INPE – National Institute for Space Research
São José dos Campos – SP – Brazil – May 16-20, 2016

CORRELATED TIME SERIES USING MIXED MODELS IN A BAYESIAN PERSPECTIVE

Roberto Molina de Souza[1], *Jorge Alberto Achcar*[2], *Glaucia Maria Bressan*[3]

[1]Federal University of Technology of Paraná, Cornélio Procópio, Brazil, rmolinasouza@utfpr.edu.br
[2]University of São Paulo, São Paulo, Brazil, achcar@fmrp.edu.br
[3]Federal University of Technology of Paraná, Cornélio Procópio, Brazil, rmolinasouza@utfpr.edu.br

Abstract: The goal of this work is to propose the application of an autoregressive model considering two data series: hospital internment and inhalable particulate material with aerodynamic diameter less than 10 μm in a city in the state of São Paulo (oct/2003 to dez/2007). In a Bayesian perspective, using MCMC methods, we consider a mixed model with random effects to capture the possible correlation between the series and residuals with a Student-t distribution. This model is more appropriate when compared to independent models with normal residuals.

keywords: Time series correlated, Bayesian Inference, Student-t residuals, hospital internment, MCMC methods.

1. INTRODUCTION

In practical problems, analysis of two time series can be conducted considering a possible correlation between them. In such cases it is common the use of bivariate distributions [1]. A simpler alternative is the use of a common random effect [2] to the both series.

Under a Bayesian perspective, using MCMC (Markov chain Monte Carlo) methods, this alternative can be applied in a simple way, while the random effect is generated directly in a hierarchical model [3].

In the next section, we present the model assuming random effects associated to the two time series. In an application, we consider two data series related to hospital internment and inhalable particulate material with aerodynamic diameter less than 10 μm in a city in the state of São Paulo (oct/2003 to dez/2007). In the last section we present conclusions and perspectives.

2. METHODOLOGY

Consider two data series, Y_1 and Y_2, measured in the same period and possibly correlated. Assuming that each serie can be represented by a p order autoregressive model, denoted by $AR(p)$, with the introduction of a latent variable, w_t (random effect), capturing the possible correlation we have:

$$\begin{aligned} Y_{1t} &= \alpha_0 + \alpha_1 Y_{1t-1} + \cdots + \alpha_p Y_{1t-p} + w_t + e_{1t} \quad (1) \\ Y_{2t} &= \beta_0 + \beta_1 Y_{2t-1} + \cdots + \beta_p Y_{2t-p} + w_t + e_{2t} \end{aligned}$$

where $\alpha_0, \alpha_1, \cdots, \alpha_p$ and $\beta_0, \beta_1, \cdots, \beta_p$ are fixed effects and w_t has a normal distribution with mean zero and variance σ_w^2. If σ_w^2 is close to 0, the randon effect capture a strong dependence; e_{1t} and e_{2t} are random errors associated to each model assuming Student-t distribution with mean zero, variance σ_1^2 and σ_2^2, respectively, and v_1 and v_2 degrees of freedom of the Student-t distribution ($v_1, v_2 > 2$).

Under a Bayesian perspective, the model (1) can be introduced using a normal distribution [4], that is:

$$\begin{aligned} \mathbf{Y}_1 \quad &| \quad \boldsymbol{\alpha}, \sigma_1^2, \zeta_{1t} \sim N\left(\mu_{1t}, \frac{\sigma_1^2}{\zeta_{1t}}\right) \quad (2) \\ \mathbf{Y}_2 \quad &| \quad \boldsymbol{\beta}, \sigma_2^2, \zeta_{2t} \sim N\left(\mu_{2t}, \frac{\sigma_1^2}{\zeta_{2t}}\right) \end{aligned}$$

where

$$\begin{aligned} \mu_{1t} &= \alpha_0 + \alpha_1 X_{t-1} + \cdots + \alpha_p X_{t-p} + w_t \\ \mu_{2t} &= \beta_0 + \beta_1 Y_{t-1} + \cdots + \beta_p Y_{t-p} + w_t \end{aligned}$$

and ζ_{1t} and ζ_{2t} are the weights relative to each subject ($t = 1, ..., n$) assuming gamma distributions:

$$\zeta_{1t} \mid \boldsymbol{\alpha}, \sigma_1^2, v_1 \sim G\left(\frac{v_1}{2}, \frac{v_1}{2}\right)$$
$$\zeta_{2t} \mid \boldsymbol{\beta}, \sigma_2^2, v_2 \sim G\left(\frac{v_2}{2}, \frac{v_2}{2}\right)$$

with mean 1 and variance $\frac{2}{v_1}$ and $\frac{2}{v_2}$. Observe that when $v_1 \to \infty$ and $v_2 \to \infty$ the variance of ζ_{1t} and ζ_{2t} tends to be too small, leading the weights to be close to 1, converging towards a normal distribution. The advantage of using these expresssions under a Bayesian approach is that ζ_{1t} and ζ_{2t} can be generated directly from a gamma distribution.

3. APLICATION

For the application of the proposed methodology, we consider two data series about hospital internment and inhalable particulate material with aerodynamic diameter less than 10 μm in a city in the state of São Paulo (oct/2003 to dez/2007). In the Figure 1, we observe positive correlation between hospital internment and inhalable particulate material.

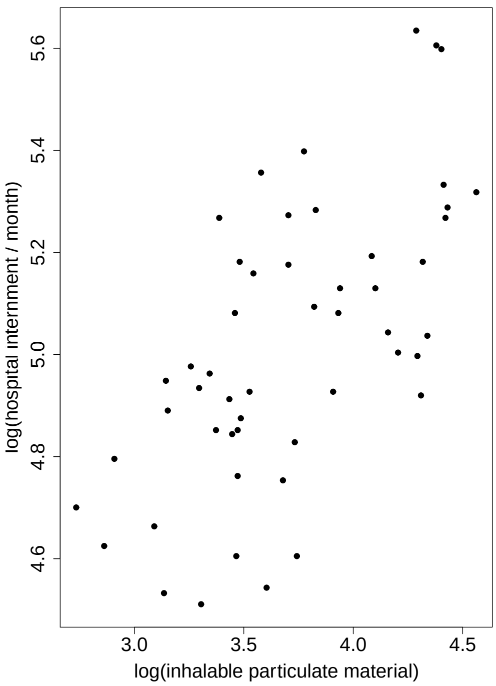

Figure 1 – Scatterplot to hospital internment and inhalable particulate material

3.1. *Prior elicitation*

Assuming $p = 1$ in (2), we consider the prior distribuions given by:

$$\alpha_0 \sim N(a_{\alpha_0}, b^2_{\alpha_0}); \alpha_1 \sim N(a_{\alpha_1}, b^2_{\alpha_1})$$
$$\beta_0 \sim N(a_{\beta_0}, b^2_{\beta_0}); \beta_1 \sim N(a_{\beta_1}, b^2_{\beta_1})$$
$$\sigma_1^{-2} \sim G(c_1, d_1); \sigma_2^{-2} \sim G(c_2, d_2)$$
$$w_t \sim N\left(0, \sigma_w^2\right)$$

Observe that we are assuming a simplified form of model (2), that is a AR(1) model. For the second stage, we have $\sigma_w^{-2} \sim G(e, f)$.

The choice of the degrees of freedom (v_1 and v_2) was based on [5], where the use of $v = 4$ degrees of freedom is a good choice in several applications as suggested in the literature. (see also [6]).

The values of the hyper-parameter were chosen to lead to an approximately non-informative prior distribution, resulting in similar estimates to the ones obtained through MLE (Maximum Likelihood Estimation). In this way, we have: $a_{\alpha_0} = a_{\alpha_1} = a_{\beta_0} = a_{\beta_1} = 0$; $b^2_{\alpha_0} = b^2_{\alpha_1} = b^2_{\beta_0} = b^2_{\beta_1} = 10^2$ and $c_1 = d_1 = c_2 = d_2 = e = f = 0.1$.

For all models, $100,000$ Gibbs samples were simulated for each parameter using the OpenBUGS software [7], from the R software (R2OpenBUGS library). From the total number of generated samples, $50,000$ were discarded (burn-in-sample) to avoid the possible influence of initial values until the chains are stable. From the $50,000$ remaining samples for each parameter, values of 50 were taken to ensure non-correlated samples. Convergence of the Gibbs sampling algorithm was observed from traceplots of the simulated Gibbs samples.

3.2. *Results*

In Table 1, we introduce the estimated parameters and DIC (Deviance information criterion) value for the models. DIC [8] is a famous discrimination criterion used to compare models. When two values of DIC are compared, smaller values of DIC imply better model.

Table 1 – Mean (standard deviation) for the estimated parameters and DIC value for the models.

Model	$(v = 4)$	$(v \to \infty)$
Parameters	Student-t	(Random Effect)
α_0	3.14 (0.69)	2.84 (0.64)
α_1	0.37 (0.14)	0.43 (0.13)
β_0	1.09 (0.38)	1.21 (0.37)
β_1	0.70 (0.10)	0.67 (0.10)
σ_1	0.14 (0.02)	0.17 (0.03)
σ_2	0.22 (0.03)	0.28 (0.03)
σ_w^2	0.05 (0.01)	0.05 (0.01)
DIC	6.8	10.5
Parameters	Student-t	(Bivariate)
α_0	3.09 (0.62)	2.92 (0.58)
α_1	0.39 (0.12)	0.41 (0.11)
β_0	0.96 (0.33)	1.23 (0.35)
β_1	0.74 (0.09)	0.66 (0.09)
σ_1	0.21 (0.03)	0.24 (0.03)
σ_2	0.29 (0.04)	0.35 (0.04)
ρ	0.61 (0.10)	0.55 (0.11)
DIC	32.9	30.7

We consider the DIC values for the two proposed models, Student-t distribution with random effects ($v = 4$ and $v \to \infty$) and two classical approaches considering bivariate Student-t distributions with the same degrees of freedom. The model considering a Student-t distribution with random effects ($v = 4$) has the lowest value of DIC, indicating the best fit between the compared models. In this way, we have the Figures 2 and 3 with the observed values versus the fitted model values. Note that we have a good fit for the proposed model.

In terms of estimated parameters, we have similar values for the models.

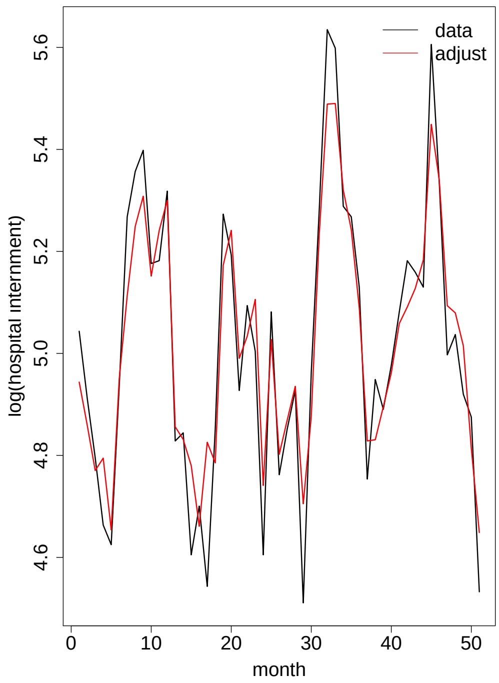

Figure 2 – Fitted model for hospital internment

4. CONCLUSION AND PERSPECTIVES

The use of Bayesian methods with the MCMC algorithm is a good alternative for the development of new statistical models in the solution of applied problems.

In this work, we proposed a model with random effects to capture a possible correlation between two series as alternative to classical method using bivariate distributions. This is useful since in this approach it is not necessary to assume a bivariate form of a specified distribution.

The perspectives of this work are: considering autoregressive models with $p > 1$; considering others structures to random effects; using the two techniques (random effects and bivariate models simultaneously) and a simulation study to compare these models in several different situations.

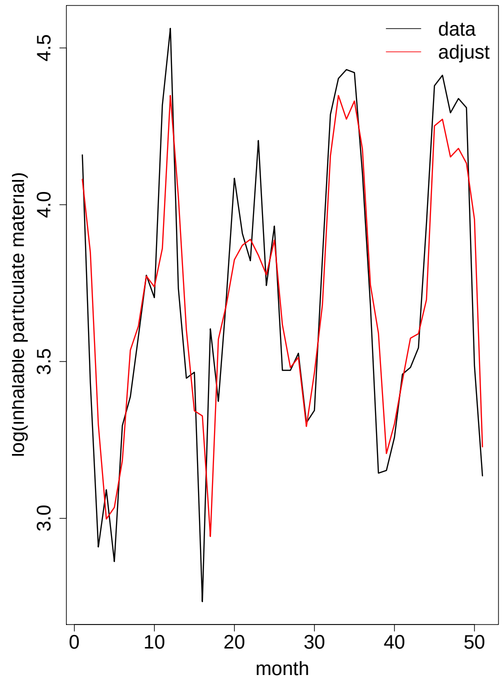

Figure 3 – Fitted model for inhalable particulate material

References

[1] B. Biller and B.L. Nelson, "Modeling and generating multivariate time-series input processes using a vector autoregressive technique," ACM Transactions on Modeling and Computer Simulation Vol. 13, No. 3, pp. 211-237, July 2003.

[2] R.A. McLean, W.L. Sanders and W. W. Stroup, "A unified approach to mixed linear models," American Statistical Associaton Vol. 45, No. 1, pp. 54-64, February 1991.

[3] A. Gleman, "Prior distributions for variance parameters in hierarchical models," Bayesian Analysis Vol. 1, No. 3, pp. 1-19, February 2006.

[4] S.K. Sahu, D.K. Dey and M.D. Branco, "A new class of multivariate skew distributions with applications to bayesian regression models," Canadian Journal of Statistics Vol. 31, No. 2, pp. 129-150, June 2003.

[5] K.L. Lange, R.J.A. Little and J.M.G. Taylor, "Robust statistical modeling using t distribution," Journal of the American Statistical Association Vol. 84, No. 408, pp. 881-896, December 1989.

[6] R.M. Souza, J.A. Achcar, E.Z. Martinez and J. Mazucheli, "The use of asymmetric distributions in average bioequivalence," Statistics in Medicine early view, January 2016.

[7] D.J. Lunn, A. Thomas, N. Best and D. Spiegelhalter, "WinBUGS - A Bayesian modelling framework: Concepts, structure, and extensibility," Statistics and Computing Vol. 10, No. 4, pp. 325-337, October 2000.

[8] D.J. Spiegelhalter, N.G. Best, B.P. A. Van Der Linde, "Bayesian measures of model complexity and fit," Journal of the Royal Statistical Society: Series B (Statistical Methodology) Vol. 64, No. 4, pp. 583-639, October 2002.