On Bayesian Hierarchical Modelling for Large Spatial Datasets

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

RAJARSHI GUHANIYOGI

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
Doctor of Philosophy

Sudipto Banerjee, Ph.D.

March, 2012

© RAJARSHI GUHANIYOGI 2012
ALL RIGHTS RESERVED

Acknowledgements

I would like to acknowledge Andrew O. Finley in the Michigan State University for his constant help and insightful inputs. I would also acknowledge Dr. James S. Hodges, Dr. Cavan Reilly and Dr. Dennis Cook for their comments which have improved the quality of work presented here.

Dedication

This thesis has been dedicated to my dearest wife Sharmistha Guha, my parents and my beloved adviser Dr. Sudipto Banerjee.

Abstract

We propose a class of fully process-based low-rank spatially-varying cross-covariance matrices that produce non-degenerate spatial processes and that effectively capture non-stationary covariances among the multiple outcomes. We provide theoretical and modeling insight into these constructions and elucidate certain implications of some common structural assumptions in building cross-covariance matrices. We also propose low rank version of cross-covariance functions using predictive process class of models, popularly employed in spatial statistics to handle large datasets. Predictive process is obtained by projecting the parent Gaussian process onto a space spanned by a set of basis functions. An efficient model to choose those basis functions and have been proposed. Being a low rank model, predictive process often loses spatial information which might lead to spurious inferences. In the Chapter of this thesis, this loss of information has been quantified and model based adjustments have been suggested. Proposed models have been validated with carefully designed simulation studies. Finally, they have been employed to analyze interesting ecological datasets. Our framework has been able to produce substantive inferential tools such as maps of non-stationary cross-covariances that constitute the premise of further mechanistic modeling and hitherto not been easily available for environmental scientists and ecologists.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Spatial data analysis

The growing popularity of Geographical Information Systems (GIS) has generated much interest in analyzing and modeling geographically referenced data. Geographical referencing depends upon the resolution of the data: when data referencing is done with respect to the coordinates of the location (e.g. latitude and longitude), we call them *point-referenced*, while data aggregated over regions in a map (e.g. mortality rates by counties or zip-codes) are called *areally-referenced* or *lattice*. In the domain of public health, due to patient confidentiality, data are usually of the latter type and are usually available as case counts or rates referenced to *areal* regions, such as counties, census-tracts or ZIP codes. On the other hand, in environmental and ecological studies researchers mostly encounter *point-referenced* spatial data.

Statistical models for spatial data are primarily concerned with explaining variation, separating spatial signals from noise and improving estimation and prediction. These models capture associations or correlations across space depending upon the type of referencing in the data. For point-referenced datasets, models customarily employ spatial processes to capture spatial associations as a function of Euclidean geometric objects such as distance and direction. These models are popular in geostatistics (see, e.g., Cressie, 1993; Banerjee et al., 2004) and provide spatial interpolation or "kriging" accounting for uncertainty in estimation and prediction.

1.2 Low Dimensional Spatial Modeling

Bayesian hierarchical models, using spatial processes, are widely recognized as versatile inferential tools for capturing the rich dependence structures underlying spatial data and for offering full inference without resorting to potentially inappropriate asymptotic paradigms. Hierarchical spatial process models are typically estimated using Markov chain Monte Carlo methods (Banerjee et al. 2004) and entail expensive matrix decompositions in every iteration of the MCMC algorithm. However, when the number of locations is huge, as mostly encountered in environmental or ecological studies, this procedure becomes infeasible. Evidently, multivariate and spatiotemporal data exacerbate the problem.

Modeling large spatial datasets has received much attention in the recent past. One approach seeks approximations to the model, or likelihood, with more computationally tractable forms. These approximations arise either as products of conditional distributions (Vecchia, 1988; Stein et al. 2004), or as spectral representations (Fuentes, 2007) or perhaps using Gaussian Markov random fields (Rue, Martino and Chopin, 2009) that lead to an `INLA` (Integrated Nested Laplace Approximation) algorithm. These approximations, however, often fail to retain the richness of the underlying process and lose their computational advantages for all but the simplest of spatial processes. Yet another approach considers compactly supported correlation functions (Furrer et al., 2006; Kaufman et al., 2009) that yield sparse correlation structures. This approach can be conveniently applied to isotropic processes, but may be less effective in capturing nonstationarity.

Arguably, a more versatile approach would pursue models especially geared towards the handling of large spatial datasets. Typically these emerge from representations of the spatial process in lower-dimensional subspaces and can easily be adapted to multivariate and/or spatiotemporal processes. These are often referred to as low-rank or reduced-rank spatial models and have been explored in different contexts (Higdon 2002; Stein, 2007, 2008; Cressie and Johannesson, 2008; Banerjee et al., 2008; Crainiceaniu et al., 2008). Many of these methods are variants of the so-called "subset of regressors" methods used in Gaussian process regressions for large data sets in machine learning (e.g. Rasmussen and Williams, 2006). The idea here is to consider a smaller set of

locations, or "knots", say $\mathcal{S}^* = \{\boldsymbol{s}_1^*, \ldots, \boldsymbol{s}_{n^*}^*\}$, where n^* is fixed to be much smaller than the number of observed sites, and to express the spatial process realizations over $\mathcal{S}$ in terms of its realizations over the smaller set of knots. It is assumed that there will be insignificant loss of spatial information in the underlying process as a result of using a smaller set of locations (the knots) that adequately covers the domain. Throughout this thesis, we consider a special class of low-rank processes called the predictive process (Banerjee et al., 2008) which arises as a conditional expectation of the original process given its realization at the knots. Although the predictive process solves the computational problem posed by large spatial datasets, it often raises different issues regarding its usage. The thesis addresses three major problems. The first is the selection of knots for the predictive process. The second is modeling low-rank cross covariances for multivariate spatial processes. The third chapter is dedicated to the exploration of tapered residual processes.

1.2.1 Adaptive Gaussian Predictive Process Models for Large Spatial Datasets

Low-rank models assume a judicious choice of *fixed* knots. A key issue in low-rank methods has always been the choice of knots, which is usually dictated by computational cost and sensitivity to choice. In practice, we often investigate sensitivity of inference to different choices of n^*, which entails separately estimating a number of low-rank models. Typically, for each n^* we use some space-covering design (e.g., Royle and Nychka, 1998) to fix the knots.

In chapter 2, our modeling framework expands existing hierarchical low-rank models, as explored in the aforementioned references, to accommodate modeling knot locations as random point patterns. In particular, we specify the knots as realizations of a log Gaussian process with a fixed number of points. We achieve this by including an additional specification in our overall spatial process modeling, allowing the observations to inform about a *good* set of locations. While our formulation applies to any low rank likelihood, we specifically work with *predictive process* models. We propose a framework that will accommodate knot selection within the same hierarchical model as the predictive process likelihood and explore what benefits, if any, such stochastic modeling of the knots will fetch. Details of knot selection can be found in Chapter 2.

1.2.2 Modeling low-rank spatially-varying cross-covariances using predictive process with application to soil nutrient data

Another computationally challenging problem arises in modeling spatial correlations within a location and across different locations among variables in large multivariate spatial data sets. For example, in ecological datasets these could be measures of species abundance, vegetation characteristics, or pollutants at each inventory or monitoring location. Given these data, ecologists are often interested in making inferences about the correlation of these biotic and/or abiotic variables within a location, and, how this within-location correlation changes across the domain. These non-stationary patterns can often reveal unmeasured covariates, deepen understanding of ecological processes, and improve prediction of the multivariate vector of outcomes at new locations within the domain. However, Bayesian computation for multivariate data involves matrix computation of high order rendering computation infeasible. To overcome this computational bottleneck, I have employed a multivariate version of the predictive process with spatially varying correlation structures. I have explored the proposed models using multivariate synthetic data and soil nutrient data, collected at La Selva Biological Station, Costa Rica, where I estimated correlations among nutrients across space.

1.2.3 On the residual spatial process from multivariate hierarchical low rank models

Finally, Chapter 4 deals with some inferential issues for the predictive process. Stein (2008) and Banerjee et al. (2008) report potential problems in prediction and inference arising from the smoothness of low-rank models. More specifically, Finley et al. (2009) demonstrate how a particular low-rank spatial process, called the predictive process, yields biased estimates of certain variance components in spatial progeny trials and offer one specific adjustment to avoid spurious inference.

Chapter 4 embarks upon characterizing and understanding biases, discusses their potential impact on spatial inference and explores remedies applicable to a wide range of hierarchical low-rank spatial process models. We show precisely how such biases arise in *any* low-rank likelihood, not just predictive processes, and impact the *smoothness* of the spatial surface. The term *spatial smoothness* has been used quite often in spatial

statistics, but has never been studied carefully. Of late, interest has been shifted to formalize the idea of spatial smoothness. Often, in spatial statistics, the main focus resides on prediction or interpolation for new points based on a finite realization. In the case of prediction when there are surrounding observations, the local behavior of the random field becomes very crucial. This local behavior of the random field is characterized by *smoothness*. For example, when abrupt changes in the realizations is observed, one would naturally think of a spatial surface which is not quite smooth. On the other hand, for a relatively even realization, a smooth spatial surface would be anticipated. This observation points out, quite expectedly, the similarity between smoothness and existence of directional derivatives of a spatial process.

Note that inference about the smoothness or the directional derivatives of a process cannot be drawn from a realization. Rather, smoothness of the spatial process has to be investigated from the model or the source of the data. Such investigations have been carried out by Banerjee et. al. (2003), who proposed a fully model-based formalization of spatial smoothness with the idea of mean square directional derivatives. Their idea offers an elegant extension of the notion of smoothness in univariate stationary processes (see Stein, 1999) to multivariate stationary and nonstationary processes. Later, this concept is used to assess gradients to curves in identifying curves that track a path through a region where the spatial surface is rapidly changing. Such boundaries are generally referred to as *wombling boundaries* and are found to be extremely helpful for decision makers in public health. For a detailed overview of the *wombling* literature, see Banerjee et. al. (2006) and references therein.

Although low rank models have been widely employed for spatial model fitting and prediction, study of their smoothness properties is less explored. It is typically observed from different simulation studies and real data analyses that low rank models oversmooth spatial surfaces. In this article, we develop theoretical results on spatial smoothness for different low rank knot based models, which have not been discussed hitherto. We also introduce a class of low rank knot based models which is able to rectify bias as in the estimation of variance components and simultaneously restricts the surface from being oversmoothed.

Chapter 2

Adaptive Gaussian Predictive Process Models for Large Spatial Datasets

2.1 The Gaussian Predictive Process

Geostatistical modeling settings typically assume, at locations $\boldsymbol{s} \in D \subseteq \Re^2$, a response variable $Y(\boldsymbol{s})$ and a $p \times 1$ vector of spatially referenced predictors $\boldsymbol{x}(\boldsymbol{s})$, associated through a spatial regression model $\mathrm{E}[Y(\boldsymbol{s}) \,|\, \boldsymbol{x}(\boldsymbol{s}), \boldsymbol{\beta}, w(\boldsymbol{s})] = \boldsymbol{x}(\boldsymbol{s})'\boldsymbol{\beta} + w(\boldsymbol{s})$. This includes a spatial process over the study region D, defined by the set $w_D = \{w(\boldsymbol{s}) : \boldsymbol{s} \in D\}$, viewed as a randomly realized surface over the region.

The $w(\boldsymbol{s})$ provides local adjustment (with structured dependence) to the mean, capturing the effect of unmeasured or unobserved covariates with a spatial pattern. In practice, the surface $Y(\boldsymbol{s})$ is only observed at a finite set of locations, say $\mathcal{S} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$. For any such finite set, $w(\mathcal{S}) = \{w(\boldsymbol{s}_i) : \boldsymbol{s}_i \in \mathcal{S}\}$ is a (partial) realization of $w(\cdot)$ over $\mathcal{S}$ and $[w_D \,|\, \mathcal{S}]$ denotes the resulting joint distribution of $w(\mathcal{S})$. The customary process specification for $w(\boldsymbol{s})$ is a zero-centered Gaussian Process with a parametric covariance function $C(\boldsymbol{s}_1, \boldsymbol{s}_2; \boldsymbol{\theta})$, denoted by $w(\boldsymbol{s}) \sim GP(0, C(\boldsymbol{s}_1, \boldsymbol{s}_2; \boldsymbol{\theta}))$, so that $[w(\mathcal{S}) \,|\, \boldsymbol{\theta}] = N(\mathbf{0}, \boldsymbol{C}(\mathcal{S}; \boldsymbol{\theta}))$ Here $\boldsymbol{C}(\mathcal{S}; \boldsymbol{\theta})$ is the $n \times n$ matrix whose (i, j)-th element is given by $\mathrm{cov}\{w(\boldsymbol{s}_i), w(\boldsymbol{s}_j)\} = C(\boldsymbol{s}_i, \boldsymbol{s}_j; \boldsymbol{\theta})$. Often we specify $C(\boldsymbol{s}_i, \boldsymbol{s}_j; \boldsymbol{\theta}) = \sigma^2 \rho(\boldsymbol{s}_i, \boldsymbol{s}_j; \boldsymbol{\phi})$,

where $\boldsymbol{\theta} = \{\sigma^2, \boldsymbol{\phi}\}$, σ^2 being a spatial variance component and $\rho(\cdot;\boldsymbol{\phi})$ a spatial correlation function.

For the vector of observed outcomes, $\boldsymbol{y} = (y(\boldsymbol{s}_1), y(\boldsymbol{s}_2), \ldots, y(\boldsymbol{s}_n))'$, with a conditionally independent Gaussian likelihood and associated priors, a hierarchical model arises with posterior distribution

$$[\boldsymbol{\beta}, \tau^2, w(\mathcal{S}), \boldsymbol{\theta} \,|\, \boldsymbol{y}, \mathcal{S}] \propto [\boldsymbol{\theta}] \times [\tau^2 \,|\, a_\tau, b_\tau] \times [\boldsymbol{\beta}] \\ \times N\left(w(\mathcal{S}) \,|\, \boldsymbol{0}, \boldsymbol{C}(\mathcal{S};\boldsymbol{\theta})\right) \times \prod_{i=1}^{n} N(y(\boldsymbol{s}_i) \,|\, \boldsymbol{x}(\boldsymbol{s}_i)'\boldsymbol{\beta} + w(\boldsymbol{s}_i), \tau^2) \,. \tag{2.1}$$

In the sequel we take $[\boldsymbol{\beta}] = N(\boldsymbol{\beta} \,|\, \boldsymbol{\mu}_\beta, \Sigma_\beta)$. Fitting of (2.1) customarily proceeds using an MCMC algorithm that generates samples from the posterior distribution.

Recently Banerjee et al. (2008) explored a class of reduced-rank spatial process models for large spatial datasets using a *fixed* set of "knots" $\mathcal{S}^* = (\boldsymbol{s}_1^*, \ldots, \boldsymbol{s}_{n^*}^*)$ with $n^* << n$, which may, but need not, be a subset of the entire collection of observed locations in $\mathcal{S}$. An optimal projection of the process $w(\boldsymbol{s})$ at a generic location $\boldsymbol{s}$, based upon its realization over $\mathcal{S}^*$ is given by the "kriging equation" $\tilde{w}(\boldsymbol{s}) = E\{w(\boldsymbol{s}) \,|\, w(\mathcal{S}^*)\}$, where $w(\mathcal{S}^*) = \{w(\boldsymbol{s}_i^*) : \boldsymbol{s}_i^* \in \mathcal{S}^*\}$. We refer to $\tilde{w}(\boldsymbol{s})$ as the *predictive process* derived from the *parent process* $w(\boldsymbol{s})$. When the parent process is a zero-centered Gaussian process with covariance function $C(\boldsymbol{s}_1, \boldsymbol{s}_2; \boldsymbol{\theta})$, we can write the predictive process as $\tilde{w}(\boldsymbol{s}) = E\{w(\boldsymbol{s}) \,|\, w(\mathcal{S}^*)\} = \boldsymbol{z}(\boldsymbol{s}, \mathcal{S}^*; \boldsymbol{\theta})' w(\mathcal{S}^*)$, where $\boldsymbol{c}(\boldsymbol{s}, \mathcal{S}^*; \boldsymbol{\theta})'$ is the $1 \times n$ vector whose j-th element is $C(\boldsymbol{s}, \boldsymbol{s}_j^*; \boldsymbol{\theta})$ and $\boldsymbol{z}(\boldsymbol{s}, \mathcal{S}^*; \boldsymbol{\theta})' = \boldsymbol{c}(\boldsymbol{s}, \mathcal{S}^*; \boldsymbol{\theta})' \boldsymbol{C}(\mathcal{S}^*; \boldsymbol{\theta})^{-1}$. Since $w(\mathcal{S}^*)$ follows a multivariate normal law with zero mean and $n^* \times n^*$ variance-covariance matrix $\boldsymbol{C}(\mathcal{S}^*; \boldsymbol{\theta})$, the predictive process is itself a nonstationary Gaussian process arising from a spatially adaptive linear transformation of the parent process over the set of knots. The elements of $\boldsymbol{z}(\boldsymbol{s}, \mathcal{S}^*; \boldsymbol{\theta})'$ comprise the coefficients of the linear transformation. Replacing $w(\boldsymbol{s})$ with $\tilde{w}(\boldsymbol{s})$ in (2.1), leads to its predictive process counterpart

$$[\boldsymbol{\beta}, \tau^2, w(\mathcal{S}^*), \boldsymbol{\theta} \,|\, \boldsymbol{y}, \mathcal{S}^*, \mathcal{S}] \propto [\boldsymbol{\theta}] \times [\tau^2 \,|\, a_\tau, b_\tau] \times N(\boldsymbol{\beta} \,|\, \boldsymbol{\mu}_\beta, \Sigma_\beta) \\ \times N(w(\mathcal{S}^*) \,|\, \boldsymbol{0}, \boldsymbol{C}(\mathcal{S}^*; \boldsymbol{\theta})) \times \prod_{i=1}^{n} N(y(\boldsymbol{s}_i) \,|\, \boldsymbol{x}(\boldsymbol{s}_i)'\boldsymbol{\beta} + \boldsymbol{z}(\boldsymbol{s}_i, \mathcal{S}^*; \boldsymbol{\theta})' w(\mathcal{S}^*), \tau^2). \tag{2.2}$$

Computational gains are achieved since matrix computations now involve the $n^* \times n^*$ matrix $\boldsymbol{C}(\mathcal{S}^*;\boldsymbol{\theta})$, where $n^* << n$. Unlike some other knot-based methods, the predictive process does not introduce additional parameters nor does it involve projecting data onto a grid. Thus, it avoids identifiability issues or spurious loss of uncertainty. Indeed, predictive process models are attractive since they are directly induced by the parent process without requiring choices of basis functions or kernels or alignment algorithms for the locations.

Rather than an approximation to the parent process, we consider the predictive process as a dimension-reducing model for large point-referenced datasets. Therefore, its parameters should be interpreted with respect to (2.2) and not (2.1). In fact, being smoother than the parent process, the predictive process tends to have lower variance which, in turn, leads to an upward bias in the nugget (see Chapter 4).

A remedy for this bias (Finley et al., 2009; Banerjee et al. 2010) is to use the process $\tilde{w}_\epsilon(\boldsymbol{s}) = \tilde{w}(\boldsymbol{s}) + \tilde{\epsilon}(\boldsymbol{s})$, where $\tilde{\epsilon}(\boldsymbol{s}) \overset{iid}{\sim} N(0, \mathrm{E}\{\mathrm{var}[w(\boldsymbol{s}) \,|\, w(\mathcal{S}^*)]\})$ and $\tilde{\epsilon}(\boldsymbol{s})$ is independent of $\tilde{w}(\boldsymbol{s})$. Now, the variance of $\tilde{w}_\epsilon(\boldsymbol{s})$ equals that of $w(\boldsymbol{s})$. For Gaussian processes, $\mathrm{E}\{\mathrm{var}[w(\boldsymbol{s}) \,|\, w(\mathcal{S}^*)]\} = \{C(\boldsymbol{s},\boldsymbol{s};\boldsymbol{\theta}) - \boldsymbol{c}(\boldsymbol{s},\mathcal{S}^*;\boldsymbol{\theta})'\boldsymbol{C}(\mathcal{S}^*;\boldsymbol{\theta})^{-1}\boldsymbol{c}(\boldsymbol{s},\mathcal{S}^*;\boldsymbol{\theta})\}$. We refer to $\tilde{w}_\epsilon(\boldsymbol{s})$ as the "bias-adjusted" predictive process. Replacing $w(\boldsymbol{s})$ with $\tilde{w}_\epsilon(\boldsymbol{s})$ in (2.1) yields a bias-adjusted predictive process. For Gaussian likelihoods, explicit marginalization over $w(\mathcal{S}^*)$ is possible. This yields the marginalized bias-adjusted predictive process counterpart of (2.2), given by

$$[\boldsymbol{\beta},\boldsymbol{\theta},\tau^2 \,|\, \boldsymbol{y},\mathcal{S}^*] \propto [\boldsymbol{\theta}] \times IG(\tau^2 \,|\, a_\tau, b_\tau) \times N(\boldsymbol{\beta} \,|\, \boldsymbol{\mu}_\beta, \Sigma_\beta) \times N\left(\boldsymbol{y} \,|\, \boldsymbol{X}\boldsymbol{\beta}, \Sigma_{\boldsymbol{y}}\left(\mathcal{S}^*,\boldsymbol{\theta},\tau^2\right)\right), \tag{2.3}$$

where $\Sigma_{\boldsymbol{y}}\left(\mathcal{S}^*,\boldsymbol{\theta},\tau^2\right) = \mathcal{C}(\mathcal{S}^*;\boldsymbol{\theta})'\boldsymbol{C}(\mathcal{S}^*;\boldsymbol{\theta})^{-1}\mathcal{C}(\mathcal{S}^*;\boldsymbol{\theta}) + \boldsymbol{D}_{\tilde{\epsilon}+\epsilon}$, $\mathcal{C}(\mathcal{S}^*;\boldsymbol{\theta})'$ is the $n \times n^*$ matrix with i-th row given by $\boldsymbol{c}(\boldsymbol{s}_i,\mathcal{S}^*;\boldsymbol{\theta})'$ and $\boldsymbol{D}_{\tilde{\epsilon}+\epsilon}$ is an $n \times n$ diagonal matrix whose i-th diagonal element is given by $\{C(\boldsymbol{s}_i,\boldsymbol{s}_i) - \boldsymbol{c}(\boldsymbol{s}_i,\mathcal{S}^*;\boldsymbol{\theta})'\boldsymbol{C}(\mathcal{S}^*;\boldsymbol{\theta})^{-1}\boldsymbol{c}(\boldsymbol{s}_i,\mathcal{S}^*;\boldsymbol{\theta})\} + \tau^2$. The dispersion matrix of $\boldsymbol{y}$ in the above model is $n \times n$, but computational benefits accrue by employing the Sherman-Woodbury-Morrison matrix identities (Henderson and Searle, 1981) which express $\Sigma_{\boldsymbol{y}}\left(\mathcal{S}^*,\boldsymbol{\theta},\tau^2\right)^{-1}$ as

$$\boldsymbol{D}_{\tilde{\epsilon}+\epsilon}^{-1} - \boldsymbol{D}_{\tilde{\epsilon}+\epsilon}^{-1}\mathcal{C}(\mathcal{S}^*;\boldsymbol{\theta})'\left[\boldsymbol{C}(\mathcal{S}^*;\boldsymbol{\theta}) + \mathcal{C}(\mathcal{S}^*;\boldsymbol{\theta})\boldsymbol{D}_{\tilde{\epsilon}+\epsilon}^{-1}\mathcal{C}(\mathcal{S}^*;\boldsymbol{\theta})'\right]^{-1}\mathcal{C}(\mathcal{S}^*;\boldsymbol{\theta})\boldsymbol{D}_{\tilde{\epsilon}+\epsilon}^{-1}$$

$$\det\left(\Sigma_{\boldsymbol{y}}\left(\mathcal{S}^*,\boldsymbol{\theta},\tau^2\right)\right)) = \frac{\det\left(\boldsymbol{D}_{\tilde{\epsilon}+\epsilon}\right)}{\det\left(\boldsymbol{C}(\mathcal{S}^*;\boldsymbol{\theta})\right)} \times \det\left(\boldsymbol{C}(\mathcal{S}^*;\boldsymbol{\theta}) + \mathcal{C}(\mathcal{S}^*;\boldsymbol{\theta})\boldsymbol{D}^{-1}_{\tilde{\epsilon}+\epsilon}\mathcal{C}(\mathcal{S}^*;\boldsymbol{\theta})'\right).$$

These expressions involve inverses and determinants that are either diagonal or $n^* \times n^*$.

2.2 Learning about the Knots

As with any existing low-rank model, knot selection is required and sensitivity to the number of knots is expected. With a fairly even distribution of data locations, one possibility is to select knots on a uniform grid overlaid on the domain. However, in general the locations are highly irregular, generating substantial areas of sparse observations where we wish to avoid placing knots, since they would be "wasted" and possibly lead to inflated predictive process variances and slower convergence. More practical space-covering designs (e.g., Royle and Nychka, 1998) or popular clustering algorithms (e.g., Kaufman and Rousseeuw 1990) can yield a representative collection of knots that better cover the domain. Finley et al. (2009) explore knot selection strategies so that the induced predictive process offers a better approximation to the parent process. They regard the predictive variance of $w(\boldsymbol{s})$, conditional upon the realization of the parent process over $\mathcal{S}^*$, as a measure of how well we approximate $w(\boldsymbol{s})$ with the predictive process $\tilde{w}(\boldsymbol{s})$ and propose a design-based framework to select knots based upon spatially averaged predictive variance

We fix the *number* of knots. This number should be "as large as possible", but is dictated by availability of computational resources and sensitivity to choice of knots. The selection procedure will need to be repeated for different choices of n^* and a final choice is made based on the run time and the stability of inference. Letting the number of knots vary using a reversible jump mcmc algorithm will give rise to convergence issues and the resultant complexities offset the computational gains from the predictive process.

Here we present a simple one-dimensional example to illustrate why estimating the knot locations can be beneficial. The $\circ$ symbols in Figure 2.1(a and b) represent observed values of y, which were drawn from a normal distribution with mean a varying frequency sine function and variance 0.01. Here the + symbol depicts the location of seven knots equally distributed along the x axis.

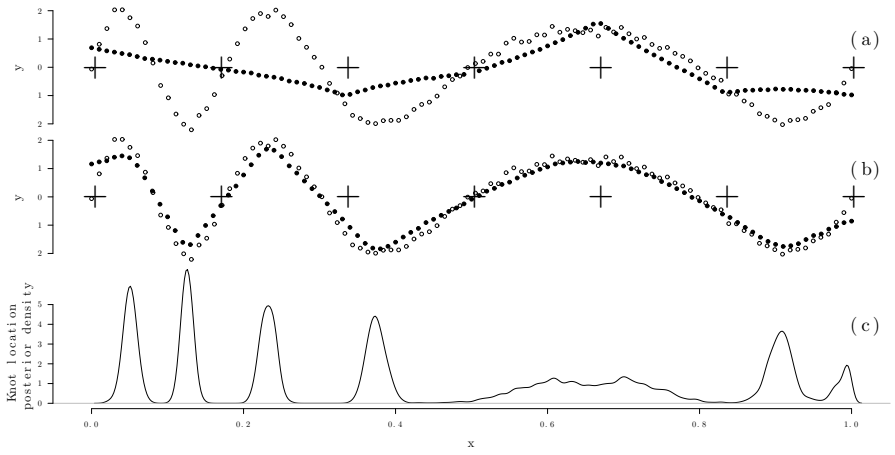


Figure 2.1: Observed data (∘) drawn from a normal distribution with a varying frequency sine function mean and variance 0.01. Knot starting locations (+) and subsequent posterior density of 5,000 knot location MCMC samples illustrated in the lower panel. Posterior predictive means of 100 new locations (•).

Given these data, two models were used to predict the values of y for 100 new x values between 0 and 1. First, we use the model in (2.3) with only an intercept in the regression and the fixed knots shown in Figure 2.1. Second, we let the knots vary by assigning a simple uniform prior $U(0,1)$ for the position of each knot on the x axis. Posterior inference for each model was based on $5,000$ post burn-in Markov chain Monte Carlo (MCMC) samples. The medians for each of the 100 posterior predictive distributions produced using the fixed and adaptive knot models is indicated by the • symbol in (a) and (b), respectively.

Prediction using the fixed knot model is based only on information at the knot locations and, as a result, produces a poor approximation of y, as seen in Figure 2.1(a). In contrast, Figure 2.1(b) shows that by allowing the knot locations to move along the x axis, and learn from the data, predictions from the adaptive model more accurately capture the variability of y. The bottom density plot Figure 2.1(c) illustrates where the adaptive knots were sampled. This plot shows that knots tend to sample at values of x that correspond to the stationary points on the sine curve, resulting in a greatly improved approximation of the original data.

2.2.1 Modeling the knots

With the number of knots fixed according to available computing resources, we allow knot locations to vary across space by modeling $\mathcal{S}^*$. We assume that the density of $\mathcal{S}^*$ is given by

$$[\mathcal{S}^* \,|\, \eta_D, n^*] = \prod_{i=1}^{n^*} \frac{\eta(\boldsymbol{s}_i^*)}{\int_{\mathcal{D}} \eta(\boldsymbol{s}) d\boldsymbol{s}} = \left(\int_{\mathcal{D}} \eta(\boldsymbol{s}) d\boldsymbol{s} \right)^{-n^*} \times \prod_{i=1}^{n^*} \eta(\boldsymbol{s}_i^*) \,, \tag{2.4}$$

$\eta(\boldsymbol{s}) = \exp(\lambda(\boldsymbol{s}))$ is an intensity function, $\eta_{\mathcal{D}} = \{\eta(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{D}\}$ and n^* is the number of knots. The density in (2.4), popularly known as log Gaussian density, emerges from a non-homogeneous Poisson process, conditional upon n^* and $\eta_{\mathcal{D}}$.

We extend (2.3), allowing the knot locations to vary over $\mathcal{D}$, by

$$\begin{aligned}[\boldsymbol{\beta}, \tau^2, \boldsymbol{\theta}_1, \mathcal{S}^*, \boldsymbol{\theta}_2 \,|\, \boldsymbol{y}, \mathcal{S}, n^*] \propto [\boldsymbol{\theta}_1] \times IG(\tau^2 \,|\, a_\tau, b_\tau) \times N(\boldsymbol{\beta} \,|\, \boldsymbol{\mu}_\beta, \Sigma_\beta) \\ \times [\eta_{\mathcal{D}}] \times [\mathcal{S}^* \,|\, \eta_{\mathcal{D}}] \times N\left(\boldsymbol{y} \,|\, \boldsymbol{X}\boldsymbol{\beta}, \Sigma_{\boldsymbol{y}}\left(\mathcal{S}^*; \boldsymbol{\theta}_1, \tau^2\right)\right) \,,\end{aligned} \tag{2.5}$$

where $\boldsymbol{\theta}_1$ now represents the process parameters in the data likelihood. Two practical approaches for modeling $[\eta_{\mathcal{D}}] \times [\mathcal{S}^* \,|\, \eta_{\mathcal{D}}]$ are outlined below.

Modeling $\eta(\boldsymbol{s})$ - a parametric model

Parametric forms can be prescribed for $\eta(\boldsymbol{s})$, such as basis representations or tiled surfaces (see, e.g., Diggle, 2003). Here, we employ a random equally weighted bivariate normal mixture, and then add priors on the parameters in the normal kernel, say $\boldsymbol{\theta}_2$. More specifically, let $\boldsymbol{\theta}_2 = \{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_m, \Sigma_\eta\}$, where the $\boldsymbol{u}_j$'s are m points in $\mathcal{D}$ and Σ_η is a common 2×2 variance covariance matrix. The intensity is $\log\{\eta(\boldsymbol{s}; \boldsymbol{\theta}_2)\} = \dfrac{1}{m} \displaystyle\sum_{j=1}^{m} N_{2\mathcal{D}}(\boldsymbol{s} \,|\, \boldsymbol{u}_j, \Sigma_\eta)$, where $N_{2\mathcal{D}}(\cdot \,|\, \boldsymbol{u}_j, \Sigma_\eta)$ denotes a bivariate normal density, truncated to $\mathcal{D}$, with mean $\boldsymbol{u}_j$ and variance-covariance matrix Σ_η. This parametric kernel specification replaces $[\eta_{\mathcal{D}}] \times [\mathcal{S}^* \,|\, \eta_{\mathcal{D}}]$ with $[\boldsymbol{\theta}_2] \times [\mathcal{S}^* \,|\, \boldsymbol{\theta}_2]$ in (2.5) (suppressing the implicit conditioning on n^*).

Prior specifications for $\boldsymbol{\theta}_2$ typically comprise a uniform support over $\mathcal{D}$ for each of the $\boldsymbol{u}_j$'s and an inverse Wishart $IW(r_\eta, \Omega_\eta)$ (e.g., Gelman et al. 2004) for Σ_η. Alternatively,

we could further parametrize $\Sigma_\eta = \sigma_\eta^2 \begin{pmatrix} 1 & \rho_\eta \\ \rho_\eta & 1 \end{pmatrix}$ and assign appropriate priors to σ_η^2 and ρ_η.

Modeling $\eta(\boldsymbol{s})$ - a log-Gaussian model

Rather than the parametric choice above, we can use a log-Gaussian process $\eta(\boldsymbol{s}) = \exp\{\alpha w_2(\boldsymbol{s})\}$ where $w_2(\boldsymbol{s})$ is a Gaussian process with zero mean, unit variance and correlation function $\rho_2(\cdot;\boldsymbol{\phi}_2)$. In Appendix A it has been shown that under some conditions the parameter α will enjoy posterior propriety. Learning of α, however, is quite poor and there is an issue regarding the choice of prior on α. We, therefore, fix $\alpha = 1$. Even after that, there remains an analytically inaccessible integral of $w_2(\boldsymbol{s})$. Matters are assuaged by a lower-dimensional representation for $w_2(\boldsymbol{s})$. One option, naturally, is the predictive process itself and we replace $w_2(\boldsymbol{s})$ by $\tilde{w}_2(\boldsymbol{s})$.

More specifically, let $\{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_m\}$ be a set of *fixed* knots and let $\tilde{w}_2(\boldsymbol{s}) = \mathrm{E}[w_2(\boldsymbol{s}) \,|\, \boldsymbol{w}_2^*]$ be the corresponding predictive process, where $\boldsymbol{w}_2^* = (w_2(\boldsymbol{u}_1), .., w_2(\boldsymbol{u}_m))'$. The corresponding hierarchical model is still described by (2.5) with $\boldsymbol{\theta}_2 = \{\boldsymbol{\phi}_2, \boldsymbol{w}_2^*\}$ and $[\boldsymbol{\theta}_2] = [\boldsymbol{\phi}_2] \times N_m(\boldsymbol{w}_2^* \,|\, \boldsymbol{0}, \boldsymbol{R}_2(\boldsymbol{\phi}_2))$, where $\boldsymbol{R}_2(\boldsymbol{\phi}_2)$ is the $m \times m$ correlation matrix with $\rho_2(\boldsymbol{u}_i, \boldsymbol{u}_j; \boldsymbol{\phi}_2)$ as the (i,j)-th element. Our experiments show that a modest value of m, usually between 10 and 30, allows adequate exploration of most domains by the knots. For a given size of n^*, increasing m beyond $\sim$10 did not substantially alter the final inference.

2.2.2 Implementation details

The parameters in (2.5) are updated using a combination of Gibbs and Metropolis steps. We first update $\boldsymbol{\beta}$ from $N(\mu_{\boldsymbol{\beta}|\cdot}, \Sigma_{\boldsymbol{\beta}|\cdot})$, with covariance matrix

$$\Sigma_{\boldsymbol{\beta}|\cdot} = (\boldsymbol{X}'\Sigma_{\boldsymbol{y}} \left(\mathcal{S}^*; \boldsymbol{\theta}_1, \tau^2\right)^{-1} \boldsymbol{X} + \Sigma_{\boldsymbol{\beta}}^{-1})^{-1},$$

and mean

$$\mu_{\boldsymbol{\beta}|\cdot} = \left(\boldsymbol{X}'\Sigma_{\boldsymbol{y}} \left(\mathcal{S}^*; \boldsymbol{\theta}_1, \tau^2\right)^{-1} \boldsymbol{X} + \Sigma_{\boldsymbol{\beta}}^{-1}\right)^{-1} (\Sigma_{\boldsymbol{\beta}}^{-1}\mu_{\boldsymbol{\beta}} + \boldsymbol{X}'\Sigma_{\boldsymbol{y}} \left(\mathcal{S}^*; \boldsymbol{\theta}_1, \tau^2\right)^{-1} \boldsymbol{y})$$

The inverse of $\Sigma_{\boldsymbol{y}}\left(\mathcal{S}^*, \boldsymbol{\theta}_1, \tau^2\right)$ is obtained from the Sherman-Woodbury-Morrison formulas.

We update $\{\boldsymbol{\theta}_1, \tau^2\}$, $\{\mathcal{S}^*\}$ and $\{\boldsymbol{\theta}_2\}$ in separate blocks. The step for $\boldsymbol{\theta}_2$ requires some clarification. This involves evaluating $[\boldsymbol{\theta}_2] \times [\mathcal{S}^* \,|\, \boldsymbol{\theta}_2]$, which involves the integral $\int_{\mathcal{D}} \eta(\boldsymbol{s}; \boldsymbol{\theta}_2) d\boldsymbol{s}$ in (2.4). We use a grid-based integration scheme. Letting $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_M \in \mathcal{D}$ be the grid of points covering $\mathcal{D}$ and each cell area equal to Δ, we approximate $\int_{\mathcal{D}} \eta(\boldsymbol{s}; \boldsymbol{\theta}_2) d\boldsymbol{s} \approx \Delta \sum_{j=1}^{M} \eta(\boldsymbol{v}_j; \boldsymbol{\theta}_2)$. The full conditional distribution for $\boldsymbol{\theta}_2 = [\{\boldsymbol{u}_j\}_{j=1}^m, \Sigma_\eta]$ depends upon how we model $\eta(\boldsymbol{s}; \boldsymbol{\theta}_2)$. For the bivariate parametric kernels, as in Section 2.2.1 it is proportional to

$$\prod_{j=1}^{m} [\boldsymbol{u}_j \,|\, \mathcal{D}] \times IW(\Sigma_\eta \,|\, r_\eta, \Omega_\eta) \times \left(\sum_{j=1}^{M} \eta(\boldsymbol{v}_j; \boldsymbol{\theta}_2) \right)^{-n^*} \times \prod_{i=1}^{n^*} \eta(\boldsymbol{s}_i^*; \boldsymbol{\theta}_2) \,. \tag{2.6}$$

A common choice for each $[\boldsymbol{u}_j \,|\, \mathcal{D}]$ is a bivariate uniform density over $\mathcal{D}$.

When the log-Gaussian process, as described in Section 2.2.1, is used to model $\mathcal{S}^*$, the only change in parameters arises in $\boldsymbol{\theta}_2$, which now comprises $\{\boldsymbol{\phi}_2, \boldsymbol{w}_2^*\}$. The full conditional distribution for $\boldsymbol{\theta}_2$ is proportional to

$$[\boldsymbol{\phi}_2] \times N_m(\boldsymbol{w}_2^* \,|\, \boldsymbol{0}, \boldsymbol{R}_2(\boldsymbol{\phi}_2)) \times \left(\int_{\mathcal{D}} \eta(\boldsymbol{s}; \boldsymbol{\theta}_2) d\boldsymbol{s} \right)^{-n^*} \times \prod_{i=1}^{n^*} \eta(\boldsymbol{s}_i^*; \boldsymbol{\theta}_2) \,. \tag{2.6$'$}$$

Evaluating (2.6′) entails approximating the integral of the intensity surface in each iteration. We conveniently take the $\boldsymbol{u}_i$'s in Section 2.2.1 over a grid and use the current estimate for $\boldsymbol{\theta}_2$ to approximate $\int_{\mathcal{D}} \eta(\boldsymbol{s}; \boldsymbol{\theta}_2) d\boldsymbol{s} \approx \Delta \sum_{j=1}^{m} \eta(\boldsymbol{u}_j; \boldsymbol{\theta}_2)$. Details regarding the choice of priors and updating steps are provided in Section 2.3 in the context of specific examples.

2.2.3 Spatial prediction, interpolation and model assessment

For predicting $Y(\boldsymbol{s}_0)$ at any location $\boldsymbol{s}_0$ in the domain, we sample from the posterior predictive distribution, $[Y(\boldsymbol{s}_0) \,|\, \boldsymbol{y}] = \int [Y(\boldsymbol{s}_0) \,|\, \boldsymbol{y}, \boldsymbol{\theta}_1, \tau^2, \mathcal{S}^*][\boldsymbol{\theta}_1, \tau^2, \mathcal{S}^* \,|\, \boldsymbol{y}]$ using *composition* (e.g., Banerjee et al. 2004). For each $\{\boldsymbol{\theta}_1^{(l)}, \tau^{2(l)}, \mathcal{S}^{*(l)}\}$, for $l = 1, 2, \ldots, L$, obtained from the posterior distribution $[\boldsymbol{\theta}_1, \tau^2, \mathcal{S}^* \,|\, \boldsymbol{y}]$, we draw $Y(\boldsymbol{s}_0)^{(l)}$ from $[Y(\boldsymbol{s}_0) \,|\, \boldsymbol{y}, \boldsymbol{\theta}_1^{(l)}, \tau^{2(l)} \mathcal{S}^{*(l)}]$.

For inference on the spatial process, $\tilde{w}_{\tilde{\epsilon}}(\boldsymbol{s}_0)$, we use posterior predictive samples from

$$[\tilde{w}_{\tilde{\epsilon}}(\boldsymbol{s}_0) \,|\, \boldsymbol{y}] = \int [\tilde{w}_{\tilde{\epsilon}}(\boldsymbol{s}_0) \,|\, w(\mathcal{S}^*), \boldsymbol{\theta}_1, \tau^2, \mathcal{S}^*][w(\mathcal{S}^*) \,|\, \boldsymbol{y}, \boldsymbol{\theta}_1, \tau^2, \mathcal{S}^*][\boldsymbol{\theta}_1, \mathcal{S}^* \,|\, \boldsymbol{y}].$$

We first sample $\{\boldsymbol{\theta}_1^{(l)}, \tau^{2(l)}, \mathcal{S}^{*(l)}\}$, for $l = 1, 2, \ldots, L$, from the posterior distribution. Next, we sample $w(\mathcal{S}^*)^{(l)}$ from $[w(\mathcal{S}^*) \,|\, \boldsymbol{y}, \boldsymbol{\theta}_1^{(l)}, \tau^{2(l)}, \mathcal{S}^{*(l)}]$ which, in fact, is a normal distribution, and finally, we sample $\tilde{w}_{\tilde{\epsilon}}(\boldsymbol{s}_0)^{(l)}$ from $[w(\boldsymbol{s}_0) \,|\, w(\mathcal{S}^*)^{(l)}, \boldsymbol{\theta}^{(l)}, \mathcal{S}^{*(l)}]$, again, a normal distribution.

We assess model performance using *independent* replicates for each observed outcome: for each $\boldsymbol{s}_i \in \mathcal{S}$, we draw $Y_{rep}(\boldsymbol{s}_i)^{(l)}$ from $N\left(\boldsymbol{x}(\boldsymbol{s}_i)'\boldsymbol{\beta}^{(l)} + \tilde{w}_\epsilon(\boldsymbol{s}_i)^{(l)}, \tau^{2(l)}\right)$, one for one for the posterior samples. Letting $\mu_{rep,i}$ and $\sigma^2_{rep,i}$ be the posterior predictive mean and variance for each $Y_{rep}(\boldsymbol{s}_i)$, we compute $G = \sum_{i=1}^n (y(\boldsymbol{s}_i) - \mu_{rep,i})^2$ and $P = \sum_{i=1}^n \sigma^2_{rep,i}$. We use D $= G + P$ (e.g., Gelfand and Ghosh, 1998) as a model selection criteria, with lower values of D indicating better models. Further, for each analysis we used a holdout set to assess each models' predictive performance by computing the mean squared prediction error (MSPE), $\frac{1}{q}\sum_{i=1}^{q}(y(\boldsymbol{s}_i) - \tilde{Y}(\boldsymbol{s}_i))^2$, where $\tilde{Y}(\boldsymbol{s}_i)$ is the predicted outcome at the i-th holdout location and q is the number of locations in the holdout set.

2.3 Illustrations

We use both a synthetic and forest inventory dataset to assess model performance with regard to learning about process parameters and predicting at new locations. Posterior inference was based on three chains of $25,000$ iterations (the first 5,000 iterations were discarded as burn-in). The samplers were coded in C++ and Fortran and leveraged Intel's Math Kernel Library threaded BLAS and LAPACK routines for matrix computations. All analyses were conducted on a Linux workstation using two Intel Nehalem quad-Xeon processors.

2.3.1 Synthetic data analysis

The synthetic dataset comprises $n = 5,500$ observations within a unit square domain with outcome values generated from $N(\beta_0\mathbf{1}, \sigma^2\boldsymbol{R}(\phi) + \tau^2\boldsymbol{I})$ with $\boldsymbol{R}(\phi)$ an $n \times n$ correlation matrix whose (i,j)-th element is $e^{-\phi d_{ij}}$, $d_{ij} = \|\boldsymbol{s}_i - \boldsymbol{s}_j\|$ and parameters given in the first column of Table 2.1. Figure 2.2(a) illustrates the spatial random effect surface interpolated over the $w(\boldsymbol{s})$'s. To facilitate model comparison using predictive performance, 500 observations were withheld to serve as a holdout set. Eleven gridded knot intensities ($n^* = (5^2, 6^2, \ldots, 15^2)$ were considered for both the non-adaptive (i.e., fixed knot) and adaptive predictive process models. For all models, the intercept parameter β_0 was given a *flat* prior and the variance parameters τ^2 and σ^2 each received inverse-Gamma $IG(2,1)$ priors. Further, assuming an exponential spatial correlation function the prior for the spatial decay parameter ϕ was a Uniform $U(3,300)$, which corresponds to support between 0.01 and 1.0 in map distance units. This is a broad range of support considering the maximum distance between any two observations is 1.4. The adaptive knot models follow the log-Gaussian parameterization of $\eta(\boldsymbol{s})$, detailed in Subsection 2.2.1, with a broad prior support of $U(3,300)$ on ϕ_2.

Results for the 25, 36, 196, and 225 knot models are detailed in Table 2.1. Here, both the non-adaptive and adaptive models produce similar estimates of β_0 across the range of knot intensities. When n^* is small, the sparse grid of knots provides an over-smoothed representation of the latent spatial surface and, as a result, the non-adaptive model is not able to accurately estimate the spatial random effect parameters σ^2 and ϕ. In contrast, even at a 25 knot intensity, the adaptive model provides better estimates of these parameters. Note, however, that the nugget, τ^2, is apparently underestimated for the adaptive model. In fact, the bias-adjustment incorporated here may tend to slightly over-fit. It is not surprising, therefore, that the adaptive knots further overcompensate for the bias, which, after all, is characterized only for the fixed-knot setting. Considering the non-adaptive and adaptive models' D and MSPE across knot intensities in Table 2.1, it is clear that knot location influences model fit and subsequent prediction. For example, with just 25 knots, the adaptive model produced D=45880. This level of fit was not achieved until the $\sim$81 knot intensity for the non-adaptive model. In addition to a consistently better model fit (i.e., lower D), the adaptive model offers considerably lower MSPE across all knot intensities. For instance, the 9.24 MSPE of the 225 knot

non-adaptive model is considerably larger than the 9.05 MSPE achieved by the 36 knot adaptive model.

Even with the reduced dimensionality afforded by the predictive process, fitting these models is time consuming. The last row in each section of Table 2.1 gives the run time for 25,000 MCMC iterations on a single non-hyperthreaded processor. The extra complexity of the adaptive predictive process sampler approximately doubles the run time across all knot intensities. Importantly, however, the adaptive model can produce an MSPE of 9.05 in 12 hours versus the non-adaptive model's substantially larger 9.24 MSPE of the 225 knot model, which requires a 28.5 hour run time. Further, the 225 knot non-adaptive model's fit, of D=37007, is achieved by the 81 knot adaptive model, which had a run time of 23.0 hours.

The plots in Figure 2.2 can help us understand the adaptive model's advantage over the fixed knot model. Figure 2.2(b) is a trace plot of knot movement for the adaptive 25 knot model. Here, the $\bullet$ symbols mark the states of one MCMC chain over the domain. The density surface associated with Figure 2.2(b) is illustrated in Figure 2.2(c), where higher values (darker shades) indicate the regions where the knots were sampled more intensely. By comparing the *true* spatial random effect surface Figure 2.2(a) with Figure 2.2(c) it is apparent that the knots tend to move to regions of extreme $w(\boldsymbol{s})$ values (as seen in the one-dimensional example offered in Section 2.2). This is a trend repeated across the adaptive predictive process models. Figures 2.2(d) and (e) were generated by interpolating over the median of each location's spatial random effect posterior distribution calculated using the non-adaptive and adaptive 25 knot predictive process models, respectively. Comparing these surfaces with Figure 2.2(a) shows that the adaptive knots provide a more detailed representation of the spatial random effect surface, hence improved model fit and predictive ability.

Given the trade-off between the non-adaptive and adaptive knot models' run time and predictive performance, we considered a hybrid non-adaptive model that used the knot locations of the 1000-th MCMC iteration of an adaptive model. Here, the choice of the 1000-th iteration was arbitrary; however, the idea was to allow enough iterations for the knots to move about the domain, while keeping the run time to a minimum. As summarized in the last column in the second row of Table 2.1, this hybrid model seems to enjoy the improved fit and lower MSPE of the adaptive model and the shorter run

Table 2.1: Predictive process candidate models' parameter posterior credible intervals 50 (2.5, 97.5), model fit criterion, and mean squared prediction error (MSPE) for the synthetic dataset. Run time is for a single chain of 25,000 iterations on a single non-hyperthreaded processor.

		Non-adaptive predictive process (i.e., fixed knots)				
	True	25	36	196	225	
β_0	1	1.62 (0.41, 2.91)	1.10 (-0.02, 2.58)	1.22 (0.56, 1.90)	1.26 (0.64, 1.91)	
σ^2	5	2.47 (1.62, 3.49)	2.91 (1.93, 4.33)	4.94 (4.04, 6.43)	5.22 (4.45, 6.41)	
τ^2	1	4.18 (3.82, 4.49)	3.91 (3.43, 4.35)	1.76 (1.28, 2.16)	1.45 (1.03, 1.79)	
ϕ	30	3.45 (3.02, 5.28)	4.26 (3.11, 7.32)	14.02 (9.76, 18.06)	15.11 (11.41, 18.75)	
G	–	25123	24725	18762	17726	
P	–	25545	25145	20068	19281	
D	–	50669	49871	38831	37007	
MSPE	–	10.37	10.22	10.11	9.24	
Run time (hours)	–	5.0	6.0	24.5	28.5	
		Adaptive predictive process				Non-adaptive with adaptive starting
	True	25	36	196	225	36
β_0	1	1.24 (1.15, 1.34)	1.29 (1.18, 1.40)	1.05 (0.83, 1.31)	1.11 (0.91, 1.52)	0.70 (0.51, 0.86)
σ^2	5	4.60 (4.37, 4.87)	4.40 (4.10, 4.70)	4.63 (4.32, 4.92)	4.69 (4.40, 5.07)	4.84 (4.50, 5.13)
τ^2	1	0.27 (0.13, 0.43)	0.30 (0.12, 0.50)	0.14 (0.08, 0.22)	0.13 (0.08, 0.26)	0.35 (0.21, 0.61)
ϕ	30	23.31 (21.13, 26.06)	22.30 (20.54, 24.51)	24.46 (22.16, 27.01)	24.08 (21.47, 26.93)	18.55 (16.65, 20.34)
ϕ_2	–	155.34 (23.11, 290.71)	160.39 (18.05, 290.93)	3.14 (3.00, 3.95)	3.20 (3.02, 3.90)	–
G	–	22722	21032	15395	14782	23025
P	–	23157	21825	17615	17186	23464
D	–	45880	42858	33010	31968	46490
MSPE	–	9.66	9.05	7.06	6.84	9.69
Run time (hours)	–	10.0	12.01	50.0	59.4	6.6

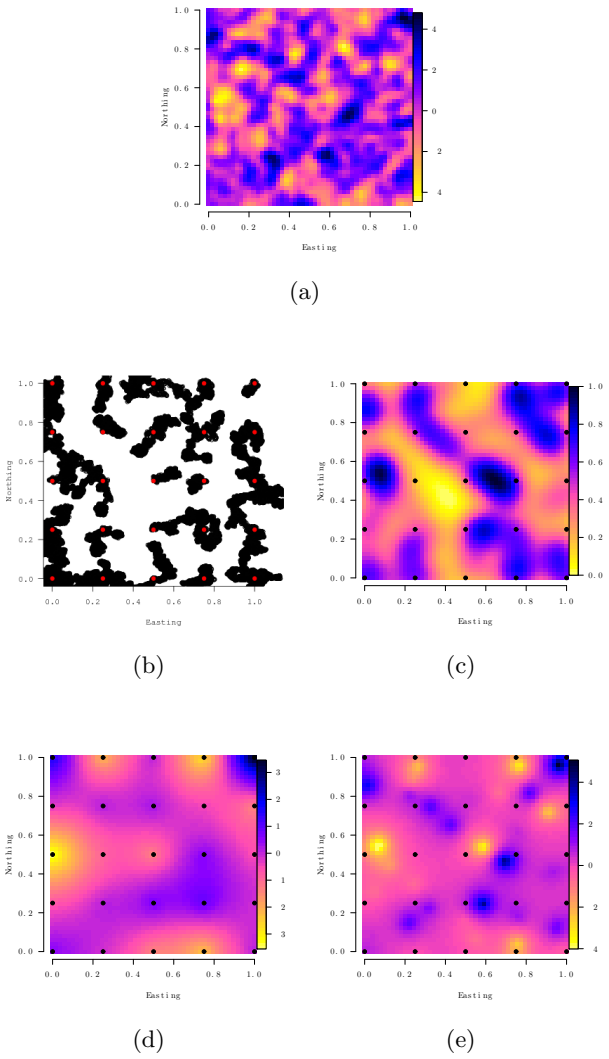


Figure 2.2: Synthetic data and associated estimates for the 25 knot predictive process models: (a) synthetic spatial random effect surface generated using 5,000 observations; (b) 25,000 MCMC iteration trace plot of the adaptive knot locations; (c) density plot associated with the MCMC iteration in (b); (d) non-adaptive predicted process model estimated spatial random effects, and (e) adaptive predicted process model estimated spatial random effects.

time advantage of the non-adaptive model.

2.3.2 Forest biomass data analysis

Spatial prediction of forest biomass is critical to many important contemporary global-, regional-, and local-scale decisions, including assessments of current carbon stock and flux, bio-feedstock for emerging bio-economies, and impact of deforestation. In the United States, the Forest Inventory and Analysis (FIA) program of the USDA Forest Service collects the data needed to support these assessments.

The program has established field plot centers in permanent locations using a sampling design that produces an equal probability sample (Bechtold and Patterson, 2005). Locations of the 7.32 m radius forested plots are determined using GPS receivers. The state of Michigan, in which the study area is located, has a sampling intensity of approximately one plot per 800 ha. On these plots, field crews recorded stem measurements for all trees with diameter at breast height (dbh; 1.37 m above the forest floor) of 12.7 cm or greater. Given these data, established allometric equations were used to estimate each plot's forest biomass per ha. Here, we model the log metric tons of forest biomass per ha. A July, 2003 mosaic of Landsat TM imagery was used to calculate tasseled cap components of brightness (TC1), greenness (TC2), and wetness (TC3) to serve as predictor variables (Huang et al., 2002). Figure 2.3(a) illustrates the georeferenced forest inventory data consisting of 6,538 forested FIA plots measured between 1999 and 2006 across the lower peninsula of Michigan.

Candidate models include a simple non-spatial regression and the non-adaptive and adaptive predictive process models. Similar to the synthetic data analysis, we considered a range of knot intensities. Knot locations were chosen by applying the `k-means` clustering algorithm to the observed locations. For the adaptive models these knot locations served as starting values. As in the synthetic data analysis, we considered an additional non-adaptive candidate model, that used the knot locations of the 1000-th MCMC iteration of an adaptive model.

Based on results from an initial variogram analysis of the non-spatial model's residuals, the priors for τ^2 and σ^2 for the non-adaptive and adaptive predictive process models followed $IG(2\ 0.5)$. Assuming an Exponential spatial correlation function the prior for the spatial decay parameter ϕ followed a $U(0.006\ \ 3)$, which corresponds to support from 1–500 km. Again, this is a broad range of support, given the maximum distance between any two plots is 460 km. For all models the regression coefficients each received

Table 2.2: Predictive process candidate models' parameter posterior credible intervals 50 (2.5, 97.5), model fit criterion, and mean squared prediction error (MSPE) for the forest biomass dataset. Run time is for a single chain of 25,000 iterations on a single non-hyperthreaded processor.

	Non-spatial	Non-adaptive predictive process			Adaptive predictive process	Non-adaptive with adaptive starting
		50	100	200	50	50
β_0	10.98 (10.95, 11.01)	10.99 (10.71, 11.26)	11.00 (10.68, 11.30)	11.01 (10.97, 11.06)	10.81 (10.69, 10.94)	10.92 (10.82, 11.04)
β_{TC1}	0.07 (0.02, 0.12)	0.03 (-0.02, 0.09)	0.03 (-0.02, 0.09)	0.05 (0.00, 0.11)	0.06 (0.00, 0.11)	0.04 (-0.02, 0.09)
β_{TC2}	-0.03 (-0.09, 0.02)	-0.00 (-0.07, 0.06)	-0.01 (-0.07, 0.06)	-0.02 (-0.08, 0.04)	-0.02 (-0.09, 0.02)	-0.01 (-0.07, 0.05)
β_{TC3}	0.45 (0.41, 0.49)	0.43 (0.39, 0.48)	0.43 (0.39, 0.48)	0.44 (0.39, 0.48)	0.45 (0.41, 0.48)	0.44 (0.39, 0.48)
σ^2	–	0.17 (0.09, 0.33)	0.14 (0.08, 0.24)	0.28 (0.16, 0.48)	1.14 (1.00, 1.32)	0.65 (0.46, 0.79)
τ^2	1.01 (0.97, 1.04)	0.93 (0.84, 0.98)	0.96 (0.87, 1.00)	0.74 (0.53, 0.86)	0.13 (0.08, 0.22)	0.52 (0.40, 0.64)
ϕ		0.016 (0.009, 0.051)	0.011 (0.007, 0.062)	0.016 (0.011, 0.023)	0.042 (0.031, 0.052)	0.05 (0.03, 0.065)
G	5935	5804	5759	5640	5751	5902
P	5939	5842	5830	5810	5777	5791
D	11874	11646	11590	11450	11529	11693
MSPE	2.00	1.98	1.96	1.94	1.96	1.96
Run time (hours)	–	9.44	19.31	37.08	19.42	9.73

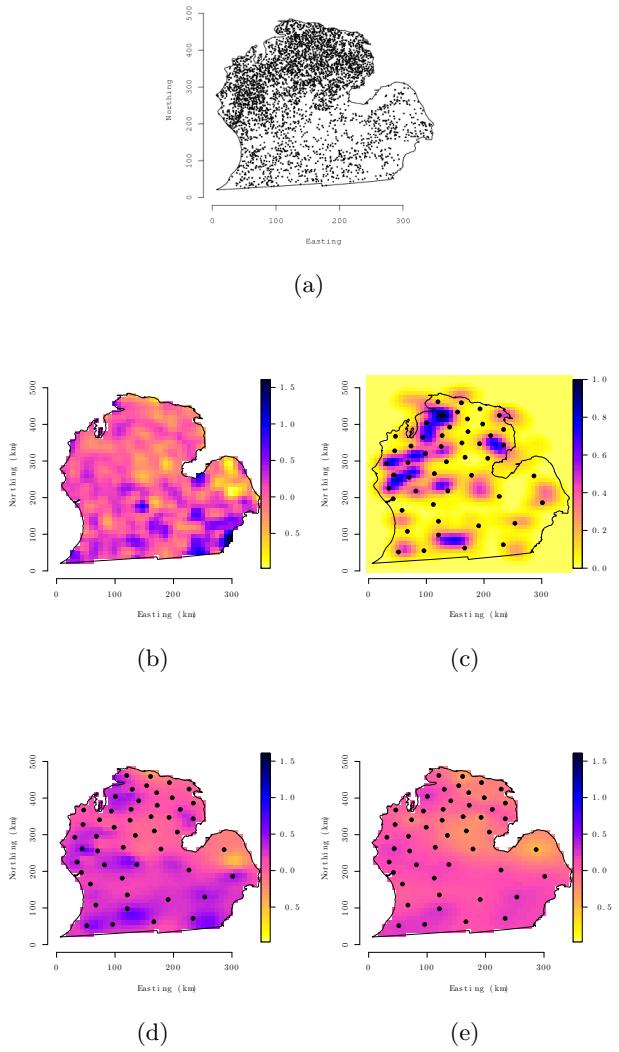


Figure 2.3: Forest biomass dataset and associated estimates for the 50 knot predictive process models: (a) location of forest invetory plots; (b) interpolated surface of the non-spatial model residuals; (c) density plot of the adaptive knot locations over 25,000 MCMC iterations; (d) adaptive predicted process model estimated spatial random effects with knot starting locations, and; (e) non-adaptive predicted process model estimated spatial random effects with knot locations.

a *flat* prior. The prior on the adaptive knot locations, $\mathcal{S}^*$, was defined by a rectangular domain that covered the extent of the irregularly shaped study area.

Here, we opted to use the parametric parameterization of $\eta(\boldsymbol{s})$ detailed in Section 2.2.1. Priors for the parameters comprising the bivariate normal mixture covariance matrix Σ_η were an $IG(2,1)$ for σ^2_η and $U(-1,1)$ for ρ_η. The mixture was evaluated over a grid of m=25 locations. We experimented with a range of m, 25–100, and found that it had negligible influence on parameter estimates and subsequent prediction.

Candidate models were assessed based on their fit to observed data, predictive performance at new locations, and run time. To assess predictive performance, 653 observations (i.e., 10%) were selected randomly to serve as a holdout set. The remaining 5,885 observations were used to fit the candidate models.

Figure 2.3(b) is an interpolated surface of the non-spatial model residuals. We would expect the fitted spatial random effects of the candidate models to look somewhat similar to this residual surface. Figure 2.3(c) provides the density plot of the adaptive knot locations over 25,000 MCMC iterations for the 50 knot model. Here, darker colors correspond to regions where the knots sampled more intensely and the $\bullet$ symbols indicate the starting location of the 50 knots. As in the synthetic data analysis, it is obvious the knots moved from the starting locations to sample at locations where the absolute value of the residual surface is large. This figure also shows the knots generally sample at locations close to the observed data (i.e., they did not sample much beyond the state's bounding polygon). Figure 2.3(d) and (e) provide interpolated surfaces of the median of spatial random effects posterior distribution for the adaptive and non-adaptive models, respectively. Here, we see the adaptive model produces spatial random effects that more closely approximate Figure 2.3(b) and hence provide improved model fit and prediction over the fixed knot model as detailed in Table 2.2.

Table 2.2 offers parameter estimates for the predictive process candidate models. Over the range of knot intensities, both the non-adaptive and adaptive predictive process models produce comparable estimates of the regression coefficients – several of which explain a significant amount of variability in log biomass. We again see a discrepancy between non-adaptive and adaptive models' estimates of the variance components. Specifically, the non-adaptive model seems to estimate a larger nugget, τ^2, and a smaller partial sill, σ^2, whereas the opposite trend is seen in the adaptive model. This trend was also observed in the synthetic data analysis and other exploratory analyses we conducted. Results suggest that $\sim$100 fixed knots are needed to produce MSPE and model

fit, D, comparable to that of the 50 knot adaptive model. However, this advantage is lessened by the near equal run times of the two models (as noted in the last row of Table 2.2). The last column in this table presents the results for the non-adaptive model that used the knot locations of the 1000-th MCMC iteration of the 50 knot adaptive model. Similar to the synthetic analysis, this hybrid model offers the improved fit and predictive ability of an adaptive knot model with a run time comparable to that of the fixed knot model.

2.4 Discussion

The current chapter integrates modeling of knots in low rank predictive process models within a hierarchical framework, thereby circumventing issues underlying the choice of "knots". Indeed, we were able to obtain essentially indistinguishable inference with fewer stochastic knots than with fixed knots. Also, our approach applies seamlessly to other low-rank models that use kernel convolutions or other nonstationary covariance structures (e.g., Higdon, 2002; Cressie and Johannesson, 2008). It also applies when $\epsilon(\tilde{\boldsymbol{s}})$ in the modified predictive process is a tapered process (e.g., Furrer et al., 2006) so that $\boldsymbol{D}_{\tilde{\epsilon}}$ is sparse (but not necessarily diagonal). Direct methods for sparse linear systems can then be employed for efficient computations (e.g., Davis, 2006). Feasibility of alternative estimation strategies such as INLA (Rue et al., 2009; Eidsvik et al., 2010) can also be explored.

Any random probability measure for $[\mathcal{S}^* \,|\, \boldsymbol{\theta}_2]$ will yield a valid hierarchical model in (2.5). If we eschew spatially informative priors for the knots, a fully non-parametric option using realizations from a Dirichlet Process may be viable. Algorithms to estimate such models have been outlined, among others, by Neal (1998). Stochastic modeling for random locations also arise when the outcome is "preferentially sampled" (Diggle et al., 2010), i.e., the process generating the outcome is not independent of the process generating the observed locations. This requires jointly modeling the process and the set of observed locations $\mathcal{S}$. Pati et al., (2011) recently proposed a hierarchical Bayesian geostatistical model for preferentially sampled data. An adaptive predictive process version of these models can be envisioned through stochastic specifications for $[\mathcal{S} \,|\, \boldsymbol{\theta}_1]$ in (2.5) that would add an additional level of hierarchy.

Chapter 3

Modeling low-rank spatially-varying cross-covariances using predictive process with application to soil nutrient data

3.1 Introduction

As discussed in the last chapter, spatial process models have, over the past decade, contributed substantially to answering increasingly complex questions encountered in the natural and environmental sciences. One area of active research is the analysis of spatially indexed datasets with multivariate outcomes measured at each location. For example, in ecological datasets these could be measures of species abundance, vegetation characteristics, or pollutants at each inventory or monitoring location. It is typically posited that there is association between the measurements at a given location. In addition, we anticipate association between measurements at different locations, which is likely to weaken as locations become farther apart but not necessarily as a function of the (Euclidean) distance between the locations. Interest in multivariate spatial statistics has

customarily centered upon interpolation and prediction of the outcomes ("multivariate kriging") while accounting for underlying associations. There is an extensive literature in multivariate spatial statistics, which is too large to be reviewed here; Gelfand and Banerjee (2010) offer a recent review. In the context of fully model-based inference, to achieve computational tractability as well as easier interpretability, usually stationary multivariate spatial processes have been prescribed. In particular, it is customarily assumed that the associations between the outcomes do not vary across locations.

Scientific queries in recent times, however, have sought to relax this assumption. Today, for example, ecologists are often interested in ascertaining associations among these biotic and/or abiotic variables within a location and how this within-location correlation changes across the domain, see, e.g., Diez and Pulliam (2007), Ovaskainen et al. (2010), and Waddle et al. (2010). The lurking nonstationary patterns in correlations among spatially indexed outcomes can reveal important unmeasured predictors, deepen understanding of the ecological processes, and improve prediction of the multivariate vector of outcomes at new locations within the domain.

Multivariate spatial process models are built either from a valid cross-covariogram or a valid cross-covariance function. We seek full and exact inference, including prediction, from such models, which requires a full distributional specification and, in particular, a full sampling distribution for the data. We take this to be a multivariate Gaussian process and so the issue becomes specification of the cross covariance function. These functions are not routine to specify since they demand that for any number of locations and any choice of these locations the resulting covariance matrix for the data be positive definite. For univariate spatial processes, Bochner's theorem offers a useful characterization for valid covariance functions (see, e.g., Gneiting and Guttorp, 2010). The analogous characterization for cross-covariance functions is given by Cramér's Theorem (see, e.g. Cramér, 1940), which does not necessarily lead to tractable forms.

Various approaches for deriving valid cross-covariances are possible; see Gelfand and Banerjee (2010) and Cressie and Wickle (2011) for recent reviews. Ver Hoef and Barry (1998) propose a moving average approach, which is similar to the popular kernel convolution approach used to create rich classes of stationary and nonstationary spatial processes (Higdon, 2002). Gaspari and Cohn (1999) and Majumdar and Gelfand (2007) use convolution of univariate covariance functions to produce valid multivariate

cross-covariances. Approaches using linear transformations of independent latent processes include the linear model of coregionalization (Grzebyk and Wackernagel, 1994; Wackernagel, 2006) and its variants (Schmidt and Gelfand, 2003; Gelfand et al., 2004; Zhang, 2007; Finley et al., 2009a). In more recent work, Gneiting, Kleiber and Schlather (2010) extend the univariate Matérn covariance functions to an interesting class of cross-covariance functions. Apanasovich and Genton (2010) offer a different approach using latent dimensions that produce a class of valid cross-covariance functions.

Most of the above literature has, however, focused upon stationary cross-covariances that do not vary across space. Some of the above methods can be adapted to accommodate spatially-varying cross-covariances but they become computationally prohibitive for even moderately sized datasets. For example, Gelfand et al. (2004) detail a spatially-varying matrix-variate Wishart process for modeling nonstationary multivariate spatial data. However, the computational burden for estimating the resulting models is prohibitive for even moderately sized spatially indexed datasets. To be specific, this approach requires explicit estimation of random effect vectors of length $nm + n(m(m+1)/2)$, where n is the number of locations and m is the number of outcomes. Iterative estimation algorithms involve matrix decompositions with operations in the order of $O(n^3m^3)$ in every iteration. Classical likelihood-based methods for estimating associations among outcomes are also difficult, although under stationarity, that is when these associations do not vary across space, an E-M algorithm can be devised (Zhang, 2007). For nonstationary models, the large parameter space and the matrix computations make these models infeasible for analyzing most datasets we encounter today on standard computing frameworks.

Our current work can be regarded as a computationally feasible alternative to spatially-varying "linear models of coregionalization" using the matrix-variate Wishart process. Our contribution is a class of low-rank spatially-varying cross-covariance functions that can be seamlessly incorporated in hierarchical models. We depart from the more traditional uses of low rank spatial models, which are typically to counter the challenges posed by spatial datasets involving a very large number of locations (e.g., Wahba, 1990; Higdon, 2002; Rasmussen and Williams, 2006; Tokdar, 2007; Stein, 2007, 2008; Cressie and Johannesson 2008; Banerjee et al., 2008, 2010; Crainiceanu et al., 2008; Finley et al., 2009b).

Unlike in the aforementioned references, the challenge in our current application is not a very large number of locations but, instead, the high-dimensional parameter space that arises from spatially-varying cross-covariances. Given the established efficacy and flexibility of latent process approaches in multivariate spatial modeling, we use spatially-varying linear transformations of independent processes to construct our cross-covariances. For added flexibility, we opt for a Bayesian paradigm where the elements of the cross-covariance matrix are treated as unknown functions of space that are modeled using spatial processes. We then use the predictive process (Banerjee et al., 2008, 2010) counterparts of these spatial processes to arrive at flexible classes of low-rank spatially-varying cross-covariance functions. A challenge is that these low rank processes are typically *degenerate* in that their realizations over the original set of locations will yield singular joint distributions. The multivariate predictive process suggests a natural adjustment or modification that makes it non-degenerate without increasing the computational burden.

From an application standpoint, we believe that our current work is the first attempt at estimating general nonstationary spatially-varying cross-covariances for multiple outcomes. We demonstrate several data analytic advantages of our flexible framework including, of course, the production of maps for covariances among the different outcomes. Not only do the low-rank processes accrue computational benefits, they also lead to improved model fit and better predictive performance.

Our methodological contributions entail hitherto unaddressed implications of certain structural assumptions in modeling spatially-varying cross-covariances. In particular, we explore two different constructions for low-rank cross-covariances based upon the predictive process. We derive relationships between the respective marginal cross-covariance matrices after the effect of the space-varying transformation has been integrated out. We show how one approach implies stronger cross-covariances *apriori* than the other, but how the within-location covariances among the multiple outcomes are the same for the two approaches.

The rest of the chapter evolves as follows. In Section 3.2 we introduce the La Selva Biological Station soil nutrients dataset and the research interest that motivates the development of methods for modeling and mapping these and similar multivariate data. The proposed extension is developed in Section 3.3. In Section 3.5, we provide details

on model implementation, prediction, and assessment of fit. Additional detail on model development is offered in Appendix B. In Section 3.6, we explore the proposed models using synthetic and La Selva Biological Station soil nutrients datasets.

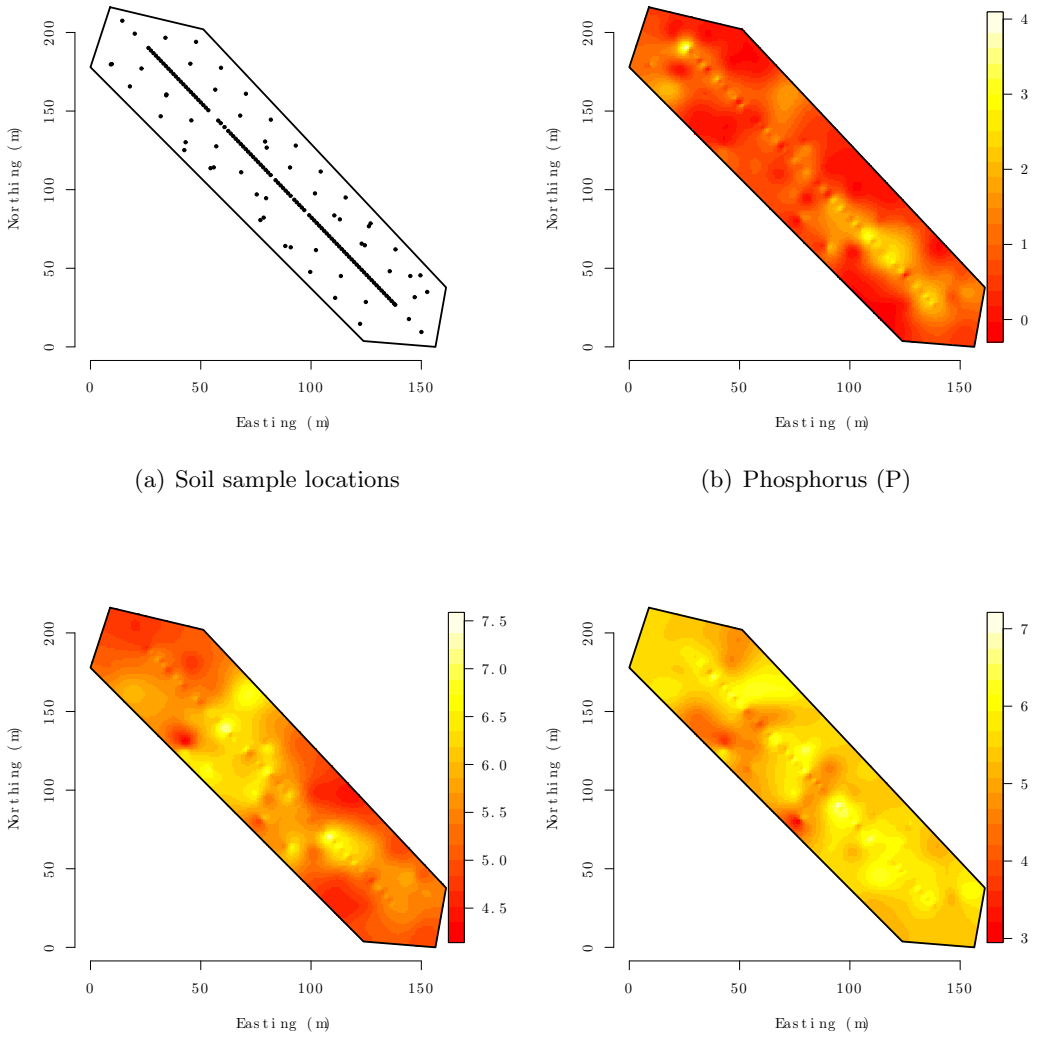
3.2 La Selva Biological Station soil nutrients dataset

Maps of soil nutrients correlations over forested domains are important for identifying limitations to tree growth and survivorship from seedling to canopy tree stages. Furthermore, spatial correlations in soil nutrients also could indicate the degree to which different elements are coupled in biogeochemical cycles and what factors influence this coupling. Thus, spatial variation in the correlation between elements can provide a window into the ecological processes that are causing spatial patterns in nutrient availability.

These and similar needs motivate the methods detailed in Section 3.3. In particular we use a spatial soil nutrients dataset from a wet tropical forest at La Selva Biological Station, Costa Rica (Holste et al., 2010) to illustrate how the proposed methods provide computationally efficient insight into spatial biogeochemical processes. Holste et al. (2010) sampled soils, n=249, at 1 m intervals along a central 1×200 m transect that is used for studies on tree seedling demography (e.g., Kobe and Vriesendorp, 2009); additional samples were taken at 10 m intervals, alternating between a distance of 10 and 20 m from the central axis and at 14 randomly chosen locations (Figure 3.1(a)).

Each sample was analyzed for a broad array of soil nutrients using standard methods. Here we focus on three aspects of soil nutrients: phosphorus (P) because it is widely believed to limit tropical forest productivity; the sum of inorganic pools of nitrogen (SN) because of its importance in light capturing pigments and enzymes involved in photosynthesis and P acquisition, and; base cations (SBC), expressed as the sum of charge equivalents of calcium, magnesium, and potassium, because these nutrients are often deficient in highly weathered tropical soils. Figure 3.1 (b-c) shows interpolated surfaces of the three soil nutrient variables.

A typical assumption in geostatistics is that of *isotropic* random fields, which means that the spatial correlation between observations from two different locations is a function of the geometric distance between the locations. The shape of the transect, as



(a) Soil sample locations

(b) Phosphorus (P)

(c) Sum of base cations (SBC)

(d) Sum of nitrogen (SN)

Figure 3.1: Sampling grid and interpolated surfaces of the observed soil nutrient outcomes

seen in Figure 3.1(a) suggests a likely violation of this assumption. Being substantially more long that it is wide, the transect allows many more observation locations along its

length than along its width. This asymmetry in information is likely to yield higher spatial ranges (that is, the distance where spatial correlation becomes negligible) along its length than along its width. In fact, even if the true underlying process were isotropic, the sampling scheme imposed by the shape of the transect would likely yield different spatial ranges from directional variograms along its length and width. One approach is to accommodate "stationary anisotropies" (Ecker and Gelfand, 1999; 2003) for each univariate outcome that characterize the anisotropy in terms of the geometry of the domain. We, on the other hand, seek to capture more general nonstationarity in the associations among our outcomes. This is motivated as follows.

Spatial covariance in nutrients could be an important signature for fundamental biogeochemical processes. For example, SN and P may covary spatially because nitrogen is an important constituent of extracellular phosphatases secreted by microbes, which in turn should increase P availability (Houlton et al., 2008). In fact, this mechanism has been hypothesized to explain the high prevalence of tree species in the tropics that fix atmospheric dinitrogen into plant available forms of SN, even though tropical forest growth is generally thought to be limited by P (Houlton et al. 2008). On the other hand, species-specific tree feedbacks on soil chemistry (e.g., Townsend et al., 2008, McCarthy-Neumann and Kobe, 2010) could decouple nutrient cycles. Characterizing spatially-varying covariance between SN and P tests the hypothesis that SN influences P availability at a local scale where the processes of nitrogen fixation and phosphatase production are manifested; in addition, understanding patterns of spatial covariance between SN and P is a first step in identifying the factors that disrupt the potential SN-P linkage.

3.3 Multivariate spatial process models

Let $\mathcal{D} \subset \Re^d$ be a connected subset of d-dimensional Euclidean space and let $\boldsymbol{s} \in \mathcal{D}$ be a generic point in $\mathcal{D}$. In our subsequent applications $d = 2$. The multivariate spatial setting envisions, at each spatial location $\boldsymbol{s}$, an $m \times 1$ outcome $\boldsymbol{y}(\boldsymbol{s}) = (y_1(\boldsymbol{s}), y_2(\boldsymbol{s}), \ldots, y_m(\boldsymbol{s}))'$ along with an $m \times p$ matrix of covariates, $\boldsymbol{X}(\boldsymbol{s})'$, whose i-th row is a $1 \times p$ vector of covariates $\boldsymbol{x}_i(\boldsymbol{s})'$. Let $\mathcal{S} = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_n\}$ be a set of n locations in $\mathcal{D}$ where the outcome and predictors have been observed. A multivariate spatial

regression model assumes, for each $\boldsymbol{s}_i \in \mathcal{S}$, that

$$\boldsymbol{y}(\boldsymbol{s}_i) = \boldsymbol{X}(\boldsymbol{s}_i)'\boldsymbol{\beta} + \boldsymbol{w}(\boldsymbol{s}_i) + \boldsymbol{\epsilon}(\boldsymbol{s}_i) \; , \tag{3.1}$$

where $\boldsymbol{w}(\boldsymbol{s}) = (w_1(\boldsymbol{s}), w_2(\boldsymbol{s}), \ldots, w_m(\boldsymbol{s}))'$ and $\boldsymbol{\epsilon}(\boldsymbol{s}) = (\epsilon_1(\boldsymbol{s}), \epsilon_2(\boldsymbol{s}), \ldots, \epsilon_m(\boldsymbol{s}))'$ are $m \times 1$ vectors of spatial effects and measurement error effects respectively. That is, $\boldsymbol{\epsilon}(\boldsymbol{s}) \sim N(\boldsymbol{0}, \boldsymbol{\Psi})$, where $\boldsymbol{\Psi}$ is customarily assumed to be a diagonal matrix with τ_j^2 as its j-th diagonal element.

3.3.1 Modeling cross-covariance functions

A critical ingredient in (3.1) is the unobserved spatial process $\boldsymbol{w}(\boldsymbol{s})$. Customarily, $\{\boldsymbol{w}(\boldsymbol{s}) \in \Re^m : \boldsymbol{s} \in \mathcal{D}\}$ is assumed to be a zero-centered m-variate Gaussian process. The process $\boldsymbol{w}(\boldsymbol{s})$ is completely specified by its *cross-covariance* function $\boldsymbol{C}_w(\boldsymbol{s}, \boldsymbol{t})$, which, for any pair of locations $\boldsymbol{s}$ and $\boldsymbol{t}$, is an $m \times m$ matrix with $\text{cov}\{w_i(\boldsymbol{s}), w_j(\boldsymbol{t})\}$ as its (i, j)-th element. The joint distribution of the $mn \times 1$ vector $\boldsymbol{w} = (\boldsymbol{w}(\boldsymbol{s}_1)', \boldsymbol{w}(\boldsymbol{s}_2)', \ldots, \boldsymbol{w}(\boldsymbol{s}_n)')'$ is a zero centered multivariate normal distribution. The variance-covariance matrix of $\boldsymbol{w}$ is denoted by $\boldsymbol{C}_w$, which is an $mn \times mn$ block matrix formed by placing $\boldsymbol{C}_w(\boldsymbol{s}_i, \boldsymbol{s}_j)$ as the (i, j)-th block. For the special case of $m = 1$, i.e, univariate Gaussian processes (as in Chapter 2), the cross-covariance matrix becomes a real-valued covariance function (see, e.g., Gneiting and Guttorp, 2010).

A valid multivariate process must ensure that $\boldsymbol{C}_w$ is positive definite (hence symmetric too), which implies that $\boldsymbol{C}_w(\boldsymbol{s}, \boldsymbol{t})$ must satisfy the following two conditions:

$$\begin{array}{ll} \text{(i)} & \boldsymbol{C}_w(\boldsymbol{s}, \boldsymbol{t}) = \boldsymbol{C}_w(\boldsymbol{t}, \boldsymbol{s})' \\ \text{(ii)} & \sum_{i=1}^n \sum_{j=1}^n \boldsymbol{u}_i' \boldsymbol{C}_w(\boldsymbol{s}_i, \boldsymbol{s}_j) \boldsymbol{u}_j > 0 \quad \forall \quad \boldsymbol{u}_i, \boldsymbol{u}_j \in \Re^m \setminus \{\boldsymbol{0}\} \; . \end{array} \tag{3.2}$$

The first condition in (3.2) ensures that $\boldsymbol{C}_w$ is symmetric, although the cross-covariance matrix function itself need not be. The second condition in (3.2) ensures that $\boldsymbol{C}_w$ is positive-definite. These must be satisfied for all integers n and any finite collection of locations $\mathcal{S} = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\} \subset \mathcal{D}$. Note that (3.2) implies that $\boldsymbol{C}_w(\boldsymbol{s}, \boldsymbol{s})$ is symmetric and positive definite. In fact, it is precisely the variance-covariance matrix for the elements of $\boldsymbol{w}(\boldsymbol{s})$ within site $\boldsymbol{s}$.

Characterizing valid cross-covariance matrix functions that satisfy (3.2) is not trivial.

Assuming stationarity, i.e. the cross-covariance function depends only upon the separation of the locations $\boldsymbol{h} = \boldsymbol{t} - \boldsymbol{s}$ so that $\boldsymbol{C}_w(\boldsymbol{s}, \boldsymbol{t}) = \boldsymbol{C}_w(\boldsymbol{h})$, the primary characterization theorem for cross-covariance functions (Cramér, 1940; Yaglom, 1987) says that real-valued functions $C_{ij}(\boldsymbol{h})$ will form the elements of a valid cross-covariance matrix if and only if each admits the cross-spectral representation $C_{ij}(\boldsymbol{h}) = \int \exp(2\pi i \boldsymbol{t}'\boldsymbol{h}) d(F_{ij}(\boldsymbol{t}))$ with respect to a positive definite measure $F(\cdot)$, i.e., where the cross-spectral matrix $M(B)$, with (i, j)-th element $F_{ij}(B)$, is positive definite for any Borel subset $B \subseteq \Re^d$. This is the analogue of Bochner's Theorem for covariance functions in univariate settings (e.g, Gneiting and Guttorp, 2010). Valid choices for $F_{ij}(\boldsymbol{t})$'s for constructing classes of cross-covariance functions have been discussed by several authors. Corollaries of the above representation lead to the approaches proposed by Gaspari and Cohn (1999) for constructing valid cross-covariance functions as convolutions of covariance functions of stationary random fields. Gelfand and Banerjee (2010) offer a review of widely different approaches for constructing cross-covariance matrix functions. Also see recent work by Gneiting, Kleiber and Schlather (2010) and Apanasovich and Genton (2010) for further theoretical insights.

3.3.2 Constructive approaches for cross-covariance functions

The cross-spectral representation above is, however, less useful for constructing non-stationary processes and, in particular, for spatially-varying covariance models. One alternative approach uses space-varying linear transformations of tractable spatial processes. We rewrite the cross-covariance matrix as $\boldsymbol{C}_w(\boldsymbol{s}, \boldsymbol{t}) = \boldsymbol{A}(\boldsymbol{s})\Theta_w(\boldsymbol{s}, \boldsymbol{t})\boldsymbol{A}(\boldsymbol{t})'$, where $\Theta_w(\boldsymbol{s}, \boldsymbol{t})$ is called the *cross-correlation* function and $\boldsymbol{A}(\boldsymbol{s})$ is an $m \times m$ matrix whose elements are functions of $\boldsymbol{s}$. The cross-correlation function must satisfy $\Theta_w(\boldsymbol{s}, \boldsymbol{s}) = \boldsymbol{I}_p$ in addition to (3.2). For any non-singular matrix $\boldsymbol{A}(\boldsymbol{s})$, it is clear that a valid cross-correlation function will produce a valid cross-covariance function. A feasible approach for building richly structured spatial process models, therefore, is to specify a simple form for $\Theta_w(\boldsymbol{s}, \boldsymbol{t})$ and then build additional structure through $\boldsymbol{A}(\boldsymbol{s})$ and $\boldsymbol{A}(\boldsymbol{t})$.

One of the simplest specifications for a cross-correlation matrix is diagonal with univariate correlation functions along its diagonal. This approach can be motivated using latent variables. Let $\boldsymbol{v}(\boldsymbol{s})$ be an $m \times 1$ vector with elements $v_i(\boldsymbol{s})$ that are independent zero-centered latent spatial processes with unit variance. That is, each

$v_i(\boldsymbol{s}) \sim GP(0, \rho_i(\cdot;\boldsymbol{\theta}_k))$ with $\text{var}\{v_i(\boldsymbol{s})\} = 1$, and

$$\text{cov}(v_i(\boldsymbol{s}), v_j(\boldsymbol{t})) = \begin{cases} \rho_i(\boldsymbol{s},\boldsymbol{t};\boldsymbol{\theta}_i) & \text{if} \quad i = j \\ 0 \quad \text{if} \quad i \neq j \ , \end{cases} \tag{3.3}$$

where $\rho_i(\cdot;\boldsymbol{\theta}_i)$ is a real-valued correlation function corresponding to the process $v_i(\boldsymbol{s})$ and $\boldsymbol{\theta}_i$ are parameters therein. Equation 3.3 implies that the cross-covariance matrix $\boldsymbol{C}_v(\boldsymbol{s},\boldsymbol{t})$ is diagonal with $\rho_i(\boldsymbol{s},\boldsymbol{t};\boldsymbol{\theta}_i)$ as the i-th diagonal entry. We assume that $\boldsymbol{w}(\boldsymbol{s}) = \boldsymbol{A}(\boldsymbol{s})\boldsymbol{v}(\boldsymbol{s})$, which yields a highly structured cross-covariance function,

$$\boldsymbol{C}_w(\boldsymbol{s},\boldsymbol{t}) = \boldsymbol{A}(\boldsymbol{s})\boldsymbol{C}_v(\boldsymbol{s},\boldsymbol{t})\boldsymbol{A}(\boldsymbol{t})' = \sum_{k=1}^{m} \boldsymbol{a}_k(\boldsymbol{s})\boldsymbol{a}_k(\boldsymbol{t})'\rho_k(\boldsymbol{s},\boldsymbol{t};\boldsymbol{\theta}_k) \ , \tag{3.4}$$

where $\boldsymbol{a}_k(\boldsymbol{s})$ is the k-th column of $\boldsymbol{A}(\boldsymbol{s})$ and $\boldsymbol{C}_v(\boldsymbol{s},\boldsymbol{t})$ coincides with the cross-correlation matrix $\Theta_w(\boldsymbol{s},\boldsymbol{t})$. The validity of $\boldsymbol{C}_w(\boldsymbol{s},\boldsymbol{t})$ as a cross-covariance function is immediate by construction.

A flexible choice for $\rho_i(\cdot;\boldsymbol{\theta}_i)$ is the Matérn correlation function that controls spatial association and smoothness (see, e.g., Stein, 1999; Gneiting and Guttorp, 2010) and is given by

$$\rho(\boldsymbol{s},\boldsymbol{t};\boldsymbol{\theta}) = \frac{1}{2^{\theta_2-1}\Gamma(\theta_2)}(\|\boldsymbol{s}-\boldsymbol{t}\|\theta_1)^{\theta_2}\mathcal{K}_{\theta_2}(\|\boldsymbol{s}-\boldsymbol{t}\|\theta_1); \ \ \theta_1 > 0, \theta_2 > 0 \ , \tag{3.5}$$

where $\boldsymbol{\theta} = \{\theta_1, \theta_2\}$, θ_1 controls the decay in spatial correlation and θ_2 is a smoothness parameter with higher values yielding smoother process realizations. Also, Γ is the usual Gamma function while $\mathcal{K}_{\theta_2}$ is a modified Bessel function of the second kind with order θ_2 and $\|\boldsymbol{s}-\boldsymbol{t}\|$ is the Euclidean distance between the locations $\boldsymbol{s}_1$ and $\boldsymbol{t}$. Covariance functions that depend upon the distance metric only are often referred to as *isotropic*. Several other choices for valid correlation functions are discussed in Banerjee et al. (2004). For $v_k(\boldsymbol{s})$ we choose isotropic Matérn functions $\rho_k(\boldsymbol{s}_1,\boldsymbol{s}_2;\boldsymbol{\theta}_k)$ with $\boldsymbol{\theta}_k = (\theta_{1k},\theta_{2k})$ for $k = 1,\ldots,m$. Note that $\rho_k(\boldsymbol{s}_1,\boldsymbol{s}_2;\boldsymbol{\theta}_k)$ is the correlation function for the k-th component of $\boldsymbol{v}(\boldsymbol{s})$ and not for the k-th element of the observed outcome $\boldsymbol{y}(\boldsymbol{s})$. Consequently, the $\boldsymbol{\theta}_k$'s do not correspond directly to $\boldsymbol{y}(\boldsymbol{s})$, but to the unobserved process $\boldsymbol{v}(\boldsymbol{s})$ that drives the spatial variation in $\boldsymbol{y}(\boldsymbol{s})$.

Attention then turns to modeling $\boldsymbol{A}(\boldsymbol{s})$, which, in practice, is unknown. A key observation in this regard is that $\boldsymbol{C}_w(\boldsymbol{s},\boldsymbol{s}) = \boldsymbol{A}(\boldsymbol{s})\boldsymbol{A}(\boldsymbol{s})'$. This suggests a few options for structurally modeling $\boldsymbol{A}(\boldsymbol{s})$. One can assume that $\boldsymbol{A}(\boldsymbol{s})$ is lower-triangular, corresponding to a Cholesky factorization of $\boldsymbol{C}_w(\boldsymbol{s},\boldsymbol{s})$. This approach has been adopted by Pourahmadi (1999) in the context of longitudinal studies but without varying covariances and also in certain versions of the so called Linear Model of Coregionalization (LMC) in geostatistics (Gelfand et al., 2004; Wackernagel, 2006; Zhang, 2007; Banerjee and Johnson, 2006; Finley et al., 2008). With the Cholesky square-root, the one-to-one correspondence between the elements of $\boldsymbol{A}(\boldsymbol{s})$ and $\boldsymbol{C}_w(\boldsymbol{s},\boldsymbol{s})$ is well-known (see, e.g., Harville, p 229), provided we insist that the diagonal elements of $\boldsymbol{A}(\boldsymbol{s})$ are greater than zero. This has obvious implications for prior specifications for the elements of $\boldsymbol{A}(\boldsymbol{s})$ in a Bayesian hierarchical model. In fact, specifying parametric forms for the elements of $\boldsymbol{A}(\boldsymbol{s})$ is awkward. For example, it is not clear what parametric forms will be appropriate nor is there a reason to believe that these elements will share some parameters. Yet, allowing each element of $\boldsymbol{A}(\boldsymbol{s})$ to have its own parameters will yield a highly over-parametrized model.

A more reasonable option is to assume that the elements of $\boldsymbol{A}(\boldsymbol{s})$ themselves arise from spatial processes. Gelfand et al. (2004) propose a matrix-variate inverse-Wishart process for modeling $\boldsymbol{A}(\boldsymbol{s})\boldsymbol{A}(\boldsymbol{s})'$ but that is computationally clumsy and prohibitive for our data. A simpler option is to assume that the sub-diagonal and the logarithm of the diagonal elements follow independent Gaussian processes. This is conceptually simpler, but will still be computationally prohibitive. Nevertheless, this specification allows us to replace these processes with flexible low-rank counterparts, which we discuss in the next section.

The lower-triangular specification for $\boldsymbol{A}(\boldsymbol{s})$ imposes some conditional independence constraints. To see how, consider $m = 2$ and two locations $\boldsymbol{s}_1$ and $\boldsymbol{s}_2$ and suppose that the elements of $\boldsymbol{A}(\boldsymbol{s})$ are fixed. Then, for $i = 1, 2$

$$w_1(\boldsymbol{s}_i) = a_{11}(\boldsymbol{s}_i)v_1(\boldsymbol{s}_i) \text{ and } w_2(\boldsymbol{s}_i) = a_{21}(\boldsymbol{s}_i)v_1(\boldsymbol{s}_i) + a_{22}(\boldsymbol{s}_i)v_2(\boldsymbol{s}_i)\ .$$

Since the process $v_1(\boldsymbol{s})$ completely determines $w_1(\boldsymbol{s})$, we can write

$$\begin{aligned}\mathrm{cov}\{w_1(\boldsymbol{s}_1), w_2(\boldsymbol{s}_2) \,|\, w_1(\boldsymbol{s}_2)\} &= \mathrm{cov}\{a_{11}(\boldsymbol{s})v_1(\boldsymbol{s}_1), a_{21}(\boldsymbol{s})v_1(\boldsymbol{s}_2) + a_{22}(\boldsymbol{s})v_2(\boldsymbol{s}_2) \,|\, v_1(\boldsymbol{s}_2)\} \\ &= a_{11}(\boldsymbol{s})a_{22}(\boldsymbol{s})\mathrm{cov}\{v_1(\boldsymbol{s}_1), v_2(\boldsymbol{s}_2) \,|\, v_1(\boldsymbol{s}_2)\} = 0 \,,\end{aligned} \tag{3.6}$$

where the last equality follows because the process $v_1(\cdot)$ is independent of $v_2(\cdot)$. This shows that $w_1(\boldsymbol{s}_1)$ and $w_2(\boldsymbol{s}_2)$ will be independent conditional upon $w_1(\boldsymbol{s}_2)$. In terms of the dispersion matrix, this conditional independence implies that the inverse of $\boldsymbol{C}_w$ has zeroes in its entries. When $\boldsymbol{A}(\boldsymbol{s})$ is random, the marginal cross-covariance is obtained by integrating out, or taking expectation with respect to, the elements of $\boldsymbol{A}(\boldsymbol{s})$. From (3.6) it is clear that the marginal cross-covariance will still be zero.

In theory, we can easily obviate restrictions such as in (3.6) by specifying $\boldsymbol{A}(\boldsymbol{s})$ to be a non-triangular square root. For example, we could set $\boldsymbol{A}(\boldsymbol{s}) = \boldsymbol{P}(\boldsymbol{s})\Lambda^{1/2}(\boldsymbol{s})$, or the symmetric version $\boldsymbol{A}(\boldsymbol{s}) = \boldsymbol{P}(\boldsymbol{s})\Lambda^{1/2}(\boldsymbol{s})\boldsymbol{P}(\boldsymbol{s})'$, where $\boldsymbol{C}_w(\boldsymbol{s},\boldsymbol{s}) = \boldsymbol{P}(\boldsymbol{s})\Lambda\boldsymbol{P}(\boldsymbol{s})'$ is the spectral decomposition for $\boldsymbol{C}_w(\boldsymbol{s},\boldsymbol{s})$. This requires further parametrization for the $m \times m$ orthogonal matrix $\boldsymbol{P}(\boldsymbol{s})$, such as in terms of the $m(m-1)/2$ *Givens* angles $\theta_{ij}(\boldsymbol{s})$ for $i = 1, \ldots, p-1$ and $j = i+1, \ldots, p$ (e.g. Daniels and Kass, 1999).

Although the conditional independence in the triangular specification may seem theoretically unnecessary, note that it is only an *apriori* assumption that does not carry over to posterior inference conditional upon the data. Our ultimate interest lies in $\boldsymbol{C}_w(\boldsymbol{s},\boldsymbol{t})$, which is robustly estimated using any bijective square-root map. Also, the number of parameters to be estimated in the Given's angle specification is the same as that for the triangular Cholesky. In practical settings these specifications matter little but the Cholesky decompositions will be numerically more stable than computing the spectral decomposition. The former is also less expensive, requiring $O(m^3/3)$ flops as compared to more than $O(4m^3/3)$ flops required by the latter. For these reasons, we opt for the Cholesky square-root in our subsequent data analysis.

3.4 Multivariate predictive process models

Once a valid cross-covariance function is specified for a multivariate spatial process, the realizations of $\boldsymbol{w}(\boldsymbol{s})$ over the finite set $\mathcal{S}$ follow a multivariate normal distribution with zero mean and variance covariance matrix $\boldsymbol{C}_w$. Without further specifications, estimating (3.1) will require matrix factorizations involving the $mn \times mn$ matrix $\boldsymbol{C}_w$. Such computations invoke linear solvers or Cholesky decompositions of complexity $O(n^3m^3)$, once every iteration, rendering them computationally prohibitive when mn is large.

A easy yet effective remedy to this problem is to use the multivariate version of predictive process models, discussed in Chapter 2. Using generic notation, let $\boldsymbol{\eta}(\boldsymbol{s}) \sim GP(\boldsymbol{0}, \boldsymbol{C}_\eta(\cdot))$ denote an $m \times 1$ zero-centered multivariate Gaussian process with cross-covariance function $\boldsymbol{C}_\eta(\boldsymbol{s}, \boldsymbol{t})$. We call $\boldsymbol{\eta}(\boldsymbol{s})$ the *parent process* and consider its realizations over an arbitrary but fixed set of knots $\mathcal{S}^* = \{\boldsymbol{s}_1^*, \ldots, \boldsymbol{s}_{n^*}^*\}$, where $n^* \ll n$. Let $\boldsymbol{\eta}^* = (\boldsymbol{\eta}(\boldsymbol{s}_1^*)', \boldsymbol{\eta}(\boldsymbol{s}_2^*)', \ldots, \boldsymbol{\eta}(\boldsymbol{s}_{n^*}^*)')'$ be the realizations of the parent process over $\mathcal{S}^*$. The *multivariate predictive process* for $\boldsymbol{\eta}(\boldsymbol{s})$ is

$$\boldsymbol{\eta}_{pp}(\boldsymbol{s}) = \mathrm{E}[\boldsymbol{\eta}(\boldsymbol{s}) \,|\, \boldsymbol{\eta}^*] = \mathrm{cov}\{\boldsymbol{\eta}(\boldsymbol{s}), \boldsymbol{\eta}^*\}\mathrm{var}\{\boldsymbol{\eta}^*\}^{-1}\boldsymbol{\eta}^* = \mathcal{C}_\eta(\boldsymbol{s}, \mathcal{S}^*)'\boldsymbol{C}_\eta^{*-1}\boldsymbol{\eta}^* \,, \tag{3.7}$$

where $\mathcal{C}_\eta(\boldsymbol{s}, \mathcal{S}^*)'$ is an $m \times n^*m$ matrix composed of the $m \times m$ blocks $\boldsymbol{C}_\eta(\boldsymbol{s}, \boldsymbol{s}_j^*)$ for $j = 1, 2, \ldots, n^*$ and $\boldsymbol{C}_\eta^*$ is the $n^*m \times n^*m$ variance-covariance matrix for $\boldsymbol{\eta}^*$, i.e., with $\boldsymbol{C}_\eta(\boldsymbol{s}_i^*, \boldsymbol{s}_j^*)$ as its (i, j)-th block. For the univariate case, $\mathcal{C}_\eta(\boldsymbol{s}, \mathcal{S}^*)$, $\boldsymbol{C}_\eta^*$ are replaced by $\boldsymbol{c}_\eta(\boldsymbol{s}, \mathcal{S}^*)$, $\boldsymbol{C}_\eta^*$ respectively.

The process $\boldsymbol{\eta}_{pp}(\boldsymbol{s})$ is *singular* or *degenerate* in that its realizations over any finite set of locations follow singular Gaussian distributions (e.g. Rao, 1973) as soon as the number of locations in that set exceeds the number of knots. In geostatistical regression models, this does not usually create problems in parameter estimation because the presence of the measurement error component ensures that the joint distribution of the outcomes is always proper. However, care is needed in situations such as ours where we use this to model low-rank cross-covariance matrices.

Since the predictive process is a conditional expectation, the variance covariance

matrix of $\boldsymbol{\eta}(\boldsymbol{s}) - \boldsymbol{\eta}_{pp}(\boldsymbol{s})$ is available in closed form as

$$\text{var}\{\boldsymbol{\eta}(\boldsymbol{s}) - \boldsymbol{\eta}_{pp}(\boldsymbol{s})\} = \boldsymbol{C}_\eta(\boldsymbol{s},\boldsymbol{s}) - \mathcal{C}_\eta(\boldsymbol{s},\mathcal{S}^*)'\boldsymbol{C}_\eta^{*-1}\mathcal{C}_\eta(\boldsymbol{s},\mathcal{S}^*) = \text{var}\{\boldsymbol{\eta}(\boldsymbol{s})\} - \text{var}\{\boldsymbol{\eta}_{pp}(\boldsymbol{s})\}\ . \tag{3.8}$$

Such a closed form expression facilitates deriving a bound of the stochastic error incurred due to the approximation of the Gaussian process through a predictive process. In fact,

Lemma 3.4.1 *Assume that (i)* $\eta_i(\boldsymbol{s}) - \eta_{pp,i}(\boldsymbol{s})$ *is a.s. bounded on* D *for* $i = 1, ..., m$. *(ii)* D *is compact. Let*
$||\eta_i - \eta_{pp,i}|| = \sup_{\boldsymbol{s}\in D}\{\eta_i(\boldsymbol{s}) - \eta_{pp,i}(\boldsymbol{s})\}$ *and* $\sigma^2_{D,i} = \sup_{\boldsymbol{s}\in D}[var\{\eta_i(\boldsymbol{s}) - \eta_{pp,i}(\boldsymbol{s})\}]$.
Then $\forall\, \epsilon > E||\eta_i - \eta_{pp,i}||$,

$$P\left\{\sup_{\boldsymbol{s}\in D}|\eta_i(\boldsymbol{s}) - \eta_{pp,i}(\boldsymbol{s})| > \epsilon\right\} \leq 2\exp\left\{-\frac{(\epsilon - E||\eta_i - \eta_{pp,i}||)^2}{\sigma^2_{D,i}}\right\} \tag{3.9}$$

Proof Proof of the result is presented in Appendix B.

This result simple tells us that as $\sigma^2_{D,i} \to 0$ the supnorm distance between the high rank Gaussian process and the predictive process tends to 0 in probability.

While approximating a parent process with a low rank model, a stochastic modeler always expects to retain some of the desirable features of the spatial correlation function. A close look at the spatial correlation function of the parent process in (3.5) reveals that $0 \leq \rho(\boldsymbol{s},\boldsymbol{t};\boldsymbol{\theta}) \leq 1$. Hitherto, no effort has been made to see whether a low rank spatial model retains such an attractive feature. Positivity of the correlation function is particularly simple to check for *Higdon's* class (see Higdon, 2002) for low rank models. In Appendix B, we present an analogous treatment for the predictive process model.

Note that (3.8) seen to be positive definite (unless the knots coincide with locations whereupon it is zero) and reveals an inherent bias in the estimates of the spatial and non-spatial variance components (e.g. Finley et al., 2009a; Banerjee et al. 2010). The expression in (3.8) suggests an effective and simple resolution to the singularity of the low rank process as well as the biased estimation: we add a vector process $\boldsymbol{\epsilon}_{\eta_{pp}}(\boldsymbol{s})$ to $\boldsymbol{\eta}_{pp}(\boldsymbol{s})$ that is independent across locations and has (3.8) as its variance-covariance matrix at

$\boldsymbol{s}$. More precisely, we write $\boldsymbol{\eta}_{mpp}(\boldsymbol{s}) = \boldsymbol{\eta}_{pp}(\boldsymbol{s}) + \boldsymbol{\epsilon}_{\eta_{pp}}(\boldsymbol{s})$ and call it a bias-adjusted or *modified predictive process*. This has cross-covariance matrix

$$\boldsymbol{C}_{\eta_{mpp}}(\boldsymbol{s},\boldsymbol{t}) = \boldsymbol{C}_{\eta_{pp}}(\boldsymbol{s},\boldsymbol{t}) + I_{\{\boldsymbol{s}=\boldsymbol{t}\}}\{\boldsymbol{C}_{\eta}(\boldsymbol{s},\boldsymbol{t}) - \boldsymbol{C}_{\eta_{pp}}(\boldsymbol{s},\boldsymbol{t})\} \; . \tag{3.10}$$

We now apply these notations to different spatial processes that arise in our subsequent modeling.

In our current application, our multivariate process is $\boldsymbol{w}(\boldsymbol{s}) = \boldsymbol{A}(\boldsymbol{s})\boldsymbol{v}(\boldsymbol{s})$ (Section 3.3.2). If $\boldsymbol{A}(\boldsymbol{s})$ had a tractable parametric form, for example when its elements are simple parametric functions of $\boldsymbol{s}$ or when each element is a constant, dimension reduction proceeds from $\boldsymbol{A}(\boldsymbol{s})\boldsymbol{v}_{mpp}(\boldsymbol{s})$, which can easily be shown to be $\mathrm{E}[\boldsymbol{w}(\boldsymbol{s}) \,|\, \boldsymbol{w}^*]$ (Appendix B). See Finley et al. (2009b; 2011) and Banerjee et al. (2010) for applications of modified predictive process models in forestry that assumed $\boldsymbol{A}(\boldsymbol{s})$ to be invariant over $\boldsymbol{s}$.

Here we depart from the existing paradigm and posit that each element of $\boldsymbol{A}(\boldsymbol{s})$ is itself a spatial process, independent of other elements of $\boldsymbol{A}(\boldsymbol{s})$ as well as those of $\boldsymbol{v}(\boldsymbol{s})$. Now we seek dimension reduction in both $\boldsymbol{A}(\boldsymbol{s})$ and $\boldsymbol{v}(\boldsymbol{s})$. A natural approach is to replace the elements of $\boldsymbol{A}(\boldsymbol{s})$ and $\boldsymbol{v}(\boldsymbol{s})$ with their predictive process counterparts. As mentioned in Section 3.3.2, although $\boldsymbol{A}(\boldsymbol{s})$ is often taken as a triangular matrix, the low-rank models we formulate do not require any specific forms.

Let $\boldsymbol{a}^*_{ij}$ be the collection of $a_{ij}(\boldsymbol{s}^*_i)$'s, $i = 1, 2, \ldots, n^*$, and let $\boldsymbol{A}^*$ be the collection of $\boldsymbol{a}^*_{ij}$'s. Then, we can define the low-rank process $\boldsymbol{w}_{pp}(\boldsymbol{s})$ as

$$\begin{aligned} \boldsymbol{w}_{pp}(\boldsymbol{s}) &= \mathrm{E}[\boldsymbol{A}(\boldsymbol{s})\boldsymbol{v}(\boldsymbol{s}) \,|\, \boldsymbol{A}^*, \boldsymbol{v}^*] = \mathrm{E}_{\boldsymbol{A}^*}[\mathrm{E}_{\boldsymbol{v}^* \,|\, \boldsymbol{A}^*}[\boldsymbol{A}(\boldsymbol{s})\boldsymbol{v}(\boldsymbol{s})]] = \mathrm{E}_{\boldsymbol{A}^*}[\mathrm{E}_{\boldsymbol{v}^*}[\boldsymbol{A}(\boldsymbol{s})\boldsymbol{v}(\boldsymbol{s})]] \\ &= \mathrm{E}[\boldsymbol{A}(\boldsymbol{s}) \,|\, \boldsymbol{A}^*]\mathrm{E}[\boldsymbol{v}(\boldsymbol{s}) \,|\, \boldsymbol{v}^*] = \boldsymbol{A}_{pp}(\boldsymbol{s})\boldsymbol{v}_{pp}(\boldsymbol{s}) = \sum_{k=1}^{m} \boldsymbol{a}_{k,pp}(\boldsymbol{s}) v_{k,pp}(\boldsymbol{s}) \; , \end{aligned} \tag{3.11}$$

where $\boldsymbol{A}_{pp}(\boldsymbol{s})$ is an $m \times m$ matrix with elements $a_{ij,pp}(\boldsymbol{s}) = \mathrm{E}[a_{ij}(\boldsymbol{s}) \,|\, \boldsymbol{a}^*_{ij}]$ and $\boldsymbol{a}_{k,pp}(\boldsymbol{s})$ is the k-th column vector of $\boldsymbol{A}_{pp}(\boldsymbol{s})$. The cross-covariance matrix for $\boldsymbol{w}_{pp}(\boldsymbol{s})$ is given by

$$\boldsymbol{C}_{w_{pp}}(\boldsymbol{s},\boldsymbol{t}) = \boldsymbol{A}_{pp}(\boldsymbol{s})\boldsymbol{C}_{v_{pp}}(\boldsymbol{s},\boldsymbol{t})\boldsymbol{A}_{pp}(t)' = \boldsymbol{A}_{pp}(\boldsymbol{s})\mathcal{C}_v(\boldsymbol{s},\mathcal{S}^*)'\boldsymbol{C}_v^{*-1}\mathcal{C}_v(\boldsymbol{t},\mathcal{S}^*)\boldsymbol{A}(t)' \; , \tag{3.12}$$

and $\boldsymbol{C}_{v_{pp}}(\boldsymbol{s},\boldsymbol{t})$ is easily shown to be diagonal, which has computational advantages.

Apparently there are two ways to construct a non-singular (or non-degenerate) modified predictive process from (3.11). The first approach is to simply use $\boldsymbol{w}_{mpp}(\boldsymbol{s})$ with cross-covariance function as in (3.10) except that η is replaced by w. Note that this corresponds to the process $\boldsymbol{w}_{mpp}(\boldsymbol{s}) = \boldsymbol{A}_{pp}(\boldsymbol{s})\boldsymbol{v}_{pp}(\boldsymbol{s}) + \boldsymbol{\epsilon}_{w_{pp}}(\boldsymbol{s})$. The second is to construct the modified predictive process versions of $\boldsymbol{A}(\boldsymbol{s})$ and $\boldsymbol{v}(\boldsymbol{s})$ and construct a low-rank spatial process. We elucidate this second approach below and explore the relationship between the resulting cross-covariance function and (3.10).

Let $\tilde{\boldsymbol{A}}(\boldsymbol{s})$ be the $m \times m$ matrix with (i,j)-th element $\tilde{a}_{ij}(\boldsymbol{s})$, which is the modified predictive process corresponding to $a_{ij}(\boldsymbol{s})$. Similarly, let $\tilde{\boldsymbol{v}}(\boldsymbol{s})$ have $\tilde{v}_i(\boldsymbol{s})$ as its i-th element, which is the modified predictive processes corresponding to $v_i(\boldsymbol{s})$. Let us now define $\tilde{\boldsymbol{w}}(\boldsymbol{s}) = \tilde{\boldsymbol{A}}(\boldsymbol{s})\tilde{\boldsymbol{v}}(\boldsymbol{s})$ and let $C_{\tilde{w}}(\boldsymbol{s},\boldsymbol{t})$ be the cross-covariance function for $\tilde{\boldsymbol{w}}(\boldsymbol{s})$.

It will be instructive to compare the *marginal* cross-covariance matrices for $\boldsymbol{w}_{mpp}(\boldsymbol{s})$ and $\tilde{\boldsymbol{w}}(\boldsymbol{s})$, given by $\mathrm{E}\left[\boldsymbol{C}_{w_{mpp}}(\boldsymbol{s},\boldsymbol{t})\right]$ and $\mathrm{E}\left[\boldsymbol{C}_{\tilde{w}}(\boldsymbol{s},\boldsymbol{t})\right]$ respectively, where the expectations are with respect to the joint distributions of $\boldsymbol{A}^*$ and $\boldsymbol{v}^*$. These can be expressed as

$$\mathrm{E}\left[\boldsymbol{C}_{\tilde{w}}(\boldsymbol{s},\boldsymbol{t})\right] = \mathrm{E}\left[\tilde{\boldsymbol{A}}(\boldsymbol{s})\boldsymbol{C}_{\tilde{v}}(\boldsymbol{s},\boldsymbol{t})\tilde{\boldsymbol{A}}(\boldsymbol{t})'\right]$$
$$\text{and} \quad \mathrm{E}\left[\boldsymbol{C}_{w_{mpp}}(\boldsymbol{s},\boldsymbol{t})\right] = \mathrm{E}\left[\boldsymbol{A}_{pp}(\boldsymbol{s})\mathrm{cov}\{\boldsymbol{v}_{pp}(\boldsymbol{s}),\boldsymbol{v}_{pp}(\boldsymbol{t})\}\boldsymbol{A}_{pp}(\boldsymbol{t})'\right] + \mathrm{cov}\{\boldsymbol{\epsilon}_{w_{pp}}(\boldsymbol{s}),\boldsymbol{\epsilon}_{w_{pp}}(\boldsymbol{t})\}\ .$$

We offer a brief outline of the relationship between the above two marginal cross-covariances, leaving the details to the Appendix B.

Using the fact that the off-diagonal elements of $\boldsymbol{A}(\boldsymbol{s})$ are zero-centered independent Gaussian processes, as are the elements of $\boldsymbol{v}(\boldsymbol{s})$, it is straightforward to derive that the marginal cross-covariance matrices above are both diagonal. Thus, the two marginal cross-covariance matrices can possibly differ only along their diagonal. If $c_{\tilde{w};i,j}(\boldsymbol{s},\boldsymbol{t})$ and $c_{w_{mpp};i,j}(\boldsymbol{s},\boldsymbol{t})$ be the (i,j)-th entry of $\boldsymbol{C}_{\tilde{w}}(\boldsymbol{s},\boldsymbol{t})$ and $\boldsymbol{C}_{w_{mpp}}(\boldsymbol{s},\boldsymbol{t})$ respectively. When $\boldsymbol{s} \neq \boldsymbol{t}$, it can be shown (see Appendix B) that the difference between the i-th diagonal element is

$$\{\mathrm{E}[\tilde{a}_{ii}(\boldsymbol{s})]\mathrm{E}[\tilde{a}_{ii}(\boldsymbol{t})] - \mathrm{E}[a_{ii,pp}(\boldsymbol{s})]\mathrm{E}[a_{ii,pp}(\boldsymbol{t})]\}\,\mathrm{cov}\{v_{i,pp}(\boldsymbol{s}), v_{i,pp}(\boldsymbol{t})\}\ . \tag{3.13}$$

If the elements of $\boldsymbol{A}(\boldsymbol{s})$ are zero-centered processes, then (3.13) is easily seen to be zero and the two cross-covariance matrices coincide. However, as noted in Section 3.3.2,

$\boldsymbol{A}(\boldsymbol{s})$, and hence $\boldsymbol{A}_{pp}(\boldsymbol{s})$, is often assumed lower-triangular with *positive* diagonal elements to ensure the one-one correspondence with the cross-covariance matrix. Their logarithms are assumed to be Gaussian processes. In that case, $\log \tilde{a}_{ii}(\boldsymbol{s}) = \log a_{pp;ii}(\boldsymbol{s}) + \tilde{\epsilon}_{ii;a}(\boldsymbol{s})$, where $e^{\tilde{\epsilon}_{ii;a}(\boldsymbol{s})}$ is log-normally distributed. When $\text{cov}\{v_{i,pp}(\boldsymbol{s}), v_{i,pp}(\boldsymbol{t})\} \geq 0$ (a reasonable assumption in spatial settings), it implies, $\text{E}\left[c_{\tilde{w};i,i}(\boldsymbol{s},\boldsymbol{t})\right] > \text{E}\left[c_{w_{mpp};i,i}(\boldsymbol{s},\boldsymbol{t})\right]$ for each i. This holds whenever $\boldsymbol{s} \neq \boldsymbol{t}$. This means that the process $\tilde{\boldsymbol{w}}(\boldsymbol{s})$ induces stronger cross-covariances between distinct locations than $\tilde{\boldsymbol{w}}_{mpp}(\boldsymbol{s})$.

When $\boldsymbol{s} = \boldsymbol{t}$, (3.13) equals zero and, again, the two cross-covariances are the same. In fact, $\text{E}\left[\boldsymbol{C}_{\tilde{w}}(\boldsymbol{s},\boldsymbol{s})\right] = \text{E}\left[\boldsymbol{C}_{w}(\boldsymbol{s},\boldsymbol{s})\right] = \text{E}\left[\boldsymbol{C}_{w_{mpp}}(\boldsymbol{s},\boldsymbol{s})\right]$. Given that our primary interest in the current application, and in most multivariate spatial settings, lies in the covariances between the outcomes within a location, the preceding results suggest that we can unambiguously work with either $\tilde{\boldsymbol{w}}(\boldsymbol{s})$ or $\tilde{\boldsymbol{w}}_{mpp}(\boldsymbol{s})$ for our low rank model. Subsequently, we work with $\tilde{\boldsymbol{w}}(\boldsymbol{s})$, which yields our low-rank counterpart of (3.1)

$$\boldsymbol{y}(\boldsymbol{s}) = \boldsymbol{X}(\boldsymbol{s})'\boldsymbol{\beta} + \tilde{\boldsymbol{w}}(\boldsymbol{s}) + \boldsymbol{\epsilon}(\boldsymbol{s}); \quad . \tag{3.14}$$

No new parameters are introduced and the cross-covariance function of the modified predictive process derives directly from that of the parent process.

3.5 Statistical inference

3.5.1 Model fitting

Let $\mathcal{S} = \{\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_n\}$ denote the set of locations where the dependent and independent variables have been observed. Our multivariate hierarchical predictive process model with spatially-varying cross-covariances can be expressed as

$$\prod_{j=1}^{m}\prod_{i=j}^{m} p(\boldsymbol{\theta}_{a;i,j}) \times \prod_{i=1}^{m} LN(\tilde{\boldsymbol{a}}_{ii} \,|\, \boldsymbol{0}, \boldsymbol{C}_{\tilde{a}}(\boldsymbol{\theta}_{a;i,i})) \times \prod_{j=1}^{m}\prod_{i=j+1}^{m} N(\tilde{\boldsymbol{a}}_{ij} \,|\, \boldsymbol{0}, \boldsymbol{C}_{\tilde{a}}(\boldsymbol{\theta}_{a;i,j}))$$
$$\times \prod_{k=1}^{m} p(\boldsymbol{\theta}_k) \times \prod_{k=1}^{m} N(\tilde{\boldsymbol{v}}_k \,|\, \boldsymbol{0}, \boldsymbol{C}_{\tilde{v}}(\boldsymbol{\theta}_k)) \times \prod_{i=1}^{n} N\left(\boldsymbol{y}(\boldsymbol{s}_i) \,|\, \boldsymbol{X}(\boldsymbol{s}_i)'\boldsymbol{\beta} + \sum_{k=1}^{m} \tilde{\boldsymbol{a}}_k(\boldsymbol{s}_i)\tilde{v}_k(\boldsymbol{s}_i), \boldsymbol{\Psi}\right), \tag{3.15}$$

where LN and N denote log-normal and normal densities respectively, $\boldsymbol{\theta}_{a;i,j}$ and $\boldsymbol{\theta}_k$ are the process parameters corresponding to $\tilde{a}_{ij}(\boldsymbol{s})$ and $\tilde{v}_k(\boldsymbol{s})$ respectively, $\boldsymbol{\Psi}$ is an $m \times m$ diagonal matrix, the diagonal elements are the residual variances ("nugget") for the corresponding dependent variable, $\boldsymbol{C}_{\tilde{a}}(\boldsymbol{\theta}_{a;i,j})$ and $\boldsymbol{C}_{\tilde{v}}(\boldsymbol{\theta}_k)$ are the variance-covariance matrices for $\tilde{\boldsymbol{a}}_{ij} = (\tilde{a}_{ij}(\boldsymbol{s}_1), \tilde{a}_{ij}(\boldsymbol{s}_2), \ldots, \tilde{a}_{ij}(\boldsymbol{s}_n))'$ (log-transformed if $i = j$) and $\tilde{\boldsymbol{v}}_k = (\tilde{v}_k(\boldsymbol{s}_1), \tilde{v}_k(\boldsymbol{s}_2), \ldots, \tilde{v}_k(\boldsymbol{s}_m))'$ respectively. These matrices are $nm \times nm$, but enjoy a low-rank structure that reaps computational benefits (Stein, 2008).

We estimate the posterior distribution arising from (3.15) using Markov chain Monte Carlo (e.g. Robert and Casella, 2010; Gelman et al. 2004; Banerjee et al. 2004). There are two options here. The first is to estimate the posterior distribution arising from (3.15) using Gibbs sampling updates for $\boldsymbol{\beta}$, the $\tilde{\boldsymbol{v}}_k$'s and, with a conjugate inverse-Wishart prior, for $\boldsymbol{\Psi}$. The remaining parameters are updated using blocked random walk Metropolis steps using multivariate normal proposals (all parameters with positive support are log-transformed). The second option is to integrate out the $\tilde{\boldsymbol{v}}_k$'s (but not the $\tilde{a}_{ij}(\boldsymbol{s}_i)$'s) from (3.15) and work with the marginalized likelihood

$$\prod_{j=1}^{m}\prod_{i=j}^{m} p(\boldsymbol{\theta}_{a;i,j}) \times \prod_{i=1}^{m} LN(\tilde{\boldsymbol{a}}_{ii} \,|\, \boldsymbol{0}, \boldsymbol{C}_{\tilde{a}}(\boldsymbol{\theta}_{a;i,i})) \times \prod_{j=1}^{m}\prod_{i=j+1}^{m} N(\tilde{\boldsymbol{a}}_{ij} \,|\, \boldsymbol{0}, \boldsymbol{C}_{\tilde{a}}(\boldsymbol{\theta}_{a;i,j})) \times \prod_{k=1}^{m} p(\boldsymbol{\theta}_k) \times N\left(\boldsymbol{y} \,|\, \boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{C}_{\tilde{\boldsymbol{w}}} + \boldsymbol{I} \otimes \boldsymbol{\Psi}\right) , \tag{3.16}$$

where $\boldsymbol{X}$ is the $nm \times p$ matrix formed by stacking the $\boldsymbol{X}(\boldsymbol{s}_i)'$'s, $\boldsymbol{C}_{\tilde{\boldsymbol{w}}}$ is the $mn \times mn$ block matrix with (i,j)-th block given by $\tilde{\boldsymbol{A}}(\boldsymbol{s}_i)\boldsymbol{C}_{\tilde{v}}(\boldsymbol{s}_i, \boldsymbol{s}_j; \boldsymbol{\theta})\tilde{\boldsymbol{A}}(\boldsymbol{s}_j)$. Using the structure of $\boldsymbol{C}_{\tilde{v}}(\boldsymbol{s}_i, \boldsymbol{s}_j; \boldsymbol{\theta})$, $\boldsymbol{C}_{\tilde{\boldsymbol{w}}}$ can be written as $\mathcal{A}\boldsymbol{P}\boldsymbol{D}_{\boldsymbol{C}_{\tilde{v}}}\boldsymbol{P}'\mathcal{A}'$, where $\mathcal{A}$ is an $mn \times mn$ block-diagonal matrix with $\boldsymbol{A}(\boldsymbol{s}_i)$ as the i-th $m \times m$ block, $\boldsymbol{D}_{\boldsymbol{C}_{\tilde{v}}}$ is another $mn \times mn$ block diagonal matrix with the k-th block being the $n \times n$ variance-covariance matrix for $\tilde{\boldsymbol{v}}_k$, $k = 1, \ldots, m$, and $\boldsymbol{P}' = [\boldsymbol{I}_m \otimes \boldsymbol{e}_1 : \boldsymbol{I}_m \otimes \boldsymbol{e}_2 : \cdots : \boldsymbol{I}_m \otimes \boldsymbol{e}_n]$, where $\boldsymbol{e}_i = (0, \ldots, \underbrace{1}_{i\text{-}th}, \ldots, 0)'$. The easy invertibility of $\boldsymbol{D}_{\tilde{\boldsymbol{C}}_v}$ allows us to use the Sherman-Woodbury-Morrison formulas (e.g. Henderson and Searle, 1981) to compute the inverse and determinant of $\boldsymbol{C}_{\tilde{\boldsymbol{w}}} + \boldsymbol{I} \otimes \boldsymbol{\Psi}$ in $O(nm^3)$ operations.

3.5.2 Prediction

Once the parameters have been estimated, inferential interest turns to spatial prediction. Here, a few situations are of interest. Let $\boldsymbol{s}_0$ be any location in the domain, where we seek to predict $\boldsymbol{y}(\boldsymbol{s}_0)$, an $m \times 1$ vector, and are given an $m \times p$ matrix of covariates $\boldsymbol{X}(\boldsymbol{s}_0)$. Let $\tilde{\boldsymbol{v}}_0 = \{\tilde{v}_k(\boldsymbol{s}_0) : k = 1, \ldots, m\}, \tilde{\boldsymbol{\theta}}_{\boldsymbol{v}} = \{\boldsymbol{\theta}_k : k = 1, \ldots, m\}, \tilde{\boldsymbol{a}}_0 = \{\tilde{a}_{i,j}(\boldsymbol{s}_0) : i, j = 1, \ldots, m, i \geq j\}$ and $\tilde{\boldsymbol{\theta}}_{\boldsymbol{a}} = \{\boldsymbol{\theta}_{a;i,j} : i, j = 1, \ldots, m, i \geq j\}$ For the unmarginalized model, prediction of $\boldsymbol{y}(\boldsymbol{s}_i)$ is achieved by sampling from the posterior predictive distribution

$$[\boldsymbol{y}(\boldsymbol{s}_0) \,|\, \boldsymbol{y}] = \int [\boldsymbol{y}(\boldsymbol{s}_0) \,|\, \boldsymbol{y}, \boldsymbol{\beta}, \tilde{\boldsymbol{a}}_0, \tilde{\boldsymbol{v}}_0, \boldsymbol{\Psi}][\tilde{\boldsymbol{v}}_0 \,|\, \tilde{\boldsymbol{\theta}}_{\boldsymbol{v}}, \boldsymbol{Y}][\tilde{\boldsymbol{a}}_0 \,|\, \tilde{\boldsymbol{\theta}}_{\boldsymbol{a}}, \boldsymbol{y}][\boldsymbol{\beta}, \tilde{\boldsymbol{\theta}}_{\boldsymbol{v}}, \tilde{\boldsymbol{\theta}}_{\boldsymbol{a}}, \boldsymbol{\Psi} \,|\, \boldsymbol{y}] \qquad (3.17)$$

The sampling is done using *composition sampling*: for $\{\boldsymbol{\beta}^{(l)}, \tilde{\boldsymbol{\theta}}_{\boldsymbol{v}}^{(l)}, \tilde{\boldsymbol{\theta}}_{\boldsymbol{a}}^{(l)}, \boldsymbol{\Psi}^{(l)}\}, l = 1, 2, \ldots, L$, drawn from the posterior distribution $[\boldsymbol{\beta}, \tilde{\boldsymbol{\theta}}_{\boldsymbol{v}}, \tilde{\boldsymbol{\theta}}_{\boldsymbol{a}}, \boldsymbol{\Psi} \,|\, \boldsymbol{Y}]$, we draw $\tilde{\boldsymbol{v}}_0^{(l)}$ from $[\tilde{\boldsymbol{v}}_0 \,|\, \boldsymbol{y}, \tilde{\boldsymbol{\theta}}_{\boldsymbol{v}}^{(l)}]$ and $\tilde{\boldsymbol{a}}_0^{(l)}$ from $[\tilde{\boldsymbol{a}}_0 \,|\, \boldsymbol{y}, \tilde{\boldsymbol{\theta}}_{\boldsymbol{a}}^{(l)}]$. Next, we draw $\boldsymbol{y}(\boldsymbol{s}_0)^{(l)}$ from $[\boldsymbol{y}(\boldsymbol{s}_0) \,|\, \boldsymbol{y}, \boldsymbol{\beta}^{(l)}, \tilde{\boldsymbol{a}}_0^{(l)}, \tilde{\boldsymbol{v}}_0^{(l)}, \boldsymbol{\Psi}^{(l)}]$. The resulting $\boldsymbol{y}(\boldsymbol{s}_0)^{(l)}$, $l = 1, 2, \ldots, L$ is a sample from the desired posterior predictive distribution. This is especially simple for Gaussian likelihoods because the distribution $[\boldsymbol{y}(\boldsymbol{s}_0) \,|\, \boldsymbol{y}, \boldsymbol{\beta}, \tilde{\boldsymbol{a}}_0, \tilde{\boldsymbol{v}}_0, \boldsymbol{\Psi}]$ is a normal distribution with mean and variance given by $\boldsymbol{X}(\boldsymbol{s}_0)'\boldsymbol{\beta} + \sum_{k=1}^{m} \tilde{\boldsymbol{a}}_k(\boldsymbol{s}_0)\tilde{v}_k(\boldsymbol{s}_0)$ and $\boldsymbol{\Psi}$ respectively. For inference on the residual process, we carry out the first two steps outlined beforehand and draw samples $\{\tilde{\boldsymbol{v}}_0^{(l)}, \tilde{\boldsymbol{a}}_0^{(l)}\}, l = 1, \ldots, L$. This in turn helps us to construct $\tilde{\boldsymbol{w}}(\boldsymbol{s}_0)^{(l)} = \tilde{\boldsymbol{A}}(\boldsymbol{s}_0)^{(l)}\tilde{\boldsymbol{v}}(\boldsymbol{s}_0)^{(l)}, l = 1, \ldots, L$ which are precisely samples from the posterior distribution $[\tilde{\boldsymbol{w}}(\boldsymbol{s}_0) \,|\, \boldsymbol{y}]$.

3.5.3 Model selection

We assess model performance and subsequent comparisons by simulating *independent* replicates for each observed outcome. Specifically, for each observed location, $\boldsymbol{s}_i$, we compute $[\boldsymbol{y}_{rep}(\boldsymbol{s}_i) \,|\, \boldsymbol{y}] = \int [\boldsymbol{y}_{rep}(\boldsymbol{s}_i) \,|\, \boldsymbol{\beta}, \{\tilde{\boldsymbol{a}}_k(\boldsymbol{s}_i)\}, \{\tilde{v}_k(\boldsymbol{s}_i)\}, \boldsymbol{\Psi}][\boldsymbol{\beta}, \{\tilde{\boldsymbol{a}}_k(\boldsymbol{s}_i)\}, \{\tilde{v}_k(\boldsymbol{s}_i)\}, \boldsymbol{\Psi} \,|\, \boldsymbol{y}]$, where the distribution of $\boldsymbol{y}_{rep}(\boldsymbol{s}_i)$ is simply the likelihood component corresponding to $\boldsymbol{y}(\boldsymbol{s}_i)$ in (3.15). Letting $\boldsymbol{\mu}_{rep,i}$ and $\boldsymbol{\Sigma}_{rep,i}$ be the posterior predictive mean and variance for each $\boldsymbol{y}_{rep}(\boldsymbol{s}_i)$, we will prefer models that will perform well under a decision-theoretic balanced loss function, penalizing both departure of replicated means from their observed values (lack of fit) and excessive uncertainty in the replicated data. Using a squared error loss function (e.g. Gelfand and Ghosh, 1998), the measures for these two

criteria are evaluated as $G = \sum_{i=1}^{n} \|\boldsymbol{y}(\boldsymbol{s}_i) - \boldsymbol{\mu}_{rep,i}\|^2$, where $\|\cdot\|$ is the standard Euclidean norm, and $P = \sum_{i=1}^{n} \text{tr}(\boldsymbol{\Sigma}_{rep,i})$. We will use the score $D = G + P$ as a model selection criteria, with lower values of D indicating better models.

3.6 Analysis of data

We used both synthetic data and the soil nutrient dataset to assess the proposed models. Both the synthetic and real datasets were kept purposely small to allow for comparison between full and reduced dimension models. For these analyses, posterior inference was based on three chains of 25,000 iterations (the first 5,000 iterations were discarded as burn-in). The samplers detailed in Section 3.5.1 were coded in C++ and Fortran and leveraged Intel's Math Kernel Library threaded BLAS and LAPACK routines for matrix computations. The updating schemes for $\tilde{a}_{ij}(\boldsymbol{s})$'s and $\tilde{v}(\boldsymbol{s})$'s are discussed in the appendix B of Chapter 5. All analyses were conducted on a Linux workstation using two Intel Nehalem hyperthreaded quad-Xeon processors. For the synthetic data analysis, sampling the three chains of 25,000 iterations required approximately 43, 29, and 18 hours for the nonstationary full, 225 knot, and 100 knot models, respectively. To deliver the same number of samples for the soil nutrients data analysis required approximately 21, 11, and 7, hours for the nonstationary full, 48 knot, and 26 knot models, respectively.

3.6.1 Synthetic data

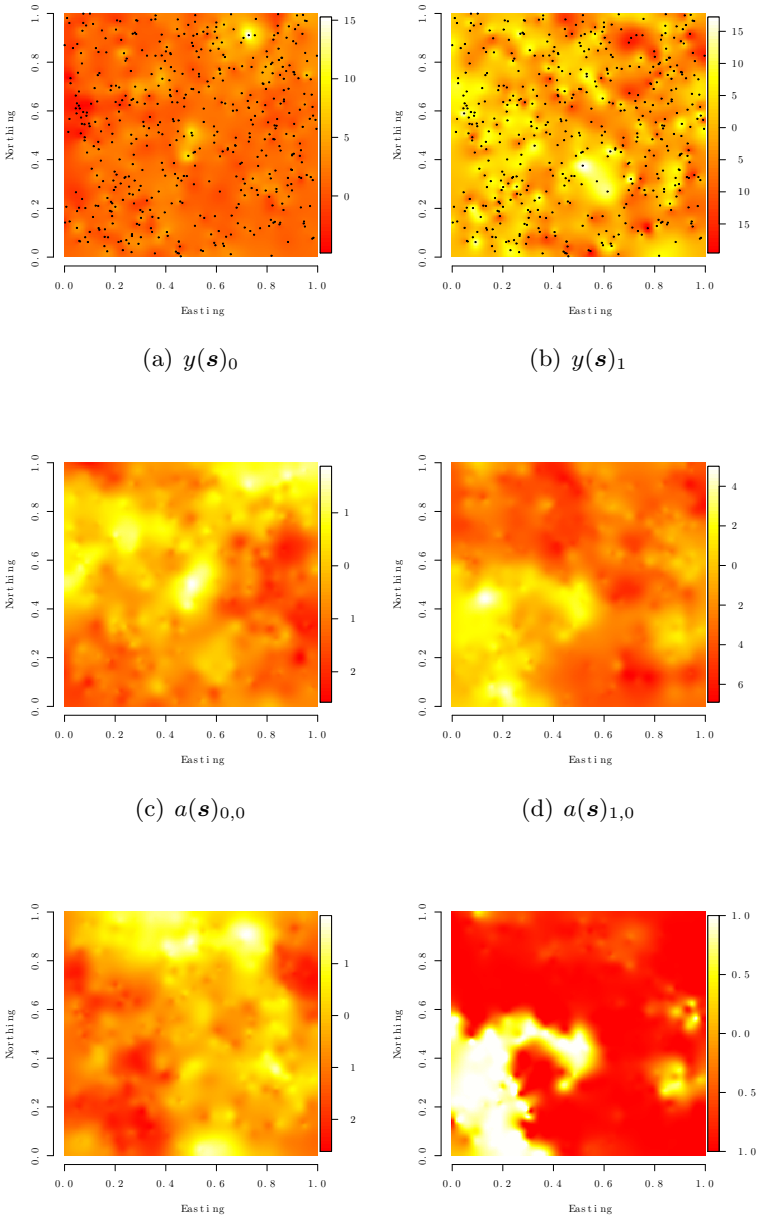
Here we use an analysis of synthetic data to explore the properties of the proposed models. The data set comprises n=500 locations distributed randomly within a 1×1 unit square domain. Two outcomes were generated at each location using (3.1). The column labeled *True* in Table 3.1 offers the mean and spatial covariance parameter values used to generate these data. The mean of each location was assumed to have a common intercept and single covariate drawn from $N(0, 1)$. The intercept and slope parameter associated with each outcome are labeled $\beta_{0,0}$, $\beta_{0,1}$, $\beta_{1,0}$, and $\beta_{1,1}$, where the first subscript refers to the outcome and the second to the corresponding column of the $\boldsymbol{X}$ matrix. An exponential spatial correlation function was assumed for all spatial processes, i.e., θ_2 was fixed at 0.5 in (3.5). Initial exploration of (3.1) and (3.14) suggested the data

poorly identified the parameters in $\boldsymbol{\theta}_1 = \{\theta_{1;k} : k = 1, \ldots, m\}$ and $\boldsymbol{\theta}_{1,\boldsymbol{a}} = \{\theta_{1,a;i,j} : i, j = 1, \ldots, m, i \geq j\}$. However, we found that specifying a common spatial range parameter for the $v_k(\boldsymbol{s})$'s and $a(\boldsymbol{s})_{ij}$'s, i.e., θ_1 and $\theta_{1,a}$, did produce data from which we could estimate these parameters. This concession still provides a sufficiently rich covariance structure.

The synthetic outcome surfaces are illustrated in Figure 3.2(a) and (b), respectively. Here too, surfaces of $\boldsymbol{A}(\boldsymbol{s})$ depict the space-varying nature of the outcomes' residual covariances. As described in Section 3.3, location specific covariance and subsequent correlation matrices can be calculated using $\boldsymbol{A}(\boldsymbol{s})$. This resulting residual spatial correlation surface is given in Figure 3.2(f) and shows regions of strong positive and negative residual spatial association between the outcomes.

Candidate models included three submodels of (3.1): stationary $\boldsymbol{w}(\boldsymbol{s}) = \boldsymbol{A}\boldsymbol{v}(\boldsymbol{s})$ ($\boldsymbol{A} = \boldsymbol{A}(\boldsymbol{s})$ that does not depende on $\boldsymbol{s}$); nonstationary $\boldsymbol{w}(\boldsymbol{s}) = \boldsymbol{A}(\boldsymbol{s})\boldsymbol{v}(\boldsymbol{s})$ (the *full* model); and the modified predictive process (nonstationary) version with $\tilde{\boldsymbol{w}}(\boldsymbol{s}) = \tilde{\boldsymbol{A}}(\boldsymbol{s})\tilde{\boldsymbol{v}}(\boldsymbol{s})$, i.e., model (3.14). Knot grids of 100 and 225 were considered for the predictive process model. Figure 3.3(a) shows the 100 knot grid. Parameter estimates and summaries of model fit are given in Table 3.1. Both the full and 225 knot predictive process nonstationary models estimate correctly the model parameters. However, several parameters are missed by the 100 knot model. Looking to goodness-of-fit, the relatively large D of the stationary model suggests the spatially invariant $\boldsymbol{A}$ is unable to accommodate the space-varying within location covariance. Interestingly, the predictive process models produced a better fit than the full model. This could be attributed to the former's enhanced adaptability to the data given the added non-stationarity of $\tilde{\boldsymbol{v}}(\boldsymbol{s})$ as opposed to its counterpart $\boldsymbol{v}(\boldsymbol{s})$. This added flexibility is achieved by $\tilde{\boldsymbol{v}}(\boldsymbol{s})$ while at the same time being more parsimonious due to the smaller number of "knots."

Comparing Figure 3.2 and 3.3, it is apparent that reducing the dimensionality from 500 observations to 100 knots does not greatly degrade the estimates of the fitted values and covariance among the $\tilde{a}_{ij}(\boldsymbol{s})$'s. As noted in Section 3.1, ecologists are often interested testing hypotheses regarding the strength, sign, and space-varying nature of the correlation among the residuals. Samples from the posterior distribution of this correlation are collected using $\rho^{(l)}_{i,j}(\boldsymbol{s}) = \tilde{a}^{(l)}_{i,j}(\boldsymbol{s})/(\tilde{a}^{(l)}_{i,i}(\boldsymbol{s})\tilde{a}^{(l)}_{j,j}(\boldsymbol{s}))^{1/2}$, for $l = 1, 2, \ldots, L$ samples, and when summarized can be used to identify locations where the correlation differs



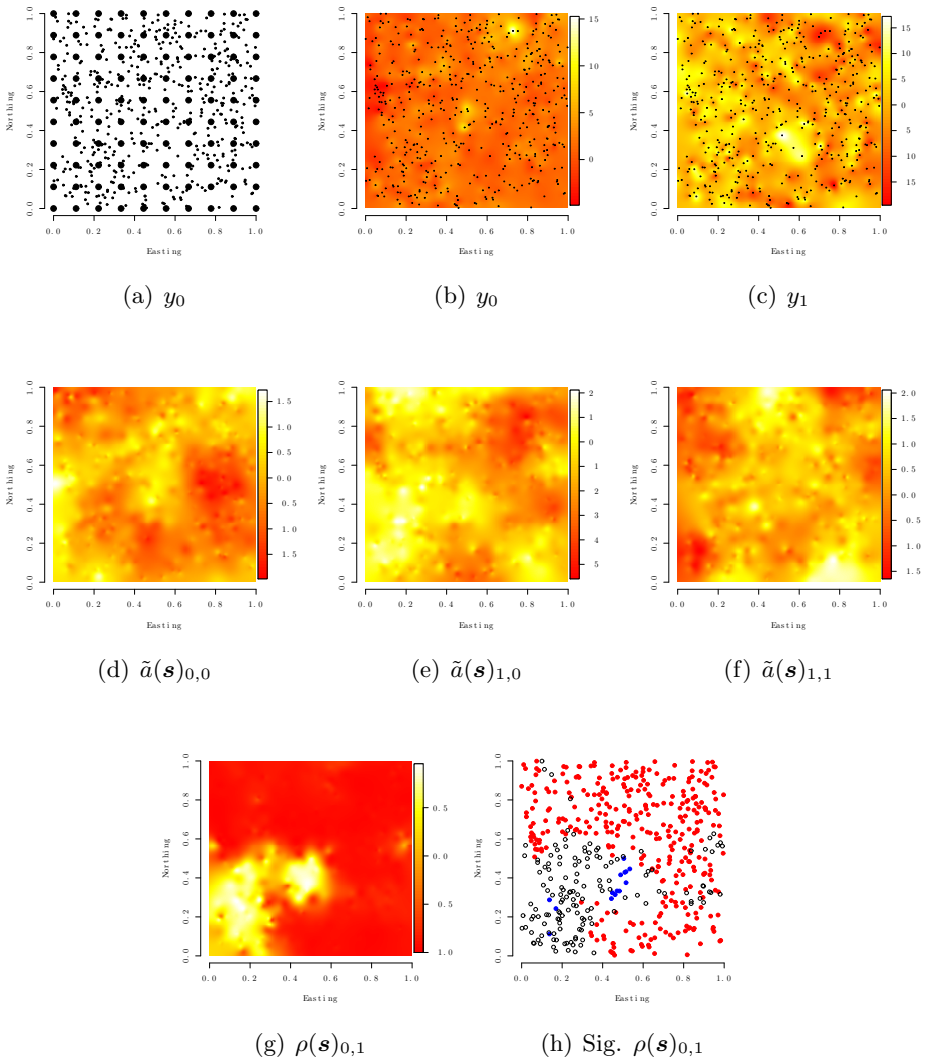
(a) $y(\boldsymbol{s})_0$ (b) $y(\boldsymbol{s})_1$

(c) $a(\boldsymbol{s})_{0,0}$ (d) $a(\boldsymbol{s})_{1,0}$

(e) $a(\boldsymbol{s})_{1,1}$ (f) $\rho(\boldsymbol{s})_{0,1}$

Figure 3.2: Interpolated surfaces of the *true* synthetic data.

from zero at some pre-specified level. The median $\rho_{0,1}(\boldsymbol{s})$ and locations where $\rho_{0,1}(\boldsymbol{s})$ differs significantly from zero are illustrated in Figures 3.3(g) and (h).



(a) y_0 (b) y_0 (c) y_1

(d) $\tilde{a}(\boldsymbol{s})_{0,0}$ (e) $\tilde{a}(\boldsymbol{s})_{1,0}$ (f) $\tilde{a}(\boldsymbol{s})_{1,1}$

(g) $\rho(\boldsymbol{s})_{0,1}$ (h) Sig. $\rho(\boldsymbol{s})_{0,1}$

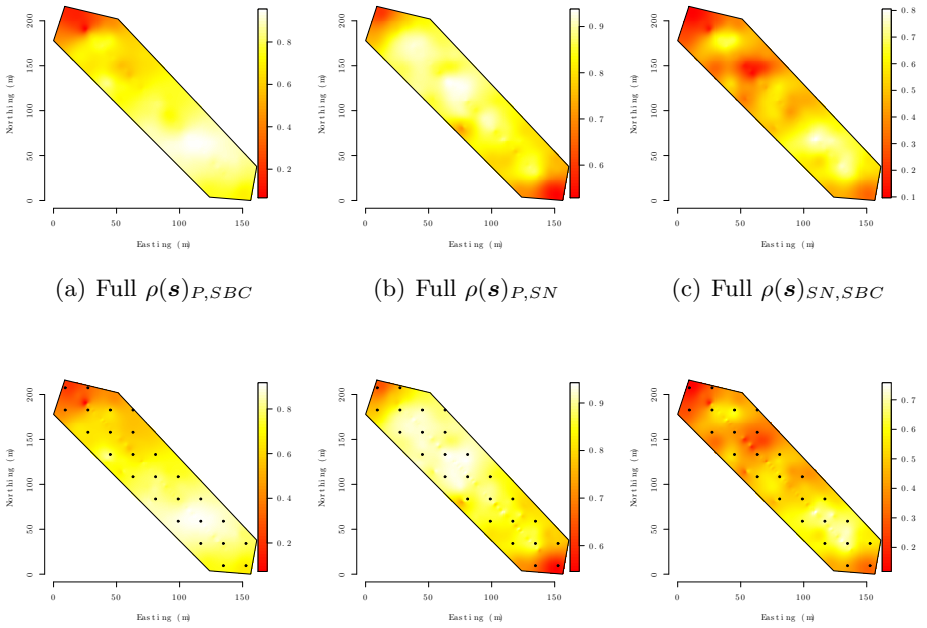
Figure 3.3: Synthetic data observed locations and 100 knot candidate model's predictive process knots, small and large points in (a), respectively. Interpolated surfaces of the 100 knot model's median posterior fitted values, space-varying elements of $\tilde{\boldsymbol{A}}(\boldsymbol{s})$, and associated residual spatial correlation. Locations in (h) depict residual spatial correlation that are significantly different from zero at the 0.1 level with negative and positive correlations identified in red and blue, respectively.

3.6.2 Analysis of soil nutrients data

Similar to the synthetic data analysis, three submodels of (3.1) were considered for the La Selva Biological Station soil nutrient dataset detailed in Section 3.2, and included the: non-predictive process and stationary $\boldsymbol{w}(\boldsymbol{s}) = \boldsymbol{A}\boldsymbol{v}(\boldsymbol{s})$; non-predictive process and nonstationary $\boldsymbol{w}(\boldsymbol{s}) = \boldsymbol{A}(\boldsymbol{s})\boldsymbol{v}(\boldsymbol{s})$ (*full* model), and; modified predictive process and nonstationary $\tilde{\boldsymbol{w}}(\boldsymbol{s}) = \tilde{\boldsymbol{A}}(\boldsymbol{s})\tilde{\boldsymbol{v}}(\boldsymbol{s})$, i.e., model (3.14). An exponential spatial correlation function was assumed for all GP's. Knot grids of 26 and 48 were considered for the predictive process model. Parameter estimates and summaries of model fit are given in Table 3.2. Both the full and predictive process nonstationary models produced nearly identical parameter estimates, with the exception of slightly longer spatial ranges for θ_1 (defined as the distance at which the correlation drops to 0.05, i.e., -log(0.05)/θ_1) in the predictive process models. Looking to goodness-of-fit, the nonstationary models provided better fit, i.e., lower D, than the stationary model. Here again, as in the synthetic example, we see that the predictive process models produce a marginally lower D, than the full nonstationary model.

The posterior distribution median for $\rho_{P,SBC}(\boldsymbol{s})$, $\rho_{P,SN}(\boldsymbol{s})$, and $\rho_{SN,SBC}(\boldsymbol{s})$ from the full and 26 knot predictive process models are illustrated in Figure 3.4, rows one and two, respectively. Again, samples from these posterior distributions are drawn via composition sampling, e.g., $\rho^{(l)}_{P,SBC}(\boldsymbol{s}) = \tilde{a}^{(l)}_{P,SBC}(\boldsymbol{s})/(\tilde{a}^{(l)}_{P,P}(\boldsymbol{s})\tilde{a}^{(l)}_{SBC,SBC}(\boldsymbol{s}))^{1/2}$ for $l = 1, 2, \ldots, L$ samples. These surfaces suggest the strength of correlation among the outcomes does vary across the domain. Specifically, there are regions of near zero correlation and regions of strong correlation, particularly between P-SBC and P-SN. Further, the similarity between the surfaces in the first and second row of Figure 3.4 suggest that dimension reduction from the full model to the 26 knot predictive process model does not oversmooth the fitted values.

More formal inference about these patterns are offered in Figure 3.5. Here, those locations with residual spatial correlations differing significantly from zero are identified by filled circle symbols and those that do not differ from zero are identified with open circle symbols. Here, we see significant positive correlation between P and SBC along the southern two-thirds of the central transect. Phosphorus and SN show a significant positive correlation along most of the transect. In contrast, very few sample locations show significant correlation between SN and SBC. Importantly, this figure shows that



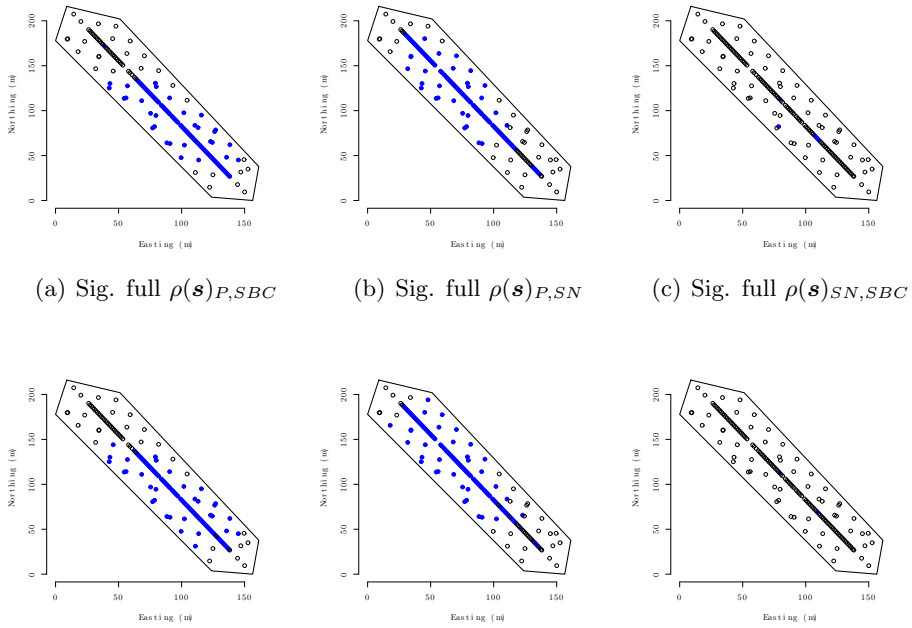
(a) Full $\rho(\boldsymbol{s})_{P,SBC}$ (b) Full $\rho(\boldsymbol{s})_{P,SN}$ (c) Full $\rho(\boldsymbol{s})_{SN,SBC}$

(d) Pred. proc. $\rho(\boldsymbol{s})_{P,SBC}$ (e) Pred. proc. $\rho(\boldsymbol{s})_{P,SN}$ (f) Pred. proc. $\rho(\boldsymbol{s})_{SN,SBC}$

Figure 3.4: Interpolated surfaces of the soil nutrient data residual spatial correlations for the nonstationary full and 26 knot predictive process models. Knot locations are overlaid on (d-f).

the 26 knot predictive process nonstationary model delivers inference comparable to that of the full nonstationary model.

Overall, these patterns indicate strong biogeochemical coupling between SN and P availability at a local scale. This supports the hypothesis that nitrogen influences the availability of P, presumably due to the importance of nitrogen as a constituent of extracellular phosphatases secreted by microbes, which in turn increases P availability (Houlton et al., 2008). Despite the local scale positive correlation between SN and P, it is important to note that at a coarser scale (i.e., across soil formations of different ages), natural soil weathering processes lead to a negative correlation between nitrogen and P (Walker and Syers 1976, Wardle et al., 2004). Younger soils are generally abundant in P and deficient in nitrogen and vice-versa for older soils. Thus the strong local positive correlations run counter to the expected negative correlation based on soil weathering



(a) Sig. full $\rho(\boldsymbol{s})_{P,SBC}$ (b) Sig. full $\rho(\boldsymbol{s})_{P,SN}$ (c) Sig. full $\rho(\boldsymbol{s})_{SN,SBC}$

(d) Sig. pred. proc. $\rho(\boldsymbol{s})_{P,SBC}$ (e) Sig. pred. proc. $\rho(\boldsymbol{s})_{P,SN}$ (f) Sig. pred. proc. $\rho(\boldsymbol{s})_{SN,SBC}$

Figure 3.5: Locations depict residual spatial correlation that are significantly different from zero at the 0.05 level with negative and positive correlations identified in red and blue, respectively, nonstationary full and 26 knot predictive process models.

processes alone. In contrast to SN and P, spatially variable correlations of base cations with SN and P suggest that local variability usurps strong coupling, perhaps due to species-specific variation among individual trees in their influence on base cations (Finzi et al., 1998).

For both the full and predictive process models, the correlation between P-SBC and P-SN are not significant at the ends of the transect. We attribute this to an *edge effect* – there are too few locations to estimate precisely the local spatial cross-covariance. This edge effect can be clearly seen in Figure 3.6, which shows the width of the residual spatial correlations' 95% credible interval increases substantially at the ends of the transect. These wider credible intervals are more likely to include zero. Therefore, we need to be cautious about making inference on the space-varying nature of association among

outcomes in regions with insufficient support. Figure 3.7 shows the nonstationary full and 26 knot predictive process models predicted outcomes at a fine spatial resolution over the transect with values equal to each location's posterior predictive distribution median. Similar to the fitted value surfaces, the values predicted using the 26 knot predictive process model are nearly indistinguishable from the full model.

3.7 Discussion and summary

Much of the existing multivariate spatial model literature has focused upon applications and situations where the cross-covariances do not vary over the spatial domain. With rapid advances in the collection and assimilation of spatial data in the natural sciences, there is a burgeoning need for models that represent more complex nonstationary behavior in cross-covariances. While theoretical tools for constructing nonstationary cross-covariance matrices are available, the resulting hierarchical models are usually computationally prohibitive – even for moderately sized datasets. In particular, they entail an explosion of process realizations resulting in a very large parameter space, which leads to severe difficulties in statistical estimation and inference.

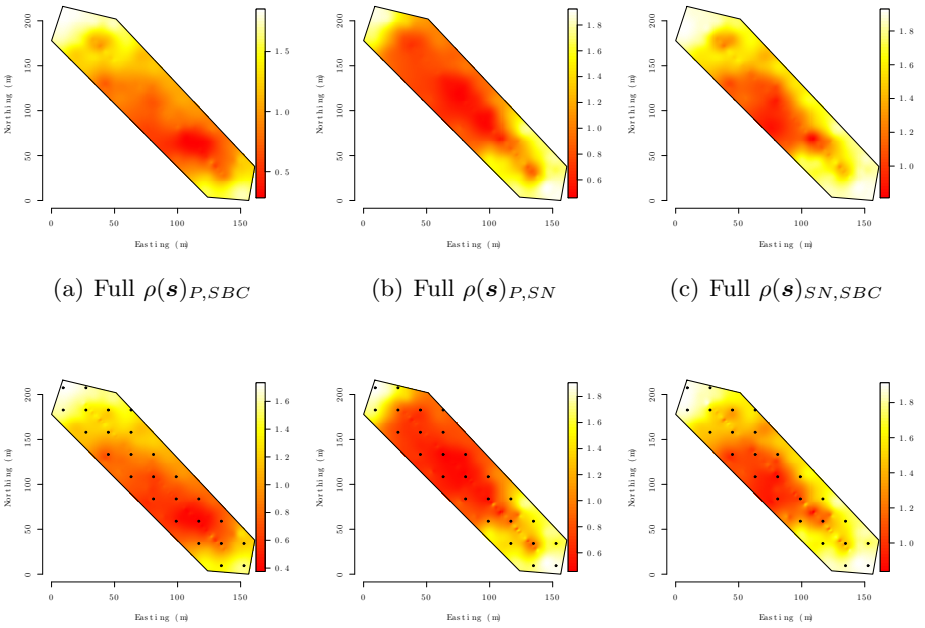
Here, we have proposed a class of low-rank spatially-varying cross-covariance functions that produce non-degenerate or non-singular multivariate spatial process models. We use a constructive approach that has been shown to be effective in stationary settings where the cross-covariances do not vary spatially. The idea is to express each component of a vector process as a spatially adaptive linear transformation of a set of independent latent processes. The coefficients of these latent processes are themselves modeled as low-rank processes. We use the Gaussian predictive process as a low rank process, which easily allows us to modify the process and deliver non-degenerate low-rank processes.

We discuss two different approaches for modeling spatially-varying cross-covariance matrices using the predictive process and explore the relationship between the marginal cross-covariances of these processes apriori. Attractively, the spatial associations within a location are seen to be identical for the two processes. Another appeal of our method is that one need not digress from the modeling objectives to think about choices of basis functions, or kernels or alignment algorithms for the locations. Further, no new parameters are introduced and the cross-covariance function of the modified predictive

process derives directly from that of the parent process.

As illustrated by La Selva Biological Station soil nutrients analysis, the spatially-varying Linear Model Coregionalization framework offers ecologists a tool for exploring the correlation of biotic and/or abiotic variables within and across large inventory or monitoring datasets. Here, the space varying residual patterns among important nutrients provided insight into the biogeochemical processes that influence soil nutrients. Importantly, the synthetic and soil nutrients analyses suggested that very little information was lost when moving to a predictive process model. This, however, is likely data-specific and we recommend that a range of predictive process knots be explored to judge the robustness of model inference. The soil nutrients analysis also identified the potential for an edge effect, which in regions with sparse observations could affect inference about the space-varying nature of the outcomes.

A possibly more relevant investigation will pursue situations where the number of outcomes is much larger than encountered here. Examples include joint modeling of a large number of tree-species. A natural adaptation of our current work will involve low-rank factor models, where we achieve dimension reduction in the number of spatial locations as well as the number of outcomes. How low-rank factor loadings can be modeled in a space-varying manner and the nature of identifiability issues this will entail, as opposed to usual factor models, also presents theoretical and methodological challenges that we seek to pursue in future work.



(a) Full $\rho(\boldsymbol{s})_{P,SBC}$ (b) Full $\rho(\boldsymbol{s})_{P,SN}$ (c) Full $\rho(\boldsymbol{s})_{SN,SBC}$

(d) Pred. proc. $\rho(\boldsymbol{s})_{P,SBC}$ (e) Pred. proc. $\rho(\boldsymbol{s})_{P,SN}$ (f) Pred. proc. $\rho(\boldsymbol{s})_{SN,SBC}$

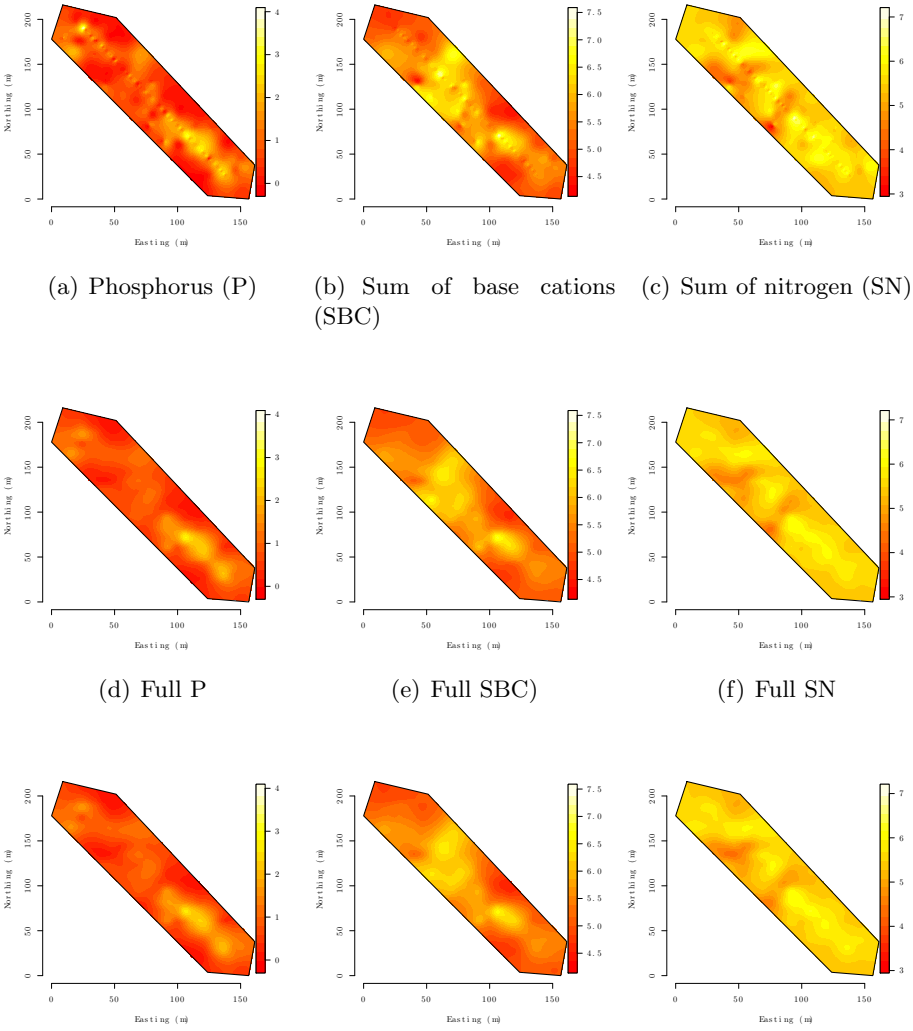
Figure 3.6: Interpolated surfaces of the width of the 95% soil nutrient data residual spatial correlations for the nonstationary full and 26 knot predictive process models. Knot locations are overlaid on (d-f).

Table 3.1: Parameter credible intervals, 50 (2.5, 97.5) percentiles, for the synthetic data analysis candidate models. Bold indicate that the 95% credible interval does not include the *true* parameter value.

			Nonstationary		
				Predictive process	
Parameter	True	Stationary	Full	225	100
$\beta_{0,0}$	1	1.42 (0.89, 2.00)	1.04 (0.87, 1.21)	1.04 (0.86, 1.22)	1.10 (0.91, 1.28)
$\beta_{0,1}$	1	1.02 (0.94, 1.11)	1.01 (0.94, 1.08)	1.00 (0.94, 1.07)	1.01 (0.94, 1.07)
$\beta_{1,0}$	1	-0.50 (-1.43, 1.11)	0.66 (0.16, 1.21)	0.60 (0.15, 1.16)	**0.45 (-0.17, 0.94)**
$\beta_{1,1}$	5	5.03 (4.89, 5.17)	5.03 (4.92, 5.14)	5.04 (4.93, 5.14)	5.04 (4.92, 5.15)
$\boldsymbol{AA}'_{0,0}$	–	2.19 (1.61, 3.13)	–	–	–
$\boldsymbol{AA}'_{1,0}$	–	-1.91 (-3.08, -1.28)	–	–	–
$\boldsymbol{AA}'_{1,1}$	–	11.19 (8.42, 15.99)	–	–	–
$\text{var}(\tilde{a}_{0,0})$	1	–	0.88 (0.53, 1.52)	0.90 (0.52, 1.49)	0.84 (0.52, 1.38)
$\text{var}(\tilde{a}_{1,0})$	5	–	6.90 (4.11, 11.92)	5.42 (3.08, 9.54)	4.82 (2.29, 9.04)
$\text{var}(\tilde{a}_{1,1})$	1	–	0.42 (0.17, 1.27)	0.41 (0.15, 1.09)	**0.45 (0.21, 0.98)**
$\theta_{1,a}$	4	–	4.66 (3.15, 6.92)	4.13 (3.07, 5.95)	3.66 (3.03, 5.35)
θ_1	6	**9.20 (6.27, 12.53)**	5.60 (3.54, 8.70)	6.59 (3.86, 9.66)	6.86 (4.00, 10.64)
$\Psi_{0,0}$	0.5	0.53 (0.37, 0.72)	0.44 (0.35, 0.53)	0.42 (0.33, 0.52)	0.42 (0.33, 0.52)
$\Psi_{1,1}$	0.5	0.37 (0.16, 0.75)	0.36 (0.19, 0.63)	0.29 (0.14, 0.53)	**0.24 (0.12, 0.46)**
$\boldsymbol{G}$		156.64	174.76	149.03	140.02
$\boldsymbol{P}$		765.43	636.63	569.88	533.37
$\boldsymbol{D}$		922.07	811.4	718.9	673.39

Table 3.2: Parameter credible intervals, 50 (2.5 97.5) percentiles, for soil nutrient data analysis candidate models.

		Nonstationary		
			Predictive process	
Parameter	Stationary	Full	48	26
$\beta_{0,P}$	0.71 (0.26, 1.35)	0.66 (0.22, 1.05)	0.76 (0.37, 1.17)	0.64 (0.33, 1.20)
$\beta_{0,SBC}$	5.38 (5.03, 6.08)	5.18 (4.86, 5.49)	5.19 (4.97, 5.45)	5.16 (4.83, 5.40)
$\beta_{0,SN}$	5.42 (4.97, 5.86)	5.53 (5.30, 5.78)	5.57 (5.35, 5.89)	5.53 (5.31, 5.73)
$\boldsymbol{AA}'_{P,P}$	0.92 (0.52, 2.29)	–	–	–
$\boldsymbol{AA}'_{SBC,P}$	0.47 (0.25, 1.23)	–	–	–
$\boldsymbol{AA}'_{SN,P}$	0.49 (0.26, 1.25)	–	–	–
$\boldsymbol{AA}'_{SBC,SBC}$	0.44 (0.27, 1.08)	–	–	–
$\boldsymbol{AA}'_{SN,SBC}$	0.19 (0.06, 0.51)	–	–	–
$\boldsymbol{AA}'_{SN,SN}$	0.39 (0.19, 1.08)	–	–	–
var($\tilde{a}_{P,P}$)	–	0.20 (0.09, 0.53)	0.22 (0.09, 0.52)	0.22 (0.08, 0.57)
var($\tilde{a}_{SBC,P}$)	–	0.24 (0.10, 0.63)	0.25 (0.10, 0.72)	0.21 (0.10, 0.54)
var($\tilde{a}_{SN,P}$)	–	0.20 (0.09, 0.50)	0.23 (0.09, 0.60)	0.23 (0.11, 0.75)
var($\tilde{a}_{SBC,SBC}$)	–	0.54 (0.18, 1.64)	0.40 (0.15, 1.05)	0.36 (0.13, 1.01)
var($\tilde{a}_{SN,SBC}$)	–	0.14 (0.06, 0.36)	0.16 (0.06, 0.43)	0.15 (0.07, 0.38)
var($\tilde{a}_{SN,SN}$)	–	1.85 (0.62, 6.11)	1.40 (0.28, 27.56)	1.77 (0.41, 10.38)
$\theta_{1,a}$	–	0.0135 (0.0125, 0.0173)	0.0134 (0.0125, 0.0177)	0.0134 (0.0125, 0.0170)
θ_1	0.0499 (0.0165, 0.0873)	0.0371 (0.0180, 0.0737)	0.0264 (0.0136, 0.0612)	0.0284 (0.0133, 0.0603)
Eff. range$_a$ m	–	222.13 (173.32, 238.97)	223.20 (169.70, 238.89)	224.36 (176.57, 239.17)
Eff. range$_v$ m	60.04 (34.31, 181.08)	80.68 (40.66, 166.33)	113.28 (48.95, 220.69)	105.64 (49.65, 225.05)
$\Psi_{P,P}$	0.21 (0.14, 0.30)	0.19 (0.13, 0.28)	0.20 (0.13, 0.29)	0.19 (0.13, 0.28)
$\Psi_{SBC,SBC}$	0.07 (0.05, 0.11)	0.06 (0.04, 0.09)	0.06 (0.04, 0.09)	0.06 (0.04, 0.09)
$\Psi_{SN,SN}$	0.15 (0.11, 0.21)	0.11 (0.07, 0.16)	0.10 (0.07, 0.15)	0.09 (0.06, 0.14)
$\boldsymbol{G}$	39.45	28.02	26.62	24.4
$\boldsymbol{P}$	92.62	79.9	80.02	77.28
$\boldsymbol{D}$	132.07	107.92	106.65	101.68



(a) Phosphorus (P) (b) Sum of base cations (SBC) (c) Sum of nitrogen (SN)

(d) Full P (e) Full SBC) (f) Full SN

(g) Pred. proc. P (h) Pred. proc. SBC (i) Pred. proc. SN

Figure 3.7: Interpolated surfaces of the observed soil nutrient outcomes (a-c) and predicted outcomes at a fine spatial resolution using the nonstationary full and 26 knot predictive process models, (d-f) and (g-i), respectively.

Chapter 4

On the residual spatial process from multivariate hierarchical low rank models

4.1 Low-rank spatial models and related biases

4.1.1 Biases in low rank models

In this chapter we extend earlier work on "low-rank" methods, such as *predictive process* models, to accomodate analysis of large spatial dataset. From Chapters 2 & 3 it is evident that given a spatial data, one would naturally tend to fit a high dimensional Bayesian hierarchical model as in (3.1). This will imply a posterior distribution of the parameters as

$$p(\boldsymbol{\Theta}, \boldsymbol{\beta}, \boldsymbol{w} \,|\, \boldsymbol{y}) \propto p(\boldsymbol{\Theta}) \times N(\boldsymbol{\beta} \,|\, \boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \times N(\boldsymbol{w} \,|\, \boldsymbol{0}, \boldsymbol{C}_w) \\ \times \prod_{i=1}^{n} N(\boldsymbol{y}(\boldsymbol{s}_i) \,|\, \boldsymbol{X}(\boldsymbol{s}_i)'\boldsymbol{\beta} + \boldsymbol{w}(\boldsymbol{s}_i), \boldsymbol{\Psi}) \,, \tag{4.1}$$

where, $\boldsymbol{\Theta} = (\boldsymbol{\Theta}_1, \boldsymbol{\Psi})$, $\boldsymbol{\Theta}_1 = (\boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_m)$. However, large datasets would render estimation of (4.1) infeasible, as discussed in Chapters 2 & 3.

Low-rank spatial process models attempt to circumvent the above computational bottleneck by replacing the *parent* process $\boldsymbol{w}(\boldsymbol{s})$ with a cheaper representation based upon a set of *knots* or *centers*, $\mathcal{S}^* = \{\boldsymbol{s}_1^*, \boldsymbol{s}_2^*, \ldots, \boldsymbol{s}_{n^*}^*\}$ with $n^* << n$ and the knots chosen to provide an adequate representation of the domain. To elucidate, suppose, without much loss of generality, that $\mathcal{S} \cap \mathcal{S}^* = \mathcal{S}^*$ with the first n^* locations in $\mathcal{S}$ acting as the knots. Note that the Gaussian likelihood with the parent process in (4.1) can be written as $N(\boldsymbol{y} \,|\, \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}_w\boldsymbol{u}, \boldsymbol{\Psi})$, where $\boldsymbol{Z}_w'$ is the $mn \times mn$ lower-triangular Cholesky square-root matrix of $\boldsymbol{C}_w$ and $\boldsymbol{u} = (u_1, u_2, \ldots, u_{mn})'$ is now an $mn \times 1$ vector such that $u_i \overset{iid}{\sim} N(0,1)$. Note that the covariance matrix of $\boldsymbol{w}$ can be expressed in terms of the submatrices of $\boldsymbol{Z}_w$ as

$$\boldsymbol{C}_w = \boldsymbol{Z}_w\boldsymbol{Z}_w' = \begin{pmatrix} \boldsymbol{Z}_{11,w}' & \boldsymbol{0} \\ \boldsymbol{Z}_{12,w}' & \boldsymbol{Z}_{22,w}' \end{pmatrix} \begin{pmatrix} \boldsymbol{Z}_{11,w} & \boldsymbol{Z}_{12,w} \\ \boldsymbol{0} & \boldsymbol{Z}_{22,w} \end{pmatrix} \tag{4.2}$$

where $\boldsymbol{Z}_{11,w}$ is $mn^* \times mn^*$, $\boldsymbol{Z}_{12,w}$ is $mn^* \times (mn - mn^*)$ and $\boldsymbol{Z}_{22,w}$ is $(mn - mn^*) \times (mn - mn^*)$. Writing $\boldsymbol{Z}_w = [\boldsymbol{Z}_{1w} : \boldsymbol{Z}_{2w}]$, where $\boldsymbol{Z}_{1w} = \begin{pmatrix} \boldsymbol{Z}_{11,w}' \\ \boldsymbol{Z}_{12,w}' \end{pmatrix}$, a low-rank model would work with the likelihood $N(\boldsymbol{y} \,|\, \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}_{1w}\boldsymbol{u}_1, \boldsymbol{\Psi})$, where $\boldsymbol{u}_1$ is an $mn^* \times 1$ vector whose components are independently and identically distributed $N(0,1)$ variables. Dimension reduction occurs because the matrix decompositions for estimating hierarchical models with this low-rank likelihood can be implemented with $mn^* \times mn^*$ (instead of $mn \times mn$) matrix decompositions.

More generally, low-rank representations replace the parent process in (4.1) with some linear combination of the form $\boldsymbol{w} \approx \boldsymbol{Z}_1\boldsymbol{u}$, where $\boldsymbol{Z}_1$ is an $mn \times mn^*$ matrix whose (i,j)-th block is given by $\boldsymbol{z}(\boldsymbol{s}_i, \boldsymbol{s}_j^*)$ and $\boldsymbol{u} = (u_1, u_2, \ldots, u_{mn^*})'$ is an $mn^* \times 1$ random vector. In semiparametric regression models using splines, the $\boldsymbol{z}(\boldsymbol{s}, \boldsymbol{s}_j^*)$'s are called the *basis functions* and u_j's the *basis coefficients*. Models for the $\boldsymbol{z}(\boldsymbol{s}, \boldsymbol{s}_j^*)$'s and u_j's include kernel convolutions motivated by an integral representation of (certain) stationary processes as a kernel convolution of Brownian motion on $\Re^2$ (e.g, Higdon, 2002; Xia and Gelfand, 2005), functional forms motivated from the specific application (Stein, 2007, 2008), estimating the basis functions using empirical estimates from the data (Cressie and Johannesson, 2008), and deriving them from optimal projections of the parent process (e.g. Banerjee et. al. 2008).

Irrespective of their precise specifications, low-rank models tend to overestimate the residual variance. This bias arises from systemic over-smoothing or model under-specification by the low-rank model when compared to the parent model. In fact, this becomes especially transparent from writing the parent likelihood and low-rank likelihood derived from (4.2) as mixed linear models (see Ruppert et. al, 2003),

$$\begin{aligned} \textbf{Parent likelihood:} \quad & \boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z}_{1w}\boldsymbol{u}_1 + \boldsymbol{Z}_{2w}\boldsymbol{u}_2 + \boldsymbol{\epsilon}_1\,;\ \boldsymbol{\epsilon}_1 \sim N(\boldsymbol{0}, \boldsymbol{I} \otimes \boldsymbol{\Psi}); \\ \textbf{Low rank likelihood:} \quad & \boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Z}_{1w}\boldsymbol{u}_1 + \boldsymbol{\epsilon}_2\,; \quad \boldsymbol{\epsilon}_2 \sim N(\boldsymbol{0}, \boldsymbol{I} \otimes \boldsymbol{\Psi})\ . \end{aligned}$$

For fixed $\boldsymbol{\beta}$ and $\boldsymbol{\Theta}_1$, the basis functions forming the columns of $\boldsymbol{Z}_{2w}$ in the parent likelihood are absorbed into the residual error in the low rank likelihood, leading to an upward bias in the estimate of the nugget. More precisely, letting $\boldsymbol{P}_{\boldsymbol{Z}_w} = \boldsymbol{Z}_w(\boldsymbol{Z}_w'\boldsymbol{Z}_w)^{-1}\boldsymbol{Z}_w'$ (the orthogonal projection matrix or "hat" matrix into the column space of $\boldsymbol{Z}_w$), standard linear model calculations reveal that the residual variability from the parent model is quantified by $(\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{Z}_w})(\boldsymbol{y} - \boldsymbol{X\beta})$, while that from the low-rank model is given by $(\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{Z}_{1w}})(\boldsymbol{y} - \boldsymbol{X\beta})$. Using the fact that $\boldsymbol{P}_{\boldsymbol{Z}_w} = \boldsymbol{P}_{\boldsymbol{Z}_{1w}} + \boldsymbol{P}_{[(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{Z}_{1w}})\boldsymbol{Z}_{2w}]}$, the excess residual variability in the low-rank likelihood appears as $(\boldsymbol{y} - \boldsymbol{X\beta})'\boldsymbol{P}_{[(\boldsymbol{I} - \boldsymbol{P}_{\boldsymbol{Z}_{1w}})\boldsymbol{Z}_{2w}]}(\boldsymbol{y} - \boldsymbol{X\beta})$.

Although this excess residual variability can be quantified as above, it is less clear how the low-rank spatial likelihood could be modified to compensate for this overestimation without adding significantly to the computational burden. Matters are complicated by the fact that expressions for the excess variability involve the unknown process parameters $\boldsymbol{\Theta}_1$, which must be estimated. To characterize this bias and provide a computationally feasible remedy for hierarchical models, it will be helpful to work with a low-rank spatial process rather than a low-rank likelihood. In this context, we remark that not all low-rank models lead to a straightforward quantification of bias. For instance, low-rank models based upon kernel convolutions (Higdon, 2002) approximate $\boldsymbol{w}(\boldsymbol{s})$ with $\boldsymbol{w}_{KC}(\boldsymbol{s}) = \sum_{j=1}^{n^*} \boldsymbol{k}(\boldsymbol{s} - \boldsymbol{s}_j^*)u_j$, where $\boldsymbol{k}(\cdot)$ is a $m \times 1$ vector of kernel functions and $u_j \overset{iid}{\sim} N(0,1)$, assumed to arise from a Brownian motion $U(\boldsymbol{v})$ on $\Re^2$. This yields

$$\boldsymbol{w}(\boldsymbol{s}) - \boldsymbol{w}_{KC}(\boldsymbol{s}) = \int \boldsymbol{k}(\boldsymbol{s} - \boldsymbol{v})dU(\boldsymbol{v}) - \sum_{j=1}^{n^*} \boldsymbol{k}(\boldsymbol{s} - \boldsymbol{s}_j^*)u_j \approx \sum_{j=n^*+1}^{\infty} \boldsymbol{k}(\boldsymbol{s} - \boldsymbol{s}_j^*)u_j\ , \quad (4.3)$$

which does not, in general, allow a closed form and may be difficult to compute accurately. Expression for the residual process is simplified for a special class of low-rank models, known as, predictive process models. The predictive process model and the bias incurred through it were discussed in Chapters 2 and 3. In the next section we propose a new class of models which can rectify the bias as well as significantly improve the predictive performance over the state of the art *modified predictive process*.

4.2 Tapered adjustment to predictive process models

We now undertake an exploration of model-based bias adjustments in low-rank models. Given the easier accessibility of predictive processes, we will restrict attention to the predictive process framework. The underlying idea, common to any strategy, is to incorporate additional spatially adaptive random effects to reckon with the oversmoothing by the reduced-rank process. We will, of course, need to ensure computational feasibility for each of the resulting processes.

The marginalized bias-adjusted predictive process, defined as modified predictive process in earlier chapters, is estimated by sampling from

$$p(\boldsymbol{\beta}, \boldsymbol{\Theta} \,|\, \boldsymbol{y}) \propto p(\boldsymbol{\Theta}) \times N(\boldsymbol{\beta} \,|\, \boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta) \times N(\boldsymbol{y} \,|\, \boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{y};mpp}) \,, \tag{4.4}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{y};mpp} = \mathcal{C}'_w \boldsymbol{C}^{*-1}_w \mathcal{C}_w + \boldsymbol{D}_{mpp}$, $\mathcal{C}'_w$ is the $mn \times mn^*$ matrix whose (i,j)-th element is $\mathcal{C}_w(\boldsymbol{s}_i, \boldsymbol{s}^*_j)$ and $\boldsymbol{D}_{mpp}$ is a block diagonal matrix with (i,i)-th block diagonal element $\boldsymbol{C}_w(\boldsymbol{s}_i, \boldsymbol{s}_i) - \mathcal{C}_w(\boldsymbol{s}_i, \mathcal{S}^*)' \boldsymbol{C}_w(\mathcal{S}^*)^{-1} \mathcal{C}_w(\boldsymbol{s}_i, \mathcal{S}^*) + \boldsymbol{\Psi}$. Estimation involves the inverse and determinant of $\boldsymbol{\Sigma}_{\boldsymbol{y};mpp}$, which can be efficiently computed using the Sherman-Woodbury-Morrison formula (e.g. Banerjee et al., 2010).

As an alternative to modified predictive process, a modification accruing computational benefits, *tapers* the process $\boldsymbol{w}(\boldsymbol{s}) - \boldsymbol{w}_{pp}(\boldsymbol{s})$. Tapered processes offer an alternative means of dimension reduction, by producing sparse spatial covariance matrices and have received much attention in the recent past (see, e.g., Furrer et al, 2006, Kaufman et al, 2009, Du et al, 2009, Sang et al, 2011). To make our presentation more clear, we initially focus on univariate ($m = 1$) tapering. In the latter part of this subsection, we will generalize our idea in multivariate processes.

The underlying idea of tapering is to use a compactly supported covariance function

(Gneiting, 1999) as a *tapering kernel* $C_\nu(\boldsymbol{s}_1, \boldsymbol{s}_2)$, which is a positive-definite function satisfying

$$C_\nu(\boldsymbol{s}_1, \boldsymbol{s}_2) = 0 \quad \text{if} \quad ||\boldsymbol{s}_1 - \boldsymbol{s}_2|| > \nu \ ,$$

where ν is the distance beyond which the covariance becomes zero. Now consider a covariance function obtained by taking product of $C_\nu(\cdot, \cdot)$ and any spatial covariance function $C_w(\cdot, \cdot)$, so $C_{tap}(\cdot, \cdot) = C_\nu(\cdot, \cdot)C_w(\cdot, \cdot)$. Tapering seeks to approximate inference by replacing covariance matrices based on $C_w(\cdot, \cdot)$ by those based on $C_{tap}(\cdot, \cdot)$. If one has enough reason to believe that there is very little or almost zero correlation among distant pair of observations, the approximation seems suitable and should incur acceptable loss of information. Even when distant pairs are highly correlated, they may be nearly independent conditional on their neighbors. This is a very interesting observation but hard to prove and is addressed throughout the geostatistical literature. Based on this principle, Vecchia (1988) proposed the likelihood approximation technique where the likelihood of an observed data $\boldsymbol{y}$ is approximated by the product of conditional densities $p(y(\boldsymbol{s}_i) \,|\, \boldsymbol{y}_{(i)})$, $\boldsymbol{y}_{(i)}$ being nearest neighbors of $\boldsymbol{s}_i$ among $\{\boldsymbol{s}_1, ..., \boldsymbol{s}_{i-1}\}$. Stein et. al. (2004) extended this idea to Restricted Maximum Likelihood Estimation (RMLE) and proposed different ways to select conditioning sets $\boldsymbol{y}_{(i)}$. A very similar idea drives the tapering approach.

From an implementation standpoint, tapering introduces a sparse structure for the dispersion matrix from the Gaussian process model. Referring to the univariate ($m = 1$) case, let $\boldsymbol{T}$ denote the $n \times n$ matrix with (i, j)-th element $C_\nu(\boldsymbol{s}_i, \boldsymbol{s}_j)$. The matrix $\boldsymbol{T}$ will have zero entries for any pair of locations separated by more than ν units and, hence, is sparse. There are choices aplenty for the tapering kernel, but the more widely used kernels use the Wendland family of tapered covariance functions (Wendland, 1995; Furrer et al. 2006). One particularly popular choice is given by $C_\nu(\boldsymbol{s}_1, \boldsymbol{s}_2) = \left(1 - \frac{h}{\nu}\right)_+^4 \left(1 + 4\frac{h}{\nu}\right)$, where $h = ||\boldsymbol{s}_1 - \boldsymbol{s}_2||$. Note that ν is typically not estimated but fixed to achieve the desired degree of sparsity in the dispersion matrix (Kaufman, 2009). In a Bayesian context, we can estimate ν using some prior distribution, but such priors will need to be strongly informative for ν to be identified. We avoid this needless complexity and work with a fixed ν in the subsequent development.

Tapered covariance structures have been used effectively for analyzing large spatial

datasets (e.g. Furrer et al. 2009). The tapered process realizations yield a dispersion matrix $\boldsymbol{C}_w \odot \boldsymbol{T}$, where $\odot$ is elementwise matrix product (also known as the Hadamard product). A standard property of the Hadamard product ensures that $\boldsymbol{C}_w \odot \boldsymbol{T}$ will be positive definite because $\boldsymbol{C}_w$ and $\boldsymbol{T}$ are. In any standard linear algebra text this property is proved using Schur Product Theorem (see Harville, pp. 458). We remark that the positive definiteness of $\boldsymbol{C}_w \odot \boldsymbol{T}$ follows directly from the fact that $\boldsymbol{C}_w \otimes \boldsymbol{T}$ is positive definite and, $\boldsymbol{E}'(\boldsymbol{C}_w \otimes \boldsymbol{T})\boldsymbol{E} = \boldsymbol{C}_w \odot \boldsymbol{T}$, where $\boldsymbol{E} = [\boldsymbol{e}_1 : \boldsymbol{e}_{n+2} : \boldsymbol{e}_{2n+3} : \cdots : \boldsymbol{e}_{n^2}]$, $\boldsymbol{e}_j$ is a $n^2 \times 1$ vector with 1 in the j th element and 0 otherwise. Since $\boldsymbol{T}$ is sparse, so is $\boldsymbol{C}_w \odot \boldsymbol{T}$; therefore, sparse matrix algorithms can be employed to estimate tapered spatial process models.

To accommodate richer underlying processes, we pursue the following strategy: we use the predictive process $\boldsymbol{w}_{pp}(\boldsymbol{s})$ as in (2.2), but taper the residual process $\boldsymbol{w}(\boldsymbol{s}) - \boldsymbol{w}_{pp}(\boldsymbol{s})$. Tapering $\boldsymbol{w}(\boldsymbol{s}) - \boldsymbol{w}_{pp}(\boldsymbol{s})$ will require a multivariate tapering kernel, where correlations need to be imposed among the componentwise tapering processes. To circumvent this problem, we pursue an alternative strategy to taper the residual process $\boldsymbol{v}(\boldsymbol{s}) - \boldsymbol{v}_{pp}(\boldsymbol{s})$. More precisely, assume $\boldsymbol{\eta}(\boldsymbol{s}) = (\eta_1(\boldsymbol{s}), ..., \eta_m(\boldsymbol{s}))'$ is a multivariate spatial random field independent of $\boldsymbol{w}(\boldsymbol{s})$ with $\eta_k(\boldsymbol{s}) \sim \mathrm{GP}(0, C_{\nu_k}(\cdot,\cdot))$ independently over $k = 1, ..., m$. Evidently, $\boldsymbol{v}(\boldsymbol{s}) - \boldsymbol{v}_{pp}(\boldsymbol{s})$ follows a Gaussian process with a covariance kernel $\boldsymbol{C}_{res,v}(\boldsymbol{s}_1, \boldsymbol{s}_2) = \boldsymbol{C}_v(\boldsymbol{s}_1, \boldsymbol{s}_2) - \mathcal{C}_v(\boldsymbol{s}_1, \mathcal{S}^*)'\boldsymbol{C}_v^{*-1}\mathcal{C}_v(\boldsymbol{s}_2, \mathcal{S}^*)$. The tapered predictive process is defined to be,

$$\boldsymbol{w}_{tap}(\boldsymbol{s}) = \boldsymbol{w}_{pp}(\boldsymbol{s}) + \boldsymbol{A}(\boldsymbol{s})[(\boldsymbol{v}(\boldsymbol{s}) - \boldsymbol{v}_{pp}(\boldsymbol{s})) \odot \boldsymbol{\eta}(\boldsymbol{s})] \tag{4.5}$$

which can easily be seen as a Gaussian process with covariance function $\boldsymbol{C}_{tap,w}(\boldsymbol{s}_1, \boldsymbol{s}_2) = \boldsymbol{A}(\boldsymbol{s}_1)\boldsymbol{C}_{res,v}(\boldsymbol{s}_1, \boldsymbol{s}_2)\boldsymbol{C}_\nu(\boldsymbol{s}_1, \boldsymbol{s}_2)\boldsymbol{A}(\boldsymbol{s}_2)' + \boldsymbol{C}_{w_{pp}}(\boldsymbol{s}_1, \boldsymbol{s}_2)$.

These specifications yield $N(\boldsymbol{y} \,|\, \boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\boldsymbol{y};tap})$ as the likelihood in (4.4), where

$$\begin{aligned} \boldsymbol{D}_{tap} &= \mathcal{A}\left[\left(\boldsymbol{C}_v - \mathcal{C}_v'\boldsymbol{C}_v^{*-1}\mathcal{C}_v\right) \odot \boldsymbol{T}\right]\mathcal{A}' \\ \Sigma_{\boldsymbol{y},tap} &= \mathcal{C}_w'\boldsymbol{C}_w^{*-1}\mathcal{C}_w + \boldsymbol{D}_{tap} + \boldsymbol{I}_n \otimes \boldsymbol{\Psi}\ , \end{aligned} \tag{4.6}$$

$\mathcal{A} = \mathrm{diag}(\boldsymbol{A}(\boldsymbol{s}_i))_{i=1}^n$, $\boldsymbol{T} = \{\mathrm{diag}(C_{\nu_k}(\boldsymbol{s}_1, \boldsymbol{s}_j))_{k=1}^m\}_{i,j=1}^n$. This dispersion matrix is $mn \times mn$, but the sparse structure can be utilized, again in conjunction with the Sherman-Woodbury-Morrison formula, to achieve substantial computational gains.

We point out that the modified predictive process did not account for the cross-covariance in the residual process $\boldsymbol{w}(\boldsymbol{s}) - \boldsymbol{w}_{pp}(\boldsymbol{s})$. Instead, it only adjusted for the variability in this residual using the independent process $\boldsymbol{\epsilon}_{w_{pp}}(\boldsymbol{s})$. The tapered approach, on the other hand, accounts for the residual spatial association as well as the variability. Putting $\nu_k = 0$ and $\nu_k = \infty \; \forall \; k$ in the tapered predictive process yields the modified predictive process and the parent Gaussian process respectively. In this respect, tapering offers a generalized modeling approach which includes parent Gaussian process and modified predictive process as special cases. In fact, adjusting ν_k's at different values, one can arrive at arbitrarily rich models. It is also evident from (4.5) that while the predictive process approximation of $\boldsymbol{w}(\boldsymbol{s})$ only offers a good approximation of the original covariance function at long distances due to the disappearance of the residual term, $\boldsymbol{w}_{tap}(\boldsymbol{s})$ aims at approximating the original covariance function both at long and short distances. In this sense, tapering enriches the approximation to the underlying parent process.

4.2.1 Dispersion matrix distances

Given the above three model-based strategies for modeling low-rank spatial covariance matrices, it may be instructive to ascertain each of their "distances" from the parent model's dispersion matrix. For this section we will assume $\boldsymbol{A}(\boldsymbol{s}) = \boldsymbol{A}$, so that $\mathcal{A} = \boldsymbol{I} \otimes \boldsymbol{A}$. Let $\boldsymbol{\Sigma}_{\boldsymbol{y};pp}$, $\boldsymbol{\Sigma}_{\boldsymbol{y};mpp}$ and $\boldsymbol{\Sigma}_{\boldsymbol{y};tap}$ be the marginal dispersion matrices corresponding to the predictive process, the modified predictive process and the tapered adjustment respectively. We use the following metric,

$$\Delta_* = \| \Sigma_{\boldsymbol{y};*} - (\boldsymbol{C}_w + \boldsymbol{I} \otimes \boldsymbol{\Psi}) \|_2 \; ,$$

where $|| * ||_2$ is the l_2 matrix norm, to compute Δ_{mpp}, Δ_{tap} and Δ_{pp}. We now have the following lemma.

Proposition 4.2.1 *Let Δ_{pp}, Δ_{mpp} and Δ_{tap} be the three metrics defined above. Then, for any fixed $\boldsymbol{\Theta}$ we have the following inequality:*

$$\Delta_{pp} \geq \Delta_{mpp} \geq \Delta_{tap} \; .$$

Proof See Appendix C.

4.3 Smoothness properties of the Low Rank Models

We will devote this section to study the smoothness properties of low rank models. The whole section describes univariate spatial processes because smoothness properties for multivariate processes follow from componentwise smoothness. It is often found in the spatial literature that the surface fitted through a low rank model is oversmoothed due to excessive borrowing of information from the neighboring locations. In this section, we would like to formally study this phenomenon. For simplicity, we will mainly focus on the class of predictive process and modifications thereof. For Gaussian processes a popular approach to study the "smoothness" or "roughness" of different processes is through the supremum metric defined by $||g|| = \sup_{\boldsymbol{s} \in D} g(\boldsymbol{s})$ (see Adler & Taylor, 2007). If $g_1(\cdot), g_2(\cdot)$ are two centered Gaussian processes, $\mathrm{E}\{||g_1||\} > \mathrm{E}\{||g_2||\}$ will imply g_2 being smoother than g_1. In fact, we have the following lemma.

Lemma 4.3.1 *Assume $w_{pp}(\boldsymbol{s}), w_{mpp}(\boldsymbol{s})$ are a.s. bounded on D. Then,*

$$E\{||w_{mpp}||\} > E\{||w_{pp}||\}$$

In other words, the predictive process is smoother than modified predictive process.

Proof See Appendix C.

Lemma 4.3.1 is quite appealing to study the smoothness properties of different Gaussian processes. It defines a paradigm under which smoothness of different processes can be studied. However, it is quite difficult to study the smoothness of the tapered adjusted process under this paradigm. We, therefore, investigate the paradigm of mean square continuity and differentiability of stochastic processes to throw further light on the smoothness of the predictive processes and their modified versions.

Mean square properties of a random field are instructive for many parametric spatial gradient problems. They also provide a way to formalize the concept of spatial smoothness by introducing the idea of derivatives of random fields. Derivatives of the random fields have been discussed in Adler (1981), Banerjee et al. (2003, 2006) and

Mardia, Kent, Goodall and Little (1996). Let $w(\boldsymbol{s})$ be a real valued spatial process; then the process $\{w(\boldsymbol{s}) : \boldsymbol{s} \in \mathcal{R}^d\}$ is L_2 continuous at $\boldsymbol{s}_0$ if $\lim_{\boldsymbol{s}\to\boldsymbol{s}_0} E\{w(\boldsymbol{s}) - w(\boldsymbol{s}_0)\}^2 = 0$. Moreover, if $w(\boldsymbol{s})$ is stationary with $\text{cov}\{w(\boldsymbol{s}), w(\boldsymbol{s}')\} = C_w(\boldsymbol{s} - \boldsymbol{s}')$ then the process $w(\boldsymbol{s})$ is mean square continuous at all sites $\boldsymbol{s}$ if C_w is continuous at $\mathbf{0}$. Analogous to the definition of mean square continuity, mean square differentiability of a process $w(\boldsymbol{s})$ demands the existence of a vector $\nabla w(\boldsymbol{s}_0)$, known as the gradient vector, such that, for any unit vector $\boldsymbol{u}$,

$$\lim_{h\to 0} \text{E}\left\{\frac{w(\boldsymbol{s}_0 + h\boldsymbol{u}) - w(\boldsymbol{s}_0)}{h} - \langle \nabla w(\boldsymbol{s}_0), \boldsymbol{u}\rangle\right\}^2 = 0 \tag{4.7}$$

Consider $\nabla w(\boldsymbol{s}_0) = (\nabla w_1(\boldsymbol{s}_0), \nabla w_2(\boldsymbol{s}_0))'$; the process $w(\cdot)$ is said to be twice mean square differentiable at $\boldsymbol{s}_0$ if each of $\nabla w_1(\boldsymbol{s}_0)$ and $\nabla w_2(\boldsymbol{s}_0)$ is mean square differentiable at $\boldsymbol{s}_0$. Let the mean square derivatives of $\nabla w_1(\boldsymbol{s}_0)$ and $\nabla w_2(\boldsymbol{s}_0)$ be $\nabla^2 w_1(\boldsymbol{s}_0) = (\nabla^2 w_{11}(\boldsymbol{s}_0), \nabla^2 w_{12}(\boldsymbol{s}_0))'$ and $\nabla^2 w_2(\boldsymbol{s}_0) = (\nabla^2 w_{21}(\boldsymbol{s}_0), \nabla^2 w_{22}(\boldsymbol{s}_0))$ respectively; we will stack them up and call the second mean square derivative of $w(\boldsymbol{s}_0)$ as $\nabla^2 w(\boldsymbol{s}_0) = ((\nabla^2 w_1(\boldsymbol{s}_0))', (\nabla^2 w_2(\boldsymbol{s}_0))')'$. Continuing in this way, we denote the m_1-th order mean square derivative by $\nabla^{m_1} w(\boldsymbol{s}_0)$, obtained by stacking all the 2^{m_1} elements of the set $\mathcal{P}_{m_1} = \left\{\nabla^{m_1} w_{i_1,\ldots,i_{m_1}}(\boldsymbol{s}_0) : (i_1, \ldots, i_{m_1}) \in \{1,2\}^{m_1}\right\}$.

With this definition, when $w(\boldsymbol{s})$ is stationary, mean square differentiability only requires the existence of the 2nd order derivative of $C_w(\cdot)$ at 0. On a similar note, existence of $C_w^{(2m_1)}(0)$ ensures m times mean square differentiability of a stationary process $w(\boldsymbol{s})$.

In the light of the aforesaid concepts it is always instructive to compare the two different bias adjusted model based strategies along with predictive process. The main result that accompanies our discussion here is stated below.

Proposition 4.3.2 *Let, $C_w(\boldsymbol{s}, \boldsymbol{t})$ be a Matern correlation function with $m_1 < \theta_2 < m_1 + 1$ and $C_\nu(\cdot)$ be $2k$ times differentiable at $\mathbf{0}$. Then*

1. *The predictive process model is infinitely mean square differentiable except at the set of knot points $\mathcal{S}^*$.*

2. *The modified predictive process is not even mean square continuous at any point.*

3. *The tapered predictive process is* $\min(m_1, k)$*-times mean square differentiable except at the set of knot points* $\mathcal{S}^*$.

Proof See Appendix C.

Proposition 4.3.2 indicates that the continuity and differentiability of the parent process $w(\cdot)$ can be retained in the tapered predictive process with an appropriate choice of the tapering kernel. Apart from this, the results can be potentially helpful for inference on spatial gradients (see Banerjee et al, 2006). The tapering approach facilitates gradient assessment of the spatial surface for large datasets quite accurately as it can ensure a similar degree of smoothness to the parent Gaussian process while being computationally more efficient. For spatial gradients, a popular choice for $C_w(\boldsymbol{s}, \boldsymbol{t})$ is the Mat'ern corelation function with $\theta_2 = \frac{3}{2}$ which yields once differentiable spatial realization. It is not very difficult to show that the Wendland tapering kernel $C_\nu(\boldsymbol{s}, \boldsymbol{t}) = (1 - \frac{\|\boldsymbol{s}-\boldsymbol{t}\|}{\nu})_+^4(1 + 4\frac{\|\boldsymbol{s}-\boldsymbol{t}\|}{\nu})$ is twice differentiable at $\mathbf{0}$ (see Appendix C). Referring to Proposition 4.3.2, it is now quite straightforward to see that the tapered predictive process yields once differentiable spatial realizations. In this sense, tapering, even with a low rank structure, offers rich modeling aspects.

4.4 Estimation and inference

4.4.1 Implementation

Estimation of the parameters in (4.4) is carried out using MCMC. Here there are two options. Either one keeps $\boldsymbol{\beta}$ unmarginalized in the model and proceeds with a combination of Gibbs and Metropolis steps, or, one can marginalize out $\boldsymbol{\beta}$ and carry out estimation of the parameters with random walk Metropolis steps. We prefer to follow the second option as it does not require storage of $\boldsymbol{\beta}$ estimates until burn-in. Assuming $\boldsymbol{\beta} \sim N(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)$ in the prior and marginalizing out $\boldsymbol{\beta}$, the full posterior distribution becomes,

$$p(\boldsymbol{\Theta} \,|\, \boldsymbol{y}) \propto p(\boldsymbol{\Theta}) \times N(\boldsymbol{y} \,|\, \boldsymbol{X}\boldsymbol{\mu}_\beta, \boldsymbol{X}\boldsymbol{\Sigma}_\beta\boldsymbol{X}' + \boldsymbol{\Sigma}_{\boldsymbol{y};*}) \,, \tag{4.8}$$

where * can be pp, mpp or tap. While carrying out MCMC steps, one needs to efficiently compute the inverse and determinant of the matrix $\boldsymbol{V}_{\boldsymbol{y},*} = \boldsymbol{X}\boldsymbol{\Sigma}_\beta\boldsymbol{X}' + \boldsymbol{\Sigma}_{\boldsymbol{y};*}$. Consider the following two systems of equations,

$$\begin{pmatrix} \boldsymbol{D}_{*,\Psi} & -\mathcal{C}_w' \\ \mathcal{C}_w & \boldsymbol{C}_w^* \end{pmatrix} \begin{pmatrix} \boldsymbol{W}_1 \\ \boldsymbol{W}_2 \end{pmatrix} = \begin{pmatrix} \boldsymbol{I} \\ \boldsymbol{0} \end{pmatrix} \tag{4.9}$$

$$\begin{pmatrix} \boldsymbol{D}_{*,\Psi} & -\boldsymbol{X} & -\mathcal{C}_w' \\ \boldsymbol{X}' & \boldsymbol{\Sigma}_\beta^{-1} & \boldsymbol{0} \\ \mathcal{C}_w & \boldsymbol{0} & \boldsymbol{C}_w^* \end{pmatrix} \begin{pmatrix} \boldsymbol{V}_1 \\ \boldsymbol{V}_2 \\ \boldsymbol{V}_3 \end{pmatrix} = \begin{pmatrix} \boldsymbol{I} \\ \boldsymbol{0} \\ \boldsymbol{0} \end{pmatrix} \tag{4.10}$$

Routine calculations reveal that solutions for $\boldsymbol{W}_1$ and $\boldsymbol{V}_1$ in (4.9), (4.10) yield $\boldsymbol{\Sigma}_{\boldsymbol{y},*}^{-1}$ and $\boldsymbol{V}_{\boldsymbol{y},*}^{-1}$ respectively. The inverse and determinant of $\boldsymbol{V}_{\boldsymbol{y},*}$ can be obtained efficiently by solving (4.9) and (4.10) through the following steps:

Step 1: Solve $\boldsymbol{W}_2$ from the $mn^* \times mn^*$ system

$$\left[\boldsymbol{C}_w^* + \mathcal{C}_w\boldsymbol{D}_{*,\Psi}^{-1}\mathcal{C}_w'\right]\boldsymbol{W}_2 = -\mathcal{C}_w\boldsymbol{D}_{*,\Psi}^{-1}.$$

Step 2: Compute $\boldsymbol{W}_1 = \boldsymbol{D}_{*,\Psi}^{-1}\left[\boldsymbol{I} + \mathcal{C}_w'\boldsymbol{W}_2\right]$.

Step 3: Solve $\boldsymbol{V}_2$ from the $p \times p$ system

$$\left[\boldsymbol{\Sigma}_\beta^{-1} + \boldsymbol{X}'\boldsymbol{W}_1\boldsymbol{X}\right]\boldsymbol{V}_2 = -\boldsymbol{X}'\boldsymbol{W}_1.$$

Step 4: Compute $\boldsymbol{V}_1 = \boldsymbol{W}_1\left[\boldsymbol{I} + \boldsymbol{X}\boldsymbol{V}_2\right]$.

Step 5: Evaluate $|\boldsymbol{V}_{\boldsymbol{y},*}|$ from

$$|\boldsymbol{V}_{\boldsymbol{y},*}| = |\boldsymbol{\Sigma}_\beta|\,|\boldsymbol{D}_{*,\Psi}|\,|\boldsymbol{C}_w^*|^{-1}\,|\boldsymbol{\Sigma}_\beta^{-1} + \boldsymbol{X}'\boldsymbol{W}_1\boldsymbol{X}|\,|\boldsymbol{C}_w^* + \mathcal{C}_w\boldsymbol{D}_{*,\Psi}^{-1}\mathcal{C}_w'|.$$

Step 1 and step 3 require evaluating the Cholesky decomposition of $\boldsymbol{C}_w^* + \mathcal{C}_w\boldsymbol{D}_{*,\Psi}^{-1}\mathcal{C}_w'$ and $\boldsymbol{\Sigma}_\beta^{-1} + \boldsymbol{X}'\boldsymbol{W}_1\boldsymbol{X}$ respectively which facilitates determinant computation. It must also be noted that $\boldsymbol{D}_{*,\Psi}$ is diagonal for mpp and sparse for tpp. This mitigates the computational burden in step 5.

Finally, the draws $\boldsymbol{\beta}^{(l)}$, $l = 1, ..., L$ after burn-in period are recovered by conjugate sampling from the normal distribution, $N(\boldsymbol{\beta}|\boldsymbol{\mu}_{\cdot|\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\cdot|\boldsymbol{\beta}})$, with

$$\boldsymbol{\Sigma}_{\cdot|\boldsymbol{\beta}} = [\boldsymbol{X}'\boldsymbol{W}_1\boldsymbol{X} + \boldsymbol{\Sigma}_{\beta}^{-1}]^{-1};\ \boldsymbol{\mu}_{\cdot|\boldsymbol{\beta}} = \boldsymbol{\Sigma}_{\cdot|\boldsymbol{\beta}}[\boldsymbol{X}'\boldsymbol{W}_1\boldsymbol{Y} + \boldsymbol{\Sigma}_{\beta}^{-1}\boldsymbol{\mu}_{\beta}]$$

using already obtained draws $\boldsymbol{\Theta}^{(l)}$. This computation is facilitated by the fact that $\boldsymbol{\Sigma}_{\cdot|\boldsymbol{\beta}}$ can be obtained directly from step 3. With some algebra one can also show that $\boldsymbol{\mu}_{\cdot|\boldsymbol{\beta}} = \boldsymbol{\Sigma}_{\beta}\boldsymbol{X}'\boldsymbol{V}_1\boldsymbol{Y} + \boldsymbol{\Sigma}_{\cdot|\boldsymbol{\beta}}\boldsymbol{\Sigma}_{\beta}^{-1}\boldsymbol{\mu}_{\beta}$. Matters are similar when $\boldsymbol{\beta}$ is assigned a flat prior, whereupon $\boldsymbol{\Sigma}_{\beta}^{-1} = \boldsymbol{0}$.

Since statistical inference from the class of predictive process models and its counterparts heavily depends on the choice of "knot" points, efficient knot selection becomes very important. In recent spatial literature, different ways of selecting knots have been proposed (see e.g. Tokdar, 2011; Finley et. al., 2009). Rather than knot selection through ad hoc criteria, Guhaniyogi et. al. (2011a) have proposed an approach of modeling the knots through point processes (see Chapter 2 in this thesis), gaining substantial benefits both in computational time and spatial inference. In this approach, however, the number of knots is fixed *apriori* and the number of knots should be as large as possible depending on the available computational resources. But in many cases, increasing the number of knots marginally improves statistical inference, while the increase in computational complexity becomes enormous. To address such scenarios, we discuss a sequential knot selection mechanism fpr predictive processes based on a sound theoretical formulation. First, we state a useful lemma,

Lemma 4.4.1 *Define* $\boldsymbol{w}_{pp,k}(\boldsymbol{s}) = E[\boldsymbol{w}(\boldsymbol{s}) \,|\, \boldsymbol{w}(\boldsymbol{s}_1^*), ..., \boldsymbol{w}(\boldsymbol{s}_k^*)]$, *and* $\boldsymbol{w}(\boldsymbol{s}) - \boldsymbol{w}_{pp,k}(\boldsymbol{s}) = \boldsymbol{w}_{res,k}(\boldsymbol{s})$ *for* $k = 1, ...$ *Then for all* $\boldsymbol{s}_1, \boldsymbol{s}_2 \in D$,

$$cov\{\boldsymbol{w}_{pp,k}(\boldsymbol{s}_1), \boldsymbol{w}_{pp,k}(\boldsymbol{s}_2)\} = cov\{\boldsymbol{w}_{pp,k-1}(\boldsymbol{s}_1), \boldsymbol{w}_{pp,k-1}(\boldsymbol{s}_2)\} + cov\{\boldsymbol{w}_{res,k-1}(\boldsymbol{s}_1), \boldsymbol{w}_{res,k-1}(\boldsymbol{s}_k^*)\}\,[var\{\boldsymbol{w}_{res,k-1}(\boldsymbol{s}_k^*)\}]^{-1}\,cov\{\boldsymbol{w}_{res,k-1}(\boldsymbol{s}_2), \boldsymbol{w}_{res,k-1}(\boldsymbol{s}_k^*)\}' \tag{4.11}$$

Proof See Appendix C

Apart from its other implications, the strength of this lemma can be harnessed to develop a sequential knot selection criterion for predictive process models. Choosing

$\boldsymbol{s}_1 = \boldsymbol{s}_2 = \boldsymbol{s}$ in (4.11), $\text{Diff}_{\boldsymbol{s},k} = \text{var}\{\boldsymbol{w}_{pp,k}(\boldsymbol{s})\} - \text{var}\{\boldsymbol{w}_{pp,k-1}(\boldsymbol{s})\}$ comes out to be

$$\text{cov}\{\boldsymbol{w}_{res,k-1}(\boldsymbol{s}), \boldsymbol{w}_{res,k-1}(\boldsymbol{s}_k^*)\}\,[\text{var}\{\boldsymbol{w}_{res,k-1}(\boldsymbol{s}_k^*)\}]^{-1}\text{cov}\{\boldsymbol{w}_{res,k-1}(\boldsymbol{s}), \boldsymbol{w}_{res,k-1}(\boldsymbol{s}_k^*)\}'$$

If the $(k-1)$ knots are sufficient to capture the features of the underlying surface, the right hand side of the above equation should be close to zero. So, given the data points $\boldsymbol{s}_1, ..., \boldsymbol{s}_n$, one can consider following algorithm to determine the number of knots,

---

- Specify locations of size N, which can either be on a grid or chosen in some other way or a combination of both. Also specify a tolerance level ϵ_{tol} and an initial set of n^* knots obtained from collecting knots at 1000-th MCMC iteration of the adaptive predictive process (Guhaniyogi et. al., 2011a).

for t `in` $1 : N$ **do**

- At the $(t+1)$-th iteration choose a location $\boldsymbol{s}^*_{n^*+t+1}$ randomly, outside (n^*+t) knot points already chosen, from the prespecified N locations.
- Compute, $\text{Diff}_{t+1} = \frac{1}{n}\sum_{i=1}^{n} \|\text{Diff}_{\boldsymbol{s}_i,t+1}\|$

if $\text{Diff}_{t+1} < \epsilon_{tol}$ **then**

break, with (n^*+t) knots as the final set of knots, Otherwise proceed as in step 2.

else

$\boldsymbol{s}^*_{n^*+t+1}$ is also included in the set of knots

end if

end for

Once the parameters have been estimated, inferential interest turns to spatial prediction. Here, a few situations are of interest. Let $\boldsymbol{s}_0$ be any location in the domain, where we seek to predict $\boldsymbol{y}(\boldsymbol{s}_0)$ based on a given matrix of predictors $\boldsymbol{X}(\boldsymbol{s}_0)'$. For the marginalized model, spatial prediction proceeds from the posterior predictive distribution

$$p(\boldsymbol{y}(\boldsymbol{s}_0)\,|\,\boldsymbol{y}) = \int p(\boldsymbol{y}(\boldsymbol{s}_0)\,|\,\boldsymbol{y}, \boldsymbol{\Theta}) p(\boldsymbol{\Theta}\,|\,\boldsymbol{y})\, d\boldsymbol{\Theta}. \tag{4.12}$$

Posterior predictive sampling is achieved using *composition* (e.g. Banerjee et al. 2004). For each $\{\boldsymbol{\Theta}^{(l)}\}$, $l = 1, 2, \ldots, L$, obtained from the posterior distribution $p(\boldsymbol{\Theta} \,|\, \boldsymbol{y})$, we draw $\boldsymbol{y}(\boldsymbol{s}_0)^{(l)}$ from $p(\boldsymbol{y}(\boldsymbol{s}_0) \,|\, \boldsymbol{y}, \boldsymbol{\Theta}^{(l)})$. The resulting $\boldsymbol{y}(\boldsymbol{s}_0)^{(l)}$, $l = 1, 2, \ldots, L$ are samples from (4.12). This is especially simple for Gaussian likelihoods because $p(\boldsymbol{y}(\boldsymbol{s}_0) \,|\, \boldsymbol{y}, \boldsymbol{\Theta})$ then turns out to be a normal distribution.

Bayesian inference is especially attractive for spatial data analysis because it facilitates full inference on the latent spatial processes $\boldsymbol{w}_*(\boldsymbol{s})$ or even $\boldsymbol{w}(\boldsymbol{s})$. For example, the posterior distribution for $\boldsymbol{w}_*(\boldsymbol{s}_0)$ for an arbitrary location $\boldsymbol{s}_0$ is given by,

$$p(\boldsymbol{w}_*(\boldsymbol{s}_0) \,|\, \boldsymbol{y}) = \int p(\boldsymbol{w}_*(\boldsymbol{s}_0) \,|\, \boldsymbol{w}^*, \boldsymbol{\Theta}) p(\boldsymbol{w}^* \,|\, \boldsymbol{y}, \boldsymbol{\Theta}) p(\boldsymbol{\Theta} \,|\, \boldsymbol{y}) \, d\boldsymbol{\Theta} \,. \tag{4.13}$$

Sampling from (4.13) is straightforward: for each posterior sample $\{\boldsymbol{\Theta}^{(l)}\}$, $l = 1, 2, \ldots, L$, we draw $\boldsymbol{w}^{*(l)}$ from $p(\boldsymbol{w}^* \,|\, \boldsymbol{y}, \boldsymbol{\Theta}^{(l)})$ and then $\boldsymbol{w}_*(\boldsymbol{s}_0)^{(l)}$ from $p(\boldsymbol{w}_*(\boldsymbol{s}_0) \,|\, \boldsymbol{w}^{*(l)}, \boldsymbol{\Theta}^{(l)})$, which are both normal distributions. The procedure for predictive processes is analogous.

4.4.2 Model selection

To compare how well the different bias adjustments perform, we adopt the posterior predictive loss approach of Gelfand and Ghosh (1998) (discussed already in Chapter 3).

As a second alternative, we will also consider the popular Deviance Information Criterion, or DIC, to rank models in terms of how well they fit the data. This criterion is the sum of the Bayesian deviance (a measure of model fit) and the (effective) number of parameters (a penalty for model complexity). The deviance is defined as $D(\boldsymbol{y};\, \boldsymbol{\beta}, \boldsymbol{\Theta}) = -2 \log p(\boldsymbol{y} \,|\, \boldsymbol{\beta}, \boldsymbol{\Theta})$ and is regarded as a measure of discrepancy between the data and the model. Spiegelhalter et al. (2002) consider two different estimates of the deviance: $D(\boldsymbol{y}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Theta}})$, where $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\Theta}}$ are posterior means for $\boldsymbol{\beta}$ and $\boldsymbol{\Theta}$ respectively, and the posterior average of the deviance itself given by $\bar{D}(\boldsymbol{y}) = (1/L) \sum_{l=1}^{L} D(\boldsymbol{y}, \boldsymbol{\beta}^{(l)}, \boldsymbol{\Theta}^{(l)})$ using posterior samples $\boldsymbol{\beta}^{(l)}$ and $\boldsymbol{\Theta}^{(l)}$. The difference $\bar{D}(\boldsymbol{y}) - D(\boldsymbol{y}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Theta}})$, it can be argued, is a measure of model complexity (denoted by p_D) and thus DIC $= 2\bar{D}(\boldsymbol{y}) - D(\boldsymbol{y}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\Theta}})$ serves as a criterion for model fit.

4.5 Illustrations

In this section, we present two simulation examples (one in univariate setting and other in bivariate setting) and one real data example. In the simulation example we have analyzed datasets simulated from the classical geostatistical model with exponential spatial correlation function. Finally, we illustrate our models with a forestry dataset on Tree Volume collected from the Zurichberg forest inventory. All the models have been implemented on `R` package. For sparse matrix operations we have used `SparseM` package in `R`. Details on `SparseM` can be found in `cran.r-project.org/web/packages/SparseM/SparseM.pdf`. While implementing Tapered Predictive Process, we have employed the Wendland taper function described in Subsection 4.2.

4.5.1 Analysis of a univariate synthetic data

We first present an analysis of a synthetic dataset to explore the properties of the proposed low-rank models. The data set comprises $n = 1100$ locations, generated randomly within a 1×1 unit square domain. We argue that restriction of the domain does not hamper the full generality of the problem, as any domain can be transformed to the 1×1 domain with an application of a suitable location-scale transformation. Since we are in the univariate set up, $\boldsymbol{A}(\boldsymbol{s}) = \sigma^2$. We generate outcome variable at sampled location using a geostatistical likelihood $N(\boldsymbol{y} \mid \beta_0\mathbf{1}, \boldsymbol{C}_w + \tau^2\boldsymbol{I})$, where $C_w(\boldsymbol{s}_1, \boldsymbol{s}_2) = \sigma^2 C_v(\boldsymbol{s}_1, \boldsymbol{s}_2)$, $C_v(\boldsymbol{s}_1, \boldsymbol{s}_2) = \exp(-\phi\|\boldsymbol{s}_1 - \boldsymbol{s}_2\|)$. Note that, $C_v(\boldsymbol{s}_1, \boldsymbol{s}_2)$ corresponds to a Matérn correlation function in (3.5) with $\theta_1 = \phi$, $\theta_2 = .5$. The column labeled *True* in Table 4.1 depicts the mean and variance parameter values used to generate the data.

Given the simulated dataset, we considered the four model-fitting methods: *non-spatial model, predictive process model, modified predictive process model* and *tapered predictive process model*. For fitting the models, we assigned a flat prior to the common intercept β_0, while all the variance parameters were assigned $IG(2, 1)$ (mean = 1) prior. Using the exponential spatial correlation, the prior for the decay parameter ϕ was a Uniform $U(2.2, 7.34)$. This yields a fairly acceptable range of support for the range, given that the maximum inter-location distance in the generated data was 1.36. We remark that a uniform prior on ϕ does not translate into a uniform prior for the range, which, for the exponential correlation function, is customarily defined as $3/\phi$.

Nevertheless, the prior range is broad enough to allow the data to drive the inference.

In estimating low rank models, we usually experiment with a varying number of knots. Typically, beyond a certain number of knots the substantive inference becomes robust. In this particular example, we found that placing just 30 knots randomly over the unit square domain was able to capture the salient features of the underlying process. For brevity, we only present the analysis with 30 knots. In this context, we would like to assert that, of late, there has been growing interest in "sensible" knot selection methods in low rank knot based models and its predictive process counterparts (see Guhaniyogi et al., 2011a; Finley et al., 2009; Tokdar, 2011). "Sensible" selection of knots in the three models of interest (predictive process, modified predictive process and tapered predictive process) will, of course, help in accruing substantial computational benefits as well as better posterior inference. But given the context of this paper, knot selection is less a matter of concern here, as any "sensible" selection of knots will improve performance for all the three models. Therefore, we opted for the simplest way of selecting knots randomly over the entire domain. While fitting tapered predictive process, we kept $\nu = .06$ which yielded approximately 1% nonzero off diagonal entries in the matrix $(\boldsymbol{C}_w - \mathcal{C}_w' \boldsymbol{C}_w^{*-1} \mathcal{C}_w) \odot \boldsymbol{T}$. The positions of the nonzero entries have been plotted in Figure 4.1.

Table 4.1 presents the 95% Bayesian credible intervals. We find that the global mean, β_0, is estimated robustly across all the models. Note that the nugget, τ^2, is significantly overestimated by the predictive process model (PP), while the two bias-adjusted models capture the true value of the nugget. The spatial variance parameter (σ^2) seems to be modestly underestimated, but we hasten to emphasize that while the data was generated from a stationary process, all the spatial models shown in Table 4.1 are nonstationary. Spatial variability, therefore, is not captured by σ^2 alone but by $\text{var}\{w_{mpp}(\boldsymbol{s})\}$ and $\text{var}\{w_{tap}(\boldsymbol{s})\}$ for modified and tapered processes respectively, which now varies by location. Finally, note that since the data is generated from a spatial process model, it is not surprising that the linear model is not able to capture the spatial variability, resulting in a grossly inflated estimate for the residual variance.

Turning to model comparisons, we find that both the posterior predictive loss metric and the DIC indicate improvements in performance for the two bias-adjusted models when compared to the predictive process model. Unsurprisingly, the spatial models

Table 4.1: The median and 95% Bayesian credible intervals for a non-spatial (i.e. ordinary linear regression) model, and four spatial models – the predictive process model (PP) and the three model-based bias adjustments – are presented for the synthetic data set. Also presented are model comparison metrics.

	True	Non-spatial	PP	Modified PP	Tapered PP
β_0	8.26	8.26 (8.15 , 8.27)	10.83 (9.29 , 12.60)	9.21(7.83 , 10.97)	8.43 (7.20 , 9.64)
σ^2	6	–	8.95 (2.68 , 15.81)	5.07 (3.44 , 7.32)	4.06 (3.12 , 5.91)
τ^2	0.5	3.59 (3.30 , 3.88)	2.20 (2.02 , 2.40)	.73 (.39 , 1.17)	0.43 (0.34 , 0.55)
ϕ	4	–	2.78 (2.32 , 3.62)	2.73 (2.23 , 5.38)	4.09 (2.61 , 5.77)
G	–	3959.95	2397.21	347.16	146.72
P	–	3943.83	2502.70	1471.05	858.04
D	–	7903.79	4899.91	1818.22	1004.76
p_D	–	1.95	31.79	731.42	1010.30
DIC	–	2509.32	2000.50	1628.88	1370.06

perform much better than the non-spatial model. Given that the tapered predictive process retains the smoothness of the geostatistical model and the tapered covariance structure is "closest" (in the sense of Lemma 4.2.1) to the full geostatistical model, it is also not very surprising that the tapered predictive process produces an overall better fit than the simple modified predictive process. A pictorial depiction of the residual spatial surfaces from all the three spatial models can be found in Figure 4.2. From the residual surfaces, it is evident that, of the two bias-adjusted processes, the tapered process is smoother than the other one. This is also theoretically justified by the fact that the tapered process will yield mean-square continuous process realizations almost everywhere, while the modified predictive process is not. Needless to say, the residual surface constructed from the predictive process model is oversmoothed, although somewhat capturing the essence of the true spatial surface.

While putting forward the two bias-adjusted models, our aim remains to capture the spatial correlation between any two observations compared to the true spatial correlation (exponential spatial correlation) between them. Figure 4.3 represents the spatial correlation constructed from the posterior estimates of the modified, tapered and predictive process models. Note that, predictive process model performs poorly in capturing spatial correlation at small distances, although it is pretty close to the true exponential correlation when the distance is large. The modified predictive process shows the same

behavior except at zero, where the correlation is always 1, as it should be. On the other hand, the tapered predictive process shows better behavior both at long and short distances. In fact, after distance= 0.6, it almost coincides with the true exponential correlation. These plots provide us with concluding evidence of better performance by using the tapered predictive process model. In the next subsection, we will investigate the models in multivariate synthetic dataset .

4.5.2 Analysis of multivariate synthetic data

This subsection presents a multivariate synthetic dataset to explore the properties of the proposed low-rank models. The data set comprises $n = 1000$ locations, generated randomly within a 1×1 unit square domain. Two outcomes were generated at each sampled location using a geostatistical likelihood $N(\boldsymbol{y} \mid \boldsymbol{\beta}\mathbf{1}_n, \boldsymbol{C}_w + \boldsymbol{I} \otimes \boldsymbol{\Psi})$ where $\boldsymbol{\Psi}$ is taken to be a diagonal matrix with diagonal entries ψ_1 and ψ_2. The mean of each location was assumed to have a common intercept denoted by $\boldsymbol{\beta} = (\beta_0, \beta_1)'$. An exponential spatial correlation function was assumed for all spatial processes, i.e., θ_2 was fixed at 0.5 in (3.5). For simplicity, we assume $\boldsymbol{A}(\boldsymbol{s}) = \boldsymbol{A} = \begin{pmatrix} A_{11} & 0 \\ A_{21} & A_{22} \end{pmatrix}$. The column labeled *True* in Table 4.2 depicts the mean and variance parameter values used to generate the data.

Given the simulated dataset, we considered three model-fitting methods: *predictive process model, modified predictive process model* and *tapered predictive process model*. For estimating the models, we assigned a flat prior to the common intercept $\boldsymbol{\beta}$ and N(0,1) prior for A_{21}, while all the variance parameters (ψ_1, ψ_2, A_{11}, A_{22}) were assigned $IG(2, 1)$ (mean = 1) prior. Using the exponential spatial correlation, priors for decay parameters ϕ_1 and ϕ_2 were Uniform $U(1, 8)$. This yields fairly acceptable ranges of support for range parameters, given that the maximum inter-location distance in the generated data was 1.36. To carry out the whole analysis, we have chosen 50 knots uniformly over the spatial domain. Figure 4.4 represents the simulated data points with knot points overlaid.

Table 4.2 presents the 95% Bayesian credible intervals. We find that the global mean, $\boldsymbol{\beta}$, is estimated robustly across all the models. Note that the nugget parameters, ψ_1 and ψ_2, are significantly overestimated by the predictive process model (PP), while

the two bias-adjusted models capture the true values of nuggets. The spatial variance parameters (A_{11}, A_{22}) seem to be consistently estimated both in modified and tapered adjusted models. The same remains true for the covariance parameter A_{21}. Predictive and modified predictive processes show a significant underestimation in range parameters ϕ_1 and ϕ_2, while 95% CI's for ϕ_1 and ϕ_2 in tapered model properly capture the true values of ϕ_1 and ϕ_2.

Table 4.2: The median and 95% Bayesian credible intervals for three spatial models – the predictive process model (PP) and the two model-based bias adjustments – are presented for the synthetic data set. Also presented are model comparison metrics.

	True	PP	Modified PP	Tapered PP
β_0	5	6.94 (3.21 , 11.37)	6.34 (4.05 , 9.44)	5.92 (4.17 , 7.97)
β_1	1	-0.82 (-4.83, 1.47)	0.48 (-0.60,1.68)	0.48 (-0.38, 1.48)
ψ_1	.5	2.42 (2.22, 2.67)	0.66 (0.38, 0.94)	0.44 (0.32, 0.57)
ψ_2	.4	1.55 (1.44, 1.69)	0.44 (0.24, 0.67)	0.31 (0.23, 0.42)
A_{11}	3	6.63 (4.78, 8.98)	3.04 (2.50, 4.32)	2.68 (2.28, 3.37)
A_{12}	.9	1.67 (0.10, 3.15)	1.03 (0.78, 1.41)	0.81 (0.63, 1.05)
A_{22}	2	3.46 (2.14, 5.25)	1.62 (1.40, 2.01)	1.60 (1.42, 1.92)
ϕ_1	4	1.43 (1.02 , 2.32)	2.35 (1.12 , 3.51)	3.27 (1.95 , 4.84)
ϕ_2	6	1.96 (1.05 , 4.77)	4.66 (2.55 , 6.60)	5.47 (3.29 , 7.28)
G	–	3906.87	361.72	214.72
P	–	4221.81	2077.43	1501.29
D	–	8128.68	2439.15	1716.02

As far as the predictive performance is concerned, we find that the posterior predictive loss metric indicates improvements in performance by the two bias-adjusted models when compared to the predictive process model. In particular, the value of G, indicating predictive fit, decreases significantly in the modified and tapered adjusted models. The tapered model being a more close approximation to the high dimensional model, it is also not very surprising that the tapered predictive process produces an overall better fit than the simple modified predictive process. A pictorial depiction of the residual spatial surfaces for outcomes 1 and 2 from all three spatial models can be found in Figures 4.5 and 4.6 respectively. As in section 4.5.1, it is evident that, of the two bias-adjusted processes, the tapered process is smoother than the other one. This is also theoretically justified by the fact that the tapered process will yield mean-square continuous process

realizations almost everywhere, while the modified predictive process is not even mean-square continuous. Needless to say, the residual surface constructed from the predictive process model is oversmoothed, although somewhat capturing the essence of the true spatial surface.

Like in the univariate synthetic analysis, one would like to see the estimated spatial correlations in the multivariate set up from the three models. Figures 4.7 and 4.8 represent spatial correlations of 100 selected points from the full data set, constructed from the posterior estimates of modified, tapered and predictive process models for the two latent processes respectively. True exponential correlation curves are overlaid. Note that the predictive process model proves to be the worst in capturing spatial correlation at small distances. On the other hand, both the modified predictive process and tapered adjusted models perform much better at capturing the true spatial correlation. The tapered adjusted model is particularly notable here as the spatial correlation constructed from this model almost coincides with the true exponential correlation (see Figure 4.8) for distances greater than 0.6 from the full model. These plots show better fit of tapered model and illustrate that even in multivariate data, the tapered model excels over its modified counterpart. Next, we will look at the performances of the different models in a real data from forestry.

4.5.3 Forestry example

Tree diameter at breast height (DBH; 1.37m above the forest floor) and volume (VOL) of the tree are used, in conjunction with other information, to assess a tree's economic and ecological value. Measuring VOL is more cumbersome and expensive as compared to measuring DBH. Therefore VOL is often measured on a small subset of those trees that also provide DBH measurements. Then a statistical model that relates VOL to DBH is used to predict VOL for the complement of this subset. Here, we illustrate how the proposed methods can help improve the predictive performance of such models.

The dataset that we analyze (hereafter referred to as Zurichberg data) has been collected from an inventory conducted in the early 1990s in parts of the Zurichberg Forest belonging to the city and Canton of Zurich (see Mandallaz, 2008). The inventoried area covered 217.9 ha, of which 17.1 ha served for the full census, in which the tree coordinates have been recorded. The inventory used systematic cluster sampling schemes where

cluster comprises five points: *central point, two points established 30m east and west of the central point; two other points each established 40m north nd south of the the cental point.* At these five points, a number of measurements, including DBH and VOL, have been collected for 4954 trees. Given this data set, a forestry scientist is interested in predicting the volume (VOL) of a tree based on the Diameter at breast height (DBH) measurement of the tree at some location.

The relationship between DBH and VOL is influenced by many individual and environmental factors. Individual factors include age, species, and genetics, whereas environmental factors include quality of soil, quantity of water and light, and competition for these resources. These factors often vary spatially across the domain. For example, a cohort of trees of the same species, age, and parentage will likely depict similar DBH and VOL. An example of an environmental factor influencing tree growth characteristics is soil productivity which can result in parent material or disturbance history. Given these unobserved covariates, we can expect some level of spatial dependence in VOL even after accounting for DBH. This analysis was based on representative 1320 mature trees from the Zurichberg dataset. Our candidate models include exactly the four models explored in the synthetic example.

First, the locations are all mapped to a $[0,1] \times [0,1]$ domain. As in the simulation experiment, we present, again for brevity, the results with only 36 knots, distributed uniformly over the domain. The priors for τ^2 and σ^2 follow $IG(2,1)$. We assume an exponential spatial correlation function, and assign a $U(0.5,6)$ prior to ϕ, which corresponds to a broad range of support given the maximum distance between any two trees is 1.22 (in the transformed scale). For all models the regression coefficient for DBH is assigned a *flat* prior.

The inference is based on three parallel chains of 5000 iterations each, discarding the first 2000 iterations as pre-convergence burn-in. The resulting posterior credible intervals with the G,P, and D scores are presented in Table 4.3. DBH is seen, quite expectedly, to have a significant impact on VOL. As seen in the synthetic example, the estimates for τ^2 from the bias-adjusted models are more reliable than those from the predictive process model. The G, P, D scores again reveal the considerable improvements in overall model fit achieved by the bias-adjusted models over the predictive process model and the non-spatial model, the improvement being most prominent in tapered

predictive process model. To understand the model fit better, we have plotted 95% predictive interval calculated from $y_{rep}(\boldsymbol{s}_i)$ for each $\boldsymbol{s}_i$ (with $\mu_{rep,i}$ shown by $\circ$) against the observed $y(\boldsymbol{s}_i)$ at each point. Figure 4.9 shows these plots for predictive, modified predictive and tapered predictive process models. It clearly shows that for both the modified and tapered predictive process models, $\mu_{rep,i}$'s lie close to the 45 degree line, implying closer alignment between the observed VOL and mean of the replicated VOL from the models. This alignment, however, is not very close at larger values of VOL. This can be attributed to the insufficient number of large VOL observations, so that spatial surfaces cannot be estimated accurately at these points. Although, the alignment is fairly indistinguishable in the two bias adjusted models, a closer introspection of the Figure 4.9 will show the tapered model excels over modified process. This phenomenon will be more prominent when the data offer more and more spatial information compared to the noise.

Table 4.3: The median and 95% Bayesian credible intervals for a non-spatial (i.e. ordinary linear regression) model, and three spatial models – the predictive process model (PP) and the two model-based bias adjustments – are presented for the forestry example. Also presented are model comparison metrics.

	Non-Spatial	PP	Modified PP	Tapered PP
β_0	-1.51 (-1.56 , -1.46)	-1.17 (-2.03 , -0.26)	-1.35 (-1.63 , -1.02)	-1.47 (-1.85 , -1.11)
β_{DBH}	0.09 (0.09 , 0.09)	0.09 (0.09 , 0.09)	0.09 (0.09 , 0.09)	0.09 (0.09 , 0.09)
σ^2	–	0.38 (0.17 , 0.81)	0.11 (0.07 , 0.18)	0.18 (0.11 , 0.33)
τ^2	0.18 (0.17 , 0.20)	0.16 (0.15 ,0.17)	0.12 (0.10 , 0.14)	0.09 (0.08 , 0.11)
ϕ	–	1.05 (0.54 , 2.51)	4.05 (2.54 , 5.81)	4.75 (2.38 , 5.95)
G	248.53	209.87	116.15	82.72
P	250.19	219.17	205.44	182.28
D	498.72	419.04	321.59	265
MSPE	0.32	0.32	0.29	0.29
p_D	3.04	18.94	126.69	140.85
DIC	5562.06	5451.89	5370.78	5390.58

Figure 4.10 displays the posterior means of the residual surface (from (4.13)) from the different models. Figure is an interpolated surface for the non-spatial model residuals. We would expect the fitted spatial random effects of the candidate models to look somewhat similar to this residual surface. It must be noted, in this respect, that the range of the legend strips are different for the non-spatial surface compared to the

other three surfaces. The reason behind this is, in non-spatial model the residuals are higher than the other three models. So, using the same legend strip for the four models will disrupt the precision of the surfaces. Therefore, we have used the same legend strips for the three spatial models. Figure reveals excessive oversmoothing by the predictive process, while the bias-adjustments compensate for such excesses. Of the two bias-adjusted models, the tapered process is smoother than the other two – this is theoretically justified as well by the fact that the tapered process will yield mean-square continuous process realization almost everywhere, while the modified predictive process is not mean-square continuous.

4.6 Conclusion and Further work

This Chapter formally explores the nature of biases in residual variability captured by low-rank spatial (geostatistical) models and their impact on the smoothness of the spatial surface. We have explained how such biases arise and show, specifically, how low-rank predictive processes can help quantify such biases. We have then proceeded to formulate a "bias-adjusted" model that attempts to ameliorate the impact of this bias and lead to improved model performance, without compromising on the degree of smoothness of the fitted spatial surface. In order to have a clear understanding of the bias, we have focussed our attention mainly on the predictive process models, which offer an easy way to quantify and subsequently rectify the bias in our proposed models. An elegant knot selection algorithm has also been proposed for predictive process models and its counterparts.

Future work will pursue an understanding of biases in *adaptive predictive process* models (Guhaniyogi et.al., 2011a), where the location of the knots are modeled stochastically through a spatial point process model. Another interesting direction entails the understanding of biases in residual variability from approximate likelihoods such as those proposed by Vecchia (1988) and Stein et al. (2004). Finally, developing anisotropic tapering, where ν may vary by direction, can lead to interesting classes of spatial models that will be useful to applied spatial modelers.

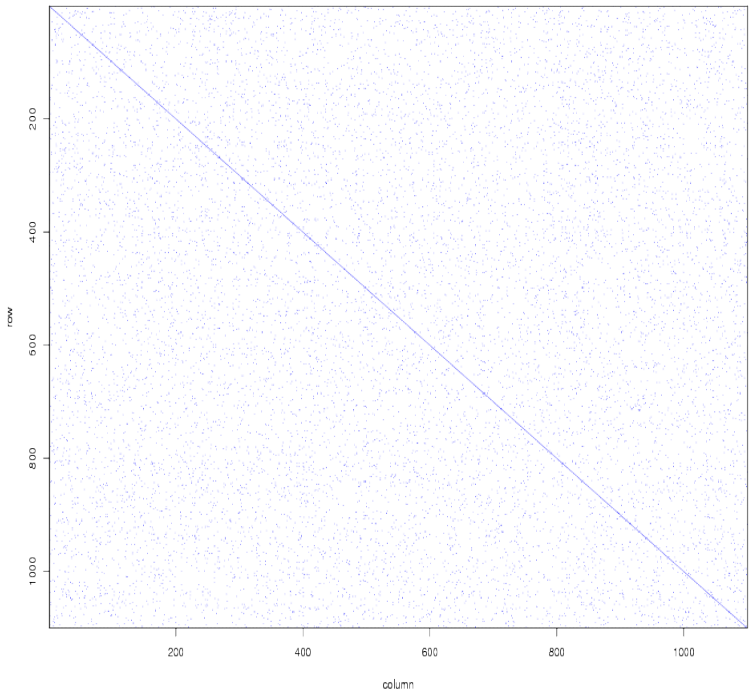
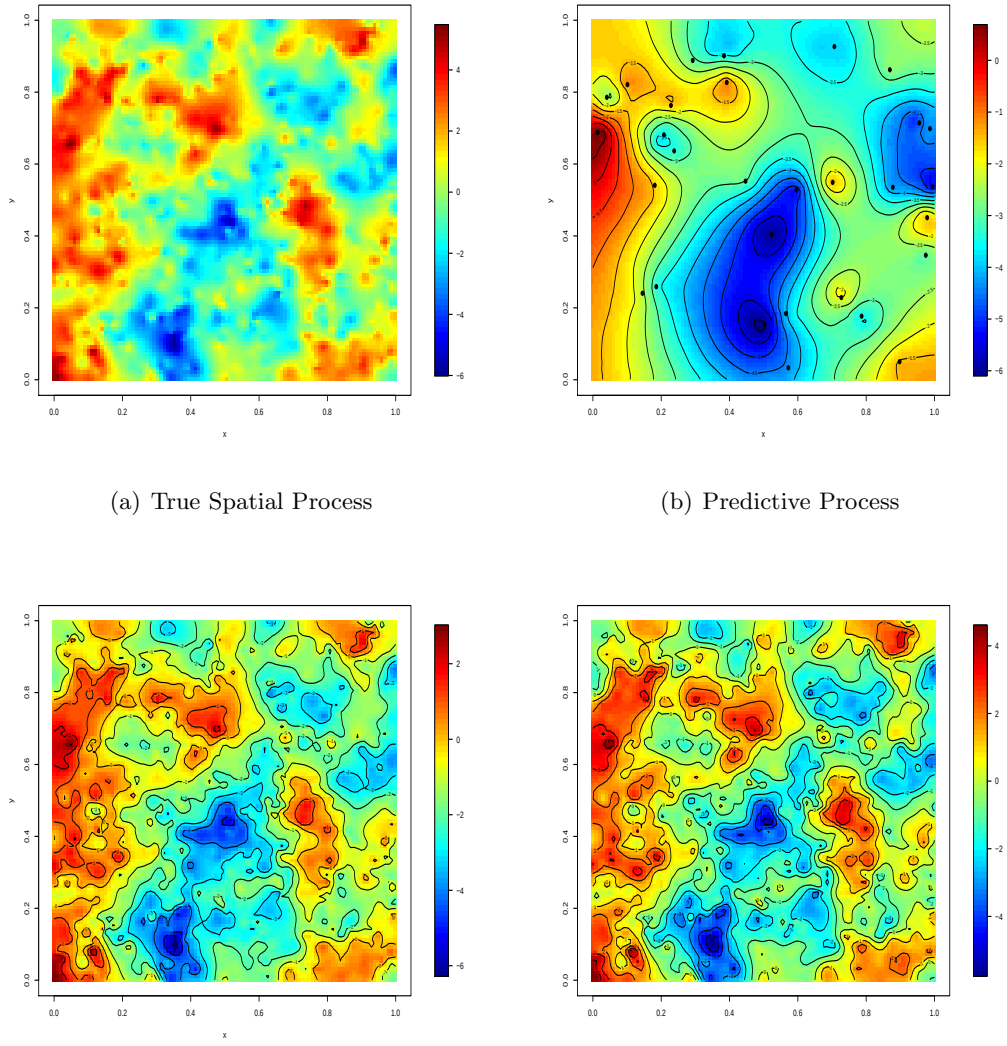


Figure 4.1: $(\boldsymbol{C}_w - \mathcal{C}_w' \boldsymbol{C}_w^{*-1} \mathcal{C}_w) \odot \boldsymbol{T}$ matrix with dots showing nonzero off-diagonal entries



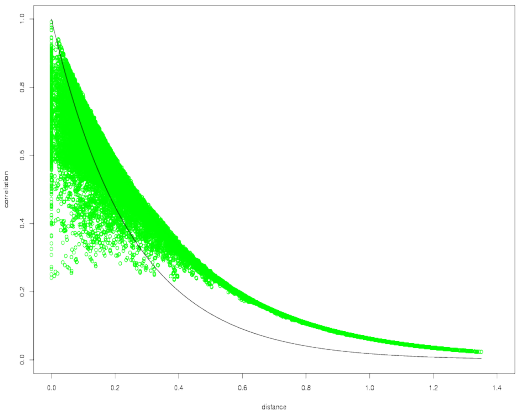
(a) True Spatial Process

(b) Predictive Process

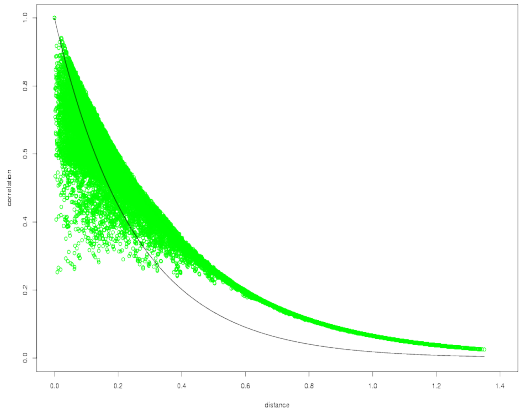
(c) Modified Predictive Process

(d) Tapered Predictive Process

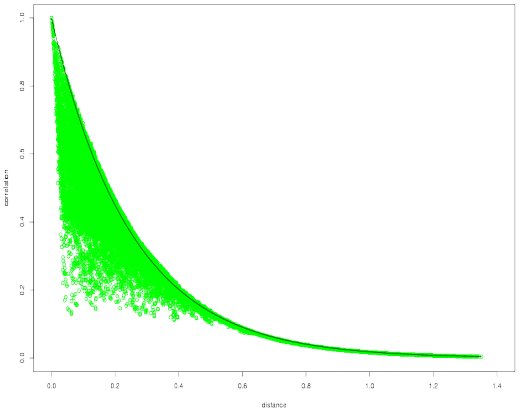
Figure 4.2: True and the estimated (posterior mean) spatial surface from the three candidate models: (a) True Spatial surface; (b) Predictive process model; (c) Modified Predictive process model; and (d) Tapered Predictive process model.



(a) Predictive Process



(b) Modified Predictive Process



(c) Tapered Predictive Process

Figure 4.3: Estimated spatial correlation: (a) Predictive process model; (b) Modified Predictive process model; and (c) Tapered Predictive process model.

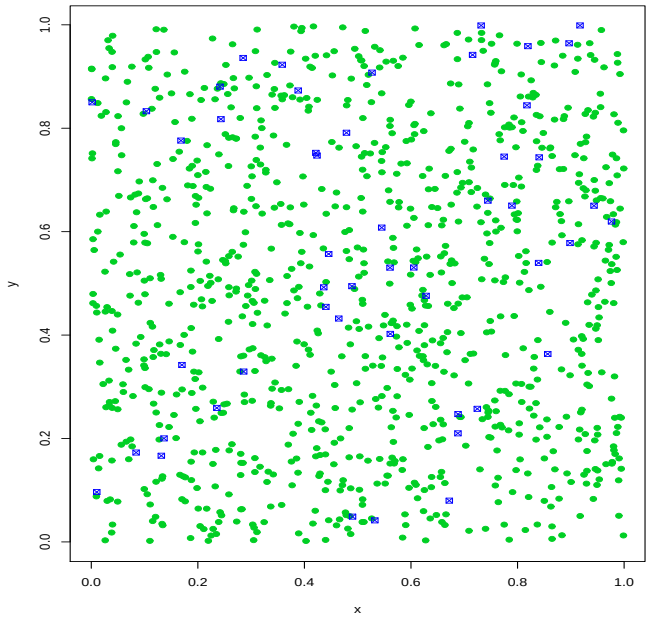
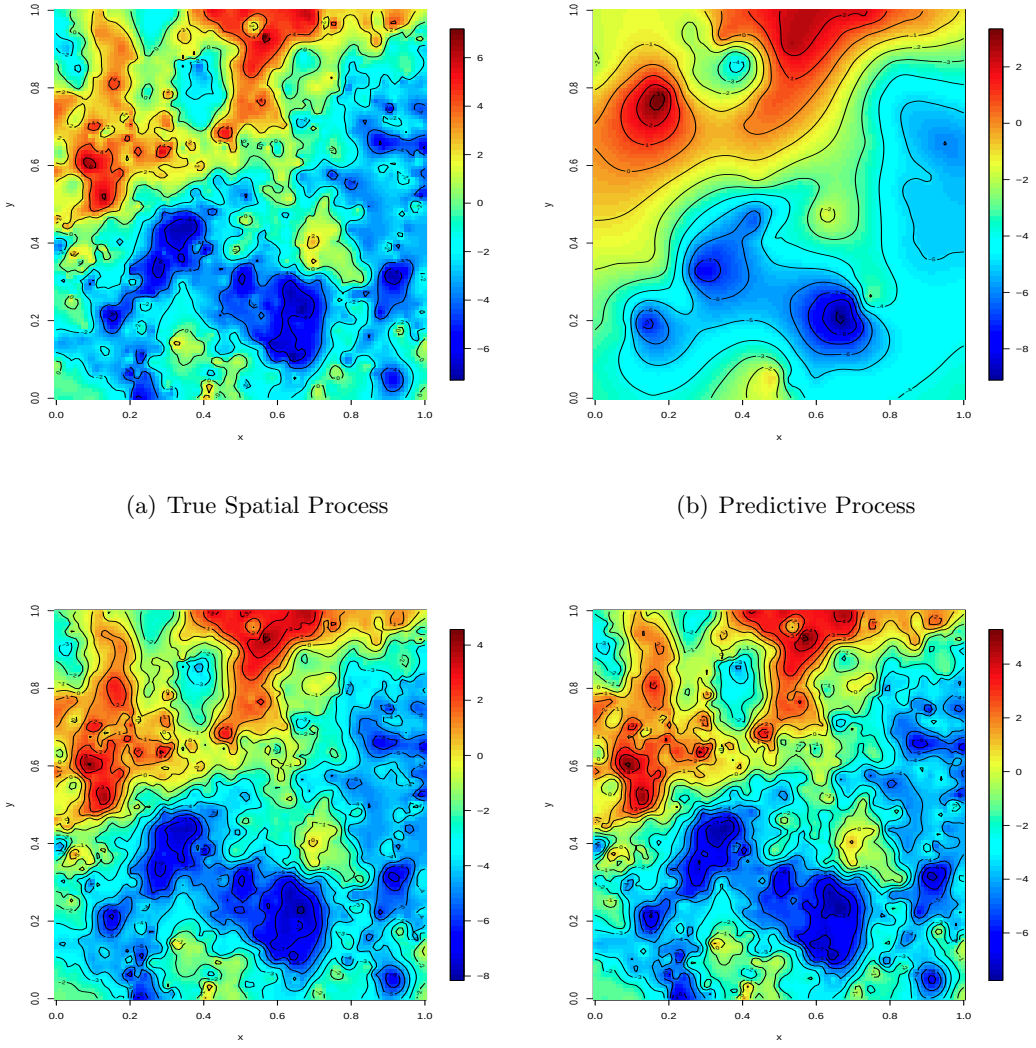


Figure 4.4: plot of simulated data points • overlaid with knots ⊠



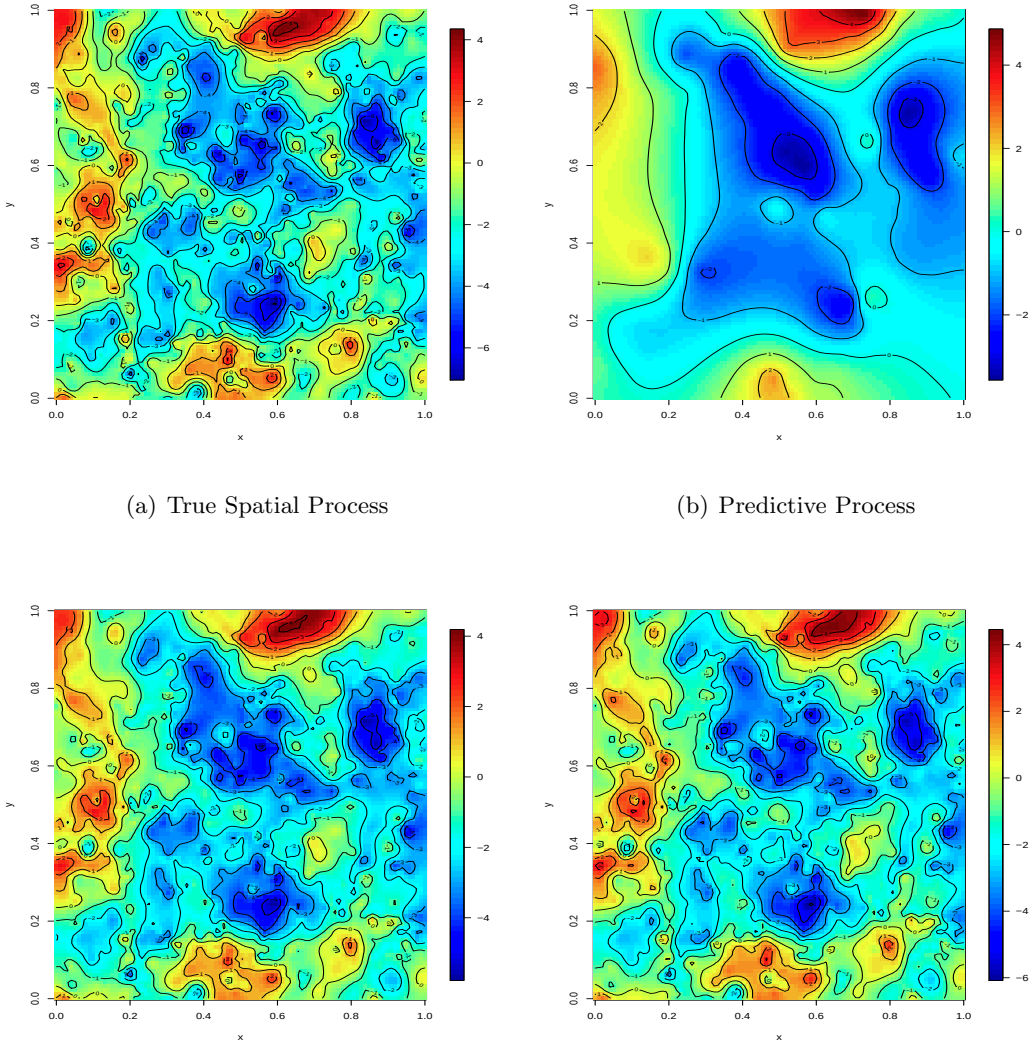
(a) True Spatial Process

(b) Predictive Process

(c) Modified Predictive Process

(d) Tapered Predictive Process

Figure 4.5: True and the estimated (posterior mean) spatial surfaces from the three candidate models for $w_1(\boldsymbol{s})$: (a) True Spatial surface; (b) Predictive process model; (c) Modified Predictive process model; and (d) Tapered Adjustment model.



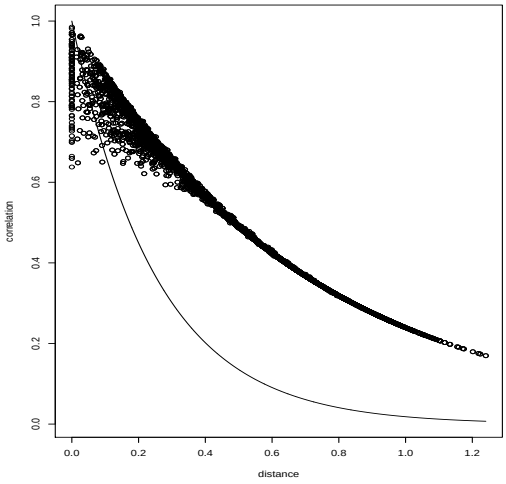
(a) True Spatial Process

(b) Predictive Process

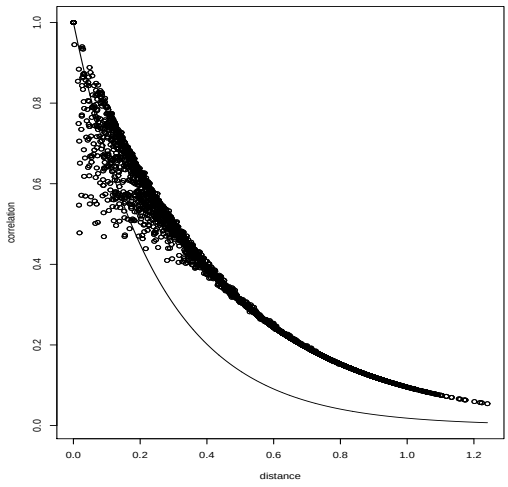
(c) Modified Predictive Process

(d) Tapered Predictive Process

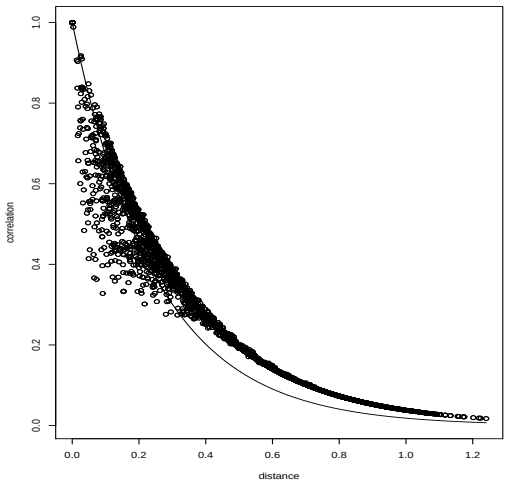
Figure 4.6: True and the estimated (posterior mean) spatial surfaces from the three candidate models for $w_2(\boldsymbol{s})$: (a) True Spatial surface; (b) Predictive process model; (c) Modified Predictive process model; and (d) Tapered Adjustment model.



(a) Predictive Process

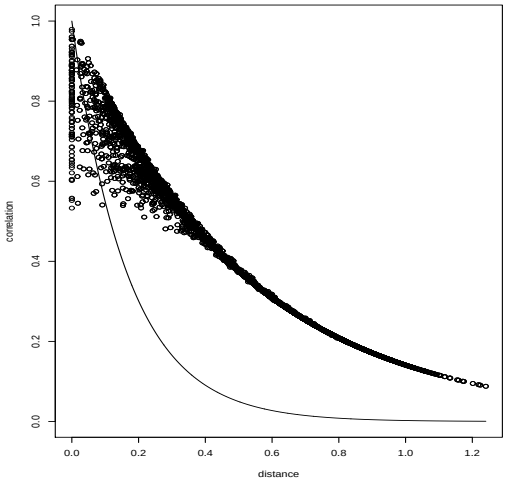


(b) Modified Predictive Process

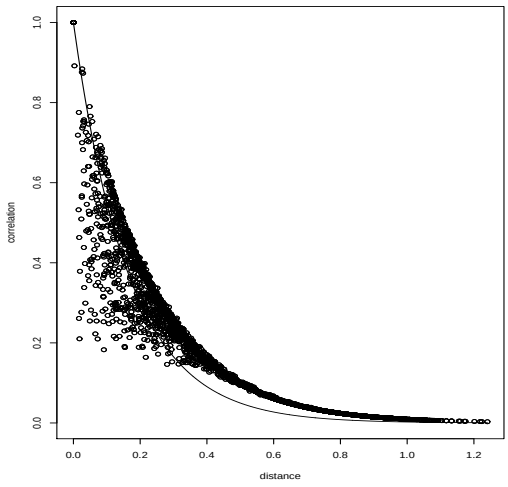


(c) Tapered Predictive Process

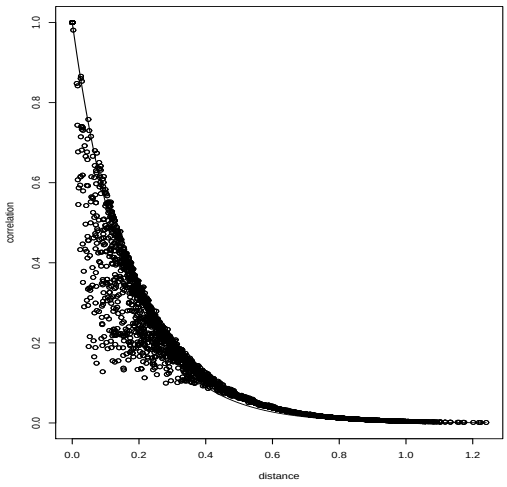
Figure 4.7: Estimated spatial correlation for the latent process $v_1(\boldsymbol{s})$ overlaid with the true exponential correlation: (a) Predictive process model; (b) Modified Predictive process model; and (c) Tapered Predictive process model.



(a) Predictive Process

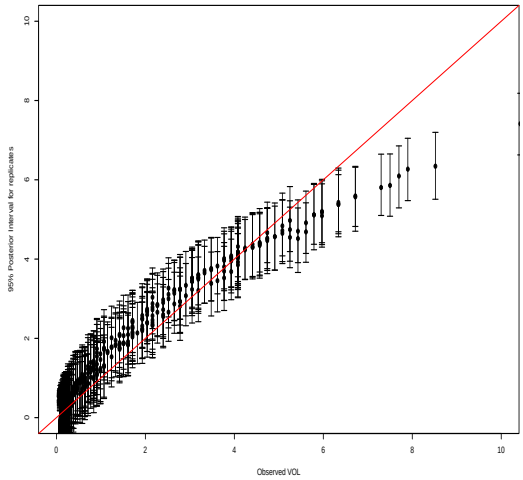


(b) Modified Predictive Process

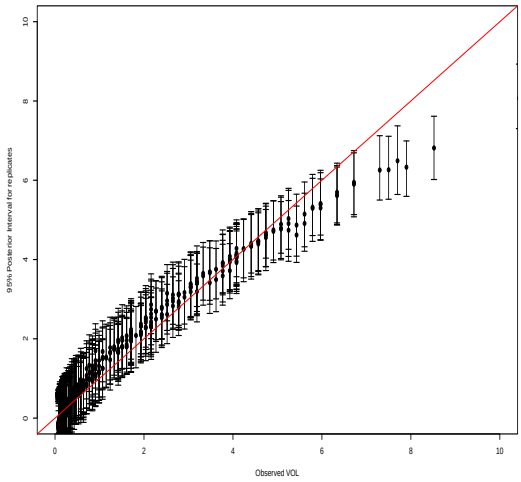


(c) Tapered Predictive Process

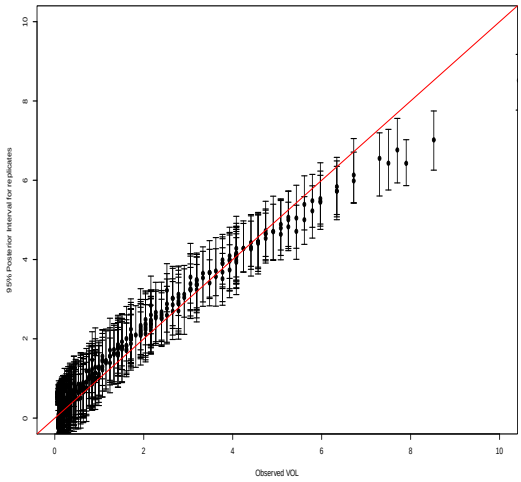
Figure 4.8: Estimated spatial correlation for the latent process $v_2(\boldsymbol{s})$ overlaid with the ture exponential correlation: (a) Predictive process model; (b) Modified Predictive process model; and (c) Tapered Predictive process model.



(a) Predictive Process

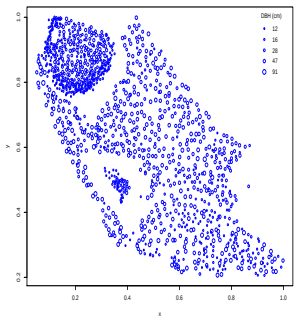


(b) Modified Predictive Process

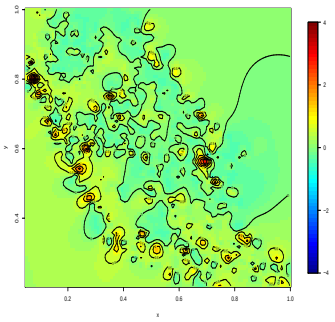


(c) Tapered Predictive Process

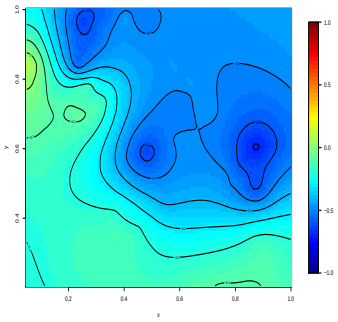
Figure 4.9: Posterior Median and the 95% CI from replicated data (plotted vs. observed VOL): (c) Predictive process model; (b) Modified Predictive process model; and (c) Tapered Predictive process model.



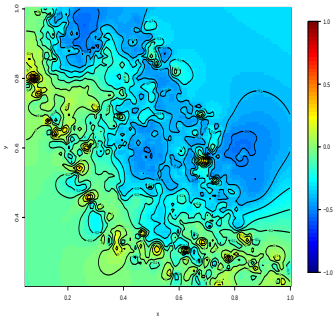
(a) Data Locations



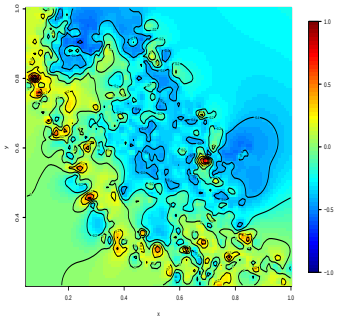
(b) Non Spatial Process



(c) Predictive Process



(d) Modified Predictive Process



(e) Tapered Predictive Process

Figure 4.10: Zurich data set and the estimated (posterior mean) spatial surface from the four candidate models:(a) Data Locations with larger circles representing larger values of DBH (cm); (b) Non-Spatial Model; (c) Predictive process model; (d) Modified Predictive process model; and (e) Tapered Predictive process model.

References

Adler, R.J. (1981). *The Geometry of Random Fields*, Chichester, U.K., John Wiley & Sons.

Adler, R.J., Taylor, J.E. (2007). *Random Fields and Geometry*, New York, U.S.A., Springer-Verlag

Apanasovich, T.V., and Genton, M.G. (2010). Cross-covariance Functions for Multivariate Random Fields Based on Latent Dimensions. *Biometrika*, **97**, 15–30.

Banerjee, S., Carlin, B.P., and Gelfand, A.E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*, Boca Raton, FL: Chapman and Hall/CRC Press.

Banerjee, S., Finley, A.O., Waldmann, P. and Ericcson, T. (2010). Hierarchical Spatial Process Models for Multiple Traits in Large Genetic Trials. *Journal of the American Statistical Association*, **105**, 506–521.

Banerjee, S. and Gelfand, A.E. (2003). On Smoothness Properties of Spatial Processes. *Journal of Multivariate Analysis*, **84**, 85–100

Banerjee, S., and Gelfand, A.E. (2006). Bayesian wombling: Curvilinear Gradient Assessment under Spatial Process Models. *Journal of the American Statistical Association*, **101**, 1487–1501

Banerjee, S., and Gelfand, A.E., and Sirmans, C.F. (2006). Directional Rates of Change under Spatial Process Models. *Journal of the American Statistical Association*, **98**, 946–954

Banerjee, S., Gelfand, A.E., Finley, A.O. and Sang, H. (2008). Gaussian Predictive Process Models for Large Spatial Datasets. *Journal of the Royal Statistical Society, Series B*, **70** 825–848.

Banerjee, S. and Johnson, G.A. (2006). Coregionalized Single- and Multi-Resolution Spatially-Varying Growth Curve Modeling With Application to Weed Growth. *Biometrics*, **61**, 617–625.

Bechtold, W.A., Patterson, P.L. (Eds.) (2005). *The enhanced Forest Inventory and Analysis National Sample Design and Estimation Procedures.* SRS-80. U.S. Department of Agriculture, Forest Service, Southern Research Station, Asheville, NC

Chilés, J.P. and Delfiner, P. (1999). *Geostatistics: Modeling Spatial Uncertainty*, New York: Wiley.

Cramér, H. (1940). On the Theory of Stationary Random Processes. *Annals of Mathematics*, **41**, 215–230.

Crainiceanu, C.M., Diggle, P.J., and Rowlingson, B. (2008). Bivariate Binomial Spatial Modeling of Loa Loa Prevalence in Tropical Africa (With Discussion). *Journal of the American Statistical Association*, **103**, 21–37.

Cressie, N.A.C. (1993). *Statistics for Spatial Data*, 2nd edition. New York: Wiley.

Cressie, N.A.C. and Johannesson, G. (2008). Spatial Prediction for Massive Datasets. *Journal of the Royal Statistical Society Series B*, **70**, 209–226.

Cressie, N.A.C. and Wikle, C.K. (2011). *Statistics for Spatio-Temporal Data.* New York: Wiley.

Daniels, M.J. and Kass, R.E. (1999). Nonconjugate Bayesian Estimation of Covariance Matrices and Its Use in Hierarchical Models. *Journal of the American Statistical Association*, **94**, 1254–1263.

Davis, T.A. (2006). *Direct Methods for Sparse Linear Systems.* Philadelphia: Society for Industrial and Applied Mathematics.

Diez, J.M. and Pulliam, H.R. (2007). Hierarchical Analysis of Species Distributions and Abundance Across Environmental Gradients. *Ecology*, **88**, 3144–3152.

Diggle, P.J. (2003). *Statistical Analysis of Spatial Point Patterns*, Second edition. Arnold, London.

Diggle, P. and Lophaven, S. (2006). Bayesian Geostatistical Design. *Scandinavian Journal of Statistics*, **33**, 53–64.

Diggle, P., Menezes, R. and Su, T. (2010). Geostatistical Inference Under Preferential Sampling. *Applied Statistics*, **59**, 191–232.

Diggle, P.J., Tawn, J.A. and Moyeed, R.A. (1998). Model-Based Geostatistics (with discussion). *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, **47**, 299–350.

Diggle, P.J. and Ribeiro, P.J. (2007). *Model-Based Geostatistics*. New York: Springer.

Du, J., Zhang, H., Mandrekar, V.S. (2009) Fixed-Domain Asymptotic Properties of Tapered Maximum Likelihood Estimators.*Ann. Statist.*, **37**, 3330–3361

Dong J., Kaufmann, R.K., Myneni, R.B., Tucker, C.J., Kauppi, P.E., Liski, J., Buermann, W., Alexeyev, V., and Hughes, M.K. (2003). Remote Sensing Estimates of Boreal and Temperate Forest Woody Biomass: Carbon Pools, Sources, and Sinks. *Remote Sensing of Environment*, **84**, 393–410.

Ecker, M.D. and Gelfand, A.E. (2003). Spatial Modeling and Prediction Under Stationary Non-Geometric Range Anisotropy. *Environmental and Ecological Statistics*, **10** 165–178.

Ecker, M.D. and Gelfand, A.E. (1999). Bayesian Modeling and Inference for Geometrically Anisotropic Spatial Data. *Mathematical Geology*, **31**, 67–83.

Eidsvik, J., Finley, A.O., Banerjee, S. and Rue, H. (2010). Approximate Bayesian Inference for Large Spatial Datasets Using Predictive Process Models. Technical Report, Norwegian University of Science and Technology.

Finley, A.O., Banerjee, S., Ek, A.R., and McRoberts, R.E. (2008). Bayesian Multivariate Process Modeling for Prediction of Forest Attributes. *Journal of Agricultural, Biological, and Environmental Statistics*, **13**, 60–83.

Finley, A.O., Sang, H., Banerjee, S., and Gelfand, A.E. (2009a). *Improving the Performance of Predictive Process Modeling for Large Datasets. Computational Statistics and Data Analysis*, **53**, 2873–2884.

Finley, A.O., Banerjee, S. and McRoberts, R.E. (2009b). Hierarchical Spatial Models for Predicting Tree Species Assemblages Across Large Domains. *Annals of Applied Statistics*, **3**, 1052–1079.

Finley, A.O., S. Banerjee, and D.W. MacFarlane. (2011). A Hierarchical Model for Quantifying Forest Variables Over Large Heterogeneous Landscapes With Uncertain Forest Areas. *Journal of the American Statistical Association*, **106**, 31–48.

Finley, A.O., Banerjee, S., Waldmann, P., and Ericsonn, T. (2009). Hierarchical Spatial Modeling of Additive and Dominance Genetic Variance for Large Spatial Trial Datasets. *Biometrics*, **61**, 441–451.

Finzi A.C., van Breemen N and Canham CD. (1998). Canopy Tree-Soil Interactions Within Temperate Forests: Species Effects on pH and Base Cations. *Ecological Applications*, **8**, 447–454.

Fuentes, M. (2007) Approximate Likelihood for Large Irregularly Spaced Spatial Data. *Journal of the American Statistical Association*, **102**, 321–331.

Furrer, R., Genton, M.G., Nychka, D. (2006) Covariance Tapering for Interpolation of Large Spatial Datasets. *Journal of Computational and Graphical Statistics*, **15**, 502–523

Furrer, R., Sain, S.R. (2009) Spatial Model Fitting for Large Datasets with Applications to Climate and Microarray Problems. (2009) *Statistics and Computing*, **19**, 113–128

Gaspari, G. and Cohn, S.E. (1999). Construction of Correlation Functions in Two and Three Dimensions. *Quarterly Journal of the Royal Metereological Society*, **125**,

723–757.

Gelman, A., Carlin, J.B., Stern, H.S., and Rubin, D.B. (2004). *Bayesian Data Analysis*. Second Edition. Boca Raton, FL: Chapman and Hall/CRC Press.

Gelfand, A.E. and Banerjee, S. (2010). Multivariate Spatial Process Models. In *Handbook of Spatial Statistics*, eds. A.E. Gelfand, P. Diggle, P. Guttorp, and M. Fuentes. Boca Raton, FL: Taylor and Francis/CRC, pp. 495–516.

Gelfand, A.E. and S.K. Ghosh. (1998). Model Choice: a Minimum Posterior Predictive Loss Approach. *Biometrika*, **85**, 1–11.

Gelfand, A.E., Schmidt, A.M., Banerjee, S., and Sirmans, C.F. (2004). Nonstationary Multivariate Process Modeling Through Spatially Varying Coregionalization (with discussion). *Test*, **13**, 263–312.

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2004). *Bayesian Data Analysis*, 2nd edition. Boca Raton, FL: Chapman and Hall/CRC Press.

Gneting, T. (2002). Compactly Supported Correlation Functions. *Journal of Multivariate Analysis*, **83**, 493–508.

Gneiting, T. and Guttorp, P. (2010). Continuous-Parameter Stochastic Process Theory. In *Handbook of Spatial Statistics*, eds. A.E. Gelfand, P. Diggle, P. Guttorp and m. Fuentes. Boca Raton, FL: Taylor and Francis/CRC, pp. 17-28.

Gneiting, T., Kleiber, W. and Schlather, M. (2010). Matérn Cross-Covariance Functions for Multivariate Random Fields. *Journal of the American Statistical Association*, **105**, 1167–1177.

Grzebyk, M. and Wackernagel, H. (1994). Multivariate Analysis and Spatial/Temporal Scales: Real and Complex Models. In *Proceedings of the XVIIth International Biometrics Conference*, Hamilton, Ontario, Canada: International Biometric Society, pp. 19–33.

Guhaniyogi, R., Finley, A.O., Banerjee, S. and Gelfand, A.E. (2011a). Adaptive Gaussian Predictive Process Models for Large Spatial Datasets. *Environmetrics*, **22**, 997–1007.

Guhaniyogi, R., Finley, A.O., Kobe, R., Banerjee, S. (2011b). Modeling and Mapping Non-Stationary Multivariate Processes for Large Spatial Datasets. Submitted to *Journal of the American Statistical Association.*

Harville, D.A. (1997) *Matrix Algebra from a Statistician's Perspective.* New York: Springer.

Henderson, H.V. and Searle, S.R. (1981). On Deriving the Inverse of a Sum of Matrices. *SIAM Review*, **23**, 53–60.

Higdon, D. (2002). Space and Space-Time Modeling Using Process Convolutions. In *Quantitative methods for current environmental issues*, eds. C. Anderson, V. Barnett, P. C. Chatwin, and A. H. El-Shaarawi, 37–56. Springer-Verlag.

Hodges, J.S., Sargent, D.J. (2001). Counting Degrees of Freedom in Hierarchical and Other Richly Parametrised Models. *Biometrika*, **88**, 367–379.

Holste E.K., Kobe R.K., Vriesendorp C.F. (2010). Seedling Growth Responses to Soil Nutrients in the Forest Understory. *Ecology*, In revision.

Houlton B.Z., Wang Y.P., Vitousek P.M., and Field C.B. (2008). A Unifying Framework for Dinitrogen Fixation in the Terrestrial Biosphere. *Nature*, **454**, 327–331.

Huang, C., Wylie, B., Homer, C., Yang, L., and Zylstra, G. (2002). Derivation of a Tasseled Cap Transformation Based on Landsat 7 at-Satellite Reflectance. *International Journal of Remote Sensing*, **8**, 1741–1748.

Kaufman L. and Rousseeuw P.J. (1990), *Finding Groups in Data: An Introduction to Cluster Analysis.* Wiley, New York.

Kaufman, C.G., Schervish, M.J. and Nychka, D.W. (2009). Covariance Tapering for Likelihood-Based Estimation in Large Spatial Data Sets. *Journal of the American Statistical Association*, **103**, 1545–1555.

Kobe, R.K. and Vriesendorp C.F. (2009). Size of sampling Unit Strongly Influences Detection of Seedling Limitation in a Wet Tropical Forest. *Ecology Letters*, **12**, 220–228.

Majumdar, A., and Gelfand, A.E. (2007). Multivariate Spatial Modeling for Geostatistical Data Using Convolved Covariance Functions. *Mathematical Geology*, **39**, 225–245.

Mandallaz, D. (2008). Sampling Techniques for Forest Inventories. Boca Raton, FL: Chapman and Hall/CRC Press.

Mardia, K.V., Kent, J.T., Goodall, C.R., and Little, J.A. (1996). Kriging and Splines With Derivative Information *Biometrika*, **83**, 207–221.

McCarthy-Neumann S. and Kobe R.K. (2010). Conspecific Plant-Soil Feedbacks Reduce Survivorship and Growth of Tropical Tree Seedlings. *Journal of Ecology*, **98**, 396–407.

Neal, R. M. (1998) Markov Chain Sampling Methods for Dirichlet Process Mixture Models. Technical Report No. 9815, Department of Statistics, University of Toronto.

Ovaskainen, O., Hottola, J., and Siitonen, J. (2010). Modeling Species Co-Occurrence by Multivariate Logistic Regression Generates New Hypotheses on Fungal Interactions. *Ecology*, **9**, 2414–2521.

Pati, D., Reich B.J. and Dunson D.B. (2011). Bayesian Geostatistical Modeling with Informative Sampling Locations. *Biometrika*, **98**, 35–48.

Pourahmadi, M. (1999). Joint Mean-Covariance Model with Applications to Longitudinal Data: Unconstrained Parameterization. *Biometrika*, **86**, 677–690.

Rao, C.R. (1973). *Linear Statistical Inference and Its Applications.* 2nd Ed. New York: John Wiley and Sons.

Rasmussen, C.E., and Williams, C.K.I. (2006). *Gaussian Processes for Machine Learning.* Cambridge, Massachusetts: The MIT Press.

Robert, C. P. and Casella, G. (2010). *An Introduction to Monte Carlo Methods with R.* New York: Spring-Verlag.

Roberts, G.O. and Rosenthal, J.S. (2009). Examples of Adaptive MCMC. *Journal of Computational and Graphical Statistics*, **18**, 349–367.

Royle, J.A. and Nychka, D. (1998). An Algorithm for the Construction of Spatial Coverage Designs with Implementation in SPLUS. *Computers and Geosciences*, **24**, 479–88

Rue, H., Martino, S.. and Chopin, N. (2009). Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations (With Discussion). *Journal of the Royal Statistical Society, Series B*, **71**, 1–35.

Ruppert, D., Wand, M.P. and Caroll, R.J. (2003). *Semiparametric Regression*. Cambridge University Press.

Sang, H. and Huang, J.Z. (2011). A Full-scale Approximation of Covariance Functions for Large Spatial Data Sets. *Journal of the Royal Statistical Society*, Series B, in press

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

Schmidt, A.M. and Gelfand, A.E. (2003). A Bayesian Coregionalization Approach for Multivariate Pollutant Data, *Journal of Geophysical Research - Atmosphere* **108, D24**, 8783.

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., Linde, A.V.D. (2002) Bayesian Measures of Model Complexity and Fit. *Journal of Royal Statistical Society:B*. **64**, 583–639.

Stein, M.L. (1999). *Interpolation of Spatial Data: Some Theory of Kriging*. New York: Springer.

Stein, M.L. (2007). Spatial Variation of Total Column Ozone on a Global Scale. *Annals of Applied Statistics*, **1**, 191–210.

Stein, M.L. (2008). A Modeling Approach for Large Spatial Datasets. *Journal of the Korean Statistical Society*, **37**, 3–10.

Stein, M.L., Chi, Z., and Welty, L.J. (2004). Approximating Likelihoods for Large Spatial Datasets. *Journal of the Royal Statistical Society, Series B*, **66**, 275–296.

Tokdar, S.T. (2007). Towards a Faster Implementation of Density Estimation with Logistic Gaussian Process Priors. *Journal of Computational and Graphical Statistics*, **16**, 633–655.

Tokdar, S.T. (2011). Adaptive Gaussian Predictive Process Approximation. Duke Statistical Science Discussion Paper 11-13.

Townsend A.R., Asner G.P., and Cleveland C.C. (2008). The Biogeochemical Heterogeneity of Tropical Soils. *Trends in Ecology and Evolution*, **23**, 424–431.

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

Ver Hoef, J.M. and Barry, R.P. (1998). Constructing and Fitting Models for Cokriging and Multivariable Spatial Prediction. *Journal of Statistical Planning and Inference*, **69**, 275–294.

Ver Hoef, J.M., Cressie, N.A.C. and Barry, R.P. (2004). Flexible Spatial Models for Kriging and Cokriging Using Moving Averages and the Fast Fourier Transform (FFT). *Journal of Computational and Graphical Statistics*, **13**, 265–282.

Wackernagel, H. (2006). *Multivariate Geostatistics: An Introduction With Applications, 3rd edition.* New York: Springer-Verlag.

Waddle, J.H., Dorazio, R.M., Walls, S.C., Rice, K.G., Beauchamp, J. Schuman, M.J., and Mazzotti, F.J. (2010). A New Parameterization for Estimating Co-occurrence of Interacting Species. *Ecological Applications*, **20**, 1467–1475.

Wahba, G. (1990). *Spline Models for Observational Data.* SIAM, Philadelphia.

Walker T.W. and Syers J.K. (1976). The Fate of Phosphorus During Pedogenesis. *Geoderma*, **15**, 1–19.

Wardle D.A., Walker L.R. and Bardgett R.D. (2004). Ecosystem Properties and Forest Decline in Contrasting Long-term Chronosequences. *Science*, **305**, 509–512.

Wendland, H. (1995). Piecewise polynomial, positive definite and Compactly Supported Radial Functions of Minimal Degree. *Advances in Computational Mathematics*, **4**, 389–396

Xia, G., Gelfand, A.E. (2005). Stationary Process Approximation for the Analysis of Large Spatial Datasets. *Technical Report, ISDS, Duke University.*

Yaglom, A.M. (1987). *Correlation Theory of Stationary and Related Random Functions*, Vol. I. New York: Springer Verlag.

Zhang, H. (2004) Inconsistent Estimation and Asymptotically Equivalent Interpolations in Model-Based Geostatistics. *Journal of the American Statistical Association*, **99**, 250–261.

Zhang, H. (2007). Maximum-Likelihood Estimation for Multivariate Spatial Linear Coregionalization Models. *Environmetrics*, **18**, 125–139.

Appendix A

Appendix for Chapter 2

proof of posterior propriety

For the sake of simplicity we assume that $[\mathcal{S}^*|\alpha] = \frac{\prod_{i=1}^{n^*} \exp\big(\alpha f(\boldsymbol{s}_i^*)\big)}{\big(\sum_{j=1}^{n^*} \exp\big(\alpha f(\boldsymbol{s}_j^*)\big)\big)^{n^*}}$, with f as a fixed function. For notational convenience, we suppress dependence of $\mathcal{C}(\mathcal{S}^*, \boldsymbol{\theta}), \boldsymbol{C}(\mathcal{S}^*, \boldsymbol{\theta})$ on $\boldsymbol{\theta}$ and denote them as $\mathcal{C}(\mathcal{S}^*)$ and $\boldsymbol{C}(\mathcal{S}^*)$ respectively. Let's also call $D^{n^*} - \{\boldsymbol{s}_1^* = \cdots = \boldsymbol{s}_{n^*}^*\}$ as $D_{\mathcal{S}^*}^{n^*}$. We will then prove posterior propriety of α under the following assumptions,
(a) $D \subset \mathcal{R}^2$ is bounded.
(b) $\mathcal{C}(\mathcal{S}^*)$ has full row rank for $\mathcal{S}^* \in D_{\mathcal{S}^*}^{n^*}$ and $\inf_{\mathcal{S}^* \in D_{\mathcal{S}^*}^{n^*}} |\mathcal{C}(\mathcal{S}^*)\mathcal{C}(\mathcal{S}^*)'| > 0$
(c) $\inf_{\mathcal{S}^* \in D_{\mathcal{S}^*}^{n^*}} ||\boldsymbol{g}^*||_2 > 0$, where $\boldsymbol{g}^* = (f(\boldsymbol{s}_i^*) - f(\boldsymbol{s}_j^*))_{1 \le i \ne j \le n^*}$, $||\cdot||_2$ is the L_2 norm.
We wish to prove posterior propriety of α for fixed ϕ and σ^2. The result can then straightforwardly be extended for priors on ϕ and σ^2 with bounded supports. Note that, the posterior distribution of $\big(\alpha, \boldsymbol{\beta}, w(\mathcal{S}^*), \mathcal{S}^*, \tau^2\big)$ given $\boldsymbol{y}$ is,

$$f\left(\alpha, \boldsymbol{\beta}, w(\mathcal{S}^*), \mathcal{S}^*, \tau^2 | \boldsymbol{y}\right) \propto N\left(\boldsymbol{y} \,|\, \boldsymbol{X}\boldsymbol{\beta} + \mathcal{C}(\mathcal{S}^*)'\boldsymbol{C}(\mathcal{S}^*)^{-1} w(\mathcal{S}^*), \tau^2 \boldsymbol{I}\right) \times$$
$$N\left(w(\mathcal{S}^*) \,|\, \boldsymbol{0}, \boldsymbol{C}(\mathcal{S}^*)\right) \times p(\boldsymbol{\beta}) \times p(\tau^2) \times p(\alpha) \times \frac{\prod_{i=1}^{n^*} \exp\left(\alpha f(\boldsymbol{s}_i^*)\right)}{\left(\sum_{j=1}^{n^*} \exp\left(\alpha f(\boldsymbol{s}_j^*)\right)\right)^{n^*}}$$

We will assign proper prior distributions on $\boldsymbol{\beta}$ and τ^2 and flat prior, $p(\alpha) \propto 1$, on α. Let,

$$L_1 = \int \cdots \int f(\alpha, \boldsymbol{\beta}, w(\mathcal{S}^*), \mathcal{S}^*, \tau^2 | \boldsymbol{y}) dw(\mathcal{S}^*)\, d\boldsymbol{\beta}\, d\alpha\, d\mathcal{S}^*\, d\tau^2 \tag{A.1}$$

Consider the inner integrals w.r.t $(w(\mathcal{S}^*), \boldsymbol{\beta})$,

$$L = \int\int N\left(\boldsymbol{y} \,|\, \boldsymbol{X}\boldsymbol{\beta} + \mathcal{C}(\mathcal{S}^*)'\boldsymbol{C}(\mathcal{S}^*)^{-1}w(\mathcal{S}^*), \tau^2\boldsymbol{I}\right) \times N\left(w(\mathcal{S}^*) \,|\, \boldsymbol{0}, \boldsymbol{C}(\mathcal{S}^*)\right) p(\boldsymbol{\beta}) dw(\mathcal{S}^*) d\boldsymbol{\beta}$$

Denote, $\boldsymbol{\Sigma}_1 = \left[\boldsymbol{C}(\mathcal{S}^*)^{-1}\mathcal{C}(\mathcal{S}^*)(\tau^2\boldsymbol{I})^{-1}\mathcal{C}(\mathcal{S}^*)'\boldsymbol{C}(\mathcal{S}^*)^{-1} + \boldsymbol{C}(\mathcal{S}^*)^{-1}\right]^{-1}$,
$\boldsymbol{\Sigma}_2 = \boldsymbol{C}(\mathcal{S}^*)^{-1}\mathcal{C}(\mathcal{S}^*)(\tau^2\boldsymbol{I})^{-1}$. Completing quadratic form with $w(\mathcal{S}^*)$ and integrating w.r.t $w(\mathcal{S}^*)$, we obtain,

$$\begin{aligned} L &= \frac{|\boldsymbol{\Sigma}_1|^{\frac{1}{2}}}{|\tau^2\boldsymbol{I}|^{\frac{1}{2}}|\boldsymbol{C}(\mathcal{S}^*)|^{\frac{1}{2}}} \times \int \exp\{-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})' \left[(\tau^2\boldsymbol{I})^{-1} - \boldsymbol{\Sigma}_2'\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2\right](\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\} p(\boldsymbol{\beta}) d\boldsymbol{\beta} \\ &\leq \frac{|\boldsymbol{\Sigma}_1|^{\frac{1}{2}}}{|\tau^2\boldsymbol{I}|^{\frac{1}{2}}|\boldsymbol{C}(\mathcal{S}^*)|^{\frac{1}{2}}} \int p(\boldsymbol{\beta}) d\boldsymbol{\beta} = \frac{|\boldsymbol{\Sigma}_1|^{\frac{1}{2}}}{|\tau^2\boldsymbol{I}|^{\frac{1}{2}}|\boldsymbol{C}(\mathcal{S}^*)|^{\frac{1}{2}}} \end{aligned} \tag{A.2}$$

The second inequality follows from the fact that $\left[(\tau^2\boldsymbol{I})^{-1} - \boldsymbol{\Sigma}_2'\boldsymbol{\Sigma}_1\boldsymbol{\Sigma}_2\right]$ is positive definite by applying Sherman-Woodbury-Morrison matrix inversion formula. Combining (A.1) and (A.2) we obtain,

$$L_1 \leq \int\int\int \frac{|\boldsymbol{\Sigma}_1|^{\frac{1}{2}}}{|\tau^2\boldsymbol{I}|^{\frac{1}{2}}|\boldsymbol{C}(\mathcal{S}^*)|^{\frac{1}{2}}} \times p(\tau^2) \times \prod_{i=1}^{n^*} \frac{\exp(\alpha f(\boldsymbol{s}_i^*))}{\sum_{j=1}^{n^*}\exp(\alpha f(\boldsymbol{s}_j^*))} d\alpha d\mathcal{S}^* d\tau^2 \tag{A.3}$$

Next, we introduce latent variables $\boldsymbol{t} = (t_1, ..., t_{n^*})$ to rewrite $\prod_{i=1}^{n^*} \frac{\exp(\alpha f(\boldsymbol{s}_i^*))}{\sum_{j=1}^{n^*}\exp(\alpha f(\boldsymbol{s}_j^*))}$ as,

$$\begin{aligned} \prod_{i=1}^{n^*} \frac{\exp(\alpha f(\boldsymbol{s}_i^*))}{\sum_{j=1}^{n^*}\exp(\alpha f(\boldsymbol{s}_j^*)} &= \prod_{i=1}^{n^*} \frac{1}{\left[1 + \sum_{j\neq i}\exp\{\alpha(f(\boldsymbol{s}_j^*) - f(\boldsymbol{s}_i^*))\}\right]} \\ &= \int_{\mathcal{R}_+^{n^*}} \prod_{i=1}^{n^*} \exp\left\{-t_i\left[1 + \sum_{j\neq i}\exp\{\alpha(f(\boldsymbol{s}_j^*) - f(\boldsymbol{s}_i^*))\}\right]\right\} d\boldsymbol{t} \end{aligned}$$

Denote $\frac{|\boldsymbol{\Sigma}_1|^{\frac{1}{2}}}{|\tau^2\boldsymbol{I}|^{\frac{1}{2}}|\boldsymbol{C}(\mathcal{S}^*)|^{\frac{1}{2}}} = K(\mathcal{S}^*, \tau^2)$, $d(-F^{t_i}(-v_{ij})) = t_i \exp(v_{ij}) \exp\{-t_i \exp(v_{ij})\} dv_{ij}$,
$dF_{\boldsymbol{t}}(\boldsymbol{v}) = \prod_{1\leq i\neq j\leq n^*} d(-F^{t_i}(-v_{ij}))$. Following algebra in the proof of Theorem 1 in

Chen et. al. (Biometrika, 2006), (A.3) is written as

$$\begin{aligned} L_1 &\leq \int\int\int p(\tau^2)K(\mathcal{S}^*,\tau^2) \int\limits_{\mathcal{R}_+^{n^*}} \exp\{-\sum_{i=1}^{n^*} t_i\} \prod_{1\leq i\neq j\leq n^*} \int\limits_{\mathcal{R}} \mathbf{1}(v_{ij} \geq \alpha\{f(\boldsymbol{s}_i^*) - f(\boldsymbol{s}_j^*)\}) \\ &\quad d(-F^{t_i}(-v_{ij}))\, d\boldsymbol{t}\, d\alpha\, d\mathcal{S}^*\, d\tau^2 \;\; ; \;\; \text{applying Fubini's theorem} \\ &= \int\int p(\tau^2)K(\mathcal{S}^*,\tau^2) \int\limits_{\mathcal{R}_+^{n^*}} \exp\{-\sum_{i=1}^{n^*} t_i\} \int\limits_{\mathcal{R}^{n^*(n^*-1)}} \int\limits_{\mathcal{R}_+} \mathbf{1}(\boldsymbol{g}^*\alpha \preceq \boldsymbol{v}) \end{aligned} \tag{A.4}$$

$\preceq$ represents elementwise inequality between vectors. $\{\boldsymbol{g}^*\alpha \preceq \boldsymbol{v}\} \subseteq \left\{\alpha \leq \frac{||\boldsymbol{v}||_2}{||\boldsymbol{g}^*||_2}\right\}$ together with (A.4) yields

$$\begin{aligned} L_1 &\leq \int\int p(\tau^2)K(\mathcal{S}^*,\tau^2) \int\limits_{\mathcal{R}_+^{n^*}} \exp\{-\sum_{i=1}^{n^*} t_i\} \int\limits_{\mathcal{R}^{n^*(n^*-1)}} \int\limits_{\mathcal{R}_+} \mathbf{1}(\alpha \leq \frac{||\boldsymbol{v}||_2}{||\boldsymbol{g}^*||_2}) \\ &= \int\int p(\tau^2)K(\mathcal{S}^*,\tau^2) \int\limits_{\mathcal{R}_+^{n^*}} \exp\{-\sum_{i=1}^{n^*} t_i\} \int\limits_{\mathcal{R}^{n^*(n^*-1)}} \frac{||\boldsymbol{v}||_2}{||\boldsymbol{g}^*||_2} \\ &= K_1 \int\int \frac{p(\tau^2)K(\mathcal{S}^*,\tau^2)}{||\boldsymbol{g}^*||_2} \int ||\boldsymbol{v}||_2 \left\{\prod_{1\leq j\neq i\leq n^*} \exp(v_{ij})\right\} \prod_{i=1}^{n^*} \left\{1 + \sum_{j\neq i} \exp(v_{ij})\right\}^{-n^*} \\ &\quad \text{After some algebra we obtain,} \\ &\leq K_1' \int\int p(\tau^2) \frac{K(\mathcal{S}^*,\tau^2)}{||\boldsymbol{g}^*||_2} d\mathcal{S}^* d\tau^2 \;\; , \;\; K_1, K_1' \text{ are constants} \end{aligned}$$

Now note that, $K(\mathcal{S}^*,\tau^2) = \frac{|\boldsymbol{C}(\mathcal{S}^*)|^{\frac{1}{2}}|\mathcal{C}(\mathcal{S}^*)(\tau^2\boldsymbol{I})^{-1}\mathcal{C}(\mathcal{S}^*)' + \boldsymbol{C}(\mathcal{S}^*)|^{-\frac{1}{2}}}{|\tau^2\boldsymbol{I}|^{\frac{1}{2}}|\boldsymbol{C}(\mathcal{S}^*)|^{\frac{1}{2}}} = \frac{1}{|\mathcal{C}(\mathcal{S}^*)\mathcal{C}(\mathcal{S}^*)' + \tau^2\boldsymbol{C}(\mathcal{S}^*)|^{\frac{1}{2}}}$.
Note that, $c_{(1)}$ = smallest eigenvalue of $\boldsymbol{C}(\mathcal{S}^*)(\mathcal{C}(\mathcal{S}^*)\mathcal{C}(\mathcal{S}^*)')^{-1} > 0$ by assumption (a).
Let's denote, $K_2(\mathcal{S}^*) = |(\mathcal{C}(\mathcal{S}^*)\mathcal{C}(\mathcal{S}^*)')|^{\frac{1}{2}}$; applying assumption (a) & (b) together,

$$\begin{aligned} L_1 &\leq \int\int \frac{1}{|\boldsymbol{I} + \tau^2\boldsymbol{C}(\mathcal{S}^*)(\mathcal{C}(\mathcal{S}^*)\mathcal{C}(\mathcal{S}^*)')^{-1}|^{\frac{1}{2}}|(\mathcal{C}(\mathcal{S}^*)\mathcal{C}(\mathcal{S}^*)')|^{\frac{1}{2}}||\boldsymbol{g}^*||_2} p(\tau^2) d\mathcal{S}^* d\tau^2 \\ &\leq \int\int \frac{1}{(1 + c_{(1)}\tau^2)^{\frac{n^*}{2}} K_2(\mathcal{S}^*)||\boldsymbol{g}^*||_2} p(\tau^2) d\tau^2 d\mathcal{S}^* \leq \int \frac{1}{K_2(\mathcal{S}^*)||\boldsymbol{g}^*||_2} d\mathcal{S}^* < \infty \end{aligned}$$

Appendix B

Appendix for Chapter 3

Marginal cross covariances for $\boldsymbol{w}_{mpp}(\boldsymbol{s})$ and $\tilde{\boldsymbol{w}}(\boldsymbol{s})$

Here we provide the details for deriving the relationships between the marginal cross-covariances for $\boldsymbol{w}_{mpp}(\boldsymbol{s})$ and $\tilde{\boldsymbol{w}}(\boldsymbol{s})$. The derivations we present work not only for triangular $\boldsymbol{A}(\boldsymbol{s})$, but more generally for any matrix $\boldsymbol{A}(\boldsymbol{s})$ whose elements are independent non-degenerate spatial processes.

Note that $\operatorname{cov}\left\{\mathrm{E}\left[\tilde{\boldsymbol{A}}(\boldsymbol{s})\tilde{\boldsymbol{v}}(\boldsymbol{s}) \,|\, \{\boldsymbol{a}_{ij}^*\}\right], \mathrm{E}\left[\tilde{\boldsymbol{A}}(\boldsymbol{t})\tilde{\boldsymbol{v}}(\boldsymbol{t}) \,|\, \{\boldsymbol{a}_{ij}^*\}\right]\right\} = 0$ because processes $a_{ij}(\cdot)$ and $v_k(\cdot)$ are independent. This yields

$$\begin{aligned}\mathrm{E}\left[\boldsymbol{C}_{\tilde{w}}(\boldsymbol{s},\boldsymbol{t})\right] &= \operatorname{cov}\{\tilde{\boldsymbol{A}}(\boldsymbol{s})\tilde{\boldsymbol{v}}(\boldsymbol{s}), \tilde{\boldsymbol{A}}(\boldsymbol{t})\tilde{\boldsymbol{v}}(\boldsymbol{t})\} = \mathrm{E}\left[\operatorname{cov}\{\tilde{\boldsymbol{A}}(\boldsymbol{s})\tilde{\boldsymbol{v}}(\boldsymbol{s}), \tilde{\boldsymbol{A}}(\boldsymbol{t})\tilde{\boldsymbol{v}}(\boldsymbol{t}) \,|\, \{\boldsymbol{a}_{ij}^*\}\}\right] \\ &\quad + \operatorname{cov}\left\{\mathrm{E}\left[\tilde{\boldsymbol{A}}(\boldsymbol{s})\tilde{\boldsymbol{v}}(\boldsymbol{s}) \,|\, \{\boldsymbol{a}_{ij}^*\}\right], \mathrm{E}\left[\tilde{\boldsymbol{A}}(\boldsymbol{t})\tilde{\boldsymbol{v}}(\boldsymbol{t}) \,|\, \{\boldsymbol{a}_{ij}^*\}\right]\right\} \\ &= \mathrm{E}\left[\operatorname{cov}\{\tilde{\boldsymbol{A}}(\boldsymbol{s})\tilde{\boldsymbol{v}}(\boldsymbol{s}), \tilde{\boldsymbol{A}}(\boldsymbol{t})\tilde{\boldsymbol{v}}(\boldsymbol{t}) \,|\, \{\boldsymbol{a}_{ij}^*\}\}\right] = \mathrm{E}\left[\tilde{\boldsymbol{A}}(\boldsymbol{s})\boldsymbol{C}_{\tilde{v}}(\boldsymbol{s},\boldsymbol{t})\tilde{\boldsymbol{A}}(\boldsymbol{t})'\right] .\end{aligned}$$

The (i,j)-th element of $\mathrm{E}\left[\boldsymbol{C}_{\tilde{w}}(\boldsymbol{s},\boldsymbol{t})\right]$ is

$$\mathrm{E}\left[c_{\tilde{w};i,j}(\boldsymbol{s},\boldsymbol{t})\right] = \mathrm{E}\left[\operatorname{tr}\left\{\tilde{\boldsymbol{a}}_{i*}(\boldsymbol{s})'\boldsymbol{C}_{\tilde{v}}(\boldsymbol{s},\boldsymbol{t})\tilde{\boldsymbol{a}}_{j*}(\boldsymbol{t})\right\}\right] = \operatorname{tr}\left\{\boldsymbol{C}_{\tilde{v}}(\boldsymbol{s},\boldsymbol{t})\mathrm{E}\left[\tilde{\boldsymbol{a}}_{j*}(\boldsymbol{t})\tilde{\boldsymbol{a}}_{i*}(\boldsymbol{s})'\right]\right\} , \tag{B.1}$$

where $c_{\tilde{w};i,j}(\boldsymbol{s},\boldsymbol{t})$ is the (i,j)-th entry of $\boldsymbol{C}_{\tilde{w}}(\boldsymbol{s},\boldsymbol{t})$, and $\tilde{\boldsymbol{a}}_{i*}(\boldsymbol{s})'$ and $\tilde{\boldsymbol{a}}_{*j}(\boldsymbol{s})$ are the i-th row and j-th column of $\tilde{\boldsymbol{A}}(\boldsymbol{s})$ respectively. Because the off-diagonal elements of $\boldsymbol{A}(\boldsymbol{s})$ are zero-centered independent Gaussian processes and $\boldsymbol{C}_{\tilde{v}}(\boldsymbol{s},\boldsymbol{t})$ is diagonal, it follows

from (B.1) that the matrix $\mathrm{E}\left[\boldsymbol{C}_{\tilde{w}}(\boldsymbol{s},\boldsymbol{t})\right]$ is diagonal with i-th diagonal element

$$\mathrm{E}\left[c_{\tilde{w};i,i}(\boldsymbol{s},\boldsymbol{t})\right] = \sum_{k=1}^{m} \mathrm{cov}\{\tilde{a}_{ik}(\boldsymbol{s}),\tilde{a}_{ik}(\boldsymbol{t})\}\mathrm{cov}\{\tilde{v}_k(\boldsymbol{s}),\tilde{v}_k(\boldsymbol{t})\} \tag{B.2}$$

$$+ \mathrm{E}[\tilde{a}_{ii}(\boldsymbol{s})]\mathrm{E}[\tilde{a}_{ii}(\boldsymbol{t})]\mathrm{cov}\{\tilde{v}_i(\boldsymbol{s}),\tilde{v}_i(\boldsymbol{t})\}\ . \tag{B.3}$$

Recall that $\boldsymbol{w}_{mpp}(\boldsymbol{s}) = \boldsymbol{w}_{pp}(\boldsymbol{s}) + \tilde{\boldsymbol{\epsilon}}_{w_{pp}}(\boldsymbol{s})$, where the "modification" $\boldsymbol{\epsilon}(\boldsymbol{s}) \overset{ind}{\sim} N(\boldsymbol{0}, \mathrm{var}\{\boldsymbol{w}(\boldsymbol{s})\} - \mathrm{var}\{\boldsymbol{w}_{pp}(\boldsymbol{s})\})$. This implies that $\mathrm{E}\left[\boldsymbol{C}_{w_{mpp}}(\boldsymbol{s},\boldsymbol{t})\right]$ can be written as

$$\begin{aligned}
\mathrm{E}\left[\boldsymbol{C}_{w_{mpp}}(\boldsymbol{s},\boldsymbol{t})\right] &= \mathrm{cov}\{\boldsymbol{w}_{mpp}(\boldsymbol{s}),\boldsymbol{w}_{mpp}(\boldsymbol{t})\} = \mathrm{cov}\{\boldsymbol{w}_{pp}(\boldsymbol{s}) + \boldsymbol{\epsilon}_{w_{pp}}(\boldsymbol{s}),\boldsymbol{w}_{pp}(\boldsymbol{t}) + \boldsymbol{\epsilon}_{w_{pp}}(\boldsymbol{t})\} \\
&= \mathrm{cov}\{\boldsymbol{w}_{pp}(\boldsymbol{s}),\boldsymbol{w}_{pp}(\boldsymbol{t})\} + \mathrm{cov}\{\boldsymbol{\epsilon}_{w_{pp}}(\boldsymbol{s}),\boldsymbol{\epsilon}_{w_{pp}}(\boldsymbol{t})\} \\
&= \mathrm{cov}\{\boldsymbol{A}_{pp}(\boldsymbol{s})\boldsymbol{v}_{pp}(\boldsymbol{s}),\boldsymbol{A}_{pp}(\boldsymbol{t})\boldsymbol{v}_{pp}(\boldsymbol{t})\} + \mathrm{cov}\{\boldsymbol{\epsilon}_{w_{pp}}(\boldsymbol{s}),\boldsymbol{\epsilon}_{w_{pp}}(\boldsymbol{t})\} \\
&= \mathrm{E}\left[\boldsymbol{A}_{pp}(\boldsymbol{s})\mathrm{cov}\{\boldsymbol{v}_{pp}(\boldsymbol{s}),\boldsymbol{v}_{pp}(\boldsymbol{t})\}\boldsymbol{A}_{pp}(\boldsymbol{t})'\right] + \mathrm{cov}\{\boldsymbol{\epsilon}_{w_{pp}}(\boldsymbol{s}),\boldsymbol{\epsilon}_{w_{pp}}(\boldsymbol{t})\}
\end{aligned}$$

Letting $c_{w_{mpp};i,j}(\boldsymbol{s},\boldsymbol{t})$ be the (i,j)-th entry of $\boldsymbol{C}_{w_{mpp}}(\boldsymbol{s},\boldsymbol{t})$, and using calculations analogous to (B.2) we find that $\mathrm{E}\left[\boldsymbol{C}_{w_{mpp}}(\boldsymbol{s},\boldsymbol{t})\right]$ is also a diagonal matrix with i-th diagonal element

$$\mathrm{E}\left[c_{w_{mpp};i,i}(\boldsymbol{s},\boldsymbol{t})\right] = \sum_{k=1}^{m} \mathrm{cov}\{a_{ik,pp}(\boldsymbol{s}),a_{ik,pp}(\boldsymbol{t})\}\mathrm{cov}\{v_{k,pp}(\boldsymbol{s}),v_{k,pp}(\boldsymbol{t})\} \tag{B.4}$$

$$+ \mathrm{E}[a_{ii,pp}(\boldsymbol{s})]\mathrm{E}[a_{ii,pp}(\boldsymbol{t})]\mathrm{cov}\{v_{i,pp}(\boldsymbol{s}),v_{i,pp}(\boldsymbol{t})\} + \mathrm{cov}\{\epsilon_{w_{pp};i}(\boldsymbol{s}),\epsilon_{w_{pp};i}(\boldsymbol{t})\}\ .$$

Now, we compare the elements in (B.2) with (3.10). When $\boldsymbol{s} \neq \boldsymbol{t}$, the difference between their i-th diagonal elements is

$$\{\mathrm{E}[\tilde{a}_{ii}(\boldsymbol{s})]\mathrm{E}[\tilde{a}_{ii}(\boldsymbol{t})] - \mathrm{E}[a_{ii,pp}(\boldsymbol{s})]\mathrm{E}[a_{ii,pp}(\boldsymbol{t})]\}\,\mathrm{cov}\{v_{i,pp}(\boldsymbol{s}),v_{i,pp}(\boldsymbol{t})\}.$$

If the elements of $\boldsymbol{A}(\boldsymbol{s})$ are zero-centered processes, then this difference is easily seen to be zero and the two cross-covariance matrices coincide.

On the other hand, as mentioned in Section 3.3.2, $\boldsymbol{A}(\boldsymbol{s})$, and hence $\boldsymbol{A}_{pp}(\boldsymbol{s})$, is often assumed lower-triangular with *positive* diagonal elements to ensure the one-one correspondence with the cross-covariance matrix. Their logarithms are assumed to be

Gaussian processes. In that case, $\log \tilde{a}_{ii}(\boldsymbol{s}) = \log a_{pp;ii}(\boldsymbol{s}) + \tilde{\epsilon}_{ii;a}(\boldsymbol{s})$, where $e^{\tilde{\epsilon}_{ii;a}(\boldsymbol{s})}$ is log-normally distributed. We can now write

$$\begin{aligned} \mathrm{E}[\tilde{a}_{ii}(\boldsymbol{s})]\mathrm{E}[\tilde{a}_{ii}(\boldsymbol{t})] &= \mathrm{E}[e^{\log a_{ii,pp}(\boldsymbol{s})+\tilde{\epsilon}_{ii,a}(\boldsymbol{s})}]\mathrm{E}[e^{\log a_{ii,pp}(\boldsymbol{t})+\tilde{\epsilon}_{ii,a}(\boldsymbol{t})}] \\ &= \mathrm{E}[a_{ii,pp}(\boldsymbol{s})]\mathrm{E}[a_{ii,pp}(\boldsymbol{t})]\mathrm{E}[e^{\tilde{\epsilon}_{ii,a}(\boldsymbol{s})}]\mathrm{E}[e^{\tilde{\epsilon}_{ii,a}(\boldsymbol{t})}] \\ &> \mathrm{E}[a_{ii,pp}(\boldsymbol{s})]\mathrm{E}[a_{ii,pp}(\boldsymbol{t})] \,, \end{aligned} \tag{B.5}$$

where the last inequality follows from the fact that the mean of a zero centered log-normal distribution is always greater than 1. When $\mathrm{cov}\{v_{i,pp}(\boldsymbol{s}), v_{i,pp}(\boldsymbol{t})\} \geq 0$ (a reasonable assumption under spatial settings), the inequality in (B.5) implies that

$$\mathrm{E}\left[c_{\tilde{w};i,i}(\boldsymbol{s},\boldsymbol{t})\right] > \mathrm{E}\left[c_{w_{mpp};i,i}(\boldsymbol{s},\boldsymbol{t})\right]$$

for each i. Since the off-diagonal elements of both cross-covariance matrices are zero, each element of $\mathrm{E}\left[\boldsymbol{C}_{\tilde{w}}(\boldsymbol{s},\boldsymbol{t})\right]$ is at least as large as the corresponding element in $\mathrm{E}\left[\boldsymbol{C}_{\tilde{w}}(\boldsymbol{s},\boldsymbol{t})\right]$.

When $\boldsymbol{s} = \boldsymbol{t}$, we have

$$\begin{aligned} \mathrm{E}\left[c_{\tilde{w};i,i}(\boldsymbol{s},\boldsymbol{s})\right] &= \sum_{k=1}^{m} \mathrm{var}\{\tilde{a}_{ik}(\boldsymbol{s})\}\mathrm{var}\{\tilde{v}_k(\boldsymbol{s})\} + \{\mathrm{E}[\tilde{a}_{ii}(\boldsymbol{s})]\}^2\mathrm{var}\{\tilde{v}_i(\boldsymbol{s})\} \\ &= \sum_{k=1}^{m} \mathrm{var}\{a_{ik}(\boldsymbol{s})\}\mathrm{var}\{v_k(\boldsymbol{s})\} + \{\mathrm{E}[\tilde{a}_{ii}(\boldsymbol{s})]\}^2\mathrm{var}\{v_i(\boldsymbol{s})\} \\ &= \sum_{k=1}^{m} \mathrm{var}\{a_{ik}(\boldsymbol{s})\}\mathrm{var}\{v_k(\boldsymbol{s})\} + \{\mathrm{E}[e^{\log a_{ii,pp}(\boldsymbol{s})+\tilde{\epsilon}_{ii,a}}]\}^2\mathrm{var}\{v_i(\boldsymbol{s})\} \\ &= \sum_{k=1}^{m} \mathrm{var}\{a_{ik}(\boldsymbol{s})\}\mathrm{var}\{v_k(\boldsymbol{s})\} + \{\mathrm{E}[e^{\log a_{ii,pp}(\boldsymbol{s})}]\mathrm{E}[e^{\tilde{\epsilon}_{ii,a}}]\}^2\mathrm{var}\{v_i(\boldsymbol{s})\} \\ &= \sum_{k=1}^{m} \mathrm{var}\{a_{ik}(\boldsymbol{s})\}\mathrm{var}\{v_k(\boldsymbol{s})\} + e^{\mathrm{var}\{\log a_{ii}(\boldsymbol{s})\}}\mathrm{var}\{v_i(\boldsymbol{s})\} \\ &= \sum_{k=1}^{m} \mathrm{var}\{a_{ik}(\boldsymbol{s})\}\mathrm{var}\{v_k(\boldsymbol{s})\} + \{\mathrm{E}[a_{ii}(\boldsymbol{s})]\}^2\mathrm{var}\{v_i(\boldsymbol{s})\} \,, \end{aligned}$$

which is the i-th diagonal element of $\mathrm{E}\left[\boldsymbol{C}_w(\boldsymbol{s},\boldsymbol{s})\right]$. From (3.10) it is clear that $\boldsymbol{C}_{w_{mpp}}(\boldsymbol{s},\boldsymbol{s}) = \boldsymbol{C}_w(\boldsymbol{s},\boldsymbol{s})$. The above equality, therefore, proves that $\mathrm{E}\left[\boldsymbol{C}_{\tilde{w}}(\boldsymbol{s},\boldsymbol{s})\right] = \mathrm{E}\left[\boldsymbol{C}_w(\boldsymbol{s},\boldsymbol{s})\right] = \mathrm{E}\left[\boldsymbol{C}_{w_{mpp}}(\boldsymbol{s},\boldsymbol{s})\right]$.

Summarizing the above results for $\boldsymbol{s} \neq \boldsymbol{t}$ and $\boldsymbol{s} = \boldsymbol{t}$, we have shown that if $\text{cov}\{v_{i,pp}(\boldsymbol{s}), v_{i,pp}(t)\} \geq 0 \, \forall \, \boldsymbol{s}, \boldsymbol{t} \in \mathcal{R}^2$ and $\forall \, i$, then $\text{E}\left[\boldsymbol{C}_{\tilde{w}}(\boldsymbol{s}, \boldsymbol{t})\right] \geq \text{E}\left[\boldsymbol{C}_{w_{mpp}}(\boldsymbol{s}, \boldsymbol{t})\right]$, where the "$\geq$" denotes elementwise inequality.

proof of lemma 3.4.1

Note that, $\sigma_D^2 = \sup_{\boldsymbol{s} \in D}(C_{\eta,i}(\boldsymbol{s}, \boldsymbol{s}) - \boldsymbol{c}_{\eta,i}(\boldsymbol{s}, \mathcal{S}^*)' \boldsymbol{C}_{\eta,i}^{*-1} \boldsymbol{c}_{\eta,i}(\boldsymbol{s}, \mathcal{S}^*))$; Since, D is compact and $C_{\eta,i}(\boldsymbol{s}, \boldsymbol{s}) - \boldsymbol{c}_{\eta,i}(\boldsymbol{s}, \mathcal{S}^*)' \boldsymbol{C}_{\eta,i}^{*-1} \boldsymbol{c}_{\eta,i}(\boldsymbol{s}, \mathcal{S}^*)$ is a continuous function in $\boldsymbol{s}$, it achieves its maximum value at a point $\boldsymbol{s}_{max,i} \in D$. Hence,

$$\sigma_{D,i}^2 = (C_{\eta,i}(\boldsymbol{s}_{max,i}, \boldsymbol{s}_{max,i}) - \boldsymbol{c}_{\eta,i}(\boldsymbol{s}_{max,i}, \mathcal{S}^*)' \boldsymbol{C}_{\eta,i}^{*-1} \boldsymbol{c}_{\eta,i}(\boldsymbol{s}_{max,i}, \mathcal{S}^*)) > 0$$

The RHS in (3.9), Therefore, makes sense. Now, applying Borell-TIS inequality, $\forall \, \epsilon > E||w_i - w_{pp,i}||$,

$$P\left\{\sup_{\boldsymbol{s} \in D} |(w_i - w_{pp,i})(\boldsymbol{s})| > \epsilon\right\} \leq 2P\left\{||w_i - w_{pp,i}|| > \epsilon\right\} \leq 2\exp\left\{-\frac{(\epsilon - E||w_i - w_{pp,i}||)^2}{\sigma_{D,i}^2}\right\}$$

Positivity of the correlation function

We prove the result in our set up. Assume, $\text{var}(w(\boldsymbol{s})) = \sigma^2 \; \forall \, \boldsymbol{s}$. For each $\boldsymbol{s}_0 \in D$, let's consider,

$$f_{\boldsymbol{s}_0}(\boldsymbol{s}) = \boldsymbol{c}_v(\boldsymbol{s}, \mathcal{S}^*)' \boldsymbol{C}_v^{*-1} \boldsymbol{c}_v(\boldsymbol{s}_0, \mathcal{S}^*)$$

$f_{\boldsymbol{s}_0}(\cdot)$ is a continuous function and $f_{\boldsymbol{s}_0}(\boldsymbol{s}_0) > 0$; therefore, given $\epsilon > 0$, $\exists \delta_{\boldsymbol{s}_0}$ s.t.,

$$||\boldsymbol{s} - \boldsymbol{s}_0|| < \delta_{\boldsymbol{s}_0} \Rightarrow |f_{\boldsymbol{s}_0}(\boldsymbol{s}) - f_{\boldsymbol{s}_0}(\boldsymbol{s}_0)| < \epsilon \tag{B.6}$$

let, $\epsilon = \min_{\boldsymbol{s}_i^* \in \mathcal{S}^*} \boldsymbol{c}_v(\boldsymbol{s}_i^*, \mathcal{S}^*)' \boldsymbol{C}_v^{*-1} \boldsymbol{c}_v(\boldsymbol{s}_i^*, \mathcal{S}^*)$, this specific choice of ϵ along with (B.6) gives

$$\boldsymbol{s} \in B(\delta_{\boldsymbol{s}_i^*}) \Rightarrow f_{\boldsymbol{s}_i^*}(\boldsymbol{s}) > 0$$

where $B(\delta_{\boldsymbol{s}_i^*}) = \{\boldsymbol{s} : ||\boldsymbol{s} - \boldsymbol{s}_i^*|| < \delta_{\boldsymbol{s}_i^*}\}$.

Lemma B.0.1 *Assume that,*

1. *knots are tightly packed so that* $\bigcup_{\boldsymbol{s}_i^* \in \mathcal{S}^*} B(\delta_{\boldsymbol{s}_i^*}) \supset D$.

2. *Covariance function in predictive process follows transitivity of sign, i.e.,* $\forall \boldsymbol{l}_1, \boldsymbol{l}_2, \boldsymbol{l}_3 \in D$

$$cov\{w_{pp}(\boldsymbol{l}_1), w_{pp}(\boldsymbol{l}_2)\} > 0, cov\{w_{pp}(\boldsymbol{l}_2), w_{pp}(\boldsymbol{l}_3)\} > 0 \Rightarrow cov\{w_{pp}(\boldsymbol{l}_1), w_{pp}(\boldsymbol{l}_3)\} > 0, \tag{B.7}$$

then, $cov\{w_{pp}(\boldsymbol{s}_1), w_{pp}(\boldsymbol{s}_2)\} > 0 \ \forall \ \boldsymbol{s}_1, \boldsymbol{s}_2 \in D$.

Proof Given any two points $\boldsymbol{s}_1, \boldsymbol{s}_2 \in D$, by Assumption 1, $\exists\, i, j$ s.t. $\boldsymbol{s}_1 \in B(\delta_{\boldsymbol{s}_i^*})$, $\boldsymbol{s}_2 \in B(\delta_{\boldsymbol{s}_j^*})$. Therefore,

$$\text{cov}\left\{w_{pp}(\boldsymbol{s}_i^*), w_{pp}(\boldsymbol{s}_1)\right\} > 0 \ , \ \text{cov}\left\{w_{pp}(\boldsymbol{s}_j^*), w_{pp}(\boldsymbol{s}_2)\right\} > 0 \tag{B.8}$$

Also, $\text{cov}\left\{w_{pp}(\boldsymbol{s}_i^*), w_{pp}(\boldsymbol{s}_j^*)\right\} = \text{cov}\left\{w(\boldsymbol{s}_i^*), w(\boldsymbol{s}_j^*)\right\} > 0$. The proof now follows by application of (B.8) and Assumption 2.

Although, Assumption 2 seems to be quite restrictive, it is found to hold in all practical cases.

Updates for $\tilde{\boldsymbol{a}}_{ij}$'s and $\tilde{\boldsymbol{v}}_k$'s

- Let, $\boldsymbol{\Sigma}_{\tilde{v}}$ be a block diagonal matrix of the order $mn \times mn$ with the k-th block diagonal entry as $\boldsymbol{\Sigma}_{\tilde{v}}(\boldsymbol{\theta}_k)$, $\tilde{\boldsymbol{v}} = (\tilde{\boldsymbol{v}}_1', ..., \tilde{\boldsymbol{v}}_m')'$, then

$$\begin{aligned} \tilde{\boldsymbol{v}} \mid \cdots &\sim N(\boldsymbol{\mu}_{\tilde{\boldsymbol{v}}}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{v}}}) \\ \boldsymbol{\Sigma}_{\tilde{\boldsymbol{v}}} &= \left[\boldsymbol{P}'\mathcal{A}'(\boldsymbol{I}_n \otimes \boldsymbol{\Psi}^{-1})\mathcal{A}\boldsymbol{P} + \boldsymbol{\Sigma}_{\tilde{v}}^{-1}\right]^{-1} \\ \boldsymbol{\mu}_{\tilde{\boldsymbol{v}}} &= \boldsymbol{\Sigma}_{\tilde{\boldsymbol{v}}}\boldsymbol{P}'\mathcal{A}'\left(\boldsymbol{I}_n \otimes \boldsymbol{\Psi}^{-1}\right)(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \end{aligned}$$

  any marginal and conditional distribution can be found out from standard results in multivariate normal.

- Further, assume $\boldsymbol{y}_i = (y_i(\boldsymbol{s}_1), ..., y_i(\boldsymbol{s}_n))'$, $\boldsymbol{X}_i$ is an $n \times p$ matrix with k-th row $\boldsymbol{x}_i(\boldsymbol{s}_k)'$ and $\tilde{\boldsymbol{V}}_j$ is a diagonal matrix with k th diagonal entry as $\tilde{v}_j(\boldsymbol{s}_k)$. Then,

$$\begin{aligned}
\tilde{\boldsymbol{a}}_{ij} \mid \cdots &\sim N(\boldsymbol{\mu}_{\tilde{\boldsymbol{a}}_{ij}}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{a}}_{ij}}), \text{ for } i > j \\
\boldsymbol{\Sigma}_{\tilde{\boldsymbol{a}}_{ij}} &= \left[\frac{\tilde{\boldsymbol{V}}_j' \tilde{\boldsymbol{V}}_j}{\psi_{ii}} + \boldsymbol{\Sigma}_{\tilde{a}}(\boldsymbol{\theta}_{a;i,j})^{-1} \right]^{-1} \\
\boldsymbol{\mu}_{\tilde{\boldsymbol{a}}_{ij}} &= \boldsymbol{\Sigma}_{\tilde{\boldsymbol{a}}_{ij}} \frac{\tilde{\boldsymbol{V}}_j'}{\psi_{ii}} \left(\boldsymbol{y}_i - \boldsymbol{X}_i \boldsymbol{\beta} - \sum_{k \neq j} \tilde{\boldsymbol{V}}_k \tilde{\boldsymbol{a}}_{ik} \right).
\end{aligned}$$

- $\tilde{\boldsymbol{a}}_{ii}$'s are updated using Metropolis steps.

proof of $\boldsymbol{w}_{pp}(\boldsymbol{s}) = \mathcal{A}\boldsymbol{v}_{pp}(\boldsymbol{s})$

Assume, $\boldsymbol{A}(\boldsymbol{s})$ is a non-random space varying function. Basic properties of the multivariate normal distribution yield

$$\begin{aligned}
\boldsymbol{w}_{pp}(\boldsymbol{s}) &= \mathrm{E}\left[\boldsymbol{w}(\boldsymbol{s}) \mid \boldsymbol{w}^*\right] = \mathrm{cov}\left\{\boldsymbol{w}(\boldsymbol{s}), \boldsymbol{w}^*\right\} \mathrm{var}^{-1}\{\boldsymbol{w}^*\}\boldsymbol{w}^* = \mathcal{A}\mathcal{C}_v'\mathcal{A}'\mathcal{A}'^{-1}\boldsymbol{C}_v^{*-1}\mathcal{A}^{-1}\mathcal{A}\boldsymbol{v}^* \\
&= \mathcal{A}\mathcal{C}_v'\boldsymbol{C}_v^{*-1}\boldsymbol{v}^* = \mathcal{A}E\left[\boldsymbol{v}(\boldsymbol{s}) \mid \boldsymbol{v}^*\right] = \mathcal{A}\boldsymbol{v}_{pp}(\boldsymbol{s}),
\end{aligned}$$

which immediately leads to the desired result.

Appendix C

Appendix for chapter 4

Proof of Lemma 4.4.1

Recall, if the block matrix $\boldsymbol{M} = \begin{pmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{B}' & \boldsymbol{D} \end{pmatrix}$ and $\boldsymbol{A}$ are both invertible, then,

$$\boldsymbol{M}^{-1} = \begin{pmatrix} \boldsymbol{A}^{-1} & \boldsymbol{0} \\ \boldsymbol{0}' & \boldsymbol{0} \end{pmatrix} + \begin{pmatrix} \boldsymbol{A}^{-1}\boldsymbol{B} \\ -\boldsymbol{I} \end{pmatrix} \boldsymbol{T}^{-1} \begin{pmatrix} \boldsymbol{B}'\boldsymbol{A}^{-1} & -\boldsymbol{I} \end{pmatrix}, \tag{C.1}$$

where, $\boldsymbol{T} = \boldsymbol{D} - \boldsymbol{B}'\boldsymbol{A}^{-1}\boldsymbol{B}$.

Let, $\mathcal{C}_{v,k}(\boldsymbol{s})$ be an $mk \times m$ matrix whose i-th block is an $m \times m$ matrix $\boldsymbol{C}_v(\boldsymbol{s}, \boldsymbol{s}_i^*)$, $i = 1, .., k$ and $\boldsymbol{C}_{v,k}^*$ be an $mk \times mk$ matrix with $(i,\ j)$ th block as $\boldsymbol{C}_v(\boldsymbol{s}_i^*, \boldsymbol{s}_j^*)$. Then,

$$\operatorname{cov}\{\boldsymbol{w}_{pp,k}(\boldsymbol{s}_1), \boldsymbol{w}_{pp,k}(\boldsymbol{s}_2)\} = \boldsymbol{A}(\boldsymbol{s}_1)\mathcal{C}_{v,k}(\boldsymbol{s}_1)'\boldsymbol{C}_{v,k}^{*-1}\mathcal{C}_{v,k}(\boldsymbol{s}_2)\boldsymbol{A}(\boldsymbol{s}_2)' =$$
$$\boldsymbol{A}(\boldsymbol{s}_1)\left(\mathcal{C}_{v,k-1}(\boldsymbol{s}_1)', \boldsymbol{C}_v(\boldsymbol{s}_1, \boldsymbol{s}_k^*)\right)' \begin{pmatrix} \boldsymbol{C}_{v,k-1}^* & \mathcal{C}_{v,k-1}(\boldsymbol{s}_k^*) \\ \mathcal{C}_{v,k-1}(\boldsymbol{s}_k^*)' & \boldsymbol{C}_v(\boldsymbol{s}_k^*, \boldsymbol{s}_k^*) \end{pmatrix}^{-1} \begin{pmatrix} \mathcal{C}_{v,k-1}(\boldsymbol{s}_2) \\ \boldsymbol{C}_v(\boldsymbol{s}_2, \boldsymbol{s}_k^*)' \end{pmatrix} \boldsymbol{A}(\boldsymbol{s}_2)'$$

Invoking (C.1), we obtain,

$$\operatorname{cov}\{\boldsymbol{w}_{pp,k}(\boldsymbol{s}_1), \boldsymbol{w}_{pp,k}(\boldsymbol{s}_2)\} = \operatorname{cov}\{\boldsymbol{w}_{pp,k-1}(\boldsymbol{s}_1), \boldsymbol{w}_{pp,k-1}(\boldsymbol{s}_2)\} +$$
$$\operatorname{cov}\{\boldsymbol{w}_{res,k-1}(\boldsymbol{s}_1), \boldsymbol{w}_{res,k-1}(\boldsymbol{s}_k^*)\} \operatorname{var}\{\boldsymbol{w}_{res,k-1}(\boldsymbol{s}_k^*)\}^{-1} \operatorname{cov}\{\boldsymbol{w}_{res,k-1}(\boldsymbol{s}_2), \boldsymbol{w}_{res,k-1}(\boldsymbol{s}_k^*)\}'$$

Proof of proposition 4.2.1

Each of the dispersion metrics in Section 4.2.1 can be expressed as

$$\Delta_{(k)} = \|\mathcal{A}[\mathcal{C}_v'\boldsymbol{C}_v^{*-1}\mathcal{C}_v + \boldsymbol{D}_{(k)} - \boldsymbol{C}_v]\mathcal{A}'\|_2 \ , k = 0, 1, 2$$

where $\boldsymbol{D}_{(0)}$ is the zero matrix so $\Delta_{(0)}$ corresponds to the predictive process and $k = 1, 2$ correspond to the modified predictive process and the tapered adjustments respectively.

Let $\boldsymbol{H}_{ij}$ denote the $m \times m$ (i, j)-th block element of $(\boldsymbol{C}_v - \mathcal{C}_v'\boldsymbol{C}_v^{*-1}\mathcal{C}_v)$; $\boldsymbol{H}_{ij}$'s remain invariant to the four models. Let $\boldsymbol{B}_{ij}^{(k)}$ denote the elements of $\boldsymbol{C}_v - \mathcal{C}_v'\boldsymbol{C}_v^{*-1}\mathcal{C}_v - \boldsymbol{D}_{(k)}$, and let $\boldsymbol{T}_{ij}$ be the block elements of $\boldsymbol{T}$ (for the tapered model). $\boldsymbol{H}_{ij}, \boldsymbol{T}_{ij}$ are diagonal. Note Then,

$$\boldsymbol{B}_{ij}^{(1)} = \boldsymbol{H}_{ij}, \quad \text{if} \quad i \neq j \ ; \quad \boldsymbol{B}_{ii}^{(1)} = \boldsymbol{0} \tag{C.2}$$

$$\boldsymbol{B}_{ij}^{(2)} = \boldsymbol{H}_{ij} \odot (\boldsymbol{I} - \boldsymbol{T}_{ij}), \quad \text{if} \quad i \neq j \ , \quad \boldsymbol{B}_{ii}^{(2)} = 0 \tag{C.3}$$

It is quite straightforward to see,

$$\boldsymbol{B}_{ij}^{(2)} \preceq \boldsymbol{B}_{ij}^{(1)} \preceq \boldsymbol{H}_{ij}, \tag{C.4}$$

where, $\boldsymbol{A}_1 \preceq \boldsymbol{A}_2$ means $\boldsymbol{A}_2$ is elementwise greater than or equal to $\boldsymbol{A}_1$.

We will state the following lemma, the proof of which follows from induction method.

Lemma C.0.2 *Let* $\boldsymbol{S} = diag\,(S_1, ..., S_m)$*, then,*

$$\|\boldsymbol{A}\boldsymbol{S}\boldsymbol{A}'\|_2^2 = \sum_{l=1}^{m} S_l^2[\sum_{k=l}^{m} a_{kl}^2]^2 + 2\sum_{l<l'} S_l S_{l'}[\sum_{k=l'}^{m} a_{kl}a_{kl'}]^2$$

Let the l-th diagonal element of $\boldsymbol{B}_{ij}^{(1)}$, $\boldsymbol{B}_{ij}^{(2)}$ and $\boldsymbol{H}_{ij}$ be given by $b_{ijl}^{(1)}$, $b_{ijl}^{(2)}$ and h_{ijl}

respectively. Applying the aforementioned lemma,

$$\begin{aligned}
\Delta_{(2)} &= \left(\sum_{i,j=1}^{m} \left\{ \sum_{l=1}^{m} b_{ijl}^{(2)2} \left[\sum_{k=l}^{m} a_{kl}^2 \right]^2 + 2 \sum_{l<l'} b_{ijl}^{(2)} b_{ijl'}^{(2)} \left[\sum_{k=l'}^{m} a_{kl} a_{kl'} \right]^2 \right\} \right)^{\frac{1}{2}} \\
\Delta_{(1)} &= \left(\sum_{i,j=1}^{m} \left\{ \sum_{l=1}^{m} b_{ijl}^{(1)2} \left[\sum_{k=l}^{m} a_{kl}^2 \right]^2 + 2 \sum_{l<l'} b_{ijl}^{(1)} b_{ijl'}^{(1)} \left[\sum_{k=l'}^{m} a_{kl} a_{kl'} \right]^2 \right\} \right)^{\frac{1}{2}} \\
\Delta_{(0)} &= \left(\sum_{i,j=1}^{m} \left\{ \sum_{l=1}^{m} h_{ijl}^{2} \left[\sum_{k=l}^{m} a_{kl}^2 \right]^2 + 2 \sum_{l<l'} h_{ijl} h_{ijl'} \left[\sum_{k=l'}^{m} a_{kl} a_{kl'} \right]^2 \right\} \right)^{\frac{1}{2}}
\end{aligned}$$

The result, now, follows from (C.4).

Proof of Lemma 4.3.1

Note that, $\mathrm{E}\left[w_{pp}(\boldsymbol{s})\right] = \mathrm{E}\left[w_{mpp}(\boldsymbol{s})\right]$, $\forall\, \boldsymbol{s} \in D$ and, $\forall\, \boldsymbol{s}, \boldsymbol{t} \in D$,

$$\begin{aligned}
\mathrm{E}\left[w_{mpp}(\boldsymbol{s}) - w_{mpp}(\boldsymbol{t})\right]^2 &= \mathrm{var}\left\{w_{mpp}(\boldsymbol{s})\right\} + \mathrm{var}\left\{w_{mpp}(\boldsymbol{t})\right\} - 2\mathrm{cov}\left\{w_{mpp}(\boldsymbol{s}), w_{mpp}(\boldsymbol{t})\right\} \\
&\geq \mathrm{var}\left\{w_{pp}(\boldsymbol{s})\right\} + \mathrm{var}\left\{w_{pp}(\boldsymbol{t})\right\} - 2\mathrm{cov}\left\{w_{pp}(\boldsymbol{s}), w_{pp}(\boldsymbol{t})\right\} \\
&= \mathrm{E}\left[w_{pp}(\boldsymbol{s}) - w_{pp}(\boldsymbol{t})\right]^2
\end{aligned}$$

The result, now, follows with a simple application of Sudakov-Fernique Inequality (see Adler & Taylor, 2007).

$C_\nu(\boldsymbol{s}, \boldsymbol{t})$ is twice differentiable at 0

Since differentiability is a limiting statement and $\nu > 0$, it is enough to prove that $(1 - \frac{\|\boldsymbol{h}\|}{\nu})^4(1 + 4\frac{\|\boldsymbol{h}\|}{\nu})$ is twice differentiable at $\mathbf{0}$. Some routine calculation yields

$$(1 - \frac{\|\boldsymbol{h}\|}{\nu})^4(1 + 4\frac{\|\boldsymbol{h}\|}{\nu}) = 1 - 10\frac{\|\boldsymbol{h}\|^2}{\nu^2} + 20\frac{\|\boldsymbol{h}\|^3}{\nu^3} - 15\frac{\|\boldsymbol{h}\|^4}{\nu^4} + 4\frac{\|\boldsymbol{h}\|^5}{\nu^5}$$

which clearly indicates that the function is twice differentiable a $\mathbf{0}$.

proof of Lemma 4.3.2

Before proving lemma 4.3.2 we will prove the following Proposition,

Proposition 1

If $Z_1(\boldsymbol{s})$ and $Z_2(\boldsymbol{s})$ are two different mean square differentiable Gaussian process then,

1. $Z_1(\boldsymbol{s}) + Z_2(\boldsymbol{s})$ and $Z_1(\boldsymbol{s}) - Z_2(\boldsymbol{s})$ are mean square differentiable.
2. If $Z_1(\boldsymbol{s})$ and $Z_2(\boldsymbol{s})$ are independent processes then $Z_1(\boldsymbol{s})Z_2(\boldsymbol{s})$ is also mean square differentiable.

proof: for any $\boldsymbol{s}$, $Z_1(\cdot)$ and $Z_2(\cdot)$ are differentiable at $\boldsymbol{s}$ means, $\exists$ functions ∇Z_1 and ∇Z_2 respectively, s.t., for any vector $\boldsymbol{u}$ with $||\boldsymbol{u}|| = 1$ we have,

$$\lim_{h\to 0} \mathrm{E}\left[\frac{Z_1(\boldsymbol{s}+h\boldsymbol{u}) - Z_1(\boldsymbol{s})}{h} - \langle \nabla Z_1(\boldsymbol{s}), \boldsymbol{u}\rangle\right]^2 = 0 \qquad \text{(C.5)}$$

$$\lim_{h\to 0} \mathrm{E}\left[\frac{Z_2(\boldsymbol{s}+h\boldsymbol{u}) - Z_2(\boldsymbol{s})}{h} - \langle \nabla Z_2(\boldsymbol{s}), \boldsymbol{u}\rangle\right]^2 = 0 \qquad \text{(C.6)}$$

Let $X_h^{(1)} = \left\{\frac{Z_1(\boldsymbol{s}+h\boldsymbol{u}) - Z_1(\boldsymbol{s})}{h} - \langle \nabla Z_1(\boldsymbol{s}), \boldsymbol{u}\rangle\right\}$, $X_h^{(2)} = \left\{\frac{Z_2(\boldsymbol{s}+h\boldsymbol{u}) - Z_2(\boldsymbol{s})}{h} - \langle \nabla Z_2(\boldsymbol{s}), \boldsymbol{u}\rangle\right\}$ and $X_h^{(3)} = \{Z_1(\boldsymbol{s}+h\boldsymbol{u}) - Z_1(\boldsymbol{s})\}$. (1) follows from the fact,

$$\begin{aligned}
&\lim_{h\to 0} \mathrm{E}\left[\frac{(Z_1 \overset{+}{-} Z_2)(\boldsymbol{s}+h\boldsymbol{u}) - (Z_1 \overset{+}{-} Z_2)(\boldsymbol{s})}{h} - \langle \nabla Z_1(\boldsymbol{s}) \overset{+}{-} \nabla Z_2(\boldsymbol{s}), \boldsymbol{u}\rangle\right]^2 \\
\leq\ &\lim_{h\to 0} 2\left\{\mathrm{E}\left[X_h^{(1)}\right]^2 + E\left[X_h^{(2)}\right]^2\right\} \\
=\ &0
\end{aligned}$$

(2) Now assume $Z_1(\boldsymbol{s})$ and $Z_2(\boldsymbol{s})$ are independent.

$$\begin{aligned}
&\lim_{h\to 0} \mathrm{E}\left[\frac{(Z_1Z_2)(\boldsymbol{s}+h\boldsymbol{u}) - (Z_1Z_2)(\boldsymbol{s})}{h} - \langle Z_2(\boldsymbol{s})\nabla Z_1(\boldsymbol{s}) + Z_1(\boldsymbol{s})\nabla Z_2(\boldsymbol{s}), \boldsymbol{u}\rangle\right]^2 \\
=\ &\lim_{h\to 0} \mathrm{E}\left[Z_1(\boldsymbol{s}+h\boldsymbol{u})X_h^{(2)} + Z_2(\boldsymbol{s})X_h^{(1)} + X_h^{(3)}\langle \nabla Z_2(\boldsymbol{s}), \boldsymbol{u}\rangle\right]^2 \\
=\ &0
\end{aligned}$$

The last step follows by direct application of *Minkowski Inequality* and independence of $\boldsymbol{Z}_1(\boldsymbol{s})$ and $\boldsymbol{Z}_2(\boldsymbol{s})$.

Proof of lemma 4.3.2

- We know, the function $H_j(\boldsymbol{s}) = \frac{\sigma^2}{2^{\theta_2-1}\Gamma(\theta_2)}(||\boldsymbol{s}-\boldsymbol{s}_j^*||\theta_1)^{\theta_2}\kappa_{\theta_2}(||\boldsymbol{s}-\boldsymbol{s}_j^*||;\theta_1)$ is totally differentiable except at $\boldsymbol{s}_j^*$. Therefore $\exists\, \Delta\boldsymbol{H}_j(\boldsymbol{s}) = (\Delta H_{j_1}(\boldsymbol{s}), \Delta H_{j_2}(\boldsymbol{s}))'$ s.t. for any vector $\boldsymbol{u}$ with $||\boldsymbol{u}|| = 1$ we have,

$$H_j(\boldsymbol{s}+h\boldsymbol{u}) = H_j(\boldsymbol{s}) + h\Delta\boldsymbol{H}_j'(\boldsymbol{s})\boldsymbol{u} + o(h),\ \ \forall j = 1,2,...,n^*,\ \ \boldsymbol{s}\in\mathcal{R}^2 - \mathcal{S}^* \quad \text{(C.7)}$$

therefore,

$$\lim_{h\to 0} X_j(\boldsymbol{s},h) = \lim_{h\to 0}\frac{H_j(\boldsymbol{s}+h\boldsymbol{u}) - H_j(\boldsymbol{s})}{h} - \Delta\boldsymbol{H}_j'(\boldsymbol{s})\boldsymbol{u} = 0.\ \boldsymbol{s}\in\mathcal{R}^2 - \mathcal{S}^* \quad \text{(C.8)}$$

Let, $\Delta\boldsymbol{H} = (\Delta\boldsymbol{H}_1(\boldsymbol{s}), \ldots, \Delta\boldsymbol{H}_{n^*}(\boldsymbol{s}))$, then,

$$\begin{aligned}
&\lim_{h\to 0} \quad \mathrm{E}\left[\frac{w_{pp}(\boldsymbol{s}+h\boldsymbol{u}) - w_{pp}(\boldsymbol{s})}{h} - \langle\Delta\boldsymbol{H}\boldsymbol{C}_v^{*-1}\boldsymbol{w}^*, \boldsymbol{u}\rangle\right]^2 \\
&= \quad \lim_{h\to 0}\mathrm{E}\left[\left\{\frac{\boldsymbol{c}_v(\boldsymbol{s}+h\boldsymbol{u})' - \boldsymbol{c}_v(\boldsymbol{s})'}{h} - \boldsymbol{u}'\Delta\boldsymbol{H}\right\}\boldsymbol{C}_v^{*-1}\boldsymbol{w}^*\right]^2 \\
&= \quad \lim_{h\to 0}\{X_1(\boldsymbol{s},h),\ldots,X_{n^*}(\boldsymbol{s},h)\}'\boldsymbol{C}_v^{*-1}\{X_1(\boldsymbol{s},h),\ldots,X_{n^*}(\boldsymbol{s},h)\} \\
&= \quad 0.
\end{aligned}$$

The last equality follows from (C.8) and the fact that $\boldsymbol{C}_v^{*-1}$ doesn't involve h. Therefore, with an appeal to the arbitrariness of $\boldsymbol{s}$ and $\boldsymbol{u}$, we conclude that Predictive Process is L_2 differentiable.

Let, $\Delta\boldsymbol{H}_j(\boldsymbol{s}) = (\Delta\boldsymbol{H}_{j_1}(\boldsymbol{s}), \Delta\boldsymbol{H}_{j_2}(\boldsymbol{s}))'$ with $\Delta\boldsymbol{H}_{j_1}(\boldsymbol{s})$ and $\Delta\boldsymbol{H}_{j_2}(\boldsymbol{s})$ being totally differentiable except at the set of knot points. Continuing in this way, we define,

$$\Delta^l\boldsymbol{H}_{j_{i_1,i_2,\ldots,i_{l-1}}}(\boldsymbol{s}) = (\Delta^l\boldsymbol{H}_{j_{i_1,i_2,\ldots,i_{l-1},1}}(\boldsymbol{s}), \Delta^l\boldsymbol{H}_{j_{i_1,i_2,\ldots,i_{l-1},2}}(\boldsymbol{s}))$$

for any l and any $(i_1,...,i_{l-1}) \in \{1,2\}^{l-1}$

Note that, in order to prove the fact that Predictive process is infinitely differentiable, it is enough to prove that Predictive process is differentiable for any $k_1 \geq 1$. Let, it be already k_1-times mean square differentiable. We will show, it is (k_1+1)-times mean square differentiable.

It is not difficult to see that $\nabla^{k_1}\tilde{w}(\boldsymbol{s}) = \Delta^{k_1}\boldsymbol{H}(\boldsymbol{s})\boldsymbol{C}_v^{*-1}\boldsymbol{w}^*$, where, $\Delta^{k_1}\boldsymbol{H}(\boldsymbol{s})$ is a $2^{k_1} \times n^*$ matrix with a typical row is of the form

$$(\Delta^{k_1}\boldsymbol{H})_{i_1,...,i_{k_1}}(\boldsymbol{s}) = (\Delta^{k_1}\boldsymbol{H}_{1_{i_1,i_2,...,i_{k_1}}}(\boldsymbol{s}), ..., \Delta^{k_1}\boldsymbol{H}_{n^*{}_{i_1,...,i_{k_1}}}(\boldsymbol{s})).$$

It is now enough to show that, $(\Delta^{k_1}\boldsymbol{H})_{i_1,...,i_{k_1}}(\boldsymbol{s})\boldsymbol{C}_v^{*-1}\boldsymbol{w}^*$ is mean square differentiable.

$$\begin{aligned}\lim_{h\to 0} X_{j_{i_1,...,i_{k_1}}}(\boldsymbol{s},h) &= \lim_{h\to 0}\frac{\Delta^{k_1}\boldsymbol{H}_{j_{i_1,i_2,...,i_{k_1}}}(\boldsymbol{s}+h\boldsymbol{u}) - \Delta^{k_1}\boldsymbol{H}_{j_{i_1,i_2,...,i_{k_1}}}(\boldsymbol{s})}{h} \\ &- \Delta^{k_1+1}\boldsymbol{H}_{j_{i_1,i_2,...,i_{k_1}}}(\boldsymbol{s})'\boldsymbol{u} = 0.\ \boldsymbol{s}\in\mathcal{R}^2 - \mathcal{S}^* \qquad \text{(C.9)}\end{aligned}$$

Now,

$$\begin{aligned}&\mathrm{E}[\{\frac{\Delta^{k_1}\boldsymbol{H}_{i_1,i_2,...,i_{k_1}}(\boldsymbol{s}+h\boldsymbol{u}) - \Delta^{k_1}\boldsymbol{H}_{i_1,i_2,...,i_{k_1}}(\boldsymbol{s})}{h} - \Delta^{k_1+1}\boldsymbol{H}_{i_1,i_2,...,i_{k_1}}(\boldsymbol{s})\} \\ &\boldsymbol{C}_v^{*-1}\boldsymbol{w}^*]^2 \\ &= \left\{X_{1_{i_1,...,i_{k_1}}}(\boldsymbol{s},h),\ldots,X_{n^*_{i_1,...,i_{k_1}}}(\boldsymbol{s},h)\right\}'\boldsymbol{C}_v^{*-1} \\ &\quad \left\{X_{1_{i_1,...,i_{k_1}}}(\boldsymbol{s},h),\ldots,X_{n^*{}_{i_1,...,i_{k_1}}}(\boldsymbol{s},h)\right\}\end{aligned}$$

Taking limit as $h \to 0$ on both sides and using (C.9) proves the (k_1+1)-th mean square differentiability of the process $\tilde{w}(\cdot)$ of the process at $\boldsymbol{s}$ in the direction $\boldsymbol{u}$. Since $\boldsymbol{s}$ and $\boldsymbol{u}$ are arbitrarily chosen, Predictive process is (k_1+1)-times mean square differentiable.

- For Modified Predictive Process, a simple calculation will yield

$$\mathrm{E}\left(w_{mpp}(\boldsymbol{s}) - w_{mpp}(\boldsymbol{s}_0)\right)^2 = 2\sigma^2\{1 - \boldsymbol{I}(\boldsymbol{s} = \boldsymbol{s}_0)\}\{1 - \boldsymbol{c}_v(\boldsymbol{s}_1, \mathcal{S}^*)'\boldsymbol{C}_v^{*-1}\boldsymbol{c}_v(\boldsymbol{s}_1, \mathcal{S}^*)\} \tag{C.10}$$

  thus, $\lim_{\boldsymbol{s}\to\boldsymbol{s}_0} \mathrm{E}\left(w_{mpp}(\boldsymbol{s}) - w_{mpp}(\boldsymbol{s}_0)\right)^2$ does not exist. Therefore, Modified Predictive Process is not L_2 continuous.

- Let's denote $\boldsymbol{z}(\boldsymbol{s}, \mathcal{S}^*)' = \boldsymbol{c}_v(\boldsymbol{s}, \mathcal{S}^*)'\boldsymbol{C}_v^{*-1}$. For Tapered Predictive Process with some little algebra it can be shown that,

$$\begin{aligned}\mathrm{E}\left[w_{tap}(\boldsymbol{s}) - w_{tap}(\boldsymbol{s}_0)\right]^2 = 2\sigma^2 - \sigma^2\{&\boldsymbol{z}(\boldsymbol{s}, \mathcal{S}^*)'\boldsymbol{c}_v(\boldsymbol{s}, \mathcal{S}^*)(1 - C_\nu(\boldsymbol{s}, \boldsymbol{s}_0)) \\ &+ C_w(\boldsymbol{s}, \boldsymbol{s}_0)C_\nu(\boldsymbol{s}, \boldsymbol{s}_0)\}\end{aligned} \tag{C.11}$$

  The L_2 continuity follows easily from the equation (C.11).

  It is well known that with matern covariance kernel with $m_1 < \theta_2 < m_1 + 1$, $w(\boldsymbol{s})$ is m_1-times differentiable anywhere. By the assumption, $C_\nu(\cdot)$ is 2k-times differentiable. Thus, Proposition 1 and (1) of Lemma 4.3.2 together yield $w_{tap}(\boldsymbol{s}) = w_{pp}(\boldsymbol{s}) + (w(\boldsymbol{s}) - w_{pp}(\boldsymbol{s}))\eta(\boldsymbol{s})$ is $\min(m_1, k)$-times Mean square differentiable except at the set of knot points.

an alternative proof of the positive definiteness of Hadamard Product

Suppose $\boldsymbol{A}$ and $\boldsymbol{B}$ be two positive definite matrices of order $n \times n$. The Kronecker product of $\boldsymbol{A}$ and $\boldsymbol{B}$, denoted by $\boldsymbol{A} \otimes \boldsymbol{B}$, is defined to be a matrix of order $n^2 \times n^2$ having (i, j)th block as $a_{ij}\boldsymbol{B}$. Thus the Kronecker product of $\boldsymbol{A}$ and $\boldsymbol{B}$ can be represented by $\boldsymbol{A} \otimes \boldsymbol{B} = (a_{ij}\boldsymbol{B})_{i,j=1}^n$. On the other hand the Hadamard product between two matrices are defined to be the elementwise product of $\boldsymbol{A}$ and $\boldsymbol{B}$; $\boldsymbol{A} \odot \boldsymbol{B} = (a_{ij}b_{ij})_{i,j=1}^n$.

Define $\boldsymbol{e}_i$ be a vector of order $n^2 \times 1$ such that,

$$\boldsymbol{e}_{ij} = \begin{cases} 1 & \text{if } j = i \\ 0 & \text{if } j \neq i. \end{cases} \tag{C.12}$$

Define , $\boldsymbol{E} = [\boldsymbol{e}_1 : \boldsymbol{e}_{n+2} : \boldsymbol{e}_{2n+3} : \boldsymbol{e}_{3n+4} : \ldots : \boldsymbol{e}_{n^2}]$

A simple matrix algebra will tell us $\boldsymbol{E}^{'}(\boldsymbol{A} \otimes \boldsymbol{B})\boldsymbol{E} = \boldsymbol{A} \odot \boldsymbol{B}$.

The proof now easily follows from the fact that $\boldsymbol{A} \otimes \boldsymbol{B}$ is positive definite whenever $\boldsymbol{A}$ and $\boldsymbol{B}$ are positive definite.