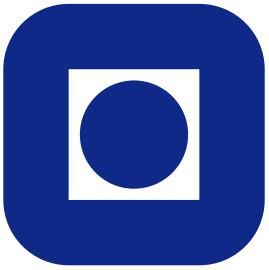
NORGES TEKNISK-NATURVITENSKAPELIGE
UNIVERSITET

Approximate Bayesian Inference for Large Spatial Datasets Using Predictive Process Models

by

Jo Eidsvik, Andrew O. Finley, Sudipto Banerjee and Håvard Rue

PREPRINT
STATISTICS NO. 9/2010



NORWEGIAN UNIVERSITY OF SCIENCE AND
TECHNOLOGY
TRONDHEIM, NORWAY

This preprint has URL http://www.math.ntnu.no/preprint/statistics/2010/S9-2010.pdf
Håvard Rue has homepage: http://www.math.ntnu.no/∼hrue
E-mail: hrue@math.ntnu.no
Address: Department of Mathematical Sciences, Norwegian University of Science and Technology, N-7491 Trondheim, Norway.

Approximate Bayesian Inference for Large Spatial Datasets Using Predictive Process Models

Jo Eidsvik

Department of Mathematical Sciences,

Norwegian University of Science and Technology, Norway

and

Andrew O Finley

Department of Forestry, Michigan State University, MI, USA

and

Sudipto Banerjee

Department of Biostatistics, University of Minnesota, MN, USA

and

Håvard Rue

Department of Mathematical Sciences,

Norwegian University of Science and Technology, Norway

May 17, 2010

Abstract

This article addresses the challenges of estimating hierarchical spatial models to large datasets. With the increasing availability of geocoded scientific data, hierarchical models involving spatial processes have become a popular method for carrying out spatial inference.

Such models are customarily estimated using Markov chain Monte Carlo algorithms that, while immensely flexible, can become prohibitively expensive. In particular, fitting hierarchical spatial models often involves expensive decompositions of dense matrices whose computational complexity increases in cubic order with the number of spatial locations. Such matrix computations are required in each iteration of the Markov chain Monte Carlo algorithm, rendering them infeasible for large spatial data sets.

This article proposes to address the computational challenges in modeling large spatial datasets by merging two recent developments. First, we use the predictive process model as a reduced-rank spatial process, to diminish the dimensionality of the model. Then we proceed to develop a computational framework for estimating predictive process models using the integrated nested Laplace approximation. We discuss settings where the first stage likelihood is Gaussian or non-Gaussian. Issues such as predictions and model comparisons are also discussed. Results are presented for synthetic data, an environmental dataset and for a large dataset on forest biomass.

Keywords: Approximate Bayesian inference; Computational statistics; Gaussian processes; Geostatistics; Laplace approximation; Predictive process model.

1 Introduction

Recent advances in Geographical Information Systems (GIS) and Global Positioning Systems (GPS) enable accurate geocoding of locations where scientific data are collected. This has encouraged formation of large spatiotemporal datasets in many fields and has generated considerable interest in statistical modelling for such data; see, for example, the books by Cressie (1993), Banerjee et al. (2004), and Schabenberger and Gotway (2004). Here, we focus upon 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 involves matrix decompositions whose complexity increases as $O(n^3)$ in the number of locations, n, hence the infeasibility or "big n" problem for large datasets.

Modelling large spatial datasets have received much attention in the recent past. Vecchia (1988)

proposed approximating the likelihood with a product of appropriate conditional distributions to obtain maximum-likelihood estimates. Stein et al. (2004) adapt this to restricted maximum likelihood estimation. Another possibility is to approximate the likelihood using spectral representations of the spatial process (Fuentes, 2007). These likelihood approximations yield a joint distribution, but not a process that facilitates spatial interpolation. Yet another approach considers compactly supported correlation functions (Furrer et al., 2006; Kaufman et al., 2008; Du et al., 2009) that yield sparse correlation structures. More efficient sparse solvers can then be employed for kriging and variance estimation, but the tapered structures may limit modeling flexibility. Also, full likelihood-based inference still requires determinant computations that may be problematic.

Rather than approximations, one could build models especially geared towards handling of large spatial datasets. These are representations of the spatial process in a lower-dimensional subspace and are often referred to as low-rank or reduced-rank spatial models (Higdon, 2002; Kamman and Wand, 2003; Stein, 2007, 2008; Cressie and Johannesson 2008; Banerjee et al., 2008; Crainiceaniu et al., 2008). Many of these methods are variants of the so-called "subset of regressors" methods used in Gaussian process regressions for large data sets in machine learning (e.g. Rasmussen and Williams, 2006). The idea here is to consider a smaller set of locations, or "knots", say $\mathcal{S}^* = \{\boldsymbol{s}_1^*, \ldots, \boldsymbol{s}_{n^*}^*\}$, where the number of knots, n^*, is *fixed* to be much smaller than the number of observed sites, and to express the spatial process realizations over n locations in terms of its realizations over the smaller set of knots. It is reasonable to assume there will be insignificant loss of spatial information in the underlying process from using a smaller set of locations – the knots – with adequate domain coverage. Subsequently, we will consider a special class of low-rank processes called the *predictive process* (Banerjee et al., 2008). This arises from a conditional expectation of the original process (often referred to as the *parent process*) given its realization over the knots.

A key issue in such methods is the number and selection of knots which is a challenging problem with choice in two dimensions more difficult than in one. The choice of n^* is governed by computational cost and sensitivity to choice. Customarily, the analysis is implemented over different choices of n^*. The selection of the sites that will act as knots is an even more complex problem

and raises the question of whether to use a subset of the observed spatial locations or a disjoint set of locations. The issue is not dissimilar to a spatial design problem (e.g. Nychka and Saltzman, 1998; Xia et al., 2006; Diggle and Lophaven, 2006). Finley et al. (2009) explored the knot selection issue for predictive processes. In practice, one must estimate predictive process models with different choices of knots to arrive at configurations yielding reliable and robust inference. Using Markov chain Monte Carlo (MCMC) for such experimentations will, however, be a daunting task and fast, accurate approximation methods will need to be explored.

In recent work Rue et al. (2009) propose an Integrated Nested Laplace Approximation (INLA) algorithm as an alternative to MCMC for latent Gaussian models. INLA presents a very versatile template for estimating latent Gaussian models by repeated use of the Laplace approximation (LA), (see Tierney and Kadane, 1986). Rue et al. (2009) use computationally effective Gaussian Markov random field approximations (Rue and Held, 2005) to deliver fast and accurate approximations to posterior marginals. Eidsvik et al. (2009) use the same Laplace techniques for irregular moderate size data from a spatial Generalized Linear Mixed Model (GLMM). Extensive studies conducted by Eidsvik et al. (2009) and Rue et al. (2009) reveal that, for a wide class of latent Gaussian models, INLA produces inference that is essentially indistinguishable from MCMC in a mere fraction of the time required by the latter. The key to succesful use of INLA, is a reasonable Gaussian approximation to the full conditional of the latent variables, including regression effects. A numerical optimization and integration routine is used for the covariance hyperparameters. The LA has been a powerful tool in statistical inference. Frequentist approaches use the LA for marginalized likelihood inference, see e.g. Breslow and Clayton (1993) and Ainsworth and Dean (2006). In the Bayesian context it has been applied for model choice using Bayes factors, but then the full conditionals are usually approximated by sampling, see e.g. Chib (1995) and Lewis and Raftery (1997). Hsiao et al (2004) use the LA for related purposes, referring to the Laplace expression by Candidate's formula.

This article presents a framework for estimating predictive process models using INLA. The remainder of the article evolves as follows. Section 2 discusses the spatial predictive process, its properties and how it is employed in hierarchical spatial GLMM context. Section 3 outlines

approximate Bayesian inference using INLA. Section 4 considers a number of simulation experiments as well as practical illustrations from fisheries and forestry. Finally, Section 5 concludes the article with a discussion and an eye towards future work.

2 Hierarchical modeling with the predictive process

2.1 The Gaussian Predictive Process

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

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

That is, the residual comprises a spatial process, $w(\boldsymbol{s})$, and an independent process, $\epsilon(\boldsymbol{s})$, often called the *nugget*. The $w(\boldsymbol{s})$ are spatial random effects, providing local adjustment (with structured dependence) to the mean, interpreted as capturing the effect of unmeasured or unobserved covariates with spatial pattern.

The customary process specification for $w(\boldsymbol{s})$ is a mean 0 Gaussian Process with covariance function, $C(\boldsymbol{s}_1, \boldsymbol{s}_2)$, denoted $GP(0, C(\boldsymbol{s}_1, \boldsymbol{s}_2))$. In applications, we often specify $C(\boldsymbol{s}_1, \boldsymbol{s}_2) = \sigma^2 \rho(\boldsymbol{s}_1, \boldsymbol{s}_2; \boldsymbol{\phi})$ 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, $\epsilon(\boldsymbol{s}) \overset{iid}{\sim} N(0, \tau^2)$ for every location $\boldsymbol{s}$. Prior distributions on the remaining parameters complete the hierarchical model. Customarily, the regression effect $\boldsymbol{\beta}$ is assigned a multivariate Gaussian prior, i.e. $\boldsymbol{\beta} \sim N(\boldsymbol{\mu}_\beta, \Sigma_\beta)$, while the latent variance component σ^2 and the nugget variance τ^2 are assigned $IG(\cdot, \cdot)$ priors. The process correlation parameter(s), $\boldsymbol{\phi}$, are usually assigned some informative priors (e.g. uniform over a finite range) based upon the underlying spatial domain.

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

hood that combines with the customary hierarchical specifications to yield a posterior distribution $\pi(\boldsymbol{\beta}, \boldsymbol{w}, \sigma^2, \tau^2, \boldsymbol{\phi} \mid \boldsymbol{Y})$ that is proportional to

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

Often a marginalized likelihood is used that is obtained by integrating out the spatial effects $\boldsymbol{w}$ and the regression coefficients $\boldsymbol{\beta}$. This yields

$$\pi(\sigma^2, \tau^2, \boldsymbol{\phi} \mid \boldsymbol{Y}) \propto \pi(\boldsymbol{\phi}) \times IG(\tau^2 \mid a_\tau, b_\tau) \times IG(\sigma^2 \mid a_\sigma, b_\sigma)$$
$$N\left(\boldsymbol{Y} \mid \boldsymbol{X}\boldsymbol{\mu}_\beta, \sigma^2 \boldsymbol{R}(\boldsymbol{\phi}) + \tau^2 \boldsymbol{I} + \boldsymbol{X}\boldsymbol{\Sigma}_\beta \boldsymbol{X}'\right), \tag{3}$$

where row i of matrix $\boldsymbol{X}$ is $\boldsymbol{x}(\boldsymbol{s}_i)'$. This marginalization over $\boldsymbol{w}$ and $\boldsymbol{\beta}$ can be interpreted as a ratio of joints and conditionals since,

$$\pi(\boldsymbol{Y} \mid \cdot) = \int_{\boldsymbol{w},\boldsymbol{\beta}} \pi(\boldsymbol{Y}, \boldsymbol{w}, \boldsymbol{\beta} \mid \cdot) d\boldsymbol{w} d\boldsymbol{\beta} = \frac{\pi(\boldsymbol{Y}, \boldsymbol{w}, \boldsymbol{\beta} \mid \cdot)}{\pi(\boldsymbol{w}, \boldsymbol{\beta} \mid \boldsymbol{Y}, \cdot)}. \tag{4}$$

In fact, we will utilize this in the LA below. The marginal posterior distribution of the spatial effects and regression parameters is given by

$$\pi(\boldsymbol{w}, \boldsymbol{\beta} \mid \boldsymbol{Y}) = \int \pi\left(\boldsymbol{w}, \boldsymbol{\beta} \mid \boldsymbol{Y}, \sigma^2, \tau^2, \boldsymbol{\phi}\right) \pi(\sigma^2, \tau^2, \boldsymbol{\phi} \mid \boldsymbol{Y}) d\sigma^2 d\tau^2 d\boldsymbol{\phi},$$

where $\pi\left(\boldsymbol{w}, \boldsymbol{\beta} \mid \boldsymbol{Y}, \sigma^2, \tau^2, \boldsymbol{\phi}\right)$ is a multivariate normal distribution.

Irrespective of whether we use (2) or (3), estimation and prediction will require matrix factorizations involving the dense $n \times n$ matrix $\boldsymbol{R}(\boldsymbol{\phi})$ which may become prohibitively expensive for large n. Recently Banerjee et al. (2008) proposed a class of knot-based spatial process models for large spatial datasets. These models consider a fixed set of "knots" $\mathcal{S}^* = (\boldsymbol{s}_1^*, \ldots, \boldsymbol{s}_{n^*}^*)$ with $n^* \ll n$, which may or may not form a subset of the entire collection of observed locations in $\mathcal{S}$. The Gaussian process $w(\boldsymbol{s})$ yields an n^*-vector of realizations over the knots, say $\boldsymbol{w}^* = (w(\boldsymbol{s}_1^*), \ldots, w(\boldsymbol{s}_{n^*}^*))'$, which follow a $N\{\boldsymbol{0}, \sigma^2 \boldsymbol{R}^*(\boldsymbol{\phi})\}$ where $\boldsymbol{R}^*(\boldsymbol{\phi}) = \{\rho(\boldsymbol{s}_i^*, \boldsymbol{s}_j^*; \boldsymbol{\phi})\}_{i,j=1}^{n^*}$ is the corresponding $n^* \times n^*$ dispersion matrix. Spatial interpolation (or "kriging") at a generic site $\boldsymbol{s}$ is executed through

$$\tilde{w}(\boldsymbol{s}) = E\{w(\boldsymbol{s}) \mid \boldsymbol{w}^*\} = \boldsymbol{r}(\boldsymbol{s}; \boldsymbol{\phi})' \boldsymbol{R}^{*-1}(\boldsymbol{\phi}) \boldsymbol{w}^*. \tag{5}$$

This yields a spatial process $\tilde{w}(\boldsymbol{s}) \sim GP\{0, \sigma^2\tilde{\rho}(\cdot)\}$ where $\tilde{\rho}(\boldsymbol{s}, \boldsymbol{s}'; \boldsymbol{\phi}) = \boldsymbol{r}(\boldsymbol{s}; \boldsymbol{\phi})'\boldsymbol{R}^{*-1}(\boldsymbol{\phi})\boldsymbol{r}(\boldsymbol{s}', \boldsymbol{\phi})$ and $\boldsymbol{r}(\boldsymbol{s}; \boldsymbol{\phi})$ is the $n^* \times 1$ vector whose j-th element is given by $\rho(\boldsymbol{s}, \boldsymbol{s}_j^*; \boldsymbol{\phi})$. We refer to $\tilde{w}(\boldsymbol{s})$ as the *predictive process* derived from the *parent process* $w(\boldsymbol{s})$. The predictive process is a spatially adaptive linear transformation of the realizations of $w(\boldsymbol{s})$ over $\mathcal{S}^*$ with $\boldsymbol{r}(\boldsymbol{s}; \boldsymbol{\phi})'\boldsymbol{R}^{*-1}(\boldsymbol{\phi})$ comprising the coefficients of the transformation. This also implies that $\tilde{w}(\boldsymbol{s})$ is non-stationary, even though $w(\boldsymbol{s})$ is not, allowing the model to adapt better to fit the data.

Replacing $w(\boldsymbol{s})$ in (1) with $\tilde{w}(\boldsymbol{s})$, we obtain the predictive process model,

$$Y(\boldsymbol{s}) = \boldsymbol{x}(\boldsymbol{s})'\boldsymbol{\beta} + \tilde{w}(\boldsymbol{s}) + \epsilon(\boldsymbol{s}). \tag{6}$$

Using (6) as the likelihood, we obtain the predictive process counterpart of (2) as

$$\begin{aligned}\pi(\boldsymbol{\phi}) \times IG(\tau^2 \,|\, a_\tau, b_\tau) \times IG(\sigma^2 \,|\, a_\sigma, b_\sigma) \times N(\boldsymbol{\beta} \,|\, \boldsymbol{\mu}_\beta, \Sigma_\beta) \\ N(\boldsymbol{w}^* \,|\, \boldsymbol{0}, \sigma^2\boldsymbol{R}^*(\boldsymbol{\phi})) \times \prod_{i=1}^{n} N(Y(\boldsymbol{s}_i) \,|\, \boldsymbol{x}(\boldsymbol{s}_i)'\boldsymbol{\beta} + \tilde{w}(\boldsymbol{s}_i), \tau^2).\end{aligned} \tag{7}$$

Dimension reduction occurs since the computations now involve evaluating the $n^* \times n^*$ matrix $\boldsymbol{R}^{*-1}(\boldsymbol{\phi})$, where n^* is chosen to be much smaller than n. Unlike other knot-based methods, the predictive process does not introduce any additional parameters nor involves projecting data onto a grid while enjoying attractive theoretical properties that justify its use as a *best approximation* for the parent process. For example, $\tilde{w}(\boldsymbol{s})$ is an orthogonal projection of $w(\boldsymbol{s})$ on an appropriate linear subspace (e.g. Stein, 1999) minimizing $E[\{w(\boldsymbol{s}) - f(\boldsymbol{w}^*)\}^2 \,|\boldsymbol{w}^*]$ over all real-valued functions $f(\boldsymbol{w}^*)$.

Rather than an approximation to the parent process, we consider the predictive process as a dimension-reducing model for large point-referenced datasets. It is crucial, therefore, that its parameters should be interpreted with respect to (6) and not (1). In fact, being smoother than the parent process, the predictive process tends to have lower variance which, in turn, leads to an upward bias in the nugget. The following inequality reflects, more formally, the shrinkage in variability for the predictive process

$$\text{var}\{w(\boldsymbol{s})\} = \text{var}\{\text{E}[w(\boldsymbol{s}) \,|\, \boldsymbol{w}^*]\} + \text{E}\{\text{var}[w(\boldsymbol{s}) \,|\, \boldsymbol{w}^*]\} \geq \text{var}\{\text{E}[w(\boldsymbol{s}) \,|\, \boldsymbol{w}^*]\} = \text{var}\{\tilde{w}(\boldsymbol{s})\}.$$

The diminished variability in $\tilde{w}(\boldsymbol{s})$ is often manifested by an overestimation of the nugget variance τ^2. Banerjee et al. (2010) explore these biases in greater detail.

Finley et al. (2009) consider modifying the predictive process by adding a heteroscedastic white-noise Gaussian process. More specifically, they propose replacing $\tilde{w}(\boldsymbol{s})$ in (6) with $\tilde{w}_\epsilon(\boldsymbol{s}) = \tilde{w}(\boldsymbol{s}) + \tilde{\epsilon}(\boldsymbol{s})$, where $\tilde{\epsilon}(\boldsymbol{s}) \overset{iid}{\sim} N(0, \sigma^2(1 - \boldsymbol{r}(\boldsymbol{s};\boldsymbol{\phi})'\boldsymbol{R}^{*-1}(\boldsymbol{\phi})\boldsymbol{r}(\boldsymbol{s};\boldsymbol{\phi})))$. Using $\tilde{w}_\epsilon(\boldsymbol{s})$ instead of $\tilde{w}(\boldsymbol{s})$ as the spatial process in (7) yields

$$\pi(\boldsymbol{\phi}) \times IG(\tau^2 \,|\, a_\tau, b_\tau) \times IG(\sigma^2 \,|\, a_\sigma, b_\sigma) \times N(\boldsymbol{\beta} \,|\, \boldsymbol{\mu}_\beta, \Sigma_\beta) \times \\ N(\boldsymbol{w}^* \,|\, \boldsymbol{0}, \sigma^2\boldsymbol{R}^*(\boldsymbol{\phi})) \times N(\tilde{w}_\epsilon \,|\, \boldsymbol{F}(\boldsymbol{\phi})\boldsymbol{w}^*, \sigma^2\boldsymbol{R}_{\tilde{\epsilon}}) \times \prod_{i=1}^{n} N(Y(\boldsymbol{s}_i) \,|\, \boldsymbol{x}(\boldsymbol{s}_i)'\boldsymbol{\beta} + \tilde{w}_\epsilon(\boldsymbol{s}_i), \tau^2), \tag{8}$$

where $\tilde{\boldsymbol{w}}_\epsilon = (\tilde{w}_\epsilon(\boldsymbol{s}_1), \ldots \tilde{w}_\epsilon(\boldsymbol{s}_n))'$, $\boldsymbol{F}(\boldsymbol{\phi}) = \mathcal{R}(\boldsymbol{\phi})'\boldsymbol{R}^{*-1}(\boldsymbol{\phi})$, where $\mathcal{R}(\boldsymbol{\phi})'$ is the $n \times n^*$ matrix whose i-th row is given by $\boldsymbol{r}(\boldsymbol{s}_i;\boldsymbol{\phi})'$, for $i = 1, \ldots, n$, and $\boldsymbol{R}_{\tilde{\epsilon}}$ is an $n \times n$ diagonal matrix with i-th diagonal element $\left\{1 - \boldsymbol{r}(\boldsymbol{s}_i;\boldsymbol{\phi})'\boldsymbol{R}^{*-1}(\boldsymbol{\phi})\boldsymbol{r}(\boldsymbol{s}_i;\boldsymbol{\phi})\right\}$. Now let $\boldsymbol{v}^* = (\boldsymbol{w}^{*'}, \boldsymbol{\beta}', \tilde{\epsilon}')'$ be the $(n^* + p + n) \times 1$ vector collecting all a priori latent Gaussian effects $\boldsymbol{v}^*$. An expression for the marginalized posterior distribution, the bias-adjusted predictive process counterpart to (3), can be obtained by integrating out $\boldsymbol{v}^*$, whereupon we have

$$\pi(\boldsymbol{\phi}) \times IG(\tau^2 \,|\, a_\tau, b_\tau) \times IG(\sigma^2 \,|\, a_\sigma, b_\sigma) \times \\ N(\boldsymbol{Y} \,|\, \boldsymbol{X}\boldsymbol{\mu}_\beta, \boldsymbol{X}\Sigma_\beta\boldsymbol{X}' + \sigma^2\mathcal{R}(\boldsymbol{\phi})\boldsymbol{R}^{*-1}(\boldsymbol{\phi})\mathcal{R}(\boldsymbol{\phi})' + \sigma^2\boldsymbol{R}_{\tilde{\epsilon}} + \tau^2\boldsymbol{I}_n). \tag{9}$$

2.2 Predictive process models with non-Gaussian likelihoods

We now consider the setting with non-Gaussian likelihoods. There are two typical non-Gaussian GLMM first stage settings: (i) binary response at locations modelled using logit or probit regression and (ii) count data at locations modeled using Poisson regression. Diggle et al. (1998) unify the use of these GLMMs in spatial data contexts. See also Lin et al. (2000), Kammann and Wand (2003) and Banerjee et al. (2004). Essentially, we construct the likelihood assuming conditional independence of the outcomes, i.e. the $Y(\boldsymbol{s}_i)$'s, which arise from an exponential family. In other words, we replace (1) with the assumption that the expected value is linear on a transformed scale,

i.e., $\eta(\boldsymbol{s}) \equiv g(E(Y(\boldsymbol{s}))) = \boldsymbol{x}(\boldsymbol{s})'\boldsymbol{\beta} + w(\boldsymbol{s})$, where $g(\cdot)$ is a suitable link function. More specifically, the resulting posterior would take a form analogous to (8):

$$\pi(\boldsymbol{\phi}) \times IG(\sigma^2 \,|\, a_\sigma, b_\sigma) \times N(\boldsymbol{\beta} \,|\, \boldsymbol{\mu}_\beta, \Sigma_\beta) \\ \times N(\boldsymbol{w}^* \,|\, \boldsymbol{0}, \sigma^2 \boldsymbol{R}^*(\phi)) \times N(\tilde{\boldsymbol{w}}_\epsilon \,|\, \boldsymbol{F}(\boldsymbol{\phi})\boldsymbol{w}^*, \sigma^2 \boldsymbol{R}_{\tilde{\epsilon}}) \times \prod_{i=1}^{n} \pi(Y(\boldsymbol{s}_i) \,|\, \eta(\boldsymbol{s}_i)), \tag{10}$$

where $\pi(Y(\boldsymbol{s}_i) \,|\, \eta(\boldsymbol{s}_i))$ belongs to the exponential family of densities. For large datasets, we insert the predictive process, $\tilde{w}(\boldsymbol{s})$, in the link function so that $\eta(\boldsymbol{s}_i) = \boldsymbol{x}(\boldsymbol{s}_i)'\boldsymbol{\beta} + \tilde{w}(\boldsymbol{s}_i)$. The bias-adjusted version would replace $\tilde{w}(\boldsymbol{s})$ with $\tilde{w}_\epsilon(\boldsymbol{s})$. Unlike with Gaussian likelihoods, analytical marginalization over the spatial and regression effects (as in (3) and (9)) is no longer possible.

Let again $\boldsymbol{v}^* = (\boldsymbol{w}^{*\prime}, \boldsymbol{\beta}', \tilde{\epsilon}')'$ be the $(n^* + p + n) \times 1$ vector, comprising the realizations of the spatial predictive process, the regression parameters and the realizations of the bias-adjustment process. The posterior $\pi(\boldsymbol{v}^*, \sigma, \boldsymbol{\phi} \,|\, \boldsymbol{Y})$ corresponding to the bias-adjusted predictive process is proportional to

$$\pi(\boldsymbol{\phi}) \times IG(\sigma^2 \,|\, a_\sigma, b_\sigma) \times N(\boldsymbol{v}^* \,|\, \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*) \\ \times \prod_{i=1}^{n} \pi(Y(\boldsymbol{s}_i) \,|\, \boldsymbol{r}(\boldsymbol{s}_i; \boldsymbol{\phi})'\boldsymbol{R}^{*-1}(\boldsymbol{\phi})\boldsymbol{w}^* + \boldsymbol{x}(\boldsymbol{s}_i)'\boldsymbol{\beta} + \tilde{\epsilon}(\boldsymbol{s}_i)), \tag{11}$$

where mean vector $\boldsymbol{\mu}^* = (\boldsymbol{0}_{n^*}, \boldsymbol{\mu}_\beta, \boldsymbol{0}_n)$, and the $(n^* + p + n) \times (n^* + p + n)$ covariance matrix

$$\boldsymbol{\Sigma}^* = \begin{bmatrix} \sigma^2 \boldsymbol{R}^*(\boldsymbol{\phi}) & \boldsymbol{0}_{n^* \times p} & \boldsymbol{0}_{n^* \times n} \\ \boldsymbol{0}_{p \times n^*} & \boldsymbol{\Sigma}_\beta & \boldsymbol{0}_{p \times n} \\ \boldsymbol{0}_{n \times n^*} & \boldsymbol{0}_{n \times p} & \sigma^2 \boldsymbol{R}_{\tilde{\epsilon}} \end{bmatrix}. \tag{12}$$

The n last diagonal entries in the covariance matrix in (12) facilitate fast evaluation routines. The canonical length n parameter vector in the GLM likelihood model can be defined by $\boldsymbol{\eta} = \boldsymbol{H}^\star \boldsymbol{v}$, where $\boldsymbol{H}^\star = [\boldsymbol{F}(\boldsymbol{\phi}), \boldsymbol{X}, \boldsymbol{I}_n]$.

General settings can be treated using these ideas with the appropriate choice of an exponential family member and a link function. For instance, with binomial data, $\pi(Y(\boldsymbol{s}_i) \,|\, \eta(\boldsymbol{s}_i)) \sim$ *Binomial*$(N(\boldsymbol{s}_i), p(\eta(\boldsymbol{s}_i)))$, where $p(\eta(\boldsymbol{s}_i))$ is the success probability at $\boldsymbol{s}_i$, defined by a link function, and where $N(\boldsymbol{s}_i)$ represents the fixed number of trials. A logit link function specifies $p(\eta(\boldsymbol{s}_i)) =$

$\exp(\eta(\boldsymbol{s}_i))/(1+\exp(\eta(\boldsymbol{s}_i)))$. In some cases, the exponential family density could also include an unknown vector of nuisance parameters, say $\boldsymbol{\psi}$. In the examples we present here, such nuisance parameters do not arise; if they did, we would simply modify our hierarchical model to accommodate a prior for $\boldsymbol{\psi}$. Moreover, a more general unimodal non-Gaussian likelihood, outside the exponential family class, would also fit into our framework.

3 Approximate Bayesian inference

3.1 The Laplace approximation for predictive process model

MCMC algorithms are the current standard for inference in hierarchical Bayesian models. The generality of MCMC allows fitting very flexible models, using full conditional Gibbs sampling schemes in conjunction with Metropolis updates when full conditionals are not directly available. One challenge with MCMC is the slow mixing that can occur, such that subsequent samples in the Markov chain are very dependent, and a huge number of MCMC iterations are required to explore the sampling space and to reduce the Monte Carlo error bounds sufficiently.

The Laplace approximation (Tierney and Kadane, 1986) was constructed for deterministic Bayesian inference, not for sampling based inference. The approach presented in Rue et al. (2009) is in the same spirit. Rather than sampling, analytical Gaussian approximations and numerical routines are applied. The Gaussian approximation is used for the latent effects, which are a priori Gaussian. In our notation from the previous section $\pi(\boldsymbol{v}^* \mid \boldsymbol{\theta}) = N(\boldsymbol{v}^* \mid \boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$, where $\boldsymbol{v}^* = (\boldsymbol{w}^{*\prime}, \boldsymbol{\beta}', \tilde{\epsilon}')'$, including the bias correction term. We now let $\boldsymbol{\theta}$ denote the covariance parameters. For the Gaussian predictive process model this would be $\boldsymbol{\theta} = (\sigma^2, \boldsymbol{\phi}, \tau^2)$, while $\boldsymbol{\theta} = (\sigma^2, \boldsymbol{\phi})$ for the predictive process GLMM formulation. The LA approach exploits a recombination of the marginals and conditionals so that

$$\begin{aligned} \pi(\boldsymbol{\theta}|\boldsymbol{Y}) &= \frac{\pi(\boldsymbol{Y}|\boldsymbol{v}^*,\boldsymbol{\theta})\pi(\boldsymbol{v}^*|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{Y})\pi(\boldsymbol{v}^*|\boldsymbol{Y},\boldsymbol{\theta})}, \\ &\propto \frac{\pi(\boldsymbol{Y}|\boldsymbol{v}^*,\boldsymbol{\theta})\pi(\boldsymbol{v}^*|\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\pi(\boldsymbol{v}^*|\boldsymbol{Y},\boldsymbol{\theta})}, \end{aligned} \tag{13}$$

where the numerator is defined by the model, while the left hand side and the denominator are

needed for posterior inference. The full conditional in the denominator of (13) is Gaussian for a Gaussian likelihood model (Section 2.1). Then, the posterior $\pi(\boldsymbol{\theta}|\boldsymbol{Y})$ in (13) can be evaluated exactly, up to a normalizing constant. For posterior inference about these covariance hyperparameters we turn to numerical methods. Notice that MCMC algorithms can also be constructed for the posterior in (13). In fact, this formula is identical to marginalized model in (9). The LA approach is in this way a marginalization method using the full conditional, rather than integrating out the latent effects, see (4).

When the likelihood model is non-Gaussian, such as in the GLMM, the full conditional $\pi(\boldsymbol{v}^*|\boldsymbol{Y},\boldsymbol{\theta})$ is no longer analytically available. Nevertheless, for standard GLMs the inference of regression effects is commonly obtained by an iterative scoring algorithm, computing the maximum likelihood estimate, and then assessing the uncertainty from the likelihood second derivatives (Hessian) at the maximum location. The LA method uses similar ideas for Bayesian spatial inference:

$$\hat{\pi}(\boldsymbol{\theta}\,|\,\boldsymbol{Y}) \propto \left.\frac{\pi(\boldsymbol{Y}\,|\,\boldsymbol{v}^*,\boldsymbol{\theta})\pi(\boldsymbol{v}^*\,|\,\boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\hat{\pi}(\boldsymbol{v}^*\,|\,\boldsymbol{Y},\boldsymbol{\theta})}\right|_{\hat{\boldsymbol{v}}^*} \tag{14}$$

where $\hat{\pi}(\boldsymbol{v}^*|\boldsymbol{Y},\boldsymbol{\theta})$ is a Gaussian approximation of

$$\pi(\boldsymbol{v}^*|\boldsymbol{Y},\boldsymbol{\theta}) \propto \pi(\boldsymbol{Y}|\boldsymbol{v}^*,\boldsymbol{\theta})\pi(\boldsymbol{v}^*|\boldsymbol{\theta}), \tag{15}$$

constructed to match the mode $\hat{\boldsymbol{v}}^*$ and the curvature at the mode of this full conditional expression. The LA gives a relative error in (14), see Tierney and Kadane (1986). The Monte Carlo error is additive, possibly giving a larger relative error in the tails.

The posterior approximation $\hat{\pi}(\boldsymbol{\theta}|\boldsymbol{Y})$ is explored by numerical routines. For models of reasonable complexity, the dimension of $\boldsymbol{\theta}$ is small (in our examples 2 or 3), and numerical routines can efficiently find the mode, assess uncertainty bounds, and so on. Returning solely the mode is identical to the empirical Bayes estimate. We construct $\hat{\pi}(\boldsymbol{\theta}|\boldsymbol{Y})$ by a deterministic scheme returning a discretized representation of the posterior. Our numerical routine is run for a parameterization with log precision parameters and logistic range. One reason for this parameterization is variance stabilization, another reason is that the surface of the approximate posterior marginal appears close to Gaussian. The posterior percentiles for variance or range parameters can be derived by a direct tranformation. The implementation is similar to Rue et al. (2009) and goes as follows:

1. Choose a starting value $\boldsymbol{\theta}$. Evaluate $\ln \hat{\pi}(\boldsymbol{\theta}|\boldsymbol{Y})$ upto a constant.

2. Perform an optimization scheme to find the mode of $\ln \hat{\pi}(\boldsymbol{\theta}|\boldsymbol{Y})$.

3. Compute the Hessian of $\ln \hat{\pi}(\boldsymbol{\theta}|\boldsymbol{Y})$ at the mode.

4. Step along the main directions away from the mode until $\ln \hat{\pi}(\boldsymbol{\theta}|\boldsymbol{Y})$ is negligible.

5. Fill in a grid (or design) of $\boldsymbol{\theta}$ values within the defined region from stepping-out.

6. Evaluate and normalize $\ln \hat{\pi}(\boldsymbol{\theta}|\boldsymbol{Y})$ on the set of nodes.

In Figure 1 we show the approximate posterior marginals for log precision and logistic range parameters (integrated over the log nugget precision in this Gaussian case). The contours are constructed by rough interpolation over the evaluation points marked as dots. Notice that the contours appear in an almost Gaussian / quadratic form. Altogether, the numerical optimization (a simplex algorithm in this case), Hessian computation, stepping-out and filling-in procedures required about N_{la} =200 evaluations of the posterior. The number of evaluation points would depend on the specific goals of an application. For instance, a standard central composite design approach in the three parameter space uses only 14 evaluation points after the optimization and Hessian computation, at the cost of a coarser approximation. For comparison we display the first 200 samples of a random-walk MCMC sampler (dashed line in Figure 1) with acceptance probability of about 0.3. The random walk pattern is very different from the regular pattern of the numerical scheme. The MCMC algorithm does not span the probability space very well in the 200 iterations shown here.

We next outline the construction of the full conditional required for the Laplace approximation under the bias-corrected predictive process model. Every evaluation of $\hat{\pi}(\boldsymbol{\theta}|\boldsymbol{Y})$ entails computing this full conditional. The GLM likelihood is $\prod_i \pi(Y(\boldsymbol{s}_i)|\eta(\boldsymbol{s}_i))$, where the GLM parameter $\boldsymbol{\eta} = \boldsymbol{H}^\star \boldsymbol{v}^\star = [\boldsymbol{F}(\boldsymbol{\phi}), \boldsymbol{X}, \boldsymbol{I}_n]\boldsymbol{v}^*$. We have prior $\boldsymbol{v}^* \sim N(\boldsymbol{\mu}^*, \boldsymbol{\Sigma}^*)$, see (11) and (12).

Under Gaussian likelihood assumptions $\pi(\boldsymbol{Y}|\boldsymbol{\eta}) = N(\boldsymbol{H}^\star \boldsymbol{v}^\star, \tau^2 \boldsymbol{I}_n)$ the full conditional for $\boldsymbol{v}^\star$

is

$$\begin{aligned}\pi(\boldsymbol{v}^\star|\boldsymbol{Y},\boldsymbol{\theta}) &\propto N(\boldsymbol{H}^\star\boldsymbol{v}^\star,\boldsymbol{T})N(\boldsymbol{\mu}^*,\boldsymbol{\Sigma}^*) \\ &\propto \exp[-\frac{1}{2}(\boldsymbol{Y}-\boldsymbol{H}^\star\boldsymbol{v}^\star)'\boldsymbol{T}^{-1}(\boldsymbol{Y}-\boldsymbol{H}^\star\boldsymbol{v}^\star)-\frac{1}{2}(\boldsymbol{v}^\star-\boldsymbol{\mu}^*)'\boldsymbol{\Sigma}^{*-1}(\boldsymbol{v}^\star-\boldsymbol{\mu}^*)] \\ &\propto \exp[-\frac{1}{2}\boldsymbol{v}^{\star\prime}\boldsymbol{Q}\boldsymbol{v}^\star+\boldsymbol{v}^{\star\prime}\boldsymbol{b}],\end{aligned} \tag{16}$$

where $\boldsymbol{T}=\tau^2\boldsymbol{I}_n$, the full conditional precision matrix $\boldsymbol{Q}=\boldsymbol{H}^{\star\prime}\boldsymbol{T}^{-1}\boldsymbol{H}^\star+\boldsymbol{\Sigma}^{*-1}$, and the canonical parameter $\boldsymbol{b}=\boldsymbol{H}^{\star\prime}\boldsymbol{T}^{-1}\boldsymbol{Y}+\boldsymbol{\Sigma}^{*-1}\boldsymbol{\mu}$. Thus, the full conditional is $\pi(\boldsymbol{v}^\star|\boldsymbol{Y},\boldsymbol{\theta})\sim N(\boldsymbol{Q}^{-1}\boldsymbol{b},\boldsymbol{Q}^{-1})$.

We can compute the required inverse and determinant of the size $(n^*+p+n)\times(n^*+p+n)$ matrix $\boldsymbol{Q}$ by utilizing the structure of $\boldsymbol{\Sigma}^*$ and $\boldsymbol{H}^\star$, and that $\boldsymbol{T}=\tau^2\boldsymbol{I}_n$ is diagonal. Note that the precision is given by

$$\begin{aligned}\boldsymbol{Q} &= \tau^{-2}\boldsymbol{H}^{\star\prime}\boldsymbol{H}^\star+\boldsymbol{\Sigma}^{*-1}=\begin{bmatrix}\boldsymbol{Q}_0 & \boldsymbol{q}\\ \boldsymbol{q}' & \boldsymbol{Q}_1\end{bmatrix}, \\ \boldsymbol{Q}_0 &= \begin{bmatrix}\tau^{-2}\boldsymbol{F}(\boldsymbol{\phi})'\boldsymbol{F}(\boldsymbol{\phi})+\sigma^{-2}\boldsymbol{R}^{*-1}(\boldsymbol{\phi}) & \tau^{-2}\boldsymbol{F}'(\boldsymbol{\phi})\boldsymbol{X}\\ \tau^{-2}\boldsymbol{X}'\boldsymbol{F}(\boldsymbol{\phi}) & \tau^{-2}\boldsymbol{X}'\boldsymbol{X}+\boldsymbol{\Sigma}_\beta^{-1}\end{bmatrix},\end{aligned} \tag{17}$$

where $\boldsymbol{q}'=\tau^{-2}(\boldsymbol{F}(\boldsymbol{\phi}),\boldsymbol{X})$, while $\boldsymbol{Q}_1=\tau^{-2}\boldsymbol{I}_n+\sigma^{-2}\boldsymbol{R}_{\tilde{\epsilon}}^{-1}$ is a size $n\times n$ diagonal matrix. When $n\gg(n^*+p)$, the matrix determinant and inverse are computed efficiently by

$$|\boldsymbol{Q}|=\begin{vmatrix}\boldsymbol{Q}_0 & \boldsymbol{q}\\ \boldsymbol{q}' & \boldsymbol{Q}_1\end{vmatrix}=|\boldsymbol{Q}_1||\boldsymbol{Q}_2|,\quad \boldsymbol{Q}_2=\boldsymbol{Q}_0-\boldsymbol{q}\boldsymbol{Q}_1^{-1}\boldsymbol{q}', \tag{18}$$

$$\boldsymbol{Q}^{-1}=\begin{bmatrix}\boldsymbol{Q}_0 & \boldsymbol{q}\\ \boldsymbol{q}' & \boldsymbol{Q}_1\end{bmatrix}^{-1}=\begin{bmatrix}\boldsymbol{Q}_2^{-1} & -\boldsymbol{Q}_1^{-1}\boldsymbol{q}'\boldsymbol{Q}_2^{-1}\\ -(\boldsymbol{Q}_1^{-1}\boldsymbol{q}'\boldsymbol{Q}_2^{-1})' & \boldsymbol{Q}_1^{-1}+\boldsymbol{Q}_1^{-1}\boldsymbol{q}'\boldsymbol{Q}_2^{-1}\boldsymbol{q}\boldsymbol{Q}_1^{-1}\end{bmatrix}, \tag{19}$$

exploiting that $\boldsymbol{Q}_1$ is diagonal. The cost of matrix inversion for the predictive process model is thus $O(n^{*3})$, since $\boldsymbol{Q}_2$ is $n^*\times n^*$, assuming $n^\star\gg p$. The main cost of inference is building $\boldsymbol{F}(\boldsymbol{\phi})'\boldsymbol{F}(\boldsymbol{\phi})$ which requires $O(nn^{*2})$.

When the likelihood model is non-Gaussian, we expand the likelihood in a quadratic form. For instance, with binomial data $\pi(Y(\boldsymbol{s}_i)|\eta(\boldsymbol{s}_i))\propto\exp(Y(\boldsymbol{s}_i)\eta(\boldsymbol{s}_i)-N(\boldsymbol{s}_i)\log(1+\exp(\eta(\boldsymbol{s}_i))))$, where $\tau(\boldsymbol{s}_i)$ is the fixed number of trials, we Taylor expand $N(\boldsymbol{s}_i)\log(1+\exp(\eta(\boldsymbol{s}_i)))$ to second

order. By expressing the result in a quadratic form of $\boldsymbol{v}^*$ we obtain

$$\log(\pi(\boldsymbol{Y}|\boldsymbol{\eta}) = -\frac{1}{2}\boldsymbol{v}^{\star\prime}\boldsymbol{T}_{\text{lin}}^{-1}\boldsymbol{v}^{\star} + \boldsymbol{v}^{\star\prime}\boldsymbol{c}_{\text{lin}} + const \tag{20}$$

where *const* does not depend on $\boldsymbol{v}^{\star}$, and with

$$\boldsymbol{T}_{\text{lin}}^{-1} = \boldsymbol{H}^{\star\prime}\boldsymbol{D}_2\boldsymbol{H}^{\star}, \quad \boldsymbol{c}_{\text{lin}} = \boldsymbol{H}^{\star\prime}\boldsymbol{D}_2(\boldsymbol{Y} - \boldsymbol{d}_1 + \boldsymbol{D}_2\boldsymbol{H}^{\star}\hat{\boldsymbol{v}}^{\star}). \tag{21}$$

These derivative expressions are defined using componentwise multiplication and division to get

$$\begin{aligned} \boldsymbol{d}_1 &= \{\boldsymbol{N} \odot \exp(\boldsymbol{H}^*\boldsymbol{v}^*)\} \oslash \{\boldsymbol{1}_n + \exp(\boldsymbol{H}^*\boldsymbol{v}^*)\}, \\ \boldsymbol{D}_2 &= \text{diag}\left(\{\boldsymbol{N} \odot \exp(\boldsymbol{H}^*\boldsymbol{v}^*)\} \oslash \{(\boldsymbol{1}_n + \exp(\boldsymbol{H}^*\boldsymbol{v}^*))^{\circledast 2}\}\right), \end{aligned} \tag{22}$$

where $\boldsymbol{1}_n$ is a $n \times 1$ vector of ones, $\boldsymbol{N} = ((N(\boldsymbol{s}_1), \ldots, N(\boldsymbol{s}_n))'$, and $\exp(\cdot)$ also works componentwise. Given the quadratic expansion the approximate full conditional is similar to (16). Five iterations are usually enough to detect the mode of the full conditional. At each iteration, the linearization point is recomputed as the mode from the previous step. Thus, this resembles a usual GLM optimization method, except that the model has a spatial predictive process representation.

3.2 The Integrated nested Laplace approximation for predictive process model

The posterior marginals for regression effects or spatial effects can be computed from the Gaussian approximation of the full conditional by numerical integration over the covariance parameters. For any effect v_j^* we have

$$\hat{\pi}(v_j^*|\boldsymbol{Y}) = \int \hat{\pi}(v_j^*|\boldsymbol{Y}, \boldsymbol{\theta})\hat{\pi}(\boldsymbol{\theta}|\boldsymbol{Y})d\boldsymbol{\theta}, \tag{23}$$

where $\hat{\pi}(v_j^*|\boldsymbol{Y}, \boldsymbol{\theta})$ is an element j of the joint approximate Gaussian. The integral is solved by numerical integration over the evaluation points for $\hat{\pi}(\boldsymbol{\theta}|\boldsymbol{Y})$. Numerical integration is usually superior to Monte Carlo integration in small dimensions like we have here. The full conditional for latent Gaussian variables is computed for every evaluation point of $\hat{\pi}(\boldsymbol{\theta}|\boldsymbol{Y})$. Thus, all entries in the integrand of (23) are readily available. This same numerical integration formula can be applied to spatial effects or a linear combination of regression parameters and spatial effects, which

also has an approximate Gaussian full conditional. Recall that the numerical integration uses a parameterization with the log precision parameters and logistic range parameter. In our experience, this parameterization means that fewer numerical evaluation points are required to estimate the integral. An empirical Bayes approach would use only one evaluation point at the posterior mode for $\boldsymbol{\theta}$.

These direct LA marginals in (23) can be improved by applying the integrated nested Laplace routine, see Rue et al (2009) and Eidsvik et al. (2009). This INLA approach allows one particular regression effect β_j, or one spatial effect to be non-Gaussian, while all remaining latent effects remain Gaussian. The INLA for posterior marginals of a latent effect v_j^* is based on

$$\pi(v_j^*|\boldsymbol{Y},\boldsymbol{\theta}) \propto \frac{\pi(\boldsymbol{Y}|\boldsymbol{v}^*,\boldsymbol{\theta})\pi(\boldsymbol{v}^*|\boldsymbol{\theta})}{\pi(\boldsymbol{v}^*_{-j}|v_j^*,\boldsymbol{Y},\boldsymbol{\theta})}, \tag{24}$$

where the latent effect v_j^* can again be a regression effect, spatial effect, or a linear combination. We will denote the INLA by $\tilde{\pi}(v_j^*|\boldsymbol{Y},\boldsymbol{\theta})$. Its computation uses (24) with a Gaussian approximation $\hat{\pi}(\boldsymbol{v}^*_{-j}|v_j^*,\boldsymbol{Y},\boldsymbol{\theta})$ in the denominator, treating v_j^* as fixed (measured). Thus, the improved approximate marginal becomes

$$\tilde{\pi}(v_j^* \mid \boldsymbol{Y},\boldsymbol{\theta}) \propto \left.\frac{\pi(\boldsymbol{Y}|\boldsymbol{v}^*,\boldsymbol{\theta})\pi(\boldsymbol{v}^* \mid \boldsymbol{\theta})}{\hat{\pi}(\boldsymbol{v}^*_{-j} \mid v_j^*,\boldsymbol{Y},\boldsymbol{\theta})}\right|_{\hat{\boldsymbol{v}}^*_{-j}}, \tag{25}$$

where the expression is evaluated at the full conditional mode, keeping v_j^* fixed. Thus, INLA uses a second round of the LA to cancel out the remaining approximate Gaussian variables $\boldsymbol{v}^*_{-j}$, and in this way provides a better approximation for the posterior marginals for spatial effects and regression effects (Rue et al., 2009). The improved approximation $\tilde{\pi}(v_j^*|\boldsymbol{Y},\boldsymbol{\theta})$ can be computed on a grid of v_j^* values, or fitted a parametric density. For instance, a Gaussian approximation requires only three evaluation points v_j^* to assign a mean and covariance, and normalize.

The posterior marginal INLA is obtained by

$$\tilde{\pi}(v_j^*|\boldsymbol{Y}) = \int \tilde{\pi}(v_j^*|\boldsymbol{Y},\boldsymbol{\theta})\hat{\pi}(\boldsymbol{\theta}|\boldsymbol{Y})d\boldsymbol{\theta}, \tag{26}$$

which is solved by numerical integration over the same evaluation points of the covariance parameters. In our experience, INLA provides a shift of LA for regression parameters, but very little for

the spatial effects. Intuitively, the non-Gaussian data has greater effect on the regression parameters, which are valid for the entire spatial domain. By assuming a Gaussian full conditional (LA) for these parameters, we could induce some bias. The spatial effects are local variables and learn effectively from only parts of the data. Then the Gaussian prior model is more dominating, and the LA is more accurate.

4 Analysis and Results

Four data sets are used to explore the candidate models' ability to estimate parameters of interest and predict at new locations. The first two are synthetic. The third has been used to understand the distribution of lake acidity and subsequent decline in trout abundance in Norway. These first data sets are moderate in size, and a full MCMC based data analysis is possible to perform. This allows comparison between MCMC and INLA, and between different predictive process models. The fourth is a large forest inventory data set used to produce estimates of forest landuse across northcentral United States. For this data set, full MCMC based inference is too computationally challenging for a modern desktop workstation. The following subsections describe these data sets and accompanying modeling details.

The LA and INLA based analyses were conducted in Matlab version 7.9. The `R` package *spBayes* was used for MCMC based analyses. In this package the higher level `R` code calls C++ and Fortran that subsequently calls BLAS (`www.netlib.org/blas`) and LAPACK (`www.netlib.org/lapack`) routines for efficient matrix computations. All analyses were conducted on a Linux workstation using two Intel Nehalem-based quad-Xeon processors. The Matlab, BLAS, and LAPACK routines were threaded and therefore leveraged multiple CPUs for matrix operations. Specifically, *spBayes* was compiled to call Intel's Math Kernel Library version 10.2 BLAS and LAPACK implementations.

4.1 Synthetic data

Data sets of Gaussian and binomial outcome variables were generated. These synthetic data are composed of 750 locations selected randomly from a $[1, 100] \times [1, 100]$ square. The eight

candidate models include a full geostatistical and three predictive process specifications for both MCMC and LA methods. The predictive process models are based on 64, 100, and 256 regular grid knot intensities covering this square.

Both Gaussian and binomial data were generated using a 750×3 covariates matrix $\boldsymbol{X}$, where the first column is the intercept and the values in the subsequent columns were randomly generated from a $N(0, 1)$. The regression coefficients were set to $\boldsymbol{\beta} = (0.1, 0.5, 1)'$. An exponential spatial correlation function $C(w(\boldsymbol{s}), w(\boldsymbol{s} + \boldsymbol{h})) = \sigma^2 \exp(-\phi|\boldsymbol{h}|)$ was used with variance, σ^2=5, and spatial correlation parameter, ϕ=0.06, which corresponds to an effective spatial range of $\sim$50 units. Here, effective spatial range is defined as the distance (in map units) at which the spatial correlation drops to 0.05. For the continuous outcome data the nugget variance, τ^2, was set to 1. A subset of 250 observations were selected randomly to serve as a hold-out set to assess predictive performance, while parameter estimates were based on the remaining 500 observations. All candidate models were fit using the same independent prior specification with each β following a $N(0, 10000)$, an $IG(2, 1)$ for the variance parameters, σ^2 and τ^2, and a broad uniform support for the spatial correlation parameter $\phi \sim U(0.03, 3)$.

Inference results of the Gaussian response model are given in Table 1. Here, 'full' refers to the MCMC or INLA method that uses all the data. When the data are Gaussian, the LA is exact, and for fixed knot configurations the small differences between MCMC and LA inference in Table 1 are caused by Monte Carlo and numerical approximation errors. These differences are sometimes visible, especially for the 2.5 and 97.5 percentiles. Considering the predictive process models, the estimates for regression parameters are captured very accurately, even with 64 knots. The distributions for covariance parameters are also quite close to the results obtained using the full data set, but the correlation range appears a little too narrow for a small number of knots, the nugget variance is slightly underestimated, while the variance in the latent process is a little overestimated. When the knot size increases to 256, the predictive process results get closer to that of full data.

The mean square prediction error (MSPE) in Table 1 is computed and summed over the hold-out dataset. The MSPE values for MCMC and Laplace show similar increase for predictive models

with few knots. This increase (about $4.5/3.5 = 30\%$ for 64 knots) is caused by the data reduction idea that plays an intrinsic role in the predictive process formulation. The main prediction differences between full size n model and predictive process models occur at hold-out sites that are close to data locations, but far from knots. More creative design of knot locations could reduce the MSPE in this case (see e.g. Finley et al., 2009).

The last row in Table 1 shows the number of operations required to deliver the inference and prediction results. The number of operations is the product of the number of evaluations and the cost of every evaluation. For the full model the main evaluation cost is matrix inversion at $O(n^3)$. For the predictive process model the main cost is $O(nn^{*2})$, which is the cost of building the required size $n^* \times n^*$ matrix. The MCMC inference was based on three MCMC chains, with unique starting values, running for N_{mcmc} =10,000 iterations. The CODA package in `R` (`www.r-project.org`) was used to diagnose convergence by monitoring mixing with the Gelman-Rubin diagnostics and autocorrelations (see, e.g., Gelman et al., 2004, Section 11.6). Acceptable convergence was diagnosed within 5,000 iterations (which were discarded as burn-in). Therefore, the parameter estimates and posterior predictive inference offered in Table 1 are based on 15,000 post burn-in samples.

For the LA approach we count the number of evaluations needed to reach a certain tolerance on the numerical approximation. The tolerance is tuned in the optimization determining the poserior mode, the step size moving out from the mode, and by the number of grid / design points used to compute the density approximation. Altogether we use about $N_{la} = 200$ posterior evaluations for the LA approach. Putting these numbers into Table 1 we see a clear reduction in the number of operations when using the predictive process models and INLA. For instance, with 64 predictive process knots, we use 60 times less operations than with the full data. Similarly, the LA approach means a factor 50 reduction in computation time, for any of the predictive process or full data models. By merging the two ideas we achieve sufficiently accurate results in moderate time. Note the operations counts are simply based on the order of the main computations. In practice the constant in front of the order will vary, and the other computations as well, depending on predictive process or full data approach, and other computer related aspects. Our counts should thus not be

taken in earnest, but regarded more as a guideline.

In Figure 2 (left column) we visualize the predictive process results with LA for our parameterization with log precision (top), logistic range parameter (center), and the log nugget precision (bottom). Notice that high log precision means small variance, so the interpretation for these parameters is opposite that of Table 1. The displays are for the three predictive process models with 64, 100, 256 knots (dashed) and the full data (solid). The dashed lines are somewhat biased to the right for precision and to the left for range. The posteriors using predictive process get closer to the full data posterior when the knot size increases. In Figure 2 (right column) we similarly show results for the regression effects. The predictive process models with various knot configurations and the full data provide almost the same posteriors, but with some small visible differences, especially for covariate 1 (Figure 2, right column, center). This might be caused by quite extreme covariates at the edge of the domain, and where the knots are not so dense.

The binomial data are simulated in the same geographic locations as for the Gaussian case. In each location we draw 10 trials with the success probability at that location, using a logit link function. In Table 2 we show results of an MCMC algorithm and the INLA approach for this synthetic data set. With the binomial reponse model differences in Table 2 are caused by the LA and numerical routine errors for the INLA approach, or by Monte Carlo error and convergence and mixing challenges for the MCMC algorithm. Two comparisons can be made here: knots intensity versus full data and MCMC versus INLA. For comparing knot intensities, we see the predictive process models are quite close to the full data results, but there is some overestimation for the scale and underestimation for the range, when using few knots. This is observed both for MCMC and INLA, and was also seen for the Gaussian response model in Table 1. The effect of covariates (β_1 and β_2) are accurately estimated with the predictive process. The β_0 parameter shows the least consistent pattern, and it has a very wide distribution. Comparing MCMC with INLA, we see that most regression parameters and the range are very similar, while the β_0 distribution again varies a lot, and the tails of the σ^2 distribution are a little different. This could be a consequence of using the LA, but could also be caused by the Markov chain staying too long in a tail, or a too rough truncation scheme for the numerical LA approach. Inference for the regression parameters was

done using the INLA approach. In this case evaluating the $\tilde{\pi}(\beta_j|\boldsymbol{Y}, \sigma^2, \phi, \tau)$ at three evaluation points and fitting a Gaussian to this improved marginal. The LA approach gives similar results as INLA, but slightly shifted up or down.

The MSPE values for the spatial effects at the hold-out set show similar tendencies as for the Gaussian data set. The predictive process models with few knots have slightly higher MSPE. The un-marginalized models used to fit the binomial outcome data required more MCMC iterations to begin adequate mixing. The MCMC based inference in Table 2 is based on N_{mcmc} =25,000 iterations. The LA approach is now only over two covariance parameters (ϕ,σ^2, no nugget), and it uses fewer numerical steps than for the Gaussian data. But, on the other hand, it takes about five iterations to compute the Gaussian approximation $\hat{\pi}(\boldsymbol{v}^*|\boldsymbol{Y}, \sigma^2, \phi)$ for the full conditional. Thus, the number of posterior evaluations is still about $N_{la} = 200$. The operations counts in Table 2 show that the INLA solution with 64 predictive process knots uses a factor 8.000 less operations than the MCMC sampling with all the data.

4.2 Lake acidification

We next study a data set originally published by Varin et al. (2005). The focus of their study was to model trout abundance in Norwegian lakes as a function of lake acidity. The data were collected during 1986 from interviews with local fishermen. Here, we use data from the southern part of Norway. The response is 'population status' of trout for each lake $i = 1, \ldots, 361$, coded as unaffected ($Y(\boldsymbol{s}_i) = 0$) or decreased/extinct ($Y(\boldsymbol{s}_i) = 1$). Lakes' northing coordinates and Acid Neutralizing Capacity (ANC) are used as covariates, along with an intercept. ANC is a measure for the overall buffering capacity against acidification for a solution.

As in the synthetic data analysis, the eight candidate models include a full geostatistical and three predictive process specifications for both MCMC and LA methods. The predictive process models are based on 54, 89, and 126 knot intensities. Table 3 shows the inference results for all candidates. For this data set we detect almost no differences between the various predictive process models. Of course there is variability in the regression parameters and range, but considering the wide confidence bounds these differences are very small. The INLA results are similar to the

MCMC, but show slight differences for the distribution of β_1 and σ^2 parameters. For the σ^2 parameter the difference in MCMC and LA seems to be driven by a heavy left tail in the MCMC results. One possible explanation is the MCMC chain stays out in the tail for too long, in the limited time of the Markov chain run. Another explanation is the truncation limits of the numerical LA approach misses this heavy tail. We constructed the INLA approximation by evaluating $\tilde{\pi}(\beta_1|\boldsymbol{Y}, \sigma^2, \phi, \tau)$ at three points and fitting a Gaussian, and thus the marginal is a mixture of Gaussians. This INLA solution is almost indistinguishable from the direct Gaussian LA for the intercept and northing, while it is visibly shifted to the left for the ANC effect(β_1). Still, the INLA using a Gaussian mixture underestimates the tail a little, as extending to a non-Gaussian INLA gives a larger tail, but not quite as large as the MCMC solution. We note the posterior distribution for the range parameter almost hits the boundary for the uniform distribution for ϕ. This indicates there is limited information about the large ranges in the data, and effects of this could possibly cause some heavy tail challenges for MCMC or LA.

Figure 3(top) shows predictions of latent effects $\boldsymbol{x}(s)\boldsymbol{\beta} + \tilde{w}(s)$. This is displayed for the full dataset using MCMC sampling (top, left) and for the 54 knots predictive process model using the LA (top, right). For the prediction results we see little difference between these two model/inference combinations. This emphasizes that small differences in the marginals for covariance parameters or the regression effects (Table 3), translate into miniscule prediction differences in Figure 3. There are some minor changes when going to the predictive process predictions, such as a slightly smoother result for some of the northern datapoints, that the knot configuration do not capture, but this is hard to distinguish in a map like Figure 3 and would hardly have much effect on decision making. Similarly, Figure 3 (bottom) shows the 95% prediction range intervals for full dataset using MCMC sampling (left) and for 54 knots using LA (right). The differences in prediction range are also small, but the full data results have wider ranges at some sites.

The MCMC results takes a few hours to compute, while LA with 54 knots takes a few seconds. If we would like to check sensitivity to the shape of the covariance model, perform cross validation, or other such high-level inference tasks, this difference in computation time becomes important. LA with predictive process makes it possible to compute the results on-line on the laptop computer.

We perform one such high-level task. We use cross-validation to compare the 54 knots predictive process model using LA with a non-spatial model (i.e. using only the regression part for the explanatory variables). This comparison is done over 10 randomized leave-37-out sets. For each of these sets we predict $\hat{Y}(s_{0,q})$ for every hold-out site $q = 1, \ldots, 37$, using the most likely outcome in the predictive distribution as the predictor. We compare the predictions with the observed data values. Table 4 shows the results summarized by a 2 by 2 table of the total number of classifications $(\hat{Y}(s_{0,q}), Y(s_{0,q})) \in (0,0)$, $(0,1)$, $(1,0)$, and $(1,1)$, where a good model gets mostly the diagonal entries $(0,0)$ and $(1,1)$. The non-spatial model has large diagonal elements, and the explanatory variables catch much of the structure in the data. Even larger diagonal elements are achieved for the spatial model, meaning the spatial residual process adds explanatory effect.

4.3 Forest land use

Here we analyze a large data set where full MCMC based inference, even using the predictive process, is prohibitively expensive to run on a modern desktop workstation. The analysis is motivated by the need for spatially explicit estimates of forest area which are useful for land use change monitoring, carbon budgeting, and ecological and timber supply forecasting. The data consist of n =12,629 Forest Inventory and Analysis (FIA) plots measured in Michigan, USA, between 1999-2006. The FIA program of the USDA Forest Service has established field plot centers in permanent locations using a sampling design that is assumed to produce a systematic equal-probability sample with a random spatial component (Bechtold and Patterson, 2005). Locations of plots are determined using GPS receivers. Plot locations are depicted in Figure 4 (top left). Each plot consists of a 7.31 m radius circular area. For each plot, $i = 1, \ldots, n$, a field crew determined the response variable value as forested ($Y(\boldsymbol{s}_i) = 1$) or non-forested ($Y(\boldsymbol{s}_i) = 0$) given the FIA definition of forest land (Bechtold and Patterson, 2005). Figure 4 (top right) is a surface interpolation of forest occupancy at the plot locations. This figure shows that northern Michigan is dominated by forest while the south is primarily not forested. A Landsat 7 ETM+ satellite image, 30×30 m spatial resolution, taken in mid-summer 2002 was tasseled cap transformed into its brightness, greenness, and wetness components (Kauth and Thomas, 1976) and used as covariates. Finley et

al. (2008) show how these components can help explain variability in the probability of forest occupancy.

Non-spatial and spatial predictive process models, using 50, 100, and 200 knot intensities, were considered in this comparison. Knot locations were chosen using a *k-means* clustering algorithm on the observed locations. Because our primary interest is in predicting forest occupancy, we compare the candidate models' using a set of 1,262 holdout (or validation) plots that were selected at random from the 12,629 FIA plots. Model parameters were estimated using the remaining 11,367 observations. Table 5 summarizes the results of parameter estimation. As noted in previous studies and seen here, the three remotely sensed covariates contribute significantly to explaining variability in the probability of forest occupancy and do not change substantially among the candidate models. The predictive process models produced very similar estimates of the spatial range, ϕ, and variance, σ^2, parameters. The median effective spatial ranges in km are 83, 76, and 73 for the 50, 100, and 200 knot model, respectively. This large spatial range is capturing the broad scale residual dependence within the forested (north) and prairie/agriculture (south) landuse patterns. These results suggest we could potentially use fewer than 50 knots; however, the minimum inter-site distance among the knots of the 50 knot model is 45.6 km. Models using fewer knots, and hence greater distance between knots, could have trouble estimating the spatial range parameter. The bottom two plots in Figure 4 depict surface interpolation of the 50 and 200 knot models' median fitted probability of forest at the observed locations. The two models produce nearly indistinguishable surfaces, both of which capture the patterns in the non-statistical surface generated using the actual observations (top right).

Because our primary interest is in predicting forest occupancy, we compare the candidate models' using the 1,262 holdout sites. Extending the zero-one prediction rule from the Norwegian Lakes dataset in the previous section, we now consider several different scoring rules to evaluate the predictive performance of the candidate models. A scoring rule provides a summary measure for evaluating a probabilistic prediction given the predictive distribution and the observed outcome. In our setting the scoring rule function is $SR(\boldsymbol{\pi}, i)$, where $\boldsymbol{\pi}$ is the vector of probabilities associated with each category (here $\boldsymbol{\pi}$ is of length $J = 2$ i.e., forest and non-forest) and

$i = Y(\boldsymbol{s}_0)$ is the observed condition at a hold-out site. Given all the holdout sites $\{\boldsymbol{s}_{0q}\}_{q=1}^{1262}$ we can calculate summary statistics of the scores, e.g., the mean score is $\widehat{SR} = \sum_{q=1}^{1262} \frac{SR(\boldsymbol{\pi}_q, i_q)}{1262}$, where $\boldsymbol{\pi}_q = \{P(Y(\boldsymbol{s}_{0q}) = 0|\boldsymbol{Y}), P(Y(\boldsymbol{s}_{0q}) = 1|\boldsymbol{Y})\}$. Gneiting and Raftery (2007) offer four scoring rules for prediction of categorical variables,

$$\text{Zero-one: } SR(\boldsymbol{\pi}, i) = \begin{cases} 1 & \text{if } \pi_i = \max\{\pi_1 \dots \pi_J\} \\ 0 & \text{if otherwise} \end{cases}$$

$$\text{Quadratic: } SR(\boldsymbol{\pi}, i) = 2\pi_i - \sum_{j=1}^{J} \pi_j^2 - 1$$

$$\text{Spherical: } SR(\boldsymbol{\pi}, i) = \frac{\pi_i}{(\sum_{j=1}^{J} \pi_j^2)^{\frac{1}{2}}}$$

$$\text{Logarithmic: } SR(\boldsymbol{\pi}, i) = \log \pi_i.$$

Following definitions in Gneiting and Raftery (2007), all the noted scoring rules are strictly *proper* but for the zero-one, which is only proper. The zero-one scoring rule uses only a portion of available information, ignoring variability in the predictive distribution and returning either a zero or one. Similarly, the logarithmic scoring rule considers only one of the probabilities in the predictive distribution. The maximum values (i.e., perfect prediction) of the different scoring rules are 1 for zero-one, 0 for quadratic, 1 for spherical, and 0 for logarithmic. The results are shown in Table 6. Here, for all rules, the spatial models offer improved predictive performance. Further, there seems to be only limited gain in predictions between the 50 and 200 knot model.

5 Conclusions

The main contribution of this paper is combining computational ideas for modeling and inference of spatial data. Together, the predictive process models and approximate Bayesian inference using LA and INLA provide very fast analysis of large spatial data sets.

Predictive process models are efficient dimension reduction techniques that builds directly on the spatial structure of the model. The predictive process models entail a selection of knot locations that covers the domain of interest. The number of knots is much smaller than the number of data sites, and, as a result, matrix computations are feasible. In terms of parameter estimation and

prediction, the predictive process model performs adequately, especially when one uses a bias correction term in the modeling. Testing results from many different predictive process models would be easy using parallel computing.

The LA combined with numerical routines provides a very fast and accurate approximation for the posterior of model parameters and for prediction in spatial models. The numerical approach offers flexibility in low parameter dimensions. In applications with a larger dimension of the covariance parameters one might still do better than empirical Bayes by allowing some evaluation points on a design surface. Moreover, in several application the covariance parameters are treated as a nuisance parameters, while the main interest is in regression or spatial effects. In such contexts a refined posterior representation of the covariance parameters is not required. If the main interest is in some function of the covariance parameters, one could reparameterize the model such that this main functional parameter becomes a key feature in the numerical assessment.

Further work might include predictive process modeling and INLA for multivariate spatial data, space-time applications, or to situations where the regression parameters change in space-time, such as the spatially varying coefficients model. The computational problems are further aggravated in these situations, and combining predictive process models and INLA appears very appealing.

References

Ainsworth, L.M. and Dean, C.B. (2006). Approximate inference for disease mapping. *Computational Statistics and Data Analysis* **50**, 2552-2570.

Banerjee, S., Carlin, B.K. and Gelfand, A.E. (2004). Hierarchical modeling and analysis for spatial data. Chapman & Hall.

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

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

Bechtold, W.A. and Patterson, P.L. (2005). The enhanced forest inventory and analysis program - national sampling design and estimation procedures. *In General Technical report SRS-80*, Asheville, NC, USDA Forest Services, Southern Research Station, **85**.

Breslow, N.E. and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association* **88**, 9-25.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of American Statistical Association* **90**, 1313-1321.

Crainiceaniu, C.M., Diggle, P.J. and Rowlingson, B. (2008). Bivariate binomial spatial modeling of Loa Loa prevalence in Tropical Africa. *Journal of the American Statistical Association* **103**, 21-37.

Cressie, N. and Johannesson, G. (2008). Fixed rank kriging for large spatial datasets. *Journal of the Royal Statistical Society, Series B* **70**, 209-226.

Diggle, P.J., Tawn, J.A. and Moyeed, R.A. (1998). Model-based geostatistics. *Journal of the Royal Statistical Society, Series C* **47**, 299-350.

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

Diggle, P.J. and Ribeiro, P.J. (2007). Model-based geostatistics. Springer.

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

Eidsvik, J., Martino, S. and Rue, H. (2009). Approximate Bayesian inference for spatial generalized linear mixed models. *Scandinavian Journal of Statistics* **36**, 1-22.

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

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

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

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

Gelfand, A.E., Banerjee, S. and Gamerman, D. (2005). Spatial process modeling for univariate and multivariate dynamic spatial data. *Environmetrics* **16**, 465-479.

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

Gelman, A., Carlin, J.B., Stern, H.S. and Rubin, D.B. (2004). Bayesian data analysis. Chapman & Hall.

Gneiting, T. and Raftery, A.E. (2007). Strictly proper scoring rules, prediction and estimation. *Journal of the American Statistical Association* **102**, 359-378.

Higdon, D. (2002). Space and space-time modeling using process convolutions. In *Quantitative methods for current environmental issues*, eds. C. Anderson, V. Barnett, P. C. Chatwin, and A. H. El-Shaarawi, 37–56. Springer-Verlag.

Hsiao, C.K., Huang, S.Y. and Chang, C.W. (2004). Bayesian marginal inference via candidate's formula. *Statistics and Computing* **14**, 59-66.

Kammann, E.E. and Wand, M.P. (2003). Geoadditive models. *Applied Statistics* **52**, 1-18.

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

Kauth, R.J. and Thomas, G.S. (1976). The tasseled cap: A graphic description of the spectral-temporal development of agricultural crops as seen by Landsat. *Proceedings of the symposium on machine processing of remotely sensed data*. West Lafayette, IN: Purdue University: 41-51.

Lewis, S.M. and Raftery, A.E. (1997). Estimating Bayesian factors via posterior simulation with the Laplace–Metropolis estimator. *Journal of the American Statistical Association* **92**, 648-655.

Nychka, D.W. and Saltzman, N. (1998). Design of air quality monitoring networks. *Case Studies in Environmental Statistics*, 51-76, Springer.

Rasmussen, C.E. and Williams, C.K.I. (2006). Gaussian processes for machine learning. Cambridge MA, MIT Press.

Rue, H. and Held, L. (2005). Gaussian Markov random fields, Theory and applications. Chapman & Hall.

Rue, H., Martino, S. and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models using integrated nested Laplace approximations. *Journal of the Royal Statistical Society, Series B* **71**, 319-392.

Schabenberger, O. and Gotway, C.A. (2004). Statistical methods for spatial data analysis. Chapman & Hall.

Stein, M.L., Chi, Z. and Welty, L.J. (2004). Approximating likelihoods for large spatial datasets. *Journal of the Royal Statistical Society, Series B* **66**, 275-296.

Stein, M.L. (2007). Spatial variation of total column ozone on a global scale. *The Annals of Applied Statistics* **1**, 191-200.

Stein, M.L. (2008). A modeling approach for large spatial datasets. *Journal of the Korean Statistical Society* **37**, 3-10.

Tierney, L. and Kadane, J.B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association* **81**, 82-86.

Varin, C., Høst, G. and Skare, Ø. (2005). Pairwise likelihood inference in spatial generalized linear mixed models. *Computational Statistics and Data Analysis* **49**, 1173-1191.

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

Xia, G., Miranda, M. and Gelfand, A.E. (2006). Approximately optimal spatial design approaches for environmental health data. *Environmetrics* **17**, 363-385.

Table 1: Synthetic Gaussian data set: Summary of candidate models' parameter estimates, 50 (2.5 97.5) percentiles, hold-out set mean squared error of prediction (MSEP), and the number of operations.

		MCMC				LA			
			Pred. proc. knots				Pred. proc. knots		
	True	Full	64	100	256	Full	64	100	256
β_0	0.1	-1.11 (-2.48, 0.27)	-1.07 (-2.56, 0.51)	-1.10 (-2.39, 0.28)	-1.12 (-2.38, 0.19)	-1.11 (-2.60, 0.42)	-1.10 (-2.47, 0.37)	-1.11 (-2.53, 0.40)	-1.13 (-2.39, 0.18)
β_1	0.5	0.46 (0.33, 0.58)	0.42 (0.29, 0.56)	0.44 (0.31, 0.57)	0.45 (0.33, 0.59)	0.46 (0.33, 0.58)	0.42 (0.29, 0.56)	0.44 (0.30, 0.57)	0.45 (0.32, 0.58)
β_2	1	0.99 (0.86, 1.11)	0.98 (0.85, 1.11)	0.99 (0.86, 1.12)	0.99 (0.87, 1.11)	0.99 (0.86, 1.11)	0.98 (0.85, 1.12)	0.99 (0.86, 1.12)	0.99 (0.86, 1.11)
ϕ	0.06	0.06 (0.03, 0.10)	0.05 (0.03, 0.08)	0.06 (0.04, 0.10)	0.07 (0.03, 0.11)	0.06 (0.03, 0.10)	0.06 (0.04, 0.08)	0.06 (0.04 0.09)	0.07 (0.04, 0.10)
σ^2	5	4.07 (2.72, 6.81)	4.67 (3.05, 7.89)	4.51 (3.08, 7.16)	4.13 (2.76, 6.78)	3.90 (2.60, 6.58)	4.55 (3.11, 6.96)	4.59 (3.03, 6.86)	4.05 (2.69, 6.30)
τ^2	1	1.25 (0.90, 1.62)	0.74 (0.24, 1.39)	0.68 (0.23, 1.30)	1.02 (0.36, 1.53)	1.23 (0.86, 1.68)	0.64 (0.29, 1.29)	0.67 (0.31, 1.30)	1.03 (0.55, 1.56)
MSPE		3.54	4.55	4.27	3.87	3.35	4.67	4.40	4.17
Operations (in billions)		1250	20	50	330	25	0.4	1.0	6.6

Table 2: Synthetic binomial data set: Summary of candidate models' parameter estimates, 50 (2.5 97.5) percentiles, hold-out set mean squared error of prediction (MSEP), and the number of operations.

		MCMC				INLA			
			Pred. proc. knots				Pred. proc. knots		
	True	Full	64	100	256	Full	64	100	256
β_0	0.1	-0.74 (-2.49, 0.33)	-1.04 (-3.36, 0.34)	-0.99 (-3.06, 0.45)	-0.83 (-2.78, 0.39)	-0.64 (-2.13, 0.40)	-1.14 (-3.52, 0.55)	-1.11 (-3.27, 0.46)	-0.79 (-2.92, 0.77)
β_1	0.5	0.39 (0.27, 0.52)	0.42 (0.27, 0.56)	0.40 (0.26, 0.55)	0.41 (0.27, 0.55)	0.39 (0.25, 0.54)	0.39 (0.25, 0.54)	0.40 (0.26, 0.54)	0.40 (0.27, 0.54)
β_2	1	1.08 (0.95, 1.22)	1.10 (0.95, 1.25)	1.09 (0.94, 1.24)	1.09 (0.95, 1.24)	1.04 (0.87, 1.22)	1.06 (0.87, 1.22)	1.07 (0.92, 1.22)	1.08 (0.93, 1.22)
ϕ	0.05	0.07 (0.04, 0.11)	0.04 (0.03, 0.07)	0.05 (0.03, 0.08)	0.06 (0.03, 0.10)	0.07 (0.04, 0.11)	0.04 (0.03, 0.07)	0.05 (0.04, 0.08)	0.06 (0.04, 0.10)
σ^2	5	4.40 (2.93, 7.90)	5.04 (3.27, 7.55)	4.90 (3.17, 7.90)	4.65 (3.01, 8.29)	4.20 (2.78 6.96)	4.74 (3.16, 6.78)	4.60 (3.07, 6.82)	4.42 (2.91, 6.97)
MSPE		3.11	4.73	3.79	3.70	3.40	4.60	4.05	3.73
Operations (in billions)		3125	50	125	820	25	0.4	1.0	6.6

Table 3: Lakes data set: Summary of candidate models' parameter estimates, 50 (2.5 97.5) percentiles.

	MCMC				INLA			
		Pred. proc. knots				Pred. proc. knots		
	Full	54	89	126	Full	54	89	126
β_0	1.72 (-1.13, 4.46)	1.58 (-1.34, 4.61)	1.73 (-1.32, 4.98)	1.76 (-1.00, 4.84)	1.59 (-1.25, 4.39)	1.50 (-1.67, 5.30)	1.64 (-1.37, 4.90)	1.62 (-1.28, 5.15)
β_1	-5.99 (-8.25, -4.39)	-6.25 (-8.95, -4.53)	-6.23 (-9.08, -4.53)	-6.20 (-8.88, -4.53)	-5.75 (-7.43, -4.26)	-5.79 (-7.36, -4.21)	-5.77 (-7.46, -4.21)	-5.77 (-7.33, -4.12)
β_2	-2.75 (-7.74, 1.57)	-2.56 (-7.75, 2.00)	-2.82 (-8.30, 1.99)	-2.80 (-8.15, 1.98)	-2.65 (-8.35, 2.14)	-2.44 (-7.77, 3.10)	-2.57 (-8.48, 2.38)	-2.57 (-7.48, 1.90)
ϕ	0.07 (0.03, 0.20)	0.07 (0.03, 0.18)	0.07 (0.03, 0.20)	0.08 (0.03, 0.22)	0.06 (0.03 , 0.20)	0.06 (0.03 , 0.18)	0.06 (0.03 , 0.23)	0.06 (0.03, 0.24)
σ^2	2.73 (0.91, 9.02)	3.12 (0.98, 11.34)	3.20 (1.04, 12.14)	3.17 (1.03, 11.69)	2.34 (0.76, 8.15)	2.52 (0.70, 8.31)	2.49 (0.76, 8.45)	2.54 (0.72, 8.77)

Table 4: Lakes data set: Predictions and observed data for holdout sets.

	Non-spatial			Spatial Pred. proc. model, 54 knots	
	$Y(s_{0,q}) = 0$	$Y(s_{0,q}) = 1$		$Y(s_{0,q}) = 0$	$Y(s_{0,q}) = 1$
$\hat{Y}(s_{0,q}) = 0$	110	32	$\hat{Y}(s_{0,q}) = 0$	122	16
$\hat{Y}(s_{0,q}) = 1$	33	195	$\hat{Y}(s_{0,q}) = 1$	21	211

Table 5: Forest landuse data set: Summary of candidate models' parameter estimates, 50 (2.5 97.5) percentiles.

		INLA		
		Pred. proc. knots		
	Non-spatial	50	100	200
β_0	4.04 (3.82, 4.28)	4.19 (2.83, 5.60)	3.86 (2.44, 5.33)	3.95 (2.23, 5.55)
β_1	-0.0044 (-0.0046, -0.0042)	-0.0049 (-0.0051, -0.0046)	-0.0049 (-0.0052, -0.0046)	-0.0049 (-0.0052, -0.0046)
β_2	0.0051 (0.0048, 0.0054)	0.0059 (0.0055, 0.0061)	0.0059 (0.0055, 0.0062)	0.0059 (0.0056, 0.0062)
β_3	-0.0027 (-0.0030, -0.0024)	-0.0033 (-0.0037, -0.0029)	-0.0034 (-0.0037, -0.003)	-0.0034 (-0.0037, -0.003)
ϕ	–	0.036 (0.033, 0.042)	0.039 (0.036, 0.045)	0.041 (0.031, 0.079)
σ^2	–	1.52 (1.40, 1.65)	1.68 (1.49, 1.89)	2.32 (1.28, 3.98)

Table 6: Forest landuse data set: Mean scoring rules based on prediction of hold-out set.

		INLA		
		Pred. proc. knots		
	Non-spatial	50	100	200
Zero-one	0.80	0.88	0.88	0.88
Quadratic	-0.29	-0.18	-0.18	-0.18
Spherical	0.84	0.90	0.90	0.90
Logarithmic	-0.45	-0.31	0.30	-0.30

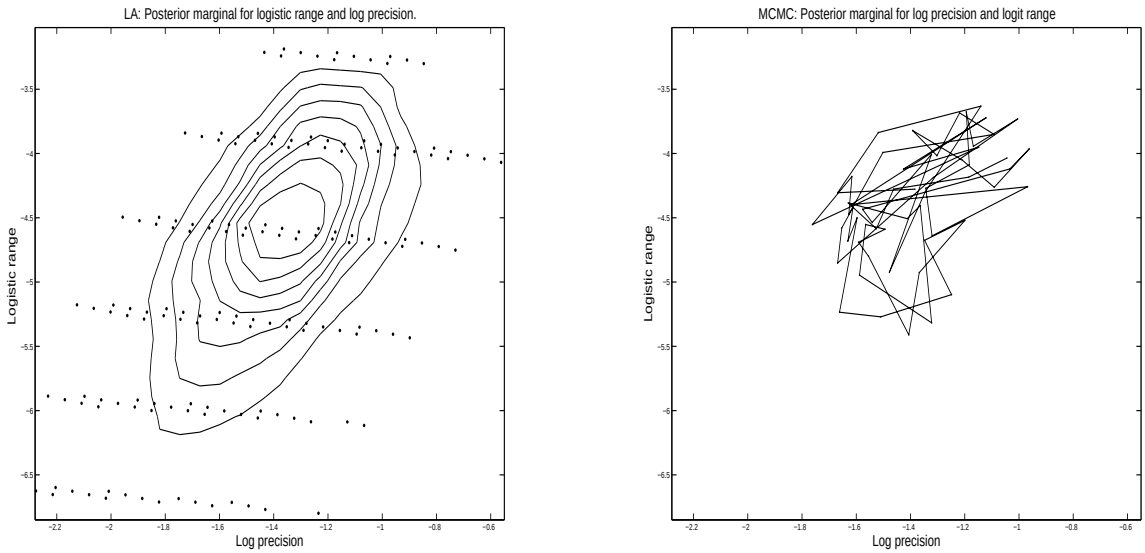


Figure 1: Illustration of posterior inference for logistic range and log precision, where nugget parameter is marginalized out. Left) The dots are evaluation points for the numerical Laplace approach. The contours are computed based on an these points. Right) The line segments show the first 200 realizations of a MCMC run.

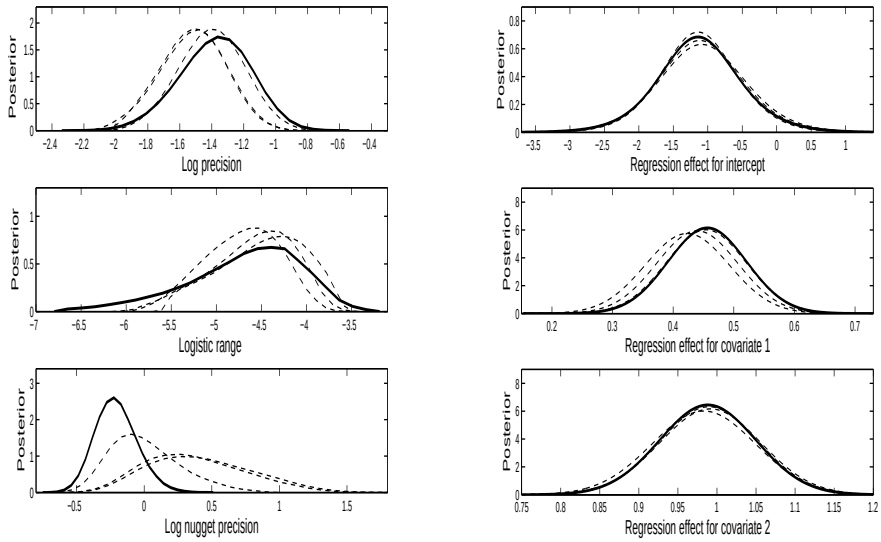


Figure 2: Synthetic Gaussian dataset: Left: LA approach for the posterior marginal of log precision (top), logistic range (center), and log nugget precition (bottom). Right: LA approach for the posterior marginals of regression parameters. The dashed curves represent three different knot sizes (64, 100, 256), while the solid curve is for the full data set (n=500).

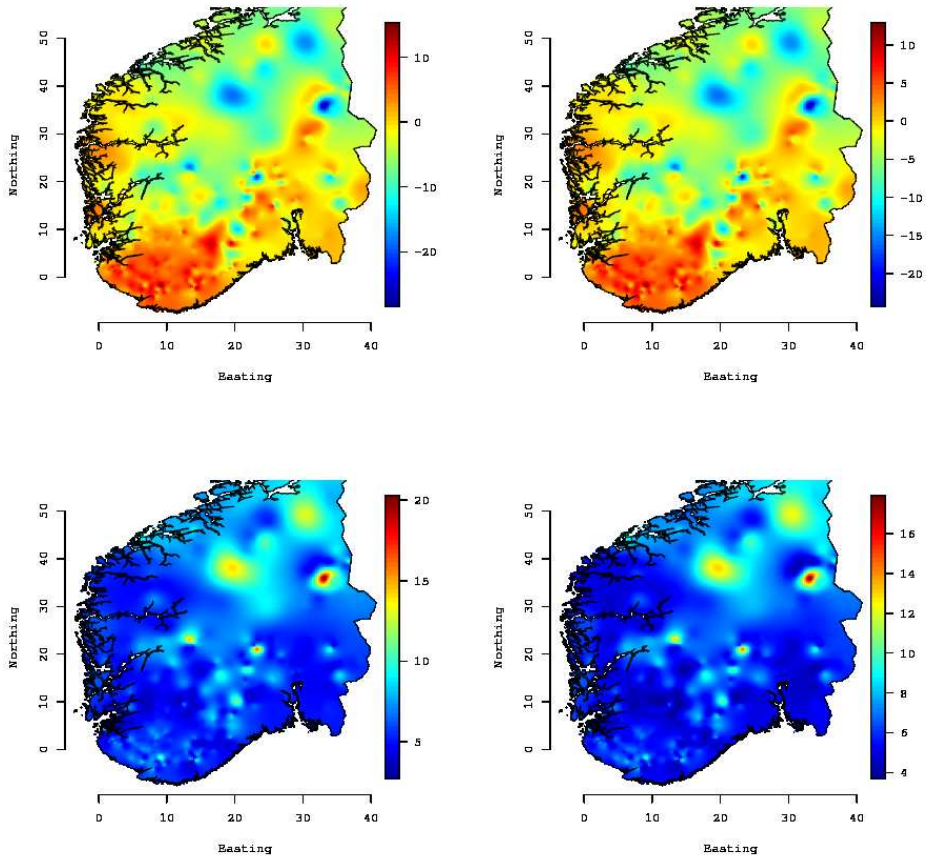


Figure 3: Lakes data set: Prediction (top) and 95% prediction range (bottom) of the latent intensity surface. Top left: MCMC median, full data set. Top right: LA median, 54 knot predictive process. Bottom left: MCMC range, full data set. Bottom right: LA range, 54 knot predictive process.

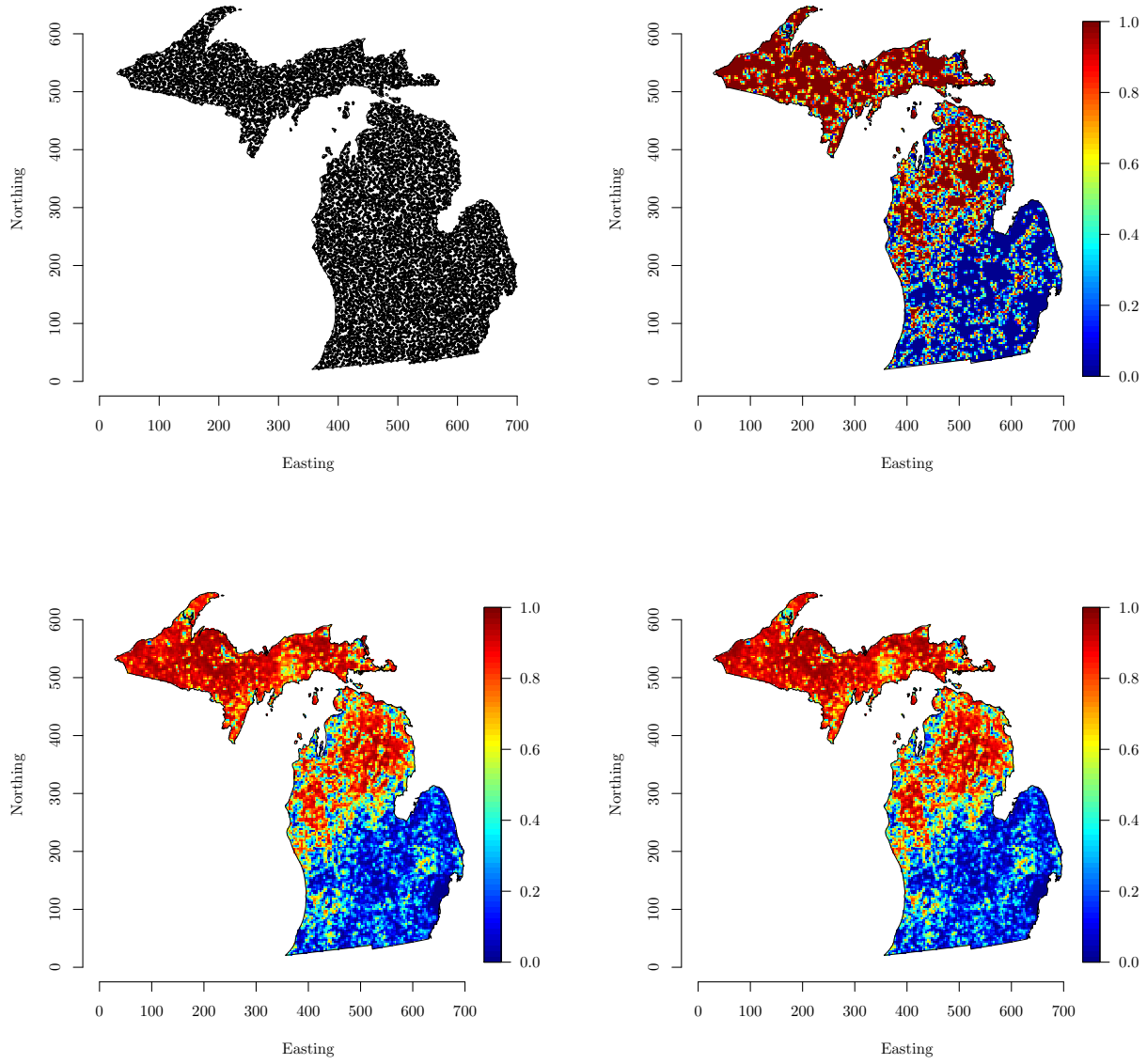


Figure 4: Forest landuse data set: Inventory data and probability of forest occupancy surfaces. Top left: forest inventory plot locations. Top right: observed forest occupancy. Bottom left: median fitted probability of forest occupancy for the 50 knot model. Bottom right: median fitted probability of forest occupancy for the 200 knot model. Map units are in kilometers.