Modelling categorized levels of rainfall

Patrícia L. C. Velozo[a,b], Mariane B. Alves[a*] and Alexandra Melo Schmidt[a]

[a]Instituto de Matemática - UFRJ

[b]GET - IME - Universidade Federal Fluminense

Abstract

We propose a dynamic model to analyze polychotomous data subject to temporal variation. In particular, we propose to model categorized levels of rainfall across time. Our model assumes that the observed category is related to an underlying latent continuous variable, which is modelled according to a power transformation of a Gaussian latent process, centered on a predictor that assigns dynamic effects to observable covariates. The inference procedure is based on the Bayesian paradigm and makes use of Markov chain Monte Carlo methods. We analyze artificial sets of data and daily measurements of rainfall in Rio de Janeiro, Brazil. When compared to the fitting of the actual observed volume of rainfall, our categorized model seems to recover well the structure of the data.

Key words: Bayesian Inference, Cumulative link model, Latent variable, Ordinal data, Probit model.

1 Introduction

In different fields of science, such as atmospheric sciences, agriculture, and hydrology, understanding and forecasting levels of precipitation over a region, across time, is a key issue. Depending on the time scale, observed values of precipitation are either equal to 0 (dry period) or equal to a positive quantity. For this reason, it is important to have statistical models that account for this property of the data. There are in the literature different proposals to model levels of rainfall. Stid (1973) proposes a model which assumes that levels of precipitation are realizations from a normal distribution that has been truncated and transformed. Sansó and Guenni (1999a)

*Responsible author. Address for correspondence: Instituto de Matemática - UFRJ Caixa Postal 68530, CEP 21945-970 Rio de Janeiro, RJ, FAX: 55 21 25627374, mariane@im.ufrj.br

propose a dynamic version of the model proposed by Stid (1973). More specifically, Sansó and Guenni (1999a) assume that levels of rainfall are a power transformation of a normal process, which, in turn, is centered on a dynamic linear predictor allowing covariates' effects to vary smoothly through time. Sansó and Guenni (1999b) extend the idea of the dynamic model to a spatio-temporal setting. De Oliveira (2004) proposes a model for rainfall fields that do not have continuous distributions, and possess a distinctive probabilistic structure that is not presented by standard random field models. His proposal is suitable for short to medium periods of time as it accounts for the zero inflation typically present in such rainfall data. Fernandes et al. (2009) pursue a different approach by assuming that observed rainfall is a realization from a mixture distribution between a variable with Bernoulli distribution, and another one assuming only positive values. They explore the exponential, gamma and lognormal distributions for the positive part of the model.

For some applications the interest lies only in predicting if it will rain or not. In this case one can propose models for precipitation occurrence by assuming, for example, a temporal logistic or probit regression. Alternatively, Hughes et al. (1999) propose a non-homogeneous hidden Markov model for relating precipitation occurrences to broad scale atmospheric circulation patterns.

Here we propose to consider that observed volumes of rainfall at each time t can be categorized into one of J categories. As pointed out by Fuentes et al. (2008), rain gauges are widely used to measure rainfall accumulation, but the information they provide is limited by their spatial and temporal resolution. Rainfall estimates are also obtained through remote senses which provide information about rainfall at locations which do not have a ground monitor. We focus on situations in which the actual volume of rainfall for some time t at a particular location is unknown. However, it is known, through different sources of information (remote sense, physical model, etc.), in which range, e.g. dry, drizzle, rain, storm, the amount of rainfall at time t is.

The multinomial distribution is a natural choice to model polychotomous data. For ordinal responses, it is usual to model the cumulative distribution function, according to the so called cumulative link models, as seen in e.g., Agresti (1990) and Congdon (2005). The choice of a link function can be arbitrary or induced by data augmentation, which is a method frequently adopted to model categorical ordinal data. The idea is to assume that the categorical response is generated by an underlying, latent, continuous variable, supposed to be divided into intervals, each of which representing a category.

Albert and Chib (1993) develop exact Bayesian inference for polychotomous data by using data augmentation. The idea is to make use of an underlying normal regression structure on latent continuous data. On the other hand, Chen and Dey (2000) use scale mixture of multivariate normal link functions to model correlated ordinal response data.

On a pure spatial setting, De Oliveira (2000) proposes a model for binary random fields by clipping a Gaussian random field at a fixed level. Higgs and Hoeting (2010) extend the approach of De Oliveira (2000) to model ordinal, categorical spatial observations. Berret and Calder (2010) develop strategies to improve the inference of the Bayesian spatial probit regression model.

In the temporal context, Carlin and Polson (1992) assume that the categorical time series is a known function of an underlying continuous process which evolves according to a state-space model. Inference is performed under the Bayesian paradigm and they concentrate on the dichotomous case. Knorr-Held (1995) proposes a dynamic version of the cumulative probit model. In particular, a multivariate autoregressive structure is assumed for the regression coefficients and threshold parameters which define each of the categories. Cargnoni et al. (1997) discuss a class of conditionally Gaussian dynamic models for non-normal, multivariate time series. They focus on multivariate time series of multinomial observations.

This paper is organized as follows. Next section proposes a model for temporal observations of categories of rainfall. Basically, we assume the latent approach of Albert and Chib (1993), but model the latent variable following Sansó and Guenni (1999a). Besides, we consider the bin boundaries that connect the latent variable with each of the J categories to be unknown. Therein we also discuss possible identifiability problems with the multinomial model. Then, in Section 3 we start by performing a simulation study to check if our proposed model is able to capture the true structure of the data when the truth is known. We provide an example with real data by analyzing observed temporal categories of rainfall in Rio de Janeiro, Brazil. As the actual volumes of rainfall are observed for this data set we also fit a model to the daily observed amount of rain and compare the predictions based on the categorized and continuous observations. Finally, Section 4 presents some concluding remarks and points to future avenues of research.

2 Proposed Model

Let $Y_t = j$ be an ordinal categorical variable indicating that the response variable is in category j at time t, which is equivalent to define a vector of variables $\boldsymbol{Y_{t\cdot}} = (Y_{t1}, \cdots, Y_{tJ})$ where $Y_{tj} = 1$ and $Y_{tr} = 0$, $r = 1, \cdots, J$ and $r \neq j$. Let π_{tj} be the probability that the response variable lies in category j at time t, that is, $\pi_{tj} = Pr(Y_{tj} = 1)$. Then, given this probability, the response variable follows a multinomial distribution:

$$\boldsymbol{Y_{t\cdot}} | \boldsymbol{\pi_{t\cdot}} \sim \text{Multinomial}\left[1, \boldsymbol{\pi_{t\cdot}}\right], \tag{2.1}$$

where $\boldsymbol{\pi_{t\cdot}} = (\pi_{t1}, \cdots, \pi_{tJ})$, $\sum_{j=1}^{J} \pi_{tj} = 1$ and $t = 1, \cdots, T$.

One reasonable way to think of a categorized variable Y_t is to consider that it has been generated from a continuous latent variable, Z_t, divided into intervals whose bin boundaries are unknown. The categorical variable is classified in category j if, and only if, the continuous variable belongs to the category j, that is

$$Y_t = j \iff \lambda_{j-1} < Z_t \leq \lambda_j, \quad j = 1, \cdots, J,$$

with $\lambda_J = \infty$. Then one can model the cumulative probability that the response variable lies in category j or below it at time t as

$$\gamma_{tj} = Pr(Y_t \leq j) = Pr(Z_t \leq \lambda_j). \tag{2.2}$$

We propose to model categories of rainfall, treating the actual volumes of rain as a latent process Z_t, to which we assign a structure based on Sansó and Guenni (1999a). Assume that Z_t is a transformation of a Gaussian latent variable ζ_t, given by:

$$\begin{aligned} Z_t &= \begin{cases} \zeta_t^{\alpha}, & \zeta_t > 0 \\ 0, & \zeta_t \leq 0. \end{cases}, \\ \zeta_t &= \boldsymbol{F}_t' \boldsymbol{\theta}_t + e_t, \quad e_t \sim N(0, V_t) \\ \boldsymbol{\theta}_t &= \boldsymbol{G}_t \boldsymbol{\theta}_{t-1} + \boldsymbol{\omega}_t, \quad \boldsymbol{\omega}_t \sim N_K(\boldsymbol{0}, \boldsymbol{W}_t), \end{aligned} \tag{2.3}$$

with $\alpha > 0$ and $\boldsymbol{F}_t$ being a vector of regressors, which may include trend and seasonal components, as well as other covariates at time t. The effects of the structural components of $\boldsymbol{F}_t$ are described through $\boldsymbol{\theta}_t$, a vector with K regression coefficients, which may vary through time according to the stochastic dynamic structure described in the bottom line of (2.3). Note that current and

past values of the state parameters $\boldsymbol{\theta}$ are related through a $K \times K$ evolution matrix $\boldsymbol{G}_t$. We assume, in particular, that $V_t = V$ and that $\boldsymbol{G}_t$ is the identity matrix of dimension K, $\forall t$. The structure in (2.3) implies that Z_t is positive and zero inflated.

It is worth noting that the inclusion of the Gaussian latent variable ζ_t implies that the link function that is implicitly assumed in the proposed formulation is a variation of a probit model, since:

$$\gamma_{tj} \quad = \quad Pr(Z_t \leq \lambda_j) = Pr(\zeta_t \leq 0) + Pr(0 < \zeta_t \leq \lambda_j^{1/\alpha}) = \Phi\left(\frac{\lambda_j^{1/\alpha} - \boldsymbol{F}_t^{'}\boldsymbol{\theta}_t}{\sqrt{V}}\right).$$

Hence, $\Phi^{-1}(\gamma_{tj}) = \rho_j - \boldsymbol{F}_t^{'}\boldsymbol{\vartheta}_t$, with $\rho_j = \dfrac{\lambda_j^{1/\alpha}}{\sqrt{V}}$ and $\boldsymbol{\vartheta}_t = \dfrac{\boldsymbol{\theta}_t}{\sqrt{V}}$, with $\Phi(.)$ denoting the cumulative standard normal distribution.

2.1 Inference Procedure

Let $\boldsymbol{y} = (y_1, \cdots, y_T)$ denote realizations of the categorical variable for T instants in time. Note that $\pi_{t1} = \gamma_{t1}$ and $\pi_{tj} = \gamma_{tj} - \gamma_{t,j-1}, j = 2, \ldots, J$, with γ_{tj} given by (2.2). Thus, following the model introduced in (2.1), the likelihood function is given by

$$\begin{aligned} l(\boldsymbol{y}|\boldsymbol{\lambda}, \boldsymbol{\theta}_0, \cdots, \boldsymbol{\theta}_T, V, \alpha) \quad &\propto \quad \prod_{t=1}^{T}\prod_{j=1}^{J} \pi_{tj}^{y_{tj}} = \prod_{t=1}^{T} \left[\Phi\left(u_{t,1}\right)\right]^{y_{t1}} \prod_{j=1}^{J} \left[\Phi\left(u_{t,j}\right) - \Phi\left(u_{t,j-1}\right)\right]^{y_{tj}} \qquad (2.4) \\ &= \quad \prod_{t=1}^{T} \left\{ 1(y_t = 1)\Phi\left(u_{t,1}\right) + \sum_{j=2}^{J} 1(y_t = j)\left[\Phi\left(u_{t,j}\right) - \Phi\left(u_{t,j-1}\right)\right] \right\}, \end{aligned}$$

with $1(.)$ denoting an indicator function and

$$u_{t,j} = \frac{\lambda_j^{1/\alpha} - \boldsymbol{F}_t^{'}\boldsymbol{\theta}_t}{\sqrt{V}}, \quad j = 1, \cdots, J; \quad t = 1, \cdots, T.$$

Examination of (2.4) (De Oliveira, 2000) shows that the model is identifiable for $\alpha \neq 1$, since if $\alpha = 1$ and if the predictor contains an intercept, then the substituition of the parameters $(\boldsymbol{\lambda}, \boldsymbol{\theta}_0, \cdots, \boldsymbol{\theta}_T, V)$ by $(\boldsymbol{\lambda}^*, \boldsymbol{\theta}_0^*, \cdots, \boldsymbol{\theta}_T^*, V^*) = (a\boldsymbol{\lambda} + c, a\boldsymbol{\theta}_0 + c\boldsymbol{e_1}, \cdots, a\boldsymbol{\theta}_T + c\boldsymbol{e_1}, a^2V)$ for any $a > 0$, $c \in \mathcal{R}$ and $\boldsymbol{e_1} = (1, 0, \ldots, 0)$ implies that

$$u_{t,j}^* \quad = \quad \frac{\lambda_j^* - \boldsymbol{F_t}^{'}\boldsymbol{\theta}_t^*}{\sqrt{V^*}} = \frac{a\lambda_j + c - a\theta_{t,1} - c - \sum_{k=2}^{K} aX_{tk}\theta_{t,k}}{\sqrt{a^2V}} = \frac{\lambda_j - \boldsymbol{F_t}^{'}\boldsymbol{\theta}_t}{\sqrt{V}} = u_{t,j},$$

thus resulting in $l(\boldsymbol{y}|\boldsymbol{\lambda}, \boldsymbol{\theta}_0, \cdots, \boldsymbol{\theta}_T, 1, V) = l(\boldsymbol{y}|\boldsymbol{\lambda}^*, \boldsymbol{\theta}_0^*, \cdots, \boldsymbol{\theta}_T^*, 1, V^*)$. If the predictor has no intercept, it still follows that, for $\alpha = 1$, $l(\boldsymbol{y}|\boldsymbol{\lambda}, \boldsymbol{\theta}_0, \cdots, \boldsymbol{\theta}_T, 1, V) = l(\boldsymbol{y}|a\boldsymbol{\lambda}, a\boldsymbol{\theta}_0, \cdots, a\boldsymbol{\theta}_T, 1, a^2V)$.

Therefore, care must be taken when assigning a prior distribution for α. It should assign very low probabilities to values of α close to 1. We return to this below when we discuss the prior distribution of α.

For computational convenience (Albert and Chib, 1993) we parameterize the likelihood in terms of the latent variables $\zeta_t, \cdots, \zeta_T$:

$$l(\boldsymbol{y}|\boldsymbol{\lambda}, \boldsymbol{\zeta}, \alpha) \propto \prod_{t=1}^{T} \left[1(y_{t1} = 1) 1(\zeta_t \leq \lambda_1^{1/\alpha}) + \sum_{j=2}^{J} 1(y_{tj} = 1) 1(\lambda_{j-1}^{1/\alpha} < \zeta_t \leq \lambda_j^{1/\alpha}) \right], \quad (2.5)$$

and hence the parameter vector to be estimated in the proposed model is $\boldsymbol{\psi} = (\boldsymbol{\theta}_0, \ldots, \boldsymbol{\theta}_T, V, \boldsymbol{\zeta}, \alpha, \boldsymbol{\lambda})$, as well as the evolution covariance matrixes $\boldsymbol{W}_t$. In particular, we assume that $\boldsymbol{W}_t$ is a diagonal matrix, implying prior independence among the components of $\boldsymbol{\theta}_t$, $\forall t$. In order to specify $\boldsymbol{W}_t$ we make use of discount factors. The choice of such discounts reflects the rate of adaptation of $\boldsymbol{\theta}_t$ to new incoming data, that is, it implies a graduate decay on the information that observations previous to time t should bring to the estimation of $\boldsymbol{\theta}_t$. For details on the specification of discount factors and the relationship between such discounts and evolution errors' variances, see West and Harrison (1997, pp. 51, 193-202).

The prior specification for the components of the parametric vector $\boldsymbol{\psi}$ is as follows: for $\boldsymbol{\theta}_0$ we assign a multivariate normal distribution with mean vector $\boldsymbol{m}_0$ and covariance matrix $\boldsymbol{C}_0$; for V we assign an inverse gamma distribution with shape a_V and scale b_V; for the exponent α we assign a gamma distribution with shape a_α and rate b_α, with these last two hyperparameters specified in such a way that the prior distribution for α presents low probability mass in the neighborhood of 1, due to the identifiability problem discussed above. Also, as our continuous variable represents rainfall, the positive part of the distribution is typically skewed. The hydrological literature suggests that a reasonable transformation to rainfall data to attain normality is the cubic root. For this reason we assign a prior distribution to α with high mass of probability around 3. Completing the prior specification, we assume that conditioned on λ_{j-1}, λ_j follows a truncated normal distribution with parameters m_{λ_j} and V_{λ_j}, defined in the interval (λ_{j-1}, ∞), for $j = 1, \ldots, J-1$. We assume prior independence among the errors e_t, thus $\zeta_1, \ldots, \zeta_T$ are conditionally independent, *a priori*, given $\boldsymbol{\theta}_t$ and V. Therefore, the joint posterior distribution for the general model, conditional on $\boldsymbol{W}_t$, is given by

$$p(\boldsymbol{\psi}|\boldsymbol{y}, \boldsymbol{W}_t) \propto l(\boldsymbol{y}|\boldsymbol{\lambda}, \boldsymbol{\zeta}, \alpha) p(\boldsymbol{\theta}_0) \prod_{t=1}^{T} \left[p(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1}, \boldsymbol{W}_t) p(\zeta_t|\boldsymbol{\theta}_t, V) \right] p(\alpha) p(V) \prod_{j=1}^{J-1} p(\lambda_j|\lambda_{j-1}).$$

The joint posterior distribution is analytically intractable and we resort to Markov chain Monte Carlo (MCMC) methods to obtain samples from the target distribution. In particular we use a hybrid Gibbs algorithm (Geman and Geman (1984); Gelfand and Smith (1990)) with some Metropolis-Hastings steps (Metropolis et al. (1953); Hastings (1970)). Appendix A provides some details about the resultant full conditional posterior distributions and proposal distributions adopted in the MCMC scheme.

2.2 Predictive Inference

Let D_0 denote the set that summarizes all the information available to a forecaster at time $t = 0$. If the model is closed to external information, the available information at each time t is given by $D_t = \{D_{t-1}, y_t\}$. In most time series applications, one aims to predict future values Y_{T+h}, $h = 1, \cdots, H$, given the information available up to time T, D_T. Let $\boldsymbol{Y}_f = (Y_{T+1}, \ldots, Y_{T+H})'$ be the future values at times $T+1, \cdots, T+H$ and define $\boldsymbol{\psi}_f$ as the collection of parameters required for the likelihood of $\mathbf{Y}_f$. Then the predictive distribution for $\mathbf{Y}_f$, under model M, is given by:

$$\begin{aligned} l(\mathbf{y}_f|D_T, M) &= \int l(\mathbf{y}_f|\boldsymbol{\psi}_f, D_T, M)p(\boldsymbol{\psi}_f|D_T, M)d\boldsymbol{\psi}_f = \int l(\mathbf{y}_f|\boldsymbol{\psi}_f, M)p(\boldsymbol{\psi}_f|D_T, M)d\boldsymbol{\psi}_f \\ &= E_{\boldsymbol{\psi}_f|D_T,M}[l(\mathbf{y}_f|\boldsymbol{\psi}_f, M)], \end{aligned} \tag{2.6}$$

with $p(\boldsymbol{\psi}_f|D_T, M)$ obtained by updating $p(\boldsymbol{\psi}|D_T, M)$ through the evolution equation in the bottom line of (2.3) and $l(\mathbf{y}_f|\boldsymbol{\psi}_f, M) = \prod_{h=1}^{H} l(y_{T+h}|\boldsymbol{\psi}_f, M)$. When looked at as a function of the model M, (2.6) gives the predictive likelihood for model M, which may be used as a criterion for model selection, see e.g. Alves et al. (2010).

Let $\boldsymbol{\psi}_m$ be the set of the parameters needed to describe the predictive likelihood of the model m and suppose that a Monte Carlo sample of size N of $p(\boldsymbol{\psi}|\boldsymbol{Y}, \boldsymbol{W})$ is available. Then the construction of a sample of $p(\boldsymbol{\psi}_m|M = m, D_T)$ follows directly and a Monte Carlo estimate for the predictive likelihood in (2.6) is given by

$$\hat{E}_{\boldsymbol{\psi}_m|M=m,D_T}\left[l(\boldsymbol{y}_f \mid \boldsymbol{\psi}_m, M = m, D_T)\right] = \frac{1}{N}\sum_{i=1}^{N}\prod_{h=1}^{H} l(y_{T+h}|\boldsymbol{\psi}_m^{(i)}, M = m, D_T) \tag{2.7}$$

When selecting among a set of proposed models based on predictive likelihoods, the specification that provides maximum value for (2.7) should be the chosen one.

3 Data analysis

In order to verify if the proposed model is able to recover the actual structure that generated the data, when that structure is known, artificial data sets were generated following (2.3). This simulation exercise is summarized in subsection 3.1. Next we fit our proposed model to a real data set, aiming at predicting categorized volumes of rainfall. We also compare the performance of the prediction under the categorized formulation with a fitting to actual volumes of rainfall, which we call continuous formulation.

3.1 Artificial data

Based on equation (2.3) we generated $L = 25$ samples, each of length $T = 169$, with $J = 4$ categories and used $K = 2$ covariates, such that $\mathbf{F}_t' = (x_{1t}, x_{2t})$, where x_1 and x_2 are the same covariates used in the analysis of the real data in section 3.2. After fixing $\boldsymbol{\theta}_0 = (-1.3, 4.0)$, $W = 0.0001$, $V = 0.1$, and $\alpha = 2$, we generated the true values for θ_{1t}, θ_{2t}, and ζ_t. The true bin boundaries were fixed at $\lambda_1 = 0.5$, $\lambda_2 = 7.5$, $\lambda_3 = 15$. Once these values were established, we obtained the observed values y_t as follows: if $z_t \in [0.0, 0.5)$ then $y_t = (1, 0, 0, 0)$, else if $z_t \in [0.5, 7.5)$ then $y_t = (0, 1, 0, 0)$, else if $z_t \in [7.5, 15.0)$ then $y_t = (0, 0, 1, 0)$, else if $z_t \in [15.0, \infty)$ then $y_t = (0, 0, 0, 1)$.

For each of the $L = 25$ samples we fitted the same model used to generate the data, and assigned the following prior distributions: for θ_{01} and θ_{02}, independent, zero mean normal distributions, each with variance 10, for V we assigned an inverse gamma distribution with infinite variance and mean equal to 0.1, whereas for α, a gamma prior distribution was asssigned with shape parameter equal to 9 and rate equal to 3. The variances of the evolution equation of the parameters of the covariates, $\mathbf{W} = diag(W_1, W_2)$, were estimated using discounting factors, and these were fixed at 0.98.

We explored three different prior specifications for the bin boundaries λ_js, $j = 1, 2, 3$. All of them assume normal distributions for λ_j, truncated on λ_{j-1}, and the parameters are as shown in Table 1. Prior distributions I and II assume the mean of the associated normal distribution for each λ_j equal to the values used to generate the data. Prior I is more concentrated around the true values used to generate the data, than prior II. On the other hand, for prior III the mean of the associated normal distributions are fixed at values greater than the ones used to generate the data, and with variance fixed at a reasonably high value.

Table 1: Mean (m_λ) and variance (V_λ) of the associated normal distributions for the bin boundaries, λ_1, λ_2, and λ_3, of the simulation study.

Prior	m_{λ_1}	m_{λ_2}	m_{λ_3}	$V_\lambda = V_{\lambda_j} \forall j = 1, 2, 3$
I	0.5	7.5	15.0	5.0
II	0.5	7.5	15.0	10.0
III	1.0	9.0	25.0	10.0

For each sample, and prior specification, we let the MCMC run for 900,000 iterations, considered the first 10,000 as burn in, and stored every 800th iteration to avoid autocorrelation among the sampled values. Convergence of the chains was checked through trace plots.

Clearly, the posterior distribution of the bin boundaries λ_js are sensitive to their prior specification. When comparing the posterior distributions obtained under priors I and II, except of λ_1, prior II provided wider ranges of the 95% the posterior credible intervals for λ_2 and λ_3 (1st and 2nd columns of Figure 1). Apparently, the posterior distribution of λ_1 is not highly affected by its prior specification, as prior III provided very similar summaries for λ_1 when compared to priors I and II. However, as we must assume $\lambda_3 > \lambda_2$, we notice that as we increase the prior mean, we tend to overestimate the true values of λ_2 and λ_3 (3rd column of Figure 1).

The variance of the observation equation, V, seems not to be sensitive to the prior specification of the λ_js (1st row of Figure 2). On the other hand, the posterior distributions obtained under prior III tend to slightly overestimate the true value of the power transformation α.

3.2 Analyzing daily categories of rainfall in Rio de Janeiro

In this subsection, two approaches are compared, both aiming to model rainfall data. In the first one it is assumed that the available information is on categorized rainfall occurrence and that the actual amount of rain is unknown, being treated as a latent process, as described in section 2. In the second approach, we follow Sansó and Guenni (1999a), and model volumes of rainfall, then we compare the resultant predictions under both approaches.

The analyzed data were made available by the Ministry of Agriculture, Livestock and Supply, National Institute of Meteorology - INMET, Brazil, and comprises daily ground observations on volumes of rainfall. We have also available daily records on average wind speed, average humidity

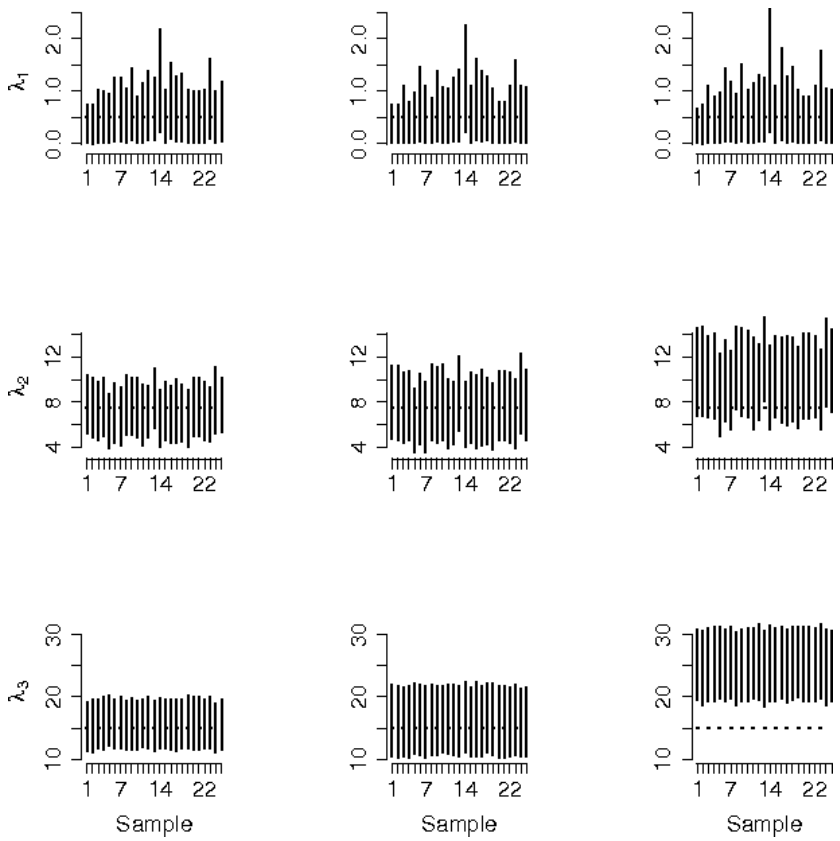

Figure 1: Each panel shows the 95% posterior sequence of credible intervals, based on each of the $L = 25$ samples, of λ_1, λ_2, and λ_3 (rows) under each of the prior specifications of Table 1 (columns). In all panels, the horizontal dotted line represents the true value of the respective λ_j.

and average temperature, but preliminary analyzes showed no significant effect of average wind speed on rainfall. We fitted the models to the period ranging from September 22, 2005 to March 19, 2006, comprising 179 observations. We held out the last $H = 10$ observations for model selection and predictive purposes, such that for the inference we had $T = 169$ observations.

Although the used data set provides information on volumes of rainfall, here we classify the observed rain occurrences in six categories, each representing different levels of rainfall. We referred to Dias and Espinosa (personal communication, 2008), who proposed five bin boundaries to rain volumes (in mm) during the Spring/ Summer months in Rio de Janeiro (0.5, 7.5, 15.0, 22.5 and 30.0). Rain volumes were divided into categories according to these bin boundaries and the resultant categorized variable is the one considered in the likelihood function in (2.5). We fitted models with discount factors for the the evolution errors' variances W_t equal to 0.95 and

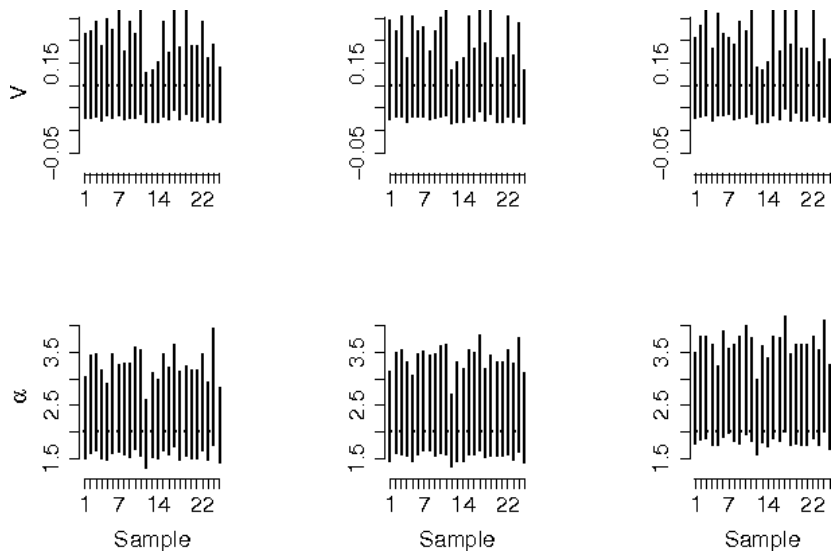

Figure 2: Each panel shows the 95% posterior credible intervals, based on each of the $L = 25$ samples, of V and α (rows) under each of the prior specifications of Table 1 (columns).

0.99. As the results were not sensitive to this choice we show the results based on 0.99. The parameters of the prior distributions for α, V and $\boldsymbol{\theta}_0$ were, respectively: $a_\alpha = 9$, $b_\alpha = 3$, $a_V = 2$, $b_V = 2$, $\boldsymbol{m}_0 = 0$, $\boldsymbol{C}_0 = 10$. In the same spirit of the simulated exercise, a sensitivity analysis has been performed to evaluate the impact of the bin boundaries' prior specification on the joint posterior distribution of $\boldsymbol{\psi}$. A grid of values was specified for the parameters of the truncated normal prior distributions for λ_j, $j = 1, \ldots, 5$, according to Table 2.

Table 2: Parameters of the prior distribution specifications for the bin boundaries λ_j, $j = 1, 2, 3, 4, 5$ for the rainfall dataset.

Set	μ_{λ_1}	μ_{λ_2}	μ_{λ_3}	μ_{λ_4}	μ_{λ_5}	$V_\lambda = V_{\lambda_j} \forall j$
1	0.5	7.5	15.0	22.5	30.0	5.0
2	0.5	7.5	15.0	22.5	30.0	10.0
3	1.0	8.0	20.0	30.0	40.0	10.0
4	0.7	6.0	12.0	20.0	35.0	10.0
5	0.7	15.0	25.0	30.0	60.0	10.0

We ran the MCMC for the different models for 900,000 iterations, considered the first 100,000 as burn in and stored every 800th iteration. The summary of the posterior distribution of the

bin boundaries, under the five different prior specifications are depicted in Figure 3. When the first and second specifications are compared, it is clear that changes in the prior variance did not substantially affect the point estimates, with the second specification providing slightly wider credibility ranges, as expected. Except for λ_1, the estimation of the remaining bin boundaries was indeed sensitive to different prior mean specification, as shown by the comparison between priors 1 and 2 and the remaining ones. The predictive likelihood estimates originated from the

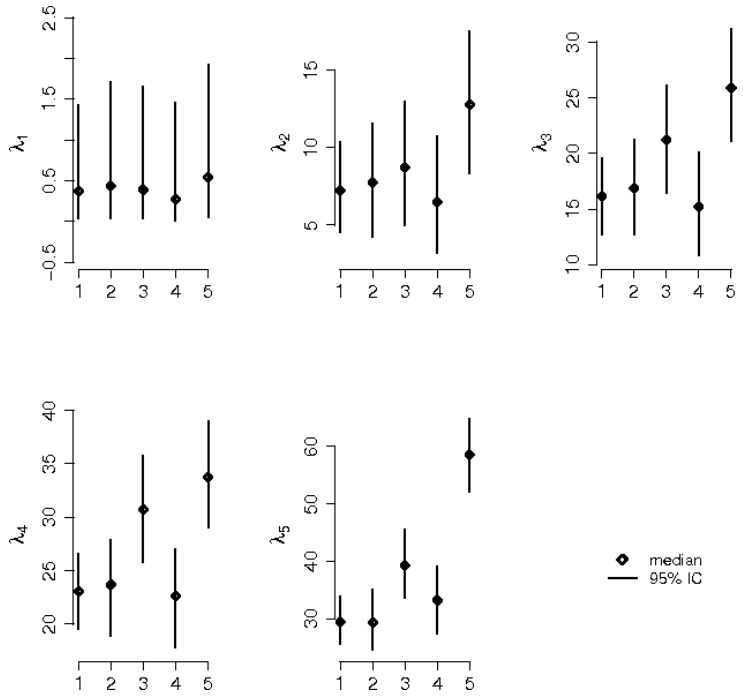

Figure 3: Summary of the posterior MCMC samples of the bin boundaries under the five different prior specifications.

five specifications above are registered in Table 3, which shows that the second prior specification provides the best predictive result, followed very closely by the third specification. The fifth prior, centered on values which are in dissonance with the experts' information, produces the worst results among the proposed specifications.

Table 3: Model comparison under the predictive likelihood estimates considering different prior specifications for the bin boundaries.

Prior 1	Prior 2	Prior 3	Prior 4	Prior 5
0.000742	0.000773	0.000772	0.000703	0.000635

In the remaining of this section, except if otherwise stated, the posterior and predictive results associated to the categorized formulation refer to the model fitted with the second prior specification. Figure 4 displays the estimated time series of precipitation (in the log scale for easy of visualization), as well as the actual observed precipitation volumes. It is clear that the latent variable Z_t captures the observed time series and its general behavior pretty well.

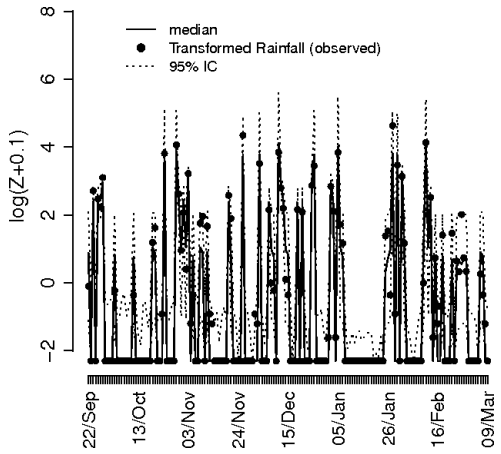

Figure 4: Observed time series (solid circles) of precipitation - in the log scale- and its estimates (solid line), according to the best proposed model under the categorized formulation, together with the 95% posterior credible interval (dotted lines).

A comparison with an analysis using the actual observed volumes of rainfall

We now compare the results of the categorized formulation to a continuous formulation, as proposed by Sansó and Guenni (1999a). Basically, following equation (2.3), we write down a likelihood function for the daily volumes of rainfall, in this case Z_t denotes the actual observed amount of rain. Thus it is only necessary to estimate the variance of the latent variable, $\boldsymbol{\zeta}$, the regression coefficients $\boldsymbol{\theta}$ and the exponent α.

Figure 5 shows the regression coefficients, which are positive for humidity and negative for temperature, under both models. The estimated coefficients exhibit quite similar temporal trajectories, regardless of the adopted approach. Figure 6 shows the summary of the posterior distribution of the variance V and the exponent α, for the five different prior specifications under the categorized formulation, as well as under the continuous formulation. When the best categorical specification (based on prior 2) is compared to the continuous formulation, it is clear that the

observational variance V concentrates on smaller values under the categorical approach, while the exponent α concentrates on smaller values under the continuous approach. The uncertainty associated to the estimation of α is significantly smaller under the continuous approach, as reflected by the 95% credibility intervals. This is probably related to the fact that more information is available when we fit the model using the actual volumes of rainfall.

Through the forecasted amount of rainfall obtained by the continuous formulation and using the bin boundaries suggested by Dias and Espinosa, we can obtain forecasts for future categorized rain categories based on the continuous model. The top line of Figure 7 exhibits the predictive median for the categorized response variable, as well as the actual observed category at each time, under the best categorized model and the continuous formulation (2nd column). Both models exhibit very similar patterns in their respective predictions. The second line of Figure 7 shows posterior and predictive medians, and their respective 95% credible limits, for the latent variables ζ_{T+h}^{α}, as well as, the actual observed rainfall volumes. The predictive point estimates provided by the continuous formulation are closer to the actual observed rain volumes, just as expected, since the volume of rainfall was assumed known in that approach. It is worth noting that, although the observed volumes were unknown under the categorized formulation, the estimation procedure under that approach was able to recover the general pattern of the continuous response, with the actual observed amounts of rainfall falling within the limits of the 95% posterior credible interval.

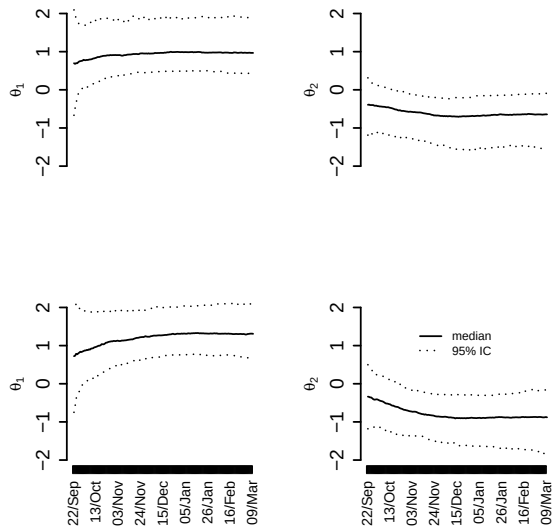

Figure 5: Posterior summary of the evolution of the regression coefficients for humidity (θ_1) and temperature (θ_2), under the categorized model (first row) and the continuous formulation (second row).

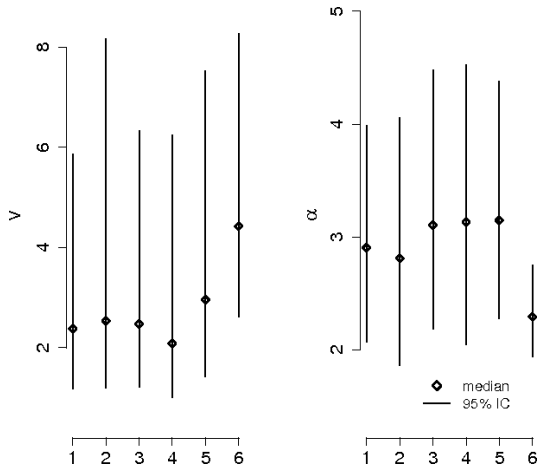

Figure 6: Summary of the posterior distribution for the variance of the evolution errors, V, and for the exponent α. In each panel, vertical lines 1 to 5 represent 95% posterior credible intervals obtained under the five different prior specifications for the categorized model and vertical line 6 represents the result under the continuous formulation.

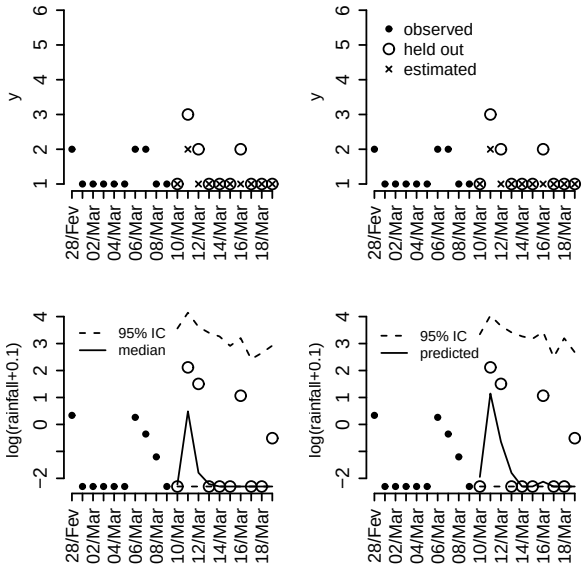

Figure 7: First row: Median of the posterior predictive distribution for the last 10 observations held out from the inference procedure. The symbol $\times$ represents the posterior median of the prediction. Second row: summary of the posterior (left panel) and posterior predictive (right panel) distributions and respective limits of the 95% posterior credibile intervals, under the categorized (left panel) and the continuous (right panel) formulations, for the volumes of rainfall (in the log scale). Solid lines are the median and dashed lines are the limits of the credible intervals. In all panels, the solid circles represent the last 10 observations which were used in the inference procedure, and the hollow circle is the actual observed category (first row) or the observed volume (second row).

4 Discussion

We proposed a model for polycothomous data that vary across time. More specifically, we concentrated on the problem of modelling observed categories of rainfall. We extended the work of Albert and Chib (1993) assuming that the underlying continuous variable follows the model proposed by Sansó and Guenni (1999a). Different from Albert and Chib (1993), we assumed the bin boundaries, that connect the categorical variable to the (latent) continuous one, as parameters to be estimated.

In section 2 we showed that we must impose some restrictions to the proposed model in order to be able to obtain estimates of the parameters of interest. Analysis of artificial data suggest that we are able to recover the true values of the parameters. The analysis of daily measurements of rainfall in Rio de Janeiro suggest that the categorized approach is able to recover reasonably well the underlying true process, when compared to the model that makes use of the actual volumes of rainfall (Section 3.2).

Following the suggestion of Dias and Espinosa (Private Communication, 2008) we assumed the bin boundaries fixed across time because we had only Spring/Summer observations. If a longer time series, covering different seasons of the year is investigated, we suggest to change the prior distribution of the λs accordingly. In this case, the MCMC described in the appendix has to be accommodated as different bin boundaries will be used for different instants in time. The proposed model might be used as the top layer of a hierarchical model which accounts for different sources of information on rainfall, e.g. ground-based measurements, remote sense, physical models, etc. Combining the information from these different sources is challenging and is a current subject of research.

Acknowledgements

Most of this work was developed while P. L. Velozo was a M.Sc. student at IM-UFRJ. Velozo is grateful to CAPES for the financial support during her M.Sc. studies. Schmidt was partially supported by CNPq and FAPERJ. The authors are grateful to Pedro L. S. Dias (IAG-USP/LNCC) and America M. Espinosa (IAG-USP) for fruitful discussions about rainfall modelling.

A Full conditional posterior distributions

In what follows, the full conditional posterior distributions based on the likelihood function in(2.5), which makes use of the latent variables ζ, are described. Let $\boldsymbol{\psi} = (V, \boldsymbol{\zeta}, \boldsymbol{\lambda}, \boldsymbol{\theta}, \alpha)$ and $\boldsymbol{\psi}_{-\beta}$ be the vector $\boldsymbol{\psi}$, except for a component β.

Full conditional distribution of the bin boundaries $\lambda_1, \cdots, \lambda_J$ The full conditional posterior distribution of $\boldsymbol{\lambda}$ is given by

$$p(\boldsymbol{\lambda}|\boldsymbol{\psi}_{-\boldsymbol{\lambda}}, \boldsymbol{W}, \boldsymbol{y}) \propto \exp\left\{-\frac{1}{2}\sum_{j=1}^{J-1}\left(\frac{\lambda_j - m_{\lambda_j}}{\sqrt{V_{\lambda_j}}}\right)^2\right\}\prod_{j=1}^{J-2}\left\{\left[1-\Phi\left(\frac{\lambda_j - m_{\lambda_{j+1}}}{\sqrt{V_{\lambda_{j+1}}}}\right)\right]^{-1}\right\} \times$$
$$\prod_{j=1}^{J-1}\left[1\left(\max\left\{\max\left\{Z_t : Y_t = j\right\}, \lambda_{j-1}\right\} < \lambda_j < \min\left\{\min\left\{Z_t : Y_t = j+1\right\}, \lambda_{j+1}\right\}\right)\right].$$

This distribution is analytically intractable, hence we use Metropolis-Hastings steps to obtain samples from it. A product of truncated normal distributions, each one centered on the current value of each cut point, is adopted as proposal density for this step, so that $q(\boldsymbol{\lambda}^p|\boldsymbol{\lambda}^c) = q_1(\lambda_1^p|\boldsymbol{\lambda}^c)\prod_{j=2}^{J-1} q_j(\lambda_j^p|\lambda_{j-1}^p, \boldsymbol{\lambda}^c)$, with

$$q_j(\lambda_j^p|\lambda_{j-1}^p, \boldsymbol{\lambda}^c) = \frac{\frac{1}{\sqrt{2\pi V_l}}\exp\left\{-\frac{1}{2}\left(\frac{\lambda_j^p - \lambda_j^c}{\sqrt{V_l}}\right)^2\right\}}{\Phi\left(\frac{\min\{Z_t : Y_t = j+1\} - \lambda_j^c}{\sqrt{V_l}}\right) - \Phi\left(\frac{\max\{\max\{Z_t : Y_t = j\}, \lambda_{j-1}^p\} - \lambda_j^c}{\sqrt{V_l}}\right)}.$$

The supports of the densities $q_1, \ldots, q_{J-1}$ are given by:

$$\max\{\max\{Z_t : Y_t = 1\}, 0\} < \lambda_1^p < \min\{Z_t : Y_t = 2\},$$
$$\max\{\max\{Z_t : Y_t = j\}, \lambda_{j-1}^p\} < \lambda_j^p < \min\{Z_t : Y_t = j+1\}, \quad j = 2, \ldots, J-1$$

and V_l is tuned to provide reasonable acceptance rates.

Full conditional distribution of the exponent α The full conditional posterior distribution of α, $p(\alpha|\boldsymbol{\psi}_{-\alpha}, \boldsymbol{W}, \boldsymbol{y})$, is proportional to $p(\alpha)l(\boldsymbol{y}|\boldsymbol{\psi})$. The domain of α is constrained because of the likelihood function in (2.5). As showed in the following lines, this parameter lies in an interval given by $\alpha \in (\max\{\boldsymbol{a}\}, \min\{\boldsymbol{b}\})$, whith $\boldsymbol{a} = (a_1, \ldots, a_{Na})$, $\boldsymbol{b} = (b_1, \ldots, b_{Nb})$ and $Na, Nb \leq T$. To determine $\boldsymbol{a}$ and $\boldsymbol{b}$, it is necessary to analyze the category to which the response variable belongs and the value of $\boldsymbol{\zeta}$. Let $q = 1, \ldots, Na$ and $s = 1, \ldots, Nb$.

If the response variable belongs to category 1 at time t, then

$$Y_{t1} = 1 \quad \Leftrightarrow \quad 0 \quad < \zeta_t^{\alpha} \quad \leq \lambda_1 \quad \text{or} \quad \zeta_t \quad \leq 0.$$

Note that if $\zeta_t \leq 0$ or $\zeta_t = 1$, the restriction does not depend on α. If $\zeta_t > 0$, then $0 < \zeta_t^{\alpha} \leq \lambda_1$, implying that $-\infty < \alpha \log(\zeta_t) \leq \log(\lambda_1)$. This last inequality implies that

$$\alpha \geq \frac{\log(\lambda_1)}{\log(\zeta_t)} = a_q, \quad \text{for } 0 < \zeta_t < 1,$$
$$\alpha \leq \frac{\log(\lambda_1)}{\log(\zeta_t)} = b_s, \quad \text{for } \zeta_t > 1.$$

If the response variable belongs to category $j > 1$ at time t, then

$$Y_{tj} = 1 \Leftrightarrow \lambda_{j-1} < \zeta_t^{\alpha} \leq \lambda_j \Leftrightarrow \log(\lambda_{j-1}) < \alpha \log(\zeta_t) \leq \log(\lambda_j).$$

Once again, notice that:

$$a_q = \frac{\log(\lambda_j)}{\log(\zeta_t)} < \alpha \leq \frac{\log(\lambda_{j-1})}{\log(\zeta_t)} = b_s, \quad \text{for } 0 < \zeta_t < 1,$$
$$a_q = \frac{\log(\lambda_{j-1})}{\log(\zeta_t)} < \alpha \leq \frac{\log(\lambda_j)}{\log(\zeta_t)} = b_s, \quad \text{for } \zeta_t > 1.$$

Note that $\max\{\boldsymbol{a}\} < \min\{\boldsymbol{b}\}$, $\forall \boldsymbol{\zeta}, \boldsymbol{\lambda}$, because the bin boundaries, $\lambda_1, \ldots, \lambda_j$, are ordered and also because, for each pair of instants $t_1, t_2 \in \{1, 2, \ldots, T\}$, with $t_1 \neq t_2$, it is true that, if $j_1 < j_2$ and $Y_{t_1} = j_1$, $Y_{t_2} = j_2$, it follows that $\zeta_{t_1} < \zeta_{t_2}$. Then, the full conditional posterior distribution of α is given by

$$p(\alpha|\boldsymbol{\psi}_{-\alpha}, \boldsymbol{W}, \boldsymbol{y}) \propto G(a_\alpha, b_\alpha) 1(\alpha \in (\max\{0, \boldsymbol{a}\}, \min\{\boldsymbol{b}\})).$$

Full conditional distribution of the latent variable ζ_t The full conditional posterior distribution of ζ_t, $t = 1, \ldots, T$, is given by $p(\zeta_t|\boldsymbol{\psi}_{-\zeta_t}, \boldsymbol{W}, \boldsymbol{y}) \propto l(\boldsymbol{y}|\boldsymbol{\psi}) p(\zeta_t|\boldsymbol{\theta}_t, V)$. Therefore, $p(\zeta_t|\boldsymbol{\psi}_{-\zeta_t}, \boldsymbol{W}, y_t = j) \propto N(\boldsymbol{F}_t'\boldsymbol{\theta}_t, V) \left[1(\zeta_t \leq \lambda_1^{1/\alpha}) 1(Y_{t1} = 1) + \sum_{j=2}^{J} 1(\lambda_{j-1}^{1/\alpha} < \zeta_t \leq \lambda_j^{1/\alpha}) 1(Y_{tj} = 1)\right]$.

Full conditional distribution of the variance V The full conditional posterior distribution of V is an inverse gamma distribution defined by

$$(V|\boldsymbol{\psi}_{-V}, \boldsymbol{W}, \boldsymbol{Y}) \sim IG\left(a_V + \frac{T}{2}, b_V + \frac{1}{2}\sum_{t=1}^{T}\left(\zeta_t - \boldsymbol{F}_t'\boldsymbol{\theta}_t\right)^2\right).$$

Full conditional distribution of the regression coefficients $\boldsymbol{\theta}$ As $(\zeta_t|\boldsymbol{\theta}_t, V) \sim N(\boldsymbol{F}_t'\boldsymbol{\theta}_t, V)$, $(\boldsymbol{\theta}_t|\boldsymbol{\theta_{t-1}}, \boldsymbol{W}) \sim N(\boldsymbol{\theta_{t-1}}, \boldsymbol{W})$ and $p(\boldsymbol{\theta}|\boldsymbol{\psi}_{-\boldsymbol{\theta}}, \boldsymbol{W}, \boldsymbol{Y}) \propto \prod_{t=1}^{T} \{p(\zeta_t|\boldsymbol{\theta}_t, V)\}\, p(\boldsymbol{\theta}|\boldsymbol{W})$, we can use the FFBS algorithm. The variance $\boldsymbol{W}$ is assumed known through discount factors.

References

Agresti, A. (1990) *Categorical Data Analysis*. Wiley Series in Probability and Mathematical Statistics.

Albert, J. H. and Chib, S. (1993) Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association*, **88**, 669–679.

Alves, M. B., Gamerman, D. and Ferreira, M. A. R. (2010) Transfer functions in dynamic generalized linear models. *Statistical Modelling*, **10**, 3–40.

Berret, C. and Calder, C. A. (2010) Data augmentation strategies for the bayesian spatial probit regression model. *Tech. Rep. 841*, Dept. of Statistics, The Ohio State University, USA.

Cargnoni, C., Müller, P. and West, M. (1997) Bayesian forecasting of multinomial time series through conditional Gaussian dynamic models. *Journal of the American Statistical Association*, **92**, 640–647.

Carlin, B. P. and Polson, N. G. (1992) Monte carlo Bayesian methods for discrete regression models and categorical time series. In *Bayesian Statistics 4*, 577–586. Oxford University Press, Oxford, UK. Bernardo, J.M., Berger, J.O., Dawid, A.P. and Smith, A.F.M. (eds).

Chen, M.-H. and Dey, D. K. (2000) A unified Bayesian approach for analysing correlated ordinal response data. *Brazilian Journal of probability and Statistics*, **14**, 87–111.

Congdon, P. (2005) *Bayesian Models for Categorical Data*. Wiley Series in Probability and Statistics.

De Oliveira, V. (2000) Bayesian prediction of clipped Gaussian random fields. *Computational Statistics & Data Analysis*, **34**, 299 – 314.

— (2004) A simple model for spatial rainfall fields. *Stochastic Environmental Research*, **18**, 131 – 140.

Fernandes, M. V. M., Schmidt, A. M. and Migon, H. S. (2009) Modelling zero-inflated spatio-temporal processes. *Statistical Modelling*, **9**, 3–25.

Fuentes, M., Reich, B. J. and Lee, G. (2008) Spatial-temporal mesoscale modelling of rainfall intensity using gage and radar data. *Annals of Applied Statistics*, **4**, 1148–1169.

Gelfand, A. E. and Smith, A. F. M. (1990) Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398–409.

Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

Hastings, W. K. (1970) Monte carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

Higgs, M. D. and Hoeting, J. A. (2010) A clipped latent variable model for spatially correlated ordered categorical data. *Comput. Stat. Data Anal.*, **54**, 1999–2011.

Hughes, J. P., Guttorp, P. and Charles, S. P. (1999) A non-homogeneous hidden Markov model for precipitation ocurrence. *Applied Statistics*, **48**, 15–30.

Knorr-Held, L. (1995) Dynamic cumulative probit models for ordinal panel-data; a Bayesian analysis by Gibbs sampling. *Tech. Rep. 386*, Ludwig-Maximilians-Universität, Munich, Germany.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953) Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087–1092.

Sansó, B. and Guenni, L. (1999a) A stochastic model for tropical rainfall at a single location. *Journal of Hydrology*, **214**, 64–73.

— (1999b) Venezuelan rainfall data analysed by using a Bayesian space-time model. *Applied Statistics*, **48**, 345–362.

Stid, C. K. (1973) Estimating the precipitation climate. *Water Resour. Res.*, **9**, 1235–2141.

West, M. and Harrison, J. (1997) *Bayesian Forecasting and Dynamic Models*. Springer Series in Statistics, second edn.