Computational Statistics and Data Analysis 71 (2014) 274–287

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

State space mixed models for binary responses with scale mixture of normal distributions links

Carlos A. Abanto-Valle [a,*], Dipak K. Dey [b]

[a] Department of Statistics, Federal University of Rio de Janeiro, Caixa Postal 68530, CEP: 21945-970, RJ, Brazil

[b] Department of Statistics, University of Connecticut, USA

ARTICLE INFO

Article history:
Received 8 April 2012
Received in revised form 13 January 2013
Accepted 15 January 2013
Available online 21 January 2013

Keywords:
Binary time series
Longitudinal data
Markov chain Monte Carlo
Particle learning
Probit
Scale mixture of normal links
Sequential Monte Carlo
State space models

ABSTRACT

A state space mixed models for binary time series where the inverse link function is modeled to be a cumulative distribution function of the scale mixture of normal (SMN) distributions. Specific inverse links examined include the normal, Student-t, slash and the variance gamma links. The threshold latent approach to represent the binary system as a linear state space model is considered. Using a Bayesian paradigm, an efficient Markov chain Monte Carlo (MCMC) algorithm is introduced for parameter estimation. The proposed methods are illustrated with real data sets. Empirical results showed that the slash inverse link fits better over the usual inverse probit link.

1. Introduction

In many areas of application of statistical modeling one encounters observations that take one of two possible forms. Such binary data are often measured with covariates or explanatory variables that either continuous or discrete or categorical. Time series of binary responses may adequately be described by Generalized linear models (McCullagh and Nelder, 1989). However, if serial correlation is present or if the observations are overdispersed, these models may not be adequate, and several approaches can be taken. Generalized linear state space models also address those problems and are treated in a paper by West et al. (1985) in a conjugate Bayesian setting. They have been subject to further research by Fahrmeir (1992), Song (2000), Carlin and Polson (1992) and Czado and Song (2008) among others.

Consider a binary time series $\{Y_t,\ t = 1, \ldots, T\}$, taking the values 0 or 1 with probability of success given by π_t and which is related with a time-varying covariates vector $\mathbf{x}_t = (x_{t1}, \ldots, x_{tk})'$ and a q-dimensional latent state variable $\boldsymbol{\theta}_t$. We consider a Generalized linear state space model framework for binary responses in the following way

$$Y_t \sim \mathcal{B}\mathrm{er}(\pi_t) \quad t = 1, \ldots, T, \tag{1}$$

$$\pi_t = F(\mathbf{x}_t'\beta + \mathbf{S}_t'\theta_t), \tag{2}$$

$$\boldsymbol{\theta}_t = \mathbf{G}_t\boldsymbol{\theta}_t + \boldsymbol{\eta}_t \quad \boldsymbol{\eta}_t \sim \mathcal{N}_q(\mathbf{0}, \mathbf{W}_t). \tag{3}$$

In the above setup the observed process $\{Y_t\}$ is described by Eqs. (1)–(2), where $\pi_t = P(Y_t = 1 \mid \boldsymbol{\theta}_t, \mathbf{x}_t, \mathbf{S}_t)$ is the conditional probability of success, $\mathbf{S}_t$ is a q-dimensional vector, $\boldsymbol{\beta}$ is a k-dimensional vector of regression coefficients and

* Corresponding author. Fax: +55 2125627374.
E-mail addresses: cabantovalle@im.ufrj.br, cabantovalle@gmail.com (C.A. Abanto-Valle).

0167-9473/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2013.01.009

$\mathbf{x}_t = (x_{t1}, \dots, x_{tk})'$ is a $k \times 1$ vector of covariates. The system process is defined as a first order Markov process in Eq. (3), where $\mathbf{G}_t$ is the $q \times q$ transition matrix, $\mathbf{W}_t$ is the covariance matrix of error term η_t, $\mathcal{B}\text{er}(\cdot)$ and $\mathcal{N}_q(.,.)$ indicate the Bernoulli and the q-dimensional normal distributions respectively. In the terminology of generalized linear models (McCullagh and Nelder, 1989), F is the inverse link function. For ease of exposition, we refer to F as the link function in this article.

A critical issue in modeling binary response data is the choice of the links. In the context of binary regression problems, the probit link is widely used in the literature. Albert and Chib (1993) using the data augmentation principle introduced the threshold latent approach to deal with the symmetric probit and Student-t links in an elegant way. Recently, Naranjo et al. (forthcoming) used the exponential power link. Other symmetric links using normal scale mixture links in a nonparametric setup are described in Basu and Mukhopadhyay (2000a,b). The binary state space model with probit link using the threshold approach (Albert and Chib, 1993) have been used by Carlin and Polson (1992) and Song (2000) without including covariates. Czado and Song (2008) introduced covariates for binary state space models with probit link and called the resulting class as binary state space mixed models (BSSMM). They justified that including regression variables is appealing as it would enable us to quantify the relationship between the probability of success and covariates.

In this paper, we extend the BSSMM with probit link (Czado and Song, 2008) by assuming the flexible class of scale mixtures of normal (SMN) links (Lange and Sinsheimer 1993; Chow and Chan 2008) and the univariate latent states follow a first order autoregressive process. Interestingly, this rich class contains as proper elements the normal (BSSMM-N), Student-t (BSSMM-T), slash (BSSMM-S) and variance gamma (BSSMM-VG) links. All these distributions have heavier tails than the normal one, and thus can be used for robust inference in these types of models. We refer to this generalization as BSSMM-SMN. Inference in the class of BSSMM-SMN is performed under a Bayesian paradigm via MCMC methods, which permits to obtain the posterior distribution of parameters by simulation starting from reasonable prior assumptions on the parameters. Using the threshold latent approach (Albert and Chib, 1993), we simulate the latent states in an efficient way by using the simulation smoother of de Jong and Shephard (1995).

The remainder of this paper is organized as follows. Section 2 gives a brief review about the SMN distributions and links. Section 3 outlines the general class of the BSSMM-SMN models as well as the Bayesian estimation procedure using MCMC methods. Particle Filtering methods are developed to calculate the predictive likelihood function. Section 5 is devoted to the application and model comparison among particular members of the BSSM-SMN models using two real data sets. Finally, some concluding remarks and suggestions for future developments are given in Section 6.

2. Scale mixture of normal distributions

A random variable Y belongs to the SMN family if it can be expressed as

$$Y = \mu + \kappa(\lambda)^{1/2} X, \tag{4}$$

where μ is a location parameter, $X \sim \mathcal{N}(0, \sigma^2)$, λ is a positive mixing random variable with *cdf* $H(\cdot \mid \boldsymbol{\nu})$ and *pdf* $h(\cdot|\boldsymbol{\nu})$, $\boldsymbol{\nu}$ is a scalar or parameter vector indexing the distribution of λ and $\kappa(\cdot)$ is a positive weight function. As in Lange and Sinsheimer (1993) and Chow and Chan (2008), we restrict our attention to the case in that $\kappa(\lambda) = 1/\lambda$. Given λ, we have $Y|\lambda \sim \mathcal{N}(\mu, \lambda^{-1}\sigma^2)$ and the pdf of Y is given by

$$f_{\text{SMN}}(y|\mu, \sigma^2, \nu) = \int_{-\infty}^{\infty} \phi(y|\mu, \lambda^{-1}\sigma^2) dH(\lambda|\boldsymbol{\nu}), \tag{5}$$

where $\phi(\cdot \mid \mu, \sigma^2)$ denotes the density of the univariate $\mathcal{N}(\mu, \sigma^2)$ distribution. From Eq. (5), we have that the *cdf* of the SMN distributions is given by

$$\begin{aligned} F_{\text{SMN}}(y|\mu, \sigma^2, \nu) &= \int_{-\infty}^{y} \int_{-\infty}^{\infty} \phi(u|\mu, \lambda^{-1}\sigma^2) dH(\lambda|\boldsymbol{\nu}) du \\ &= \int_{-\infty}^{\infty} \Phi\left(\frac{\lambda^{1/2}[y-\mu]}{\sigma}\right) dH(\lambda|\boldsymbol{\nu}), \end{aligned} \tag{6}$$

where $\Phi(\cdot)$ is the *cdf* of the standard normal distribution. The notation $Y \sim \text{SMN}(\mu, \sigma^2, \boldsymbol{\nu}, H)$ will be used when Y has *pdf* (5) and *cdf* (6), as before $H(\cdot \mid \boldsymbol{\nu})$ is the cdf of the mixing variable λ. As was mentioned above, the SMN family constitutes a class of thick-tailed distributions including the normal, the Student-t, the Slash and variance gamma distributions, which are obtained respectively by choosing the mixing variables as: $\lambda = 1$, $\lambda \sim \mathcal{G}(\frac{\nu}{2}, \frac{\nu}{2})$, $\lambda \sim \mathcal{B}\text{e}(\nu, 1)$ and $\lambda \sim \mathcal{IG}(\frac{\nu}{2}, \frac{\nu}{2})$, where $\mathcal{G}(.,.)$, $\mathcal{B}\text{e}(.,.)$ and $\mathcal{IG}(.,.)$ denote the gamma, beta and inverse gamma distributions respectively.

3. Binary responses state space mixed models with normal scale mixture links

In this section we introduce the BSSM with SMN links using a latent variable representation in order to develop an efficient MCMC algorithm for parameter estimation.