Bayesian Analysis (2011) **6**, Number 4, pp. 691–726

Product partition models with correlated parameters

João V. D. Monteiro*, Renato M. Assunção† and Rosangela H. Loschi*

Abstract. In sequentially observed data, Bayesian partition models aim at partitioning the entire observation period into disjoint clusters. Each cluster is an aggregation of sequential observations and a simple model is adopted within each cluster. The main inferential problem is the estimation of the number and locations of the clusters. We extend the well-known product partition model (PPM) by assuming that observations within the same cluster have their distributions indexed by correlated and different parameters. Such parameters are similar within a cluster by means of a Gibbs prior distribution. We carried out several simulations and real data set analyses showing that our model provides better estimates for all parameters, including the number and position of the temporal clusters, even for situations favoring the PPM. A free and open source code is available.

Keywords: Change point, Gibbs prior, MCMC, Temporal correlation

1 Introduction

Sequentially observed data can be analyzed by partitioning the observation period into disjoint contiguous segments, called here temporal clusters. The moment when a cluster ends and another one starts is called a change point. Within each one of these clusters, it is usually assumed that the data are independent and identically distributed random variables with a certain distribution such as the normal or Poisson distribution. As soon as the number of disjoint clusters and their locations are established, the analysis is quite simple since it is reduced to a univariate analysis within each cluster. The major difficulty in this approach is to make inference about the unknown number of clusters and their locations.

Some approaches consider that the number of change points is known and fixed

*Division of Biostatistics, School of Public Health, University of Minnesota, Minneapolis, MN, monte092@umn.edu

†Departamento de Estatstica, LESTE e CRISP - UFMG, Belo Horizonte, Brasil assuncao@est.ufmg.br

‡Departamento de Estatistica, ICEx, UFMG, Belo Horizonte, Brasil loschi@est.ufmg.br

DOI:10.1214/11-BA626

(Chen and Lee 1995; Geweke and Terui 1993; Hawkins 2001). Bayesian approaches for a single change point problem were developed in Menzefricke (1981), Hsu (1982), Smith (1975), and Carlin *et al.* (1992). A major advance was the product partition model (PPM) developed by Hartigan (1990) (see also Barry and Hartigan 1992) which generalizes all these Bayesian approaches since it considers the number of change points as unknown. Barry and Hartigan (1993) and Crowley (1997) applied the PPM to identify multiple change points in normally distributed data. They did not fully explore the posterior distribution, obtaining only the posterior means for the expected value at each time point. Later, Loschi and Cruz (2005a) extended the PPM by providing a method to obtain the posterior distributions for the positions and number of change points as well as the posterior probability of each instant being a change point. Another approach to obtain such posteriors was provided by Fearnhead (2006) and Fearnhead and Liu (2007) which provide algorithms based on filtering for multiple change point problems. As an important extension of the PPM, Quintana and Iglesias (2003) present a decision-theoretic formulation to specific change point problems, such as outlier detection, using the PPM. This approach is also used for Tarantola *et al.* (2008) for row effects models. A similar approach in the context of market risk measuring is also presented in Bormetti *et al.* (2010). Quintana and Iglesias (2003) also proved that PPM generalizes the Dirichlet process, which has been intensively used for clustering analysis (Müller and Quintana 2010) and density estimation (Escobar and West 1995). Quintana (2006) establishes connections between the PPM and other models that induce a partition structure. More recently, the PPM was used in the spatial context by Hegarty and Barry (2008) for Bayesian disease mapping, and in survival analysis by Demarqui *et al.* (2008) that applied it to estimate the time grid in piecewise exponential models. Ruggeri and Sivaganesan (2005) and Booth *et al.* (2008) suggest other approaches on multiple change point identification. In the spatio-temporal setting, Majumdar *et al.* (2005) introduced a change point model attempting to capture changes in both the temporal and spatial associations.

The popularity of PPM is justified by its flexibility to analyze change point or clustering problems. However, its formulation assumes a common parameter indexing the distributions of the observations into the same temporal cluster. Furthermore, it also assumes independence among the common parameters associated with different temporal clusters. This approach may lead to an inaccurate identification of the number of clusters if these assumptions are not closely followed by the data. For example, whenever the time series has an underlying trend, the usual PPM overestimates the number of clusters.

In this paper, we extend the PPM to a hierarchical Bayesian model for clustering analysis in the temporal context. We also assume independence among parameters in different temporal clusters, but contrary to what is assumed in the PPM, we consider that the observations in the same cluster have their distributions indexed by different parameters. Although different, the parameters are similar for observations within a given cluster. This is done by adopting a Gibbs distribution as the prior specification for the canonical parameters. As a result, the parameters within the same temporal cluster are correlated.

One important advantage in allowing similar parameters within a temporal cluster is that, rather than having an unknown dimension, the dimension of the parameter vector is fixed. This facilitates the numerical procedures used to sample from the posterior distribution. As we show in our examples, it also makes the interpretation of the results more straightforward and accurate.

We obtain several probabilistic results related to the prior specification for the random partition that defines the position of the change points. We focus on normally and Poisson distributed data but the method can be easily generalized to a larger class of distributions such as the exponential family.

In Section 2, we introduce our clustering model for Poisson and normal data. We also present some probabilistic results related to the prior specification for the random partition and the connections between the proposed model and PPM. In Section 3, we present examples using simulated data sets. In Section 4, we analyze some case studies in order to illustrate the use of the proposed model. Finally, we close the paper with some final comments and conclusions in Section 5. Technical proofs are presented in the Appendix.

2 Model specification

In this section we review the PPM introduced by Barry and Hartigan (1992) and present our approach for the partition model. We also present some results related to the prior specifications for the partition and number of clusters.

Throughout this paper we assume the following notation. Consider an observation $\mathbf{y} = (y_1, \ldots, y_n)$ of the vector $\mathbf{Y} = (Y_1, \ldots, Y_n)$ composed of sequentially observed random variables. Let $I = \{1, \ldots, n\}$ be the index set. A *temporal cluster* $\mathcal{C}_j$ is the subset $\mathcal{C}_j = \{i_{j-1} + 1, \ldots, i_j\}$ of I, $i_k \in I$, $k = 1, \ldots, c$, such that $0 = i_0 < i_1 <$

$\ldots < i_c = n$. Let $\pi = \{\mathcal{C}_1, \ldots, \mathcal{C}_c\}$ be a random partition of the set I into $C = c$ contiguous temporal clusters inducing the partition $\mathbf{y}_{\mathcal{C}_1} = (y_{i_0+1}, \ldots, y_{i_1}), \ldots, \mathbf{y}_{\mathcal{C}_c} = (y_{i_{c-1}+1}, \ldots, y_n)$ in $\mathbf{Y}$.

2.1 The Product Partition Model

We review briefly the PPM introduced by Hartigan (1990) (see also Barry and Hartigan 1992). Assume that, given $\mu_1, \cdots, \mu_n$, $\mu_i \in \mathbb{R}$, the random variables $Y_1, \ldots, Y_n$ are independent and $Y_i|\mu_i \sim f(y_i \mid \mu_i)$, $\forall\ i \in I$. In the PPM it is assumed that, given a partition $\pi = \{\mathcal{C}_1, \ldots, \mathcal{C}_c\}$ and $c \in I$, there are common parameters $\mu_{\mathcal{C}_j}$, $j = 1, \ldots, c$, that is, $\mu_{\mathcal{C}_j} = \mu_{i_{j-1}+1} = \cdots = \mu_j$, which indexes the conditional density of $\mathbf{Y}_{\mathcal{C}_j}$. Additionally, it is assumed that $\mu_{\mathcal{C}_1}, \ldots, \mu_{\mathcal{C}_c}$ are independent, with $\mu_{\mathcal{C}_j}$ having (block) prior density $\pi(\mu_{\mathcal{C}_j})$.

Denote by $G_{\mathcal{C}_j}$, $j \in I$ the prior cohesion associated with the block $\mathcal{C}_j$. The prior cohesions are non-negative numbers, not all equal to zero. In the contiguous clusters case, the set of $G_{\mathcal{C}_j}$ values can be interpreted as the transition probabilities in a Markov chain with state space defined by the possible endpoints of the clusters in π. Thus, $G_{\mathcal{C}_j}$ denotes the probability of having a change at instant i_j, given that a change takes place at i_{j-1}. Barry and Hartigan (1992) establish that the joint distribution of $(Y_1, \ldots, Y_n, \pi)$ follows the PPM if

(i) the prior distribution of π is the following product distribution

$$P(\pi = \{\mathcal{C}_1, \ldots, \mathcal{C}_c\}) = \frac{\prod_{j=1}^{c} G_{\mathcal{C}_j}}{\sum_{\mathcal{C}} \prod_{j=1}^{l} G_{\mathcal{C}_j}}, \tag{1}$$

in which $\mathcal{C}$ is the set of all possible partitions of the set I.

(ii) conditional on $\pi = \{\mathcal{C}_1, \ldots, \mathcal{C}_c\}$, the sequence $Y_1, \ldots, Y_n$ has the joint density given by

$$f(\mathbf{y}|\pi = \{\mathcal{C}_1, \ldots, \mathcal{C}_c\}) = \prod_{j=1}^{c} f(\mathbf{y}_{\mathcal{C}_j}),$$

where $f(\mathbf{y}_{\mathcal{C}_j}) = \int f(\mathbf{y}_{\mathcal{C}_j}|\mu_{\mathcal{C}_j})\pi(\mu_{\mathcal{C}_j})d\mu_{\mathcal{C}_j}$ denotes the data factor of the cluster $\mathcal{C}_j$.

Under the PPM, the posteriors of μ_k, C and π are given, respectively, by

$$\begin{aligned} P(\pi = \{\mathcal{C}_1, \ldots, \mathcal{C}_c\} | \mathbf{y}) &\propto \prod_{j=1}^{c} G_{\mathcal{C}_j} f(\mathbf{y}_{\mathcal{C}_\mathbf{j}}), \\ P(C = c | \mathbf{y}) &\propto \sum_{\mathcal{C}_c} \prod_{j=1}^{c} G_{\mathcal{C}_j} f(\mathbf{y}_{\mathcal{C}_j}), \\ f(\mu_k | \mathbf{y}) &= \sum_{i=0}^{k-1} \sum_{j=k}^{n} r^*_{[ij]} f(\mu_k | y_{i+1}, \ldots, y_j), \end{aligned}$$

where $\mathcal{C}_c$ is the set of all possible partitions of I in $C = c$ clusters. The posterior for μ_k, $k = 1, \ldots, n$, is a mixture of posterior-by-block distributions of μ_k where the mixing measure is the posterior probability $r^*_{[ij]}$ of block $[ij]$ being into the partition π. The probability $r^*_{[ij]}$ is known as the block $[ij]$ posterior relevance.

2.2 Proposed model

As in the PPM model, assume that, given $\mu_1, \cdots, \mu_n$, $\mu_i \in \mathbb{R}$, the random variables $Y_1, \ldots, Y_n$ are independently distributed and such that $Y_i | \mu_i \sim f(y_i \mid \mu_i)$, $\forall\, i \in I$. The likelihood function when the data set is partitioned into $C = c$ temporal clusters is given by:

$$f(\mathbf{y} | \boldsymbol{\mu}, \pi = \{\mathcal{C}_1, \ldots, \mathcal{C}_c\}) = \prod_{j=1}^{c} \prod_{k=i_{j-1}+1}^{i_j} f(y_k \mid \mu_k). \tag{2}$$

Since we expect temporal correlation among data, we connected the μ_i's by means of the prior specification for $\boldsymbol{\mu}$ conditioned on the partition π. Rather than fixing a constant value for all μ_i's within a cluster, we assume that they vary smoothly within the cluster. We believe this is a more realistic assumption. Many latent and unobserved factors vary in time implying differences, even if small, between the observations' distributions. As a consequence, we expect differences between the y_i's distribution parameters, even if they are close in time and share many characteristics. If similarity between a pair of instants is high, this difference may be so small as to make it undetectable statistically or irrelevant in practice. Conceptually it seems reasonable to allow for prior differences between any two moments in time and let the data determine the inference.

A prior that incorporates smooth variation between items is the intrinsic conditional autoregressive (ICAR) model, a Markov model introduced by Besag *et al.* (1991) to estimate relative risks in spatially located small areas or to evaluate the effects of covariates acting as exposure measurements surrogates. In the ICAR model, spatial dependence is expressed conditionally by requiring that the random effect in a given area, given the values in all other areas, depends only on a small set of neighboring values. This spatial prior distribution is improper and it has been extremely successful in the spatial statistics literature. It has been extended into several directions to include space-time generalized linear models, spatial survival models, spatially-varying parameters models, and generalized additive models. A thorough review of applications in spatial problems can be found in Banerjee *et al.* (2004) and Rue and Held (2005).

Our prior distribution is a one-dimensional version of the ICAR model. Given $\pi = \{\mathcal{C}_1, \ldots, \mathcal{C}_c\}$ and the hyperparameter $\tau_{\boldsymbol{\mu}}$, we assume that $\boldsymbol{\mu}$ has the following Gibbs distribution:

$$f(\boldsymbol{\mu}|\pi = \{\mathcal{C}_1, \ldots, \mathcal{C}_c\}, \tau_\mu) \propto \tau_\mu^{(n-c)/2} \exp\left\{-\frac{\tau_\mu}{2}\sum_{i=1}^{n-1}\delta_i(\mu_i - \mu_{i+1})^2\right\}\prod_{i=1}^{n} 1_S(\mu_i), \quad (3)$$

where $1_S(\mu_i)$ is the indicator function assuming 1 if $\mu_i \in S$ and 0 otherwise, the indicator function δ_i is equal to 0 if the observations i and $i+1$ do not belong to the same temporal cluster, and equal to 1, otherwise. The set $S \subseteq \mathbb{R}$ is the parametric space for μ_i, $i = 1, \ldots, n$. The presence of the support S in this prior density is necessary to avoid pathological behavior of the posterior, as we show in Section 2.3.

This prior distribution puts more probability mass on configurations with similar μ_i values within clusters. The existence of different distributions between clusters is akin to the existence of islands in the spatial context and this explains the rather unusual exponent $n - c$ of the precision τ_μ in (3) (Hodges *et al.* 2003; Knorr-Held 2003). The degree of similarity is controlled by the hyperparameter τ_μ. Let the $n-1$ dimensional vector $\boldsymbol{\mu}_{-i}$ be the $\boldsymbol{\mu}$ vector without the i-th entry. Suppose that μ_i, $1 < i < n$, belongs to a cluster with at least two observations. Then

$$f\left(\mu_i \mid \boldsymbol{\mu}_{-i}, \pi = \{\mathcal{C}_1, \ldots, \mathcal{C}_c\}, \tau_\mu\right) \propto N\left(\overline{\mu}_i, (\tau_\mu n_i)^{-1}\right) 1_S(\mu_i), \quad (4)$$

where $\overline{\mu}_i = (\delta_{i-1}\mu_{i-1} + \delta_i\mu_{i+1})/n_i$ and $n_i = \delta_{i-1} + \delta_i$. The larger the hyperparameter τ_μ, the more tightly clustered are the within cluster values.

To complete the model specification, we need the prior for τ_μ and, most importantly, for the random partition π. The prior for the random partition is built by noticing the equivalence of π to the random vector $(\delta_1, \ldots, \delta_{n-1})$:

$$\begin{aligned} \pi &= \{\mathcal{C}_1, \ldots, \mathcal{C}_c\} \\ &\Leftrightarrow \{\delta_1 = 1, \ldots, \delta_{i_1} = 0, \delta_{i_1+1} = 1, \ldots, \delta_{i_2} = 0, \ldots, \delta_{n-1} = 1\}. \quad (5) \end{aligned}$$

Let us suppose that $p \in (0,1)$ is the probability of two adjacent observations belonging to the same temporal cluster. Assuming that the Bernoulli random variables δ_i are independent, one option for the prior distribution of π may be given by:

$$P(\pi = \{\mathcal{C}_1, \dots, \mathcal{C}_c\} | p) = p^{n-c}(1-p)^{c-1}, \tag{6}$$

where $c = n - \sum_{i=1}^{n-1} \delta_i$ denotes the number of temporal clusters in the data set.

It is noteworthy that, given p, the prior distribution for π is a product distribution. In fact, it is a product of truncated geometric distributions. This implies that the prior distribution for π is the same as that defined by Barry and Hartigan (1992) assuming the Yao's prior cohesion (Yao 1984).

In Section 2.4, we discuss several properties of the prior distribution in (6) that help us to select appropriate hyperparameters in specific applications. The distribution in (6) typically renders excellent results in practice (see the simulated and case studies sections). However, it is not the only option for the partition prior. Other priors may be more appropriate in some particular applications and hence we provide additional options in the last section. For the sake of definiteness, we will continue introducing our model in the present section using (6).

For the hyperparameters, we assume that p and τ_μ are independent with $p \sim \text{Beta}(a,b)$ and $\tau_\mu \sim \text{Gamma}(r,s)$, where a, b, r and s are known, positive and real numbers.

Let $\boldsymbol{\theta} = (\boldsymbol{\mu}, \pi, p, \tau_\mu)$. The posterior distribution of $\boldsymbol{\theta}$ is given by:

$$\begin{aligned} P(\boldsymbol{\theta}|\mathbf{y}) &\propto f(\mathbf{Y}|\boldsymbol{\mu}) f(\boldsymbol{\mu}|\tau_\mu, \pi) P(\pi|p) f(p) f(\tau_\mu) \\ &\propto \left[\prod_{j=1}^{c} \prod_{k=i_{j-1}+1}^{i_j} f(y_k \mid \mu_k)\right] \tau_\mu^{(n-c+2r-2)/2} p^{\,n+a-c-1}(1-p)^{c+b-2} \\ &\times \exp\left\{-\tau_\mu \left(\sum_{i=1}^{n-1} \delta_i (\mu_i - \mu_{i+1})^2 + s\right)\right\} \prod_{i=1}^{m} 1_S(\mu_i). \end{aligned} \tag{7}$$

The prior distribution of $\boldsymbol{\mu}$ in (3) is improper (Banerjee *et al.* 2004). The use of improper priors is acceptable as long as we get a proper posterior distribution. In general they arise in reference or objective Bayesian analysis (see Robert 2007, for a more detailed discussion) but are frequently used whenever there is no prior information available. The next proposition provides conditions under which the posterior is proper. Its proof is presented in Appendix A.

Proposition 1. *Assume the previous prior specifications for the components of $\boldsymbol{\theta}$ and that, given $\mu_1, \ldots, \mu_n$, the random variables $Y_1, \ldots Y_n$ are independent and such that $Y_i \mid \mu_i \sim f(y_i \mid \mu_i)$. If $g(\boldsymbol{\mu}) = \prod_{i=1}^n f(y_i \mid \mu_i) 1_S(\mu_i)$ is an integrable function w.r.t. $\boldsymbol{\mu}$, then the posterior distribution of $\boldsymbol{\theta}$ given in (7) is proper.*

The posterior distribution of π under the proposed model is a product distribution, as it is the case for such posterior distribution under PPM (Barry and Hartigan 1992) described in Section 2.1. In a general setting, consider a product prior distribution for the partition π, say, assume $P(\pi = \{\mathcal{C}_1, \ldots, \mathcal{C}_c\}) \propto \prod_{j=1}^c G_{\mathcal{C}_j}$, where $G_{\mathcal{C}_j}$, $j = 1, \ldots, c$, is a non-negative number. Assume that the μ's in different clusters are independent but that they are correlated whenever in the same cluster, such that $f(\boldsymbol{\mu}|\pi = \{\mathcal{C}_1, \ldots, \mathcal{C}_c\}) = \prod_{j=1}^c f(\mu_{i_{j-1}+1}, \ldots, \mu_{i_j})$. Thus, the posterior for π is given by:

$$P(\pi = \{\mathcal{C}_1, \ldots, \mathcal{C}_c\}|\mathbf{y}) \propto \prod_{j=1}^{c} G_{\mathcal{C}_j} f(\mathbf{y}_{\mathcal{C}_j}),$$

where $f(\mathbf{y}_{\mathcal{C}_j}) = \int \ldots \int \prod_{k=i_{j-1}+1}^{i_j} f(y_k|\mu_k) f(\mu_{i_{j-1}+1}, \ldots, \mu_{i_j}) d\mu_{i_{j-1}+1} \ldots d\mu_{i_j}$. Therefore, the data factor $f(\mathbf{y}_{\mathcal{C}_j})$ is more general than that assumed in the usual PPM (Barry and Hartigan, 1992, 1993).

2.3 Poisson and Normal cases

In this section we present two particular applications of the proposed model, when the sample distributions are the Poisson and normal ones. These are important special cases because many data analyses assume these distributions for the observed data.

Assume that, given $\lambda_1, \cdots, \lambda_n$, the random variables $Y_1, \ldots, Y_n$ are independent and such that $Y_i|\lambda_i \sim \text{Poisson}(\lambda_i), \lambda_i > 0$. Consider the canonical parameter $\mu_i = \ln(\lambda_i)$. Let $\boldsymbol{\mu} = (\mu_1, \ldots \mu_n)$. The likelihood function when the data set is partitioned into $C = c$ temporal clusters is given by:

$$f(\mathbf{y}|\boldsymbol{\mu}, \pi = \{\mathcal{C}_1, \ldots, \mathcal{C}_c\}) = \prod_{j=1}^{c} \prod_{k=i_{j-1}+1}^{i_j} \left[\frac{\exp\{-e^{\mu_k} + y_k \mu_k\}}{y_k!} \right]. \tag{8}$$

The Poisson case illustrates the need for the introduction of the support set S in the prior distribution (3). Let S^* be the implied support set for $\lambda_i = \exp(\mu_i)$, $i = 1, \ldots, n$. If we allow 0 to be an accumulation point of S^* we are led to an improper posterior if any $y_i = 0$. This is clearly an undesirable property that can be readily remedied if we assume that there exists $\epsilon > 0$ such that $\lambda_i \in S^* = (\epsilon, \infty)$ for all $i = 1, \ldots, n$. An important point is that, for the posterior to be proper, ϵ must merely exist. We do not need to know its value in any specific application. In practice this assumption will always be valid since ϵ can be taken arbitrarily small. Indeed, we can only think of mathematically pathological examples where the assumption that such $\epsilon > 0$ exists it is not true. It is hard to imagine a situation, even one with many observed zeros, in which we can not imagine an extremely small lower bound for the λ_i parameters. The mere adoption of extremely small values such as $\epsilon = 10^{-30}$ suffices to make the posterior a proper distribution.

Let $\boldsymbol{\theta} = (\boldsymbol{\mu}, \pi, p, \tau_\mu)$ and $S^* = (\epsilon, \infty)$, which implies that $S = (\log(\epsilon), \infty)$. The posterior distribution of $\boldsymbol{\theta}$ is given by:

$$\begin{aligned}
P(\boldsymbol{\theta}|\mathbf{y}) &\propto f(\mathbf{Y}|\boldsymbol{\mu}) f(\boldsymbol{\mu}|\tau_\mu, \pi) P(\pi|p) f(p) f(\tau_\mu) \\
&\propto \prod_{j=1}^{c} \prod_{k=i_{j-1}+1}^{i_j} \left[\frac{\exp\left\{-e^{\mu_k} + y_k \mu_k\right\}}{y_k!} \right] \tau_\mu^{(n-c+2r-2)/2} p^{\,n+a-c-1} (1-p)^{c+b-2} \\
&\times \exp\left\{ -\tau_\mu \left(\sum_{i=1}^{n-1} \delta_i (\mu_i - \mu_{i+1})^2 + s \right) \right\} \prod_{i=1}^{n} 1_S(\mu_i)\,. \qquad (9)
\end{aligned}$$

Corollary 1. *Assume the previous prior specifications for the components of* $\boldsymbol{\theta}$ *with* $S^* = (\epsilon, \infty)$ *where* $\epsilon > 0$, *and conditionally independent Poisson distributed data* Y_i. *The posterior distribution of* $\boldsymbol{\theta}$ *given in (9) is proper.*

Proof: Let $\lambda_i = e^{\mu_i}$ and consider the function $g(\boldsymbol{\mu}) = \prod_{i=1}^{n} \exp\{-e^{\mu_i} + y_i \mu_i\} 1^*_S(\lambda_i)$. Let $Z \subseteq \{1, \ldots, n\}$ be the set of indexes such that $y_i = 0$. Then $g(\boldsymbol{\mu})$ is the product of two factors, one collecting the indexes in Z, and the other collecting the remaining indexes. For those indexes such that $y_i > 0$, the corresponding factor in $g(\boldsymbol{\mu})$ is the product of kernels of the Gamma pdfs which is an integrable function w.r.t. $\boldsymbol{\mu}$. Consider now the indexes in Z and its corresponding factor $\prod_{i \in Z} [e^{\lambda_i} \lambda_i]^{-1} 1_{S^*}(\lambda_i)$ in $g(\boldsymbol{\mu})$. It is clear that $\int_\epsilon^\infty [e^{\lambda_i} \lambda_i]^{-1} d\lambda_i < \infty$. The result then follows from Proposition 1.

Using such an ICAR prior in the spatial context, Ghosh *et al.* (1998) proved that the posterior distribution is proper for the related parameters but their result is not as strong as needed for data analysis. For example, for Poisson data, the posterior is not proper if there is at least one zero count in the sample. We have been able to prove that a proper posterior is obtained including the zero count case by assuming a constraint in the prior dominion.

Assume now that, given $\boldsymbol{\mu} = (\mu_1, \ldots \mu_n)$ and τ_y, the random variables $Y_1, \ldots, Y_n$

are independent and such that $Y_i|\mu_i,\tau_y \sim \text{Normal}\left(\mu_i,{\tau_y}^{-1}\right)$. With the same previous prior specifications and assuming that $S=\mathbb{R}$ and $\tau_y \sim \text{Gamma}(t,u)$, the posterior distribution for $\boldsymbol{\theta}=(\boldsymbol{\mu},\tau_y,\pi,p,\tau_\mu)$ is given by:

$$\begin{aligned} P(\boldsymbol{\theta}|\mathbf{y}) &\propto f(\mathbf{Y}|\boldsymbol{\mu},\tau_y)f(\boldsymbol{\mu}|\tau_\mu,\pi)P(\pi|p)f(\tau_y)f(p)f(\tau_\mu) \\ &\propto \exp\left\{-\frac{\tau_y}{2}\left(\sum_{i=1}^{n}(y_i-\mu_i)^2+2u\right)\right\}\exp\left\{-\tau_\mu\left(\sum_{i=1}^{n-1}\delta_i(\mu_i-\mu_{i+1})^2+s\right)\right\} \\ &\times \tau_y^{(n+2t-2)/2}\tau_\mu^{(n-c+2r-2)/2}\; p^{\,n+a-c-1}(1-p)^{c+b-2}. \end{aligned} \qquad (10)$$

Corollary 2 establishes that the joint posterior distribution for normally distributed data in (10) is proper without any restriction in the parameter space, despite having assumed an improper prior for $\boldsymbol{\mu}$.

Corollary 2. *Assume that, given $\mu_1,\ldots,\mu_n$ and τ_y, $Y_1,\ldots,Y_n$ are independent and that $Y_i|\mu_i,\tau_y\sim\text{Normal}(\mu_i,{\tau_y}^{-1})$, $i=1,\ldots,n$. With the stated prior specification for $\boldsymbol{\theta}$ and for $S=\mathbb{R}$ the posterior distribution (10) of $\boldsymbol{\theta}$ is proper.*

Proof: The proof follows from Proposition 1 by noticing that

$$\begin{aligned} g(\boldsymbol{\mu}) &= s^{-(n-c+2r)/2}\left[1/2\left(\sum_{i=1}^{n}(y_i-\mu_i)^2+2u\right)\right]^{2/(n+2t)} \\ &= s^{-(n-c+2r)/2}u^{(n+2t)/2}\left[1+\frac{1}{2t}\left(\boldsymbol{\mu}-\mathbf{y}\right)^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}-\mathbf{y})\right]^{-(n+2t)/2}, \end{aligned} \qquad (11)$$

where $\boldsymbol{\Sigma}^{-1}$ is an $n\times n$ matrix such that $\boldsymbol{\Sigma}^{-1}=\text{diag}\{t/u,\ldots,t/u\}$, is the kernel of a multivariate Student-t distribution with location parameter $\mathbf{y}\in\mathbb{R}^n$, scale matrix $\boldsymbol{\Sigma}\in\mathbb{R}^+\times\mathbb{R}^+$ and degree of freedom $2t>0$. Consequently, the function $g(\boldsymbol{\mu})$ is integrable with respect to $\boldsymbol{\mu}$.

Since the posterior distributions of $\boldsymbol{\theta}$ in (9) and (10) do not have a closed form, we use Markov chain Monte Carlo (MCMC) methods. We use the standard Metropolis-Hastings algorithm to sample from the posterior distributions of the $\mu_i's$. To sample from the posterior of π, we take into consideration the binary characteristic of δ_i and its relationship with π. Then, we use the Gibbs sampling schemes introduced by Barry and Hartigan (1993) and Loschi and Cruz (2005a). The posteriors for the other parameters are approximated by a Gibbs sampler. We have no need for methods such as reversible jump MCMC since our parameter vector has a fixed dimension, in contrast with PPM (for the PPM approach for the normal case, see Barry and Hartigan 1993).

2.4 On the prior specification for the random partition π

In this section, we derive some properties concerning the number and the composition of the clusters that are instrumental for the prior specification in practice. The following lemma summarizes some well known results related to C.

Lemma 1. *If $P(\pi = \{\mathcal{C}_1, \dots, \mathcal{C}_c\} \mid p) = p^{n-c}(1-p)^{c-1}$ then it follows that:*

(i) the prior distribution of C, given p, is

$$P(C = c \mid p) = \binom{n-1}{c-1} p^{n-c}(1-p)^{c-1}, \quad c = 1, \dots, n; \tag{12}$$

(ii) if $p \sim \mathrm{Beta}(a, b)$, then the prior distribution of C is:

$$P(C = c) = \binom{n-1}{c-1} \frac{\Gamma(a+b)\Gamma(a+n-c)\Gamma(b+c-1)}{\Gamma(a)\Gamma(b)\Gamma(n+a+b-1)}, \quad c = 1, \dots, n; \tag{13}$$

(iii) if $p \sim \mathrm{Beta}(a, b)$, then the expectation and the variance of C is, respectively,

$$\begin{aligned} &\mathrm{E}(C) = n - (n-1)\tfrac{a}{a+b}, \\ &\mathrm{Var}(C) = (n-1)\left[\tfrac{ab(n-2)}{(a+b)^2(a+b+1)} + \tfrac{a}{a+b} + \tfrac{a^2}{(a+b)^2}\right]. \end{aligned} \tag{14}$$

The prior specification for π can be tuned by exploring the dependence of the prior mean and variance of C on a and b. If we set $a = b$, it follows that $\mathrm{E}(C) = (n+1)/2$ and $\mathrm{Var}(C) = (n-1)4^{-1}\left[(n-2)(2a+1)^{-1} + 3\right]$. As the common value of a and b increases, the variance of C decreases converging to $3(n-1)/4$. This prior specification of C with $a = b$ implies that around 50% of the observations are expected to be change points. Therefore, the special case of $a = b$ in the prior for p stimulates a large number of clusters. In the important particular case when $a = b = 1$, C has the discrete uniform in the set $\{1, \dots, n\}$, which is a common prior specification for C. Although this stimulates a priori a large number of clusters, it can be extremely effective in practice, as we show in our simulated and real data illustrations in Sections 3 and 4.

If b is constant and $a \to 0$, then $\mathrm{E}(C) \to n$ and $\mathrm{Var}(C) \to 0$. In other words, we are eliciting *a priori* that all the observations are in different clusters with probability one. If a is constant and $b \to 0$, then $\mathrm{E}(C) \to 1$ and $\mathrm{Var}(C) \to 2(n-1)$. In this case, we are anticipating a small number of change points in the series. Notice that the prior uncertainty about this number of clusters tends to increase as the sample size increases.

In Proposition 2, we give additional results concerning the prior specification for π. Let N_j be the number of observations in the temporal cluster which contains y_j and $\mathcal{B}(x,y)=\Gamma(x)\Gamma(y)/\Gamma(x+y)$.

Proposition 2. *If $P(\pi=\{\mathcal{C}_1,\ldots,\mathcal{C}_c\}\mid p)=p^{n-c}(1-p)^{c-1}$, then, for all $j\leq(n+1)/2$, we have that*

(i) the conditional prior distribution $P\left(N_j=k\mid p\right)$ is given by:

$$\begin{array}{lll} k(1-p)^2p^{k-1} & \textit{if} & 1\leq k<j, \\ p^{k-1}(1-p)+(j-1)p^{k-1}(1-p)^2 & \textit{if} & j\leq k<(n-j+1), \\ 2p^{k-1}(1-p)+[n-(k+1)](1-p)^2p^{k-1} & \textit{if} & (n-j+1)\leq k<n, \\ (1-p)^{n-1} & \textit{if} & k=n; \end{array} \tag{15}$$

(ii) if $p\sim$ Beta(a,b), then the prior distribution $P\left(N_j=k\right)$ is given by:

$$\begin{array}{lr} k\mathcal{B}(a+k-1,b+2)/\mathcal{B}(a,b), & 1\leq k<j, \\ \left(\mathcal{B}(a+k-1,b+1)+(j-1)\mathcal{B}(a+k-1,b+2)\right)/\mathcal{B}(a,b), & j\leq k<n-j+1, \\ \left(2\mathcal{B}(a+k-1,b+1)+[n-(k+1)]\mathcal{B}(a+k-1,b+2)\right)/\mathcal{B}(a,b), & n-j+1\leq k<n, \\ \mathcal{B}(a,b+n-1)/\mathcal{B}(a,b), & k=n. \end{array} \tag{16}$$

The proof of Proposition 2 can found in Appendix C. Notice that there is a symmetric relationship between observations that have the same distance from the middle. Thus, the results in Proposition 2 also follow for $j>(n+1)/2$.

2.5 Limitations of the use of improper priors

An important limitation of using improper priors in change point problems is the lack of interpretability of the posterior results, as discussed in Girón *et al.* (2007) and Moreno *et al.* (2005). A simplified example showing clearly the difficulties one can find is the following. Suppose conditionally that we observe independently and normally distributed random variables $Y_1,\ldots,Y_n$ with known variance $\sigma_y^2=1$. Under the PPM, assume the Jeffreys's prior for the common means $\mu_{\mathcal{C}_j}$ within the temporal cluster $\mathcal{C}_j$.

For illustration consider the following two partitions, $\pi_1=\{1,\ldots,n\}$, implying no change, and $\pi_n=\{\{1\},\ldots,\{n\}\}$, implying that every observation is a change point.

Let $\bar{y} = \sum_i y_i/n$. Then,

$$\begin{aligned} P(\pi_1 \mid \mathbf{y}) &\propto P(\pi_1) \int \left(\frac{1}{2\pi}\right)^{n/2} \exp\left\{-\frac{1}{2}\sum_i (y_i - \mu)^2\right\} d\mu \\ &\propto P(\pi_1) \left(\frac{1}{2\pi}\right)^{(n-1)/2} \exp\left\{-\frac{1}{2}\sum_i y_i^2 + \frac{n}{2}\bar{y}^2\right\}, \\ P(\pi_n \mid \mathbf{y}) &\propto P(\pi_n) \int \ldots \int \prod_i \frac{1}{2\pi} \exp\left\{-\frac{1}{2}(y_i - \mu_i)^2\right\} d\mu_1 \ldots d\mu_n \\ &\propto P(\pi_n)\,. \end{aligned}$$

Suppose that we observe $y_1 = \cdots = y_n = 0$, an empirical evidence favoring π_1. Thus,

$$R = \frac{P(\pi_1 \mid \mathbf{y} = \mathbf{0})}{P(\pi_n \mid \mathbf{y} = \mathbf{0})} = \frac{P(\pi_1)}{P(\pi_n)} \left(\frac{1}{2\pi}\right)^{(n-1)/2} \quad .$$

Assuming the general prior in (6) with a beta distribution for p, we find

$$P(\pi_1) = \frac{\Gamma(a+b)\Gamma(a+n-1)}{\Gamma(a)\Gamma(a+b+n-1)}, \tag{17}$$

and

$$P(\pi_n) = \frac{\Gamma(a+b)\Gamma(b+n-1)}{\Gamma(b)\Gamma(a+b+n-1)}. \tag{18}$$

If $a = b$, we have $R < 1$ for all n and goes to zero as n increases. This is also true when we assume a discrete uniform prior distribution for the random partition π. In particular, if $P(\pi_1) = P(\pi_n) = 1/2$, we have the posterior probability of π_1 always smaller than that of π_n and the ratio R going to zero as n increases. This is clearly undesirable.

Considering our proposed model, the situation is different. The prior distribution of $\boldsymbol{\mu}$ under π_n is constant, as the usual objective prior, and therefore the posterior probability $P(\pi_n \mid \mathbf{y}) \propto P(\pi_n)$, as in the PPM. However, given π_1, the prior distribution of $\boldsymbol{\mu}$ is not constant and assumes the following expression

$$f(\boldsymbol{\mu} \mid \tau_\mu, \pi_1) \propto \tau_\mu^{(n-1)/2} \exp\left\{-\frac{\tau_\mu}{2}\sum_{i=1}^{n-1} (\mu_i - \mu_{i-1})^2\right\}.$$

Consequently, we have that

$$P(\pi_1 \mid \mathbf{y}) \propto \tau_\mu^{(n-1)/2} \det(A)^{1/2} \exp\left\{-\frac{1}{2}\left(\mathbf{y}^t\mathbf{y} - \mathbf{y}^t \left(A^{-1}\right)^t \mathbf{y}\right)\right\} P(\pi_1)\,,$$

where A is the $n \times n$ covariance matrix

$$A = \begin{bmatrix} \tau_y + \tau_\mu & -\tau_\mu & 0 & \dots & 0 \\ -\tau_\mu & \tau_y + 2\tau_\mu & -\tau_\mu & \dots & 0 \\ 0 & -\tau_\mu & \tau_y + 2\tau_\mu & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \tau_y + \tau_\mu \end{bmatrix}.$$

Considering the general prior distribution for the random partition and the same beta prior distribution for p as before, we have that

$$R = \frac{P(\pi_1 \mid \mathbf{y} = \mathbf{0})}{P(\pi_n \mid \mathbf{y} = \mathbf{0})} = \frac{P(\pi_1)}{P(\pi_n)} \tau_\mu^{(n-1)/2} |A|^{0.5}, \tag{19}$$

where $P(\pi_1)$ and $P(\pi_n)$ are given in (17) and (18), respectively.

Figure 1 shows that the behavior of R calculated in (19) depends on τ_μ and n. In this plot, we considered $\tau_y = 1$ and $P(\pi_1) = P(\pi_n)$, which includes the cases of $a = b$ as well as the discrete uniform prior distribution for the random partition. We considered

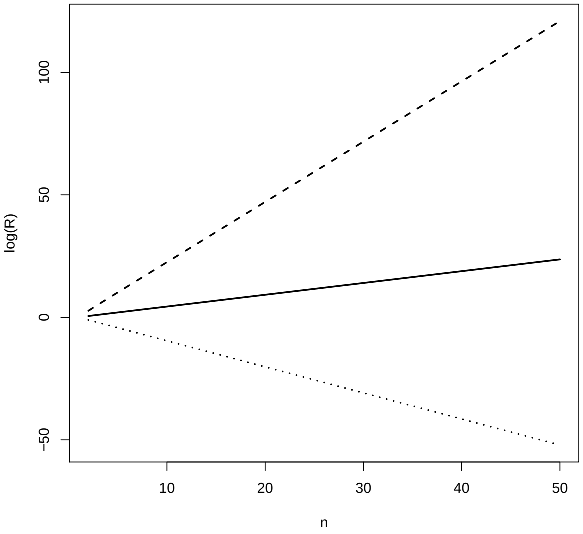


Figure 1: Logarithm of the ratio $R = P(\pi_1 \mid \mathbf{y} = \mathbf{0})/P(\pi_n \mid \mathbf{y} = \mathbf{0})$ versus the sample size n, for $\tau_\mu = 0.1$ (dotted line), $\tau_\mu = 1.0$ (solid line), and $\tau_\mu = 10.0$ (dashed line).

three different values for τ_μ (0.1, 1.0, and 10.0). When we assume *a priori* that the μ's within clusters are quite similar (that is, $\tau_\mu = 1.0$ or 10.0), the posterior probability for π_1 is higher than that for π_n, and the ratio increases with n. For the case where the degree of similarity among the μ's within clusters is not high (that is, $\tau_\mu = 0.1$), the posterior probability for π_1 is smaller than that for π_n and the ratio decreases with n. This last case is an undesirable behavior, as in the PPM. The more evidence favoring π_1 the data provide, the more the inference favors π_n.

However, this last situation, with τ_μ much smaller than τ_y, is one that is not likely to be considered in practical change point analysis. To be useful, a change point model should consider that $\tau_\mu \geq \tau_y$ as one expects to have more residual variability of the y's around their means μ than the variation of these μ's within a cluster. Different clusters should show large differences between their μ's but, within a cluster, we anticipate little variation of the successive μ's with respect to that on the data. This is the situation of the first two cases, where our model behaves properly. These two prior specifications presume a much smaller difference between successive μ's within a cluster than the variability of the y's.

Although we have not proved that our model is pitfall free, we have shown that, in constrast with the PPM in some specific cases, it has the behavior that one expects from a good inference tool. Our proposed model is less susceptible to the inconvenient interpretation issues that affect the PPM. Furthermore, in the following simulation study we show in some examples how the proposed model performs better than a PPM in identifying the number of clusters and their locations.

3 Analysis of simulated Poisson data

We ran a Monte Carlo study to evaluate repeated sampling properties of our proposed model (hereafter, called PM) and PPM. We consider only the more interesting case of Poisson distributed data. Three scenarios were considered, two favoring PPM and one favoring our PM. Scenario 1 contains data sets that do not have any change points, with n i.i.d. random values following a Poisson distribution. We considered two cases, data with mean equal to 10 or data with mean 40. In Scenario 2 we have two change points, at $i = 21$ and $i = 41$. Observations within the same temporal cluster are assumed to be i.i.d. We also considered two different cases, the first one with λ_i changing from 10 in the first cluster to $\lambda_i = 17$ in the second, and $\lambda_i = 25$ in the third. The second case had λ_i varying little, from 10 to 12 and then to 15.

As before, Scenario 3 has two change points at $i = 21$ and $i = 41$ but, in contrast, it is assumed that observations within the same temporal cluster have different distributions. In the first case, we assumed that the independent observations followed the Poisson distribution with mean $8 + 0.25(i-1)$ within the first cluster, with constant mean equal to 22 in the second cluster, and with mean equal to $5(i-40)$ in the third cluster. In the second case, we considered the means in the three clusters equal to $15 - 0.25i$, changing to $13 + 0.75(i-20)$, and then to the constant 22.

For all cases, we generated 100 independent series of size $n = 60$. For each series, the MCMC specifications were: 21,000 updates taking every 30^{th} observation to avoid serial correlation. The MCMC initial values were the known simulation true values making burn-in unnecessary. The algorithm for the proposed model was implemented in C++ and it can be obtained from the authors upon request. All scenarios were performed in a PC, Intel Core 2 Duo 2.26 GHz processor, 3 GB RAM. For our proposed model, it took approximately 55 minutes to run one case (100 series) while it took the PPM 10 minutes to run the same 100 series.

As prior specifications, we assume that $p \sim \text{Beta}(1,1)$ and $\tau_\mu \sim \text{Gamma}\,(2\times 10^4, 10)$. Notice that this prior expects a priori half of the observations to be change points, a clear overestimation. However, it will work extremely well delivering a posterior distribution concentrated around the true partition adopted, in contrast with the results for the traditional PPM.

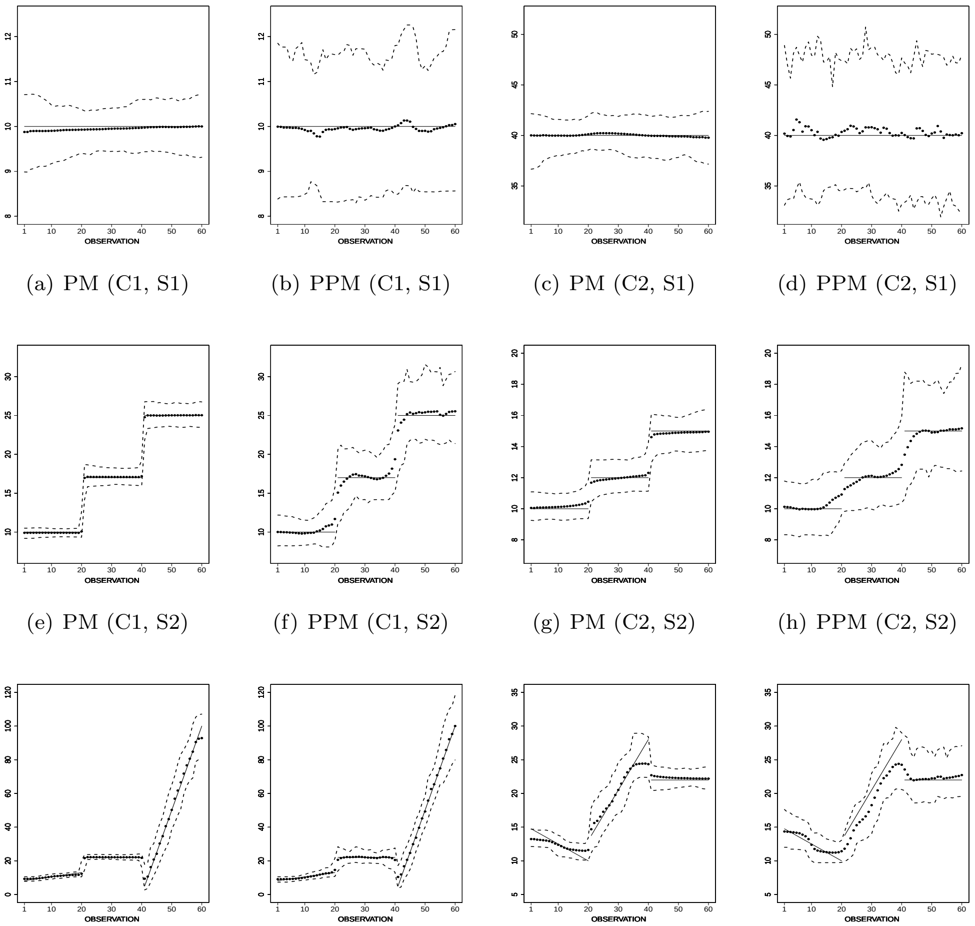
Since $E(\tau_\mu^{-1})$ and $\text{Var}(\tau_\mu^{-1})$ are close to zero, there is a strong prior belief that the $\mu's$ within the same temporal cluster are similar to each other, This ensures that the prior assumptions for PM are close to those in the PPM, which assumes equal μ's within the clusters. Furthermore, by assuming a non-informative prior for p, we are overestimating the number of clusters in the partition by a large amount ($E(C) = 30.50$ and $\text{Var}(C) = 18.15$). Following Loschi and Cruz (2005b) and in order to make the models comparable, we also assume Gamma distributions for the common rates $\lambda_{[i_{j-1}i_j]}$ in PPM with little prior information, considering $\lambda_{[i_{j-1}i_j]} \sim \text{Gamma}(1.010, 0.001)$.

Figure 2 summarizes the information about the posterior means of λ, at each instant, for both models, obtained from the replications of the series. It shows the average and the interval obtained considering the 2.5% and 97.5% percentiles of the posterior means for the λ_i's. Such an interval is named the equal-tailed 95% interval. In Scenario 1, case 1, on average, the posterior means of the λ_i's are close to the true values of the λ_i's. However, under PM, the estimates present smoother behavior over time and the equal-tailed 95% interval for the posterior means is shorter than under PPM. The tighter interval shows that the posterior means obtained are less influenced by different samples under PM than under PPM. Similar conclusions can be drawn for case 2. The posterior means for the λ_i's under the PPM are more distant from the true values whenever data variability is high. Moreover, the constant behaviour observed thorough time is not well captured by the product estimates.

Although Scenario 2 favors the PPM, the proposed model provides better estimates for the $\lambda_i's$. Under PM, the equal-tailed 95% intervals are shorter and the averages of the posterior means of the $\lambda_i's$ are closer to the true ones. More importantly, the estimates near the change points are much closer to the true values under PM than under PPM. For Scenario 3, case 1, on average, the models are comparable. A slight difference concerns the last two observations which are less well modeled under PM. For the first two temporal clusters, the equal-tailed 95% intervals are shorter under PM while in the last temporal cluster this is slightly reversed. For case 2, both models estimate the $\lambda_i's$ modestly well. Although PM is again better than PPM, both models poorly estimate the $\lambda_i's$ around the change points 21 and 41. Also, PM shows inefficiency in estimating the $\lambda's$ near the first observations. PPM does not work well in the second temporal cluster, underestimating the $\lambda's$.

Figure 3 shows the average of the posterior probabilities that each observation is a change point under the proposed model and the PPM. In scenario 1, a good model should assign low probability that an observation is a change point, as PM does, in contrast with PPM. Case 2 shows a puzzling pattern for PPM. For scenarios 2 and 3, PM assigns probabilities higher than 0.2 for observations 21 and 41 being change points, and lower probabilities for all the other observations. This is behavior not shared by PPM. We advance that this may be explained by a larger effect of untypical observations in the PPM than in the PM. This latter model seems to be more flexible to incorporate the variation in the data. In case 2 of scenario 3, both models have a poorer performance compared to their performance in case 1. On average, the posterior probability that $i = 41$ is a change point is greater than that for $i = 21$, due to the smaller difference $\lambda_{21} - \lambda_{20}$ than $\lambda_{41} - \lambda_{40}$.

Similarly, for scenario 3, case 1, on average, the posterior probabilities that an observation is a change point under PM are much closer to what one expects from a good model than the results for PPM. Note the increasing trend in these probabilities for the final i indexes under PPM. This shows that PPM is more influenced by the increasing trend in rates in the third temporal cluster than PM. For case 2, on average, PM presents for almost all non-change points smaller posterior probabilities than PPM



(a) PM (C1, S1) (b) PPM (C1, S1) (c) PM (C2, S1) (d) PPM (C2, S1)

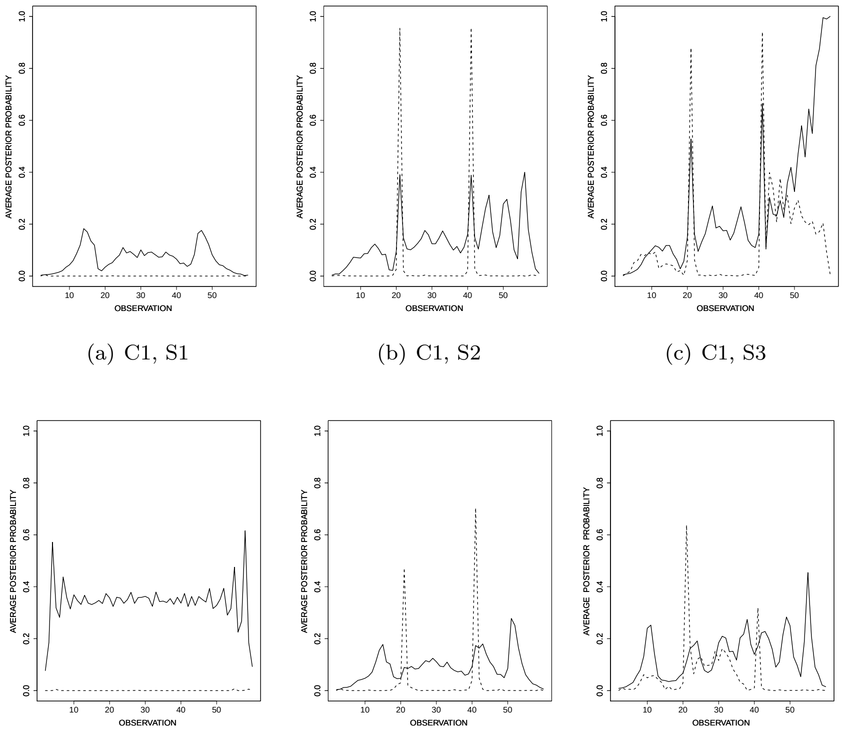
(e) PM (C1, S2) (f) PPM (C1, S2) (g) PM (C2, S2) (h) PPM (C2, S2)

(i) PM (C1, S3) (j) PPM (C1, S3) (k) PM (C2, S3) (l) PPM (C2, S3)

Figure 2: Average of the posterior means (dotted line) of the $\lambda_i' s$, 95% equal-tailed intervals (dashed lines) and true parameters (solid line) for the product partition (PPM) and proposed (PM) models, all scenarios (S) and cases (C).

while this order is reversed for the change points 21 and 41.

Table 1 gives some descriptive statistics for the posterior means of C under PM and PPM and it shows that PM performs better than PPM in the identification of the number of clusters. For Scenario 1, under PM, 75% of the posterior means of C are smaller than 1.01, while PPM overestimates C by a large amount. For Scenario 2, PPM



(a) C1, S1 (b) C1, S2 (c) C1, S3

(d) C2, S1 (e) C2, S2 (f) C2, S3

Figure 3: Average of the posterior probabilities that each observation is a change point, proposed model (dashed line) and the PPM (solid line), for all scenarios(S) and cases(C), Poisson data.

again overestimates C in both cases. PM correctly identifies the number of clusters in case 1 (50% of the posterior means are between 3.011 and 3.047), but it underestimates C in case 2 since the posterior mean is 2.21, on average. The detection of change points in case 2 is more difficult than in case 1 and this explains why both models have the average and the median of the posterior means of C smaller in case 2 than in case 1. For scenario 3, both models overestimate the number of temporal clusters, with PPM having a poorer performance. This behavior is due to the changing rates within the temporal clusters.

Table 1: Summaries for posterior means of C, Simulated Poisson data.

Scen.	Case	Model	Q1	Median	Mean	Q3
1	1	Proposed	1.0010	1.0060	1.0210	1.0130
		PPM	4.9770	5.0390	4.8920	5.0700
	2	Proposed	1.0010	1.0040	1.0280	1.0100
		PPM	20.5600	21.0700	21.0700	21.5500
2	1	Proposed	3.0110	3.0210	3.0630	3.0470
		PPM	8.4310	8.9030	8.8570	9.1750
	2	Proposed	2.1400	2.3360	2.4130	2.6410
		PPM	5.9480	6.086	6.0530	6.1820
3	1	Proposed	7.4520	8.1490	8.0760	8.5220
		PPM	16.0900	16.5900	16.6700	17.2300
	2	Proposed	3.4000	4.0890	4.1200	4.5980
		PPM	8.3800	8.6640	8.7460	9.0020

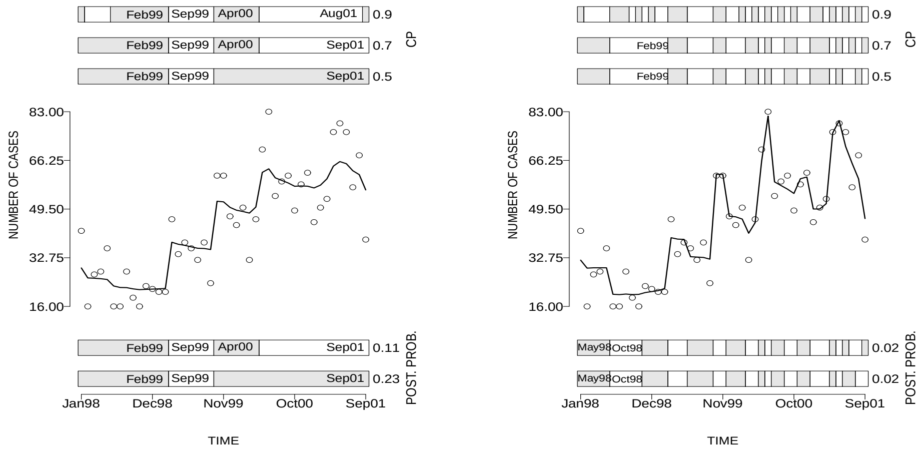
4 Case Studies

In this section, we apply the proposed model to two data sets. In the first application, we consider a series of crime counts and hence we use the proposed model for Poisson data. In the second application, we consider a time series of Brazilian household energy consumption and we use the proposed model for normally distributed data.

4.1 Case 1: Violent Crimes Data

The rate of violent crimes in urban centers in Brazil had been increasing since the 1980s and the citizens' trust in the police efficacy decreased, mainly after some acts of human rights violation by the police namely Candelária Massacre, in Rio de Janeiro, July 1993, the Carandiru Massacre, an October 1992 event in a São Paulo prison, and the Massacre of Vigário Geral slum, in Rio de Janeiro, August 1993. Crime is also an important issue in Belo Horizonte, a 2.4 million inhabitants city, Minas Gerais State capital. To restore the public confidence in the police and to modernize the police system the Military Police Command of Minas Gerais State introduced some changes to improve the police service. Particularly, since 1997, the Military Police Command is providing public specialized courses (which include Public Relations and Sociology) for police, through Fundação Jõao Pinheiro and Universidade Federal de Minas Gerais, and in the late 1990s, introduced a new program for crime reduction - "Policing with Results". The central idea is to help the police to be converted from a quasi military organization into a public service (Ward 2000).

Our interest is to estimate the rate of violent crimes in each month and to verify if the "Policing with Results" program produced a change by decreasing this rate. As part of a program to monitor crime statistics, we analyzed the series of counts of violent crimes in a Belo Horizonte neighborhood recorded monthly from January 1998, to September 2001. The data is plotted in Figure 4 jointly with the posterior estimates for the rate of crimes.



(a) Proposed Model

(b) PPM

Figure 4: Temporal Clusters, posterior means (solid line) of the $\lambda_i' s$ and data (circles), Violent crimes data.

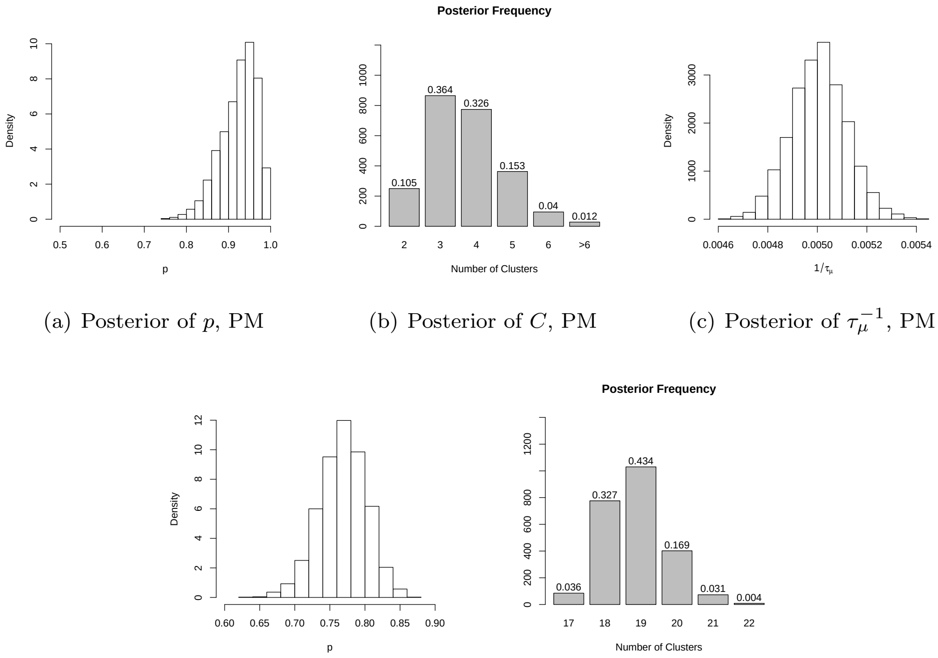
We assume that, given the underlying rates, the number of violent crime incidents are independent and distributed according to the Poisson distribution. For PPM, we adopt the gamma distribution, the natural conjugate prior distribution, for the rate of crimes $\lambda_{[ij]}$ within each cluster. Within the cluster, the observations have equal rates. We do not have a precise information about the rate of violent crimes and hence we assume that $\lambda_{[ij]} \sim \mathcal{G}(1.01, 0.001)$. For PM, we expect similar, but not exactly equal, rates of violent crimes within the same temporal cluster. Hence, we assume that $\tau_\mu \sim \text{Gamma}(2000, 10)$ implying a prior expectation that τ_μ^{-1} is around 0.005. A small number of changes is expected in the series. Thus, we also assume that the probability p has a beta prior distribution with parameters $\alpha = 1.2$ and $\beta = 0.24$, which means that the expected prior probability of a change is 16.7% and that $E(C) = 7.4$.

Table 2 and Figure 5 present the posterior distributions of p, τ_μ^{-1}, and C for both models. The posterior distributions for p and C under the two models have unique modes and are asymmetric. The posterior probability of a change $(1 - p)$ under the proposed model is smaller than we have assumed in the prior evaluation. It is much higher for PPM than for PM, with posterior expectation equal to 0.23. Similar behavior is observed for the number of temporal clusters C. PM infers that the data sequence, most probably, has 3 or 4 change points, with posterior probabilities 0.36 and 0.33, respectively. Inference based on PPM points out that, most probably, there are 19 clusters in the series with posterior probability 0.43. The same conclusion can be drawn from the posterior means. The posterior of τ_μ^{-1} is concentrated in small values implying that the parameters μ_i within each temporal cluster have similar values in the posterior evaluation. The posterior shows that the parameters μ_i are as similar as we expected *a priori*.

Table 2: Posterior summaries, Violent crime data.

	Parameter	Mean	St. Dev.	HPD: 95%	
Proposed Model	p	0.9244	0.0429	0.8396	0.9908
	C	3.6956	1.0615	2.0000	6.0000
	τ_μ^{-1}	0.0050	0.0001	0.0047	0.0052
PPM	p	0.7683	0.0335	0.6972	0.8269
	C	18.8438	0.8787	17.0000	20.0000

Figure 4 shown previously also presents the posterior most probable partitions and the posterior means for the λ_i's under both models. The estimates for $\lambda_i's$ are smoother than the ones obtained under the PPM. Also, the estimates for the λ_i's provided by the proposed model are less influenced by more extreme observations. Figure 4 also shows the most probable partitions *a posteriori*. The two bars on the bottom of each graph represent the two most probable partitions π. On the right hand side, the numbers on the extreme right give the posterior probability of each one of these partitions. The temporal clusters are represented by blocks of alternating white and gray colors.



(a) Posterior of p, PM (b) Posterior of C, PM (c) Posterior of τ_μ^{-1}, PM

(d) Posterior of p, PPM (e) Posterior of C, PPM

Figure 5: Posterior distributions for p, C and τ_μ^{-1}, Violent crimes data.

Within each cluster, we print the change point observation index. Moreover, the three bars in the top of Figure 4 represent a different attempt to summarize the posterior distribution of the partitions. CP is a fixed cut probability such that if the probability that observations i and $i+1$ belong to the same temporal cluster is higher than CP, then observations i and $i+1$ are considered to belong to the same temporal cluster. The right hand side in Figure 4 shows the results for PPM using CP equal to 0.5, 0.7, and 0.9.

The most probable partition under PM has posterior probability equal to 0.23 and it indicates that the data sequence is divided into 3 temporal clusters with changes occurring in March and October 1999. This same partition was obtained using $CP = 0.5$. The second most probable partition has posterior probability equal to 0.11 and it is the same as the one we obtain using $CP = 0.7$. This partition has three change points taking place in March and October 1999, and in May 2000. Under PPM, both the most probable partition π as well as the partitions based on the CP cut points indicate a large number of clusters in the data sequence. However, in this case, the two most probable partitions occur with posterior probability 0.02, which is very small, and indicates a large degree of uncertainty about π.

In conclusion, we observe that the rate of violent crimes in the analyzed region of Belo Horizonte is considerably high and experiences many changes in its behavior. A positive effect of the Policing with Results program could be the reduction of the rate of increase of the rate of violent crimes observed after May 2000.

4.2 Case 2: Consumption of Electric Energy Data

We turn now to the second case study. In 2001, Brazil experienced an energy crisis that led the government to adopt some policies to decrease the electric energy consumption. Figure 6 shows the time series of the Brazilian energy consumption in GWh/1000 and the posterior means, recorded monthly from January 1997 to December 2007 (source: www.ipeadata.gov.br). The goal is to verify if such policies produced a change in the behavior of Brazilian people concerning the use of power energy.

Since we do not introduce parameters to explicitly capture trends and seasonality in the time series, we expect a large number of changes in addition to one associated with the energy saving policy adopted by the government. We also expect the mean consumption within the same temporal cluster to be similar. Therefore, we take as prior specifications for τ_μ, τ_y and p the distributions Gamma$(10, 1)$, Gamma$(1, 1)$ and Beta$(5, 1)$, respectively. The prior expectation for the probability of a change is small and equal to 0.167, and the expected prior number of clusters in the time series is 22.

Somewhat surprisingly, the posterior inference was quite different from the prior we used. Table 3 and Figure 7 present the posterior distributions of p, τ_μ^{-1}, τ_y^{-1} and C for both models. The posterior distribution of p is left-skewed. This implies that the posterior probability $1 - p$ of a change is smaller than we have assumed in the prior evaluation. In fact, the posterior expectation of $1 - p$ is 0.0134 and the number of clusters resulted much smaller then we assumed *a priori*. It is most probable that the data sequence experiences only one change with probability 76.8%. Similar conclusion is drawn considering the posterior mean of C. The posterior of τ_μ^{-1} is concentrated in small values (around 0.0474) indicating that the parameters μ_i within each temporal cluster are more similar than we expected a priori (the prior mean is 0.11). The posterior distribution of τ_y^{-1} is also concentrated in small values, indicating that the consumption

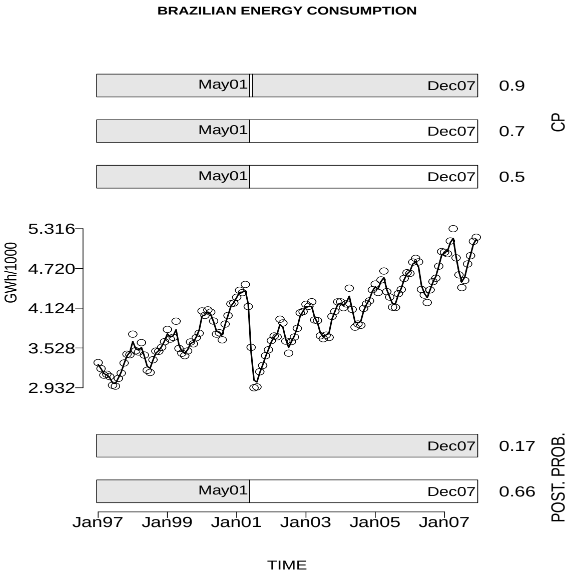
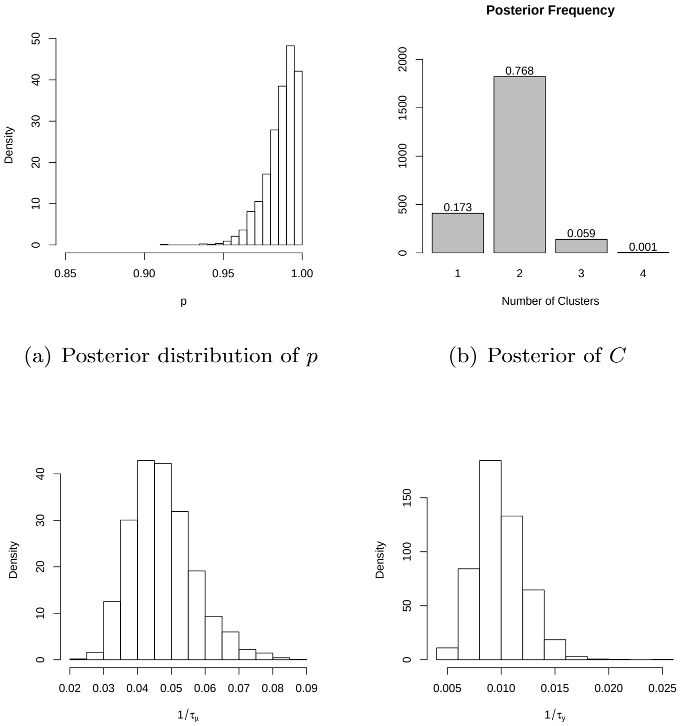


Figure 6: Temporal Clusters, posterior means (solid line) of the $\mu_i's$ and data (circles), Consumption of energy data.

of energy has small variability.

Table 3: Posterior summaries, Consumption of energy data.

Parameter	Mean	St. Dev.	HPD: 95%	
p	0.9866	0.0101	0.9671	0.9999
C	1.8880	0.4717	1.0000	3.0000
τ_y^{-1}	0.0099	0.0022	0.0059	0.0143
τ_μ^{-1}	0.0474	0.0094	0.0302	0.0660



(a) Posterior distribution of p

(b) Posterior of C

(c) Posterior of τ_μ^{-1}

(d) Posterior of τ_y^{-1}

Figure 7: Posterior distributions for p, C, τ_μ^{-1} and τ_y^{-1}, Consumption of energy data.

From Figure 6 we also perceive that the posterior means of the μ_i's follow the time series seasonality and indicate a change in the time series behavior around May 2001.

Notice also that the most probable partition, which occurs with probability 66.0%, as well as the partitions obtained using the $CP = 0.5$ and $CP = 0.7$ cut points, indicate that there is only one change point taking place in June 2001. Such decrease in the consumption of energy coincides with the energy crisis in Brazil and could be explained by the policy adopted by the Brazilian government for saving energy.

5 Conclusions

We extend the product partition model by introducing a hierarchical Bayesian model for clustering detection in the temporal setting. The new model assumes independence of the parameters into different temporal clusters but, contrary to PPM assumptions, it considers that the observations within the same temporal cluster have their distributions indexed by different and correlated parameters. Such correlation was introduced into the model by means of a Gibbs prior distribution for the parameters. Since such a prior is improper, we gave sufficient conditions under which the resulting posteriors are proper. We considered in particular the cases where the observed data are Poisson or normally distributed.

As one reviewer of this paper pointed out, there are limitations in using improper priors in change point problems. Even having proper posterior distributions, we can obtain undesirable paradoxical posterior results in some situations. These problems are especially acute in the PPM, as shown in Section 2.5. We have not proved that our model is completely free of these potential interpretation problems. However, we showed that our proposed model is less susceptible to those issues if the prior specifications are in agreement with the usual assumptions of change point models.

Some probabilistic results related to the prior specification for the partition, such as the distribution for the number of observations in the temporal cluster which contains a specific observation, were also obtained. We showed by means of simulations and real data illustrations that this simple prior distribution produces excellent results. However, if the user finds that (6) is not an appropriate prior distribution for the partition π, another possible approach is to consider a uniform prior for the random partition π with $P(\pi = \{\mathcal{C}_1, \ldots, \mathcal{C}_c\}) = 1/2^{n-1}$, implying that $C \sim \text{Binomial}(n-1, 0.5)$. Hence, we have around 50% of the observations expected to be change points, a likely overestimation.

Another alternative to (6) is to choose the uniform distribution $P(C = c) = 1/(n-1)$ for $c = 1, \ldots, n$. Hence, the configurations have different probabilities. For example, the single partition with only one cluster (the entire data sequence) or the single partition with n clusters (each observation is a cluster) receives the same prior probability as the large set of partitions with $c = n/2$ temporal clusters. In this approach, we avoid altogether the use of a prior distribution for p, defining the prior distribution of π conditioned on the number of clusters C. It follows that

$$P(\pi = \{\mathcal{C}_1, \ldots, \mathcal{C}_c\}|C = c) = \frac{1}{n}\binom{n-1}{c-1}^{-1},\ c = 1, \ldots, n.$$

Partitions with c and $n-c+1$ temporal clusters have the same prior probabilities. Note that partitions with $c \approx n/2$ (number of clusters close to half observations) are the least probable.

We presented three Monte Carlo studies using simulated Poisson data. We compared the results obtained by using the proposed and the product partition models. We concluded that PM presented better performance than PPM in all scenarios analyzed, including those that follow the PPM assumptions. In particular, PM provided better estimates for the number of temporal clusters and for the position of change points, while PPM overestimates the number of temporal clusters. PM also presented better estimates for the Poisson means. We also analyzed two data sequences based on non-simulated datasets and in both cases PM performed very well.

In summary, the proposed model provided better estimates for all parameters, including the number and positions of the temporal clusters, even for scenarios which favor the PPM (Scenarios 1 and 2). We conclude that by assuming non identical but correlated parameters within the same temporal cluster, we end up with better inference about the temporal clusters in the time series without burdening the computation of the posteriors.

A possible explanation for the worse performance of the PPM when compared to the proposed model is the assumptions considered in its construction. Since the PPM is built assuming constant cluster-specific parameters, high (similarly, for small) observed values in the time series - that commonly occur, for instance, when the rate in the Poisson model is high - tend to be identified as outliers, that is, as change points that do not truly occur. The proposed model is more flexible and depending on the degree of similarity assumed *a priori* for the parameters in the same cluster such observations are not atypical and, therefore, can not represent a change. Because of this flexibility the results provided by the proposed model is not as influenced by such atypical observations as the PPM is.

Appendix A: Proof of Proposition 1

We show here that the joint posterior distribution in (7) is a proper distribution. Integrating with respect to p and τ_μ successively, we obtain that

$$\begin{aligned} P(\pi = \{\mathcal{C}_1, \dots, \mathcal{C}_c\}, \boldsymbol{\mu}|\mathbf{y}) &\leq K_1 \left(\prod_{i=1}^{n} f(y_i \mid \mu_i) 1_S(\mu_i)\right) \Gamma\left(\sum_{i=1}^{n-1} \delta_i + a\right) \\ &\times \Gamma\left(n-1-\sum_{i=1}^{n-1} \delta_i + b\right) \Gamma\left(\frac{\sum_{i=1}^{n-1} \delta_i + 2r}{2}\right) \quad (20) \\ &\times \left[s + \sum_{i=1}^{n-1} \delta_i(\mu_i - \mu_{i+1})^2\right]^{-\left(\sum_{i=1}^{n-1} \delta_i + 2r\right)/2}, \quad (21) \end{aligned}$$

where $K_1 > 0$ is a constant that does not depend on $\pi = \{\mathcal{C}_1, \dots, \mathcal{C}_c\}$ and $\boldsymbol{\mu}$.

Denote by $h(\boldsymbol{\mu})$ the function collecting all terms in (21) that depend on $\boldsymbol{\mu}$ and let

$$g(\boldsymbol{\mu}) = \left(\prod_{i=1}^{n} f(y_i \mid \mu_i) 1_S(\mu_i)\right) s^{-\left(\sum_{i=1}^{n-1} \delta_i + 2r\right)/2}. \quad (22)$$

Since $\sum_{i=1}^{n-1} \delta_i(\mu_i - \mu_{i+1})^2 > 0$ and $(\sum_{i=1}^{n-1} \delta_i + 2r)/2 > 0$, then $h(\boldsymbol{\mu}) \leq g(\boldsymbol{\mu})$ for all $\boldsymbol{\mu} \in \mathbb{R}^n$. It also follows that

$$\int_S \cdots \int_S g(\boldsymbol{\mu}) d\mu_1 \dots d\mu_n = C(\boldsymbol{\delta}, r, s) \int_S \cdots \int_S \prod_{i=1}^{n} f(y_i \mid \mu_i) d\mu_1 \dots d\mu_n, \quad (23)$$

where $C(\boldsymbol{\delta}, r, s) = s^{\left(\sum_{i=1}^{n-1} \delta_i + 2r\right)/2}$.

By hypothesis, we have that $\int_S f(y_i \mid \mu_i) d\mu_i < \infty$, $\forall i = 1, \dots, n$. Then, the right term in (23) is finite. Since the bounding function $g(\boldsymbol{\mu})$ is integrable, the function $h(\boldsymbol{\mu})$ is also integrable. Consequently, we have that

$$\begin{aligned} P(\pi = \{\mathcal{C}_1, \dots, \mathcal{C}_c\}|\mathbf{Y}) &\leq K_1 \Gamma\left(\sum_{i=1}^{n-1} \delta_i + a\right) \Gamma\left(n-1-\sum_{i=1}^{n-1} \delta_i + b\right) \\ &\times \Gamma\left(\frac{\sum_{i=1}^{n-1} \delta_i + 2r}{2}\right) K_2(\boldsymbol{\delta}, r, s), \end{aligned}$$

where $K_1 > 0$ is a constant that does not depend on π. Since $\pi = \{\mathcal{C}_1, \dots, \mathcal{C}_c\}$ is a function of a finite set of $\delta_i's$, it follows that:

$$\sum_{\delta_1=0}^{1} \cdots \sum_{\delta_{n-1}=0}^{1} P(\pi = \{\mathcal{C}_1, \dots, \mathcal{C}_c\} | \mathbf{Y}) < \infty \, .$$

Appendix B: Proof of Proposition 2

Denote by $[i;j]$, the temporal cluster $\{i, i+1, \dots j\}$, where $i, j \in I$ and $i \leq j$ and let x be the number of observations in the temporal cluster that contains y_j and that appear before y_j.

For the first case, since $k < j$, all temporal clusters of size k that contain the observation y_j can be written as $[j-x; j+k-(x+1)]$, where $0 \leq x \leq k-1$. All these clusters have the same pattern: δ_{j-x}, …, $\delta_{j+k-(x+2)} = 1$, and $\delta_{j-x-1} = \delta_{j+k-(x+1))} = 0$. A similar pattern also happens for the extreme cases. For instance, if $x = 0$, the temporal cluster of interest is $[j; j+k-1]$, and $\delta_l = 1$ for $k-1$ observations. Since $1 \leq k < j$, the lower limit of this temporal cluster is such that $j \geq 2$.

Since $j < \lfloor (n+1)/2 \rfloor$, the number of observations after the y_j is $n - j + 1 \geq \lfloor (n+1)/2 \rfloor$. Thus, it follows that:

$$\begin{aligned} &k < j \text{ and } j \leq \lfloor (n+1)/2 \rfloor \Rightarrow k < \lfloor (n+1)/2 \rfloor \Rightarrow k \leq \lfloor n/2 \rfloor \, , \\ &n-j+1 \geq \lfloor (n+1)/2 \rfloor \Rightarrow n-j+1 \geq \lfloor n/2 \rfloor \Rightarrow n-j+1 > k. \end{aligned}$$

Consequently, the upper limit is such that $j+k-1 \leq n-1$. That is, in the extreme case $x = 0$, the upper limit $j+k-1$ for the temporal cluster can be the penultimate observation at the most. Thus, since the lower limit of the temporal cluster is at least 2 and the upper limit is $n-1$ at the most, we have that δ_{j-1} and δ_{j+k-1} are equal to 0 and also there are $k-1$ δ_i's equal to 1, say, $\delta_j = 1, \dots, \delta_{j+k-2} = 1$.

A similar argument holds for the other extreme case when $x = k-1$ and the temporal cluster of interest is $[j-k+1; j]$. Since $k < j$, the lower limit is such that $j-k+1 > 1$, which means that the lower limit of such a cluster is at least in the second observation. Since $j < \lfloor (n+1)/2 \rfloor$, we also have that the upper limit of the temporal cluster never can reach the last observation. Therefore, it follows that δ_{j-k} and δ_j are equal to 0, and there are $k-1$ $\delta_i's$ equal to 1, say, $\delta_{j-k+1} = \ldots = \delta_{j-1} = 1$.

Each of these temporal clusters occurs with probability $(1-p)^2 p^{k-1}$. Consequently, since there are k temporal clusters with k observations which contain y_j,

$$P\left(N_j | \, p\right) = k(1-p)^2 p^{k-1} \text{ if } \quad 1 \leq k < j.$$

Consider now that $j \leq k < n - j + 1$. Then, all temporal clusters of size k that contain y_j are such that $[j - x; j + k - (x + 1)]$, where $0 \leq x \leq j - 1$.

There are two different patterns for temporal clusters in this situation. In one pattern we observed $k - 1$ variables δ_i assuming value 1 and two others assuming value zero, say, $\delta_{j-x} = \ldots = \delta_{j+k-(x+2)} = 1$, and $\delta_{j-x-1} = \delta_{j+k-(x+1)} = 0$. In the other pattern, we have $k - 1$ $\delta_i's$ equal to 1, say, $\delta_{j-x} = \ldots = \delta_{j+k-(x+2)} = 1$, and $\delta_{j+k-(x+1)} = 0$.

Let us consider the extreme cases for clusters with the first kind of pattern. If $x = j - 1$, the temporal cluster of interest is $[1; k]$. Since $k < n - j + 1$, we have that the upper limit of such a temporal cluster is $j < n$. Thus, in this case, there are $k - 1$ variables δ_i which assume values equal to 1, say, $\delta_1 = 1, \ldots, \delta_{k-1} = 1$ and one that is such that $\delta_k = 0$. Consequently, this event occurs with probability $p^{k-1}(1 - p)$. On the other hand, for the second kind of pattern, if $x = 0$ the temporal cluster is $[j; j + k - 1]$. If $j = 1$, the only possible cluster is $[1; k]$. Then, let us consider $2 \leq j < n - j + 1$. Since $k < n - j + 1$ the upper limit is such that $j + k - 1 \leq n - 1$, which means that the upper limit of the temporal cluster is the penultimate observation at the most which implies that $\delta_{j+k-1} = 0$. Also, it follows that the lower limit is such that $j \geq 2$, that is, $\delta_{j-1} = 0$.

If $x = j - 2$ the temporal cluster of interest is $[2, k + 1]$. Since $k < n - j + 1$, the upper limit of this temporal cluster is such that $k + 1 < n$, which means that the upper limit of the temporal cluster can be, at the most, at the penultimate observation. Thus, δ_{k+1} and δ_1 are equal to 0. This particular cluster occurs with probability to $p^{k-1}(1 - p)^2$. Moreover, for all the other cases, say, $j - 2 < x < 1$, the same pattern is observed. Consequently, there are $j - 1$ temporal clusters of size k which contain the y_j and that have the same pattern observed for the case $x = 0$ and $x = j - 2$. Therefore, for situation 2, it follows that

$$P\left(N_j = k|p\right) = p^{k-1}(1 - p) + (j - 1)p^{k-1}(1 - p)^2 \text{ if } \quad j \leq k < (n - j + 1).$$

Assume now that $(n - j + 1) \leq k < n$. Then, all temporal clusters of size k that contain the y_j are such that $[j - x; j + k - (x + 1)]$, where $k - (n - j + 1) \leq x \leq j - 1$.

Notice that here we also have two different patterns for the temporal clusters. In one pattern $\delta_{j-x} = \ldots = \delta_{j+k-(x+2)} = 1$, and $\delta_{j-x-1} = \delta_{j+k-(x+1)} = 0$. In the other pattern, $\delta_{j-x} = \ldots = \delta_{j+k-(x+2)} = 1$ and $\delta_{j+k-(x+1)} = 0$.

Let us consider the extreme cases in which $x = k - (n - j + 1)$ and the temporal cluster of interest is $[n - k + 1; n]$. Since $k < n$ the lower limit is such that $n - k + 1 > 1$, which means $\delta_{n-k} = 0$. There are $k-1$ $\delta_i's$ equal to 1, say, $\delta_{n-k+1} = 1, \ldots, \delta_{n-1} = 1$. For the other extreme case, when $x = j-1$, the temporal cluster is $[1; k]$. Since $k < n$ the upper limit can reach the penultimate observation at the most, so that $\delta_k = 0$. As a consequence, it follows that $\delta_1 = \ldots = \delta_{k-1} = 1$. Thus, this temporal cluster occurs with probability $p^{k-1}(1-p)$.

The pattern changes when $k-(n-j+1) < x < j-1$. For instance, if $x = k-(n-j)$ the temporal cluster is $[n-k; n-1]$. Consider $k < n-1$ (if $k = n-1$ the temporal cluster has a pattern mentioned before). Then, the lower limit of the temporal cluster is such that $n-k > 1$. Thus, the observation $n-k$ is, at least, equal to 2, that is, $\delta_{n-k-1} = 0$. Similarly, for $n-1$, we have that $\delta_{n-1} = 0$. In all these cases, there are $k-1$ $\delta_i's$ equal to 1, say, $\delta_{n-k} = \ldots =, \delta_{n-2} = 1$.

On the other hand, if $x = j-2$, the temporal cluster is $[2, k+1]$. Since $k < n-1$ the upper limit can reach, at the most, the observation $n-1$. Then, $\delta_{k+1} = \delta_1 = 0$. Therefore, there are $k-1$ $\delta_i's$ equal to 1, say, $\delta_2 = \ldots =\delta_k = 1$. It follows that this temporal cluster occurs with probability $p^{k-1}(1-p)^2$.

All temporal clusters such that $k-(n-j+1)+1 < x < j-2$ have the same pattern observed for $x = k-(n-j)$ and $x = j-2$. Therefore, there are $(j-2)-[k-(n-j)]+1 = n-(k+1)$ temporal clusters with such pattern. It follows that $P(N_j = k|p) = 2p^{k-1}(1-p)+$ $[n-(k+1)](1-p)^2p^{k-1}$ if $(n-j+1) \le k < n$. The proof is concluded by noticing that, for $k = n$, $\delta_1 = \ldots = \delta_{n-1} = 0$. Thus, $P(N_j = k|p) = (1-p)^{k-1}$ if $k = n$.

References

Banerjee, S., Carlin, B. P. and Gelfand, A. E., *Hierarchical Modeling and Analysis for Spatial Data*, Chapman & Hall, New York (2004).

Barry, D. and Hartigan, J., Product partition models for change point problems, *The Annals of Statistics* **20**(1), 260–279 (1992).

Barry, D. and Hartigan, J., A Bayesian analysis for change point problems, *Journal of the American Statistical Association* **88**(421), 309–319 (1993).

Besag, J., York, J., and Mollié, A. Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, **43**, 1–20 (1991).

Booth, J. G., Casella, G. and Holbert, J. P., Clustering using objective functions and stochastic search, *Journal of the Royal Statistical Society B* **70**, 119–139 (2008).

Bormetti, G., De Giuli, M. E., Delpini, D. and Tarantola, C., Bayesian Value-at-Risk with product partition models. *Quantitative Finance* (2010) (DOI:10.1080/14697680903512786).

Carlin, B. P., Gelfand, A. E. and Smith, A., Hierarchical Bayesian analysis of change point problems, *Applied Statististics* **41**, 389–405 (1992).

Chen, C. W. S. and Lee, J. C., Bayesian inference of threshold autoregressive models, *Journal of Time Series Analysis* **16**, 483–492 (1995).

Crowley, E. M., Product partition models for normal means, *Journal of the American Statistical Association* **92**(437), 192–198 (1997).

Demarqui, F. N., Loschi, R. H., and Colosimo, E. A., Estimating the grid of time-points for the piecewise exponential model. *Lifetime Data Analysis* **14,** 333–356, (2008).

Escobar, M. D. and West, M., Bayesian density estimation and inference using mixtures, *Journal of the American Statistical Association* **90**(430),577–588 (1995).

Fearnhead, P., Exact and efficient Bayesian inference for multiple changepoint problems, *Statistics and Computing* **16**(2), 203–213 (2006).

Fearnhead, P. and Liu, Z., Online inference for multiple changepoint problems, *Journal of the Royal Statistical Society, Series B* **69**, 589–605 (2007).

Geweke, J. and Terui, N., Bayesian threshold autoregressive models for nonlinear time series, *Journal of Time Series Analysis* **14**(5), 441–454 (1993).

Girón, F. J., Moreno, E. and Casella, G., Objective Bayesian analysis of multiple changepoints for linear models, *Bayesian Statistics 8*, J.M. Bernardo, M.J. Bayarri, J.O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds. Oxford University Press, 1–27 (2007).

Ghosh, M., Natarajan, K., Stroud, T. W. F., Carlin, B. P., Generalized linear models for small-area estimation. *Journal of the American Statistical Association*, **93**, 273–282 (1998).

Hartigan, J., Partition models, *Communication in Statistics -Theory and Method* **19**(8), 2745–2756, (1990).

Hawkins, D. M., Fitting multiple change-point models to data, *Computational Statistics & Data Analysis* **37**(3), 323–341 (2001).

Hegarty, A. and Barry, D., Bayesian disease mapping using product partition models, *Statistics in Medicine* **27**, 3868–3893 (2008).

Hodges J. S., Carlin, B. P., and Fan, Q. On the Precision of the Conditionally Autoregressive Prior in Spatial Models. *Biometrics*, 59, 317–322 (2003).

Hsu, D. A., A Bayesian robust detection of shift in the risk structure of stock market returns, *Journal of American Statistical Association* **77**(2), 29–39 (1982).

Knorr-Held L., Some remarks on Gaussian Markov random field models for disease mapping. In *Highly Structured Stochastic Systems*, P. Green, N. Hjort and S. Richardson (eds), 260–264. Oxford: Oxford University Press (2003)

Loschi, R. H. and Cruz, F. R. B., Extension to the product partition model: Computing the probability of a change, *Computational Statistics and Data Analysis* **48**(2), 255–268 (2005a).

Loschi, R. H. and Cruz, F. R. B., Bayesian identification of multiple change points in Poisson data, *Advances in Complex Systems* **8**(4), 465–482 (2005b).

Majumdar, A., Gelfand, A. E. and Banerjee, S., Spatio-temporal change-point modeling. *Journal of Statistical Planning and Inference* **130**, 149–166, (2005).

Menzefricke, U., A Bayesian analysis of a change in the precision of a sequence of independent normal random variables at an unknown time point, *Applied Statistics* **30**(2), 141–146 (1981).

Moreno, E., Casella, G. and Garcia-Ferrer, A., An objective Bayesian analysis of the change point problem, *Stochastic Environmental Research and Risk Assessment*, **19**, 191–204 (2005).

Müller, P. and Quintana, F., Random partition models with regression on covariates. *Journal of Statistical Planning and Inference* **140**(10), 2801–2808 (2010).

Quintana, F. A., A predictive view of Bayesian clustering, *Journal of Statistical Planning and Inference* **136**(8), 2407–2429 (2006).

Quintana, F. A. and Iglesias, P. L., Bayesian clustering and product partition models, *Journal of the Royal Statistical Society B* **65**(2), 557–574 (2003).

Robert, C.P., *The Bayesian Choice: From decision-theoretic foundations to computational implementation*, Second edition, Springer, New York (2007).

Rue, H., and Held, L. *Gaussian Markov Random Fields: Theory and Applications*. New York: Chapman & Hall (2005).

Ruggeri, F. and Sivaganesan, S., On modeling change points in non-homogeneous Poisson processes, *Statistical Inference for Stochastic Processes* **8**, 311–329 (2005).

Smith, A. F. M., A Bayesian approach to inference about a change-point in a sequence of random variables, *Biometrika* **62**(2), 407–416 (1975).

Tarantola, C., Consonni, G. and Dallaportas, P., Bayesian clustering for row effects models, *Journal of Statistical Planning and Inference* **138**, 2223–2235 (2008).

Ward, H. H., *Police reform in Latin America: Current efforts in Argentina, Brazil and Chile*, Unpublished manuscript. Woodrow Wilson Center for Scholars, Latin America Program, December, (2000).

Yao, Y. C., Estimation of a noisy discrete-time step function: Bayes and empirical Bayes approaches, *The Annals of Statistics* **12**(4), 1434–1447 (1984).

Acknowledgments

We would like to express our gratitude to the editor and two anonymous referees. Their constructive criticisms and suggestions contributed definitively to improvement of the paper. In particular, we thank the referee that pointed out the problems with the use of improper priors in hierarchical change point problems. We also thank professors Sudipto Banerjee and Bradley P. Carlin for comments and suggestions in earlier versions of this paper. J. V. D. Monteiro was supported in part by NFS/DMS grant 1106609. R.M. Assunção received support from CNPq and FAPEMIG, Brazilian agencies for research support. R.H.Loschi's research was partially supported by CNPq, grants 473163/2010-1, 306085/2009-7, 3004505/2006-4.