for $j\neq 0$. Here $\alpha := 2\nu+1$ and $c := C\sigma^2\phi^{2\nu}$ with C defined in (5.1.3). Following the same argument, it is not hard to see that the mean squared error of the best linear unbiased estimate (based on data in $\mathbb{R}^d$) is of order $\delta^{2\nu/(2\nu+d)}$. Stein [1999] proved this for observations on the whole line (with a typo in the expression (44) of Stein [1999]). He also conjectured that the above expression for the mean-square error holds for data on any finite interval. We conduct simulations in Section 5.3.4 with the nugget effect to corroborate this.

5.3 Simulations

5.3.1 Set-up

The preceding results help explain the behavior of the inference from (5.1.1) as the sample size increases within a fixed domain. Here, we present some simulation experiments to illustrate statistical inference for finite samples. We simulate data sets based on (5.1.1) in a unit square setting $\nu=1/2$ and $\sigma^2=1$. We pick three different values of the nugget, $\tau^2\in\{0,0.2,0.8\}$, and choose the decay parameter ϕ so that the effective spatial range is 0.15, 0.4 or 1, i.e., the correlation decays to 0.05 at a distance of 0.15, 0.4 or 1 units. Therefore, we consider $3\times3=9$

different parameter settings. For each parameter setting, we simulate 1000 realizations of the Gaussian process over $n = 1600$ observed locations. The observed locations are chosen from a perturbed grid. We construct a 67×67 regular grid with coordinates from 0.005 to 0.995 in increments of 0.015 in each dimension. We add a uniform $[-0.005, 0.005]^2$ perturbation to each grid point to ensure at least 0.005 units separation from its nearest neighbor. We then choose $n = 1600$ locations out of the perturbed grid. Codes for studies in this Section are available on `https://github.com/LuZhangstat/nugget_consistency`.

5.3.2 Likelihood comparisons

Theorem 5.2.2 suggests that it is difficult to distinguish between the two Matérn models with measurement error when their microergodic parameters $\{\kappa, \tau^2\}$ are close to each other. This property should be reflected in the behavior of the likelihood function for a large finite sample. To see this, we plot interpolated maps of the log-likelihood among different grids of parameter values. We consider the three values of τ_0^2 in Section 5.3.1 and $\phi_0 = 7.49$, which implies an effective spatial range of approximately 0.4 units, and pick $n = 900$ observations from the first realization generated from (5.1.1). This yields three different data sets corresponding to the three values of τ_0^2. We map the negative one-half of the log-likelihood in (5.2.1).

The interpolated maps of the log-likelihood are provided in Fig. 5.1 as a function of (τ^2, ϕ) in the first two rows and of (σ^2, ϕ) in the third row. The first column presents cases with $\tau_0 = 0$, while the second and the third columns are for $\tau_0 = 0.2$ and 0.8, respectively. The grid for ϕ ranges from 2.5 to 30 so that the effective spatial ranges between 0.1 and 1.2. We specify the range of τ^2 and σ^2 to be $(0.0, 1.0)$ and $(0.2, 4.2)$, respectively, so that the pattern of the log-likelihood map around the true generating values of parameters can be captured. All the interpolated maps, including the contour lines, are drawn to the same scale.

The first row of Fig. 5.1 corresponds to $\sigma^2 = \sigma_0^2 = 1$, the second row corresponds to $\kappa = \kappa_0$ and the third row corresponds to $\tau^2 = \tau_0^2$. In the first row, we observe that similar log-likelihoods are located along parallel lines $\phi + \tau^2 = Const$. This suggests that one can identify the maximum with either a fixed ϕ or τ^2 when $\sigma^2 = \sigma_0^2$. In the second row, we find that contours for high log-likelihood values are situated around the actual generating value of nugget, supporting the identifiability of the nugget, as provided in Theorem 5.2.2. The log-likelihood along the ϕ-axis has a flat tail as ϕ decreases when fixing the nugget, which indicates having the same value of the microergodic parameter $\kappa = \sigma^2\phi^{2\nu}$ can result in equivalent probability measures (Theorem 5.2.2). The third row reveals that the log-likelihood now has close values along the curve $\sigma^2\phi = Const$, thereby corroborating Theorem 5.2.2.

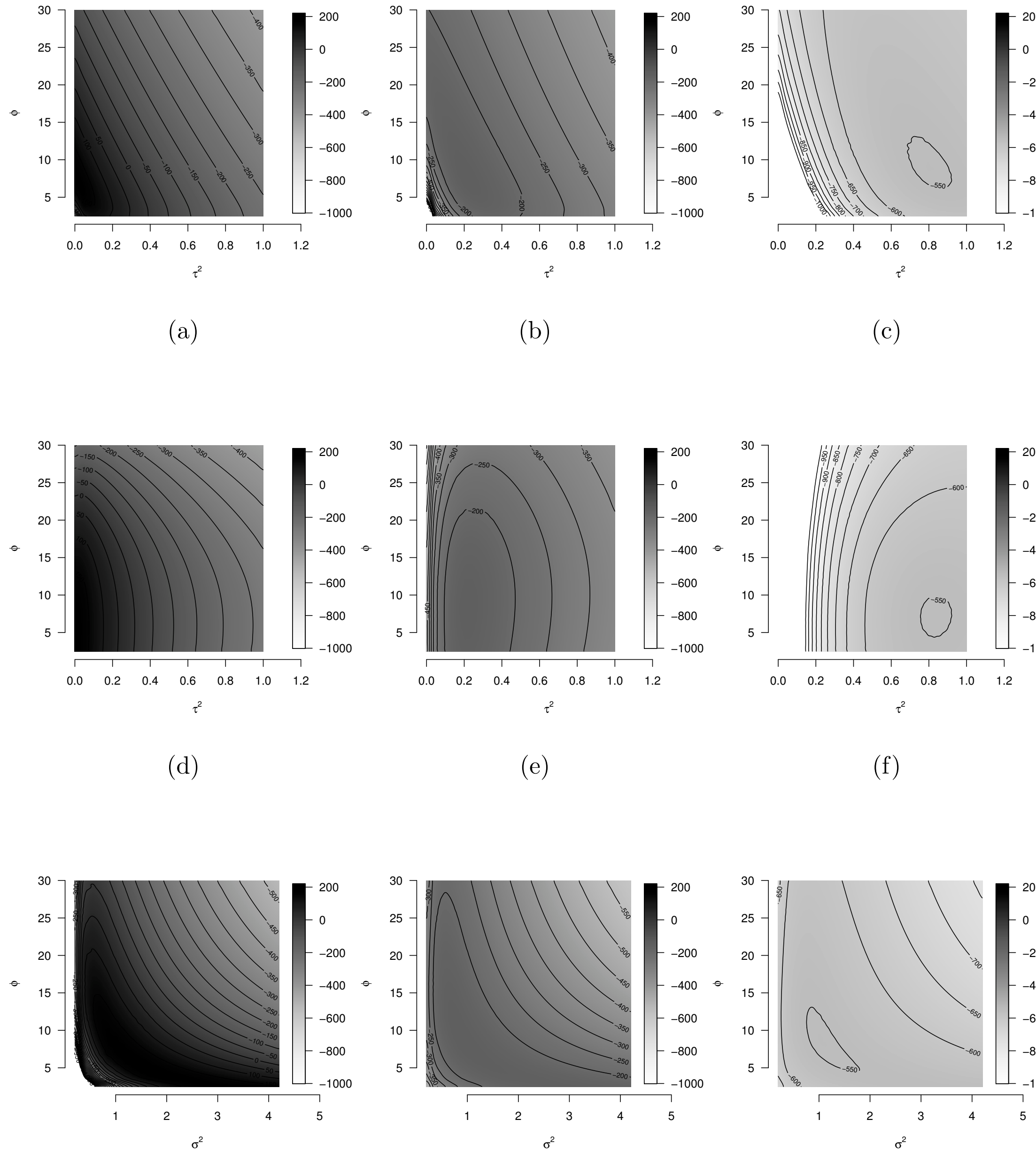
5.3.3 Parameter estimation

We use maximum likelihood estimators to illustrate the asymptotic properties of the parameter estimates. To find the maximum likelihood estimators of $\{\sigma^2, \tau^2, \phi, \kappa\}$, we use the log of the profile likelihood for ϕ and $\eta = \tau^2/\sigma^2$, given by

$$\begin{aligned}\log\{\mathcal{PL}(\phi, \eta)\} \propto & -\frac{1}{2}\log[\det\{\rho(\phi) + \eta\mathbf{I}_n\}] - \frac{n}{2} \\ & - \frac{n}{2}\log\left[\frac{1}{n}\mathbf{y}^\top\{\rho(\phi) + \eta\mathbf{I}_n\}^{-1}\mathbf{y}\right] \end{aligned} \tag{5.3.1}$$

where $\log\{\mathcal{PL}(\phi, \eta)\} = \log[\sup_{\sigma^2}\{\mathcal{L}(\sigma^2, \phi, \eta)\}]$, $\rho(\phi)$ is the correlation matrix of the underlying process $w(\cdot)$ over observed locations χ_n. We optimize (5.3.1) to obtain maximum likelihood estimators $\widehat{\phi}$ and $\widehat{\eta}$. The maximum likelihood estimator for σ^2 is $\widehat{\sigma}_n^2 = \mathbf{y}^\top\{\rho(\widehat{\phi}) + \widehat{\eta}\mathbf{I}_n\}^{-1}\mathbf{y}/n$. Calculations were executed using the R function `optimx` using the Broyden-Fletcher-Goldfarb-Shanno algorithm [Fletcher, 2013] with $\phi > 0$ and $\eta > 0$, and $\eta = 0$ for models without a nugget.

We calculate estimators for $\{\tau^2, \phi, \sigma^2, \kappa\}$ for each realization with sample sizes 400, 900 and 1600. For each parameter setting and sample size, there are 1000 estimators for $\{\tau^2, \phi, \sigma^2\}$ and κ. Figure 5.2 depicts the histograms for the maximum likelihood estimators for τ^2, ϕ, σ^2 and κ obtained from simulations with the parameter setting $\{\phi_0, \tau_0^2\} = \{7.49, 0.2\}$. There is an obvious shrinkage of the variance of estimators for τ^2 and κ as we increase the sample size from 400 to 1600. We also observe that their distribution becomes more symmetric with an increasing sample size. In contrast, the variance of the estimators for σ^2 and ϕ do not have

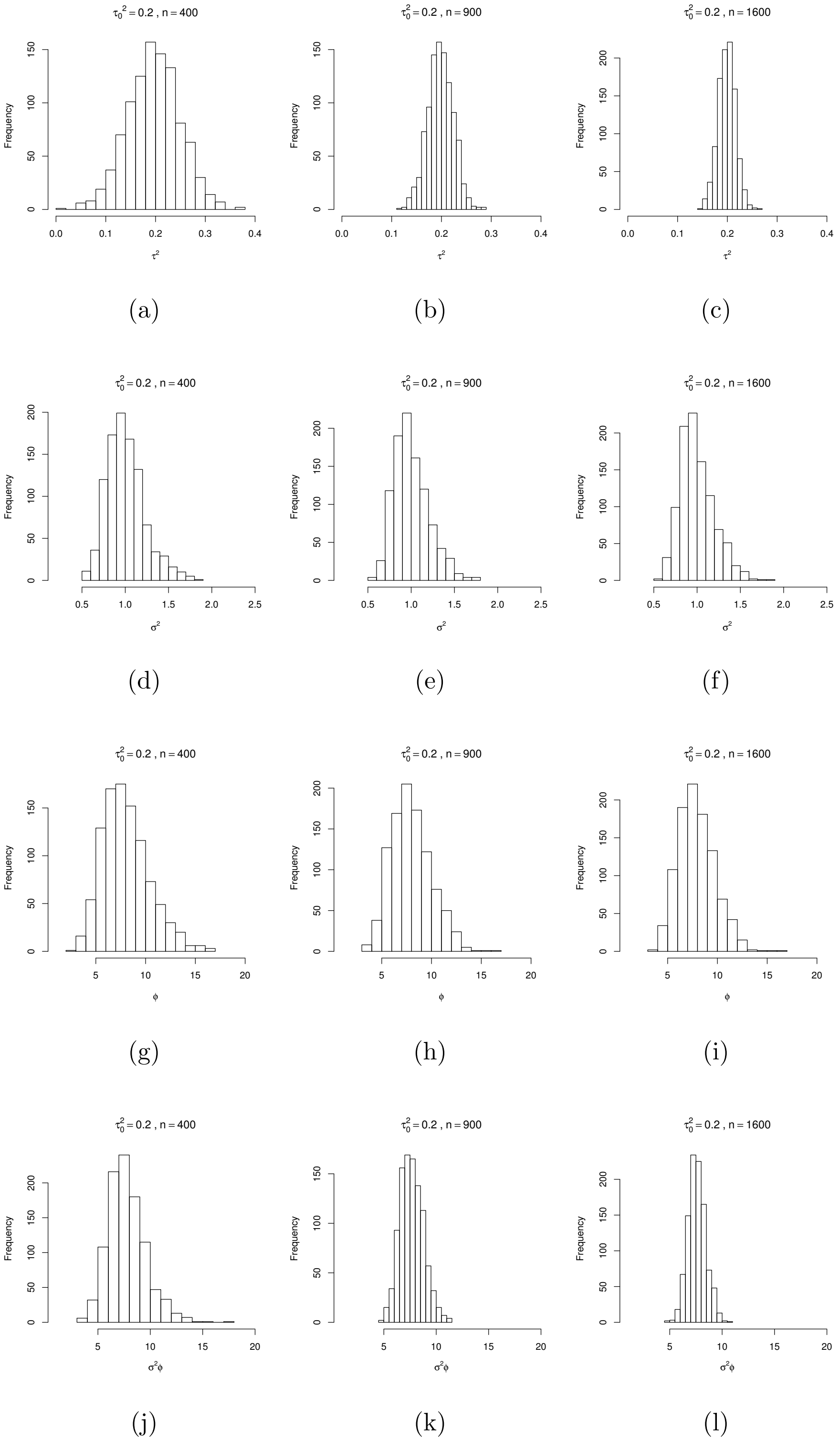


(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 5.1: Interpolated maps of log-likelihoods. Darker shades indicate higher values. Panels (a)–(c) correspond to $\sigma^2 = \sigma_0^2 = 1$, (d)–(f) correspond to $\sigma^2\phi = \phi_0 = 7.49$, and (g)–(i) correspond to $\tau_0 = \tau_0^2$. The columns correspond to $\tau_0 = 0.0$, $\tau_0 = 0.2$, and $\tau_0 = 0.8$.



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

(j) (k) (l)

Figure 5.2: Histograms of τ^2 (a)–(c), σ^2 (d)–(f), ϕ (g)–(i) and $\kappa = \sigma^2\phi^{2\nu}$ (j)–(l) for simulation with $\phi_0 = 7.49, \tau_0^2 = 0.2$.

a significant decrease as sample size increases. This is supported by the infill asymptotic results. The maximum likelihood estimators for τ^2 and κ are consistent and asymptotically normal. The maximum likelihood estimators for ϕ and σ^2 are not consistent and, hence, their variances do not decrease to zero with increasing sample size.

Table 5.1–5.4 list percentiles, biases, and sample standard deviations for the estimates of τ^2, ϕ, σ^2 and κ for each of the 9 parameter settings and offer further insights about the finite sample inference. When the spatial correlation is strong (ϕ is small), $\widehat{\tau}^2$ tends to be more precise, while $\widehat{\sigma}^2$ tends to have more variability. Unsurprisingly, the measurement error is easily distinguished from a less variable latent process $w(\cdot)$. Highly correlated realizations of $w(\cdot)$ results in less precise inference for σ^2. If the nugget is larger, then the estimators for ϕ, σ^2 and κ are less precise; the presence of measurement error weakens the precision of the estimates.

5.3.4 Interpolation

We use the kriging estimator in (5.2.19) and its mean squared prediction error (MSPE) in (5.2.20) to explore spatial interpolation in the presence of the nugget. We use (5.2.19) to predict the underlying process $w(\cdot)$ over unobserved locations. From Theorem 8 in Chapter 3 of Stein [1999], we expect a clear trend of convergence for $d = 1$. Let $\nu = 1/2$, $\tau_0^2 = 0.2$, $\sigma_0^2 = 1.0$ and $\phi_0 = 7.49$. We use (5.1.1) to generate observations over $12,000$ randomly picked locations in $[0, 1]$. We compute the MSPE using 3 hold-out points $\{0.25, 0.5, 0.75\} \in [0, 1]$ for different subsets of the data with sample sizes ranging from 500 to $12,000$. Figure 5.3(a) shows that the MSPE tends to approach 0 as sample size increases. This corroborates Stein's conjecture that the underlying process $w(\cdot)$ in (5.1.1) can be consistently estimated on a finite interval.

Next, we use the simulated data set with $n = 1600$ locations over the unit square used in Section 5.3.3. We calculate the MSPE using (5.2.20) and (5.2.21) over a 50×50 regular grid of locations over $[0, 1]^2$. This is repeated for different data sets with sample sizes

Table 5.1: Estimates of τ^2: percentiles, bias, and sample standard deviation(SD)

τ_0^2	ϕ_0	n	5%	25%	50%	75%	95%	BIAS	SD
0.200	19.972	400	0.000	0.111	0.189	0.269	0.382	-0.007	0.112
		900	0.102	0.159	0.197	0.235	0.289	-0.004	0.056
		1600	0.141	0.175	0.199	0.221	0.252	-0.002	0.035
	7.489	400	0.110	0.162	0.197	0.232	0.281	-0.003	0.053
		900	0.157	0.181	0.198	0.216	0.238	-0.002	0.025
		1600	0.170	0.187	0.199	0.211	0.227	-0.001	0.017
	2.996	400	0.152	0.177	0.196	0.217	0.248	-0.003	0.029
		900	0.173	0.188	0.199	0.212	0.227	0.000	0.017
		1600	0.182	0.191	0.200	0.208	0.219	0.000	0.012
0.800	19.972	400	0.321	0.619	0.777	0.903	1.090	-0.047	0.229
		900	0.615	0.725	0.792	0.861	0.974	-0.009	0.110
		1600	0.682	0.746	0.795	0.841	0.910	-0.006	0.069
	7.489	400	0.582	0.714	0.789	0.859	0.974	-0.015	0.114
		900	0.689	0.752	0.794	0.835	0.897	-0.006	0.065
		1600	0.725	0.768	0.799	0.826	0.869	-0.003	0.044
	2.996	400	0.662	0.738	0.789	0.845	0.931	-0.007	0.081
		900	0.720	0.766	0.797	0.828	0.871	-0.004	0.047
		1600	0.737	0.775	0.799	0.823	0.856	-0.002	0.036

Table 5.2: Estimates of ϕ: percentiles, bias, and sample standard deviation(SD)

τ_0^2	ϕ_0	n	5%	25%	50%	75%	95%	BIAS	SD
0.000	19.972	400	16.151	18.355	19.992	21.798	25.003	0.223	2.708
		900	16.706	18.642	20.072	21.548	23.928	0.182	2.185
		1600	17.077	18.800	20.041	21.403	23.557	0.144	1.968
	7.489	400	5.237	6.680	7.643	8.830	10.792	0.324	1.672
		900	5.430	6.722	7.659	8.655	10.382	0.280	1.511
		1600	5.520	6.730	7.664	8.687	10.245	0.255	1.450
	2.996	400	1.584	2.489	3.297	4.315	5.859	0.479	1.339
		900	1.605	2.468	3.316	4.298	5.792	0.463	1.299
		1600	1.624	2.490	3.259	4.279	5.613	0.448	1.281
0.200	19.972	400	13.626	17.185	20.058	23.260	28.138	0.358	4.427
		900	15.117	17.938	20.059	22.188	26.097	0.221	3.321
		1600	15.749	18.328	19.972	21.728	25.02	0.158	2.779
	7.489	400	4.596	6.271	7.757	9.377	12.430	0.535	2.364
		900	5.081	6.521	7.820	9.179	11.572	0.480	1.998
		1600	5.195	6.557	7.774	9.079	11.391	0.410	1.838
	2.996	400	1.436	2.291	3.244	4.415	6.725	0.563	1.707
		900	1.534	2.383	3.243	4.269	6.405	0.48	1.518
		1600	1.570	2.420	3.217	4.208	6.130	0.453	1.424
0.800	19.972	400	11.804	16.533	20.359	24.806	33.859	1.315	6.932
		900	14.650	17.405	20.077	23.065	27.831	0.490	4.175
		1600	15.340	17.911	20.197	22.544	26.195	0.396	3.352
	7.489	400	3.878	6.029	7.754	9.866	14.034	0.670	3.038
		900	4.468	6.266	7.745	9.317	12.249	0.475	2.402
		1600	4.691	6.430	7.735	9.142	11.663	0.405	2.157
	2.996	400	1.259	2.281	3.279	4.723	7.385	0.681	1.975
		900	1.443	2.364	3.249	4.38	7.199	0.603	1.771
		1600	1.479	2.382	3.216	4.263	6.591	0.509	1.602

Table 5.3: Estimates of σ^2: percentiles, bias, and sample standard deviation(SD)

τ_0^2	ϕ_0	n	5%	25%	50%	75%	95%	BIAS	SD
0.000	19.972	400	0.835	0.928	0.992	1.063	1.172	-0.004	0.103
		900	0.859	0.938	0.997	1.063	1.155	0.001	0.091
		1600	0.865	0.942	0.998	1.057	1.151	0.002	0.087
	7.489	400	0.721	0.860	0.976	1.109	1.374	0.000	0.198
		900	0.724	0.872	0.980	1.104	1.344	0.001	0.192
		1600	0.733	0.871	0.978	1.111	1.356	0.002	0.189
	2.996	400	0.527	0.700	0.905	1.217	1.856	0.014	0.446
		900	0.532	0.708	0.900	1.216	1.843	0.010	0.427
		1600	0.537	0.705	0.914	1.204	1.845	0.011	0.423
0.200	19.972	400	0.735	0.890	1.012	1.127	1.280	0.009	0.167
		900	0.830	0.928	1.001	1.085	1.203	0.008	0.114
		1600	0.860	0.941	1.000	1.071	1.170	0.008	0.097
	7.489	400	0.706	0.848	0.978	1.129	1.435	0.006	0.22
		900	0.732	0.855	0.972	1.128	1.373	0.002	0.203
		1600	0.731	0.857	0.970	1.116	1.374	0.000	0.195
	2.996	400	0.527	0.700	0.905	1.217	1.856	0.014	0.446
		900	0.532	0.708	0.900	1.216	1.843	0.010	0.427
		1600	0.537	0.705	0.914	1.204	1.845	0.011	0.423
0.800	400	19.972	0.653	0.874	1.025	1.208	1.531	0.050	0.265
		900	0.761	0.911	1.014	1.110	1.257	0.011	0.149
		1600	0.826	0.931	1.009	1.085	1.197	0.009	0.113
	7.489	400	0.640	0.848	1.004	1.174	1.487	0.027	0.263
		900	0.701	0.862	0.990	1.146	1.421	0.016	0.225
		1600	0.710	0.860	0.985	1.129	1.413	0.012	0.215
	2.996	400	0.482	0.715	0.955	1.254	1.916	0.047	0.482
		900	0.517	0.720	0.950	1.240	1.874	0.044	0.462
		1600	0.524	0.735	0.968	1.250	1.839	0.045	0.449

Table 5.4: Estimates of κ: percentiles, bias, and sample standard deviation(SD)

τ_0^2	ϕ_0	n	5%	25%	50%	75%	95%	BIAS	SD
0.000	19.972	400	17.200	18.596	19.752	21.117	23.197	-0.045	1.881
		900	18.098	19.221	19.957	20.798	21.974	0.035	1.177
		1600	18.764	19.457	19.973	20.531	21.399	0.039	0.805
	7.489	400	6.538	7.092	7.499	7.943	8.568	0.032	0.619
		900	6.903	7.236	7.500	7.784	8.146	0.018	0.387
		1600	7.061	7.317	7.491	7.680	7.979	0.013	0.280
	2.996	400	2.666	2.869	3.004	3.158	3.369	0.018	0.213
		900	2.780	2.915	3.001	3.103	3.254	0.012	0.142
		1600	2.841	2.935	3.000	3.077	3.191	0.011	0.106
0.200	19.972	400	11.760	16.227	20.111	24.691	31.242	0.677	6.052
		900	14.827	17.806	19.879	22.566	26.735	0.313	3.693
		1600	16.421	18.434	19.943	21.624	24.404	0.186	2.528
	7.489	400	5.116	6.546	7.552	8.825	11.045	0.268	1.802
		900	5.999	6.843	7.605	8.404	9.645	0.177	1.110
		1600	6.197	7.033	7.585	8.141	9.085	0.105	0.850
	2.996	400	2.010	2.546	3.040	3.533	4.322	0.092	0.716
		900	2.282	2.706	3.028	3.343	3.900	0.055	0.493
		1600	2.434	2.779	3.012	3.292	3.724	0.040	0.384
0.800	19.972	400	8.846	15.161	20.858	28.202	47.108	3.314	12.319
		900	12.700	16.839	20.077	24.320	31.399	0.830	5.715
		1600	14.846	17.751	20.215	22.941	26.997	0.530	3.888
	7.489	400	4.080	5.980	7.677	9.679	13.537	0.591	2.929
		900	5.084	6.394	7.626	8.923	10.918	0.269	1.808
		1600	5.598	6.675	7.622	8.546	10.030	0.169	1.361
	2.996	400	1.708	2.444	3.093	3.849	5.432	0.259	1.175
		900	1.999	2.626	3.114	3.666	4.534	0.185	0.789
		1600	2.210	2.712	3.086	3.478	4.210	0.129	0.618

varying between 400 and 1600. Figure 5.3(b) shows that the MSPE decreases as sample size increases. This trend still holds when the predictor is formed under misspecified models, a finding similar to those in Kaufman and Shaby [2013] without the nugget. If ν is fixed at the true generating value, then predictions under any parameter setting are consistent and asymptotically efficient with no nugget effect. The proof in Kaufman and Shaby [2013] is based on Stein [1993], hence their results do not carry over to our setting due to the discontinuity in our covariogram at 0. (This technical difficulty was also pointed out by Yakowitz and Szidarovszky [1985, p.38]). However, their results suggest empirical studies to explore the asymptotic properties of interpolation.

To compare with results in Kaufman and Shaby [2013, Section 2.3], we examine two ratios

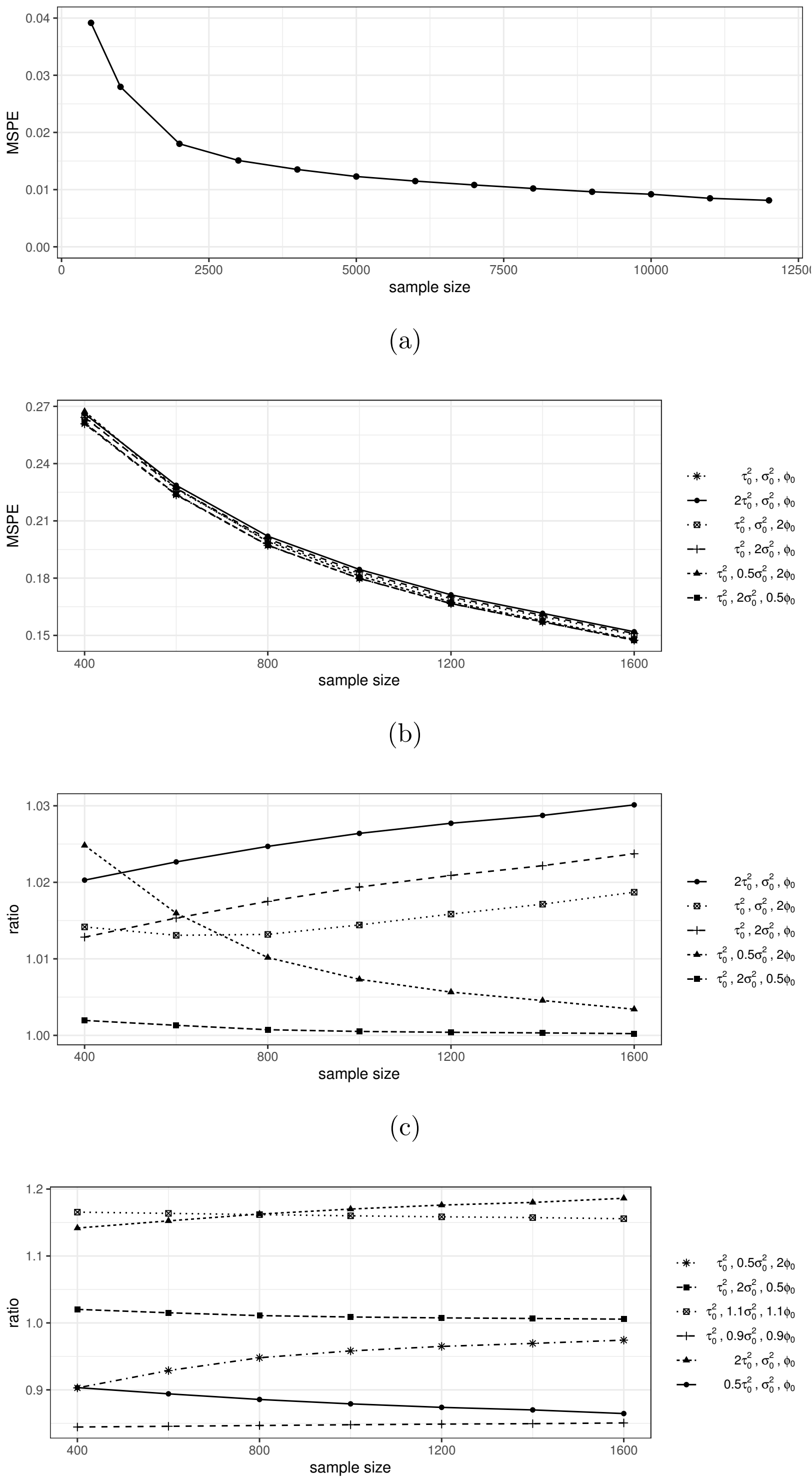
$$\text{i)}\quad \frac{\mathrm{Var}_{\tau_0^2,\sigma_0^2,\phi_0}\{\widehat{z}_n(\tau_1^2,\sigma_1^2,\phi_1)-w_0\}}{\mathrm{Var}_{\tau_0^2,\sigma_0^2,\phi_0}\{\widehat{z}_n(\tau_0^2,\sigma_0^2,\phi_0)-w_0\}},\quad \text{and ii)}\quad \frac{\mathrm{Var}_{\tau_1^2,\sigma_1^2,\phi_1}\{\widehat{z}_n(\tau_0^2,\sigma_1^2,\phi_1)-w_0\}}{\mathrm{Var}_{\tau_0^2,\sigma_0^2,\phi_0}\{\widehat{z}_n(\tau_0^2,\sigma_1^2,\phi_1)-w_0\}}.$$

Figure 5.3(c) compares the ratio defined by i). This ratio tends to approach 1 only when $\tau_1^2 = \tau_0^2$ and $\kappa = \kappa_0$. Unlike the case with no nugget, asymptotic efficiency is only observed when the estimator is fitted under models with Gaussian measures equivalent to the generating Gaussian measure. Figure 5.3(d) plots the ratio defined by ii). As in Fig. 5.3(c), this ratio also tends to approach 1 only when $\tau_1^2 = \tau_0^2$, $\kappa = \kappa_0$. Based on our simulation study, we posit that the asymptotic efficiency and asymptotically correct estimation of MSPE hold only when $\tau_1^2 = \tau_0^2$, $\kappa = \kappa_0$.

5.4 Discussion

We have developed insights into inference under infill asymptotics of Gaussian process parameters in the context of spatial or geostatistical analysis in the presence of the nugget effect. Our work in this Chapter can be regarded as an extension of similar investigations without the nugget effect.

We have discussed the complications in establishing consistency and asymptotic efficiency in parameter estimation and spatial prediction due to the discontinuity introduced by the



(a)

(b)

(c)

(d)

Figure 5.3: The MSPE for $w(\cdot)$ at (a) unobserved locations with study domain $[0, 1]$ (b) a 50×50 grid over $[0, 1]^2$. The ratio of mean square predict error (ratio) for testing asymptotic efficiency (c) and asymptotically correct estimation of MSPE (d)

nugget. Tools in standard spectral analysis no longer work in this scenario. Understanding the behavior of such processes will enhance our understanding of identifiability of process parameters. For example, the failure to consistently estimate certain (non-microergodic) parameters can also be useful for Bayesian inference where we can conclude that the effect of the likelihood will never overwhelm the prior when calculating the posterior distribution of non-microergodic parameters.

We anticipate the current manuscript to generate further research in variants of geostatistical models with the nugget. For example, it is conceivable that these results will lead to asymptotic investigations of covariance-tapered models that too have been investigated without the nugget by Wang et al. [2011]. One can also explore whether some results, such as Theorem 2 in Kaufman and Shaby [2013] where ϕ is estimated, will hold for the Matérn model with the nugget. Our simulations also suggest further research in asymptotic efficiency provided in Theorem 3 of Kaufman and Shaby [2013] in the presence of the nugget. With recent interest in scaleable Gaussian process models, we can investigate asymptotic properties of approximations indicated on the lines of Vecchia [1988] and Section 10.5.3 in Zhang [2012]. Finally, we point out that the conditions in Assumptions 5.2.7 and 5.2.8 about the eigenvalues estimates are quite expected and their rigorous proofs will constitute future research. In particular, a rigorous proof of Assumption 5.2.8 is challenging and will be of interest in general kernel methods and bandit problems.

APPENDIX A

Appendix

A.1 Multivariate Response NNGP Model with Misalignment

The conjugacy in conjugate multivariate response model is violated with misalignment. Therefore, for datasets with misalignment, we need to drop data on the location with misalignment and use the "cleaned up" data to obtain quick inference through the conjugate model. However, for NNGP based response model, we can provide a modified algorithm that can utilize the "dropped" data to improve the inference of the conjugate multivariate response NNGP model.

Assume $\mathcal{S}$ is the set of observed locations, where at least one response is recorded, $\mathcal{M}_i \subset \mathcal{S}$ is the set of locations where the ith response has not been observed. Let $\mathcal{M} = \cup\{\mathcal{M}_i\}_{i=1}^q$ and assume $\mathcal{M}$ has n_m locations, then $\mathcal{R} = \mathcal{S} \setminus \mathcal{M}$ is the observed n_r locations with no misalignment. We label $\mathbf{Y}_{\mathcal{R}}$ and $\mathbf{X}_{\mathcal{R}}$ as the responses and design matrix over $\mathcal{R}$. The posterior distributions $p(\mathbf{\Sigma} \,|\, \mathbf{Y}_{\mathcal{R}})$ and $p(\boldsymbol{\beta} \,|\, \mathbf{\Sigma}, \mathbf{Y}_{\mathcal{R}})$ are given in (3.2.5) using $\mathcal{R}$ instead of $\mathcal{S}$. For $\mathbf{s} \in \mathcal{M}$, we use footnote os to denote the index of observed responses on $\mathbf{s}$, the vector of observed responses on location $\mathbf{s}$ is $\mathbf{y}(\mathbf{s})_{os}$ and the corresponding coefficient matrix is $\boldsymbol{\beta}_{os}$. Then, the posterior distribution of $\mathbf{\Sigma}$ given all observed data is

$$\begin{aligned} p(\mathbf{\Sigma} \,|\, \mathbf{Y}_{\mathcal{R}}, \{\mathbf{y}(\mathbf{s})_{os}\}_{\mathbf{s}\in\mathcal{M}}) &\propto \{\int p(\{\mathbf{y}(\mathbf{s})_{os}\}_{\mathbf{s}\in\mathcal{M}} \,|\, \mathbf{\Sigma}, \boldsymbol{\beta}, \mathbf{Y}_{\mathcal{R}}) p(\boldsymbol{\beta} \,|\, \mathbf{\Sigma}, \mathbf{Y}_{\mathcal{R}}) d\boldsymbol{\beta}\} p(\mathbf{\Sigma} \,|\, \mathbf{Y}_{\mathcal{R}}) \\ &\propto p(\{\mathbf{y}(\mathbf{s})_{os}\}_{\mathbf{s}\in\mathcal{M}} \,|\, \mathbf{\Sigma}, \mathbf{Y}_{\mathcal{R}}) p(\mathbf{\Sigma} \,|\, \mathbf{Y}_{\mathcal{R}}) \ . \end{aligned} \tag{A.1.1}$$

Now consider the formulation of $p(\{\mathbf{y}(\mathbf{s})_{os}\}_{\mathbf{s}\in\mathcal{M}} \,|\, \mathbf{\Sigma}, \mathbf{Y}_{\mathcal{R}})$. Let $\mathcal{R}$ be the reference set for the response processes. Define the m nearest neighbor of $\mathbf{s}$ in $\mathcal{R}$ as $N_m(\mathbf{s})$ and $n_{\mathcal{R}}$ as the number of locations in $\mathcal{R}$, we have

$$p(\{\mathbf{y}(\mathbf{s})_{os}\}_{\mathbf{s}\in\mathcal{M}} \,|\, \boldsymbol{\Sigma}, \boldsymbol{\beta}, \mathbf{Y}_{\mathcal{R}}) = \prod_{\mathbf{s}\in\mathcal{M}} \mathrm{N}(\mathbf{y}(\mathbf{s})_{os} \,|\, \mathrm{vec}[\mathbf{x}(\mathbf{s})^\top\boldsymbol{\beta} + \mathbf{L}_{\mathbf{s}}^\top\{\mathbf{Y}_{\mathcal{R}} - \mathbf{X}_{\mathcal{R}}\boldsymbol{\beta}\}]_{os}, \mathbf{D}_{\mathbf{s}}) \ ,$$

where $\mathbf{L}_{\mathbf{s}}$ is a $n_{\mathcal{R}} \times 1$ vector whose i-th element is zero if $\mathbf{s}_i \notin N_m(\mathbf{s})$. Define the index of nonzero elements in $\mathbf{L}_{\mathbf{s}}$ as Pa[$\mathbf{s}$], then we have

$$\begin{aligned} \mathbf{L}_{\mathbf{s}}[\mathrm{Pa}[\mathbf{s}]] &= [\mathbf{C}(\mathbf{s}, N_m(\mathbf{s}))\mathbf{C}(N_m(\mathbf{s}), N_m(\mathbf{s}))^{-1}]^\top \ , \\ \mathbf{D}_{\mathbf{s}} &= [\mathbf{C}(\mathbf{s},\mathbf{s}) - \mathbf{C}(\mathbf{s}, N_m(\mathbf{s}))\mathbf{C}(N_m(\mathbf{s}), N_m(\mathbf{s}))^{-1}\mathbf{C}(N_m(\mathbf{s}), \mathbf{s})]\boldsymbol{\Sigma}_{[os,os]} \end{aligned} \tag{A.1.2}$$

where $\mathbf{C}$ is defined as $\rho_\psi(\mathbf{s},\mathbf{s}') + (\alpha^{-1} - 1)\delta_{\mathbf{s}=\mathbf{s}'}$ with δ denoting a Dirac's delta function, $\boldsymbol{\Sigma}_{[os,os]}$ denotes the sub-matrix extracted from $\boldsymbol{\Sigma}$ with row and column index os. Let $\boldsymbol{\beta} \,|\, \boldsymbol{\Sigma}, \mathbf{Y}_{\mathcal{R}} \sim \mathrm{MN}(\boldsymbol{\mu}^*, \mathbf{V}^*, \boldsymbol{\Sigma})$ and define $\mathbf{V}_{pr} = \boldsymbol{\Sigma} \otimes \mathbf{V}^*$, by intergrating out $\boldsymbol{\beta}$ we have

$$\{\mathbf{y}(\mathbf{s})_{os}\}_{\mathbf{s}\in\mathcal{M}} \,|\, \boldsymbol{\Sigma}, \mathbf{Y}_{\mathcal{R}} \sim \mathrm{N}(\mathbf{H}_1\mathrm{vec}(\boldsymbol{\mu}^*) + \mathbf{H}_2\mathrm{vec}(\mathbf{Y}_{\mathcal{R}}), \mathrm{diag}(\{\mathbf{D}_{\mathbf{s}}\}_{\mathbf{s}\in\mathcal{M}}) + \mathbf{H}_1\mathbf{V}_{pr}\mathbf{H}_1^\top) \ , \tag{A.1.3}$$

where

$$\begin{aligned} \mathbf{H}_1 &= \{\mathbf{I}_{q[os,:]} \otimes \mathbf{x}(\mathbf{s})^\top\}_{\mathbf{s}\in\mathcal{M}} - \mathbf{H}_2[\mathbf{I}_q \otimes \mathbf{X}_{\mathcal{R}}] \ , \\ \mathbf{H}_2 &= \{\mathbf{I}_{q[os,:]} \otimes \mathbf{L}_{\mathbf{s}}^\top\}_{\mathbf{s}\in\mathcal{M}} \ . \end{aligned} \tag{A.1.4}$$

Here, $\mathbf{I}_{q[os,:]}$ is the osth row of a $q \times q$ identity matrix, thus $\mathbf{H}_2$ provides the index of observed response and weights of neighbors for $\mathbf{s} \in \mathcal{M}$. Once q and p are relatively small, we can use matrix determinant lemma and Sherman-Morrison-Woodbury formulas

$$\begin{aligned} &\det(\mathrm{diag}(\{\mathbf{D}_{\mathbf{s}}\}_{\mathbf{s}\in\mathcal{M}}) + \mathbf{H}_1\mathbf{V}_{pr}\mathbf{H}_1^\top) = \det(\mathbf{V}_{update})\det(\mathbf{V}_{pr}) \prod_{\mathbf{s}\in\mathcal{M}} \det(\mathbf{D}_{\mathbf{s}}) \ ; \\ &(\mathrm{diag}(\{\mathbf{D}_{\mathbf{s}}\}_{\mathbf{s}\in\mathcal{M}}) + \mathbf{H}_1\mathbf{V}_{pr}\mathbf{H}_1^\top)^{-1} = \mathrm{diag}(\{\mathbf{D}_{\mathbf{s}}^{-1}\}_{\mathbf{s}\in\mathcal{M}}) - \\ &\qquad\qquad \mathrm{diag}(\{\mathbf{D}_{\mathbf{s}}^{-1}\}_{\mathbf{s}\in\mathcal{M}})\mathbf{H}_1\mathbf{V}_{update}^{-1}\mathbf{H}_1^\top\mathrm{diag}(\{\mathbf{D}_{\mathbf{s}}^{-1}\}_{\mathbf{s}\in\mathcal{M}}) \ , \end{aligned} \tag{A.1.5}$$

with $\mathbf{V}_{update} = \mathbf{V}_{pr}^{-1} + \mathbf{H}_1^\top\mathrm{diag}(\{\mathbf{D}_{\mathbf{s}}^{-1}\}_{\mathbf{s}\in\mathcal{M}})\mathbf{H}_1$ to facilitate the calculation of $p(\{\mathbf{y}(\mathbf{s})_{os}\}_{\mathbf{s}\in\mathcal{M}} \,|\, \boldsymbol{\Sigma}, \mathbf{Y}_{\mathcal{R}})$. Plugging the above two equations into the log-likelihood of $\{\mathbf{y}(\mathbf{s})_{os}\}_{\mathbf{s}\in\mathcal{M}} \,|\, \boldsymbol{\Sigma}, \mathbf{Y}_{\mathcal{R}}$

$$\begin{aligned} \log\{p(\{\mathbf{y}(\mathbf{s})_{os}\}_{\mathbf{s}\in\mathcal{M}} \,|\, \boldsymbol{\Sigma}, \mathbf{Y}_{\mathcal{R}})\} &= -\frac{1}{2}\log\{\det(\mathrm{diag}(\{\mathbf{D}_{\mathbf{s}}\}_{\mathbf{s}\in\mathcal{M}}) + \mathbf{H}_1(\boldsymbol{\Sigma} \otimes \mathbf{V}^*)\mathbf{H}_1^\top)\} - \\ &\frac{1}{2}\{\boldsymbol{\mu}_{\mathcal{M}}^\top(\mathrm{diag}(\{\mathbf{D}_s\}_{s\in\mathbf{M}}) + \mathbf{H}_1(\boldsymbol{\Sigma} \otimes \mathbf{V}^*)\mathbf{H}_1^\top)^{-1}\mu_{\mathcal{M}} \ , \\ \mu_{\mathcal{M}} &= \mathbf{H}_1\mathrm{vec}(\boldsymbol{\mu}^*) + \mathbf{H}_2\mathrm{vec}(\mathbf{Y}_{\mathcal{R}}) \ , \end{aligned} \tag{A.1.6}$$

then we can use the likelihood and (A.1.1) to conduct MCMC algorithm for $\mathbf{\Sigma}$.

Since $\mathbf{\Sigma}$ is positive-definite, we represent $\mathbf{\Sigma}$ through $\mathbf{L}\mathbf{L}^\top$ and update $\mathbf{L}$ in the MCMC chain instead. The MCMC update requires transforming the prior by the Jacobian $2^q \prod_{i=1}^q \mathbf{L}_{ii}^{q-i+1}$ to account for the map between $\mathbf{L}$ and $\mathbf{\Sigma}$. Benefit from an informative prior $\mathbf{\Sigma} \,|\, \mathbf{Y}_R$, we can estimate the covariance matrix of the posterior distribution of elements in $\mathbf{L}$ through $\mathbf{\Sigma} \,|\, \mathbf{Y}$, and design a Gaussian distribution with a covariance matrix equals 2.38^2 times the estimated covariance matrix as the proposal distribution. The Cholesky decomposition of $E(\mathbf{\Sigma} \,|\, \mathbf{Y}_R) = \mathbf{\Psi}/(\nu^* - q - 1)$ also serves as a good initial value for $\mathbf{L}$.

The posterior inference of $\boldsymbol{\beta}$ is more straightforward after obtaining the samples of $\mathbf{\Sigma} \,|\, \mathbf{Y}_\mathcal{R}, \{\mathbf{y}(\mathbf{s})_{os}\}_{\mathbf{s}\in\mathcal{M}}$. We formulate $\boldsymbol{\beta} \,|\, \mathbf{\Sigma}$, $\mathbf{Y}_\mathcal{R} \,|\, \boldsymbol{\beta}, \mathbf{\Sigma}$ and $\{\mathbf{y}(\mathbf{s})_{os}\}_{\mathbf{s}\in\mathcal{M}} \,|\, \boldsymbol{\beta}, \mathbf{\Sigma}$ as the following augmented linear system,

$$\underbrace{\begin{bmatrix} \mathrm{vec}(\mathbf{D}_\mathcal{R}^{-\frac{1}{2}}(\mathbf{I} - \mathbf{A}_\mathcal{R})\mathbf{Y}_\mathcal{R}) \\ \{\mathbf{y}(\mathbf{s})_{os}\}_{\mathbf{s}\in\mathcal{M}} - \mathbf{H}_2\mathrm{vec}(\mathbf{Y}_\mathcal{R}) \\ \mathrm{vec}(\mathbf{L}_r^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}}) \end{bmatrix}}_{\mathbf{Y}^*} = \underbrace{\begin{bmatrix} \mathbf{I}_q \otimes \mathbf{D}_\mathcal{R}^{-\frac{1}{2}}(\mathbf{I} - \mathbf{A}_\mathcal{R})\mathbf{X}_\mathcal{R} \\ \mathbf{H}_1 \\ \mathbf{I}_q \otimes \mathbf{L}_r^{-1} \end{bmatrix}}_{\mathbf{X}^*} \underbrace{\begin{bmatrix} \boldsymbol{\beta}_1 \\ \vdots \\ \boldsymbol{\beta}_q \end{bmatrix}}_{\mathrm{vec}(\boldsymbol{\beta})} + \underbrace{\begin{bmatrix} \boldsymbol{\eta}_1 \\ \boldsymbol{\eta}_2 \\ \boldsymbol{\eta}_3 \end{bmatrix}}_{\boldsymbol{\eta}} , \tag{A.1.7}$$

where $\mathbf{A}_\mathcal{R}$ and $\mathbf{D}_\mathcal{R}$ are the $\mathbf{A}_\mathcal{K}$ and $\mathbf{D}_\mathcal{K}$ defined in (3.2.12) with reference set being $\mathcal{R}$, and $\boldsymbol{\eta}$ follows a zero-centered Gaussian distribution with covariance matrix

$$\mathbf{V}_{\boldsymbol{\eta}} = \begin{bmatrix} \mathbf{\Sigma} \otimes \mathbf{I}_\mathcal{R} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathrm{diag}(\{\mathbf{\Sigma}_{[os,os]}\}_{\mathbf{s}\in\mathcal{M}}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{\Sigma} \otimes \mathbf{I}_p \end{bmatrix} . \tag{A.1.8}$$

The posterior distribution of $\mathrm{vec}(\boldsymbol{\beta}) \,|\, \mathbf{\Sigma}, \mathbf{Y}_\mathcal{R}, \{\mathbf{y}(\mathbf{s})_{os}\}_{\mathbf{s}\in\mathcal{M}}$ follows $\mathrm{MVN}(\boldsymbol{\mu}^*_{\boldsymbol{\beta}}, \mathbf{V}^*_{\boldsymbol{\beta}})$ with

$$\mathbf{V}^*_{\boldsymbol{\beta}} = (\mathbf{X}^{*\top}\mathbf{V}_{\boldsymbol{\eta}}^{-1}\mathbf{X}^*)^{-1} \;,\; \boldsymbol{\mu}^*_{\boldsymbol{\beta}} = \mathbf{V}^*_{\boldsymbol{\beta}}(\mathbf{X}^{*\top}\mathbf{V}_{\boldsymbol{\eta}}^{-1}\mathbf{Y}^*) \;. \tag{A.1.9}$$

The posterior prediction over unobserved and misaligned locations Taking $\mathcal{R}$ as the reference set, the full conditional posterior distribution of a new location $\mathbf{u}$ is

$$\mathbf{y}(\mathbf{u}) \,|\, \mathbf{Y}_\mathcal{R}, \mathbf{\Sigma}, \boldsymbol{\beta} \sim \mathrm{MVN}(\mathbf{x}(\mathbf{u})^\top\boldsymbol{\beta} + \tilde{\mathbf{A}}_\mathbf{u}^\top[\mathbf{Y}_\mathcal{R} - \mathbf{X}_\mathcal{R}\boldsymbol{\beta}], \tilde{\mathbf{D}}_\mathbf{u}\mathbf{\Sigma}) \;, \tag{A.1.10}$$

where $\tilde{\mathbf{A}}_{\mathbf{u}}$ is a $n_{\mathcal{R}} \times 1$ vector whose i-th element is zero if $\mathbf{s}_i \notin N_m(\mathbf{u})$. Define the index of nonzero elements in $\tilde{\mathbf{A}}_{\mathbf{u}}$ as Pa[$\mathbf{s}$], then

$$\begin{aligned} \tilde{\mathbf{A}}_{\mathbf{u}}[\text{Pa}[\mathbf{u}]] &= \{\mathbf{C}(\mathbf{u}, N_m(\mathbf{u}))[\mathbf{C}(N_m(\mathbf{u}), N_m(\mathbf{u})) + (\alpha^{-1} - 1)\mathbf{I}_m]^{-1}\}^\top , \\ \tilde{\mathbf{D}}_{\mathbf{u}} &= \alpha^{-1} - \tilde{\mathbf{A}}_{\mathbf{u}}[\text{Pa}[\mathbf{u}]]^\top \mathbf{C}(N_m(\mathbf{u}), \mathbf{u}) . \end{aligned} \tag{A.1.11}$$

For $\mathbf{s} \in \mathcal{M}$, label the index of unobserved responses at $\mathbf{s}$ by us, by the definition of NNGP,

$$\mathbf{y}(\mathbf{s})_{us} \,|\, \mathbf{Y}_{\mathcal{R}}, \{\mathbf{y}(\mathbf{s})_{os}\}_{\mathbf{s} \in \mathcal{M}}, \boldsymbol{\Sigma}, \boldsymbol{\beta} = \mathbf{y}(\mathbf{s})_{us} \,|\, \mathbf{Y}_{\mathcal{R}}, \mathbf{y}(\mathbf{s})_{os}, \boldsymbol{\Sigma}, \boldsymbol{\beta} \tag{A.1.12}$$

follows a Gaussian distribution with mean $\text{E}\{\mathbf{y}(\mathbf{s}) \,|\, \mathbf{Y}_{\mathcal{R}}\}_{us} - \boldsymbol{\Sigma}_{[us,os]}\boldsymbol{\Sigma}^{-1}_{[os,os]}\{\mathbf{y}(\mathbf{s})_{os} - \text{E}(\mathbf{y}(\mathbf{s}) \,|\, \mathbf{Y}_{\mathcal{R}})_{os}\}$ and covariance matrix $\tilde{\mathbf{D}}_{\mathbf{s}}(\boldsymbol{\Sigma}_{[us,us]} - \boldsymbol{\Sigma}_{[us,os]}\boldsymbol{\Sigma}^{-1}_{[os,os]}\boldsymbol{\Sigma}_{[os,us]})$ where $\text{E}\{\mathbf{y}(\mathbf{s}) \,|\, \mathbf{Y}_{\mathcal{R}}\} = \mathbf{x}(\mathbf{s})^\top \boldsymbol{\beta} + \tilde{\mathbf{A}}_{\mathbf{s}}^\top[\mathbf{Y}_{\mathcal{R}} - \mathbf{X}_{\mathcal{R}}\boldsymbol{\beta}]$. The following gives the detailed algorithm:

Algorithm A.1: Obtaining inference of $\{\boldsymbol{\beta}, \boldsymbol{\Sigma}\}$ and predictions for conjugate multivariate response NNGP with misalignment.

```
1. Obtain $\boldsymbol{\mu}^*$, $\mathbf{V}^*$, $\boldsymbol{\Psi}^*$ and $\nu^*$ in $\boldsymbol{\Sigma} \,|\, \mathbf{Y}_{\mathcal{R}} \sim \text{IW}(\boldsymbol{\Psi}^*, \nu^*)$ through step 1 in Algorithm 3.1.
2. Generate posterior samples of $\boldsymbol{\Sigma} \,|\, \mathbf{Y}_{\mathcal{R}}, \{\mathbf{y}(\mathbf{s})_{os}\}_{\mathbf{s} \in \mathcal{M}}$ through MCMC algorithm.
    (a) Take the Cholesky decomposition of $E(\boldsymbol{\Sigma} \,|\, \mathbf{Y}_{\mathcal{R}})$ as the starting point $\mathbf{L}^{(0)}$ of the MCMC chains
    (b) Design proposal distribution for elements of $\mathbf{L}^{(l)}$ as a multivariate Gaussian with $2.38^2$ times the covariance
        estimated from $\text{IW}(\boldsymbol{\Psi}^*, \nu^*)$
    (c) Construct $\mathbf{L}_{\mathbf{s}}$ and $\mathbf{D}_{\mathbf{s}}$ for $\mathbf{s} \in \mathcal{M}$ as described in (A.1.2)    $\mathcal{O}(n_m m^3)$
    (d) Construct $\mathbf{H}_1$, $\mathbf{H}_2$ in (A.1.4) and calculate $\mu_{\mathcal{M}}$ in (A.1.6)    $\mathcal{O}(n_m m^2)$
    (e) For $l$ in $1 : L$
        i. Propose new $\boldsymbol{\Sigma}^* = \mathbf{L}^*\mathbf{L}^{*\top}$ based on $\boldsymbol{\Sigma}^{(l-1)} = \mathbf{L}^{(l)}\mathbf{L}^{(l)\top}$
        ii. Calculate the likelihood of the new proposed $\boldsymbol{\Sigma}^*$ and $\boldsymbol{\Sigma}^{(l)}$ given $\mathbf{F}^{(l)}$
            • Obtain the Cholesky decomposition $\mathbf{L}_{update}$ of $\mathbf{V}_{update}$ in (A.1.5)    $\mathcal{O}(p^2q^2n_m)$
            • Generate $u = \mathbf{L}^{-1}_{update}\mathbf{H}_1\text{diag}(\{\mathbf{D}_{\mathbf{s}}^{-1}\}_{\mathbf{s} \in \mathcal{M}})\mu_{\mathcal{M}}$    $\mathcal{O}(pq^2n_m)$
            • Calculate log-likelihood $l(\boldsymbol{\Sigma} \,|\, \mathbf{Y}_{\mathbf{R}}, \{\mathbf{y}(\mathbf{s})_{os}\}_{\mathbf{s} \in \mathcal{M}})$ through    $\mathcal{O}(n_m)$
              $l(\boldsymbol{\Sigma} \,|\, \mathbf{Y}_{\mathbf{R}}) - \frac{1}{2}(\text{logdet}(\mathbf{V}_{update}) + \text{logdet}(\mathbf{V}_{pr}) - \sum_{s \in \mathcal{M}} \text{logdet}(\mathbf{D}_{\mathbf{s}}) + \mu_{\mathcal{M}}^\top \text{diag}(\{\mathbf{D}_{\mathbf{s}}^{-1}\}_{\mathbf{s} \in \mathcal{M}})\mu_{\mathcal{M}} - u^\top u)$
            • Add the log of the Jacobian $2^q \prod_{i=1}^q \mathbf{L}_{ii}$ to the log-likelihood from last step with $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^\top$
        iii. Accept the new $\boldsymbol{\Sigma}^*$ as $\boldsymbol{\Sigma}^{(l)}$ with the probability of the ratio of the likelihood of $\boldsymbol{\Sigma}^*$ and $\boldsymbol{\Sigma}^{(l)}$. Let $\boldsymbol{\Sigma}^{(l)} =$
             $\boldsymbol{\Sigma}^{(l-1)}$ when the new proposal is rejected.
3. Generate posterior samples of $\boldsymbol{\beta}$
    (a) Construct $\mathbf{A}_{\mathcal{R}}$ and $\mathbf{D}_{\mathcal{R}}$ in (A.1.7)    $\mathcal{O}(n_r m^3)$
    (b) Construct $\mathbf{X}^*$ and $\mathbf{Y}^*$ in (A.1.7)    $\mathcal{O}(n_r(m+1)(p+q))$
    (c) For each $\boldsymbol{\Sigma}^{(l)}$ after burn-in
```

```
            i. Construct $\mathbf{V}_\eta$ in (A.1.8) and $\mathbf{V}_{\boldsymbol{\beta}}^*$, $\boldsymbol{\mu}_{\boldsymbol{\beta}}^*$ in (A.1.9)    $\mathcal{O}((p+q)q^2n)$
            ii. Generate vec($\boldsymbol{\beta}^{(l)}$) from MVN($\boldsymbol{\mu}_{\boldsymbol{\beta}}^*, \mathbf{V}_{\boldsymbol{\beta}}^*$)
    4. Generate posterior predictive samples of unobserved responses on $\mathcal{U}$ and $\mathcal{M}$.
        (a) Obtain $\tilde{\mathbf{A}}_{\mathbf{u}}$,$\tilde{\mathbf{D}}_{\mathbf{u}}$ in (A.1.11) for all $\mathbf{u} \in \mathcal{U}$.    $\mathcal{O}(n'm^3)$
        (b) For each pair of $\boldsymbol{\beta}^{(l)}, \boldsymbol{\Sigma}^{(l)}$
            i. Generate $\mathbf{y}(\mathbf{u}) \,|\, \boldsymbol{\beta}^{(l)}, \boldsymbol{\Sigma}^{(l)}, \mathbf{Y}_{\mathcal{R}}$ for $\mathbf{u} \in \mathcal{U}$ through (A.1.10)    $\mathcal{O}(q^2n')$
            ii. Generate $\mathbf{y}(\mathbf{s})_{us} \,|\, \boldsymbol{\beta}^{(l)}, \boldsymbol{\Sigma}^{(l)}, \mathbf{y}(\mathbf{s})_{os}$ for $\mathbf{s} \in \mathcal{M}$ through (A.1.12)    $\mathcal{O}(q^2n_m)$
```

We estimate the hyper-parameter set $\{\psi, \alpha\}$ through Algorithm 3.3, where in step 1 we use $\mathcal{S}_{-k}$ to denote the location of $\mathcal{R}$ without $\mathcal{S}_k$.

Bibliography

M. Abramowitz and A. Stegun. *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables.* Dover, 1965.

Ethan Anderes. On the consistent separation of scale and variance for Gaussian random fields. *The Annals of Statistics*, 38(2):870–893, 2010.

Sudipto Banerjee. On geodetic distance computations in spatial modeling. *Biometrics*, 61 (2):617–625, 2005.

Sudipto Banerjee. High-Dimensional Bayesian Geostatistics. *Bayesian Analysis*, 12:583–614, 2017.

Sudipto Banerjee and Anindya Roy. *Linear algebra and matrix analysis for statistics.* CRC Press, Boca Raton, FL, 2014.

Sudipto Banerjee, Bradley P Carlin, and Alan E Gelfand. *Hierarchical modeling and analysis for spatial data.* CRC Press, Boca Raton, FL, 2014.

Douglas Bates and Dirk Eddelbuettel. Fast and elegant numerical linear algebra using the RcppEigen package. *Journal of Statistical Software*, 52(5):1–24, 2013. URL `http://www.jstatsoft.org/v52/i05/`.

Mikhail Belkin. Approximation beats concentration? An approximation view on inference with smooth radial kernels. In *Conference On Learning Theory, COLT 2018*, pages 1348–1361, 2018.

Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral B Shah. Julia: A fresh approach to numerical computing. *SIAM review*, 59(1):65–98, 2017. URL `https://doi.org/10.1137/141000671`.

Christopher Bishop. *Pattern Recognition and Machine Learning*. Springer-Verlag, New York, NY, 2006.

Gilles Bourgault and Denis Marcotte. Multivariable variogram and its application to the linear model of coregionalization. *Mathematical Geology*, 23(7):899–928, 1991.

Philip J Brown, Nhu D Le, and James V Zidek. Multivariate spatial interpolation and exposure to air pollutants. *Canadian Journal of Statistics*, 22(4):489–509, 1994.

Huann-Sheng Chen, Douglas G. Simpson, and Zhiliang Ying. Infill asymptotics for a stochastic process model with measurement error. *Statistica Sinica*, pages 141–156, 2000.

Jean-Paul Chiles and Pierre Delfiner. *Geostatistics: modeling spatial uncertainty*, volume 497. John Wiley & Sons, 2009.

Noel A. C. Cressie and Christopher K. Wikle. *Statistics for Spatio-temporal Data*. Wiley series in probability and statistics. Hoboken, N.J. Wiley, 2011. ISBN 978-0-471-69274-4. URL `http://opac.inria.fr/record=b1133266`.

A. Datta, S. Banerjee, A. O. Finley, N. A. S. Hamm, and M. Schaap. Non-separable Dynamic Nearest-Neighbor Gaussian Process Models for Large spatio-temporal Data With an Application to Particulate Matter Analysis. *Annals of Applied Statistics*, 10:1286–1316, 2016. URL `http://dx.doi.org/10.1214/16-AOAS931`.

A. Datta, S. Banerjee, A. O. Finley, and A. E. Gelfand. Hierarchical Nearest-Neighbor Gaussian Process Models for Large Geostatistical Datasets. *Journal of the American Sta-*

tistical Association, 111:800–812, 2016a. URL http://dx.doi.org/10.1080/01621459.2015.1044091.

A. Datta, S. Banerjee, A. O. Finley, N. A. S. Hamm, and M. Schaap. Non-separable Dynamic Nearest-Neighbor Gaussian Process Models for Large spatio-temporal Data With an Application to Particulate Matter Analysis. *Annals of Applied Statistics*, 10:1286–1316, 2016b. URL http://dx.doi.org/10.1214/16-AOAS931.

Shanshan Ding and R Dennis Cook. Dimension folding PCA and PFC for matrix-valued predictors. *Statistica Sinica*, 24(1):463–492, 2014.

J. Du, H. Zhang, and V. S. Mandrekar. Fixed-domain Asymptotic Properties of Tapered Maximum Likelihood Estimators. *Annals of Statistics*, 37:3330–3361, 2009.

Friedhelm Eicker. Asymptotic normality and consistency of the least squares estimators for families of linear regressions. *The Annals of Mathematical Statistics*, pages 447–456, 1963.

N. Etemadi. Convergence of weighted averages of random variables revisited. *Proceedings of the American Mathematical Society*, 134(9):2739–2744, 2006.

Andrew Finley, Abhirup Datta, and Sudipto Banerjee. *spNNGP: Spatial Regression Models for Large Datasets using Nearest Neighbor Gaussian Processes*, 2017. URL https://CRAN.R-project.org/package=spNNGP. R package version 0.1.1.

Andrew O Finley, Sudipto Banerjee, and Bradley P Carlin. spBayes: an R package for univariate and multivariate hierarchical point-referenced spatial models. *Journal of statistical software*, 19(4):1, 2007.

Andrew O Finley, Abhirup Datta, Bruce C Cook, Douglas C Morton, Hans E Andersen, and Sudipto Banerjee. Efficient algorithms for Bayesian Nearest Neighbor Gaussian Processes. *Journal of Computational and Graphical Statistics*, 28(2):401–414, 2019.

Roger Fletcher. *Practical methods of optimization.* John Wiley & Sons, 2013.

David Chin-Lung Fong and Michael Saunders. LSMR: An iterative algorithm for sparse least-squares problems. *SIAM Journal on Scientific Computing*, 33(5):2950–2971, 2011.

R. Furrer, M. G. Genton, and D. Nychka. Covariance Tapering for Interpolation of Large Spatial Datasets. *Journal of Computational and Graphical Statistics*, 15:503–523, 2006.

Dani Gamerman and Ajax RB Moreira. Multivariate spatial regression models. *Journal of multivariate analysis*, 91(2):262–281, 2004.

Alan Gelfand, Alexandra Schmidt, Sudipto Banerjee, and Sirmans C. Nonstationary multivariate process modeling through spatially varying coregionalization. *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research*, 13(2):263–312, 2004. URL `http://EconPapers.repec.org/RePEc:spr:testjl:v:13:y:2004:i:2:p:263-312`.

Alan E Gelfand, Peter Diggle, Peter Guttorp, and Montserrat Fuentes. *Handbook of spatial statistics*. CRC press, Boca Raton, FL, 2010.

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis, 3rd Edition*. Chapman & Hall/CRC Texts in Statistical Science. Chapman & Hall/CRC, 2013.

Marc G Genton and William Kleiber. Cross-covariance functions for multivariate geostatistics. *Statistical Science*, pages 147–163, 2015.

Tilmann Gneiting and Adrian E Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

Gene H. Golub and Charles F. Van Loan. *Matrix Computations, 4th Edition*. Johns Hopkins University Press, 2012.

J. Guinness. Permutation Methods for Sharpening Gaussian Process Approximations. *arXiv preprint arXiv:1609.05372*, 2016.

Heikki Haario, Eero Saksman, and Johanna Tamminen. Componentwise adaptation for high dimensional MCMC. *Computational Statistics*, 20(2):265–273, 2005.

Matthew J Heaton, Abhirup Datta, Andrew O Finley, Reinhard Furrer, Joseph Guinness, Rajarshi Guhaniyogi, Florian Gerber, Robert B Gramacy, Dorit Hammerling, Matthias Katzfuss, et al. A case study competition among methods for analyzing large spatial data. *Journal of Agricultural, Biological and Environmental Statistics*, 24(3):398–425, 2019.

M.J. Heaton, A. Datta, A.O. Finley, R. Furrer, R. Guhaniyogi, F. Gerber, D. Hammerling, M. Katzfuss, F. Lindgren, D. Nychka, and A. Zammit-Mangion. Methods for Analyzing Large Spatial Data: A Review and Comparison. *arXiv:1710.05013*, 2017. URL `https://arxiv.org/abs/1710.05013`.

J. W. Jerome. Asymptotic estimates of the n-widths in Hilbert space. *Proceedings of the American Mathematical Society*, 33(2):367–372, 1972.

Matthias Katzfuss. Bayesian nonstationary modeling for very large spatial datasets. *Environmetrics*, 24:189–200, 2013.

Matthias Katzfuss. A multi-resolution approximation for massive spatial datasets. *Journal of the American Statistical Association*, 112:201–214, 2017. doi: 10.1080/01621459.2015.1123632. URL `http://dx.doi.org/10.1080/01621459.2015.1123632`.

Matthias Katzfuss and Joseph Guinness. A General Framework for Vecchia Approximations of Gaussian Processes. *arXiv preprint arXiv:1708.06302*, 2017.

C. G. Kaufman and B. A. Shaby. The role of the range parameter for estimation and prediction in geostatistics. *Biometrika*, 100(2):473–484, 2013.

S. L. Lauritzen. *Graphical Models*, chapter Chapter 2, 3, 5. Clarendon Press, Oxford, United Kingdom, 1996.

Nhu Le, Li Sun, and James V Zidek. Spatial prediction and temporal backcasting for environmental fields having monotone data patterns. *Canadian Journal of Statistics*, 29 (4):529–554, 2001.

Nhu D Le and James V Zidek. *Statistical analysis of environmental space-time processes.* Springer Science & Business Media, 2006.

Nhu D Le, Weimin Sun, and James V Zidek. Bayesian multivariate spatial interpolation with data missing by design. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(2):501–510, 1997.

Finn Lindgren, Havard Rue, and Johan Lindstrom. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73 (4):423–498, 2011. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2011.00777.x. URL `http://dx.doi.org/10.1111/j.1467-9868.2011.00777.x`.

Hedibert Freitas Lopes, Esther Salazar, and Dani Gamerman. Spatial Dynamic Factor Analysis. *Bayesian Analysis*, 3(4):759 – 792, 2008.

Albert W Marshall and Ingram Olkin. Matrix versions of the Cauchy and Kantorovich inequalities. *Aequationes Mathematicae*, 40(1):89–93, 1990.

B. Matérn. *Spatial Variation.* Springer-Verlag, 1986.

Qiaozhen Mu, Maosheng Zhao, and Steven W Running. MODIS global terrestrial evapotranspiration (ET) product (NASA MOD16A2/A3). *Algorithm Theoretical Basis Document, Collection*, 5, 2013.

K.P. Murphy. *Machine Learning: A probabilistic perspective.* The MIT Press, Cambridge, MA, 2012.

Akihiko Nishimura and Marc A Suchard. Prior-preconditioned conjugate gradient method for accelerated Gibbs sampling in "large n & large p" sparse Bayesian regression. *arXiv preprint arXiv:1810.12437*, 2018.

D. Nychka, C. Wikle, and J. A. Royle. Multiresolution models for nonstationary spatial covariance functions. *Statistical Modelling*, 2(4):315–331, 2002.

Douglas Nychka, Soutir Bandyopadhyay, Dorit Hammerling, Finn Lindgren, and Stephan Sain. A Multiresolution Gaussian Process Model for the Analysis of Large Spatial Datasets. *Journal of Computational and Graphical Statistics*, 24(2):579–599, 2015. doi: 10.1080/10618600.2014.914946. URL `http://dx.doi.org/10.1080/10618600.2014.914946`.

R Ramon Solano, K Didan, A Jacobson, and A Huete. Modis Vegetation Index User's Guide. *The University of Arizona: Tucson, AZ, USA*, 2010.

Q. Ren and S. Banerjee. Hierarchical factor models for large spatially misaligned datasets: A low-rank predictive process approach. *Biometrics*, 69:19–30, 2013.

Paulo J. Ribeiro Jr and Peter J. Diggle. *geoR: a package for geostatistical analysis*, June 2012. URL `https://cran.r-project.org/web/packages/geoR`. R package version 1.7-4.

Gareth O Roberts and Jeffrey S Rosenthal. Examples of adaptive MCMC. *Journal of Computational and Graphical Statistics*, 18(2):349–367, 2009.

Havard Rue and Leonard Held. *Gaussian Markov Random Fields: Theory and Applications*. Monographs on statistics and applied probability. Chapman and Hall/CRC Press, Boca Raton, FL, 2005. ISBN 1-584-88432-0. URL `http://opac.inria.fr/record=b1119989`.

Havard Rue, Sara Martino, and Nicolas Chopin. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71(2):319–392, 2009. ISSN 1467-9868. doi: 10.1111/j.1467-9868.2008.00700.x. URL `http://dx.doi.org/10.1111/j.1467-9868.2008.00700.x`.

Mary Lai O Salvaña and Marc G Genton. Nonstationary cross-covariance functions for multivariate spatio-temporal random fields. *Spatial Statistics*, page 100411, 2020.

H. Sang and J. Z. Huang. A Full Scale Approximation of Covariance Functions for Large Spatial Data Sets. *Journal of the Royal Statistical society, Series B*, 74:111–132, 2012.

Gabriele Santin and Robert Schaback. Approximation of eigenfunctions in kernel-based spaces. *Advances in Computational Mathematics*, 42(4):973–993, 2016.

Robert Schaback. Error estimates and condition numbers for radial basis function interpolation. *Advances in Computational Mathematics*, 3(3):251–264, 1995.

O. Schabenberger and C. A. Gotway. *Statistical Methods for Spatial Data Analysis.* Chapman and Hall/CRC Press, Boca Raton, FL, first edition, 2004.

Alexandra M Schmidt and Alan E Gelfand. A Bayesian coregionalization approach for multivariate pollutant data. *Journal of Geophysical Research: Atmospheres*, 108(D24), 2003.

Stan Development Team. RStan: the R interface to Stan, 2016. URL `http://mc-stan.org/`. R package version 2.14.1.

M. L. Stein. A simple condition for asymptotic optimality of linear predictions of random fields. *Statistics & Probability Letters*, 17(5):399–404, 1993.

M. L. Stein. *Interpolation of Spatial Data: Some Theory for Kriging.* Springer, first edition, 1999.

M. L. Stein, Z. Chi, and L. J. Welty. Approximating Likelihoods for Large Spatial Data Sets. *Journal of the Royal Statistical society, Series B*, 66:275–296, 2004.

Michael L Stein. Asymptotically efficient prediction of a random field with a misspecified covariance function. *The Annals of Statistics*, pages 55–63, 1988.

J.R. Stroud, M. L. Stein, and S. Lysen. Bayesian and Maximum Likelihood Estimation for Gaussian Processes on an Incomplete Lattice. *Journal of Computational and Graphical Statistics*, 26:108–120, 2017. URL `http://dx.doi.org/10.1080/10618600.2016.1152970`.

Damien Sulla-Menashe and Mark A Friedl. User guide to collection 6 MODIS land cover (MCD12Q1 and MCD12C1) product. *USGS: Reston, VA, USA*, pages 1–18, 2018.

Weimin Sun, Nhu D Le, James V Zidek, and Rick Burnett. Assessment of a Bayesian multivariate interpolation approach for health impact studies. *Environmetrics: The official journal of the International Environmetrics Society*, 9(5):565–586, 1998.

Y. Sun, B. Li, and M.G. Genton. Geostatistics for large datasets. In J.M. Montero, E. Porcu, and M. Schlather, editors, *Advances And Challenges In Space-time Modelling Of Natural Events*, pages 55–77. Berlin Heidelberg: Springer-Verlag, 2011.

Daniel Taylor-Rodriguez, Andrew O Finley, Abhirup Datta, Chad Babcock, Hans Erik Andersen, Bruce D Cook, Douglas C Morton, and Sudipto Banerjee. Spatial factor models for high-dimensional and large spatial data: An application in forest variable mapping. *Statistica Sinica*, 29(3):1155–1180, 2019.

Robert L Thorndike. Who belongs in the family. In *Psychometrika*. Citeseer, 1953.

A. V. Vecchia. Estimation and Model Identification for Continuous Spatial Processes. *Journal of the Royal Statistical society, Series B*, 50:297–312, 1988.

Hans Wackernagel. *Multivariate Geostatistics*. Springer-Verlag, Berlin, 3 edition, 2003.

Daqing Wang, Wei-Liem Loh, et al. On fixed-domain asymptotics and covariance tapering in Gaussian random field models. *Electronic Journal of Statistics*, 5:238–269, 2011.

SJ Yakowitz and F Szidarovszky. A comparison of kriging with nonparametric regression methods. *Journal of Multivariate Analysis*, 16(1):21–53, 1985.

Ozgür Yeniay and Atill Goktas. A comparison of partial least squares regression with other prediction methods. *Hacettepe Journal of Mathematics and Statistics*, 31(99):99–101, 2002.

Zhiliang Ying. Asymptotic properties of a maximum likelihood estimator with data from a Gaussian process. *Journal of Multivariate Analysis*, 36(2):280–296, 1991.

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

Hao Zhang. Asymptotics and computation for spatial statistics. In *Advances and Challenges in Space-time Modelling of Natural Events*, pages 239–252. Springer, 2012.

Hao Zhang and Dale L Zimmerman. Towards reconciling two asymptotic frameworks in spatial statistics. *Biometrika*, 92(4):921–936, 2005.

Lu Zhang, Abhirup Datta, and Sudipto Banerjee. Practical Bayesian Modeling and Inference for Massive Spatial Datasets On Modest Computing Environments. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 12(3):197–209, 2019. doi: 10.1002/sam.11413. URL https://doi.org/10.1002/sam.11413.