LIII Simpósio Brasileiro de Pesquisa Operacional
João Pessoa - PB, 3 a 5 de novembro de 2021

A Comparison of Product Partition Model Applications to Univariate Multiple Change-Point Analysis

Rosangela H. Loschi
Departamento de Estatística – UFMG
Av. Antônio Carlos, 6627, Pampulha, 31270-901
Belo Horizonte – MG
loschi@est.ufmg.br

Ricardo C. Pedroso
Departamento de Estatística – UFMG
Av. Antônio Carlos, 6627, Pampulha, 31270-901
Belo Horizonte – MG
ricardocunhap@gmail.com

RESUMO

O modelo partição produto (PPM) é um modelo amplamente usado para detecção de múltiplos pontos de mudança. Neste caso, o PPM tradicional considera uma única partição aleatória que indica os grupos de observações para os quais os valores dos parâmetros do modelo probabilístico são idênticos. Em modelos multiparamétricos, se parâmetros diferentes mudam em momentos diferentes, um modelo de partição única não identifica quais parâmetros sofreram essas mudanças. Para resolver este problema, modelos de múltiplas partições podem ser considerados. Comparamos o desempenho de um modelo de múltiplas partições com algumas abordagens tradicionais baseadas no PPM de partição única, para identificar mudanças na média e na variância de sequências univariadas de observações normais. O ajuste do modelo é feito através de um método amostrador de Gibbs parcialmente colapsado. Aplicamos o modelo a uma série de dados reais financeiros. Os resultados mostram como estruturas com múltiplas partições podem enriquecer a análise de problemas de ponto de mudança.

PALAVRAS CHAVE. Pontos de mudança. Modelo partição produto. Amostrador de Gibbs. Múltiplas partições.

ABSTRACT

The product partition model (PPM) is a widely used approach for multiple change-point detection. The traditional PPM considers a single random partition that indicates the clusters of observations for which the sampling model parameter values are identical. In multiparametric models, if different parameters change at different times, a single partiton model does not identify the

parameters that experienced those changes. To solve this problem, multipartition models may be considered.have been proposed. We compare the performances of the new multipartition model with some traditional single partition PPM-based approaches to identify changes in the mean and variance of univariate sequences of Normal observations. The model estimation is made through a partially collapsed Gibbs sampler method. We apply the models to a real dataset and the results show that multipartition structures may enriche the analysis of change-point problems.

KEYWORDS. Change-point. Product partition model. Gibbs sampler. Multipartition.

Paper topics: EST&MP, SIM

1. Introduction

The main goals when addressing multiple change-point problems are to estimate the number and the positions of the changes. Estimation of the parameters within the clusters may be also of interest. The product partition model (PPM) was first applied to detect change-points by [Barry e Hartigan, 1992]. They assume a single partition ρ to indicate the change-point positions. Later, [Barry e Hartigan, 1993] applied the PPM to detect multiple changes in the means of a sequence of Normal observations with unknown constant variance. One of the greatest contributions of their work is the proposed Gibbs sampler scheme to sample from the random partition ρ. Extending the work of [Barry e Hartigan, 1993], [Loschi e Cruz, 2005] proposed a model to indentify changes in the mean or variance of univariate sequences of Normal data. Other extensions of the PPM to detect changes in several structural parameters can be found in [Loschi e Cruz, 2005], [Loschi et al., 2010] and references therein.

Despite being a competitive model to the identification of multiple change-points, the model proposed by [Loschi e Cruz, 2005] fails to identify the parameter associated to the each change. In financial data, for example, some events may produce changes in a return volatility but not in the mean return. Under single partition PPM-based models, we only obtain the posterior distribution of the random partition, that indicates the positions when changes occurred. However, one or more parameters may experience changes at different times. Recently, [Pedroso et al., 2021] proposed a general PPM-based model to change-point detection in different parameters of multiparameric models. They consider a multipartition structure that allows to identify the changes of each parameter separatelly.

In this work, we compare the performances of the single partition models proposed by [Barry e Hartigan, 1993] and [Loschi e Cruz, 2005] to the Normal data specification of the multipartition model proposed by [Pedroso et al., 2021]. We briefly present the PPM (Section 2) and describe the models (Sections 3 and 4). We analyze a real dataset application where changes in the mean and variance are found at different times (Section 5).

2. PPM for multiple change-point problems

The PPM is a probabilistic model for cluster identification introduced by [Hartigan, 1990]. This model assumes that the set of observations $\boldsymbol{X} = (X_1, \ldots, X_n)$ is partitioned into a random number b of disjoint subsets of the observation indexes $\boldsymbol{I} = \{1, \ldots, n\}$, determined by the random partition $\rho = \{S_1, \ldots, S_b\}$, such that $\cup_{j=1}^{b} S_j = \boldsymbol{I}$. PPM assumes a prior product distribution for ρ given by

$$p\left(\rho = \{S_1, \ldots, S_b\}\right) \propto \prod_{j=1}^{b} c(S_j) \tag{1}$$

where $c(S_j)$ is a non-negative cohesion assigned to the subset S_j of $\boldsymbol{I}$, which represents the prior knowledge about the similarity level of the elements in S_j. PPM also assumes that, given ρ, the clusters of observations $\boldsymbol{X}_{S_1}, \ldots, \boldsymbol{X}_{S_b}$, where $\boldsymbol{X}_{S_j} = \{X_i, i \in S_j\},\ j = 1, \ldots, b$, are independent.

PPM was first applied to multiple change-point identification by [Barry e Hartigan, 1992]. They consider that $X_1, \ldots, X_n$ is an univariate sequence of observations conditionally independent, given the sequence of unkown parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)$, with independent conditional marginal densities $f(X_i|\theta_i)$, $i = 1, \ldots, n$. They assume the number of changes and their locations both unkown, and change-point process is modeled by supposing that $\boldsymbol{\theta}$ is partitioned into contiguous subsequences (clusters) of equal parameter values. In this case, there exists a partition $\rho = \{S_1, \ldots, S_b\}$ of the set of indexes $\boldsymbol{I}$ that may be also denoted by

$$\rho = \{i_0, i_1, \ldots, i_b\}, \quad 0 = i_0 < i_1 < \cdots < i_b = n,$$

such that, given ρ, there exist the common parameters $\theta_{S_1}, \ldots, \theta_{S_b}$ such that

$$\theta_i = \theta_{S_j} \quad \text{for} \quad i \in S_j = \{i_{j-1} + 1, i_{j-1} + 2, \ldots, i_j\}, \quad j = 1, 2, \ldots, b.$$

The points $i_1, \ldots, i_b$ are the end points of the clusters $S_1, \ldots, S_b$, respectively. The first point of each cluster is called a change-point. The cluster parameters are assumed independent with prior distributions $f_{S_j}(\theta_{S_j})$, $j = 1, \ldots, b$. Under these assumptions, the likelihood and posterior distribution for $(\boldsymbol{\theta}, \rho)$ are respectively given by

$$f(\boldsymbol{X} \mid \boldsymbol{\theta}, \rho) = \prod_{j=1}^{b} \prod_{i \in S_j} f(X_i \mid \theta_{S_j}).$$

and

$$f(\boldsymbol{\theta}, \rho \mid \boldsymbol{X}) \propto \prod_{j=1}^{b} \left(\prod_{i \in S_j} f(X_i \mid \theta_{S_j}) \right) f_{S_j}(\theta_{S_j})\, c(S_j).$$

3. Single partition models for multiple change-point identification in Normal data

In this section, we describe the PPM-based models for change-point analysis proposed by [Barry e Hartigan, 1993] and [Loschi e Cruz, 2005]. These models were developed to change-point identification in Normal data. Both models consider a single partition structure. We refer to these models as BH93 and LCIA05, respectively.

3.1. The BH93 model

The BH93 model is an application of the PPM for multiple change-point identification, as proposed by [Barry e Hartigan, 1992], to detect mean changes in Normal observations with unkown constant variance. Assume that observations $X_1, \ldots, X_n$ are independent given the sequence of parameters $\mu_1, \ldots, \mu_n, \sigma^2$, that is, $X_i \mid \mu_i, \sigma^2 \overset{ind}{\sim} N(\mu_i, \sigma^2)$, $i = 1, \ldots, n$. Assume the prior distribution for the cluster parameters $\mu_{S_j} \sim N(\mu_0, \sigma_0^2/n_j)$, $j = 1, \ldots, b$, where $n_j = \#S_j$, the cardinality of S_j. The prior cohesions follow the parametric approach suggested by [Yao, 1984], which considers the probability p that a change occurs at any instant in the sequence, that is,

$$c(S_j) = \begin{cases} (1-p)^{n_j-1}p & \text{if } \; j = 1, 2, \ldots, b-1, \\ (1-p)^{n_j-1} & \text{if } \; j = b. \end{cases} \tag{2}$$

To sample from the random partition ρ, a fixed dimension representation of ρ is considered. Let $\boldsymbol{U} = (U_1, \ldots, U_{n-1})$ be a vector of auxiliary variables such that for $i = 1, \ldots, n-1$,

$$U_i = \begin{cases} 1 & \text{if } \; \mu_i = \mu_{i+1}, \\ 0 & \text{if } \; \mu_i \neq \mu_{i+1}. \end{cases} \tag{3}$$

Thus, $\boldsymbol{U}$ is a vector of random variables assuming values in $\{0, 1\}$, indicating whether or not a change-point occurred at each time i, $i = 1, \ldots, n-1$. The implementation of the BH93 model is available in the `R` package `bcp` ([Erdman e Emerson, 2007]).

3.2. The LCIA05 model

[Loschi e Cruz, 2005] extend PPM to multiple change-point detection in both the mean and variance of univariate sequences of Normal observations. It assumes that given the sequences of unkown parameters $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ and $\boldsymbol{\sigma} = (\sigma_1^2, \ldots, \sigma_n^2)$, the observations $X_1, \ldots, X_n$ are independent with $X_i|\mu_i, \sigma_i^2 \overset{ind}{\sim} N(\mu_i, \sigma_i^2)$, $i = 1, \ldots, n$. Given a partition $\rho = \{S_1, \ldots, S_b\}$, there exists $(\mu_{S_j}, \sigma_{S_j}^2)$, $j = 1, \ldots, b$, such that $(\mu_i, \sigma_i^2) = (\mu_{S_j}, \sigma_{S_j}^2)$ for $i \in S_j$, $i = 1, \ldots, n$. The joint prior distribution of $(\mu_{S_j}, \sigma_{S_j}^2)$ is given by a Normal-Inverse-Gamma distribution

$$\mu_{S_j}|\sigma_{S_j}^2 \sim N(m, v\sigma_{S_j}^2) \quad \text{and} \quad \sigma_{S_j}^2 \sim IG(a/2, d/2), \tag{4}$$

that is a conjugate prior for the Normal model with mean and variance unknown. Similar to BH93 model, the prior cohesions follow the parametric model described in (2) and the fixed dimension representation of ρ described in (3).

4. Multipartition model for multiple change-point identification in Normal data

A general development of a multipartition change-point model to change-point identification in multiparametric probabilistic models was introduced by [Pedroso et al., 2021]. This work also presents a specification of the general model to detect changes in mean and variance of Normal data, which is described next. We refer to this model as BMCP.

Consider the sequence of random variables $\boldsymbol{X} = (X_1, \ldots, X_n)$ and the sequences of unknown parameters $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)$ and $\boldsymbol{\sigma} = (\sigma_1^2, \ldots, \sigma_n^2)$. Assume that $X_i \mid \boldsymbol{\mu}, \boldsymbol{\sigma} \overset{ind}{\sim} N(\mu_i, \sigma_i^2)$, $i = 1, \ldots, n$. In addition, change-points in $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are assumed to occur independently, at unknown and possibly different positions. Let ρ_1 and ρ_2 be the random partitions of I that induce contiguous clusters in $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$, respectively. Denote by $\mu_{j_1}^\star$ the common mean into the cluster S_{j_1}, $j_1 = 1, \ldots, b_1$ and $\sigma_{j_2}^{2\star}$ the common variance for observations into the cluster S_{j_2}, $j_2 = 1, \ldots, b_2$.

Conditionally on ρ_1 and ρ_2, assume that X_i, $i \in S_j^\star = S_{j_1} \cap S_{j_2}$, are independent and identically distributed with $X_i \mid \mu_{j_1}^\star, \sigma_{j_2}^{2\star} \overset{iid}{\sim} N(\mu_{j_1}^\star, \sigma_{j_2}^{2\star})$ and observations in different clusters are independent. Then, the likelihood function is given by

$$f(\boldsymbol{X} \mid \boldsymbol{\mu}, \boldsymbol{\sigma}, \rho_1, \rho_2) = \prod_{j_1=1}^{b_1} \prod_{j_2 | S_j^\star \neq \emptyset} \left(\frac{1}{2\pi\sigma_{j_2}^{2\star}} \right)^{n_j^\star/2} \exp\left\{ -\sum_{i \in S_j^\star} \frac{(X_i - \mu_{j_1}^\star)^2}{2\sigma_{j_2}^{2\star}} \right\}, \tag{5}$$

where $\{j_k \mid S_j^\star \neq \emptyset\}$ denotes the set of values j_k for which $S_j^\star \neq \emptyset$, for $k = 1, 2$. Given ρ_1 and ρ_2, assume that $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are independent and the structural parameters in different clusters are also independent, with prior distributions

$$\begin{aligned} \mu_{j_1}^\star &\overset{iid}{\sim} N(\mu_0, \sigma_0^2), \quad j_1 = 1, ..., b_1, \\ \sigma_{j_2}^{2\star} &\overset{iid}{\sim} IG(a/2, d/2), \quad j_2 = 1, ..., b_2. \end{aligned} \tag{6}$$

For each random partition ρ_1 and ρ_2, assume the independent product partition distributions given in (1). The cohesion proposed by [Yao, 1984] is considered to quantify how strongly we believe the components of $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are to co-cluster *a priori*. That is, for $k = 1, 2$, assume the cohesions

$$c_k(S_{j_k}) = \begin{cases} (1 - p_k)^{n_{j_k} - 1} p_k & \text{if } \ j = 1, 2, \ldots, b_k - 1, \\ (1 - p_k)^{n_{j_k} - 1} & \text{if } \ j = b_k. \end{cases} \tag{7}$$

To complete the model specification, assume *a priori* that $p_k \overset{ind}{\sim} Beta(\alpha_k, \beta_k), \qquad k = 1, 2.$

5. Real data application

The Mexican Peso/US Dollar exchange rate we analyse in this section is available at www.federalreserve.gov, and is presented in Figure 1. This data set is composed by daily records of Mexican Peso/US Dollar exchange rate from January 2007 to December 2012, a total of $n = 1,510$ observations. The presence of regime changes in this time series was analysed by Martínez e Mena [2014] fitting a nonparametric change point detection model. *A priori*, we consider $(\mu_0, \sigma_0^2, a, d) = (0, 10^6, 0.002, 0.002)$ for the BMCP model, $(m, v, a, d) = (0, 10, 0.002, 0.002)$ for the LCIA05 model and $(p_0, w_0) = (0.05, 10^{-6})$ for the BH93 model.

The BMCP indicates a greater number of changes in the mean than the LCIA05 model. The estimated posterior distribution of the ρ_1 and ρ_2 are too flat, not providing strong evidence about the respective change point locations. In this cases, the probability of a change (Figure 3) may be a more useful result. For example, we could determine the true partition as the one composed by the instants with probability of a change greater than some probability threshold. The variance estimates under the LCIA05 are overestimated if comapared with the sample moving variance (Figure 4(e)) and are clearly affected by the mean changes. The Under the BMCP model, variance estimates are closer to the sample moving variance and this model estimates possible change-points associated to higher and lower variance clusters.

The BH93 model indicates the higher number of mean changes and these changes are concentrated in the higher variance clusters indicated by the BMCP model. The variance estimate for the BH93 model is closer to the average sample moving variance but it is limited to a constant

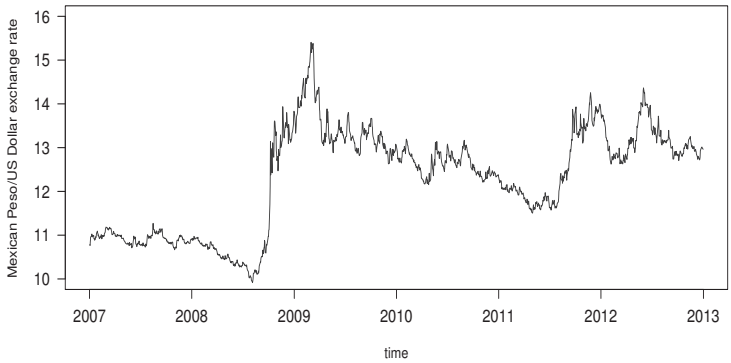
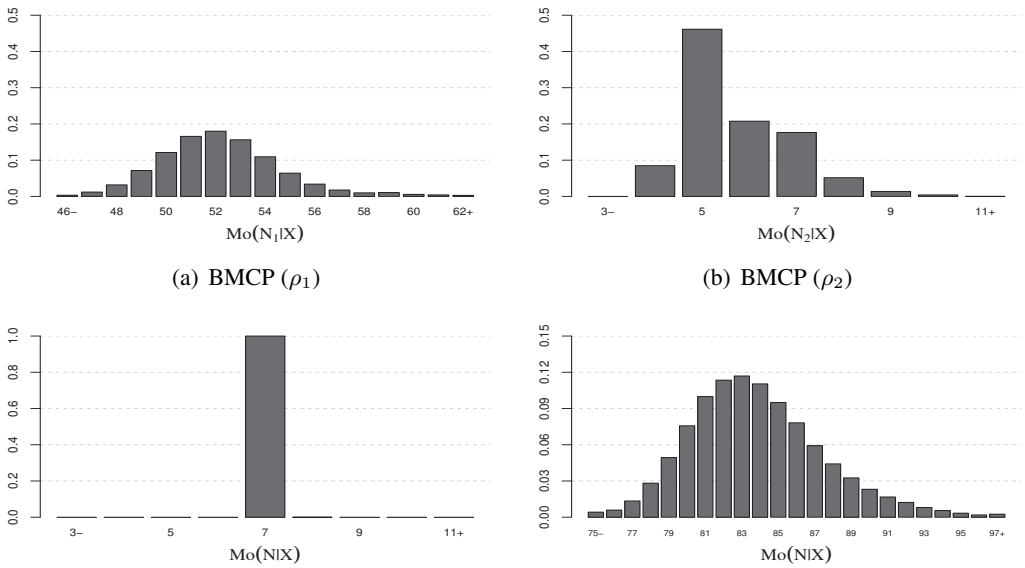


Figura 1: Daily records of Mexican Peso/US Dollar exchange rate from January 2007 to December 2012.



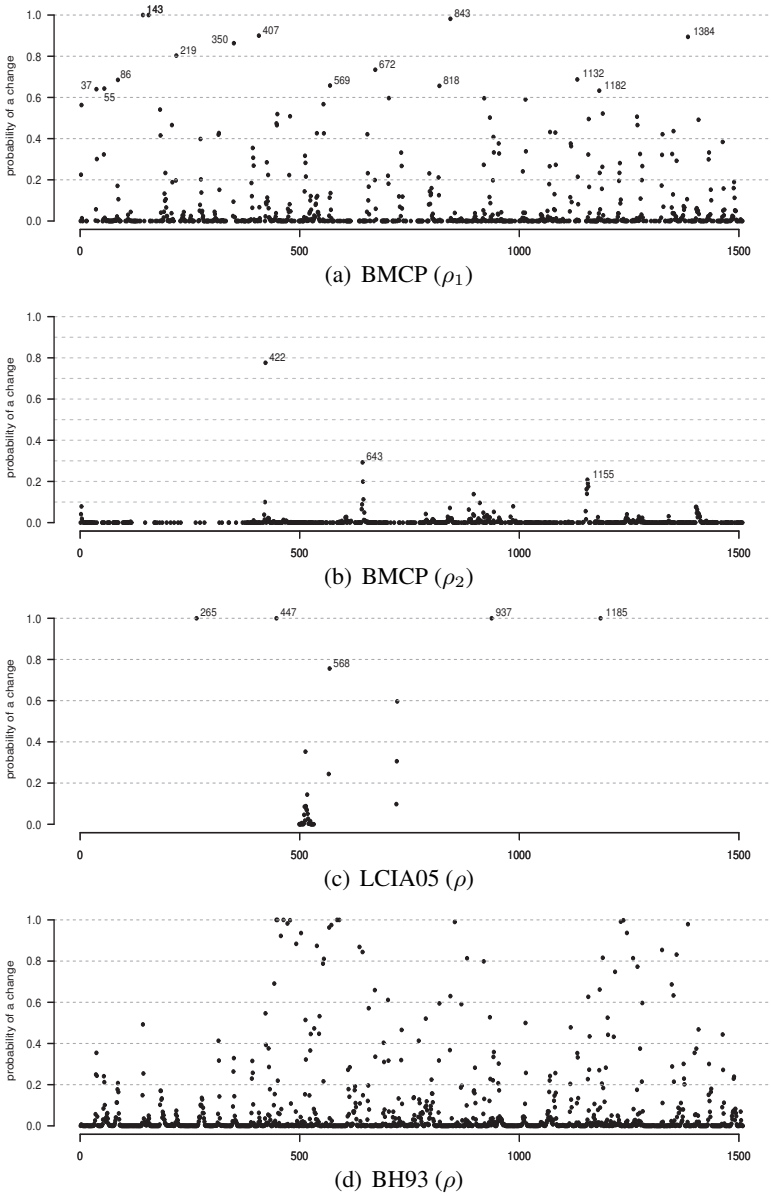
(a) BMCP (ρ_1)

(b) BMCP (ρ_2)

(c) LCIA05 (ρ)

(d) BH93 (ρ)

Figura 2: Posterior distributions of the number of changes for the Mexican Peso/US Dollar dataset.



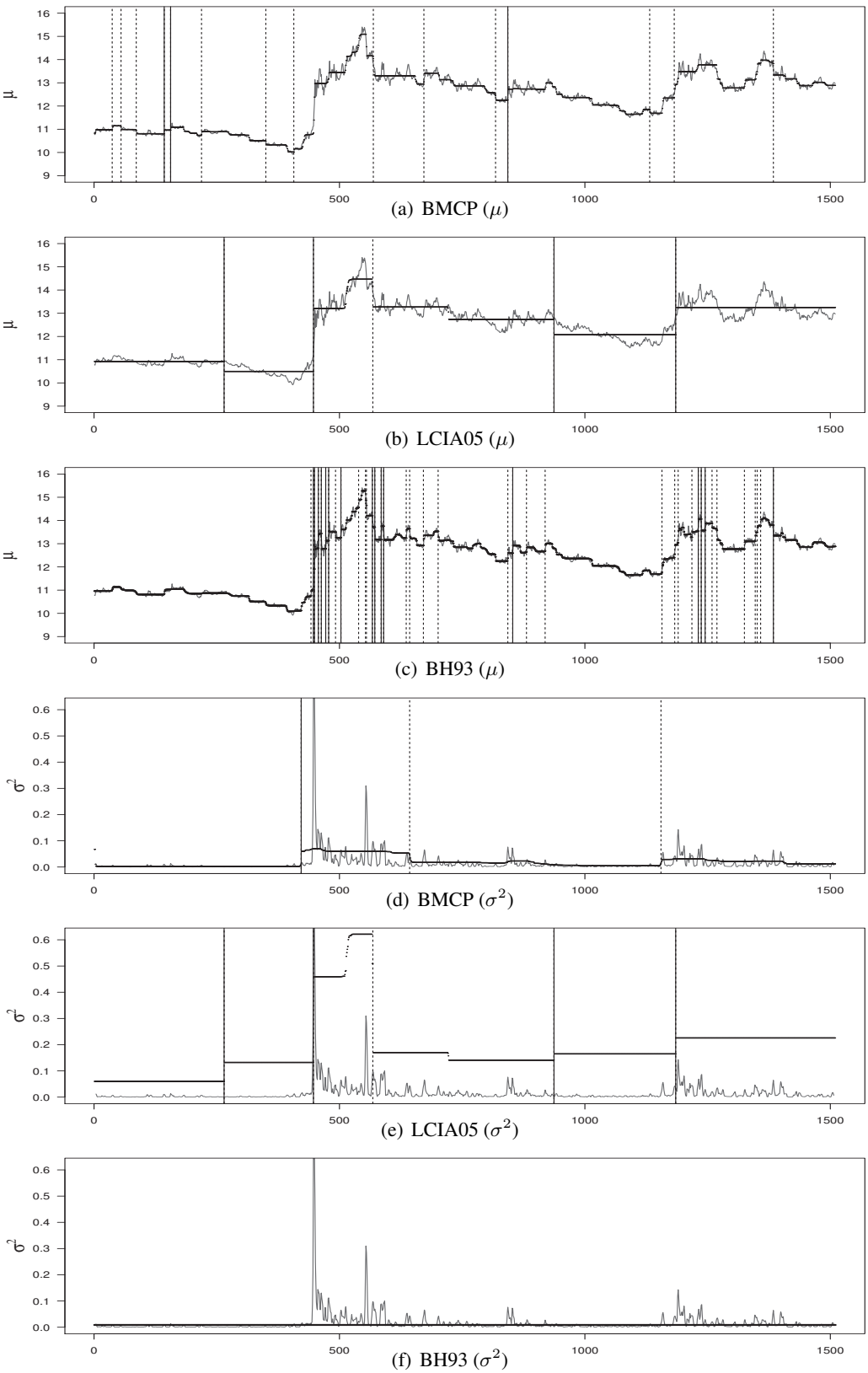
(a) BMCP (ρ_1)

(b) BMCP (ρ_2)

(c) LCIA05 (ρ)

(d) BH93 (ρ)

Figura 3: Posterior probability of each position to be a change for the Mexican Peso/US Dollar dataset. The labeled positions are those with probability greater than 0.6 in (a,c) and greater than 0.2 in (b).



(a) BMCP (μ)

(b) LCIA05 (μ)

(c) BH93 (μ)

(d) BMCP (σ^2)

(e) LCIA05 (σ^2)

(f) BH93 (σ^2)

Figura 4: Product estimates (black dots) for $\boldsymbol{\mu}$ (a-c) and $\boldsymbol{\sigma}$ (d-f) for the Mexican Peso/US Dollar dataset. The gray lines represent the observed data in (a-c) and the moving sample variance, calculated over ranges of length seven, in (d-f). The vertical lines in represent the positions with probability of being a change greater than 0.6 (dashed lines) and greater than 0.9 (solid lines).

variance assumption that is not appropriate to this dataset analysis, according to the results provided by the multipartition model.

6. Conclusions

In this study, we compare single and multipartition PPM-based models for Normal data. We found some evidence that multipartition strucutres may provide improved estimates in multiple change-point problems, allowing us to identify the positions and the number of changes that occurred, as well as which parameters have changed along the data sequence. Analyzing the real dataset, the multipartition model indicates changes in the mean and variance at different times. The multipartition model also provided better variance estimates than the single partition models. These evidences indicate that multiparition models may be more efficient in multiple change-point analysis.

Acknowledgements

This work was funded by Coordenação de Aperfeiçoamento de Pessoal de Ensino Superior (CAPES), Conselho Nacional de Desenvolvimento Científico e Tecnologico (CNPq) and Fundação de Amparo à Pesquisa do Estado de Minas Gerais (FAPEMIG).

Referências

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

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

Erdman, C. e Emerson, J. W. (2007). bcp: an r package for performing a Bayesian analysis of change point problems. *Journal of Statistical Software*, 23(3):1–13.

Hartigan, J. A. (1990). Partition models. *Communications in statistics-Theory and methods*, 19(8): 2745–2756.

Loschi, R. H., Pontel, J. G., e Cruz, F. R. (2010). Multiple change-point analysis for linear regression models. *Chilean Journal of Statistics*, 1(2):93–112.

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

Martínez, A. F. e Mena, R. H. (2014). On a nonparametric change point detection model in Markovian regimes. *Bayesian Analysis*, 9(4):823–858.

Pedroso, R. C., Loschi, R. H., e Quintana, F. A. (2021). Multipartition model for multiple change point identification.

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