Semi-parametric Bayesian change-point model based on the Dirichlet process

Gianluca Mastrantonio

1Department of Mathematical Science, Politecnico di Torino & International Association for Research Seismic Precursors (iAReSP)

Abstract

In this work we introduce a semi-parametric Bayesian change-point model, defining its time dynamic as a latent Markov process based on the Dirichlet process. We treat the number of change point as a random variable and we estimate it during model fitting. Posterior inference is carried out using a Markov chain Monte Carlo algorithm based on a marginalized version of the proposed model.

The model is illustrated using simulated examples and two real datasets, namely the coal-mining disasters, that is a widely used dataset for illustrative purpose, and a dataset of indoor radon recordings. With the simulated examples we show that the model is able to recover the parameters and number of change points, and we compare our results with the ones of the-state-of-the-art models, showing a clear improvement in terms of change points identification. The results obtained on the coal-mining disasters and radon data are coherent with previous literature.

1 Introduction

A change-point model is a mixture-type model used to infer changes in a time series subjected to random shifts in its characteristics/features. This means that the data can be broken down into segments and each segment follows a statistical model with different parameters. The time when a segment ends is called change point and the segment is often referred to as regime or state. The inference based on change-point models focuses on two major issues: i) the estimate of number and locations of the change-points; ii) the choice of the best statistical model for each segment.

The change-point literature, starting from [1], and [2], is by now fairly extensive in both frequentist [3, 4] and Bayesian framework [5, 6, 7]. In the former model estimation can be difficult since the likelihood function becomes rapidly intractable as the number of change points increases [for a discussion see ?]. On the other hand, in the more recently developed Bayesian models, the estimation procedures, generally based on Markov chain Monte Carlo (MCMC) algorithms, is always feasible, raising attention to this modelling approach. Among the existing Bayesian models, the most commonly used is the one proposed by [8] [see for example [9, 10, 11]].

In [12] a time series is modelled introducing a latent realization of a discrete time series, that denotes the regime membership, with temporal evolution ruled by a first order Markov process. The change-point model is then obtained assuming a transition matrix constrained so that regimes are visited in a non-reversible sequence; the model can then be seen as a constrained hidden Markov model (HMM) [for an extensive introduction on the HMM, see ?].

In [13], the Bayes factor is used to assess the number of segments through an off-line procedure. Informational criteria, such as the Bayes factor, AIC and BIC, has been criticized [14] since they often suggest different models and it is not always clear which one is the most trustworthy. In a Bayesian setting, we can replace the information criteria with a fully probabilistic on-line model choice, that
can be based on the reversible-jumps [? ] or Dirichlet process (DP) [? ].

The reversible-jump is a Markov chain Monte Carlo (MCMC) algorithm that simulates from posterior distributions defined on spaces of varying dimensions and it can be used to perform model choice. Its implementation requires a mapping function between model parameters that is not always straightforward to define and it has a great impact on the ability of the MCMC to explore the target distribution [? ]. On the other hand, the DP can be used as a prior for an infinite set of parameters, it allows to perform model choice in mixture-based models [? ] and, generally, it leads to MCMC algorithms straightforward to implement.

In this work we propose a semi-parametric extension of [? ] based on the DP, which address issue i) in a fully probabilistic setting, allowing an on-line model choice, while the second issue is left to future developments and considered out of the scope of this work.

Prior to this work, [? ] and [? ] dealt with [? ] extensions DP based. Both of them have flaws that make their use problematic. In [? ], as also noted by [? ], no temporal evolution in the latent allocation dynamic is considered, a regime can always be revisited and the model reduces to a mixture. In [? ] there is not a clear and rigorous formalization of the underlying DP, there are incorrect computations of some full conditionals and the proposed MCMC algorithm updates the latent allocations in a way that easily leads to the identification of the wrong number of regimes (more details on these issues are given in the Appendix). The model of [? ] is close to our proposal and then, together with the one of [? ], are considered as our main competitor.

In this paper we explain how to use the DP to build a semi-parametric extension of [? ], giving a rigorous formalization of the entire procedure. Semi-parametric HMMs based on the DP has been previously proposed, see for example by [? ] and [? ], but here, due to the peculiar transition matrix, these approaches cannot be used. We propose to use the DP to obtain countably infinite distributions, each one with only two possible outcomes and where the probabilities of the outcomes are related to the stick-breaking weights [? ]. This approach allows us to treat the number of segments as random and to estimate it during model fitting. Our specification of the model induces issues in the regime labeling that are solved by using a collapsed Gibbs sampler [? ] that marginalizes over the DP weights; the sampling algorithm is partially based on an birth and death MCMC.

Our proposal is applied to simulated datasets and two real ones. The formers are used to show how the proposed MCMC algorithm is able to recover model parameters, number and positions of the change points. Our results are compared with the ones of [? ] and [? ] and we show that a great improvement in terms of change points identification is achieved. The models are then applied to one of the most used test-dataset in change-point studies, i.e. the coal-mining disasters data [see for example ? ? ]. The results we obtain are consistent with the one of [? ], but, under our model, we are able to give a measure of uncertainly on the number of latent change points. In the last example a time series of Italian indoor radon measurement [? ] is analyzed. Radon emissions are characterized by a non-stationary temporal pattern with periodic components [? ] at different time scales [? ] and changes in mean, variability and trend. Radon concentration is considered a possible earthquake precursor [? ] since have been observed that, prior to strong earthquakes, abrupt changes in the time series characteristics occur. The segmentation of radon data is a first step to try to understand its connection with geodynamic activity. To the best of our knowledge, in the literature have never been proposed a model-based method to segment a radon time series while, for example, wavelet transformation [? ] and testing procedures [? ] have been exploited. We show that our model identifies reasonable change points and with sojourn time in a regime of about a day, that was also observed in previous studies [see for example ? ], proving that change-point models can be used to infer changes in a radon time series.

The paper is organized as follows. In Section 2.1 we introduce the DP, In Section 2.2 we formalize the model of Chib and in Section 2.3 we show our proposal. The MCMC algorithm is shown in Section 3 while Section 4 contains the simulated and real data examples. The paper ends with a discussion in Section 5. In the Appendix we highlight what we believe are the problematic aspects and unclear points of the model and MCMC implementation proposed by [? ].
2 The semi-parametric change-point model

Before the model specification, we introduce the DP.

2.1 The Dirichlet process

The DP is a stochastic process defined over a measurable space $(\Theta, \mathcal{B})$ and it is a random probability measure on a space of distribution functions, i.e. a drawn from a DP is a random discrete distribution, it depends on a scaling parameter $\beta > 0$ and a base distribution $H$ over $\Theta$; the density of $H$ will be indicated with $h(\cdot)$. By definition $G$ is DP distributed with parameters $(\beta, H)$, i.e. $G|\beta, H \sim DP(\beta, H)$, if for any finite partition $\{A_k\}_{k=1}^K$ of $\Theta$ such that $\cup_{k=1}^K A_k \equiv \Theta$ and $A_k \cap A_k' = \{\emptyset\}$ if $k \neq k'$, we have

\[
\begin{align*}
(G(A_1), G(A_2), ... G(A_K))|\beta, H & \sim \\
Dir(\beta H(A_1), \beta H(A_2), ..., \beta H(A_K)),
\end{align*}
\]

where $Dir(\cdot, \cdot, \cdot)$ indicates the Dirichlet distribution. Since

\[
\begin{align*}
(G(A), 1 - G(A))'|\beta, H & \sim \\
Dir(\beta H(A), \beta (1 - H(A))) & \equiv B(\beta H(A), \beta (1 - H(A)));
\end{align*}
\]

where $B(\cdot, \cdot)$ is the beta distribution, mean and variance of $G(A)$ can be easily computed:

\[
\begin{align*}
E(G(A)) = H(A), & \quad Var(G(A)) = \frac{H(A)(1 - H(A))}{\beta + 1}. \quad (1)
\end{align*}
\]

From (1) we see that $H$ is the expected shape of $G$ while $\beta$ controls the degree of variability.

[? ] gives an explicit representation of $G$, that is called the stick-breaking process or stick-breaking representation; If

\[
G = \sum_{k \in \mathbb{N}} \tau_k \delta_{\theta_k},
\]

is DP distributed, then

\[
\begin{align*}
\pi_k & \sim B(1, \beta), \\
\tau_k & = \frac{k - 1}{k} \prod_{l=1}^{k-1} (1 - \pi_l), \\
\theta_k & \sim H,
\end{align*}
\]

(2)

where $\delta$ is a point mass function, $\{\tau_k\}_{k \in \mathbb{N}}$ is the set of weights and $\{\theta_k\}_{k \in \mathbb{N}}$ the set of atoms of the DP. Notice that $\tau_k > 0$, $\sum_{k \in \mathbb{N}} \tau_k = 1$ and $G$ is then a discrete distribution. Sets $\{\tau_k\}_{k \in \mathbb{N}}$ and $\{\pi_k\}_{k \in \mathbb{N}}$ are often written as $\{\tau_{\theta_k}\}_{k \in \mathbb{N}}$ and $\{\pi_{\theta_k}\}_{k \in \mathbb{N}}$ to stress their connection with the DP atoms $\{\theta_k\}_{k \in \mathbb{N}}$. For computational purposes [see for example ? ] a drawn from a DP is frequently parametrized using $\{\tau_k, \theta_k\}_{k=1}^\infty$.

The discrete nature of $G$, with its countably infinite atoms and weights, makes the use of the DP convenient to extend semi-parametrically mixture-based models, where the couples atom-weight are potential sets of parameters (the atoms) and mixture probabilities (the weights); details can be found in [? ], [? ], [? ] or [? ].

2.2 The model of [? ]

In this section we introduce the hierarchical model of [? ]. Let $y = \{y_t\}_{t=1}^T$ be an observed time series. At the first level the conditional density of $y_{|1}, \{y_j\}_{j=1}^{t-1}$ is assumed to depend on a vector of parameters

\[
\text{We assume } \{y_j\}_{j=1}^0 \equiv \{\emptyset\}.
\]
\( \theta_{s_t} \in \Theta \), indexed by a discrete latent random variable \( s_t \in \{1, 2, \ldots, K^*\} \) that indicates the regime membership, i.e. if \( s_t = k \) then \( y_t \) belongs to the \( k^{th} \) regime: notice that \( \theta_{s_t} \equiv \theta_k \) if \( s_t = k \). At the second level \( \{s_t\}_{t=1}^T \) is a Markov process, with starting point \( s_1 = 1 \), ruled by a \( K^* \times K^* \) constrained one-step ahead transition matrix

\[
P = \begin{pmatrix}
\pi_1 & 1 - \pi_1 & 0 & 0 & \cdots & 0 \\
0 & \pi_2 & 1 - \pi_2 & 0 & \cdots & 0 \\
0 & 0 & \pi_3 & 1 - \pi_3 & \cdots & 0 \\
0 & 0 & 0 & \pi_4 & \cdots & 0 \\
\cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\
0 & 0 & 0 & 0 & \cdots & 1 \\
\end{pmatrix}.
\] (3)

Since the lower diagonal elements of \( P \) are zeros, a regime left cannot be visited again.

Let \( f(\cdot) \) indicate a density function and \( I(\cdot, \cdot) \) the indicator function, we can then write

\[
f(s_t | s_{t-1} = k, \{\pi_k\}_{k=1}^{K^*}) = 
\pi_k I(s_t, k) + (1 - \pi_k)I(s_t, k+1), \ t = 2, \ldots, T,
\]
\[
s_1 = 1,
\]
where it is assumed that \( s_t \in \{k, k+1\} \) if \( s_{t-1} = k \).

\[\text{[? ] assumes beta distributions with the same set of parameters for all the elements of } P \text{ and then,}
\]
\[\text{letting } H \text{ be a prior distribution, the model can be written as}
\]
\[
f(y|\{\theta_k\}_{k=1}^{K^*}, \{s_t\}_{t=1}^T) = \prod_{t=1}^T \prod_{k=1}^{K^*} f(y_t | \{y_j\}_{j=1}^{t-1}, \theta_k) I(s_t, k),
\]
\[
f(s_t | s_{t-1} = k, \{\pi_k\}_{k=1}^{K^*}) = 
\pi_k I(s_t, k) + (1 - \pi_k)I(s_t, k+1), \ t = 2, \ldots, T,
\]
\[
s_1 = 1,
\]
\[
\pi_k \sim B(\alpha, \beta), \ k = 1, \ldots, K^*,
\]
\[
\theta_k \sim H, \ k = 1, \ldots, K^*.
\]

The model described above can be seen as an HMM with constrained transition matrix. The number of rows of \( P \), that is equal to the number of regimes, must be set \textit{a priori} (see equation (3)) and an off-line procedure is needed to assess the value of \( K^* \). With our proposal we are going to extend the model of [? ] allowing an on-line model choice.

### 2.3 The semi-parametric extension

Our extension starts with the introduction of an equivalent specification of the model of Chib that is obtained by substituting the latent process \( \{s_t\}_{t=1}^T \) with \( \{\psi_t \in \Theta\}_{t=1}^T \), assuming the following time evolution:

\[
f(\psi_t | \psi_{t-1} = \theta_k, \{\pi_k\}_{k=1}^{K^*}) \sim 
\pi_k I(\psi_t, \theta_k) + (1 - \pi_k)I(\psi_t, \theta_{k+1}), \ t = 2, \ldots, T,
\]
\[
\psi_1 = \theta_1.
\] (6)

assuming \( \psi_t \in \{\theta_k, \theta_{k+1}\} \) if \( \psi_{t-1} = \theta_k \).

Notice that equations (4) and (5) are equivalent to equations (6) and (7) since \( f(\psi_t = \theta_k | \psi_{t-1} = \theta_{k'}) = f(s_t = k | s_{t-1} = k') \) and \( \psi_t = \theta_k \) if and only if \( s_t = k \).

Semi-parametric extensions for mixture-based models are generally defined by taking \( K^* \to \infty \) and assuming a DP based prior for the probability structure of \( \{\psi_t\}_{t=1}^T \) [see for example \( ? \ ? \ ? \)]. Here
we propose the following. First notice that each row of \( P_\theta \) sums to 1, i.e. is a vector of probabilities, with only two non-zero values. We assume \( G = \sum_{k=1}^{\infty} \tau_k \delta_{\theta_k} \sim DP(\beta, H) \) and we define distributions \( G_{\theta_k}'s \), with \( k \in \mathbb{N} \), as follows:

\[
G_{\theta_k} = \frac{\tau_k}{1 - \sum_{l=1}^{k-1} \tau_l} \delta_{\theta_k} + \left( 1 - \frac{\tau_k}{1 - \sum_{l=1}^{k-1} \tau_l} \right) \delta_{\theta_{k+1}}, \quad k \in \mathbb{N}.
\] (8)

In our model \( G_{\theta_k} \) is used as distribution for the \( k^{th} \) row of \( P_\theta \). Notice that \( \frac{\tau_k}{1 - \sum_{l=1}^{k-1} \tau_l} \) is equal to the beta distributed weight \( \pi_k \) (see equation (2)), and then (8) can be written equivalently as

\[
G_{\theta_k} = \pi_k \delta_{\theta_k} + (1 - \pi_k) \delta_{\theta_{k+1}}, \quad k \in \mathbb{N},
\] (9)

where by definition, see Section 2.1,

\[
\pi_k \sim B(1, \beta),
\]

\[
\theta_k \sim H.
\] (10)

We have then distributions based on the DP, one for each row of the infinite-dimensional transition matrix \( P_\theta \). Notice that the atoms in the regimes are tied by construction, i.e. the atom of \( [P_\theta]_{i,i+1} \) is equal to the one of \( [P_\theta]_{i+1,i+1} \). We can then write

\[
\psi_t | \psi_{t-1} = \theta_k, \{ \theta_k, \pi_k \}_{k \in \mathbb{N}} \sim G_{\theta_k}, \quad t = 2, \ldots, T,
\]

\[
\psi_1 = \theta_1.
\]

The model is then

\[
f(y | \{ \psi_t \}_{t=1}^T) = \prod_{t=1}^{T} f(y_t | \{ y_j \}_{j=1}^{t-1}, \psi_t),
\]

\[
\psi_t | \psi_{t-1} = \theta_k, \{ \theta_k, \pi_k \}_{k \in \mathbb{N}} \sim G_{\theta_k}, \quad t = 2, \ldots, T,
\]

\[
\psi_1 = \theta_1,
\]

\[
\pi_k | \beta \sim B(1, \beta), \quad k \in \mathbb{N},
\]

\[
\theta_k | H \sim H, \quad k \in \mathbb{N},
\]

or, introducing the discrete time series \( \{ s_t \}_{t=1}^T \), it can be equivalently stated as

\[
f(y | \{ \theta_k \}_{k \in \mathbb{N}}, \{ s_t \}_{t=1}^T) = \prod_{t=1}^{T} \prod_{k \in \mathbb{N}} f(y_t | \{ y_j \}_{j=1}^{t-1}, \theta_k)^I(s_t, k),
\]

\[
f(s_t | s_{t-1} = k, \{ \pi_k \}_{k \in \mathbb{N}}) = \pi_k I(s_t, k) + (1 - \pi_k) I(s_t, k + 1), \quad t = 2, \ldots, T,
\]

\[
s_1 = 1,
\]

\[
\pi_k | \beta \sim B(1, \beta), \quad k \in \mathbb{N},
\]

\[
\theta_k | H \sim H, \quad k \in \mathbb{N}.
\]

This model is an infinite-dimensional extension of the one shown at the end of Section 2.2.

As in the standard DP based mixture models, the number \( K \) of unique values that \( \psi \) (or \( s \)) assumes, is used as an estimate of the number of segments of the observed time series. Notice that \( H \) acts as the prior distribution of \( \theta_k \).
3 The MCMC algorithm

From equation (9) and matrix \( P_0 \) we see that regimes are visited in increasing order, e.g. after regime \( k \), regime \( k + 1 \) is visited and this can produce an inefficient MCMC algorithm. Then, to avoid the problem, we marginalized over the vector of DP weights. This strategy is often adopted [see for example ? ? ? ] since the resulting process defines a prior over a partition of the data that no more depends on the labels. Let \( n_{i,j} = \sum_{t=1}^{T-1} \delta(s_t, i) \delta(s_{t+1}, i) \), that is the number of self-transitions in the \( i^{th} \) regime between time \( j \) and \( j' \). After marginalization we obtain the following for the dynamic of \( s_t \):

\[
f(s_t = i | s_{t-1} = k, s_{t-1}, \ldots, s_1, \beta) =
\begin{cases}
\frac{n_k^{1(t-1)} + 1}{n_k^{1(t-1)} + 1 + \beta} & \text{if } i = k, \\
\frac{n_k^{1(t-1)} + \beta}{n_k^{1(t-1)} + 1 + \beta} & \text{if } i = k + 1,
\end{cases}
\]

\( s_1 = 1 \).

We want to remark that now regimes are visited in increasing order only to simplify the notation, but any regimes re-labeling are equivalent.

The conditional distribution of \( s_t \) depends on the count \( n_k^{1(t-1)} \) and parameter \( \beta \) and the process \( s_t \) is no more Markovian. The probability of \( s_t = k | s_{t-1} = k, s_{t-1}, \ldots, s_1, \beta \), i.e. \( s_t \) assumes the same value of \( s_{t-1} \), increases with \( n_k^{1(t-1)} \) meaning that, if at time \( t \) an observation is allocated to the previously observed regime \( k \), at time \( t + 1 \) the probability to belong to the same regime increases; i.e. the process has the self reinforcement property [? ]. Parameter \( \beta \) can be interpreted noticing that when there is only one observation in the \( k^{th} \) regime, i.e. \( n_k^{1(t-1)} = 0 \), the odd to move to a new regime at time \( t + 1 \) is \( \beta \).

The model is then

\[
f(y | \{ \theta_k \}_{k \in \mathbb{N}}, \{ s_t \}_{t=1}^T) = \prod_{t=1}^T \prod_{k \in \mathbb{N}} f(y_t | \{ y_j \}_{j=1}^{t-1}, \theta_k)^{(s_t, k)},
\]

\[
f(s_t = i | s_{t-1} = k, s_{t-2}, \ldots, s_1, \beta) =
\begin{cases}
\frac{n_k^{1(t-1)} + 1}{n_k^{1(t-1)} + 1 + \beta} & \text{if } i = k, \\
\frac{n_k^{1(t-1)} + \beta}{n_k^{1(t-1)} + 1 + \beta} & \text{if } i = k + 1,
\end{cases}
\]

\( s_1 = 1 \),

\( \theta_k \mid H \sim H, k \in \mathbb{N} \).

Under this setting the MCMC updates of \( \beta \) and \( \theta_k \) are simple and we show, in the next paragraphs, how to implement them. The update of \( \{ s_t \}_{t=1}^T \) will be discussed in more details since it needs to be more carefully implemented to obtain and efficient algorithm.

The update of \( \beta \) Let \( f(\beta) \) be a prior distribution, then the full conditional of \( \beta \) is proportional to \( f(s_1, \ldots, s_T | \beta) f(\beta) \). Using (11) we can find that

\[
f(s_1, \ldots, s_T | \beta) =
\left[ \prod_{i=1}^{K-1} \beta \prod_{j=0}^{n_i^{1:T-1}(1+j)} \frac{\prod_{j=0}^{n_i^{1:T-1}(1+j)} \prod_{j=0}^{n_i^{1:T-1}(1+j)} \Gamma(\beta + 1) \Gamma(n_i^{1:T} + 1)}{\Gamma(n_i^{1:T} + 1 + \beta + 1 - I(i, K))} \right],
\]

and using relation \( a(a+1) \ldots (a + m - 1) = \frac{\Gamma(a+m)}{\Gamma(a)} \), (12) can be expressed as

\[
\beta^{K-1} \prod_{i=1}^{K} \frac{\Gamma(\beta + 1) \Gamma(n_i^{1:T} + 1)}{\Gamma(n_i^{1:T} + 1 + \beta + 1 - I(i, K))}.
\]
The full conditional is normal-inverse gamma.

\[ \beta^{K-1} \prod_{i=1}^{K} \frac{\Gamma(\beta+1)\Gamma(n_i^{1:T} + 1)}{\Gamma(n_i^{1:T} + 1 + \beta + 1) - I(i, K)} f(\beta). \tag{14} \]

To the best of our knowledge, there is not a prior distribution \( f(\beta) \) that let us express (14) in a closed form from which sampling is easy and a sample of \( \beta \) must be draw using a Metropolis-Hastings step.

**The update of \( \theta_k \)** The full conditional of \( \theta_k \) is proportional to

\[ \prod_{t=1}^{T} f(y_t | \{y_j\}_{j=1}^{t-1}, \theta_k) I(s_t, k) h(\theta_k), \tag{15} \]

The functional form of (15) depends on how we specify \( f(y_t | \{y_j\}_{j=1}^{t-1}, \theta_k) \) and \( H \). As an example, if

\[ f(y_t | \{y_j\}_{j=1}^{t-1}, \theta_k) \equiv f(y_t | \theta_k), \]

with \( Y_t | \theta_k \sim N(\mu_k, \sigma_k^2) \) and \( \theta_k = \{\mu_k, \sigma_k^2\} \), then if \( H \) is the product of a normal distribution over \( \mu_k \) and inverse gamma over \( \sigma_k^2 \), likelihood and prior are conjugate and the full conditional is normal-inverse gamma [?].

### 3.1 The update of \( \{s_t\}_{t=1}^{T} \)

It is generally preferable to update jointly as many random variables as possible [?]. Unfortunately, we are unable to find a way to sample from the joint full condition of \( \{s_t\}_{t=1}^{T} \) and then a different approach must be used. A simple solution is the univariate update of each component \( s_t \) but, experimenting with simulated data, we notice that this leads to unsatisfactory results in terms of MCMC chain mixing since, for example, redundant states with similar \( \theta_k \)'s are created and the distribution of \( K \) is generally entirely concentrated on a single value. We solved the aforementioned problems by combining the univariate update with other updates:

- **the split update** (or birth move) - we propose a new change point at time \( t \);
- **the merge update** (or death move) - we propose to merge consecutive regimes.

At each MCMC iteration only one of them is performed, choosing randomly with assign probabilities. We assume that before the MCMC update of \( s_t \) is performed, it have value \( k \) and, to simplify notation, after each MCMC step regimes are relabelled so to \( s_1 = 1 \) and \( s_t \in \{s_{t-1}, s_{t-1} + 1\} \).

**The single-component update** Let \( n_i^{-t} = n_i^{1:(t-1)} + n_i^{(t+1):T} \), i.e. the number of self transition in the \( i^{th} \) regime without taking into account transitions that involve \( s_t \), and let \( * \) indicates a new regime. We have to sample \( s_t \) only if \( s_t \neq s_{t-1} \) or \( s_t \neq s_{t+1} \), otherwise \( s_t = s_{t+1} = s_{t-1} \) with probability 1, then

- with probability proportional to

\[ \frac{\beta}{\beta + 1} f(y_t | \{y_j\}_{j=1}^{t-1}, \theta_*) \]

\( s_t = *; \)

- if \( t \neq 1, s_t \) can be equal to \( s_{t-1} \) with probability proportional to

\[ \frac{n_{s_{t-1}}^{-t} + 1}{n_{s_{t-1}}^{-t} + 1 + \beta + 1} f(y_t | \{y_j\}_{j=1}^{t-1}, \theta_{s_{t-1}}); \]

\[ \text{or} \]

\[ \frac{n_{s_{t-1}}^{-t} + 1}{n_{s_{t-1}}^{-t} + 1 + \beta + 1} f(y_t | \{y_j\}_{j=1}^{t-1}, \theta_{s_{t-1}}); \]

\[ \text{or} \]

\[ \frac{n_{s_{t-1}}^{-t} + 1}{n_{s_{t-1}}^{-t} + 1 + \beta + 1} f(y_t | \{y_j\}_{j=1}^{t-1}, \theta_{s_{t-1}}); \]
• if $t \neq T$, $s_t$ can be equal to $s_{t+1}$ with probability proportional to
\[
\frac{n_{s_{t+1}} - 1}{n_{s_{t+1}} + 1 + \beta + 1 - I(s_{t+1}, K)} f(y_t | \{y_j\}_{j=1}^{t-1}, \theta_{s_{t+1}});
\]

• if $n_k^{1:T} = 0$, then $s_{t-1} \neq k \neq s_{t+1}$, and $s_t$ can be equal to $k$ with full conditional
\[
\propto \frac{\beta}{\beta + 1} f(y_t | \{y_j\}_{j=1}^{t-1}, \theta_k).
\]

**The split update**  Let $S_- = \{ s_{t'} : s_{t'} = k, t' < t \}$ and $S^+ = \{ s_{t'} : s_{t'} = k, t' \geq t \}$, let $n_{S_-}$ and $n_{S^+}$ be the number of self transitions in the two subsets and let
\[
\gamma_c(n) = \frac{\Gamma(\beta + 1)\Gamma(n + 1)}{\Gamma(n + 1 + \beta + 1 - c)};
\]
then

• $s_t = k$ for all $s_t \in S_- \cup S^+$ with probability
\[
\propto \gamma_{l(k,K)}(n_k^{1:T}) \prod_{t:s_t \in S_- \cup S^+} f(y_t | \{y_j\}_{j=1}^{t-1}, \theta_k);
\]

• $s_t = *$ for all $s_t \in S_-$ and $s_t = k$ for all $s_t \in S^+$ with probability
\[
\propto \beta \gamma_0(n_{S_-}) \gamma_{l(k,K)}(n_{S^+}) \times \prod_{t:s_t \in S_-} f(y_t | \{y_j\}_{j=1}^{t-1}, \theta_*) \prod_{t:s_t \in S^+} f(y_t | \{y_j\}_{j=1}^{t-1}, \theta_k);
\]

• $s_t = k$ for all $s_t \in S_-$ and $s_t = *$ for all $s_t \in S^+$ with probability
\[
\propto \beta \gamma_0(n_{S_-}) \gamma_{l(k,K)}(n_{S^+}) \times \prod_{t:s_t \in S_-} f(y_t | \{y_j\}_{j=1}^{t-1}, \theta_k) \prod_{t:s_t \in S^+} f(y_t | \{y_j\}_{j=1}^{t-1}, \theta_*).
\]

**Merge update**  Let $S_j = \{ s_t : s_t = j \}$, then, for $k = 1, \ldots, K$:

• $s_t = *$ for all $s_t$ in $S_k$ with probability
\[
\beta \gamma_0(n_{k-1,k-1})^{1-I(k,1)} \gamma_0(n_{k,k-1}) \gamma_{l(k,K)}(n_{k+1,k+1})^{1-I(k,K)} \times \prod_{t:s_t \in S_k} f(y_t | \{y_j\}_{j=1}^{t-1}, \theta_*);
\]

• if $k \neq 1$, then $s_t = k - 1$ for all $s_t$ in $S_k$ with probability
\[
\gamma_0(n_{k-1,k-1}^{1:T} + n_{k,k}^{1:T}) \gamma_{l(k,K)}(n_{k+1,k+1}^{1:T})^{1-I(k,K)} \times \prod_{t:s_t \in S_k} f(y_t | \{y_j\}_{j=1}^{t-1}, \theta_{k-1});
\]

• if $k \neq K$, then $s_t = k + 1$ for all $s_t$ in $S_k$ with probability
\[
\gamma_0(n_{k+1,k+1}^{1:T} + n_{k,k}^{1:T}) \gamma_{l(k,K)}(n_{k,k}^{1:T} + n_{k+1,k+1}^{1:T} + 1) \times \prod_{t:s_t \in S_k} f(y_t | \{y_j\}_{j=1}^{t-1}, \theta_{k+1});
\]
s_t = k for all s_t in S_k with probability
\begin{equation}
\beta \gamma \sigma_0(n_{k+1,k+1}^{1:T})^{-1-t(k,K)} \gamma(n_{k,k}^{1:T}) \times 
\prod_{t:s_t \in S_k} f(y_t|y_{t-1}, \theta_k).
\end{equation}

MCMC mixing is improved if, choosing randomly, the univariate and split updates are performed starting from the first to the last time or from the last to the first.

4 Examples

In this section we compare the results of our model with the ones of \cite{?} and \cite{?} on simulated datasets and real ones. Using simulated datasets we test the ability of the models in recovering the right number of latent regimes and parameters, then the models are estimated on a standard change-point problem, and real ones. Using simulated datasets we test the ability of the models in recovering the right number of coal-mining disasters, and on the radon data. We implement the model of \cite{?}, introduced in Section 2.2, assuming prior (10) for the transition probabilities and the same priors over \{\alpha, \beta, \sigma^2\} that reduces to our specification if \alpha = 1, see equation (11). Then, in this section, the model of \cite{?} is implemented using the MCMC algorithm that the authors proposed, assuming \alpha = 1 and under the same priors over \beta and likelihood parameters of our proposal; model choice is performed using BIC. The model of \cite{?}, with respect to our approach, starts from a different specification of the latent process \(s_t\) (for details see the Appendix) with

\begin{equation}
f(s_t = i | s_{t-1} = k, s_1, \ldots, s_{t-1}, \beta, \alpha) =
\begin{cases}
\frac{\partial n_k^{1:(t-1)+\alpha}}{\partial n_k^{1:(t-1)+\alpha + \beta}} & \text{if } i = k, \\
\frac{\partial \beta}{\partial \beta} & \text{if } i = k + 1,
\end{cases}
\end{equation}

that reduces to our specification if \alpha = 1, see equation (11). Then, in this section, the model of \cite{?} is implemented using the MCMC algorithm that the authors proposed, assuming \alpha = 1 and under the same priors over \beta and likelihood parameters of our proposal. In all the examples posterior inference is carried out using 130000 iterations, burnin 80000 and thin 10, using 5000 posterior values for inferential purposes, with an half normal prior for \beta with variance parameter \(\sigma_\beta^2\). We chose between the single component, split and merge updates with probabilities respectively equal to 0.5, 0.25 and 0.25, while with probability 0.5 we perform the univariate or split update starting from the first to the last time or from the last to the first.

We indicate our proposal as Model1, the model of \cite{?} as Model2 and the one of \cite{?} as Model3.

The \(R\) \cite{?} source codes that can be used to replicate the results of the simulated and coal-mining disasters examples are available online in a GitHub repository\footnote{https://github.com/GianlucaMastrantonio/Change-Point} while, due to a confidentiality issue, only the \(R\) functions used to analyze the radon example are available.

4.1 Simulated data

We simulate datasets under two schemes, both with \(T = 1500\), 7 regimes and assuming conditional independence between the \(y_i\)’s and with \(Y_t|\theta_k \sim N(\mu_k, \sigma^2_k)\). In the first set the change points are \(\xi = \{\xi_k\} = \{50, 250, 650, 750, 1000, 1400, 1500\}\), and the parameters are \(\mu = \{\mu_k\}_{k=1}^7 = \{0, 5.2, -2, 0, 2, 10\}\) and \(\sigma^2 = \{\sigma^2_k\}_{k=1}^7 = \{1, 2, 1, 0.5, 1, 3, 5\}\) while in the other \(\xi = \{50, 250, 900, 950, 1100, 1400, 1500\}\) with \(\mu = \{0, 5.2, 2, 2, 2, 10\}\) and \(\sigma^2 = \{1, 2, 1, 0.1, 1, 15, 5\}\). For each scheme 100 datasets are simulated; two examples of simulated time series, one for each scheme, are plotted in Figures 1 and 2.

The set of parameters of the first scheme are chosen so to have regimes of short (1 and 4) and long (3 and 6) length, overlapping distributions on adjacent regimes (2-3 and 4-5-6), well separated ones (1-2 and 3-4) and different values of variability. In the second scheme we are mainly interested in the evaluation of how the models behave when a short regime (the fourth) is in between two regimes (the third and fifth) that have the same density parameters.
We assume a normal prior for $\mu_k$ with 0 mean and variance 1000, while the prior over $\sigma^2_k$ is inverse gamma with shape and rate parameters both equal to 1. Here $\sigma^2_\beta$ is set to be 1000 and through a simulation we evaluated that it induces a prior over $K$ that puts the central 90% of probability mass between 741 and 1477. For Model2, we estimated model with $K^* \in \{4, 5, \ldots, 10\}$.

**First scheme**  Under our proposal the maximum a posteriori (MAP) estimate of $K$ is 7 in 96 datasets, in 3 is 8 and in 1 is 9. We measure the agreement between the true partition and the one found by our model, i.e. the MAP classification, through the Rand Index (RI) [?], that is an index that ranges between 0 and 1, with 0 indicating that the partitions, the true one and the MAP, do not agree in any pair of points and 1 in case of perfect agreement [for details see ?]. Among datasets, the minimum value reached by the RI is 0.986 and in 27 of them is exactly 1. Model2 identifies 7 regimes in 99 datasets and 8 in 1, with minimum RI equal to 0.969, that is lower than the one found by our model, and it is exactly 1 in 27 datasets. On the other hand, Model3 identifies always 1 regime and we will give a justification of this result in the Appendix.

For Model1 and Model2 we show in Tables 1 and 2 the posterior estimates and credible intervals (CIs) for the simulated dataset depicted in Figure 1. Results under Model1 and Model2 are quite similar and both estimate well the parameters, i.e. the true values are inside the associated CIs, but only with our proposal we have an estimate of the posterior distribution of $K$, shown in Figure 3, and we can evaluate the uncertainty on the estimated number of regimes. The CI of $\beta$ is $[0.076, 0.366]$ for Model1 and $[0.0730, 0.352]$ for Model2.

**Second scheme**  Under Model1, the MAP estimate of $K$ is 7 in 75 datasets, while is 8 in 6, 9 in 1, 6 in 1 and 5 in 17. As in the first scheme, Model3 identifies always 1 regime while Model2 estimates 7 regimes in 23 datasets and 5 in 66. As we expected, when a posterior sample of $K$ is 5, regimes 3, 4 and 5 are generally merged.

In terms of RI Model1 and Model2 have similar minimum and maximum values, i.e. for both the maximum is 0.999 while Model1 has minimum 0.717 and Model2 0.714. The main difference is in the distribution of the RI across simulated datasets, since our proposal has median RI that is equal to
0.995 while the model of [?] has median value 0.725, showing that our proposal tends to perform better.

For the simulated dataset plotted in Figure 2, we show in Tables 3 and 4 the parameters estimates for Model 1 and Model 2, where we can see a good agreement between the posterior CIs; for both models the values $\sigma_3^2$ and $\sigma_4^2$ are the only one not inside the associated CIs. $\beta$ has CI [0.0778 0.370] in Model 1 and [0.0760 0.369] under Model 2. The posterior distribution of $K$ is shown in Figure 4.

### 4.2 Coal-mining disasters

Our first real data application is devoted to the analysis of one of the most analysed dataset in change-point literature [see ? ], the annual number of coal-mining disasters in Britain, during the period 1851-1962. Here $y_t \in \mathbb{N}, \theta_k = \lambda_k$ and, following [? ], we assume $f(y_t | \{y_j\}_{j=1}^{t-1}, \lambda_k) = f(y_t | \lambda_k)$ with $Y_t | \lambda_k \sim \text{Pois}(\lambda_k)$. The data are plotted in Figure 5.

We assume $\lambda_k \sim G(2,1)$ while the variance parameter of the half normal prior on $\beta$ is set to 0.1 and, again through a simulation, we evaluated that this induces a prior over $K$ that puts the central 90% percent of total mass of probabilities between 1 and 10. On this dataset the Model 2 is tested with $K$ between 1 and 4.

The MAP estimate of $K$ is 1 under Model 3, i.e. there is not a segmentation, while Model 2 chooses $K = 2$ that is also the value found in [? ]. Our proposal has MAP estimate of $K$ equal to 2 and the associated posterior distribution is showed in Figure 6. We wanted to point out that since our proposal has a non-parametric specification of the latent allocation structure, it is not surprising that with little data, as in this example, the posterior of $K$ has a lot of variability.

For Model 1, posterior mean estimates of $\lambda_1$ and $\lambda_2$ are $\hat{\lambda}_1 = 3.045$ and $\hat{\lambda}_2 = 0.923$ with, respectively, CIs [2.544 3.648] and [0.711 1.166]; on the other hand, under Model 2 they are $\hat{\lambda}_1 = 3.084$ and $\hat{\lambda}_2 = 0.933$ with, respectively, CIs [2.587 3.688] and [0.7223 1.186]. In both models the CI of $\xi_1$ is [1886 1896] with MAP estimate 1890 while the CI of $\beta$ is [0.053 1.017] for Model 1 and [0.025 0.45] for Model 2.
4.3 Indoor radon data

Radon is a colorless and odorless inert noble gas generated by the radioactive decay of Radium (Ra226) in the decay chain of Uranium (U238) [? ]. A time series of radon concentration is characterized by daily and annual periodic components with about daily changes in mean level, variance and temporal trend [? ]. Radon concentration is considered a proxy of geodynamic activity since many authors [see ? ? , and references therein] proved that prior to a powerful geodynamic event, such as an earthquake, a radon time series can show abrupt and out of ordinary changes [? ? ]; this connection between radon anomalies and geodynamic events makes relevant the understanding of the radon time series dynamic.

In this example we use the radon data owned by the International Association for Research Seismic Precursors (iAReSP) [? ? ]. The iAReSP, with the Tellus project [? ], that consists of a network of radon recording stations, aims to understand what happens to the radon concentration during the phase preceding an earthquake. At the present moment we have data only from one station and on a limited time window. More precisely our data are the mean radon counts over ten minutes, observed between November 18\textsuperscript{th} 2015, 8:00, and November 28\textsuperscript{th} 2015, 17:50, having then 1500 observations. Data are recorded in central Italy, in the town of Pizzoli, close to the city of L’aquila, 803 m above sea level, using an ionization chamber with continuous measurement of Alfa particles produced by the decay of radon stable isotope \textsuperscript{222}RN [? ]. In the observational period no major earthquakes were observed.

Here we show some preliminary results that prove the ability/potentiality of the model in the segmentation of a radon time series. As said in the beginning of this section, radon data presents a daily periodicity that is stable in time [see ? ]; in other words changes in the time series do not affect its amplitude. This characteristic of the radon data cannot be expressed in our model which assumes that all parameters change between regimes and then, prior to the model fitting, we decompose the time series into seasonal, trend and irregular components using the approach of [? ], implemented in the \textit{stl} function of \textit{R} and the (daily) seasonal component is subtracted to eliminate the periodicity. The resulting time series has mean \( \approx 5080.515 \) and variance \( \approx 1124047 \) and, to avoid possible numerical stability problems that such large numbers may raise, we standardize the data; the resulting time series is plotted in Figure 7. To take into account changes in mean level, variance and temporal trend
the following is assumed: $Y_t | \theta_k \sim N(\mu_{0,k} + \mu_{1,k} t, \sigma_k^2)$. Parameters $\mu_{0,k}$ and $\mu_{1,k}$ have normal priors with 0 mean and variance 1000 while $\sigma_k^2 \sim IG(1,1)$. In this example $T$ is 1500, as in the simulated ones, and then we use the same prior on $\beta$.

A posteriori, Model3 estimates only 1 regime while our proposal put 99% of the mass of probability on $K = 8$ and the remaining on $K = 9$, the posterior means and CIs of parameters and change points can be seen in Table 5 while Figure 7 shows the MAP classification. Model2 estimates 8 regimes. The main differences between our proposal and Model2 can be seen in the estimates of parameters and change points of the first two regimes, see Tables 5 and 6. Posterior mean estimates, CIs and change points of the other regimes, are similar. The CI of $\beta$ is [0.078, 0.331] under Model1 and [0.082, 0.368] under Model2.

Model1 and Model2 found a clear and reasonable segmentation of the data showing that there are almost daily changes in the radon emission features. This last finding is coherent with previous studies, see for example [? ].

5 Discussion

In this work we proposed a semi-parametric formalization of the standard change-point model of [? ]. In our extension, the first order latent Markov process, ruled by a constrained one-step transition matrix, is substituted by a stochastic process based on the stick-breaking representation of the DP. We suggested to draw samples from the posterior distribution using a marginalized version of the proposed model and we showed how to compute the full conditionals needed to implement the MCMC algorithm.

To assess the ability of the model in recovering the number and locations of change points, we used simulated examples. Then we make inference in one of the most analysed dataset in the literature and on a new one. We showed that our proposal outperformed the ones of [? ] and [? ] in terms of change points estimates.

The future will find us enriching the model including covariates information and modelling multiple time series subjected to individual and concurrent shifting in their features. We will also use the model to analyzed a longer times series of radon data to possibly acquire an early signal of major earthquake
Table 1: Simulated example - first scheme - Model1: posterior means (ˆ) and credible intervals (CI) of \( \mu_k, \sigma_k^2 \) and \( \xi_k \) computed using the subset of posterior samples that has \( K = 7 \).

### Acknowledgements

The author wishes to thank Giovanna Jona Lasinio and Antonello Maruotti for assistance and comments that greatly improved the manuscript.

### Appendix

Here we discuss the problematic aspects and errors in the computation of the full conditionals in [? ] that, in our opinion, justify the need of a more rigorous formalized model as the one we are proposing in this work.

To obtain a semi-parametric extension of [? ], [? ] substitute the matrix \( P \) with

\[
\hat{P} = \begin{pmatrix}
\pi_{11} & \pi_{12} & \pi_{13} & \ldots & \pi_{1K^*} \\
0 & \pi_{22} & \pi_{23} & \ldots & \pi_{2K^*} \\
0 & 0 & \pi_{33} & \ldots & \pi_{3K^*} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \ldots & \pi_{K^*K^*}
\end{pmatrix},
\]

and they assume

\[
[\hat{P}]_k|\beta \sim Dir\left(\frac{\beta}{K^* - k + 1}, \ldots, \frac{\beta}{K^* - k + 1}\right).
\]

They integrate out the elements of \( \hat{P} \) and taking the limit as \( K^* \) goes to infinity, the following time dynamic for \( s_t \) is obtained:

\[
f(s_t = i|s_{t-1} = k, s_1, \ldots, s_{t-1}, \beta) = \begin{cases}
\frac{n_k^{(t-1)}}{n_k^{(t-1)} + \beta} & \text{if } i = k, \\
\frac{n_k^i}{n_k^{(t-1)} + \beta} & \text{if } i = k + 1.
\end{cases}
\]
Table 2: Simulated example - first scheme - Model2: posterior means (ˆ) and credible intervals (CI) of \( \mu_k \), \( \sigma_k^2 \) and \( \xi_k \).

The authors state that, as \( K^* \rightarrow \infty \), each row of \( \tilde{P} \) follows a DP. Since the DPs are independent it is not clear how the rows share the same set of DP atoms since there must be a way to tie them as in the hierarchical Dirichlet process of [? ], i.e. the atom associated to \( \pi_{ij} \), with \( j \geq i \), must be equal to the one of \( \pi_{hj} \), for all \( j \geq h \). This was also the problem of the infinite hidden Markov model of [? ], that is close to the proposal of [? ], but then the work of [? ] shows how to solve this problem.

The joint density of \( \{s_1, \ldots, s_T|\beta\} \) Equation (17) reduces to our specification if \( \alpha = 1 \), see equation (11). Then the joint density of \( s_1, \ldots, s_T|\alpha = 1, \beta \) derived from [? ] and \( s_1, \ldots, s_T|\beta \) derived in this work must be the same. [? ] write

\[
f(s_t = i|s_{t-1} = k, s_1, \ldots, s_{t-1}, \beta, \alpha) =
\begin{cases}
\frac{n_i^{1:(t-1)} + \alpha}{n_k^{1:(t-1)} + \alpha + \beta} & \text{if } i = k, \\
\frac{\beta}{n_k^{1:(t-1)} + \alpha + \beta} & \text{if } i = k + 1.
\end{cases}
\] (17)

(18) is different from our (13). Equation (18) implicitly assumes the existence of a new observation at time \( T + 1 \) belonging to the \( (K + 1)^{th} \) regime, and if we multiply (13) by \( f(s_{T+1} = K + 1|s_T = K, s_1, \ldots, s_{T-1}) = \frac{\beta}{n_k^{1:T+1} + \alpha + \beta} \), we obtain (18).
\begin{table}[h]
\centering
\begin{tabular}{|c|ccc|}
\hline
& $\hat{\mu}_k$ & $\hat{\sigma}_k^2$ & $\hat{\xi}_k$ \\
& CI & CI & CI \\
1 & 0.101 & 0.715 & 50 \\
& [-0.138 0.336] & [0.493 1.095] & [50 50] \\
2 & 5.008 & 1.974 & 250 \\
& [4.816 5.199] & [1.627 2.420] & [250 251] \\
3 & 1.972 & 1.123 & 900 \\
& [1.891 2.050] & [1.014 1.257] & [900 903] \\
4 & 2.08 & 0.163 & 957 \\
& [1.966 2.188] & [0.114 0.245] & [950 959] \\
5 & 1.935 & 1.118 & 1101 \\
& [1.758 2.107] & [0.896 1.432] & [1098 1101] \\
6 & 1.887 & 14.833 & 1400 \\
& [1.448 2.332] & [12.638 17.546] & [1399 1400] \\
7 & 10.171 & 5.376 & \\
& [9.720 10.625] & [4.088 7.161] & \\
\hline
\end{tabular}
\caption{Simulated example - second scheme - Model 1: posterior means ($\hat{\cdot}$) and credible intervals (CI) of $\mu_k$, $\sigma_k^2$ and $\xi_k$ computed using the subset of posterior samples that has $K = 7$.}
\end{table}

**The full conditional of $s_t$** When [? ] show how to sample from the full conditional of $s_t$, they derive the following probabilities:

\[
f(s_{t+1} = k + 1 | s_t = k, s_{t+2}, \ldots s_T, \alpha, \beta) = \\
\frac{\beta}{n_k^{(t-1)} + 1 + \beta + \alpha},
\]

(19)

\[
f(s_{t+1} = k + 1 | s_t = k + 1, s_{t+2}, \ldots s_T, \alpha, \beta) = \\
\frac{n_k^{(t+1):T} + \alpha}{n_{k+1}^{(t+1):T} + \beta + \alpha},
\]

(20)

Just for example let assume $s_t = s_{t-1} = s_{t-2} = s_{t-3} = k$ and $s_{t-4} = k - 1$; then $n_k^{(t-1)} = 2$. Indeed we can find $n_k^{(t-1)}$ only if we know $\{s_{t-3}, s_{t-2}, s_{t-1}\}$ and then it is not possible that $n_k^{(t-1)}$ appears in equation (19) if none of them is in the conditioning set of (19). As further example, let assume $s_{t+2} = i + 1$. In this case it is trivial to demonstrate that $f(s_{t+1} = k + 1 | s_t = k + 1, s_{t+2} = k + 1, \ldots s_T, \alpha, \beta) = 1$ and not $\frac{n_{k+1}^{(t+1):T} + \alpha}{n_k^{(t+1):T} + \beta + \alpha}$, as in equation (20).

**The MCMC algorithm** In the MCMC algorithm proposed in [? ], during the model fitting the number of occupied regimes can only decrease or remain the same. The major implication is the possibility that the MCMC can never reach the stationary distribution and then we cannot sample from the posterior. To see more clearly why, let suppose that the posterior distribution of $K$, the number of change points, is entirely concentrated on $d$, i.e. $f(K = d | \{y_t\}_{t=1}^T) = 1$, and we initialize the MCMC with $K + c$ different regimes ($c \in \mathbb{N}$). It can happen that at the $b^{th}$ iteration, before the chain has reached its stationary distribution, $K$ assumes value $d - 1$. Then, since $K$ can only decrease or remain constant, after the $b^{th}$ iteration, it will never assume value $d$, the chain will never reach its stationary distribution and we cannot have samples from the posterior of interest. Notice that, given samples from the MCMC algorithm, we have no way to tell if they are from the posterior distribution or not [? ].

Moreover, even if the algorithm has reached its stationary distribution, there is always a non zero
|   | $\hat{\mu}_k$ | $\hat{\sigma}^2_k$ | $\xi_k$ |
|---|---|---|---|
| 1 | 0.099 | 0.712 | 50 |
|    | [-0.138 0.340] | [0.496 1.071] | [50 50] |
| 2 | 5 | 1.978 | 250 |
|    | [4.808 5.199] | [1.637 2.419] | [250 251] |
| 3 | 1.973 | 1.123 | 900 |
|    | [1.895 2.058] | [1.008 1.257] | [899 905] |
| 4 | 2.077 | 0.162 | 956 |
|    | [1.963 2.190] | [0.111 0.238] | [947 959] |
| 5 | 1.937 | 1.115 | 1101 |
|    | [1.767 2.113] | [0.89 1.42] | [1096 1101] |
| 6 | 1.897 | 14.808 | 1400 |
|    | [1.458 2.341] | [12.699 17.392] | [1399 1400] |
| 7 | 10.159 | 5.386 | 17 |
|    | [9.693 10.623] | [4.079 7.218] |

Table 4: Simulated example - second scheme - Model2: posterior means (\(\hat{\cdot}\)) and credible intervals (CI) of $\mu_k$, $\sigma^2_k$ and $\xi_k$.

probability that two regimes can be merged (see the results of the second scheme in Section 4.1) and then, if we run the algorithm for enough iterations, eventually this will happen and the number of change points decreases with no possibility to increase again. As a consequence, as the iterations go toward infinity, the number of regimes tends to 1. This last consideration explains why in the examples of Section 4, the algorithm of [?] finds always 1 regime.
Figure 5: Coal-mining disasters data: the height of the bar represents the count in the associated year. The vertical dashed line separates the regimes identified by the MAP estimate of Model1.

Figure 6: Coal-mining disasters data: posterior distribution of $K$. 
Figure 7: Indoor radon concentration data: the vertical dashed line separates the regimes identified by the MAP estimate of Model1.

| Regime | $\mu_{0,k}$ | $\mu_{1,k}$ | $\sigma_{k}^2$ | $\xi_k$ |
|--------|-------------|-------------|---------------|--------|
| 1      | -0.19       | 0           | 0.027         | 367    |
|        | [-0.225, -0.155] | [0.000, 0.001] | [0.024, 0.032] | [366, 371] |
| 2      | 1.429       | -0.003      | 0.039         | 470    |
|        | [0.856, 1.987] | [-0.004, -0.002] | [0.030, 0.052] | [468, 471] |
| 3      | 5.824       | -0.011      | 0.032         | 614    |
|        | [5.448, 6.218] | [-0.012, -0.011] | [0.026, 0.041] | [604, 627] |
| 4      | -1.136      | 0           | 0.021         | 881    |
|        | [-1.311, -0.953] | [-1.311, -0.953] | [0.018, 0.025] | [865, 893] |
| 5      | -9.571      | 0.01        | 0.027         | 1022   |
|        | [-10.360, -8.943] | [0.009, 0.010] | [0.022, 0.035] | [1004, 1027] |
| 6      | -2.042      | 0.002       | 0.041         | 1183   |
|        | [-2.862, -1.243] | [0.001, 0.003] | [0.034, 0.051] | [1181, 1187] |
| 7      | -18.73      | 0.016       | 0.069         | 1315   |
|        | [-20.302, -17.178] | [0.015, 0.018] | [0.055, 0.089] | [1309, 1325] |
| 8      | 31.758      | -0.022      | 0.046         |        |
|        | [30.886, 32.641] | [-0.022, -0.021] | [0.037, 0.056] |        |

Table 5: Indoor radon concentration data: posterior means ($\hat{\cdot}$) and credible intervals (CI) of $\mu_{0,k}$, $\mu_{1,k}$, $\sigma_{k}^2$ and $\xi_k$ computed using the subset of posterior samples that has $K = 8$. 
| Regime | $\hat{\mu}_{0,k}$ | $\hat{\mu}_{1,k}$ | $\hat{\sigma}^2_k$ | $\hat{\xi}_k$ |
|--------|------------------|------------------|------------------|--------------|
|        | CI               | CI               | CI               | CI           |
| 1      | -0.758           | 0.448            | 1.495            | 1            |
|        | [-44.558 44.088] | [-42.921 44.201] | [0.262 39.986]   | [1 2]        |
| 2      | -0.23            | 0.001            | 0.03             | 477          |
|        | [-0.263 -0.197]  | [0.001 0.001]    | [0.027 0.034]    | [469 521]    |
| 3      | 5.94             | -0.012           | 0.033            | 613          |
|        | [5.245 6.525]    | [-0.013 -0.010]  | [0.026 0.046]    | [604 629]    |
| 4      | -1.142           | 0                | 0.021            | 883          |
|        | [-1.318 -0.952]  | [-1.318 -0.952]  | [0.018 0.025]    | [865 895]    |
| 5      | -9.61            | 0.01             | 0.028            | 1021         |
|        | [-10.584 -8.911] | [0.009 0.011]    | [0.022 0.037]    | [988 1027]   |
| 6      | -2.017           | 0.002            | 0.041            | 1183         |
|        | [-2.912 -1.253]  | [0.001 0.003]    | [0.032 0.052]    | [1180 1188]  |
| 7      | -18.712          | 0.016            | 0.068            | 1317         |
|        | [-20.190 -17.031]| [0.015 0.018]    | [0.053 0.087]    | [1308 1334]  |
| 8      | 31.776           | -0.022           | 0.046            | 1183         |
|        | [30.865 32.631]  | [-0.022 -0.021]  | [0.038 0.056]    |              |

Table 6: Indoor radon concentration data. Posterior means (\(\hat{\cdot}\)) and credible intervals (CI) of $\mu_{0,k}$, $\mu_{1,k}$, $\sigma^2_k$ and $\xi_k$. 