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

Anonymous Author(s)
Affiliation
Address
email

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. 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.

1 Introduction

Traditional change point detection algorithms often handle the cases where the occurrent frequency of the change points is relatively consistent across the signal. For example, the popular narrowest-over-threshold (NOT) [1] algorithm is suitable under the setting where the different segments between the change points have comparable lengths, while the stepwise marginal likelihood (SML) method [5] works the best to identify the frequent change points. However, it is often case that the gaps between the consecutive change points are dramatically different, while only ones with certain distances are of interest. In this paper, we develop a computational efficient Bayesian model selection algorithm for identifying the change points under this specific setting.

The inconsistent gaps between the change points can be observed from the signals generated by the magnetic resonance imaging guided radiation therapy (MRgRT). One problem with MRgRT is that when radiation traveled in the magnetic field, the dose level can be significantly enhanced near the boundaries between different tissues or organs in human bodies. The Duke Mid-sized Optical-CT System (DMOS) (See Figure A.1 in Appendix) was developed to identify the dose changes near the region of this boundary artifact. The right panel in Figure A.1 contains one radiation dose, 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.

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 A.1) which makes the boundary detection extremely challenging. Figures A.2 in Appendix shows the change points in the MRgRT data identified by the popular NOT [1] and the SML [5] algorithms, respectively. It can be

Submitted to 31st Conference on Neural Information Processing Systems (NIPS 2017). Do not distribute.

seen that neither of the algorithms correctly identify the true boundaries. This motivate us to propose a new algorithm in detecting the systematic changes when the segment length can have dramatic differences.

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 enforce 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 [6, 7, 8, 13, 14, 16, 18, 19].

To perserve the positive detection rate of the change points and reduce the false detection rate of the non-change points, 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 [2, 3, 4, 17] and non-local priors [11], whereas the convergence rate is faster under the non-local priors.

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, which allows the algorithms to be implemented in parallel.

2 Bayesian multiple change points detection

2.1 Probability model

Suppose there are p_0 true change points $t_1 \leqslant, \ldots, \leqslant t_{p_0}$ 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 $n_j = \tau_{j+1} - \tau_j$, $n_I = \min_{j=0,\ldots,K_n-1} n_j$, $n_I \leqslant \lambda$. 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 $l \geqslant \tau_1$ 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 [9], the non-local moment prior and the inverse moment prior in [11].

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

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

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

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

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

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

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

The two marginal likelihoods indicate 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*

$$\Pr(\mathbf{Y}_n|M_j) = O_p(a_{n_j}) \tag{2}$$

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

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

Hence when $n_I/\log(n) \to c$, $0 < c \leqslant \infty$, $n_I \leqslant \lambda$, 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.

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 $\mathcal{T}(p)$, that is $\Pr\{\mathbf{Y}_n | \mathcal{T}(p)\}$ as

$$\prod_{\tau_j \notin \mathcal{T}(p)} \prod_{l=\tau_j}^{\tau_{j+1}-1} \exp\{-(Y_l - \bar{Y}_{\tau_j})^2\} \prod_{\tau_k \in \mathcal{T}(p)} \int \prod_{l=\tau_k}^{\tau_{k+1}-1} \exp\{-(Y_l - \bar{Y}_{\tau_k} - \mu)^2\} \pi(\mu) d\mu.$$

We can estimate the locations and the number of change points in two steps: (i) for any given p, we obtain $\hat{\mathcal{T}}(p)$ using the procedure described in the previous section, and (ii) we estimate p_0 by $\hat{p}$ via maximizing $\Pr\{\mathbf{Y}_n | \hat{\mathcal{T}}(p)\}$ with respect to p. In contrast to the procedure in [5] 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 $\Pr(|t_j - \tau_k| \leqslant n_I) = 1 - O_p[\min\{\exp(-n_I\delta^2), a_{n_I}\}]$. 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 [15], because by the Laplace approximation, we can write

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

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 [15].

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

Theorem 2. *Assume that $n_I/\log(n) \to c$, $0 < c \leqslant \infty$, $n_I^{1/2}\delta/\sigma \to \infty$, $\limsup_{n\to\infty} n_I/\lambda < 1/2$, 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 $\tau_k \in \mathcal{H}_c(n_I)$, $\Pr(|t_j - \tau_k| \leqslant n_I) = 1 - O_p[\min\{\exp(-n_I\delta^2), a_{n_I}\}]$, Then,*

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

and furthermore,

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

and

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

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 of $\Pr(\mathbf{Y}_n|M_j)$ and $\Pr(\mathbf{Y}_n|M_j)^{-1}$ 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 $O(nK_n)$.

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

$$(t_j/n, j = 1, \ldots, p_0) = (0.1, 0.13, 0.15, 0.23, 0.25, 0.40, 0.44, 0.65, 0.76, 0.78, 0.81).$$

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. To save the space, we only present the results by using the non-local inverse moment prior. We take $n_I = \{\log(n)\}^{1.5}h$, where $h \geqslant 0.5$ generally works well in the simulations.

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 [12], WBS [7], NOT with normal or heavy-tail distributions [1] and SML [5]. Table 1 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. 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.

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. In the experiments (not show) NOT ignores both the change points with small signal

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	$\widehat{p}-p_0$							$d(\mathcal{G}_n\|\widehat{\mathcal{G}}_n)$	$d(\widehat{\mathcal{G}}_n\|\mathcal{G}_n)$
		$\leqslant -3$	-2	-1	0	1	2	$\geqslant 3$		
$N(0,1)$	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)

noise ratio and the spike signals with small segment length. On the other hand, SML is sensitive to the extreme value. 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	$\widehat{p}-p_0$				
	$\leqslant -1$	0	1	2	$\geqslant 3$
BMS	31	276	113	67	13
NOT	387	32	20	11	50
SML	0	0	0	0	500

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 [10]. Throughout the implementation, we use the non-local inverse moment prior $\pi_I(\mu) = s\nu^{q/2}/\Gamma\{q/(2s)\}\mu^{-(q+1)}\exp\left\{-(\mu^2/\nu)^{-s}\right\}$ 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 1 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 1.

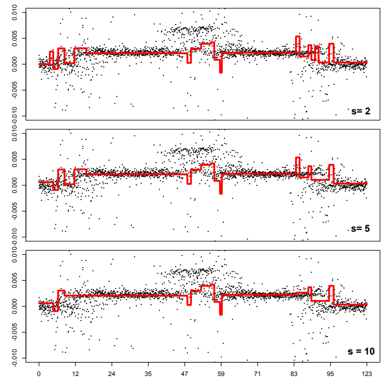

Figure 1: Change points detection when using different $s = 2, 5, 10$

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 2 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.

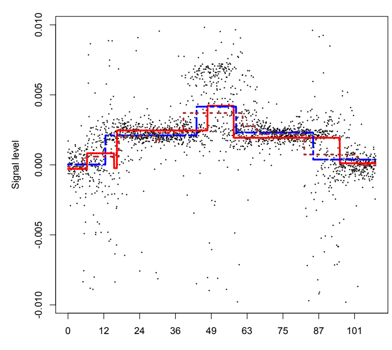

Figure 2: Change points detection after restricting the data in the range of (-0.01, 0.01). BMS: the red line solid; NOT: the blue dash line; SML: the brown two dash linear.

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.

References

[1] 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.

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

[3] Caterina Conigliani and Anthony O'hagan. Sensitivity of the fractional bayes factor to prior distributions. *Canadian Journal of Statistics*, 28(2):343–352, 2000.

[4] Fulvio De Santis and Fulvio Spezzaferri. Consistent fractional bayes factor for nested normal linear models. *Journal of statistical planning and inference*, 97(2):305–321, 2001.

[5] Chao Du, Chu-Lan Michael Kao, and SC Kou. Stepwise signal extraction via marginal likelihood. *Journal of the American Statistical Association*, 111(513):314–330, 2016.

[6] Peter Eiauer and Peter Hackl. The use of MOSUMS for quality control. *Technometrics*, 20(4):431–436, 1978.

[7] Piotr Fryzlewicz. Wild binary segmentation for multiple change-point detection. *Annals of Statistics*, 42(6):2243–2281, 2014.

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

[9] Harold Jeffreys. *The theory of probability*. Oxford University Press, 1998.

[10] Fei Jiang, Guosheng Yin, and Francesca Dominici. Bayesian multiple change points detection with non-local priors. *R code*, 2017.

[11] 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*, 72(2):143–170, 2010.

[12] Rebecca Killick, Paul Fearnhead, and IA Eckley. Optimal detection of change points with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598, 2012.

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

[14] Marc Lavielle and Carenne Ludeña. The multiple change-points problem for the spectral distribution. *Bernoulli*, 6(5):845–869, 2000.

[15] Yue S Niu and Heping Zhang. The screening and ranking algorithm to detect dna copy number variations. *Annals of Applied Statistics*, 6(3):1306, 2012.

[16] Philip Preuss, Ruprecht Puchstein, and Holger Dette. Detection of multiple structural breaks in multivariate time series. *Journal of the American Statistical Association*, 110(510):654–668, 2015.

[17] Stephen G Walker. Modern bayesian asymptotics. *Statistical Science*, pages 111–117, 2004.

[18] Yi-Ching Yao. Approximating the distribution of the maximum likelihood estimate of the change-point in a sequence of independent random variables. *Annals of Statistics*, 15(3), 1987.

[19] 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*, 78(4):895–916, 2015.

Appendix

A Additional Figures

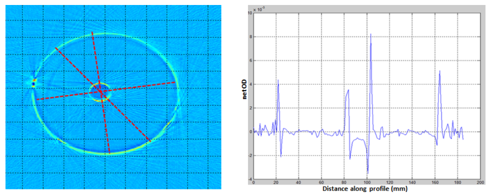

Figure A.1: Reconstructed image of a slice in a cylindrical dosimeter with a cavity in the middle (left) and a typical line profile through the center of the cavity (right). The radiation enter the dosimeter from the hole on the left of the cylindrical dosimeter. The cylindrical rotates 360 degree so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

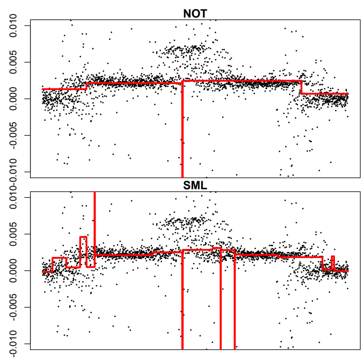

Figure A.2: The change point detection results from the SML with normal prior and the maximal number of change points to be 30 and NOT methods for the dose level changes in the DMOS system.

B Proofs

Let $g'(\mu)$ and $g''(\mu)$ denote the first and second derivatives of a generic function $g(\mu)$ with respect to μ respectively, and further define the utility function as

$$U_k(\mu) = -\sum_{l=\tau_k}^{\tau_{k+1}-1} (Y_l - \bar{Y}_{\tau_k} - \mu)^2.$$

The following conditions are imposed for the theoretical derivations.

(A1) Assume μ to be in a closed set of points in $\mathbb{R}$.

(A2) Assume $\pi(\mu)$ to be a continuous density function with bounded first and second derivatives.

(A3) Assume that $\Pr\{\mathbf{Y}_n | \mathcal{T}(p)\}$ has a unique maximizer in the neighborhood of $\mathcal{T}_0(p_0)$.

Lemma 1. *Assume that τ_k is a change point for which the mean of $Y_l - \bar{Y}_{\tau_k}$ satisfies $|\mu_{k0}| > \delta$ for $\delta > 0$, $n_k^{1/2}\delta/\sigma \to \infty$, then there is a constant $D > 0$ such that*

$$\lim_{n\to\infty} \Pr\left[\frac{\int \prod_{l=\tau_k}^{\tau_{k+1}-1} \exp\{-(Y_l - \bar{Y}_{\tau_k} - \mu)^2\}\pi(\mu)d\mu}{\prod_{l=\tau_k}^{\tau_{k+1}-1} \exp\{-(Y_l - \bar{Y}_{\tau_k})^2\}} > \exp(Dn_k\delta^2)\right] = 1.$$

Proof: By the definition of $U_k(\mu)$, we can write

$$\frac{\int \prod_{l=\tau_k}^{\tau_{k+1}-1} \exp\{-(Y_l - \bar{Y}_{\tau_k} - \mu)^2\}\pi(\mu)d\mu}{\prod_{l=\tau_k}^{\tau_{k+1}-1} \exp\{-(Y_l - \bar{Y}_{\tau_k})^2\}} = \frac{\int \exp\{U_k(\mu)\}\pi(\mu)d\mu}{\exp\{U_k(0)\}}.$$

We first define $\mathcal{N}_\delta(\mu_{k0}) = \{\mu : |\mu - \mu_{k0}| < \delta\}$ and $\mathcal{N}_\delta^c(\mu_{k0})$ as its compliment, and then show

$$\lim_{n\to\infty} \Pr\left[\sup_{\mu\in\mathcal{N}_\delta^c(\mu_{k0})} \{U_k(\mu) - U_k(\mu_{k0})\} < -Dn_k\delta^2\right] = 1. \tag{4}$$

Note that

$$\begin{aligned} & U_k(\mu) - U_k(\mu_{k0}) \\ = \;& n_k\left\{(\mu_{k0}^2 - \mu^2) - n_k^{-1}(\mu_{k0} - \mu)\sum_{l=\tau_k}^{\tau_{k+1}-1} 2(Y_l - \bar{Y}_{\tau_k})\right\} \\ = \;& n_k\{(\mu_{k0}^2 - \mu^2) - 2(\mu_{k0} - \mu)\mu_{k0} + O_p(n_k^{1/2}|\mu_{k0} - \mu|\sigma_k)\} \\ = \;& n_k\{-(\mu - \mu_{k0})^2\} + O_p(n_k^{1/2}|\mu_{k0} - \mu|\sigma_k) \\ \leqslant \;& -n_k\delta^2 + O_p(n_k^{1/2}|\mu_{k0} - \mu|\sigma_k) \\ = \;& -n_k\delta^2/2 - n_k\delta^2/2 + O_p(n_k^{1/2}|\mu_{k0} - \mu|\sigma_k). \end{aligned}$$

As $n_k^{1/2}\delta/\sigma \to \infty$, we have $-n_k\delta^2/2 + O_p(n_k^{1/2}|\mu_{k0} - \mu|\sigma_k) < 0$ with probability 1, and thus

$$\lim_{n\to\infty} \Pr\left[\sup_{\mu\in\mathcal{N}_\delta^c(\mu_{k0})} \{U_k(\mu) - U_k(\mu_{k0})\} < -Dn_k\delta^2\right] = 1.$$

When τ_k is a change point, let $\mu = 0$, because $|\mu_{k0}| > \delta$, we have

$$\lim_{n\to\infty} \Pr\left[\frac{\exp\{U_k(0)\}}{\exp\{U_k(\mu_{k0})\}} < \exp(-Dn_k\delta^2)\right] = 1. \tag{5}$$

By the Laplace approximation,

$$\int \exp\{U_k(\mu)\}\pi(\mu)d\mu = O_p[-U_k''(\widetilde{\mu})^{-1/2}\exp\{U_k(\widetilde{\mu})\}\pi(\widetilde{\mu})], \tag{6}$$

where $\widetilde{\mu}$ is the maximizer of $U_k(\mu) + \log\{\pi(\mu)\}$, and $U_k''(\widetilde{\mu}) = O_p(n_j)$. Let $\widehat{\mu}$ be the maximizer of $U_k(\mu)$, and then

$$\begin{aligned} 0 &= L_k'(\widetilde{\mu}) + \partial\log\pi(\widetilde{\mu})/\partial\mu \\ &= L_k''(\mu^*)(\widetilde{\mu} - \widehat{\mu}) + \partial\log\pi(\widetilde{\mu})/\partial\mu, \end{aligned}$$

where μ^* is a point on the line segment between $\widetilde{\mu}$ and $\widehat{\mu}$. As $\pi(\mu)$ has two bounded derivatives by condition (A2), $L_k''(\mu^*) = O_p(n_j)$, we have $\widetilde{\mu} - \widehat{\mu} = O_p(n_j^{-1})$. Therefore, (6) can be written as

$$\begin{aligned} \int \exp\{U_k(\mu)\}\pi(\mu)d\mu &= O_p[-U_k''(\widehat{\mu})^{-1/2}\exp\{U_k(\widehat{\mu})\}\pi(\widehat{\mu})] \\ &= O_p[-U_k''(\mu_{k0})^{-1/2}\exp\{U_k(\mu_{k0})\}\pi(\mu_{k0})], \end{aligned} \tag{7}$$

where the last equality holds because $\widehat{\mu}$ is the least squared estimator. This implies

$$\frac{\exp\{U_k(\mu_{k0})\}}{\int \exp\{U_k(\mu)\}\pi(\mu)d\mu} = O_p(n_k^{1/2}), \tag{8}$$

which in conjunction with (5) leads to

$$\lim_{n\to\infty} \Pr\left[\frac{\exp\{U_k(0)\}}{\int \exp\{U_k(\mu)\}\pi(\mu)d\mu} < \exp(-Dn_k\delta^2)\right] = 1.$$

By condition (A2) and the boundedness of $U_k(\mu)$, we have

$$\lim_{n\to\infty} \Pr\left[\frac{\int \exp\{U_k(\mu)\}\pi(\mu)d\mu}{\exp\{U_k(0)\}} > \exp(Dn_k\delta)\right] = 1,$$

which completes the proof.

Lemma 2. *Let $\pi(\mu) = \pi_L(\mu)$ be a local prior, and assume that τ_j is not a change point, i.e., $\mu_{j0} = 0$, then*

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

Proof: By the definition of $U_j(\mu)$, we can write

$$\frac{\int \exp\{U_j(\mu)\}\pi(\mu)d\mu}{\exp\{U_j(0)\}} = \frac{\int \prod_{l=\tau_j}^{\tau_{j+1}-1} \exp\{-(Y_l - \bar{Y}_{\tau_j} - \mu)^2\}\pi(\mu)d\mu}{\prod_{l=\tau_j}^{\tau_{j+1}-1} \exp\{-(Y_l - \bar{Y}_{\tau_j})^2\}}.$$

Using the same argument as that leading to (7) with $\mu_{j0} = 0$, we have

$$\int \exp\{U_j(\mu)\}\pi(\mu)d\mu = O_p[-U''(0)^{-1/2} \exp\{U_j(0)\}\pi(0)].$$

Since $U''(0)^{-1/2} = O_p(n_j^{-1/2})$, and $\pi(0)$ is a bounded density, we have

$$\frac{\int \exp\{U_j(\mu)\}\pi(\mu)d\mu}{\exp\{U_j(0)\}} = O_p(n_j^{-1/2}).$$

Lemma 3. *Let*

$$\pi(\mu) = \pi_M(\mu) \equiv \frac{\mu^{2v}}{C_M}\pi_b(\mu),$$

where C_M is a normalizing constant, $\pi_b(\mu)$ with $\pi_b(0) > 0$ is the base prior density with $2v$ finite moments, and bounded first two derivatives in the neighborhood around 0. Assume that τ_j is not a change point, i.e., $\mu_{j0} = 0$, then

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

Proof: We can write

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

Let $h(\mu) = U_j(\mu) + \log\pi(\mu) = 2v\log(\mu) + \log\{\pi_b(\mu)\} + U_j(\mu)$, and let $\tilde{\mu}$ be the maximizer of $h(\mu)$, then we have

$$2v/\tilde{\mu} + \pi_b'(\tilde{\mu})/\pi_b(\tilde{\mu}) + L_j'(\tilde{\mu}) = 0.$$

If we expand $L_j'(\tilde{\mu})$ around $\hat{\mu}$, the least squared estimator for μ_{j0}, the above equality can be rewritten as

$$2v/n + n^{-1}\tilde{\mu}\pi_b'(\tilde{\mu})/\pi_b(\tilde{\mu}) + L_j''(\mu^*)\tilde{\mu}(\tilde{\mu} - \hat{\mu}) = 0,$$

where μ^* is the point on the line segment between $\hat{\mu}$ and $\tilde{\mu}$. Therefore,

$$\begin{aligned} O_p(n^{-1}) &= \tilde{\mu}(\tilde{\mu} - \hat{\mu}) \\ &= (\tilde{\mu} - \hat{\mu})^2 + \hat{\mu}(\tilde{\mu} - \hat{\mu}) \\ &= (\tilde{\mu} - \hat{\mu} + \hat{\mu}/2)^2 - \hat{\mu}^2/4. \end{aligned}$$

Along with the fact that $\hat{\mu} = O_p(n_j^{-1/2})$, we have $\tilde{\mu} - \hat{\mu} = O_p(n_j^{-1/2})$, and $\tilde{\mu} = O_p(n^{-1/2})$. Next, by the Laplace expansion, we have

$$\int \exp\{h(\mu)\}du = O_p(\{2v/\tilde{\mu}^2 - U_j''(\tilde{\mu})\}^{-1/2} \exp[2v\log(\tilde{\mu}) + \log\{\pi_b(\tilde{\mu})\} + U_j(\tilde{\mu})]),$$

and also

$$\begin{aligned} n^{-1}U_j(\widetilde{\mu}) - n^{-1}U_j(0) &= n^{-1}U_j'(\mu^{**})\widetilde{\mu} \\ &= n^{-1}\{U_j'(\widehat{\mu}) + U_j'(\mu^{**}) - U_j'(\widehat{\mu})\}\widetilde{\mu} \\ &= O_p(\mu^{**} - \widehat{\mu})\widetilde{\mu} \\ &= O_p(n^{-1}), \end{aligned} \tag{9}$$

where μ^{**} is on the line segment between $\widetilde{\mu}$ and 0. Thus,

$$|U_j(\widetilde{\mu}) - U_j(0)| = O_p(1).$$

As a result,

$$\begin{aligned} & \frac{\int \prod_{l=\tau_j}^{\tau_{j+1}-1} \exp\{-(Y_l - \bar{Y}_{\tau_j} - \mu)^2\}\pi(\mu)d\mu}{\prod_{l=\tau_j}^{\tau_{j+1}-1} \exp\{-(Y_l - \bar{Y}_{\tau_j})^2\}} \\ = \; & \frac{\int \exp\{U_j(\mu)\}\pi(\mu)d\mu}{\exp\{U_j(0)\}} \\ = \; & \frac{\int \exp\{h(u)\}du}{\exp\{U_j(0)\}} \\ = \; & O_p(\{2v/\widetilde{\mu}^2 - U_j''(\widetilde{\mu})\}^{-1/2}\exp[2v\log(\widetilde{\mu}) + \log\{\pi_M(\widetilde{\mu})\} + U_j(\widetilde{\mu}) - U_j(0)]) \\ = \; & O_p(n_j^{-1/2}\widetilde{\mu}^{2v}) \\ = \; & O_p(n_j^{-1/2-v}), \end{aligned}$$

where the last equality holds by the fact that $\widetilde{\mu} = O_p(n^{-1/2})$. This proves the result.

Lemma 4. *Let*

$$\pi(\mu) = \pi_I(\mu) \equiv \frac{s\nu^{q/2}}{\Gamma\{q/(2s)\}}\mu^{-(q+1)}\exp\left\{-\left(\frac{\mu^2}{\nu}\right)^{-s}\right\}.$$

Assume that τ_j is not a change point, i.e., $\mu_{j0} = 0$, then

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

Proof: We first write

$$\int \prod_{l=\tau_j}^{\tau_{j+1}-1} \exp\{-(Y_l - \bar{Y}_{\tau_j} - \mu)^2\}\pi(\mu)d\mu = c\int \exp\left\{U_j(\mu) - \mu^{-2s}\nu^s - (q+1)\log(\mu)\right\}d\mu,$$

for a constant c. Let $h(\mu) = U_j(\mu) - \mu^{-2s} - (q+1)\log(\mu)$, and assume $\widetilde{\mu}$ is the maximizer of $h(\mu)$, then we have

$$U_j'(\widetilde{\mu}) + 2s\widetilde{\mu}^{-2s-1}\nu^s - (q+1)\widetilde{\mu}^{-1} = U_j'(\mu^*)(\widetilde{\mu} - \widehat{\mu}) + 2s\widetilde{\mu}^{-2s-1}\nu^s - (q+1)\widetilde{\mu}^{-1} = 0,$$

where μ^* is on the line segment between $\widetilde{\mu}$ and $\widehat{\mu}$. The above equality yields

$$n_j\widetilde{\mu}^{2s+2}(1 - \widehat{\mu}/\widetilde{\mu}) = \frac{2s\nu^s - (q+1)\widetilde{\mu}^{2s}}{-U_j''(\mu^*)/n_j}, \tag{10}$$

which implies $\widetilde{\mu} = O_p(n_j^{1/(2s+2)})$.

From (10), we have $n\widetilde{\mu}^{2s+1}(\widetilde{\mu} - \widehat{\mu}) = O_p(1)$, which leads to

$$\widetilde{\mu} - \widehat{\mu} = O_p\{n_j^{-(4s+3)/(2s+2)}\}. \tag{11}$$

Following (30) in [11] and using our notation, we obtain

$$\int \exp\{h(\mu)\}du \quad = \quad O_p\left[\left\{\frac{(4s^2+2s)^{2s+2}}{\widetilde{\mu}} - U_j''(\widetilde{\mu})\right\}^{-1/2}|\widetilde{\mu}|^{-q-1}\exp\{-\widetilde{\mu}^{-2s}\nu^s + U_j(\widetilde{\mu})\}\right].$$

Expanding $U_j(\tilde{\mu})$ around the least squared estimator $\hat{\mu}$, we have

$$\begin{aligned} U_j(\tilde{\mu}) &= U_j(\hat{\mu}) + 1/2 U_j''(\tilde{\mu} - \hat{\mu})^2 \\ &= U_j(\hat{\mu}) + o_p(1) \\ &= U_j(0) + O_p(1), \end{aligned}$$

where the second equality follows (11), and the last equality follows the same argument as that leading to (9). Therefore, we have

$$\begin{aligned} & \frac{\int \prod_{l=\tau_j}^{\tau_{j+1}-1} \exp\{-(Y_l - \bar{Y}_{\tau_j} - \mu)^2\}\pi(\mu)d\mu}{\prod_{l=\tau_j}^{\tau_{j+1}-1} \exp\{-(Y_l - \bar{Y}_{\tau_j})^2\}} \\ = & \frac{\int \exp\{U_j(\mu)\}\pi(\mu)d\mu}{\exp\{U_j(0)\}} \\ = & \frac{\int \exp\{h(u)\}du}{\exp\{U_j(0)\}} \\ = & O_p\left[\left\{\frac{(4s^2+2s)^{2s+2}}{\tilde{\mu}} - U_j''(\tilde{\mu})\right\}^{-1/2} |\tilde{\mu}|^{-q-1} \exp\{-\tilde{\mu}^{-2s}\nu^s + U_j(\tilde{\mu}) - U_j(0)\}\right] \\ = & O_p\{\exp(-n_j^{s/(s+1)})\}. \end{aligned}$$

This proves the result.

Lemma 5. *Assume $p_0 = 1$ and τ_k is the only true change point. As $n_k^{1/2}\delta/\sigma \to \infty$, $\Pr(M_k|\mathbf{Y}_n) - 1 = O_p\{K_n a_{n_I} \exp(-n_I\delta^2)\}$. Hence when $n_I/\log(n) \to c$, $0 < c \leqslant \infty$, $n_I \leqslant \lambda$, we have $\Pr(M_k|\mathbf{Y}_n) \xrightarrow{p} 1$.*

Proof: First we can write

$$\Pr(M_k|\mathbf{Y}_n) = \left\{1 + \sum_{j\neq k}^{K_n} \frac{\Pr(\mathbf{Y}_n|M_j)}{\Pr(\mathbf{Y}_n|M_k)}\right\}^{-1}. \tag{12}$$

To show $\Pr(M_k|\mathbf{Y}_n) - 1 \to 0$, it is equivalent to showing

$$\sum_{j=1, j\neq k}^{K_n} \frac{\Pr(\mathbf{Y}_n|M_j)}{\Pr(\mathbf{Y}_n|M_p)} \to 0.$$

Note that

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

where

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

and

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

As shown in [11], A is a Bayes factor whose convergence rate is $O_p(a_{n_j})$. For B, first note that the data in $[\tau_k, \tau_{k+1})$ are generated from the model with mean μ_{k0} such that $|\mu_{k0}| > \delta$. Hence, we have

$$B = O_p\{\exp(-n_k\delta^2)\},$$

where the last equality holds by Lemma 1 that

$$\lim_{n\to\infty}\left(\Pr\left[\frac{\prod_{l=\tau_k}^{\tau_{k+1}-1}\exp\{-(Y_l-\bar{Y}_{\tau_k})^2\}}{\int\prod_{l=\tau_k}^{\tau_{k+1}-1}\exp\{-(Y_l-\bar{Y}_{\tau_k}-\mu)^2\}\pi(\mu)d\mu}<\exp(-Dn_k\delta^2)\right]\right)=1,$$

where D is a constant. Combining the convergence rates for A and B, we have

$$AB = O_p\{a_{n_k}\exp(-n_k\delta^2)\}. \tag{13}$$

Thus,

$$\frac{\Pr(\mathbf{Y}_n|M_j)}{\Pr(\mathbf{Y}_n|M_k)} = AB = O_p\{a_{n_I}\exp(-n_I\delta^2)\}, \tag{14}$$

and

$$\sum_{j=1}^{K_n}\frac{\Pr(\mathbf{Y}_n|M_j)}{\Pr(\mathbf{Y}_n|M_k)} - 1 = O_p\{K_n a_{n_I}\exp(-n_I\delta^2)\}$$

Plugging this result into (12), we have

$$\Pr(M_k|\mathbf{Y}_n)\xrightarrow{p}1,$$

which completes the proof.

Proof of Theorem 1

First we can write

$$\sum_{M_k\in\mathcal{M}}\Pr(M_k|\mathbf{Y}_n) = \left\{1+\frac{\sum_{M_j\notin\mathcal{M}}\Pr(\mathbf{Y}_n|M_j)}{\sum_{M_k\in\mathcal{M}}\Pr(\mathbf{Y}_n|M_k)}\right\}^{-1}. \tag{15}$$

Note that

$$\frac{\sum_{M_j\notin\mathcal{M}}\Pr(\mathbf{Y}_n|M_j)}{\sum_{M_k\in\mathcal{M}}\Pr(\mathbf{Y}_n|M_k)} \leqslant \frac{\sum_{M_j\notin\mathcal{M}}\Pr(\mathbf{Y}_n|M_j)}{\Pr(\mathbf{Y}_n|M_k)},$$

for $M_k\in\mathcal{M}$. Hence by the same argument as that leading to Lemma 5, we have

$$\frac{\sum_{M_j\notin\mathcal{M}}\Pr(\mathbf{Y}_n|M_j)}{\sum_{M_k\in\mathcal{M}}\Pr(\mathbf{Y}_n|M_k)} = O_p\{K_n a_{n_I}\exp(-n_I\delta^2)\},$$

and

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

This proves the result.

Proof of Proposition 1

Following [15], we define x as an n_I-flat point so that there is no change-point in $(x-n_I, x+n_I)$. Let $\mathcal{F}$ be the set of all n_I-flat points, then

$$\begin{aligned}
& \Pr\left\{\left(\bigcap_{t\in\mathcal{T}_0(p_0)} R_t > C\right)\bigcap\left(\bigcap_{\tau\in\mathcal{F}} R_\tau < C\right)\right\} \\
= \;& 1-\Pr\left\{\left(\bigcup_{\tau\in\mathcal{T}_0(p_0)} R_\tau > C\right)\bigcup\left(\bigcup_{\tau\in\mathcal{F}} R_\tau < C\right)\right\} \\
= \;& 1-\Pr\left\{\left(\max_{t\in\mathcal{T}_0(p_0)} R_t > C\right)\bigcup\left(\min_{\tau\in\mathcal{F}} R_\tau < C\right)\right\} \\
\geqslant \;& 1-\left\{\Pr\left(\max_{t\in\mathcal{T}_0(p_0)} R_t > C\right)+\Pr\left(\min_{\tau\in\mathcal{F}} R_\tau < C\right)\right\}.
\end{aligned}$$

For each $\tau \in \mathcal{T}_0(p_0)$,

$$\Pr(R_\tau < C) = O\{\exp(-n_I\delta^2)\},$$

which holds by Lemma 1. Furthermore, for $\tau \in \mathcal{F}$,

$$\Pr(R_\tau > C) = O(a_{n_I}),$$

which holds by Lemmas 2 to 4. Hence,

$$\begin{aligned}
&\Pr\left\{\left(\bigcap_{t\in\mathcal{T}_0(p_0)} R_t > C\right)\bigcap\left(\bigcap_{\tau\in\mathcal{F}} R_\tau < C\right)\right\}\\
\geqslant\quad & 1 - O[\min\{\exp(-n_I\delta^2), a_{n_I}\}]
\end{aligned}$$

By Lemma 3 in [15], for any $t \in \mathcal{T}_0(p_0)$ we have a $\tau \in \mathcal{H}_c(n_I)$ such that $\Pr\{t \in (\tau - n_I, \tau + n_I)\} = 1 - O[\min\{\exp(-n_I\delta^2), a_{n_I}\}]$.

Proof of Theorem 2

We first show that for a given p, $\widehat{\mathcal{T}}(p)$ is the maximizer of $\Pr\{\mathbf{Y}_n|\mathcal{T}(p)\}$. Based on the BMS procedure, $\widehat{\mathcal{T}}(p)$ is the maximizer of $\sum_{M_k\in\mathcal{M}}\Pr(M_k|\mathbf{Y}_n)$, where $\mathcal{M} = \{M_k, \tau_k \in \mathcal{T}(p)\}$. Since we impose the uniform prior on M_k, $\widehat{\mathcal{T}}(p)$ is the maximizer of

$$\begin{aligned}
&\sum_{M_k\in\mathcal{M}}\Pr(\mathbf{Y}_n|M_k) \qquad (16)\\
=\quad & D_n\sum_{M_k\in\mathcal{M}}\prod_{j=1,j\neq k}^{K}\prod_{l=\tau_j}^{\tau_{j+1}-1}\exp\{-(Y_l-\bar{Y}_{\tau_j})^2\}\int\prod_{l=\tau_k}^{\tau_{k+1}-1}\exp\{-(Y_l-\bar{Y}_{\tau_k}-\mu)^2\}\pi(\mu)d\mu\\
=\quad & D_n\prod_{j=1}^{K}\prod_{l=\tau_j}^{\tau_{j+1}-1}\exp\{-(Y_l-\bar{Y}_{\tau_j})^2\}\sum_{M_k\in\mathcal{M}}\frac{\int\prod_{l=\tau_k}^{\tau_{k+1}-1}\exp\{-(Y_l-\bar{Y}_{\tau_k}-\mu)^2\}\pi(\mu)d\mu}{\prod_{l=\tau_k}^{\tau_{k+1}-1}\exp\{-(Y_l-\bar{Y}_{\tau_k})^2\}}
\end{aligned}$$

where D_n is a constant depending on n. Further note that

$$\begin{aligned}
&\Pr\{\mathbf{Y}_n|\mathcal{T}(p)\} \qquad (17)\\
=\quad & \prod_{\tau_j\notin\mathcal{T}(p)}\prod_{l=\tau_j}^{\tau_{j+1}-1}\exp\{-(Y_l-\bar{Y}_{\tau_j})^2\}\prod_{\tau_k\in\mathcal{T}(p)}\int\prod_{l=\tau_k}^{\tau_{k+1}-1}\exp\{-(Y_l-\bar{Y}_{\tau_k}-\mu)^2\}\pi(\mu)d\mu\\
=\quad & \prod_{j=1}^{K}\prod_{l=\tau_j}^{\tau_{j+1}-1}\exp\{-(Y_l-\bar{Y}_{\tau_j})^2\}\prod_{\tau_k\in\mathcal{T}(p)}\frac{\int\prod_{l=\tau_k}^{\tau_{k+1}-1}\exp\{-(Y_l-\bar{Y}_{\tau_k}-\mu)^2\}\pi(\mu)d\mu}{\prod_{l=\tau_k}^{\tau_{k+1}-1}\exp\{-(Y_l-\bar{Y}_{\tau_k})^2\}}.
\end{aligned}$$

Comparing (17) and (16), clearly they have the same optimizer, and thus $\widehat{\mathcal{T}}(p)$ is the maximizer of (17). Hence, our BMS procedure results in the estimators $\hat{p}$ and $\widehat{\mathcal{T}}(\hat{p})$ that maximize $\Pr\{\mathbf{Y}_n|\mathcal{T}(p)\}$.

Next, let $\mathcal{E}_1$ be the event that at least one j such that $t_j \in (\tau_k, \tau_{k+1})$, and $\hat{t}_i \neq \tau_k$, $\hat{t}_i \neq \tau_{k+1}$ for all i and $\tau_k, \tau_{k+1} \in \mathcal{H}_c(n_I)$, $\hat{t}_i \in \widehat{\mathcal{T}}(\hat{p})$ that maximizes $\Pr\{\mathbf{Y}_n|\mathcal{T}(p)\}$. Following the similar arguments as those in [5], we show that the probability of $\mathcal{E}_1$ goes to 0. Suppose that $\widehat{\mathcal{T}}(\hat{p})$ is such an estimate. Consider the first case where $(t_j - \tau_k + 1)(\tau_{k+1} - \tau_k + 1)^{-1} = O(1)$; that is, t_j is bounded away from τ_k. Then, we can choose a set of change points that

$$\begin{aligned}
\widetilde{\mathcal{T}}(\hat{p}+1) &\equiv \{\tilde{\tau}_1, \ldots, \tilde{\tau}_{\hat{p}+1}\}\\
&= \{\hat{t}_1, \ldots, \hat{t}_i, \tau_{k+1}, \hat{t}_{i+1}, \hat{t}_{\hat{p}}\}.
\end{aligned}$$

Then

$$\frac{\Pr\{\mathbf{Y}_n|\widehat{\mathcal{T}}(\hat{p})\}}{\Pr\{\mathbf{Y}|\widetilde{\mathcal{T}}(\hat{p}+1)\}} = \frac{\prod_{l=\tau_{k+1}}^{\tau_{k+2}-1}\exp\{-(Y_l-\bar{Y}_{\tau_{k+1}})^2\}}{\int\prod_{l=\tau_{k+1}}^{\tau_{k+2}-1}\exp\{-(Y_l-\bar{Y}_{\tau_{k+1}}-\mu)^2\}\pi(\mu)d\mu}.$$

Because $\lim_{n\to\infty}\sup n_I/\lambda < 1/2$, there is an N, such that for all $n > N$, $t_{j+1} > \tau_{k+2}$, and hence there is no change point within (τ_{k+1}, τ_{k+2}). This prevents the situation where there are more than one change points in between τ_k and τ_{k+2}. Further for $n > N$,

$$
\begin{aligned}
E(Y_l - \bar{Y}_{\tau_{k+1}}) &= (\tau_{k+1} - \tau_k + 1)^{-1} E\left\{\sum_{s=\tau_k}^{t_j}(Y_l - Y_s) + \sum_{s=t_j+1}^{\tau_{k+1}}(Y_l - Y_s)\right\} \\
&\geqslant (t_j - \tau_k + 1)(\tau_{k+1} - \tau_k + 1)^{-1}\delta.
\end{aligned}
$$

Therefore, by Lemma 1,

$$
\frac{\Pr\{\mathbf{Y}_n|\widehat{\mathcal{T}}(\widehat{p})\}}{\Pr\{\mathbf{Y}|\widetilde{\mathcal{T}}(\widehat{p}+1)\}} = O_p\{\exp(-n_I\delta^2)\}.
$$

If $(t_j - \tau_k + 1)(\tau_{k+1} - \tau_k + 1)^{-1} = o(1)$, we define

$$
\begin{aligned}
\widetilde{\mathcal{T}}(\widehat{p}+1) &\equiv \{\widetilde{\tau}_1, \ldots, \widetilde{\tau}_{\widehat{p}+1}\} \\
&= \{\widehat{t}_1, \ldots, \widehat{t}_i, \tau_k, \widehat{t}_{i+1}, \widehat{t}_{\widehat{p}}\},
\end{aligned}
$$

and then

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

The first term in the last equation is of order $O_p\{\exp(-n_I\delta^2)\}$ by Lemma 1, and the last two terms are of order $O_p(1)$ because $(t_j - \tau_k + 1)(\tau_{k+1} - \tau_k + 1)^{-1} = o(1)$. Therefore,

$$
\Pr\left[\frac{\Pr\{\mathbf{Y}_n|\widehat{\mathcal{T}}(\widehat{p})\}}{\Pr\{\mathbf{Y}|\widetilde{\mathcal{T}}(\widehat{p}+1)\}} > 1|\mathcal{E}_1\right] \leqslant E\left[\frac{\Pr\{\mathbf{Y}_n|\widehat{\mathcal{T}}(\widehat{p})\}}{\Pr\{\mathbf{Y}|\widetilde{\mathcal{T}}(\widehat{p}+1)\}}\right] = O\{\exp(-n_I\delta^2)\}.
$$

As $\widehat{\mathcal{T}}(\widehat{p})$ is the maximizer of $\Pr\{\mathbf{Y}_n|\mathcal{T}(p)\}$, $\Pr\left[\frac{\Pr\{\mathbf{Y}_n|\widehat{\mathcal{T}}(\widehat{p})\}}{\Pr\{\mathbf{Y}|\widetilde{\mathcal{T}}(\widehat{p}+1)\}} > 1\right] = 1$ because $\widehat{\mathcal{T}}(\widehat{p})$ is the maximizer of $\Pr\{\mathbf{Y}_n|\widehat{\mathcal{T}}(\widehat{p})\}$ and it is unique by condition (A3). By the Bayes rule, we have

$$
\Pr\left[\mathcal{E}_1|\frac{\Pr\{\mathbf{Y}_n|\widehat{\mathcal{T}}(\widehat{p})\}}{\Pr\{\mathbf{Y}|\widetilde{\mathcal{T}}(\widehat{p}+1)\}} > 1\right] \leqslant O\{\exp(-n_I\delta^2)\}. \tag{18}
$$

Hence, $\Pr(\widehat{t}_i = \tau_k \text{ or } \widehat{t}_i = \tau_{k+1}) = 1 - O\{\exp(-n_I\delta^2)\}$. Further note that τ_k and τ_{k+1} are in the n_I-neighborhood of t_j, and thus for any t_j, there is a $\widehat{t}_i$ such that

$$
\Pr\{t_j \in (\widehat{t}_i - n_I, \widehat{t}_i + n_I)\} = 1 - O\{\exp(-n_I\delta^2)\}.
$$

Since it holds for any j, we can write

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

Next we show that for any $\widehat{t}_i$ there is a t_j in the n_I-neighborhood of $\widehat{t}_i$. Define $\mathcal{E}_2$ as the event that there is at least one $\widehat{t}_i$ such that no t_j in the n_I-neighborhood of $\widehat{t}_i$. Let $\widehat{\mathcal{T}}(\widehat{p})$ be such an estimate that

$\hat{t}_i$ is the kth candidate points and $(\hat{t}_i, \tau_{k+1})$, $(\tau_{k-1}, \hat{t}_i)$ do not contain t_j for all j. Then, we define a new set of change points by deleting $\hat{t}_i$,

$$\widetilde{\mathcal{T}}(\hat{p}-1) = \{\hat{t}_1, \ldots, \hat{t}_{i-1}, \hat{t}_{i+1}, \hat{t}_{\hat{p}}\}.$$

Then,

$$\frac{\Pr\{\mathbf{Y}_n|\widehat{\mathcal{T}}(\hat{p})\}}{\Pr\{\mathbf{Y}|\widetilde{\mathcal{T}}(\hat{p}-1)\}} = \frac{\prod_{l=\hat{t}_i}^{\tau_{k+1}-1} \exp\{-(Y_l - \bar{Y}_{\hat{t}_i} - \mu)^2\}\pi(\mu)d\mu}{\prod_{l=\hat{t}_i}^{\tau_{k+1}-1} \exp\{-(Y_l - \bar{Y}_{\hat{t}_i})^2\}} = O_p(a_{n_I})$$

by Lemmas 2–4. Therefore, using the same argument as that leading to (18), we have $\Pr(\mathcal{E}_2|\frac{\Pr\{\mathbf{Y}_n|\widehat{\mathcal{T}}(\hat{p})\}}{\Pr\{\mathbf{Y}|\widetilde{\mathcal{T}}(\hat{p}-1)\}} > 1) = O(a_{n_I})$. For any $\hat{t}_i$, there exists a t_j such that

$$\Pr\{\hat{t}_i \in (t_j - n_I, t_j + n_I)\} = 1 - O(a_{n_I}).$$

It holds for any $\hat{t}_i$, and thus we have

$$\Pr\left[\sup_{t_j \in \mathcal{T}_0(p_0)} \inf_{\hat{t}_j \in \widehat{\mathcal{T}}(\hat{p})} |(\hat{t}_j - t_j)/n| < n_I/n\right] = 1 - O(a_{n_I})$$

Because $(t_j - n_I, t_j + n_I)$ contains only one estimate by the definition of $\widehat{\mathcal{T}}(\hat{p})$ that $|\hat{t}_{j+1} - \hat{t}_j| > \lambda$, we have $\Pr(\hat{p} = p_0) = 1 - O_p(\max\{\exp(-n_I\delta^2), a_{n_I}\})$ by using the same arguments as those leading to Theorem 3.3 in [5]. $\square$