Spatial Statistics 4 (2013) 14–44

Contents lists available at SciVerse ScienceDirect

Spatial Statistics

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

Hierarchical statistical modeling of big spatial datasets using the exponential family of distributions

Aritra Sengupta [a,*], Noel Cressie [a,b]

[a] Department of Statistics, The Ohio State University, United States

[b] National Institute for Applied Statistics Research Australia, University of Wollongong, Australia

ARTICLE INFO

Article history:
Received 2 October 2012
Accepted 21 February 2013
Available online 8 March 2013

Keywords:
EM algorithm
Empirical Bayes
Geostatistical process
Maximum likelihood estimation
MCMC
SRE model

ABSTRACT

Big spatial datasets are very common in scientific problems, such as those involving remote sensing of the earth by satellites, climate-model output, small-area samples from national surveys, and so forth. In this article, our interest lies primarily in very large, non-Gaussian datasets. We consider a hierarchical statistical model consisting of a conditional exponential-family model for the data and an underlying (hidden) geostatistical process for some transformation of the (conditional) mean of the data model. Within this hierarchical model, dimension reduction is achieved by modeling the geostatistical process as a linear combination of a fixed number of spatial basis functions, which results in substantial computational speed-ups. These models do not rely on specifying a spatial-weights matrix, and no assumptions of homogeneity, stationarity, or isotropy are made. Our approach to inference using these models is empirical-Bayesian in nature. We develop maximum likelihood (ML) estimates of the unknown parameters using Laplace approximations in an expectation–maximization (EM) algorithm. We illustrate the performance of the resulting empirical hierarchical model using a simulation study. We also apply our methodology to analyze a remote sensing dataset of aerosol optical depth.

* Corresponding author.
E-mail addresses: sengupta.11@osu.edu, atrsrv@gmail.com (A. Sengupta), ncressie@uow.edu.au (N. Cressie).

2211-6753/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.spasta.2013.02.002

1. Introduction

Big spatial datasets are very common in scientific problems, such as those involving remote sensing of the earth by satellites, climate-model output, small-area samples from national surveys, and so forth. In this article, our interest lies primarily in datasets that are very large and non-Gaussian in form. We consider a hierarchical statistical model consisting of two levels. At the first level, we have an exponential-family model for the data given a spatial process and parameters (which we call the data model). At the second level, we assume a geostatistical process given parameters (which we call the process model), for some transformation of the mean of the data model.

The exponential family of distributions include commonly used continuous and discrete distributions; for a detailed review, see McCullagh and Nelder (1989, Section 2.2.2). All members of the exponential family have a density or probability mass function that can be written as:

$$p(z|\gamma) = \exp\{(z\gamma - b(\gamma))/\tau^2 - c(z,\tau)\}, \tag{1}$$

where γ is called the canonical parameter or the natural parameter, $b(\gamma)$ is a function that depends only on γ, $c(z,\tau)$ is a function independent of γ, and τ is a scaling constant. The representation above is called the canonical form, or the natural form, of the exponential family.

Here, and in what follows, we use the notation $[A|B]$ to denote the conditional probability distribution of A given B. Suppose we have data, $Z_1, \ldots, Z_n$, coming from a member of the exponential family such that $\{[Z_i|\gamma_1, \ldots, \gamma_n] : i = 1, \ldots, n\}$ are mutually independent, and $[Z_i|\gamma_1, \ldots, \gamma_n] \equiv [Z_i|\gamma_i]$, where $[Z_i|\gamma_i]$ has density given by (1). Then one may proceed by modeling a transformation of the expectation of $[Z_i|\gamma_i]$, namely $E(Z_i|\gamma_i) = b'(\gamma_i)$, as

$$g(E(Z_i|\gamma_i)) = \mathbf{X}_i^\top \boldsymbol{\beta}, \tag{2}$$

where $g(\cdot)$ is the link function, $\mathbf{X}_i$ denotes a p-dimensional vector of known covariates, and $\boldsymbol{\beta}$ is a p-dimensional vector of regression coefficients. There are a lot of possible choices for $g(\cdot)$. The maximum likelihood (ML) estimator of $\boldsymbol{\beta}$ can be obtained via iteratively reweighted least squares. For a detailed review of the literature on GLMs, see McCullagh and Nelder (1989) or McCulloch et al. (2001).

When $Z_1, \ldots, Z_n$ are associated with locations in space, the assumption of independence is doubtful. A way to extend the framework above, that takes into account spatial variability, is to replace γ in (1) with a spatial process, $\{Y(\mathbf{s}) : \mathbf{s} \in D\}$, where D is the spatial domain of interest. The covariance between $Y(\mathbf{s})$ and $Y(\mathbf{u})$, for $\mathbf{s}, \mathbf{u} \in D$, is defined as:

$$C_Y(\mathbf{s}, \mathbf{u}) \equiv \mathrm{cov}(Y(\mathbf{s}), Y(\mathbf{u})).$$

Now consider spatial data $Z(\mathbf{s}_1), \ldots, Z(\mathbf{s}_n)$ from a GLM such that $\{[Z(\mathbf{s}_i)|Y(\cdot)] : i = 1, \ldots, n\}$ are mutually independent, and

$$g(E(Z(\mathbf{s}_i)|Y(\cdot))) = Y(\mathbf{s}_i); \quad i = 1, \ldots, n, \tag{3}$$

where $g(\cdot)$ is the link function. The hierarchical modeling framework defined above yields a spatial version of the GLM framework; it was proposed by Diggle et al. (1998), who assumed a Gaussian model for $Y(\cdot)$ and a prior distribution on its parameters. See also Omre and Tjelmeland (1997) for an exposition of the same framework for solving complex problems in petroleum geostatistics.

Lindley and Smith (1972) introduced a Bayesian-linear-model framework, where conditional and prior distributions come from a multivariate Gaussian distribution. In the spatial context, Omre (1987) defined Bayesian kriging for the linear model; for further extensions see Cressie (1993, Sec. 3.4.4). Besag et al. (1991) showed how a spatial model for counts in small areas could be decomposed hierarchically, where the hidden process $Y(\cdot)$ was used to model the spatial dependence. They assumed that the counts were (conditionally) Poisson distributed, and that the log means were a Gaussian spatial process, specifically a Gaussian Markov Random Field (MRF) known as the conditional autoregressive (CAR) model. However, a simultaneous autoregressive (SAR) model, or a geostatistical model could also be used. Indeed Diggle et al. (1998) employed spatial generalized linear mixed models (GLMMs) for spatially dependent non-Gaussian variables observed potentially anywhere in D, and they assumed