Austrian Journal of Statistics
2024, Volume 53, 114–140.
http://www.ajs.or.at/
doi:10.17713/ajs.v53i4.1890

Hidden Markov Mixtures for Change Detection in Unevenly Spaced Time Series

Marta Cristina C. Bianchi
Instituto de Matemática e Estatística,
Universidade Federal de Goiás,
Goiânia, Brazil.

Vinícius D. Mayrink
Departamento de Estatística, ICEx,
Universidade Federal de Minas Gerais,
Belo Horizonte, Brazil.

Abstract

This work tackles sequential data change-point detection, a research area with various applications in different fields. It focuses on analyzing sequential data such that the distance between locations of consecutive observations is not fixed. The models developed here for change detection extend a Hidden Markov Mixture approach originally designed to handle irregular spacing to improve the identification of atypical values. These models consider the probability of Markov dependence explained by the distance between locations. Bayesian inference is carried out via Gibbs Sampling. Informative priors for the dependency structure are crucial to identify clusters. The models adapt these priors to enable change identification in a general setting. Two mixture models are formulated: one for mean changes and another for mean or variance changes. Post-processing strategies are suggested to categorize observations among the components to facilitate change identification by the mixtures. These strategies are based on maximum posterior probability and consider the uncertainty associated with classifications. Models and clustering performances are evaluated in simulation studies including a Monte Carlo scheme. The analysis also explores three real illustrations. Finally, the proposed approaches are compared to existing clustering and change detection methods available in the literature.

Keywords: Bayesian inference, breakpoint, clustering, Markov-switching model, serial data.

1. Introduction

A change point is a structural shift in ordered observations pattern, over time or space. Literature offers diverse methods for detecting such changes, some of which are addressed in Fearnhead (2005), Ratkovic and Eng (2010), Fearnhead and Liu (2007), Fearnhead and Rigaill (2019), Mira and Petrone (1996), Martinez and Mena (2014), and Haynes, Fearnhead, and Eckley (2017). In the present work, the analysis of irregularly spaced series is the central focus. Here, the time or spatial gaps between consecutive observations are not fixed. As a result, dependence can be explained by the distance between adjacent locations. Ignoring this dependence can determine bias and information loss. Irregular spacing may arise due to different reasons: the presence of missing values or the occurrence pattern of the phenomenon of interest. Natural disasters (e.g. earthquakes and floods) often generate this type of series.

The positions of genes along a chromosome is another example.

Various fields employ irregular series analysis. Koumar and Cejka (2023) examines some of the more than 35 million irregular series from a real ISP network, showcasing their utility in network traffic analysis. Miller (2019) deals with irregularly recorded paleoclimatic data due to dating challenges, like CO_2 concentrations and surface temperatures, collected sporadically and non-contemporaneously. Li, Liao, Peng, and Hwang (2015) addresses with telecom synchronization measurement data often unevenly spaced due to non-uniform data bit distribution. There are many other examples in fields such as astronomy, climatology, high-frequency finance, and signal processing. The present study introduces Bayesian models for change detection in this context. These models are developed as extensions of a Hidden Markov Gaussian Mixture (Mayrink and Gonçalves 2020) originally proposed to identify atypical observations in genetic data.

The Bayesian modeling for change point detection started with Chernoff and Zacks (1964), Smith (1965), Carlin, Gelfand, and Smith (1992), Stephens (1994), and Whittaker and Frühwirth-Schnatter (1994). The Product Partition Model (PPM) was pivotal in this domain, evolving into a popular approach for change detection. Initial efforts to identify multiple change points via PPM were made by Hartigan (1990), Barry and Hartigan (1992), and Barry and Hartigan (1993). In this case, the inference, via Markov Chain Monte Carlo (MCMC), can be computationally expensive for large datasets due to the stochastic partitioning. Various works have advanced in this direction, including Loschi and Cruz (2005), which extend the PPM by providing a method to obtain posterior distributions for the number of change points, as well as the probability of each instance being a change point.

Another Bayesian approach for identifying change points was formulated by Chib (1998). In this case, a reparameterization is proposed for a general change point model, assuming a known number of regimes and treating the process as a Hidden Markov Chain of discrete states. Each state of a latent variable indicates a distinct regime, and the probability of changing regimes depends on the current regime. Markovian dependence is present and remains constant throughout the series. Models for sequential data developed as extensions of Chib's method are referred to as Hidden Markov Models, also known as Markov-switching or Markov Mixture models. The proposals in this paper are constructed under this approach and will be referred to as Hidden Markov Mixture (HMM). The main feature here is the use of the Markovian structure to describe how the data transitions between the unobserved regimes. In Mayrink and Gonçalves (2020), the authors define a finite Gaussian HMM that adapts the dependency structure of Chib to handle an irregularly spaced series. The Markov dependence is treated as uncertain, meaning that it exists with some probability modeled as a function of distances between neighboring observations. Informative priors establish the Markovian structure such that consecutive and close observations tend to be assigned to the same component (or adjacent components). The model discourages classifying near observations into distant components. Other methods building upon Chib's work encompass articles such as Gales and Yong (2007), Meigenier, Bonastre, Fredouille, and Merlin (2000), and Beal and Krishnamurthy (2006); they address speech recognition and gene clustering problems. For foundational concepts on this approach, from both classical and Bayesian perspectives, see Fruhwirth-Schnatter (2006).

Numerous studies address the grouping of time series data while considering spatial dependence. A common strategy to incorporate this dependence involves employing Markov random fields in mixture models. This approach finds widespread application in spatial data, as demonstrated in Fernández and Green (2002) and Vincent, Risser, and Ciuciu (2010). Another approach involves developing models that combine Dirichlet Process (DP) to accommodate uncertainty in the number of clusters, along with Gaussian Process (GP) to handle temporal dependence. Recently, McDowell, Manandhar, Vockley, Schmid, Reddy, and Engelhardt (2018) constructed such a model by assuming that observations within a cluster are generated from a specific GP, with a mean obtained via DP and a covariance kernel specific to each cluster, defined as a function of the Euclidean distance between locations. Furthermore,

Teixeira, Assuncao, and Loschi (2019) work with PPM in the spatiotemporal context, and Page and Quintana (2016) propose a spatial version.

The detection of change points in series using clustering mixture models is an active field in Bayesian studies. In this context, a change point can be characterized as the location at which a component change occurred. Mixture modeling facilitates accommodating various distribution patterns, and in a clustering mixture, regimes are recurrent. Parameter values vary across different components, allowing better adaptation to cyclic changes in the series. In change point models, structural parameters are estimated, with no possibility of identifying a return to a previously identified regime in earlier locations. Works in this domain include Broët and Richardson (2006), identifying changes in series of genetic data using spatially correlated mixtures; Zhu and Melnykov (2022) developing an asymmetric matrix-variate mixture capable of simultaneously estimating change points across all data groups; and Melnykov and Maitra (2010) providing a retrospective on important finite mixtures for clustering. However, finite mixture models, including those developed here, do not consider uncertainty in the number of clusters, requiring pre-specification of this quantity and subsequent analyses to determine the appropriate value. In this regard, the Bayesian nonparametric extension for infinite mixture models (using the DP) is common in the literature. This approach allows simultaneous estimation of the number of components and parameters. See, for example, Ko, Chong, and Ghosh (2015); Dufays (2016).

The development of methods for identifying multiparametric changes in a series is also a relevant issue. The aim is to identify locations where changes may occur at distinct time points for various parameters. Examples of such methods include the multiple partition PPM (Pedroso, Loschi, and Quintana 2023), and the model in Peluso, Chib, and Mira (2019) where groups of structural parameters follow distinct DP priors, with a different multiple-change process for each, under the structure of Chib (1998).

The allocation of observations to clusters in grouping models is commonly done using post-processing methods based on Maximum *a posteriori* (MAP) estimation, or the Posterior Similarity Matrix (PSM) in Fritsch and Ickstadt (2009). The MAP seeks the optimal clustering that maximizes the posterior. Specifically, in finite mixture models, the term MAP can be used as the criterion that assigns an observation to the component k that maximizes the vector of posterior classification probabilities across all components. The PSM provides probabilities that two observations belong to the same cluster. Methods based on the PSM yield a single optimal estimate of the partition, disregarding the associated uncertainty. Bayesian mixtures with a varying number of clusters often employ PSM-based approaches to estimate the optimal configuration, as classification probabilities are influenced by label swapping during MCMC (Fritsch and Ickstadt 2009).

The contributions of this work are as follows. (*i*) development of HMMs for change detection in irregular series through adaptations of priors defined in a foundational model proposed to identify atypical observations. The proposed HMMs harness the original model's ability to handle spatial dependence as uncertain, thus contributing to change point analysis in irregular series. The stochastic modeling flexibility for dependence structure allows us to determine whether there is dependence between a pair of adjacent observations and quantify such dependence. (*ii*) one of the proposed models leverages the structure of components with distinct and ordered means (Mayrink and Gonçalves 2020), and with the contribution of choosing appropriate informative hyperparameters, defines an HMM for change point detection focused on the means. (*iii*) a second model is constructed to identify changes in mean and/or variance separately. This is achieved by carefully choosing priors and imposing constraints. In the multiparametric proposal, it is crucial to allow some components to have equal (or close) means (but different variances), unlike the reference model. (*iv*) propose clustering methods based on MAP criterion, enabling the quantification of the classification uncertainty. (*v*) evaluate the methodology through simulations with Monte Carlo (MC) scheme across various scenarios. (*vi*) illustrate the models through real applications in finance and astronomy, which have not been previously analyzed for change point detection assuming irregular spacing. (*vii*)

compare the proposals with existing approaches for clustering and change point detection.

The outline of this paper is as follows. Section 2 describes the Hidden Markovian Bayesian mixture model. Subsequently, this section presents the proposed models for change detection in irregular series as extensions of the base model. Section 3 details the post-processing classification methods. Section 4 shows the MC studies for various scenarios of artificial irregular series. Section 5 compares the proposed models with other methods from the literature for clustering and change detection. Section 6 encompasses the real applications. Finally, Section 7 summarizes the main conclusions.

2. Gaussian hidden Markov mixture

The models developed here share a similar structure, particularly utilizing the spatial dependence to address the irregular spacing. These models diverge in terms of specifying informative priors tailored to achieve specific goals.

Let X_i be a real random variable associated with the i-th location; $X_1, \cdots, X_n$ is a random sample. Consider $K > 1$ (fixed and known) representing the number of mixture components. Define $\boldsymbol{q}_0 = (q_{01}, q_{02}, \cdots, q_{0K})^\top \in \mathbb{R}^{K+}$ as an unknown vector of probabilities, with $\sum_{k=1}^{K} q_{0k} = 1$. Let $\boldsymbol{Q} = (q_{k1k2})$, $k_1, k_2 = 1, \cdots, K$, represent the transition matrix of a first-order Hidden Markov Chain with discrete time, having K latent states, where each state indicates a mixture component. Denote $\boldsymbol{q}_k = (q_{k1}, q_{k2}, \cdots, q_{kK})^\top \in \mathbb{R}^{K+}$, with $\sum_{k'=1}^{K} q_{kk'} = 1$, as the probability vector in the k-th row of $\boldsymbol{Q}$. Define the classification vector $\boldsymbol{Z}_i = (Z_{i1}, \cdots, Z_{iK})^\top \in \{0,1\}^K$. In this case, $Z_{ik} = 1$, if X_i belongs to the k-th component, and $Z_{ik} = 0$, otherwise. An HMM with Gaussian components of unknown means and variances μ_k and σ_k^2, for $k = 1, \cdots, K$, is defined as follows:

$$
\begin{aligned}
(X_i|Z_i) &\sim N(\mu_k, \sigma_k^2), \quad \text{independent} \quad \forall i; && (1)\\
(\boldsymbol{Z}_1|\boldsymbol{q}_0) &\sim \text{Mult}(1, \boldsymbol{q}_0); && (2)\\
(\boldsymbol{Z}_i|Z_{i-1,k} = 1, \rho_i, \boldsymbol{q}_0, \boldsymbol{Q}) &\sim (1-\rho_i)\,\text{Mult}(1, \boldsymbol{q}_0) + \rho_i\,\text{Mult}(1, \boldsymbol{q}_k), \quad i = 2, \cdots, n. && (3)
\end{aligned}
$$

Here, "Mult" indicates the Multinomial distribution. The X_i's are not marginally (with respect to Z) independent. The dependence is defined at a lower level of the model by Expression (3), a discrete mixture representing the absence or presence of first-order Markovian dependence. The Markovian structure represented by $\text{Mult}(1, \boldsymbol{q}_k)$ stochastically determines the classification of the observation X_i, given that the model classified X_{i-1} in the k-th component. The Expression (3) reflects that the Markovian dependence is the same for any pair of consecutive observations, however, it is uncertain, being present with probability ρ_i. Let Φ be the cumulative distribution function of the standard normal distribution and define d_i as the transformed distance on the scale $(0, 1]$ between neighboring locations $i-1$ and i. The following priors are adopted.

$$
\begin{aligned}
(\mu_k, \sigma_k^2) &\sim \text{NIG}(m_k, v_k, u_{1k}, u_{2k}); && (4)\\
\boldsymbol{q}_k &\sim \text{Dir}(\boldsymbol{\alpha}_k), \quad k = 1, \cdots, K, \quad \text{independent}; && (5)\\
\boldsymbol{q}_0 &\sim \text{Dir}(\boldsymbol{\alpha}_0); && (6)\\
(\rho_i|\boldsymbol{\beta}) &= \Phi(\beta_0 + \beta_1 d_i); && (7)\\
\boldsymbol{\beta} = (\beta_0, \beta_1)^\top &\sim \text{N}_2(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0). && (8)
\end{aligned}
$$

The terms "NIG" and "Dir" indicate the Normal-Inverse-Gamma and Dirichlet distributions. Denote $\boldsymbol{\beta} = (\beta_0, \beta_1)^\top \in \mathbb{R}^2$. In terms of hyperparameters, consider v_k, $u_{2k} \in \mathbb{R}^+$, $m_k \in \mathbb{R}$, $u_{1k} > 2$ is positive real, $\boldsymbol{\alpha}_0$, $\boldsymbol{\alpha}_k \in \mathbb{R}^{K+}$, $\boldsymbol{\mu}_0 \in \mathbb{R}^2$, and $\boldsymbol{\Sigma}_0$ is a positive definite variance-covariance matrix (2×2). Equation 7 introduces spatial dependence into the HMM by defining the probability ρ_i of Markovian dependence as a function (probit) of the transformed distance d_i between neighboring observations. For computational implementation, auxiliary

Bernoulli (Ber) random variables denoted as W_i, $i = 2, \cdots, n$, are also introduced. One can write

$$(\boldsymbol{Z}_i | Z_{i-1,k} = 1, W_i = 0, \boldsymbol{q}_0) \sim \text{Mult}(1, \boldsymbol{q}_0); \quad (9)$$

$$(\boldsymbol{Z}_i | Z_{i-1,k} = 1, W_i = 1, \boldsymbol{Q}) \sim \text{Mult}(1, \boldsymbol{q}_k); \quad (10)$$

$$(W_i | \rho_i) \sim \text{Ber}(\rho_i). \quad (11)$$

Set $W_1 = 0$ almost surely, since there is no observation before X_1. Furthermore, $(\boldsymbol{Z}_1 | \boldsymbol{Z}_0, W_1 = 0, \boldsymbol{q}_0) := (\boldsymbol{Z}_1 | \boldsymbol{q}_0)$ in the scenario of an absence of Markovian dependence. The next section discusses informative priors.

2.1. Proposed prior specifications

The purpose of this section is to verify and justify the suitability of the priors for achieving the goal of change point detection. In this article, two HMM's are proposed. The first one is intended for identifying changes in mean, also known as the Separated Means Component model (HMM$_S$). The second model addresses changes in mean and/or variance, and will be referred to as the Grouped Means Component model (HMM$_G$). The priors to deal with the spatial structure are the same in both models. However, the priors for the means and variances of the mixture components, as well as the strategy adopted for handling label reassignment, differ between the models.

Label switching

The label-switching problem can arise in mixture models when there are multiple possible permutations of labels assigned to the mixture components, resulting in the same probability distribution. Label swapping does not affect the statistical properties or the quality of model fitting, but it can lead to confusion in interpreting estimates.

HMM$_S$: It addresses the permutation problem by imposing a constraint in which the Gaussian components have distinct and ordered means (Mayrink and Gonçalves 2020). This ordering also defines the concept of neighborhood among the components, which is crucial for handling spatial dependence. Consequently, the model fits mixtures with distinct component means, justifying the acronym HMM$_S$. The approach enables the identification of changes focused on the mean.

HMM$_G$: It allows fitting Gaussian components with equal or close means. This enables handling situations where the densities of components significantly overlap. To achieve this capability, the HMM$_G$ groups the means of these components into a set called "grouped mean components". Within each group, different variances are defined. Consequently, a change in the mean is obtained as the location between consecutive observations generated by different components belonging to distinct groups. In cases where these components belong to the same group of means, only the variance change is determined. An additional parameter, $R \in \mathbb{N}^*$, fixed and known, where $1 < R < K$, defines the number of component groups. The components are allocated to each group using the k-means clustering method (Hartigan and Wong 1979).

The flexibility to identify changes in mean and/or variance by the HMM$_G$ is enabled through the treatment of label switching, and this is facilitated by the appropriate choice of priors for means and variances. To carry out the relabeling, consider the r-th collection of grouped means as $g_r = \{\mu_{r1}, \cdots, \mu_{rs_r}\}$, where s_r is the known size of the collection, and $1 \leq s_r < K$. The sizes of the groups are included in the vector $\boldsymbol{s} = (s_1, \cdots, s_R)$, with $\sum_{r=1}^{R} s_r = K$. The labels of the mixture components are established in two steps. First, the grouping of means is ordered. Then, the label is determined within each group by the restricted ordering of variances, σ^2_{rj}, $j = 1, \cdots, s_r$, for each collection r of means, $r = 1, \cdots, R$. This implies that components within the same group exhibit distinct variances. Thus, the constraints in the

HMM_G are defined by:

$$\begin{aligned} &\mu_{11} \leq \cdots \leq \mu_{1s_1} < \mu_{21} \leq \cdots \leq \mu_{2s_2} < \mu_{R1} \leq \cdots \leq \mu_{Rs_R}; \\ &\sigma^2_{11} < \cdots < \sigma^2_{1s_1}; \quad \sigma^2_{21} < \cdots < \sigma^2_{2s_2}; \quad \cdots \quad \sigma^2_{R1} < \cdots < \sigma^2_{Rs_R}. \end{aligned} \tag{12}$$

The constraints in (12) enable the model to identify changes in mean and variance individually. For instance, consider two components of an HMM_G grouped in the same collection. In this case, a component change signifies a variance-only change. Now, consider two components grouped in separate collections. Here, a change in component indicates a change in mean.

This labeling strategy allows an understanding of the model's degree of uncertainty in changing the mean. The components being labeled in ascending order of prior variances, within the same group of means, implies that the adjacent component to the right has higher uncertainty about the mean. The first component of each group has the lowest variability and the separation of means is slightly higher "between" than "within" groups. These aspects imply that a small overlap is configured between the first and last components of adjacent groups. Consequently, a group of means change-point only occurs when there is strong evidence from the data.

Restrictions on the parameters of the HMMs are enforced during the MCMC. The re-labeling by mean restriction post-processing in HMM_S showed slightly inferior results to the restriction enforced during the MCMC in extreme mixture scenarios (components with significant overlap). The results were analogous (post and during MCMC) in other scenarios. Tests conducted on the HMM_S with few samples demonstrated that ordering of means produced analogous re-labeling results compared to the "data-based relabeling" by Rodriguez and Walker (2014), where the optimal permutations are defined as those that minimize a k-means type loss function between the cluster centroids and the observed data. Another k-means type alternative by Malsiner-Walli, Frühwirth-Schnatter, and Grün (2017) is an interesting option for future analyses focused on the label-switching issue. Other post-processing methods, to be applied after sampling, proved inefficient in our study. This includes sorting only variances, mixture component weights, or using the methods by Stephens (2000b) and Papastamoulis and Iliopoulos (2010).

Priors to handle the spatial dependence

The choice of informative priors dealing with the dependence structure is crucial to empower the model with the ability to identify clusters and achieve the desired spatial association. Under less informative settings, the HMM might behave similarly to a common mixture model. Alternatively, in the analysis of large datasets, inference could be dominated by the likelihood. The $\boldsymbol{\beta}$ is responsible for reflecting the desired relationship $d_i \times \rho_i$. The $\boldsymbol{\mu}_0$ in (8) indicates the expected behavior "d_i increases and ρ_i decreases". The result is a smoothly decreasing curve, where $d_i = 0.5$ corresponds to $\rho_i = 0.5$. The smoothness is important to differentiate the degree of Markovian dependence across different distances.

Figure 1 illustrates some patterns. The choice of the prior expected behavior for $d_i \times \rho_i$ depends on the application. The HMMs proposed by Mayrink and Gonçalves (2017) and Mayrink and Gonçalves (2020) use the option in Panel (a). In the genetic application motivating these studies, few $d_i < 0.2$ or $d_i > 0.8$ are present. As a consequence, Panel (a) suggests a faster decrease of ρ_i when $d_i \in (0.2, 0.8)$. Modifying the the shape of this curve enables the model to handle specific irregular spacing. If most observations have near neighbors (small d_i's), and assuming informative prior for $\boldsymbol{\beta}$ imposing the curve in (a), then many $\rho_i \approx 1$ (strong dependence) is obtained. This behavior is not reasonable since it poses difficulties to model reclassification, thus leading to convergence issues in the MCMC. In this situation, it is recommended to use one of the patterns in Panels $(b-d)$. In the present study, the choice $\boldsymbol{\mu}_0 = (0.5, -1)$ yields the best results for the proposed applications. Moreover, it is crucial to assume an informative prior for $\boldsymbol{\beta}$, with $\Sigma_{0_{11}} = \Sigma_{0_{22}} = 0.0001$, to ensure the desired posterior behavior. The choice of not fixing $\boldsymbol{\beta}$ allows some flexibility for the analyst in expressing

uncertainty about the relationship $d_i \times \rho_i$. For simplicity, β_0 and β_1 are assumed independent ($\Sigma_{0_{12}} = \Sigma_{0_{21}} = 0$).

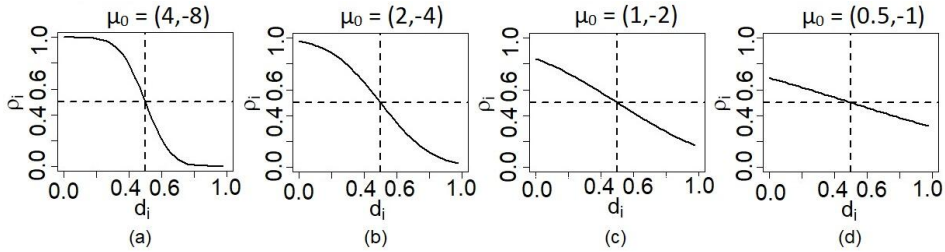


Figure 1: Prior relationship between the distance d_i and the dependence probability ρ_i for different $\boldsymbol{\mu}_0$

The modeling in Mayrink and Gonçalves (2017) defines $d_i \in (0, 1]$, which is motivated by computational reasons, given that large d_i's are observed in their genetic application. The same strategy is considered in Mayrink and Gonçalves (2020) and it remains important to be used in the HMMs proposed here. The scale of distances varies between applications and the $(0, 1]$ transformation can be viewed as a standard strategy to allow a broad use and understanding of the model's behavior. Exploring the distances on a fixed scale, regardless of the application, simplifies the selection of priors that reflect spatial association. The choice of whether or not to apply the transformation does not alter the modeling structure but can impact inference.

The HMM in Mayrink and Gonçalves (2020) employs Markovian dependence to identify observations in just two components, those with lower and higher means. The proposed adaptation aiming to identify change points retains the model's clustering capability and enables change detection. However, it does not prioritize one component over the other in the mixture. In most change-point applications, there is no reason to favor allocation to any component.

The Dirichlet prior of $\boldsymbol{q_0}$ reflects the absence of Markovian dependence. Specify $\boldsymbol{\alpha}_0 = (T/K)\mathbf{1}_{(K\times 1)}$. The term $\mathbf{1}_{(K\times 1)}$ is a $(K \times 1)$ vector of 1's. The value T must be chosen by the user and it is recommended to be of greater magnitude than the sample size size. This way $\boldsymbol{\alpha}_0$ suggests that all weights are equal to enforce similar allocation priority among the K components. Small T dictates a weak prior, which would be overshadowed by the likelihood. This holds particularly true when the series is large. On the other hand, the parameters in $\boldsymbol{\alpha}_k$ of the Dirichlet for $\boldsymbol{q}_k$, $k = 1, \cdots, K$, are defined to favor the formation of clusters. In other words, the chosen $\boldsymbol{\alpha}_k$ should facilitate the classification of consecutive and near observations into the same mixture component. Adhering to the expectations in the presence of spatial dependence, the $\boldsymbol{\alpha}_k$ should discourage the model from allocating adjacent observations into distant components. Values of $\boldsymbol{\alpha}_k$ can be arranged in the k-th row of a matrix $\boldsymbol{\alpha}_{K\times K}$, which would be configured with higher values in the main diagonal.

The strategy for determining the Markovian structure given by $\boldsymbol{\alpha}_{K\times K}$ is as follows. First, the labels of the hidden states are defined in accordance with the label-switching strategy adopted among mixture components. Each model has a specific re-labeling strategy (to be discussed later). Gaussian components with close means (neighbors) correspond to adjacent states. The sum of any row in $\boldsymbol{\alpha}_{K\times K}$ is T. The highest values are on the main diagonal, and the second highest values are arranged in the sub- and super-diagonals. This configuration is necessary to impose clustering power on the model, promoting classifications in nearby components for adjacent serial observations. The remaining elements in the row have equal and small values, meaning that distant components have similar and reduced priority. To achieve a reasonable clustering behavior, it is recommended to allocate at least 50% of the sum of weights to the elements on the main diagonal of $\boldsymbol{\alpha}_{K\times K}$. The analyst must be aware that setting extremely high values in this main diagonal leads to a situation where the model is not flexible enough to allocate near observations in different components. As a consequence, convergence issues

will occur in the MCMC.

An MC sensitivity analysis for the HMMs, under different choices for $\boldsymbol{\alpha}_{K\times K}$, indicated that the HMM$_G$ requires large percentages of the sum $T = 2{,}000$ (inspired by the main references) in the main diagonal as opposed to the HMM$_S$. Specifically, based on the data examined in this study, we selected the Markovian structure $\boldsymbol{\alpha}^{50\%}$ for the HMM$_S$ and $\boldsymbol{\alpha}^{75\%}$ for the HMM$_G$. The overridden value refers to the percentage considered for the magnitude on the main diagonal. In addition, assume allocating 22% (for $\boldsymbol{\alpha}^{50\%}$) and 10% (for $\boldsymbol{\alpha}^{75\%}$) of the sum 2,000 to the elements in the sub- and super-diagonals. The Supplementary Material presents this sensitivity analysis in detail.

The HMMs proposed in this work are flexible in terms of choosing the prior uncertainty level about the parameters. The decision to increase or decrease the initial information level on $\boldsymbol{\beta}$, $\boldsymbol{q}_0$, and $\boldsymbol{q}_k$ depends on the application and the available sample size. Fixing parameters to build these models would result in the loss of such flexibility.

Prior distributions for means and variances

The proposed methodology for constructing priors for the means took into consideration an assessment of the degree of proximity among the means of the mixture components. In the HMM$_G$, the priors are centered in closer locations of the mixture domain than those of the HMM$_S$. This is justified by the fact that the HMM$_G$ allows components with equal (or close) means. Regarding the prior for the variances, a single vague prior was maintained for the HMM$_S$, as this model does not intend to analyze changes in variance. However, to aid the identification of different variances by the HMM$_G$, it became necessary to adopt an informative prior for each variance. Further details regarding this aspect are described for each model in sequence.

HMM$_S$: Gaussian priors for μ_k (HMM$_S$) are proposed with different levels of variability to assess the model's behavior. A criterion based on the overlap between adjacent component densities was employed, quantified by the intersection area between these pairs (Inman and Bradley 1989). Gaussian densities (symmetric) with high overlap indicate closely positioned means, while low overlap suggests more distant means. Results from a sensitivity analysis of the HMM$_S$, across different data scenarios, indicate that the choice for a 10% overlap configures a moderate separation of the means. It is recommended to adopt such separation, and thus avoid greater overlaps that impair the distinct means requirement. Further details can be found in the Supplementary Material. The hyperparameters of the NIG distribution defined for (μ_k, σ_k^2) are determined to reach the desired overlap rate. The m_k were chosen to be equally spaced, taking into account the scale of the observations. The v_k were defined such that the variability of the Gaussian priors for μ_k's yield the target overlap rate. Regarding the variance σ_k^2, a single vague prior is adopted in the HMM$_S$; set $u_{1k} = 2.1$ and $u_{2k} = 1.1$, then $E(\sigma_k^2) = 1$ and $Var(\sigma_k^2) = 10$, which follows a similar approach to that used in Mayrink and Gonçalves (2020).

HMM$_G$: The strategy here is similar to the one defined for the HMM$_S$. However, the HMM$_G$ allows for components with equal (or close) means to be accommodated. In this case, adopting a large intersection between two adjacent Gaussian prior densities is an interesting configuration. Sensitivity analyses performed on a simulated sample under various scenarios, with overlap values of $10\%, 15\%, 20\%, 25\%$, and 30%, revealed that a minimum overlap of 25% is necessary for MCMC convergence, particularly for the means. The 25% overlap is also considered for the real applications. In the HMM$_G$, it became evident that the use of a single vague prior for variances leads to a lack of MCMC convergence. This issue arises due to component swapping caused by the weak prior distinction for the σ_k^2's. The inclusion of distinct informative distributions for variances is necessary to aid the model in component identification, considering that components are labeled by the restricted ordering of variances within each mean group; see Expression (12). A reasonable option is to define K distinct

priors, one for each variance, given that in practice the number of groups and their cardinality (R and $\boldsymbol{s}$) are unknown. That way, K Inverse-Gammas with ordered means are specified. One has $E(\sigma_1^2) = 0.5$, and the subsequent expectations are uniformly spaced, ensuring that the mean of the last component is smaller than the sample variance $S_V = \sum_{i=1}^{n}(X_i - \bar{X})^2/(n-1)$[1]. The S_V is an upper limit for $E(\sigma_K^2)$. The spacing between means is determined by $S_V/(K+2)$, which prevents $E(\sigma_K^2)$ from being too close to the upper limit. For the prior variability, consider $Var(\sigma_k^2) = 1$. See the Supplementary Material for further details. Malsiner-Walli *et al.* (2017) control the overlap and relative position between the means of the components on the mixture-of-mixtures modeling by specifying hyperparameters based on the decomposition of total, intra, and inter-group variability. A similar strategy could be adapted to our HMMs; this topic is left for future work.

2.2. Bayesian inference

A Gibbs Sampling was employed to sample from the joint posterior distribution of all unknown quantities defined in the models. The algorithm is implemented with a block structure to improve mixing and convergence. The block (Z, W) requires a BFFS method (Backward-Filtering-Forward-Sampling) developed in Mayrink and Gonçalves (2017). This entails sampling (Z, W) in the forward direction, while the parameter values of their respective marginal distributions are recursively obtained in the backward direction, from n to 1. As a result, the procedure accounts for the Markovian dependence in the series in both future and past directions. The MCMC implementation encompassing all complete conditionals, remains consistent with that used in Mayrink and Gonçalves (2020). Nevertheless, the rejection sampling strategy is replaced by mean (and variance) reordering of the components to address the label-switching issue.

The MCMC has 6,000 iterations and *burn-in* period of 3,000. Regarding initial values, set $\boldsymbol{q}_0^{(0)} = (1/K)\ \mathbf{1}_{(K\times 1)}$ and $\boldsymbol{Q}^{(0)} = (1/K)\mathbf{1}_{(K\times K)}$, where $\mathbf{1}_{(l_1\times l_2)}$ is a $l_1 \times l_2$ matrix of 1's. Starting values for μ_k, σ_k^2, and $\boldsymbol{\beta}$ were defined using their respective prior means. In assessing convergence, two tests were applied to the chains of μ_k's, in addition to observing traceplots and posterior densities. The `coda` package (Plummer, Best, Cowles, and Vines 2006) was utilized, specifically remploying the `geweke.diag` (Geweke 1992) and `heidel.diag` (Heidelberger and Welch 1983) functions for convergence diagnostics. A lack of convergence was assumed if the chain failed both tests.

The quantities K and R of the HMMs are treated as fixed, yet in practice, they are rarely known. We adopt the strategy of commencing the analysis with K being estimated by a standard Bayesian Gaussian mixture model (without dependence) fitted by the `bmixnorm` function from the `R` package `bmixture` (Mohammadi and Salehi-Rad 2012; Mohammadi, Salehi-Rad, and Wit 2013). This function employs transdimensional MCMC sampling based on a birth-death approach (Stephens 2000a). Particularly for the HMM$_G$, knowledge of the quantity R, representing the number of groups of Gaussian components with proximate (or equal) means, is imperative. A reasonable estimate for this element is garnered from the modes in the data histogram of the application. The estimate $\hat{R}$ is determined by the number of peaks detected in the predictive density of a simple Gaussian mixture with K components (`bmixture`). The peak points (modes) are determined through an optimization method in the package `Splus2R` (Constantine and Hesterberg 2021). Malsiner-Walli *et al.* (2017) propose a more systematic approach for determining the number of group components and the total group count. Implementing this strategy is beyond the scope of our study.

In the present study, K and R are chosen in two stages. First, initial values are determined as previously explained. Subsequently, the HMM of interest is fitted under the initial candidate of K (and R), as well as under adjacent options ($K-1$, $K+1$, $R-1$, $R+1$). The final model is then selected using goodness-of-fit measurements and model selection criteria for clustering. Furthermore, in the HMM$_G$, the allocation of the K components among the R collections of

[1] If $S_V < 0.5$, the prior mean for σ_1^2 must be redefined to maintain the scale.

grouped means $g_r^{(t)} = \{\mu_{r1}, \cdots, \mu_{rs_r}\}^{(t)}$, $r = 1, \cdots, R$, is performed at each MCMC iteration t using the k-means clustering method (Hartigan and Wong 1979) on the means $\mu_1^{(t)}, \cdots, \mu_K^{(t)}$, with the number of centroids determined by R. The cardinality of each group in each iteration constitutes the coordinates of the vector $\boldsymbol{s}^{(t)}$. Once the grouping $g_1^{(t)}, \cdots, g_R^{(t)}$ is obtained, the label switching strategy (12) is applied. The estimate of $\boldsymbol{s}$ is the most frequent posterior configuration $\tilde{s}$ among those visited throughout the MCMC.

3. Classification method

In an HMM, a change point detection should examine whether there is a classification change, regarding mixture components, between consecutive observations in the series. Hence, a strategy is to first establish a criterion to allocate each observation to one of the mixture components and subsequently analyze alterations throughout the series. Similar to Mayrink and Gonçalves (2017) and Mayrink and Gonçalves (2020), classifying a data point into the k-th mixture component can be achieved based on the corresponding posterior probability. Given an (approximate) sample of size M from the posterior distribution of Z, and the observation at position i, the mentioned posterior probability is given by $p_{ik} = (1/M)\sum_{m=1}^{M} \mathcal{I}(Z_{i,k}^{(m)} = 1)$, where $\boldsymbol{p}_i = (p_{i1}, \cdots, p_{iK})^\top$, with $\sum_{k=1}^{K} p_{ik} = 1$ for all i, represents the vector of classification probabilities for the i-th observation. The term $\mathcal{I}(.)$ is an indicator function. In the present study, the i-th observation is assigned to the component k, if p_{ik} takes the highest value within $\boldsymbol{p}_i$. In the literature, this strategy is referred to as the Maximum *a posteriori* (MAP) method. Particularly, for HMM_S, a mean change point is defined as the location between two adjacent observations, where the first is classified in one component and the second is allocated in another one.

The chosen criterion selects the component with the highest probability, even in situations where multiple components have large probabilities, a scenario prone to classification errors. In the case of a Gaussian mixture, this classification uncertainty tends to be higher for those components that have substantial overlap. Given that the HMM_G permits fitting components with completely (or nearly completely) overlapping densities, the adoption of the simple MAP criterion may lead to errors in identifying changes in variance.

The frequent situation where "multiple components exhibit high probabilities", in the HMM_G, has motivated the proposal of a certainty-level classification criterion to assess changes in variance. The idea is to segregate observations into two groups based on the level of certainty in their allocation, here referred to as "decisive" and "doubtful". For an observation i that satisfies $\max(\boldsymbol{p}_i) > 0.5$, its allocation is designated to the component with the maximum p_{ik} with a high level of certainty, termed a "decisive" classification. Conversely, the criterion accounts for the two components having the first and second largest probabilities. In this case, the model is not certain about the allocation, termed as "doubtful" classification. In a simulation study, under the "doubtful" category, a classification is assumed reasonable or satisfactory when one of the two high-probability components corresponds to the real one for that observation; classification error is obtained otherwise. The performance of the classification criteria, under different certainty levels, was studied using MC samples. Details are available in the Supplementary Material. In brief, the criterion "$\max(\boldsymbol{p}_i) > 0.5$" (henceforth denoted $P_{0.5}$) yielded the best results.

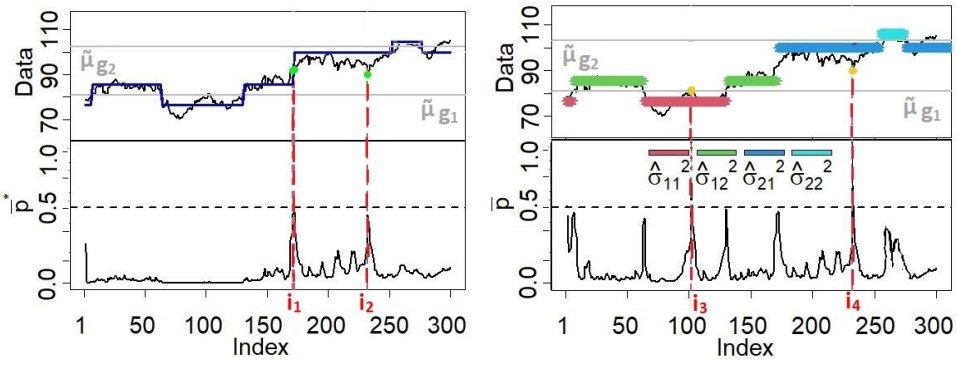
In an HMM_G, to distinguish change points between mean and variance, two steps are necessary. First, each observation is assigned to one of the K components using the criterion $\max(\boldsymbol{p}_i)$. Then, based on the posterior estimate of the vector of sizes for the R collections of grouped means, $\hat{\boldsymbol{s}} = (\hat{s}_1, \cdots, \hat{s}_R)^\top$, the components are allocated to the corresponding collections following the constraints to avoid label-switching in (12). For example, if $K = 7$, $R = 3$, and $\boldsymbol{s} = (3, 2, 2)$, then components 1 to 3 belong to collection 1, 4 and 5 are related to collection 2, and the rest correspond to collection 3. If two adjacent observations in the series are classified into different groups, then a change in the mean occurs. If they are classified in

the same group, but under different components, it configures a change in variance.

After applying the classification method, it is possible to measure the certainty level for the classification of each observation. The model's uncertainty in allocating observation i to component k' is evaluated through $\bar{p}_{ik'} \equiv 1 - p_{ik'}$ for the purpose of studying mean changes under HMM_S, or variance changes under HMM_G. If the interest lies in analyzing mean changes under HMM_G, the classification certainty level is not directly associated with allocating an observation to a component, but to a set of grouped means. In this context, the model's uncertainty is calculated as $\bar{p}^*_{ir'} \equiv 1 - \sum_{k \in g_{r'}} p_{ik}$, where $g_{r'}$ defines the components belonging to the r'-th collection of means. The lower this probability, the lower the model's uncertainty that the classification is correct.

Figure 2 illustrates the classifications provided by HMM_G, as well as allocation uncertainty. Panel (a) displays the classification concerning the mean. The blue line indicates the posterior mean of the component in which the observation was allocated. The horizontal gray lines represent the median of each collection of grouped means. A mean change occurs when the blue line moves away from one gray line and aligns with another. For example, in Panel (a), there is only one mean change point, corresponding to observation i_1. Although not depicted here, it is worth noting that when using the HMM_S, the type of graph in Panel (a) does not feature the gray lines, and a change in the level of the blue line indicates a mean change.

Figure 2 (b) highlights the variance-based classifications produced by HMM_G. Asterisks of different colors are superimposed onto the series, representing classifications into distinct components. The y-axis value signifies the posterior mean of the component to which the observation was classified. Panel (b) indicates a mean change by the shift in the asterisk series, moving away from one gray line. Variance change is denoted by a color alteration of the asterisk sequence. There are 5 evident variance changes and 1 mean change in the graph. Panel (b) displays only one of the two possible classifications at the doubtful level. Specifically, for the i-th observation, the shown allocation corresponds to the component with the highest value in $\boldsymbol{p}_i$. The Supplementary Material discusses the efficiency of assuming the first or the second largest probabilities.



(a) Mean-based classification (b) Variance-based classification

Figure 2: Classifications and change-points from HMM_G $\substack{K=4 \\ R=2}$ with $\boldsymbol{s} = (2, 2)$. Black line: data series. Gray lines $\tilde{\mu}_{g_1}$ and $\tilde{\mu}_{g_2}$: median of grouped means collections. Panel (a): posterior mean by mean-based classification (in blue). Panel (b): posterior mean by variance-based classification (colored asterisks show distinct component classification). Subpanels $\bar{p}^*$ and $\bar{p}$ show classification uncertainty for mean and variance, respectively. Peaks at $\bar{p}^*$ for i_1 and i_2 (i_1 local mean change). Peaks at i_3 and i_4: doubtful classifications for variance.

Generally, the classification uncertainty is high at a change point, resulting in spikes in the graph of $\bar{p}$ and $\bar{p}^*$. Regarding changes in mean groups, Panel (a) indicates a sequence of observations classified under g_1, 1 to $i_1 - 1$, with low uncertainty. However, upon reaching the first observation classified in g_2 at i_1 (the change point), the model tends to be less confident. For subsequent observations, the uncertainty decreases again due to the model clustering nature (consecutive observations from nearby locations tend to be under the same

component). When the model fit produces a high number of change points, one will observe an erratic trajectory for $\bar{p}$ and/or $\bar{p}^*$ (many spikes), as seen in Panel (b). Changes in asterisk colors within the same group level (trajectory near one gray line) indicate changes in variance. Note the presence of spikes of $\bar{p}$ at these locations.

Spikes in $\bar{p}^*$ or $\bar{p}$ do not only occur at change points. A spike under an observation literally indicates a higher classification uncertainty provided by the criterion. In Panel (a), even though the observations from i_1 to 300 are classified under the same group g_2, there is a spike located at i_2. Observe that data point i_2 is slightly lower than the other values in the data sequence, sufficient for the model to be uncertain about allocating i_2 to g_1 or g_2. Nevertheless, the chosen criterion indicates g_2 due to the clustering nature of HMM$_G$. In Panel (b), observations i_3 and i_4 do not determine variance changes. However, they exhibit a high classification uncertainty. Additionally, the classifications at i_3 and i_4 are deemed "doubtful", with $\bar{p} > 0.5$.

In addition to MAP, other clustering methods (Stephens 2000b; Kaufman and Rousseeuw 1990) were explored in our study. Their results are consistent in estimating similar partitions.

4. Simulation studies

This section showcases simulations conducted to assess the performance of HMMs in change-point detection within artificial data scenarios of series characterized by irregular spacing. An MC scheme is applied with 50 replications. The MCMC settings are the same as those presented in Section 2.2. The value of K (and R) was assumed to be known, as established in the artificial data generation. Convergence issues were not observed for any MC sample.

4.1. Artificial data

A total of 7 different scenarios are studied, 3 generated via HMM$_S$ and 4 via HMM$_G$. Among them, both purely artificial scenarios and scenarios motivated by real applications are considered. Each of the simulated series comprises a total of $n = 1{,}000$ irregularly spaced observations. For the purely artificial cases, the d_i's were generated between $(0, 1]$ using the Uniform distribution. Figure 3 (a, c) display the d_i's. Panel (c) shows that low and high values are scattered throughout the series. For each series, approximately 90% of the distances are small < 0.5 (strong dependence). The remaining d_i's establish weak to moderate dependencies. These distances were then randomly allocated to the series to form the irregular spacing. Concerning the spatial dependence, assume $\boldsymbol{\beta} = (4, -8)$ reflecting the behavior in Figure 1 (a). The transition probability matrix $\boldsymbol{\alpha}_{K\times K}$ is defined to impose strong clustering power. In this case, the probability that adjacent observations are allocated in the same component (main diagonal) is 0.95, while the probability of transitioning to an adjacent component is 0.02. These settings were chosen to ensure the presence of clusters in the series and prevent consecutive observations from being classified into non-adjacent components. In the absence of serial dependence, assume $\boldsymbol{q}_0 = (1/K)\mathbf{1}_{(K\times 1)}$.

The degree of proximity between means of the components was considered to build the data scenarios. In the HMM$_S$, assume different overlap levels between components. Denote the scenario by C_A^K with $K = 8$ and overlap rate $A = \{0.05, 0.10, 0.25\}$. Observations were generated between [0, 30]. This scale allows up to a maximum of 8 components under study to be accommodated, while simultaneously achieving the desired overlap rate. The means of the Gaussian components were determined based on the data scale, and the variances were chosen to achieve the target overlap. In the HMM$_G$, two scenarios are referred to as "purely artificial". They are denoted by $C_{\boldsymbol{s}=(2,4)}$ and $C_{\boldsymbol{s}=(3,3)}$. The other two cases are simulated based on the features from two real applications discussed later in this paper (IPC Mexico and Solar flares). Here, denote $C^*_{\boldsymbol{s}=(1,2,2)}$ and $C^*_{\boldsymbol{s}=(3,1,1)}$. In purely artificial scenarios, the components within each group have equal means. Then $R = 2$ was considered, and the component parameters were chosen to reflect data distribution with distinct bimodality

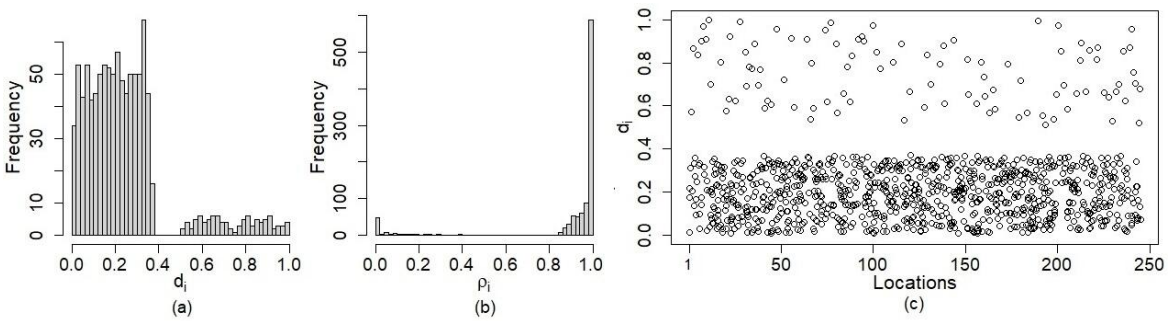
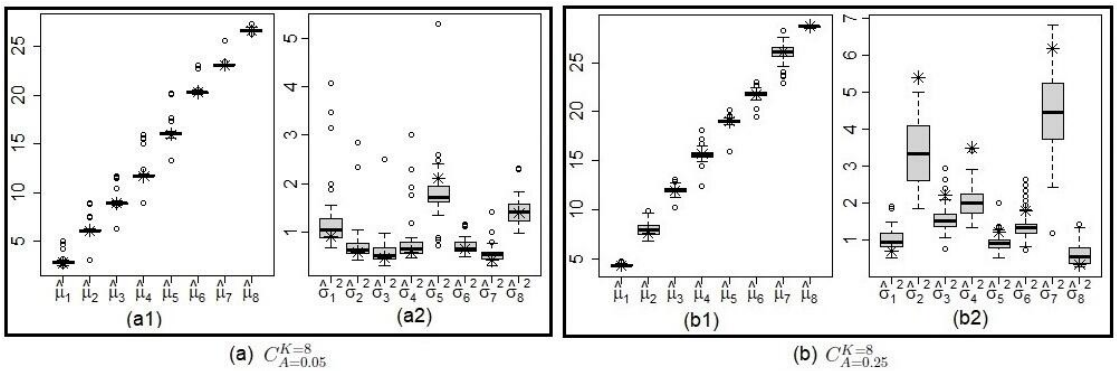


Figure 3: Distances between adjacent observations in a purely artificial series. Panel (a): Histogram of the d_i's. Panel (b): Markovian dependence probabilities ρ_i. Panel (c): Distances d_i versus location.

(groups with moderate means separation). In the scenarios mimicking real data, the groups are characterized by components having close means (not identical).

4.2. Performance of HMM$_S$

The MC study examines how HMM$_S$ behaves under data scenarios with low and high overlap ($A = 0.05$ or 0.25). Preliminary studies with a small number of samples from $C^{K=8}_{A=0.10}$ did not uncover any inference or classification challenges. MC results demonstrate that the model exhibited fitting stability. Figures 4 $(a1)$ and $(b1)$ indicate that the boxplots of $\hat{\mu}_k$ are positioned in increasing order, following the imposed increasing and restricted ordering of the means of the HMM$_S$ components. Figure 4 and Table 1 (prior structure for HMM$_S$) present reasonable outcomes for the estimation of means and variances of the components, with a noticeable lower MC variability in mean estimation compared to variance estimation. However, the issue of underestimating variances, particularly among components with higher true variances and under scenarios with higher overlap rates, stands out. For these cases, Figures 4 $(a2)$ and $(b2)$ depict the true values positioned above the boxplots. The error for estimating σ^2_k can reach up to 80% (Table 1).



(a) $C^{K=8}_{A=0.05}$ (b) $C^{K=8}_{A=0.25}$

Figure 4: MC study for the HMM$_S$, with $\boldsymbol{\alpha}^{50\%}$ and 10% prior overlap. Boxplots exhibit the posterior means for μ_k or σ^2_k. Asterisks indicate true values.

Table 1 indicates a classification error that is not excessively high (with a small MC interquartile range). Results reveal that HMM$_S$ provides a better fit for data arising from mixtures with low overlap rates. As a consequence, one can observe lower percentages of errors for μ_k and σ^2_k, as well as a lower classification error. However, when exploring data with higher overlap in mixtures, the MC results are less favorable. The next section develops a similar analysis for the HMM$_G$.

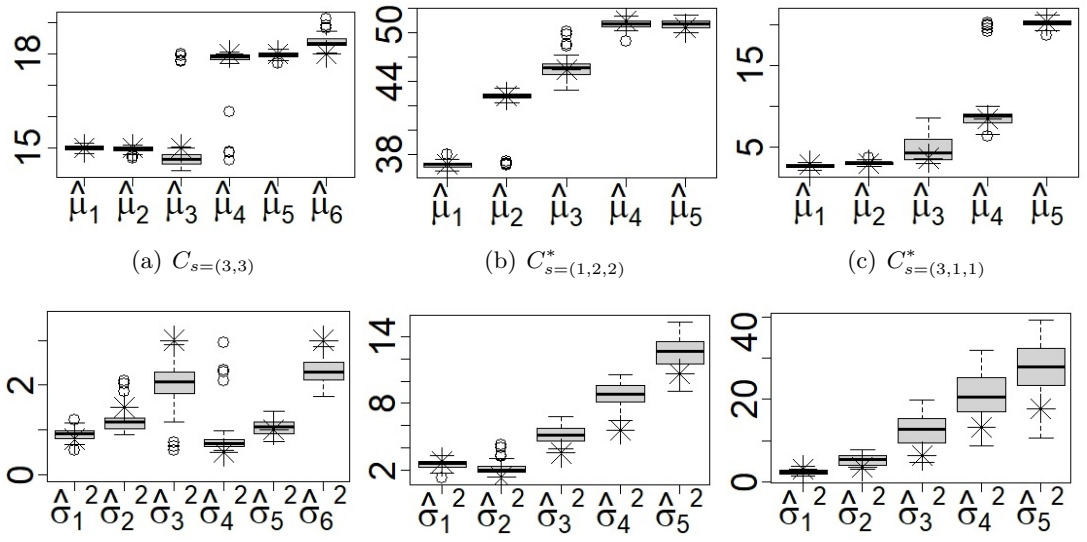
Table 1: MC study: HMM$_S$, $\boldsymbol{\alpha}^{50\%}$, 10% overlap. Error: 95% HPD without the true value. "Error Tot. μ_k" (or "σ_k^2") is the sum of errors across all components. "Max. Error σ_k^2" indicates the highest variance error (component in parentheses). Last column: MC median of the % of misclassified observations (interquartile range in parentheses).

Artificial data	Monte Carlo results (%)			
	Error Tot. μ_k	Error Tot. σ_k^2	Max. Error σ_k^2	Clas. Error
$C_{A=0.05}^{K=8}$	12.5	8.5	20.0 (5)	2.9 (0.9)
$C_{A=0.25}^{K=8}$	20.3	38.8	80.0 (4)	16.5 (2.6)

4.3. Performance of HMM$_G$

Now, the HMM$_G$ is explored in an MC scheme with four scenarios (two purely artificial and two inspired by real data). Figure 5 highlights that under the chosen prior configuration for the HMM$_G$, the estimates of μ_k and σ_k^2 align with the true values (asterisk). In Panels $(a-c)$, note that the positions of the boxplots tend to express the group of means. Panels $(d-f)$ demonstrate an ascending order of boxplot for σ_k^2's within a group, reflecting the imposed model constraint of ordering variances within each cluster. Furthermore, lower variability is observed for means than for variances, similar to what was evidenced for the HMM$_S$. The patterns obtained from scenarios assuming groups of equal means (denoted by C) and scenarios inspired by real data (C^*) are similar, even considering issues with variance estimation. For equal means scenarios, a tendency to underestimate the higher-value variances within each group is observed (asterisk below the boxplot), contrasting with the occurrence of overestimation in scenarios based on real data (asterisk above the boxplot).

The performance analysis of the HMMs, in addition to goodness-of-fit, requires the evaluation of classifications due to their clustering purpose. Table 2 summarizes the number of classification errors, both for variance (component-wise classification) and for the group of means. The results are similar for any of the scenarios under study. The lower values (below 8%) of classification errors validate the use of the max($\boldsymbol{p}_i$) criterion for the mean. However, for variance identification, this criterion can provide more than 50% of errors.



(a) $C_{s=(3,3)}$ (b) $C^*_{s=(1,2,2)}$ (c) $C^*_{s=(3,1,1)}$

(d) $C_{s=(3,3)}$ (e) $C^*_{s=(1,2,2)}$ (f) $C^*_{s=(3,1,1)}$

Figure 5: MC study for HMM$_G$ with $\boldsymbol{\alpha}^{75\%}$, $Var(\sigma_k^2) = 1$ and overlap of 25%. Posterior estimates for μ_k and σ_k^2 (asterisk indicates true value). Panels $(a-c)$: $\hat{\mu}_k$. Panels $(d-f)$: $\hat{\sigma}_k^2$. Remark: $C_{s=(2,4)}$ (Supplementary Material) resembles $C_{s=(3,3)}$.

The MC study reveals a similarity among the p_{ik} in $\boldsymbol{p}_i$, mainly for the components within the same group r. This justifies the tendency of the criterion max($\boldsymbol{p}_i$) to correctly classify

Table 2: MC Study for HMM_G under $\boldsymbol{\alpha}^{75\%}$, $Var(\sigma_k^2) = 1$ and overlap of 25%. Evaluating mean-based and variance-based classification errors using the $\max(\boldsymbol{p}_i)$ criterion. Percentages refer to MC medians.

Classification type	Classification errors by scenarios (%)			
	$C_{s=(2,4)}$	$C_{s=(3,3)}$	$C^*_{s=(1,2,2)}$	$C^*_{s=(3,1,1)}$
Means group	4.6	5.8	8.0	4.5
Variance	50.6	52.4	30.4	45.1

observations with respect to the group, but indiscriminately classify the observation into any component within a group (possible variance classification error). Generally, classification probabilities are below 0.5 for misclassified observations. However, correctly allocated data points, with criterion $\max(\boldsymbol{p}_i)$, have the highest p_{ik} for their original component k, which may exceed 0.5. Based on the information obtained from the classification probability vector, several criteria were proposed at two levels of certainty for variance classification (decisive and doubtful). The adopted criterion $P_{0.5}$ yielded the best results in the examined scenarios. See details in the Supplementary Material.

Table 3: MC study: HMM_G, $\boldsymbol{\alpha}^{75\%}$, $Var(\sigma_k^2) = 1$, and 25% overlap. Assessing the $P_{0.5}$ certainty-level classification. Median MC values are expressed in % (the higher, the best).

Performance measures	Scenarios			
	$C_{s=(2,4)}$	$C_{s=(3,3)}$	$C^*_{s=(1,2,2)}$	$C^*_{s=(3,1,1)}$
1. % Decisive Obs.	35.05	26.55	85.69	54.97
2. % Decisive Correctness	71.50	67.00	75.89	72.22
3. % Decisive Error in Neighbor	85.38	88.32	90.68	92.87
4. % Doubtful Correctness	67.40	73.20	68.43	75.53
5. % Doubtful Error in Neighbor	38.24	42.87	40.30	49.80
6. % Doubtful Correctness in Neighbor	97.13	98.53	98.49	99.32
7. % Total Correctness	69.00	72.90	74.69	72.79

Table 3 presents results about the efficiency of the $P_{0.5}$ criterion in HMM_G. The following quantities are analyzed: percentage of decisive classifications (row 1), percentage of correctness among decisive and doubtful classifications (rows 2 and 4), percentage of misclassification allocating "decisive observations" to an adjacent component (row 3), percentage of misclassification allocating "doubtful observations" to an adjacent component (row 5), percentage of well-classified doubtful observation where one of the components is true and the other is a neighbor (row 6), and number of correct classifications (decisive + doubtful classifications) (row 7). These measurements are important to evaluate the classification performance of the HMM. An error due to allocation to a component close to the true one is less critical than classifying into distant components.

Table 3 shows that the amount of decisive classification is not small (26%-86%). The lowest value is associated with the scenario of component groups with equal means, $C_{s=(3,3)}$, with high overlap among components forming the groups. The higher percentages are obtained in scenarios of mixtures with lower overlap between components, which is attributed to the presence of mean groups with only 1 or 2 components. Decisive correctness is notably substantial ($\geq 67\%$). Misclassification tends to occur towards adjacent components for decisive errors ($> 85\%$). This behavior is not similar for doubtful errors ($< 50\%$). The model's ability to produce doubtful classifications that at least include the true component is also evident ($> 67\%$ in row 4, $> 97\%$ in row 6). Regarding the total correctness, the adopted certainty-level criterion outperforms ($\geq 69\%$) the $\max(\boldsymbol{p}_i)$. These results reveal that the HMM_G has reasonable variance-based classification capacity, as it tends to at least identify the σ_k^2 from a component adjacent to the true one.

5. Model comparison

This section is focused on a comparison between the proposed HMMs and 4 approaches from the literature to cluster or detect change points. Here, the HMM_S and HMM_G will be denoted as M_1, and their application depends on the scenario under study. The alternative approaches are represented as follows. The M_2 corresponds to the Bayesian mixture of Gaussian components for clustering, without spatial dependence, and with a Dirichlet process prior to incorporating uncertainty about K. This option is implemented in the `profRegr` function of the `R` package `PreMiuM` (Liverani, Hastie, Azizi, Papathomas, and Richardson 2015). The M_3 refers to the spatial partition product model present in the `sppm` function of the `R` package `ppmSuite` (Page and Quintana 2016). The M_4 is the reference partition product model by Barry and Hartigan (1993), implemented in the `R` package `bcp` (Erdman and Emerson 2007). Lastly, the classical non-parametric approach M_5 identifies change-points of arbitrary nature in series, without spatial dependence, using the `cpt.np` function of the `R` package `changepoint.np` (Haynes *et al.* 2017). The methodologies related to M_2, M_3, and M_4 allow for the identification of mean changes.

An MC analysis is conducted in 4 scenarios of irregular series. One scenario involves components with means having an intermediate level of separation, $C^{K=8}_{A=0.10}$ ($M_1 = \text{HMM}_S$), and three scenarios with grouped means ($M_1 = \text{HMM}_G$). For the HMMs, assume K and R to be known. The competing models are applied under their respective default configurations, with the MCMC settings remaining the same as in Section 2.2 for Bayesian cases. In this study, for M_4 (standard PPM), change points are considered to be observations with high posterior probability (> 0.5). For change detection using M_2 and M_3, a post-processing clustering method is employed. Here, consider a robust version of the *k-means* known as *Partitioning Around Medoids* (PAM) developed by Kaufman and Rousseeuw (1990). This method is based on an estimated similarity matrix (PSM). Changes are identified by M_1 using both PAM and the proposed classification criterion. The output of the `R` functions of M_4 and M_5 does not provide sufficient information to apply the PAM.

Three measurements are used to evaluate the similarity between estimated and true partitions. Consider the *Adjusted Rand Index* (ARI) in Hubert and Arabie (1985), being a pair-counting measurement, along with the *Adjusted Mutual Information* (AMI) and *Normalized Variation of Information* (NVI) classified as information theory-based measurements (Vinh, Epps, and Bailey 2009). To calculate them, use the functions `arandi`, `AMI`, `NVI` from the `R` package `aricode` (Chiquet, Rigaill, and Sundqvist 2022). Additionally, one can compare the detected change points with the true configuration using the *True Positive Rate* (TPR) calculated for binary classifications. The values of these measurements are bounded in $[0, 1]$. Here, the value $(1 - \text{NVI})$ is used to ensure a uniform interpretation of all three measurements. Higher values indicate greater similarity between estimated and true clustering.

The model comparison is developed through two distinct approaches: MC analysis of mean changes, and analysis of “mean and variance" changes. The first approach accounts for all 5 models, while the second one is focused on M_1 and M_5. Note that M_5 estimates change points of any nature, i.e., its corresponding package does not provide sufficient information to classify a change as related to the mean or the variance. This is a critical contrast with respect to M_1 (HMM_G).

In the analysis of Bayesian models, consider only the convergent MC replications. Our proposed HMMs converged in all replications under study, however, M_2, M_3, and M_4 indicated a lack of convergence for some samples. Convergence issues were detected in a maximum of 18% of the replications for at least one of the models in each scenario. The M_2 is an exception with convergence issues for 40% of the replications in $C^{K=8}_{A=0.10}$ (mixture components with separated means).

Figure 6 shows the number of mean changes produced by each model in relation to the true quantity. Panels $(b - d)$ indicate that M_2 and M_3 identify a higher number of changes, more prominently in the real data-inspired scenarios in (c) and (d). The case with components

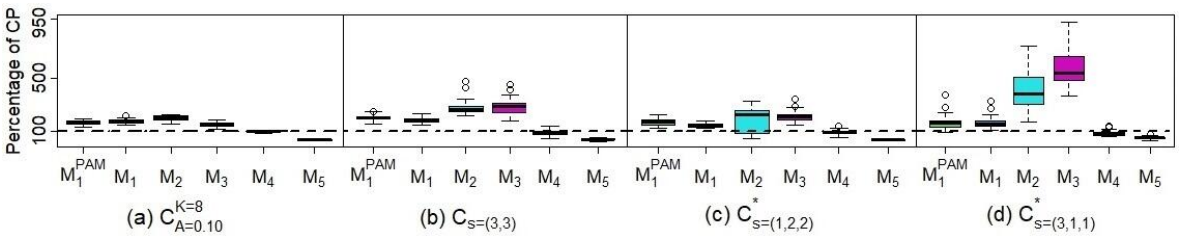


Figure 6: MC study comparing models. Estimated number of mean changes. Dashed line indicates the true number. Change points generated by M_5 are associated with changes of any nature (μ_k or σ_k^2).

having separated means, Panel (a), exhibits boxplots with lower variability and concentrated near the level 100 (the true number of changes). Let M_1^{PAM} denote the HMM model under the PAM method. This option provides boxplots suggesting a similar number of changes across all panels. Overall, the HMMs tend to detect more changes than the original quantity. However, this number does not appear to deviate significantly from the true value (except for M_2 and M_3). The M_4 and M_5 detect the lowest number of changes (they underestimate the true quantity).

Table 4: Model comparison (MC study). Similarity (ARI, AMI, NVI) between estimated and true partitions. TPR due to binary response (change: yes/no) for each location. Scenarios in each column: $1 = C^{K=8}_{A=0.10}$, $2 = C_{s=(3,3)}$, $3 = C^*_{s=(1,2,2)}$, $4 = C^*_{s=(3,1,1)}$. Remark: changes from M_5 can be related to the mean or variance.

Change	Model	ARI				AMI				1-NVI				TPR			
		1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
μ_k	M_1^{PAM}	0.84	0.76	0.68	0.84	0.86	0.65	0.69	0.77	0.76	0.49	0.53	0.64	0.94	0.79	0.82	0.82
		(0.05)	(0.05)	(0.15)	(0.13)	(0.04)	(0.05)	(0.15)	(0.15)	(0.04)	(0.05)	(0.09)	(0.15)	(0.03)	(0.07)	(0.07)	(0.07)
	M_1	0.80	0.76	0.82	0.89	0.80	0.81	0.85	0.87	0.90	0.87	0.89	0.88	0.94	0.78	0.78	0.73
		(0.09)	(0.10)	(0.08)	(0.14)	(0.07)	(0.06)	(0.04)	(0.10)	(0.03)	(0.04)	(0.03)	(0.08)	(0.03)	(0.08)	(0.07)	(0.11)
μ_k and σ_k^2	M_1	-	0.49	0.75	0.55	-	0.58	0.79	0.65	-	0.79	0.86	0.76	-	0.60	0.70	0.59
		-	(0.10)	(0.09)	(0.12)	-	(0.07)	(0.06)	(0.06)	-	(0.02)	(0.03)	(0.03)	-	(0.07)	(0.05)	(0.08)
	M_5	-	0.31	0.34	0.46	-	0.53	0.53	0.62	-	0.63	0.63	0.66	-	0.03	0.02	0.03
		-	(0.09)	(0.10)	(0.21)	-	(0.05)	(0.06)	(0.12)	-	(0.04)	(0.05)	(0.11)	-	(0.02)	(0.03)	(0.03)

In Figure 6 and Table 4, the comparison $M_1 \times M_1^{\text{PAM}}$ demonstrates that the classification criteria, proposed for the HMMs (Section 3), yields results of similar or superior quality to the PAM method across the scenarios. Thus, for comparison against competing models, the results from the HMMs will be subjected to the classification criteria proposed for them in this study. In the analysis of "mean and variance" changes, only scenarios involving mixtures with grouped mean components are investigated. As depicted in Table 4, overall, the HMM$_G$ (M_1) exhibits notably superior performance compared to M_5, mainly in the detection of true change points (TPR). Only for the scenario $C^*_{s=(3,1,1)}$ the M_5 shows performance comparable to M_1 in terms of partition similarity measures (ARI and AMI).

Figure 7 reveals that M_1 generally exhibits superior or equivalent performance compared to the other models in identifying mean changes. This can be seen in terms of the correct detection of locations with changes and the partition quality of the series. It's important to emphasize that none of the scenarios indicated a single model having the best behavior across all aspects. Furthermore, the scenario $C^{K=8}_{A=0.10}$ (separated means) was the only one where the majority of models demonstrated similar performance, particularly in terms of partition quality. The M_2 stood out from the others by presenting some measurements with a larger MC variability for the real data-inspired scenarios, notably in Figure 7 [1, 3], panel in row 1 and column 3, [2, 4], [3, 4], and [4, 4].

The M_2 and M_3 exhibit performance close to that of the HMMs in terms of true change

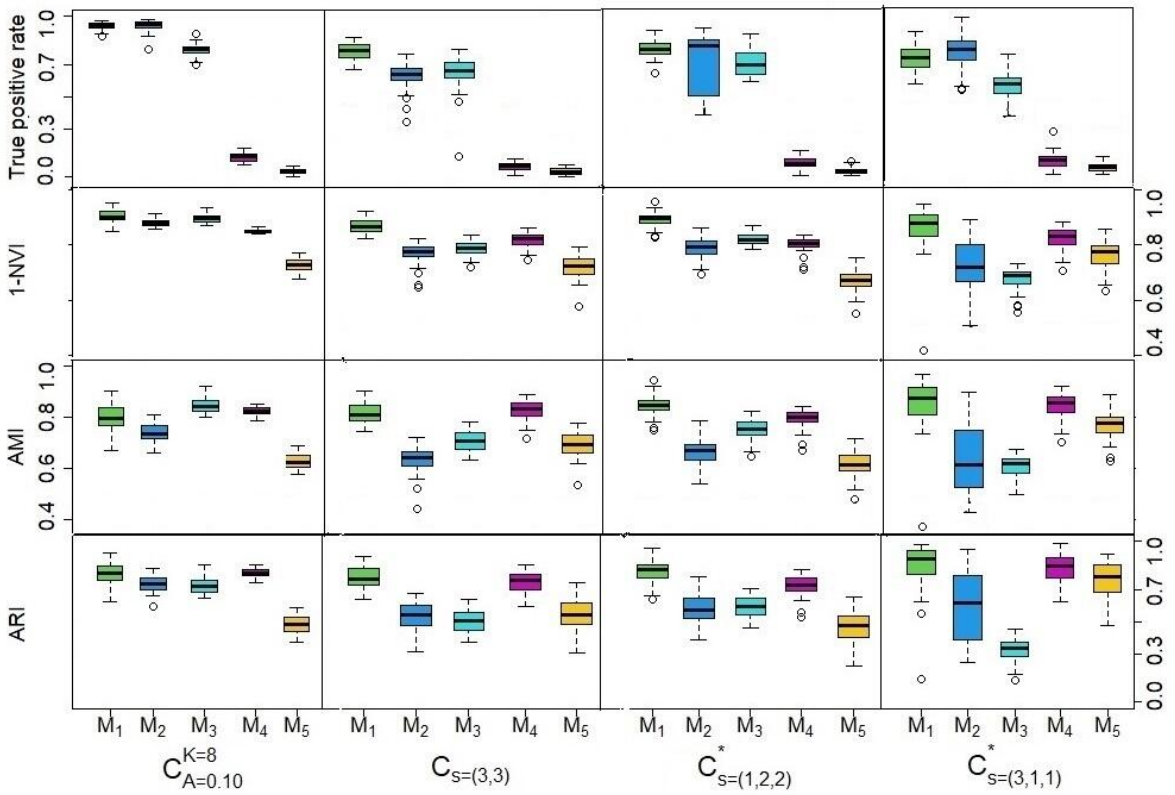


Figure 7: Model comparison with boxplots (MC study). Similarity (ARI, AMI, NVI) between estimated and true partitions. TPR due to binary response (change: yes/no) for each location. Scenarios in each column: $1 = C^{K=8}_{A=0.10}$, $2 = C_{s=(3,3)}$, $3 = C^{*}_{s=(1,2,2)}$, $4 = C^{*}_{s=(3,1,1)}$. Remark: changes from M_5 can be related to the mean or variance.

point detection. This suggests that, in terms of efficiency in locating the changes, these two options can be considered interesting alternatives to M_1. However, it's important to note that M_2 demonstrated unstable performance in some aspects for real data-inspired scenarios. Furthermore, M_2 and M_3 tend to estimate a larger number of change points than the competitors (Figure 6). When examining the partition quality, the M_4 is the one displaying a good performance closer to that of M_1. This indicates that M_4 is a strong competitor when the focus is on the overall partition of the series.

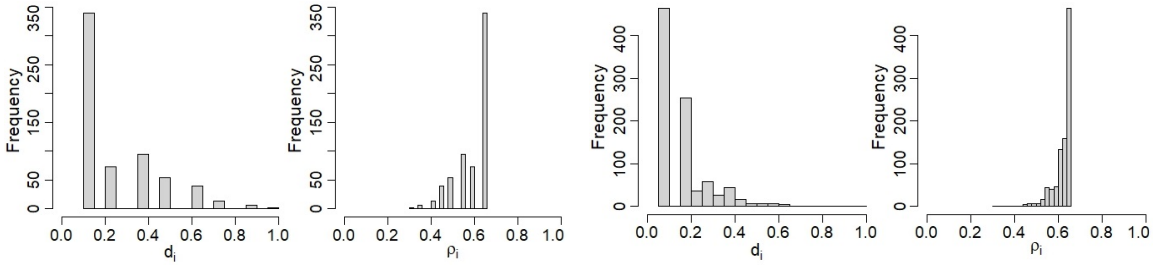
6. Real applications

In this section, the HMMs are applied to 3 real daily series. Two of them are related to stock market indices: Ibovespa (Brazil) and IPC Mexico. The third time series corresponds to astronomical data concerning solar explosions durations (in seconds). The financial indices represent adjusted closing prices observed from January 01 (2018) to August 12 (2021). These series were extracted from the *Yahoo Finance*[2] website using the `getSymbols` function from the `R` package `quantmod` (Jeffrey and Joshua 2020). The period for the solar explosion data spans from August 12 (2008) to August 12 (2021). The dataset includes days with multiple explosions, then the average duration is adopted for the analysis (the median provides a similar series). The astronomical data is available through the NASA website[3] accessed October 30 (2021). The observations in these applications are rescaled to align with the magnitude explored in the simulated studies (financial data is divided by 1000, explosion duration is divided by 100).

[2] https://finance.yahoo.com
[3] https://hesperia.gsfc.nasa.gov/fermi/gbm/qlook/fermi_gbm_flare_list.txt

The irregular spacing in both financial series is highly similar, characterized by the absence of observations on weekends and holidays, as well as the presence of a few missing data points. In the solar data, there can be periods with no recorded explosions ranging from 2 to 278 days in sequence, thus indicating irregular spacing. These missing observations can happen for several reasons, depending on the mechanism to register data or external conditions. To ensure a greater diversity of distinct degrees of dependence, 30% of the observations are randomly removed from each series. Please note that this removal strategy reduces the sample size, therefore, fewer observations are available to fit the models (i.e. such a strategy is not conducive to the effective implementation of the HMMs). After the alteration, we have $n = 623$, 636, and 926, corresponding to 47%, 48%, and 19% of the total days, for Ibovespa, IPC Mexico, and solar explosions, respectively. The analysis of these series without data removal can be found in the Supplementary Material. Conclusions align with those in this paper, except for the solar data showing better component classification with the full series.

For the financial series, the d_i's rescale to $(0, 1]$ was performed by dividing each d_i by the maximum distance. The financial series exhibit a similar pattern of irregular spacing, encompassing almost the entire interval $(0, 1]$. For the series of average explosion duration, the rescale has two steps: (i) calculate $\log(d_i)$'s and (ii) divide them by the maximum log distance. The log preserves the differences among the smaller distances and almost equalizes the larger values; see Figure 8. This choice was motivated by the presence of a few large d_i's viewed as outliers in the series. The relationship $d_i \times \rho_i$, defining the strength of Markov dependence, is influenced by these outliers imposing that most d_i's are below 0.5 associated with large ρ_i. The log scale ensures that the distribution of d_i's covers different parts of the interval (0, 1). Figure 8 indicates that the ρ_i's are spread across $(0.3, 0.7)$, with a prevalence of the stronger dependence ($\rho_i \approx 0.7$).



(a) Ibovespa (b) Solar explosions

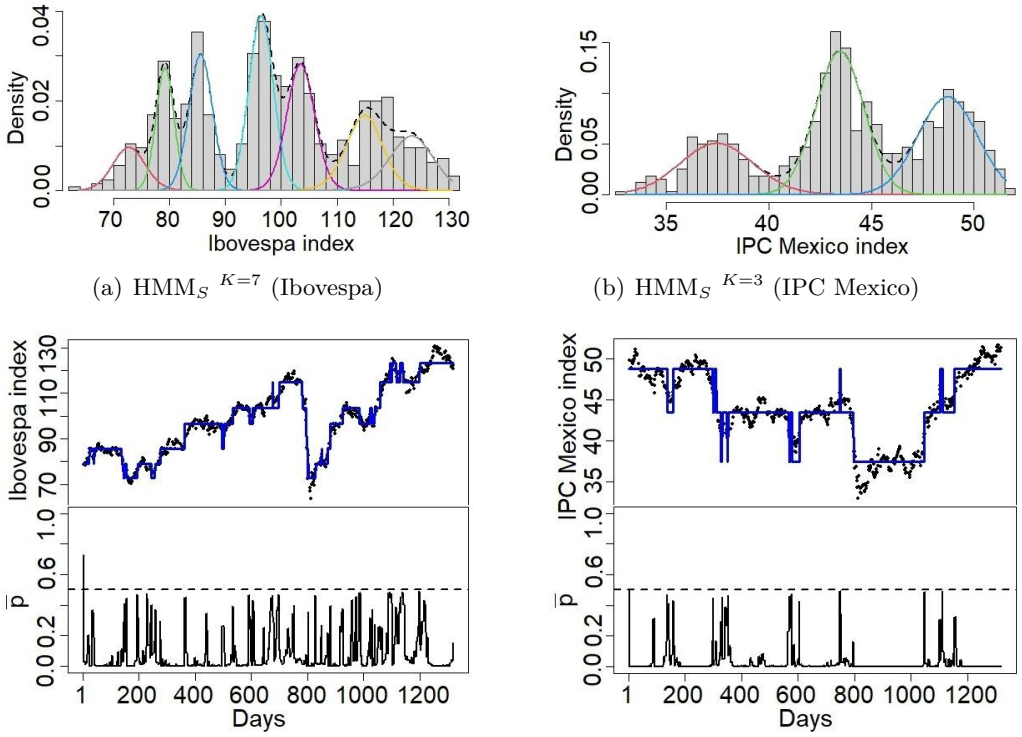
Figure 8: Histograms of d_i's and corresponding Markovian dependence probabilities ρ_i's in the HMMs related to real applications. The behavior of IPC Mexico resembles Panel (a).

This study applies the HMMs, for the 3 datasets, assuming different K and R, as described in Section 2.2. Configurations showing a lack of convergence, in the chains of μ_k or σ_k^2, were excluded from the analysis. Figure 11 (a) shows an asymmetric and unimodal behavior for the solar explosion dataset. This aspect motivates the use of the HMM_G for this particular application. The option $R = 1$ is determined assuming the criterion based on the number of histogram modes. However, we would like to handle the mode region and the tail region, therefore, consider $R = 2$ and $R = 3$ here. The analysis indicates that the HMM_G with $R = 3$ places more emphasis on changes in the mean, while the $R = 2$ case only assesses changes in the tail. Due to the multimodal feature of both financial series histograms, it makes sense to assume the HMM_S for them; details can be found in the Supplementary Material.

Model evaluation relies on the following goodness-of-fit measurements: Deviance Information Criterion (DIC), Widely Applicable Information Criterion (WAIC), Log Pseudo Marginal Likelihood (LPML), and a clustering-based model selection criterion known as Integrated Completed Likelihood (ICL). Further information on these criteria can be found in Spiegelhalter, Best, Carlin, and der Linde (2002), Watanabe (2010), Gelman, Carlin, Stern, and Rubin (2003), and Biernacki, Celeux, and Govaert (2000). For all cases, we establish that lower values indicate a better fit.

Figure 9 exhibits results of the selected HMM_S and their respective classifications produced for the financial series. Panels (a) and (b) highlight the good fit of the predictive density to the data distribution. In Panels (c) and (d), note that the blue line (posterior mean) follows (as expected) the trajectory of the series. Some occurrences of singleton clusters can be observed in the results. For instance, in Panel (d), around the days 320 and 355 (with IPC index ≈ 41) there is no change in the mean except for two isolated observations. This phenomenon may be associated with observations located in the tails of adjacent mixture components.

Figure 10 shows the HMM_G fitted to the IPC Mexico data. In Panel (b), the blue line indicates the posterior mean of μ_k related to the component where the observation is allocated. Note that the blue line tracks the series path, with all classifications exhibiting low uncertainty ($\bar{p}^* < 0.5$) for most days, especially between 400 and 1100. In Panel (c), it can be seen that the analysis provides variance classifications with low uncertainty for a large portion of the series (e.g. $\bar{p} \approx 0$ between 800 and 1000). Nearly all observations are categorized as decisive. The HMM_G, selected for the Ibovespa via information criteria, proves ineffective in adhering to the data; see details in the Supplementary Material. Thus, the $\text{HMM}_S^{\ K=7}$ is the preferred model for the Ibovespa.

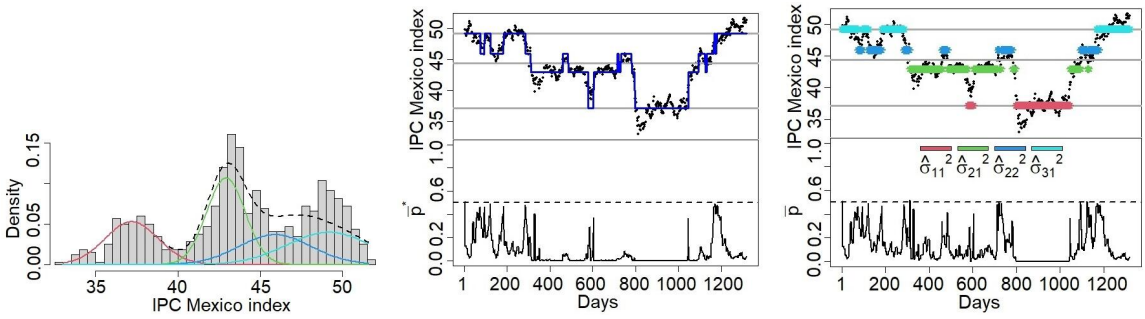


(a) $\text{HMM}_S^{\ K=7}$ (Ibovespa) (b) $\text{HMM}_S^{\ K=3}$ (IPC Mexico)

(c) Mean-based classif. (Ibovespa) (d) Mean-base classif. (IPC Mexico)

Figure 9: Model fit and mean-based classifications from HMM_S. Panels (a, b): histogram with estimated mixture density (dashed line), and estimated densities of components (solid lines). Panels (c, d): data series (in black) and posterior mean of μ_k (in blue). Subpanels $\bar{p}$ represent the degree of uncertainty in mean-based classification.

Table 5 demonstrates that the HMM_G applied to the explosion dataset performs as effectively as the applications to financial data concerning classification certainty for the mean. At least 96% of the observations are classified with low uncertainty. In terms of variance, the models classify at most 40% of the observations as decisive. The HMM_G with a greater number of components, applied to the explosion data, is the least effective in classifying variance, with a percentage of "decisive" between $\approx 11\%$ and $\approx 22\%$. Nonetheless, these models are indicated as good alternatives in terms of goodness-of-fit. According to the study, the $\text{HMM}_G^{\ \substack{K=5 \\ R=3}}$ is a reasonable choice for the application. This conclusion is obtained when accounting for both "classification power" and goodness-of-fit. For the chosen HMM_G in each series, no



(a) HMM_G $\frac{K=4}{R=3}$ fit (b) Mean-based classification (c) Variance-based classification

Figure 10: HMM_G fitted to the IPC Mexico. Panel (a): histogram, estimated mixture (dashed line), and components (solid lines). Panel (b): Data series (black points) and estimated μ_k (blue line). Panel (c): estimated σ_K^2. Uncertainty levels are expressed in Panels (b) and (c).

issues with estimating the components group size vector $\boldsymbol{s}$ were evident, which is crucial for classification. The outcomes provided $P(\hat{s}) >> 0.5$, indicating a strong conviction about component grouping.

Table 5: HMM_G fitted with different K and R to the Solar explosions data. The last two columns indicate the percentage of observations classified with low uncertainty. A favorable result is suggested when a high value is obtained in the last three columns, and a low value is detected for ICL, DIC, LPML, and WAIC.

K	R	ICL	DIC	LPML	WAIC	$P(\hat{s})$	$\bar{p}^* \leq 0.5$ (%)	$\bar{p} \leq 0.5$ (%)
4	2	6364.02	4978.73	2491.52	2494.55	1.00	100.00	40.71
	3	6359.49	4977.26	2490.93	2493.95	0.99	87.26	41.47
5	2	6407.32	4923.90	2463.74	2466.52	1.00	100.00	35.64
	3	6407.13	4922.07	2462.74	2465.51	1.00	98.06	36.18
6	2	6520.89	4891.64	2446.85	2449.37	1.00	100.00	22.79
	3	6517.22	4889.44	2445.60	2448.14	0.56	96.98	23.43
7	2	6607.02	4870.71	2435.68	2438.04	0.98	100.00	15.77
	3	6608.18	4870.84	2435.74	2438.02	0.47	96.11	15.77
8	2	6684.45	4858.66	2429.35	2431.66	0.91	100.00	11.45
	3	6683.52	4858.73	2429.49	2431.83	0.44	97.19	11.45

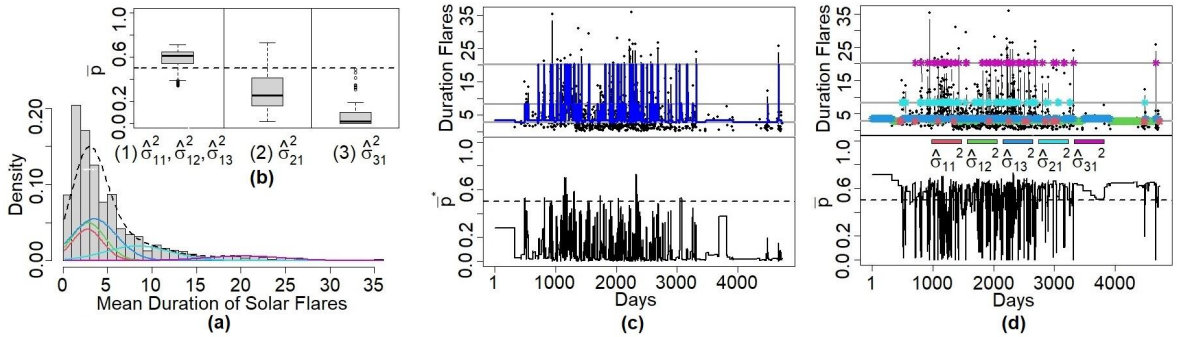


Figure 11: HMM_G $\frac{K=5}{R=3}$ fitted to the solar data. Panel (a): estimated mixture. Panel (b): uncertainty level in variance classification by group. Panel (c): data (in black) and posterior mean of μ_k (in blue) with uncertainty level ($\bar{p}^*$). Panel (d): data (in black), variance classifications (colored asterisks), and uncertainty level ($\bar{p}$).

Figures 10 and 11 (a) present goodness-of-fit results related to the IPC Mexico and Solar datasets. The HMM fit for the solar data exhibits significant overlap among the Gaussian

components. As a result, upon observing the estimated densities in Figure 11 (a), it becomes evident that the number of components allocated to each of the mean groups can be $\hat{s} = (3, 1, 1)$. For the IPC Mexico, the estimate is $\hat{s} = (1, 2, 1)$.

In contrast with the financial series, Figure 11 (d) indicates that, for the solar data, the uncertainty level in variance classification is ≈ 0.5 for most observations. In Panel (b), it is evident that only classifications corresponding to the first three components, Group g_1, with red, green, and dark blue asterisks in (d), tends to be designated as "doubtful" ($\bar{p} > 0.5$). However, the model does not exhibit excessive uncertainty in this type of classification, given that the median of $\bar{p}$ in (b) is ≈ 0.6 with a small interquartile range. One possible reason for the model's lower certainty in variance classification is the high overlap between mixture components within the same group.

The HMMs applied to the financial data estimated $\approx 1.0 - 7.1\%$ of the series as mean changes and $\approx 1.4 - 1.6\%$ as variance-only changes. The solar data exhibited higher volatility, with 22% of the series estimated as mean changes, and 9% as variance-only changes. The initial period of the COVID-19 pandemic (days 788 to 809, corresponding to February 27 to March 19 (2020) are detected as change points in the financial series. This aspect can be seen in Figure 9 (c, d), and Figure 10 (b, c).

In summary, both HMM_S and HMM_G seem suitable for the IPC Mexico data. A brief comparison study (HMM_S vs. HMM_G) is included in the Supplementary Material. The choice of the HMM is left to the analyst. The study must account for the research goal, which may involve detecting alterations solely in the mean or extending to variance, alongside other factors such as the number of changes and the desired level of model certainty. Emphasizing classification accuracy is important, while still ensuring goodness-of-fit.

7. Conclusions

This work introduced Bayesian mixture models with Markovian dependence, inspired by Mayrink and Gonçalves (2020), to identify change points in irregularly spaced series. Two models are proposed: one for identifying mean changes (HMM_S), and one for detecting changes in the mean and/or variance (HMM_G). These proposals incorporate a dependence structure to handle the irregular spacing. They allow inferences about the existence of Markovian dependence between neighboring observations. This aspect stands out as a contribution to change point detection in irregular series. Specifically, the HMM_G is designed to group mixture components with equal or close means but having distinct variances. Here, a group change is the focus rather than a simple mean change. A variance change involves alternating components within groups.

Classification methods were proposed for assigning observations to mixture components. They are based on the MAP approach. Simulation studies showed that the posterior mean of μ_k reasonably follows the series path. The HMMs yield accurate mean classifications for nearly all observations. In scenarios with high overlap between components, the occurrence of singleton clusters was more frequent. This phenomenon is not well-addressed in the change-point literature, making it an interesting topic for future improvements. The MAP criterion was not effective for classifying variance in the HMM_G due to the high overlap in the mixture. Thus, a certainty-level classification method was proposed. The simulation studies revealed a substantial number of "decisive" classifications (high accuracy rate). Even when making mistakes, the tendency is to classify into adjacent components. The same can be said about the category "doubtful".

The simulation results also demonstrate the HMM's efficiency in terms of goodness-of-fit and inference for μ_k. Some estimation difficulty was noted for σ_k^2, especially for components with high variability (high overlap). MC studies indicated that the HMM_G requires priors to establish a strong Markovian structure. In contrast, a less strong dependence is recommended for the HMM_S (lower overlap). Inconsistencies in fitting can occur depending on the overlap

level and the strength of dependence imposed *a priori*.

The HMMs were also compared with change point detection methods from the literature. The selected methods include cases assuming independence or spatial dependence between observations. In this context, for mean change detection, the HMMs demonstrated superiority or equivalence to the competitors. However, for identifying changes in both the mean and variance, HMM_G exhibited notably superior performance compared to the non-parametric competitor.

The real data analysis suggested that both HMM's were suited for the chosen financial series. Nevertheless, the HMM_S is an interesting choice for the Ibovespa, due to its flexibility in tracking the rising trajectory. The HMM_G was fitted to the solar data due to the need for multiple highly overlapping components to accommodate the asymmetry. Here, a reasonable number of observations was obtained with "decisive" and "doubtful" classifications. The "doubtful" cases are associated (allocated with a moderate uncertainty) with components having high overlap. To select the HMM, in addition to goodness-of-fit, one should consider the research goal (mean or mean/variance changes), the expected number of changes, and the level of classification uncertainty.

Acknowledgements

The authors would like to thank CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, Brazil), CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico, Brazil), and FAPEMIG (Fundação de Amparo à Pesquisa do Estado de Minas Gerais, Brazil) for supporting this research. The authors also thank an anonymous referee for constructive comments to improve this paper.

References

Barry D, Hartigan JA (1992). "Product Partition Models for Change Point Problems." *The Annals of Statistics*, **20**(1), 260–279. doi:2242159.

Barry D, Hartigan JA (1993). "A Bayesian Analysis for Change Point Problem." *Journal of the American Statistical Associations*, **88**(421), 309–319. doi:10.2307/2290726.

Beal MJ, Krishnamurthy P (2006). "Gene Expression Time Course Clustering with Countably Infinite Hidden Markov Models." In *arXiv preprint arXiv:*1206.6824.

Biernacki C, Celeux G, Govaert G (2000). "Assessing a Mixture Model for Clustering with the Integrated Completed Likelihoods." *IEEE Transctions on Pattern Analysis and Machine Intelligence*, **22**(7), 719–725. doi:10.1109/34.865189.

Broët P, Richardson S (2006). "Detection of Gene Copy Number Changes in CGH Microarrays Using a Spatially Correlated Mixture Model." *Bioinformatics*, **22**(8), 911–918. doi:10.1093/bioinformatics/btl035.

Carlin PB, Gelfand AE, Smith AFM (1992). "Hierarchical Bayesian Analysis of Changepoint Problems." *Journal of the Royal Statistical Society, Series C*, **41**(2), 389–405. doi:10.2307/2347570.

Chernoff H, Zacks S (1964). "Estimating the Current Mean of a Normal Distribution which is Subjected to Changes in Time." *The Annals of Mathematical Statistics*, **35**(3), 999–1018. URL https://www.jstor.org/stable/2238232.

Chib S (1998). "Estimation and Comparison of Multiple Change-Point Models." *Journal of Econometrics*, **86**, 221–241. doi:10.1016/S0304-4076(97)00115-2.

Chiquet J, Rigaill G, Sundqvist M (2022). *aricode: Efficient Computations of Standard Clustering Comparison Measures*. R package version 1.0.2. URL https://CRAN.R-project.org/package=aricode.

Constantine W, Hesterberg T (2021). *splus2R: Supplemental S-PLUS Functionality in R*. R package version 1.3.3. URL https://CRAN.R-project.org/package=splus2R.

Dufays A (2016). "Infinite-State Markov-Switching for Dynamic Volatility." *Journal of Financial Econometrics*, **14**, 418–460. doi:10.1093/jjfinec/nbv017.

Erdman C, Emerson JW (2007). "bcp: An R Package for Performing a Bayesian Analysis of Change Point Problems." *Journal of Statistical Software*, **23**(3), 1–13. doi:10.18637/jss.v023.i03.

Fearnhead P (2005). "Exact Bayesian Curve Fitting and Signal Segmentation." *Transactions on Signal Processing*, **53**(6), 2160–2166. doi:10.1109/TSP.2005.847844.

Fearnhead P, Liu Z (2007). "On-line Inference for Multiple Changepoint Problems." *Journal of the Royal Statistical Society, Series B*, **69**(4), 589–605. doi:10.1111/j.1467-9868.2007.00601.x.

Fearnhead P, Rigaill G (2019). "Changepoint Detection in the Presence of Outliers." *Journal of the American Statistical Association*, **114**(525), 169–183. doi:10.1080/01621459.2017.1385466.

Fernández C, Green PJ (2002). "Modelling Spatially Correlated Data via Mixtures: A Bayesian Approach." *Journal of the Royal Statistical Society, Series B*, **64**(4), 805–826. doi:10.1111/1467-9868.00362.

Fritsch A, Ickstadt K (2009). "Improved Criteria for Clustering Based on the Posterior Similarity Matrix." *Bayesian Analysis*, **4**(2), 367–392. doi:10.1214/09-BA414.

Fruhwirth-Schnatter S (2006). *Finite Mixture and Markov Switching Models*. Springer, New York.

Gales M, Yong S (2007). "The Application of Hidden Markov Models in Speech Recognition." *Foundations and Trends in Signal Processing*, **1**, 195–304. doi:10.1561/2000000004.

Gelman A, Carlin JB, Stern HS, Rubin DB (2003). *Baysian Data Analysis*. Chapman and Hall/CRC, London.

Geweke J (1992). "Evaluating the Accuracy of Sampling-Based Approaches to Calculating Posterior Moments." In *Bayesian Statistics 4: Proceedings of the Fourth Valencia International Meeting*, pp. 641–649. Clarendon Press, Oxford, UK.

Hartigan JA (1990). "Partition Models." *Communications in Statistics*, **19**, 2745–2756. doi:10.1080/03610929008830345.

Hartigan JA, Wong MA (1979). "Algorithm AS 136: A K-means Clustering Algorithm." *Applied Statistics*, **28**, 100–108. doi:10.2307/2346830.

Haynes K, Fearnhead P, Eckley IA (2017). "A Computationally Efficient Nonparametric Approach for Change Point Detection." *Statistics and Computing*, **27**(5), 1293–1305. doi:10.1007/s11222-016-9687-5.

Heidelberger P, Welch PD (1983). "Simulation Run Length Control in the Presence of an Initial Transient." *Operations Research*, **31**(6), 1109–1144. doi:10.1287/opre.31.6.1109.

Hubert L, Arabie P (1985). "Comparing Partitions." *Journal of Classification*, **2**, 193–218. doi:10.1007/BF01908075.

Inman HF, Bradley EL (1989). "The Overlapping Coefficient as a Measure of Agreement Between Probability Distributions and Point Estimation of the Overlap of Two Normal Densities." *Communications in Statistics - Theory and Methods*, **18**(10), 3851–3874. doi:10.1080/03610928908830127.

Jeffrey AR, Joshua MU (2020). *quantmod: Quantitative Financial Modelling Framework*. R Package Version 0.4.18. URL https://CRAN.R-project.org/package=quantmod.

Kaufman L, Rousseeuw PJ (1990). "Partitioning around Medoids (Program PAM)." In *Finding Groups in Data: An Introduction to Cluster Analysis*, pp. 68–125. John Wiley and Sons, Hoboken.

Ko S, Chong T, Ghosh P (2015). "Dirichlet Process Hidden Markov Multiple Change-Point Model." *Bayesian Analysis*, **10**, 275–296. doi:10.1214/14-BA910.

Koumar J, Cejka T (2023). "Unevenly Spaced Time Series from Network Traffic." In *2023 7th Network Traffic Measurement and Analysis Conference (TMA)*, pp. 1–4. IEEE. doi:10.23919/TMA58422.2023.10198988.

Li M, Liao CS, Peng HM, Hwang JK (2015). "MTIE and TDEV Analysis of Unevenly Spaced Time Series Data." *International Journal of Modelling and Simulation*, **26**(4), 303–308. doi:10.1080/02286203.2006.11442381.

Liverani S, Hastie DI, Azizi L, Papathomas M, Richardson S (2015). "PReMiuM: An R Package for Profile Regression Mixture Models Using Dirichlet Processes." *Journal of Statistical Software*, **64**(7), 1–30. doi:10.18637/jss.v064.i07.

Loschi RH, Cruz FRB (2005). "Extension to the Product Partition Model: Computing the Probability of a Change." *Computational Statistics and Data Analysis*, **48**(2), 255–268. doi:10.1016/j.csda.2004.03.003.

Malsiner-Walli G, Frühwirth-Schnatter S, Grün B (2017). "Identifying Mixtures of Mixtures Using Bayesian Estimation." *Journal of Computational and Graphical Statistics*, **26**(2), 285–295. doi:10.1080/10618600.2016.1200472.

Martinez AF, Mena RH (2014). "On a Nonparametric Change Point Detection Model in Markovian Regimes." *Bayesian Analysis*, **9**(4), 823–858. doi:10.1214/14-BA878.

Mayrink VD, Gonçalves FB (2017). "A Bayesian Hidden Markov Mixture Model to Detect Overexpressed Chromosome Regions." *Journal of the Royal Statistical Society, Series C*, **66**, 387–412. doi:10.1111/rssc.12178.

Mayrink VD, Gonçalves FB (2020). "Identifying Atypically Expressed Chromosome Regions Using RNA-Seq Data." *Statistical Methods and Applications*, **29**, 619–649. doi:10.1007/s10260-019-00496-4.

McDowell IC, Manandhar D, Vockley CM, Schmid AK, Reddy TE, Engelhardt BE (2018). "Clustering Gene Expression Time Series Data Using an Infinite Gaussian Process Mixture Model." *PLoS Computational Biology*, **14**(1), e1005896. doi:10.1371/journal.pcbi.1005896.

Meigenier S, Bonastre JF, Fredouille C, Merlin T (2000). "Evolutive HMM for Multi-Speaker Tracking System." In *Proceedings 2000 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pp. 1201–1204. doi:10.1109/ICASSP.2000.859181.

Melnykov V, Maitra R (2010). "Finite Mixture Models and Model-Based Clustering." *Statistical Surveys*, **4**, 80–116. doi:10.1214/09-SS053.

Miller JI (2019). "Testing Cointegrating Relationships Using Irregular and Non-contemporaneous Series with an Application to Paleoclimate Data." *Journal of Time Series Analysis*, **40**(6), 936–950. doi:10.1111/jtsa.12469.

Mira A, Petrone S (1996). "Bayesian Hierarchical Nonparametric Inference for Change Point Problems." *Bayesian Statistics*, **5**, 693–703. doi:10.1093/oso/9780198523567.003.0049.

Mohammadi A, Salehi-Rad MR (2012). "Bayesian Inference and Prediction in an M/G/1 with Optional Second Service." *Computational Statistics - Simulation and Computation*, **41**(3), 419–435. doi:10.1080/03610918.2011.588358.

Mohammadi A, Salehi-Rad MR, Wit EC (2013). "Using Mixture of Gamma Distributions for Bayesian Analysis in an M/G/1 with Optional Second Service." *Computational Statistics*, **28**(2), 683–700. doi:10.1007/s00180-012-0323-3.

Page GL, Quintana FA (2016). "Spatial Product Partition Models." *Bayesian Analysis*, **11**(1), 265–298. doi:10.1214/15-BA971.

Papastamoulis P, Iliopoulos G (2010). "An Artificial Allocations Based Solution to the Label Switching Problem in Bayesian Analysis of Mixtures of Distributions." *Journal of Computational and Graphical Statistics*, **19**(2), 313–331. doi:10.1198/jcgs.2010.09008.

Pedroso RC, Loschi RH, Quintana FA (2023). "Multipartition Model for Multiple Change Point Identification." *TEST*, **32**, 759–783. doi:10.1007/s11749-023-00851-4.

Peluso S, Chib S, Mira A (2019). "Semiparametric Multivariate and Multiple Change-Point Modeling." *Bayesian Analysis*, **14**(3), 727–751. doi:10.1214/18-BA1125.

Plummer M, Best N, Cowles K, Vines K (2006). "CODA: Convergence Diagnosis and Output Analysis for MCMC." *R News*, **6**(1), 7–11. URL https://journal.r-project.org/articles/RN-2006-002/.

Ratkovic MT, Eng KH (2010). "Finding Jumps in Otherwise Smooth Curves: Identifying Critical Events in Political Processes." *Political Analysis*, **18**(1), 57–77. doi:10.1093/pan/mpp032.

Rodriguez CE, Walker SG (2014). "Label Switching in Bayesian Mixture Models: Deterministic Relabelling Strategies." *Journal of Computational and Graphical Statistics*, **23**(1), 25–45. doi:10.1080/10618600.2012.735624.

Smith AFM (1965). "A Bayesian Approach to Inference about a Change-Point in a Sequence of Random Variables." *Biometrika*, **62**(2), 407–416. doi:10.1093/biomet/62.2.407.

Spiegelhalter DJ, Best NG, Carlin BP, der Linde AV (2002). "Bayesian Measures of Model Complexity and Fit." *Journal of the Royal Statistical Society, Series B*, **64**(4), 583–639. doi:10.1111/1467-9868.00353.

Stephens DA (1994). "Bayesian Retrospective Multiple-Changepoint Identification." *Journal of the Royal Statistical Society, Series C*, **43**(1), 159–178. doi:10.2307/2986119.

Stephens M (2000a). "Bayesian Analysis of Mixture Models with an Unknown Number of Components: An Alternative to Reversible Jump Methods." *Annals of Statistics*, **28**(1), 40–74. URL https://www.jstor.org/stable/2673981.

Stephens M (2000b). "Dealing with Label Switching in Mixture Models." *Journal of the Royal Statistical Society, Series B*, **62**(4), 795–809. doi:10.1111/1467-9868.00265.

Teixeira LV, Assuncao RM, Loschi RH (2019). "Bayesian Space-Time Partitioning by Sampling and Pruning Spanning Trees." *Journal of Machine Learning Research*, **20**(85), 1–35. URL http://jmlr.org/papers/v20/16-615.html.

Vincent T, Risser L, Ciuciu P (2010). "Spatially Adaptive Mixture Modeling for Analysis of fMRI Time Series." *IEEE Transactions on Medical Imaging*, **29**(4), 1059–1074. doi:10.1109/TMI.2010.2042064.

Vinh NX, Epps J, Bailey J (2009). "Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance Necessary?" In *ICML '09: Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 1073–1080. Association for Computing Machinery, New York, USA. doi:10.1145/1553374.1553511.

Watanabe S (2010). "Asymptotic Equivalence of Bayes Cross Validation and Widely Applicable Information Criterion in Singular Learning Theory." *Journal of Machine Learning Research*, **11**, 3571–3594. URL https://dl.acm.org/doi/10.5555/1756006.1953045.

Whittaker J, Frühwirth-Schnatter S (1994). "A Dynamic Changepoint Model for Detecting the Onset of Growth in Bacteriological Infections." *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **43**(4), 625–640. doi:10.2307/2986261.

Zhu X, Melnykov Y (2022). "On Finite Mixture Modeling of Change-Point Processes." *Journal of Classification*, **39**(1), 3–22. doi:10.1007/s00357-021-09385-6.

Affiliation:

Marta Cristina C. Bianchi,
Instituto de Matemática e Estatística,
Universidade Federal de Goiás,
Rua Jacarandá, Chácaras Califórnia, Goiânia, GO, Brazil, 74001-970.
E-mail: marta_bianchi@ufg.br
URL: http://lattes.cnpq.br/8404301614834264

Vinícius D. Mayrink,
Departamento de Estatística, ICEx,
Universidade Federal de Minas Gerais,
Av. Antônio Carlos, 6627, Belo Horizonte, MG, Brazil, 31270-901.
E-mail: vdm@est.ufmg.br
URL: https://www.est.ufmg.br/~vdinizm

http://www.ajs.or.at/
http://www.osg.or.at/
Submitted: 2024-02-16
Accepted: 2024-06-19