Title

Bayesian Model Selection Approach to Multiple Change-Points Detection with Non-Local Prior Distributions

Journal

ACM Transactions on Knowledge Discovery from Data, 13(5)

ISSN

1556-4681

Authors

Jiang, Fei
Yin, Guosheng
Dominici, Francesca

Publication Date

2019-10-31

DOI

10.1145/3340804

Peer reviewed

Bayesian Model Selection Approach to Boundary Detection with Non-Local Priors

Fei Jiang, Guosheng Yin

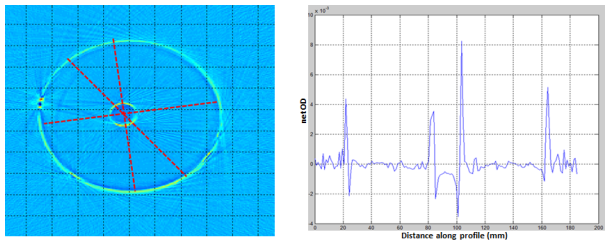
Abstract—We propose a Bayesian model selection (BMS) boundary detection procedure using non-local prior distributions for a sequence of data with multiple systematic mean changes. By using the non-local priors in the Bayesian model selection framework, the BMS method can effectively suppress the non-boundary spike points with large instantaneous changes. Further, we speed up the algorithm by reducing the multiple change points to a series of single change point detection problems. We establish the consistency of the estimated number and locations of the change points under various prior distributions. From both theoretical and numerical perspectives, we show that the non-local inverse moment prior leads to the fastest convergence rate in identifying the true change points on the boundaries. Extensive simulation studies are conducted to compare the BMS with existing methods, and our method is illustrated with application to the magnetic resonance imaging guided radiation therapy data.

Index Terms—Bayes factor; Bayesian model selection; Inverse moment prior; Local prior; Marginal likelihood; Moment prior; Parallel computing.

1 Introduction

Radiation therapy is a popular cancer treatment method to deliver a highly conformal beam to the target volume. The recent efforts to visualize the patient's anatomy at the time of radiation treatment led to the development of the magnetic resonance imaging guided radiation therapy (MRgRT). However, when radiation is delivered in the magnetic field, the dose level can be significantly enhanced near the boundaries between different tissues or organs in human bodies. Mimicking the environment in the human body, the Duke Mid-sized Optical-CT System (DMOS) was utilized to identify the dose changes near the region of the boundary artifact. Figure 1 is a single slice of the simplified dosimeter model, which contains only one cavity in the middle.

Fig. 1: Reconstructed image of a slice in a cylindrical dosimeter with a cavity in the middle (left) and a typical line profile through the center of the cavity (right). The radiation enter the dosimeter from the hole on the left of the cylindrical dosimeter. The cylindrical rotates 360 degree so that the radiation can enter from different direction.



- *Fei Jiang was with the Department of Department of Statistics and Actuarial Science, University of Hong Kong. Email: feijiang@hku.hk*
- *Guosheng Yin was with the Department of Department of Statistics and Actuarial Science, University of Hong Kong. Email: gyin@hku.hk*

Manuscript received 00; revised 00, 2018.

In practice, the inner structure of the dosimeter can be complicated. The right panel in Figure 1 contains one signal of the dose change, where we can observe that the boundaries on and inside the dosimeter can be distinguished by noticeable peaks of the radiation dose levels. In the experiment, multiple radiations enter the cylindrical dosimeter from different directions, and a sequence of data ordered by their distances to the sources was collected. Because the dosimeter is a circle and the cavity is in the middle, the radiation from different directions would hit the boundary at similar distances away from their sources in the dosimeter. Hence the entire data sequence, merged from the multiple sequences, often shows systematic mean changes at the boundaries (see the background data points as shown in Figure 2). This pattern motivates the boundary detection by using the change points detection algorithms.

However, in the MRgRT data, radiations in certain directions may experience temporary changes at non-boundary locations, which may result from the abnormal status of the DMOS system rather than the true dose changes. The temporary change points, appear in the data sequence as the spike points, often mixed up with the ones on the boundary (the peak locations in Figure 1) which makes the boundary detection extremely challenging. Figures 2 shows the change points in the MRgRT data identified by the popular NOT [14] and the SML [3] algorithms, respectively. It can be seen that neither of the algorithms correctly identify the true boundaries. The detection procedures are misled by spike points, which does not fall on the true boundaries. In general, NOT is suitable under the setting where the different segments have comparable length, while the SML works the best to identify the frequent and irregular change points. However, in the MRgRT data, the superfluous change points that do not indicate systematic changes could deviate the purpose of identifying the true boundaries. This motivate us to propose an new algorithm in detecting the systematic

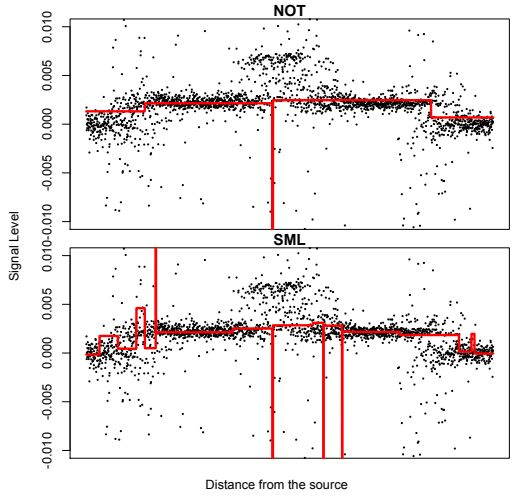


Fig. 2: The change point detection results from the SML with normal prior and the maximal number of change points to be 30 and NOT methods for the dose level changes in the DMOS system.

changes when the segment length can have dramatic differences.

Extensive research has been conducted on change point detection problems. Existing approaches often rely on optimizing certain objective functions to detect the locations of the change points, such as by maximizing the likelihood [1, 2] or the marginal likelihood [3]. The total number of change points can be estimated by minimizing the Bayesian information criterion (BIC); for example, see [4, 5, 6, 7]. In addition, penalty methods have also been developed to detect the locations and estimate the number of change points [8, 9, 10]. These approaches obtain the optimal solutions by taking into account the entire sample and, as a result, the computational burden can be immense for a large data set. On the other hand, several methods based on the local data are advocated to alleviate the computational burden, such as the screening [11], and the likelihood ratio scan statistics [12]. In addition, [13] proposed the wild binary segmentation (WBS) and [14] developed the narrowest-over-threshold (NOT) methods, both of which achieve consistency and enjoy computational efficiency under their specific model assumptions. [3] introduced a Bayesian stepwise marginal likelihood (SML) change point detector, which reduces the the computational load by dynamic programming.

As shown in the preliminary analysis of the MRgRT data, the local control of the discovery is crucial. To avoid picking the spike points, we reinforce minimal distances between the change points. Moreover, we adopt the computational efficient local scan routine and propose a systematic two-stage procedure to speed up the change point detection procedure. More specifically, the local scan method first identifies the candidate points with minimal distance based on the local data, and then optimizes an utility function to obtain the estimates for the locations and the total number of change points. Because the change points are defined based on the mean changes in two consecutive segments, the local data are sufficient to detect the systemic changes [15, 1, 16, 17, 13, 18, 19, 12].

To define the utility function for the change point detection, we first investigate the criteria for identifying the change points in the typical frequentist NOT and Bayesian SML methods. For simplicity, we focus on the problem with one change point under the normal assumption with a constant variance. Assume the data sequence is $X_1, \ldots, X_n$, and the $(n_1 + 1)$ observation is a change point, $n_1/n < 1$. Let $\widehat{\theta}_1, \widehat{\theta}_2$, and $\widehat{\theta}$ be the maximum likelihood estimators based on the first n_1, last $n - n_1$ and entire n samples, respectively. The NOT method distinguishes the change and non-change points through the logarithm of the likelihood ratio statistics, which is proportional to

$$L_n = -\sum_{i=1}^{n_1}(X_i - \widehat{\theta}_1)^2 - \sum_{j=n_1+1}^{n}(X_i - \widehat{\theta}_2)^2 + \sum_{i=1}^{n}(X_i - \widehat{\theta})^2.$$

On the other hand, the SML method uses the Bayesian marginal likelihood ratio and yields the detection statistics proportional to $L_n - O_p(\log n)$. Clearly, both the NOT and SML statistics go to infinity at the rate of $O_p(n)$ in detecting the true change points. However, for the non-change points, the SML statistics converges to $-\infty$ at the rate of $O_P\{\log(n)\}$ and the NOT statistics approaches to a constant in probability [20].

This asymmetric phenomenon has been discussed in the frequentist hypothesis testing literature. However, because a frequentist test does not make a conclusion on accepting the null hypothesis, the asymmetric convergence feature is less problematic [21]. On the contrary, the Bayesian hypothesis testing procedures rely on the probabilities of the null and alternative hypotheses, and thus it is critical to control the rate of the convergence under the null. To this end, a class of priors, namely local alternative priors, have been studied; for example, see [22], [23], [24], [25], [26], [27], [28], [28], [29]. However, the local priors, as shown in the SML method, suffer from the shortcoming that the evidence in favor of true null models is accumulating at a much slower rate compared with that in favor of true alternative models. To balance the convergence of the Bayes factor, [21] introduced the non-local priors, which largely improve the speed of the accumulation of the evidence in favor of true null models.

Hence, viewing the advantages of the Bayesian method under the null model, we utilize a Bayesian marginal likelihood function as the utility, and propose a new Bayesian model selection (BMS) procedure for identifying change points. In particular, we show that the selection consistency is achieved by choosing both the local and non-local priors, whereas the convergence rate is faster under the non-local priors, which justifies the previous conclusion from the preliminary analysis.

Our BMS procedure is cast in the model selection framework, which is faster than the dynamic programming introduced under the SML framework. For example, for a single sequence in the MRgRT data, BMS takes roughly 1.3 seconds and SML takes 3.1 seconds when the maximum

number of change points is capped at 100. The major factor that contributes to the efficiency of BMS algorithm is the fact that BMS reduces the search space dramatically by selecting a small set of candidate change-points. Further, once the candidate points are selected, BMS only needs to evaluate two consecutive segments at a time. In contrast, the dynamic programming not only assesses two consecutive segments, but also computes the utility function for the larger segment formed by the merging of these two smaller segments. When multiple potential change-points are present, the different ways of merging adjacent segments can grow significantly.

The rest of this paper is organized as follows. In Section 2, we describe the model and theoretical properties of the BMS method in line with the computational algorithms. In Section 3, we evaluate the BMS method and discuss the selection of tuning parameters. In Section 4, the BMS is applied to analyze the MRgRT data for illustration. Section 5 concludes with some remarks.

2 Bayesian multiple change points detection

2.1 Probability model

Suppose there are p_0 true change points $t_1 <, \ldots, < t_{p_0}$ among n observations $\mathbf{Y}_n = \{Y_1, \ldots, Y_n\}$. As a convention, let $t_0 = 1$ and $t_{(p_0+1)} = n + 1$. Denote $\lambda_j = t_{j+1} - t_j$ and $\lambda = \min_{j=0,\ldots,p_0} \lambda_j$. We consider a set of K_n candidate points $\tau_1, \ldots, \tau_{K_n}$, with $\tau_0 = 1$ and $\tau_{K_n+1} = n + 1$, while postponing the discussion on the candidate set selection to Section 2.4. Define $n_j = \tau_{j+1} - \tau_j$, $n_I = \min_{j=0,\ldots,K_n-1} n_j$, $n_I \leq \lambda$. Let $\mathcal{H}(n_I) = \{\tau_j : j = 1, \ldots, K_n, |\tau_{j+1} - \tau_j| > n_I\}$ denote the set of all candidate points and $\mathcal{T}_0(p_0) = \{t_j : j = 1, \ldots, p_0\}$ be the set of true change points. The specification of the candidate points allows the BMS method to be implemented in a lower dimensional space with the most influential points. It also guarantees that there are a sufficient number of non-change points surrounding the true ones so that the consistency conditions are met. The probability model takes the form of

$$Y_l = \nu_{\tau_j} + \epsilon_l, \;\; l \in [\tau_j, \tau_{j+1}),$$

where the random errors ϵ_l are independent with mean zero and variance σ_j^2. Further, we define $\sigma = \max_{j=0,\ldots,p_0} \sigma_j$.

For ease of exposition, we first consider the situation where the locations of the candidate change points are given and $\mathcal{T}_0(p_0) \subset \mathcal{H}(n_I)$. Define $\bar{Y}_{\tau_j} = n_{j-1}^{-1} \sum_{l=\tau_{j-1}}^{\tau_j - 1} Y_l$, which is the sample average for the $(j-1)$th segment $[\tau_{j-1}, \tau_j)$, $j > 1$. Suppose the candidate point τ_k is not a change point, then the points in $[\tau_k, \tau_{k+1})$ have the same mean as those in $[\tau_{k-1}, \tau_k)$. If τ_k is a change point, we expect a mean shift between the segments $[\tau_k, \tau_{k+1})$ and $[\tau_{k-1}, \tau_k)$. Hence, we can formulate the model and prior distribution for $l \geq \tau_1$ as follows:

$$\begin{aligned} Y_l - \bar{Y}_{\tau_k} &= \mu_k + \xi_l, \quad l \in [\tau_k, \tau_{k+1}), \\ \mu_k &\sim \pi(\mu_k), \text{ if } \tau_k \text{ is a change point,} \\ \mu_k &= 0, \text{ with probability 1, if } \tau_k \text{ is an } n_I\text{-flat point,} \end{aligned}$$

where $\pi(\cdot)$ is a prior density and ξ_l is a mean-zero error. Note that the observations in the first segment are unchanged. Here the n_I-flat point is defined as a non-change point which is at least n_I apart from any change points. We require the n_I distance between the true change points and the flat ones so that there are sufficient neighborhood samples to achieve the estimation consistency.

Let μ_{k0} be the true value of μ_k, and we assume $|\mu_{k0}| > \delta$, where $\delta > 0$ is the lower bound of the μ_{k0}, for the k's with $\tau_k \in \mathcal{T}_0(p_0)$. The prior distribution on μ_k is crucial for determining the convergence rate of the BML procedure. We explore three types of priors: the local prior [30], the non-local moment prior and the inverse moment prior in [21].

$$\begin{aligned} \text{Local:} \quad \pi_L(\mu) &= N(0, \omega^2) \\ \text{Moment:} \quad \pi_M(\mu) &= \mu^{2v}/C_M 1/\sqrt{2\pi} \exp(-\mu^2/2) \\ \text{Inverse moment:} \quad \pi_I(\mu) &= s\nu^{q/2}/\Gamma(q/2s)\mu^{-(q+1)} \\ &\quad \times \exp\left\{-(\mu^2/\nu)^{-s}\right\}, \end{aligned}$$

where C_M is the normalizing constant.

Let M_k represent the model that τ_k is the sole change point. We define the marginal likelihood with the Gaussian kernel as

$$\begin{aligned} \Pr(\mathbf{Y}_n | M_k) &= \prod_{j=1, j\neq k}^{K_n} \prod_{l=\tau_j}^{\tau_{j+1}-1} \exp\{-(Y_l - \bar{Y}_{\tau_j})^2\} \\ &\quad \times \int \prod_{l=\tau_k}^{\tau_{k+1}-1} \exp\{-(Y_l - \bar{Y}_{\tau_k} - \mu)^2\} \pi(\mu) d\mu. \end{aligned}$$

The posterior model probability of M_k given $\mathbf{Y}_n$ is

$$\begin{aligned} \Pr(M_k | \mathbf{Y}_n) &= \frac{\Pr(\mathbf{Y}_n | M_k) \Pr(M_k)}{\sum_{j=1}^{K_n} \Pr(\mathbf{Y}_n | M_j) \Pr(M_j)} \\ &= \frac{\Pr(\mathbf{Y}_n | M_k)}{\sum_{j=1}^{K_n} \Pr(\mathbf{Y}_n | M_j)}, \end{aligned} \tag{1}$$

when M_j assumes a non-informative uniform prior, $j = 1, \ldots, K_n$. Note that it is not necessary for $\mathbf{Y}_n$ to be normally distributed to ensure the selection consistency in detecting mean changes. The Gaussian kernel serves as a utility function, which tends to be large when the distances between the true and the hypothetical segment means are small. Hence, as $n \to \infty$, $\Pr(M_k | \mathbf{Y}_n)$ approaches 1 when τ_k is indeed a change point and the τ_j's $(j \neq k)$ are n_I-flat points.

2.2 Change point detection

We start with the simplest case that there is only one mean shift in the data, i.e., $p_0 = 1$ is fixed *a priori*. We select the candidate point τ_k associated with the largest $\Pr(M_k | \mathbf{Y}_n)$ among all the candidates, which corresponds to the largest marginal likelihood $\Pr(\mathbf{Y}_n | M_k)$. It can be shown that

$$\Pr(M_k | \mathbf{Y}_n) = \left\{ 1 + \sum_{j \neq k}^{K_n} \frac{\Pr(\mathbf{Y}_n | M_j)}{\Pr(\mathbf{Y}_n | M_k)} \right\}^{-1},$$

where

$$\begin{aligned}&\frac{\Pr(\mathbf{Y}_n|M_j)}{\Pr(\mathbf{Y}_n|M_k)}\\ =\;&\frac{\int\prod_{l=\tau_j}^{\tau_{j+1}-1}\exp\{-(Y_l-\bar{Y}_{\tau_j}-\mu)^2\}\pi(\mu)d\mu}{\prod_{l=\tau_j}^{\tau_{j+1}-1}\exp\{-(Y_l-\bar{Y}_{\tau_j})^2\}}\\ &\times\left[\frac{\int\prod_{l=\tau_k}^{\tau_{k+1}-1}\exp\{-(Y_l-\bar{Y}_{\tau_k}-\mu)^2\}\pi(\mu)d\mu}{\prod_{l=\tau_k}^{\tau_{k+1}-1}\exp\{-(Y_l-\bar{Y}_{\tau_k})^2\}}\right]^{-1},\end{aligned}$$

is the ratio of the Bayes factors for an n_I-flat point and a change point. The ratio indicates that the selection consistency is fully determined by the evidence in favor of $\mu_k \sim \pi(\mu_k)$ and that of $\mu_j = 0$ for $j \neq k$. The product on the right hand side converges to 0 when n_I grows to ∞ with the sample size. The candidate points retain enough samples in the neighborhood to guarantee the convergency.

For the case with multiple change points ($p_0 > 1$), we select the points associated with the p_0 largest $\Pr(M_k|\mathbf{Y}_n)$, and the selection consistency is presented as follows.

Theorem 1. *Let $\mathcal{M} = \{M_k, \tau_k \in \mathcal{T}_0(p_0)\}$. If the Bayes factor satisfies*

$$\frac{\int\prod_{l=\tau_j}^{\tau_{j+1}-1}\exp\{-(Y_l-\bar{Y}_{\tau_j}-\mu)^2\}\pi(\mu)d\mu}{\prod_{l=\tau_j}^{\tau_{j+1}-1}\exp\{-(Y_l-\bar{Y}_{\tau_j})^2\}} = O_p(a_{n_j}) \quad (2)$$

for $\tau_j \notin \mathcal{T}_0(p_0)$, $a_{n_j} = o_p(1)$, and $n_I^{1/2}\delta/\sigma \to \infty$, then

$$\sum_{M_k\in\mathcal{M}}\Pr(M_k|\mathbf{Y}_n) = 1 + O_p\{K_n a_{n_I}\exp(-n_I\delta^2)\}. \quad (3)$$

Hence when $n_I/\log(n) \to c$, $0 < c \leq \infty$, $n_I \leq \lambda$, we have

$$\sum_{M_k\in\mathcal{M}}\Pr(M_k|\mathbf{Y}_n) \xrightarrow{P} 1.$$

The proof of Theorem 1 is delineated in the Appendix. Clearly, the selection consistency depends on the convergence rate of a_{n_I}, which is determined by the choice of the prior $\pi(\cdot)$. Lemmas 2–4 in the Appendix show that $a_{n_j} = n_j^{-1/2}$ when $\pi(\cdot) = \pi_L(\mu)$; $a_{n_j} = n_j^{-v-1/2}$ if $\pi(\cdot) = \pi_M(\mu)$ and $a_{n_j} = \exp\{-n_j^{s/(s+1)}\}$ if $\pi(\cdot) = \pi_I(\mu)$. Hence, the selection consistency is achieved at the fastest rate using the non-local inverse moment prior. Lemmas 2–4 follow by using the similar arguments as those in [21], and we additionally show the above convergence can be achieved without using the exact likelihood.

The convergence in (3) contains the effects from two aspects: The order a_{n_I} represents the speed of vanishing for the Bayes factors at the n_I-flat points and the order $\exp(-n_I\delta^2)$ is the inverse of the growth rate for the Bayes factor at the true change points. It worth mentioning that the NOT method does not benefit from the information provided by the n_I-flat points locally, while the SML method only takes $n_I^{-1/2}$ order of improvement on the convergence. Hence, although BMS, NOT, and SML are asymptotically equivalent, the finite sample performance can be quite different, as shown in our simulation studies.

2.3 Number of change points

When p_0 is unknown, let $\mathcal{T}(p)$ be the set containing p points from the procedure described in Section 2.2. We define the marginal likelihood given $\mathcal{T}(p)$ as

$$\begin{aligned}&\Pr\{\mathbf{Y}_n|\mathcal{T}(p)\}\\ =\;&\prod_{\tau_j\notin\mathcal{T}(p)}\prod_{l=\tau_j}^{\tau_{j+1}-1}\exp\{-(Y_l-\bar{Y}_{\tau_j})^2\}\\ &\times\prod_{\tau_k\in\mathcal{T}(p)}\int\prod_{l=\tau_k}^{\tau_{k+1}-1}\exp\{-(Y_l-\bar{Y}_{\tau_k}-\mu)^2\}\\ &\times\pi(\mu)d\mu. \quad (4)\end{aligned}$$

We can estimate the locations and the number of change points in two steps: (i) for any given p, we obtain $\widehat{\mathcal{T}}(p)$ using the procedure described in the previous section, and (ii) we estimate p_0 by $\widehat{p}$ via maximizing $\Pr\{\mathbf{Y}_n|\widehat{\mathcal{T}}(p)\}$ with respect to p. In contrast to the procedure in [3] which simultaneously estimates the locations and the number of change points by maximizing the marginal likelihood with respect to $\mathcal{T}(p)$ and p, our BMS splits the estimation procedure into a scanning step as described in Section 2.2 and an optimization step. This scanning–optimization mechanism reduces the computational burden substantially, because the optimization in step (ii) is merely implemented in a single dimension.

2.4 Candidate points selection

Previous discussions rely upon a critical assumption that the candidate points are specified in advance, which, however, is often infeasible in practice. To facilitate the implementation of the BMS, we need to find a candidate set $\mathcal{H}_c(n_I)$ that is close to $\mathcal{H}(n_I)$. For the selection consistency of the change points, we require for each t_j there is a $\tau_k \in \mathcal{H}_c(n_I)$, such that $\Pr(|t_j - \tau_k| \leq n_I) = 1 - O_p[\min\{\exp(-n_I\delta^2), a_{n_I}\}]$. Define

$$R_i = \frac{\int\prod_{l=i}^{i+n_I-1}\exp\{-(Y_l-\bar{Y}_i-\mu)^2\}\pi(\mu)d\mu}{\prod_{l=i}^{i+n_I-1}\exp\{-(Y_l-\bar{Y}_i)^2\}},$$

where $\bar{Y}_i = (n_I-1)^{-1}\sum_{j=i-n_I}^{i-1}Y_j$. By the argument similar to that in Lemma 1, R_i goes to infinity when i is a true change point, and $R_i \to 0$ in probability, when i is an n_I-flat point. Hence, the value of R_i can distinguish a change point from a set of n_I-flat points. To further eliminate the non-change points that are also not n_I-flat, we implement the non-maximum suppression that removes away the points which do not give the largest R_i's in their n_I-neighborhood. More specifically, the screening procedure of selecting candidate points is described as follows.

Algorithm 1 : **Screening**

(1) For each i in $[n_I, n - n_I]$, compute R_i.
(2) If $R_i = \max\{R_j : j \in (i - n_I, i + n_I]\}$, i is selected as a candidate point.
(3) After scanning through the data sequence, a set of K_n candidate points, denoted as $\mathcal{H}_c(n_I)$, is obtained.

This screening algorithm is comparable to that in [11], because by the Laplace approximation, we can write

$$\begin{aligned} R_i &= \frac{D_n \prod_{l=i}^{i+n_I-1} \exp\{-(Y_l - \bar{Y}_i - \mu^*)^2\}\pi(\mu^*)}{\prod_{l=i}^{i+n_I-1} \exp\{-(Y_l - \bar{Y}_i)^2\}} \\ &\quad \times\{1 + o_p(1)\} \\ &= D_n \exp\left\{ 2 \left(\sum_{l=i}^{i+n_I-1} Y_l - \sum_{j=i-n_I}^{i-1} Y_j \right) \right. \\ &\quad \left. \times\ \mu^* - n_I \mu^{*2} \right\} \pi(\mu^*)\{1 + o_p(1)\}, \end{aligned}$$

where D_n is a constant of order $O_p(n_I^{-1/2})$ and μ^* is the maximizer of $-\sum_{l=i}^{i+n_I-1}(Y_l - \bar{Y}_i - \mu)^2 + \log\pi(\mu)$. Clearly, the magnitude of the leading term in R_i is strongly associated with $n_I^{-1}\left(\sum_{l=i}^{i+n_I-1} Y_l - \sum_{j=i-n_I}^{i-1} Y_j\right)$, which is the local diagnosis function with $h = n_I$ in [11].

Next, we show the screening procedure identifies a set $\mathcal{H}_c(n_I)$ that would lead to the change point consistency.

Proposition 1. *Assume $n_I^{1/2}\delta/\sigma \to \infty$. For each $t_j \in \mathcal{T}_0(p_0)$, there is a $\tau \in \mathcal{H}_c(n_I)$ such that $\Pr\{t_j \in (\tau - n_I, \tau + n_I)\} = 1 - O[\min\{\exp(-n_I\delta^2), a_{n_I}\}]$.*

In theory, $i = t_j$ maximizes R_i in the n_I-neighborhood of t_j asymptotically. By selecting the local maximal R_i in the screening procedure, $\mathcal{H}_c(n_I)$ would cover the n_I-neighborhood of $\mathcal{T}_0(p_0)$ as $n \to \infty$. Also the condition $n_I^{1/2}\delta/\sigma \to \infty$ indicates that the effect size cannot be too small in order to find the candidate points around the true change points. After selecting the candidate points, we perform the refinement step to identify the locations and the total number of change points.

Algorithm 2 : **Refinement**

Scanning

(1) Compute $\Pr(\mathbf{Y}_n|M_k)$ by scanning over all the candidate points in $\mathcal{H}_c(n_I)$.
(2) For each p, we obtain a set of change points $\widehat{\mathcal{T}}(p)$ corresponding to the p largest $\Pr(\mathbf{Y}_n|M_k)$, $k = 1, \ldots, K_n$.

Optimization

(3) Select $\widehat{p}$ that maximizes $\Pr\{\mathbf{Y}_n|\widehat{\mathcal{T}}(p)\}$.

Theorem 2. *Assume that $n_I/\log(n) \to c$, $0 < c \leq \infty$, $n_I^{1/2}\delta/\sigma \to \infty$, $\limsup_{n\to\infty} n_I/\lambda < 1/2$, and equation (2) holds. Let $\mathcal{H}_c(n_I)$ be the set containing candidate points such that $|\tau_{k+1} - \tau_k| > n_I$, and for each t_j there is a $\tau_k \in \mathcal{H}_c(n_I)$, $\Pr(|t_j - \tau_k| \leq n_I) = 1 - O_p[\min\{\exp(-n_I\delta^2), a_{n_I}\}]$, Then,*

$$\Pr(\widehat{p} = p_0) = 1 - O_p[\max\{\exp(-n_I\delta^2), a_{n_I}\}],$$

and furthermore,

$$\begin{aligned} &\Pr\left\{ \sup_{\widehat{t}_j \in \widehat{\mathcal{T}}(\widehat{p})} \inf_{t_j \in \mathcal{T}_0(p_0)} |(\widehat{t}_j - t_j)/n| \leq n_I/n \right\} \\ &= 1 - O\{\exp(-n_I\delta^2)\}. \end{aligned}$$

and

$$\begin{aligned} &\Pr\left\{ \sup_{t_j \in \mathcal{T}_0(p_0)} \inf_{\widehat{t}_j \in \widehat{\mathcal{T}}(\widehat{p})} |(\widehat{t}_j - t_j)/n| < n_I/n \right\} \\ &= 1 - O(a_{n_I}). \end{aligned}$$

Theorem 2 shows that BMS controls both the over- and under-segmentation errors. The intrinsic rationale is that for any $\mathcal{T}(p)$ different from $\mathcal{T}_0(p_0)$, there is at least a chosen point $\tau \in \mathcal{T}(p)$ whose n_I-neighborhood does not contain true change points. Then the likelihood ratio $\Pr\{\mathbf{Y}_n|\mathcal{T}(p)\}/\Pr\{\mathbf{Y}_n|\mathcal{T}_0(p_0)\}$ goes to 0 with probability 1, because the ratio contains at least one

$$\frac{\int \prod_{l=\tau_j}^{\tau_{j+1}-1} \exp\{-(Y_l - \bar{Y}_{\tau_j} - \mu)^2\}\pi(\mu)d\mu}{\prod_{l=\tau_j}^{\tau_{j+1}-1} \exp\{-(Y_l - \bar{Y}_{\tau_j})^2\}},$$

or

$$\frac{\prod_{l=\tau_k}^{\tau_{k+1}-1} \exp\{-(Y_l - \bar{Y}_{\tau_k})^2\}}{\int \prod_{l=\tau_k}^{\tau_{k+1}-1} \exp\{-(Y_l - \bar{Y}_{\tau_k} - \mu)^2\}\pi(\mu)d\mu}$$

for $\tau_k \in \mathcal{T}_0(p_0)$ and $\tau_j \notin \mathcal{T}_0(p_0)$, which converges to 0 in probability by Lemma 1 and (2).

For a given p, each multiplicand in $\Pr\{\mathbf{Y}_n|\widehat{\mathcal{T}}(p)\}$ can be obtained and stored through the scanning step. The optimization step essentially picks the maximal $\Pr\{\mathbf{Y}_n|\widehat{\mathcal{T}}(p)\}$ among a set of known quantities, which avoids intensive optimization procedures. Because the computational time for $\Pr(\mathbf{Y}_n|M_k)$ grows at the speed of $O(n)$ for $k = 1, \ldots, K_n$, the computational time for the refinement stage grows with the sample size at the speed of $O(nK_n)$.

2.5 An online change point prediction extension

As shown in the proof of Proposition 1 in the Appendix, R_i and R_j can be separated by a constant C for a given n_I, where i is a change point and j is an n_I-flat point. We can estimate the value of C from existing samples. When the new samples arrive, we use this cutoff value to identify new change points in an online detection procedure.

Algorithm 3 : **Online change point detection**

(1) For τ in $\widehat{\mathcal{T}}(\widehat{p})$, compute R_τ and let $C = \min_{\tau \in \widehat{\mathcal{T}}(\widehat{p})} R_\tau$.
(2) For each observation j in the new sample of size n^*, compute R_j.
(3) If the set $\mathcal{N} = \{R_j \geq C, j = 1, \ldots, n^*\}$ is empty, we claim the new sample does not contain any change point.
(4) If the set $\mathcal{N}$ is not empty, we first select the point associated with the largest R_j. Suppose the kth observation is selected, then delete all the points j satisfying $|j - k| < n_I$.
(5) Pick the point associated with the largest R_j in the remaining samples.
(6) Continue this process until no point satisfies $R_j \geq C$.

This procedure essentially selects all the points with $R_j \geq C$ while the minimum distance between two consecutive points is greater than n_I.

3 SIMULATION

3.1 The sequence without spikes

To evaluate the performance of the proposed BMS method in the setting without spike points, we generate data from two different models. Model I takes the form of

$$Y_i = \mathbf{h}^{\mathrm{T}}\mathbf{J}(x_i) + \sigma\epsilon_i$$

where the error term $\epsilon_i \sim N(0,1)$, $\sigma = 0.5$, and $\mathbf{h} = (2.01, -2.51, 1.51, -2.01,$
$2.51, -2.11, 1.05, 2.16, -1.56, 2.56, -2.11)^{\mathrm{T}}$ with $p_0 = 11$. We set $\mathbf{J}(x_i) = \{(1 + sgn(nx_i - t_j))/2, j = 1, \ldots, p_0\}^{\mathrm{T}}$, and the x_i's are equally spaced points on [0, 1]. The true change points are

$$\begin{aligned} & (t_j/n, j = 1, \ldots, p_0) \\ = \; & (0.1, 0.13, 0.15, 0.23, 0.25, 0.40, 0.44, 0.65, 0.76, 0.78, 0.81). \end{aligned}$$

The random errors are generated from three distributions: the standard normal distribution $N(0,1)$; Student's t distribution with 5 degrees of freedom $t(5)$, which is standardized to have a unit variance; and the log-normal distribution $LN(0,1)$, which is the exponential of the standard normal distribution and also standardized to have a unit variance. Model II considers a heteroscedastic error term across the segments. The data generating model is

$$Y_i = \mathbf{h}^{\mathrm{T}}\mathbf{J}(t_i) + \sigma\epsilon_i \prod_{j=1}^{\mathbf{1}^{\mathrm{T}}\mathbf{J}(t_i)} v_j$$

where $(v_j, j = 1, \ldots, 11) = (1, 0.5, 3, 2/3, 0.5, 3, 2/3, 0.5, 3, 2/3, 0.5)$. Other model specifications remain the same as those in model I. We define the over- and under-segmentation errors as $d(\widehat{\mathcal{G}}_n|\mathcal{G}_n)$ and $d(\mathcal{G}_n|\widehat{\mathcal{G}}_n)$ respectively,

$$d(\widehat{\mathcal{G}}_n|\mathcal{G}_n) = \sup_{b\in\mathcal{G}_n}\inf_{a\in\widehat{\mathcal{G}}_n}|a-b|, \quad d(\mathcal{G}_n|\widehat{\mathcal{G}}_n) = \sup_{b\in\widehat{\mathcal{G}}_n}\inf_{a\in\mathcal{G}_n}|a-b|.$$

For the BMS procedure, we consider three different priors for $\pi(\cdot)$, corresponding to the local prior, non-local moment prior and non-local inverse moment priors. We take $n_I = \{\log(n)\}^{1.5}h$, where $h \geq 0.5$ generally works well in the simulations. Figure 3 presents the relationship between the maximum of the over- and under-segmentation errors, $|\widehat{p} - p_0|$ and the value of h with sample size 1000, which indicates $h = 0.65$ leading to the smallest segmentation errors.

Furthermore, we assess the performance of BMS using different priors under model I with a normal error, when p_0 is not prespecified. In Figure 4, we present the selection error which is defined as the maximum of the number of selected change points that are not in $\mathcal{T}_0(p_0)$ and the number of true change points that are not in $\widehat{\mathcal{T}}(\widehat{p})$. The tuning parameters are calibrated to yield the smallest segmentation error and $|\widehat{p} - p_0|$ on average for each prior. Clearly, both the selection error and $|\widehat{p} - p_0|$ decrease with the increasing sample size, and the prior $\pi_I(\cdot)$ leads to the best convergence among the three priors.

For a comprehensive comparison with existing methods, we consider BMS under the non-local inverse moment prior $\pi_I(\cdot)$ with $q = v = 2$ and $s = 6$, PELT [31], WBS [13], NOT

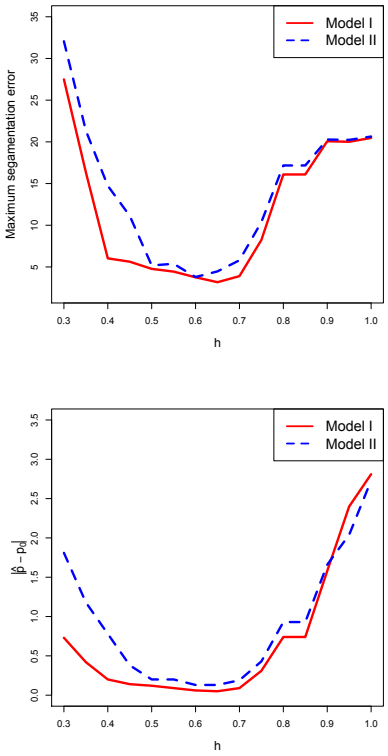


Fig. 3: The effect of the tuning parameter h in the minimum distance between the candidate points $n_I = \log(n)^{1.5}h$: the maximum segmentation error (upper panel) and $|\widehat{p} - p_0|$ (bottom panel) versus h over 100 simulations with sample size $n = 1000$.

with normal or heavy-tail distributions [14] and SML [3]. We use the default values of the tuning parameters in the packages for the comparative methods. Table 1 (at the end of the text) summarizes the numerical results under model I and model II with normal, Student's t, and log-normal error distributions and their heteroscedastic counterparts, respectively. On average, BMS performs the best in selecting the number of change points and balancing both over- and under-segmentation errors. The WBS method performs the best when the error distribution is normal, while its performance is unsatisfactory when the error distribution deviates from the normal. It is expected that the performances of WBS, PELT, and SML deteriorate when the errors do not follow a normal distribution, because all the three procedures heavily rely on the parametric model assumptions, and hence they are not robust to model mis-specifications. In contrast, both BMS and NOT behave well under various error distributions. Also, NOT and SML perform the best in controlling the over-segmentation errors, while the resulting estimator $\widehat{p}$ tends to be larger than the true p_0. On the other hand, the BMS allows for slightly larger over-segmentation errors in order to maintain $\widehat{p}$ to be more concentrated around p_0.

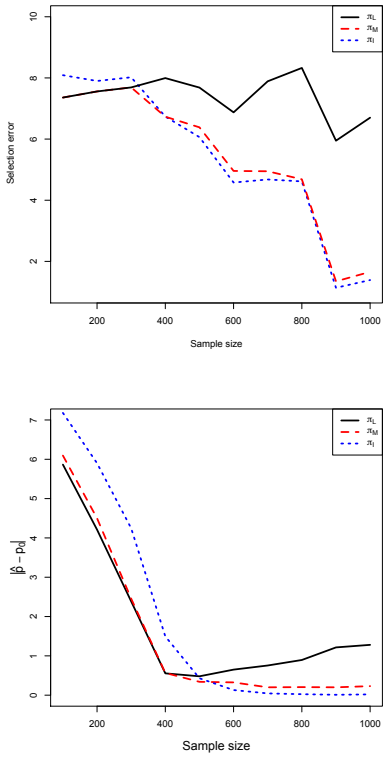


Fig. 4: The selection error (upper panel) and $|\widehat{p} - p_0|$ (lower panel) averaged over 500 simulations under three different prior distributions: the local prior π_L, non-local moment prior π_M, and non-local inverse moment prior π_I.

3.2 The sequence with spike points.

In addition, we illustrate the features of the BMS, NOT and SML methods on the sequences contaminated with spike points. Assume the noises are normal, we generate 500 sequences each contains $n = 1000$ points with mean changes at 0.01 and -0.01 on the 400 and 440's observations, respectively. We set the noise standard deviation to be 0.002. Further, we generate 10 random samples uniformly in the range of $(-0.07, -0.08)$ and $(0.07, 0.08)$, and add them to the original sequence at random locations to form up the spike points. Note that we choose these parameters to mimic the real data setting. We implement BMS, NOT, and SML on the simulated samples. In BMS, we select $n_I = 12$, which is the largest integer that smaller than $0.65\{\log(n)\}^{1.5}$.

From Table 2, we can seen that the BMS is insensitive to the spike points with the smallest $|\widehat{p} - p_0|$ on average. Further, we present three simulation results in Figure 5. It clear that NOT ignores both the change points with small signal noise ratio and the spike signals with small segment length. That is because NOT is constructed to work in the settings where the segments have comparative lengths. On the other hand, SML is sensitive to the extreme value. That is because SML is developed to treat frequent and irregular change points. To this end, BMS is the most suitable procedure for the MRgRT data, because on the one hand it reinforces the minimal segment length to avoid the identification of the spike signal; on the other hand it retains small minimal segment length to detect change points with short distance.

TABLE 2: Comparison results averaged over 500 simulations among the BMS, NOT and SML methods on the data sequences with spike points.

Method	$\widehat{p} - p_0$				
	≤ -1	0	1	2	≥ 3
BMS	31	276	113	67	13
NOT	387	32	20	11	50
SML	0	0	0	0	500

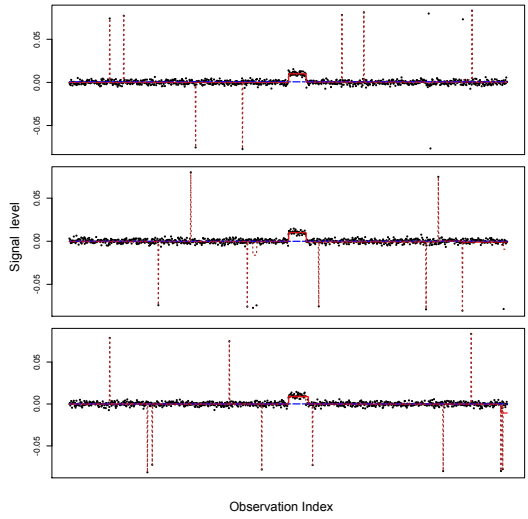


Fig. 5: Change points detection for 3 simulated sequence with spike points. BMS: the red line solid; NOT: the blue dash line; SML: the brown two dash linear.

3.3 The evaluation of the online algorithms

To further evaluate the prediction using the online change point detection algorithm, we consider model I with a normal error. We first simulate 40 samples with one change point at the 21st observation, where the effect size between two consecutive segments range from 0 to 6. We show the correct selection proportion of the change points versus the effect size in Figure 6. As the smallest effect size in the training sample is 2.1, the correct selection proportion is small when the effect size is less than 2, and it exceeds 60% when the effect size increases to 2.2. This demonstrates that the online change point detection can only identify the change points associated with effect sizes larger than the minimum in the training sample.

For online multiple change points detection, we simulate data from model II with a normal error and sample sizes from 200 to 300, where the change points are $(t_j/n, j = 1, \ldots, 4) = (0.20, 0.40, 0.6, 0.85)$, $\mathbf{h} = (3, 1.5, -2, -1.5)$, and $(v_j, j = 1, \ldots, 4) = (1, 1, 0.5, 0.5)$. This choice of $\mathbf{h}$ guarantees that the effect sizes from two consecutive segments are larger than the minimal effect size in the original sample, and in turn ensures that the online change point detection is feasible to detect the change points in

the new sample. The lower panel of Figure 6 shows the relation between the segmentation error and the new sample size, which indicats that the segmentation errors are well controlled in the new samples.

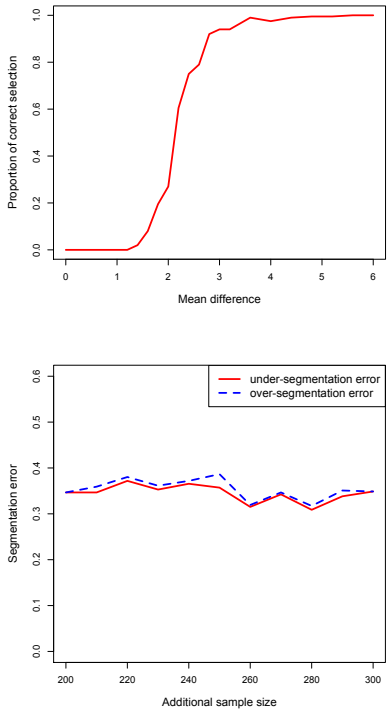


Fig. 6: Predictive results for online change point detection averaged over 500 simulations. The upper panel exhibits the relationship between the proportion of correct selection with the mean difference (effect size), and the lower panel shows the under- and over-segmentation errors for different additional sample sizes.

4 MRgRT DATA

We illustrate the BMS method with the application to the MRgRT data which contains 2265 data points ordered by the distances from the sources of the radiations dose. The R code for implementing the method can be downloaded from our GitHub repository [32]. Throughout the implementation, we use the non-local inverse moment prior $\pi_I(\mu) = s\nu^{q/2}/\Gamma\{q/(2s)\}\mu^{-(q+1)}\exp\{-(\mu^2/\nu)^{-s}\}$ with $q = 2, \nu = 2$. We set $n_I = 13$, which is the largest integer that smaller than $0.65\{\log(n)\}^{1.5}$.

We first vary s from 2 to 10. Figure 8 shows that when s is small, we identify more change points than the true ones, and as s grows, the number of identified change points decreases. This phenomenon is consistent with the result in Lemma 4 that the convergence rate for the non-local prior is $O_p\{\exp(-n_I^{s/(1+s)})\}$. When s is small, the Bayes factor vanishes slowly and hence the algorithm picks redundant change points. When s is sufficiently large, the convergence rate approaches to $O_p\{\exp(-n_I)\}$, and hence the algorithm eliminates the flat points more effectively. In the following analysis, we select $s = 10$, with which the algorithm provides the best result in Figure 8. Furthermore, we vary h from 0.35 to 0.65. Figure 7 shows that when h is 0.35, the BMS algorithm yields irregular results with higher effects from the spike points. When h is greater than 0.5, the algorithm gives smoother results, which are protected from the instantaneous changes.

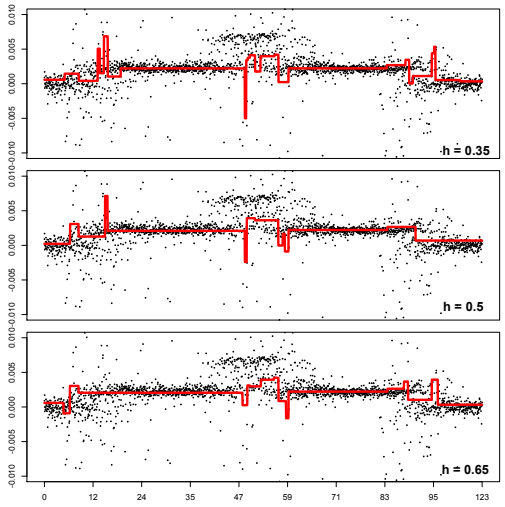


Fig. 7: Change points detection when using different $h = 0.35, 0.5, 0.65$

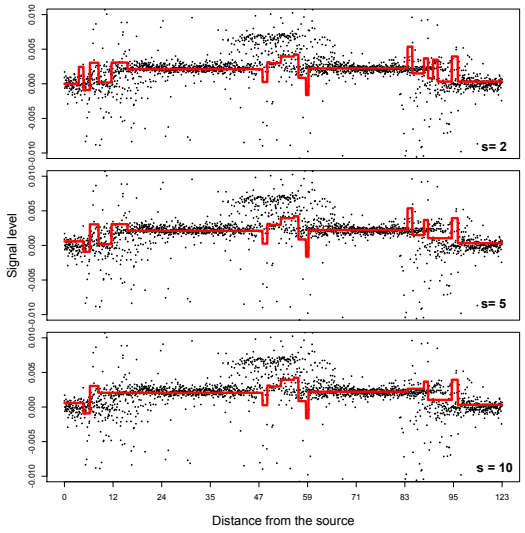


Fig. 8: Change points detection when using different $s = 2, 5, 10$

Interestingly, after we remove the spike points and kept the data within the range of $(-0.01, 0.01)$, we implement the NOT, SML, and BMS on the truncated sequence. Figures 9 shows that the results from the three methods are overlapped. The NOT and BMS methods have similar results, and both outperform SML. This implies that removing the spike points improves the boundary detection accuracy for all the three methods. However, the spike remover is infeasible in practice, because the locations and the magnitudes of the spike are difficult to track in human bodies.

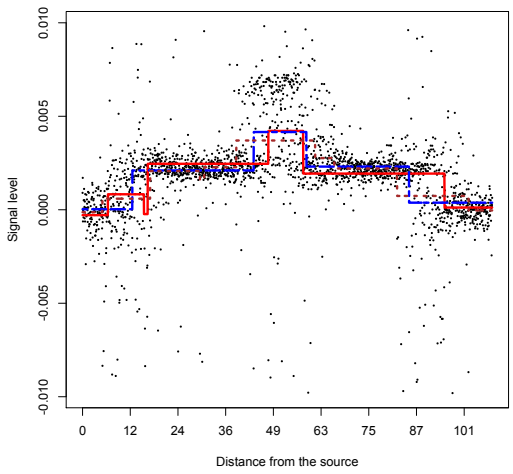


Fig. 9: Change points detection after restricting the data in the range of (-0.01, 0.01). BMS: the red line solid; NOT: the blue dash line; SML: the brown two dash linear.

5 Conclusion

We propose the BMS method that consistently identifies multiple mean changes in a data sequence. The BMS method removes the flat points effectively without sacrificing the detection accuracy. Further, our method is particularly useful when the data sequence contains spike points, which are not of interest. We apply the BMS to analyze the MRgRT data, for which the NOT, SMT and other methods fail to but the BMS algorithm correctly detects the mean changes boundaries. We explore the BMS performance with different tuning parameters, and the resulting patterns are consistent with the theoretical properties. Moreover, we demonstrate that the BMS is robust to the error distributions by evaluating the detection procedures on the sequence with different random errors.

Due to the patterns in the motivating data, our discussion mainly focuses on detecting the mean changes. For extension to analyze changes in the variance and correlation structures, one needs to redefine the marginal likelihood according to specific model assumptions. Furthermore, the selection of the prior requires further investigation to guarantee the consistency of the procedure.

References

[1] Yi-Ching Yao, "Approximating the distribution of the maximum likelihood estimate of the change-point in a sequence of independent random variables," *The Annals of Statistics*, vol. 15, no. 3, pp. 1321–1328, 1987.

[2] Lajos Horváth, "The maximum likelihood method for testing changes in the parameters of normal observations," *The Annals of Statistics*, vol. 21, no. 2, pp. 671–680, 1993.

[3] Chao Du, Chu-Lan Michael Kao, and SC Kou, "Stepwise signal extraction via marginal likelihood," *Journal of the American Statistical Association*, vol. 111, no. 513, pp. 314–330, 2016.

[4] Yi-Ching Yao, "Estimating the number of change-points via schwarz' criterion," *Statistics & Probability Letters*, vol. 6, no. 3, pp. 181–189, 1988.

[5] Yi-Ching Yao and ST Au, "Least-squares estimation of a step function," *Sankhyā: Indian Journal of Statistics, Series A*, vol. 51, no. 3, pp. 370–381, 1989.

[6] Nancy R Zhang and David O Siegmund, "A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data," *Biometrics*, vol. 63, no. 1, pp. 22–32, 2007.

[7] Changliang Zou, Guosheng Yin, Long Feng, and Zhaojun Wang, "Nonparametric maximum likelihood approach to multiple change-point problems," *The Annals of Statistics*, vol. 42, no. 3, pp. 970–1002, 2014.

[8] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight, "Sparsity and smoothness via the fused lasso," *Journal of the Royal Statistical Society: Series B*, vol. 67, no. 1, pp. 91–108, 2005.

[9] Robert Tibshirani and Pei Wang, "Spatial smoothing and hot spot detection for CGH data using the fused lasso," *Biostatistics*, vol. 9, no. 1, pp. 18–29, 2008.

[10] Leif Boysen, Angela Kempe, Volkmar Liebscher, Axel Munk, and Olaf Wittich, "Consistencies and rates of convergence of jump-penalized least squares estimators," *The Annals of Statistics*, vol. 31, no. 1, pp. 157–183, 2009.

[11] Yue S Niu and Heping Zhang, "The screening and ranking algorithm to detect dna copy number variations," *The Annals of Applied Statistics*, vol. 6, no. 3, pp. 1306–1326, 2012.

[12] Chun Yip Yau and Zifeng Zhao, "Inference for multiple change points in time series via likelihood ratio scan statistics," *Journal of the Royal Statistical Society: Series B*, vol. 78, no. 4, pp. 895–916, 2015.

[13] Piotr Fryzlewicz, "Wild binary segmentation for multiple change-point detection," *The Annals of Statistics*, vol. 42, no. 6, pp. 2243–2281, 2014.

[14] Rafal Baranowski, Yining Chen, and Piotr Fryzlewicz, "Narrowest-over-threshold detection of multiple change-points and change-point-like features," *arXiv preprint arXiv:1609.00293*, 2016.

[15] Peter Eiauer and Peter Hackl, "The use of MOSUMS for quality control," *Technometrics*, vol. 20, no. 4, pp. 431–436, 1978.

[16] Marc Lavielle and Carenne Ludeña, "The multiple change-points problem for the spectral distribution," *Bernoulli*, vol. 6, no. 5, pp. 845–869, 2000.

[17] Joseph Glaz, Joseph I Naus, Sylvan Wallenstein, Sylvan Wallenstein, and Joseph I Naus, *Scan Statistics*, Springer, 2001.

[18] Philip Preuss, Ruprecht Puchstein, and Holger Dette, "Detection of multiple structural breaks in multivariate time series," *Journal of the American Statistical Association*, vol. 110, no. 510, pp. 654–668, 2015.

[19] Claudia Kirch and Birte Muhsal, "A MOSUM procedure for the estimation of multiple random change points," *Preprint*, 2014.

[20] AM Walker, "On the asymptotic behaviour of posterior distributions," *Journal of the Royal Statistical Society.*

Series B, vol. 31, no. 1, pp. 80–88, 1969.

[21] Valen E Johnson and David Rossell, "On the use of non-local prior densities in Bayesian hypothesis tests," *Journal of the Royal Statistical Society: Series B*, vol. 72, no. 2, pp. 143–170, 2010.

[22] Stephen G Walker, "Modern Bayesian asymptotics," *Statistical Science*, pp. 111–117, 2004.

[23] Francesco Bertolino, Walter Racugno, and Elías Moreno, "Bayesian model selection approach to analysis of variance under heteroscedasticity," *The Statistician*, pp. 503–517, 2000.

[24] Caterina Conigliani and Anthony O'Hagan, "Sensitivity of the fractional bayes factor to prior distributions," *Canadian Journal of Statistics*, vol. 28, no. 2, pp. 343–352, 2000.

[25] Fulvio De Santis and Fulvio Spezzaferri, "Consistent fractional bayes factor for nested normal linear models," *Journal of Statistical Planning and Inference*, vol. 97, no. 2, pp. 305–321, 2001.

[26] Parhasarathi Lahiri, "Model selection," IMS, 2001.

[27] Juan Antonio Cano, Mathieu Kessler, and Elías Moreno, "On intrinsic priors for nonnested models," *Test*, vol. 13, no. 2, pp. 445–463, 2004.

[28] Elías Moreno, "Objective bayesian methods for one-sided testing," *Test*, vol. 14, no. 1, pp. 181–198, 2005.

[29] George Casella and Elías Moreno, "Objective bayesian variable selection," *Journal of the American Statistical Association*, 2012.

[30] Harold Jeffreys, *The Theory of Probability*, Oxford University Press, 1998.

[31] Rebecca Killick, Paul Fearnhead, and IA Eckley, "Optimal detection of change points with a linear computational cost," *Journal of the American Statistical Association*, vol. 107, no. 500, pp. 1590–1598, 2012.

[32] Fei Jiang, Guosheng Yin, and Francesca Dominici, "Bayesian model selection approach to boundary detection with non-local priors," https://github.com/homebovine/BCP.git, 2018.

TABLE 1: Comparison results averaged over 200 simulations among the BMS, PELT, WBS, NOT and SML methods under model I and model II under different error distributions: the standard normal distribution $N(0,1)$, Student's $t(5)$, and log-normal $LN(0,1)$ with constant variances; and the corresponding distributions with heteroscedastic variances. Standard deviations are in parentheses.

Error Distribution	Method	$\widehat{p}-p_0$							$d(\mathcal{G}_n\|\widehat{\mathcal{G}}_n)$	$d(\widehat{\mathcal{G}}_n\|\mathcal{G}_n)$
		≤ -3	-2	-1	0	1	2	≥ 3		
$N(0,1)$	BMS	0	0	1	197	2	0	0	2.41 (6.06)	1.96 (3.94)
	PELT	0	1	37	162	0	0	0	0.91 (1.19)	6.32 (11.92)
	WBS	0	0	0	194	6	0	0	1.22 (4.13)	0.86 (0.79)
	NOT	0	0	0	192	7	1	0	1.93 (8.04)	0.75 (0.80)
	SML	0	0	0	132	52	13	3	12.94 (42.98)	0.78 (0.90)
$t(3)$	BMS	0	0	8	190	2	0	0	2.15 (5.76)	2.83 (7.01)
	PELT	0	4	31	165	0	0	0	0.95 (1.03)	6.24 (12.24)
	NOT	0	0	3	184	3	6	4	7.57 (27.70)	1.51 (2.57)
	SML	0	0	0	42	34	44	80	40.13 (53.68)	0.88 (0.87)
$LN(0,1)$	BMS	0	0	12	180	6	1	2	3.69 (12.12)	3.11 (6.89)
	PELT	1	2	21	135	15	23	3	12.10 (29.63)	7.22 (13.32)
	NOT	0	1	4	183	7	1	4	6.06 (26.32)	1.18 (4.45)
	SML	0	0	0	0	0	4	196	111.77 (52.05)	0.73 (1.33)
Heterosced $N(0,1)$	BMS	0	0	13	176	8	3	0	3.69 (7.08)	3.88 (7.36)
	PELT	0	0	31	169	0	0	0	1.49 (1.53)	6.15 (11.44)
	NOT	0	0	0	150	23	21	6	7.52 (12.60)	1.66 (1.68)
	SML	0	0	0	119	55	20	6	6.75 (23.98)	1.37 (1.52)
Heterosced $t(5)$	BMS	0	0	14	181	4	1	0	2.79 (4.87)	4.21 (8.17)
	PELT	0	4	35	159	2	0	0	1.50 (1.83)	7.76 (13.48)
	NOT	0	1	5	179	10	2	3	8.15 (25.01)	2.36 (4.70)
	SML	0	0	0	43	22	41	94	26.74 (35.42)	1.27 (1.59)
Heterosced $LN(0,1)$	BMS	0	0	20	173	7	0	0	3.73 (11.72)	4.20 (7.85)
	PELT	0	2	21	142	22	12	1	6.07 (16.09)	8.10 (13.93)
	NOT	0	0	4	183	7	4	2	6.32 (24.67)	1.42 (3.70)
	SML	0	0	0	2	1	0	197	68.05 (47.15)	0.87 (1.67)