Ann Inst Stat Math (2015) 67:943–961
DOI 10.1007/s10463-014-0481-x

Change-point model selection via AIC

Yoshiyuki Ninomiya

Received: 21 April 2013 / Revised: 21 June 2014 / Published online: 24 August 2014

Abstract Change-point problems have been studied for a long time not only because they are needed in various fields but also because change-point models contain an irregularity that requires an alternative to conventional asymptotic theory. The purpose of this study is to derive the AIC for such change-point models. The penalty term of the AIC is twice the asymptotic bias of the maximum log-likelihood, whereas it is twice the number of parameters, $2p_0$, in regular models. In change-point models, it is not twice the number of parameters, $2m + 2p_m$, because of their irregularity, where m and p_m are the numbers of the change-points and the other parameters, respectively. In this study, the asymptotic bias is shown to become $6m + 2p_m$, which is simple enough to conduct an easy change-point model selection. Moreover, the validity of the AIC is demonstrated using simulation studies.

Keywords Brownian motion · Functional central limit theorem · Information criterion · Irregularity · Random walk · Structural change

1 Introduction

Model selection by a testing procedure in change-point problems, including the estimation of the number of change-points, has been studied for a long time (e.g., Vostrikova

This research was partially supported by a Grant-in-Aid for Scientific Research (20700252) from the Ministry of Education, Culture, Sports, Science and Technology of Japan. The author thanks Professor Jonathan Taylor and Professor David Siegmund for their insightful and helpful comments.

Y. Ninomiya (✉)
Institute of Mathematics for Industry, Kyushu University, 744 Moto-oka, Nishi-ku, Fukuoka 819-0395, Japan
e-mail: nino@imi.kyushu-u.ac.jp

1981; Haccou and Meelis 1988; Inclan and Tiao 1994; Bai and Perron 1998; Aue et al. 2009). This is not only because many important applications require such an approach but also because the change-point as a parameter has an irregularity. For example, conventional asymptotic theory does not hold for the change-point, and so a specific theory as an alternative is needed for change-point problems (see e.g., Csörgő and Horváth 1997).

In this study, we consider model selection by an information criterion for change-point problems. This topic originated in the studies of Yao (1988) and Jones and Dey (1995). They, respectively, proposed a naive Bayes information criterion (BIC, Schwarz 1978) and a naive Akaike's information criterion (AIC, Akaike 1973) by ignoring the irregularity of the change-points. That is, their respective penalty terms are $(m + p_m) \log n$ and $2m + 2p_m$, where m, p_m, and n are the number of the change-points, the number of the other parameters, and the data size, respectively. Other than these information criteria, information criteria obtained by omitting all penalties for change-points have also been used (e.g., Chen and Gupta 1997). The first known article proposing an information criterion considering the irregularity of change-points is that of Siegmund (2004). He treated a model used in mapping quantitative trait loci (QTL) for genetic linkage analysis, which is closely related to an independent Gaussian sequence with a change in mean, and derived a BIC from the Bayes factor of the model. The result was generalized by Zhang and Siegmund (2007) for an independent Gaussian sequence with multiple changes in mean. Hannart and Naveau (2012) also derived a BIC-type criterion for a Bayesian change-point model from its Bayes factor. In such BICs, the penalty terms are different than conventional penalty terms owing to the irregularity of the models.

This study aims to derive the AIC for general change-point models based on the original definition of the AIC to consider the irregularity of change-points. The models are the ones usually adopted in change-point problems and are particularly different than the model in Siegmund (2004) or Hannart and Naveau (2012) and generalizations of the model in Zhang and Siegmund (2007). The penalty term of the AIC is twice the asymptotic bias of the maximum log-likelihood from the expected log-likelihood, whereas this is $2p_0$ in regular models, where p_0 is the number of parameters. In Sect. 2, we show that the penalty term for the change-point model depends on the expected value of the maximum of a random walk with a negative drift. Furthermore, we show that the penalty becomes $6m + 2p_m$ (not $2m + 2p_m$) under the condition considered by Csörgő and Horváth (1997) (Sect. 1.5). In Sect. 3, we demonstrate the validity of the AIC using a simulation study.

2 Main results

2.1 Independent sequence

For an independent multivariate sequence $\{\boldsymbol{x}_i,\ 1 \leq i \leq n\}$, let us consider a change-point model with m change-points $k^{(1)}, \ldots, k^{(m)}$ whose distribution belongs to a parametric family. For simplicity, we consider the exponential family as the parametric family like in Csörgő and Horváth (1997) (Sect. 1.5), that is, the probability function

of $\boldsymbol{x}_i$ for this model is

$$\exp\{\boldsymbol{\theta}^{(j)\mathrm{T}}\boldsymbol{T}(\cdot) + S(\cdot) - A(\boldsymbol{\theta}^{(j)})\} \quad \text{when} \quad k^{(j-1)} + 1 \le i \le k^{(j)} \tag{1}$$

for $1 \le j \le m+1$, where $k^{(0)} = 0$ and $k^{(m+1)} = n$. Let $\boldsymbol{\theta}^* = (\boldsymbol{\theta}^{*(1)\mathrm{T}}, \ldots, \boldsymbol{\theta}^{*(m+1)\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{k}^* = (k^{*(1)}, \ldots, k^{*(m)})^{\mathrm{T}}$ be the true values of $\boldsymbol{\theta} = (\boldsymbol{\theta}^{(1)\mathrm{T}}, \ldots, \boldsymbol{\theta}^{(m+1)\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{k} = (k^{(1)}, \ldots, k^{(m)})^{\mathrm{T}}$, and hereafter we similarly use * to denote the true values of parameters . We assume that

$$\boldsymbol{\theta}^{*(1)} \ne \boldsymbol{\theta}^{*(2)} \ne \cdots \ne \boldsymbol{\theta}^{*(m+1)}, \tag{2}$$

and that $\boldsymbol{\theta}^*$ and $\boldsymbol{k}^*$ are unknown. In addition, we assume that $\boldsymbol{\theta}^{*(1)}, \ldots, \boldsymbol{\theta}^{*(m+1)}$ are in the interior of the parameter set for the model, which is included in the natural parameter space for the family, and that $\partial^2 A(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathrm{T}}$ is strictly positive-definite in the parameter set. These assumptions ensure the asymptotic normality for the maximum likelihood estimator of $\boldsymbol{\theta}$ (see e.g., van der Vaart 1998, Sect. 4.2). For the purpose of the later asymptotic theory, we assume that $\lim_{n\to\infty} k^{*(j)}/n = \kappa^{(j)}$ for $1 \le j \le m$, where $0 < \kappa^{(1)} < \cdots < \kappa^{(m)} < 1$. This assumption means that the change-points are far enough from each other for a large sample size, and this is a common assumption in change-point analysis (see e.g., Csörgő and Horváth 1997).

Let $\hat{\boldsymbol{k}}_{\boldsymbol{x}}$ and $\hat{\boldsymbol{\theta}}_{\boldsymbol{x}}$ be the maximum likelihood estimators of $\boldsymbol{k}^*$ and $\boldsymbol{\theta}^*$ based on $\boldsymbol{x} = (\boldsymbol{x}_1^T, \ldots, \boldsymbol{x}_n^T)^T$, and $f(\boldsymbol{x}|\boldsymbol{k}^*, \boldsymbol{\theta}^*)$ be the joint probability function of $\boldsymbol{x}$. Model selection can be approached by trying to reduce twice the Kullback–Leibler divergence (Kullback and Leibler 1951) of $f(\boldsymbol{y}|\boldsymbol{k}^*, \boldsymbol{\theta}^*)$ and $f(\boldsymbol{y}|\hat{\boldsymbol{k}}_{\boldsymbol{x}}, \hat{\boldsymbol{\theta}}_{\boldsymbol{x}})$,

$$2\mathrm{KL}\{f(\boldsymbol{y}|\boldsymbol{k}^*, \boldsymbol{\theta}^*), f(\boldsymbol{y}|\hat{\boldsymbol{k}}_{\boldsymbol{x}}, \hat{\boldsymbol{\theta}}_{\boldsymbol{x}})\} = 2E_{\boldsymbol{y}}\{\log f(\boldsymbol{y}|\boldsymbol{k}^*, \boldsymbol{\theta}^*)\} - 2E_{\boldsymbol{y}}\{\log f(\boldsymbol{y}|\hat{\boldsymbol{k}}_{\boldsymbol{x}}, \hat{\boldsymbol{\theta}}_{\boldsymbol{x}})\},$$

where $\boldsymbol{y}$ is a copy of $\boldsymbol{x}$, in other words, $\boldsymbol{y}$ is distributed according to the distribution of $\boldsymbol{x}$ and is independent of $\boldsymbol{x}$, and $E_{\boldsymbol{y}}$ denotes the expectation with respect to $\boldsymbol{y}$. Because the first term on the right-hand side does not depend on the model, we need only consider the second term. A simple estimator of the second term is $-2\log f(\boldsymbol{x}|\hat{\boldsymbol{k}}_{\boldsymbol{x}}, \hat{\boldsymbol{\theta}}_{\boldsymbol{x}})$, but it is an underestimator. Then, in AIC-type information criteria, minimization of its bias correction is considered,

$$\begin{aligned}
&-2\log f(\boldsymbol{x}|\hat{\boldsymbol{k}}_{\boldsymbol{x}}, \hat{\boldsymbol{\theta}}_{\boldsymbol{x}}) + 2E_{\boldsymbol{x}}[\log f(\boldsymbol{x}|\hat{\boldsymbol{k}}_{\boldsymbol{x}}, \hat{\boldsymbol{\theta}}_{\boldsymbol{x}}) - E_{\boldsymbol{y}}\{\log f(\boldsymbol{y}|\hat{\boldsymbol{k}}_{\boldsymbol{x}}, \hat{\boldsymbol{\theta}}_{\boldsymbol{x}})\}] \\
&= -2\log f(\boldsymbol{x}|\hat{\boldsymbol{k}}_{\boldsymbol{x}}, \hat{\boldsymbol{\theta}}_{\boldsymbol{x}}) + 2E\{\log f(\boldsymbol{x}|\hat{\boldsymbol{k}}_{\boldsymbol{x}}, \hat{\boldsymbol{\theta}}_{\boldsymbol{x}}) - \log f(\boldsymbol{y}|\hat{\boldsymbol{k}}_{\boldsymbol{x}}, \hat{\boldsymbol{\theta}}_{\boldsymbol{x}})\} \\
&= -2\log f(\boldsymbol{x}|\hat{\boldsymbol{k}}_{\boldsymbol{x}}, \hat{\boldsymbol{\theta}}_{\boldsymbol{x}}) + 2E\{\log f(\boldsymbol{x}|\hat{\boldsymbol{k}}_{\boldsymbol{x}}, \hat{\boldsymbol{\theta}}_{\boldsymbol{x}}) - \log f(\boldsymbol{x}|\hat{\boldsymbol{k}}_{\boldsymbol{y}}, \hat{\boldsymbol{\theta}}_{\boldsymbol{y}})\} \\
&= -2\log f(\boldsymbol{x}|\hat{\boldsymbol{k}}_{\boldsymbol{x}}, \hat{\boldsymbol{\theta}}_{\boldsymbol{x}}) + 2E[\sup_{(\boldsymbol{k},\boldsymbol{\theta})} L_{\boldsymbol{x}}(\boldsymbol{k}, \boldsymbol{\theta}) - L_{\boldsymbol{x}}\{\mathrm{argsup}_{(\boldsymbol{k},\boldsymbol{\theta})} L_{\boldsymbol{y}}(\boldsymbol{k}, \boldsymbol{\theta})\}],
\end{aligned} \tag{3}$$

where E denotes the expectation with respect to both $\boldsymbol{x}$ and $\boldsymbol{y}$, and $L_{\boldsymbol{x}}(\boldsymbol{k}, \boldsymbol{\theta}) = \log f(\boldsymbol{x}|\boldsymbol{k}, \boldsymbol{\theta}) - \log f(\boldsymbol{x}|\boldsymbol{k}^*, \boldsymbol{\theta}^*)$. However, the expectation in (3) cannot be obtained explicitly, and so we use its asymptotic evaluation in the same way as done for the

AIC, that is,

$$-2\log f(\boldsymbol{x}|\hat{\boldsymbol{k}}_{\boldsymbol{x}}, \hat{\boldsymbol{\theta}}_{\boldsymbol{x}}) + 2E\{b(\boldsymbol{k}^*, \boldsymbol{\theta}^*)\} \tag{4}$$

is considered in place of (3), where $b(\boldsymbol{k}^*, \boldsymbol{\theta}^*)$ is the limit to which $\sup_{(\boldsymbol{k},\boldsymbol{\theta})} L_{\boldsymbol{x}}(\boldsymbol{k}, \boldsymbol{\theta}) - L_{\boldsymbol{x}}\{\mathrm{argsup}_{(\boldsymbol{k},\boldsymbol{\theta})} L_{\boldsymbol{y}}(\boldsymbol{k}, \boldsymbol{\theta})\}$ converges in distribution, and then we call $E\{b(\boldsymbol{k}^*, \boldsymbol{\theta}^*)\}$ an asymptotic bias of the maximum log-likelihood. Here, we take $\sup_{(\boldsymbol{k},\boldsymbol{\theta})}$ and $\mathrm{argsup}_{(\boldsymbol{k},\boldsymbol{\theta})}$ in a set of $(\boldsymbol{k}, \boldsymbol{\theta})$ such that $L_{\boldsymbol{x}}(\boldsymbol{k}, \boldsymbol{\theta})$ is $O_{\mathrm{P}}(1)$ or positive. If there is no change-point, $E\{b(\boldsymbol{k}^*, \boldsymbol{\theta}^*)\} = E\{b(\boldsymbol{\theta}^*)\}$ becomes the number of different parameters in $\boldsymbol{\theta}$.

Let us define vector $A'(\boldsymbol{\theta}^{*(j)})$ to be $\partial A(\boldsymbol{\theta}^{(j)})/\partial \boldsymbol{\theta}^{(j)}|_{\boldsymbol{\theta}^{(j)}=\boldsymbol{\theta}^{*(j)}}$, $B_1^{(j)}(\boldsymbol{\theta}^*) = A(\boldsymbol{\theta}^{*(j+1)}) - A(\boldsymbol{\theta}^{*(j)}) - (\boldsymbol{\theta}^{*(j+1)} - \boldsymbol{\theta}^{*(j)})^{\mathrm{T}} A'(\boldsymbol{\theta}^{*(j)})$, $B_2^{(j)}(\boldsymbol{\theta}^*) = A(\boldsymbol{\theta}^{*(j)}) - A(\boldsymbol{\theta}^{*(j+1)}) - (\boldsymbol{\theta}^{*(j)} - \boldsymbol{\theta}^{*(j+1)})^{\mathrm{T}} A'(\boldsymbol{\theta}^{*(j+1)})$, and let $Q_{k,\boldsymbol{x}}^{(j)}$ be

$$I_{\{k<k^{*(j)}\}} \sum_{i=k+1}^{k^{*(j)}} [(\boldsymbol{\theta}^{*(j+1)} - \boldsymbol{\theta}^{*(j)})^{\mathrm{T}} \{\boldsymbol{T}(\boldsymbol{x}_i) - A'(\boldsymbol{\theta}^{*(j)})\} - B_1^{(j)}(\boldsymbol{\theta}^*)]$$

$$+ I_{\{k>k^{*(j)}\}} \sum_{i=k^{*(j)}+1}^{k} [(\boldsymbol{\theta}^{*(j)} - \boldsymbol{\theta}^{*(j+1)})^{\mathrm{T}} \{\boldsymbol{T}(\boldsymbol{x}_i) - A'(\boldsymbol{\theta}^{*(j+1)})\} - B_2^{(j)}(\boldsymbol{\theta}^*)].$$

Note that $B_1^{(j)} > 0$ and $B_2^{(j)} > 0$ because of the convexity of A, and so $Q_{k,\boldsymbol{x}}^{(j)}$ is a two-sided random walk with a negative drift and origin $k^{*(j)}$. We can then obtain the following theorem, whose derivation may be found in Appendix A.

Theorem 1 *Suppose that $\boldsymbol{x}$ is distributed according to the probability function (1) and that conditions (2) are satisfied. Then, the asymptotic bias in (4) is given by*

$$E\{b(\boldsymbol{k}^*, \boldsymbol{\theta}^*)\} = \sum_{j=1}^{m} E(sup_k Q_{k,\boldsymbol{x}}^{(j)} + Q^{(j)}_{argsup_k Q_{k,\boldsymbol{y}}^{(j)},\boldsymbol{x}}) + p_m, \tag{5}$$

where p_m is the number of different parameters in $\boldsymbol{\theta}$.

We can regard $E(\sup_k Q_{k,\boldsymbol{x}}^{(j)} + Q^{(j)}_{\mathrm{argsup}_k Q_{k,\boldsymbol{y}}^{(j)},\boldsymbol{x}})$ $(1 \le j \le m)$ and p_m as the biases for the change-point $k^{(j)}$ $(1 \le j \le m)$ and the other parameters $\boldsymbol{\theta}$, respectively. Because an advantage of using an information criterion is the ease of its execution in comparison with testing procedures, it is important to evaluate the asymptotic bias (5) explicitly. In a similar approach to that by Csörgő and Horváth (1997) (Sect. 1.5), we consider the condition

$$\boldsymbol{\theta}^{*(j+1)} - \boldsymbol{\theta}^{*(j)} = \alpha_n^{-1/2} \Delta_{\boldsymbol{\theta}^*}^{(j)} \quad \text{and} \quad O(1) \ne \alpha_n = o(n) \tag{6}$$

for $1 \le j \le m$ to obtain the following theorem, where $\Delta_{\boldsymbol{\theta}^*}^{(j)}$ is a constant vector. Under this type of condition, the asymptotic behavior of the change-point estimator changes

(see e.g., Dümbgen 1991), and we obtain the following theorem in place of Theorem 1. The derivation of this theorem may be found in Appendix B.

Theorem 2 *Under the condition* (6)*, the asymptotic bias in* (4) *is given by*

$$E\{b(\boldsymbol{k}^*, \boldsymbol{\theta}^*)\} = 3m + p_m.$$

Thus, the AIC for the change-point model is given by

$$\mathrm{AIC} = -2 \log f(\boldsymbol{x}|\hat{\boldsymbol{k}}_{\boldsymbol{x}}, \hat{\boldsymbol{\theta}}_{\boldsymbol{x}}) + 6m + 2p_m. \tag{7}$$

We can see that the penalty for each change-point is three times the penalty for each of the other parameters.

Remark 1 In regular models, the asymptotic bias of the maximum log-likelihood under a condition such as (6) is the same as the one without the condition. On the other hand, in change-point models, as can be seen from Theorems 1 and 2, the bias under the condition (6) is different than the one without the condition, and so it is needed for obtaining the AIC to decide whether we assume the condition. Here, we adopt to assume it because the condition provides an easy and explicit asymptotic bias, which is important when an information criterion is constructed, as mentioned before. A more significant reason is as follows. For the case when a structural change is clearly present, the maximum log-likelihood for the model with the change is clearly larger than that without the change, and so the model with the change will be selected whether or not we use the asymptotic bias under the condition. Meanwhile, an accurate evaluation is needed for the case when the structural change is not so large. In the light of this notion, we assume the condition to evaluate the bias for the change-point model close to the no change-point model.

Remark 2 Ninomiya (2005) derived the same AIC only for an independent Gaussian sequence with changes in mean. However, the bias evaluation in Ninomiya (2005) is only an approximation, and so we can say that the above theorems justify it theoretically. If we use the same approximation method as in Ninomiya (2005), which is not theoretically justified and does not require conditions such as (6), the evaluated bias can be obtained as in the following remark. The evaluated biases vary by the underlying distribution of the sequence and the type of the changing parameter.

Remark 3 Let us consider approximating the random walk by a Brownian motion with the same mean and variance as those of the random walk, that is, we replace $Q_k^{(j)}$ with $V_{k-k^{*(j)}}(c_1^{(j)}, c_2^{(j)}, \sigma_1^{(j)}, \sigma_2^{(j)})$. Here,

$$V_s(c_1, c_2, \sigma_1, \sigma_2) = \begin{cases} -c_1|s| + \sigma_1 W_s & (s \le 0) \\ -c_2|s| + \sigma_2 W_s & (k > 0), \end{cases} \tag{8}$$

$c_1^{(j)} = B_1^{(j)}(\boldsymbol{\theta}^*)$, $c_2^{(j)} = B_2^{(j)}(\boldsymbol{\theta}^*)$, $\sigma_1^{(j)} = \{(\boldsymbol{\theta}^{*(j)} - \boldsymbol{\theta}^{*(j+1)})^{\mathrm{T}} A''(\boldsymbol{\theta}^{*(j)})(\boldsymbol{\theta}^{*(j)} - \boldsymbol{\theta}^{*(j+1)})\}^{1/2}$ and $\sigma_2^{(j)} = \{(\boldsymbol{\theta}^{*(j)} - \boldsymbol{\theta}^{*(j+1)})^{\mathrm{T}} A''(\boldsymbol{\theta}^{*(j+1)})(\boldsymbol{\theta}^{*(j)} - \boldsymbol{\theta}^{*(j+1)})\}^{1/2}$, where

$\{W_s\}_{s\in\boldsymbol{R}}$ is a two-sided standard Brownian motion with $E(W_s) = 0$ and $\mathrm{Var}(W_s) = |s|$, and $A''(\boldsymbol{\theta}^{*(j)}) = \partial^2 A(\boldsymbol{\theta}^{(j)})/\partial\boldsymbol{\theta}^{(j)}\partial\boldsymbol{\theta}^{(j)\mathrm{T}}|_{\boldsymbol{\theta}^{(j)}=\boldsymbol{\theta}^{*(j)}}$. Then, we can evaluate (5) as follows:

$$E\{b(\boldsymbol{k}^*, \boldsymbol{\theta}^*)\} \approx \sum_{j=1}^{m} \frac{c_1^{(j)2}\sigma_2^{(j)4} + c_1^{(j)}c_2^{(j)}\sigma_1^{(j)2}\sigma_2^{(j)2} + c_2^{(j)2}\sigma_1^{(j)4}}{2c_1^{(j)}c_2^{(j)}(c_1^{(j)}\sigma_2^{(j)2} + c_2^{(j)}\sigma_1^{(j)2})} + p_m.$$

The derivation can be easily obtained from the proof of Theorem 2. This result is applicable to model selection by substituting some estimators of $c_1^{(j)}$, $c_2^{(j)}$, $\sigma_1^{(j)}$, and $\sigma_2^{(j)}$ for them.

Remark 4 Here, we consider the exponential family as the parametric family for simplicity, but the AIC in (7) can be extended to a more general parametric family under some regularity conditions satisfying the asymptotic normality of $\hat{\boldsymbol{\theta}}_{\boldsymbol{k}^*,\boldsymbol{x}}$ and $\hat{\boldsymbol{\theta}}_{\boldsymbol{k},\boldsymbol{x}} - \hat{\boldsymbol{\theta}}_{\boldsymbol{k}^*,\boldsymbol{x}} = O_{\mathrm{P}}(||\boldsymbol{k} - \boldsymbol{k}^*||/n)$, where $\hat{\boldsymbol{\theta}}_{\boldsymbol{k},\boldsymbol{x}}$ is the maximum likelihood estimator of $\boldsymbol{\theta}^*$ based on $\boldsymbol{x}$ when change-points are $\boldsymbol{k}$. Actually, defining $Q_{k,\boldsymbol{x}}^{(j)}$ by $I_{\{k<k^{*(j)}\}}\sum_{i=k+1}^{k^{*(j)}}\{g_i(\boldsymbol{\theta}^{*(j+1)}) - g_i(\boldsymbol{\theta}^{*(j)})\} + I_{\{k>k^{*(j)}\}}\sum_{i=k^{*(j)}+1}^{k}\{g_i(\boldsymbol{\theta}^{*(j)}) - g_i(\boldsymbol{\theta}^{*(j+1)})\}$ in this case, we can obtain Theorem 1, where $g_i(\cdot)$ is the log-likelihood function for $\boldsymbol{x}_i$. Under the condition (6), $2\alpha_n$ times the expectations of $g_i(\boldsymbol{\theta}^{*(j+1)}) - g_i(\boldsymbol{\theta}^{*(j)})$ for $i \in [k^{*(j-1)} + 1, k^{*(j)}]$ and $g_i(\boldsymbol{\theta}^{*(j)}) - g_i(\boldsymbol{\theta}^{*(j+1)})$ for $i \in [k^{*(j)} + 1, k^{*(j+1)}]$ converge to $\Delta_{\boldsymbol{\theta}^*}^{(j)T} J(\boldsymbol{\theta}^{*(j)})\Delta_{\boldsymbol{\theta}^*}^{(j)}$, and α_n times their variances also converge to $\Delta_{\boldsymbol{\theta}^*}^{(j)T} J(\boldsymbol{\theta}^{*(j)})\Delta_{\boldsymbol{\theta}^*}^{(j)}$, where $J(\boldsymbol{\theta}^{*(j)})$ is the Fisher information matrix at $\boldsymbol{\theta}^{*(j)}$. From this, letting $\sigma^{(j)} = \{\Delta_{\boldsymbol{\theta}^*}^{(j)T} J(\boldsymbol{\theta}^{*(j)})\Delta_{\boldsymbol{\theta}^*}^{(j)}\}^{1/2}$, $\sup_k Q_{k,\boldsymbol{x}}^{(j)}$ and $Q^{(j)}_{\mathrm{argsup}_k Q_{k,\boldsymbol{y}}^{(j)},\boldsymbol{x}}$ converge to $\sup_s V_s(\sigma^{(j)2}/2, \sigma^{(j)2}/2, \sigma^{(j)}, \sigma^{(j)})$ and $V_{\mathrm{argsup}_s V_s'(\sigma^{(j)2}/2,\sigma^{(j)2}/2,\sigma^{(j)},\sigma^{(j)})}(\sigma^{(j)2}/2, \sigma^{(j)2}/2, \sigma^{(j)}, \sigma^{(j)})$, respectively, where V_s is the random process defined in (8) and V_s' is its copy. Then, we can obtain Theorem 2 also in this case.

2.2 Auto-regressive sequence

To investigate whether the results of the previous section hold for dependent sequences, we consider the example of a change-point model in an auto-regressive sequence. We define this sequence by $\{x_i,\ 1 \le i \le n\}$ satisfying

$$x_i = \boldsymbol{\theta}^{(j)T}\boldsymbol{z}_i + \epsilon_i \quad \text{when} \quad k^{(j-1)} + 1 \le i \le k^{(j)}, \tag{9}$$

for $1 \le j \le m+1$, where $\boldsymbol{\theta}^{(j)} = (\theta_1^{(j)}, \ldots, \theta_p^{(j)})^T$, $\boldsymbol{z}_i = (x_{i-1}, \ldots, x_{i-p})^T$, $k^{(0)} = 0$, and $k^{(m+1)} = n$. We assume that $\{\epsilon_i,\ 1 \le i \le n\}$ is an independent Gaussian sequence with mean 0 and unknown variance θ_0, and that $1 + \sum_{h=1}^{p}\theta_h^{(j)}a^h \ne 0$ for all $a \in \boldsymbol{C}$ such that $|a| < 1$. Letting $\boldsymbol{\theta} = (\theta_0, \boldsymbol{\theta}^{(1)T}, \ldots, \boldsymbol{\theta}^{(m+1)T})^T$ and denoting its true value by $\boldsymbol{\theta}^* = (\theta_0^*, \boldsymbol{\theta}^{*(1)T}, \ldots, \boldsymbol{\theta}^{*(m+1)T})^T$, we can obtain the following corollary, whose derivation may be found in Appendix C.

Corollary 1 *For model (9) satisfying conditions (2), Theorem 1 holds using* $Q_{k,\boldsymbol{x}}^{AR(j)}$ *defined by*

$$I_{\{k<k^{*(j)}\}} \sum_{i=k+1}^{k^{*(j)}} [(\boldsymbol{\theta}^{*(j+1)} - \boldsymbol{\theta}^{*(j)})^T \boldsymbol{z}_i \epsilon_i - \{(\boldsymbol{\theta}^{*(j+1)} - \boldsymbol{\theta}^{*(j)})^T \boldsymbol{z}_i\}^2/2]/\theta_0^* + I_{\{k>k^{*(j)}\}} \sum_{i=k^{*(j)}+1}^{k} [(\boldsymbol{\theta}^{*(j)} - \boldsymbol{\theta}^{*(j+1)})^T \boldsymbol{z}_i \epsilon_i - \{(\boldsymbol{\theta}^{*(j)} - \boldsymbol{\theta}^{*(j+1)})^T \boldsymbol{z}_i\}^2/2]/\theta_0^*$$

in place of $Q_{k,\boldsymbol{x}}^{(j)}$. *In addition, Theorem 2 holds under the condition (6).*

By combining this result with that of the previous subsection, we can treat an auto-regressive sequence in which both the auto-regressive coefficients $\boldsymbol{\theta}^{(j)}$ and the variance θ_0 change.

3 Simulation study

We investigated the performance of the AIC in (7) by conducting simulations. For comparison, we also considered the criterion proposed by Jones and Dey (1995), which has been used in some applications (e.g., Hurrell and Trenberth 1997). This criterion uses $2m + 2p_m$ as the penalty term without considering the irregularity of the change-points. We denote this naive criterion by $\mathrm{AIC}_{\mathrm{naive}}$ to distinguish it from the AIC in (7).

Consider two simple models: an independent Gaussian sequence with one change in variance and an auto-regressive sequence with one change in coefficient, such that

$$x_i = \begin{cases} \mu + \sigma^{(1)}\epsilon_i \\ \mu + \sigma^{(2)}\epsilon_i \end{cases} \quad \text{and} \quad x_i = \begin{cases} a^{(1)}x_{i-1} + \sigma\epsilon_i & (1 \le i \le k) \\ a^{(2)}x_{i-1} + \sigma\epsilon_i & (k+1 \le i \le n) \end{cases}, \tag{10}$$

where $x_0 = 0$ and $\{\epsilon_i,\ 1 \le i \le n\}$ is an independent Gaussian sequence with mean 0 and variance 1. The penalty terms of the AIC in (7) and $\mathrm{AIC}_{\mathrm{naive}}$ become $6 \times 1 + 2 \times 3 = 12$ and $2 \times 1 + 2 \times 3 = 8$, respectively, because there are one change-point and three regular parameters. To investigate whether these penalty terms provide a sufficiently accurate approximation for the bias of twice the maximum log-likelihood, we evaluate them numerically for several sets of true values of parameters and data sizes of the above model. The results are given in Table 1. We can see that these values lie close to 12, and are certainly much closer to 12 than to 8.

To understand the influence of model misspecification, we evaluate the bias under a misspecified model. Let us consider the first model in (10). As the distribution of ϵ_i, we use a mixture distribution whose components are $N(0, 1)$ and $U(-\sqrt{3}, \sqrt{3})$ as the true one, while $N(0, 1)$ is assumed in the model, where we denote the continuous uniform distribution on $[a, b]$ by $U(a, b)$. The results are given in Table 2. In a part of the table, we set $\sigma^{*(1)} = \sigma^{*(2)}$, which means that not only the distribution but also the

Table 1 Bias of twice the maximum log-likelihood

			n: 50	n: 100	n: 200	n: 400
First model (μ^*: 0, k^*: $n/2$)	$\sigma^{*(1)}$: 0.95	$\sigma^{*(2)}$: 1.05	12.79	11.67	11.33	11.23
	$\sigma^{*(1)}$: 0.9	$\sigma^{*(2)}$: 1.1	12.80	11.66	11.65	11.84
	$\sigma^{*(1)}$: 0.85	$\sigma^{*(2)}$: 1.15	13.10	12.00	11.95	12.10
	$\sigma^{*(1)}$: 0.8	$\sigma^{*(2)}$: 1.2	13.10	12.69	12.46	12.34
	$\sigma^{*(1)}$: 0.75	$\sigma^{*(2)}$: 1.25	13.28	12.77	12.67	12.49
Second model (σ^*: 1, k^*: $n/2$)	$a^{*(1)}$: 0	$a^{*(2)}$: 0.2	11.08	11.77	12.42	13.03
	$a^{*(1)}$: 0	$a^{*(2)}$: 0.4	11.36	11.83	12.25	12.30
	$a^{*(1)}$: 0	$a^{*(2)}$: 0.6	12.01	12.25	12.43	12.37
	$a^{*(1)}$: 0.2	$a^{*(2)}$: 0.4	11.36	11.95	12.66	13.04
	$a^{*(1)}$: 0.2	$a^{*(2)}$: 0.6	12.00	12.16	12.37	12.24
	$a^{*(1)}$: 0.2	$a^{*(2)}$: 0.8	13.18	13.37	13.55	13.41

These values are obtained by Monte Carlo simulation with 10,000 repetitions using the models (10), that is, $\sum_{h=1}^{10,000}[\sup_{(\boldsymbol{k},\boldsymbol{\theta})} L_{\boldsymbol{x}^{[2h-1]}}(\boldsymbol{k},\boldsymbol{\theta}) - L_{\boldsymbol{x}^{[2h-1]}}\{\mathrm{argsup}_{(\boldsymbol{k},\boldsymbol{\theta})} L_{\boldsymbol{x}^{[2h]}}(\boldsymbol{k},\boldsymbol{\theta})\}]/10{,}000$, where $\{\boldsymbol{x}^{[h]} = (x_1^{[h]},\ldots,x_n^{[h]}) \mid 1 \le h \le 20{,}000\}$ are 20,000 independent sets of random samples generated using the models (10)

Table 2 Bias of twice the maximum log-likelihood under model misspecification

			n: 50	n: 100	n: 200	n: 400
$\sigma^{*(1)}$: 1.0	$\sigma^{*(2)}$: 1.0	ρ: 0.1	12.06	11.06	10.75	10.71
		ρ: 0.3	11.08	9.95	9.61	9.56
		ρ: 0.5	9.71	8.75	8.52	8.47
$\sigma^{*(1)}$: 0.9	$\sigma^{*(2)}$: 1.1	ρ: 0.1	12.11	11.16	11.07	11.32
		ρ: 0.3	11.09	10.00	9.87	10.13
		ρ: 0.5	9.78	8.94	8.79	9.02
$\sigma^{*(1)}$: 0.8	$\sigma^{*(2)}$: 1.2	ρ: 0.1	12.13	11.74	11.75	11.78
		ρ: 0.3	11.36	10.58	10.55	10.35
		ρ: 0.5	9.94	9.44	9.31	9.17

These values are obtained by Monte Carlo simulation with 10,000 repetitions using the first model in (10) in the same way as in Table 1. As the distribution of ϵ_i in (10), a mixture distribution whose components are $N(0, 1)$ with weight $1 - \rho$ and $U(-\sqrt{3}, \sqrt{3})$ with weight ρ is used as the true one, while $N(0, 1)$ is assumed in the model

number of change-points are misspecified in the model. The true bias becomes smaller in comparison with our evaluation as the model becomes more misspecified; however, we can say that the influence is negligible if the model is not heavily misspecified even when the number of change-points is misspecified.

To check the performances of the AIC in (7) and $\mathrm{AIC}_{\mathrm{naive}}$ simply, we evaluate the rate of selecting one change against no changes by the AIC and $\mathrm{AIC}_{\mathrm{naive}}$ for the data simulated according to the models (10). The results are given in Table 3. First, let us

Table 3 Rate of selecting the one change-point model versus the no change-point model by the $\mathrm{AIC}_{\mathrm{naive}}$ and AIC in (7)

			n: 50 (%)	n: 100 (%)	n: 200 (%)	n: 400 (%)
First model (μ^*: 0, k^*: $n/2$)	$\sigma^{*(1)}$: 0.9	$\mathrm{AIC}_{\mathrm{naive}}$	59.8	73.6	86.8	96.6
	$\sigma^{*(2)}$: 1.1	AIC	15.9	25.8	45.7	73.6
	$\sigma^{*(1)}$: 0.8	$\mathrm{AIC}_{\mathrm{naive}}$	80.7	95.1	99.7	100.0
	$\sigma^{*(2)}$: 1.2	AIC	39.3	70.8	95.2	100.0
	$\sigma^{*(1)}$: 1	$\mathrm{AIC}_{\mathrm{naive}}$	49.2	56.7	62.8	69.0
	$\sigma^{*(2)}$: 1	AIC	9.3	11.5	14.0	15.9
Second model (σ^*: 1, k^*: $n/2$)	$a^{*(1)}$: 0	$\mathrm{AIC}_{\mathrm{naive}}$	51.1	75.6	93.8	99.6
	$a^{*(2)}$: 0.4	AIC	14.3	36.5	69.7	95.2
	$a^{*(1)}$: 0.2	$\mathrm{AIC}_{\mathrm{naive}}$	52.5	78.2	94.9	99.8
	$a^{*(2)}$: 0.6	AIC	16.1	39.8	74.3	97.2
	$a^{*(1)}$: 0.3	$\mathrm{AIC}_{\mathrm{naive}}$	27.6	37.1	44.8	52.3
	$a^{*(2)}$: 0.3	AIC	3.2	5.0	6.8	8.6

These values are obtained by Monte Carlo simulations with 10,000 repetitions using the models (10)

consider the values for the one change-point model when the data size is small, which will be close to the no change-point model. In this case, we cannot say whether it is better to select no changes or one change from the viewpoint of the original purpose of the AIC because the original purpose is not to identify the true model, but to provide a distribution close to the true one. Note that, we will show the clear superiority of the AIC in (7) from this viewpoint in next simulation study. Next, let us consider the values for the one change-point model when the data size is large, which will be far from the no change-point model. In this case, one change may be desirable, and we can see that both the AIC and $\mathrm{AIC}_{\mathrm{naive}}$ can select one change with a high probability. Finally, let us see the values for the no change-point model. In this case, we definitely want to select no change. However, the values for the $\mathrm{AIC}_{\mathrm{naive}}$ are too large.

Let us compare the AIC in (7) and $\mathrm{AIC}_{\mathrm{naive}}$ under more realistic situations. Consider an independent exponential sequence such that

$$x_i = \lambda^{(j)}\epsilon_i \quad \text{when } k^{(j-1)} + 1 \le i \le k^{(j)} \tag{11}$$

for $1 \le j \le m+1$, where $k^{(0)} = 0$ and $k^{(m+1)} = n$. We assume that ϵ_i $(1 \le i \le n)$ is independently distributed according to $Ex(1)$, where we denote the exponential distribution with mean λ by $Ex(\lambda)$. We consider three as the true number of the change-points and randomly determine the true change-point $k^{*(j)}$ and the true amount of change $\lambda^{*(j+1)}/\lambda^{*(j)}$ using uniform distributions for $1 \le j \le 3$. For such sequences, we evaluate the Kullback–Leibler divergence between the true and estimated distributions and the rate of selecting an m change-points model. The results are given in Table 4. The values by a naive BIC (Yao 1988) are included only for comparison. We

Table 4 Average Kullback–Leibler (K–L) divergence between the true and estimated distributions and rate of selecting the m change-points model by the $\mathrm{AIC}_{\mathrm{naive}}$, AIC in (7), and BIC

			K–L	≤ 1(%)	2 (%)	3 (%)	4 (%)	≥ 5(%)
n: 100	ξ: 2.0	$\mathrm{AIC}_{\mathrm{naive}}$	10.60	0.0	0.7	64.6	26.3	8.4
		AIC	9.71	1.0	4.7	90.2	3.9	0.2
		BIC	9.86	2.1	6.8	89.0	2.1	0.1
n: 100	ξ: 1.5	$\mathrm{AIC}_{\mathrm{naive}}$	9.94	0.3	4.0	64.0	24.4	7.4
		AIC	9.77	6.0	18.4	72.3	3.2	0.1
		BIC	10.17	9.6	23.1	65.7	1.6	0.0
n: 200	ξ: 1.5	$\mathrm{AIC}_{\mathrm{naive}}$	9.64	0.0	0.4	45.1	32.1	22.5
		AIC	7.86	0.5	3.5	87.6	7.6	0.8
		BIC	7.96	2.2	6.8	88.8	2.2	0.1
n: 200	ξ: 1.0	$\mathrm{AIC}_{\mathrm{naive}}$	9.07	0.2	3.2	45.9	30.0	20.8
		AIC	8.17	4.8	18.1	70.8	5.7	0.7
		BIC	9.09	11.4	28.9	58.3	1.3	0.1
n: 400	ξ: 1.0	$\mathrm{AIC}_{\mathrm{naive}}$	10.46	0.0	0.4	26.1	29.1	44.5
		AIC	7.65	0.8	4.9	80.7	11.1	2.5
		BIC	7.77	3.7	12.5	82.2	1.6	0.1
n: 400	ξ: 0.5	$\mathrm{AIC}_{\mathrm{naive}}$	10.11	0.4	4.4	26.4	29.2	39.6
		AIC	8.15	7.6	29.5	53.7	7.8	1.4
		BIC	9.12	21.2	42.5	35.5	0.8	0.0

These values are obtained by Monte Carlo simulation with 10,000 repetitions using the model (11). The three true change-points $k^{*(1)}$, $k^{*(2)}$, and $k^{*(3)}$ are randomly determined using the uniform distribution under the restriction where $k^{*(1)}$, $k^{*(2)} - k^{*(1)}$, $k^{*(3)} - k^{*(2)}$, and $n - k^{*(3)}$ are larger than $0.1n$, and the true amount of change $\lambda^{*(j+1)}/\lambda^{*(j)}$ is a realization of $2^{u_{1,j}(\xi + u_{2,j})}$ for $1 \leq j \leq 3$, where $u_{1,j}$ and $u_{2,j}$ are independent random variables distributed according to the discrete uniform distribution on $\{-1, 1\}$ and $U(0, 1)$, respectively

can say that the AIC in (7) is superior to the other criteria from the viewpoint of the original purpose of the AIC because the AIC provides a smaller average divergence than the others in every setting. Let us examine the rates to check their performances in more detail. The $\mathrm{AIC}_{\mathrm{naive}}$ tends to select too many change-points when compared to the AIC, especially when the sample size is large. If the amount of change is large, the rate of selecting three changes by the BIC is sometimes higher than that by the AIC, while the difference is small. Otherwise, the BIC tends to select too few change-points when compared to the AIC. Thus, we can also say from the rates that the AIC in (7) provides reasonable model selection in comparison with the $\mathrm{AIC}_{\mathrm{naive}}$ and BIC.

In Table 5, we evaluate such divergences and rates under model misspecification for reference. Comparing to the case without model misspecification, every criterion tends to select fewer change-points. This is because the true bias becomes small under model misspecification, and so this result is consistent with that in Table 2. In addition, even in this case, the AIC in (7) provides a smaller average divergence than the others, and so we can say that the AIC is superior to the other criteria.

Table 5 Average Kullback–Leibler (K–L) divergence between the true and estimated distributions and rate of selecting the m change-points model by the $\mathrm{AIC}_{\mathrm{naive}}$, AIC in (7), and BIC under model misspecification

		K–L		≤1 (%)	2 (%)	3 (%)	4 (%)	≥5 (%)
n: 200	ξ: 1.5	$\mathrm{AIC}_{\mathrm{naive}}$	6.91	0.0	0.4	62.0	26.2	11.3
		AIC	5.83	0.6	2.6	93.9	2.8	0.2
		BIC	6.17	2.3	5.8	91.3	0.6	0.0
n: 200	ξ: 1.0	$\mathrm{AIC}_{\mathrm{naive}}$	6.55	0.3	3.6	60.8	25.0	10.3
		AIC	6.37	5.6	17.5	74.6	2.3	0.1
		BIC	7.53	12.6	27.9	59.1	0.4	0.0

These values are obtained by Monte Carlo simulation with 10,000 repetitions using the model (11). As the distribution of ϵ_i in (11), a mixture distribution whose components are $Ex(1)$ with weight 0.7 and $U(0, 2)$ with weight 0.3 is used as the true one, while $Ex(1)$ is assumed in the model. The three true change-points and the true amounts of changes are randomly determined in the same way as in Table 4

4 Conclusion

Recently, various information criteria have been proposed by generalizing and modifying the idea of the AIC, e.g., RIC (Foster and George 1994), DIC (Spiegelhalter et al. 2002), FIC (Claeskens and Hjort 2003), and the AIC for the LASSO (Zou et al. 2007). For change-point models, however, there was not even a proper AIC, in spite of a great demand of these models in various fields such as econometrics (e.g., Hsu 1979; Hamilton 1989; Garcia and Perron 1996; Chen and Gupta 1997) and biometrics (e.g., Avery and Henderson 1999; Siegmund 2004; Zhang and Siegmund 2007; Ninomiya and Yoshimoto 2008). In this study, the proper AIC for change-point models has been derived explicitly by evaluating the asymptotic bias of the maximum log-likelihood. It has been shown that each change-point requires 6 as the penalty in the AIC, while each of the other parameters requires 2. We can then conduct an easy change-point model selection. In addition, it has been shown by a simulation study that the asymptotic evaluation approximates the bias accurately, and the model selection by the AIC is reasonable. We, therefore, recommend the use of the AIC for applications, as described by Hurrell and Trenberth (1997) for example, which currently use the $\mathrm{AIC}_{\mathrm{naive}}$, a naive AIC proposed by Jones and Dey (1995).

5 Appendix: Mathematical proofs

5.1 Proof of theorem 1

Let $\hat{\boldsymbol{\theta}}_{\boldsymbol{k},\boldsymbol{x}}$ be the maximum likelihood estimator of $\boldsymbol{\theta}^*$ based on $\boldsymbol{x}$ when change-points are $\boldsymbol{k} = (k^{(1)}, \ldots, k^{(m)})^T$, i.e., $\hat{\boldsymbol{\theta}}_{\boldsymbol{k},\boldsymbol{x}} = \mathrm{argsup}_{\boldsymbol{\theta}} L_{\boldsymbol{x}}(\boldsymbol{k}, \boldsymbol{\theta})$, and let $\hat{L}_{\boldsymbol{x}\boldsymbol{y}}(\boldsymbol{k})$ be $L_{\boldsymbol{x}}(\boldsymbol{k}, \hat{\boldsymbol{\theta}}_{\boldsymbol{k},\boldsymbol{y}})$. Then, we can see that $b(\boldsymbol{k}^*, \boldsymbol{\theta}^*)$ in the theorems is the limit to which

$$\mathrm{sup}_{\boldsymbol{k}\in K}\, \hat{L}_{\boldsymbol{x}\boldsymbol{x}}(\boldsymbol{k}) - \hat{L}_{\boldsymbol{x}\boldsymbol{y}}\{\mathrm{argsup}_{\boldsymbol{k}\in K}\, \hat{L}_{\boldsymbol{y}\boldsymbol{y}}(\boldsymbol{k})\}$$

converges in distribution, where K is a set of $\boldsymbol{k}$ such that $\hat{L}_{\boldsymbol{xx}}(\boldsymbol{k})$ is $O_{\mathrm{P}}(1)$ or positive, that is, $P\{\hat{L}_{\boldsymbol{xx}}(\boldsymbol{k}) > -M\}$ does not converge to 0 for some $M > 0$. In addition, we use $A_{\boldsymbol{\theta}}$, $A'_{\boldsymbol{\theta}}$, $A''_{\boldsymbol{\theta}}$, and $\boldsymbol{T}_{\boldsymbol{x}}$ in place of $A(\boldsymbol{\theta})$, $A'(\boldsymbol{\theta})$, $A''(\boldsymbol{\theta})$, and $\boldsymbol{T}(\boldsymbol{x})$ to save space. We use this notation throughout the appendix.

Let us consider the case $\boldsymbol{k} - \boldsymbol{k}^* = O(1)$ and the case $\boldsymbol{k} - \boldsymbol{k}^* \neq O(1)$, separately. In the case $\boldsymbol{k} - \boldsymbol{k}^* = O(1)$, $\hat{\boldsymbol{\theta}}_{\boldsymbol{k},\boldsymbol{x}} = \hat{\boldsymbol{\theta}}_{\boldsymbol{k}^*,\boldsymbol{x}} + O_{\mathrm{P}}(n^{-1}) = \boldsymbol{\theta}^* + O_{\mathrm{P}}(n^{-1/2})$, and so we have

$$\begin{aligned}
&\log f(\boldsymbol{x}|\boldsymbol{k}, \hat{\boldsymbol{\theta}}_{\boldsymbol{k},\boldsymbol{x}}) - \log f(\boldsymbol{x}|\boldsymbol{k}, \boldsymbol{\theta}^*) \\
&\quad = \sum_{j=1}^{m+1} (k^{(j)} - k^{(j-1)})(\boldsymbol{\theta}^{*(j)} - \hat{\boldsymbol{\theta}}^{(j)}_{\boldsymbol{k},\boldsymbol{x}})^{\mathrm{T}} A''_{\boldsymbol{\theta}^{*(j)}} (\boldsymbol{\theta}^{*(j)} - \hat{\boldsymbol{\theta}}^{(j)}_{\boldsymbol{k},\boldsymbol{x}})/2 + o_{\mathrm{P}}(1) \\
&\quad = \sum_{j=1}^{m+1} (k^{*(j)} - k^{*(j-1)})(\boldsymbol{\theta}^{*(j)} - \hat{\boldsymbol{\theta}}^{(j)}_{\boldsymbol{k}^*,\boldsymbol{x}})^{T} A''_{\boldsymbol{\theta}^{*(j)}} (\boldsymbol{\theta}^{*(j)} - \hat{\boldsymbol{\theta}}^{(j)}_{\boldsymbol{k}^*,\boldsymbol{x}})/2 + o_{\mathrm{P}}(1) \\
&\quad = \log f(\boldsymbol{x}|\boldsymbol{k}^*, \hat{\boldsymbol{\theta}}_{\boldsymbol{k}^*,\boldsymbol{x}}) - \log f(\boldsymbol{x}|\boldsymbol{k}^*, \boldsymbol{\theta}^*) + o_{\mathrm{P}}(1), \qquad (12)
\end{aligned}$$

where $k^{(0)} = 0$ and $k^{(m+1)} = n$. The first and third equalities are obtained from a Taylor expansion, and the second one is obtained from $(\boldsymbol{\theta}^{*(j)} - \hat{\boldsymbol{\theta}}^{(j)}_{\boldsymbol{k}^*,\boldsymbol{x}}) - (\boldsymbol{\theta}^{*(j)} - \hat{\boldsymbol{\theta}}^{(j)}_{\boldsymbol{k},\boldsymbol{x}}) = O_{\mathrm{P}}(n^{-1})$ and $(k^{*(j)} - k^{*(j-1)}) - (k^{(j)} - k^{(j-1)}) = O(1)$ together with $\boldsymbol{\theta}^{*(j)} - \hat{\boldsymbol{\theta}}^{(j)}_{\boldsymbol{k},\boldsymbol{x}} = O_{\mathrm{P}}(n^{-1/2})$ and $\boldsymbol{\theta}^{*(j)} - \hat{\boldsymbol{\theta}}^{(j)}_{\boldsymbol{k}^*,\boldsymbol{x}} = O_{\mathrm{P}}(n^{-1/2})$. Therefore, it follows that

$$\begin{aligned}
&\log f(\boldsymbol{x}|\boldsymbol{k}, \hat{\boldsymbol{\theta}}_{\boldsymbol{k},\boldsymbol{x}}) - \log f(\boldsymbol{x}|\boldsymbol{k}^*, \hat{\boldsymbol{\theta}}_{\boldsymbol{k}^*,\boldsymbol{x}}) \\
&\quad = \log f(\boldsymbol{x}|\boldsymbol{k}, \boldsymbol{\theta}^*) - \log f(\boldsymbol{x}|\boldsymbol{k}^*, \boldsymbol{\theta}^*) + o_{\mathrm{P}}(1) \\
&\quad = \sum_{j=1}^{m} \Bigg[I_{\{k^{(j)} < k^{*(j)}\}} \sum_{i=k^{(j)}+1}^{k^{*(j)}} \{(\boldsymbol{\theta}^{*(j+1)} - \boldsymbol{\theta}^{*(j)})^{T} \boldsymbol{T}_{\boldsymbol{x}_i} - A_{\boldsymbol{\theta}^{*(j+1)}} + A_{\boldsymbol{\theta}^{*(j)}}\} \\
&\qquad + I_{\{k^{(j)} > k^{*(j)}\}} \sum_{i=k^{*(j)}+1}^{k^{(j)}} \{(\boldsymbol{\theta}^{*(j)} - \boldsymbol{\theta}^{*(j+1)})^{T} \boldsymbol{T}_{\boldsymbol{x}_i} - A_{\boldsymbol{\theta}^{*(j)}} + A_{\boldsymbol{\theta}^{*(j+1)}}\} \Bigg] + o_{\mathrm{P}}(1) \\
&\quad = \sum_{j=1}^{m} Q^{(j)}_{k^{(j)},\boldsymbol{x}} + o_{\mathrm{P}}(1) = O_{\mathrm{P}}(1). \qquad (13)
\end{aligned}$$

In addition, it holds that

$$\log f(\boldsymbol{x}|\boldsymbol{k}^*, \hat{\boldsymbol{\theta}}_{\boldsymbol{k}^*,\boldsymbol{x}}) - \log f(\boldsymbol{x}|\boldsymbol{k}^*, \boldsymbol{\theta}^*) = \chi^2_{p_m}/2 + o_{\mathrm{P}}(1) = O_{\mathrm{P}}(1), \qquad (14)$$

where $\chi^2_{p_m}$ is a random variable distributed according to the χ^2 distribution of degree p_m. From (13) and (14), we have

$$\hat{L}_{\boldsymbol{xx}}(\boldsymbol{k}) = O_{\mathrm{P}}(1) \quad \text{when} \quad \boldsymbol{k} - \boldsymbol{k}^* = O(1). \qquad (15)$$

On the other hand, we can obtain

$$P\{\hat{L}_{\boldsymbol{xx}}(\boldsymbol{k}) > -M\} \to 0 \quad \text{when} \quad \boldsymbol{k} - \boldsymbol{k}^* \neq O(1) \tag{16}$$

for arbitrary $M > 0$, where $\boldsymbol{k} - \boldsymbol{k}^* \neq O(1)$ means that $|k^{(j)} - k^{*(j)}| \to \infty$ as $n \to \infty$ for some j. To derive it, let us consider the case $0 < k^{*(j')} - k^{(j')} \neq O(1)$ and $k^{(j)} - k^{*(j)} = O(1)$ for $j \neq j'$ as an example. In this case, we have

$$\begin{aligned}
&\log f(\boldsymbol{x}|\boldsymbol{k}, \hat{\boldsymbol{\theta}}_{\boldsymbol{k},\boldsymbol{x}}) - \log f(\boldsymbol{x}|\boldsymbol{k}^*, \hat{\boldsymbol{\theta}}_{\boldsymbol{k}^*,\boldsymbol{x}}) \\
&= \sum_{i=k^{(j'-1)}+1}^{k^{(j')}} \{(\hat{\boldsymbol{\theta}}_{\boldsymbol{k},\boldsymbol{x}}^{(j')} - \hat{\boldsymbol{\theta}}_{\boldsymbol{k}^*,\boldsymbol{x}}^{(j')})^T \boldsymbol{T}_{\boldsymbol{x}_i} - A_{\hat{\boldsymbol{\theta}}_{\boldsymbol{k},\boldsymbol{x}}^{(j')}} + A_{\hat{\boldsymbol{\theta}}_{\boldsymbol{k}^*,\boldsymbol{x}}^{(j')}} \} \\
&\quad + \sum_{i=k^{(j')}+1}^{k^{*(j')}} \{(\hat{\boldsymbol{\theta}}_{\boldsymbol{k},\boldsymbol{x}}^{(j'+1)} - \hat{\boldsymbol{\theta}}_{\boldsymbol{k}^*,\boldsymbol{x}}^{(j')})^T \boldsymbol{T}_{\boldsymbol{x}_i} - A_{\hat{\boldsymbol{\theta}}_{\boldsymbol{k},\boldsymbol{x}}^{(j'+1)}} + A_{\hat{\boldsymbol{\theta}}_{\boldsymbol{k}^*,\boldsymbol{x}}^{(j')}} \} \\
&\quad + \sum_{i=k^{*(j')}+1}^{k^{(j'+1)}} \{(\hat{\boldsymbol{\theta}}_{\boldsymbol{k},\boldsymbol{x}}^{(j'+1)} - \hat{\boldsymbol{\theta}}_{\boldsymbol{k}^*,\boldsymbol{x}}^{(j'+1)})^T \boldsymbol{T}_{\boldsymbol{x}_i} - A_{\hat{\boldsymbol{\theta}}_{\boldsymbol{k},\boldsymbol{x}}^{(j'+1)}} + A_{\hat{\boldsymbol{\theta}}_{\boldsymbol{k}^*,\boldsymbol{x}}^{(j'+1)}}\} + O_{\mathrm{P}}(1) \\
&= \sum_{i=k^{(j'-1)}+1}^{k^{(j')}} [(\hat{\boldsymbol{\theta}}_{\boldsymbol{k},\boldsymbol{x}}^{(j')} - \hat{\boldsymbol{\theta}}_{\boldsymbol{k}^*,\boldsymbol{x}}^{(j')})^T \{\boldsymbol{T}_{\boldsymbol{x}_i} - A'_{\hat{\boldsymbol{\theta}}_{\boldsymbol{k}^*,\boldsymbol{x}}^{(j')}} \} - B_{\hat{\boldsymbol{\theta}}_{\boldsymbol{k},\boldsymbol{x}}^{(j')}, \hat{\boldsymbol{\theta}}_{\boldsymbol{k}^*,\boldsymbol{x}}^{(j')}}] \\
&\quad + \sum_{i=k^{(j')}+1}^{k^{*(j)}} [(\hat{\boldsymbol{\theta}}_{\boldsymbol{k},\boldsymbol{x}}^{(j'+1)} - \hat{\boldsymbol{\theta}}_{\boldsymbol{k}^*,\boldsymbol{x}}^{(j')})^T \{\boldsymbol{T}_{\boldsymbol{x}_i} - A'_{\hat{\boldsymbol{\theta}}_{\boldsymbol{k}^*,\boldsymbol{x}}^{(j')}} \} - B_{\hat{\boldsymbol{\theta}}_{\boldsymbol{k},\boldsymbol{x}}^{(j'+1)}, \hat{\boldsymbol{\theta}}_{\boldsymbol{k}^*,\boldsymbol{x}}^{(j')}}] \\
&\quad + \sum_{i=k^{*(j')}+1}^{k^{(j'+1)}} [(\hat{\boldsymbol{\theta}}_{\boldsymbol{k},\boldsymbol{x}}^{(j'+1)} - \hat{\boldsymbol{\theta}}_{\boldsymbol{k}^*,\boldsymbol{x}}^{(j'+1)})^T \{\boldsymbol{T}_{\boldsymbol{x}_i} - A'_{\hat{\boldsymbol{\theta}}_{\boldsymbol{k}^*,\boldsymbol{x}}^{(j'+1)}}\} - B_{\hat{\boldsymbol{\theta}}_{\boldsymbol{k},\boldsymbol{x}}^{(j'+1)}, \hat{\boldsymbol{\theta}}_{\boldsymbol{k}^*,\boldsymbol{x}}^{(j'+1)}}] + O_{\mathrm{P}}(1),
\end{aligned} \tag{17}$$

where $B_{\boldsymbol{\theta}^\dagger,\boldsymbol{\theta}^\ddagger} = A_{\boldsymbol{\theta}^\dagger} - A_{\boldsymbol{\theta}^\ddagger} - (\boldsymbol{\theta}^\dagger - \boldsymbol{\theta}^\ddagger)^T A'_{\boldsymbol{\theta}^\ddagger}$. It holds from the central limit theorem that $\sum_{i=k^{(j'-1)}+1}^{k^{(j')}} (\hat{\boldsymbol{\theta}}_{\boldsymbol{k},\boldsymbol{x}}^{(j')} - \hat{\boldsymbol{\theta}}_{\boldsymbol{k}^*,\boldsymbol{x}}^{(j')})^T \{\boldsymbol{T}_{\boldsymbol{x}_i} - A'_{\hat{\boldsymbol{\theta}}_{\boldsymbol{k}^*,\boldsymbol{x}}^{(j')}} \} + \sum_{i=k^{(j')}+1}^{k^{*(j)}} (\hat{\boldsymbol{\theta}}_{\boldsymbol{k},\boldsymbol{x}}^{(j'+1)} - \hat{\boldsymbol{\theta}}_{\boldsymbol{k}^*,\boldsymbol{x}}^{(j')})^T \{\boldsymbol{T}_{\boldsymbol{x}_i} - A'_{\hat{\boldsymbol{\theta}}_{\boldsymbol{k}^*,\boldsymbol{x}}^{(j')}} \} + \sum_{i=k^{*(j')}+1}^{k^{(j'+1)}} (\hat{\boldsymbol{\theta}}_{\boldsymbol{k},\boldsymbol{x}}^{(j'+1)} - \hat{\boldsymbol{\theta}}_{\boldsymbol{k}^*,\boldsymbol{x}}^{(j'+1)})^T \{\boldsymbol{T}_{\boldsymbol{x}_i} - A'_{\hat{\boldsymbol{\theta}}_{\boldsymbol{k}^*,\boldsymbol{x}}^{(j'+1)}}\}$ is $O_{\mathrm{P}}\{(k^{*(j')} - k^{(j')})^{1/2}\}$ because $\hat{\boldsymbol{\theta}}_{\boldsymbol{k},\boldsymbol{x}}^{(j')} - \hat{\boldsymbol{\theta}}_{\boldsymbol{k}^*,\boldsymbol{x}}^{(j')}$ and $\hat{\boldsymbol{\theta}}_{\boldsymbol{k},\boldsymbol{x}}^{(j'+1)} - \hat{\boldsymbol{\theta}}_{\boldsymbol{k}^*,\boldsymbol{x}}^{(j'+1)}$ are $O_{\mathrm{P}}\{n^{-1}(k^{*(j')} - k^{(j')})\}$, and it holds that $\sum_{i=k^{(j'-1)}+1}^{k^{(j')}} B_{\hat{\boldsymbol{\theta}}_{\boldsymbol{k},\boldsymbol{x}}^{(j')}, \hat{\boldsymbol{\theta}}_{\boldsymbol{k}^*,\boldsymbol{x}}^{(j')}} + \sum_{i=k^{(j')}+1}^{k^{*(j)}} B_{\hat{\boldsymbol{\theta}}_{\boldsymbol{k},\boldsymbol{x}}^{(j'+1)}, \hat{\boldsymbol{\theta}}_{\boldsymbol{k}^*,\boldsymbol{x}}^{(j')}} + \sum_{i=k^{*(j')}+1}^{k^{(j'+1)}} B_{\hat{\boldsymbol{\theta}}_{\boldsymbol{k},\boldsymbol{x}}^{(j'+1)}, \hat{\boldsymbol{\theta}}_{\boldsymbol{k}^*,\boldsymbol{x}}^{(j'+1)}}$ is positive by its definition and is not $o_{\mathrm{P}}(k^{*(j')} - k^{(j')})$ because $\hat{\boldsymbol{\theta}}_{\boldsymbol{k},\boldsymbol{x}}^{(j'+1)} - \hat{\boldsymbol{\theta}}_{\boldsymbol{k}^*,\boldsymbol{x}}^{(j')}$ is not $o_{\mathrm{P}}(1)$. Then, it follows that

$$P\{\log f(\boldsymbol{x}|\boldsymbol{k}, \hat{\boldsymbol{\theta}}_{\boldsymbol{k},\boldsymbol{x}}) - \log f(\boldsymbol{x}|\boldsymbol{k}^*, \hat{\boldsymbol{\theta}}_{\boldsymbol{k}^*,\boldsymbol{x}}) > -M\} \to 0, \tag{18}$$

and we can say from (14) and (18) that (16) holds. Thus, from (15) and (16), we obtain that $K = \{\boldsymbol{k} \mid (\boldsymbol{k} - \boldsymbol{k}^*) = O(1)\}$, and so (13) and (14) hold. Therefore, we have

$$\sup_{\boldsymbol{k}\in K} \hat{L}_{\boldsymbol{xx}}(\boldsymbol{k}) = \sum_{j=1}^{m} \sup_{k\in K^{(j)}} Q_{k,\boldsymbol{x}}^{(j)} + \chi_{p_m}^2/2 + o_{\mathrm{P}}(1), \tag{19}$$

where $K^{(j)} = \{k^{(j)} \mid \boldsymbol{k} \in K\}$, and

$$\operatorname{argsup}_{\boldsymbol{k}\in K} \hat{L}_{\boldsymbol{yy}}(\boldsymbol{k}) = \tilde{\boldsymbol{k}}_{\boldsymbol{y}} + o_{\mathrm{P}}(1), \tag{20}$$

where $\tilde{k}_{\boldsymbol{y}}^{(j)} = \operatorname{argsup}_{k\in K^{(j)}} Q_{k,\boldsymbol{y}}^{(j)}$ and $\tilde{\boldsymbol{k}}_{\boldsymbol{y}} = (\tilde{k}_{\boldsymbol{y}}^{(1)}, \ldots, \tilde{k}_{\boldsymbol{y}}^{(m)})$. Recalling that $\boldsymbol{y}$ is a copy of $\boldsymbol{x}$, it follows that $\tilde{k}_{\boldsymbol{y}}^{(j)} - k^{*(j)} = O_{\mathrm{P}}(1)$ and $\hat{\boldsymbol{\theta}}_{\tilde{\boldsymbol{k}}_{\boldsymbol{y}},\boldsymbol{y}}^{(j)} = \hat{\boldsymbol{\theta}}_{\boldsymbol{k}^*,\boldsymbol{y}}^{(j)} + O_{\mathrm{P}}(n^{-1}) = \boldsymbol{\theta}^{*(j)} + O_{\mathrm{P}}(n^{-1/2})$ for all j. Using these and applying Taylor expansion, we have

$$\begin{aligned}
&\log f(\boldsymbol{x}|\tilde{\boldsymbol{k}}_{\boldsymbol{y}}, \boldsymbol{\theta}^*) - \log f(\boldsymbol{x}|\tilde{\boldsymbol{k}}_{\boldsymbol{y}}, \hat{\boldsymbol{\theta}}_{\tilde{\boldsymbol{k}}_{\boldsymbol{y}},\boldsymbol{y}}) \\
&= -\sum_{j=1}^{m+1} (\hat{\boldsymbol{\theta}}_{\tilde{\boldsymbol{k}}_{\boldsymbol{y}},\boldsymbol{y}}^{(j)} - \boldsymbol{\theta}^{*(j)})^T \Big[\sum_{i=\tilde{k}_{\boldsymbol{y}}^{(j-1)}+1}^{\tilde{k}_{\boldsymbol{y}}^{(j)}} \{\boldsymbol{T}_{\boldsymbol{x}_i} - A'_{\boldsymbol{\theta}^{*(j)}}\} \Big] \\
&\quad + \sum_{j=1}^{m+1} (\tilde{k}_{\boldsymbol{y}}^{(j)} - \tilde{k}_{\boldsymbol{y}}^{(j-1)}) (\hat{\boldsymbol{\theta}}_{\tilde{\boldsymbol{k}}_{\boldsymbol{y}},\boldsymbol{y}}^{(j)} - \boldsymbol{\theta}^{*(j)})^T A''_{\boldsymbol{\theta}^{*(j)}} (\hat{\boldsymbol{\theta}}_{\tilde{\boldsymbol{k}}_{\boldsymbol{y}},\boldsymbol{y}}^{(j)} - \boldsymbol{\theta}^{*(j)})/2 + o_{\mathrm{P}}(1) \\
&= -\sum_{j=1}^{m+1} (\hat{\boldsymbol{\theta}}_{\tilde{\boldsymbol{k}}_{\boldsymbol{y}},\boldsymbol{y}}^{(j)} - \boldsymbol{\theta}^{*(j)})^T \Big[\sum_{i=k^{*(j-1)}+1}^{k^{*(j)}} \{\boldsymbol{T}_{\boldsymbol{x}_i} - A'_{\boldsymbol{\theta}^{*(j)}}\} \Big] \\
&\quad + \sum_{j=1}^{m+1} (k^{*(j)} - k^{*(j-1)}) (\hat{\boldsymbol{\theta}}_{\tilde{\boldsymbol{k}}_{\boldsymbol{y}},\boldsymbol{y}}^{(j)} - \boldsymbol{\theta}^{*(j)})^T A''_{\boldsymbol{\theta}^{*(j)}} (\hat{\boldsymbol{\theta}}_{\tilde{\boldsymbol{k}}_{\boldsymbol{y}},\boldsymbol{y}}^{(j)} - \boldsymbol{\theta}^{*(j)})/2 + o_{\mathrm{P}}(1) \\
&= -\sum_{j=1}^{m+1} (\hat{\boldsymbol{\theta}}_{\boldsymbol{k}^*,\boldsymbol{y}}^{(j)} - \boldsymbol{\theta}^{*(j)})^T \Big[\sum_{i=k^{*(j-1)}+1}^{k^{*(j)}} \{\boldsymbol{T}_{\boldsymbol{x}_i} - A'_{\boldsymbol{\theta}^{*(j)}}\} \Big] \\
&\quad + \sum_{j=1}^{m+1} (k^{*(j)} - k^{*(j-1)}) (\hat{\boldsymbol{\theta}}_{\boldsymbol{k}^*,\boldsymbol{y}}^{(j)} - \boldsymbol{\theta}^{*(j)})^T A''_{\boldsymbol{\theta}^{*(j)}} (\hat{\boldsymbol{\theta}}_{\boldsymbol{k}^*,\boldsymbol{y}}^{(j)} - \boldsymbol{\theta}^{*(j)})/2 + o_{\mathrm{P}}(1) \\
&= \sum_{j=1}^{m+1} \boldsymbol{N}^{(j)} \boldsymbol{N}'^{(j)} + \chi_{p_m}^{2'}/2 + o_{\mathrm{P}}(1),
\end{aligned} \tag{21}$$

where $\tilde{k}_{\boldsymbol{y}}^{(0)} = 0$ and $\tilde{k}_{\boldsymbol{y}}^{(m+1)} = n$. Here, $\chi_{p_m}^{2'}$ is another random variable distributed according to the χ^2 distribution of degree p_m, and $\boldsymbol{N}^{(j)}$ and $\boldsymbol{N}'^{(j)}$ are two independent random vectors distributed according to normal distributions $N(\boldsymbol{0}, A''^{-1}_{\boldsymbol{\theta}^{*(j)}})$ and

$N(\mathbf{0}, A''_{\boldsymbol{\theta}^{*(j)}})$, respectively, where $\mathbf{0}$ is a zero-vector. In addition, we obtain

$$\begin{aligned}
&\log f(\boldsymbol{x}|\boldsymbol{k}^*, \boldsymbol{\theta}^*) - \log f(\boldsymbol{x}|\tilde{\boldsymbol{k}}_{\boldsymbol{y}}, \boldsymbol{\theta}^*) \\
&= \sum_{j=1}^{m} \Bigg[I_{\{\tilde{k}_{\boldsymbol{y}}^{(j)} < k^{*(j)}\}} \sum_{i=\tilde{k}_{\boldsymbol{y}}^{(j)}+1}^{k^{*(j)}} \{(\boldsymbol{\theta}^{*(j)} - \boldsymbol{\theta}^{*(j+1)})^T \boldsymbol{T}_{\boldsymbol{x}_i} - A_{\boldsymbol{\theta}^{*(j)}} + A_{\boldsymbol{\theta}^{*(j+1)}}\} \\
&\quad + I_{\{\tilde{k}_{\boldsymbol{y}}^{(j)} > k^{*(j)}\}} \sum_{i=k^{*(j)}+1}^{\tilde{k}_{\boldsymbol{y}}^{(j)}} \{(\boldsymbol{\theta}^{*(j+1)} - \boldsymbol{\theta}^{*(j)})^T \boldsymbol{T}_{\boldsymbol{x}_i} - A_{\boldsymbol{\theta}^{*(j+1)}} + A_{\boldsymbol{\theta}^{*(j)}}\} \Bigg] + o_{\mathrm{P}}(1) \\
&= \sum_{j=1}^{m} Q^{(j)}_{\mathrm{argsup}_{k \in K^{(j)}} Q^{(j)}_{k,\boldsymbol{y},\boldsymbol{x}}} + o_{\mathrm{P}}(1). \qquad (22)
\end{aligned}$$

From (20), (21), and (22), it follows that

$$\begin{aligned}
&-\hat{L}_{\boldsymbol{xy}}\{\mathrm{argsup}_{\boldsymbol{k} \in K} \hat{L}_{\boldsymbol{yy}}(\boldsymbol{k})\} = -\hat{L}_{\boldsymbol{xy}}(\tilde{\boldsymbol{k}}_{\boldsymbol{y}}) + o_{\mathrm{P}}(1) \\
&= \sum_{j=1}^{m} Q^{(j)}_{\mathrm{argsup}_{k \in K^{(j)}} Q^{(j)}_{k,\boldsymbol{y},\boldsymbol{x}}} + \sum_{j=1}^{m+1} \boldsymbol{N}^{(j)} \boldsymbol{N}'^{(j)} + \chi^2_{p_m}/2 + o_{\mathrm{P}}(1). \qquad (23)
\end{aligned}$$

From (19) and (23), we can obtain the theorem.

5.2 Proof of theorem 2

To begin with, we check that (19), (20), and (23) also hold under the condition (6). Let β_n be an increasing sequence satisfying $o(\alpha_n) \neq \beta_n = O(n)$. When $\boldsymbol{k} - \boldsymbol{k}^* = O(\beta_n)$, $\hat{\boldsymbol{\theta}}_{\boldsymbol{k},\boldsymbol{x}} = \hat{\boldsymbol{\theta}}_{\boldsymbol{k}^*,\boldsymbol{x}} + O_{\mathrm{P}}\{n^{-1} \sum_{j=1}^{m} (\boldsymbol{\theta}^{*(j+1)} - \boldsymbol{\theta}^{*(j)}) \beta_n\} = \hat{\boldsymbol{\theta}}_{\boldsymbol{k}^*,\boldsymbol{x}} + O_{\mathrm{P}}(n^{-1} \alpha_n{}^{-1/2} \beta_n)$. Using this relation, we can obtain the following in the same way as in the proof of Theorem 1. First, we consider the case where $\boldsymbol{k} - \boldsymbol{k}^* = O(\alpha_n)$. In this case, (12), (13), and (14) hold, and so (15) holds using $\boldsymbol{k} - \boldsymbol{k}^* = O(\alpha_n)$ in place of $\boldsymbol{k} - \boldsymbol{k}^* = O(1)$. Next, we consider the case where $0 < k^{*(j')} - k^{(j')} \neq O(\alpha_n)$ and $k^{(j)} - k^{*(j)} = O(\alpha_n)$ for $j \neq j'$ as an example. In this case, (17) holds, $\sum_{i=k^{(j'-1)}+1}^{k^{(j')}} (\hat{\boldsymbol{\theta}}^{(j')}_{\boldsymbol{k},\boldsymbol{x}} - \hat{\boldsymbol{\theta}}^{(j')}_{\boldsymbol{k}^*,\boldsymbol{x}})^T (\boldsymbol{T}_{\boldsymbol{x}_i} - A'_{\hat{\boldsymbol{\theta}}^{(j')}_{\boldsymbol{k}^*,\boldsymbol{x}}}) + \sum_{i=k^{(j')}+1}^{k^{*(j)}} (\hat{\boldsymbol{\theta}}^{(j'+1)}_{\boldsymbol{k},\boldsymbol{x}} - \hat{\boldsymbol{\theta}}^{(j')}_{\boldsymbol{k}^*,\boldsymbol{x}})^T (\boldsymbol{T}_{\boldsymbol{x}_i} - A'_{\hat{\boldsymbol{\theta}}^{(j')}_{\boldsymbol{k}^*,\boldsymbol{x}}}) + \sum_{i=k^{*(j')}+1}^{k^{(j'+1)}} (\hat{\boldsymbol{\theta}}^{(j'+1)}_{\boldsymbol{k},\boldsymbol{x}} - \hat{\boldsymbol{\theta}}^{(j'+1)}_{\boldsymbol{k}^*,\boldsymbol{x}})^T (\boldsymbol{T}_{\boldsymbol{x}_i} - A'_{\hat{\boldsymbol{\theta}}^{(j'+1)}_{\boldsymbol{k}^*,\boldsymbol{x}}}) = O_{\mathrm{P}}\{\alpha_n^{-1/2} (k^{*(j')} - k^{(j')})^{1/2}\}$ and $\sum_{i=k^{(j'-1)}+1}^{k^{(j')}} B_{\hat{\boldsymbol{\theta}}^{(j')}_{\boldsymbol{k},\boldsymbol{x}}, \hat{\boldsymbol{\theta}}^{(j')}_{\boldsymbol{k}^*,\boldsymbol{x}}} + \sum_{i=k^{(j')}+1}^{k^{*(j)}} B_{\hat{\boldsymbol{\theta}}^{(j'+1)}_{\boldsymbol{k},\boldsymbol{x}}, \hat{\boldsymbol{\theta}}^{(j')}_{\boldsymbol{k}^*,\boldsymbol{x}}} + \sum_{i=k^{*(j')}+1}^{k^{(j'+1)}} B_{\hat{\boldsymbol{\theta}}^{(j'+1)}_{\boldsymbol{k},\boldsymbol{x}}, \hat{\boldsymbol{\theta}}^{(j'+1)}_{\boldsymbol{k}^*,\boldsymbol{x}}} \neq o_{\mathrm{P}}\{\alpha_n^{-1} (k^{*(j')} - k^{(j')})\}$, and so we can say that (16) holds using $\boldsymbol{k} - \boldsymbol{k}^* \neq O(\alpha_n)$ in place of $\boldsymbol{k} - \boldsymbol{k}^* \neq O(1)$. Thus, we obtain $K = \{\boldsymbol{k} \mid (\boldsymbol{k} - \boldsymbol{k}^*) = O(\alpha_n)\}$, and so (19), (20), (21), (22), and (23) all hold.

Thus, we consider the case $\boldsymbol{k} - \boldsymbol{k}^* = O(\alpha_n)$. Let $\{W_s\}_{s \in \boldsymbol{R}}$ be a two-sided standard Brownian motion with $E(W_s) = 0$ and $\mathrm{var}(W_s) = |s|$, and let $\sigma^{(j)} =$

$\{\Delta_{\boldsymbol{\theta}^*}^{(j)T} A''(\boldsymbol{\theta}^{*(j)})\Delta_{\boldsymbol{\theta}^*}^{(j)}\}^{1/2}$. Under the condition (6), $Q^{(j)}_{k^{*(j)}+[s\alpha_n],\boldsymbol{x}}$ can be written as

$$I_{\{s<0\}}\alpha_n^{-1/2}\sum_{i=k^{*(j)}+[s\alpha_n]+1}^{k^{*(j)}}\{\Delta_{\boldsymbol{\theta}^*}^{(j)T}(\boldsymbol{T}_{\boldsymbol{x}_i}-A'_{\boldsymbol{\theta}^{*(j)}})\}-|s|\alpha_n B_{\boldsymbol{\theta}^{*(j+1)},\boldsymbol{\theta}^{*(j)}}$$
$$+I_{\{s>0\}}\alpha_n^{-1/2}\sum_{i=k^{*(j)}+1}^{k^{*(j)}+[s\alpha_n]}\{-\Delta_{\boldsymbol{\theta}^*}^{(j)T}(\boldsymbol{T}_{\boldsymbol{x}_i}-A'_{\boldsymbol{\theta}^{*(j+1)}})\}-|s|\alpha_n B_{\boldsymbol{\theta}^{*(j)},\boldsymbol{\theta}^{*(j+1)}},$$

and all of the terms $\Delta_{\boldsymbol{\theta}^*}^{(j)T} A''_{\boldsymbol{\theta}^{*(j+1)}}\Delta_{\boldsymbol{\theta}^*}^{(j)}$, $2\alpha_n B_{\boldsymbol{\theta}^{*(j+1)},\boldsymbol{\theta}^{*(j)}}$, and $2\alpha_n B_{\boldsymbol{\theta}^{*(j)},\boldsymbol{\theta}^{*(j+1)}}$ converge to $\sigma^{(j)2}$. It then follows from the functional central limit theorem that $Q^{(j)}_{k^{*(j)}+[s\alpha_n],\boldsymbol{x}}$ converges in distribution to $V_s(\sigma^{(j)2}/2, \sigma^{(j)2}/2, \sigma^{(j)}, \sigma^{(j)})$, where

$$V_s(c_1,c_2,\sigma_1,\sigma_2)=\begin{cases}-c_1|s|+\sigma_1 W_s & (s\le 0)\\ -c_2|s|+\sigma_2 W_s & (s>0).\end{cases}$$

Therefore, we have

$$\sup_k Q^{(j)}_{k,\boldsymbol{x}} \xrightarrow{d} \sup_s V_s(\sigma^{(j)2}/2, \sigma^{(j)2}/2, \sigma^{(j)}, \sigma^{(j)}) \tag{24}$$

and

$$Q^{(j)}_{\mathrm{argsup}_k Q^{(j)}_{k,\boldsymbol{y}},\boldsymbol{x}} \xrightarrow{d} V_{\mathrm{argsup}_s V'_s(\sigma^{(j)2}/2,\sigma^{(j)2}/2,\sigma^{(j)},\sigma^{(j)})}(\sigma^{(j)2}/2, \sigma^{(j)2}/2, \sigma^{(j)}, \sigma^{(j)}), \tag{25}$$

where V'_s is a copy of V_s.

For distributions about the maximum of a Brownian motion with a negative drift, the results by Bhattacharya and Brockwell (1976) or Shepp (1979) can be used. First, using the equality

$$P\{\sup_{s>0}(W_s-cs)>a\}=\exp(-2ac)$$

for arbitrary $a, c>0$, we have

$$\begin{aligned}&E\{\sup_s V_s(c_1,c_2,\sigma_1,\sigma_2)\}\\ &=\int_0^\infty \mathrm{P}\{\sup_s V_s(c_1,c_2,\sigma_1,\sigma_2)>a\}\mathrm{d}a\\ &=\int_0^\infty [\exp(-2c_1a/\sigma_1^2)+\exp(-2c_2a/\sigma_2^2)-\exp\{-2(c_1\sigma_2^2+c_2\sigma_1^2)a/\sigma_1^2\sigma_2^2\}]\mathrm{d}a\\ &=(c_1^2\sigma_2^4+c_1c_2\sigma_1^2\sigma_2^2+c_2^2\sigma_1^4)/\{2c_1c_2(c_1\sigma_2^2+c_2\sigma_1^2)\}.\end{aligned} \tag{26}$$

Next, we use the property that the probability function of $\mathrm{argsup}_s V_s(c_1, c_2, \sigma_1, \sigma_2)$ is

$$f^{\mathrm{argsup}V}(s|c_1, c_2, \sigma_1, \sigma_2) = \begin{cases} g(-s|c_1/\sigma_1, c_2\sigma_1/\sigma_2^2) & (s < 0) \\ g(s|c_2/\sigma_2, c_1\sigma_2/\sigma_1^2) & (s > 0), \end{cases}$$

where $g(s|a_1, a_2)$ is

$$2a_1(a_1 + 2a_2)\exp\{2a_2(a_1 + a_2)s\}\Phi\{-(a_1 + 2a_2)s^{1/2}\} - 2a_1^2\Phi(-a_1 s^{1/2}).$$

Then, we have

$$\begin{aligned} &E\{V_{\mathrm{argsup}_s V'_s(c_1,c_2,\sigma_1,\sigma_2)}(c_1, c_2, \sigma_1, \sigma_2)\} \\ &= \int_0^\infty sg(s|c_1/\sigma_1, c_2\sigma_1/\sigma_2^2)c_1 \mathrm{d}s + \int_0^\infty sg(s|c_2/\sigma_2, c_1\sigma_2/\sigma_1^2)c_2 \mathrm{d}s \\ &= c_2(2c_1\sigma_2^2 + c_2\sigma_1^2)\sigma_1^4/\{2c_1(c_1\sigma_2^2 + c_2\sigma_1^2)^2\} \\ &\quad + c_1(2c_2\sigma_1^2 + c_1\sigma_2^2)\sigma_2^4/\{2c_2(c_1\sigma_2^2 + c_2\sigma_1^2)^2\} \\ &= (c_1^2\sigma_2^4 + c_1c_2\sigma_1^2\sigma_2^2 + c_2^2\sigma_1^4)/\{2c_1c_2(c_1\sigma_2^2 + c_2\sigma_1^2)\}. \end{aligned} \tag{27}$$

Here, the second equality holds because

$$\int_0^\infty tg(t|a_1, a_2)\mathrm{d}t \quad = a_2(2a_1 + a_2)/\{2a_1^2(a_1 + a_2)^2\}$$

from Stryhn (1996). By setting $(c_1, c_2, \sigma_1, \sigma_2) = (\sigma^{(j)2}/2, \sigma^{(j)2}/2, \sigma^{(j)}, \sigma^{(j)})$ in (26) and (27), we have

$$\begin{aligned} &3/2 = E\{\sup_s V_s(\sigma^{(j)2}/2, \sigma^{(j)2}/2, \sigma^{(j)}, \sigma^{(j)})\} \\ &= E\{V_{\mathrm{argsup}_s V'_s(\sigma^{(j)2}/2,\sigma^{(j)2}/2,\sigma^{(j)},\sigma^{(j)})}(\sigma^{(j)2}/2, \sigma^{(j)2}/2, \sigma^{(j)}, \sigma^{(j)})\}. \end{aligned} \tag{28}$$

From (24), (25), and (28), we can obtain the theorem.

5.3 Proof of corollary

Similar to the independent case, when $\boldsymbol{k} - \boldsymbol{k}^* = O(1)$ without the condition (6), it follows that $\hat{\boldsymbol{\theta}}_{\boldsymbol{k},\boldsymbol{x}} = \hat{\boldsymbol{\theta}}_{\boldsymbol{k}^*,\boldsymbol{x}} + O_{\mathrm{P}}(n^{-1})$ and

$$
\begin{aligned}
&\log f(\boldsymbol{x}|\boldsymbol{k},\boldsymbol{\theta}^*) - \log f(\boldsymbol{x}|\boldsymbol{k}^*,\boldsymbol{\theta}^*) \\
&= \sum_{j=1}^{m}\Bigg[\mathrm{I}_{\{k^{(j)}<k^{*(j)}\}}\sum_{i=k^{(j)}+1}^{k^{*(j)}}\{(x_i-\boldsymbol{\theta}^{*(j)T}\boldsymbol{z}_i)^2-(x_i-\boldsymbol{\theta}^{*(j+1)T}\boldsymbol{z}_i)^2\}/(2\theta_0^*) \\
&\quad + I_{\{k^{(j)}>k^{*(j)}\}}\sum_{i=k^{*(j)}+1}^{k^{(j)}}\{(x_i-\boldsymbol{\theta}^{*(j+1)T}\boldsymbol{z}_i)^2-(x_i-\boldsymbol{\theta}^{*(j)T}\boldsymbol{z}_i)^2\}/(2\theta_0^*)\Bigg]+o(1) \\
&= \sum_{j=1}^{m} Q_{k,\boldsymbol{x}}^{AR(j)}+o(1).
\end{aligned}
$$

We can then easily obtain Theorem 1 in the same way.

Under the condition (6), $Q_{k^*+[s\alpha_n],\boldsymbol{x}}^{AR(j)}$ can be written as

$$
\begin{aligned}
&I_{\{s<0\}}\Bigg[\alpha_n^{-1/2}\sum_{i=k^{*(j)}+[s\alpha_n]+1}^{k^{*(j)}}\Delta_{\boldsymbol{\theta}^*}^{(j)T}\boldsymbol{z}_i\epsilon_i/\theta_0^* - \alpha_n^{-1}\sum_{i=k^{*(j)}+[s\alpha_n]+1}^{k^{*(j)}}(\Delta_{\boldsymbol{\theta}^*}^{(j)T}\boldsymbol{z}_i)^2/(2\theta_0^*)\Bigg] \\
&\quad + I_{\{s>0\}}\Bigg[\alpha_n^{-1/2}\sum_{i=k^{*(j)}+1}^{k^{*(j)}+[s\alpha_n]}-\Delta_{\boldsymbol{\theta}^*}^{(j)T}\boldsymbol{z}_i\epsilon_i/\theta_0^* - \alpha_n^{-1}\sum_{i=k^{*(j)}+1}^{k^{*(j)}+[s\alpha_n]}(\Delta_{\boldsymbol{\theta}^*}^{(j)T}\boldsymbol{z}_i)^2/(2\theta_0^*)\Bigg].
\end{aligned}
$$

By applying the functional central limit theorem for martingales to $\alpha_n^{-1/2}\sum \Delta_{\boldsymbol{\theta}^*}^{(j)T}\boldsymbol{z}_i\epsilon_i/\theta_0^*$, applying the uniform law of large numbers to $\alpha_n^{-1}\sum(\Delta_{\boldsymbol{\theta}^*}^{(j)T}\boldsymbol{z}_i)^2/(2\theta_0^*)$, and using the stationarity of $\boldsymbol{x}$, it follows that $Q_{k^*+[s\alpha_n],\boldsymbol{x}}^{AR(j)}$ converges in distribution to

$$
\{\mathrm{var}(\Delta_{\boldsymbol{\theta}^*}^{(j)T}\boldsymbol{z}_i)/\theta_0^*\}^{1/2}W_s - |s|\mathrm{var}(\Delta_{\boldsymbol{\theta}^*}^{(j)T}\boldsymbol{z}_i)/(2\theta_0^*).
$$

We can thus obtain Theorem 2.

References

Akaike, H. (1973). "Information Theory and an Extension of the Maximum Likelihood Principle", In B.N. Petrov and F. Csaki (Eds.) 2nd International Symposium on Information Theory, (pp. 267–281). Budapest: Akademiai Kiado.

Aue, A., Hörmann, S., Horváth, L., Reimherr, M. (2009). Break detection in the covariance structure of multivariate time series models. *The Annals of Statistics*, *37*, 4046–4087.

Avery, P. J., Henderson, D. A. (1999). Detecting a changed segment in DNA sequences. *Applied Statistics*, *48*, 489–503.

Bai, J., Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica*, *66*, 47–78.

Bhattacharya, P.K., Brockwell, P.J. (1976). The minimum of an additive process with applications to signal estimation and storage theory. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, *37*, 51–75.

Chen, J., Gupta, A.K. (1997). Testing and locating variance changepoints with application to stock prices. *Journal of the American Statistical Association*, *92*, 739–747.

Claeskens, G., Hjort, N.L. (2003). Focused information criterion (with discussion). *Journal of the American Statistical Association*, *98*, 900–916.
Csörgő, M., Horváth, L. (1997). *Limit Theorems in Change-Point Analysis*. New York:Wiley.
Dümbgen, L. (1991). The asymptotic behavior of some nonparametric change-point estimates. *The Annals of Statistics*, *19*, 1471–1495.
Foster, D.P., George, E.I. (1994). The risk inflation criterion for multiple regression. *The Annals of Statistics*, *22*, 1947–1975.
Garcia, R., Perron, P. (1996). An analysis of the real interest rate under regime shifts. *Review of Economics and statistics*, *78*, 111–125.
Haccou, P., Meelis, E. (1988). Testing for the number of change points in a sequence of exponential random variables. *Journal of Statistical Computation and Simulation*, *30*, 285–298.
Hamilton, J.D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica*, *57*, 357–384.
Hannart, A., Naveau, P. (2012). An improved Bayesian information criterion for multiple change-point models. *Technometrics*, *54*, 256–268.
Hsu, D.A. (1979). Detecting shifts of parameters in gamma sequences with applications to stock price and air traffic flow analysis. *Journal of the American Statistical Association*, *74*, 31–40.
Hurrell, J.W., Trenberth, K.E. (1997). Spurious trends in satellite MSU temperatures from merging different satellite records. *Nature*, *386*, 164–167.
Inclan, C., Tiao, G.C. (1994). Use of cumulative sums of squares for retrospective detection of changes of variance. *Journal of the American Statistical Association*, *89*, 913–923.
Jones, R.H., Dey, I. (1995). Determining one or more change points. *Chemistry and Physics of Lipids*, *76*, 1–6.
Kullback, S., Leibler, R.A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*, 79–86.
Ninomiya, Y. (2005). Information criterion for gaussian change-point model. *Statistics and Probability Letters*, *72*, 237–247.
Ninomiya, Y., Yoshimoto, A. (2008). Statistical method for detecting structural change in the growth process. *Biometrics*, *64*, 46–53.
Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464.
Shepp, L.A. (1979). The joint density of the maximum and its location for a wiener process with drift. *Journal of Applied Probability*, *16*, 423–427.
Siegmund, D.O. (2004). Model selection in irregular problems: applications to mapping quantitative trait loci. *Biometrika*, *91*, 785–800.
Spiegelhalter, D.J., Best, N.G., Carlin, B.P., van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B*, *64*, 1–34.
Stryhn, H. (1996). The location of the maximum of asymmetric two-sided Brownian motion with triangular drift. *Statistics and Probability Letters*, *29*, 279–284.
van der Vaart, A.W. (1998). *Asymptotic Statistics* Cambridge: Cambridge University Press.
Vostrikova, L.J. (1981). Detecting 'Disorder' in multidimensional random processes. *Soviet Mathematics Doklady*, *24*, 55–59.
Yao, Y.C. (1988). Estimating the number of change-points via Schwarz's criterion. *Statistics and Probability Letters*, *6*, 181–189.
Zhang, N.H., Siegmund, D.O. (2007). A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, *63*, 22–32.
Zou, H., Hastie, T., Tibshirani, R. (2007). On the "Degrees of Freedom" of the LASSO. *The Annals of Statistics*, *35*, 2173–2192.