Computational Statistics and Data Analysis 90 (2015) 1–14

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Data augmentation and parameter expansion for independent or spatially correlated ordinal data

Erin M. Schliep[a,*], Jennifer A. Hoeting[b]

[a] Duke University, Durham, NC, USA
[b] Colorado State University, Fort Collins, CO, USA

ARTICLE INFO

Article history:
Received 28 November 2013
Received in revised form 29 March 2015
Accepted 30 March 2015
Available online 17 April 2015

Keywords:
Data augmentation
Parameter expansion
Ordinal data
Probit model
Spatial correlation

ABSTRACT

Data augmentation and parameter expansion can lead to improved iterative sampling algorithms for Markov chain Monte Carlo (MCMC). Data augmentation allows for simpler and more feasible simulation from a posterior distribution. Parameter expansion accelerates convergence of iterative sampling algorithms by increasing the parameter space. Data augmentation and parameter-expanded data augmentation MCMC algorithms are proposed for fitting probit models for independent ordinal response data. The algorithms are extended for fitting probit linear mixed models for spatially correlated ordinal data. The effectiveness of data augmentation and parameter-expanded data augmentation is illustrated using the probit model and ordinal response data, however, the approach can be used broadly across model and data types.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

The primary goal of data augmentation is to construct an efficient and iterative sampling algorithm by introducing latent or unobserved variables into the model. The approach first became popular within deterministic algorithms for maximizing likelihood functions or posterior densities using the expectation–maximization (EM) algorithm (Dempster et al., 1977). The work of Tanner and Wong (1987) popularized data augmentation within the literature of stochastic algorithms by developing the method for posterior sampling. The schemes are used to make simulating from the posterior distribution simpler and more feasible, and thus, improve the speed of iterative simulation (Swendsen and Wang, 1987). Constructing a data augmentation algorithm is somewhat of an art because data augmentation algorithms must be carefully developed for each type of model (Van Dyk and Meng, 2001).

Parameter-expanded data augmentation is a method of introducing additional parameters into the model through the augmented data (Liu et al., 1998). The extra parameters are introduced without distorting the original observed data model. The increase, or expanded, parameter space can speed up convergence of the algorithm.

In this work we propose a set of parameter-expanded data augmentation algorithms for modeling ordinal data using the probit model (Sections 2 and 3). In Section 4 we extend the algorithms for the probit linear mixed model for spatially correlated ordinal response data and apply them using a model for biotic integrity of wetlands in Colorado. We conclude with a discussion in Section 5.

[*] Correspondence to: Department of Statistical Science, Duke University, Box 90251, Durham, NC, 27708, USA. Tel.: +1 919 684 4210.
E-mail address: erin.schliep@duke.edu (E.M. Schliep).

http://dx.doi.org/10.1016/j.csda.2015.03.020
0167-9473/© 2015 Elsevier B.V. All rights reserved.

2. Data augmentation

The goal of data augmentation strategies within the Bayesian framework is to weaken the dependence between draws from the posterior distribution within a Markov chain Monte Carlo algorithm (Liang et al., 2011). Chains with lower parameter dependence have better mixing and faster convergence. Models with an elaborate hierarchy structure or a high-dimensional parameter space can greatly benefit from the approach. Data augmentation techniques have been shown to increase the conditional variability of the parameters of interest given the observed (or augmented) data (e.g., Besag and Green, 1993). This leads to larger jumps between draws of the parameters within the chain resulting in a faster and more efficient exploration of the parameter space. A second benefit of data augmentation is that it can also lead to known, closed-form posterior full conditional distributions. In some cases this can alleviate the need for Metropolis–Hastings steps within a Gibbs sampling algorithm. Many data augmentation strategies have been used to facilitate Bayesian analysis (e.g. Tanner and Wong, 1987; Royle et al., 2007; Cauchemez et al., 2004). In this work we focus on latent variable models for ordinal response data. We use data augmentation to improve the performance of the MCMC algorithm.

2.1. *Probit model for ordinal data*

The probit model is a common approach for modeling ordinal data. For a Bayesian analysis, the posterior distribution of the probit model parameters given the response data is not available in closed form. Here, we consider a K-class ordinal model where the observable response data, $\boldsymbol{Y}$, are such that $\boldsymbol{Y} \in \{1, 2, \ldots, K\}$. Let $\boldsymbol{\lambda} = [\lambda_0, \lambda_1, \ldots, \lambda_k]$ be a non-decreasing vector of thresholds where $\lambda_0 = -\infty$ and $\lambda_K = \infty$. Under the probit model we define the density of the ith observable Y_i, in terms of the p-vector of covariates for the ith observation, $\boldsymbol{X}_i$, the p-vector of coefficients $\boldsymbol{\beta}$, and thresholds, $\boldsymbol{\lambda}$, as

$$P(Y_i = k) = \Phi\left(\lambda_k - \boldsymbol{X}_i'\boldsymbol{\beta}\right) - \Phi\left(\lambda_{k-1} - \boldsymbol{X}_i'\boldsymbol{\beta}\right), \tag{1}$$

where Φ is the CDF of the standard normal distribution.

Define the observed data as $\boldsymbol{y} = [y_1, \ldots, y_n]$, where n is the number of observations in the dataset. In the Bayesian framework using Markov chain Monte Carlo (MCMC), we need to sample from the posterior distribution of $(\boldsymbol{\beta}, \boldsymbol{\lambda})$ given the observed data, $\boldsymbol{y}$. Using a Gibbs sampler we sample iteratively from the full conditional distributions of $\boldsymbol{\beta}$ and $\boldsymbol{\lambda}$ separately (Algorithm 1, hereafter referred to as A.1). Neither of the full conditional distributions are available in closed form requiring Metropolis–Hastings steps which can be challenging to tune. This motivates the desire for a better algorithm for sampling from the posterior distribution.

Algorithm 1. Gibbs algorithm for ordinal data.

```
1. Sample $\boldsymbol{\lambda}^t$ from $p(\boldsymbol{\lambda}^t \mid \boldsymbol{y}, \boldsymbol{\beta}^{t-1})$ using a Metropolis–Hastings step.
2. Sample $\boldsymbol{\beta}^t$ from $p(\boldsymbol{\beta}^t \mid \boldsymbol{y}, \boldsymbol{\lambda}^t)$ using a Metropolis–Hastings step.
```

2.2. *Data augmentation for the probit model for ordinal data*

The probit model can alternatively be defined using latent variables and is an application of data augmentation (Albert and Chib, 1993). Assume the latent variable, defined by Z_i, is modeled as

$$Z_i = \boldsymbol{X}_i'\boldsymbol{\beta} + \epsilon_i, \tag{2}$$

where $\epsilon_i \sim N(0, 1)$. We can define the relationship between the latent variable Z_i and the observable random variable Y_i as

$$\begin{aligned} P(Y_i = k) &= P(\lambda_{k-1} < Z_i \le \lambda_k) = P(Z_i \le \lambda_k) - P(Z_i < \lambda_{k-1}) \\ &= \Phi\left(\lambda_k - \boldsymbol{X}'\boldsymbol{\beta}\right) - \Phi\left(\lambda_{k-1} - \boldsymbol{X}_i'\boldsymbol{\beta}\right) \end{aligned} \tag{3}$$

for $k = 1, \ldots, K$. Under this specification, the augmented model (2) is equivalent to the original probit model in (1).

In general, data augmentation is advantageous if the conditional distributions of the data-augmented model are easier to sample from than the conditional distributions of the model containing only the observed data. For the probit model, the data-augmented model (2) has advantages over the original parameterization (1) if sampling from both $p(\boldsymbol{\beta}, \boldsymbol{\lambda}|\boldsymbol{Z})$ and $p(\boldsymbol{Z}|\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{\lambda})$ is easier than sampling directly from $p(\boldsymbol{\beta}, \boldsymbol{\lambda}|\boldsymbol{y})$.

Using the augmented data approach, there are several possible sampling algorithms for drawing inference. One approach is to use a three-step Gibbs sampler where $\boldsymbol{Z}, \boldsymbol{\lambda}, \boldsymbol{\beta}$ are all drawn from their full conditional distributions (Albert and Chib, 1993). This algorithm, referred to as the *data-augmented Gibbs sampler*, is given in Algorithm 2.

Algorithm 2. Data-augmented Gibbs sampler for ordinal data.

```
1. Draw $\boldsymbol{Z}^t$ from $p(\boldsymbol{Z} \mid \boldsymbol{y}, \boldsymbol{\beta}^{t-1}, \boldsymbol{\lambda}^{t-1})$.
2. Draw $\boldsymbol{\lambda}^t$ from $p(\boldsymbol{\lambda} \mid \boldsymbol{y}, \boldsymbol{Z}^t, \boldsymbol{\beta}^{t-1})$.
3. Draw $\boldsymbol{\beta}^t$ from $p(\boldsymbol{\beta} \mid \boldsymbol{y}, \boldsymbol{Z}^t, \boldsymbol{\lambda}^t)$.
```

Download English Version:

https://daneshyari.com/en/article/416283

Download Persian Version:

https://daneshyari.com/article/416283

Daneshyari.com