On Some Computational, Modeling and Design Issues in Bayesian Analysis of Spatial Data

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

Qian Ren

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

Sudipto Banerjee

October, 2012

Acknowledgements

First and foremost I want to thank my thesis advisor, Dr. Sudipto Banerjee, for his warm encouragement and thoughtful guidance. It has been an honor to be his Ph.D. student. The joy and enthusiasm he has for his research was contagious and motivational for me.

Special thanks to my committee, Dr. Snigdhansu Chatterjee, Dr. Lynn E. Eberly, Dr. James Hodges and Dr. Julian Wolfson for their support, guidance and helpful suggestions. Their guidance has served me well and I owe them my heartfelt appreciation.

I would specially like to thank Dr. Lan Wang, who is not available this time, but was a member on my committee since my Plan B presentation. My warm thanks are due to Dr. Andrew Finley for reviewing my paper and giving me valuable advice. I also like to thank Dr. Paul Delamater for being patient and answering my tedious questions and Barb Zweber for kindly reviewing my thesis.

Lastly, I would like to thank my family for all their love and encouragement. For my mother who has supported me in all my pursuits. And most of all for my loving wife, Jia, whose faithful support during the final stages of this Ph.D. is so appreciated.

Dedication

This dissertation is dedicated to my wonderful family. Particularly to my understanding wife, Jia, who has put up with these many years of research, and to our precious daughter Alyssa, who is the joy of our lives. I must also thank my loving mother who have given me her fullest support.

Abstract

My research on Bayesian spatial analysis can be divided into three challenges: computing, methodology (modeling) and experimental design. My first exploration in research is to find an alternative to Markov chain Monte Carlo (MCMC) for the Bayesian hierarchical model. Variational Bayesian (VB) method would be a choice to tackle the massive computational burden for large spatial data analysis. We discuss applying VB to spatial analysis, especially to the multivariate spatial cases. Different VB algorithms are developed and applied to simulated and real examples.

When the number of the locations and the dimension of the outcome variables are large, models with feature of dimension reduction are essential in the real applications. Low-rank spatial processes and factor analysis models are merged together to capture the associations among the variables as well as the strength of spatial correlation for each variable. We also develop stochastic selection of the latent factors by utilizing certain identifiability characterizations for the spatial factor model. A MCMC algorithm is developed for estimation, which also deals with the spatial misalignment problem.

In many of the spatial applications (environmental epidemiology, for instance), parameter estimation is the most important objective in the study. Even with carefully constructed models and computing technique, it is always a challenge to handle the large spatial data set. Bayesian experimental design may help us to get the desired information from a spatial survey study with a sample size that can be analyzed by most available software. The problem of finding the optimum experimental design for the purpose of performing one or more hypothesis tests is considered in the context of spatial analysis. The Bayesian decision theoretic approach is used to arrive at several new optimality criteria for this purpose. Different approaches to achieving this goal are explored, including additive weighted loss and convex approximation. Simulated annealing algorithm (SAA) is applied to real examples to find the optimum design based on our objective function.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Bayesian Spatial Analysis

Analysis of geographically referenced data has generated considerable interest in many scientific disciplines, such as health, environment, geology, agronomy and others; see, for example, the books by Cressie (1993), Møller (2003), Banerjee et al. (2004), and Schabenberger and Gotway (2004) for a variety of methods and applications. Such studies are becoming more and more common, due to the availability of low cost Geographic Information System (GIS) and Global Positioning Systems (GPS), which enable accurate geocoding of locations where scientific data are collected.

Spatial data are widely modeled using spatial processes that assume, for a study region D, a collection of random variables $\{y(\mathbf{s}) : \mathbf{s} \in D\}$ where $\mathbf{s}$ indexes the points in D. This set is viewed as a randomly realized surface over D which, in practice, is only observed at a finite set of locations in $\mathcal{S} = \{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_n\}$. For point-referenced spatial data that are assumed to be normally distributed (perhaps after suitable transformation), we employ a Gaussian spatial process to specify the joint distribution for an arbitrary number and choice of locations in D. In several instances, inference focuses upon three major aspects, to: (i) estimate associations among the variables, (ii) estimate the strength of spatial association for each variable, and (iii) predict the outcomes at arbitrary locations.

Both estimation and prediction require evaluating the Gaussian likelihood, and hence

evaluating an $n \times n$ matrix, irrespective of whether we use the marginalized or unmarginalized approach. While explicit inversion is replaced with faster linear solvers, likelihood evaluation remains expensive for big n. In this case, we focus on the setting where the number of locations yielding observations is too large for fitting desired hierarchical spatial random effects models. Full inference and accurate assessment of uncertainty often requires Markov chain Monte Carlo (MCMC) methods (Banerjee et al., 2004). However, such estimation involves matrix decompositions whose complexity increases as $O(n^3)$ in the number of locations n at every iteration of the MCMC algorithm, rendering them infeasible for large data sets. This problem is further aggravated when we have a vector of random effects at each location (multivariate settings) or when we have spatiotemporal random effects.

Increasingly in spatial data settings there is need for analyzing multivariate measurements obtained at spatial locations. Such data settings arise when several spatially-dependent outcomes are recorded at each spatial location:

- A primary example is data taken at environmental monitoring stations where measurements on levels of several pollutants (e.g., ozone, $PM_{2.5}$, nitric oxide, carbon monoxide, etc.) are typically taken.
- In atmospheric modeling, at a given site we may observe surface temperature, precipitation, and wind speed.
- In a study of ground level effects of nuclear explosives, soil and vegetation contamination in the form of plutonium and americium concentrations at sites have been collected.
- In examining commercial real estate markets, for an individual property at a given location data includes both selling price and total rental income.
- In forestry, investigators seek to produce spatially explicit predictions of multiple forest attributes using a multi-source forest inventory approach.

In each of these settings, we anticipate both dependence between measurements at a particular location, and association between measurements across locations.

Two different aspects are explored in this thesis to analyze the large multivariate spatial data sets. One is computational, in which VB method is applied to univariate

and multivariate spatial data sets. Section 1.2 reviews the literature of VB method and examines the use of variational methods for obtaining lower bounds on the likelihood. Bayesian central limit theorem (BCLT) is also discussed with which the VB results are compared in Chapter 2. The second is model-based and uses an adaptive low-rank spatial factor model developed in Chapter 3 to lower the dimension of both the outcome vector and the number of locations. Latent variable (factor) models are usually used to address the former, while low-rank spatial processes offer a rich and flexible modeling option for dealing with a large number of locations. In Section 1.3 the properties of the predictive process is explained.

Chapter 4 deals with a different issue. In many instances, inference is sought for the regression slope coefficients in the presence of spatial correlation. A carefully designed study would be able to capture statistically significant regression slopes and may help avoid unnecessarily large data sets. Simulated annealing algorithm (SAA) is a method of combinatorial optimization to find the optimal design with the minimum Bayesian risk. All the technical details about this algorithm are presented in Section 1.4.

1.2 Variational Bayesian (VB) Methods and Bayesian Central Limit Theorem

Variational methods are a faster alternative to MCMC that delivers approximate inference for hierarchical spatial models. Variational Bayesian methods, also called ensemble learning, are a family of techniques for approximating intractable integrals arising in Bayesian inference and machine learning. The idea is to transform the Bayesian inference problem from one of high-dimensional integration to one of optimization. Variational methods, which have been used extensively in Bayesian machine learning for several years, provide a lower bound on the marginal likelihood which can be computed efficiently. One can make use of such bounds to derive a variational approximation to the posterior distribution. Variational methods for lower bounding probabilities have been explored by several researchers in the past decade. Hinton and van Camp (1993) proposed an early approach for Bayesian learning of one-hidden-layer neural networks using variational approximations. Neal and Hinton (1998) presented a generalisation of

Expectation maximization (EM) which made use of Jensen's inequality to allow partial E-steps. Jordan et al. (1998) reviewed variational methods in a general context. Variational Bayesian methods have also been widely applied to various models with latent variables (Waterhouse et al., 1995; MacKay, 1997; Bishop, 1999; Attias, 2000; Ghahramani and Beal, 2000). The structural EM algorithm for scoring discrete graphical models (Friedman, 1998) is closely related to variational methods.

Variational methods have their origins in the 18^{th} century with the work of Euler, Lagrange, and others on the calculus of variations (Gelfand and Fomin, 1963). Here, we define a functional as a mapping that takes a function as input instead of a variable and returns the value of the functional as the output. An example would be the entropy $H(p) = -\int p(y) \ln p(y)\, dy$, which takes a probability density function $p(y)$ as the input and returns a quantity value.

Many problems can be expressed in terms of an optimization problem in which the quantity being optimized is a functional. The solution is obtained by exploring all possible functions to find the one that maximizes (or minimizes) the functional. Since it is unusual that a closed form solution can be found, variational methods naturally focus on approximations to the optimal solutions. In the case of applications to probabilistic inference, the mean field approximation is used.

Consider how variational optimization can be applied to the Bayes inference problem. Let $\mathbf{y}$ denote the observed variables and $\boldsymbol{\theta}$ denote the unobserved parameters. We assume a prior distribution $p(\boldsymbol{\theta})$ for parameter $\boldsymbol{\theta}$. Then the marginal likelihood $p(\mathbf{y}) = \int p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\, d\boldsymbol{\theta}$ can be bounded below using any distribution over the parameter $\boldsymbol{\theta}$. To see how, let $q(\boldsymbol{\theta})$ be any probability density function on $\boldsymbol{\theta}$. Then, writing $p(\mathbf{y}) = \dfrac{p(\mathbf{y}, \boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathbf{y})}$, we have $\log p(\mathbf{y}) = \log p(\mathbf{y}, \boldsymbol{\theta}) - \log p(\boldsymbol{\theta}|\mathbf{y}) = \log \dfrac{p(\mathbf{y}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})} + \log \dfrac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathbf{y})}$. Multiplying both sides by $q(\boldsymbol{\theta})$ and integrating with respect to $\boldsymbol{\theta}$, we obtain

$$\begin{aligned}\log p(\mathbf{y}) = \int q(\boldsymbol{\theta}) \log p(\mathbf{y})\, \mathrm{d}\boldsymbol{\theta} &= \int q(\boldsymbol{\theta}) \log \frac{p(\mathbf{y}, \boldsymbol{\theta})}{q(\boldsymbol{\theta})}\, \mathrm{d}\boldsymbol{\theta} + \int q(\boldsymbol{\theta}) \log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathbf{y})}\, \mathrm{d}\boldsymbol{\theta} \\ &= \mathcal{L}(q) + KL(q\|p) \geq \mathcal{L}(q),\end{aligned}$$

where $\mathcal{L}(q)$ is a function of $\mathbf{y}$ and $KL(q\|p)$ is the Kullback-Liebler (KL) distance from $q(\boldsymbol{\theta})$ to $p(\boldsymbol{\theta}|\mathbf{y})$. Since $KL(q||p)$ satisfies Gibb's inequality (Mackay, 2003) it is always nonnegative, hence $\mathcal{L}(q)$ is a lower bound for the log marginal likelihood. Thus, to

find a $q(\boldsymbol{\theta})$ that approximates $p(\boldsymbol{\theta}|\mathbf{y})$ well, we can either maximize $\mathcal{L}(q)$ or minimize $KL(q\|p)$. Let $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m)$ and $\mathcal{Q} = \{q(\boldsymbol{\theta}) : q(\boldsymbol{\theta}) = \prod_{i=1}^{m} q_i(\boldsymbol{\theta}_i)\}$, where each $\boldsymbol{\theta}_i$ can be a scalar or vector. Then $\mathcal{L}(q)$ for $q(\boldsymbol{\theta}) \in \mathcal{Q}$ can be written as:

$$\mathcal{L}(q) = \int \prod_{i=1}^{m} q_i(\boldsymbol{\theta}_i) \log p(\mathbf{y}, \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta} - \int q(\boldsymbol{\theta}) \log q(\boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}.$$

Using variational calculus we can now show that the optimal $q_i^*(\boldsymbol{\theta}_i)$, which maximizes $\mathcal{L}(q)$, is given by $\log q_i^*(\boldsymbol{\theta}_i) = \mathrm{E}_{j \neq i}[\log p(\mathbf{y}, \boldsymbol{\theta})] + \text{constant}$, where $\mathrm{E}_{j \neq i}[\log p(\mathbf{y}, \boldsymbol{\theta})]$ is the expectation of $\log p(\mathbf{y}, \boldsymbol{\theta})$ over $\prod_{j \neq i} q_j(\boldsymbol{\theta}_j)$. Then (for details see Appendix A):

$$q_i^*(\boldsymbol{\theta}_i) = \frac{\exp\left\{\mathrm{E}_{j \neq i}[\log p(\mathbf{y}, \boldsymbol{\theta})]\right\}}{\int \exp\left\{\mathrm{E}_{j \neq i}[\log p(\mathbf{y}, \boldsymbol{\theta})]\right\} \, \mathrm{d}\boldsymbol{\theta}_i}. \tag{1.1}$$

Equation (1.1) represents a set of consistent conditions for the maximum of the lower bound subject to the factorization constraint. However, it does not represent an explicit solution because the right hand side of (1.1) depends on the expectation computed with respect to the other parameters $\boldsymbol{\theta}_j$. So we must initialize the distribution of all the $\boldsymbol{\theta}_j$ and then cycle through them iteratively. Each parameter's distribution is updated in turn with a revised function given by (1.1) and evaluated using the current estimate of the distribution function for all other parameters. Convergence is guaranteed because the bound is convex with respect to each of the factors $q_i(\boldsymbol{\theta}_i)$ (Attias, 2000).

Some of our subsequent comparisons will be made with the Bayesian central limit theorem which is a large-sample approximation for posterior distributions (Carlin and Louis, 1996). Let $f(\mathbf{y} \,|\, \boldsymbol{\theta})$ be the likelihood for the n observations $\mathbf{y} = (y_1, \ldots, y_n)'$, and suppose $p(\boldsymbol{\theta})$ is prior for $\boldsymbol{\theta}$. Although the prior maybe improper, as long as the posterior distribution $p(\boldsymbol{\theta} \,|\, \mathbf{y}) \propto f(\mathbf{y} \,|\, \boldsymbol{\theta}) p(\boldsymbol{\theta})$ is proper and its mode exists, we could have, as $n \to \infty$, $p(\boldsymbol{\theta} \,|\, \mathbf{y}) \sim N\left(\hat{\boldsymbol{\theta}}_m, \mathbf{H}^{-1}(\hat{\boldsymbol{\theta}}_m)\right)$, where $\hat{\boldsymbol{\theta}}_m$ is the posterior mode and the matrix $\mathbf{H} = -\dfrac{\partial^2 \log p(\boldsymbol{\theta} \,|\, \mathbf{y})}{\partial \theta_i \partial \theta_j}$ is the negative of the *Hessian matrix* of $\log p(\boldsymbol{\theta} \,|\, \mathbf{y})$. The estimator of the asymptotic variance is the negative of the inverse Hessian matrix estimated at the posterior mode $\hat{\boldsymbol{\theta}}_m$. To compute estimates of the parameters using the BCLT, we use the R built-in function called nlminb. This function does constrained and unconstrained optimizations using PORT routines, allowing us to estimate the posterior mode numerically. Subsequently, we use the R function hessian, from the package numDeriv to

calculate a numerical approximation to the Hessian matrix of the log posterior function at the estimated posterior mode.

1.3 Predictive Process Model

Geostatistical settings typically assume, at locations $\mathbf{s} \in D \subseteq \Re^2$, an outcome $Y(\mathbf{s})$ along with a $p \times 1$ vector of spatially referenced predictors, $\mathbf{x}(\mathbf{s})$, which are associated through a spatial regression model,

$$Y(\mathbf{s}) = \mathbf{x}(\mathbf{s})' \boldsymbol{\beta} + \sigma w(\mathbf{s}) + \epsilon(\mathbf{s}), \tag{1.2}$$

where σ is an unknown parameter describing the variability of the latent random process $w(\mathbf{s})$. The residual, after adjusting for predictors, comprises a spatial process, $w(\mathbf{s})$, capturing spatial association, and an independent process, $\epsilon(\mathbf{s})$, often called the *nugget*. The $w(\mathbf{s})$ is a spatial random effect, providing local adjustment (with structured dependence) to the mean, sometimes interpreted as capturing the effect of unmeasured or unobserved covariates with spatial pattern, while $\epsilon(\mathbf{s})$ captures measurement error and/or micro-scale variation.

The customary process specification for $w(\mathbf{s})$, at location $\mathbf{s} \in D$, is a Gaussian process with 0 mean, unit variance and correlation function $\rho(\mathbf{s}, \mathbf{t}; \boldsymbol{\phi})$ that depends on additional parameters $\boldsymbol{\phi}$. This process is viewed as a randomly realized surface over D which, in practice, is only observed at a finite set of locations in D. For point-referenced spatial data that are assumed (perhaps after suitable transformation) to be normally distributed, a Gaussian spatial process is employed to specify the joint distribution for an arbitrary number of and arbitrary choice of locations $\mathcal{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$ in D. For the rest of the paper, the random process $w(\mathbf{s})$ may be defined in a different way in different chapters.

Modelling large spatial datasets has received much attention in the recent past. One approach seeks approximations for the spatial process using kernel convolutions, moving averages, low-rank splines or basis functions (e.g., Wikle and Cressie, 1999; Lin et al., 2000; Higdon, 2001; Ver Hoef et al., 2004; Xia and Gelfand, 2006; Kamman and Wand, 2003; Paciorek, 2007; Banerjee et al., 2008). Essentially, these methods replace the random process with an approximation that represents the realizations in

a lower-dimensional subspace. A second approach seeks to approximate the likelihood either by working in the spectral domain of the spatial process and avoiding the matrix computations (Stein, 1999; Fuentes, 2007; Paciorek, 2007) or by forming a product of appropriate conditional distributions to approximate the likelihood (e.g., Vecchia, 1988; Jones and Zhang, 1997; Stein et al., 2004). A concern is the adequacy of the resultant likelihood approximation. Expertise is required to tailor and tune a suitable spectral density estimate or a sequence of conditional distributions and they do not easily adapt to multivariate processes. Also, the spectral density approaches seem best suited to stationary covariance functions on a (near-)regular lattice of a directly observed Gaussian process. Another approach either replaces the process (random field) model by a Markov random field (Cressie, 1999) or approximates the random field model by a Markov random field (Rue and Tjelmeland, 2002; Rue and Held, 2005). Recently Lindgren et al. (2010) derived a method for explicit Markov representations of the Matérn covariance family using a class of stochastic partial differential equations (SPDE). This approach can be extended to Matérn fields on manifolds, non-stationary covariance structures (Paciorek and Schervish, 2006), oscillating covariance functions (Bolin and Lindgren, 2011) and non-separable space-time models. Adapting these approaches to more complex hierarchical spatial models involving multivariate processes (e.g., Wackernagel, 2003; Gelfand et al., 2004) and spatially varying regressions (Gelfand et al., 2003) is potentially problematic.

Now consider a set of fixed "knots" $\mathcal{S}^* = \{\mathbf{s}_1^*, \ldots, \mathbf{s}_{n^*}^*\}$, $n^* \leq n$, which are usually fixed and may, but need not, form a subset of the observed locations $\mathcal{S}$. The Gaussian process defined before implies that $\mathbf{w}^* = (w(\mathbf{s}_1^*), \ldots, w(\mathbf{s}_{n^*}^*))'$ follows $N_{n^*}(\mathbf{0}, \mathbf{D}^*(\boldsymbol{\phi}))$, where $\mathbf{D}^*(\boldsymbol{\phi})$ is the $n^* \times n^*$ covariance matrix whose (i,j)-th element is $\rho(\mathbf{s}_i^*, \mathbf{s}_j^*; \boldsymbol{\phi})$. The spatial interpolation or "kriging" function at a site $\mathbf{s}_0$ is given by $\tilde{w}(\mathbf{s}_0) = \mathrm{E}[w(\mathbf{s}_0) \mid \mathbf{w}^*] = \mathbf{d}(\mathbf{s}_0; \boldsymbol{\phi})' \mathbf{D}^*(\boldsymbol{\phi})^{-1} \mathbf{w}^*$, where $\mathbf{d}(\mathbf{s}_0; \boldsymbol{\phi})$ is an $n^* \times 1$ vector whose i-th element is $\rho(\mathbf{s}_0, \mathbf{s}_i^*; \boldsymbol{\phi})$. This defines the predictive process $\tilde{w}(\mathbf{s}) \sim GP(0, \tilde{\rho}(\cdot; \boldsymbol{\phi}))$ derived from the parent process $w(\mathbf{s})$, where $\tilde{\rho}(\mathbf{s}_i, \mathbf{s}_j; \boldsymbol{\phi}) = \mathbf{d}(\mathbf{s}_i; \boldsymbol{\phi})' \mathbf{D}^*(\boldsymbol{\phi})^{-1} \mathbf{d}(\mathbf{s}_j; \boldsymbol{\phi})$.

The realizations of $\tilde{w}(\mathbf{s})$ are precisely the kriged predictions conditional upon a realization of $w(\mathbf{s})$ over $\mathcal{S}^*$. The process is completely specified given the covariance function of the parent process and $\mathcal{S}^*$. This process is nonstationary regardless of whether $w(\mathbf{s})$ is. Furthermore, the joint distribution associated with the realizations at

any set of locations in D is nonsingular if and only if the set has at most n^* locations. Since $\tilde{w}(\mathbf{s}) = \mathbf{d}(\mathbf{s};\boldsymbol{\phi})'\mathbf{D}^*(\boldsymbol{\phi})^{-1}\mathbf{w}^*$, $\tilde{w}(\mathbf{s})$ is a spatially varying linear transformation of $\mathbf{w}^*$. The dimension reduction is seen immediately. The n random effect $\{w(\mathbf{s}_i), i = 1, \ldots, n\}$ are replaced with only the n^* random effects in $\mathbf{w}^*$; we can work with an n^* dimensional joint distribution involving only $n^* \times n^*$ matrices.

The predictive process's variance is systematically lower than the variance of the parent process $w(\mathbf{s})$ at any location $\mathbf{s}$. This follows immediately since we have

$$\begin{aligned}\text{var}\{\mathbf{w}(\mathbf{s})\} &= \text{E}[\text{var}(w(\mathbf{s}) \,|\, \mathbf{w}^*)] + \text{var}\{\text{E}(w(\mathbf{s}) \,|\, \mathbf{w}^*)\} \\ &= \text{E}[\text{var}(w(\mathbf{s}) \,|\, \mathbf{w}^*)] + \text{var}\{\tilde{w}(\mathbf{s})\} > \text{var}\{\tilde{w}(\mathbf{s})\}.\end{aligned}$$

In practical implementations, this often reveals itself in overestimation of the nugget variance in predictive process versions of models. Indeed, Banerjee et al. (2008) observed that while predictive process models employing a few hundred knots excelled in estimating most parameters in several complex high-dimensional models for datasets involving thousands of data points, reducing this upward bias in the nugget variance was especially problematic.

Recently, Finley et al. (2009) proposed a modified predictive process model to remedy the problem. A *modified predictive process* is defined as $f(\mathbf{s}) \,|\, \tilde{w}(\mathbf{s}) \sim N\left(\tilde{w}(\mathbf{s}), \sigma_f^2(\mathbf{s})\right)$, where $\sigma_f^2(\mathbf{s}) = \text{var}\{w(\mathbf{s})\} - \tilde{\rho}(\mathbf{s}, \mathbf{s}, \boldsymbol{\phi})$, is a process with spatially adaptive variances. Now, it is easy to see that $\text{var}\{w(\mathbf{s})\} = \text{var}\{f(\mathbf{s})\}$, as desired.

1.4 Simulated Annealing Algorithm

The problem faced in searching for optimal sampling designs can be outlined as follows: for a fixed sample size n and the objective function $O(S, \mathbf{X}_n)$, find design $S = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\} \subset D$ and the corresponding matrix of predictors $\mathbf{X}_n$, which minimize O. Clearly, assuming the covariates are known everywhere in D, $\mathbf{X}_n$ is a function of design S. Therefore, the objective function is denoted as $O(S)$.

A sampling design on a discrete design region is usually easier to implement. In the rest of the paper, D is assumed to be a finite set of locations with a fine grid of size N, and S to be a subset of D with sample size n. When considering a fine grid, it amounts to choosing the best combination of n locations among all N points of the grid. When

N is large, the complete enumeration of all combinations is not feasible even for a small sample size n. A SAA can be used to improve an initially chosen design (e.g., starting from a completely random design). SAA was developed from the statistical mechanics of the annealing process (e.g., Kirkpatrick, 1984). Its name derives from an analogy with the annealing process by which a molten metal, on slow cooling, tends to a global energetic minimum.

The algorithm includes the following steps (e.g., van Groenigen and Stein, 1998; Lark, 2002):

1. Choose a design S_0, then compute $O(S_0)$. Set the initial values for the "cooling factor" T_0 and "distance factor" h_{max_0}, respectively.

2. For each point in design S_k, move this point h units in a randomly selected direction, where $h \sim unif[0; h_{max_k}]$. Here h_{max_k} is the distance that each point in S_k can at most move to. If the perturbation takes the point to a location outside the study region, then the point is returned to its original location and a new perturbation is generated at random until the perturbed point falls into the design region. Call this new design S_k^*.

3. Calculate $O(S_k^*)$. Then the transition between S_k and S_k^* is accepted with the following probability:

$$P_T(S_k \to S_k^*) \begin{cases} 1 & \text{if } O(S_k^*) \leq O(S_k) \\ \exp\left\{\frac{O(S_k)-O(S_k^*)}{T}\right\} & \text{if } O(S_k^*) > O(S_k) \end{cases} . \tag{1.3}$$

4. Repeat steps in 2 and 3 until all the points in S_k have been updated.

5. Reduce the "cooling factor" and "distance factor" by $T_{k+1} = \alpha_T T_k$, $h_{max_{k+1}} = \alpha_h h_{max_k}$.

6. Repeat 2-5 enough times to get a good estimate of the optimal design.

The constant T_k in equation (1.3) is of particular importance. Making T_k smaller reduces the probability of a given transition. Compared to a Boltzman distribution, it is seen that T_k is analogous to the temperature of a system. Reducing T_k is therefore a "cooling" step. If the cooling schedule is well-chosen, then the system is expected to

converge to a value $O(S)$ close to the global minimum (Lark, 2002). Cooling the system too fast will cause convergence to a poor local minimum, while cooling the system too slowly will require a lot of computational time.

In SAA, T_0 is set such that 95% or more of the perturbations will be accepted before the first cooling step. A geometric cooling schedule is generally used with α_T between 0.99 and 0.9. h_{max_0} is usually set at half the length of the region to be sampled; and α_h is set such that the final h_{max} is 0.5 unit (Li, 2009). The reason for using smaller and smaller h_{max} is that, when we approach a global optimum, it becomes less and less likely that a large perturbation will improve the objective function.

The main shortcomings of SAA are well known, particularly the need of tedious tuning of the parameters and the uncertainty of finding the global optimum in practical applications (Zhu and Stein, 2005). The theoretical results point out that a logarithmic decrease of the temperature is necessary to ensure asymptotic convergence, a condition that is never met in practice (Romary et al., 2012). So multiple starting points should be explored to increase the chance of finding a global minimum or, more realistically, a better local minimum.

1.5 Overview

Chapter 2 discusses the use of VB methods as an alternative to MCMC to approximate the posterior distributions of complex spatial models. Variational methods, which have been used extensively in Bayesian machine learning for several years, provide a lower bound on the marginal likelihood, which can be computed efficiently. VB algorithms are developed in several models especially emphasizing their use in multivariate spatial analysis. In addition, estimation and model comparisons from VB methods are demonstrated by using simulated data as well as environmental data sets. Inference from MCMC is used to compare with VB results.

Chapter 3 deals with jointly modeling a large number of geographically referenced outcomes observed over a very large number of locations. We seek to capture associations among the variables as well as the strength of spatial association for each variable. In addition, these are reckoned with the common setting where not all the variables

have been observed over all locations, which leads to *spatial misalignment*. The framework also pursues stochastic selection of the latent factors without resorting to complex computational strategies (such as reversible jump algorithms) by using certain identifiability characterizations for the spatial factor model. An MCMC algorithm is developed for estimation that also deals with the spatial misalignment problem. The full posterior distribution of the missing values (along with model parameters) is recovered in a Bayesian predictive framework and various additional modeling and implementation issues are also discussed. The method is illustrated with simulation experiments and an environmental data set.

Chapter 4 is devoted to the problem of designing an optimal experiment for the purpose of testing a single or multiple hypotheses in spatial survey studies. The experimental design includes two goals: one is sample size determination, and the other one is to find the optimal design (a set of locations) on the map assuming a known sample size. The problem of using Bayesian assurance to determine the sample size is discussed. Alternatively, Bayesian risk of a single hypothesis testing for the slope parameters is used as a design criterion in the spatial regression model, and the additive weighted risks for multiple hypothesis testing. SAA is developed to search for an optimal design among all possible designs on a fine grid. Both single and multiple hypothesis testing examples are demonstrated.

Chapter 2

Variational Bayesian Methods for Spatial Data Analysis

2.1 Introduction

Geostatistical settings typically assume, at locations $\mathbf{s} \in D \subseteq \Re^2$, an outcome $Y(\mathbf{s})$ along with a $p \times 1$ vector of spatially referenced predictors, $\mathbf{x}(\mathbf{s})$, which are associated through a spatial regression model,

$$Y(\mathbf{s}) = \mathbf{x}(\mathbf{s})' \boldsymbol{\beta} + w(\mathbf{s}) + \epsilon(\mathbf{s}). \tag{2.1}$$

The residual, after adjusting for predictors, comprises a spatial process, $w(\mathbf{s})$, capturing spatial association, and an independent process, $\epsilon(\mathbf{s})$, often called the *nugget*. The $w(\mathbf{s})$ is spatial random effect, providing local adjustment (with structured dependence) to the mean, sometimes interpreted as capturing the effect of unmeasured or unobserved covariates with spatial pattern, while $\epsilon(\mathbf{s})$ captures measurement error and/or micro-scale variation.

The customary process specification for $w(\mathbf{s})$ is a mean 0 Gaussian Process with covariance function $C(\mathbf{s}_1, \mathbf{s}_2)$, denoted $GP(0, C(\mathbf{s}_1, \mathbf{s}_2))$. In applications, we often specify $C(\mathbf{s}_1, \mathbf{s}_2) = \sigma^2 \rho(\mathbf{s}_1, \mathbf{s}_2; \boldsymbol{\phi})$ where $\rho(\cdot; \boldsymbol{\phi})$ is a correlation function and $\boldsymbol{\phi}$ includes decay and smoothness parameters, yielding a constant process variance. In any event, $\epsilon(\mathbf{s}_i) \overset{iid}{\sim} N(0, \tau^2)$ for any collection of locations $\mathcal{S} = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n\}$. Prior distributions on the remaining parameters complete the hierarchical model. Customarily, $\boldsymbol{\beta}$ is assigned a

multivariate Gaussian prior, i.e. $\boldsymbol{\beta} \sim N(\boldsymbol{\mu}_\beta, \Sigma_\beta)$, while the variance components σ^2 and τ^2 are assigned *Inverse Gamma* (IG) priors. The process correlation parameter(s), $\boldsymbol{\phi}$, are usually assigned informative priors (e.g., uniform over a finite range) based on the underlying spatial domain.

With n locations, say $\mathcal{S} = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n\}$, the process realizations are collected into an $n \times 1$ vector, say $\mathbf{w} = (w(\mathbf{s}_1), \ldots, w(\mathbf{s}_n))'$, which follows a multivariate normal distribution with mean $\mathbf{0}$ and dispersion matrix $\sigma^2\mathbf{R}(\boldsymbol{\phi})$ with $\rho(\mathbf{s}_i, \mathbf{s}_j; \boldsymbol{\phi})$ being the (i, j)-th element of $\mathbf{R}(\boldsymbol{\phi})$. Letting $\mathbf{Y} = (Y(\mathbf{s}_1), ..., Y(\mathbf{s}_n))'$ be the $n \times 1$ vector of observed responses, we obtain a Gaussian likelihood that combines with the hierarchical specification to yield a posterior distribution $p(\boldsymbol{\beta}, \mathbf{w}, \sigma^2, \tau^2, \boldsymbol{\phi} \,|\, \mathbf{Y})$ that is proportional to

$$p(\boldsymbol{\phi}) \times IG(\tau^2 \,|\, a_\tau, b_\tau) \times IG(\sigma^2 \,|\, a_\sigma, b_\sigma) \times N(\boldsymbol{\beta} \,|\, \boldsymbol{\mu}_\beta, \Sigma_\beta) \\ \times N(\mathbf{w} \,|\, \mathbf{0}, \sigma^2\mathbf{R}(\boldsymbol{\phi})) \times \prod_{i=1}^{n} N(Y(\mathbf{s}_i) \,|\, \mathbf{x}(\mathbf{s}_i)'\boldsymbol{\beta} + w(\mathbf{s}_i), \tau^2). \tag{2.2}$$

Estimation of (2.2) customarily proceeds using an MCMC algorithm. Often a marginalized likelihood is used that is obtained by integrating out the spatial effects $\mathbf{w}$. This yields the posterior distribution $p(\boldsymbol{\beta}, \sigma^2, \tau^2, \boldsymbol{\phi} \,|\, \mathbf{Y})$ that is proportional to

$$p(\boldsymbol{\phi}) \times IG(\tau^2 \,|\, a_\tau, b_\tau) \times IG(\sigma^2 \,|\, a_\sigma, b_\sigma) \\ \times N(\boldsymbol{\beta} \,|\, \boldsymbol{\mu}_\beta, \Sigma_\beta) \times N\left(\mathbf{Y} \,|\, \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{R}(\boldsymbol{\phi}) + \tau^2\mathbf{I}_n\right), \tag{2.3}$$

where $\mathbf{X}$ is the matrix of regressors with the i-th row given by $\mathbf{x}(\mathbf{s}_i)'$ and $\mathbf{I}_n$ is the $n \times n$ identity matrix. With Gaussian likelihood we can recover the posterior distribution of the spatial effects $\mathbf{w}$ using the posterior samples of $\{\boldsymbol{\beta}, \sigma^2, \tau^2, \boldsymbol{\phi}\}$. This is achieved via *composition sampling* from the full conditional distribution of $\mathbf{w}$ derived from (2.2); see Banerjee et al. (2004) for details. In fact we can integrate out $\boldsymbol{\beta}$ from (2.3) as well. The new likelihood function follows multivariate normal distribution with mean $\mathbf{X}\boldsymbol{\mu}_\beta$ and variance $\Sigma_\mathbf{Y} + \mathbf{X}\Sigma_\beta\mathbf{X}'$, where $\Sigma_\mathbf{Y} = \sigma^2\mathbf{R}(\boldsymbol{\phi}) + \tau^2\mathbf{I}_n$. If flat prior is assigned to $\boldsymbol{\beta}$, the likelihood reduces to a singular multivariate normal distribution with mean $\mathbf{0}$ and precision matrix $\Sigma_\mathbf{Y}^{-1/2}(\mathbf{I}_n - \mathbf{P}_\mathbf{V})\Sigma_\mathbf{Y}^{-1/2}$, where $\mathbf{V} = \Sigma_\mathbf{Y}^{-1/2}\mathbf{X}$ and $\mathbf{P}_\mathbf{V} = \mathbf{V}(\mathbf{V}'\mathbf{V})^{-1}\mathbf{V}'$. Even though the covariance matrix of $p(\mathbf{Y} \,|\, \sigma^2, \tau^2, \boldsymbol{\phi})$ does not exist, the posterior distribution for $\{\sigma^2, \tau^2, \boldsymbol{\phi}\}$ is proper. Once the posterior samples

from $p(\sigma^2, \tau^2, \boldsymbol{\phi} \mid \mathbf{Y})$, $\{\sigma^{2(l)}, \tau^{2(l)}, \boldsymbol{\phi}^{(l)}\}_{l=1}^{L}$, have been obtained, the posterior samples of $\boldsymbol{\beta}$ are recovered by drawing for each $l = 1, \ldots, L$ from a multivariate normal distribution with mean $(\mathbf{X}'\Sigma_{\mathbf{Y}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\Sigma_{\mathbf{Y}}^{-1}\mathbf{Y}$ and variance $(\mathbf{X}'\Sigma_{\mathbf{Y}}^{-1}\mathbf{X})^{-1}$. It is hard to apply VB method to this likelihood due to the difficulty of finding closed form approximation to any parameter's posterior distribution.

In this chapter, we describe a framework for using variational methods as a faster alternative to MCMC that delivers approximate inference for hierarchical spatial models. Variational Bayesian (VB) methods, also called ensemble learning, are a family of techniques for approximating intractable integrals arising in Bayesian inference and machine learning. The idea is to transform the Bayesian inference problem from one of high-dimensional integration to one of optimization.

We explore VB methods for estimating univariate and multivariate spatial models. In the next section we review the VB framework for Bayesian inference. Section 2.2 describes applying VB to Bayesian linear regression. A closed form expression for each updating step is calculated and the limit of the process is derived. Sections 2.3 and 2.4 focus on univariate and multivariate spatial models. The VB algorithms are derived and importance sampling is used in the algorithms to estimate the expectation of functions of the parameters, which do not have closed forms. Section 2.5 presents some comparative studies with simulated and real data sets for different spatial models. We compare the performance of VB methods to the performance of MCMC and the Bayesian central limit theorem (BCLT). Section 2.6 concludes this chapter with a summary.

2.2 VB for Bayesian linear regression

The procedure of the VB algorithm is best illustrated with a relatively simple example. We apply the VB algorithm to a Bayesian linear regression model with the conjugate Normal-Inverse-Gamma (*NIG*) prior. The posterior distribution is accessible in closed form, which helps us assess the VB method against an analytical benchmark. Letting $\mathbf{Y}$ be an $n \times 1$ vector of outcomes, we write $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\mathbf{X}$ is the $n \times p$ matrix of regressors, $\boldsymbol{\beta}$ is the slope vector of regression coefficients and $\boldsymbol{\epsilon}$ is an $n \times 1$ vector of Gaussian errors, $\boldsymbol{\epsilon} \sim MVN(\mathbf{0}, \sigma^2\mathbf{I}_n)$. For known values of $\boldsymbol{\mu}_\beta$, $\mathbf{V}_\beta$, a, and b, assume a

NIG prior for $\boldsymbol{\beta}$ and σ^2 as follows,

$$p(\boldsymbol{\beta},\sigma^2) = p(\boldsymbol{\beta}|\sigma^2)p(\sigma^2) = N(\boldsymbol{\beta}\,|\,\boldsymbol{\mu}_\beta,\sigma^2\mathbf{V}_\beta)\times IG(\sigma^2\,|\,a,b) = NIG(\boldsymbol{\beta},\sigma^2\,|\,\boldsymbol{\mu}_\beta,\mathbf{V}_\beta,a,b)$$
$$= \frac{b^a}{(2\pi)^{p/2}|\mathbf{V}_\beta|^{1/2}\Gamma(a)}\left(\frac{1}{\sigma^2}\right)^{a+p/2+1}\exp\left\{-\frac{1}{\sigma^2}\left[b+\frac{\boldsymbol{\beta}-\boldsymbol{\mu}_\beta)'\mathbf{V}_\beta^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu}_\beta}{2}\right]\right\}.$$

The likelihood is

$$p(\mathbf{Y}|\boldsymbol{\beta},\sigma^2) = N(\mathbf{Y}\,|\,\mathbf{X}\boldsymbol{\beta},\sigma^2\mathbf{I}_n) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2}\exp\left\{-\frac{1}{2\sigma^2}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})\right\}.$$

Assuming $q(\boldsymbol{\beta},\sigma^2) = q_\beta(\boldsymbol{\beta})q_{\sigma^2}(\sigma^2)$ and the approximate distributions $q_i^{(t)}(\boldsymbol{\theta}_i)$ for all the parameters are known at iteration t, we apply (1.1) to this model to obtain the iterative solutions: for the slope parameter $\boldsymbol{\beta}$, $q^{(t+1)}(\boldsymbol{\beta}) \sim N\left(\boldsymbol{\mu}^*,(\delta^2)^{(t+1)}\mathbf{V}^*\right)$, where $(\delta^2)^{(t+1)} = \left[\int q^{(t)}(\sigma^2)/\sigma^2\,\mathrm{d}\sigma^2\right]^{-1}$, $\mathbf{V}^* = (\mathbf{V}_\beta^{-1}+\mathbf{X}'\mathbf{X})^{-1}$ and $\boldsymbol{\mu}^* = \mathbf{V}^*(\mathbf{V}_\beta^{-1}\boldsymbol{\mu}_\beta+\mathbf{X}'\mathbf{Y})$; for the variance parameter σ^2, $q^{(t+1)}(\sigma^2) \sim IG\left(a^*+\frac{p}{2},\frac{2b^*+p(\delta^2)^{(t+1)}}{2}\right)$ where $a^* = a+\frac{n}{2}$ and $b^* = b+\frac{1}{2}\left(\boldsymbol{\mu}_\beta'\mathbf{V}_\beta^{-1}\boldsymbol{\mu}_\beta+\mathbf{Y}'\mathbf{Y}-\boldsymbol{\mu}^{*\prime}\mathbf{V}^{*-1}\boldsymbol{\mu}^*\right)$. Note that this algorithm only needs the starting value of $(\delta^2)^{(0)} = \mathrm{E}^{(0)}(1/\sigma^2)$. We do not have to calculate the expectation by specifying $q^{(0)}(\sigma^2)$, but give an initial value to $(\delta^2)^{(0)}$ directly.

Thus using the distribution of σ^2 at iteration $t+1$, we find

$$(\delta^2)^{(t+2)} = \left\{\int\frac{q^{(t+1)}(\sigma^2)}{\sigma^2}\,\mathrm{d}\sigma^2\right\}^{-1} = \frac{2b^*+p(\delta^2)^{(t+1)}}{2a^*+p}. \tag{2.4}$$

Defining $\lim_{t\to+\infty}(\delta^2)^{(t)} = \delta^2$ and taking limit on both sides of (2.4), we obtain $\delta^2 = \frac{b^*}{a^*}$. So when $t\to\infty$,

$$\frac{2b^*+p(\delta^2)^{(t+1)}}{2} \to \frac{2b^*+p\delta^2}{2} = \frac{b^*}{a^*}\left(a^*+\frac{p}{2}\right).$$

The approximate posterior distributions are, for $\boldsymbol{\beta}$, a multivariate normal centered at $\boldsymbol{\mu}^*$ with variance $\frac{b^*}{a^*}\mathbf{V}^*$, and for σ^2, an *Inverse Gamma* with parameters $a^*+\frac{p}{2}$ and $\frac{b^*}{a^*}\left(a^*+\frac{p}{2}\right)$. The joint posterior distribution for $\boldsymbol{\beta}$ and σ^2 with the conjugate NIG prior is $NIG(\boldsymbol{\beta},\sigma^2\,|\,\boldsymbol{\mu}^*,\mathbf{V}^*,a^*,b^*)$. The exact marginal posterior distributions are $p(\boldsymbol{\beta}\,|\,\mathbf{Y}) \sim MVSt_{2a^*}\left(\boldsymbol{\mu}^*,\frac{b^*}{a^*}\mathbf{V}^*\right)$ and $p(\sigma^2\,|\,\mathbf{Y}) \sim IG(a^*,b^*)$, where $MVSt$ denotes the *multivariate Student t distribution*:

$$MVSt_\nu = \frac{\Gamma\left(\frac{\nu+p}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)\pi^{p/2}|\nu\Sigma|^{1/2}}\left[1+\frac{(\boldsymbol{\beta}-\boldsymbol{\mu})'\Sigma^{-1}(\boldsymbol{\beta}-\boldsymbol{\mu})}{\nu}\right]^{-\frac{\nu+p}{2}},$$

with $\nu = 2a^*$, $\Sigma = \frac{b^*}{a^*}\mathbf{V}^*$ and $\boldsymbol{\mu} = \boldsymbol{\mu}^*$.

Both the true and the VB estimated marginal posterior distribution for $\boldsymbol{\beta}$ have the same mean $\boldsymbol{\mu}^*$ and scale parameter $\frac{b^*}{a^*}\mathbf{V}^*$. However, the posterior variance of $\boldsymbol{\beta}$ estimated from VB is smaller because the *Student t* distribution has a heavier tail than the normal distribution. It is easier to see this when $p = 1$. The variance of the *Student t* distribution is $\frac{a^*}{a^*-1} > 1$, while it is 1 for a standard normal distribution.

A similar situation arises for the marginal posterior distribution of σ^2. An inverse gamma random variable's mean and mode can be estimated as the ratio of its scale and shape parameter when the latter is large. Here, when the sample size n, and thus $a^* = a + \frac{n}{2}$, are large enough, the approximate posterior mean and mode of σ^2 from VB are the same as the exact true posterior because $\frac{b^*}{a^*}\left(a^* + \frac{p}{2}\right) / \left(a^* + \frac{p}{2}\right) = \frac{b^*}{a^*}$. But the approximate posterior variance of σ^2 from VB is smaller due to a larger shape parameter: $a^* + p/2 > a^*$. For both parameters, when sample size $n \to \infty$, the posterior variance estimates from VB have limits which equal the true values, i.e., $\frac{a^*}{a^*-1} \to 1$ and $\frac{a^*+p/2}{a^*} \to 1$. These results are further explored using a simulated example in Section 2.5.1. For datasets with reasonable sample sizes, the difference between the VB estimate and the true posterior is very small. Thus the VB approach offers a very good approximation for the true posterior distribution in this simple model.

2.3 VB for univariate spatial regression

2.3.1 Spatial regression model treating the spatial random effect as a hidden variable

We assume a univariate dependent variable $Y(\mathbf{s})$ observed at a generic location $\mathbf{s}$ along with a $p \times 1$ vector of spatially referenced regressors $\mathbf{x}(\mathbf{s})$ over a set of locations. The hierarchical model is given in (2.2). Note that $w(\mathbf{s})$ provides local adjustment (with structured dependence) to the mean. Assuming stationarity, the correlation depends on the separation $\rho(\mathbf{s}_i - \mathbf{s}_j; \boldsymbol{\phi})$, while under isotropy it depends only on the distance $d_{ij} = \|\mathbf{s}_i - \mathbf{s}_j\|$ and we write $\rho(d_{ij}; \boldsymbol{\phi})$. Here we choose $\rho(d_{ij}; \phi) = \exp(-\phi d_{ij})$, the exponential correlation function, to demonstrate the VB approach.

Typically we set a prior distribution for the decay parameter ϕ relative to the size of the spatial domain; for instance, by setting the prior mean to a value that implies

a spatial range of approximately a certain fraction of the maximum intersite distance. Here we put a uniform prior on ϕ based on earlier findings (Best et al., 2000; Stein, 1999). For τ^2 and σ^2 the conjugate priors *Inverse Gamma* are chosen. And the prior of $\boldsymbol{\beta}$ is specified as flat, which corresponds to setting $\boldsymbol{\mu}_\beta = \mathbf{0}$ and $\Sigma_\beta \to \infty$ in (2.2).

When applying the VB algorithm to find the estimated posterior distributions, we have to specify $q_i^{(0)}(\boldsymbol{\theta}_i)$, where $\boldsymbol{\theta} = \{\boldsymbol{\beta}, \tau^2, \sigma^2, \phi\}$. In fact, instead of giving explicit distributions $q_i^{(0)}(\boldsymbol{\theta}_i)$, as in the linear regression example we only need the starting values for the expectation of some functions of the parameters and latent variables to initiate the algorithm corresponding to (1.1) as in the linear regression example. The kinds of functions that are needed depends on the statistical models and the order in which the parameters are updated in the algorithm. In the present model, these functions are $1/\tau^2$, $\mathbf{w}$ and $\mathbf{R}(\phi)^{-1}$. After the estimates for the posterior distributions of VB have been updated in the algorithm, the expectation of these functions are recalculated at each iteration. For the univariate spatial model treating $\mathbf{w}$ as hidden, we have Algorithm 1 (in Appendix B) to find the VB estimates for the posterior distributions. In Algorithm 1, $\mathrm{Tr}(\cdot)$ denotes the trace function, and $\boldsymbol{\mu}_{\mathbf{w}}^{(i-1)}$, and $\mathbf{V}_{\mathbf{w}}^{(i-1)}$ are the expectation and covariance matrix for $\mathbf{w} \sim q^{(i-1)}(\mathbf{w})$ respectively.

With the expectation of $\mathbf{R}(\phi)^{-1}$ under $q_\phi^{(t)}(\phi)$, we can update the approximate distributions of parameters $\boldsymbol{\beta}$, τ^2, σ^2 and the latent variable $\mathbf{w}$ using closed form expressions. However the density function (B.1) is not analytically tractable, so importance sampling is proposed to approximate $E^{(t)}\left(\mathbf{R}(\phi)^{-1}\right)$. Denote function (B.1) as $g(\phi)$. Then for a function $f(\phi)$ of ϕ

$$\begin{aligned}
\mathrm{E}(f(\phi)) &= \frac{\int f(\phi) g(\phi)\, d\phi}{\int g(\phi)\, d\phi} = \frac{\int f(\phi) \frac{g(\phi)}{p_I(\phi)} p_I(\phi)\, d\phi}{\int \frac{g(\phi)}{p_I(\phi)} p_I(\phi)\, d\phi} \\
&\approx \frac{\frac{1}{N}\sum_{i=1}^N f(\phi_i) W(\phi_i)}{\frac{1}{N}\sum_{i=1}^N W(\phi_i)} = \sum_{i=1}^N f(\phi_i) W^*(\phi_i), \qquad (2.5)
\end{aligned}$$

where $\phi_i \overset{iid}{\sim} p_I(\phi)$, $W(\phi_i) = g(\phi_i)/p_I(\phi_i)$ and $W^*(\phi_i) = \frac{W(\phi_i)}{\sum_{i=1}^N W(\phi_i)}$. The density $p_I(\phi)$ is called the *importance function*, and is chosen to be a common distribution from which it is easy to draw samples. The *weight function* $W(\phi_i)$ is the ratio of $g(\phi_i)$ and $p_I(\phi_i)$. After normalization for $W(\phi_i)$, $W^*(\phi_i)$ is treated as the weight for the $f(\phi_i)$ in the sum estimating the expectation (Carlin and Louis, 1996). We can specify $p_I(\phi)$ as any

distribution on the support of ϕ, but in general the spatial decay parameter is weakly identifiable, so here we choose $p_I(\phi)$ to be uniform for convenience. Then $W(\phi_i) = g(\phi_i)$ and $W^*(\phi_i) = g(\phi_i)/\sum_{i=1}^{N} g(\phi_i)$. Because the distributions of the parameters other than ϕ only depend on $\mathrm{E}\left(\mathbf{R}(\phi)^{-1}\right)$, using the estimated expectation from importance sampling in (2.5) allows the VB algorithm to proceed toward convergence. A simulated example based on this model is illustrated in Section 2.5.2.

2.3.2 Marginal spatial regression model

The model introduced in the previous section treats $\mathbf{w}$ as a hidden variable, whose distribution is updated with the distribution of the other parameters. In this section, VB is used to deal with the marginal spatial model in (2.3). Different ways of grouping parameters in this model result in different posterior approximations. We tried two ways of grouping parameters. One is updating the approximate joint distribution of σ^2, τ^2 and ϕ, so that $p(\boldsymbol{\theta} \mid \mathbf{Y}) \simeq q(\sigma^2, \tau^2, \phi)q(\boldsymbol{\beta})$ in the algorithm. The other one uses the reparameterization $r = \sigma^2/\tau^2$ and τ^2 instead of σ^2 and τ^2. In this case $p(\boldsymbol{\theta} \mid \mathbf{Y}) \simeq q(r, \phi)q(\tau^2)q(\boldsymbol{\beta})$. Details of the second method are shown in this section.

The likelihood of the marginal model in (2.3) is $MVN(\mathbf{Y}|\mathbf{X}\boldsymbol{\beta}, \tau^2\mathbf{C})$, where $\mathbf{C} = \mathbf{I}_n + r\mathbf{R}(\phi)$. Then, $\mathbf{C}$ is a function of ϕ and r. The prior distributions of the parameters are the same as in Section 2.5.2 for $\boldsymbol{\beta}$, τ^2 and ϕ, while a uniform prior is assigned to r. To initiate the VB algorithm for the marginal spatial model we need to give starting values for the expectation of $1/\tau^2$ and $\mathbf{C}(r, \phi)^{-1}$. Applying (1.1) to the spatial model (2.3), we have Algorithm 2 (see Appendix B for details) to find the VB estimates for the posterior distributions.

Now we have closed form expressions for the approximate distributions of parameters $\boldsymbol{\beta}$ and τ^2 given a value of $\mathrm{E}(\mathbf{C}^{-1})$. Importance sampling is again used to calculate the expectation of $\mathbf{C}^{-1}$ which then allows the VB algorithm to complete the iteration. After the VB algorithm converges, approximate marginal posterior distributions of ϕ and r can be obtained by using univariate numerical integration (quadrature).

2.4 Multivariate spatial model

2.4.1 Regression model and VB algorithm

Here we extend the univariate case discussed in Section 2.3 to the multivariate spatial regression model. In this setting each site $\mathbf{s}$ offers an $m \times 1$ response vector, $\mathbf{Y}(\mathbf{s}) = (Y_1(\mathbf{s}), \ldots, Y_m(\mathbf{s}))'$, along with a $p \times m$ spatially referenced predictor matrix $\mathbf{X}(\mathbf{s})$. Further, $\mathbf{w}(\mathbf{s}) = (w_1(\mathbf{s}), \ldots, w_m(\mathbf{s}))'$ is an $m \times 1$ zero-centered *Multivariate Gaussian Process*, denoted by $\mathbf{w}(\mathbf{s}) \sim MVGP(\mathbf{0}, \mathbf{K}(\cdot, \cdot; \boldsymbol{\phi}))$ capturing spatial variation. The multivariate Gaussian process is completely specified by an $m \times m$ *cross-covariance* matrix function $\mathbf{K}(\mathbf{s}, \mathbf{s}^*; \boldsymbol{\phi}) = \{\text{cov}(w_i(\mathbf{s}), w_j(\mathbf{s}^*))\}_{i,j=1}^m$ whose (i, j)-th element is the covariance between $w_i(\mathbf{s})$ and $w_j(\mathbf{s}^*)$, with $\boldsymbol{\phi}$ being parameters that control the correlation decay and smoothness of the process. So in total, the $mn \times 1$ vector $\mathbf{w} = (\mathbf{w}(\mathbf{s}_1)', \ldots, \mathbf{w}(\mathbf{s}_n)')'$ is distributed as a multivariate normal distribution $\mathbf{w} \sim MVN(0, \Sigma_{\mathbf{w}}(\boldsymbol{\phi}))$. Here $\Sigma_{\mathbf{w}}(\boldsymbol{\phi}) = [\mathbf{K}(\mathbf{s}_i, \mathbf{s}_j; \boldsymbol{\phi})]_{i,j=1}^n$ is the $mn \times mn$ matrix with $\mathbf{K}(\mathbf{s}_i, \mathbf{s}_j; \boldsymbol{\phi})$ forming the (i, j)-th $m \times m$ block. The Gaussian likelihood combines with the hierarchical specification to yield a posterior distribution $p(\boldsymbol{\beta}, \mathbf{w}, \boldsymbol{\Psi}, \boldsymbol{\phi} \,|\, \mathbf{Y})$ that is proportional to:

$$p(\boldsymbol{\phi}) \times MVN(\boldsymbol{\beta} \,|\, \boldsymbol{\mu}_\beta, \Sigma_\beta) \times \prod_{i=1}^{m} IG(\Psi_i \,|\, a_i, b_i)$$
$$\times MVN(\mathbf{w} \,|\, \mathbf{0}, \Sigma_{\mathbf{w}}(\boldsymbol{\phi})) \times \prod_{i=1}^{n} MVN(\mathbf{Y}(\mathbf{s}_i) \,|\, \mathbf{X}(\mathbf{s}_i)'\boldsymbol{\beta} + \mathbf{w}(\mathbf{s}_i), \boldsymbol{\Psi}), \qquad (2.6)$$

where $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients, which is customarily assigned a multivariate Gaussian prior, $\boldsymbol{\beta} \sim N(\boldsymbol{\mu}_\beta, \Sigma_\beta)$, and $\boldsymbol{\Psi}$ is an $m \times m$ covariance matrix assumed to be diagonal with diagonal elements Ψ_i, $i = 1, \ldots, m$, which are assigned $IG(a_i, b_i)$ priors.

We need to carefully choose $\mathbf{K}(\mathbf{s}, \mathbf{s}^*; \boldsymbol{\phi})$ so that $\Sigma_{\mathbf{w}}(\boldsymbol{\phi})$ is symmetric and positive definite. Modeling $\mathbf{K}(\mathbf{s}, \mathbf{s}^*; \boldsymbol{\phi})$ is indeed more demanding than choosing real-valued covariance functions in univariate spatial modeling that are characterized by Bochner's Theorem (Cressie, 1999). In the multivariate setting, we require that, for an arbitrary number and choice of locations, the resulting $\Sigma_{\mathbf{w}}(\boldsymbol{\phi})$ be symmetric and positive definite. Note that the cross-covariance matrix function need not be symmetric or positive definite but must satisfy $\mathbf{K}(\mathbf{s}, \mathbf{s}^*; \boldsymbol{\phi}) = \mathbf{K}'(\mathbf{s}^*, \mathbf{s}; \boldsymbol{\phi})$ so that $\Sigma_{\mathbf{w}}(\boldsymbol{\phi})$ is symmetric. In the limiting

sense, as $\mathbf{s}^* \to \mathbf{s}$, $\mathbf{K}(\mathbf{s},\mathbf{s}^*;\boldsymbol{\phi}) = [\text{cov}(w_i(\mathbf{s}), w_j(\mathbf{s}))]_{i,j=1}^m$ becomes the symmetric and positive definite variance-covariance matrix of $\mathbf{w}(\mathbf{s})$ within site $\mathbf{s}$. A theorem by Cramér (see e.g., Chilés and Delfiner (1999)) characterizes cross-covariance functions, akin to Bochner's theorem for univariate covariance functions, but using Cramér's result in practical modeling is not trivial.

To develop a computationally feasible and sufficiently rich multivariate spatial model, we adopt a constructive approach through *coregionalization* models (Wackernagel, 2003). Let $\tilde{\mathbf{w}}(\mathbf{s}) = (\tilde{w}_1(\mathbf{s}), \ldots, \tilde{w}_m(\mathbf{s}))'$ be an $m \times 1$ process with m independent zero-centered spatial processes with unit variance, that is, each $\tilde{w}_i(\mathbf{s}) \sim GP(0, \rho(\cdot,\cdot))$ with $\text{var}(\tilde{w}_i(\mathbf{s})) = 1$, $\text{cov}(\tilde{w}_i(\mathbf{s}), \tilde{w}_i(\mathbf{s}^*)) = \rho_i(\mathbf{s},\mathbf{s}^*,\boldsymbol{\phi}_i)$ and $\text{cov}(\tilde{w}_i(\mathbf{s}), \tilde{w}_j(\mathbf{s}^*)) = 0$ whenever $i \neq j$ (irrespective of how close $\mathbf{s}$ and $\mathbf{s}^*$ are), where $\rho_i(\cdot;\boldsymbol{\phi}_i)$ is a correlation function associated with $\tilde{w}_i(\mathbf{s})$, and $\boldsymbol{\phi}_i$ are spatial parameters. This yields a diagonal cross-covariance matrix function $\tilde{\mathbf{K}}(\mathbf{s},\mathbf{s}^*;\boldsymbol{\phi}) = \text{diag}[\rho_i(\mathbf{s},\mathbf{s}^*;\boldsymbol{\phi}_i)]_{i=1}^m$ with $\boldsymbol{\phi} = \{\boldsymbol{\phi}_i\}_{i=1}^m$. It is easy to verify that $\tilde{\mathbf{K}}(\mathbf{s},\mathbf{s}^*;\boldsymbol{\phi})$ is a valid cross-covariance matrix.

To build rich covariance structures, we assume the process $\mathbf{w}(\mathbf{s}) = \mathbf{A}(\mathbf{s})\tilde{\mathbf{w}}(\mathbf{s})$ to be a linear transformation of $\tilde{\mathbf{w}}(\mathbf{s})$, where $\mathbf{A}(\mathbf{s})$ is a space-varying transfer matrix that is nonsingular for all $\mathbf{s}$. Then the cross-covariance matrix function of $\mathbf{w}(\mathbf{s})$ is $\mathbf{K}(\mathbf{s},\mathbf{s}^*;\boldsymbol{\phi}) = \mathbf{A}(\mathbf{s})\tilde{\mathbf{K}}(\mathbf{s},\mathbf{s}^*;\boldsymbol{\phi})\mathbf{A}(\mathbf{s}^*)'$. In fact, $\tilde{\mathbf{K}}(\mathbf{s},\mathbf{s};\boldsymbol{\phi}) = \mathbf{I}_m$, so that $\mathbf{K}(\mathbf{s},\mathbf{s};\boldsymbol{\phi}) = \mathbf{A}(\mathbf{s})\mathbf{A}(\mathbf{s})'$. Therefore $\mathbf{A}(\mathbf{s}) = \mathbf{K}^{1/2}(\mathbf{s},\mathbf{s})$ is identified as a Cholesky square-root of $\mathbf{K}(\mathbf{s},\mathbf{s})$ and can be taken to be lower-triangular without loss of generality. Since $\tilde{\mathbf{K}}(\mathbf{s},\mathbf{s}^*;\boldsymbol{\phi})$ is a valid cross-covariance matrix, so is $\mathbf{K}(\mathbf{s},\mathbf{s}^*;\boldsymbol{\phi})$. The covariance matrix of $\mathbf{w}(\mathbf{s},\mathbf{s}^*;\boldsymbol{\phi})$, $\Sigma_{\mathbf{w}} = [\mathbf{K}(\mathbf{s}_i,\mathbf{s}_j;\boldsymbol{\phi})]_{i,j=1}^n = [\mathbf{A}(\mathbf{s}_i)\tilde{\mathbf{K}}(\mathbf{s}_i,\mathbf{s}_j;\boldsymbol{\phi})\mathbf{A}(\mathbf{s}_j)']_{i,j=1}^n$ is

$$[\oplus_{i=1}^n \mathbf{A}(\mathbf{s}_i)][\oplus_{k=1}^m \rho_k(\mathbf{s}_i,\mathbf{s}_j;\boldsymbol{\phi}_k)]_{i,j=1}^n[\oplus_{i=1}^n \mathbf{A}(\mathbf{s}_i)'] = \mathscr{A}\Sigma_{\tilde{\mathbf{w}}}\mathscr{A}' \,,$$

where $\oplus$ is the "diagonal" or direct-sum matrix operator. Thus, $\oplus_{k=1}^m \rho_k(\mathbf{s}_i,\mathbf{s}_j;\boldsymbol{\phi}_k)$ is an $m \times m$ diagonal matrix with $\rho_k(\mathbf{s}_i,\mathbf{s}_j;\boldsymbol{\phi}_k)$ as its diagonals, while $\mathscr{A}$ is a block-diagonal matrix with the i-th diagonal block being $\mathbf{A}(\mathbf{s}_i)$. Since $\tilde{\mathbf{K}}(\mathbf{s}_i,\mathbf{s}_j;\boldsymbol{\phi})$ is a valid cross-covariance, $\Sigma_{\tilde{\mathbf{w}}}$ is positive-definite and so is $\Sigma_{\mathbf{w}}$.

Stationary cross-covariance functions necessarily imply $\mathbf{A}(\mathbf{s})$ is independent of space. Here, since the cross-covariance is a function of the separation between sites, we have $\mathbf{K}(\mathbf{s},\mathbf{s};\boldsymbol{\phi}) = \mathbf{K}(\mathbf{0};\boldsymbol{\phi})$, so that $\mathbf{A}(\mathbf{s}) = \mathbf{A} = \mathbf{K}^{1/2}(\mathbf{0};\boldsymbol{\phi})$. In such case, $\mathscr{A} = \mathbf{I}_n \otimes \mathbf{A}$ and $\Sigma_{\mathbf{w}} = (\mathbf{I}_n \otimes \mathbf{A})\Sigma_{\tilde{\mathbf{w}}}(\mathbf{I}_n \otimes \mathbf{A}')$. Denote the observed $nm \times 1$ outcome vector by

$\mathbf{Y} = (\mathbf{Y}(\mathbf{s}_1)', \ldots, \mathbf{Y}(\mathbf{s}_n)')'$, the $nm \times 1$ spatial random effect as $\tilde{\mathbf{w}} = (\tilde{\mathbf{w}}(\mathbf{s}_1)', \ldots, \tilde{\mathbf{w}}(\mathbf{s}_n)')'$ and the $nm \times p$ matrix of regressors as $\mathbf{X} = [\mathbf{X}(\mathbf{s}_i)']_{i=1}^n$. We can then cast the data model and priors, i.e. the posterior into the following generic template:

$$p(\boldsymbol{\phi}) \times p(\mathbf{A}) \times MVN(\boldsymbol{\beta} \,|\, \boldsymbol{\mu}_\beta, \Sigma_\beta) \times \prod_{i=1}^{m} IG(\Psi_i \,|\, a_i, b_i)$$
$$\times MVN(\tilde{\mathbf{w}} \,|\, \mathbf{0}, \Sigma_{\tilde{\mathbf{w}}}(\boldsymbol{\phi})) \times \prod_{i=1}^{n} MVN(\mathbf{Y}(\mathbf{s}_i) \,|\, \mathbf{X}(\mathbf{s}_i)'\boldsymbol{\beta} + \mathbf{A}\tilde{\mathbf{w}}(\mathbf{s}_i), \boldsymbol{\Psi}). \quad (2.7)$$

Customarily, we let $\boldsymbol{\beta}$ have a flat prior, which corresponds to setting $\boldsymbol{\mu}_\beta = \mathbf{0}$ and letting $\Sigma_\beta \to \infty$. The measurement error dispersion covariance matrix $\boldsymbol{\Psi}$ could be assigned an inverse-Wishart prior, although one usually assumes independence of measurement errors for different response measurements in each site and thus define $\boldsymbol{\Psi}^{-1}$ to be the diagonal matrix with $\delta_i^2 = 1/\Psi_i$ the i-th diagonal element. Each δ_i^2 is given the Gamma prior, $G(a_i, b_i)$. The specific form of $\mathscr{A}$ will depend upon the exact form of $\mathbf{A}$, which is the square root of $K(\mathbf{0})$. Let $\mathbf{A}$ be a lower triangular matrix and assign a inverse-Wishart$(df, \mathbf{S})$ prior to $\mathbf{A}\mathbf{A}'$. Finally, recall $\Sigma_{\tilde{\mathbf{w}}} = [\tilde{\mathbf{K}}(\mathbf{s}_i - \mathbf{s}_j; \boldsymbol{\phi})]_{i,j=1}^n, \boldsymbol{\phi} = \{\boldsymbol{\phi}_k\}_{k=1}^m$, which is a function of $\boldsymbol{\phi}$; one needs to assign prior to $\boldsymbol{\phi}$. Here we assume the exponential correlation function $\rho_k(\mathbf{s}, \mathbf{s}^*; \boldsymbol{\phi}_k) = \exp(-\phi_k \|\mathbf{s} - \mathbf{s}^*\|)$, in which ϕ_k is the spatial decay parameter and receives a uniform prior distribution. The effective range (i.e., the distance at which the correlation drops to 0.05) is determined by $-\log(0.05)/\phi$. We want to set the support of ϕ to allow for a reasonable effective range estimate. Applying (1.1) to the multivariate spatial model, we need to set starting values for the expectations of $\boldsymbol{\Psi}^{-1}$, $\mathbf{A}$, $\boldsymbol{\phi}$, $\tilde{\mathbf{w}}$, $\Sigma_{\tilde{\mathbf{w}}}(\boldsymbol{\phi})^{-1}$ and $\mathbf{A}'\boldsymbol{\Psi}^{-1}\mathbf{A}$. The derivations can be found in Appendix B; and the details for a typical iteration is shown in Algorithm 3.

The approximate posterior densities for $\boldsymbol{\beta}$, $\tilde{\mathbf{w}}$ and the δ_j^2 are common distributions, and only the hyper-parameters of these distribution functions need to be updated. Importance sampling is used to calculate the expectation of the function of $\boldsymbol{\phi}$ and $\mathbf{A}$ to complete the iteration of the VB algorithm. After the algorithm converges the approximate posterior distribution for $\boldsymbol{\phi}$ and $\mathbf{A}$ can be estimated using numeric integration.

2.4.2 Posterior predictive inference

Often analysts wish to produce surface plots of $\tilde{\mathbf{w}}$ to assess model fit or identify missing regressors. If we integrate over $\tilde{\mathbf{w}}$ to get the marginal models it only can be recovered

in a posterior predictive fashion,

$$p(\tilde{\mathbf{w}}|\mathbf{Y}) = \int p(\tilde{\mathbf{w}}|\boldsymbol{\theta}, \mathbf{Y}) p(\boldsymbol{\theta}|\mathbf{Y})\, d\boldsymbol{\theta}\ , \tag{2.8}$$

where $\boldsymbol{\theta}$ denotes all the parameters of the marginal model and $p(\tilde{\mathbf{w}}|\boldsymbol{\theta}, \mathbf{Y})$ follows multivariate normal distribution with mean $\left[\Sigma_{\tilde{\mathbf{w}}}^{-1} + \mathscr{A}'\left(\mathbf{I}_n \otimes \boldsymbol{\Psi}^{-1}\right)\mathscr{A}\right]^{-1} \mathscr{A}'(\mathbf{I}_n \otimes \boldsymbol{\Psi}^{-1})(\mathbf{Y} - \mathbf{X}\beta)$ and variance $\left[\Sigma_{\tilde{\mathbf{w}}}^{-1} + \mathscr{A}'(\mathbf{I}_n \otimes \boldsymbol{\Psi}^{-1})\mathscr{A}\right]^{-1}$. Subsequently, the posterior estimates of these realizations can be mapped with contours to produce image and contour plots of the spatial processes.

Let $\{\mathbf{s}_{0l}\}_{l=1}^{n^*}$ be a collection of n^* locations where we seek to predict the spatial random effect $\tilde{\mathbf{w}}^*$. In particular we want to compute the posterior mean $\mathrm{E}(\tilde{\mathbf{w}}^*|\mathbf{Y})$ where $\tilde{\mathbf{w}}^* = (\tilde{\mathbf{w}}(\mathbf{s}_{01})', \ldots, \tilde{\mathbf{w}}(\mathbf{s}_{0n^*})')'$. Note that

$$p(\tilde{\mathbf{w}}^* \,|\, \mathbf{Y}) = \int p(\tilde{\mathbf{w}}^* \,|\, \tilde{\mathbf{w}}, \boldsymbol{\theta}, \mathbf{Y}) p(\tilde{\mathbf{w}}, \boldsymbol{\theta} \,|\, \mathbf{Y})\, d\boldsymbol{\theta}\, d\tilde{\mathbf{w}}\ .$$

The joint distribution of $\tilde{\mathbf{w}}^*$ and $\tilde{\mathbf{w}}$ is a multivariate normal distribution, namely:

$$\begin{pmatrix} \tilde{\mathbf{w}} \\ \tilde{\mathbf{w}}^* \end{pmatrix} \sim MVN\left(\begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix}, \begin{pmatrix} \Sigma_{\tilde{\mathbf{w}}} & \Sigma_{\tilde{\mathbf{w}}, \tilde{\mathbf{w}}^*} \\ \Sigma_{\tilde{\mathbf{w}}^*, \tilde{\mathbf{w}}} & \Sigma_{\tilde{\mathbf{w}}^*} \end{pmatrix}\right)\ ,$$

where $\Sigma_{\tilde{\mathbf{w}}} = [\oplus_{k=1}^m \rho_k(\mathbf{s}_i, \mathbf{s}_j; \boldsymbol{\phi}_k)]_{i,j=1}^n$, $\Sigma_{\tilde{\mathbf{w}}}^* = [\oplus_{k=1}^m \rho_k(\mathbf{s}_{0i}, \mathbf{s}_{0j}; \boldsymbol{\phi}_k)]_{i,j=1}^{n^*}$ and $\Sigma'_{\tilde{\mathbf{w}}, \tilde{\mathbf{w}}^*} = [\oplus_{k=1}^m \rho_k(\mathbf{s}_{0i}, \mathbf{s}_j; \boldsymbol{\phi}_k)]_{i=1,j=1}^{n^*,n}$. Therefore, $p(\tilde{\mathbf{w}}^* \,|\, \tilde{\mathbf{w}}, \boldsymbol{\theta}, \mathbf{Y})$ is $MVN(\mu_{\tilde{\mathbf{w}}^* \,|\, \tilde{\mathbf{w}}}, \Sigma_{\tilde{\mathbf{w}}^* \,|\, \tilde{\mathbf{w}}})$, where

$$\mu_{\tilde{\mathbf{w}}^* \,|\, \tilde{\mathbf{w}}} = \Sigma'_{\tilde{\mathbf{w}}, \tilde{\mathbf{w}}^*} \Sigma_{\tilde{\mathbf{w}}}^{-1} \tilde{\mathbf{w}} \text{ and } \Sigma_{\tilde{\mathbf{w}}^*|\tilde{\mathbf{w}}} = \Sigma_{\tilde{\mathbf{w}}^*} - \Sigma'_{\tilde{\mathbf{w}}, \tilde{\mathbf{w}}^*} \Sigma_{\tilde{\mathbf{w}}}^{-1} \Sigma_{\tilde{\mathbf{w}}, \tilde{\mathbf{w}}^*}.$$

In the hidden variable model the posterior distribution of $\tilde{\mathbf{w}}$ can be estimated by $q(\tilde{\mathbf{w}})$. So $q(\tilde{\mathbf{w}})q(\boldsymbol{\theta})$ can be used as the approximation of $p(\tilde{\mathbf{w}}, \boldsymbol{\theta}|\mathbf{Y})$. Notice that the conditional distribution of $\tilde{\mathbf{w}}^*$ given $\tilde{\mathbf{w}}$ only depends on spatial parameters $\boldsymbol{\phi}$. Then the conditional expectation of $\tilde{\mathbf{w}}^*$ given $\mathbf{Y}$ is

$$\begin{aligned}
\mathrm{E}(\tilde{\mathbf{w}}^*|\mathbf{Y}) &= \int \tilde{\mathbf{w}}^* p(\tilde{\mathbf{w}}^*|\tilde{\mathbf{w}}, \boldsymbol{\phi}) p(\tilde{\mathbf{w}}, \boldsymbol{\theta}|\mathbf{Y})\, d\boldsymbol{\theta}\, d\tilde{\mathbf{w}}\, d\tilde{\mathbf{w}}^* \\
&\simeq \int \tilde{\mathbf{w}}^* p(\tilde{\mathbf{w}}^*|\tilde{\mathbf{w}}, \boldsymbol{\phi}) q(\tilde{\mathbf{w}}) q(\boldsymbol{\theta})\, d\boldsymbol{\theta}\, d\tilde{\mathbf{w}}\, d\tilde{\mathbf{w}}^* \\
&= \int \Sigma'_{\tilde{\mathbf{w}}, \tilde{\mathbf{w}}^*} \Sigma_{\tilde{\mathbf{w}}}^{-1} \tilde{\mathbf{w}} q(\tilde{\mathbf{w}}) q(\boldsymbol{\phi})\, d\boldsymbol{\phi}\, d\tilde{\mathbf{w}} = \left[\int \Sigma'_{\tilde{\mathbf{w}}, \tilde{\mathbf{w}}^*} \Sigma_{\tilde{\mathbf{w}}}^{-1} q(\boldsymbol{\phi})\, d\boldsymbol{\phi}\right] \boldsymbol{\mu}_q(\tilde{\mathbf{w}}) \\
&= \left[\mathrm{E}_q(\Sigma'_{\tilde{\mathbf{w}}, \tilde{\mathbf{w}}^*} \Sigma_{\tilde{\mathbf{w}}}^{-1})\right] \boldsymbol{\mu}_q(\tilde{\mathbf{w}})\ .
\end{aligned}$$

The other way to approximate $\tilde{\mathbf{w}}^*$'s posterior expectation uses the posterior distribution of the other parameters as (2.8). Then $\mathrm{E}(\tilde{\mathbf{w}}^*|\mathbf{Y})$ is

$$\begin{aligned}
\mathrm{E}(\tilde{\mathbf{w}}^*|\mathbf{Y}) &= \int \tilde{\mathbf{w}}^* p(\tilde{\mathbf{w}}^*|\tilde{\mathbf{w}}, \boldsymbol{\phi}) p(\tilde{\mathbf{w}}|\boldsymbol{\theta}, \mathbf{Y}) p(\boldsymbol{\theta}|\mathbf{Y})\, d\boldsymbol{\theta}\, d\tilde{\mathbf{w}}\, d\tilde{\mathbf{w}}^* \\
&= \int \Sigma'_{\tilde{\mathbf{w}},\tilde{\mathbf{w}}^*} \Sigma_{\tilde{\mathbf{w}}}^{-1} \tilde{\mathbf{w}} p(\tilde{\mathbf{w}}|\boldsymbol{\theta}, \mathbf{Y}) p(\boldsymbol{\theta}|\mathbf{Y})\, d\boldsymbol{\theta}\, d\tilde{\mathbf{w}} \simeq \int \Sigma'_{\tilde{\mathbf{w}},\tilde{\mathbf{w}}^*} \Sigma_{\tilde{\mathbf{w}}}^{-1} \boldsymbol{\mu}_p(\tilde{\mathbf{w}}) q(\boldsymbol{\theta})\, d\boldsymbol{\theta} \\
&= \mathrm{E}_q \left(\Sigma'_{\tilde{\mathbf{w}},\tilde{\mathbf{w}}^*} \Sigma_{\tilde{\mathbf{w}}}^{-1} \boldsymbol{\mu}_p(\tilde{\mathbf{w}}) \right),
\end{aligned}$$

where $\boldsymbol{\mu}_p(\tilde{\mathbf{w}}) = \left[\Sigma_{\tilde{\mathbf{w}}}^{-1} + \mathscr{A}'(\mathbf{I}_n \otimes \boldsymbol{\Psi}^{-1}) \mathscr{A} \right]^{-1} \mathscr{A}'(\mathbf{I}_n \otimes \boldsymbol{\Psi}^{-1})(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$.

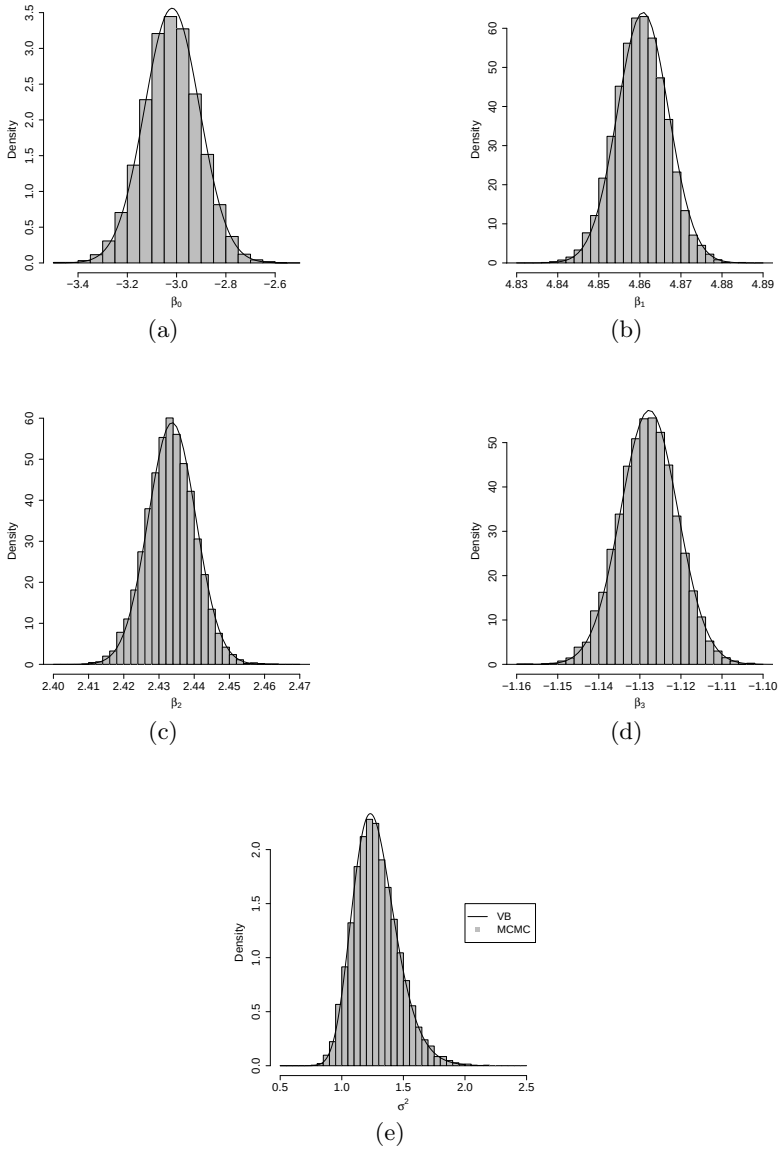


Figure 2.1: Linear regression example: (a)-(d) compare the posterior distributions of $\boldsymbol{\beta}$ from VB with the MCMC samples. (e) compares the the posterior distribution of σ^2 from VB with MCMC samples. The solid line is the approximate posterior distribution from VB and the histogram is the samples from MCMC.

2.5 Illustrations

2.5.1 Simulated example of simple linear regression

We begin by illustrating the VB algorithm to estimate the posteriors in the simple Bayesian conjugate linear model introduced in Section 2.2. The data comprises 100 observations generated from an ordinary linear model with slope parameters $\boldsymbol{\beta}' = (-3.15, 4.86, 2.44, -1.13)$ and unit variance $\sigma^2 = 1$. The regressor $\mathbf{X}$ is generated randomly with each element following uniform (-30, 30) except the first column being 1.

Given the same prior specification and synthetic data, VB and MCMC were used to estimate the posterior. One MCMC chain was run for 5000 iterations (after 2000 burn-in). Following the algorithm specified in Section 2.2, the VB algorithm converges within 5 iterations. Table 2.1 compares the posterior means and variances obtained from VB and MCMC. As discussed in Section 2.2 we can see that the estimated posterior variance from VB is smaller than MCMC, but quite close. The posterior densities of the parameters are shown in Figure 2.1. The histograms represent the MCMC samples and the solid lines are the estimated posterior densities computed using VB. The true posterior distributions are not shown in Figure 2.1 due to the similarity with the VB estimates.

		β_0	β_1	β_2	β_3	σ^2
MCMC	mean	-3.0205	4.8607	2.4338	-1.1276	1.2811
	variance	1.31e-02	3.94e-05	4.74e-05	4.98e-05	3.4e-02
VB	mean	-3.0183	4.8608	2.4337	-1.1277	1.2550
	variance	1.25e-02	3.88e-05	4.58e-05	4.85e-05	3.2e-02

Table 2.1: The posterior means and variances of the parameters from VB and MCMC in ordinary linear model.

2.5.2 Simulated example of univariate spatial model

To assess the proposed VB algorithm's utility for spatial models, we generated data from the univariate model specified in Section 2.3.1. The simulated data set involves 50 locations within a 100×100 square. The Matérn correlation function (Stein, 1999)

with $\upsilon = 0.5$ was used to produce the data's spatial dependence structure. Fixing υ at 0.5 reduces the Matérn function to the familiar exponential correlation function, $\text{cov}(\mathbf{w}(\mathbf{s}), \mathbf{w}(\mathbf{s}^*)) = \rho(\mathbf{s} - \mathbf{s}^*; \phi) = \exp(-\phi\|\mathbf{s} - \mathbf{s}^*\|)$. The data set was simulated with the following parameters: $\boldsymbol{\beta}' = (150, 10)$, $\tau^2 = 20$, $\sigma^2 = 50$ and $\phi = 0.1$. Gaussian process with exponential correlation function $\rho(\mathbf{s}_1 - \mathbf{s}_2; \xi) = \exp(-\xi\|\mathbf{s}_1 - \mathbf{s}_2\|)$ is utilized to generate spatially structured explanatory variables, which is very common in real applications such as elevation and temperature. Then the regressor matrix $\mathbf{X}$ has a column generated from the Gaussian process with mean 0 and $\xi = 1$ except the first column fixed to be 1 to indicate intercept.

As discussed in Section 2.3.1 and 2.3.2 there are two ways to apply VB. One is treating $\mathbf{w}$ as a hidden variable and updating it along with the other parameters. The other way is integrating $\mathbf{w}$ out from the likelihood and using the marginal distribution. When the VB marginal model was used, we found that different ways of grouping parameters and updating schemes result in different posterior approximations.

The BCLT estimates for the posterior percentiles using the R functions nlminb and hessian (for asymptotic posterior standard errors) are shown in Table 2.2 as well as the percentiles estimated from both MCMC and three VB methods. To find the BCLT estimates, all the positive parameters σ^2, τ^2 and ϕ are transformed using log function.

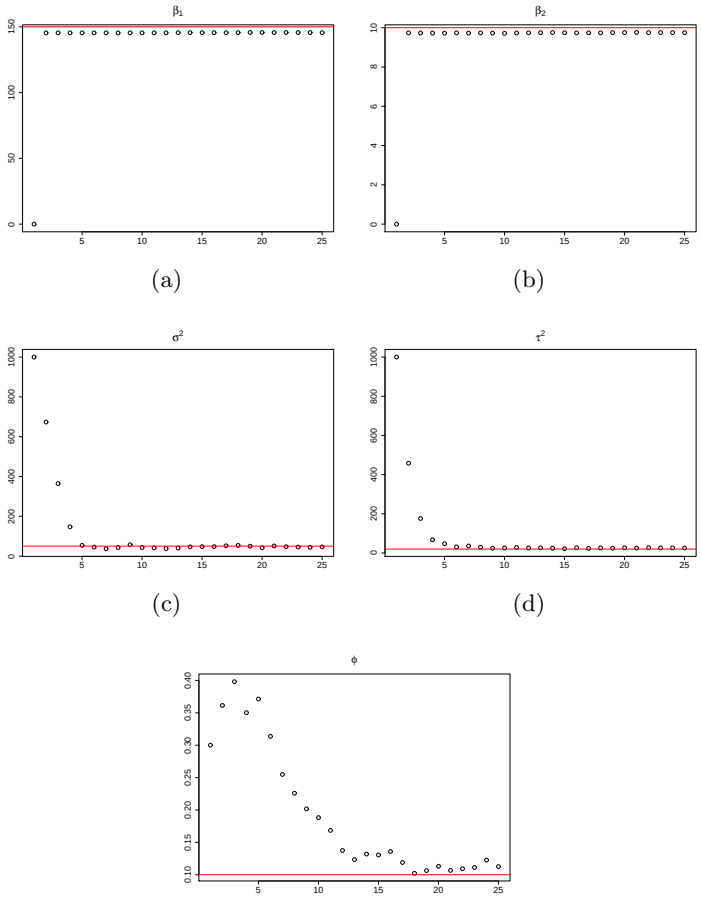
The Metropolis-Hastings algorithm was used for MCMC. Here, posterior inferences were based on 12,000 samples after discarding the initial 3,000 samples for burn-in. The VB algorithms were run until the hyper-parameters of the posterior distributions converged. Because of different parametrizations, not all the parameters in the table were directly updated in the algorithm for all the models (e.g., r in hidden variable model), whereupon the importance sampling resampling method (Rubin, 1987) was utilized to produce the posterior samples. (The estimates of these parameters are displayed in boldface in Table 2.2). Figure 2.2 offers the trace plots for the VB marginal model assuming $p(\boldsymbol{\theta} \mid \mathbf{Y}) \simeq q(\sigma^2, \tau^2, \phi)q(\boldsymbol{\beta})$, in which we can see the VB algorithm converges within 10 iterations. Both MCMC and VB algorithms are running very fast for univariate models, so we do not bother providing the running time here.

In Table 2.2, the VB marginal model assuming $p(\boldsymbol{\theta}|\mathbf{Y}) \simeq q(\sigma^2, \tau^2, \phi)q(\boldsymbol{\beta})$ provides the closest posterior percentiles estimates compared to MCMC. The BCLT and VB marginal model with $p(\boldsymbol{\theta} \mid \mathbf{Y}) \simeq q(r, \phi)q(\tau^2)q(\boldsymbol{\beta})$ also provide good posterior estimates,

Parameter (True)	MCMC Estimates	Marginal Model $q(\sigma^2, \tau^2, \phi)q(\boldsymbol{\beta})$	Marginal Model $q(r, \phi)q(\tau^2)q(\boldsymbol{\beta})$
$\beta_1 = 150$	145.78 (141.68, 150.91)	145.55 (142.27, 148.82)	145.56 (142.23, 148.89)
$\beta_2 = 10$	9.76 (8.90, 10.61)	9.75 (8.94, 10.56)	9.74 (8.94, 10.55)
$\sigma^2 = 50$	45.82 (20.64, 95.42)	43.89 (20.49, 89.32)	**47.87 (22.52, 104.23)**
$\tau^2 = 20$	27.97 (13.22, 56.18)	27.03 (13.13, 55.23)	23.29 (16.30, 34.90)
$\sigma^2/\tau^2 = 2.5$	1.67 (0.46, 5.00)	**1.68 (0.44, 5.00)**	2.31 (1.16, 5.38)
$\phi = 0.1$	0.092 (0.026, 0.774)	0.096 (0.03, 0.71)	0.09 (0.02, 0.50)
Parameter (True)	Treating $\mathbf{w}$ as Hidden Variable	Bayesian central limit using nlminb	Simple Bayesian linear regression
$\beta_1 = 150$	145.53 (144.16, 146.91)	145.64 (142.07, 149.21)	145.31 (143.03, 147.59)
$\beta_2 = 10$	9.75 (9.21, 10.29)	9.76 (8.96, 10.55)	9.77 (8.87, 10.67)
$\sigma^2 = 50$	49.20 (34.44, 73.74)	42.39 (22.30, 80.57)	
$\tau^2 = 20$	24.58 (17.21, 36.84)	24.34 (11.92, 49.71)	
$\sigma^2/\tau^2 = 2.5$	**1.76 (1.04, 3.06)**	1.74 (0.57, 5.32)	
$\phi = 0.1$	0.12 (0.07, 0.25)	0.10 (0.04, 0.29)	

Table 2.2: Posterior percentiles (50%, 2.5% and 97.5%) of MCMC, three VB methods, Bayesian central limit estimate and simple Bayesian linear regression. The percentiles calculated using importance sampling resampling method are shown in boldface.

while some the estimates for 95% confidence intervals are different from the MCMC and VB marginal models assuming $p(\boldsymbol{\theta} \mid \mathbf{Y}) \simeq q(\sigma^2, \tau^2, \phi)q(\boldsymbol{\beta})$. Note that the high-dimension optimization is not always stable and maybe sensitive to starting values. In multivariate spatial models, using nlminb to find the BCLT estimates could give problematic results, which are shown in Section 2.5.3. The coverage of the 95% confidence intervals of the VB hidden variable method is smaller than the others. So treating $\mathbf{w}$ as hidden results in independence of parameters' posterior distributions. Ignoring strong correlation between some of these parameters may cause the reduced width. With the regressor containing spatially structured explanatory variables in the simulated spatial model, both the slope parameter $\boldsymbol{\beta}$ and spatial parameters $\boldsymbol{\phi}$ are well estimated. So the spatial effects are identified without difficulty using VB method.



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

Figure 2.2: The trace plot of VB marginal model assuming $p(\boldsymbol{\theta}|\mathbf{Y}) \simeq q(\sigma^2, \tau^2, \phi)q(\boldsymbol{\beta})$. The sub figure (a)-(e) are the mean of β_1, β_2, σ^2, τ^2 and ϕ respectively.

Figure 2.3 illustrates the posterior distribution of parameters obtained from each of the candidate models. Here, the VB marginal model assuming $p(\boldsymbol{\theta} \mid \mathbf{Y}) \simeq q(\sigma^2, \tau^2, \phi)q(\boldsymbol{\beta})$, solid line, closely approximates the histograms of MCMC-based posterior distributions. The posterior estimates of VB model that treats $\mathbf{w}$ as hidden, dotted line, is much narrower than MCMC or marginal VB model.

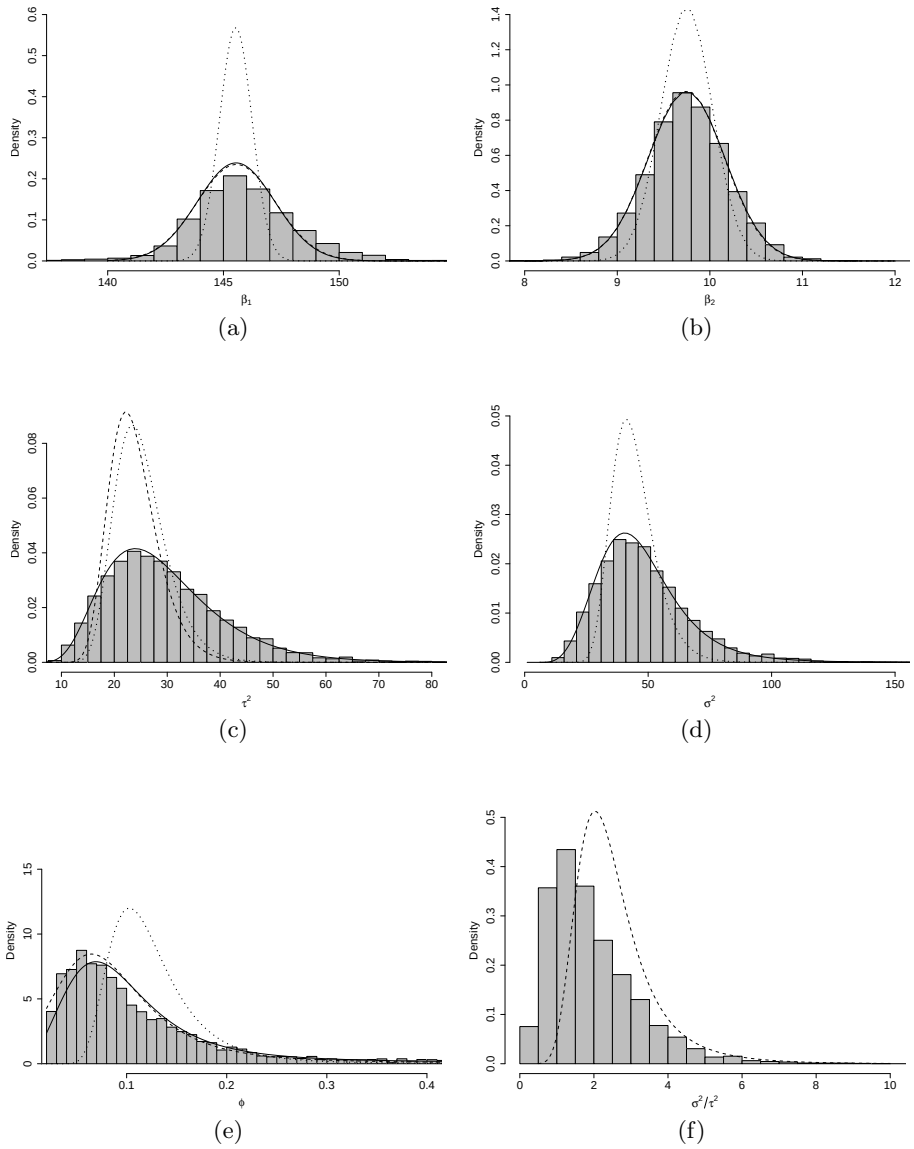


Figure 2.3: The posterior distributions got from different methods: MCMC(histogram); VB treating $\mathbf{w}$ as the hidden variable (dotted line for $\boldsymbol{\beta}$, σ^2, τ^2 and ϕ); VB marginal model with $p(\boldsymbol{\theta} \mid \mathbf{Y}) \simeq q(\sigma^2, \tau^2, \phi) q(\boldsymbol{\beta})$ (solid line for $\boldsymbol{\beta}$, σ^2, τ^2 and ϕ); VB marginal model with $p(\boldsymbol{\theta} \mid \mathbf{Y}) \simeq q(r, \phi) q(\tau^2) q(\boldsymbol{\beta})$ (dashed line for $\boldsymbol{\beta}$, τ^2, r and ϕ). (Notice that not all the distributions can be estimated in all the models.)

2.5.3 Simulated example of multivariate spatial model

Here we explore the performance of VB in a multivariate spatial model. The synthetic data set was generated from a stationary, isotropic, non-separable bivariate process (i.e., m=2). The exponential correlation function was used to produce the data's spatial dependence structure, in which $\rho(\mathbf{s}-\mathbf{s}^*;\phi) = \exp(-\phi\|\mathbf{s}-\mathbf{s}^*\|)$. Thus we take $\tilde{\mathbf{K}}(\mathbf{s}-\mathbf{s}^*;\boldsymbol{\phi}) = \text{diag}[\rho_i(\mathbf{s}-\mathbf{s}^*;\phi_i)]_{i=1}^2$ where $\boldsymbol{\phi} = (\phi_1, \phi_2)$. The multivariate process was simulated with the following parameters:

$$\boldsymbol{\beta}' = (1,-2,1,2), \quad \mathbf{K}(0) = \begin{pmatrix} 1 & -2 \\ -2 & 8 \end{pmatrix}, \quad \boldsymbol{\Psi} = \begin{pmatrix} 9 & 0 \\ 0 & 2 \end{pmatrix}, \quad \boldsymbol{\phi} = \begin{pmatrix} 0.6 \\ 0.1 \end{pmatrix}.$$

This yields $\mathbf{A} = \mathbf{K}^{1/2} = \begin{pmatrix} 1 & 0 \\ -2 & 2 \end{pmatrix}$. In this multivariate cases, we also use a Gaussian process with exponential correlation function $\rho(\mathbf{s}_1-\mathbf{s}_2;\xi) = \exp(-\xi\|\mathbf{s}_1-\mathbf{s}_2\|)$ to generate spatially structured explanatory variables. The regressor for each element of $\mathbf{Y}(\mathbf{s})$ is a 2×1 vector including an intercept and a covariate which is generated from the Gaussian process with mean 0 and $\xi = 0.05$. Then $\mathbf{x}(\mathbf{s})$ is a 4×2 block diagonal matrix with the diagonal blocks being the 2×1 regressor vectors. The above specifications describe a multivariate process with independent non-spatial variance for each response surfaces and a strong negative cross-correlation between the spatial processes. 150 observations were generated by using the parameters above in model (2.7). Figure 2.4 illustrates interpolated surfaces of the simulated response $\mathbf{Y}$ (a and b); the posterior mean of the spatial effects $\mathbf{w}$ (c and d); and the predicted random spatial effects $\text{E}(\mathbf{w}^*)$ (e and f).

MCMC algorithms using the spMvLM function of the spBayes package in R took approximately 1.5 hours to deliver its entire inferential output involving 20,000 iterations, including 5,000 samples for burn-in, on a 2.8GHz AMD Athlon processor with 2.0 GB of RAM running under Windows. Under the same conditions, the VB algorithm took about 0.5 hours to converge. For the VB method, the expectations of both $\mathbf{w}$ and $\mathbf{w}^*$ were calculated using the method introduced in Section 2.4.2, and $q(\tilde{\mathbf{w}})$ is the estimate for the true posterior of $\tilde{\mathbf{w}}$.

Percentiles of the posterior distributions estimated using VB, MCMC and BCLT are listed in Table 2.3. Here, again, we find the 95% confidence interval coverage for VB which treats $\mathbf{w}$ as a hidden variable to be smaller than that from MCMC for all the

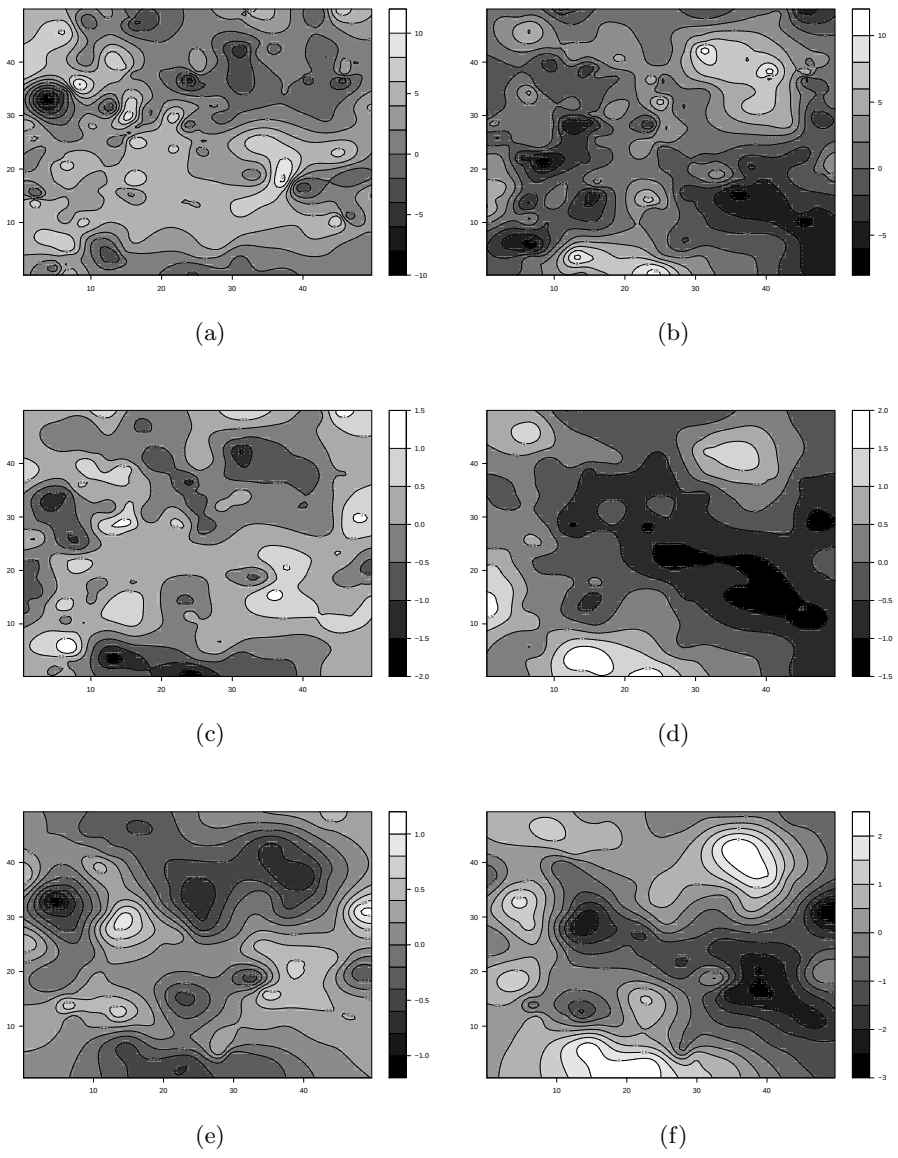


Figure 2.4: (a) and (b) are interpolated surfaces of the first and second response variable. (c) and (d) are the interpolated surface of the recovered random spatial effects from VB method $\mathrm{E}[\mathbf{w}|Data]$. (e) and (f) are the interpolated surface of the predicted random spatial effects $\mathrm{E}[\mathbf{w}^*|Data]$ from VB method.

Parameter (True)	VB	MCMC	Bayesian central limit
$\boldsymbol{\beta}_{1,0} = 1$	1.34 (1.18, 1.50)	1.27 (0.61 1.99)	1.22
$\boldsymbol{\beta}_{1,1} = -2$	-1.60 (-1.74, -1.47)	-1.71 (-2.32, -1.06)	-1.78
$\boldsymbol{\beta}_{2,0} = 1$	-0.31 (-0.38, -0.24)	-0.34 (-1.95, 1.36)	-0.43
$\boldsymbol{\beta}_{2,1} = 2$	1.99 (1.93, 2.05)	2.25 (1.34, 3.24)	2.27
$\mathbf{A}_{1,1} = 1$	1.07 (0.74 1.47)	1.32, (0.76, 2.71)	0.92
$\mathbf{A}_{2,1} = -2$	-1.75 (-2.09, -1.42)	-1.58, (-2.67, -0.23)	-1.26
$\mathbf{A}_{2,2} = 2$	2.05 (1.71, 2.44)	2.49 (1.51, 3.53)	2.49
$\Psi_1 = 9$	7.95 (6.41, 10.03)	7.61 (2.32, 10.14)	8.15
$\Psi_2 = 2$	2.17 (1.36, 4.26)	2.21 (0.86, 4.75)	1.97
$\phi_1 = 0.6$	0.43 (0.31, 0.66)	0.88 (0.18, 2.80)	3.0
$\phi_2 = 0.1$	0.16 (0.13, 0.23)	0.16 (0.06, 0.67)	0.22

Table 2.3: Percentiles (50%, 2.5% and 97.5%) of the posterior distribution of the parameters of VB, MCMC estimate. BCLT can only provide posterior mode. $\boldsymbol{\beta}$ subscripts refer to the response variable and parameter, respectively. Subscripts on $\mathbf{A}$ and $\boldsymbol{\Psi}$ refer to the covariance matrix element. Subscripts on the spatial range parameters, $\boldsymbol{\phi}$, refer to the response variable.

parameters. This trend was also seen for the univariate spatial model. The parameters with positive support were transformed using logarithm when applying BCLT. But the estimated hessian matrix was not positive definite. This result may because of the posterior estimate for ϕ_1 equaling its upper limits. The estimate stays the same even when we tried different initial values. So the posterior percentiles were not able to provided in Table 2.3. Similarly in multivariate cases, both the slope parameter $\boldsymbol{\beta}$ and spatial parameters $\boldsymbol{\phi}$ are well estimated with the regressor containing spatially structured explanatory variables.

2.5.4 Model selection

The generalized template introduced in Section 2.4 suggests several potential models. Here we consider five stationary process models of increasing complexity on the same synthetic data set introduced in Section 2.5.3. Our focus is on alternative specifications

Model	Parameters	G	P	$G+P$	DIC
Model 1	τ^2	2940.47	3015.96	5956.44	1548.53
Model 2	ϕ, σ^2, τ^2	2919.17	6194.75	9113.93	1611.27
Model 3	ϕ, $\boldsymbol{\sigma}_m^2$, $\boldsymbol{\Psi}$	1326.33	2119.56	3445.89	1438.11
Model 4	ϕ, $\mathbf{A}$, $\boldsymbol{\Psi}$	1344.25	2041.40	3385.65	1432.61
Model 5	$\boldsymbol{\phi}_m$, $\mathbf{A}$, $\boldsymbol{\Psi}$	1320.25	2026.12	**3346.37**	**1415.59**

Table 2.4: Synthetic data model comparison using DIC and minimum posterior predictive approach. For each model un-marginalized scores were calculated from 1000 samples.

of $\mathbf{A}$ and $\tilde{\mathbf{w}}$ in (2.7). For each model, we assume an isotropic spatial process that can be modeled with the exponential correlation function.

A simple linear regression model(no random effect) is

$$\text{Model 1: } \mathbf{A}\tilde{\mathbf{w}} = \mathbf{0} \ .$$

This model would suffice in the presence of negligible extraneous variation beyond what is explained by the model's regressors.

The next three spatial models impose separable association structures. For each model, $\Sigma_{\tilde{\mathbf{w}}} = [\tilde{\mathbf{K}}(\mathbf{s}_i - \mathbf{s}_j; \boldsymbol{\phi})]_{i,j=1}^n, \boldsymbol{\phi} = \{\phi\}_{k=1}^m$ implies the response variables share a common spatial decay parameter. The first, and simplest, of these models assume common spatial variance (i.e., σ^2) and a common pure error variance term (i.e., τ^2),

$$\text{Model 2: } \mathbf{A} = \sigma\mathbf{I}_m \text{ and } \boldsymbol{\Psi} = \tau^2\mathbf{I}_m.$$

The next model extends Model 2 to allow response specific spatial and pure error variance terms,

$$\text{Model 3: } \mathbf{A} = \text{diag}[\sigma_i]_{i=1}^m \text{ and } \boldsymbol{\Psi} = \text{diag}[\psi_i]_{i=1}^m.$$

Where Model 3 assumes independence among the response surfaces spatial variance. Model 4 explicitly models the off-diagonal element in the cross-covariance matrix $\mathbf{K}$,

$$\text{Model 4: } \mathbf{A} \text{ and } \boldsymbol{\Psi} = \text{diag}[\psi_i]_{i=1}^m$$

where, recall, $\mathbf{A}$ is the square root of the $m \times m$ cross-covariance matrix. The fifth model is the non-separable form of Model 4, allowing response specific spatial range terms,

$$\text{Model 5: } \mathbf{A}, \boldsymbol{\Psi} = \text{diag}[\psi_i]_{i=1}^m \text{ and } \boldsymbol{\phi} = \{\phi_k\}_{k=1}^m.$$

We fit the five competing models to the synthetic data in Section 2.5.3 using VB. After convergence, the posterior samples of parameters are drawn from the posterior distributions. Model selection was made using the deviance information criterion (DIC) (Spiegelhalter et al., 2002) and a posterior predictive criterion that balances goodness of fit and predictive variance under a squared error loss function, presented by Gelfand and Ghosh (1998). This assigns a score to each model that is the sum of two terms, P and G. Similar to DIC, the model with smaller $G + P$ value is preferred.

Table 2.4 provides DIC and posterior predictive loss approach for candidate models. Based on DIC and the effective number of parameters, Model 5 is the most parsimonious of the five, which is consistent with using the $G + P$ criterion. It is common that the notoriously ill-defined $\boldsymbol{\phi}$ does not contribute much to the model distinctions in formal model fit comparisons. Rather, we might look to the parameter estimates to determine if there is an advantage to use a more complicated model. There is a strong distinction between estimates for ϕ_1 and ϕ_2 in Table 2.3. Therefore we would conclude that Model 5 is preferred for this data set.

2.5.5 Forest inventory data analysis and results

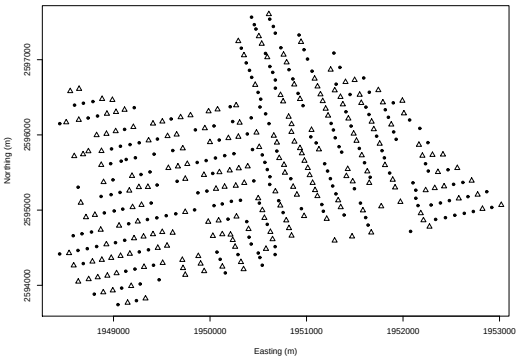
Spatially explicit estimates of forest biomass are important for quantifying forest carbon dynamics, forecasting wood availability, and a host of other forest and environmental management activities. Here we generate such maps using data from permanent georeferenced forest inventory plots on the USDA Forest Service Bartlett Experimental Forest (BEF) in Bartlett, New Hampshire. Total tree biomass on each of 415 forest inventory plots were apportioned into tree bole, branches, and foliage. For this illustration our interest is in bole and non-bole (i.e., branches+foliage) biomass. Given the known area of each inventory plot and number of measured trees, we express these quantities as metric tons of total bole and non-bole biomass per hectare. Satellite imagery and other remotely sensed variables have proved useful regressors for predicting forest biomass.

One summer 2002 of 30×30 Landsat 7 ETM+ satellite imagery was acquired for the BEF. The image was transformed to tasseled cap components of brightness (1), greenness (2), and wetness (3) using data reduction techniques. The three resulting spectral variables are labeled TC1, TC2, and TC3. In addition to these spectral variables, digital elevation model data was used to produce a 30×30 elevation (ELEV) and slope (SLOPE) layer for the BEF (see Finley et al. (2008) for more details). Using a geographic information system, these regressors were associated with the biomass response variables at each inventory plot location to form the $415 \cdot 2 \times 6 \cdot 2$ $\mathbf{X}$ and $415 \cdot 2 \times 1$ $\mathbf{Y}$ regressor matrix and response vector, respectively.

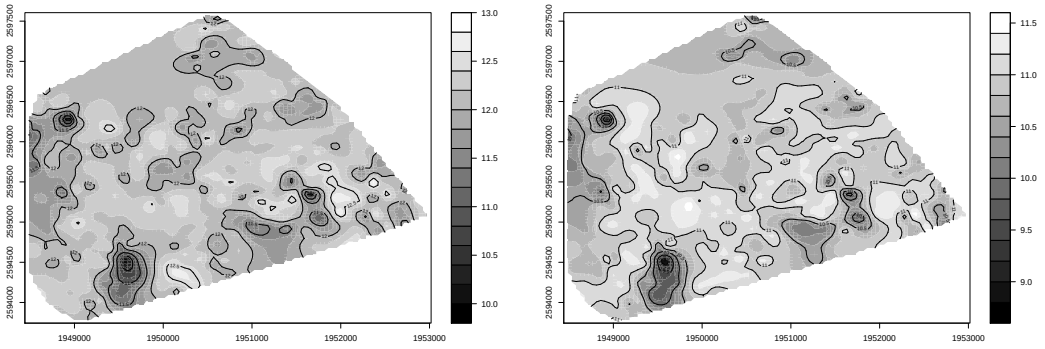
To demonstrate parameter estimation and prediction, we randomly selected 200 inventory plots for model construction and left the remaining 215 for subsequent predictive mapping. For reference, the 200 model points in Figure 2.5(a) are used to produce an interpolated surface of the biomass for each of the two categories, Figure 2.5(b) and 2.5(c).

As in the previous illustration, we fit a non-separable spatial regression with full spatial and non-spatial diagonal cross-covariance matrices, $\mathbf{K}$ and $\mathbf{\Psi}$. Further, we assume that spatial dependence can be modeled with the simple exponential correlation function. This specification corresponds to Model 5 in section 2.5.4. The inverse-Wishart prior is used for $\mathbf{K}$ and Inverse-Gamma prior for the diagonal elements of $\mathbf{\Psi}$. The noninformative prior on the spatial range parameters $\boldsymbol{\phi}$ corresponds to a range which allows for an effective spatial range to cover the maximum distance between the locations. Three MCMC chains were run for 10,000 iterations. The three chains allowed for dispersed parameter starting values. Chain mixing occurred within 1,000 iterations; Therefore, 27,000 samples were retained for posterior analysis.

The VB and MCMC estimates for the Forest inventory data are provided in Table 2.5. In comparison to the MCMC based estimates, the narrower credible intervals estimated by the VB model suggest that several regression coefficients are significant at the 0.05 level. Both methods give similar estimates for other parameters. The significance of the off-diagonal element $\mathbf{A}_{2,1}$ suggests that there is positive spatial association between the conditional response surface. The spatial range estimates in Table 2.5 do not support a distinction between the responses' spatial dependence structure, therefore, the separable form of this model might be considered.



(a)



(b) (c)

Figure 2.5: (a) Forest inventory plots across the Bartlett Experimental Forest. The 415 plots were divided randomly into 200 plots used for parameter estimation denoted with solid dot symbols (•) and the remaining 215 used for prediction marked with triangle symbols (△). Plots (b) and (c) are interpolated surfaces of biomass per hectare of the bole, and non-bole, respectively.

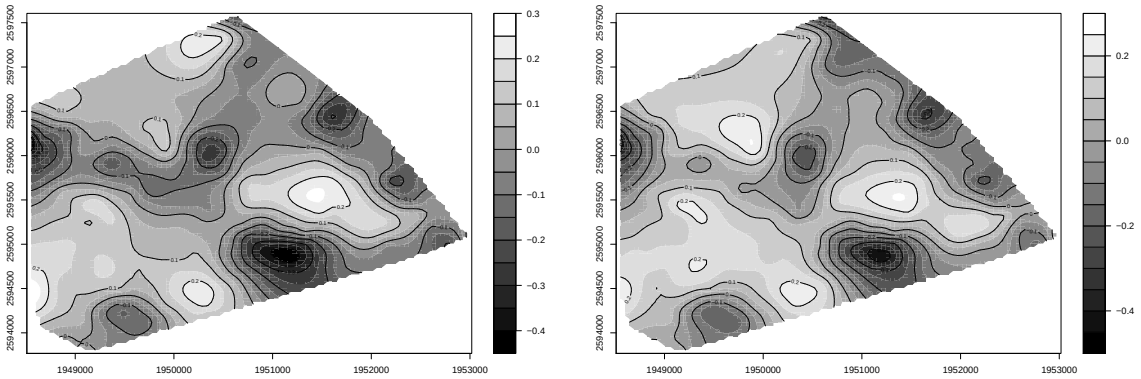
Parameter	VB	MCMC
$\boldsymbol{\beta}_{1,0}$	8.93 (7.48, 10.38)	8.95 (5.94, 11.95)
$\boldsymbol{\beta}_{1,ELEV}$	1.56e-05 (1.55e-05, 1.57e-05)	-3.67e-05 (-1.45e-03, 1.15e-03)
$\boldsymbol{\beta}_{1,SLOPE}$	-4.22e-03 (-4.24e-03, -4.20e-03)	-4.47e-03 (-1.88e-02, 9.40e-03)
$\boldsymbol{\beta}_{1,TC1}$	1.466e-02 (1.458e-02, 1.474e-02)	1.47e-02 (-1.16e-02, 3.98e-02)
$\boldsymbol{\beta}_{1,TC2}$	2.84e-04 (2.47e-04, 3.20e-04)	3.40e-04 (-1.57e-02, 1.69e-02)
$\boldsymbol{\beta}_{1,TC3}$	1.653e-02 (1.645e-02, 1.660e-02)	1.66e-02 (-6.09e-03, 3.98e-02)
$\boldsymbol{\beta}_{2,0}$	7.68 (6.09, 9.27)	7.72 (4.54, 10.87)
$\boldsymbol{\beta}_{2,ELEV}$	6.25e-05 (6.24e-05, 6.26e-05)	-6.75e-05 (-1.49e-03, 1.33e-03)
$\boldsymbol{\beta}_{2,SLOPE}$	-1.18e-03 (-1.21e-03, -1.16e-03)	-1.19e-03 (-1.58e-02, 1.30e-02)
$\boldsymbol{\beta}_{2,TC1}$	2.01e-02 (2.00e-02, 2.02e-02)	2.05e-02 (-6.38e-03, 4.77e-02)
$\boldsymbol{\beta}_{2,TC2}$	-3.11e-03 (-3.15e-03, -3.07e-03)	-3.39e-03 (-2.05e-02, 1.35e-02)
$\boldsymbol{\beta}_{2,TC3}$	1.66e-02 (1.65e-02, 1.67e-02)	1.65e-02 (-8.02e-03, 4.06e-02)
$\mathbf{A}_{1,1}$	0.30 (0.25, 0.34)	0.34 (0.27, 0.45)
$\mathbf{A}_{2,1}$	0.16 (0.037, 0.24)	0.24 (0.14, 0.35)
$\mathbf{A}_{2,2}$	0.28 (0.23, 0.31)	0.29 (0.24, 0.36)
Ψ_1	0.0790 (0.0654, 0.0965)	0.0599 (0.0458, 0.0787)
Ψ_2	0.0710 (0.0588, 0.0868)	0.0635 (0.0490, 0.0829)
ϕ_1	0.0028 (0.0021, 0.0038)	0.0021 (0.0013, 0.0067)
ϕ_2	0.0013 (0.0012, 0.0016)	0.0013 (0.0012, 0.0017)

Table 2.5: Percentiles ($50\%, 2.5\%$ and 97.5%) of the posterior distribution of the parameters of VB methods and MCMC. $\boldsymbol{\beta}$ subscripts refer to the response variable and parameter, respectively. Subscripts on $\mathbf{A}$ and $\boldsymbol{\Psi}$ refer to the covariance matrix element. Subscripts on the spatial range parameters, $\boldsymbol{\phi}$, refer to the response variable. Summaries in MCMC generated from three chains of 4500 samples.

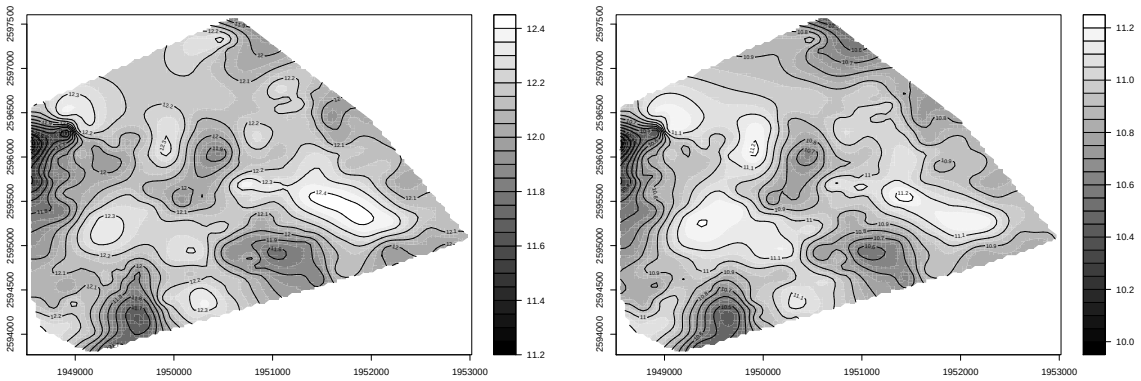
Given parameters' posterior distributions calculated using the VB algorithm, we can now turn to prediction of the holdout set's locations using the argument in Section 2.4.2. The interpolated surfaces of the predicted random spatial effects $\mathbf{w}^*$ and biomass $\mathbf{Y}^*$ are shown in Figure 2.6 (a) and (b), respectively.

2.6 Conclusion & Discussion

Our interest was to explore the utility of VB as a tool to fit spatial models. Using commonly accepted MCMC based methods as a baseline, we assessed our proposed VB algorithm's convergence and ability to summarize parameters' posterior distribution. We also compared the VB method with BCLT estimates, which gives good results in univariate models, but not for multivariate cases. VB methods can provide precise posterior estimates for the parameters in a relative shorter time compared to MCMC especially for multivariate spatial models, which require massive computational time due to big matrix decomposition. We proposed VB algorithm to fit both unmarginalized and marginalized likelihoods (as in (2.2) and (2.3)). The unmarginalized model offers the advantage of closed form expressions for $\boldsymbol{\beta}$, τ^2 and σ^2. However, these computational benefits come at a cost. Specifically, by treating $\mathbf{w}$ as hidden, the estimated posterior distribution of several important parameters are falsely precise. This narrowing of the posterior distributions was illustrated using both the univariate and multivariate models. The marginal models, which rely on importance sampling, showed slower convergence but more closely approximated the posterior distributions obtained with MCMC based methods.



(a)



(b)

Figure 2.6: (a) Interpolated surfaces of the predicted random spatial effects for biomass per hectare of the bole (left plot) and non-bole (right plot), $\mathrm{E}[\mathbf{w}^*|Data]$. (b) Interpolated surfaces of the posterior predictive distributions for biomass per hectare of the bole (left plot) and non-bole (right plot), $\mathrm{E}(\mathbf{Y}^*|Data)$.

Chapter 3

Hierarchical Factor Models for Large Spatially Misaligned Data: A Low-rank Predictive Process Approach

3.1 Introduction

With enhanced capabilities in collecting, storing and accessing geographically referenced data sets, spatial analysts today frequently encounter data comprising a large number of variables across a very large number of locations. In several instances, inference focuses upon three major tasks: (i) estimate associations among the variables, (ii) estimate the strength of spatial association for each variable, and (iii) predict the outcomes at arbitrary locations.

Modeling multiple geographically referenced outcomes proceeds from two different premises. One approach (Royle and Berliner, 1999; Gelfand et al., 2004; Cressie and Wikle, 2011) considers a conditional regression-like approach where the marginal distribution of the first outcome is specified, followed by the conditional distribution of the second outcome given the first, and so on. While elegant and often easily interpretable, the approach is more suitable when there is a natural "ordering" of the outcomes that

would suggest the sequence for constructing the conditional distributions.

Settings that lack such information are more appropriately handled with joint modeling for the set of outcomes to avoid the explosion in models emerging from alternate ordering schemes. Conditional inference can subsequently proceed from the joint distribution by conditioning on the relevant variables. The challenge in the joint modeling approach is to specify valid multivariate spatial processes using matrix-valued cross-covariance functions (e.g., Banerjee et al., 2004, Ch. 7). Gelfand et al. (2004) offer a detailed comparison of both approaches for multivariate spatial data. Here we focus upon the joint modeling approach.

The linear model of coregionalization (LMC) developed by Matheron (1982) is perhaps among the most general models for multivariate spatial data analysis. In LMC, the spatial behavior of the outcomes is assumed to arise from a linear combination of independent latent processes operating at different spatial scales (Wackernagel, 2003; Chilés and Delfiner, 1999). The idea is not very different from latent factor analysis (FA) models for multivariate data analysis (e.g., Anderson, 2003) except that in the LMC the number of latent processes is usually taken to be the same as the number of outcomes.

For estimating LMC's, an $m \times m$ covariance matrix has to be estimated for each spatial scale (see, e.g., Lark and Papritz, 2003; Castrignanó et al., 2005, 2000; Zhang, 2007; Finley et al., 2008), where m is the number of outcomes. When m is large (e.g., $m > 5$ and 300 spatial locations), obtaining such estimates is expensive, which restricts the usage of additional spatial scales in the model. In other multivariate process modeling strategies such as these presented by Schmidt and Gelfand (2003) and Gelfand et al. (2004), m latent spatial processes are presented, but only an $m \times m$ lower triangular matrix is associated with these processes. However, high dimensional outcomes can still be computationally prohibitive for these models.

When the number of independent latent processes in an LMC is taken to be fewer than the number of outcomes (e.g., Grzebyk and Wackernagel, 1994; Zhang, 2007), we obtain a spatial factor model. These have been widely deployed for modeling multivariate spatial data. Wang and Wall (2003) studied multivariate indicators of cancer risk across counties in Minnesota using one common spatial factor. Liu et al. (2005) extended the idea to hierarchical models using more than one factor with the assumption

that each underlying factor explains its own unique set of observed/measured variables. Christensen and Amemiya (2001, 2002, 2003) developed semiparametric latent variable models for rectangular grids, which they refer to as the shift-factor analysis method. Hogan and Tchernis (2004) fitted a one-factor spatial model and compared the results using different forms of spatial dependence through the single factor. Minozzo and Fruttini (2004) applied log-linear spatial factor analysis to geo-referenced frequency counts adopting the classical proportional covariance model to the latent factors.

Our current work is similar to the aforementioned articles in its use of independent latent spatial processes as underlying factors. However, we propose three methodological innovations within this framework. First, we use a multivariate low-rank spatial process to achieve dimension reduction over space. Critically, we need to work with irregularly spaced locations that do not necessarily lie on a grid, nor can they be easily projected onto a grid. While there are several choices here, we deploy the multivariate Gaussian predictive process (Banerjee et al., 2008, 2010). The method is closely related to kernel-convolutions, splines, and low-rank kriging (see, e.g., Wikle and Cressie, 1999; Ver Hoef et al., 2004; Kamman and Wand, 2003; Paciorek, 2007; Cressie and Johannesson, 2008), all of which attempt to facilitate computation through lower dimensional process representations. The predictive process can be applied to any spatial correlation function and maintains the richness of desired hierarchical spatial modeling specifications using a set of locations (or knots).

Second, we do not fix the number of factors but model it stochastically. We do so differently from some existing approaches. For example, Lopes and West (2004) addressed this problem by constructing proposals using the results of a preliminary MCMC run under each model. Such an approach has high computational demands, becoming infeasible as the sample size and potential number of factors increase. Dunson (2006) introduced a model averaging method for factor analysis, but the construction of the factor selection makes it hard to apply spatial models. Chen and Dunson (2003) proposed a Bayesian random effect selection method which is similar to what we propose but, instead of selecting random variables, we choose the underlying latent *processes* that capture spatial dependence. We call our model an *adaptive* spatial factor model. We avoid complex computational strategies such as reversible jump algorithms and build our adaptive models by using some key identifiability results, hitherto largely unaddressed,

to construct hierarchical models.

Third, we reckon with spatial misalignment in the context of spatial factor analysis. Misalignment occurs frequently in spatial data when not all variables have been observed at all locations. Put differently, the sets of observed locations for the different outcomes are not identical (either because some are missing or the outcomes have been collected by different monitoring sets). Assuming that all covariates are available at a location, we want to estimate the functional relationship between the covariates and the outcomes at that location – even if all the outcomes have not been observed there. We also seek to predict the outcomes at any arbitrary location in the domain, thereby estimating the response surfaces for each outcome.

Our motivating application pertains to ambient air quality assessment in California. The deleterious impact of air pollution upon health and quality of life is widely recognized as a major environmental issue (e.g., Dominici et al., 2006). Spatial interpolation of air pollutants plays a crucial role in assessing and monitoring ambient air quality and modeling multiple pollutants can capture associations within and across different locations, which can enhance predictive performance. Our dataset includes five commonly encountered pollutants observed over 300 monitoring stations across California and is spatially misaligned in the aforementioned manner. Estimating fully-specified joint models for such data will be exorbitant in terms of computing, which is why multivariate spatial modeling has rarely been undertaken in such frameworks.

The remainder of this article is organized as follows. Section 3.2 describes the features of the LMC and discusses model construction, identifiability issues, stochastic selection, and prior specification. Section 3.3 outlines the proposed class of low-rank adaptive spatial factor models and how we handle misaligned data and carry out inference. Section 3.4 illustrates the analysis for two simulated data sets and one air quality monitoring data set. Finally, Section 3.5 concludes the chapter with a summary and an eye toward future work.

3.2 Model Construction

3.2.1 LMC Model Structure and Specification

LMC consists of decomposing the set of original second-order stationary outcomes into a set of reciprocally orthogonal regionalized factors. Suppose, for a study region $D \subseteq \mathbb{R}^d$, an $m \times 1$ process $\mathbf{Z}(\mathbf{s}) = (Z_1(\mathbf{s}), \ldots, Z_m(\mathbf{s}))'$ is a second-order stationary process. Then, for all $\mathbf{s}, \mathbf{h} \in \mathbb{R}^d$ and $i, j = 1, \ldots, m$, we have $\mathrm{E}[Z_i(\mathbf{s})] = \mu_i$ and $\mathrm{cov}\{Z_i(\mathbf{s}), Z_j(\mathbf{s}+\mathbf{h})\} = C_{ij}(\mathbf{h})$. The matrix-valued function $\mathbf{C}(\mathbf{h}) = \{C_{ij}(\mathbf{h})\}$ is called the multivariate cross-covariance. Generally for $i \neq j$, a change in the order of the variables or a change in the sign of the separation vector $\mathbf{h}$ changes the values of the cross-covariances. If both the sequence and the sign are changed, we would have the same value $C_{ij}(\mathbf{h}) = C_{ji}(-\mathbf{h})$, which implies $\mathbf{C}(\mathbf{h}) = \mathbf{C}(-\mathbf{h})'$. $\mathbf{C}(\mathbf{h})$ must also be a positive definite function. That is, for any finite set of spatial locations $\mathbf{s}_1, \ldots, \mathbf{s}_n \in D$ and any vectors $\mathbf{a}_i \in \mathbb{R}^m, i = 1, \ldots, n$, $\mathrm{var}\{\sum_i^n \mathbf{a}_i'\mathbf{Z}(\mathbf{s}_i)\} = \sum_{i,j}^n \mathbf{a}_i'\mathbf{C}(\mathbf{s}_i - \mathbf{s}_j)\mathbf{a}_j \geq 0$. Specifying a valid cross-covariance function is less straightforward due to this constraint, but several spectral and constructive approaches (Ver Hoef and Barry, 1998; Chilés and Delfiner, 1999) have been proposed and used in multivariate spatial analysis.

The most straightforward form for achieving a valid cross-covariance model is the so-called intrinsic specification, $\mathbf{C}(\mathbf{h}) = \rho(\mathbf{h})\mathbf{T}$, where $\mathbf{T}$ is an $m \times m$ positive definite matrix and $\rho(\mathbf{h})$ is a univariate correlation function. The limitation here is that each $Z_i(\mathbf{s})$ has the same spatial correlation parameters. In particular, this means that all the outcome would have the same strength in spatial association over the domain. An extension (see, e.g., Wackernagel, 2003) specifies the cross-covariance as $\mathbf{C}(\mathbf{h}) = \sum_{k=1}^r \rho_k(\mathbf{h}; \boldsymbol{\phi}_k)\mathbf{T}_k$, where for each k, $\mathbf{T}_k$ is a rank-one positive semi-definite matrix and $\rho_k(\mathbf{h}; \boldsymbol{\phi}_k)$ is a correlation function that depends on additional parameters $\boldsymbol{\phi}_k$. Here, r is the total number of different spatial correlation functions in the multivariate cross-covariance.

The spectral decomposition yields $\mathbf{T}_k = \xi_k \mathbf{u}_k \mathbf{u}_k'$, where $\mathbf{u}_k$ is the normalized (i.e. $\|\mathbf{u}_k\| = 1$) eigenvector of $\mathbf{T}_k$ corresponding to the *only* positive eigenvalue ξ_k of $\mathbf{T}_k$. Since $\mathbf{T}_k$ is symmetric with rank one, all its other eigenvalues are zero. This implies that there is an $m \times 1$ vector, $\boldsymbol{\lambda}_k = \sqrt{\xi_k}\mathbf{u}_k$, which is the square root of $\mathbf{T}_k$ such that $\boldsymbol{\lambda}_k\boldsymbol{\lambda}_k' = \mathbf{T}_k$. Although $\|\mathbf{u}_k\| = 1$, it still has two possible directions. To build a one-to-one transformation between $\mathbf{T}_k$ and $\boldsymbol{\lambda}_k$, we need to impose some further constraints on

$\boldsymbol{\lambda}_k$ by, for instance, restricting the first element of $\boldsymbol{\lambda}_k$ to be positive.

Let $w_k(\mathbf{s})$, $k = 1, \ldots, r$, be independently distributed univariate stationary Gaussian processes, each with unit variance and a parametric correlation function. We write $w_k(\mathbf{s}) \sim GP(0, \rho_k(\cdot; \boldsymbol{\phi}_k))$ with assumption $\text{var}\{w_k(\mathbf{s})\} = 1$, $\text{cov}\{w_k(\mathbf{s}), w_k(\mathbf{t})\} = \rho_k(\mathbf{s}, \mathbf{t}; \boldsymbol{\phi}_k)$ and $\text{cov}\{w_k(\mathbf{s}), w_l(\mathbf{t})\} = 0$ whenever $k \neq l$ for all $\mathbf{s}$ and $\mathbf{t}$ (even when $\mathbf{s} = \mathbf{t}$). Here $\rho_k(\cdot; \boldsymbol{\phi}_k)$ is a correlation function associated with $w_k(\mathbf{s})$, and $\boldsymbol{\phi}_k$ includes the spatial decay and smoothness parameters. We can easily see that the multivariate cross-covariance function for the process $w_k(\mathbf{s})\boldsymbol{\lambda}_k$ is $\rho_k(\cdot; \boldsymbol{\phi}_k)\mathbf{T}_k$ and hence, for $\sum_{k=1}^r w_k(\mathbf{s})\boldsymbol{\lambda}_k$, the function is $\sum_{k=1}^r \rho_k(\cdot; \boldsymbol{\phi}_k)\mathbf{T}_k$.

We often assume that the mean of the outcomes arises linearly in the predictors so that the mean of the j-th outcome $Y_j(\mathbf{s})$ is modeled as $\mathbf{x}_j(\mathbf{s})'\boldsymbol{\beta}_j$, where $\mathbf{x}_j(\mathbf{s})$ is a $p_j \times 1$ vector of predictors, assumed to be known at location $\mathbf{s}$, and $\boldsymbol{\beta}_j$ is the corresponding $p_j \times 1$ vector of slopes. Let $\mathbf{Y}(\mathbf{s})$ be the $m \times 1$ vector of outcomes with j-th element $Y_j(\mathbf{s})$ and let $\mathbf{X}(\mathbf{s})'$ be an $m \times p$ block diagonal matrix, where $p = \sum_{j=1}^m p_j$ and the j-th diagonal block is given by $\mathbf{x}_j(\mathbf{s})'$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1', \boldsymbol{\beta}_2', \ldots, \boldsymbol{\beta}_m')'$ is a $p \times 1$ vector of slopes. Then, the spatial factor model is

$$\mathbf{Y}(\mathbf{s}) = \mathbf{X}(\mathbf{s})'\boldsymbol{\beta} + \sum_{k=1}^{r} w_k(\mathbf{s})\boldsymbol{\lambda}_k + \boldsymbol{\epsilon}(\mathbf{s}) = \mathbf{X}(\mathbf{s})'\boldsymbol{\beta} + \boldsymbol{\Lambda}\mathbf{w}(\mathbf{s}) + \boldsymbol{\epsilon}(\mathbf{s}), \tag{3.1}$$

where $\boldsymbol{\Lambda} = (\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_r)$ is a $m \times r$ matrix with k-th column $\boldsymbol{\lambda}_k$, $\mathbf{w}(\mathbf{s}) = (w_1(\mathbf{s}), \ldots, w_r(\mathbf{s}))'$, and $\boldsymbol{\epsilon}(\mathbf{s})$ is an $m \times 1$ vector of measurement errors distributed as $N(\mathbf{0}, \boldsymbol{\Psi})$. The measurement error variance $\boldsymbol{\Psi}$ can be any $m \times m$ positive definite matrix but is usually assumed diagonal with elements ψ_j^2 for $j = 1, \ldots, m$ along the diagonal.

3.2.2 Identifiability in the Spatial Factor Model

Model (3.1) is similar to FA models with the factor loading matrix $\boldsymbol{\Lambda}$ and latent factor $\mathbf{w}(\mathbf{s})$, except that $\mathbf{w}(\mathbf{s})$ is now a multivariate stochastic process positing spatial dependence. As is well-known (see, e.g., Anderson, 2003), orthogonal FA models must be further constrained to ensure identifiability. A widely used approach is to fix certain elements of $\boldsymbol{\Lambda}$ to constant values, usually to zeroes, such as restricting $\boldsymbol{\Lambda}$ to be an upper or lower triangular matrix with strictly positive diagonal elements (Lopes and West, 1999).

In LMC's (see, e.g., Schmidt and Gelfand, 2003; Finley et al., 2008; Gelfand et al., 2004), a lower-triangular $\mathbf{\Lambda}$ with positive diagonal elements identifies the covariances among the outcomes within a location because $\mathbf{C}(\mathbf{0}) = \mathbf{\Lambda}\mathbf{\Lambda}'$. But what we really seek to model is $\mathbf{C}(\mathbf{h})$. Let $\mathbf{P}$ be an $r \times r$ orthogonal matrix, so that $\mathbf{P}'\mathbf{P} = \mathbf{P}\mathbf{P}' = \mathbf{I}_r$. The random effect term in (3.1) can be written as $\mathbf{\Lambda}\mathbf{w}(\mathbf{s}) = \mathbf{\Lambda}\mathbf{P}\mathbf{P}'\mathbf{w}(\mathbf{s}) = \bar{\mathbf{\Lambda}}\bar{\mathbf{w}}(\mathbf{s})$, where $\bar{\mathbf{\Lambda}} = \mathbf{\Lambda}\mathbf{P}$ and $\bar{\mathbf{w}}(\mathbf{s}) = \mathbf{P}'\mathbf{w}(\mathbf{s})$. For non-spatial or traditional FA, the elements of $\mathbf{w}(\mathbf{s})$ are uncorrelated and the model is invariant to any orthogonal transformation because $\mathbf{w}(\mathbf{s})$ and $\bar{\mathbf{w}}(\mathbf{s})$ have identical cross-covariances and $\mathbf{\Lambda}\mathbf{\Lambda}' = \bar{\mathbf{\Lambda}}\bar{\mathbf{\Lambda}}'$. One can, therefore, obtain an infinite number of equivalent matrices of factor loadings by simply applying orthogonal transformations.

Matters are subtly different when $\mathbf{w}(\mathbf{s})$ is a spatial process. The cross-covariance matrix $\text{cov}\{\mathbf{w}(\mathbf{s}), \mathbf{w}(\mathbf{t})\} = \mathbf{\Gamma}(\mathbf{s}, \mathbf{t}; \boldsymbol{\phi})$, is now diagonal (and not an identity matrix as in traditional FA) with k-th diagonal element $\rho_k(\mathbf{s}, \mathbf{t}; \boldsymbol{\phi}_k)$. Orthogonal transformations can, therefore, alter the distribution of $\bar{\mathbf{w}}(\mathbf{s})$. To be precise, now $\text{cov}\{\bar{\mathbf{w}}(\mathbf{s}), \bar{\mathbf{w}}(\mathbf{t})\} = \mathbf{P}'\mathbf{\Gamma}(\mathbf{s}, \mathbf{t}; \boldsymbol{\phi})\mathbf{P}$, which is neither necessarily diagonal nor equal to $\mathbf{\Gamma}(\mathbf{s}, \mathbf{t}; \boldsymbol{\phi})$. Therefore, spatial factor models are not necessarily invariant to *any* orthogonal transformation.

Our primary motivation for using a $\mathbf{\Lambda}$ without any prespecified 0 elements is stochastic modeling for the number of factors (see Section 3.2.3). Such a model would automatically identify the latent spatial processes and, hence, the corresponding columns of $\mathbf{\Lambda}$ that are retained in (3.1). Specifying $\mathbf{\Lambda}$ to be lower triangular is now problematic because it is unlikely that such a structure will be retained throughout the stochastic selection process. On the other hand, using a $\mathbf{\Lambda}$ that is identifiable but that does not restrict certain elements to be zero will avoid such awkwardness.

In fact, we argue (see Appendix C for details) that only two groups of orthogonal transformations lead to non-identifiability in spatial factor models. The first transformation $\mathbf{P}$ satisfies $\mathbf{P}'\mathbf{\Gamma}(\mathbf{s}, \mathbf{t}; \boldsymbol{\phi})\mathbf{P} = \mathbf{\Gamma}(\mathbf{s}, \mathbf{t}; \boldsymbol{\phi})$. Such a $\mathbf{P}$ must be diagonal with 1's and -1's only. It is a special *reflector* (i.e. $\mathbf{P}^2 = \mathbf{I}$) and is obtained as a product of elementary reflectors of the form $\mathbf{I} - 2\mathbf{e}_i\mathbf{e}_i'$, where $\mathbf{e}_i$ is the i-th canonical vector ($\mathbf{P}_1$ in (3.2)). This orthogonal matrix $\mathbf{P}$ alters the sign of the columns in $\mathbf{\Lambda}$. To identify $\mathbf{\Lambda}$ (or have a one to one relationship between $\mathbf{\Lambda}$ and $\mathbf{\Lambda}\mathbf{\Lambda}'$), we need to specify one element in each column of $\mathbf{\Lambda}$ as positive or negative. Without losing generality, we could set the first

row of $\mathbf{\Lambda}$ to be positive.

$$\mathbf{P}_1 = \begin{pmatrix} \ddots & & & & \\ & 1 & & & \\ & & \ddots & & \\ & & & -1 & \\ & & & & \ddots \end{pmatrix}, \ \mathbf{P}_2 = \begin{pmatrix} \ddots & & & & \\ & 0 & & 1 & \\ & & \ddots & & \\ & 1 & & 0 & \\ & & & & \ddots \end{pmatrix} \tag{3.2}$$

Permutation matrices constitute the second group of orthogonal transformations that lead to non-identifiability in spatial factor models ($\mathbf{P}_2$ in (3.2)). A permutation matrix switches the elements of $\mathbf{w}(\mathbf{s})$ and the corresponding columns of $\mathbf{\Lambda}$ simultaneously, so the distributions of $\bar{\mathbf{\Lambda}}\bar{\mathbf{w}}(\mathbf{s})$ and $\mathbf{\Lambda}\mathbf{w}(\mathbf{s})$ are the same. To address such identifiability issues, we impose some constraints on $\rho_k(\mathbf{h};\boldsymbol{\phi}_k)$, more specifically on $\boldsymbol{\phi}_k$ (Zhang, 2007). For simplicity we consider the exponential correlation function $\rho_k(\mathbf{h};\phi_k) = \exp(-\phi_k\|\mathbf{h}\|)$, which has a spatial decay parameter ϕ_k as the only unknown parameter. We require the range parameter ϕ_k, $k = 1,\ldots,r$, to be ordered as $\phi_1 < \phi_2 \cdots < \phi_r$ or $\phi_1 > \phi_2 \cdots > \phi_r$. Simulation studies (see Sections 3.4.1 and 3.4.2) reveal that without this constraint, parameter estimation becomes problematic.

3.2.3 Adaptive Bayesian Factor Model

Determining the number of spatial factors (r) is challenging due to the lack of rigorous theoretical results that hint at the data's ability to inform about r. Generally, previous work (Webster et al., 1994) employing the LMC assumes $r < m$ is fixed. The choice of the number of spatial processes and their respective scales is a critical point in geostatistical models. In the applications of LMC to multivariate spatial analysis (e.g., Webster et al., 1994; Castrignanó et al., 2000, 2005; Buttafuoco et al., 2010), the spatial correlation functions $\rho_k(\mathbf{h};\boldsymbol{\phi}_k)$, the parameters $\boldsymbol{\phi}_k$ and r are obtained from empirical estimates of the auto and cross-variograms prior to any modeling. This approach ignores the uncertainty in the estimates of the spatial parameters and may yield dubious inference.

We adapt earlier work by Kuo and Mallick (1998) and Chen and Dunson (2003) to propose an approach for selecting spatial processes corresponding to different spatial

correlation parameters using a Bayesian hierarchical model. Indicator variables $\boldsymbol{\delta} = \{\delta_1, \ldots, \delta_r\}$, where each δ_k is supported at two points 1 and 0, are introduced in model (3.1) to yield:

$$\mathbf{Y}(\mathbf{s}) = \mathbf{X}(\mathbf{s})'\boldsymbol{\beta} + \sum_{k=1}^{r} \delta_k w_k(\mathbf{s})\boldsymbol{\lambda}_k + \boldsymbol{\epsilon}(\mathbf{s}). \tag{3.3}$$

When $\delta_k = 1$, we include the k-th spatial random process $w_k(\mathbf{s})$ and the corresponding $\boldsymbol{\lambda}_k$ in the model. Otherwise, we omit the k-th spatial scale. This yields 2^r submodels. We call (3.3) the *adaptive Bayesian factor model*, in which the unknown parameters are estimated together with which factors will be retained.

3.2.4 Prior Specification

We place a multivariate normal prior on the slope parameter $\boldsymbol{\beta}$ with mean $\boldsymbol{\mu}_{\boldsymbol{\beta}}$ and variance $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$. Often, a flat prior $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} = \mathbf{O}$ is used. Each diagonal element of $\boldsymbol{\Psi}$ is assigned an Inverse Gamma (IG) distribution. With $\psi_i^2 \sim \text{IG}(a, b)$, the prior for $\boldsymbol{\Psi}$ becomes $\pi(\boldsymbol{\Psi} \,|\, a, b) \propto \prod_{i=1}^{m} \left(\psi_i^2\right)^{-a-1} \exp\{-b/\psi_i^2\}$ with hyperparameters a and b. A customary choice is to use $a = 2$, which suggests a distribution with infinite variance and a mean of b (often gleaned from a semivariogram).

For simplicity, we assume independent priors for $\boldsymbol{\delta}$ and $\boldsymbol{\Lambda}$ (Kuo and Mallick, 1998). The indicator variables δ_k, $k = 1, \ldots, r$, are taken a priori to be independent, with $p(\delta_k = 1 \,|\, \omega) = \omega$. We regard ω as unknown and assume that it has a uniform prior on $(0, 1)$. Imposing constraints on $\boldsymbol{\Lambda}$ only requires a minor modification in the derivation of the full conditional distribution. Here, we take independent priors such as $\Lambda_{jk} \sim N(0, c_0^2)$. The first row of $\boldsymbol{\Lambda}$ is restricted to be positive, that is, $\Lambda_{1k} \sim N(0, c_0^2) I(\Lambda_{1k} > 0)$ for $k = 1, \ldots, r$, where $I(\cdot)$ is the indicator function. Berger and Pericchi (2001) suggest caution in the use of such priors in generic hierarchical models because the outcome of the model selection process can be quite sensitive to their vagueness. However, this seems to be less of an issue in dynamic factor models once the loading matrix is made identifiable (Lopes and West, 2004), whereupon inference and model selection are robust to the prior specifications. Similar results are obtained in random effect selection models (Chen and Dunson, 2003; Cai and Dunson, 2006). In our current context, restricting the first row of $\boldsymbol{\Lambda}$ to be positive, hence assigning a truncated normal prior to each element of the first row of $\boldsymbol{\Lambda}$, ensures robust model selection and related inference to c_0^2.

One also needs to assign priors on $\boldsymbol{\phi} = \{\boldsymbol{\phi}_k\}_{k=1}^r$. The prior for $\boldsymbol{\phi}$ depends upon the choice of correlation functions. Quite remarkably, the spatial process parameters are not consistently estimable and the effect of the prior does not disappear with increasing amounts of data (Zhang, 2004). Hence, prior information becomes an even more delicate issue. Typically, we set prior distributions for the range parameters relative to the size of their domains. In this chapter, an exponential correlation function is used and the prior for the range parameter is specified on a support with upper and lower limits denoted as $\phi_u = -\log(0.01)/d_{min}$ and $\phi_l = -\log(0.05)/d_{max}$, where d_{min} and d_{max} are the minimum and maximum distances across all the locations (Wang and Wall, 2003). Due to identifiability issues discussed in Section 3.2.2, we construct a joint distribution for the ϕ_k's to ensure ordering. In particular, we set

$$\pi(\boldsymbol{\phi}) = \pi(\phi_1)\pi(\phi_2 \,|\, \phi_1) \cdots \pi(\phi_k \,|\, \phi_{k-1}, \ldots, \phi_1) \cdots \pi(\phi_r \,|\, \phi_{r-1}, \ldots, \phi_1),$$

where $\pi(\phi_1)$ is a uniform density with support (ϕ_l, ϕ_u) and subsequently, for $k = 2, 3, \ldots, r$,

$$\pi(\phi_k \,|\, \phi_{k-1}, \ldots, \phi_1) \propto \exp\left\{-\frac{c_k}{\phi_k - \phi_{k-1}}\right\} I(\phi_u > \phi_k > \phi_{k-1}) \,. \tag{3.4}$$

Here the hyperparameters $c_k > 0$ control the shape of the distribution and the separation of the ϕ_k's. In all our subsequent analyses, we fix $c_k = 2k$ as a reasonable choice that delivers robust inference. For any finite domain this prior is proper (i.e. integrable). To offer some additional insight, let $\phi_k - \phi_{k-1} \in (0, hc_k)$ and $h = 4$. Numerical integration yields $p(\phi_k - \phi_{k-1} < c_k/2 \,|\, \phi_{k-1}) < 0.01$. This implies that ϕ_k is unlikely to appear in $(\phi_{k-1}, \phi_{k-1} + c_k/2)$ unless there is a strong mode from the likelihood. So (3.4) can efficiently separate the ϕ_k's. On the other hand, $\dfrac{p(3c_k \le \phi_k - \phi_{k-1} < 4c_k \,|\, \phi_{k-1})}{p(2c_k \le \phi_k - \phi_{k-1} < 3c_k \,|\, \phi_{k-1})} = 1.1$, which indicates that $\pi(\phi_k \,|\, \phi_{k-1})$ becomes informative when ϕ_{k-1} and ϕ_k are away from each other.

An alternative specification is $\pi(\boldsymbol{\phi}) \propto I(\phi_l < \phi_1) I(\phi_1 < \phi_2) \cdots I(\phi_r < \phi_u)$. This is somewhat simpler than (3.4), but encounters problems in practical implementation. Here, the posteriors for ϕ_k and ϕ_{k+1}, while they theoretically obey $\phi_k < \phi_{k+1}$, can become very close to each other. In that case, one of the two latent processes becomes redundant, yet stochastic factor selection keeps both processes. The specification in (3.4) avoids this situation.

3.3 Predictive Process Factor Models

While FA is a powerful tool for summarizing multivariate outcomes and conducting dimension reduction on the number of outcomes, the spatial FA is still prohibitive with a large number of locations. A popular model-based approach for dimension reduction over space uses low-rank or fixed-rank representations for $\mathbf{w}(\mathbf{s})$. Their likelihood resembles linear-mixed models and can be estimated using standard algorithms with some minor adaptations.

Here we consider one such representation that projects the spatially associated latent factors $\mathbf{w}(\mathbf{s})$ onto a lower-dimensional subspace determined by a partial realization of the process $\mathbf{w}(\mathbf{s})$ over a manageable set of locations called "knots". Unlike several other low-rank methods, the predictive process does not introduce additional parameters or kernels. Also, some approaches require empirical estimates of the data's covariance structure. This may be challenging here as the variability in the "data" is assumed to be a sum of unobserved factors. Therefore, we cannot use variograms on the data to isolate empirical estimates of its spatial covariance structure from that of the individual factors.

3.3.1 Model Construction

A Gaussian predictive process (see, e.g., Banerjee et al., 2008; Finley et al., 2009; Banerjee et al., 2010) uses a set of fixed "knots" $\mathcal{S}^* = \{\mathbf{s}_1^*, \ldots, \mathbf{s}_{n^*}^*\}$, $n^* \leq n$, which are usually fixed and may, but need not, form a subset of the observed locations $\mathcal{S} = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\}$. The Gaussian process defined in Section 3.2.1 implies that $\mathbf{w}_k^* = (w_k(\mathbf{s}_1^*), \ldots, w_k(\mathbf{s}_{n^*}^*))'$ follows $N_{n^*}(\mathbf{0}, \mathbf{D}_k^*(\boldsymbol{\phi}_k))$, where $\mathbf{D}_k^*(\boldsymbol{\phi}_k)$ is the $n^* \times n^*$ covariance matrix whose (i,j)-th element is $\rho_k(\mathbf{s}_i^*, \mathbf{s}_j^*; \boldsymbol{\phi}_k)$. The spatial interpolation or "kriging" function at a site $\mathbf{s}_0$ is given by $\tilde{w}_k(\mathbf{s}_0) = \mathrm{E}[w_k(\mathbf{s}_0) \,|\, \mathbf{w}_k^*] = \mathbf{d}_k(\mathbf{s}_0; \boldsymbol{\phi}_k)' \mathbf{D}_k^*(\boldsymbol{\phi}_k)^{-1} \mathbf{w}_k^*$, where $\mathbf{d}_k(\mathbf{s}_0; \boldsymbol{\phi}_k)$ is an $n^* \times 1$ vector whose i-th element is $\rho_k(\mathbf{s}_0, \mathbf{s}_i^*; \boldsymbol{\phi}_k)$. This defines the predictive process $\tilde{w}_k(\mathbf{s}) \sim GP(0, \tilde{\rho}_k(\cdot; \boldsymbol{\phi}_k))$ derived from the parent process $w_k(\mathbf{s})$, where $\tilde{\rho}_k(\mathbf{s}_i, \mathbf{s}_j; \boldsymbol{\phi}_k) = \mathbf{d}_k(\mathbf{s}_i; \boldsymbol{\phi}_k)' \mathbf{D}_k^*(\boldsymbol{\phi}_k)^{-1} \mathbf{d}_k(\mathbf{s}_j; \boldsymbol{\phi}_k)$.

The predictive process underestimates the variance of the parent process $w_k(\mathbf{s}_0)$ at any location $\mathbf{s}_0$ because $\mathrm{var}\{w_k(\mathbf{s}_0)\} - \mathrm{var}\{\tilde{w}_k(\mathbf{s}_0)\} = \mathrm{var}\{w_k(\mathbf{s}_0) - \tilde{w}_k(\mathbf{s}_0)\} \geq 0$. This

veracity is an immediate consequence of the definition of $\tilde{w}_k(\mathbf{s})$ as a conditional expectation. This means that the estimated $\boldsymbol{\Psi}$ from the predictive process model must roughly capture the same amount of variability as the estimated $\boldsymbol{\Psi} + \boldsymbol{\Lambda}\left(\mathbf{I}_r - \mathrm{E}[\tilde{\boldsymbol{\Gamma}}(\mathbf{s}_0)]\right)\boldsymbol{\Lambda}'$ from (3.1), where $\tilde{\boldsymbol{\Gamma}}(\mathbf{s}_0)$ is an $r \times r$ diagonal matrix with k-th diagonal element $\tilde{\rho}_k(\mathbf{s}_0, \mathbf{s}_0; \boldsymbol{\phi}_k)$.

Defining $f_k(\mathbf{s}) \,|\, \tilde{w}_k(\mathbf{s}) \sim N(\tilde{w}_k(\mathbf{s}), \sigma^2_{f_k}(\mathbf{s}))$, where $\sigma^2_{f_k}(\mathbf{s}) = 1 - \tilde{\rho}_k(\mathbf{s}, \mathbf{s}, \boldsymbol{\phi}_k)$, we easily obtain $\mathrm{var}\{f_k(\mathbf{s})\} = \mathrm{var}\{w_k(\mathbf{s})\} = 1$, as desired. Now, replace $\mathbf{w}(\mathbf{s})$ with $\mathbf{f}(\mathbf{s}) = (f_1(\mathbf{s}), \ldots, f_r(\mathbf{s}))'$ in (3.3) and define $\mathbf{f}_k = (f_k(\mathbf{s}_1), \ldots, f_k(\mathbf{s}_n))'$, $\mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_r)$ and $\mathbf{w}^* = (\mathbf{w}(\mathbf{s}_1^*)', \ldots, \mathbf{w}(\mathbf{s}_{n^*}^*)')'$. This yields a posterior distribution for all the parameters and latent factors $p(\mathbf{F}, \mathbf{w}^*, \boldsymbol{\beta}, \boldsymbol{\Psi}, \boldsymbol{\phi}, \boldsymbol{\Lambda}, \boldsymbol{\delta}, \omega \,|\, \mathbf{Y})$ proportional to

$$\begin{aligned}
&\pi(\omega) \times \pi(\boldsymbol{\phi}) \times \prod_{k=1}^{r} Ber(\delta_k \,|\, \omega) \times N_m(\boldsymbol{\lambda}_k \,|\, \mathbf{0}, c_0^2\mathbf{I}_m)\mathbf{I}(\boldsymbol{\Lambda}_{1k} > 0) \times N_p(\boldsymbol{\beta} \,|\, \boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}) \\
&\quad \times \prod_{j=1}^{m} IG(\psi_j^2 \,|\, a, b) \times \prod_{k=1}^{r} N_n\left(\mathbf{f}_k \,|\, \mathbf{D}_k(\boldsymbol{\phi}_k)\mathbf{D}_k^*(\boldsymbol{\phi}_k)^{-1}\mathbf{w}_k^*, \boldsymbol{\Sigma}_{\mathbf{f}_k}(\boldsymbol{\phi}_k)\right) \qquad (3.5) \\
&\quad \times \prod_{k=1}^{r} N_{n^*}(\mathbf{w}_k^* \,|\, \mathbf{0}, \mathbf{D}_k^*(\boldsymbol{\phi}_k)) \times \prod_{i=1}^{n} N_m\left(\mathbf{Y}(\mathbf{s}_i) \,|\, \mathbf{X}(\mathbf{s}_i)'\boldsymbol{\beta} + \sum_{k=1}^{r} \delta_k f_k(\mathbf{s}_i)\boldsymbol{\lambda}_k, \boldsymbol{\Psi}\right),
\end{aligned}$$

where $\mathbf{D}_k(\boldsymbol{\phi}_k) = (\mathbf{d}_k(\mathbf{s}_1), \ldots, \mathbf{d}_k(\mathbf{s}_n))'$ and $\boldsymbol{\Sigma}_{\mathbf{f}_k}$ is an $n \times n$ diagonal matrix with i-th diagonal element $\sigma^2_{f_k}(\mathbf{s}_i)$. $\pi(\omega)$ is a uniformly distribution on (0, 1) and $\pi(\boldsymbol{\phi})$ is defined in (3.4). Estimation of (3.5) proceeds using MCMC sampling (see Appendix C for details).

We offer some remarks on knot selection. With evenly distributed locations, knots on a uniform grid (see, e.g., Diggle and Lophaven, 2006) may suffice. With irregular locations, space-covering designs (e.g., Royle and Nychka, 1998) yield a more representative set. In general, different knots selection methods still produce robust results in predictive process models (see, e.g., Finley et al., 2009; Banerjee et al., 2010). And the modified predictive process model $f_k(\mathbf{s})$ is even less sensitive than $\tilde{w}_k(\mathbf{s})$ to such choices. Here, we use the K-means clustering algorithm (Hartigan and Wong, 1979) to arrive at a set of knots.

3.3.2 Handling Missing Observations

Missing outcomes arise frequently in many environmental applications. The outcomes may be missing for a variety of unintended reasons: non-response, equipment failure,

lack of collection and so on. Bayesian analysis often proceeds from data augmentation (DA) (Tanner, 1993) or multiple imputation (MI) (Rubin, 1976) to handle the incomplete data problem. Both methods impute the missing values within a Gibbs step and then use the complete data set. Instead, here we condition on the latent factors, which simplifies the MCMC algorithm and reduces the computational burden. Conditional on the latent factor $\mathbf{f}(\mathbf{s})$, the outcomes are independent of each other, which yields a likelihood depending only upon observed data. The missing values can, then, be recovered from the posterior predictive distribution.

Let $\mathbf{Y}(\mathbf{s})$ be the $m \times 1$ vector of measured and unmeasured outcomes at site $\mathbf{s}$. Suppose the measured and unmeasured elements of $\mathbf{Y}(\mathbf{s})$ have indices $i_1, \ldots, i_{d_\mathbf{s}}$ and $i_{d_\mathbf{s}+1}, \ldots, i_m$ respectively, where $d_\mathbf{s}$ is the number of observed outcomes at $\mathbf{s}$. Let $\mathbf{v}_j, j = 1, \ldots, m$, be an $m \times 1$ vector whose j-th element is 1 and the rest are all 0. We can then construct matrices $\mathbf{R}_1(\mathbf{s})_{d_\mathbf{s} \times m}$ and $\mathbf{R}_2(\mathbf{s})_{(m-d_\mathbf{s}) \times m}$, which can extract the observed and unobserved elements when multiplied by $\mathbf{Y}(\mathbf{s})$. More precisely, $\mathbf{R}_1(\mathbf{s}) = (\mathbf{v}_{i_1}, \ldots, \mathbf{v}_{i_{d_\mathbf{s}}})'$ and $\mathbf{R}_2(\mathbf{s}) = (\mathbf{v}_{i_{d_\mathbf{s}+1}}, \ldots, \mathbf{v}_{i_m})'$.

Observe that $\mathbf{Y}_o(\mathbf{s}) = \mathbf{R}_1(\mathbf{s})\mathbf{Y}(\mathbf{s})$ and $\mathbf{Y}_u(\mathbf{s}) = \mathbf{R}_2(\mathbf{s})\mathbf{Y}(\mathbf{s})$ consist of the observed and unobserved elements respectively. Multiplying both sides of model (3.3) by $\mathbf{R}_1(\mathbf{s})$ shows that the likelihood $p(\mathbf{Y}_o(\mathbf{s}) \,|\, \mathbf{f}(\mathbf{s}), \boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\Lambda}, \boldsymbol{\Psi})$ follows a multivariate normal distribution with variance $\mathbf{R}_1(\mathbf{s})\boldsymbol{\Psi}\mathbf{R}_1(\mathbf{s})'$ and mean $\mathbf{R}_1(\mathbf{s})\left[\mathbf{X}(\mathbf{s})'\boldsymbol{\beta} + \sum_{k=1}^{r} \delta_k f_k(\mathbf{s})\boldsymbol{\lambda}_k\right]$. Notice that $\mathbf{R}_1(\mathbf{s})\boldsymbol{\Psi}\mathbf{R}_1(\mathbf{s})'$ is a $d_\mathbf{s} \times d_\mathbf{s}$ diagonal matrix with l-th diagonal element $\psi_{i_l}^2$ for $l = 1, \ldots, d_\mathbf{s}$. This ensures that $\mathbf{R}_1(\mathbf{s})\boldsymbol{\Psi}\mathbf{R}_1(\mathbf{s})'$ is a valid covariance matrix, and so is $\mathbf{R}_2(\mathbf{s})\boldsymbol{\Psi}\mathbf{R}_2(\mathbf{s})'$ for similar reasons. See Appendix C for implementation details.

3.3.3 Prediction and Predictive Model Comparison

Spatial analysis for datasets with missing values customarily entail three types of inference: (i) recover missing values, (ii) predict outcome $\mathbf{Y}(\mathbf{t}) = (Y_1(\mathbf{t}), \ldots, Y_m(\mathbf{t}))'$ at any arbitrary location $\mathbf{t}$, and (iii) generate replicates for the observed data for use in model assessment and selection (Gelfand and Ghosh, 1998).

To recover the missing values, we use composition sampling to draw $\mathbf{Y}_u$ from $p(\mathbf{Y}_u \,|\, \mathbf{Y}_o)$, where $\mathbf{Y}_o = (\mathbf{Y}_o(\mathbf{s}_1)', \ldots, \mathbf{Y}_o(\mathbf{s}_n)')'$ and $\mathbf{Y}_u = (\mathbf{Y}_u(\mathbf{s}_1)', \ldots, \mathbf{Y}_u(\mathbf{s}_n)')'$. Note

that $p(\mathbf{Y}_u \,|\, \mathbf{Y}_o)$ equals

$$\int \prod_{i=1}^{n} p(\mathbf{Y}_u(\mathbf{s}_i) \,|\, \mathbf{f}(\mathbf{s}_i), \boldsymbol{\theta}) p(\mathbf{F}, \boldsymbol{\theta} \,|\, \mathbf{Y}_o) \mathrm{d}\mathbf{F} \mathrm{d}\boldsymbol{\theta}$$
$$= \int \prod_{i=1}^{n} N\left(\mathbf{Y}_u(\mathbf{s}_i) \,|\, \mathbf{R}_2(\mathbf{s}_i)[\mathbf{X}(\mathbf{s}_i)'\boldsymbol{\beta} + \tilde{\boldsymbol{\Lambda}}\mathbf{f}(\mathbf{s}_i)], \mathbf{R}_2(\mathbf{s}_i)\boldsymbol{\Psi}\mathbf{R}_2(\mathbf{s}_i)'\right) p(\mathbf{F}, \boldsymbol{\theta} \,|\, \mathbf{Y}_o) \mathrm{d}\mathbf{F} \mathrm{d}\boldsymbol{\theta},$$

where $\tilde{\boldsymbol{\Lambda}} = (\delta_1 \boldsymbol{\lambda}_1, \ldots, \delta_r \boldsymbol{\lambda}_r)$. Using the samples $\{\mathbf{F}^{(l)}, \boldsymbol{\theta}^{(l)}\}$ from $p(\mathbf{F}, \boldsymbol{\theta} \,|\, \mathbf{Y}_o)$, the missing value at $\mathbf{s}$ is recovered by drawing $\mathbf{Y}_u^{(l)}(\mathbf{s})$ for each, $l = 1, \ldots, L$, from an $(m - d_{\mathbf{s}}) \times 1$ multivariate normal distribution with mean $\mathbf{R}_2(\mathbf{s})\left[\mathbf{X}(\mathbf{s})'\boldsymbol{\beta}^{(l)} + \tilde{\boldsymbol{\Lambda}}^{(l)}\mathbf{f}^{(l)}(\mathbf{s})\right]$ and variance $\mathbf{R}_2(\mathbf{s})\boldsymbol{\Psi}^{(l)}\mathbf{R}_2(\mathbf{s})'$.

Prediction of the outcomes at unsampled or ungauged sites is often a major study objective. Using notation defined earlier, we have

$$p(\mathbf{Y}(\mathbf{t}) \,|\, \mathbf{Y}_o) = \int p(\mathbf{Y}(\mathbf{t}) \,|\, \mathbf{f}(\mathbf{t}), \boldsymbol{\theta}) p(\mathbf{f}(\mathbf{t}) \,|\, \mathbf{F}, \boldsymbol{\phi}, \mathbf{Y}_o) p(\mathbf{F}, \boldsymbol{\theta} \,|\, \mathbf{Y}_o) \mathrm{d}\mathbf{F} \mathrm{d}\boldsymbol{\theta} \mathrm{d}\mathbf{f}(\mathbf{t})$$
$$= \int \begin{array}{l} N\left(\mathbf{Y}(\mathbf{t}) \,|\, \mathbf{X}(\mathbf{t})'\boldsymbol{\beta} + \tilde{\boldsymbol{\Lambda}}\mathbf{f}(\mathbf{t}), \boldsymbol{\Psi}\right) \prod_{k=1}^{r} N\left(f_k(\mathbf{t}) \,|\, \mu_{f_k(\mathbf{t}) \,|\, \mathbf{f}_k}, \sigma^2_{f_k(\mathbf{t}) \,|\, \mathbf{f}_k}\right) \\ \quad \times p(\mathbf{F}, \boldsymbol{\theta} \,|\, \mathbf{Y}_o) \mathrm{d}\mathbf{F} \mathrm{d}\boldsymbol{\theta} \mathrm{d}\mathbf{f}(\mathbf{t}) \end{array}$$

where $\sigma^2_{f_k(\mathbf{t}) \,|\, \mathbf{f}_k} = 1 - \mathbf{d}_k(\mathbf{t}; \boldsymbol{\phi}_k)'\mathbf{D}_k^*(\boldsymbol{\phi}_k)^{-1}\mathbf{D}_k(\boldsymbol{\phi}_k)'\mathbf{G}_k(\boldsymbol{\phi}_k)^{-1}\mathbf{D}_k(\boldsymbol{\phi}_k)\mathbf{D}_k^*(\boldsymbol{\phi}_k)^{-1}\mathbf{d}_k(\mathbf{t}; \boldsymbol{\phi}_k)$ and $\mu_{f_k(\mathbf{t}) \,|\, \mathbf{f}_k} = \mathbf{d}_k(\mathbf{t}; \boldsymbol{\phi}_k)'\mathbf{D}_k^*(\boldsymbol{\phi}_k)^{-1}\mathbf{D}_k(\boldsymbol{\phi}_k)'\mathbf{G}_k(\boldsymbol{\phi}_k)^{-1}\mathbf{f}_k$. Here, $\mathbf{G}_k(\boldsymbol{\phi}_k)$ is the variance of $\mathbf{f}_k$, which equals $\mathbf{D}_k(\boldsymbol{\phi}_k)\mathbf{D}_k^*(\boldsymbol{\phi}_k)^{-1}\mathbf{D}_k(\boldsymbol{\phi}_k)' + \boldsymbol{\Sigma}_{\mathbf{f}_k}$. For prediction, we would first sample $f_k(\mathbf{t})^{(l)}$ from $N\left(f_k(\mathbf{t}) \,|\, \mu_{f_k(\mathbf{t}) \,|\, \mathbf{f}_k}, \sigma^2_{f_k(\mathbf{t}) \,|\, \mathbf{f}_k}\right)$ using posterior samples $\mathbf{f}_k^{(l)}$ and $\boldsymbol{\phi}_k^{(l)}$. Then, $\mathbf{Y}(\mathbf{t})^{(l)}$ is drawn from an $m \times 1$ multivariate normal distribution with mean $\mathbf{X}(\mathbf{t})'\boldsymbol{\beta}^{(l)} + \tilde{\boldsymbol{\Lambda}}^{(l)}\mathbf{f}(\mathbf{t})^{(l)}$ and variance $\boldsymbol{\Psi}^{(l)}$.

Finally we turn to generating replicated observations for $\mathbf{Y}_o$. This is easily achieved by first sampling from $p(\mathbf{w}^*, \boldsymbol{\theta} \,|\, \mathbf{Y}_o)$. The replicate $\mathbf{Y}_o^{rep}(\mathbf{s})^{(l)}$ is generated from a $d_{\mathbf{s}} \times 1$ multivariate normal distribution with mean $\mathbf{R}_1(\mathbf{s})\left[\mathbf{X}(\mathbf{s})'\boldsymbol{\beta}^{(l)} + \tilde{\boldsymbol{\Lambda}}^{(l)}\tilde{\mathbf{w}}^{(l)}(\mathbf{s})\right]$ and variance $\mathbf{R}_1(\mathbf{s})\left[\boldsymbol{\Psi}^{(l)} + \tilde{\boldsymbol{\Lambda}}^{(l)}\boldsymbol{\Sigma}_{\mathbf{f}}(\mathbf{s})(\boldsymbol{\phi}^{(l)})\tilde{\boldsymbol{\Lambda}}^{(l)'}\right]\mathbf{R}_1(\mathbf{s})'$, where $\boldsymbol{\Sigma}_{\mathbf{f}}(\mathbf{s})$ is a diagonal matrix with k-th diagonal element $\sigma^2_{f_k}(\mathbf{s})$. Then $\mu_i^{rep} = \mathrm{E}[Y_{oi}^{rep} \,|\, \mathbf{Y}_o]$ and $\sigma_i^{2rep} = \mathrm{var}\{Y_{oi}^{rep} \,|\, \mathbf{Y}_o\}$ can be calculated, where i indexes the observed data. Gelfand and Ghosh (1998) present a posterior predictive criterion that balances goodness of fit and predictive variance under a squared error loss function. This assigns a score to each model that is the sum of two terms, P and G, where $G = \sum_i (\mu_i^{rep} - Y_{oi})^2$ is an error sum of squares and represents

goodness-of-fit, while $P = \sum_i \sigma_i^{2rep}$ represents predictive variance and acts as a penalty term. The model with smaller $G + P$ value is preferred.

3.4 Illustrations

We executed our adaptive spatial factor models using the `R` programming language. The most demanding model (12×1 vector of outcomes across 1000 locations) took approximately 48 hours to deliver its entire inferential output from $20,000$ MCMC iterations, including $4,000$ samples for burn-in, on a 3.10-GHz Intel i5-2400 processor with 4.0 Gbytes of RAM. The statistics of Gelman and Rubin (1992) was used to assess chain convergence alongside visual inspection of the trace plots and empirical autocorrelation functions.

3.4.1 Simulation Study One

The objective of this simulation study is to explore different specifications for $\mathbf{\Lambda}$ and demonstrate model identifiability. Adaptive factor modeling and predictive processes are not employed here. The data generating model uses a 2×2 factor loading matrix $\mathbf{\Lambda}$ and a spatial range parameter $\boldsymbol{\phi}$ whose first element is smaller than the second. A synthetic data set comprising 250 locations within a $[0, 50] \times [0, 50]$ square was generated from (3.1) with $r = m = 2$. An isotropic exponential correlation function, $\rho_k(\mathbf{s}-\mathbf{t}; \phi_k) = \exp(-\phi_k \|\mathbf{s}-\mathbf{t}\|)$ for $k = 1, 2$, was used to produce a spatially dependent bivariate random field. The bivariate outcomes, $\mathbf{Y}(\mathbf{s})$, were simulated with the following parameters:

$$\boldsymbol{\beta} = \begin{pmatrix} 5 \\ 10 \end{pmatrix}, \quad \mathbf{\Lambda} = \begin{pmatrix} 3 & 1 \\ -2 & 2 \end{pmatrix}, \quad \mathbf{\Psi} = \begin{pmatrix} 2 & 0 \\ 0 & 5 \end{pmatrix}, \quad \text{and} \quad \boldsymbol{\phi} = \begin{pmatrix} 0.1 \\ 0.6 \end{pmatrix},$$

where $\boldsymbol{\beta}$ is the mean vector for $\mathbf{Y}(\mathbf{s})$, $\mathbf{X}(\mathbf{s})$ is the identity matrix of dimension 2, and $\mathbf{C}(\mathbf{0}) = \mathbf{\Lambda}\mathbf{\Lambda}' = \begin{pmatrix} 10 & -4 \\ -4 & 8 \end{pmatrix}$. Note the negative cross-correlation between the two spatial processes. Also, the correlation decays six times faster in the second process than in the first, yielding spatial ranges $-\log(0.05)/\phi_1 \approx 30$ and $-\log(0.05)/\phi_2 \approx 5$ units for the first and second process respectively.

Three models are estimated and compared. The first restricts $\mathbf{\Lambda}$ to be lower-triangular, as in the LMC setting, but does not order the ϕ_i's apriori. The other

two models use factor loadings with positive elements in the first row of $\mathbf{\Lambda}$ and uses the prior in (3.4) for $\boldsymbol{\phi}$. One of these two models assumes $\phi_1 < \phi_2$, reflecting the true ordering, while the other reverses the inequality and results in a misspecified model apriori. However, because the factor loadings are now identifiable, the estimated $\mathbf{\Lambda}$ contains permuted columns according to the order of the ϕ_i's. This is attractive in practice – one need not ascertain the correct ordering of the range parameters apriori. We arrange them in the same order as the true values while presenting estimates in Table 3.1.

In the first model $\Lambda_{21} = 0$, ϕ_i's are assigned uniform priors over (ϕ_l, ϕ_u) without any further restrictions. Priors for the two other models are assigned as discussed in Section 3.2.4 with hyper-parameters $\boldsymbol{\mu}_{\boldsymbol{\beta}} = \mathbf{0}$, $\mathbf{\Sigma}_{\boldsymbol{\beta}}^{-1} = \mathbf{O}$, $c_0^2 = 10$, $\phi_l = 0.047$, $\phi_u = 25$, $a = 2$ and $b = 5$. Posterior inferences for the latter two models are consistent. This indicates the importance of ordering constraints; without them, the posterior distributions for $\boldsymbol{\phi}$ and $\mathbf{\Lambda}$ can produce multiple modes as the MCMC chain could jump between the possible specifications (Lopes and West, 2004). The parameter Λ_{21} (boldface in Table 3.1), which was set as 0 in the lower triangular setting, is significantly different from 0 in both the models using general factor loadings. Thus, we can capture the underlying cross-covariance structure without restricting the loading matrix, but by imposing an ordering on the spatial decay parameters. In addition, our proposed models have a lower posterior predictive model comparison score ($D = 14,202$) than the lower triangular model ($D = 14,941$). This, again, demonstrates a preference for using general factor loadings.

3.4.2 Simulation Study Two

Now we demonstrate the adaptive spatial factor model using a simulated data set with missing values. We generated 1000 locations in a $[0, 30] \times [0, 30]$ square. At each location, we simulated a 12×1 vector of outcomes, $\mathbf{Y}(\mathbf{s})$, from (3.1) with three factors and outcome-specific intercepts as the only regressors. Each spatial process, $w_k(\mathbf{s})$, was generated from an isotropic exponential correlation function with decay parameter $\boldsymbol{\phi} = \{\phi_1, \phi_2, \phi_3\} = \{0.1, 0.8, 1.5\}$ respectively. After the data was generated, we allowed a quarter of the locations to retain all the outcomes, while randomly omitting 2/3 of the outcomes from the remaining locations. This produced a dataset with approximately 50% missingness and $\mathbf{Y}_o$ is a vector of 6000 long.

True Parameter	Lower Triangular Setting	General factor loadings	
		Increasing ϕ_k	Decreasing ϕ_k
$\beta_1 = 5$	4.77 (1.22, 7.49)	5.00 (3.29, 6.98)	5.58 (3.77, 7.85)
$\beta_2 = 10$	10.38 (8.65, 12.73)	10.15 (8.38, 11.60)	9.68 (7.73, 11.16)
$C(\mathbf{0})_{1,1} = 10$	8.59 (4.76, 15.08)	10.82 (6.53, 18.54)	10.65 (6.23, 17.66)
$C(\mathbf{0})_{2,1} = -4$	-5.64 (-9.99, -2.93)	-5.34 (-10.84, -1.82)	-5.11 (-9.88, -1.67)
$C(\mathbf{0})_{2,2} = 8$	12.17 (6.76, 17.54)	12.99 (7.15, 19.43)	12.83 (7.15, 18.42)
$\Psi_1 = 2$	3.65 (2.61, 4.88)	2.30 (1.19, 3.66)	2.35 (1.15, 3.55)
$\Psi_2 = 5$	3.27 (0.97, 8.38)	3.33 (1.06, 7.99)	3.21 (0.99, 7.99)
$\phi_1 = 0.1$	0.06 (0.05, 0.12)	0.07 (0.05, 0.13)	0.07 (0.05, 0.14)
$\phi_2 = 0.6$	0.95 (0.25, 2.15)	1.17 (0.52, 2.58)	1.15 (0.49, 2.44)
$\Lambda_{1,1} = 3$	2.93 (2.18, 3.88)	3.11 (2.32, 4.15)	3.07 (2.31, 4.06)
$\Lambda_{1,2} = -2$	-1.93 (-2.80, -1.17)	-2.53 (-3.42, -1.79)	-2.48 (-3.33, -1.68)
$\Lambda_{2,1} = 1$	**Fixed at 0**	**1.02 (0.56, 1.68)**	**1.01 (0.55, 1.78)**
$\Lambda_{2,2} = 2$	2.89 (1.84, 3.47)	2.57 (1.33, 3.22)	2.60 (1.34, 3.23)

Table 3.1: Posterior percentiles (50%, 2.5% and 97.5%) estimated for the parameters in different model specifications.

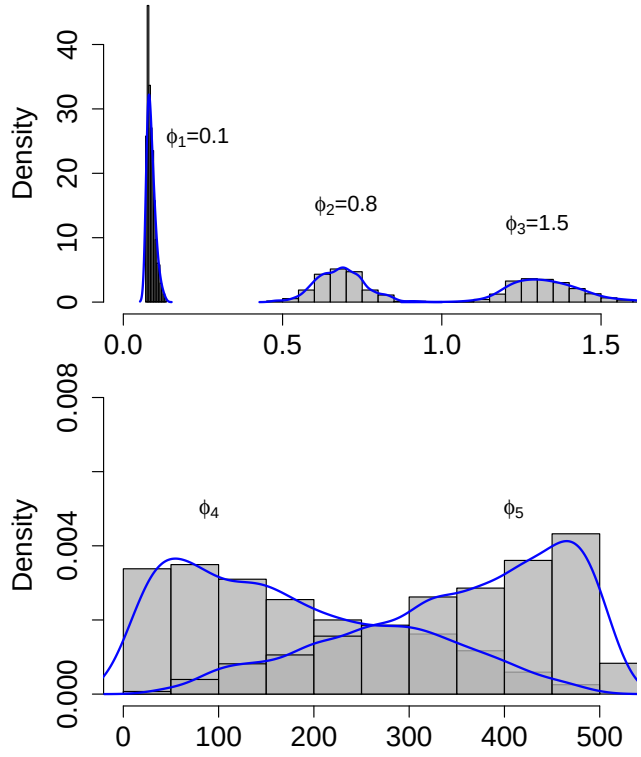


Figure 3.1: Histograms of the posterior distributions for $\boldsymbol{\phi}$.

We applied the Gibbs sampling algorithm described in Section 3.3 for estimation. For the predictive process, we selected 200 knots using a K-means clustering algorithm (Hartigan and Wong, 1979). The maximum number of latent factors, r, was taken to be 5. For comparison, we also fit spatial factor models using a fixed number of latent factors for the same data set. The prior distributions resemble those in Section 3.4.1, except that now $\phi_l = 0.07$ and $\phi_u = 678$ to better reflect the domain. Also, for the adaptive model we used independent Bernoulli priors for $\boldsymbol{\delta}$ with $p(\delta_k = 1 \,|\, \omega) = \omega$, $k = 1, \ldots, r$ and $\omega \sim U(0, 1)$.

Stochastic selection produced the final model with three latent factors based upon the highest posterior probabilities. Figure 3.1 illustrates the posterior distributions for $\boldsymbol{\phi}$. The spatial parameters in the active processes are ϕ_1, ϕ_2, ϕ_3 (corresponding to $\delta_k = 1$) and their posteriors are well identified. The true values are all included

in the central 95% credible intervals. The posteriors for ϕ_4 and ϕ_5 only reflect prior information. The wide range of their posteriors suggest that there is very little posterior learning and the MCMC sampler is sampling from the prior. The marginal posterior density for ω reveals substantial posterior learning shown in Figure 3.2.

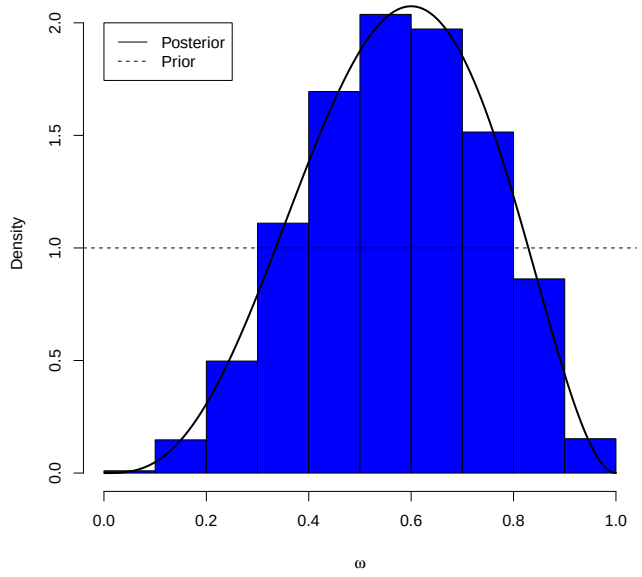


Figure 3.2: Histograms of the posterior distributions for ω.

	Number of Factors				
Criterion	$r = 1$	$r = 2$	Adaptive Factor	$r = 4$	$r = 5$
G	69567	53893	45982	46011	46036
P	71698	60173	53758	53880	53939
D	141265	114066	**99740**	99891	99975

Table 3.2: Simulation study two: Model comparison criteria.

Spatial factor models with fixed number of latent factors equaling 1, 2, 4 and 5 are

also fitted. The posterior predictive criterion discussed in Section 3.3.3 is presented for model assessment and the scores are compared in Table 3.2. The adaptive spatial factor model with $D = 99740$ (boldface in Table 3.2) seems to be the winner here. There is overwhelming evidence that the regression model with random processes $f_1(\mathbf{s})$, $f_2(\mathbf{s})$ and $f_3(\mathbf{s})$ is optimal. This, pleasantly, agrees with the true specification.

3.4.3 Air Monitor Value Data

We illustrate the adaptive spatial factor model with an air pollutant data obtained from Environmental Protection Agency (EPA) databases. This data set comprises five pollutants: carbon monoxide (CO), nitrogen dioxide (NO_2), ozone (O_3), particulate matter with diameter < 2.5 micrometers (PM2.5) and particulate matter with diameter < 10 micrometers (PM10). Concentrations (ppm) measured by monitoring equipments are used for CO, NO_2 and O_3 and (μg/m^3) for PM2.5 and PM10. Measurements were collected from 316 air monitors in California. Here we use the annual average of the monitor values recorded in 2008. All five outcomes were normalized and standardized to mean 0 and variance 1. Each pollutant has its own intercept term. Elevation in kilometers is used as the other predictor and is depicted as a contour image in the top-left panel of Figure 3.3. The instruments only monitor some of the five pollutants at most of the sites, so the data set has about 53% of observations missing. The longitude and latitude of the monitors are transformed to Easting and Northing in kilometer units. For the predictive process, 50 knots were selected using a K-means clustering algorithm. The interpolations for the air pollutants and the locations of monitor sites ($\cdot$) appear in Figure 3.3.

Visually, the air pollutants exhibit spatial varying concentrations as well as the associations between them at a given location. We seek to model both kinds of dependence. It is not clear whether it is appropriate to assume a single factor underlying the different air pollutants or if additional factors need to be introduced. To address this question, we repeated the approach described in Section 3.3 and implemented it in the same manner as in Section 3.4.2. The prior for $\boldsymbol{\phi}$ has the lower and upper limit (0.002, 136). The maximum possible number of factors is taken to be 3. We decide to use 4,000 iterations for a burn-in period, and then a further 16,000 iterations to draw 4000 samples with equal space.

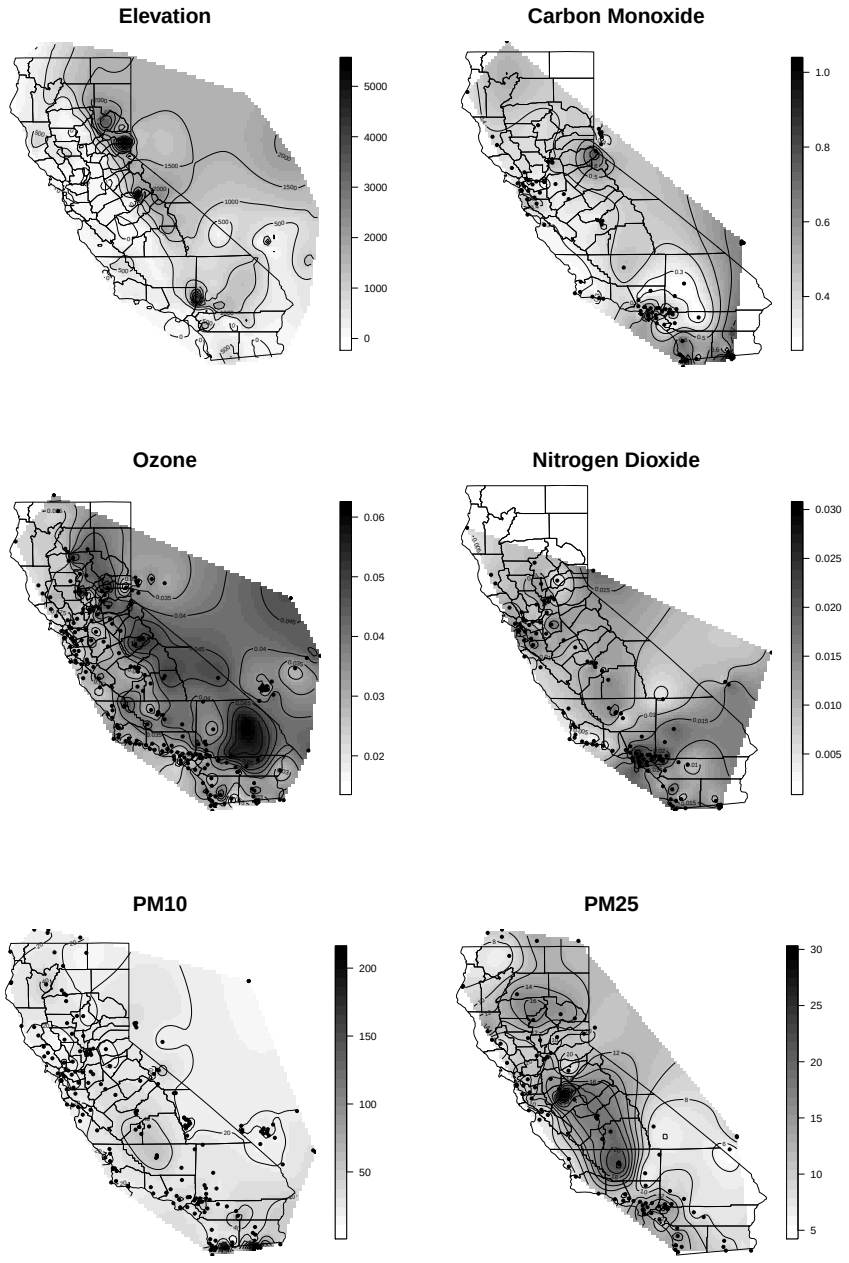


Figure 3.3: Interpolation of air pollutants measured on monitor sites across California.

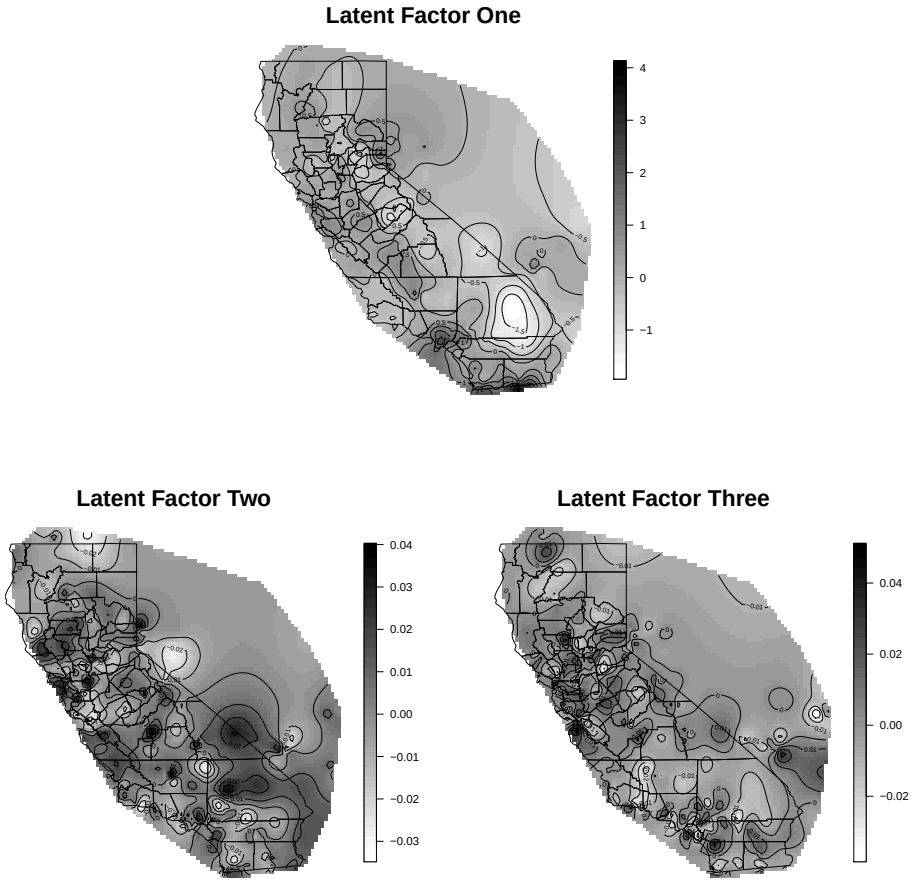


Figure 3.4: Interpolation of latent spatial factors across California.

Parameter	50% (2.5%, 97.5%)	Parameter	50% (2.5%, 97.5%)	Parameter	50% (2.5%, 97.5%)
$\beta_{1,0}$	-0.76 (-1.37, -0.29)	Λ_{11}	0.85, (0.58, 1.39)	$\rho_{1,2}$	0.60 (0.44, 0.80)
$\beta_{2,0}$	-0.61 (-1.23, -0.14)	Λ_{21}	0.88 (0.64, 1.47)	$\rho_{1,3}$	-0.45 (-0.70, -0.27)
$\beta_{3,0}$	0.04 (-0.24, 0.41)	Λ_{31}	-0.57 (-0.98, -0.36)	$\rho_{1,4}$	0.50 (0.33, 0.68)
$\beta_{4,0}$	-0.30 (-0.69, 0.05)	Λ_{41}	0.68, (0.49, 1.02)	$\rho_{1,5}$	0.39 (0.19, 0.64)
$\beta_{5,0}$	-0.27 (-0.72, 0.07)	Λ_{51}	0.55 (0.27, 1.00)	$\rho_{2,3}$	-0.48 (-0.73, -0.28)
$\beta_{1,Elev}$	-0.02 (-0.50, 0.45)	ψ_1^2	0.52 (0.35, 0.77)	$\rho_{2,4}$	0.52, (0.37, 0.70)
$\beta_{2,Elev}$	0.33 (-0.24, 0.91)	ψ_2^2	0.44 (0.30, 0.63)	$\rho_{2,5}$	0.41 (0.20, 0.66)
$\beta_{3,Elev}$	0.40 (0.21, 0.58)	ψ_3^2	0.59 (0.44, 0.79)	$\rho_{3,4}$	-0.39 (-0.60, -0.23)
$\beta_{4,Elev}$	0.02 (-0.28, 0.33)	ψ_4^2	0.63 (0.45, 0.85)	$\rho_{3,5}$	-0.30 (-0.55, -0.14)
$\beta_{5,Elev}$	-0.12 (-0.43, 0.16)	ψ_5^2	0.86 (0.62, 1.19)	$\rho_{4,5}$	0.33 (0.15, 0.55)
ϕ_1	0.018 (0.004, 0.039)				
ϕ_2	43 (5.1, 110)				
ϕ_3	101 (33, 136)				

Table 3.3: The posterior credible intervals estimated for the parameters in Air pollutants data set. The subscripts 1-5 in $\boldsymbol{\beta}$, $\boldsymbol{\Psi}$, $\boldsymbol{\rho}$ and the row index of $\boldsymbol{\lambda}_1$ refer to CO, NO2, O_3, PM10 and PM25 respectively. Subscripts on $\boldsymbol{\phi}$ refer to the three spatial range parameters.

The probability assigned to the model with one latent factor is 1 in all 4000 simulations, suggesting that one factor is sufficient. The interpolation of the latent spatial processes are displayed in Figure 3.4. The panel at the top demonstrates the only active process ($\delta_1 = 1$), which presents some characteristics for the spatial concentrations of the air pollutants. The panels at the bottom do not show significant association with the data and only demonstrate the prior information. For each air pollutant $Y_j(\mathbf{s})$, the percentage of the variance explained by the spatial factor $f_k(\mathbf{s})$, is simply $\delta_k \Lambda_{jk}^2 / \left(\sum_{k=1}^{r} \delta_k \Lambda_{jk}^2 + \psi_j^2 \right)$, $j = 1, \ldots, 5$, which is $(0.58, 0.64, 0.36, 0.42, 0.26)$ in this analysis. Overall, about 45% of the variation is explained by the first latent factor. NO_2 and CO are more closely related to the latent factor, while PM2.5 is weakly explained by it.

Table 3.3 presents the posterior inferences of $\boldsymbol{\beta}$, $\boldsymbol{\Psi}$, $\boldsymbol{\phi}$, $\boldsymbol{\Lambda}$ and the correlation $\boldsymbol{\rho} =$

$\{\rho_{i,j}\}$, $i, j = 1, \ldots, 5$, $j > i$, among air pollutants, which is estimated from $\mathbf{\Lambda\Lambda}' + \mathbf{\Psi}$. Only CO and NO_2 reveal significant (negative) intercept coefficients. The slope parameters seem to suggest that elevation has a significant impact only upon Ozone among the five pollutants. The posterior inference for ϕ_1 has median 0.018, which corresponds to the effective range (i.e., the distance at which the correlation drops to 0.05) roughly 160 kilometers. The posterior inferences for ϕ_2 and ϕ_3 indicate weak spatial dependence. So only the first column of $\mathbf{\Lambda}$ is presented in Table 3.3.

3.5 Summary

We have addressed the problem of modeling large multivariate spatial data, where dimension reduction is sought both in the number of outcomes and in the number of spatial locations. The former is achieved with fewer number of factors, while the latter is achieved using a knot-based predictive process. Model-based strategies (Guhaniyogi et al., 2011) exist for choosing knots but recent findings, including our own explorations here, suggests that simple space-covering or clustering algorithms usually deliver robust inference. Our *adaptive* model lets the number of factors be stochastic and allows the data to drive the inference.

One can use either the predictive process $\tilde{w}_k(\mathbf{s})$ or its "modification" $f_k(\mathbf{s})$ in the adaptive spatial factor models but $f_k(\mathbf{s})$ adapts better to the data and, hence, is less sensitive to the knots. Although the substantive inference is quite robust to both these processes, some subtle differences are seen in the estimation algorithm with regard to transition among models. This is related to the strength of the spatial random field and is explained in Appendix C.

Given the widespread use of `R` as a statistical language, we presented computational benchmarks on `R` running on fairly standard architectures. In multivariate spatial analysis, "large" refers to the size of $m \times n$, where m is the number of outcomes and n is the number of locations. With `R` running on standard architectures, $mn > 1000$ is usually deemed exorbitant without dimension reduction. Substantial computational gains accrue from using lower-level languages (`C/C++`) on shared memory systems.

Alternative approaches for modeling large spatial datasets include an SPDE/GMRF approach proposed by Lindgren et al. (2011) that uses explicit Markov representations

of the Matérn covariance family using a class of stochastic partial differential equations. Rather than MCMC, they use a faster Integrated Nested Laplace Approximation (INLA) algorithm for Bayesian inference. Another approach, termed covariance tapering (Furrer et al., 2006), relies upon compactly supported correlation functions to produce sparse covariance matrices containing only a moderate number of nonzero elements. How effective these alternative approaches will be with dynamic spatial factor models is yet to be ascertained.

Space-time or dynamic factor modeling using low-rank processes can also be envisioned. Misalignment can now occur both over space and over time to yield data matrices that are highly irregular. Our formulation can, nevertheless, be easily adapted to such settings. Also, while we attended only to point-referenced data here, adaptive space-time factor models could be used for multivariate regionally aggregated data (Jin et al., 2007) as well. Here, usually the number of regions is not onerous, so dimension reduction is relevant over the number of outcomes (FA) and time (a temporal predictive process).

Chapter 4

Bayesian Experimental Design for Spatial Data Based on Hypothesis Testing

4.1 Introduction

Until now this thesis has focused upon the analysis of large point-referenced datasets. We have discussed two approaches. The first was computational – we saw how a variational Bayes algorithm can provide fast approximate Bayesian inference for univariate and multivariate spatial data. The second approach was model-based – we proposed an adaptive spatial factor model that made use of a low rank predictive process to achieve dimension reduction in situations where we had a large number of spatial outcomes over a large number of variables.

In this chapter, we take a somewhat different view for a specific problem. In many instances, inference is sought for the spatial regression slope coefficients in the presence of spatial correlation. This has been addressed by Hodges and Reich (2010). But how large a sample will be needed to ensure that we will be able to capture statistically significant regression slopes? Ascertaining the sample size and the locations where the outcomes should be observed may help avoid unnecessarily large datasets.

Methods for spatial data analysis typically focus upon understanding and predicting spatial outcomes. Due to enhanced capabilities in collecting, storing and accessing geographically referenced data, studies focus upon understanding the relationship between spatially referenced variables (e.g., Dominici et al., 2006; Hogan and Tchernis, 2004). To be more specific, consider the typical geostatistical setting, where $S = \{\mathbf{s}_1, \ldots, \mathbf{s}_n\} \in D$ is a finite set of n locations within a connected and convex subset $D \subseteq \Re^d$ and $\mathbf{y} = (y(\mathbf{s}_1), \ldots, y(\mathbf{s}_n))'$ is an $n \times 1$ vector of an outcome $y(\mathbf{s})$ observed at those n locations in S. Customarily, $\mathbf{y}$ is assumed to follow a multivariate normal distribution with mean $\mathbf{X}_n\boldsymbol{\theta}$ and a variance-covariance matrix $\boldsymbol{\Sigma}_\mathbf{y} = \tau^2\mathbf{I}_n + \sigma^2\mathbf{R}(\boldsymbol{\phi})$, where $\mathbf{X}_n = \{\mathbf{x}(\mathbf{s}_1), \ldots, \mathbf{x}(\mathbf{s}_n)\}'$ is an $n \times p$ matrix of predictors, $\boldsymbol{\theta}$ is a $p \times 1$ vector of slopes, $\mathbf{I}_n$ is the $n \times n$ identity matrix and $\mathbf{R}(\boldsymbol{\phi})$ is the spatial correlation matrix with additional process parameters $\boldsymbol{\phi}$. $\mathbf{X}_n$ is highlighted as a function of the design S using n locations assuming the covariates are known everywhere on D. The parameters that relate to the variance of the outcomes, $\{\tau^2, \sigma^2, \boldsymbol{\phi}\}$ are assumed known.

The problem we undertake in this chapter can be described in terms of an *analysis objective* and a *design objective*. Our analysis objective summarizes what we seek to achieve through the modeling. Here, we wish to test for departure from linear null hypotheses of the form $\mathbf{c}'\boldsymbol{\theta} \in \Theta_0$ in the presence of spatially correlated outcomes. The design objective is to find a sample size n and a set of locations S such that we are assured of meeting the analysis objectives with a specified probability or a fixed level of certainty. This falls within the category of spatial design problems. So, what distinguishes our current work from the existing literature?

The spatial sampling design literature proposes different criteria depending upon the analysis objective. Analysis objectives more commonly addressed include one or both of spatial prediction and estimation of spatial covariance. McBratney et al. (1981); Yfantis et al. (1987); Cressie et al. (1990); Müller (2005) derived optimal designs for prediction assuming that the covariance structure is completely known. The design criteria are considered here usually either the average kriging variance or the maximum kriging variance over the region of interest, or some modification thereof.

Criteria for estimation of spatial covariances are considered by Lark (2002); Müller and Zimmerman (1999); Irvine et al. (2007), in which different approximations of the covariance matrix of the estimated spatial process parameters are used. Zhu and Stein

(2005), who considered minimization of the average expected length of predictive intervals, focused instead upon criteria directly related to quality of estimation, such as minimizing the determinant of the asymptotic covariance matrix of the process parameters. Zimmerman (2006) argued that the primary design objectives are largely antithetical and thus lead to quite different "optimal" designs.

Other articles that address spatial prediction and spatial covariance estimation as an integrated problem include Bayesian approaches that specify prior distributions over covariance functions; see, e.g., Fuentes (2007); Diggle and Lophaven (2006). The approach presented here uses Bayesian risk as a loss function for single and multiple hypothesis testing on slope parameters. The primary objective is to find the most informative set of locations that minimize the Bayesian risk.

The impact of spatial correlations on the regression parameters has been studied recently by Hodges and Reich (2010). The design questions we ask are:

1. What is the sample size needed for testing a hypotheses concerning the regression slopes in the presence of spatial correlation?

2. What is the optimum placement of locations in the study region?

These questions can certainly be extended to multiple hypothesis testing situations.

A formal theory of Bayesian optimum experimental design dates back at least to Kiefer (1959). Lindley (1972) presented a two-part decision theoretic approach to experimental design. Lindley's approach involves specification of a suitable utility function (or loss function) reflecting the purpose and costs of the experiment; the best design is selected to maximize the expected utility (or minimize the expected loss). Pilz (1991) covers Bayesian design and estimation in linear models, although from a rather mathematical perspective. More literature on Bayesian optimum experimental designs and their different applications can be found (e.g., Chaloner and Verdinelli, 1995; Clyde and Chaloner, 1996; Müller, 1999).

Toman (1996) discussed a Bayesian experimental design for multiple hypothesis testing which is similar to what we propose here but, instead of assuming independence among outcomes as Toman did, we consider spatially correlated outcomes and the challenges they present. The difficulties faced in this project include: first, properly handling the complex spatial correlation in optimal experimental design problems and

understanding how that would affect the design; second, finding the optimum design on the spatial domain (map). The latter relates to the spatial sampling design. However, the analysis objective here is very different from what has been addressed in the literature.

Bayesian assurance approach (O'Hagan and Stevens, 2001) to determine the sample size is discussed in Section 4.2. The design criteria of spatial optimum experimental design based on Bayesian decision theory are studied in Section 4.3. Applications of this method to real data sets are in Section 4.4. Finally, Section 4.5 concludes the chapter with a summary and possible future directions.

4.2 Bayesian Assurance

In Bayesian sample size determination, goal or objective functions can be used to establish correspondences between different approaches (Inoue et al., 2005). One example is the desired rate for correctly identifying a hypothesis as true or false. Examples of goal functions include power, information, mean squared prediction error, size of confidence interval or probability interval, and classification error. O'Hagan and Stevens (2001) introduced the concept of Bayesian assurance, which is the probability of meeting the analysis objective. The Bayesian assurance can be seen as somewhat analogous to the frequentist concept of "power" of a hypothesis.

4.2.1 Bayesian Assurance and Frequentist Power

The familiar frequentist approach to sample size determination is reviewed first, which is easily cast in the two stages (analysis and design) framework. A particularly simple case assumes that the outcome variable is distributed as $y_i \overset{iid}{\sim} N(\theta, \sigma^2)$ for $i = 1, 2, \ldots, n$. The variance parameter σ^2 is assumed known. So $\bar{y} = \sum_{i=1}^{n} y_i/n$ follows a normal distribution with mean θ and variance σ^2/n. Now, consider the classical hypothesis testing problem: $H_0 : \theta = \theta_0$ against the alternative $H_1 : \theta = \theta_1 > \theta_0$. The null hypothesis H_0 will be rejected if $\bar{y} > \theta_0 + \frac{\sigma}{\sqrt{n}} z_{1-\alpha}$, where $\Phi(\cdot)$ is the standard normal cumulative density function and, hence, $\Phi(z_\alpha) = \alpha$. Based on this decision rule, the power π for a given testing critical difference $\Delta = \theta_1 - \theta_0$ for a specified alternative θ_1

is

$$\pi(\Delta, n) = P(Rej\ H_0 \,|\, H_1) = \Phi\left(\sqrt{n}\frac{\Delta}{\sigma} + z_\alpha\right). \tag{4.1}$$

Requiring the procedure to have a power of at least $1-\beta$, the ubiquitous sample size formula $n = (z_\alpha + z_\beta)^2 \left(\frac{\sigma}{\Delta}\right)^2$ is obtained. For two-sided alternatives, one simply replaces α by $\alpha/2$ in (4.1).

Now, let's consider the Bayesian approach. In the analysis stage, a prior $N(\theta_1, \tau^2)$ is assigned to θ, where $\tau^2 = \frac{\sigma^2}{n_0}$ and n_0 reflects the precision of the prior relative to the data. The posterior distribution of θ is given by $p(\theta \,|\, \bar{y}) = N\left(\theta \,|\, \frac{n_0}{n+n_0}\theta_1 + \frac{n}{n+n_0}\bar{y}, \frac{\sigma^2}{n+n_0}\right)$. The hypothesis test is $H_0 : \theta < \theta_0$ and the alternative $H_1 : \theta \geq \theta_0$. Let $A_\alpha(\theta_0, \theta_1)$ be the rejection region which is defined as a set of $\bar{y}$ satisfying $P(\theta < \theta_0 \,|\, \bar{y}) < \alpha$, that is, equation $A_\alpha(\theta_0, \theta_1) = \left\{\bar{y} : \bar{y} > \theta_0 - \frac{n_0}{n}(\theta_1 - \theta_0) - \sqrt{\left(1 + \frac{n_0}{n}\right)}\frac{\sigma}{\sqrt{n}}z_\alpha\right\}$. If a flat prior is used in this stage, i.e., $n_0 \to 0$, the rejection region becomes

$$A_\alpha(\theta_0, \theta_1) = \left\{\bar{y} : \bar{y} > \theta_0 - \frac{\sigma}{\sqrt{n}}z_\alpha\right\}, \tag{4.2}$$

which is the same as frequentist case.

In the design stage, a design prior $N(\theta_1, \sigma^2/n_d)$ for θ is specified to determine the Bayesian assurance. n_d reflects the precision of design prior relative to the data. Then, the marginal distribution of $\bar{y}$ is $N\left(\bar{y} \,|\, \theta_1, \left(\frac{1}{n} + \frac{1}{n_d}\right)\sigma^2\right)$. According to the definition of the rejection region in (4.2), the Bayesian assurance is calculated as:

$$\gamma(\Delta, n) = P_{\bar{y}}(A_\alpha(\theta_0, \theta_1)) = \Phi\left(\sqrt{\frac{n_d}{n+n_d}}\left[\sqrt{n}\left(\frac{\Delta}{\sigma}\right) + z_\alpha\right]\right). \tag{4.3}$$

If an exact prior is used in the design stage, i.e., $n_d \to \infty$, the Bayesian risk $\gamma(\Delta, n)$ becomes $\Phi\left(\sqrt{n}\frac{\Delta}{\sigma} + z_\alpha\right)$ and the Bayesian assurance curve coincide with the classical power curve.

Instead of using an exact design prior, there is another approach that can provide the frequentist result. This avoids the use of two different priors at the analysis and design stages. Suppose that true distribution of y_i follows $N(\theta_1, \sigma^2)$. Then, $\bar{y} \sim N(\theta_1, \sigma^2/n)$. The Bayesian assurance can then be calculated using this distribution as

$$\gamma(\Delta, n) = P_{\bar{y}}(A_\alpha(\theta_0, \theta_1)) = \Phi\left(\sqrt{n}\frac{\Delta}{\sigma} + z_\alpha\right),$$

which is the same as frequentist power.

4.2.2 The Problem with the Bayesian Assurance

Even though frequentist power can be recovered as a special case, Bayesian assurance is problematic in more general settings where the same prior is used for both the design and analysis stages. Assuming a prior $N(\theta_1, \sigma^2/n_0)$ is assigned to θ for both stages and the marginal distribution of $\bar{y}$ is used, the Bayesian assurance is

$$\gamma(\Delta, n) = \Phi\left(\sqrt{\frac{n_0}{n}}\left[\sqrt{n+n_0}\left(\frac{\Delta}{\sigma}\right) + z_\alpha\right]\right). \tag{4.4}$$

There are several properties that separate γ from the frequentist power π. First of all, when $n \to \infty$,

$$\lim_{n\to\infty} \gamma(\Delta, n) = \Phi\left(\sqrt{n_0}\frac{\Delta}{\sigma}\right),$$

which is an increasing function of Δ, while on the other hand, $\lim_{n\to\infty} \pi(\Delta, n) = 1$. Second, when $n \to 0$ or more realistically $n/n_0 \to 0$, the result is

$$\begin{aligned}
\gamma(\Delta, n) &\simeq \Phi\left(\sqrt{\frac{n_0}{n}}\left[\sqrt{n_0}\left(1 + \frac{n}{2n_0}\right)\left(\frac{\Delta}{\sigma}\right) + z_\alpha\right]\right) \\
&= \Phi\left(\sqrt{\frac{n_0}{n}}\left[\sqrt{n_0}\left(\frac{\Delta}{\sigma}\right) + z_\alpha\right] + \frac{\sqrt{n}}{2}\left(\frac{\Delta}{\sigma}\right)\right) \\
&= \begin{cases} 1 & \text{if } \sqrt{n_0}\left(\frac{\Delta}{\sigma}\right) + z_\alpha > 0 \\ 1/2 & \text{if } \sqrt{n_0}\left(\frac{\Delta}{\sigma}\right) + z_\alpha = 0 \\ 0 & \text{if } \sqrt{n_0}\left(\frac{\Delta}{\sigma}\right) + z_\alpha < 0 \end{cases}.
\end{aligned} \tag{4.5}$$

It is easy to see from (4.5) that the Bayesian assurance cannot be an increasing function with respect to sample size in general. But $\pi(\Delta, n)$ is clearly an increasing function of n with limit α when $n \to 0$.

If we take a derivative of $\gamma(\Delta, n)$ with respect to n using (4.3), the result is:

$$\frac{\partial\gamma}{\partial n} = -\frac{c(n)}{2n^{3/2}}\left(\frac{n_0}{\sqrt{n+n_0}}\frac{\Delta}{\sigma} + z_\alpha\right),$$

where $c(n) = \varphi\left(\sqrt{n_0}\left[\sqrt{1+\frac{n_0}{n}}\left(\frac{\Delta}{\sigma}\right) + z_\alpha\sqrt{\frac{1}{n}}\right]\right) > 0$. Here, $\varphi(\cdot)$ is the standard normal distribution function. This first order derivative is bigger than 0 when n is large due to the negative z_α (α is usually small in practice). Using a similar argument, when $n/n_0 \to 0$, the result is:

$$\frac{\partial\gamma}{\partial n} \begin{cases} > 0 & \text{if } \sqrt{n_0}\left(\frac{\Delta}{\sigma}\right) + z_\alpha \le 0 \\ < 0 & \text{if } \sqrt{n_0}\left(\frac{\Delta}{\sigma}\right) + z_\alpha > 0 \end{cases},$$

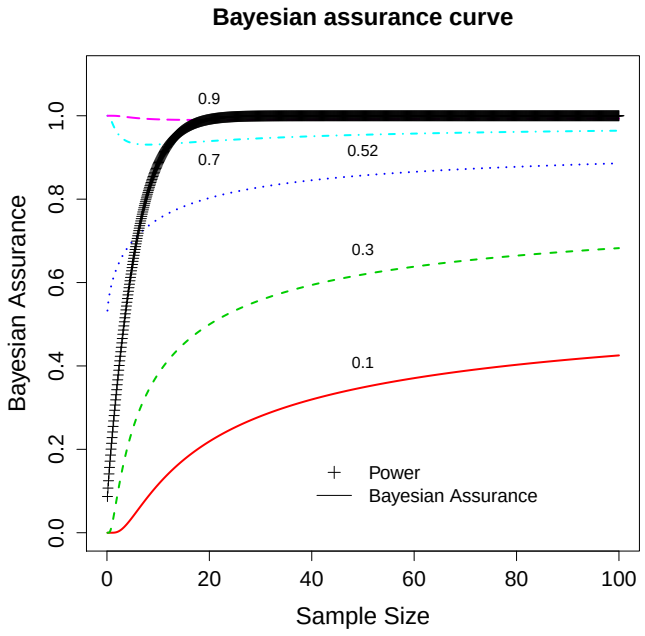


Figure 4.1: Bayesian assurance curves for different values of Δ.

which proves $\gamma(\Delta, n)$ is not an increasing function with respect to n in general.

Figure 4.1 demonstrates the Bayesian assurance curve at different Δ values. Without loss of generality, σ is specified to be 1 and α to be 0.05. The Bayesian assurance curve with $\Delta = 0.52$ is specially chosen to satisfy $\sqrt{n_0}\left(\frac{\Delta}{\sigma}\right) + z_\alpha = 0$. A power curve is also plotted in Figure 4.1 using function (4.1) with parameter $\Delta = 0.9$, $\sigma = 1$ and $\alpha = 0.05$ to compare with Bayesian assurance curves.

4.3 Bayesian Decision Theoretic Approach

Decision-theoretic approaches to experimental design assume that the investigator is a rational decision maker choosing an action that minimizes the loss of the possible consequences averaging with respect to all of the relevant unknowns. The process of experimentation followed by inference/decision making proceeds in time order; it is easier to solve the optimal decision problem in reverse time order. Suppose there are k tests of interest given in the form $H_{0j} : \mathbf{c}_j'\boldsymbol{\theta} \in \Theta_{0j}$ versus $H_{1j} : \mathbf{c}_j'\boldsymbol{\theta} \in \Theta_{1j}$, $j = 1, \ldots, k$,

	Accept H_0	Accept H_1
H_0 is True	0	K
H_1 is True	1	0

Table 4.1: Loss Function; No loss is incurred with a correct decision, but a loss of 1 is incurred if H_0 is not rejected when in fact H_1 is true, and a loss of K is incurred if H_0 is rejected when in fact H_0 is True.

where $\mathbf{c}_j$ is a $p \times 1$ vector. We assume that those two hypothesis testings are mutually exclusive. In the analysis stage, in which the decision about multiple comparisons is made, the data $\mathbf{y}$ are known and a Bayesian decision between H_{0j} and H_{1j} is based on their posterior probabilities. At the time of choosing the experimental design, Bayesian loss function is averaged with respect to the prior probability on $\boldsymbol{\theta}$ and the conditional sampling distribution of $\mathbf{y}$ given $\boldsymbol{\theta}$.

4.3.1 Model, Loss Function, and Bayesian Risk

Let's consider a single hypothesis test (i.e., $k = 1$) first. The decision of interest is either a selection of H_0 (denoted $d = 0$) or not (denoted $d = 1$) (superscript j is dropped for clarity.). If $\mathbf{c}'\boldsymbol{\theta} \in \Theta_0$, then the decision $d = 0$ is appropriate, while if $\mathbf{c}'\boldsymbol{\theta} \in \Theta_1$, then $d = 1$ is better. The "$0 - 1 - K$" loss function $L(\boldsymbol{\theta}, d)$ for traditional hypothesis testing problems is shown in Table 4.1. It was shown that the Bayesian decision rule $\delta(\mathbf{y})$ to make decision $d = 1$ when the posterior probability of H_0 is less than $1/(1 + K)$, and to make decision $d = 0$ otherwise (Berger, 1985). Thus, in classical terminology, the rejection region of the Bayesian test is

$$A_1 = \{\mathbf{y} : P(\Theta_0 \,|\, \mathbf{y}) < 1/(1 + K)\}. \tag{4.6}$$

By properly choosing $K = (1 - \alpha)/\alpha$, the rejection region is of level α.

The objective is to find sample size n, the best predictor $\mathbf{X}_n$ and hence the corresponding design S. Because the observation vector $\mathbf{y}$ is not available, it is necessary to average over both $\mathbf{y}$ and $\boldsymbol{\theta}$ to obtain the Bayesian risk r, which is defined as $r = \int L(\theta, \delta(\mathbf{y})) f(\mathbf{y} \,|\, \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta}$, where $f(\mathbf{y} \,|\, \boldsymbol{\theta})$ and $\pi(\boldsymbol{\theta})$ denote the likelihood of $\mathbf{y}$ and the prior distribution of $\boldsymbol{\theta}$. The hypothesis test $H_0 : \mathbf{c}'\boldsymbol{\theta} \leq \omega$ versus $H_1 : \mathbf{c}'\boldsymbol{\theta} > \omega$

is considered and $\pi(\boldsymbol{\theta}) = N_p(\boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_\theta)$. Then, the posterior distribution $p(\boldsymbol{\theta} \,|\, \mathbf{y})$ becomes $N_p\left(\boldsymbol{\theta} \,|\, \boldsymbol{\mu}_{\theta\,|\,\mathbf{y}}, \boldsymbol{\Sigma}_{\theta\,|\,\mathbf{y}}\right)$, where $\boldsymbol{\Sigma}_{\theta\,|\,\mathbf{y}} = \left(\mathbf{N}_n + \boldsymbol{\Sigma}_\theta^{-1}\right)^{-1}$ for $\mathbf{N}_n = \mathbf{X}_n' \boldsymbol{\Sigma}_{\mathbf{y}}^{-1} \mathbf{X}_n$ and $\boldsymbol{\mu}_{\theta\,|\,\mathbf{y}} = \boldsymbol{\Sigma}_{\theta\,|\,\mathbf{y}} \left(\mathbf{X}_n' \boldsymbol{\Sigma}_{\mathbf{y}}^{-1} \mathbf{y} + \boldsymbol{\Sigma}_\theta^{-1} \boldsymbol{\mu}_\theta\right)$. The rejection region is calculated from (4.6) as:

$$A_1 = \left\{ \mathbf{y} : u > \mathbf{c}' \boldsymbol{\Sigma}_{\theta\,|\,\mathbf{y}} \mathbf{N}_n \boldsymbol{\mu}_\theta - z_\alpha \sqrt{\mathbf{c}' \boldsymbol{\Sigma}_{\theta\,|\,\mathbf{y}} \mathbf{c}} - \Delta \right\},$$

where $u = \mathbf{c}' \boldsymbol{\Sigma}_{\theta\,|\,\mathbf{y}} \mathbf{X}_n' \boldsymbol{\Sigma}_{\mathbf{y}}^{-1} \mathbf{y}$ and $\Delta = \mathbf{c}' \boldsymbol{\mu}_\theta - \omega$. The marginal distribution of u is normal with mean $\mathbf{c}' \boldsymbol{\Sigma}_{\theta\,|\,\mathbf{y}} \mathbf{N}_n \boldsymbol{\mu}_\theta$ and variance $\sigma_u^2 = \mathbf{c}' \boldsymbol{\Sigma}_\theta \left(\mathbf{N}_n^{-1} + \boldsymbol{\Sigma}_\theta\right)^{-1} \boldsymbol{\Sigma}_\theta \mathbf{c}$. Let $r(\delta, \mathbf{N}_n)$ be the Bayesian risk, which is

$$\Phi\left(\Delta/\xi\right) + (1+K) \int_{-z_\alpha}^{+\infty} \frac{t}{\sqrt{2\pi}} \Phi(-v) \exp\left\{ -\frac{(tv - o)^2}{2} \right\} \mathrm{d}v - \Phi(t z_\alpha + o), \tag{4.7}$$

where $\xi = \sqrt{\mathbf{c}' \boldsymbol{\Sigma}_\theta \mathbf{c}}$, and $o = \Delta/\sigma_u$. The quantity t is the square root of the ratio of two variances, $t = \sqrt{\mathbf{c}' \boldsymbol{\Sigma}_{\theta\,|\,\mathbf{y}} \mathbf{c} / \sigma_u^2}$. Note that the quantity o can be written as a function of t, namely $\frac{\Delta}{\xi}\sqrt{t^2+1}$ (see details in Appendix D).

It is interesting to examine Bayesian risk $r(\delta, \mathbf{X}_n)$ as a function of $\mathbf{X}_n$ through $\mathbf{N}_n$. And, the Bayesian risk is a function of $\mathbf{N}_n$ only through t. One thing that needs to be highlighted here is that vague prior for $\boldsymbol{\theta}$ is not appropriate for experimental design or sample size determination. As $\boldsymbol{\Sigma}_\theta^{-1} \to \mathbf{O}$, $\sigma_u^2 \to \mathbf{c}' \boldsymbol{\Sigma}_\theta \mathbf{c}$, which implies $t \to 0$, and $o \to \Delta/\xi$, so the Bayesian risk r approaches 0, which is not a function of $\mathbf{X}_n$ any more.

4.3.2 Optimal Design for a Single Test

The Bayesian optimal design will be the matrix $\mathbf{N}_n$ which produces the smallest value of $r(\delta, \mathbf{N}_n)$ as given in (4.7). The relationship between $r(\delta, \mathbf{N}_n)$ and t will be key in the derivation of the optimal $\mathbf{N}_n$ and sample size n. Taking a partial derivative of $r(\delta, \mathbf{N}_n)$ with respect to t, the result is:

$$\frac{\partial r(\delta, \mathbf{N}_n)}{\partial t} = \frac{K+1}{2\pi(t^2+1)} \exp\left\{ -\frac{\Delta^2}{2\xi^2} \right\} \exp\left\{ -\frac{(t^2+1)(z_\alpha + \dot{o})^2}{2} \right\} > 0, \tag{4.8}$$

where $\dot{o} = \frac{\Delta}{\xi}\sqrt{\frac{t^2}{t^2+1}}$ is the derivative of o with respect to t (see Appendix D for details). The first order derivative of $r(\delta, \mathbf{N}_n)$ is always positive. Notice that, when $t \to 0$, (4.8) becomes $(K+1)\varphi(z_\alpha)\varphi\left(\Delta/\xi\right)$. It can be concluded that the Bayesian risk $r(\delta, \mathbf{N}_n)$ is an increasing function with respect to t.

It is of further interest to examine t as a function of $\mathbf{N}_n$. t can be rewritten as:

$$t = \sqrt{\frac{\xi^2}{\mathbf{c}'\boldsymbol{\Sigma}_\theta \left(\mathbf{N}_n^{-1} + \boldsymbol{\Sigma}_\theta\right)^{-1} \boldsymbol{\Sigma}_\theta \mathbf{c}} - 1} = \sqrt{\frac{\xi^2}{\xi^2 - \mathbf{c}' \left(\mathbf{N}_n + \boldsymbol{\Sigma}_\theta^{-1}\right)^{-1} \mathbf{c}} - 1}. \tag{4.9}$$

It can be proved that t is a decreasing function of $\mathbf{N}_n$ (Marshall and Olkin, 1979, 463-464) and $\mathbf{N}_n$ is a non-decreasing function of n (see proof in Appendix D). So t is a non-increasing function with respect to sample size n, and so is $r(\delta, \mathbf{N}_n)$. When the overall sample size is reasonably large, $\mathbf{c}' \left(\mathbf{N}_n + \boldsymbol{\Sigma}_\theta^{-1}\right)^{-1} \mathbf{c}$ becomes small, and t will be close to 0 as well. As discussed in Section 4.3.1, the Bayesian risk r approaches 0 when the sample size n is large.

Formally, solutions to the sample size determination under certain constraints exist since Bayesian risk is a monotonic function of n and approaches 0 when n is large. In many clinical studies, the matrix of predictors $\mathbf{X}_n$ is only a function of n and no design problem is involved. In such cases, the covariance matrix $\boldsymbol{\Sigma}_\mathbf{y}$ is much simpler than correlated cases, such as, longitudinal and spatial analysis. The Bayesian risk can be calculated directly from (4.7) using numerical methods. The smallest n that satisfies the constraint would be the desired sample size (Inoue et al., 2005).

For a given sample size n, $\mathbf{c}'\boldsymbol{\Sigma}_\theta \left[\mathbf{N}_n^{-1} + \boldsymbol{\Sigma}_\theta\right]^{-1} \boldsymbol{\Sigma}_\theta \mathbf{c}$ is a concave function with respect to $\mathbf{N}_n$ (Marshall and Olkin, 1979, 468-472). Here, t is a decreasing convex function of the quantity $\mathbf{c}'\boldsymbol{\Sigma}_\theta \left(\mathbf{N}_n^{-1} + \boldsymbol{\Sigma}_\theta\right)^{-1} \boldsymbol{\Sigma}_\theta \mathbf{c}$. So r is a convex function with respect to $\mathbf{N}_n$ and the optimal design is unique for a given n. The proof can be found in Toman (1996, Theorem 1). The matrix $\mathbf{N}_n$ that minimizes $r(\delta, \mathbf{N}_n)$ is the unique Bayesian design; that is, the design minimizes the trace of $\mathbf{C} \left(\mathbf{N}_n + \boldsymbol{\Sigma}_\theta^{-1}\right)^{-1}$, where $\mathbf{C} = \mathbf{c}\mathbf{c}'$.

If σ^2 is set at 0, the model becomes an ordinary linear regression and $\boldsymbol{\Sigma}_\mathbf{y} = \tau^2 \mathbf{I}_n$. It is very interesting to see the impact of spatial correlation to the experimental design. For ordinary linear regression, $\mathbf{N}_n = \dfrac{\mathbf{X}_n'\mathbf{X}_n}{\tau^2}$, while $\mathbf{N}_n = \dfrac{\mathbf{X}_n'\mathbf{X}_n}{\tau^2} - \dfrac{\mathbf{X}_n' \left(\mathbf{I}_n + \tau^2 \mathbf{R}^{-1}(\boldsymbol{\phi})/\sigma^2\right)^{-1} \mathbf{X}_n}{\tau^2}$ for spatial regression. Since $t(\mathbf{N}_n)$ is a decreasing function of $\mathbf{N}_n$ and Bayesian risk is an increasing function of t, the spatial model yields larger Bayesian risk than ordinary linear regression. So the experimental design for the data set, which is expected to have some level of correlation but fails to consider it in the design stage, would result in smaller sample size.

4.3.3 The Optimal Design for Multiple Hypothesis Testing

A natural extension for the "$0-1-K$" loss function can be constructed for the k-decision problem, which combines the loss function for each single test (Toman, 1996). This specification leads to

$$L_m(\boldsymbol{\theta}, \mathbf{d}) = \sum_{j=1}^{k} w_j L_j(\boldsymbol{\theta}, d_j), \tag{4.10}$$

where w_j are the appropriate weights representing the relative importance of each decision satisfying $\sum_{j=1}^{k} w_j = 1$, L_j and d_j are the loss function and decision for the j-th hypothesis test respectively, and $\mathbf{d}$ is the collection for all the individual decisions. Similar loss function was used in Müller et al. (2004) assuming all the w_j's are the same. The Bayesian risk for multiple tests becomes

$$r_m(\boldsymbol{\delta}, \mathbf{N}_n) = \sum_{j=1}^{k} w_j r_j(\delta^j, \mathbf{N}_n), \tag{4.11}$$

where r_j and δ^j are Bayesian risk and the decision rule for the j-th hypothesis test respectively, and $\boldsymbol{\delta}$ is the collection of all the δ^js. Unfortunately, $r_m(\boldsymbol{\delta}, \mathbf{N}_n)$ is not, in general, a convex or even quasi-convex function of $\mathbf{N}_n$. Sufficient conditions for the optimal design appear to be difficult to obtain because the direct Lagrangian is intractable. So $r_m(\boldsymbol{\delta}, \mathbf{N}_n)$ is approximated by a convex function, which will be optimized to yield a candidate design.

The Bayesian risk $r_m(\boldsymbol{\delta}, \mathbf{N}_n)$ depends on the design matrix $\mathbf{N}_n$ only through t_js (defined in (4.9) for the j-th hypothesis test). When the overall sample size is reasonably large, t_j will be small. This suggests approximating the Bayesian risks $r_j(\delta^j, \mathbf{N}_n)$ by a first-order Taylor series expansion around the value $t_j = 0$ yielding

$$\begin{aligned} r_j(\delta^j, \mathbf{N}_n) &\cong (K+1)\varphi(z_\alpha)\varphi\left(\Delta_j/\xi_j\right)\sqrt{\frac{\xi_j^2}{\xi_j^2 - \mathbf{c}_j'\left(\mathbf{N}_n + \boldsymbol{\Sigma}_\theta^{-1}\right)^{-1}\mathbf{c}_j} - 1} \\ &\cong (K+1)\varphi(z_\alpha)\varphi\left(\Delta_j/\xi_j\right)\sqrt{\frac{\mathbf{c}_j'\left(\mathbf{N}_n + \boldsymbol{\Sigma}_\theta^{-1}\right)^{-1}\mathbf{c}_j}{\xi_j^2}} \end{aligned} \tag{4.12}$$

and

$$r_m(\boldsymbol{\delta}, \mathbf{N}_n) \cong (K+1)\varphi(z_\alpha) \sum_{j=1}^{k} w_j \varphi\left(\Delta_j/\xi_j\right) \sqrt{\frac{\mathbf{c}_j' \left(\mathbf{N}_n + \boldsymbol{\Sigma}_\theta^{-1}\right)^{-1} \mathbf{c}_j}{\xi_j^2}}$$
$$\leq (K+1)\varphi(z_\alpha) \sqrt{\text{trace}\left[\mathbf{B}\left(\mathbf{N}_n + \boldsymbol{\Sigma}_\theta^{-1}\right)^{-1}\right]}, \tag{4.13}$$

where $\mathbf{B} = k \sum_{j=1}^{k} w_j^2 \frac{\varphi\left(\Delta_j/\xi_j\right)^2}{\xi_j^2} \mathbf{c}_j \mathbf{c}_j'$. The Cauchy-Schwarz inequality is used here to find this upper bound for (4.12). The equality holds when the Bayesian risks for each single test after being multiplied by the weight w_j are the same. Note that this approximate optimality criterion is a form of the ψ criterion (Chaloner, 1984). The smaller the Δ_j is, the more weight would be given in the computation of the optimal design.

Although there is a rich literature about how to find the optimal experimental design for the continuous design problem (see, e.g., Chaloner, 1984; Fedorov, 1972; Pilz, 1991), in general it is not possible to find optimal designs in a spatial application. It is practically impossible to have multiple measurements at one location (e.g., soil sampling) or allow multiple sampling locations to be arbitrarily close (e.g., precipitation). SAA, which has been discussed in detail in Section 1.4, is developed to find the best sampling design on a discrete design region.

4.4 Application

In this section, the methodology described above is applied to generate optimal sampling designs to the air pollutant data sets obtained in California from Environmental Protection Agency (EPA) databases. The data set is a multivariate spatial data set which comprises five pollutants: carbon monoxide (CO), nitrogen dioxide (NO_2), ozone (O_3), particulate matter with diameter < 2.5 micrometers (PM2.5) and particulate matter with diameter < 10 micrometers (PM10). The data set has been analyzed in Section 3.4.3 using model (3.5). The covariates include an intercept for each air pollutant and elevation in kilometers. Of particular interest is designing the monitor sets for ozone. Other air pollutants can be used as predictors. The design problem will be discussed under different scenarios.

4.4.1 Study One

This example deals with the optimum design of ozone monitoring stations based on testing the hypothesis whether the elevation has a significant effect on ozone. The design matrix $\mathbf{X}_n$ includes a column of 1s and and a column of the elevation. Here, a real elevation map in California including San Francisco Bay and some ocean area is used (Figure 4.2). The domain is from -122.7 to -119.2 in longitude and 36.73 to 39.84 in latitude. It is discretized on a 90×60 grid. The black spots in Figure 4.2 are the current 79 ozone monitoring stations. The range of elevation in Figure 4.2 is from -0.02 to 3.78 kilometer from sea level.

The variance parameters estimated in Section 3.4.3 are used in this design problem and treated as known. They are specified as follows: $\sigma^2 = 0.32$, $\tau^2 = 0.59$ and spatial range parameter $\phi = 0.018$. The objective function (4.7) is used in order to find the best design based on a single test with $\mathbf{c} = (0, 1)'$ and $\omega = 0$. Its approximation, equation (4.12) (equation (4.13) is the same as (4.12) for a single hypothesis test), is also calculated to compare with (4.7). The prior distribution for the slope parameter is normal with mean 0.5 and variance 5.

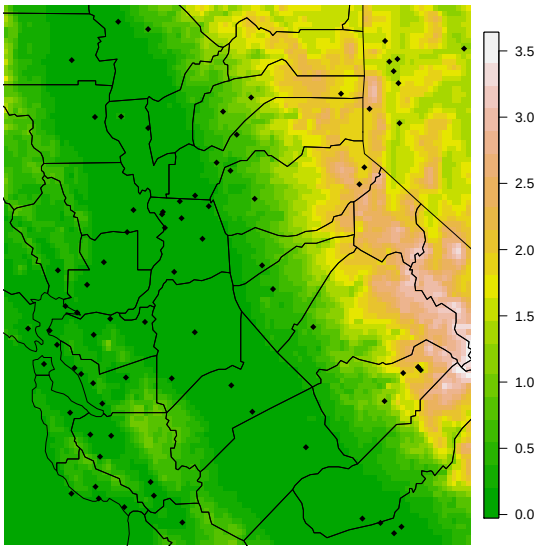


Figure 4.2: Elevation map with current ozone monitor sets.

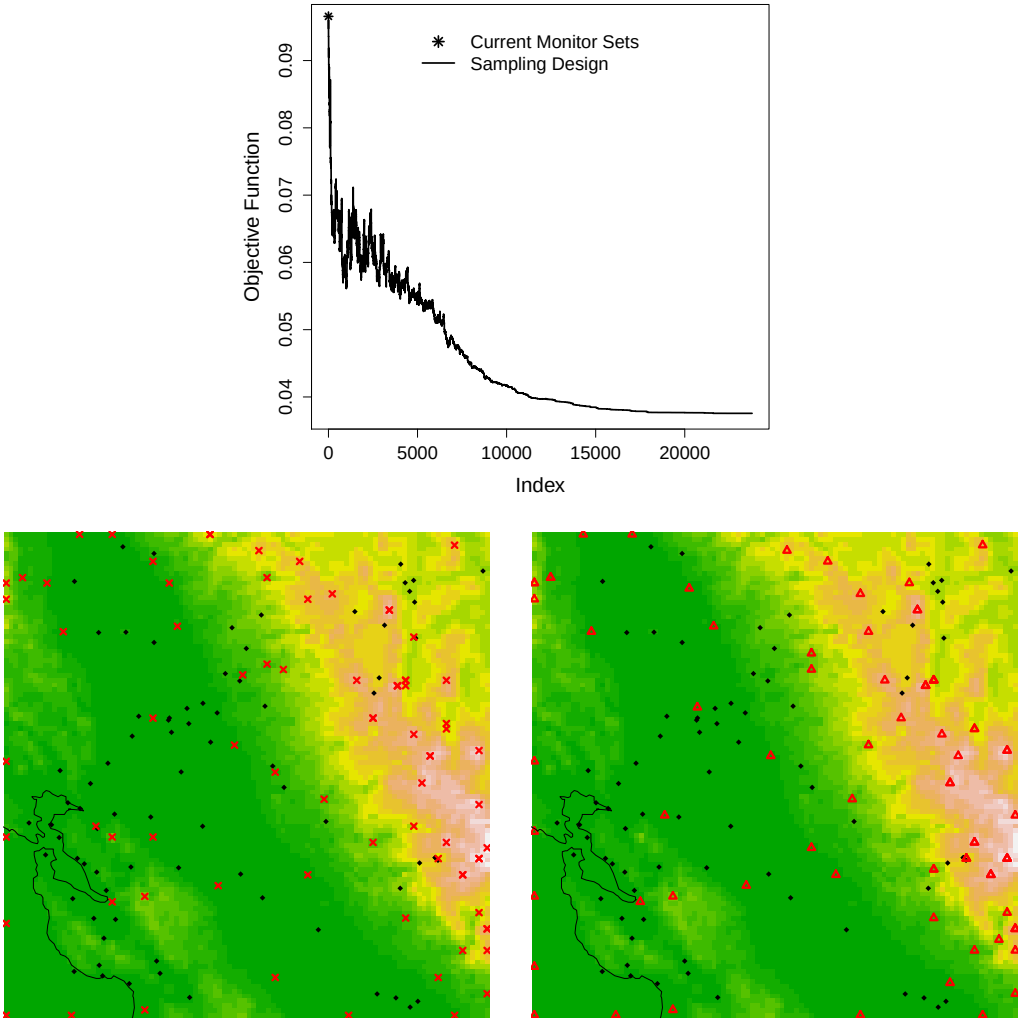


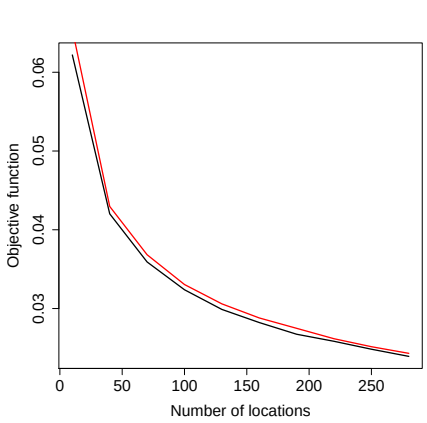
Figure 4.3: Optimal experimental design results. The top graph depicts the decrease of the Bayesian risk along the algorithm; the bottom two plots represent the best design with different initial locations with the black dots being the current stations. Lower left: the current monitor sets are used as initial design, red $\times$'s are "optimal" design points; lower right: randomly sampled points as initial design, red $\triangle$'s are "optimal" design points.

Figure 4.3 presents the designs provided by the SAA with sample size equal to 79, which is the number of current ozone monitor sets on this map. The top plot is the curve of the Bayesian risk. The profile of this curve is typical of an SAA: when the temperature is high, worsening states are more likely to be accepted than when it is low. The starting point in this plot represents the Bayesian risk value for the current monitor network, which is about three times bigger than the final optimal design.

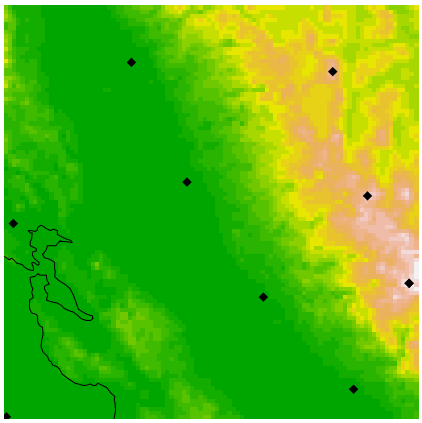
There are two different initial designs demonstrated in the lower figures. One uses the current monitor sets and the other one uses a randomly generated set of locations. The red symbols represent the optimal designs generated by SAA. In this example the claim can be made that Bayesian risk has been properly optimized. Even though the lower two figures present different designs, they share a similar pattern on the map. Many local optimum points are captured by both designs, and the rest of the points cover uniformly the flat area or boundaries.

To better understand the Bayesian risk and the corresponding design as a function of the sample size, SAA was run multiple times for each sample size to test the same hypothesis. The initial design for each run was randomly generated. In Figure 4.4(a), the Bayesian risk is plotted with respect to sample sizes. The black line is the minimum Bayesian risk within 10 runs, while the red line is the maximum value. The Bayesian risk decreases quickly from 10 locations to 100 locations then slows down. Sample size can be determined based on this curve. A threshold value can be specified, and the smallest sample size whose Bayesian risk is smaller than this value is considered. Note that adding more points is still useful, if the budget allows it due to the possibly missed measurements. The approximation for the Bayesian risk, equation (4.12) was also calculated using the optimum design at each sample size, however, the values are too close to the Bayesian risk, which cannot be identified in the plot in any way. So we don't bother plotting it. Equation (4.12) is clearly good to use as the objective function for each single hypothesis test, which is much easer to calculate.

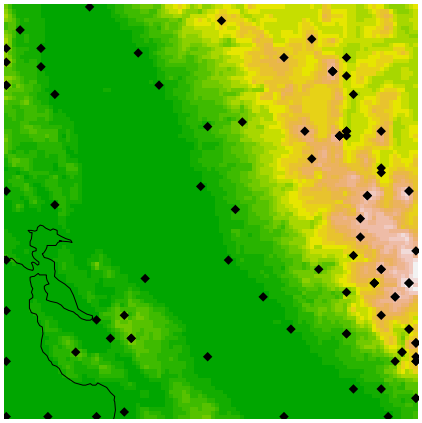
One important thing that can be learned from Figure 4.4(a) is the variability of the optimum design generated by SAA according to different initial designs. It can be seen from the plot that within 10 runs the difference between maximum and minimum values at any sample size are relatively small, which indicates the starting designs in



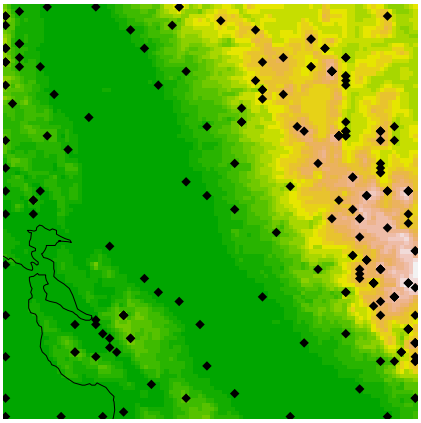
(a)



(b) 10 locations, $O(S) = 0.062$

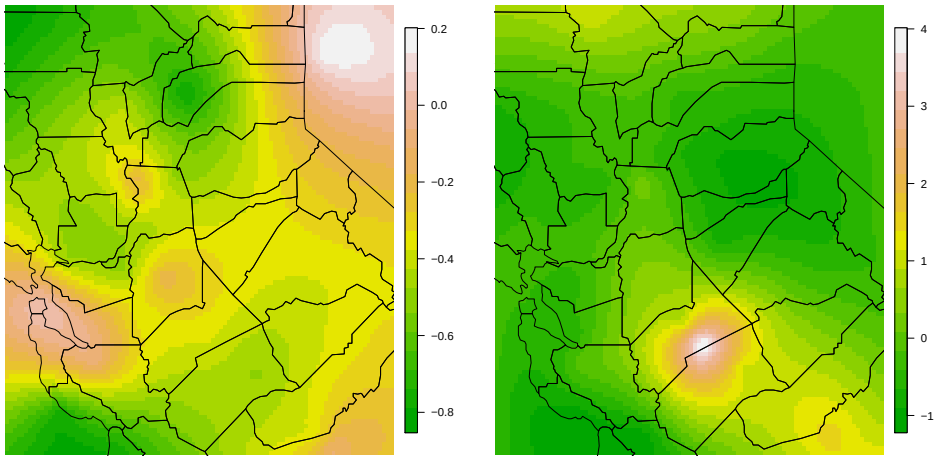


(c) 100 locations, $O(S) = 0.030$



(d) 190 locations, $O(S) = 0.027$

Figure 4.4: Bayesian risk as a function of the sample size (a.) and several optimal designs of different sizes(b., c., d.) generated by SAA. The black line in (a.) is the minimum Bayesian risk within 10 SAA runs, while the red line is the maximum.



(a) NO_2 (b) PM2.5

Figure 4.5: The interpolation of two air pollutants in the study domain.

this example have very limited influence to the sample size determination. Three resulting designs are depicted in Figure 4.4 b., c., and d., respectively, for 10, 100, 190 locations. The black curve at the lower left of those plots is the coastline. The evolution of the designs when the number of the locations increases is mainly characterized by two patterns: an improvement in the spatial coverage of area where there is elevation optimum and increased density of the locations in the flat areas and boundaries.

4.4.2 Study Two

The optimal experimental design for multiple hypothesis testing is studied in this section. The same domain in California is used. Two predictors are added to the model. One is nitrogen dioxide (NO_2), shown in Figure 4.5(a), the other one is particulate matter with diameter < 2.5 micrometers (PM2.5), shown in Figure 4.5(b). The values of those two covariates on the grid are, in fact, predicted from the measurements of its monitor network using `gstat` package in `R`. The slope parameter $\boldsymbol{\theta}$ is a 4×1 vector with a intercept and three slope coefficients.

The hypothesis tests interested in this model is on the three slope coefficients (k=3). c_j is a 4×1 vector with $(j+1)$-th element being 1 and the rest 0, w_j is specified to be

0 for $j = 1, 2, 3$. The prior distribution of $\boldsymbol{\theta}$ is a multivariate normal distribution with mean $\boldsymbol{\mu}_\theta = (1, 1, 1, 1)'$ and variance $\boldsymbol{\Sigma}_\theta = 5 \times \mathbf{I}_4$. The other parameters that are needed to calculate the objective function (4.12), are specified the same as in Section 4.4.1. An attempt was also made to use (4.13) as the objective function, but due to the different Bayesian risk of each single test, (4.13) is not a good approximation. The equal weight ($w_j = 1/3$) for each single test is specified. The curve of the objective function generated in SAA is shown in Figure 4.6 for the sample size of 100. The profile of this curve is similar to the top plot in Figure 4.3, which indicates good convergence.

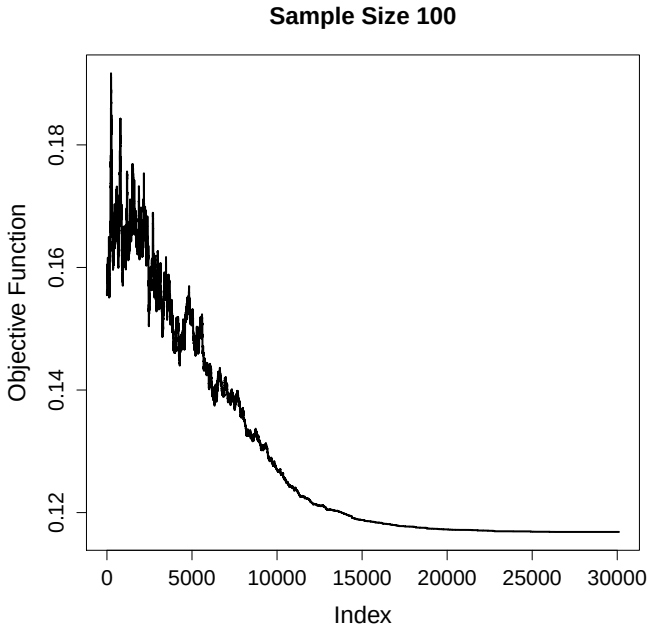
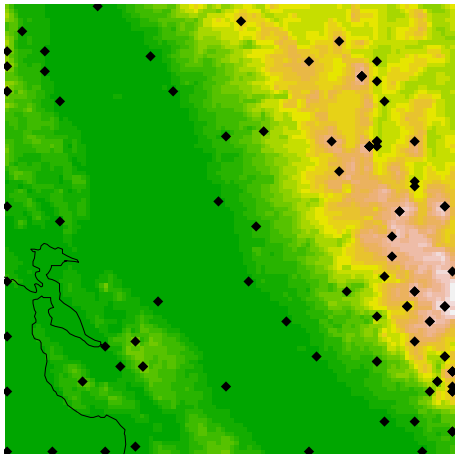


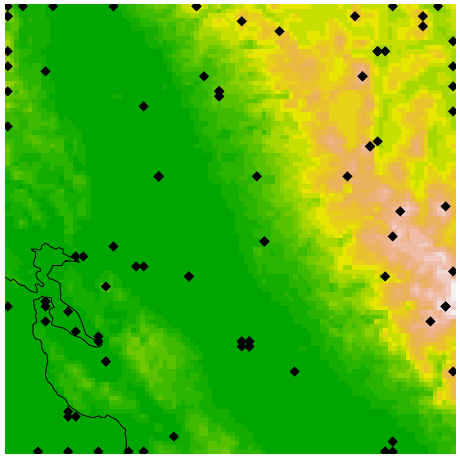
Figure 4.6: The curve of the objective function generated in SAA of multiple hypothesis testing case.

Figure 4.7 presents optimal design provided by the SAA for multiple hypothesis tests with the sample size of 100. Figure 4.7(a) is the optimal design for the single hypothesis test with the same sample size, which is in fact the lower left plot in Figure 4.4 (it is included here for comparison). The optimum design is plotted on three interpolations of the covariates: elevation in Figure 4.7(b), NO_2 in Figure 4.7(c), PM2.5 in Figure 4.7(d).

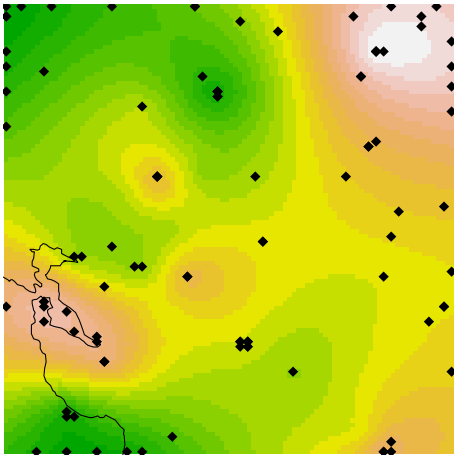
The optimal design for the multiple tests is trying to capture the optimum points for all the covariates, which can be clearly identified from the figures. Comparing Figure 4.7(b), 4.7(c) and 4.7(d) to Figure 4.7(a), the optimal design in multiple hypothesis test balances between each single test according to their Bayesian risk values and the weights assigned to them.



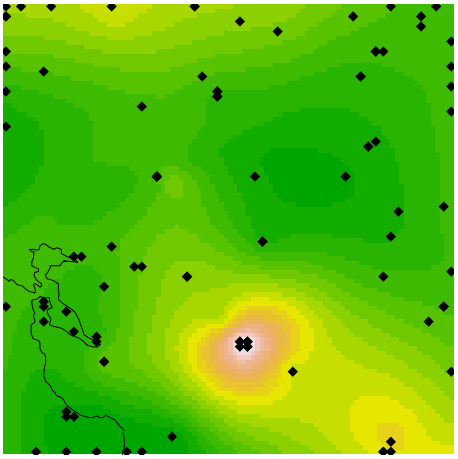
(a) Elevation: Single Testing



(b) Elevation: Multiple Testing



(c) NO_2: Multiple Testing



(d) PM2.5: Multiple Testing

Figure 4.7: The optimum sample design generated in SAA of multiple hypothesis testing case.

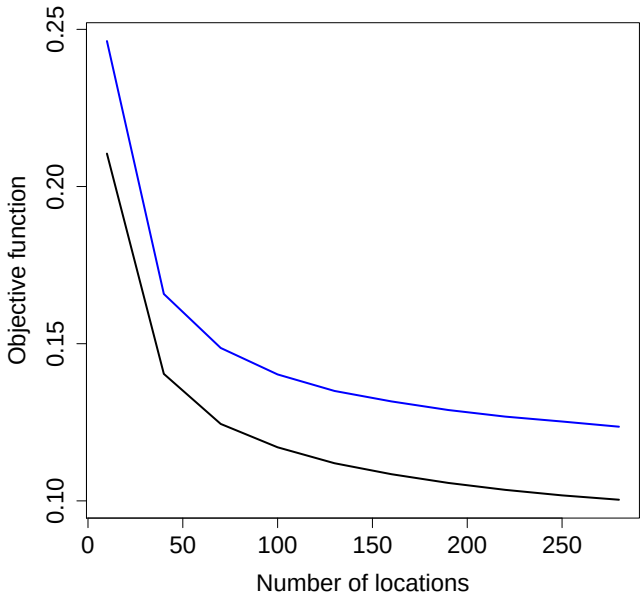


Figure 4.8: Bayesian risk as a function of the sample size for multiple hypothesis case.

Similar to the single hypothesis test, the relationship between the Bayesian risk of the optimal design and sample size is studied and plotted in Figure 4.8. The black line represents the objective function (4.12), while the blue line represents equation (4.13) using the same optimal design. The initial designs are randomly generated. Due to the large difference of the Bayesian risks among all the single hypothesis tests, the Cauchy-Schwarz inequality cannot provide a good approximation for (4.12). The Bayesian risk decreases quickly from 10 locations to 100 locations then slows down. Sample size can then be determined based on this curve for multiple hypothesis testing.

4.5 Conclusion

In this article several new optimality criteria based on Bayesian risk have been introduced specifically for the problem of single and multiple hypothesis tests. The optimal design method for the problem of generating sampling designs for survey studies of

spatial datasets has been discussed, where the objective is to estimate the slope parameters. Assuming known spatial correlation and the uncertainty about the measurement error, optimal sampling schemes can be generated. An air pollutant data set over part of California areas was used to demonstrate the method, but it is general and can be straightforwardly extended to the study of other spatial phenomena in a straightforward manner and to larger spatial scales.

The effect of the dimension of the sampling design on the Bayesian risk of the optimal design generated by SAA was studied. The plot of the curve that represents the optimal value (found by SAA) of the Bayesian risk as a function of the locations, leads to the sample size determination. By setting a cut-off value, the smallest sample size that satisfies the constant is chosen. And the corresponding design is the optimum design that is desired.

Deliberately, the problem of variogram estimation and the uncertainty caused by estimating the variance and spatial correlation parameters were not considered. There are more interesting questions haven't addressed yet, for instance, how the strength of spatial correlation affect the sample size determination and design.

References

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. Wiley, New York, NY, third edition.

Attias, H. (2000). Variational bayesian framework for graphical models. In *In proceeding: Advances in Neural Information Processing Systems*. MIT Press, Cambridge, MA.

Banerjee, S., Carlin, B. P., and Gelfand, A. E. (2004). *Hierarchical Modeling and Analysis for Spatial Data*. CHAPMAN & HALL/CRC.

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

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

Berger, J. O. (1985). *Statistics Statistical Decision Theory and Bayesian Analysis*. Springer, New York, 2nd edition.

Berger, J. O. and Pericchi, L. R. (2001). Objective bayesian methods for model selection: Introduction and comparison. *Lecture Notes-Monograph Series* **38,** 135–207.

Best, N., Marshall, C., and Thomas, A. (2000). Spatial modeling using winbugs and geobugs, short course. Brisbane.

Bishop, C. M. (1999). Latent variable models. In Jordan, M. I., editor, *Learning in Graphical Models*, pages 371–403, Cambridge, MA, USA. MIT Press.

Bolin, D. and Lindgren, F. (2011). Spatial models generated by nested stochastic partial differential equations, with an application to global ozone mapping. *The Annals of Applied Statistics* **5 (1),** 523–550.

Buttafuoco, G., Castrignan, A., Colecchia, A. S., and Ricca, N. (2010). Delineation of management zones using soil properties and a multivariate geostatistical approach. *The Italian Journal of Agronomy* **4,** 323–332.

Cai, B. and Dunson, D. B. (2006). Bayesian covariance selection in generalized linear mixed models. *Biometrics* **62,** 446–457.

Carlin, B. P. and Louis, T. A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hall, London.

Castrignanó, A., Cherubini, C., abd M. Castore, C. G., Mucci, G. D., and Molinari, M. (2005). Using multivariate geostatistics for describing spatial relationships among some soil properties. In *ISTRO Conference Brno*.

Castrignanó, A., Goovaerts, P., Lulli, L., and Bragato, G. (2000). A geostatistical approach to estimate probability of occurrence of tuber melanosporum in relation to some soil properties. *Geoderma* **98,** 95–113.

Chaloner, K. (1984). Optimal bayesian experimental design for linear models. *The Annals of Statistics* **12,** 283–300.

Chaloner, K. and Verdinelli, I. (1995). Bayesian experimental design: A review. *Statistical Science* **10,** 273–304.

Chen, Z. and Dunson, D. B. (2003). Random effects selection in linear mixed models. *Biometrics* **59 (4),** 762–769.

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

Christensen, W. and Amemiya, Y. (2003). Modeling and prediction for multivariate spatial factor analysis. *Journal of Statistics and Inferense* **115,** 543 – 564.

Christensen, W. F. and Amemiya, Y. (2001). Generalized shifted-factor analysis method for multivariate geo-referenced data. *Mathematical Geology* **33,** 801 – 824.

Christensen, W. F. and Amemiya, Y. (2002). Latent variable analysis of multivariate spatial data. *Journal of the American Statistical Association* **97,** 302 – 317.

Clyde, M. and Chaloner, K. (1996). The equivalence of constrained and weighted designs in multiple objective design problems. *Journal of the American Statistical Association* **91,** 1236–1244.

Cressie, N., Gotway, C. A., and Grondona, M. O. (1990). Spatial prediction from networks. *Chemometrics and Intelligent Laboratory Systems* **7,** 251–271.

Cressie, N. A. C. (1999). *Statistics for Spatial Data*. Wiley.

Cressie, N. A. C. and Johannesson, G. (2008). Fixed rank kriging for very large spatial data sets. *Mathematical Geology* **70,** 209 – 226.

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

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

Dominici, F., Peng, R. D., Bell, M. L., Pham, L., McDermott, A., Zeger, S. L., and Samet, J. M. (2006). Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. *The Journal Of The American Medical Association* **295,** 1127–1134.

Dunson, D. B. (2006). Efficient bayesian model averaging in factor analysis. Isds discussion paper, Duke University.

Fedorov, V. V. (1972). *Theory of optimal experiments*. Translated and edited by W.J. Studden and E.M. Klimko, Academic Press, New York.

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* **1,** 60–83.

Finley, A. O., Sang, H., Banerjee, S., and Gelfand, A. E. (2009). Improving the performance of predictive process modeling for large datasets. *Computational Statistics & Data Analysis* **53,** 2873 – 2884.

Friedman, N. (1998). The bayesian structural em algorithm. In *Fourteenth Conference on Uncertainty in Artificial Intelligence*, pages 129–138. Morgan Kaufmann.

Fuentes, M. (2007). Approximate likelihood for large irregularly spaced spatial data. *Journal of the American Statistical Association* **102(477),** 321–331.

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

Gelfand, A. E. and Ghosh, S. K. (1998). Model choice: A minimum posterior predictive loss approach. *Biometrika* **85,** 1 – 11.

Gelfand, A. E., Kim, H.-J., Sirmans, C., and Banerjee, S. (2003). Spatial modelling with spatially varying coefficient processes. *Journal of the American Statistical Association* **98,** 378–396.

Gelfand, A. E., Schmidt, A. M., Banerjee, S., and Sirmans, C. F. (2004). Nonstationary multivariate process modeling through spatially varying coregionalization. *TEST: An Official Journal of the Spanish Society of Statistics and Operations Research* **13,** 263–312.

Gelfand, I. M. and Fomin, S. (1963). *Calculus of Variations*. rentice-Hall, Inc, Englewood CliKs, New Jersey.

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

Ghahramani, Z. and Beal, M. J. (2000). Variational inference for bayesian mixtures of factor analysers. In *In Advances in Neural Information Processing Systems 12*, pages 449–455. MIT Press.

Grzebyk, M. and Wackernagel, H. (1994). Multivariate analysis and spatial/temporal scales: Real and complex models. In *Proceedings of the XVIIth International Biometrics Conference*.

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

Hartigan, J. A. and Wong, M. A. (1979). A k-means clustering algorithm. *Applied Statistics* **28,** 100–108.

Higdon, D. (2001). Space and space time modeling using process convolutions. Technical report, Duke University.

Hinton, G. E. and van Camp, D. (1993). Keeping the neural networks simple by minimizing the description length of the weights. In *COLT '93: Proceedings of the sixth annual conference on Computational learning theory*, pages 5–13, New York, NY, USA. ACM Press.

Hodges, J. S. and Reich, B. J. (2010). Adding spatially-correlated errors can mess up the fixed effect you love. *The American Statistician* **64,** 325–334.

Hogan, J. W. and Tchernis, R. (2004). Bayesian factor analysis for spatially correlated data, with application to summarizing area-level material deprivation from census data. *Journal of the American Statistical Association* **99,** 314 – 324.

Inoue, L. Y. T., Berry, D. A., and Parmigiani, G. (2005). Relationship between bayesian and frequentist sample size determination. *American Statistician* **59,** 79–87.

Irvine, K. M., Gitelman, A. I., and Hoeting, J. A. (2007). Spatial designs and properties of spatial correlation: Effects on covariance estimation. *Journal of Agricultural Biological and Environmental Statistics* **12,** 450–469.

Jin, X., Banerjee, S., and Carlin, B. P. (2007). Order-free co-regionalized areal data models with application to multiple-disease mapping. *Journal Of The Royal Statistical Society Series B* **69 (5),** 817–838.

Jones, R. and Zhang, Y. (1997). Models for continuous stationary space-time processes. In *Modeling Longitudinal and Spatially Correlated Data: Methods, Applications and Future Directions*. Springer-Verlag.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1998). An introduction to variational methods for graphical models. *Machine Learning* **37,** 183–233.

Kamman, E. and Wand, M. (2003). Geoadditive models. *Applied Statistics* **52,** 1–18.

Kiefer, J. (1959). Optimum experimental designs. *Journal of the Royal Statistical Society, Series B* **21,** 272–319.

Kirkpatrick, S. (1984). Optimization by simulated annealing: Quantitative studies. *Journal of Statistical Physics* **34,** 975–986.

Kuo, L. and Mallick, B. (1998). Variable selection for regression models. *The Indian Journal of Statistics* **60,** 65–81.

Lark, R. (2002). Optimized spatial sampling of soil for estimation of the variogram by maximum likelihood. *Geoderma* **105,** 49–80.

Lark, R. and Papritz, A. (2003). Fitting a linear model of coregionalization for soil properties using simulated annealing. *Geoderma* **115,** 245–260.

Li, J. (2009). *Spatial multivariate design in the plane and on stream networks*. PhD thesis, University of Iowa.

Lin, X., Wahba, G., Xiang, D., Gao, F., Klein, R., and Klein, B. (2000). Smoothing spline anova models for large data sets with bernoulli observations and the randomized gacv. *Annals of Statistics* **28,** 1570–1600.

Lindgren, F., Lindström, J., and Rue, H. (2010). An explicit link between gaussian fields and gaussian markov random fields, the spde approach. *Preprints in Mathematical Sciences* **2010:3,** Lund University.

Lindgren, F., Rue, H., and Lindstrm, J. (2011). An explicit link between gaussian fields and gaussian markov random fields: the stochastic partial differential equation

approach. *Journal of the Royal Statistical Society - Series B: Statistical Methodology* **73,** 423–498.

Lindley, D. V. (1972). *Bayesian Statistics, A Review*. SIAM, Philadelphia.

Liu, X., Wall, M. M., and Hodges, J. S. (2005). Generalized spatial structural equation models. *Biostatistics* **6,** 539 –557.

Lopes, H. F. and West, M. (1999). Model uncertainty in factor analysis. Technical report, Institute of Statistics and Decision Sciences, Duke University.

Lopes, H. F. and West, M. (2004). Bayesian model assessment in factor analysis. *Statistica Sinica* **14,** 41 – 67.

Mackay, J. C. D. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.

MacKay, J. D. (1997). Ensemble learning for hidden markov models. Technical report, Cavendish Laboratory, University of Cambridge.

Marshall, A. W. and Olkin, I. (1979). *Inequalities: Theory of Majorization and Its Applications*. Academic Press.

Matheron, G. (1982). Pour une analyse krigeante des donnes regionalises. *Centre de Geostatistique* page N 732.

McBratney, A. B., Webster, R., and Burgess, T. M. (1981). The design of optimal sampling schemes for local estimation and mapping of regionalized variables - i. *Computers & Geosciences* **7,** 331–334.

Minozzo, M. and Fruttini, D. (2004). Loglinear spatial factor analysis: an application to diabetes mellitus complications. *Environmetrics* **15,** 423–434.

Müller, P. (1999). *Bayesian Statistics*, chapter Simulation based optimal design (with discussion). Clarendon Press, Oxford.

Müller, P., Parmigiani, G., Robert, C., and Rousseau, J. (2004). Optimal sample size for multiple testing: the case of gene expression microarrays. *Journal of the American Statistical Association* **99,** 990–1001.

Müller, W. (2005). A comparison of spatial design methods for correlated observations. *Environmetrics* **16,** 495–505.

Müller, W. G. and Zimmerman, D. L. (1999). Optimal designs for variogram estimation. *Environmetrics* **10,** 23–37.

Neal, R. and Hinton, G. E. (1998). *A View Of The Em Algorithm That Justifies Incremental, Sparse, And Other Variants*. Kluwer Academic Publishers.

O'Hagan, A. and Stevens, J. W. (2001). Bayesian assessment of sample size for clinical trials of cost-effectiveness. *Medical Decision Making* **21,** 219–230.

Paciorek, C. J. (2007). Computational techniques for spatial logistic regression with large datasets. *Computational Statistics and Data Analysis* **51(8),** 3631–3653.

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

Pilz, J. (1991). *Bayesian Estimation and Experimental Design in Linear Regression Models*. Wiley, New York.

Romary, T., Malherbe, L., and de Foupuet, C. (2012). Optimal spatial design for air quality measurrment surveys. *submitted to Environmetrics* .

Royle, J. A. and Berliner, L. M. (1999). A hierarchical approach to multivariate spatial modeling and prediction. *Journal of Agricultural Biological and Environmental Statistics* **4,** 29–56.

Royle, J. A. and Nychka, D. (1998). An algorithm for the construction of spatial coverage designs with implementation in splus. *Computers & Geosciences* **24,** 479–488.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63(3),** 581–592.

Rubin, D. B. (1987). A noniterative sampling/importance resampling alternative to the data augmentation algorithm for creating a few imputations when fractions of missing information are modest: the sir algorithm. *Journal of the American Statistical Association* **82,** 543–546.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. Chapman and Hall/CRC Press.

Rue, H. and Tjelmeland, H. (2002). Fitting gaussian markov random fields to gaussian fields. *Scandinavian Journal of Statistics* **29,** 31–49.

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

Smith, M. and Kohn, R. (2002). Parsimonious covariance matrix estimation for longitudinal data. *Journal of the American Statistical Association* **97,** 1141–1153.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Linde, A. v. d. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B* **64(4),** 583–639.

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

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

Tanner, M. A. (1993). *Tools for statistical inference. Methods for the exploration of posterior distributions and likelihood functions*. Springer-Verlag, New York.

Toman, B. (1996). Bayesian experimental design for multiple hypothesis testing. *Journal of the American Statistical Association* **91,** 185– 190.

van Groenigen, J. and Stein, A. (1998). Constrained optimization of spatial sampling using continuous simulated annealing. *Journal of Environmental Quality* **27,** 1078–1086.

Vecchia, A. (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). Consructing and fitting models for cokriging and multivariable spatial prediction. *Journal of Statistical Planning and Inference* **69,** 275–294.

Ver Hoef, J. M., Cressie, N., and Barry, R. P. (2004). Flexible spatial models based on the fast fourier transform (fft) for cokriging. *Journal of Computational and Graphical Statistics* **13,** 265–282.

Wackernagel, H. (2003). *Multivariate Geostatistics: An Introduction With Applications*. Springer-Verlag Telos, Berlin.

Wang, F. and Wall, M. M. (2003). Generalized common spatial factor model. *Biostatistics* **4,** 569–582.

Waterhouse, S., MacKay, D., and Robinson, A. J. (1995). Bayesian methods for mixtures of experts. In Weiss, Y., Schölkopf, B., and Platt, J., editors, *Advances in Neural Information Processing Systems*, pages 351–357, Cambridge, MA. MIT Press.

Webster, R., Atteias, O., and Duboiss, J. P. (1994). Coregionalization of trace metals in the soil in the swiss jura. *European Journal of Soil Science* **45,** 205–218.

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

Xia, G. and Gelfand, A. E. (2006). Stationary process approximation for the analysis of large spatial datasets. Technical report, Duke University.

Yfantis, E. A., Flatman, G. T., and Behar, J. V. (1987). Efficiency of kriging estimation for square, triangular, and hexagonal grids. *Mathematical Geology* **19,** 183–205.

Zhang, H. (2004). Inconsistent estimation and asymptotically equal 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.

Zhu, Z. and Stein, M. (2005). Spatial sampling design for parameter estimation of the covariance function. *Journal of Statistical Planning and Inference* **134,** 583–603.

Zimmerman, D. L. (2006). Optimal network design for spatial prediction, covariance parameter estimation, and empirical prediction. *Environmetrics* **17,** 635–652.

Appendix A

Appendix for Chapter 1

VB Estimator

We wish to maximize the function:

$$\mathcal{L}(q) = \int q(\boldsymbol{\theta}) \log \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\, d\boldsymbol{\theta},$$

with respect to each factorized distribution in turn. $\mathcal{L}$ is a functional, i.e. $\mathcal{L} = \int f(\boldsymbol{\theta}, q(\boldsymbol{\theta}))\, d\boldsymbol{\theta}$. Hence to maximize $\mathcal{L}$ we need to turn to the calculus of variations. Let

$$\mathcal{Q} = \left\{ q(\boldsymbol{\theta}) : q(\boldsymbol{\theta}) = \prod_{i=1}^{m} q_i(\boldsymbol{\theta}_i) \right\} .$$

Then $\mathcal{L}(q)$ for $q(\boldsymbol{\theta}) \in \mathcal{Q}$ can be written as :

$$\mathcal{L}(q) = \int g\left(\boldsymbol{\theta}_i, q_i(\boldsymbol{\theta}_i)\right)\, d\boldsymbol{\theta}_i \ ,$$

where:

$$g\left(\boldsymbol{\theta}_i, q_i(\boldsymbol{\theta}_i)\right) = \int f(\boldsymbol{\theta}, q(\boldsymbol{\theta}))\, d\boldsymbol{\theta}_{j\neq i} \ . \tag{A.1}$$

From variational calculus the maximum of $\mathcal{L}(q)$ is the solution of the Euler-Lagrange differential equation:

$$\frac{\partial}{\partial q_i(\boldsymbol{\theta}_i)} \left[g\left(\boldsymbol{\theta}_i, q_i(\boldsymbol{\theta}_i)\right) \right] - \frac{d}{d\boldsymbol{\theta}_i} \left\{ \frac{\partial}{\partial \dot{q}_i(\boldsymbol{\theta}_i)} \left[g\left(\boldsymbol{\theta}_i, q_i(\boldsymbol{\theta}_i)\right) \right] \right\} = 0 \ , \tag{A.2}$$

where the second term is zero, in the case that g does not depend on $\dot{q}_i(\boldsymbol{\theta}_i)$. Using equation (A.1), equation (A.2) can be written as:

$$
\begin{aligned}
0 &= \frac{\partial}{\partial q_i(\boldsymbol{\theta}_i)} \int q(\boldsymbol{\theta}) \log \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})}\, d\boldsymbol{\theta}_{j\neq i} \\
&= \frac{\partial}{\partial q_i(\boldsymbol{\theta}_i)} \left[\int \prod_j q_j(\boldsymbol{\theta}_j) \log p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\, d\boldsymbol{\theta}_{j\neq i} - \int \prod_j q_j(\boldsymbol{\theta}_j) \sum_j \log q_j(\boldsymbol{\theta}_j)\, d\boldsymbol{\theta}_{j\neq i} \right] \\
&= \int \prod_{j\neq i} q_j(\boldsymbol{\theta}_j) \log p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\, d\boldsymbol{\theta}_{j\neq i} - \int \prod_{j\neq i} q_j(\boldsymbol{\theta}_j) \sum_j \log q_j(\boldsymbol{\theta}_j)\, d\boldsymbol{\theta}_{j\neq i} - \int \prod_j q_j(\boldsymbol{\theta}_j) \frac{1}{q_i(\boldsymbol{\theta}_i)}\, d\boldsymbol{\theta}_{j\neq i} \\
&= \int \prod_{j\neq i} q_j(\boldsymbol{\theta}_j) \log p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\, d\boldsymbol{\theta}_{j\neq i} - \log q_i(\boldsymbol{\theta}_i) - \int \prod_{j\neq i} q_j(\boldsymbol{\theta}_j) \left(1 + \sum_{j\neq i} \log q_j(\boldsymbol{\theta}_j) \right) d\boldsymbol{\theta}_{j\neq i}\,.
\end{aligned}
$$

Hence $\log q_i(\boldsymbol{\theta}_i) = \int \prod_{j\neq i} q_j(\boldsymbol{\theta}_j) \log p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\, d\boldsymbol{\theta}_{j\neq i} + \text{constant}.$

Appendix B

Appendix for Chapter 2

VB Algorithm for Univariate Spatial conditional Regression Model

Algorithm 1 Algorithm to carry out VB estimation for univariate spatial models treating spatial random effects as a latent variable.

Specify hyper parameters of the prior distributions for σ^2, τ^2 and ϕ.

Give initial values to the expectation of $1/\tau^2$, ϕ, $\mathbf{w}$ and $\mathbf{R}(\phi)^{-1}$: $\mathrm{E}^{(0)}(1/\tau^2) = (1/\tau^2)^{(0)}$, $\mathrm{E}^{(0)}(\phi) = \phi^{(0)}$, $\boldsymbol{\mu}_{\mathbf{w}}^{(0)} = \mathbf{0}$ and $\mathrm{E}^{(0)}(\mathbf{R}(\phi)^{-1}) = \mathbf{R}\left(\boldsymbol{\phi}^{(0)}\right)^{-1}$.

for $i = 1$ to t **do**

Step 1: Update the distribution of $\boldsymbol{\beta} \sim MVN\left(\boldsymbol{\mu}_{\beta}^{(i)}, \mathbf{V}_{\beta}^{(i)}\right)$, where
$\mathbf{V}_{\beta}^{(i)} = \left[\mathrm{E}^{(i-1)}\left(1/\tau^2\right)\right]^{-1}(\mathbf{X}'\mathbf{X})^{-1}$ and $\boldsymbol{\mu}_{\beta}^{(i)} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{w}}^{(i-1)})$.

Step 2: Update the distribution of $\tau^2 \sim IG$ with parameters $a_\tau + \frac{n}{2}$ and
$b_\tau + \frac{1}{2}\left[\mathrm{Tr}\left(\mathbf{V}_{\mathbf{w}}^{(i-1)}\right) + p\mathrm{E}^{(i-1)}\left(1/\tau^2\right) + \left(\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{w}}^{(i-1)}\right)'\left(\mathbf{I}_n - \mathbf{H}\right)\left(\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{w}}^{(i-1)}\right)\right]$,
where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$. Then, calculate $m_{\tau^2}^{(i)} = \mathrm{E}^{(i)}(1/\tau^2)$.

Step 3: Update the distribution of $\sigma^2 \sim IG$ with parameters $a_\sigma + \frac{n}{2}$ and
$b_\sigma + \frac{1}{2}\left\{\mathrm{Tr}\left[\mathrm{E}^{(i-1)}\left(\mathbf{R}\left(\phi\right)^{-1}\right)\mathbf{V}_{\mathbf{w}}^{(i-1)}\right] + \boldsymbol{\mu}_{\mathbf{w}}^{(i-1)'}\mathrm{E}^{(i-1)}\left(\mathbf{R}\left(\phi\right)^{-1}\right)\boldsymbol{\mu}_{\mathbf{w}}^{(i-1)}\right\}$. Then, calculate $m_{\sigma^2}^{(i)} = \mathrm{E}^{(i)}(1/\sigma^2)$.

Step 4: Update the distribution of $\mathbf{w} \sim MVN\left(\boldsymbol{\mu}_{\mathbf{w}}^{(i)}, \mathbf{V}_{\mathbf{w}}^{(i)}\right)$, where
$\mathbf{V}_{\mathbf{w}}^{(i)} = \left[m_{\sigma^2}^{(i)}\mathrm{E}^{(i-1)}\left(\mathbf{R}\left(\phi\right)^{-1}\right) + m_{\tau^2}^{(i)}\mathbf{I}_n\right]^{-1}$ and
$\boldsymbol{\mu}_{\mathbf{w}}^{(i)} = m_{\tau^2}^{(i)}\left[m_{\sigma^2}^{(i)}\mathrm{E}^{(i-1)}\left(\mathbf{R}\left(\phi\right)^{-1}\right) + m_{\tau^2}^{(i)}\mathbf{I}_n\right]^{-1}\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\mu}_{\beta}^{(i)}\right)$.

Step 5: Update the distribution of ϕ which is proportional to

$$|\mathbf{R}(\phi)|^{-\frac{1}{2}}\exp\left\{-\frac{m_{\sigma^2}^{(i)}\left[\mathrm{Tr}\left(\mathbf{R}(\phi)^{-1}\mathbf{V}_{\mathbf{w}}^{(i)}\right) + \boldsymbol{\mu}_{\mathbf{w}}^{(i)'}\mathbf{R}(\phi)^{-1}\boldsymbol{\mu}_{\mathbf{w}}^{(i)}\right]}{2}\right\} \tag{B.1}$$

and calculate $\mathrm{E}^{(i)}\left(\mathbf{R}(\phi)^{-1}\right)$.

end for

VB Algorithm for Univariate Marginal Spatial Regression Model

Algorithm 2 Algorithm to carry out VB estimation for marginal univariate spatial model.

Specify hyper parameters of the prior distribution for τ^2, r and ϕ.

Give initial values to the expectation of $1/\tau^2$, ϕ, r and $\mathbf{C}(\phi, r)^{-1}$: $\mathrm{E}^{(0)}(1/\tau^2) = (1/\tau^2)^{(0)}$, $\mathrm{E}^{(0)}(r) = r^{(0)}$, $\mathrm{E}^{(0)}(\phi) = \phi^{(0)}$ and $\mathrm{E}^{(0)}(\mathbf{C}(\phi, r)^{-1}) = \mathbf{C}\left(\boldsymbol{\phi}^{(0)}, r^{(0)}\right)^{-1}$.

for $i = 1$ to t **do**

Step 1: Update the distribution of $\boldsymbol{\beta} \sim MVN\left(\boldsymbol{\mu}_\beta^{(i)}, \mathbf{V}_\beta^{(i)}\right)$

$\mathbf{V}_\beta^{(i)} = \left[\mathrm{E}^{(i-1)}\left(1/\tau^2\right)\right]^{-1}\left[\mathbf{X}'\mathrm{E}^{(i-1)}\left(\mathbf{C}^{-1}\right)\mathbf{X}\right]^{-1}$ and

$\boldsymbol{\mu}_\beta^{(i)} = \left[\mathbf{X}'\mathrm{E}^{(i-1)}\left(\mathbf{C}^{-1}\right)\mathbf{X}\right]^{-1}\mathbf{X}'\mathrm{E}^{(i-1)}\left(\mathbf{C}^{-1}\right)\mathbf{Y}$.

Step 2: Update the distribution of $\tau^2 \sim IG$ with parameters $a_\tau + \frac{n}{2}$ and

$b_\tau + \frac{1}{2}\left[\mathrm{Tr}\left(\mathbf{X}'\mathrm{E}^{(i-1)}\left(\mathbf{C}^{-1}\right)\mathbf{X}\mathbf{V}_\beta^{(i)}\right) + \left(\mathbf{X}\boldsymbol{\mu}_\beta^{(i)} - \mathbf{Y}\right)'\mathrm{E}^{(i-1)}\left(\mathbf{C}^{-1}\right)\left(\mathbf{X}\boldsymbol{\mu}_\beta^{(i)} - \mathbf{Y}\right)\right]$;

calculate $\mathrm{E}^{(i)}(1/\tau^2)$.

Step 3: Update the joint distribution of ϕ and r, which is proportional to

$$|\mathbf{C}|^{-\frac{1}{2}} \times \exp\left\{\mathrm{E}^{(i)}\left(1/\tau^2\right)\left[-\frac{\mathrm{Tr}\left(\mathbf{X}'\mathbf{C}^{-1}\mathbf{X}\mathbf{V}_\beta^{(i)}\right) + \left(\mathbf{X}\boldsymbol{\mu}_\beta^{(i)} - \mathbf{Y}\right)'\mathbf{C}^{-1}\left(\mathbf{X}\boldsymbol{\mu}_\beta^{(i)} - \mathbf{Y}\right)}{2}\right]\right\}$$

and calculate $\mathrm{E}^{(i)}\left(\mathbf{C}^{-1}\right)$.

end for

Derivations of the VB Algorithm for Multivariate Spatial Model

The prior distributions of the parameters are specified as: $\boldsymbol{\beta} \sim flat$, $\mathbf{A}\mathbf{A}' \sim Inverse - Wishart(df, \mathbf{S})$, $\delta_i^2 = \Psi_i^{-1} \sim Gamma(a_i, b_i)$, $\phi_i \sim Uniform(0.06, 3)$. Then the marginal likelihood is

$$p(\mathbf{Y}) = \int p(\mathbf{Y} \mid \tilde{\mathbf{w}}, \boldsymbol{\theta}) p(\tilde{\mathbf{w}} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta})\, d\tilde{\mathbf{w}}\, d\boldsymbol{\theta},$$

where $\boldsymbol{\theta} = (\boldsymbol{\beta}, \mathbf{A}, \boldsymbol{\Psi}, \boldsymbol{\phi})$. Let Q be

$$
\begin{aligned}
Q =& \ln\left[p(\mathbf{Y}, \tilde{\mathbf{w}}, \boldsymbol{\theta})\right] \\
=& \ln p(\mathbf{Y} \,|\, \tilde{\mathbf{w}}, \mathbf{A}, \boldsymbol{\Psi}, \boldsymbol{\phi}) + \ln p(\tilde{\mathbf{w}} \,|\, \boldsymbol{\phi}) + \ln p(\boldsymbol{\Psi}) + \ln p(\mathbf{A}\mathbf{A}') + \ln p(\boldsymbol{\phi}) \\
=& \frac{n}{2}\sum_i \ln \delta_i^2 - \frac{(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathscr{A}\tilde{\mathbf{w}})' (\mathbf{I}_n \otimes \boldsymbol{\Psi})^{-1} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathscr{A}\tilde{\mathbf{w}})}{2} - \frac{1}{2}\ln|\Sigma_{\tilde{\mathbf{w}}}| - \frac{\tilde{\mathbf{w}}'{\Sigma_{\tilde{\mathbf{w}}}}^{-1}\tilde{\mathbf{w}}}{2} \\
&+ \sum_i \left[(a_i - 1)\ln\delta_i^2 - \frac{\delta_i^2}{b_i}\right] - \frac{\mathrm{Tr}\left[(\mathbf{A}\mathbf{A}')^{-1}\mathbf{S}\right]}{2} - \frac{df + m + 1}{2}\ln|\mathbf{A}\mathbf{A}'| + c,
\end{aligned}
$$

where c is constant. Assuming the densities of all the parameters at t iteration are known,then the distribution function of $\boldsymbol{\beta}$ at next iteration is,

$$
\begin{aligned}
q^{(t+1)}(\boldsymbol{\beta}) \propto& \exp\left[\int d\boldsymbol{\theta}_\beta q^{(t)}(\boldsymbol{\theta}_\beta) Q\right] \quad (\boldsymbol{\theta}_\beta \text{ means all the parameters except } \boldsymbol{\beta}) \\
\propto& \exp\left\{\int q^{(t)}(\tilde{\mathbf{w}}, \mathbf{A}, \boldsymbol{\Psi}) \left[-\frac{(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathscr{A}\tilde{\mathbf{w}})' (\mathbf{I}_n \otimes \boldsymbol{\Psi})^{-1} (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathscr{A}\tilde{\mathbf{w}})}{2}\right] d\mathbf{A}\, d\tilde{\mathbf{w}}\, d\boldsymbol{\Psi}\right\} \\
\propto& \exp\left\{\int q^{(t)}(\tilde{\mathbf{w}}, \mathbf{A}) \left[-\frac{(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathscr{A}\tilde{\mathbf{w}})' \left[\mathbf{I}_n \otimes \mathrm{E}^{(t)}\left(\boldsymbol{\Psi}^{-1}\right)\right] (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathscr{A}\tilde{\mathbf{w}})}{2}\right] d\mathbf{A}\, d\tilde{\mathbf{w}}\right\} \\
\propto& \exp\left\{\int q^{(t)}(\mathbf{A}) \left[-\frac{\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathscr{A}\boldsymbol{\mu}_{\tilde{\mathbf{w}}}^{(t)}\right)' \left[\mathbf{I}_n \otimes \mathrm{E}^{(t)}\left(\boldsymbol{\Psi}^{-1}\right)\right] \left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathscr{A}\boldsymbol{\mu}_{\tilde{\mathbf{w}}}^{(t)}\right)}{2}\right] d\mathbf{A}\right\} \\
\propto& \exp\left\{\left[-\frac{\boldsymbol{\beta}'\mathbf{X}' \left[\mathbf{I}_n \otimes \mathrm{E}^{(t)}(\boldsymbol{\Psi}^{-1})\right] \mathbf{X}\boldsymbol{\beta} - 2\boldsymbol{\beta}'\mathbf{X}' \left[\mathbf{I}_n \otimes \mathrm{E}^{(t)}\left(\boldsymbol{\Psi}^{-1}\right)\right] \left(\mathbf{Y} - \mathrm{E}^{(t)}(\mathscr{A})\boldsymbol{\mu}_{\tilde{\mathbf{w}}}^{(t)}\right)}{2}\right]\right\} \\
\sim& MVN\left(\boldsymbol{\mu}_\beta^{(t+1)}, \mathbf{V}_\beta^{(t+1)}\right),
\end{aligned}
$$

where

$$
\begin{aligned}
\boldsymbol{\mu}_\beta^{(t+1)} &= \left\{\mathbf{X}' \left[\mathbf{I}_n \otimes \mathrm{E}^{(t)}\left(\boldsymbol{\Psi}^{-1}\right)\right] \mathbf{X}\right\}^{-1} \mathbf{X}' \left[\mathbf{I}_n \otimes \mathrm{E}^{(t)}\left(\boldsymbol{\Psi}^{-1}\right)\right] \left[\mathbf{Y} - \mathrm{E}^{(t)}(\mathscr{A})\boldsymbol{\mu}_{\tilde{\mathbf{w}}}^{(t)}\right] \quad \text{and} \\
\mathbf{V}_\beta^{(t+1)} &= \left\{\mathbf{X}' \left[\mathbf{I}_n \otimes \mathrm{E}^{(t)}\left(\boldsymbol{\Psi}^{-1}\right)\right] \mathbf{X}\right\}^{-1}.
\end{aligned}
$$

To update the distribution of $\tilde{\mathbf{w}}$, we have:

$$
\begin{aligned}
q^{(t+1)}(\tilde{\mathbf{w}}) &\propto \exp\left\{\int q^{(t)}(\mathbf{A},\boldsymbol{\Psi}) q^{(t+1)}(\boldsymbol{\beta}) \left[-\frac{(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}-\mathscr{A}\tilde{\mathbf{w}})'\left(\mathbf{I}_n\otimes\boldsymbol{\Psi}\right)^{-1}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}-\mathscr{A}\tilde{\mathbf{w}})}{2}\right] d\boldsymbol{\Psi}\, d\mathbf{A}\right\} \\
&\quad \times \exp\left\{\int q^{(t)}(\boldsymbol{\phi})\left(-\frac{\tilde{\mathbf{w}}'{\Sigma_{\tilde{\mathbf{w}}}}^{-1}\tilde{\mathbf{w}}}{2}\right) d\boldsymbol{\phi}\right\} \\
&\propto \exp\left\{\int q^{(t)}(\mathbf{A})\left[-\frac{\left(\mathbf{Y}-\mathbf{X}\boldsymbol{\mu}_\beta^{(t+1)}-\mathscr{A}\tilde{\mathbf{w}}\right)'\left[\mathbf{I}_n\otimes \mathrm{E}^{(t)}\left(\boldsymbol{\Psi}^{-1}\right)\right]\left(\mathbf{Y}-\mathbf{X}\boldsymbol{\mu}_\beta^{(t+1)}-\mathscr{A}\tilde{\mathbf{w}}\right)}{2}\right] d\mathbf{A}\right\} \\
&\quad \times \exp\left\{-\frac{\tilde{\mathbf{w}}'\mathrm{E}^{(t)}(\Sigma_{\tilde{\mathbf{w}}}^{-1})\tilde{\mathbf{w}}}{2}\right\} \\
&\quad \sim MVN\left(\mu_{\tilde{\mathbf{w}}}^{(t+1)}, \mathbf{V}_{\tilde{\mathbf{w}}}^{(t+1)}\right),
\end{aligned}
$$

where

$$
\begin{aligned}
\mathbf{V}_{\tilde{\mathbf{w}}}^{(t+1)} &= \left\{\mathrm{E}^{(t)}\left[\mathscr{A}'\left(\mathbf{I}_n\otimes\mathrm{E}^{(t)}\left(\boldsymbol{\Psi}^{-1}\right)\right)\mathscr{A}\right]+\mathrm{E}^{(t)}\left(\Sigma_{\tilde{\mathbf{w}}}^{-1}\right)\right\}^{-1} \\
\boldsymbol{\mu}_{\tilde{\mathbf{w}}}^{(t+1)} &= \mathbf{V}_{\tilde{\mathbf{w}}}^{(t+1)}\mathrm{E}^{(t)}(\mathscr{A}')\left[\mathbf{I}_n\otimes\mathrm{E}^{(t)}\left(\boldsymbol{\Psi}^{-1}\right)\right]\left(\mathbf{Y}-\mathbf{X}\boldsymbol{\mu}_\beta^{(t+1)}\right).
\end{aligned}
$$

Notice that to get the updated hyper-parameters for the density function of $\tilde{\mathbf{w}}$ we need to calculate the expectation of $\Sigma_{\tilde{\mathbf{w}}}^{-1}$, $\mathscr{A}$ and $\mathscr{A}'\left[\mathbf{I}_n\otimes\mathrm{E}^{(t)}\left(\boldsymbol{\Psi}^{-1}\right)\right]\mathscr{A}$. Next, we can update the distribution of $\boldsymbol{\Psi}$.

$$
\begin{aligned}
q^{(t+1)}(\boldsymbol{\Psi}) &\propto \exp\left\{\frac{n}{2}\sum_i \ln\delta_i^2+\sum_i\left[(a_i-1)\ln\delta_i^2-\frac{\delta_i^2}{b_i}\right]\right\} \\
&\quad \times \exp\left\{\int q^{(t)}(\mathbf{A}) q^{(t+1)}(\boldsymbol{\beta},\tilde{\mathbf{w}})\left[-\frac{(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}-\mathscr{A}\tilde{\mathbf{w}})'\left(\mathbf{I}_n\otimes\boldsymbol{\Psi}\right)^{-1}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}-\mathscr{A}\tilde{\mathbf{w}})}{2}\right] d\mathbf{A}\, d\boldsymbol{\beta}\, d\tilde{\mathbf{w}}\right\} \\
&\propto \exp\left\{\sum_i\left[(a_i+n/2-1)\ln\delta_i^2-\frac{\delta_i^2}{b_i}\right]-\int q^{(t)}(\mathbf{A})\frac{\mathrm{Tr}[\mathbf{B}^{(t+1)}]}{2}\, d\mathbf{A}\right\} \\
&\quad \times \exp\left\{\int q^{(t)}(\mathbf{A})\left[-\frac{\left(\mathbf{Y}-\mathbf{X}\boldsymbol{\mu}_\beta^{(t+1)}-\mathscr{A}\boldsymbol{\mu}_{\tilde{\mathbf{w}}}^{(t+1)}\right)'\left(\mathbf{I}_n\otimes\boldsymbol{\Psi}^{-1}\right)\left(\mathbf{Y}-\mathbf{X}\boldsymbol{\mu}_\beta^{(t+1)}-\mathscr{A}\boldsymbol{\mu}_{\tilde{\mathbf{w}}}^{(t+1)}\right)}{2}\right] d\mathbf{A}\right\} \\
&\propto \exp\left\{\sum_i\left[(a_i+n/2-1)\ln\delta_i^2-\frac{\delta_i^2}{b_i}\right]-\frac{\mathrm{Tr}\left[\left(\mathbf{I}_n\otimes\boldsymbol{\Psi}^{-1}\right)\mathbf{D}^{(t+1)}\right]}{2}\right\},
\end{aligned}
$$

where $\mathbf{B}^{(t+1)} = \left[\left(\mathbf{X}\mathbf{V}_{\beta}^{(t+1)}\mathbf{X}' + \mathscr{A}\mathbf{V}_{\tilde{\mathbf{w}}}^{(t+1)}\mathscr{A}'\right)\left(\mathbf{I}_n \otimes \mathbf{\Psi}^{-1}\right)\right]$ and

$$\begin{aligned}\mathbf{D}^{(t+1)} = &\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\mu}_{\beta}^{(t+1)}\right)\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\mu}_{\beta}^{(t+1)}\right)' - 2\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\mu}_{\beta}^{(t+1)}\right)\boldsymbol{\mu}_{\tilde{\mathbf{w}}}^{(t+1)'}\mathrm{E}^{(t)}(\mathscr{A})' \\ &+ \mathrm{E}^{(t)}\left(\mathscr{A}\boldsymbol{\mu}_{\tilde{\mathbf{w}}}^{(t+1)}\boldsymbol{\mu}_{\tilde{\mathbf{w}}}^{(t+1)'}\mathscr{A}'\right) + \mathbf{X}\mathbf{V}_{\beta}^{(t+1)}\mathbf{X}' + \mathrm{E}^{(t)}\left(\mathscr{A}\mathbf{V}_{\tilde{\mathbf{w}}}^{(t+1)}\mathscr{A}'\right) .\end{aligned}$$

Since the measurement errors are assumed to be independent, $\mathbf{\Psi}^{-1}$ is a diagonal matrix, so is $\mathbf{I}_n \otimes \mathbf{\Psi}^{-1}$, with diagonal elements δ_i^2, $i = 1, \ldots, m$. The trace of $(\mathbf{I}_n \otimes \mathbf{\Psi}^{-1})\mathbf{D}^{(t+1)}$ only depends on the diagonal elements of $\mathbf{D}^{(t+1)}$. And it can be written as $\sum_{i=1}^{m} d_i^{(t+1)}\delta_i^2$, where

$$d_i^{(t+1)} = \sum_{j=0}^{n-1} \mathbf{D}_{jm+i,jm+i}^{(t+1)}. \tag{B.2}$$

Then the distribution of δ_i is

$$q^{(t+1)}(\delta_i^2) \sim Gamma\left(a_i + n/2, \left(\frac{1}{b_i} + d_i^{(t+1)}\right)^{-1}\right).$$

The distribution of spatial correlation parameter $\boldsymbol{\phi}$ at $t+1$ iteration is

$$\begin{aligned}q^{(t+1)}(\boldsymbol{\phi}) \propto &|\Sigma_{\tilde{\mathbf{w}}}|^{-\frac{1}{2}} \exp\left\{\int -\frac{\tilde{\mathbf{w}}'{\Sigma_{\tilde{\mathbf{w}}}}^{-1}\tilde{\mathbf{w}}}{2}\, d\tilde{\mathbf{w}}\right\} \times \prod_i \mathbf{I}(\phi_i \in (0.06, 3)) \\ \propto &|\Sigma_{\tilde{\mathbf{w}}}|^{-\frac{1}{2}} \exp\left\{-\frac{\mathrm{Tr}\left(\boldsymbol{\mu}_{\tilde{\mathbf{w}}}^{(t+1)'}\Sigma_{\tilde{\mathbf{w}}}^{-1}\boldsymbol{\mu}_{\tilde{\mathbf{w}}}^{(t+1)} + \Sigma_{\tilde{\mathbf{w}}}^{-1}\mathbf{V}_{\tilde{\mathbf{w}}}^{(t+1)}\right)}{2}\right\} \times \prod_i \mathbf{I}(\phi_i \in (0.06, 3)).\end{aligned}$$

The last parameter that needs to be updated is $\mathbf{A}$,

$$\begin{aligned}q^{(t+1)}(\mathbf{A}) \propto &\exp\left\{\frac{-\mathrm{Tr}\left[(\mathbf{A}\mathbf{A}')^{-1}\mathbf{S}\right]}{2}\right\} |\mathbf{A}\mathbf{A}'|^{-(df+m+1)/2} \\ &\times \exp\left\{\int q^{(t+1)}(\mathbf{\Psi}, \boldsymbol{\beta}, \tilde{\mathbf{w}})\left[-\frac{(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathscr{A}\tilde{\mathbf{w}})'\left(\mathbf{I}_n \otimes \mathbf{\Psi}\right)^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta} - \mathscr{A}\tilde{\mathbf{w}})}{2}\right] d\mathbf{\Psi}\, d\boldsymbol{\beta}\, d\tilde{\mathbf{w}}\right\} \\ &\propto |\mathbf{A}\mathbf{A}'|^{-(df+m+1)/2} \exp\left\{-\frac{Tr\left[(\mathbf{A}\mathbf{A}')^{-1}\mathbf{S} + \mathscr{A}\mathbf{V}_{\tilde{\mathbf{w}}}^{(t+1)}\mathscr{A}'\mathrm{E}^{(t+1)}\left(\mathbf{I}_n \otimes \mathbf{\Psi}^{-1}\right)\right]}{2}\right\} \\ &\times \exp\left\{-\frac{\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\mu}_{\beta}^{(t+1)} - \mathscr{A}\boldsymbol{\mu}_{\tilde{\mathbf{w}}}^{(t+1)}\right)'\mathrm{E}^{(t+1)}\left(\mathbf{I}_n \otimes \mathbf{\Psi}^{-1}\right)\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\mu}_{\beta}^{(t+1)} - \mathscr{A}\boldsymbol{\mu}_{\tilde{\mathbf{w}}}^{(t+1)}\right)}{2}\right\} .\end{aligned}$$

VB Algorithm for Multivariate Spatial Model

Algorithm 3 Algorithm to carry out VB estimation for multivariate spatial models.

Specify hyper parameters of prior distributions for $\boldsymbol{\Psi}^{-1}$, $\mathbf{A}$ and $\boldsymbol{\phi}$.

Give initial values to the expectation of $\boldsymbol{\Psi}^{-1}$, $\mathbf{A}$, $\boldsymbol{\phi}$, $\tilde{\mathbf{w}}$, $\Sigma_{\tilde{\mathbf{w}}}(\boldsymbol{\phi})^{-1}$ and $\mathbf{A}'\boldsymbol{\Psi}^{-1}\mathbf{A}$: $\mathrm{E}^{(0)}(\delta_i^2) = \delta_i^{2(0)}$, $\mathrm{E}^{(0)}(\mathbf{A}) = \mathbf{A}^{(0)}$, $\mathrm{E}^{(0)}(\boldsymbol{\phi}) = \boldsymbol{\phi}^{(0)}$, $\boldsymbol{\mu}_{\tilde{\mathbf{w}}}^{(0)} = \mathbf{0}$, $\mathrm{E}^{(0)}(\mathbf{A}'\boldsymbol{\Psi}^{-1}\mathbf{A}) = \mathbf{A}^{(0)\prime}\boldsymbol{\Psi}^{-1(0)}\mathbf{A}^{(0)}$ and $\mathrm{E}^{(0)}(\Sigma_{\tilde{\mathbf{w}}}(\boldsymbol{\phi})^{-1}) = \Sigma_{\tilde{\mathbf{w}}}\left(\boldsymbol{\phi}^{(0)}\right)^{-1}$.

for $i = 1$ to t **do**

Step 1: Update the distribution of $\boldsymbol{\beta} \sim MVN(\boldsymbol{\mu}_\beta^{(i)}, \mathbf{V}_\beta^{(i)})$, where
$\mathbf{V}_\beta^{(i)} = \left\{\mathbf{X}'\left[\mathbf{I}_n \otimes \mathrm{E}^{(i-1)}(\boldsymbol{\Psi}^{-1})\right]\mathbf{X}\right\}^{-1}$ and
$\boldsymbol{\mu}_\beta^{(i)} = \left\{\mathbf{X}'\left[\mathbf{I}_n \otimes \mathrm{E}^{(i-1)}(\boldsymbol{\Psi}^{-1})\right]\mathbf{X}\right\}^{-1}\mathbf{X}'\left[\mathbf{I}_n \otimes \mathrm{E}^{(i-1)}(\boldsymbol{\Psi}^{-1})\right]\left[\mathbf{Y} - \mathrm{E}^{(i-1)}(\mathscr{A})\boldsymbol{\mu}_{\tilde{\mathbf{w}}}^{(i-1)}\right]$.

Step 2: Update the distribution of $\tilde{\mathbf{w}} \sim MVN(\boldsymbol{\mu}_{\tilde{\mathbf{w}}}^{(i)}, \mathbf{V}_{\tilde{\mathbf{w}}}^{(i)})$, where
$\mathbf{V}_{\tilde{\mathbf{w}}}^{(i)} = \left\{\mathrm{E}^{(i-1)}\left[\mathscr{A}'\left(\mathbf{I}_n \otimes \mathrm{E}^{(i-1)}(\boldsymbol{\Psi}^{-1})\right)\mathscr{A}\right] + \mathrm{E}^{(i-1)}\left(\Sigma_{\tilde{\mathbf{w}}}^{-1}\right)\right\}^{-1}$ and
$\boldsymbol{\mu}_{\tilde{\mathbf{w}}}^{(i)} = \mathbf{V}_{\tilde{\mathbf{w}}}^{(i)}\mathrm{E}^{(i-1)}(\mathscr{A})'\left[\mathbf{I}_n \otimes \mathrm{E}^{(i-1)}(\boldsymbol{\Psi}^{-1})\right]\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\mu}_\beta^{(i)}\right)$.

Step 3: Update the distribution of $\delta_j^2 \sim Gamma\left(\frac{n}{2} + a_j, \left(\frac{1}{b_j} + d_j^{(i)}\right)^{-1}\right)$, where
$\mathrm{E}^{(i)}(\delta_j^2) = \left(\frac{n}{2} + a_j\right)\left(\frac{1}{b_j} + d_j^{(i)}\right)^{-1}$ and $d_j^{(i)}$ is defined in (B.2).

Step 4: Update the distribution of $\boldsymbol{\phi}$, which is proportional to

$$\sqrt{|\Sigma_{\tilde{\mathbf{w}}}^{-1}|}\exp\left\{-\frac{\boldsymbol{\mu}_{\tilde{\mathbf{w}}}^{(i)\prime}\Sigma_{\tilde{\mathbf{w}}}^{-1}\boldsymbol{\mu}_{\tilde{\mathbf{w}}}^{(i)} + \mathrm{Tr}\left(\mathbf{V}_{\tilde{\mathbf{w}}}^{(i)}\Sigma_{\tilde{\mathbf{w}}}^{-1}\right)}{2}\right\}$$

and calculate $\mathrm{E}^{(i)}\left(\Sigma_{\tilde{\mathbf{w}}}(\boldsymbol{\phi})^{-1}\right)$ using importance sampling.

Step 5: Update the distribution of $\mathbf{A}$, which is proportional to

$$\exp\left\{-\frac{\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\mu}_\beta^{(i)} - \mathscr{A}\boldsymbol{\mu}_{\tilde{\mathbf{w}}}^{(i)}\right)'\left[\mathbf{I}_n \otimes \mathrm{E}^{(i)}\left(\boldsymbol{\Psi}^{-1}\right)\right]\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\mu}_\beta^{(i)} - \mathscr{A}\boldsymbol{\mu}_{\tilde{\mathbf{w}}}^{(i)}\right) + \mathrm{Tr}\left[(\mathbf{A}\mathbf{A}')^{-1}\mathbf{S}\right]}{2}\right\}$$
$$\times\ |\mathbf{A}\mathbf{A}'|^{-(df+m+1)/2}$$

and calculate $\mathrm{E}^{(i)}(\mathbf{A})$, $\mathrm{E}^{(i)}\left(\mathbf{A}\mathrm{E}^{(i)}(\boldsymbol{\Psi})\mathbf{A}'\right)$ using importance sampling.

end for

Appendix C

Appendix for Chapter 3

Proof for Identification Issue

We are looking for $r \times r$ orthogonal matrix $\mathbf{P}$, which satisfies

$$\mathbf{P}'\mathbf{\Gamma}\mathbf{P} = \mathbf{\Gamma}, \tag{C.1}$$

where $\mathbf{\Gamma}$ is a diagonal matrix with the k-th diagonal element $\Gamma_{kk} = \rho_k$. Function (C.1) holds for any $\mathbf{\Gamma}$ with $\rho_k \in (0, 1)$. Let $\mathbf{V} = \mathbf{P}'\mathbf{\Gamma}\mathbf{P}$, then

$$V_{ij} = \sum_{k=1}^{r} P_{ki}\rho_k P_{kj} = \begin{cases} 0 & \text{if } i \neq j \\ \rho_i & \text{if } i = j \end{cases} . \tag{C.2}$$

V_{ii} is assumed to be differentiable as a function of ρ_k's within some domain and all of the ρ_k's are different. By taking the derivative of V_{ii} in equation (C.2) with respect to ρ_k, we have

$$\frac{dV_{ii}}{d\rho_k} = P_{ki}^2 = \begin{cases} 0 & \text{if } k \neq i \\ 1 & \text{if } k = i \end{cases} . \tag{C.3}$$

It is easy to derive from (C.3) that $P_{ii} = 1$ or -1 and $P_{ki} = 0$ for $k \neq i$. This orthogonal matrix is denoted as $\mathbf{P}_1$. We like to highlight here that if two diagonal elements of $\mathbf{\Gamma}$, for instance, ρ_k and ρ_l are the same, equation (C.3) has to be rewritten as:

$$\frac{dV_{ii}}{d\rho_k} = P_{ki}^2 + P_{li}^2 = \begin{cases} 0 & \text{if } k \neq i \\ 1 & \text{if } k = i \end{cases} .$$

Then, our argument is not valid any more. So it is important to guarantee that none of the diagonal element in $\mathbf{\Gamma}$ is identical.

If we relax the constraint for $\mathbf{P}'\mathbf{\Gamma}\mathbf{P}$ and only require $\mathbf{V}$ to be a diagonal matrix, function (C.2) can be rewritten as:

$$V_{ij} = \sum_{k=1}^{r} P_{ki}\rho_k P_{kj} = \begin{cases} 0 & \text{if } i \neq j \\ V_{ii} & \text{if } i = j \end{cases} . \tag{C.4}$$

Similarly, taking the derivative of V_{ij} in equation (C.4) with respect to ρ_k when $i \neq j$, we have

$$\frac{dV_{ij}}{d\rho_k} = P_{ki}P_{kj} = 0, \text{ for } k = 1, \ldots, r.$$

Since $\sum_{l=1}^{r} P_{li}^2 = 1$, at least one of the P_{li}'s does not equal 0. Assuming $P_{ki} \neq 0$, then $\forall j \neq i$, we have $P_{kj} = 0$. This implies each row of $\mathbf{P}$ has at most one non-zero element. Because $\mathbf{P}$ is a full rank orthogonal matrix, each row of $\mathbf{P}$ would have exactly one non-zero element with value 1 or -1 and the column index of the this non-zero element can not be the same.

If all the non-zero elements in $\mathbf{P}$ equal 1, this orthogonal matrix specified above is a permutation matrix, denoted as $\mathbf{P}_2$. In general, the multiplication of $\mathbf{P}_1$ and $\mathbf{P}_2$ would satisfy the looser restriction. So matrix $\mathbf{P}_1$ and $\mathbf{P}_2$ are the only two transformations that can cause identification issues in spatial factor models.

Gibbs Sampling Scheme

For a spatial data set with missing values, the hierarchical specification to yield a posterior distribution $p(\mathbf{F}, \mathbf{w}^*, \boldsymbol{\beta}, \mathbf{\Psi}, \boldsymbol{\phi}, \mathbf{\Lambda}, \boldsymbol{\delta}, \omega \,|\, \mathbf{Y}_o)$ in Section 3.1 can be written as:

$$\begin{aligned} &\prod_{i=1}^{n} N_{d_{\mathbf{s}_i}}\left(\mathbf{Y}_o(\mathbf{s}_i) \,|\, \mathbf{R}_1(\mathbf{s}_i)\left[\mathbf{X}(\mathbf{s}_i)'\boldsymbol{\beta} + \sum_{k=1}^{r}\delta_k f_k(\mathbf{s}_i)\boldsymbol{\lambda}_k\right], \mathbf{R}_1(\mathbf{s}_i)\mathbf{\Psi}\mathbf{R}_1(\mathbf{s}_i)'\right) \\ &\quad \times \prod_{k=1}^{r} N_n(\mathbf{f}_k \,|\, \mathbf{D}_k(\boldsymbol{\phi}_k)\mathbf{D}_k^*(\boldsymbol{\phi}_k)^{-1}\mathbf{w}_k^*, \mathbf{\Sigma}_{\mathbf{f}_k}(\boldsymbol{\phi}_k)) \times N_p(\boldsymbol{\beta} \,|\, \boldsymbol{\mu}_{\boldsymbol{\beta}}, \mathbf{\Sigma}_{\boldsymbol{\beta}}) \\ &\quad \times \prod_{k=1}^{r} N_{n^*}(\mathbf{w}_k^* \,|\, \mathbf{0}, \mathbf{D}_k^*(\boldsymbol{\phi}_k)) \times N_m(\boldsymbol{\lambda}_k \,|\, \mathbf{0}, c_0^2\mathbf{I}_m)\mathbf{I}(\mathbf{\Lambda}_{1k} > 0) \\ &\quad \times \prod_{j=1}^{m} IG(\Psi_j^2 \,|\, a, b) \times \prod_{k=1}^{r} Ber(\delta_k \,|\, \omega) \times \pi(\boldsymbol{\phi}) \times \pi(\omega), \end{aligned} \tag{C.5}$$

where $\mathbf{Y}_o = (\mathbf{Y}_o(\mathbf{s}_1)', \ldots, \mathbf{Y}_o(\mathbf{s}_n)')'$.

Define $\boldsymbol{\Sigma}^*(\mathbf{s}) = \mathbf{R}_1(\mathbf{s})'[\mathbf{R}_1(\mathbf{s})\boldsymbol{\Psi}\mathbf{R}_1(\mathbf{s})']^{-1}\mathbf{R}_1(\mathbf{s})$ and notice that $\boldsymbol{\Sigma}^*(\mathbf{s})$ is an $m \times m$ diagonal matrix with j-th diagonal element ψ_j^{-2} if $Y_j(\mathbf{s})$ is observed, the rest 0. Then sampling proceeds by first updating $\boldsymbol{\beta}$ from a $N_p(\tilde{\boldsymbol{\mu}}_{\boldsymbol{\beta}}, \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}})$ with:

$$\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}} = \left[\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} + \sum_{i=1}^{n} \mathbf{X}(\mathbf{s}_i)\boldsymbol{\Sigma}^*(\mathbf{s}_i)\mathbf{X}(\mathbf{s}_i)'\right]^{-1}$$
$$\tilde{\boldsymbol{\mu}}_{\boldsymbol{\beta}} = \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{\beta}}\left[\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu}_{\boldsymbol{\beta}} + \sum_{i=1}^{n} \mathbf{X}(\mathbf{s}_i)\boldsymbol{\Sigma}^*(\mathbf{s}_i)\left(\mathbf{Y}(\mathbf{s}_i) - \tilde{\boldsymbol{\Lambda}}\mathbf{f}(\mathbf{s}_i)\right)\right],$$

where $\tilde{\boldsymbol{\Lambda}} = (\delta_1\boldsymbol{\lambda}_1, \ldots, \delta_r\boldsymbol{\lambda}_r)$. The full conditional distribution for $\mathbf{f}(\mathbf{s})$ is independent multivariate normal distribution with the following parameters

$$\tilde{\boldsymbol{\Sigma}}_{\mathbf{f}}(\mathbf{s}) = \left[\boldsymbol{\Sigma}_{\mathbf{f}}(\mathbf{s})^{-1} + \tilde{\boldsymbol{\Lambda}}'\boldsymbol{\Sigma}^*(\mathbf{s})\tilde{\boldsymbol{\Lambda}}\right]^{-1}$$
$$\tilde{\boldsymbol{\mu}}_{\mathbf{f}}(\mathbf{s}) = \tilde{\boldsymbol{\Sigma}}_{\mathbf{f}}(\mathbf{s})\left[\tilde{\boldsymbol{\Lambda}}'\boldsymbol{\Sigma}^*(\mathbf{s})(\mathbf{Y}(\mathbf{s}_i) - \mathbf{X}(\mathbf{s}_i)'\boldsymbol{\beta}) + \boldsymbol{\Sigma}_{\mathbf{f}}(\mathbf{s})^{-1}\tilde{\mathbf{w}}(\mathbf{s})\right],$$

where $\boldsymbol{\Sigma}_{\mathbf{f}}(\mathbf{s})$ is an $r \times r$ diagonal matrix whose k-th diagonal element is $\sigma^2_{f_k}(\mathbf{s})$. $\tilde{\mathbf{w}}(\mathbf{s}) = (\tilde{w}_1(\mathbf{s}), \ldots, \tilde{w}_r(\mathbf{s}))'$ is a function of $\mathbf{w}^*$ and $\boldsymbol{\phi}$. $\mathbf{w}_k^*$ is also updated from a $N_{n^*}(\boldsymbol{\mu}_{\mathbf{w}_k^*}, \boldsymbol{\Sigma}_{\mathbf{w}_k^*})$ where

$$\boldsymbol{\Sigma}_{\mathbf{w}_k^*} = \mathbf{D}_k^*(\boldsymbol{\phi})\left[\mathbf{D}_k^*(\boldsymbol{\phi}) + \mathbf{D}_k(\boldsymbol{\phi}_k)'\boldsymbol{\Sigma}_{\mathbf{f}_k}(\boldsymbol{\phi}_k)^{-1}\mathbf{D}_k(\boldsymbol{\phi}_k)\right]^{-1}\mathbf{D}_k^*(\boldsymbol{\phi})$$
$$\boldsymbol{\mu}_{\mathbf{w}_k^*} = \boldsymbol{\Sigma}_{\mathbf{w}_k^*}\mathbf{D}_k^*(\boldsymbol{\phi})^{-1}\mathbf{D}_k(\boldsymbol{\phi}_k)'\boldsymbol{\Sigma}_{\mathbf{f}_k}(\boldsymbol{\phi}_k)^{-1}\mathbf{f}_k.$$

Each element of $\mathbf{Y}_o(\mathbf{s})$ is independent conditional on $\mathbf{f}(\mathbf{s})$, leading to another likelihood. Let $n_j \times 1$ vector $\mathbf{Y}_o^j$ denote the collection of the observed $Y_j(\mathbf{s})$ across all the locations, then $\boldsymbol{\mu}_o^j(\boldsymbol{\beta})$ is the mean vector of $\mathbf{Y}_o^j$, where n_j, $j = 1, \ldots, m$, are the total number of the observed $Y_j(\mathbf{s})$. Construct $n_j \times r$ matrix $\mathbf{F}_o^j$ out of $\mathbf{F}$ corresponding to the observations. If $Y_j(\mathbf{s}_i)$ is observed, the i-th row of $\mathbf{F}$ would be retained in $\mathbf{F}_o^j$. With the symbols defined, the likelihood in the first line of (C.5) can be rewritten as

$$\prod_{j=1}^{m} N_{n_j}\left(\mathbf{Y}_o^j \,|\, \boldsymbol{\mu}_o^j(\boldsymbol{\beta}) + \mathbf{F}_o^j\tilde{\boldsymbol{\lambda}}^j, \psi_j^2\mathbf{I}_{n_j}\right), \tag{C.6}$$

where $\tilde{\boldsymbol{\lambda}}^j = (\delta_1\Lambda_{j1}, \ldots, \delta_r\Lambda_{jr})'$. Replace the likelihood in (C.5) with (C.6) and let $\mathbf{R}_{\boldsymbol{\delta}}$ be an $r \times r$ diagonal matrix with k-th diagonal element δ_k. The full conditionals for

the j-th row of $\mathbf{\Lambda}$ follows $N_r(\mathbf{m}_j, \mathbf{K}_j)$, where $\mathbf{K}_j^{-1} = c_0^{-2}\mathbf{I}_r + \psi_j^{-2}\left(\mathbf{R}_{\boldsymbol{\delta}}\mathbf{F}_o^{j\prime}\mathbf{F}_o^j\mathbf{R}_{\boldsymbol{\delta}}\right)$ and $\mathbf{m}_j = \psi_j^{-2}\mathbf{K}_j\mathbf{R}_{\boldsymbol{\delta}}\mathbf{F}_o^{j\prime}\left(\mathbf{Y}_o^j - \boldsymbol{\mu}_o^j\right)$ for $j = 2, \ldots, m$. For $\boldsymbol{\lambda}^1$, it is a normal distribution truncated at 0 with mean $\mathbf{m}_1$ and variance $\mathbf{K}_1$. The full conditional distributions for ψ_j^2 reduces to a set of m independent Inverse Gamma with shape parameter $a + n_j/2$ and scale parameter $b + \frac{1}{2}(\mathbf{Y}_o^j - \boldsymbol{\mu}_o^j - \mathbf{F}_o^j\tilde{\boldsymbol{\lambda}}^j)'(\mathbf{Y}_o^j - \boldsymbol{\mu}_o^j - \mathbf{F}_o^j\tilde{\boldsymbol{\lambda}}^j)$ for $j = 1, \ldots, m$.

To obtain the posterior density for $\boldsymbol{\delta}$, we sample δ_k with $k = 1, \ldots, r$, preferable in random order from the full conditional distribution given $\boldsymbol{\delta}_{-k}$ and other parameters, where $\boldsymbol{\delta}_{-k} = \delta_1, \ldots, \delta_{k-1}, \delta_{k+1}, \ldots, \delta_r$ (Kuo and Mallick, 1998). The posterior for δ_k is $Bernoulli(p_k)$ with $p_k = \dfrac{\omega L(\delta_k = 1)}{\omega L(\delta_k = 1) + (1-\omega)L(\delta_k = 0)}$, where $L(\cdot)$ is the marginalized likelihood. To simulate δ_k, we use exactly the same algorithm as in Smith and Kohn (2002). Generate u from a uniform distribution on [0, 1]. Let δ_k^{old} denote the current value of δ_k, then

a If $\delta_k^{old} = 1$ and $u > 1 - \omega$, then $\delta_k^{new} = 1$.

b If $\delta_k^{old} = 1$ and $u < 1 - \omega$ then generate δ_k^{new} from probability

$$\Pr(\delta_k^{new} = 1) = \frac{L(\delta_k = 1)}{L(\delta_k = 1) + L(\delta_k = 0)}. \tag{C.7}$$

c If $\delta_k^{old} = 0$ and $u > \omega$, then $\delta_k^{new} = 0$.

d If $\delta_k^{old} = 0$ and $u < \omega$ then generate δ_k^{new} from probability

$$\Pr(\delta_k^{new} = 1) = \frac{L(\delta_k = 1)}{L(\delta_k = 1) + L(\delta_k = 0)}.$$

The posterior distribution for ω only depends on $\boldsymbol{\delta}$ and follows beta distribution with parameters $g_{\boldsymbol{\delta}} + 1$ and $r - g_{\boldsymbol{\delta}} + 1$, where $g_{\boldsymbol{\delta}} = \sum_{k=1}^r \delta_k$.

The remaining spatial parameter $\boldsymbol{\phi}$ is updated using Metropolis steps. To do so, we need to transform each element of $\boldsymbol{\phi}$ to have support equal to the whole real line. A straightforward solution here is to use $g(\phi_k) = \log\left(\frac{\phi_k - L_{\phi_k}}{U_{\phi_k} - \phi_k}\right)$, a transformation having Jacobian $(\phi_k - L_{\phi_k})(U_{\phi_k} - \phi_k)$, where L_{ϕ_k} and U_{ϕ_k} are the lower and upper bound of ϕ_k. Then the conditional distribution for ϕ_k is proportional to

$$(|\mathbf{\Sigma}_{\mathbf{f}_k}(\boldsymbol{\phi}_k)||\mathbf{D}_k^*(\boldsymbol{\phi}_k)|)^{-1/2}\exp\left\{-\frac{\mathbf{f}_k'\mathbf{\Sigma}_{\mathbf{f}_k}(\boldsymbol{\phi}_k)^{-1}\mathbf{f}_k + \mathbf{w}_k^{*\prime}\mathbf{D}_k^*(\boldsymbol{\phi}_k)^{-1}\mathbf{w}_k^*}{2}\right\}\pi(\phi_k \mid \phi_{k-1})\pi(\phi_{k+1} \mid \phi_k).$$

Discussion on Model Selection

We address some subtler issues regarding the behavior of the MCMC algorithm for the adaptive spatial factor model. In particular, we observe that the chains corresponding to the δ_k's eventually (i.e. after adequate burn-in) tend to stick to either 1 or 0 depending upon whether the k-th factor is included or excluded from the model. This phenomenon is not uncommon in stochastic model selection algorithms and does not necessarily reflect poor stability of the MCMC algorithm. In fact, a plausible explanation can be deduced from the strength of the spatial random field as we explain below.

The motivation for using $f_k(\mathbf{s})$ is to remedy the over-smoothing caused by $\tilde{w}_k(\mathbf{s})$ in low rank models. Recall the developments in Section 3.1. When the spatial random field is weak, i.e. the spatial correlation diminishes rapidly, it follows from the basic properties of the predictive process that $\tilde{w}_k(\mathbf{s}) \approx 0$, $\text{cov}\{w_k(\mathbf{s}), w_k(\mathbf{t})\} \approx 0$ and $\text{var}(f_k(\mathbf{s}) \,|\, \tilde{w}_k(\mathbf{s})) \approx 1$. This means that the kth latent factor behaves like white noise and, hence, is redundant in the model. Therefore, the stochastic selection algorithm will attempt to exclude this factor.

To achieve its goals, the algorithm compares the likelihood of models with different numbers of latent factors (δ_k being 1 or 0) in (C.7). The likelihoods are expected to be close to each other so as to allow transitions among the models (i.e., movements in the chains corresponding to the δ_k's). Let us consider a specific situation. Assume that the maximum number of factors is $r = 3$ and consider two models: $\boldsymbol{\delta} = c(1,1,1)$ (model 1) and $\boldsymbol{\delta} = c(1,1,0)$ (model 2). Also assume that ϕ_1 and ϕ_2 are small (i.e., strong spatial field), while ϕ_3 is very large (i.e., weak spatial field). Then, in model 1, $\text{var}\{\mathbf{Y}(\mathbf{s})\} = \boldsymbol{\Lambda}\text{var}\{\mathbf{f}(\mathbf{s})\}\boldsymbol{\Lambda}' = \boldsymbol{\Lambda}\boldsymbol{\Lambda}' = \sum_{i=1}^{3} \boldsymbol{\lambda}_i\boldsymbol{\lambda}_i'$ and

$$\text{cov}\{\mathbf{Y}(\mathbf{s}), \mathbf{Y}(\mathbf{t})\} = \sum_{i=1}^{3} \text{cov}\{\tilde{w}_i(\mathbf{s}), \tilde{w}_i(\mathbf{t})\}\boldsymbol{\lambda}_i\boldsymbol{\lambda}_i' \simeq \sum_{i=1}^{2} \text{cov}\{\tilde{w}_i(\mathbf{s}), \tilde{w}_i(\mathbf{t})\}\boldsymbol{\lambda}_i\boldsymbol{\lambda}_i',$$

since $\text{cov}\{\tilde{w}_3(\mathbf{s}), \tilde{w}_3(\mathbf{t})\} \simeq 0$. In model 2, $\text{var}\{\mathbf{Y}(\mathbf{s})\} = \sum_{i=1}^{2} \boldsymbol{\lambda}_i\boldsymbol{\lambda}_i'$ and

$$\text{cov}\{\mathbf{Y}(\mathbf{s}), \mathbf{Y}(\mathbf{t})\} = \boldsymbol{\Lambda}_2\text{cov}\{\tilde{\mathbf{w}}(\mathbf{s}), \tilde{\mathbf{w}}(\mathbf{t})\}\boldsymbol{\Lambda}_2' = \sum_{i=1}^{2} \text{cov}\{\tilde{w}_i(\mathbf{s}), \tilde{w}_i(\mathbf{t})\}\boldsymbol{\lambda}_i\boldsymbol{\lambda}_i',$$

where $\boldsymbol{\Lambda}_2 = (\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2)$. Here, $\text{cov}\{\mathbf{Y}(\mathbf{s}), \mathbf{Y}(\mathbf{t})\}$ is approximately the same for the two models, while $\text{var}\{\mathbf{Y}(\mathbf{s})\}$ is different, which yields different likelihoods.

If we use $\tilde{w}_k(\mathbf{s})$ instead of $f_k(\mathbf{s})$, then $\text{var}\{\tilde{w}_3(\mathbf{s})\}$ and $\text{cov}\{\tilde{w}_3(\mathbf{s}), \tilde{w}_3(\mathbf{t})\}$ are close to 0 when ϕ_3 is large. Hence, the likelihoods are also similar and the aforementioned transition among models is more frequent. Irrespective of whether we use $f_k(\mathbf{s})$ or $\tilde{w}_k(\mathbf{s})$, the algorithm will select factors having strong spatial dependence. Keeping in mind the inferential benefits associated with the modified predictive process $f_k(\mathbf{s})$ for moderate to strong spatial fields, we advocate its use although $\tilde{w}_k(\mathbf{s})$ can also be considered in applications.

Appendix D

Appendix for Chapter 4

Derivation of The Bayesian Risk Function

The rejection region is determined from (4.6)

$$\begin{aligned}
A_1 &= \left\{\mathbf{y} : P(\mathbf{c}'\boldsymbol{\theta} \leq \omega \,|\, \mathbf{y}) < \alpha)\right\} \\
&= \left\{\mathbf{y} : P\left(\frac{\mathbf{c}'\boldsymbol{\theta} - \mathbf{c}'\boldsymbol{\mu}_{\theta\,|\,\mathbf{y}}}{\sqrt{\mathbf{c}'\boldsymbol{\Sigma}_{\theta\,|\,\mathbf{y}}\mathbf{c}}} \leq \frac{\omega - \mathbf{c}'\boldsymbol{\mu}_{\theta\,|\,\mathbf{y}}}{\sqrt{\mathbf{c}'\boldsymbol{\Sigma}_{\theta\,|\,\mathbf{y}}\mathbf{c}}}\,\middle|\,\mathbf{y}\right) < \alpha\right\} \\
&= \left\{\mathbf{y} : \Phi\left(\frac{\omega - \mathbf{c}'\boldsymbol{\mu}_{\theta\,|\,\mathbf{y}}}{\sqrt{\mathbf{c}'\boldsymbol{\Sigma}_{\theta\,|\,\mathbf{y}}\mathbf{c}}}\right) < \alpha\right\} \\
&= \left\{\mathbf{y} : \omega - \mathbf{c}'\boldsymbol{\Sigma}_{\theta\,|\,\mathbf{y}}\left(\mathbf{X}_n'\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathbf{y} + \boldsymbol{\Sigma}_{\theta}^{-1}\boldsymbol{\mu}_{\theta}\right) < z_\alpha\sqrt{\mathbf{c}'\boldsymbol{\Sigma}_{\theta\,|\,\mathbf{y}}\mathbf{c}}\right\} \\
&= \left\{\mathbf{y} : \mathbf{c}'\boldsymbol{\Sigma}_{\theta\,|\,\mathbf{y}}\mathbf{X}_n'\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathbf{y} > \omega - z_\alpha\sqrt{\mathbf{c}'\boldsymbol{\Sigma}_{\theta\,|\,\mathbf{y}}\mathbf{c}} - \mathbf{c}'\boldsymbol{\Sigma}_{\theta\,|\,\mathbf{y}}\boldsymbol{\Sigma}_{\theta}^{-1}\boldsymbol{\mu}_{\theta}\right\} \\
&= \left\{\mathbf{y} : \mathbf{c}'\boldsymbol{\Sigma}_{\theta\,|\,\mathbf{y}}\mathbf{X}_n'\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathbf{y} > \omega - \mathbf{c}\boldsymbol{\mu}_{\theta} - z_\alpha\sqrt{\mathbf{c}'\boldsymbol{\Sigma}_{\theta\,|\,\mathbf{y}}\mathbf{c}} + \mathbf{c}'\boldsymbol{\Sigma}_{\theta\,|\,\mathbf{y}}\mathbf{N}_n\boldsymbol{\mu}_{\theta}.\right\}
\end{aligned}$$

Define $\Delta = \mathbf{c}\boldsymbol{\mu}_\theta - \omega$ and $u = \mathbf{c}'\boldsymbol{\Sigma}_{\theta\,|\,\mathbf{y}}\mathbf{X}_n'\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathbf{y}$, the rejection region is

$$A_1 = \left\{\mathbf{y} : u > \mathbf{c}'\boldsymbol{\Sigma}_{\theta\,|\,\mathbf{y}}\mathbf{N}_n\boldsymbol{\mu}_\theta - z_\alpha\sqrt{\mathbf{c}'\boldsymbol{\Sigma}_{\theta\,|\,\mathbf{y}}\mathbf{c}} - \Delta\right\}.$$

With this result, the Bayesian risk $r(\delta, \mathbf{N}_n)$ can then be calculated as:

$$\begin{aligned}
&\iint \left\{ I(\mathbf{y} \in A_0) I(\mathbf{c}'\boldsymbol{\theta} \in \Theta_1) + K I(\mathbf{y} \in A_1) I(\mathbf{c}'\boldsymbol{\theta} \in \Theta_0) \right\} p(\boldsymbol{\theta}, \mathbf{y}) \mathrm{d}\mathbf{y} \mathrm{d}\boldsymbol{\theta} \\
=&\int \left\{ I(\mathbf{y} \in A_0) P(\mathbf{c}'\boldsymbol{\theta} \in \Theta_1 \,|\, \mathbf{y}) + K I(\mathbf{y} \in A_1) P(\mathbf{c}'\boldsymbol{\theta} \in \Theta_0 \,|\, \mathbf{y}) \right\} m(\mathbf{y}) \mathrm{d}\mathbf{y} \\
=&\int \left\{ [1 - I(\mathbf{y} \in A_1)][1 - P(\mathbf{c}'\boldsymbol{\theta} \in \Theta_0 \,|\, \mathbf{y})] + K I(\mathbf{y} \in A_1) P(\mathbf{c}'\boldsymbol{\theta} \in \Theta_0 \,|\, \mathbf{y}) \right\} m(\mathbf{y}) \mathrm{d}\mathbf{y} \\
=&P(\mathbf{c}'\boldsymbol{\theta} \in \Theta_1) + (1 + K) \int I(\mathbf{y} \in A_1) P(\mathbf{c}'\boldsymbol{\theta} \in \Theta_0 \,|\, \mathbf{y}) m(\mathbf{y}) \mathrm{d}\mathbf{y} - P(\mathbf{y} \in A_1), \qquad \text{(D.1)}
\end{aligned}$$

where A_0 is the complementary set of A_1. To complete the calculation, we need the marginal distribution of u. It is a normal distribution with mean $\mathbf{c}'\boldsymbol{\Sigma}_{\theta\,|\,\mathbf{y}}\mathbf{N}_n\boldsymbol{\mu}_\theta$ and variance $\mathrm{var}\{u\} = \sigma_u^2$, where

$$\begin{aligned}
\sigma_u^2 &= \mathbf{c}'\boldsymbol{\Sigma}_{\theta\,|\,\mathbf{y}}\mathbf{X}_n'\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\left(\boldsymbol{\Sigma}_{\mathbf{y}} + \mathbf{X}_n'\boldsymbol{\Sigma}_\theta\mathbf{X}_n\right)\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathbf{X}_n\boldsymbol{\Sigma}_{\theta\,|\,\mathbf{y}}\mathbf{c} \\
&= \mathbf{c}'\boldsymbol{\Sigma}_{\theta\,|\,\mathbf{y}}\left(\mathbf{N}_n + \mathbf{N}_n\boldsymbol{\Sigma}_\theta\mathbf{N}_n\right)\boldsymbol{\Sigma}_{\theta\,|\,\mathbf{y}}\mathbf{c} \\
&= \mathbf{c}'\left(\mathbf{N}_n + \boldsymbol{\Sigma}_\theta^{-1}\right)^{-1}\mathbf{N}_n\boldsymbol{\Sigma}_\theta\left(\boldsymbol{\Sigma}_\theta^{-1} + \mathbf{N}_n\right)\boldsymbol{\Sigma}_{\theta\,|\,\mathbf{y}}\mathbf{c} \\
&= \mathbf{c}'\boldsymbol{\Sigma}_\theta\left(\mathbf{N}_n^{-1} + \boldsymbol{\Sigma}_\theta\right)^{-1}\boldsymbol{\Sigma}_\theta\mathbf{c}.
\end{aligned}$$

Now we can find $P(\mathbf{y} \in A_1)$ as:

$$\begin{aligned}
&P_{\mathbf{y}}\left\{u > \mathbf{c}'\boldsymbol{\Sigma}_{\theta\,|\,\mathbf{y}}\mathbf{N}_n\boldsymbol{\mu}_\theta - \Delta - z_\alpha\sqrt{\mathbf{c}'\boldsymbol{\Sigma}_{\theta\,|\,\mathbf{y}}\mathbf{c}}\right\} \\
=&P_{\mathbf{y}}\left\{\frac{u - \mathbf{c}'\boldsymbol{\Sigma}_{\theta\,|\,\mathbf{y}}\mathbf{N}_n\boldsymbol{\mu}_\theta}{\sigma_u} > \frac{-\Delta - z_\alpha\sqrt{\mathbf{c}'\boldsymbol{\Sigma}_{\theta\,|\,\mathbf{y}}\mathbf{c}}}{\sigma_u}\right\} \\
=&\Phi\left(\frac{\Delta + z_\alpha\sqrt{\mathbf{c}'\boldsymbol{\Sigma}_{\theta\,|\,\mathbf{y}}\mathbf{c}}}{\sigma_u}\right) = \Phi\left(o + tz_\alpha\right).
\end{aligned}$$

Nest, we would calculate $P(\mathbf{c}'\boldsymbol{\theta} \in \Theta_0 \,|\, \mathbf{y})$

$$\begin{aligned}
P(\mathbf{c}'\boldsymbol{\theta} \in \Theta_0 \,|\, \mathbf{y}) =&P(\mathbf{c}'\boldsymbol{\theta} < \omega \,|\, \mathbf{y}) = P\left(\frac{\mathbf{c}'\boldsymbol{\theta} - \mathbf{c}'\boldsymbol{\mu}_{\theta\,|\,\mathbf{y}}}{\sqrt{\mathbf{c}'\boldsymbol{\Sigma}_{\theta\,|\,\mathbf{y}}\mathbf{c}}} < \frac{\omega - \mathbf{c}'\boldsymbol{\mu}_{\theta\,|\,\mathbf{y}}}{\sqrt{\mathbf{c}'\boldsymbol{\Sigma}_{\theta\,|\,\mathbf{y}}\mathbf{c}}}\right) \\
=&\Phi\left(\frac{\omega - \mathbf{c}'\boldsymbol{\mu}_{\theta\,|\,\mathbf{y}}}{\sqrt{\mathbf{c}'\boldsymbol{\Sigma}_{\theta\,|\,\mathbf{y}}\mathbf{c}}}\right) = \Phi\left(\frac{\mathbf{c}'\boldsymbol{\Sigma}_{\theta\,|\,\mathbf{y}}\mathbf{N}_n\boldsymbol{\mu}_\theta - \Delta - u}{\sqrt{\mathbf{c}'\boldsymbol{\Sigma}_{\theta\,|\,\mathbf{y}}\mathbf{c}}}\right).
\end{aligned}$$

Then the integral in (D.1) is

$$\begin{aligned}
&\int I(\mathbf{y} \in A_1) P(\mathbf{c}'\boldsymbol{\theta} \in \Theta_0 \,|\, \mathbf{y}) m(\mathbf{y}) \mathrm{d}\mathbf{y} \\
&= \int I\left(u > \mathbf{c}'\boldsymbol{\Sigma}_{\theta\,|\,\mathbf{y}}\mathbf{N}_n\boldsymbol{\mu}_\theta - \Delta - z_\alpha\sqrt{\mathbf{c}'\boldsymbol{\Sigma}_{\theta\,|\,\mathbf{y}}\mathbf{c}}\right) \Phi\left(\frac{\mathbf{c}'\boldsymbol{\Sigma}_{\theta\,|\,\mathbf{y}}\mathbf{N}_n\boldsymbol{\mu}_\theta - \Delta - u}{\sqrt{\mathbf{c}'\boldsymbol{\Sigma}_{\theta\,|\,\mathbf{y}}\mathbf{c}}}\right) p(u)\mathrm{d}u \\
&= \int_{-z_\alpha}^{+\infty} \Phi(-v) \frac{\sqrt{\mathbf{c}'\boldsymbol{\Sigma}_{\theta\,|\,\mathbf{y}}\mathbf{c}}}{\sqrt{2\pi\sigma_u^2}} \exp\left\{-\frac{\left(v\sqrt{\mathbf{c}'\boldsymbol{\Sigma}_{\theta\,|\,\mathbf{y}}\mathbf{c}} - \Delta\right)^2}{2\sigma_u^2}\right\} \mathrm{d}v \\
&= \int_{-z_\alpha}^{+\infty} \frac{t\Phi(-v)}{\sqrt{2\pi}} \exp\left\{-\frac{(tv-o)^2}{2}\right\} \mathrm{d}v,
\end{aligned}$$

where $v = \left(\mathbf{c}'\boldsymbol{\Sigma}_{\theta\,|\,\mathbf{y}}\mathbf{N}_n\boldsymbol{\mu}_\theta - \Delta - u\right) / \sqrt{\mathbf{c}'\boldsymbol{\Sigma}_{\theta\,|\,\mathbf{y}}\mathbf{c}}$ and $t = \sqrt{\mathbf{c}'\boldsymbol{\Sigma}_{\theta\,|\,\mathbf{y}}\mathbf{c}/\sigma_u^2}$. It is easy to find out the $P(\mathbf{c}'\boldsymbol{\theta} \in \Theta_1) = \Phi\left(\Delta/\xi\right)$. So the Bayesian risk is obtained here:

$$\Phi\left(\Delta/\xi\right) + (1+K) \int_{-z_\alpha}^{+\infty} \frac{t}{\sqrt{2\pi}} \Phi(-v) \exp\left\{-\frac{(tv-o)^2}{2}\right\} \mathrm{d}v - \Phi(tz_\alpha + o). \tag{D.2}$$

The Derivative of the Bayesian Risk for a Single Test

Now define

$$\dot{o} = \frac{\partial o}{\partial t} = \frac{\Delta}{\xi} \frac{t}{\sqrt{t^2+1}}$$

and take a derivative of $r(\delta, \mathbf{N}_n)$ with respect to t, we get

$$
\begin{aligned}
&\frac{\partial r(\delta, \mathbf{N}_n)}{\partial t} \\
=&-(z_\alpha+\dot{o})\varphi(tz_\alpha+o)+\frac{K+1}{\sqrt{2\pi}}\int_{-z_\alpha}^{+\infty}\Phi(-v)\left[1-t(v-\dot{o})(tv-o)\right]\exp\left\{-\frac{(tv-o)^2}{2}\right\}\mathrm{d}v \\
=&-(z_\alpha+\dot{o})\varphi(tz_\alpha+o)+\frac{K+1}{\sqrt{2\pi}}\int_{-z_\alpha}^{+\infty}\Phi(-v)\exp\left\{-\frac{(tv-o)^2}{2}\right\}\mathrm{d}v \\
&+\frac{K+1}{\sqrt{2\pi}}\int_{-z_\alpha}^{+\infty}(v-\dot{o})\Phi(-v)\mathrm{d}\left(e^{-\frac{(tv-o)^2}{2}}\right) \\
=&\frac{K+1}{\sqrt{2\pi}}(v-\dot{o})\Phi(-v)\exp\left\{-\frac{(tv-o)^2}{2}\right\}\Big|_{-z_\alpha}^{+\infty}+\frac{K+1}{\sqrt{2\pi}}\int_{-z_\alpha}^{+\infty}\Phi(-v)\exp\left\{-\frac{(tv-o)^2}{2}\right\}\mathrm{d}v \\
&-\frac{K+1}{\sqrt{2\pi}}\int_{-z_\alpha}^{+\infty}\left[\Phi(-v)-(v-\dot{o})\varphi(-v)\right]\exp\left\{-\frac{(tv-o)^2}{2}\right\}\mathrm{d}v-(z_\alpha+\dot{o})\varphi(tz_\alpha+o) \\
=&-\frac{K+1}{\sqrt{2\pi}}(-z_\alpha-\dot{o})\Phi(z_\alpha)\exp\left\{-\frac{(tz_\alpha+o)^2}{2}\right\}-(z_\alpha+\dot{o})\varphi(tz_\alpha+o) \\
&+\frac{K+1}{\sqrt{2\pi}}\int_{-z_\alpha}^{+\infty}(v-\dot{o})\varphi(-v)\exp\left\{-\frac{(tv-o)^2}{2}\right\}\mathrm{d}v \\
=&\frac{K+1}{2\pi}\int_{-z_\alpha}^{+\infty}(v-\dot{o})\exp\left\{-\frac{v^2}{2}\right\}\exp\left\{-\frac{(tv-o)^2}{2}\right\}\mathrm{d}v \\
=&\frac{K+1}{2\pi}\int_{-z_\alpha}^{+\infty}(v-\dot{o})\exp\left\{-\frac{(t^2+1)\left(v^2-2\frac{to}{t^2+1}v+\frac{o^2}{t^2+1}\right)}{2}\right\}\mathrm{d}v \\
=&\frac{K+1}{2\pi}\int_{-z_\alpha}^{+\infty}(v-\dot{o})\exp\left\{-\frac{(t^2+1)\left(v^2-2\dot{o}v+\dot{o}^2+\frac{\Delta^2}{\xi^2(t^2+1)}\right)}{2}\right\}\mathrm{d}v \\
=&\frac{K+1}{2\pi}\exp\left\{-\frac{\Delta^2}{2\xi^2}\right\}\int_{-z_\alpha}^{+\infty}(v-\dot{o})\exp\left\{-\frac{(t^2+1)\,(v-\dot{o})^2}{2}\right\}\mathrm{d}v \\
=&\frac{K+1}{2\pi(t^2+1)}\exp\left\{-\frac{\Delta^2}{2\xi^2}\right\}\exp\left\{-\frac{(t^2+1)(z_\alpha+\dot{o})^2}{2}\right\}
\end{aligned}
$$

Derivation for $\mathbf{N}_n$ as an Non-Decreasing Function of n

If one more observation is included in the study, $\mathbf{N}_{n+1}$ can be written as

$$
\mathbf{N}_{n+1}=(\mathbf{X}_n',\mathbf{x}_{n+1})\begin{pmatrix}\mathbf{\Sigma_y} & \mathbf{v}\\ \mathbf{v}' & \sigma^2+\tau^2\end{pmatrix}^{-1}\begin{pmatrix}\mathbf{X}_n\\ \mathbf{x}_{n+1}'\end{pmatrix},
$$

where $\mathbf{x}_{n+1}$ the predictor for the new observation and $\mathbf{v}$ the covariance between the first n observations and the new one. Define $\lambda^2 = \sigma^2 + \tau^2 - \mathbf{v}'\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathbf{v}$. Using blockwise inverse for the covariance matrix, we have

$$\begin{aligned}
\mathbf{N}_{n+1} &= (\mathbf{X}_n', \mathbf{x}_{n+1}) \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{y}}^{-1} + \lambda^{-2}\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathbf{v}\mathbf{v}'\boldsymbol{\Sigma}_{\mathbf{y}}^{-1} & -\lambda^{-2}\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathbf{v} \\ -\lambda^{-2}\mathbf{v}'\boldsymbol{\Sigma}_{\mathbf{y}}^{-1} & \lambda^{-2} \end{pmatrix} \begin{pmatrix} \mathbf{X}_n \\ \mathbf{x}_{n+1}' \end{pmatrix} \\
&= \mathbf{X}_n'\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathbf{X}_n + \lambda^{-2}\left(\mathbf{X}_n'\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathbf{v}\mathbf{v}'\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathbf{X}_n - \mathbf{X}_n'\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathbf{v}\mathbf{x}_{n+1}' - \mathbf{x}_{n+1}\mathbf{v}'\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathbf{X}_n + \mathbf{x}_{n+1}\mathbf{x}_{n+1}'\right) \\
&= \mathbf{N}_n + \lambda^{-2}\left(\mathbf{X}_n'\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathbf{v} - \mathbf{x}_{n+1}\right)\left(\mathbf{X}_n'\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathbf{v} - \mathbf{x}_{n+1}\right)'.
\end{aligned}$$

Notice that both $\mathbf{X}_n'\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathbf{v}$ and $\mathbf{x}_{n+1}$ are $p\times 1$ vector. Since $\left(\mathbf{X}_n'\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathbf{v} - \mathbf{x}_{n+1}\right)\left(\mathbf{X}_n'\boldsymbol{\Sigma}_{\mathbf{y}}^{-1}\mathbf{v} - \mathbf{x}_{n+1}\right)'$ is a non-negative definite $p \times p$ matrix, $\mathbf{N}_n$ is an non-decreasing function with respect to n.