Computational Statistics and Data Analysis 55 (2011) 3197–3217

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Variational Bayesian methods for spatial data analysis

Qian Ren [a], Sudipto Banerjee [a,*], Andrew O. Finley [b,c], James S. Hodges [a]

[a] Division of Biostatistics, University of Minnesota, MN, USA
[b] Department of Forestry, Michigan State University, MI, USA
[c] Department of Geography, Michigan State University, MI, USA

ARTICLE INFO

Article history:
Received 1 October 2010
Received in revised form 26 May 2011
Accepted 29 May 2011
Available online 13 June 2011

Keywords:
Bayesian inference
Gaussian process
Hierarchical models
Markov chain Monte Carlo
Spatial process models
Variational Bayesian

ABSTRACT

With scientific data available at geocoded locations, investigators are increasingly turning to spatial process models for carrying out statistical inference. However, fitting spatial models often involves expensive matrix decompositions, whose computational complexity increases in cubic order with the number of spatial locations. This situation is aggravated in Bayesian settings where such computations are required once at every iteration of the Markov chain Monte Carlo (MCMC) algorithms. In this paper, we describe the use of Variational Bayesian (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. We provide results for the variational updates in several models especially emphasizing their use in multivariate spatial analysis. We demonstrate estimation and model comparisons from VB methods by using simulated data as well as environmental data sets and compare them with inference from MCMC.

1. Introduction

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 $\{w(\mathbf{s}) : \mathbf{s} \in D\}$ 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 $\mathcal{S} = \{\mathbf{s}_1, \mathbf{s}_2, \ldots, \mathbf{s}_n\}$. 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.

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,

$$Y(\mathbf{s}) = \mathbf{x}(\mathbf{s})'\boldsymbol{\beta} + w(\mathbf{s}) + \epsilon(\mathbf{s}). \tag{1}$$

* Corresponding author.
E-mail addresses: renxx014@umn.edu (Q. Ren), sudiptob@biostat.umn.edu (S. Banerjee), finleya@msu.edu (A.O. Finley), hodge003@umn.edu (J.S. Hodges).

0167-9473/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2011.05.021

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 $\mathbf{Y} = (Y(\mathbf{s}_1), \ldots, Y(\mathbf{s}_n))'$ 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 $p(\boldsymbol{\beta}, \mathbf{w}, \sigma^2, \tau^2, \boldsymbol{\phi} \mid \mathbf{Y})$ that is proportional to

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

Estimation of (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 $p(\boldsymbol{\beta}, \sigma^2, \tau^2, \boldsymbol{\phi} \mid \mathbf{Y})$ that is proportional to

$$p(\boldsymbol{\phi}) \times IG(\tau^2 \mid a_\tau, b_\tau) \times IG(\sigma^2 \mid a_\sigma, b_\sigma) \times N(\boldsymbol{\beta} \mid \boldsymbol{\mu}_\beta, \Sigma_\beta) \times N(\mathbf{Y} \mid \mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{R}(\boldsymbol{\phi}) + \tau^2\mathbf{I}_n), \tag{3}$$

where $\mathbf{X}$ is the matrix of regressors with the ith 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); see Banerjee et al. (2004) for details. In fact we can integrate out $\boldsymbol{\beta}$ from (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 $p(\mathbf{Y} \mid \sigma^2, \tau^2, \boldsymbol{\phi})$ 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.

Evidently, 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.

This computational burden is exacerbated in multivariate settings with several spatially dependent outcome variables. 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. We will subsequently illustrate with a forestry dataset.

Here, 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 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 or when we have spatiotemporal random effects.

Modeling 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) and Banerjee et al. (2008)). Essentially, these methods replace the process $w(\mathbf{s})$ with an approximation $\tilde{w}(\mathbf{s})$ 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) and 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 directly observed Gaussian process. Another approach either replaces the process (random field) model by a Markov random field (Cressie, 1993) or approximates the random field model by a Markov random field (Rue and Tjelmeland, 2002; Rue and Held, 2005). Recently Lindgern 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) and Gelfand et al. (2004)) and spatially varying regressions (Gelfand et al., 2003) is potentially problematic.

In this paper 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. 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 generalization of 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.

We explore VB methods for estimating univariate and multivariate spatial models. In the next section we review the VB framework for Bayesian inference. Section 3 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 4 and 5 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 6 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 7 concludes the paper with a summary.

2. Bayesian estimation

2.1. Review of Variational Bayesian methods

Variational methods have their origins in the 18th 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 $H(p) = -\int p(y) \ln p(y)\, \mathrm{d}y$, which takes a probability density function $p(y)$ as the input and returns the 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. Usually no closed form solution can be found. Therefore, 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 Bayesian 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 $p(\mathbf{y}) = \int p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}$ 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 $p(\mathbf{y}) = \frac{p(\mathbf{y},\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathbf{y})}$, we have $\log p(\mathbf{y}) = \log p(\mathbf{y}, \boldsymbol{\theta}) - \log p(\boldsymbol{\theta}|\mathbf{y}) = \log \frac{p(\mathbf{y},\boldsymbol{\theta})}{q(\boldsymbol{\theta})} + \log \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{\theta}|\mathbf{y})}$. Multiplying both sides by $q(\boldsymbol{\theta})$ and integrating with respect to $\boldsymbol{\theta}$, we obtain

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

where $\mathcal{L}(q)$ is a function of $\mathbf{y}$ and $KL(q \parallel p)$ is the Kullback–Leibler (KL) distance from $q(\boldsymbol{\theta})$ to $p(\boldsymbol{\theta}|\mathbf{y})$. Since $KL(q \parallel p)$ 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 $KL(q \parallel p)$. 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 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):

$$q_i^*(\boldsymbol{\theta}_i) = \frac{\exp\{\mathrm{E}_{j\neq i}[\log p(\mathbf{y}, \boldsymbol{\theta})]\}}{\int \exp\{\mathrm{E}_{j\neq i}[\log p(\mathbf{y}, \boldsymbol{\theta})]\} \, \mathrm{d}\boldsymbol{\theta}_i}. \tag{4}$$

Eq. (4) 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 (4) 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 (4) 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).

2.2. Bayesian central limit theorem

Some of our subsequent comparisons will be made with the BCLT which is a large-sample approximation for posterior distributions (Carlin and Louis, 1996). Let $f(\mathbf{y} \mid \boldsymbol{\theta})$ 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 $p(\boldsymbol{\theta} \mid \mathbf{y}) \propto f(\mathbf{y} \mid \boldsymbol{\theta})p(\boldsymbol{\theta})$ is proper and its mode exists, we have, as $n \to \infty$, $p(\boldsymbol{\theta} \mid \mathbf{y}) \sim N(\hat{\boldsymbol{\theta}}_m, \mathbf{H}^{-1}(\hat{\boldsymbol{\theta}}_m))$, where $\hat{\boldsymbol{\theta}}_m$ is the posterior mode and the matrix $\mathbf{H} = -\frac{\partial^2 \log p(\boldsymbol{\theta}|\mathbf{y})}{\partial\theta_i \partial\theta_j}$ is the negative of the *Hessian matrix* of $\log p(\boldsymbol{\theta} \mid \mathbf{y})$. 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.

3. 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,

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

The likelihood is

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

Assuming $q(\boldsymbol{\beta}, \sigma^2) = q_\beta(\boldsymbol{\beta})q_{\sigma^2}(\sigma^2)$ and the approximate distributions for all the parameters are known at iteration t, we apply (4) to this model to obtain the following iterative solutions:

$$q^{(t+1)}(\boldsymbol{\beta}) \sim N(\boldsymbol{\mu}^*, (\zeta^2)^{(t+1)}\mathbf{V}^*), \quad \text{where } (\zeta^2)^{(t+1)} = \left[\int \mathrm{d}\sigma^2 q^{(t)}(\sigma^2)/\sigma^2\right]^{-1}, \qquad \mathbf{V}^* = (\mathbf{V}_\beta^{-1} + \mathbf{X}'\mathbf{X})^{-1} \quad \text{and}$$

$$\boldsymbol{\mu}^* = \mathbf{V}^*(\mathbf{V}_\beta^{-1}\boldsymbol{\mu}_\beta + \mathbf{X}'\mathbf{Y}); \qquad q^{(t+1)}(\sigma^2) \sim IG\left(a^* + \frac{p}{2}, \frac{2b^* + p \times (\zeta^2)^{(t+1)}}{2}\right),$$

where $a^* = a + \frac{n}{2}$ and $b^* = b + \frac{1}{2}(\boldsymbol{\mu}_\beta'\mathbf{V}_\beta^{-1}\boldsymbol{\mu}_\beta + \mathbf{Y}'\mathbf{Y} - \boldsymbol{\mu}^{*\prime}\mathbf{V}^{*-1}\boldsymbol{\mu}^*)$. Notice that this algorithm only needs the starting value for $(\zeta^2)^{(0)} = \mathrm{E}^{(0)}(1/\sigma^2)$. We do not have to calculate the expectation by specifying $q^{(0)}(\sigma^2)$, but give an initial value to $(\zeta^2)^{(0)}$ directly.

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

$$(\zeta^2)^{(t+2)} = \left\{ \int \mathrm{d}\sigma^2 \frac{q^{(t+1)}(\sigma^2)}{\sigma^2} \right\}^{-1} = \frac{2b^* + p \times (\zeta^2)^{(t+1)}}{2a^* + p}. \tag{5}$$

Defining $\lim_{t\to+\infty}(\zeta^2)^{(t)} = \zeta^2$ and taking limit on both sides of (5), we obtain $\zeta^2 = \frac{b^*}{a^*}$. So when $t \to \infty$,

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

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 $\frac{b^*}{a^*}(a^* + \frac{p}{2})$. The joint posterior distribution for $\boldsymbol{\beta}$ and σ^2 with the conjugate *NIG* prior is $NIG(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{\mu}^*, \mathbf{V}^*, a^*, b^*)$. The exact marginal posterior distributions are $p(\boldsymbol{\beta} \mid \mathbf{Y}) \sim MVSt_{2a^*}(\boldsymbol{\mu}^*, \frac{b^*}{a^*}\mathbf{V}^*)$ and $p(\sigma^2 \mid \mathbf{Y}) \sim IG(a^*, b^*)$, 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 $\frac{b^*}{a^*}(a^* + \frac{p}{2})/(a^* + \frac{p}{2}) = \frac{b^*}{a^*}$. 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 6.1. For data sets 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.

4. VB for univariate spatial regression

4.1. 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). 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).

When applying the VB algorithm to find the estimated posterior distributions, we have to specify $q_i^{(0)}(\boldsymbol{\theta}_i)$. In fact, instead of giving explicit distributions to $q_i^{(0)}(\boldsymbol{\theta}_i)$, we only need the starting values for the expectation of some functions of the parameters and latent variables to initiate the algorithm corresponding to (4) as in the linear regression example. Which functions 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 have been updated in the algorithm, the expectation of these functions are calculated at each iteration. For the univariate spatial model treating $\mathbf{w}$ as hidden, we have Algorithm 1 to find the VB estimates for the posterior distributions. In Algorithm 1, $\mathrm{Tr}(\cdot)$ denotes the trace function, and $\boldsymbol{\mu}_{\mathbf{w}}^{(t-1)}$, and $\mathbf{V}_{\mathbf{w}}^{(t-1)}$ are the expectation and covariance matrix for $\mathbf{w} \sim q^{(t-1)}(\mathbf{w})$ respectively.

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 $t = 1$ to T **do**
Step 1: Update the distribution of $\boldsymbol{\beta} \sim MVN\left(\boldsymbol{\mu}_{\beta}^{(t)}, \mathbf{V}_{\beta}^{(t)}\right)$, where
$\mathbf{V}_{\beta}^{(t)} = \left[\mathrm{E}^{(t-1)}\left(1/\tau^2\right)\right]^{-1}(\mathbf{X}'\mathbf{X})^{-1}$ and $\boldsymbol{\mu}_{\beta}^{(t)} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{w}}^{(t-1)})$.
Step 2: Update the distribution of $\tau^2 \sim IG$ with parameters $a_\tau + \frac{n}{2}$ and
$b_\tau + \frac{1}{2}\left[\mathrm{Tr}\left(\mathbf{V}_{\mathbf{w}}^{(t-1)}\right) + p\mathrm{E}^{(t-1)}\left(1/\tau^2\right) + \left(\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{w}}^{(t-1)}\right)'(\mathbf{I}_n - \mathbf{H})\left(\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{w}}^{(t-1)}\right)\right]$, where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.
calculate $m_{\tau^2}^{(t)} = \mathrm{E}^{(t)}(1/\tau^2)$.
Step 3: Update the distribution of $\sigma^2 \sim IG$ with parameters $a_\sigma + \frac{n}{2}$ and
$b_\sigma + \frac{1}{2}\left\{\mathrm{Tr}\left[\mathrm{E}^{(t-1)}\left(\mathbf{R}(\phi)^{-1}\right)\mathbf{V}_{\mathbf{w}}^{(t-1)}\right] + \boldsymbol{\mu}_{\mathbf{w}}^{(t-1)\prime}\mathrm{E}^{(t-1)}\left(\mathbf{R}(\phi)^{-1}\right)\boldsymbol{\mu}_{\mathbf{w}}^{(t-1)}\right\}$;
calculate $m_{\sigma^2}^{(t)} = \mathrm{E}^{(t)}(1/\sigma^2)$.
Step 4: Update the distribution of $\mathbf{w} \sim MVN\left(\boldsymbol{\mu}_{\mathbf{w}}^{(t)}, \mathbf{V}_{\mathbf{w}}^{(t)}\right)$, where
$\mathbf{V}_{\mathbf{w}}^{(t)} = \left[m_{\sigma^2}^{(t)}\mathrm{E}^{(t-1)}\left(\mathbf{R}(\phi)^{-1}\right) + m_{\tau^2}^{(t)}\mathbf{I}_n\right]^{-1}$ and
$\boldsymbol{\mu}_{\mathbf{w}}^{(t)} = m_{\tau^2}^{(t)}\left[m_{\sigma^2}^{(t)}\mathrm{E}^{(t-1)}\left(\mathbf{R}(\phi)^{-1}\right) + m_{\tau^2}^{(t)}\mathbf{I}_n\right]^{-1}\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\mu}_{\beta}^{(t)}\right)$.
Step 5: Update the distribution of ϕ which is proportional to

$$|\mathbf{R}(\phi)|^{-\frac{1}{2}}\exp\left\{-\frac{m_{\sigma^2}^{(t)}\left[\mathrm{Tr}\left(\mathbf{R}(\phi)^{-1}\mathbf{V}_{\mathbf{w}}^{(t)}\right) + \boldsymbol{\mu}_{\mathbf{w}}^{(t)\prime}\mathbf{R}(\phi)^{-1}\boldsymbol{\mu}_{\mathbf{w}}^{(t)}\right]}{2}\right\} \tag{6}$$

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

With the expectation of $\mathbf{R}(\phi)^{-1}$ under $q_{\boldsymbol{\phi}}^{(t)}(\boldsymbol{\phi})$, 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 distribution function (6) is not analytically tractable, so importance sampling is proposed to approximate $E^{(i)}(\mathbf{R}(\phi)^{-1})$. Denote function (6) as $g(\phi)$. Then for a function $f(\phi)$ of ϕ

$$\mathrm{E}(f(\phi)) = \frac{\int f(\phi)g(\phi)\,\mathrm{d}\phi}{\int g(\phi)\,\mathrm{d}\phi} = \frac{\int f(\phi)\frac{g(\phi)}{p_I(\phi)}p_I(\phi)\,\mathrm{d}\phi}{\int \frac{g(\phi)}{p_I(\phi)}p_I(\phi)\,\mathrm{d}\phi} \approx \frac{\frac{1}{N}\sum_{i=1}^{N} f(\phi_i)W(\phi_i)}{\frac{1}{N}\sum_{i=1}^{N} W(\phi_i)} = \sum_{i=1}^{N} f(\phi_i)W^*(\phi_i), \tag{7}$$

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 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 to estimate the expectation for $f(\phi_i)$ in the sum (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 $\mathrm{E}(\mathbf{R}(\phi)^{-1})$, using the estimated expectation from importance sampling in (7) allows the VB algorithm to proceed toward convergence. After the VB algorithm converges, importance sampling resampling method (Rubin, 1987) is used to simulate samples of ϕ, which is proportional to (6). Inferences about $p(\phi \mid \mathbf{Y})$ can be made based on these samples. A simulated example using this model is illustrated in Section 6.2.

4.2. 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 (3). Different ways of grouping parameters (mean field approximation) 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 (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 4.1 for $\boldsymbol{\beta}$, τ^2 and ϕ, while a uniform prior is assigned to r. To initiate the VB algorithm for marginal spatial model we need to give starting values for the expectation of $1/\tau^2$ and $\mathbf{C}(r,\phi)^{-1}$. Applying (4) to the spatial model (3), we have Algorithm 2 to find the VB estimates for the posterior distributions.

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 $t = 1$ to T **do**
Step 1: Update the distribution of $\boldsymbol{\beta} \sim MVN\left(\boldsymbol{\mu}_\beta^{(t)}, \mathbf{V}_\beta^{(t)}\right)$
$\mathbf{V}_\beta^{(t)} = \left[\mathrm{E}^{(t-1)}\left(1/\tau^2\right)\right]^{-1}\left[\mathbf{X}'\mathrm{E}^{(t-1)}\left(\mathbf{C}^{-1}\right)\mathbf{X}\right]^{-1}$ and $\boldsymbol{\mu}_\beta^{(t)} = \left[\mathbf{X}'\mathrm{E}^{(t-1)}\left(\mathbf{C}^{-1}\right)\mathbf{X}\right]^{-1}\mathbf{X}'\mathrm{E}^{(t-1)}\left(\mathbf{C}^{-1}\right)\mathbf{Y}$.
Step 2: Update the distribution of $\tau^2 \sim IG$ with parameters $a_\tau + \frac{n}{2}$ and
$b_\tau + \frac{1}{2}\left[\mathrm{Tr}\left(\mathbf{X}'\mathrm{E}^{(t-1)}\left(\mathbf{C}^{-1}\right)\mathbf{X}\mathbf{V}_\beta^{(t)}\right) + \left(\mathbf{X}\boldsymbol{\mu}_\beta^{(t)} - \mathbf{Y}\right)'\mathrm{E}^{(t-1)}\left(\mathbf{C}^{-1}\right)\left(\mathbf{X}\boldsymbol{\mu}_\beta^{(t)} - \mathbf{Y}\right)\right]$;
calculate $\mathrm{E}^{(t)}(1/\tau^2)$.
Step 3: Update the joint distribution of ϕ and r, which is proportional to

$$|\mathbf{C}|^{-\frac{1}{2}} \times \exp\left\{\mathrm{E}^{(t)}\left(1/\tau^2\right)\left[-\frac{\mathrm{Tr}\left(\mathbf{X}'\mathbf{C}^{-1}\mathbf{X}\mathbf{V}_\beta^{(t)}\right) + \left(\mathbf{X}\boldsymbol{\mu}_\beta^{(t)} - \mathbf{Y}\right)'\mathbf{C}^{-1}\left(\mathbf{X}\boldsymbol{\mu}_\beta^{(t)} - \mathbf{Y}\right)}{2}\right]\right\}$$

and calculate $\mathrm{E}^{(t)}(r)$, $\mathrm{E}^{(t)}(\phi)$ and $\mathrm{E}^{(t)}\left(\mathbf{C}^{-1}\right)$.
end for

Now we have closed form expressions for the approximate posterior distributions of parameters $\boldsymbol{\beta}$ and τ^2 given $\mathrm{E}(\mathbf{C}^{-1})$. Importance sampling is again used to calculate the expectation of $\mathbf{C}^{-1}$ which then allows the VB algorithm to complete. After the VB algorithm converges, approximate marginal posterior distributions of ϕ and r can be obtained by using importance sampling resampling method (Rubin, 1987). Since more and more information about the posterior distribution of r is learned from each VB iteration, $\mathrm{E}^{(T)}(r)$ estimates the true posterior mean better and better. The *importance function*, $p_I(r)$, can be assigned to a distribution centered at $\mathrm{E}^{(T)}(r)$ to save computational time. The same technique is used for multivariate spatial models in Section 5.1.

5. Multivariate spatial model

5.1. Regression model and VB algorithm

Here we extend the univariate case discussed in Section 4 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 $\mathbf{K}(\mathbf{s},\mathbf{s}^*;\boldsymbol{\phi}) = \{\mathrm{cov}(w_i(\mathbf{s}), w_j(\mathbf{s}^*))\}_{i,j=1}^m$ 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 $p(\boldsymbol{\beta}, \mathbf{w}, \boldsymbol{\Psi}, \boldsymbol{\phi} \mid \mathbf{Y})$ that is proportional to:

$$p(\boldsymbol{\phi}) \times MVN(\boldsymbol{\beta} \mid \boldsymbol{\mu}_\beta, \Sigma_\beta) \times \prod_{i=1}^m IG(\Psi_i \mid a_i, b_i) MVN(\mathbf{w} \mid \mathbf{0}, \Sigma_{\mathbf{w}(\boldsymbol{\phi})}) \times \prod_{i=1}^n MVN(\mathbf{Y}(\mathbf{s}_i) \mid \mathbf{X}(\mathbf{s}_i)'\boldsymbol{\beta} + \mathbf{w}(\mathbf{s}_i), \boldsymbol{\Psi}), \tag{8}$$

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, 1993). 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 $\mathbf{s}^* \to \mathbf{s}$, $\mathbf{K}(\mathbf{s}, \mathbf{s}^*; \boldsymbol{\phi}) = [\mathrm{cov}(w_i(\mathbf{s}), w_j(\mathbf{s}))]_{i,j=1}^m$ 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 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 $\mathrm{var}(\tilde{w}_i(\mathbf{s})) = 1$, $\mathrm{cov}(\tilde{w}_i(\mathbf{s}), \tilde{w}_i(\mathbf{s}^*)) = \rho_i(\mathbf{s}, \mathbf{s}^*, \boldsymbol{\phi}_i)$ and $\mathrm{cov}(\tilde{w}_i(\mathbf{s}), \tilde{w}_j(\mathbf{s}^*)) = 0$ 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 $\tilde{\mathbf{K}}(\mathbf{s}, \mathbf{s}^*; \boldsymbol{\phi}) = \mathrm{diag}[\rho_i(\mathbf{s}, \mathbf{s}^*; \boldsymbol{\phi}_i)]_{i=1}^m$ 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 $\mathbf{w}(\mathbf{s}, \mathbf{s}^*; \boldsymbol{\phi})$, $\Sigma_{\mathbf{w}} = [\mathbf{K}(\mathbf{s}_i, \mathbf{s}_j; \boldsymbol{\phi})]_{i,j=1}^n$ is

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

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 ith 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 into the following generic template:

$$p(\boldsymbol{\phi}) \times p(\mathbf{A}) \times MVN(\boldsymbol{\beta} \mid \boldsymbol{\mu}_\beta, \Sigma_\beta) \times \prod_{i=1}^m IG(\Psi_i \mid a_i, b_i) MVN(\tilde{\mathbf{w}} \mid \mathbf{0}, \Sigma_{\tilde{\mathbf{w}}}(\boldsymbol{\phi})) \times \prod_{i=1}^n MVN(\mathbf{Y}(\mathbf{s}_i) \mid \mathbf{X}(\mathbf{s}_i)'\boldsymbol{\beta} + \mathbf{A}\tilde{\mathbf{w}}(\mathbf{s}_i), \boldsymbol{\Psi}). \tag{9}$$

Customarily, we let $\boldsymbol{\beta}$ have a flat prior, which corresponds to setting $\boldsymbol{\mu}_\beta = \mathbf{0}$ and letting $\Sigma_\beta \to \infty$. An Inverse Wishart prior could be assigned to covariance matrix $\boldsymbol{\Psi}$ of measurement error, 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 $\delta_i^2 = 1/\Psi_i$ being the ith 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 $\mathbf{K}(\mathbf{0})$. 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 (4) 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 details for a typical iteration is shown in Algorithm 3 and the derivations are in Appendix B.

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 expectations of the functions of $\boldsymbol{\phi}$ and $\mathbf{A}$ to complete the VB iteration. After the algorithm converges, the approximate posterior samples for $\boldsymbol{\phi}$ and $\mathbf{A}$ are obtained using importance sampling resampling method (Rubin, 1987). As we discussed in Section 4.2, the *importance function* can be assigned to a distribution centered at $\mathrm{E}^{(T)}(\boldsymbol{\phi})$ and $\mathrm{E}^{(T)}(\mathbf{A})$ respectively to save computational time. Inferences about $p(\boldsymbol{\phi} \mid \mathbf{Y})$ and $p(\mathbf{A} \mid \mathbf{Y})$ are made based on these samples.

5.2. 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,

$$p(\tilde{\mathbf{w}}|\mathbf{Y}) = \int p(\tilde{\mathbf{w}}|\boldsymbol{\theta}, \mathbf{Y}) p(\boldsymbol{\theta}|\mathbf{Y})\, \mathrm{d}\boldsymbol{\theta}, \tag{10}$$

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

Specify hyper-parameters of prior distributions for Ψ^{-1}, $\mathbf{A}$ and $\boldsymbol{\phi}$.
Give initial values to the expectation of $\Psi^{-1}, \mathbf{A}, \boldsymbol{\phi}, \tilde{\mathbf{w}}, \Sigma_{\tilde{\mathbf{w}}}(\boldsymbol{\phi})^{-1}$ and $\mathbf{A}'\Psi^{-1}\mathbf{A}$: $\mathrm{E}^{(0)}(\delta_i^2) = \delta_i^{2(0)}, \mathrm{E}^{(0)}(\mathbf{A}) = \mathbf{A}^{(0)}, \mathrm{E}^{(0)}(\boldsymbol{\phi}) = \boldsymbol{\phi}^{(0)}$, $\boldsymbol{\mu}_{\tilde{\mathbf{w}}}^{(0)} = \mathbf{0}$, $\mathrm{E}^{(0)}(\mathbf{A}'\Psi^{-1}\mathbf{A}) = \mathbf{A}^{(0)\prime}\Psi^{-1(0)}\mathbf{A}^{(0)}$ and $\mathrm{E}^{(0)}(\Sigma_{\tilde{\mathbf{w}}}(\boldsymbol{\phi})^{-1}) = \Sigma_{\tilde{\mathbf{w}}}\left(\boldsymbol{\phi}^{(0)}\right)^{-1}$.
for $t = 1$ to T **do**
Step 1: Update the distribution of $\boldsymbol{\beta} \sim MVN(\boldsymbol{\mu}_\beta^{(t)}, \mathbf{V}_\beta^{(t)})$, where
$\mathbf{V}_\beta^{(t)} = \left\{\mathbf{X}'\left[\mathbf{I}_n \otimes \mathrm{E}^{(t-1)}(\Psi^{-1})\right]\mathbf{X}\right\}^{-1}$ and
$\boldsymbol{\mu}_\beta^{(t)} = \left\{\mathbf{X}'\left[\mathbf{I}_n \otimes \mathrm{E}^{(t-1)}(\Psi^{-1})\right]\mathbf{X}\right\}^{-1}\mathbf{X}'\left[\mathbf{I}_n \otimes \mathrm{E}^{(t-1)}(\Psi^{-1})\right]\left[\mathbf{Y} - \mathrm{E}^{(t-1)}(\mathscr{A})\boldsymbol{\mu}_{\tilde{\mathbf{w}}}^{(t-1)}\right]$.
Step 2: Update the distribution of $\tilde{\mathbf{w}} \sim MVN(\boldsymbol{\mu}_{\tilde{\mathbf{w}}}^{(t)}, \mathbf{V}_{\tilde{\mathbf{w}}}^{(t)})$, where
$\mathbf{V}_{\tilde{\mathbf{w}}}^{(t)} = \left\{\mathrm{E}^{(t-1)}\left[\mathscr{A}'\left(\mathbf{I}_n \otimes \mathrm{E}^{(t-1)}(\Psi^{-1})\right)\mathscr{A}\right] + \mathrm{E}^{(t-1)}\left(\Sigma_{\tilde{\mathbf{w}}}^{-1}\right)\right\}^{-1}$ and
$\boldsymbol{\mu}_{\tilde{\mathbf{w}}}^{(t)} = \mathbf{V}_{\tilde{\mathbf{w}}}^{(t)}\mathrm{E}^{(t-1)}(\mathscr{A})'\left[\mathbf{I}_n \otimes \mathrm{E}^{(t-1)}(\Psi^{-1})\right]\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\mu}_\beta^{(t)}\right)$.
Step 3: Update the distribution of $\delta_j^2 \sim Gamma\left(\frac{n}{2} + a_j, \left(\frac{1}{b_j} + d_j^{(t)}\right)^{-1}\right)$, where $\mathrm{E}^{(t)}(\delta_j^2) = \left(\frac{n}{2} + a_j\right)\left(\frac{1}{b_j} + d_j^{(t)}\right)^{-1}$ and $d_j^{(t)}$ is defined in Appendix B.
Step 4: Update the distribution of $\boldsymbol{\phi}$, which is proportional to
$$\sqrt{|\Sigma_{\tilde{\mathbf{w}}}^{-1}|}\exp\left\{-\frac{\boldsymbol{\mu}_{\tilde{\mathbf{w}}}^{(t)\prime}\Sigma_{\tilde{\mathbf{w}}}^{-1}\boldsymbol{\mu}_{\tilde{\mathbf{w}}}^{(t)} + \mathrm{Tr}\left(\mathbf{V}_{\tilde{\mathbf{w}}}^{(t)}\Sigma_{\tilde{\mathbf{w}}}^{-1}\right)}{2}\right\}$$
and calculate $\mathrm{E}^{(t)}(\boldsymbol{\phi})$ and $\mathrm{E}^{(t)}\left(\Sigma_{\tilde{\mathbf{w}}}(\boldsymbol{\phi})^{-1}\right)$ using importance sampling.
Step 5: Update the distribution of $\mathbf{A}$, which is proportional to
$$\exp\left\{-\frac{\left(\mathbf{Y}-\mathbf{X}\boldsymbol{\mu}_\beta^{(t)}-\mathscr{A}\boldsymbol{\mu}_{\tilde{\mathbf{w}}}^{(t)}\right)'\left[\mathbf{I}_n\otimes\mathrm{E}^{(t)}(\Psi^{-1})\right]\left(\mathbf{Y}-\mathbf{X}\boldsymbol{\mu}_\beta^{(t)}-\mathscr{A}\boldsymbol{\mu}_{\tilde{\mathbf{w}}}^{(t)}\right)+\mathrm{Tr}\left[(\mathbf{A}\mathbf{A}')^{-1}\mathbf{S}\right]}{2}\right\}$$
$$\times\exp\left\{-\frac{Tr\left[\mathscr{A}\mathbf{V}_{\tilde{\mathbf{w}}}^{(t)}\mathscr{A}'\left(\mathbf{I}_n\otimes\mathrm{E}^{(t)}(\Psi^{-1})\right)\right]}{2}\right\}\times|\mathbf{A}\mathbf{A}'|^{-(df+m+1)/2}$$
and calculate $\mathrm{E}^{(t)}(\mathbf{A})$, $\mathrm{E}^{(t)}\left(\mathbf{A}\mathrm{E}^{(t)}(\Psi)\mathbf{A}'\right)$ using importance sampling.
end for

where $\boldsymbol{\theta}$ denotes all the parameters of the marginal model. the conditional distribution $p(\tilde{\mathbf{w}}|\boldsymbol{\theta}, \mathbf{Y})$ is

$$MVN([\Sigma_{\tilde{\mathbf{w}}}^{-1} + \mathscr{A}'(\mathbf{I}_n \otimes \boldsymbol{\Psi}^{-1})\mathscr{A}]^{-1}\mathscr{A}'(\mathbf{I}_n \otimes \boldsymbol{\Psi}^{-1})(\mathbf{Y} - \mathbf{X}\beta),\ [\Sigma_{\tilde{\mathbf{w}}}^{-1} + \mathscr{A}'(\mathbf{I}_n \otimes \boldsymbol{\Psi}^{-1})\mathscr{A}]^{-1}).$$

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

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

The distribution of $\tilde{\mathbf{w}}^*$ conditional on $\tilde{\mathbf{w}}$ 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 the distribution $p(\tilde{\mathbf{w}}^* \mid \tilde{\mathbf{w}}, \boldsymbol{\theta}, \mathbf{Y})$ is $MVN(\mu_{\tilde{\mathbf{w}}^*|\tilde{\mathbf{w}}}, \Sigma_{\tilde{\mathbf{w}}^*|\tilde{\mathbf{w}}})$, where

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

In the hidden variable model the posterior distribution of $\tilde{\mathbf{w}}$ is estimated by $q(\tilde{\mathbf{w}})$. So $q(\tilde{\mathbf{w}})q(\boldsymbol{\theta})$ is utilized 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 dependents on spatial parameters $\boldsymbol{\phi}$. Then the conditional expectation of $\tilde{\mathbf{w}}^*$ given $\mathbf{Y}$ is

Table 1
The posterior means and variances of the parameters from VB and MCMC in ordinary linear model.

		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

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

The other way to approximate $\tilde{\mathbf{w}}^*$'s posterior expectation only uses the posterior distribution of the parameters as (10). Then $\mathrm{E}(\tilde{\mathbf{w}}^*|\mathbf{Y})$ is

$$\begin{aligned}
\mathrm{E}(\tilde{\mathbf{w}}^*|\mathbf{Y}) &= \int \tilde{\mathbf{w}}^* p(\tilde{\mathbf{w}}^*|\tilde{\mathbf{w}}, \boldsymbol{\phi}) p(\tilde{\mathbf{w}}|\boldsymbol{\theta}, \mathbf{Y}) p(\boldsymbol{\theta}|\mathbf{Y})\, \mathrm{d}\boldsymbol{\theta}\mathrm{d}\tilde{\mathbf{w}}\mathrm{d}\tilde{\mathbf{w}}^* \\
&= \int \Sigma'_{\tilde{\mathbf{w}},\tilde{\mathbf{w}}^*} \Sigma_{\tilde{\mathbf{w}}}^{-1} \tilde{\mathbf{w}} p(\tilde{\mathbf{w}}|\boldsymbol{\theta}, \mathbf{Y}) p(\boldsymbol{\theta}|\mathbf{Y})\, \mathrm{d}\boldsymbol{\theta}\mathrm{d}\tilde{\mathbf{w}} \simeq \int \Sigma'_{\tilde{\mathbf{w}},\tilde{\mathbf{w}}^*} \Sigma_{\tilde{\mathbf{w}}}^{-1} \boldsymbol{\mu}_p(\tilde{\mathbf{w}}) q(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta} \\
&= \mathrm{E}_q(\Sigma'_{\tilde{\mathbf{w}},\tilde{\mathbf{w}}^*} \Sigma_{\tilde{\mathbf{w}}}^{-1} \boldsymbol{\mu}_p(\tilde{\mathbf{w}})), \quad \text{where } \boldsymbol{\mu}_p(\tilde{\mathbf{w}}) = [\Sigma_{\tilde{\mathbf{w}}}^{-1} + \mathscr{A}'(\mathbf{I}_n \otimes \boldsymbol{\Psi}^{-1})\mathscr{A}]^{-1}\mathscr{A}'(\mathbf{I}_n \otimes \boldsymbol{\Psi}^{-1})(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}).
\end{aligned}$$

6. Illustrations

6.1. Simulated example of simple linear regression

We begin by illustrating the ability of the VB algorithm to estimate the posteriors in the simple Bayesian conjugate linear model introduced in Section 3. 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 fixed to be 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 3, the VB algorithm converges within 5 iterations. Table 1 compares the posterior means and variances obtained from VB and MCMC. As discussed in Section 3 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 Fig. 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 Fig. 1 due to the similarity with the VB estimates.

6.2. 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 4.1. 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, $\mathrm{cov}(\mathbf{w}(\mathbf{s}), \mathbf{w}(\mathbf{s}^*)) = \rho(\mathbf{s} - \mathbf{s}^*; \phi) = \exp(-\phi\|\mathbf{s} - \mathbf{s}^*\|)$. The data set was simulated with the following parameters: $\boldsymbol{\beta}' = (150, 10)$, $\tau^2 = 20$, $\sigma^2 = 50$, $\phi = 0.1$. 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 Sections 4.1 and 4.2 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 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 standard errors) are shown in Table 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 3000 samples for burn-in. The VB algorithms were run until the hyper-parameters of the posterior distributions converged. Because of different parameterizations, not all the parameters in the table were directly updated

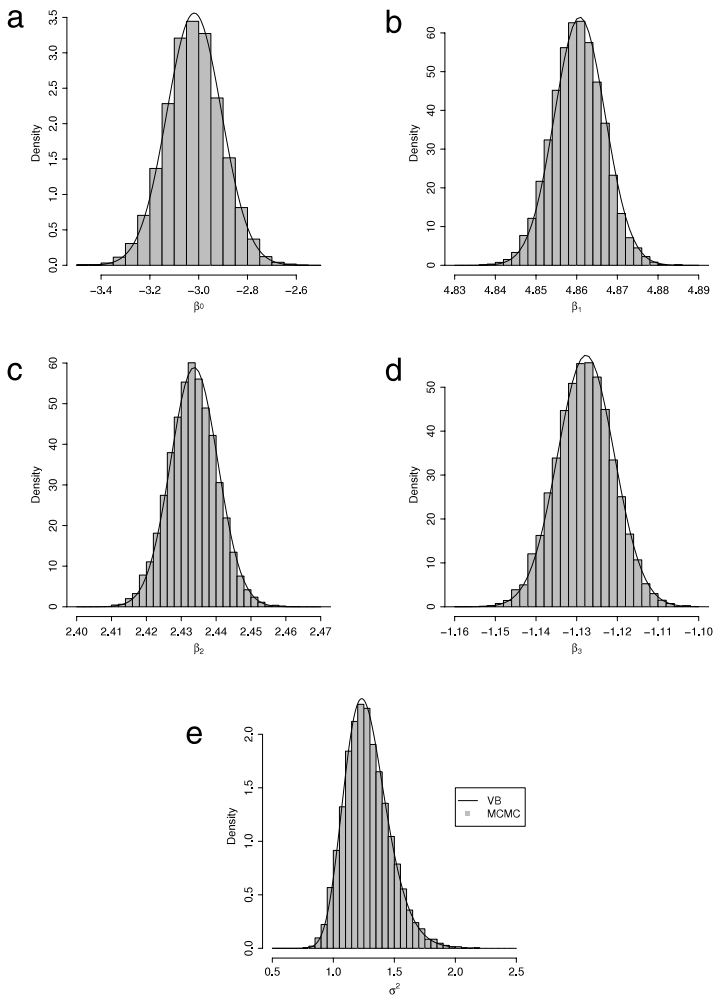

Fig. 1. Linear regression example: (a)–(d) compare the posterior distributions of $\boldsymbol{\beta}$ from VB with the MCMC samples. (e) compares the posterior distribution of σ^2 from VB with MCMC samples. The solid line is the approximate posterior distribution from VB and the histogram is the samples from MCMC.

Table 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.

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)	

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 product the posterior samples for the parameters. (The estimates of these parameters are displayed in boldface in Table 2.) Fig. 2 offers the trace plots for the VB marginal model assuming $p(\boldsymbol{\theta} \mid \mathbf{Y}) \simeq$

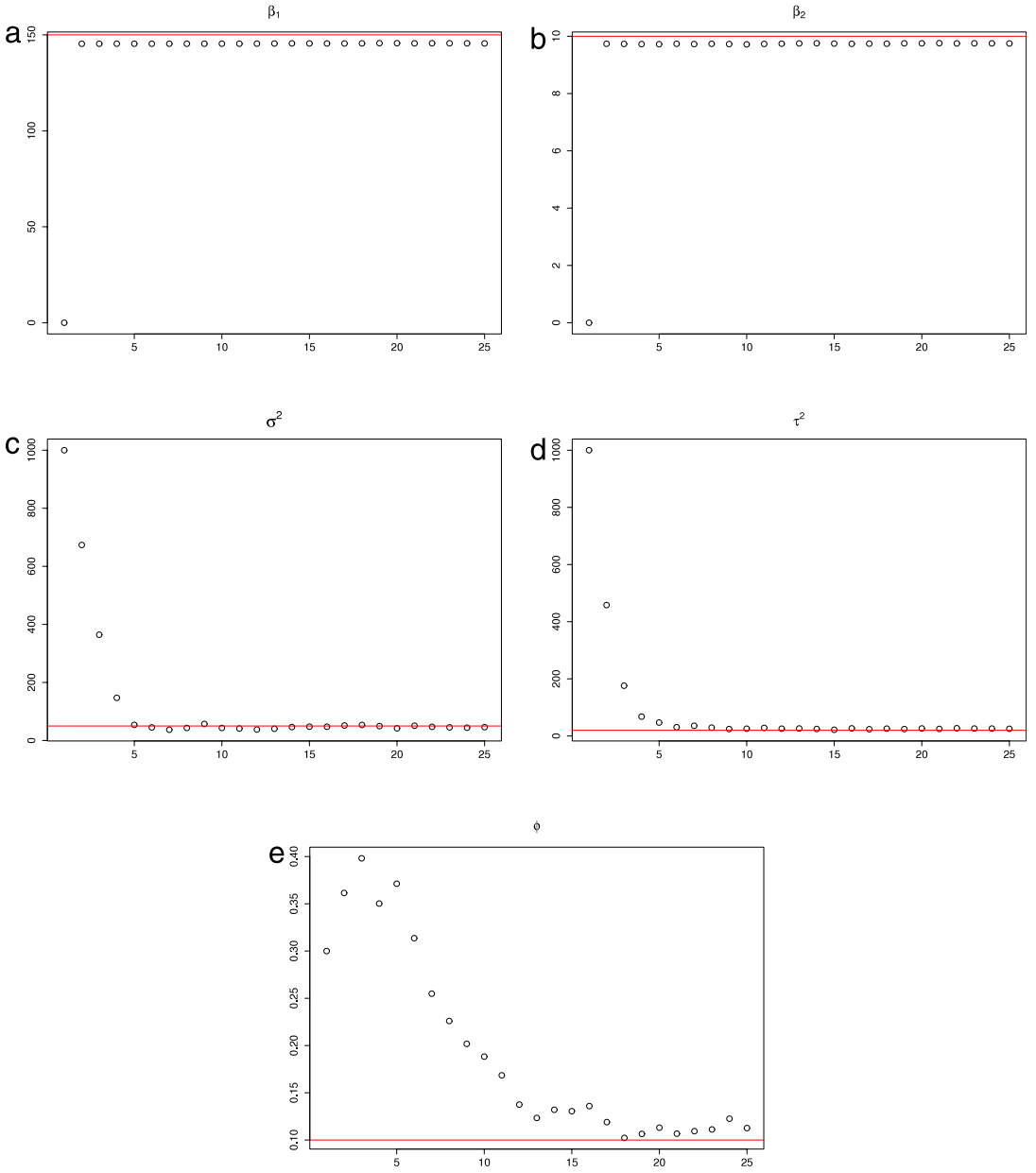

Fig. 2. The trace plot of VB marginal model assuming $p(\boldsymbol{\theta}|\mathbf{Y}) \simeq q(\sigma^2, \tau^2, \phi)q(\boldsymbol{\beta})$. The subfigure (a)–(e) are the mean of β_1, β_2, σ^2, τ^2 and ϕ respectively.

$q(\sigma^2, \tau^2, \phi)q(\boldsymbol{\beta})$, 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, 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, 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 sensitive to starting values. In multivariate spatial models, using `nlminb` to find the BCLT estimates would give problematic results, which are shown in Section 6.3. The coverage of the 95% confidence intervals of the VB hidden variable method is smaller than 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 shrinkage. 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.

Fig. 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

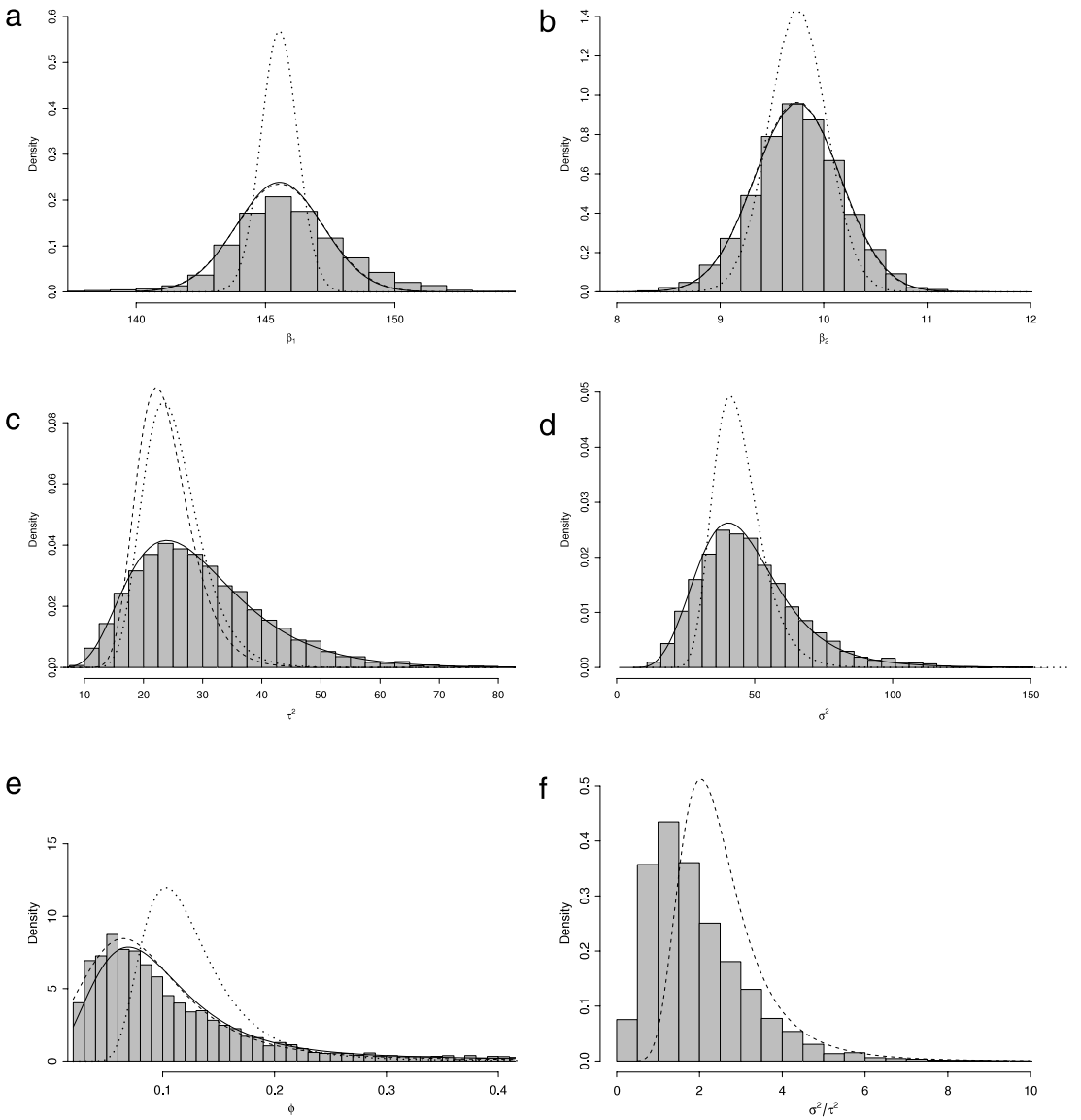

Fig. 3. The posterior distributions got from different methods: MCMC(histogram); VB treating $\mathbf{w}$ as the hidden variable (dotted line for $\boldsymbol{\beta}$, σ^2, τ^2 and ϕ); VB marginal model with $p(\boldsymbol{\theta} \mid \mathbf{Y}) \simeq q(\sigma^2, \tau^2, \phi)q(\boldsymbol{\beta})$ (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.)

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.

6.3. 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 $\tilde{\mathbf{K}}(\mathbf{s} - \mathbf{s}^*; \boldsymbol{\phi}) = \mathrm{diag}[\rho_i(\mathbf{s} - \mathbf{s}^*; \phi_i)]_{i=1}^2$ where $\boldsymbol{\phi} = (\phi_1, \phi_2)$. The multivariate process was simulated with the following parameters:

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

This yields $\mathbf{A} = \mathbf{K}^{1/2} = \begin{pmatrix} 1 & 0 \\ -2 & 2 \end{pmatrix}$. In 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

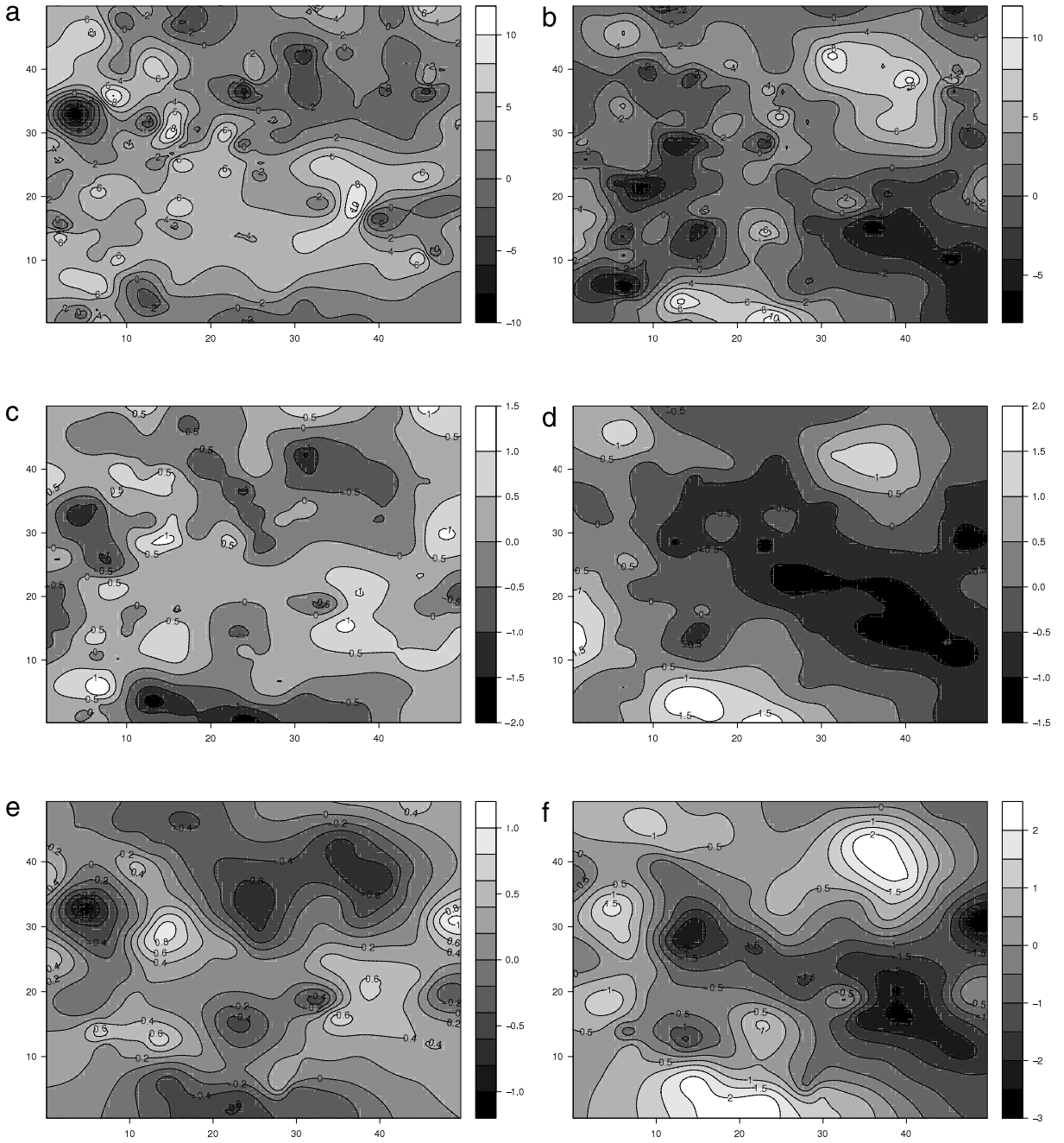

Fig. 4. (a) and (b) are interpolated surfaces of the first and second response variable. (c) and (d) are the interpolated surface of the recovered random spatial effects from VB method E[$\mathbf{w}|Data$]. (e) and (f) are the interpolated surface of the predicted random spatial effects E[$\mathbf{w}^*|Data$] from VB method.

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 $\xi = 0.05$. The above specifications describe a multivariate process with independent non-spatial variance among the response surfaces and a strong negative cross-correlation between the spatial processes. 150 observations were generated by using the parameters above in model (9). Fig. 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 E($\mathbf{w}^*$) (e and f).

MCMC algorithms using the spMvLM function of the spBayes package in R took approximately 1.5 h to deliver its entire inferential output involving 20,000 iterations, including 5000 samples for burn-in, on a 2.8 GHz AMD Athlon processor with 2.0 GB of RAM running under Windows. Under the same conditions, the VB algorithm took about 0.5 h to converge. For the VB method, the expectations of both $\mathbf{w}$ and $\mathbf{w}^*$ were calculated using the method introduced in Section 5.2, and $q(\tilde{\mathbf{w}})$ is the estimate for the true posterior of $\tilde{\mathbf{w}}$.

Table 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.

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

Percentiles of the posterior distributions estimated using VB, MCMC and BCLT are listed in Table 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 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 we tried different initial values. So the posterior percentiles were not able to provided in Table 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.

6.4. Model selection

The generalized template introduced in Section 5 suggests several potential models. Here we consider five stationary process models of increasing complexity on the same synthetic data set introduced in Section 6.3. Our focus is on the alternative specifications of $\mathbf{A}$ and $\tilde{\mathbf{w}}$ within (9). 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

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),

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

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

Model 3: $\mathbf{A} = \text{diag}[\sigma_i]_{i=1}^{m}$ 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 $\mathbf{K}$,

Model 4: $\mathbf{A}$ 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,

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

We fit the five competing models to the synthetic data in Section 6.3 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. G is an error sum of squares and represents goodness-of-fit, while P represents predictive variance and acts as a penalty term. For underfitted models, predictive variances will tend to be large and thus so will P; but also for overfitted models we expect inflated predictive variances, again making P large. Eventually, complexity is penalized and a parsimonious choice with lower $G + P$ is encouraged.

Table 4
Synthetic data model comparison using DIC and minimum posterior predictive approach. For each model unmarginalized scores were calculated from 1000 samples.

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	$\phi, \boldsymbol{\sigma}_m^2, \boldsymbol{\Psi}$	1326.33	2119.56	3445.89	1438.11
Model 4	$\phi, \mathbf{A}, \boldsymbol{\Psi}$	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 4 provides DIC and posterior predictive loss approach for candidate models. Based on DIC and the actual 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 3. Therefore we would conclude that Model 5 is preferred for this data set.

6.5. 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 + foliaged) 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 date 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 Fig. 5(a) are used to produce an interpolated surface of the biomass for each of the two categories, Fig. 5(b) and (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 $\boldsymbol{\Psi}$. Further, we assume that spatial dependence can be modeled with the simple exponential correlation function. This specification corresponds to Model 5 in Section 6.4. The Inverse Wishart prior is used for $\mathbf{K}$ and Inverse Gamma prior for the diagonal elements of $\boldsymbol{\Psi}$. 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 1000 iterations; Therefore, 27,000 samples were retained for posterior analysis.

The VB and MCMC estimates for the Forest inventory data are provided in Table 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 5 do not support a distinction between the responses' spatial dependence structure, therefore, the separable form of this model might be considered.

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 5.2. The interpolated surfaces of the predicted random spatial effects $\mathbf{w}^*$ and biomass $\mathbf{Y}^*$ are shown in Fig. 6(a) and (b), respectively.

7. 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 expensive matrix decomposition. We proposed VB algorithm to fit both unmarginalized and marginalized likelihoods (as in (2) and (3)). The unmarginalized model offers the advantage of closed form expressions for $\boldsymbol{\beta}$, τ^2 and σ^2. However, these computational

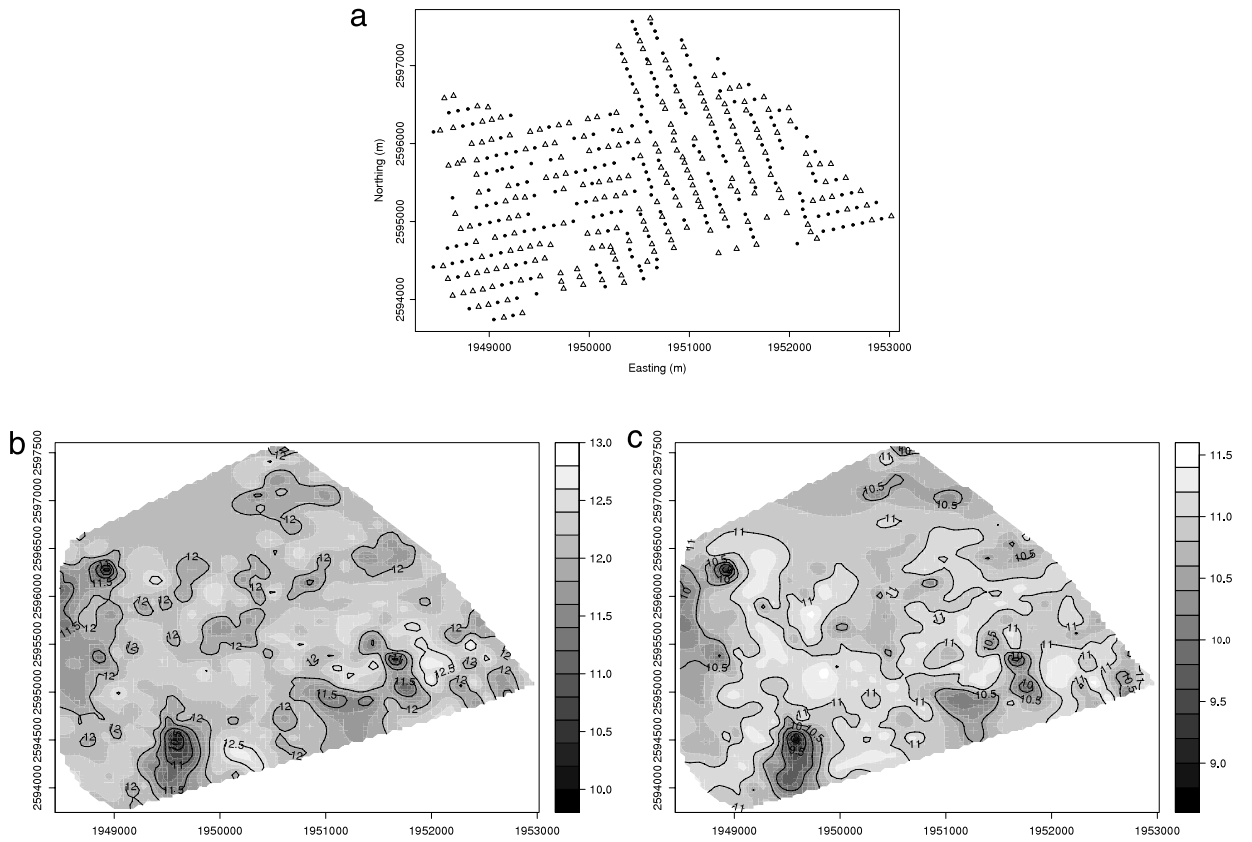

Fig. 5. (a) Forest inventory plots across the Bartlett Experimental Forest. The 415 plots were divided randomly into 200 plots used for parameter estimation denoted with solid dot symbols (•) and the remaining 215 used for prediction marked with triangle symbols (△). Plots (b) and (c) are interpolated surfaces of biomass per hectare of the bole, and non-bole, respectively.

Table 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.

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)

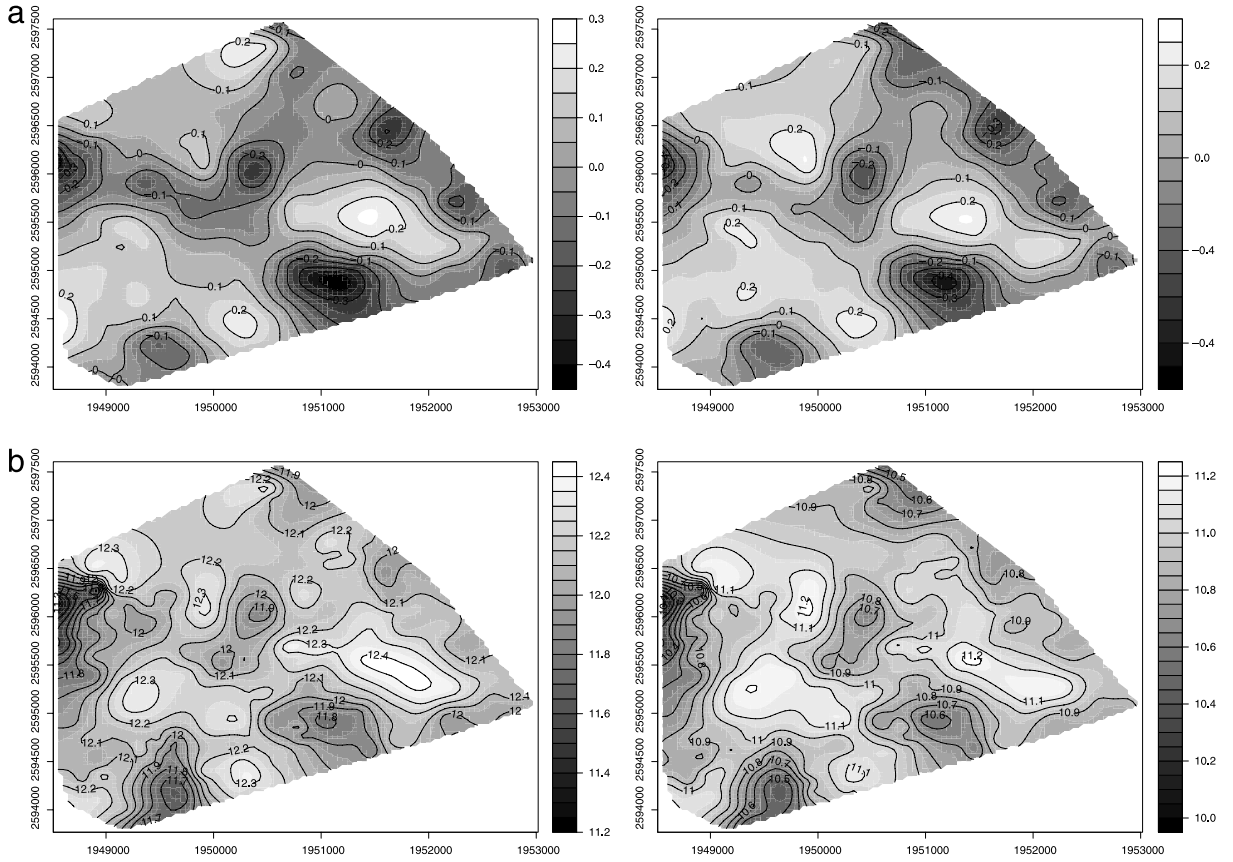

Fig. 6. (a) Interpolated surfaces of the predicted random spatial effects for biomass per hectare of the bole (left plot) and non-bole (right plot), $\mathrm{E}[\mathbf{w}^*|Data]$. (b) Interpolated surfaces of the posterior predictive distributions for biomass per hectare of the bole (left plot) and non-bole (right plot), $\mathrm{E}(\mathbf{Y}^*|Data)$.

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.

Appendix A. Variational calculation

We wish to maximize the function:

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

with respect to each factorized distribution in turn. $\mathcal{L}$ is a functional, i.e. $\mathcal{L} = \int f(\boldsymbol{\theta}, q(\boldsymbol{\theta}))\, \mathrm{d}\boldsymbol{\theta}$. 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:

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

where:

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

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

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

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

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

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

Appendix B. VB algorithm for multivariate spatial Bayesian model treating w as hidden variable

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

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

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

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

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,

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

where

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

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

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

where

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

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 $\mathscr{A}'[\mathbf{I}_n\otimes\mathrm{E}^{(t)}(\Psi^{-1})]\mathscr{A}$. Next, we can update the distribution of Ψ.

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

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

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

Since the measurement errors are assumed to be independent, Ψ^{-1} is a diagonal matrix, so is $\mathbf{I}_n\otimes\Psi^{-1}$, with diagonal elements δ_i^2, $i=1,\ldots,m$. The trace of $(\mathbf{I}_n\otimes\Psi^{-1})\mathbf{D}^{(t+1)}$ 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 $d_i^{(t+1)}=\sum_{j=0}^{n-1}\mathbf{D}_{jm+i,jm+i}^{(t+1)}$. 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

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

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

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

References

Attias, H., 2000. A variational bayesian framework for graphical models. In: Advances in Neural Information Processing Systems, vol. 12. MIT Press, Cambridge, MA, pp. 209–215.

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

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

Best, N., Marshall, C., Thomas, A., Nov. 30–Dec. 1 2000. Spatial modeling using winbugs and geobugs, short course. Brisbane.

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

Bolin, D., 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.

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

Chilés, J., Delfiner, P., 1999. Geostatistics: Modeling Spatial Uncertainty. Wiley, New York.

Cressie, N., 1993. Statistics for Spatial Data, 2nd ed.. Wiley, New York.

Finley, O.A., Banerjee, S., Ek, R.A., Mcroberts, E.R., 2008. Bayesian multivariate process modeling for prediction of forest attributes. Journal of Agricultural, Biological, and Environmental Statistics 13, 60–83.

Friedman, N., 1998. The bayesian structural em algorithm. In: Fourteenth Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann, pp. 129–138.

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

Gelfand, I.M., Fomin, S., 1963. Calculus of Variations. Rentice-Hall, Inc., Englewood CliKs, New Jersey.

Gelfand, A., Ghosh, S., 1998. Model choice: a minimum posterior predictive loss approach. Biometrika 85, 1–11.

Gelfand, A., Kim, H., Sirmans, C., Banerjee, S., 2003. Spatial modelling with spatially varying coefficient processes. Journal of the American Statistical Association 98, 387–396.

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

Ghahramani, Z., Beal, M.J., 2000. Variational inference for bayesian mixtures of factor analysers. In: Advances in Neural Information Processing Systems, vol. 12. MIT Press, pp. 449–455.

Higdon, D., 2001. Space and space time modeling using process convolutions. Tech. Rep., ISDS Duke University.

Hinton, G.E., 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. ACM Press, New York, NY, USA, pp. 5–13.

Jones, R., Zhang, Y., 1997. Models for continuous stationary space-time processes. In: Diggle, P., Warren, W., Wolfinger, R. (Eds.), Modeling Longitudinal and Spatially Correlated Data: Methods, Applications and Future Directions. Springer-Verlag, New York.

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

Kamman, E., Wand, M., 2003. Geoadditive models. Journal of the Royal Statistical Society: Series C (Applied Statistics) 52 (1), 1–18.

Lindgern, F., Lindström, J., 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.

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

MacKay, J.D., 1997. Ensemble learning for hidden markov models. Tech. Rep., Cavendish Laboratory, University of Cambridge.

MacKay, D.J.C., 2003. Information Theory, Inference, and Learning Algorithms. Cambridge University Press.

Neal, R., Hinton, G.E., 1998. A View of the Em Algorithm that Justifies Incremental, Sparse, and Other Variants. Kluwer Academic Publishers.

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

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

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., Held, L., 2005. Gaussian Markov Random Fields: Theory and Applications. Chapman and Hall/CRC Press, Boca Raton, FL.

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

Spiegelhalter, D.J., Best, N.G., Carlin, B.P., 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., 1999. Interpolation of Spatial Data: Some Theory of Kriging. Springer, New York.

Stein, M., Chi, Z., Welty, L., 2004. Approximating likelihoods for large spatial datasets. Journal of the Royal Statistical Society, Series B 66, 275–296.

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., Cressie, N., 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, 3rd ed.. Springer-Verlag, New York.

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

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

Xia, G., Gelfand, A., 2006. Stationary process approximation for the analysis of large spatial datasets. Tech. Rep., ISDS, Duke University.