Available online at www.sciencedirect.com

ScienceDirect

Journal of Multivariate Analysis 99 (2008) 2154–2171

Journal of Multivariate Analysis

www.elsevier.com/locate/jmva

Simultaneous change point analysis and variable selection in a regression problem

Y. Wu

Department of Mathematics and Statistics, York University, 4700 Keele Street, Toronto, Ontario, Canada M3J 1P3

Received 25 June 2007
Available online 12 February 2008

Abstract

In this paper, an information-based criterion is proposed for carrying out change point analysis and variable selection simultaneously in linear models with a possible change point. Under some weak conditions, this criterion is shown to be strongly consistent in the sense that with probability one, it chooses the smallest true model for large n. Its byproducts include strongly consistent estimates of the regression coefficients regardless if there is a change point. In case that there is a change point, its byproducts also include a strongly consistent estimate of the change point parameter. In addition, an algorithm is given which has significantly reduced the computation time needed by the proposed criterion for the same precision. Results from a simulation study are also presented.

AMS 2000 subject classifications: 62J05; 62F12

Keywords: Linear models; Variable selection; Change point analysis; Consistency; Computation

1. Introduction

A statistical model is convenient for representing the observed phenomenon. In practice, a linear regression model is often used to describe the data due to the fact that (a) it has been extensively studied in the literature because of its simplicity; and (b) it frequently mimics the real world well since it is well known that a nonlinear function can be approximated well by a linear function locally.

E-mail address: wuyh@mathstat.yorku.ca.

0047-259X/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.jmva.2008.02.010

Consider the following linear regression model

$$y_i = \boldsymbol{x}_i'\boldsymbol{\eta} + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ are vectors of explanatory variables, $\boldsymbol{\eta}$ is a vector of unknown regression parameters, and $\varepsilon_1, \varepsilon_2, \ldots,$ are independently distributed random variables. When the dimension of $\boldsymbol{\eta}$ is not small, by the rule of parsimony and for the better prediction, there is a need to find out those $\boldsymbol{x}_i$'s that are extraneous to $y_1, y_2, \ldots,$ which is equivalent to finding those elements of $\boldsymbol{\eta}$ that are zeros. A statistical analysis may be more effective if those independent variables with zero regression coefficients are not included in the study. There is considerable literature on this problem; see the book on model selection by McQuarrie and Tsai [6] or the review paper by Rao and Wu [10] among others.

However there may be a change point in the model. A change point problem occurs in many statistical applications in the areas including medical and health sciences, life science, meteorology, engineering, financial econometrics and risk management. By the fact that the statistical models are not homogeneous when there is a change point, to detect all change points are of great importance in statistical applications. If there exists a change point, it is harmful to make a statistical analysis without any consideration of the existence of this change point and the results derived from such an analysis may be misleading. The task of change point analysis is to find change points when they do exist. There are rich literature on change point analysis (see Csörgő and Horváth [4] and Chen and Gupta [3] among others). It is noted that many research articles in change point analysis mainly contributed to finding change points.

To allow a possible change point in the model (1), we modify it as follows: For $i = 1, \ldots, n, \ldots,$

$$y_i = (\mu_1 + \alpha_1 z_i + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_1)I(z_i \le \xi) + (\mu_2 + \alpha_2 z_i + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_2)I(z_i > \xi) + \varepsilon_i, \tag{2}$$

where $\mu_1, \mu_2, \alpha_1, \alpha_2, \boldsymbol{\beta}_1 = (\beta_{11}, \ldots, \beta_{1p})^{\mathrm{T}}$, and $\boldsymbol{\beta}_2 = (\beta_{21}, \ldots, \beta_{2p})^{\mathrm{T}}$ are unknown regression parameters, ξ is an unknown change point parameter, $\{(z_i, \boldsymbol{x}_i^{\mathrm{T}})\}$ is a sequence of explanatory variables with $z_i \in (\kappa_L, \kappa_U)$ for any i and $-\infty \le \kappa_L < \kappa_U \le \infty$ are known, $\varepsilon_1, \ldots, \varepsilon_n$ are independently and identically distributed random variables with mean 0 and variance σ^2, and $I(\cdot)$ is the indicator function. There exists a change point if there is a $\xi \in (\kappa_L, \kappa_U)$ such that

$$|\mu_1 - \mu_2| + |\alpha_1 - \alpha_2| + \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\| \ne 0, \tag{3}$$

where $\|\boldsymbol{a}\|$ denotes the Euclidean norm of a vector $\boldsymbol{a}$. When there is no change point, we put $\mu_1 = \mu_2 \equiv \mu$, $\alpha_1 = \alpha_2 \equiv \alpha$, and $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 \equiv \boldsymbol{\beta}$. The task in this paper consists of the following two parts:

(1) Check if there exists $\xi \in (\kappa_L, \kappa_U)$ such that (3) holds. If yes, estimate ξ;
(2) Find out all the nonzero components of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ or $\boldsymbol{\beta}$ and estimate them and other regression coefficients.

In other words, we need to carry out change point analysis, perform variable selection and estimate the regression coefficients simultaneously. Note that there will be $2^{2(p+2)}$ possible submodels in the frame of (2) for a fixed ξ, since each component of $\boldsymbol{\beta}_i, i = 1, 2$, may be zero.

One way to fulfill this task is to first carry out change point analysis on the model (2) including all independent variables and then implement variable selection. Thus change point analysis and variable selection can be carried out in two steps. However, by simulation study this procedure sometimes does not perform well. We will modify this two-step procedure in this paper. As an alternative approach, we will also consider to carry out change point analysis and variable selection simultaneously. An information-based criterion will be proposed, which will be shown to perform well via simulation study. Its limiting behavior will also be studied. For convenience, we only assume that there is at most one change point.

The paper is arranged as follows: In Section 2, we propose an information-based criterion for carrying out change point analysis and variable selection simultaneously. There are two subsections there. In Section 2.1, we consider the special case that the candidate models are nested. In Section 2.2, we discuss the general case and propose a modified criterion. The limiting behavior of the criteria are also derived there. The criteria are shown to be strongly consistent in the sense that with probability one, they choose the smallest true model for large n. In Section 3, we will give an algorithm for carrying out change point analysis and variable selection simultaneously. Some simulation results are presented in Section 4. Proofs of the theorems in Section 2 are given in the Appendix.

The following notations are used throughout the rest of this paper. Let $\boldsymbol{a} = (a_1, \ldots, a_\ell)^{\mathrm{T}}$ be an $\ell \times 1$ vector, $B = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_\ell)^{\mathrm{T}}$ be an $\ell \times m$ matrix, and $\mathcal{G}_k = \{j_1, \ldots, j_k\}$ with $1 \le j_1 < \ldots < j_k \le \ell$ be an index set. Denote the projection matrix onto the space spanned by the column vectors of B by P_B and if $B^{\mathrm{T}} = B$ and $\ell = m$, denote all the eigenvalues of B by $\lambda_1(B) \ge \ldots \ge \lambda_m(B)$. Write the number of elements in $\mathcal{G}_k$ by $|\mathcal{G}_k|$ and let $\boldsymbol{a}(\mathcal{G}_k) = (a_{j_1}, \ldots, a_{j_k})^{\mathrm{T}}$ and $B(\mathcal{G}_k) = (\boldsymbol{b}_{j_1}, \ldots, \boldsymbol{b}_{j_k})^{\mathrm{T}}$. In some cases $\mathcal{G}_k$ will be abbreviated as k for convenience. This should not cause confusion.

2. The criterion

Let $\mathcal{J}$ be a subset of $\{1, \ldots, p\}$, and $\mathfrak{J}$ be the set containing all such $\mathcal{J}$'s. Denote

$$\boldsymbol{\vartheta}_{1,\mathcal{J}} = (\mu_1, \alpha_1, \boldsymbol{\beta}_1^{\mathrm{T}}(\mathcal{J}))^{\mathrm{T}}, \qquad \boldsymbol{\vartheta}_{2,\mathcal{J}} = (\mu_2, \alpha_2, \boldsymbol{\beta}_2^{\mathrm{T}}(\mathcal{J}))^{\mathrm{T}}, \qquad \boldsymbol{\vartheta}_{\mathcal{J}} = (\mu, \alpha, \boldsymbol{\beta}^{\mathrm{T}}(\mathcal{J}))^{\mathrm{T}},$$
$$\mathcal{G}_{\xi,l} = \{i : z_i \le \xi\}, \qquad \mathcal{G}_{\xi,r} = \{i : z_i > \xi\}.$$

Note that n has been suppressed in $\mathcal{G}_{\xi,l}$ and $\mathcal{G}_{\xi,r}$.

Consider the model (2). We denote $\boldsymbol{y}_n = (y_1, \ldots, y_n)^{\mathrm{T}}$, and $\boldsymbol{\varepsilon}_n$ and $\boldsymbol{z}_n$ are defined similarly. $\mathbf{1}_n$ denotes an $n \times 1$ vector of 1's, $X_n = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^{\mathrm{T}}$, and $X_{n,\mathcal{J}} = (\boldsymbol{x}_1(\mathcal{J}), \ldots, \boldsymbol{x}_n(\mathcal{J}))^{\mathrm{T}}$. We also denote $Z_n = (\mathbf{1}_n, \boldsymbol{z}_n, X_n)$, and $Z_{n,\mathcal{J}} = (\mathbf{1}_n, \boldsymbol{z}_n, X_{n,\mathcal{J}})$. The subscript n will be suppressed if there is no confusion.

For $\kappa_L < \xi < \kappa_U$, $\mathcal{J}_1 \in \mathfrak{J}$ and $\mathcal{J}_2 \in \mathfrak{J}$, we define

$$\begin{aligned}\ell^{(\xi)}_{n,\mathcal{J}_1,\mathcal{J}_2} = {} & \sum_{i \in \mathcal{G}_{\xi,l}} (y_i - \hat{\mu}_{1,\mathcal{J}_1,\xi} - \hat{\alpha}_{1,\mathcal{J}_1,\xi} z_i - \boldsymbol{x}_i^{\mathrm{T}}(\mathcal{J}_1)\widehat{\boldsymbol{\beta}}_{1,\xi}(\mathcal{J}_1))^2 \\ & + \sum_{i \in \mathcal{G}_{\xi,r}} (y_i - \hat{\mu}_{2,\mathcal{J}_2,\xi} - \hat{\alpha}_{2,\mathcal{J}_2,\xi} z_i - \boldsymbol{x}_i^{\mathrm{T}}(\mathcal{J}_2)\widehat{\boldsymbol{\beta}}_{2,\xi}(\mathcal{J}_2))^2, \end{aligned} \tag{4}$$

where $\hat{\boldsymbol{\vartheta}}_{1,\mathcal{J}_1,\xi} = (\hat{\mu}_{1,\mathcal{J}_1,\xi}, \hat{\alpha}_{1,\mathcal{J}_1,\xi}, \widehat{\boldsymbol{\beta}}^{\mathrm{T}}_{1,\xi}(\mathcal{J}_1))^{\mathrm{T}}$ and $\hat{\boldsymbol{\vartheta}}_{2,\mathcal{J}_2,\xi} = (\hat{\mu}_{2,\mathcal{J}_2,\xi}, \hat{\alpha}_{2,\mathcal{J}_2,\xi}, \widehat{\boldsymbol{\beta}}^{\mathrm{T}}_{2,\xi}(\mathcal{J}_2))^{\mathrm{T}}$ are the least squares (LS) estimators of $\boldsymbol{\vartheta}_{1,\mathcal{J}_1}$ and $\boldsymbol{\vartheta}_{2,\mathcal{J}_2}$ based on $(\boldsymbol{y}_n(\mathcal{G}_{\xi,l}), Z_{n,\mathcal{J}_1}(\mathcal{G}_{\xi,l}))$ and $(\boldsymbol{y}_n(\mathcal{G}_{\xi,r}), Z_{n,\mathcal{J}_2}(\mathcal{G}_{\xi,r}))$, respectively. For $\mathcal{J} \in \mathfrak{J}$, we also define

$$\ell_{n,\mathcal{J}} = \sum_{i=1}^{n} (y_i - \hat{\mu}_{\mathcal{J}} - \hat{\alpha}_{\mathcal{J}} z_i - \boldsymbol{x}_i^{\mathrm{T}}(\mathcal{J})\widehat{\boldsymbol{\beta}}(\mathcal{J}))^2,$$

where $\hat{\boldsymbol{\vartheta}}_{\mathcal{J}} = (\hat{\mu}_{\mathcal{J}}, \hat{\alpha}_{\mathcal{J}}, \widehat{\boldsymbol{\beta}}^{\mathrm{T}}(\mathcal{J}))^{\mathrm{T}}$ is the LS estimator of $\boldsymbol{\vartheta}_{\mathcal{J}}$ based on $(\boldsymbol{y}_n, Z_{n,\mathcal{J}})$. Note that n has been suppressed in the notations above. Hence, we have

$$\begin{cases} [Z_{n,\mathcal{J}_1}(\mathcal{G}_{\xi,l})]\hat{\boldsymbol{\vartheta}}_{1,\mathcal{J}_1,\xi} = P_{Z_{n,\mathcal{J}_1}(\mathcal{G}_{\xi,l})}\boldsymbol{y}_n(\mathcal{G}_{\xi,l}), \\ [Z_{n,\mathcal{J}_2}(\mathcal{G}_{\xi,r})]\hat{\boldsymbol{\vartheta}}_{2,\mathcal{J}_2,\xi} = P_{Z_{n,\mathcal{J}_2}(\mathcal{G}_{\xi,r})}\boldsymbol{y}_n(\mathcal{G}_{\xi,r}), \\ Z_{n,\mathcal{J}}\hat{\boldsymbol{\vartheta}}_{\mathcal{J}} = P_{Z_{n,\mathcal{J}}}\boldsymbol{y}_n. \end{cases} \tag{5}$$

By Assumption (A) given later, it can be shown that $\hat{\boldsymbol{\vartheta}}_{1,\mathcal{J}_1,\xi}$, $\hat{\boldsymbol{\vartheta}}_{2,\mathcal{J}_2,\xi}$, and $\hat{\boldsymbol{\vartheta}}_{\mathcal{J}}$ are unique for large n.

Motivated by Rao and Wu [9], let

$$\mathrm{SITC}^{(\xi)}_{n,\mathcal{J}_1,\mathcal{J}_2} = \ell^{(\xi)}_{n,\mathcal{J}_1,\mathcal{J}_2} + [q(|\mathcal{J}_1|+2) + q(|\mathcal{J}_2|+2)]C_n, \tag{6}$$

$$\mathrm{SITC}_{n,\mathcal{J}} = \ell_{n,\mathcal{J}} + q(|\mathcal{J}|+2)C_n, \tag{7}$$

where $q(v)$ is a strictly increasing function of v and C_n is a function of only n. It is noted that the second terms in both (6) and (7) are the penalties on the use of models involving more parameters.

For carrying out change point analysis and variable selection simultaneously, we propose the following criterion based on $\mathrm{SITC}^{(\xi)}_{n,\mathcal{J}_1,\mathcal{J}_2}$ and $\mathrm{SITC}_{n,\mathcal{J}}$: If

$$\min_{\kappa_L<\xi<\kappa_U,\ \mathcal{J}_1\in\mathfrak{J},\ \mathcal{J}_2\in\mathfrak{J}} \mathrm{SITC}^{(\xi)}_{n,\mathcal{J}_1,\mathcal{J}_2} < \min_{\mathcal{J}\in\mathfrak{J}} \mathrm{SITC}_{n,\mathcal{J}},$$

then we conclude there is a change point and the parameter estimates are given by

$$(\hat{\xi}, \widehat{\mathcal{J}}_1, \widehat{\mathcal{J}}_2, \hat{\boldsymbol{\vartheta}}_{1,\widehat{\mathcal{J}}_1,\hat{\xi}}, \hat{\boldsymbol{\vartheta}}_{2,\widehat{\mathcal{J}}_2,\hat{\xi}}) = \arg \min_{\kappa_L<\xi<\kappa_U,\ \mathcal{J}_1\in\mathfrak{J},\ \mathcal{J}_2\in\mathfrak{J}} \mathrm{SITC}^{(\xi)}_{n,\mathcal{J}_1,\mathcal{J}_2},$$

otherwise, there is no change point and the parameter estimates are given by

$$(\widehat{\mathcal{J}}, \hat{\boldsymbol{\vartheta}}_{\widehat{\mathcal{J}}}) = \arg \min_{\mathcal{J}\in\mathfrak{J}} \mathrm{SITC}_{n,\mathcal{J}}.$$

We name this criterion as the criterion SITC. It is noted that ITC stands for "*i*nformation *t*heoretic *c*riterion" in the literature. Here we add "S" to "ITC" to reflect that the criterion ITC is used to carry out change point analysis and variable selection *s*imultaneously.

Remark 1. Let $\{z_{(1)}, \ldots, z_{(n)}\}$ be the order statistics of $\{z_1, \ldots, z_n\}$. It is noted that if n, $\mathcal{J}_1$ and $\mathcal{J}_2$ are fixed, $\mathrm{SITC}^{(\xi)}_{n,\mathcal{J}_1,\mathcal{J}_2}$ remains unchanged for $z_{(j)} \le \xi < z_{(j+1)}$.

Remark 2. When there are more than one change point, a step-wise or forward search may be employed. One may also follow the approach in Pan and Chen [7].

For deriving the limiting behavior of Criterion SITC, we need to make the following assumptions.

(A) For any $\kappa_L \le \xi_1 < \xi_2 \le \kappa_U$, there exist two constants c_1 and c_2 such that

$$\begin{aligned} 0 < c_1 n &\le \lambda_p \left\{[Z_n(\mathcal{G}_{\xi_1,r} \cap \mathcal{G}_{\xi_2,l})]^{\mathrm{T}} Z_n(\mathcal{G}_{\xi_1,r} \cap \mathcal{G}_{\xi_2,l})\right\} \\ &\le \lambda_1 \left\{[Z_n(\mathcal{G}_{\xi_1,r} \cap \mathcal{G}_{\xi_2,l})]^{\mathrm{T}} Z_n(\mathcal{G}_{\xi_1,r} \cap \mathcal{G}_{\xi_2,l})\right\} \le c_2 n, \quad \text{for large } n. \end{aligned}$$

Remark 3. Assumption (A) describes essentially the behavior of the explanatory variables. It is noted that for Assumption (A) to hold, a necessary condition is that for any subinterval (ξ_1, ξ_2) of (κ_L, κ_U), the number of $z_i \in (\xi_1, \xi_2)$ for $1 \le i \le n$ tends to ∞ as $n \to \infty$. This condition is met almost surely if $z_1, z_2, \ldots,$ are independently and identically distributed (iid.) such that $P(\xi_1 < z_1 < \xi_2) > 0$ for any $\kappa_L \le \xi_1 < \xi_2 \le \kappa_U$. Write $\tilde{\boldsymbol{x}}_i = (z_i, \boldsymbol{x}_i^{\mathrm{T}})^{\mathrm{T}}$, $i = 1, 2, \ldots$. Assume that in addition, $\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_n, \ldots$ are independently and identically distributed such that $E\tilde{\boldsymbol{x}}_1\tilde{\boldsymbol{x}}_1^{\mathrm{T}} > 0$ (positive definite). Then these assumptions are sufficient for (A) to hold almost surely, which can be easily verified by the strong law of large numbers. For ease of notation, we will treat $z_1, \ldots, z_n$ and $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ as deterministic in this paper. There is no essential complication with random z_i's and $\boldsymbol{x}_i$'s.

(B) $\varepsilon_1, \ldots, \varepsilon_n$ are independently and identically distributed random variables with mean 0 and variance σ^2.

Remark 4. The Assumption (B) can be weakened. For example, we may instead assume that $\varepsilon_1, \ldots, \varepsilon_n$ are independently distributed random variables with mean 0 and satisfying the moment conditions

$$0 < c^2 \le E(\varepsilon_i^2), \quad E(|\varepsilon_i|^3) \le \tau^3 < \infty$$

for all $1 \le i \le n$. For convenience, the Assumption (B) will be made throughout the rest of this paper.

(C) For $\kappa_L < \xi < \kappa_U$, there exists $M > 0$ such that

$$[\boldsymbol{\varepsilon}(\mathcal{G}_{\xi,l})]^{\mathrm{T}} P_{Z_{\xi,l}} \boldsymbol{\varepsilon}(\mathcal{G}_{\xi,l}) \le M \log\log(n), \quad \text{a.s.,}$$
$$[\boldsymbol{\varepsilon}(\mathcal{G}_{\xi,r})]^{\mathrm{T}} P_{Z_{\xi,r}} \boldsymbol{\varepsilon}(\mathcal{G}_{\xi,r}) \le M \log\log(n), \quad \text{a.s.}$$

Remark 5. If ε_i, $i = 1, \ldots, n$, are normally distributed with zero mean and common variance $\sigma^2 > 0$, then (C) holds true under Assumption (A). By applying some theorems (e.g. Theorem 5.7, Theorem 7.2) in Petrov [8], (C) holds true under some weak conditions on $\{\varepsilon_i\}$ and Z_n, e.g., (A) and (B) plus that $E(|\varepsilon_1|^{2+\iota}) < \infty$ for a constant $\iota > 0$ and that for any $\xi_1 < \xi_2$ and any jth column vector $\boldsymbol{z}_{j,\xi_1,\xi_2}$ of $Z_n(\mathcal{G}_{\xi_1,r} \cap \mathcal{G}_{\xi_2,l})$ with $j \ge 2$, the elements $z_{j,\xi_1,\xi_2}^1, \ldots, z_{j,\xi_1,\xi_2}^{|\mathcal{G}_{\xi_1,r}\cap\mathcal{G}_{\xi_2,l}|}$ of $\boldsymbol{z}_{j,\xi_1,\xi_2}$ satisfy the condition

$$\sum_{i=1}^{|\mathcal{G}_{\xi_1,r}\cap\mathcal{G}_{\xi_2,l}|} |z_{j,\xi_1,\xi_2}^i|^{2+\delta} = O\left[(\boldsymbol{z}_{j,\xi_1,\xi_2}^{\mathrm{T}}\boldsymbol{z}_{j,\xi_1,\xi_2})^{(2+\delta)/2}/[\log(\boldsymbol{z}_{j,\xi_1,\xi_2}^{\mathrm{T}}\boldsymbol{z}_{j,\xi_1,\xi_2})]^{1+\delta}\right],$$

for $1 \le j \le p$ and some $\delta > 0$, which can be shown by following the proofs of Lemmas 3.4–3.5 in Shao and Wu [11]

(D) $C_n/n \to 0$ and $C_n/\log\log(n) \to \infty$.

Throughout the rest of this paper, we assume that $\mathcal{J}_1 = \mathcal{J}_2 = \mathcal{J}$, i.e., the ith components of both $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are zeros or nonzeros simultaneously for convenience. Thus, for a fixed ξ, there are 2^{p+4} possible submodels in the frame of (2). For simplifying notations given previously, "$\mathcal{J}_1, \mathcal{J}_2$" is replaced by "$\mathcal{J}$", e.g., $\mathrm{SITC}_{n,\mathcal{J}_1,\mathcal{J}_2}^{(\xi)}$ in (6) is now written as $\mathrm{SITC}_{n,\mathcal{J}}^{(\xi)}$. It is noted that without this assumption, similar results as those obtained in the following two subsections can be shown to still hold under the Assumptions (A)–(D).

2.1. A special case that the candidate models are nested

In this subsection, we consider the case that the candidate models are nested, i.e., there are p alternative models $\mathfrak{M} = \{\mathcal{M}_1, \ldots, \mathcal{M}_p\}$ such that under Model $\mathcal{M}_k$, the last $p - k$ components of both $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are zeros but $\beta_{1k} \neq 0$ and $\beta_{2k} \neq 0$ if there exists a change point, and the last $p - k$ components of $\boldsymbol{\beta}$ are zeros and $\beta_k \neq 0$ if there is no change point.

Recall that if $\mathcal{J} = (1, \ldots, k)^{\mathrm{T}}$, $\mathcal{J}$ is abbreviated as k. For carrying out change point analysis and variable selection simultaneously here, the criterion SITC is reduced to as follows: If

$$\min_{\kappa_L < \xi < \kappa_U,\ 1 \le k \le p} \mathrm{SITC}^{(\xi)}_{n,k} < \min_{1 \le k \le p} \mathrm{SITC}_{n,k},$$

then we conclude there exists a change point and the parameter estimates are given by

$$(\hat{\xi}, \widehat{k}, \hat{\boldsymbol{\vartheta}}_{1,\widehat{k},\hat{\xi}}, \hat{\boldsymbol{\vartheta}}_{2,\widehat{k},\hat{\xi}}) = \arg \min_{\kappa_L < \xi < \kappa_U,\ 1 \le k \le p} \mathrm{SITC}^{(\xi)}_{n,k},$$

otherwise, there is no change point and the parameter estimates are given by

$$(\widehat{k}, \hat{\boldsymbol{\vartheta}}_{\widehat{k}}) = \arg \min_{1 \le k \le p} \mathrm{SITC}_{n,k}.$$

Consider the following two scenarios:

(S1) There is a change point at ξ_0 and hence (3) holds. For the true parameter vectors $\boldsymbol{\vartheta}_{1,0} = (\mu_{1,0}, \alpha_{1,0}, \boldsymbol{\beta}^{\mathrm{T}}_{1,0})^{\mathrm{T}}$ and $\boldsymbol{\vartheta}_{2,0} = (\mu_{2,0}, \alpha_{2,0}, \boldsymbol{\beta}^{\mathrm{T}}_{2,0})^{\mathrm{T}}$, $\beta_{1k_0,0}\beta_{2k_0,0} \neq 0$ and the last $p - k_0$ components of both $\boldsymbol{\beta}_{1,0}$ and $\boldsymbol{\beta}_{2,0}$ are zeros.

(S2) There is no change point. For the true parameter vector $\boldsymbol{\vartheta}_0 = (\mu_0, \alpha_0, \boldsymbol{\beta}^{\mathrm{T}}_0)^{\mathrm{T}}$, $\beta_{k_0,0} \neq 0$ and the last $p - k_0$ components of $\boldsymbol{\beta}_0$ are zeros.

The limiting behavior of Criterion SITC is given in the following two theorems:

Theorem 2.1. *Suppose that the Assumptions* (A)–(D) *hold. Under the scenario* (S1) *, for any* ξ_1 *and* ξ_2 *such that* $\kappa_L < \xi_1 < \xi_0$ *and* $\xi_0 < \xi_2 < \kappa_U$*, we have that with probability one,*

$$\mathrm{SITC}^{(\xi_0)}_{n,k_0} < \min_{\kappa_L < \xi \le \xi_1,\ 1 \le k \le p} \mathrm{SITC}^{(\xi)}_{n,k}, \tag{8}$$

$$\mathrm{SITC}^{(\xi_0)}_{n,k_0} < \min_{\xi_2 \le \xi < \kappa_U,\ 1 \le k \le p} \mathrm{SITC}^{(\xi)}_{n,k}, \tag{9}$$

$$\mathrm{SITC}^{(\xi_0)}_{n,k_0} < \min_{1 \le k \le p} \mathrm{SITC}_{n,k}, \tag{10}$$

$$\mathrm{SITC}^{(\xi_0)}_{n,k_0} < \min_{k \neq k_0} \mathrm{SITC}^{(\xi_0)}_{n,k} \tag{11}$$

for large n.

Theorem 2.2. *Suppose that the Assumptions* (A)–(D) *hold. Under the scenario* (S2) *, we have that with probability one,*

$$\mathrm{SITC}_{n,k_0} < \min_{\kappa_L < \xi < \kappa_U,\ 1 \le k \le p} \mathrm{SITC}^{(\xi)}_{n,k}, \tag{12}$$

$$\mathrm{SITC}_{n,k_0} < \min_{k \neq k_0} \mathrm{SITC}_{n,k} \tag{13}$$

for large n.

Proofs of these two theorems are given in the appendix.

By Theorems 2.1 and 2.2, it can be seen that if Assumptions (A)–(D) hold, we have that $\hat{\xi} \to \xi_0$, a.s. and $\hat{k} \to k_0$, a.s., and hence $\hat{\boldsymbol{\beta}}_1 \to \boldsymbol{\beta}_{1,0}$, a.s. and $\hat{\boldsymbol{\beta}}_2 \to \boldsymbol{\beta}_{2,0}$, a.s. under the scenario (S1), or that $\hat{k} \to k_0$, a.s., and hence $\hat{\boldsymbol{\beta}} \to \boldsymbol{\beta}_0$, a.s. under the scenario (S2).

2.2. *The general case*

In this section, the general case is considered, in which there are $2^{(p+4)}$ candidate models for a fixed ξ since each pair of β_{1j} and β_{2j} may be zero for $j = 1, \ldots, p$. Hence, when SITC is applied for performing a change point analysis and variable selection simultaneously, it is equivalent to the following:

> For true $\boldsymbol{\beta}_{1,0}$ and $\boldsymbol{\beta}_{2,0}$ ($\boldsymbol{\beta}_{1,0} = \boldsymbol{\beta}_{2,0}$ if there is no change point), rearranging their elements such that their first k_0 elements are not zeros and the rest elements are zeros. The columns of the design matrix Z_n are also rearranged accordingly. It ends with an equivalent regression model whose smallest true model is one of the submodels $\{\mathcal{M}_1, \ldots, \mathcal{M}_p\}$ given in the above subsection, and then, the simplified Criterion SITC can be applied. Select the model with the smallest SITC value among the SITC values computed based on all rearrangements and then reverse the corresponding rearrangement. Since the Assumptions (A)–(C) do not change under the rearrangement, the estimated model is still consistent under the assumptions (A)–(D) and hence the simultaneous correct change point detection and variable selection are eventually achieved.

However, this approach is involved with a huge amount of computation if p is large. We propose to modify the criterion SITC by adapting the kick-one-off approach (see Rao and Wu [9]) such that the computation amount can be reduced.

For a p-vector $\boldsymbol{a} = (a_1, \ldots, a_p)^{\mathrm{T}}$, denote

$$\boldsymbol{a}(-j) = (a_1, \ldots, a_{j-1}, a_{j+1}, \ldots, a_p)^{\mathrm{T}}, \quad 1 \le j \le p.$$

Consider the model (2). Write

$$X_{n,-j} = (\boldsymbol{x}_1(-j), \ldots, \boldsymbol{x}_n(-j))^{\mathrm{T}}, \quad \text{and} \quad Z_{n,-j} = (\mathbf{1}_n, \boldsymbol{z}_n, X_{n,-j}), \quad 1 \le j \le p.$$

Denote

$$\boldsymbol{\vartheta}_{1,-j} = (\mu_1, \alpha_1, \boldsymbol{\beta}_1^{\mathrm{T}}(-j))^{\mathrm{T}}, \quad \boldsymbol{\vartheta}_{2,-j} = (\mu_2, \alpha_2, \boldsymbol{\beta}_2^{\mathrm{T}}(-j))^{\mathrm{T}}, \quad \boldsymbol{\vartheta}_{-j} = (\mu, \alpha, \boldsymbol{\beta}^{\mathrm{T}}(-j))^{\mathrm{T}}.$$

For $\kappa_L < \xi < \kappa_U$, we define

$$\begin{aligned} \ell_{n,-j}^{(\xi)} = & \sum_{i \in \mathcal{G}_{\xi,l}} (y_i - \hat{\mu}_{1,-j,\xi} - \hat{\alpha}_{1,-j,\xi} z_i - \boldsymbol{x}_i^{\mathrm{T}}(-j) \widehat{\boldsymbol{\beta}}_{1,\xi}(-j))^2 \\ & + \sum_{i \in \mathcal{G}_{\xi,r}} (y_i - \hat{\mu}_{2,-j,\xi} - \hat{\alpha}_{2,-j,\xi} z_i - \boldsymbol{x}_i^{\mathrm{T}}(-j) \widehat{\boldsymbol{\beta}}_{2,\xi}(-j))^2, \end{aligned} \tag{14}$$

where $\hat{\boldsymbol{\vartheta}}_{1,-j,\xi} = (\hat{\mu}_{1,-j,\xi}, \hat{\alpha}_{1,-j,\xi}, \widehat{\boldsymbol{\beta}}_{1,\xi}^{\mathrm{T}}(-j))^{\mathrm{T}}$ and $\hat{\boldsymbol{\vartheta}}_{2,-j,\xi} = (\hat{\mu}_{2,-j,\xi}, \hat{\alpha}_{2,-j,\xi}, \widehat{\boldsymbol{\beta}}_{2,\xi}^{\mathrm{T}}(-j))^{\mathrm{T}}$ are the LS estimators of $\boldsymbol{\vartheta}_{1,-j}$ and $\boldsymbol{\vartheta}_{2,-j}$ based on $(\boldsymbol{y}_n(\mathcal{G}_{\xi,l}), Z_{n,-j}(\mathcal{G}_{\xi,l}))$ and $(\boldsymbol{y}_n(\mathcal{G}_{\xi,r}), Z_{n,-j}(\mathcal{G}_{\xi,r}))$, respectively. We also define

$$\ell_{n,-j} = \sum_{i=1}^{n} (y_i - \hat{\mu}_{-j} - \hat{\alpha}_{-j} z_i - \boldsymbol{x}_i^{\mathrm{T}}(-j) \widehat{\boldsymbol{\beta}}(-j))^2,$$

where $\hat{\boldsymbol{\vartheta}}_{-j} = (\hat{\mu}_{-j}, \hat{\alpha}_{-j}, \widehat{\boldsymbol{\beta}}^{\mathrm{T}}(-j))^{\mathrm{T}}$ is the LS estimator of $\boldsymbol{\vartheta}_{-j}$ based on $(\boldsymbol{y}_n, Z_{n,-j})$. Note that n has been suppressed in the notations above. Under Assumption (A), it can be shown that $\hat{\boldsymbol{\vartheta}}_{1,-j,\xi}$, $\hat{\boldsymbol{\vartheta}}_{2,-j,\xi}$, and $\hat{\boldsymbol{\vartheta}}_{-j}$ are unique for large n.

In light of Rao and Wu [9], let

$$\mathrm{KSITC}^{(\xi)}_{n,-j} = \ell^{(\xi)}_{n,-j} - \ell^{(\xi)}_{n} - 2[q(p) - q(p-1)]C_n, \tag{15}$$

$$\mathrm{KSITC}_{n,-j} = \ell_{n,-j} - \ell_{n} - [q(p) - q(p-1)]C_n, \tag{16}$$

and let $\widehat{\mathcal{J}}^{(n)}$ and $\widehat{\mathcal{J}}^{(n)}_{\xi}$, $\kappa_L < \xi < \kappa_U$, be subsets of $\{1, \ldots, p\}$, which are defined respectively by

$$j \notin \widehat{\mathcal{J}}^{(n)} \quad \text{if } \mathrm{KSITC}_{n,-j} \le 0 \quad \text{and} \quad j \in \widehat{\mathcal{J}}^{(n)} \quad \text{if } \mathrm{KSITC}_{n,-j} > 0, \quad j = 1, \ldots, p,$$

and

$$j \notin \widehat{\mathcal{J}}^{(n)}_{\xi} \quad \text{if } \mathrm{KSITC}^{(\xi)}_{n,-j} \le 0 \quad \text{and} \quad j \in \widehat{\mathcal{J}}^{(n)}_{\xi} \quad \text{if } \mathrm{KSITC}^{(\xi)}_{n,-j} > 0, \quad j = 1, \ldots, p.$$

For simultaneous change point detection and variable selection, we propose the following criterion: If

$$\min_{\kappa_L < \xi < \kappa_U} \mathrm{SITC}^{(\xi)}_{n,\widehat{\mathcal{J}}^{(n)}_{\xi}} < \mathrm{SITC}_{n,\widehat{\mathcal{J}}^{(n)}}, \tag{17}$$

then we conclude there is a change point and the parameter estimates are given by

$$(\hat{\xi}, \widehat{\mathcal{J}}^{(n)}_{\hat{\xi}}, \hat{\boldsymbol{\vartheta}}_{1,\widehat{\mathcal{J}}^{(n)}_{\hat{\xi}}}, \hat{\boldsymbol{\vartheta}}_{2,\widehat{\mathcal{J}}^{(n)}_{\hat{\xi}}}) = \arg \min_{\kappa_L < \xi < \kappa_U} \mathrm{SITC}^{(\xi)}_{n,\widehat{\mathcal{J}}^{(n)}_{\xi}},$$

otherwise, there is no change point and the parameter estimates are given by $(\widehat{\mathcal{J}}^{(n)}, \hat{\boldsymbol{\vartheta}}_{\widehat{\mathcal{J}}^{(n)}})$. We name this criterion as Criterion KSITC. It is noted that "K" stands for the first letter of *k*ick-one-off approach.

Consider the following two scenarios:

(S3) There is a change point at ξ_0 and hence (3) holds. For the true parameter vectors $\boldsymbol{\vartheta}_{1,0} = (\mu_{1,0}, \alpha_{1,0}, \boldsymbol{\beta}^{\mathrm{T}}_{1,0})^{\mathrm{T}}$ and $\boldsymbol{\vartheta}_{2,0} = (\mu_{2,0}, \alpha_{2,0}, \boldsymbol{\beta}^{\mathrm{T}}_{2,0})^{\mathrm{T}}$, $\beta_{1j,0} \times \beta_{2j,0} \ne 0$ for any $j \in \mathcal{J}_0$ and $|\beta_{1j,0}| + |\beta_{2j,0}| = 0$ for any $j \notin \mathcal{J}_0$, where $\mathcal{J}_0$ is an index set.

(S4) There is no change point. For the true parameter vector $\boldsymbol{\vartheta}_0 = (\mu_0, \alpha_0, \boldsymbol{\beta}^{\mathrm{T}}_0)^{\mathrm{T}}$, $\beta_{j,0} \ne 0$ if and only $j \in \mathcal{J}_0$, where $\mathcal{J}_0$ is an index set.

The limiting behavior of Criterion KSITC is given in the following two theorems:

Theorem 2.3. *Suppose that the Assumptions* (A)–(D) *hold. Under the scenario* (S3) *, for any ξ_1 and ξ_2 such that $\kappa_L < \xi_1 < \xi_0$ and $\xi_0 < \xi_2 < \kappa_U$, we have that with probability one,*

$$\mathrm{SITC}^{(\xi_0)}_{n,\mathcal{J}_0} < \min_{\kappa_L < \xi \le \xi_1} \mathrm{SITC}^{(\xi)}_{n,\widehat{\mathcal{J}}^{(n)}_{\xi}},$$

$$\mathrm{SITC}^{(\xi_0)}_{n,\mathcal{J}_0} < \min_{\xi_2 \le \xi < \kappa_U} \mathrm{SITC}^{(\xi)}_{n,\widehat{\mathcal{J}}^{(n)}_{\xi}},$$

$$\mathrm{SITC}^{(\xi_0)}_{n,\mathcal{J}_0} < \mathrm{SITC}_{n,\widehat{\mathcal{J}}^{(n)}},$$

$$\mathrm{SITC}^{(\xi_0)}_{n,\mathcal{J}_0} < \min_{\mathcal{J} \ne \mathcal{J}_0} \mathrm{SITC}^{(\xi_0)}_{n,\mathcal{J}},$$

when n is large.

Theorem 2.4. *Suppose that the Assumptions* (A)–(D) *hold. Under the scenario* (S4) , *we have that with probability one,*

$$\mathrm{SITC}_{n,\mathcal{J}_0} < \min_{\kappa_L<\xi<\kappa_U} \mathrm{SITC}^{(\xi)}_{n,\widehat{\mathcal{J}}^{(n)}_{\xi}},$$
$$\mathrm{SITC}_{n,\mathcal{J}_0} \le \mathrm{SITC}_{n,\widehat{\mathcal{J}}^{(n)}},$$

when n is large.

Proofs of these two theorems are similar to the proofs of Theorems 2.1 and 2.2 and hence are omitted.

By Theorems 2.3 and 2.4, it can be seen that if Assumptions (A)–(D) hold, we have that $\hat{\xi} \to \xi_0$, a.s. and $\widehat{\mathcal{J}}^{(n)}_{\hat{\xi}} \to \mathcal{J}_0$, a.s., and hence $\hat{\boldsymbol{\beta}}_1 \to \boldsymbol{\beta}_{1,0}$, a.s. and $\hat{\boldsymbol{\beta}}_2 \to \boldsymbol{\beta}_{2,0}$, a.s. under the scenario (S3), and that $\widehat{\mathcal{J}}^{(n)} \to \mathcal{J}_0$, a.s., and hence $\hat{\boldsymbol{\beta}} \to \boldsymbol{\beta}_0$, a.s. under the scenario (S4).

Remark 6. If $\mathcal{J}_1$ and $\mathcal{J}_2$ are allowed to be different, (15) and (16) need to be modified accordingly. The similar theoretic results can be shown to hold under the Assumptions (A)–(D).

Remark 7. If p is large compared to n, the criteria proposed in this paper need to be modified. One way to do it is to replace the penalty term by a function which depends on the number of regression parameters under consideration and also their magnitudes. Some such penalty functions can be found in Fan and Li [5] among others.

3. An algorithm for simultaneously carrying out change point analysis and variable selection

When the sample size is large, it may not be practical to check every data point in search of a change point. In the following, we propose an algorithm for carrying out change point analysis and variable selection simultaneously, which takes advantage of the maximization stage in Criterion KSITC proposed previously.

Without loss of generality, we may assume that $z_1 \le z_2 \le \cdots \le z_{n-1} \le z_n$. Otherwise, we can replace $\{z_i\}$ by the order statistics $\{z_{(i)}\}$. We then rearrange $\{y_i\}$, $\{\boldsymbol{x}_i\}$ and $\{\varepsilon_i\}$ accordingly. In view of Remark 1, by this assumption, we now need to find if there is an $m_0 < n$ such that

$$y_i = \mu_1 + \alpha_1 z_i + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_1 + \varepsilon_i, \quad i = 1, \ldots, m_0,$$
$$y_i = \mu_2 + \alpha_2 z_i + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_2 + \varepsilon_i, \quad i = m_0 + 1, \ldots, n.$$

Otherwise,

$$y_i = \mu + \alpha z_i + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \ldots, n.$$

For convenience, we replace z_i in (17) by i as follows:

$$\mathrm{SITC}^{(z_i)}_{n,\widehat{\mathcal{J}}^{(n)}_{z_i}} \equiv \mathrm{SITC}^{(i)}_{n,\widehat{\mathcal{J}}^{(n)}_{i}}. \tag{18}$$

The proposed algorithm consists of the following seven steps:

Step 1: Compute $\mathrm{SITC}_{n,\widehat{\mathcal{J}}^{(n)}}$ in (17).

Step 2: Set $l_1 = 2 \times (p + 2)$ and $l_2 = n - l_1$. Select ℓ_1 points $\{h_{1,j},\ j = 2, \ldots, \ell_1 + 1\}$ from $\{l_1 + 1, l_1 + 2, \ldots, l_2 - 2, l_2 - 1\}$ such that $l_1 < h_{1,2} < \ldots < h_{1,(\ell_1+1)} < l_2$. Let $h_{1,1} = l_1$ and $h_{1,(\ell_1+2)} = l_2$.

Remark 8. One way to select $\{h_{1,j},\ j = 2, \ldots, \ell + 1\}$ is as follows:

(a) Equally divide the interval (l_1, l_2) into $\ell_1 + 1$ subintervals;
(b) $h_{1,2}, \ldots, h_{1,(\ell_1+1)}$ are then set as the closest integers to the end points of these subintervals excluding l_1 and l_2.

Step 3: Compute $\text{SITC}^{(h)}_{n,\widehat{\mathcal{J}}^{(n)}_h}$ in (18) for $h \in \{h_{1,j},\ j = 2, \ldots, \ell_1 + 1\}$. Find

$$m_1 = \arg \min_{2 \le j \le \ell_1+1} \text{SITC}^{(h_{1,j})}_{n,\widehat{\mathcal{J}}^{(n)}_{h_{1,j}}}.$$

Set $q = 2$.

Step 4: Let $h_{q,1} = h_{(q-1),(m_{q-1}-1)}$, $h_{q,4} = h_{(q-1),m_{q-1}}$, and $h_{q,7} = h_{(q-1),(m_{q-1}+1)}$. Choose two integers $h_{q,2}$ and $h_{q,3}$ from the interval $(h_{q,1}, h_{q,4}]$ and choose another two integers $h_{q,5}$ and $h_{q,6}$ from the interval $[h_{q,4}, h_{q,7})$. Note that $h_{q,i},\ i = 2, 3, 5, 6$, may be chosen by the method given in Remark 8.

Step 5: Compute $\text{SITC}^{(h)}_{n,\widehat{\mathcal{J}}^{(n)}_h}$ in (18) for $h \in \{h_{q,j}, j = 2, \ldots, 6\}$. Find

$$m_q = \arg \min_{2 \le j \le 6} \text{SITC}^{(h_{q,j})}_{n,\widehat{\mathcal{J}}^{(n)}_{h_{q,j}}}.$$

Step 6: If $\widehat{\mathcal{J}}^{(n)}_{h_{q,m_q}} = \widehat{\mathcal{J}}^{(n)}_{h_{(q-1),m_{q-1}}}$ and $m_q = m_{q-1}$, set $\widehat{\mathcal{J}}^{(n)}_{\text{CP}} = \widehat{\mathcal{J}}^{(n)}_{h_{q,m_q}}$ and $\hat{h}_{\text{CP}} = m_q$ and then proceed to Step 7. If only $\widehat{\mathcal{J}}^{(n)}_{h_{q,m_q}} = \widehat{\mathcal{J}}^{(n)}_{h_{(q-1),m_{q-1}}}$, set $\widehat{\mathcal{J}}^{(n)}_{\text{CP}} = \widehat{\mathcal{J}}^{(n)}_{h_{(q-1),m_{q-1}}}$ and $q = q + 1$ and then proceed to Step 6a and if only $m_q = m_{q-1}$, set $\hat{h}_{\text{CP}} = m_{q-1}$ and $q = q + 1$ and then proceed to Step 6b. Otherwise, set $q = q + 1$ and go back to Step 4.

Step 6a: Only change point analysis will be carried out. We repeat Steps 4–6 until $m_q = m_{q-1}$, where $\widehat{\mathcal{J}}^{(n)}_h$, $\widehat{\mathcal{J}}^{(n)}_{h_{(q-1),m_{q-1}}}$ and $\widehat{\mathcal{J}}^{(n)}_{h_{q,j}}$ are all replaced by $\widehat{\mathcal{J}}^{(n)}_{\text{CP}}$. Then set $\hat{h}_{\text{CP}} = m_{q-1}$ and proceed to Step 7.

Step 6b: Only variable selection will be implemented. We find $\widehat{\mathcal{J}}^{(n)}_{\hat{h}_{\text{CP}}}$ such that

$$j \notin \widehat{\mathcal{J}}^{(n)}_{\hat{h}_{\text{CP}}} \quad \text{if } \text{KSITC}^{(\hat{h}_{\text{CP}})}_{n,-j} \le 0 \quad \text{and} \quad j \notin \widehat{\mathcal{J}}^{(n)}_{\hat{h}_{\text{CP}}} \quad \text{if } \text{KSITC}^{(\hat{h}_{\text{CP}})}_{n,-j} > 0, \quad j = 1, \ldots, p.$$

Set $\widehat{\mathcal{J}}^{(n)}_{\text{CP}} = \widehat{\mathcal{J}}^{(n)}_{\hat{h}_{\text{CP}}}$ and then proceed to Step 7. Note that $\text{KSITC}^{(\xi)}_{n,-j}$ is defined in (15).

Step 7: If

$$\text{SITC}^{(h_{\text{CP}})}_{n,\widehat{\mathcal{J}}^{(n)}_{\text{CP}}} < \text{SITC}_{n,\widehat{\mathcal{J}}^{(n)}},$$

then there exists a change point and the parameter estimates are given by

$$(\hat{h}_{\text{CP}}, \widehat{\mathcal{J}}^{(n)}_{\text{CP}}, \hat{\boldsymbol{\vartheta}}_{1,\widehat{\mathcal{J}}^{(n)}_{\hat{h}_{\text{CP}}}}, \hat{\boldsymbol{\vartheta}}_{2,\widehat{\mathcal{J}}^{(n)}_{\hat{h}_{\text{CP}}}}) \quad \text{or} \quad (z_{\hat{h}_{\text{CP}}}, \widehat{\mathcal{J}}^{(n)}_{\text{CP}}, \hat{\boldsymbol{\vartheta}}_{1,\widehat{\mathcal{J}}^{(n)}_{\hat{h}_{\text{CP}}}}, \hat{\boldsymbol{\vartheta}}_{2,\widehat{\mathcal{J}}^{(n)}_{\hat{h}_{\text{CP}}}}).$$

Otherwise, there is no change point and the parameter estimates are given by $(\widehat{\mathcal{J}}^{(n)}, \hat{\boldsymbol{\vartheta}}_{\widehat{\mathcal{J}}^{(n)}})$. It is noted that $\hat{\boldsymbol{\vartheta}}_{1,\widehat{\mathcal{J}}^{(n)}_{\hat{h}_{\text{CP}}}}$, $\hat{\boldsymbol{\vartheta}}_{2,\widehat{\mathcal{J}}^{(n)}_{\hat{h}_{\text{CP}}}}$ and $\hat{\boldsymbol{\vartheta}}_{\widehat{\mathcal{J}}^{(n)}}$ are the byproducts of the algorithm.

This algorithm is named ALG in this paper.

Remark 9. By removing variable selection part in the algorithm above, it can also be applied in change point analysis. It will reduce the computing time significantly when the sample size is large.

Simulation comparison of the performances of the ALG and the KSITC will be given in the next section. For carrying change point analysis and variable selection simultaneously, the ALG only needs at most half of the time needed by KSITC for the same precision. The time saving is even more significant for large sample size.

4. A simulation study

We first modify the two step procedure mentioned in the introduction section. We propose to add one more step, i.e., change point analysis will be carried out once more based on selected variables. It will be called the three-step procedure.

In this section, by computer simulations, we verify the small-sample performances of the three-step procedure, the KSITC and the ALG. The regression model (2) is considered here, i.e.,

$$y_i = (\mu_1 + \alpha_1 z_i + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_1) I(z_i \le \xi) + (\mu_2 + \alpha_2 z_i + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_2) I(z_i > \xi) + \varepsilon_i, \\ i = 1, \ldots, n.$$

In our simulation, we first generate a sequence of independent variables $\{z_i,\ i = 1, \ldots, n\}$, which are uniformly distributed on $(0, 3)$. As mentioned in Section 3, without of loss of generality, we order z_i and still name the ordered observations as $\{z_i\}$. Secondly we generate independent random vectors $\boldsymbol{x}_i,\ i = 1, \ldots, n$ such that $\boldsymbol{x}_i$ is $N(\boldsymbol{0}, I_p)$ distributed, where I_p is the $p \times p$ identity matrix. We then generate independently and identically distributed $N(0, 1)$ random variables $\varepsilon_1, \ldots, \varepsilon_n$. We also adopt the notation m_0 introduced there, i.e., for $m_0 < n$,

$$y_i = \mu_1 + \alpha_1 z_i + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_1 + \varepsilon_i, \quad i = 1, \ldots, m_0,$$
$$y_i = \mu_2 + \alpha_2 z_i + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}_2 + \varepsilon_i, \quad i = m_0 + 1, \ldots, n,$$

and for $m_0 = n$,

$$y_i = \mu + \alpha z_i + \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \ldots, n.$$

The following settings are used in the simulation studies:

(1) $n = 100$, $p = 10$, $\boldsymbol{\vartheta}_1 = (1, 1, 1, 0, 0, 1, 2, -1, 0, 0, 1.2, 0)^{\mathrm{T}}$ and

$$\boldsymbol{\vartheta}_2 = (1, 1.1, 1, 0, 0, -1, 1.8, 1, 0, 0, 1, 0)^{\mathrm{T}},$$

- $m_0 = 30, 50$, or 70;
- $m_0 = n$ and $\boldsymbol{\vartheta} = \boldsymbol{\vartheta}_1$.

This setting is used for the results reported in the Table 1.

(2) $n = 150$, $p = 17$, $\boldsymbol{\vartheta}_1 = (1,\ 1,\ 1,\ 0,\ 0,\ 1,\ 2,\ -1,\ 0,\ 0,\ 1.2,\ 0,\ 1,\ 0,\ -1,\ 0,\ 1)^{\mathrm{T}} \in \mathbb{R}^{17}$ and $\boldsymbol{\vartheta}_2 = (1,\ 1.1,\ 1,\ 0,\ 0,\ -1,\ 1.8,\ 1,\ 0,\ 0,\ 1,\ 0,\ 1.2,\ 0,\ 1.3,\ 0,\ 1.5)^{\mathrm{T}} \in \mathbb{R}^{17}$ [the first 12 elements of these $\boldsymbol{\vartheta}_1$ and $\boldsymbol{\vartheta}_2$ are the same as the $\boldsymbol{\vartheta}_1$ and $\boldsymbol{\vartheta}_2$ in the setting (1)],

- $m_0 = 45,\ 75$, or 105;
- $m_0 = n$ and $\boldsymbol{\vartheta} = \boldsymbol{\vartheta}_1$.

This setting is used for the results reported in the Table 2.

(3) $n = 300$, $p = 17$, $\boldsymbol{\vartheta}_1$ and $\boldsymbol{\vartheta}_2$ are the same as in the setting (2),

- $m_0 = 45, 90, 150, 210$, or 255;

Table 1
The entries are the numbers of correct variable selection with $|\hat{m}_0 - m_0| \le 0, 1, 2, 3$, and 4 respectively based on 1000 simulations

m_0	Method	$\hat{m}_0 = m_0$	$\|\hat{m}_0 - m_0\| \le 1$	$\|\hat{m}_0 - m_0\| \le 2$	$\|\hat{m}_0 - m_0\| \le 3$	$\|\hat{m}_0 - m_0\| \le 4$
$30 = 30\%n$	(I)	435 (746)	504 (857)	549 (946)	557 (968)	564 (981)
	(II)	453 (782)	505 (867)	553 (958)	559 (977)	564 (986)
	(III)	748 (788)	839 (881)	921 (968)	932 (979)	942 (990)
	(IV)	764 (792)	855 (884)	938 (972)	943 (977)	956 (990)
$50 = 50\%n$	(I)	244 (395)	408 (672)	490 (808)	593 (992)	594 (996)
	(II)	265 (439)	407 (681)	493 (822)	593 (993)	594 (996)
	(III)	432 (452)	660 (693)	785 (829)	942 (994)	944 (997)
	(IV)	437 (455)	655 (680)	782 (817)	951 (993)	954 (997)
$70 = 70\%n$	(I)	472 (811)	489 (868)	497 (908)	510 (939)	514 (958)
	(II)	480 (852)	499 (896)	506 (925)	514 (949)	518 (968)
	(III)	838 (894)	871 (931)	888 (949)	904 (967)	913 (977)
	(IV)	726 (812)	742 (833)	768 (861)	768 (861)	802 (901)
$m_0 = n$	(I)	843 (992)	843 (992)	843 (992)	843 (992)	843 (992)
	(II)	843 (995)	843 (995)	843 (995)	843 (995)	843 (995)
	(III)	849 (1000)	849 (1000)	849 (1000)	849 (1000)	849 (1000)
	(IV)	849 (1000)	849 (1000)	849 (1000)	849 (1000)	849 (1000)

For comparison, the numbers of only the correct change point detection are given in the parentheses. The sample size is $n = 100$.

- $m_0 = n$ and $\boldsymbol{\vartheta} = \boldsymbol{\vartheta}_1$.
  This setting is used for the results reported in the Table 3.

Here $\boldsymbol{\vartheta}_1 = (\mu_1, \alpha_1, \boldsymbol{\beta}_1^{\mathrm{T}})^{\mathrm{T}}$ and $\boldsymbol{\vartheta}_2 = (\mu_2, \alpha_2, \boldsymbol{\beta}_2^{\mathrm{T}})^{\mathrm{T}}$.

We then compute y_i as follows:

$$y_i = (1\ z_i\ \boldsymbol{x}_i^{\mathrm{T}})\boldsymbol{\vartheta}_1 I(i \le m_0) + (1 z_i \boldsymbol{x}_i^{\mathrm{T}})\boldsymbol{\vartheta}_2 I(i > m_0) + \varepsilon_i, \quad i = 1, \ldots, n.$$

It can be seen that if $m_0 < n$, there is a change point at m_0 in the series. If $m_0 = n$, there does not exist a change point.

In all the simulation studies of this section, $C_n = \log(n)$. The following methods have been used to carry out change point analysis and variable selection simultaneously:

Method I: Carry out change point analysis first and then perform variable selection. This is the two-step procedure.

Method II: Carry out change point analysis first and then perform variable selection. Finally carry out change point analysis based on selected variables. This is the three-step procedure.

Method III: Carry out change point analysis and variable selection simultaneously by Criterion KSITC.

Method IV: Carry out change point analysis and variable selection simultaneously using the algorithm ALG given in Section 3.

The simulation results are presented in the Tables 1–3. In these tables, the entries are the numbers of correct variable selection with $|\hat{m}_0 - m_0| \le 0,\ 1,\ 2,\ 3$, and 4 respectively based on 1000 simulations. For comparison, the numbers of only $|\hat{m}_0 - m_0| \le 0,\ 1,\ 2,\ 3$, and 4 are given in the parentheses. The sample sizes are $n = 100,\ 150$ and 300, respectively.

From the simulation results reported in these three tables, it can be seen that the three-step procedure, SITC and ALG outperform the two-step procedure. The three-step procedure performs slightly better than the two-step procedure. Both KSITC and ALG performs much

Table 2
The entries are the numbers of correct variable selection with $|\hat{m}_0 - m_0| \leq 0, 1, 2, 3$, and 4 respectively based on 1000 simulations

m_0	Method	$\hat{m}_0 = m_0$	$\lvert\hat{m}_0 - m_0\rvert \leq 1$	$\lvert\hat{m}_0 - m_0\rvert \leq 2$	$\lvert\hat{m}_0 - m_0\rvert \leq 3$	$\lvert\hat{m}_0 - m_0\rvert \leq 4$
	(I)	134 (259)	348 (671)	521 (955)	539 (998)	539 (999)
$45 = 30\%n$	(II)	147 (272)	360 (678)	523 (962)	540 (1000)	540 (1000)
	(III)	251 (266)	641 (674)	921 (969)	947 (1000)	947 (1000)
	(IV)	261 (277)	676 (706)	931 (970)	956 (1000)	956 (1000)
	(I)	393 (701)	489 (874)	510 (918)	561 (1000)	561 (1000)
$75 = 50\%n$	(II)	398 (724)	505 (899)	526 (938)	561 (1000)	561 (1000)
	(III)	686 (720)	864 (905)	901 (944)	956 (1000)	956 (1000)
	(IV)	669 (724)	852 (917)	855 (920)	932 (1000)	932 (1000)
	(I)	125 (244)	334 (639)	521 (983)	522 (988)	523 (992)
$105 = 70\%n$	(II)	144 (277)	344 (653)	522 (986)	522 (990)	523 (993)
	(III)	261 (271)	627 (645)	954 (989)	957 (992)	959 (994)
	(IV)	355 (373)	932 (978)	932 (978)	939 (985)	944 (990)
	(I)	780 (946)	780 (946)	780 (946)	780 (946)	780 (946)
$m_0 = n$	(II)	784 (985)	784 (985)	784 (985)	784 (985)	784 (985)
	(II)	825 (1000)	825 (1000)	825 (1000)	825 (1000)	825 (1000)
	(III)	825 (1000)	825 (1000)	825 (1000)	825 (1000)	825 (1000)

For comparison, the numbers of only the correct change point detection are given in the parentheses. The sample size is $n = 150$.

Table 3
The entries are the numbers of correct variable selection with $|\hat{m}_0 - m_0| \leq 0, 1, 2, 3$, and 4 respectively based on 1000 simulations

m_0	Method	$\hat{m}_0 = m_0$	$\lvert\hat{m}_0 - m_0\rvert \leq 1$	$\lvert\hat{m}_0 - m_0\rvert \leq 2$	$\lvert\hat{m}_0 - m_0\rvert \leq 3$	$\lvert\hat{m}_0 - m_0\rvert \leq 4$
	(I)	476 (741)	578 (935)	611 (995)	612 (997)	612 (1000)
$45 = 15\%n$	(II)	466 (737)	573 (932)	611 (998)	612 (999)	612 (1000)
	(III)	726 (739)	919 (937)	980 (999)	981 (1000)	981 (1000)
	(IV)	747 (758)	960 (976)	984 (1000)	984 (1000)	984 (1000)
	(I)	229 (349)	550 (823)	596 (894)	663 (1000)	663 (1000)
$90 = 30\%n$	(II)	243 (368)	556 (836)	600 (900)	663 (1000)	663 (1000)
	(III)	359 (370)	815 (838)	878 (904)	973 (1000)	973 (1000)
	(IV)	372 (382)	847 (866)	969 (990)	979 (1000)	979 (1000)
	(I)	654 (959)	684 (1000)	684 (1000)	684 (1000)	684 (1000)
$150 = 50\%n$	(II)	655 (960)	684 (1000)	684 (1000)	684 (1000)	684 (1000)
	(III)	936 (960)	976 (1000)	976 (1000)	976 (1000)	976 (1000)
	(IV)	941 (960)	981 (1000)	981 (1000)	981 (1000)	981 (1000)
	(I)	398 (590)	623 (935)	636 (952)	649 (976)	655 (987)
$210 = 70\%n$	(II)	401 (596)	626 (941)	638 (955)	650 (977)	658 (989)
	(III)	593 (606)	921 (946)	936 (961)	954 (980)	965 (991)
	(IV)	603 (613)	933 (954)	954 (975)	964 (985)	972 (993)
	(I)	416 (653)	616 (985)	619 (990)	620 (992)	626 (1000)
$255 = 85\%n$	(II)	436 (692)	618 (985)	620 (989)	620 (990)	626 (1000)
	(III)	693 (708)	963 (986)	966 (989)	967 (990)	976 (1000)
	(IV)	693 (711)	963 (988)	963 (988)	965 (990)	974 (1000)
	(I)	850 (978)	850 (978)	850 (978)	850 (978)	850 (978)
$m_0 = n$	(II)	855 (997)	855 (997)	855 (997)	855 (997)	855 (997)
	(III)	871 (1000)	871 (1000)	871 (1000)	871 (1000)	871 (1000)
	(IV)	871 (1000)	871 (1000)	871 (1000)	871 (1000)	871 (1000)

For comparison, the numbers of only the correct change point detection are given in the parentheses. The sample size is $n = 300$.

better than the other two procedures. By considering the time saved by applying ALG for the sample of large size, one may use ALG instead of others.

It can also be observed from Tables 1–3 that the performances of all methods have also varied with m_0 for fixed n. Their performances may be improved by also allowing penalties to vary with the location in the series, an approach used in Chen, Gupta and Pan [2].

Acknowledgments

The author is deeply grateful to an associate editor and a referee whose comments and suggestions have contributed to the improvement of the paper. The author's research is partially supported by the Natural Sciences and Engineering Research Council of Canada.

Appendix

The following lemmas are needed in the proofs of Theorems 2.1 and 2.2.

Lemma A.1. *If A and B are two $\ell \times \ell$ nonsingular matrices, then we have the following elementary linear algebra formula*

$$(A+B)^{-1} = A^{-1} - A^{-1}(A^{-1}+B^{-1})^{-1}A^{-1}.$$

See Lemma 3.2 *in Shao and Wu* [11] *for proof.*

Lemma A.2. *Let $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_p$ be ℓ-vectors and write $G_p = B_p^{\mathrm{T}} B_p$, where $B_p = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_p)$. Let $\mathcal{J}_k = \{j_1, \ldots, j_k\}$ be a subset of $\{1, \ldots, p\}$, $1 \le k \le p$ and denote $G_p(\mathcal{J}_k) = B_p(\mathcal{J}_k)^{\mathrm{T}} B_p(\mathcal{J}_k)$. If there exist constants η_1 and η_2 such that*

$$0 < \eta_1 \le \lambda_p(G_p) \le \lambda_1(G_p) \le \eta_2,$$

then

$$\eta_1 \le \lambda_k(G_p(\mathcal{J}_k)) \le \lambda_1(G_p(\mathcal{J}_k)) \le \eta_2.$$

See Lemma A.1 *in Bai, Rao and Wu* [1] *for proof.*

Proof of Theorem 2.1. By Assumptions (A)–(D), (11) can be proved following the proof of Theorem 3.1 in Rao and Wu [9]. To show (8)–(10) hold, it is easy to see that we only need to show that with probability one,

$$\mathrm{SITC}_{n,k_0}^{(\xi_0)} < \min_{\kappa_L < \xi \le \xi_1} \mathrm{SITC}_{n,k}^{(\xi)}, \quad \text{for } k \ge k_0, \tag{A.1}$$

$$\mathrm{SITC}_{n,k_0}^{(\xi_0)} < \min_{\kappa_L < \xi \le \xi_1} \mathrm{SITC}_{n,k}^{(\xi)}, \quad \text{for } k < k_0, \tag{A.2}$$

$$\mathrm{SITC}_{n,k_0}^{(\xi_0)} < \min_{\xi_2 \le \xi < \kappa_U} \mathrm{SITC}_{n,k}^{(\xi)}, \quad \text{for } k \ge k_0, \tag{A.3}$$

$$\mathrm{SITC}_{n,k_0}^{(\xi_0)} < \min_{\xi_2 \le \xi < \kappa_U} \mathrm{SITC}_{n,k}^{(\xi)}, \quad \text{for } k < k_0, \tag{A.4}$$

$$\mathrm{SITC}_{n,k_0}^{(\xi_0)} < \mathrm{SITC}_{n,k}, \quad \text{for } k \ge k_0, \tag{A.5}$$

$$\mathrm{SITC}_{n,k_0}^{(\xi_0)} < \mathrm{SITC}_{n,k}, \quad \text{for } k < k_0, \tag{A.6}$$

when n is large.

We first show that (A.1)–(A.4) hold for large n. We split the proof into two parts as follows:

Part 1. $k \geq k_0$ and $\kappa_L < \xi \leq \xi_1 < \xi_0$ or $\xi_0 < \xi_2 \leq \xi < \kappa_U$.

We now show that (A.1) holds true. By the definition of $\mathrm{SITC}_{n,k}^{(\xi)}$, we have

$$\begin{aligned}\mathrm{SITC}_{n,k}^{(\xi)} - \mathrm{SITC}_{n,k}^{(\xi_0)} &= \ell_{n,k}^{(\xi)} - \ell_{n,k_0}^{(\xi_0)} + 2[q(k+2) - q(k_0+2)]C_n,\\ &= \ell_{n,k}^{(\xi)} - \ell_{n,k}^{(\xi_0)} + \ell_{n,k}^{(\xi_0)} - \ell_{n,k_0}^{(\xi_0)} + 2[q(k+2) - q(k_0+2)]C_n. \qquad \text{(A.7)}\end{aligned}$$

Denote $\boldsymbol{w}_i = (1, z_i, \boldsymbol{x}_i^{\mathrm{T}})^{\mathrm{T}}$ for $1 \leq i \leq n$. By (2) and (5), we have

$$\begin{aligned}\ell_{n,k}^{(\xi)} - \ell_{n,k}^{(\xi_0)} &= \sum_{i:\, z_i \leq \xi} (y_i - \boldsymbol{w}_i^{\mathrm{T}} \hat{\boldsymbol{\vartheta}}_{1,k,\xi})^2 + \sum_{i:\, z_i > \xi} (y_i - \boldsymbol{w}_i^{\mathrm{T}} \hat{\boldsymbol{\vartheta}}_{2,k,\xi})^2\\ &\quad - \left[\sum_{i:\, z_i \leq \xi_0} (y_i - \boldsymbol{w}_i^{\mathrm{T}} \hat{\boldsymbol{\vartheta}}_{1,k,\xi_0})^2 + \sum_{i:\, z_i > \xi_0} (y_i - \boldsymbol{w}_i^{\mathrm{T}} \hat{\boldsymbol{\vartheta}}_{2,k,\xi_0})^2\right]\\ &= [\boldsymbol{y}(\mathcal{G}_{\xi,l})]^{\mathrm{T}} \left(I - P_{Z_{n,k}(\mathcal{G}_{\xi,l})}\right) \boldsymbol{y}(\mathcal{G}_{\xi,l}) + [\boldsymbol{y}(\mathcal{G}_{\xi,r})]^{\mathrm{T}} \left(I - P_{Z_{n,k}(\mathcal{G}_{\xi,r})}\right) \boldsymbol{y}(\mathcal{G}_{\xi,r})\\ &\quad - \Big\{[\boldsymbol{y}(\mathcal{G}_{\xi_0,l})]^{\mathrm{T}} (I - P_{Z_{n,k}(\mathcal{G}_{\xi_0,l})}) \boldsymbol{y}(\mathcal{G}_{\xi_0,l}) + [\boldsymbol{y}(\mathcal{G}_{\xi_0,r})]^{\mathrm{T}}\\ &\quad \times (I - P_{Z_{n,k}(\mathcal{G}_{\xi_0,r})}) \boldsymbol{y}(\mathcal{G}_{\xi_0,r})\Big\}. \qquad \text{(A.8)}\end{aligned}$$

For convenience, we denote $Z_{n,k}(\mathcal{G}_{\xi,r})$ by $Z_{k,\xi,r}$, $Z_{n,k}(\mathcal{G}_{\xi,l})$ by $Z_{k,\xi,l}$, and $Z_{n,k}(\mathcal{G}_{\xi,r} \cap \mathcal{G}_{\xi_0,l})$ by Z_{k,ξ,ξ_0} here. In view of (2), it follows that

$$\boldsymbol{y}(\mathcal{G}_{\xi,r}) = Z_{k,\xi,r} \boldsymbol{\vartheta}_{2,0} + (Z_{k,\xi,\xi_0}^{\mathrm{T}} \quad 0_{(2+k) \times |\mathcal{G}_{\xi_0,r}|})^{\mathrm{T}} (\boldsymbol{\vartheta}_{1,0} - \boldsymbol{\vartheta}_{2,0}) + \boldsymbol{\varepsilon}(\mathcal{G}_{\xi,r}),$$

where $0_{a \times b}$ denotes an $a \times b$ matrix with zero elements. Hence by (2) and (A.8),

$$\begin{aligned}\ell_{n,k}^{(\xi)} - \ell_{n,k}^{(\xi_0)} &= [\boldsymbol{\varepsilon}(\mathcal{G}_{\xi,l})]^{\mathrm{T}} (I - P_{Z_{k,\xi,l}}) \boldsymbol{\varepsilon}(\mathcal{G}_{\xi,l})\\ &\quad + (\boldsymbol{\vartheta}_{1,0} - \boldsymbol{\vartheta}_{2,0})^{\mathrm{T}} Z_{k,\xi,\xi_0}^{\mathrm{T}} [I - Z_{k,\xi,\xi_0} (Z_{\xi,\xi_0}^{\mathrm{T}} Z_{\xi,\xi_0} + Z_{k,\xi_0,r}^{\mathrm{T}} Z_{k,\xi_0,r})^{-1} Z_{k,\xi,\xi_0}^{\mathrm{T}}]\\ &\quad \times Z_{k,\xi,\xi_0} (\boldsymbol{\vartheta}_{1,0} - \boldsymbol{\vartheta}_{2,0})\\ &\quad + 2(\boldsymbol{\vartheta}_{1,0} - \boldsymbol{\vartheta}_{2,0})^{\mathrm{T}} (Z_{k,\xi,\xi_0}^{\mathrm{T}} \quad 0_{(2+k) \times |\mathcal{G}_{\xi_0,r}|})^{\mathrm{T}} (I - P_{Z_{k,\xi,r}}) \boldsymbol{\varepsilon}(\mathcal{G}_{k,\xi,r})\\ &\quad + [\boldsymbol{\varepsilon}(\mathcal{G}_{\xi,r})]^{\mathrm{T}} (I - P_{Z_{k,\xi,r}}) \boldsymbol{\varepsilon}(\mathcal{G}_{\xi,r})\\ &\quad - [\boldsymbol{\varepsilon}(\mathcal{G}_{\xi_0,l})]^{\mathrm{T}} (I - P_{Z_{k,\xi_0,l}}) \boldsymbol{\varepsilon}(\mathcal{G}_{\xi_0,l}) - [\boldsymbol{\varepsilon}(\mathcal{G}_{\xi_0,r})]^{\mathrm{T}} (I - P_{Z_{k,\xi_0,r}}) \boldsymbol{\varepsilon}(\mathcal{G}_{\xi_0,r})\\ &= (\boldsymbol{\vartheta}_{1,0} - \boldsymbol{\vartheta}_{2,0})^{\mathrm{T}} Z_{k,\xi,\xi_0}^{\mathrm{T}} [I - Z_{k,\xi,\xi_0} (Z_{k,\xi,\xi_0}^{\mathrm{T}} Z_{k,\xi,\xi_0} + Z_{k,\xi_0,r}^{\mathrm{T}} Z_{k,\xi_0,r})^{-1} Z_{k,\xi,\xi_0}^{\mathrm{T}}]\\ &\quad \times Z_{k,\xi,\xi_0} (\boldsymbol{\vartheta}_{1,0} - \boldsymbol{\vartheta}_{2,0})\\ &\quad + 2(\boldsymbol{\vartheta}_{1,0} - \boldsymbol{\vartheta}_{2,0})^{\mathrm{T}} (Z_{k,\xi,\xi_0}^{\mathrm{T}} 0_{(2+k) \times |\mathcal{G}_{\xi_0,r}|})^{\mathrm{T}} (I - P_{Z_{k,\xi,r}}) \boldsymbol{\varepsilon}(\mathcal{G}_{\xi,r}) - [\boldsymbol{\varepsilon}(\mathcal{G}_{\xi,l})]^{\mathrm{T}}\\ &\quad \times P_{Z_{k,\xi,l}} \boldsymbol{\varepsilon}(\mathcal{G}_{\xi,l}) - [\boldsymbol{\varepsilon}(\mathcal{G}_{\xi,r})]^{\mathrm{T}} P_{Z_{k,\xi,r}} \boldsymbol{\varepsilon}(\mathcal{G}_{\xi,r})\\ &\quad + [\boldsymbol{\varepsilon}(\mathcal{G}_{\xi_0,l})]^{\mathrm{T}} P_{Z_{k,\xi_0,l}} \boldsymbol{\varepsilon}(\mathcal{G}_{\xi_0,l}) + [\boldsymbol{\varepsilon}(\mathcal{G}_{\xi_0,r})]^{\mathrm{T}} P_{Z_{k,\xi_0,r}} \boldsymbol{\varepsilon}(\mathcal{G}_{\xi_0,r}). \qquad \text{(A.9)}\end{aligned}$$

By Assumption (C), we have for any $\xi \in (\kappa_L,\ \kappa_U)$,

$$\begin{aligned}&[\boldsymbol{\varepsilon}(\mathcal{G}_{\xi,l})]^{\mathrm{T}} P_{Z_{\xi,l}} \boldsymbol{\varepsilon}(\mathcal{G}_{\xi,l}) \leq M \log\log(n), \text{ a.s.}, \quad [\boldsymbol{\varepsilon}(\mathcal{G}_{\xi,r})]^{\mathrm{T}} P_{Z_{\xi,r}} \boldsymbol{\varepsilon}(\mathcal{G}_{\xi,r}) \leq M \log\log(n), \text{ a.s.},\\ &[\boldsymbol{\varepsilon}(\mathcal{G}_{\xi_0,l})]^{\mathrm{T}} P_{Z_{\xi_0,l}} \boldsymbol{\varepsilon}(\mathcal{G}_{\xi_0,l}) \leq M \log\log(n), \text{ a.s.}, \quad [\boldsymbol{\varepsilon}(\mathcal{G}_{\xi_0,r})]^{\mathrm{T}} P_{Z_{\xi_0,r}} \boldsymbol{\varepsilon}(\mathcal{G}_{\xi_0,r})\\ &\quad \leq M \log\log(n), \text{ a.s..}\end{aligned}$$

By Assumptions (A)–(C), Lemma A.2 and Cauchy–Schwarz inequality, it can be shown that there exists $M_1 > 0$ such that for any ξ and $k \geq k_0$,

$$|(\boldsymbol{\vartheta}_{1,0} - \boldsymbol{\vartheta}_{2,0})^{\mathrm{T}}(Z^{\mathrm{T}}_{k,\xi,\xi_0} \quad 0_{(2+k)\times|\mathcal{G}_{\xi_0,r}|})^{\mathrm{T}}(I - P_{Z_{\xi,r}})\boldsymbol{\varepsilon}(\mathcal{G}_{\xi,r})| \leq M_1\sqrt{n\log\log(n)}, \quad \text{a.s.}$$

Since $\xi \leq \xi_1 < \xi_0$, in view of Assumption (A) and Lemma A.1, there exists $\omega_1 > 0$ such that for large n and $k \geq k_0$,

$$\begin{aligned}
&(\boldsymbol{\vartheta}_{1,0} - \boldsymbol{\vartheta}_{2,0})^{\mathrm{T}} Z^{\mathrm{T}}_{k,\xi,\xi_0}[I - Z_{k,\xi,\xi_0}(Z^{\mathrm{T}}_{k,\xi,\xi_0} Z_{\xi,\xi_0} + Z^{\mathrm{T}}_{k,\xi_0,r} Z_{k,\xi_0,r})^{-1} Z^{\mathrm{T}}_{k,\xi,\xi_0}]\\
&\quad Z_{k,\xi,\xi_0}(\boldsymbol{\vartheta}_{1,0} - \boldsymbol{\vartheta}_{2,0})\\
&\quad = (\boldsymbol{\vartheta}_{1,0} - \boldsymbol{\vartheta}_{2,0})^{\mathrm{T}}[(Z^{\mathrm{T}}_{k,\xi,\xi_0} Z_{k,\xi,\xi_0})^{-1} + (Z^{\mathrm{T}}_{k,\xi_0,r} Z_{k,\xi_0,r})^{-1}]^{-1}(\boldsymbol{\vartheta}_{1,0} - \boldsymbol{\vartheta}_{2,0})\\
&\quad \geq \omega_1 |\boldsymbol{\vartheta}_{1,0} - \boldsymbol{\vartheta}_{2,0}| n.
\end{aligned}$$

Note that $|\boldsymbol{\vartheta}_{1,0} - \boldsymbol{\vartheta}_{2,0}| = |\mu_1 - \mu_2| + |\alpha_1 - \alpha_2| + \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\| \neq 0$. Hence with probability one, there exists a constant $\gamma_1 > 0$ such that for any $\xi \leq \xi_1$ and $k \geq k_0$,

$$\ell^{(\xi)}_{n,k} - \ell^{(\xi_0)}_{n,k} > \gamma_1 n \tag{A.10}$$

for large n. In addition, by Assumption (C) and (5), we have

$$\ell^{(\xi_0)}_{n,k} \leq M \log\log(n), \quad \text{a.s., for } k \geq k_0. \tag{A.11}$$

Thus by Assumption (D), (A.7) and (A.10), with probability one, (A.1) holds for large n.

It can be similarly shown that with probability one, (A.3) holds for large n. The details are omitted.

Part 2. $k < k_0$ and $\kappa_L < \xi \leq \xi_1 < \xi_0$ or $\xi_0 < \xi_2 \leq \xi < \kappa_U$.

Note that

$$\begin{aligned}
\mathrm{SITC}^{(\xi)}_{n,k} - \mathrm{SITC}^{(\xi_0)}_{n,k} &= \ell^{(\xi)}_{n,k} - \ell^{(\xi_0)}_{n,k_0} + 2[q(k+2) - q(k_0+2)]C_n\\
&= \ell^{(\xi)}_{n,k} - \ell^{(\xi)}_{n,k_0} + \ell^{(\xi)}_{n,k_0} - \ell^{(\xi_0)}_{n,k_0} + 2[q(k+2) - q(k_0+2)]C_n.
\end{aligned} \tag{A.12}$$

It is obvious that

$$\ell^{(\xi)}_{n,k} \geq \ell^{(\xi)}_{n,k_0}.$$

If $\kappa_L < \xi \leq \xi_1 < \xi_0$, by (A.10), with probability one,

$$\ell^{(\xi)}_{n,k_0} - \ell^{(\xi_0)}_{n,k_0} > \gamma_1 n$$

for large n. Hence by Assumption (D) and (A.12), with probability one, (A.2) follows for large n.

It can be similarly shown that with probability one, (A.4) holds for large n. The details are omitted.

To this point, we have shown that with probability one, (A.1)–(A.4) hold for large n. Now we show that with probability one, (A.5) and (A.6) hold for large n.

Note that

$$\begin{aligned}
\mathrm{SITC}_{n,k} - \mathrm{SITC}^{(\xi_0)}_{n,k_0} &= \ell_{n,k} - \ell^{(\xi_0)}_{n,k_0} + q(k+2) - 2q(k_0+2)C_n > 0\\
&= \ell_{n,k} - \ell^{(\xi_0)}_{n,k} + \ell^{(\xi_0)}_{n,k} - \ell^{(\xi_0)}_{n,k_0} + [q(k+2) - 2q(k_0+2)]C_n.
\end{aligned} \tag{A.13}$$

We first show that with probability one, (A.5) holds for large n.

Mimicking to the proof of (A.10), it can be shown that with probability one, there exists a constant $\gamma_2 > 0$ such that for $k \geq k_0$,

$$\ell_{n,k} - \ell_{n,k_0}^{(\xi_0)} > \gamma_2 n$$

for large n under Assumptions (A)–(C). Hence by Assumption (D) and (A.11), with probability one, (A.5) follows for large n.

We now show that with probability one, (A.6) holds for large n.

It is easy to see that $\ell_{n,k} \geq \ell_{n,k}^{(\xi_0)}$. By Rao and Wu [9], it can be shown that with probability one, for any $k < k_0$, there exists $\gamma_3 > 0$ such that

$$\ell_{n,k}^{(\xi_0)} - \ell_{n,k_0}^{(\xi_0)} > \gamma_3 n,$$

for large n. In view of Assumption (D) and (A.13), with probability one, (A.6) holds for large n. □

Proof of Theorem 2.2. By Assumptions (A)–(D) and Rao and Wu [9], with probability one, (13) holds for large n. We then only need to show that with probability one, for $1 \leq k \leq p$,

$$\mathrm{SITC}_{n,k_0} < \min_{\kappa_L < \xi < \kappa_U} \mathrm{SITC}_{n,k}^{(\xi)} \tag{A.14}$$

for large n.

Consider the following two cases:

Case I. $k \geq k_0$.

By Rao and Wu [9] and (A.11), with probability one, there exists $M_I > 0$ such that

$$|\ell_{n,k}^{(\xi)} - \ell_{n,k_0}| \leq M_I \log\log(n)$$

for large n when Assumptions (A)–(C) hold. Hence, by Assumption (D) and the fact that $q(k+2) > q(k_0+2)$, it follows that with probability one, for $k \geq k_0$,

$$\begin{aligned}\mathrm{SITC}_{n,k}^{(\xi)} - \mathrm{SITC}_{n,k_0} &= \ell_{n,k}^{(\xi)} - \ell_{n,k_0} + [2q(k+2) - q(k_0+2)]C_n,\\ &= O(\log\log(n)) + [2q(k+2) - q(k_0+2)]C_n > 0\end{aligned}$$

for large n, i.e., (A.14) holds for large n.

Case II. $k < k_0$.

Note that

$$\begin{aligned}\mathrm{SITC}_{n,k}^{(\xi)} - \mathrm{SITC}_{n,k_0} &= \ell_{n,k}^{(\xi)} - \ell_{n,k_0} + [2q(k+2) - q(k_0+2)]C_n\\ &= \ell_{n,k}^{(\xi)} - \ell_{n,k_0}^{(\xi)} + \ell_{n,k_0}^{(\xi)} - \ell_{n,k_0} + [2q(k+2) - q(k_0+2)]C_n.\end{aligned}$$

By Rao and Wu [9], it can be shown that with probability one, there exists a $\gamma_I > 0$ such that

$$\ell_{n,k}^{(\xi)} - \ell_{n,k_0}^{(\xi)} > \gamma_I n \tag{A.15}$$

for large n under Assumptions (A)–(C). In view of (5), by Assumption (C), with probability one, there exists $M_{II} > 0$ such that

$$|\ell_{n,k_0}^{(\xi)} - \ell_{n,k_0}| \leq M_{II} \log\log(n) \tag{A.16}$$

for large n. Hence, by Assumption (D) and (A.15) and (A.16), with probability one, for large n, (A.14) follows when $k < k_0$. □

References

[1] Z.D. Bai, C.R. Rao, Y. Wu, Model selection with data-oriented penalty, Journal of Statistical Planning and Inference 77 (1999) 103–117.

[2] J. Chen, A.K. Gupta, J. Pan, Information criterion and change point problem for regular models, Sankhyā Part 2 (2006) 252–282.

[3] J. Chen, A.K. Gupta, Parametric Statistical Change Point Analysis, Birkháuser, 2000.

[4] M. Csörgő, L. Horváth, Limit Theorems in Change-Point Analysis, Wiley, Chichester, 1997.

[5] J. Fan, R. Li, Statistical challenges with high dimensionality: Feature selection in knowledge discovery, in: Proceedings of the International Congress of Mathematicians, Madrid, August 22–30, vol. III, 2006, pp. 595–622.

[6] A. McQuarrie, C.L. Tsai, Regression And Time Series Model Selection, World Scientific, Singapore, 1998.

[7] J. Pan, J. Chen, Application of modified information criterion to multiple change point problems, Journal of Multivariate Analysis 97 (2006) 2221–2241.

[8] V.V. Petrov, Limit Theorems of Probability Theory, Oxford Science Publications, 1995.

[9] C.R. Rao, Y. Wu, A strongly consistent procedure for model selection in regression problem, Biometrika 76 (1989) 369–374.

[10] C.R. Rao, Y. Wu, On model selection (with discussion), in: IMS Lecture Notes, vol. 38, 2001, pp. 1–64.

[11] Q. Shao, Y. Wu, A consistent procedure for determining the number of clusters in regression clustering, Journal of Statistical Planning and Inference 135 (2005) 461–476.