Computational Statistics and Data Analysis 56 (2012) 478–490

Contents lists available at SciVerse ScienceDirect

Computational Statistics and Data Analysis

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

Data augmentation strategies for the Bayesian spatial probit regression model

Candace Berrett[a,*], Catherine A. Calder[b]

[a] Department of Statistics, Brigham Young University, Provo, UT, USA

[b] Department of Statistics, The Ohio State University, Columbus, OH, USA

ARTICLE INFO

Article history:
Received 22 June 2010
Received in revised form 30 August 2011
Accepted 31 August 2011
Available online 14 September 2011

Keywords:
Conditional and marginal data augmentation
MCMC
Binary data
Latent variable methods
Spatial statistics

ABSTRACT

The well known latent variable representation of the Bayesian probit regression model due to Albert and Chib (1993) allows model fitting to be performed using a simple Gibbs sampler. In addition, various types of dependence among categorical outcomes not explained by covariate information can be accommodated in a straightforward manner as a result of this latent variable representation of the model. One example of this is the spatial probit regression model for spatially-referenced categorical outcomes. In this setting, commonly used covariance structures for describing residual spatial dependence in the normal linear model setting can be imbedded into the probit regression model. Capturing spatial dependence in this way, however, can negatively impact the performance of MCMC model-fitting algorithms, particularly in terms of mixing and sensitivity to starting values. To address these computational issues, we demonstrate how the non-identifiable spatial variance parameter can be used to create data augmentation MCMC algorithms. We compare the performance of several non-collapsed and partially collapsed data augmentation MCMC algorithms through a simulation study and an analysis of land cover data.

1. Introduction

There has been a recent emphasis in the spatial statistics literature on the development of methods that accommodate large data sets. In the Bayesian setting, these methods include dimension reduction techniques (e.g., Higdon, 2002; Xu et al., 2005; Calder, 2007; Banerjee et al., 2008), integrated nested Laplacian approximations (Rue et al., 2009), and covariance tapering (see recent work by Shaby and Ruppert, in press, for a Bayesian treatment of this technique). In this paper, instead of focusing on model adjustments to accommodate large data sets or approximations to full Bayesian inference procedures, we investigate strategies for efficient Markov chain Monte Carlo (MCMC) algorithms for a particular class of spatial probit regression models. In particular, we introduce various data augmentation strategies for the Bayesian probit regression model for spatially-referenced categorical response variables. Such data frequently arise, for example, in analyses of satellite-derived land cover data, where measured surface reflectances are classified into one of several distinct categories (e.g., forest, agriculture, or urban). Using such data, researchers may want to better understand the economic, social, political, and demographic factors associated with observed land cover patterns while accounting for (residual) spatial dependence. While the MCMC strategies discussed here are not necessarily designed to overcome computational challenges associated with massive data sets, in high dimensional settings having efficient algorithms is clearly desirable.

Data augmentation/latent variable methods have been widely recognized for facilitating model fitting in the Bayesian probit regression model. First proposed by Albert and Chib (1993) for independent binary and multi-category data, the

[*] Corresponding author.
E-mail addresses: cberrett@stat.byu.edu (C. Berrett), calder@stat.osu.edu (C.A. Calder).

0167-9473/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2011.08.020

latent variable representation of the Bayesian probit regression model allows model fitting to be performed using a simple Gibbs sampler and, for more than two categories, also allows the so-called assumption of irrelevant alternatives required by the logistic regression model to be relaxed (Hausman and Wise, 1978). To improve the efficiency of the Gibbs sampler in this setting, Imai and van Dyk (2005) propose introducing a working parameter (defined in Section 3.1) into the model and compare various data augmentation strategies resulting from different treatments of the working parameter. We build on this work by investigating the efficiency of modified and extended versions of these algorithms for the spatial probit regression model, focusing on the special case of binary response variables. These algorithms include the one previously proposed by De Oliveira (2000), which we discuss further in Section 3.1.

Before describing our proposed methodology, we note that there are a large number of alternatives to the spatial probit regression model for analyzing spatially-referenced binary data. Indicator kriging methods (Switzer, 1977; Journel, 1983) have been proposed in the literature, but have been widely criticized due to their weak theoretical predictive properties (Cressie, 1993). Other kriging methods (i.e., best linear unbiased spatial prediction) for binary data include disjunctive and probability kriging (Cressie, 1993). Model-based approaches that directly specify the spatial dependence among binary response variables are the autologistic model (Besag, 1972) and the spatial generalized linear model (GLM; McCullagh and Nelder, 1989 and Albert and McShane, 1995). The spatial probit regression model featured in this paper is a particular type of spatial GLM. Alternatively, spatial generalized linear mixed models (GLMMs) capture spatial dependence through the introduction of an underlying Gaussian process (Diggle et al., 1998). For a review of various versions of spatial logistic GLMMs, as well as a discussion of the specification of spatial dependence in this class of models, we refer the reader to Paciorek (2007).

The remainder of this paper is organized as follows. In Section 2, we review the Albert and Chib (1993) data augmentation strategy for the Bayesian probit regression model. We then discuss the extension of this methodology to the spatially-dependent setting. In Section 3 we propose three MCMC strategies for sampling from the Bayesian spatial probit regression model. We compare our model-fitting algorithms through a simulation study in Section 4 and through an analysis of satellite-derived land cover observations over Southeast Asia in Section 5. The paper concludes with a summary of our findings and a discussion of future research possibilities in Section 6.

2. The Bayesian probit regression model

2.1. Albert and Chib's data augmentation strategy

Consider $\{Y_i, \mathbf{x}_i : i = 1, \dots, n\}$, a collection of n binary response variables Y_i and corresponding $k \times 1$ vectors of covariates $\mathbf{x}_i$. The probit GLM relating the covariates to the binary response assumes that the Y_i are conditionally independent given a $k \times 1$ vector of regression coefficients $\boldsymbol{\beta}$ and can be written as

$$\begin{aligned} &Y_i|\boldsymbol{\beta} \sim \text{Bernoulli}(p_i) \\ &\Phi^{-1}(p_i) = \mathbf{x}_i'\boldsymbol{\beta}, \end{aligned} \tag{1}$$

where $\Phi^{-1}(\cdot)$ denotes the inverse standard normal cumulative distribution function. In the Bayesian setting, a prior distribution must be specified for the unknown parameter $\boldsymbol{\beta}$. Unlike the normal linear regression model where the normal distribution is a conjugate prior for the regression coefficients, a conjugate prior is not available for $\boldsymbol{\beta}$ in (1). Thus, inference on $\boldsymbol{\beta}$ typically requires deterministic integration, which is not feasible if k is large, or a simulation-based approach such as MCMC.

In order to facilitate the use of the Gibbs sampler, a particular class of MCMC algorithms, in the Bayesian probit regression model, Albert and Chib (1993) propose the following data augmentation representation of the model. They introduce a collection of latent variables $\tilde{\mathbf{Z}} = (\tilde{Z}_1, \dots, \tilde{Z}_n)'$ and take

$$Y_i = \begin{cases} 1, & \text{if } \tilde{Z}_i > 0 \\ 0, & \text{if } \tilde{Z}_i \le 0 \end{cases} \tag{2}$$

where

$$\tilde{\mathbf{Z}} \sim \text{N}(\mathbf{X}\tilde{\boldsymbol{\beta}}, \sigma^2\mathbf{I}_n), \tag{3}$$

$\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)'$ is an $n \times k$ matrix of covariates, $\mathbf{I}_n$ is the $n \times n$ identity matrix, and σ^2 is a variance parameter. (The $\tilde{}$ notation used here distinguishes identifiable and non-identifiable parameters, and we follow this notational convention to aid our discussion of data augmentation strategies in Section 3.) Taking $\sigma^2 = 1$, setting $Z_i = \tilde{Z}_i/\sigma$ and $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}/\sigma$, and integrating out the Z_i, it is straightforward to show that Albert and Chib's model specification is equivalent to the probit GLM given in (1). Introducing the latent $\tilde{Z}_i$ and taking the prior on $\boldsymbol{\beta}$ to be $\text{N}\left(m_\beta, C_\beta\right)$ facilitates model fitting via the following Gibbs sampler:

Step 1: sample $\mathbf{Z}|\mathbf{Y}, \boldsymbol{\beta}, \sigma^2 = 1$
Step 2: sample $\boldsymbol{\beta}|\mathbf{Z}, \mathbf{Y}, \sigma^2 = 1$.

Each of these steps involves drawing from known distributions from which sampling is easily implemented; see Albert and Chib (1993) for details.