Computational Statistics and Data Analysis 56 (2012) 1362–1380

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

Approximate Bayesian inference for large spatial datasets using predictive process models

Jo Eidsvik [a,*], Andrew O. Finley [b], Sudipto Banerjee [c], Håvard Rue [a]

[a] Department of Mathematical Sciences, NTNU, Trondheim, Norway
[b] Department of Forestry, Michigan State University, MI, USA
[c] Department of Biostatistics, University of Minnesota, MN, USA

ARTICLE INFO

Article history:
Received 21 December 2010
Received in revised form 24 October 2011
Accepted 28 October 2011
Available online 4 November 2011

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

ABSTRACT

The challenges of estimating hierarchical spatial models to large datasets are addressed. 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 datasets. The computational challenges in analyzing large spatial datasets are considered by merging two recent developments. First, the predictive process model is used as a reduced-rank spatial process, to diminish the dimensionality of the model. Then a computational framework is developed for estimating predictive process models using the integrated nested Laplace approximation. The settings where the first stage likelihood is Gaussian or non-Gaussian are discussed. Issues such as predictions and model comparisons are also discussed. Results are presented for synthetic data and several environmental datasets.

1. Introduction

The advent of geographic information systems have led to accurate geocoding of locations where massive amounts of scientific data are collected. This has generated considerable interest in statistical modeling 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.

Modeling large spatial datasets has 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 like 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, see e.g. Furrer et al. (2006), Kaufman et al. (2008), Du et al. (2009) and Sang and Huang (2012), that yield

* Corresponding author. Tel.: +47 90127472; fax: +47 73593524.
E-mail address: joeid@math.ntnu.no (J. Eidsvik).

0167-9473/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2011.10.022

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, see Higdon (2002), Kammann and Wand (2003), Stein (2007, 2008), Cressie and Johannesson (2008), Banerjee et al. (2008) and Crainiceanu et al. (2008). Many of these methods are variants of the so-called "subset of regressors" methods used in Gaussian process regressions for large datasets 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*, see 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. As such, the predictive process model is a dimension reduction technique that requires no additional tuning parameters in the modeling.

A key issue in predictive process modeling 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^* and knot locations. The issue is not dissimilar to a spatial design problem, e.g. Nychka and Saltzman (1998), Xia et al. (2006) and Diggle and Lophaven (2006). The standard method is to experiment with different knot configurations. 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, see 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 successful 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), Ainsworth and Dean (2006) and Evangelou et al. (2011). 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.

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

In this section we will present the predictive process models for Gaussian processes and for GLMMs. Our exposition is meant to facilitate the use of approximate Bayes inference methods applied to these models in Section 3.

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 $GP(0, C(\boldsymbol{s}_1, \boldsymbol{s}_2))$. 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 spatial 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 $\boldsymbol{\beta}$