Published in final edited form as:

J Stat Plan Inference. 2016 March 1; 170: 106–116. doi:10.1016/j.jspi.2015.09.008.

Consistent Model Selection in Segmented Line Regression

Jeankyung Kim and

Department of Statistics, Inha University, 253 Yonghyundong, Namgu, Incheon, 402-751, Republic of Korea

Hyune-Ju Kim

Department of Mathematics, Syracuse University, 215 Carnegie Building, Syracuse, NY 13244-1150, U.S.A.

Abstract

The Schwarz criterion or Bayes Information Criterion (BIC) is often used to select a model dimension, and some variations of the BIC have been proposed in the context of change-point problems. In this paper, we consider a segmented line regression model with an unknown number of change-points and study asymptotic properties of Schwarz type criteria in selecting the number of change-points. Noticing the overestimating tendency of the traditional BIC observed in some empirical studies and being motivated by asymptotic behavior of the modified BIC proposed by Zhang and Siegmund (2007), we consider a variation of the Schwarz type criterion that applies a harsher penalty equivalent to the model with one additional unknown parameter per segment. For the segmented line regression model without the continuity constraint, we prove the consistency of the number of change-points selected by the criterion with such type of a modification and summarize the simulation results that support the consistency. Further simulations are conducted for the model with the continuity constraint, and we empirically observe that the asymptotic behavior of this modified version of BIC is comparable to that of the criterion proposed by Liu, Wu, and Zidek (1997).

Keywords

Bayes Information Criterion; Segmented line regression; Model selection

1 Introduction

A main concern in regression model selection is how to select the “best” set of independent variables, and two major approaches of the model selection are hypothesis testing and information criteria approaches. In the context of change-point problems, both approaches have been applied to select the number of change-points, and their analytic and empirical properties have been investigated by many researchers. One of widely used information criteria is the Bayes Information Criterion (BIC) proposed by Schwarz (1978). This Schwarz

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

criterion selects the model dimension by finding the Bayes solution that maximizes a posterior probability of the model, and Schwarz (1978) derived the following criterion by evaluating the leading terms of its asymptotic expansion:

$$SC(p) = \sup_{\theta_p} \log(\text{lik}(\theta_p)) - \frac{p}{2} \log n = \log(\text{lik}(\hat{\theta}_p)) - \frac{p}{2} \log n,$$

where $\text{lik}(\theta_p)$ is the likelihood function of θ_p for the model with dimension p and $\hat{\theta}_p$ is the maximum likelihood estimator of θ_p . As in general information criteria, the Schwarz criterion has two parts, the log of the maximized likelihood function and the penalty function that penalizes for the model dimension, and the method selects the model that maximizes $SC(p)$. Note that its validity is established in Schwarz (1978) for “the case of independent, identically distributed observations, and linear models.”

Yao (1988) studied the problem to select the number of change-points in means of normally distributed random variables, where the total number of unknown parameters for the model with k change-points is $p = 2(k + 1)$. For the number of change-points estimated by minimizing $-SC(p)$ with $p = 2k$, Yao (1988) proved its consistency. Lee (1997) considered a similar type of a criterion to select the number of change-points in a sequence of random variables from an exponential family distribution. Under some mild conditions on spacings of successive change-points, Lee (1997) proved the consistency of the number of change-points estimated by the Schwarz type criterion whose penalty term is greater than $2k(1 + \varepsilon_0) \log n$ for some $\varepsilon_0 > 0$. Zhang and Siegmund (2007) noted that the usage of the Schwarz criterion “is not theoretically justified” in their situation due to irregularities in the likelihood function and proposed a modified BIC derived as an asymptotic approximation of the Bayes factor to determine the number of change-points in means of normally distributed random variables. For other types of modifications and applications for detecting mean changes, see Ninomiya (2005), Pan and Chen (2006), and Hannart and Naveau (2012).

In the context of segmented line regression, similar approaches have been proposed to select the number of change-points. Kim et al. (2000) proposed the permutation test to select the number of change-points in the segmented line regression model where segments are assumed to be continuous at change-points, called the joinpoint regression model in their paper. Kim et al. (2009) considered the traditional BIC,

$$BIC(k) = -\frac{2}{n} SC(2k) = \log(RSS_k/n) + 2k \frac{\log n}{n},$$

where RSS_k is the residual sum of squares for the model with k change-points, and compared its performance with those of the permutation test procedure of Kim et al. (2000) and the method based on generalized cross validation used in MARS of Friedman (1991). Note that the penalty term of $2k \frac{\log n}{n}$ is chosen based on $2k + 3$ unknown parameters for the joinpoint regression model with k change-points. Liu et al. (1997) considered a general segmented line regression model allowing a discontinuity at the change-point and non-

Gaussian errors, proposed a penalty term with a bigger order than that of $BIC(k)$, and proved the consistency of the dimension selected by minimizing their criterion:

$$MIC(k) = \log(RSS_k / (n - p^*)) + p^* \frac{c_0 (\log n)^{2+\delta_0}}{n},$$

where $p^* = p^*(k) = (k + 1)p + k$ for the model with k change-points and p covariates and c_0 and δ_0 are positive constants. Two Bayesian model selection methods based on the Bayes factor and a Bayesian version of BIC were developed in Tiwari et al. (2005) who investigated their empirical properties via simulations and compared their performances with that of the permutation procedure of Kim et al. (2000). Martinez-Beneito et al. (2011) also proposed a Bayesian model selection method that provides posterior probabilities and is flexible to work with Poisson count data.

This paper is motivated from empirical results where the traditional BIC indicated a tendency to over-estimate the number of change-points (See Table 1 of Kim et al. (2009), Table 1 of Zhang and Siegmund (2007)). When the argument of Zhang and Siegmund (2007) is applied to segmented line regression, the penalty of the modified BIC is harsher than that of the traditional BIC, asymptotically corresponding to one additional unknown parameter per segment under some conditions, and this motivated us to consider a BIC type criterion whose penalty is $4k \frac{\log n}{n}$ for the segmented line regression model without the continuity constraint and $3k \frac{\log n}{n}$ for the model with the continuity constraint. Note that for segmented line regression with k change-points, the number of unknown parameters is $3k + 3$ for the model without the continuity constraint and $2k + 3$ for the model with the continuity constraint. Let

$$BIC_d(k) = \log(RSS_k/n) + PE_d(k) = \log(RSS_k/n) + dk \frac{\log n}{n}, \quad (1)$$

for penalty coefficients, d . Then the traditional BICs are BIC_3 for the unconstrained model and BIC_2 for the constrained model.

Our interest in this paper is on asymptotic behavior of BIC_d , a simple model selection criterion whose penalty term has the same order as that of the traditional BIC, and we focus on asymptotic properties of BIC_4 for the model without the continuity constraint and BIC_3 for the model with the continuity constraint. In Section 2, we formally introduce the unconstrained model and prove the consistency of the model dimension selected by BIC_4 for the unconstrained model with Gaussian errors. This result provides a consistent model selection criterion that imposes an asymptotically milder penalty than MIC does. In Section 3, we present the results of a simulation study where we compare the performance of BIC_4 with those of BIC_3 and MIC. Section 4 includes empirical results and discussion on the constrained case where the segments are constrained to be continuous at the change-points. Further discussion is presented in Section 5.

2 Selection Methods and Consistency: Unconstrained Model

Suppose that we observe $(x_1, y_1), \dots, (x_n, y_n)$ and consider a segmented line regression model such that

$$y_i = \beta_{j,0} + \beta_{j,1}x_i + \epsilon_i, \quad \text{if } \tau_{j-1} < x_i \leq \tau_j \quad (j=1, \dots, \kappa+1), \quad (2)$$

where κ is the unknown number of change-points, the τ 's are unknown change-points with $\tau_0 = \min_i x_i - \frac{1}{n}$ and $\tau_{\kappa+1} = \max_i x_i$, and the ϵ_i are independent $N(0, \sigma^2)$.

Let us consider the Schwarz type criterion (1) defined above and estimate κ as

$$\hat{\kappa} = \operatorname{argmin}_{0 \leq k \leq K} BIC_d(k),$$

where K is a pre-determined maximum number of change-points. For the segmented line regression model without the continuity constraint, (2), we consider BIC_3 as the traditional BIC based on $3k + 3$ unknown parameters, and study the large sample properties of BIC_4 , which is a modified version motivated by Zhang and Siegmund (2007).

Liu et al. (1997) considered a segmented line regression model that allows a discontinuity in the mean function and non-Gaussian errors, and their main idea to prove the consistency of $\hat{\kappa}$ was that the difference between the variances estimated under the true model and a model with a larger number of change-points is in the order of $\frac{(\log n)^2}{n}$, which seemed to have motivated their choice of the penalty term. Thus, arguments used in Liu et al. (1997) can not be applied in our situation where the penalty term is a constant multiple of $\frac{\log n}{n}$. Yao (1988) proved the consistency of the number of change-points estimated by the traditional BIC, BIC_2 , in the context of simple mean changes. He used properties of a normal distribution to show that the difference between the variances estimated under the true model and the model with a larger number of change-points than the true model, can be bounded above by the difference between the penalty terms of the two models. However, his argument to bound the difference does not work in our situation, and the upper bound in our case is achieved with $d = 4$ using properties of a chi-square distribution.

Rewrite the model (2) as

$$y_i = \beta_{j,0} + \beta_{j,1}x_i + \epsilon_i = \mathbf{x}_i' \boldsymbol{\beta}_j + \epsilon_i \quad \text{if } x_i \in (\tau_{j-1}, \tau_j], \quad (j=1, \dots, \kappa+1), \quad (3)$$

where $\mathbf{x}_i = (1, x_i)'$, $\boldsymbol{\beta}_j = (\beta_{j,0}, \beta_{j,1})'$ and the ϵ_i are independent and normally distributed with mean zero and variance σ^2 . Without loss of generality, we assume that $0 < x_i < 1$ for $i = 1, \dots, n$. Note that this model can be expressed as

$$\mathbf{y} = \sum_{j=1}^{\kappa+1} X(\tau_{j-1}, \tau_j) \boldsymbol{\beta}_j + \boldsymbol{\epsilon},$$

where $\mathbf{y} = (y_1, y_2, \dots, y_n)'$, $X = (\mathbf{1}, \mathbf{x})$ with $\mathbf{x} = (x_1, x_2, \dots, x_n)'$ and $\mathbf{1} = (1, 1, \dots, 1)'$, and $X(a, \eta) = I(a, \eta)X$ with $I(a, \eta) = \text{diag}(1_{\{x_1 \in (a, \eta]\}}, \dots, 1_{\{x_n \in (a, \eta]\}})$ for any $a < \eta$. Note that $1_A = 1$ if A is true and 0 otherwise.

Let $\mathcal{D}_k = \{(t_1, \dots, t_k)' : t_0 < t_1 < \dots < t_k < t_{k+1}\}$ with $t_0 = 0$ and $t_{k+1} = 1$. For a change-point vector, $\mathbf{t}_k = (t_1, \dots, t_k)' \in \mathcal{D}_k$, denote the design matrix and the corresponding hat matrix as

$$\mathbf{X}(\mathbf{t}_k) = (X(t_0, t_1), X(t_1, t_2), \dots, X(t_k, t_{k+1}))$$

and

$$H(\mathbf{t}_k) = \mathbf{X}(\mathbf{t}_k) \left(\mathbf{X}(\mathbf{t}_k)' \mathbf{X}(\mathbf{t}_k) \right)^{-1} \mathbf{X}(\mathbf{t}_k)' = \sum_{j=1}^{k+1} H(t_{j-1}, t_j),$$

where $H(a, \eta)$ is the projection matrix based on $X(a, \eta)$. Then the sum of squared residuals at given $\mathbf{t}_k$ is

$$S_n(\mathbf{t}_k) = \mathbf{y}' (I - H(\mathbf{t}_k)) \mathbf{y} = \sum_{j=1}^{k+1} S_n(t_{j-1}, t_j),$$

where $S_n(a, \eta) = \mathbf{y}'(I(a, \eta) - H(a, \eta))\mathbf{y}$, and thus the sum of squared residuals for the model with k change-points is obtained as

$$RSS_k = \inf_{\mathbf{t}_k} S_n(\mathbf{t}_k) = S_n(\hat{\tau}_k).$$

We also denote the true values of κ , σ^2 and $\boldsymbol{\tau}$ as κ_0 , σ_0^2 and $\boldsymbol{\tau}^0 = (\tau_1^0, \tau_2^0, \dots, \tau_{\kappa_0}^0)'$, respectively.

Assumption 1

The regressor variable x has a positive and continuous density function in any small neighborhoods of $\boldsymbol{\tau}^0$. Also x is independent of the error, ε .

For the case of nonrandom design points, the following assumption can replace Assumption 1. Note that the data spacing should be of order $O(1/n)$ to satisfy this assumption.

Assumption 1'

Let $\{\delta_n\}$ be a sequence of constants for which $O(1/n) \leq \delta_n = o(1)$. The number of data points in any small neighborhoods of τ_r^0 , ($r = 1, \dots, \kappa_0$) with volume δ_n is of order at least δ_n .

Theorem 1

Suppose that the segmented line regression model without the continuity constraint (3) satisfies one of Assumption 1 or 1'. Let $k < \kappa_0$. Then, for any constant $d > 0$,

$$P(BIC_d(k) > BIC_d(\kappa_0)) \rightarrow 1$$

as $n \rightarrow \infty$.

Proof

Let d be an arbitrary positive constant. By Lemma 5.3 and Lemma 5.4(i) of Liu et. al.(1997), it is true that for some $C > 0$,

$$\hat{\sigma}_k^2 = \frac{RSS_k}{n} = \frac{S_n(\hat{\tau}_k)}{n} > \sigma_0^2 + C$$

with probability going to 1 as $n \rightarrow \infty$. Since $\hat{\sigma}_{\kappa_0}^2 = \frac{RSS_{\kappa_0}}{n}$ converges to σ_0^2 in probability (that is, $\hat{\sigma}_{\kappa_0}^2 \xrightarrow{p} \sigma_0^2$), it is true that

$$BIC_d(k) - BIC_d(\kappa_0) = \log \left(\frac{\hat{\sigma}_k^2}{\hat{\sigma}_{\kappa_0}^2} \right) + \frac{d(k - \kappa_0)}{n} \log n > 0$$

with probability going to 1 as $n \rightarrow \infty$.

Theorem 1 guarantees that for any $d > 0$, $\hat{\kappa}$ based on BIC_d is consistent from the below. For the consistency of $\hat{\kappa}$ from the above, we now get the result for $d = 4$ under the assumption that the mean function of the regression model is discontinuous.

In order to prove that for $d = 4$ and for $k > \kappa_0$, $P(BIC_d(k) > BIC_d(\kappa_0)) \rightarrow 1$ as $n \rightarrow \infty$, we first state the following lemmas whose proofs are given in Appendix. In the lemmas and

Theorem 2, we let $\delta_n = \frac{(\log n)^2}{n}$, and for $1 \leq r \leq \kappa_0$, let

$$B_r(n) = \left\{ \mathbf{t}_k \in \mathcal{D}_k : |t_j - \tau_r^0| > \delta_n \text{ for } 1 \leq j \leq k \right\}.$$

Lemma 1

Suppose that the segmented line regression model without the continuity constraint (3) satisfies one of Assumption 1 or 1'. Then, for $k > \kappa_0$,

$$P(\hat{\tau}_k \in \cup_{r=1}^{\kappa_0} B_r(n)) \rightarrow 0$$

as $n \rightarrow \infty$.

Let $\xi_{k+3\kappa_0}$ be the ordered set of $\{t_k, \tau^0, \tau^0 - \delta_n \mathbf{1}, \tau^0 + \delta_n \mathbf{1}\}$. Let $T_r = (\tau_j^0 - \delta_n, \tau_j^0 + \delta_n]$ for $r = 1, \dots, \kappa_0$. Let $D_0 = (0, \tau_1^0 - \delta_n]$, $D_{\kappa_0} = (\tau_{\kappa_0}^0 + \delta_n, 1]$, and for $r = 1, \dots, \kappa_0 - 1$, let $D_r = (\tau_r^0 + \delta_n, \tau_{r+1}^0 - \delta_n]$. Let

$$G_{n,1} = \sum_{(\xi_{l=1}, \xi_l] \subset \cup_{r=1}^{\kappa_0} T_r} \epsilon' H(\xi_{l-1}, \xi_l) \epsilon,$$

$$G_{n,2} = \sum_{(\xi_{l=1}, \xi_l] \subset \cup_{r=0}^{\kappa_0} D_r} \epsilon' H(\xi_{l-1}, \xi_l) \epsilon.$$

Lemma 2

Suppose that the segmented line regression model without the continuity constraint (3) satisfies one of Assumption 1 or 1'. Then, for large n ,

$$G_{n,1} = O_p(\log \log n).$$

Lemma 3

Suppose that the segmented line regression model without the continuity constraint (3) satisfies one of Assumption 1 or 1'. Then, for $k > \kappa_0$ and any $t_k \in \cap_{r=1}^{\kappa_0} B_r(n)^c$, it is true that for any $\delta > 0$,

$$P(G_{n,2} \leq (k - \kappa_0 - 1) 4(1 + \delta) \sigma_0^2 \log n) \rightarrow 1$$

as $n \rightarrow \infty$.

Theorem 2

Suppose that the segmented line regression model without the continuity constraint (3) satisfies one of Assumption 1 or 1'. Let $k > \kappa_0$. Then, for any constant $d \geq 4$,

$$P(BIC_d(k) > BIC_d(\kappa_0)) \rightarrow 1$$

as $n \rightarrow \infty$.

Proof

Without loss of generality, we suppose $k = \kappa_0 + 1$. Since $\hat{\sigma}_{\kappa_0}^2 \xrightarrow{p} \sigma_0^2$ and $\hat{\sigma}_k^2 \xrightarrow{p} \sigma_0^2$, it is true that for any small $\delta > 0$,

$$\begin{aligned}
& BIC_d(k) - BIC_d(\kappa_0) \\
&= \log \left(\frac{S_n(\hat{\tau}_k)}{n} \right) - \log \left(\frac{S_n(\hat{\tau}_{\kappa_0})}{n} \right) + \frac{d(k-\kappa_0)}{n} \log n \\
&= \log \left(1 - \frac{\hat{\sigma}_{\kappa_0}^2 - \hat{\sigma}_k^2}{\hat{\sigma}_{\kappa_0}^2} \right) + \frac{d(k-\kappa_0)}{n} \log n \\
&\geq (1+\delta) \left(\frac{\hat{\sigma}_k^2 - \hat{\sigma}_{\kappa_0}^2}{\hat{\sigma}_{\kappa_0}^2} \right) + \frac{d(k-\kappa_0)}{n} \log n
\end{aligned}$$

with probability going to 1 as $n \rightarrow \infty$.

Since $P(\hat{\tau}_k \in \cup_{r=1}^{\kappa_0} B_r(n)) \rightarrow 0$ by Lemma 1, we now may focus on $\mathbf{t}_k$'s in $\cap_{r=1}^{\kappa_0} B_r(n)^c$. Let $\xi_{k+3\kappa_0}$ be the ordered set of $\{\mathbf{t}_k, \boldsymbol{\tau}^0, \boldsymbol{\tau}^0 - \delta_n \mathbf{1}, \boldsymbol{\tau}^0 + \delta_n \mathbf{1}\}$. Then

$$\begin{aligned}
& S_n(\mathbf{t}_k) - S_n(\hat{\tau}_{\kappa_0}) \\
&\geq S_n(\xi_{k+3\kappa_0}) - S_n(\boldsymbol{\tau}^0) \\
&\geq - \sum_{l=1}^{k+3\kappa_0+1} \epsilon' H(\xi_{l-1}, \xi_l) \epsilon \\
&= - \left\{ \sum_{\xi_{l-1}, \xi_l \in \cup_{r=1}^{\kappa_0} B_r} \epsilon' H(\xi_{l-1}, \xi_l) \epsilon + \sum_{(\xi_{l-1}, \xi_l) \in \cup_{r=0}^{\kappa_0} D_r} \epsilon' \mathbf{H}(\xi_{l-1}, \xi_l) \epsilon \right\} \\
&= - (G_{n,1} + G_{n,2}),
\end{aligned}$$

where $\xi_0 = 0$ and $\xi_{k+3\kappa_0+1} = 1$.

By Lemma 2 and Lemma 3, the following holds for any $\mathbf{t}_k \in \cap_{r=1}^{\kappa_0} B_r(n)^c$: it is true that for any $\delta > 0$,

$$S_n(\mathbf{t}_k) - S_n(\hat{\tau}_{\kappa_0}) > -(\delta + (k - \kappa_0 - 1)4(1+\delta))\sigma_0^2 \log n$$

and $\hat{\sigma}_{\kappa_0}^2 > \sigma_0^2(1 - \delta)$, with probability going to 1 as $n \rightarrow \infty$. Therefore, for any $\delta > 0$,

$BIC_d(k) - BIC_d(\kappa_0) > -(1+\delta)(\delta + (k - \kappa_0 - 1)4(1+\delta))\sigma_0^2 \log n / n\sigma_0^2(1 - \delta) + d(k - \kappa_0) \log n / n$ with probability going to 1 as $n \rightarrow \infty$. For $d \geq 4$, we can choose δ small enough such that

$$(1+\delta)(\delta + (k - \kappa_0 - 1)4(1+\delta)) / (1 - \delta) < d(k - \kappa_0),$$

and thus

$$BIC_d(k) - BIC_d(\kappa_0) > 0$$

with probability going to 1 as $n \rightarrow \infty$.

Remark

Theorem 1 and Theorem 2 imply that the consistency of $\hat{\kappa}_d$ is achieved by BIC_d for $d \geq 4$. Because the penalty term of MIC for the one independent variable case with $p = 1$,

$(2k+1) \frac{c_0(\log n)^{2+\delta_0}}{n}$, is larger than $4k \frac{\log n}{n}$ for large n , the result in this section indicates that the consistent model selection can be performed with a criterion whose penalty is asymptotically milder than that of MIC under the assumption of Gaussian errors.

3 Simulations: Unconstrained Model

This section summarizes simulations conducted to study asymptotic behavior of BIC_4 whose consistency is proved in Section 2 and to compare its performance with those of BIC_3 , the traditional BIC, and MIC whose consistency is established in Liu et al. (1997). For each case of the model parameters chosen, 300 replications of data sets are simulated, and $P(\hat{\kappa} < \kappa_0)$, $P(\hat{\kappa} = \kappa_0)$, and $P(\hat{\kappa} > \kappa_0)$ are estimated where κ_0 is the true number of change-points and $\hat{\kappa}$ ($0 \leq \hat{\kappa} \leq K$) is the estimate obtained by each selection method. The value of K was set as 4, and $\bar{3}$ or $\bar{6}$ represents a repeating decimal.

Tables 1, 2, and 3 summarize the results for unconstrained model cases. Tables 1 and 2 consider the cases with one change-point for two different x configurations, fixed x in Table 1 and random x in Table 2. Table 3 is for the case with two change-points where x is fixed. Note that MIC of Liu et al. (1997) uses the penalty term, $\text{PE}(k) = p^* c_0 (\ln n)^{2+\delta_0}/n = (2k+1)(0.299)(\ln n)^{2.1}/n$ for the model with one independent variable, which is based on $p^* = (k+1)p+k = 2k+1$ with $p = 1$ in their equation (2.3). According to Liu et al. (1997), $\delta_0 = 0.1$ is arbitrarily chosen and then $c_0 = 0.299$ is chosen by forcing the penalty term of MIC to be equal to that of the Schwarz criterion. For the model with one independent variable, the penalty term of MIC is set to be equal to that of BIC_2 at $n = 20$, but MIC is not exactly same as BIC_2 when $n = 20$ because MIC used an unbiased estimator of σ^2 while BIC_2 uses the maximum likelihood estimator of σ^2 under the Gaussian error assumption. When the full version of MIC is compared to BIC_d , the penalty term of MIC is similar to that of BIC_3 when $n = 30$ or 50 and to that of BIC_4 when $n = 150$ or 200 , respectively, which is empirically indicated in Tables 1, 2, and 3. For n very large, MIC imposes a penalty harsher than that of BIC_4 .

Results in Tables 1, 2, and 3 indicate that in the unconstrained case, the consistency of $\hat{\kappa}$ estimated by BIC_4 or MIC, which is theoretically proved, is empirically supported with $P(\hat{\kappa} = \kappa_0)$ approaching one as n increases or the effect size increases. Based on Theorem 1, a choice of d related to under-fitting is not much of an issue with n large, and our interest is on the over-fitting probabilities of these selection methods when n is large. The over-fitting probabilities of BIC_4 and MIC converge to zero as n increases, but the over-fitting probability of BIC_3 , the traditional BIC, is shown to be considerably larger than those of BIC_4 and MIC even for large n , although they seem to decrease as n increases.

In small sample situations, however, performances of these selection criteria depend more heavily on the effect size, the change size relative to the variability. For example, when the sample size is small and the effect size is small (for example, see the cases with $\sigma = 0.1$ and $n = 30$ or 50 in Table 1), the traditional BIC, BIC_3 , shows the highest probability of correct selection (PCP), while BIC_4 shows a very conservative performance. However, when the effect size is large enough to have the PCP of BIC_4 greater than 0.80, BIC_4 achieves the

highest PCP even in small sample cases (for example, see the cases with $\sigma = 0.05$ and $n = 30$ or 50 in Table 1). Regardless of the sample size, the over-fitting probabilities of the traditional BIC, BIC_3 , is significantly larger than that of BIC_4 . The performance of MIC is similar to that of BIC_3 when $n = 30, 50$ and to that of BIC_4 when $n = 150, 200$.

4 Selection Methods in Constrained Model

Recall the segmented line regression model introduced in (2):

$$y_i = \beta_{j,0} + \beta_{j,1}x_i + \epsilon_i, \quad \text{if } \tau_{j-1} < x_i \leq \tau_j \quad (j=1, \dots, \kappa+1).$$

When the segments are assumed to be continuous at the change-points,

$$\beta_{j,0} + \beta_{j,1}\tau_j = \beta_{j+1,0} + \beta_{j+1,1}\tau_j,$$

for $j = 1, \dots, \kappa$, and this model can also be represented as

$$y_i = \beta_0 + \beta_1x_i + \delta_1(x_i - \tau_1)^+ + \dots + \delta_\kappa(x_i - \tau_\kappa)^+ + \epsilon_i, \quad (4)$$

where $a^+ = \max(0, a)$.

For this segmented line regression model with the continuity constraint, called a joinpoint regression model in Kim et al. (2000), the total number of unknown parameters is $2k + 3$ for the model with k change-points, and so the traditional BIC uses $d = 2$ as in the simple mean change case. We call this traditional BIC as BIC_2 . Similarly as in the unconstrained case, we consider BIC_3 as a modified version in the constrained case.

Because Lemma 5.3 of Liu et al. (1997) also holds for the segmented line regression model with the continuity constraint, Theorem 1 holds for the constrained case as well. However, the arguments used to prove Theorem 2, more specifically the arguments used in the proof of Lemma 1, can not be used to prove that the over-fitting probability of BIC_3 converges to zero in the constrained case, and in this section, we investigate asymptotic behavior of BIC_3 via simulations. Although detailed proofs are provided only for the unconstrained model case, Liu et al. (1997) indicated the consistency of MIC in the constrained model case as well and the comparison of BIC_3 and MIC is expected to provide some insights on the consistency of BIC_3 .

Table 4, 5, and 6 include simulation results for constrained model cases, and we used the same simulation setting as in Tables 1, 2, and 3, except that in Table 4, $K = 3$ is used due to an excessive computation time. Tables 4, 5, and 6 include results for BIC_2 that is the traditional BIC for the constrained model, BIC_3 that we anticipate to be consistent, and MIC whose properties are studied in Liu et al. (1997). Note that MIC of Liu et al. (1997) uses the same penalty term for both constrained and unconstrained models, which was $(2k + 1)c_0(\ln n)^{2+\delta_0}/n = (2k + 1)(0.299)(\ln n)^{2.1}/n$ for the model with one independent variable.

The results summarized in Tables 4, 5, and 6 empirically support that the under-fitting probabilities of BIC_2 , BIC_3 , and MIC approaching zero as n increases. The over-fitting

probabilities of BIC_3 and MIC seem to converge to zero as the sample size increases and/or the effect size gets larger. Similarly as in the unconstrained case, the traditional BIC for the constrained case, BIC_2 , tends to overestimate the true number of change-points (κ), especially when the effect size is large, and the probability of overestimating κ is significantly different from zero even for $n = 150$ or 200 in most of cases, although it seems to decrease as n increases. Via numerical study not reported here, we also observed that the residual sum of squares of the constrained model fit change usually at a slower speed than those of the unconstrained model fit do as the number of change-points increases, and thus we expect the consistency of BIC_d to be achieved in the constrained model case with a less severe penalty than that of BIC_4 whose consistency was proved in the unconstrained model case.

5 Discussion

In this paper, we considered information based selection criteria to estimate the number of change-points in segmented line regression, and proved the consistency of the number of change-points estimated by BIC_4 for the unconstrained model with normal errors. The simulation results summarized in Tables 1, 2, and 3 for the unconstrained model support the theoretical results that (i) the under-fitting probability of BIC_d converges to zero for any $d > 0$ (Theorem 1) and (ii) the over-fitting probability of BIC_4 converges to zero (Theorem 2), while the traditional BIC, BIC_3 , indicates a non-negligible over-fitting tendency compared to those of BIC_4 and MIC.

For the constrained case where the line segments are assumed to be continuous at the change-points, we examined empirical properties of BIC_3 that is expected to improve large sample performance of the traditional BIC, BIC_2 . The fact that the under-fitting probability of BIC_d converges to zero for any $d > 0$, is empirically justified in Tables 4, 5, and 6. Significantly higher over-fitting probabilities of the traditional BIC, BIC_2 , is also observed even for large n , compared to those of BIC_3 and MIC.

In general, a relatively large penalty, that is, large d , will reduce the probability of over-estimating the number of change-points, but it will raise a risk of under-estimation. In that regards, our interest in this paper was on the smallest d for which BIC_d produces a consistent estimator $\hat{\kappa}$. In the normally distributed sequence of random variables, the traditional BIC, BIC_2 , produces a consistent estimator of κ according to Yao (1988), although its over-estimating tendency is observed even for $n = 497$ in Table 1 of Zhang and Siegmund (2007). Although the consistency of $\hat{\kappa}$ in the unconstrained case was proved only for $d \geq 4$ in our paper, BIC_d with $d < 4$ may provide a consistent model selection, and we hope future research to improve our results in this paper.

Another related future research problem is to identify a selection procedure that selects the best model for various sample sizes and effect sizes. In our simulation study, we observed that the traditional BIC performs better when the effect size is small, and in our future research, we plan to investigate small sample properties of various selection criteria. Any one type of criterion may not provide a uniform optimality, and a revised selection method

such as a weighted BIC, $\sum_d w_d BIC_d$, where the weight depends on effect sizes, sample size, etc., may achieve a relatively high probability of correct selection.

BIC_4/BIC_3 studied in this paper for the unconstrained/constrained model cases, respectively, were motivated from the modified BIC of Zhang and Siegmund (2007) that are derived as an asymptotic approximation of the Bayes factor. The original form of the Bayes factor could be used as a selection measure, instead of its asymptotic approximation. However, we have observed in our preliminary simulation study that even the performance of MBIC derived for the segmented line regression models depend on the x -configurations, and thus it may not be straightforward to study analytic properties of the criterion using the original Bayes factor.

Acknowledgements

A part of H.-J. Kim's research was conducted during her visit at National Cancer Institute. J. Kim's research was supported by a research grant from Inha University.

Appendix: Proofs of Lemmas

Lemma A.1

For any $\delta > 0$,

$$P \left(\sup_{\alpha < \eta} \epsilon' H(\alpha, \eta) \epsilon > 4\sigma_0^2 (1+\delta) \log n \right) \rightarrow 0$$

as $n \rightarrow \infty$.

Proof

Let X be a χ^2 distributed random variable with k degrees of freedom. For some constant $c_k > 0$,

$$P(X > x) = c_k x^{\frac{k}{2}-1} e^{-\frac{x}{2}} + o\left(x^{\frac{x}{2}-1} e^{-\frac{x}{2}}\right)$$

as $x \rightarrow \infty$. For given $\alpha < \eta$, $\frac{\epsilon' H(\alpha, \eta) \epsilon}{\sigma_0^2}$ has a χ^2 distribution with 2 degrees of freedom. So, it is true that for some constant $c_0 > 0$ and for $\delta > 0$,

$$\begin{aligned} & P \left(\sup_{\alpha < \eta} \epsilon' H(\alpha, \eta) \epsilon > 4\sigma_0^2 (1+\delta) \log n \right) \\ &= P \left(\max_{x_i < x_j} \epsilon' H(x_i, x_j) \epsilon > 4\sigma_0^2 (1+\delta) \log n \right) \\ &\leq c_0 \frac{n(n-1)}{2} n^{-2-2\delta} \\ &\rightarrow 0 \end{aligned}$$

as n gets large.

Lemma A.2

For $1 \leq j \leq n-2$,

$$\max_{2 \leq k \leq n-j} \epsilon' H(x_j, x_{j+k}) \epsilon = O_p(\log \log n),$$

and for $3 \leq j \leq n$,

$$\max_{2 \leq k \leq j-1} \epsilon' H(x_{j-k}, x_j) \epsilon = O_p(\log \log n).$$

Proof

For $2 \leq k \leq n-j$, the rank of $H(x_j, x_{j+k})$ is 2, and so there exists $2 \times n$ matrix $Q(k)$ with full row rank 2, such that

$$\epsilon' H(x_j, x_{j+k}) \epsilon = \epsilon' Q(k)' Q(k) \epsilon = \sum_{l=1}^2 u(k)_l^2,$$

where $Q(k)' = (q(k)_1, q(k)_2)$ and $u(k)_l = q(k)_l' \epsilon$, $l = 1, 2$. Then

$\sum_{l=1}^2 q(k)_l' q(k)_l = \text{tr}(H(x_j, x_{j+k})) = 2$. For each k and integer i , let $a_{lki} = \sqrt{k} q(k)_l$, so that

$$\epsilon' H(x_j, x_{j+k}) \epsilon = \frac{S_{1k}^2 + S_{2k}^2}{k},$$

where $S_{lk} = \sum_{i=j+1}^{j+k} a_{lki} \epsilon_i$. Then by Theorem 1 of Lai and Wei(1982),

$$\max_{2 \leq k \leq n-j} \epsilon' H(x_j, x_{j+k}) \epsilon = O_p(\log \log n).$$

Similarly, we can show that

$$\max_{2 \leq k \leq j-1} \epsilon' H(x_{j-k}, x_j) \epsilon = O_p(\log \log n).$$

Proof of Lemma 1

It is good enough to consider the case of $k = \kappa_0 + 1$. For any $t_k \in B_r(n)$, let $\xi_{k+\kappa_0+1}$ be an ordered set of $\{t_k, \tau_1^0, \dots, \tau_{r-1}^0, \tau_r^0 - \delta_n, \tau_r^0 + \delta_n, \tau_{r+1}^0, \dots, \tau_{\kappa_0}^0\}$. Then

$$S_n(t_k) - S_n(\tau^0, 1) \geq S_n(\xi_{k+\kappa_0+1}) - S_n(\tau^0) \geq S_n(\xi_{k+\kappa_0+1}) - \epsilon'.$$

For $(\xi_{l-1}, \xi_l] \not\subset (\tau_r^0 - \delta_n, \tau_r^0 + \delta_n]$,

$$S_n(\xi_{l-1}, \xi_l) = \epsilon' (I(\xi_{l-1}, \xi_l) - H(\xi_{l-1}, \xi_l)) = \sum_{i \in \{i: x_i \in (\xi_{l-1}, \xi_l]\}} \epsilon_i^2 - O_p(\log n)$$

by Lemma A.1.

For $1 \leq r \leq \kappa_0$, let $A_r(n) = \{i: \tau_r^0 - \delta_n < x_i \leq \tau_r^0 + \delta_n\}$. Then

$$\begin{aligned} S_n(\tau_r^0 - \delta_n, \tau_r^0 + \delta_n) &= \sum_{i \in A_r(n)} (E(y_i) + \epsilon_i - \hat{y}_i)^2 \\ &= \sum_{i \in A_r(n)} \left\{ \epsilon_i^2 + (E(y_i) - \hat{y}_i)^2 + 2(E(y_i) - \hat{y}_i) \epsilon_i \right\} \end{aligned}$$

Since $E(y)$ is discontinuous at τ_r^0 while $\hat{y}$ is continuous on $(\tau_r^0 - \delta_n, \tau_r^0 + \delta_n)$, for some constant c_0 ,

$$(E(y_i) - \hat{y}_i)^2 > c_0$$

for $i \in I_r \subset A_r(n)$ for which $\#(I_r)$ is of order at least $n\delta_n$. So by the weak law of large numbers, there exists $c_1 > 0$ such that

$$S_n(\tau_r^0 - \delta_n, \tau_r^0 + \delta_n) - \sum_{i \in A_r(n)} \epsilon_i^2 > c_1 \delta_n n = c_1 (\log n)^2.$$

Then

$$S_n(\xi_{k+\kappa_0+1}) - \epsilon' \epsilon > -O_p(\log n) + c_1 (\log n)^2 > 0$$

for large n .

Proof of Lemma 2

The number of observations in subintervals $(\tau_r^0 - \delta_n, \tau_r^0)$ and $(\tau_r^0, \tau_r^0 + \delta_n)$ is of order $(\log n)^2$ by Assumption 1 or 1'. Then, by Lemma A.1,

$$G_{n,1} = O_p\left(\log\left((\log n)^2\right)\right) = O_p(\log \log n).$$

Proof of Lemma 3

For $t_k = (t_1, \dots, t_k) \in \cap_{r=1}^{\kappa_0} B_r(n)^c$,

$$\# \{t_j: t_j \in \cup_{r=0}^{\kappa_0} D_r\} \leq k - \kappa_0.$$

Suppose that $k - \kappa_0$ such points, $\xi'_1, \dots, \xi'_{k-\kappa_0}$, are in D_r for some r . Then

$$G_{n,2} = \epsilon' H(0, \tau_1^0 - \delta_n) \epsilon + \epsilon' H(\tau_{\kappa_0}^0 + \delta_n, \mathbf{1}) \epsilon + \sum_{j \neq r} \epsilon' H(\tau_j^0 + \delta_n, \tau_{j+1}^0 - \delta_n) \epsilon + \sum_{(\xi_{1-1}, \xi_1] \subset D_r} \epsilon' H(\xi_{1-1}, \xi_1) \epsilon.$$

Since $\epsilon' H(a, \eta) \epsilon$ for given (a, η) is $O_p(1)$, the first three terms of the right-hand side are $O_p(1)$. So,

$$G_{n,2} = O_p(1) + \epsilon' H(\tau_r^0 + \delta_n, \xi'_1) \epsilon + \epsilon' H(\xi'_{k-\kappa_0}, \tau_{r+1}^0 - \delta_n) \epsilon + \sum_{j=1}^{k-\kappa_0-1} \epsilon' H(\xi'_j, \xi'_{j+1}) \epsilon.$$

By Lemma A.2, the second and the third terms of the right-hand side of the above equation are of order $\log \log n$. Then, by Lemma A.1, it is true that for any $\delta > 0$,

$$\sum_{j=1}^{k-\kappa_0-1} \epsilon' H(\xi'_j, \xi'_{j+1}) \epsilon \leq (k - \kappa_0 - 1) 4(1+\delta) \sigma_0^2 \log n,$$

with probability going to 1 as $n \rightarrow \infty$. If the points $\xi'_1, \dots, \xi'_{k-\kappa_0}$ are located in distinct D_r 's, we can get an even smaller upper bound.

References

- Friedman JH. Multivariate adaptive regression splines. *Annals of Statistics*. 1991; 19:1–67.
- Hannart A, Naveau P. An improved Bayesian information criterion for multiple change-point models. *Journal of the American Statistical Association*. 2012; 54:256–268.
- Kim H-J, Fay MP, Feuer EJ, Midthune DN. Permutation tests for joinpoint regression with applications to cancer rates. *Statistics in Medicine*. 2000; 19:335–351. [PubMed: 10649300]
- Kim H-J, Yu B, Feuer EJ. Selecting the number of change-points in segmented line regression. *Statistica Sinica*. 2009; 19(2):597–609. [PubMed: 19738935]
- Lee C-B. Estimating the number of change points in exponential families distributions. *Scandinavian Journal of Statistics*. 1997; 24:201–210.
- Liu J, Wu S, Zidek JV. On segmented multivariate regression. *Statistica Sinica*. 1997; 7:497–525.
- Martinez-Beneito MA, Garcia-Donato G, Salmeron D. A Bayesian join-point regression model with an unknown number of break-points. *The Annals of Applied Statistics*. 2011; 5:2150–2168.
- Ninomiya Y. Information criterion for Gaussian change-point model. *Statistics & Probability Letters*. 2005; 72:237–247.
- Pan J, Chen J. Application of modified information criterion to multiple change point problems. *Journal of Multivariate Analysis*. 2006; 97:2221–2241.
- Schwarz G. Estimating the dimension of a model. *The Annals of Statistics*. 1978; 6:461–464.

- Tiwari RC, Cronin KA, Davis W, Feuer EJ, Yu B. Bayesian model selection for join point regression with application to age-adjusted cancer rates. *Applied Statistics*. 2005; 54:919–939.
- Yao YC. Estimating the number of change-points via Schwarz criterion. *Statistics and Probability Letters*. 1988; 6:181–189.
- Zhang NR, Siegmund DO. A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*. 2007; 63:22–32. [PubMed: 17447926]

Highlights

- Segmented line regression models, constrained and unconstrained, are studied.
- Consistency of the estimated number of change-points is investigated.
- Simulation studies are conducted to assess the performance of the proposed criteria.

Table 1

Unconstrained model with $k_0 = 1$ and $x_i = i/n$ for $i = 1, \dots, n$. $E(Y|x) = (1 + x)I(x \leq 0.5) + (1.35 + 0.5x)I(x > 0.5)$

n	Prob	$\sigma = 0.1$			$\sigma = 0.05$		
		BIC ₃	BIC ₄	MIC	BIC ₃	BIC ₄	MIC
30	$\hat{P}(\hat{\kappa} < k_0)$	—	—	—	0.030	—	—
	$\hat{P}(\hat{\kappa} = k_0)$	0.476	0.763	0.446	—	0.123	0.026
	$\hat{P}(\hat{\kappa} > k_0)$	0.360	0.210	—	0.723	0.830	0.730
		—	—	0.383	—	—	—
50	$\hat{P}(\hat{\kappa} < k_0)$	0.163	0.026	0.170	0.246	0.046	0.243
	$\hat{P}(\hat{\kappa} = k_0)$	—	0.670	0.450	0	0.010	—
	$\hat{P}(\hat{\kappa} > k_0)$	0.386	0.320	0.510	—	—	0.003
		—	0.010	0.040	0.853	0.973	0.910
75	$\hat{P}(\hat{\kappa} < k_0)$	0.526	—	—	—	—	—
	$\hat{P}(\hat{\kappa} = k_0)$	—	—	—	0.146	0.016	0.086
	$\hat{P}(\hat{\kappa} > k_0)$	0.086	—	—	—	—	—
		—	—	—	—	—	—
150	$\hat{P}(\hat{\kappa} < k_0)$	0.193	0.493	0.346	0	—	—
	$\hat{P}(\hat{\kappa} = k_0)$	—	—	—	0.923	0.003	0.003
	$\hat{P}(\hat{\kappa} > k_0)$	0.746	0.506	0.623	—	0.980	—
		0.060	0	0.030	0.076	0.016	0.976
200	$\hat{P}(\hat{\kappa} < k_0)$	—	—	0.130	0	0	0
	$\hat{P}(\hat{\kappa} = k_0)$	0.043	0.146	—	—	0.990	0.990
	$\hat{P}(\hat{\kappa} > k_0)$	0.900	—	0.863	0.953	0.010	0.010
		—	0.846	—	—	—	—
200	$\hat{P}(\hat{\kappa} < k_0)$	0.056	—	0.006	0.046	—	—
	$\hat{P}(\hat{\kappa} = k_0)$	—	0.006	—	—	—	—
	$\hat{P}(\hat{\kappa} > k_0)$	0.006	—	—	—	—	—
		—	0.040	—	0	0	0
200	$\hat{P}(\hat{\kappa} < k_0)$	0.006	—	0.043	—	—	—
	$\hat{P}(\hat{\kappa} = k_0)$	—	0.956	—	0.963	0.996	0.996
	$\hat{P}(\hat{\kappa} > k_0)$	0.963	—	0.953	—	—	—
		0.030	0.003	0.003	0.036	0.003	0.003

Table 2

Unconstrained model with $k_0 = 1$ and $x_i \sim N(0.5, 1)$ for $i = 1, \dots, n$. $E(Y|x) = (1 + x)I(x \leq 0.5) + (1.35 + 0.5x)I(x > 0.5)$

n	Prob	$\sigma = 0.2$			$\sigma = 0.1$		
		BIC ₃	BIC ₄	MIC	BIC ₃	BIC ₄	MIC
30	$\hat{P}(\hat{\kappa} < k_0)$	—	0.300	—	0	—	0
	$\hat{P}(\hat{\kappa} = k_0)$	0.083	—	0.076	0.760	0.006	0.760
	$\hat{P}(\hat{\kappa} > k_0)$	0.703	0.666	—	0.240	—	0.240
		—	0.033	0.706	—	0.943	—
50	$\hat{P}(\hat{\kappa} < k_0)$	—	—	—	—	0.050	—
	$\hat{P}(\hat{\kappa} = k_0)$	0.016	0.076	0.023	0	0	0
	$\hat{P}(\hat{\kappa} > k_0)$	0.830	—	—	0.980	—	—
		—	0.896	0.893	0.020	0.913	—
150	$\hat{P}(\hat{\kappa} < k_0)$	—	—	—	—	—	—
	$\hat{P}(\hat{\kappa} = k_0)$	0.153	0.026	0.083	0.843	—	0.086
	$\hat{P}(\hat{\kappa} > k_0)$	—	—	—	0.156	—	—
		0	0	0	0	0	0
	$\hat{P}(\hat{\kappa} < k_0)$	—	0.990	0.990	—	—	—
	$\hat{P}(\hat{\kappa} = k_0)$	0.953	0.010	0.010	0.956	0.996	0.996
	$\hat{P}(\hat{\kappa} > k_0)$	—	—	—	—	—	—
		0.046	—	—	0.043	0.003	0.003

Table 3

Unconstrained model with $k_0 = 2$ and $x_i = i/n$ for $i = 1, \dots, n$. $E(Y|x) = (1 + x)I(x \leq 0.5) + (1.35 + 0.5x)I(0.5 < x \leq 0.7) + (0.75 + 1.5x)I(x > 0.7)$

n	Prob	$\sigma = 0.05$			$\sigma = 0.02$		
		BIC ₃	BIC ₄	MIC	BIC ₃	BIC ₄	MIC
30	$\hat{P}(\hat{\kappa} < k_0)$	0.590	—	0.590	—	—	—
	$\hat{P}(\hat{\kappa} = k_0)$	—	0.873	—	0.026	0.113	0.026
	$\hat{P}(\hat{\kappa} > k_0)$	0.243	—	0.243	—	—	—
		—	0.103	—	0.673	0.793	0.686
		0.166	—	0.166	0.300	—	—
			0.023			0.093	0.286
50	$\hat{P}(\hat{\kappa} < k_0)$	—	0.890	—	0	0	0
	$\hat{P}(\hat{\kappa} = k_0)$	0.636	—	0.723	—	—	—
	$\hat{P}(\hat{\kappa} > k_0)$	—	0.106	0.250	0.823	0.966	0.893
		0.306	—	—	—	—	—
		—	0.003	0.026	0.176	0.033	0.106
		0.056					
150	$\hat{P}(\hat{\kappa} < k_0)$	0.190	—	—	0	0	0
	$\hat{P}(\hat{\kappa} = k_0)$	0.760	0.386	0.376	—	0.990	0.990
	$\hat{P}(\hat{\kappa} > k_0)$	0.050	—	—	0.946	0.010	0.010
			0.606	0.616	—		
			—	—	0.053		
			0.006	0.006			

Table 4

Constrained model with $k_0 = 1$ and $x_i = i/n$ for $i = 1, \dots, n$. $E(Y|x) = 1 + x - 0.5(x - 0.5)^+$

n	Prob	$\sigma = 0.1$			$\sigma = 0.05$		
		BIC ₂	BIC ₃	MIC	BIC ₂	BIC ₃	MIC
30	$\hat{P}(\hat{\kappa} < k_0)$	—	0.810	—	—	—	—
	$\hat{P}(\hat{\kappa} = k_0)$	0.613	—	0.803	0.046	0.176	0.156
	$\hat{P}(\hat{\kappa} > k_0)$	0.340	0.186	—	0.880	—	0.830
		—	—	0.193	—	0.813	—
50	$\hat{P}(\hat{\kappa} < k_0)$	0.046	0.003	—	0.073	0.010	0.013
				0.003			
	$\hat{P}(\hat{\kappa} = k_0)$	—	0.730	0.770	—	0.040	0.060
	$\hat{P}(\hat{\kappa} > k_0)$	0.456	0.270	0.230	0.006	—	—
75	$\hat{P}(\hat{\kappa} < k_0)$	—	0	0	0.920	0.953	0.936
					—	—	—
	$\hat{P}(\hat{\kappa} = k_0)$	0.513	—	—	0.073	0.006	0.003
	$\hat{P}(\hat{\kappa} > k_0)$	0.030	—	—			
150	$\hat{P}(\hat{\kappa} < k_0)$	—	—	—	0	—	—
					—	—	—
	$\hat{P}(\hat{\kappa} = k_0)$	0.306	0.646	0.756	0.936	0.003	0.013
	$\hat{P}(\hat{\kappa} > k_0)$	0.650	0.350	—	—	0.990	—
200	$\hat{P}(\hat{\kappa} < k_0)$	—	—	0.243	—	—	0.983
				0	0.063	0.006	—
	$\hat{P}(\hat{\kappa} = k_0)$	0.043	0.003	—			0.003
	$\hat{P}(\hat{\kappa} > k_0)$	—	—	—			
150	$\hat{P}(\hat{\kappa} < k_0)$	0.100	0.280	—	0	0	0
		—	—	0.493	—	—	1
	$\hat{P}(\hat{\kappa} = k_0)$	0.873	0.716	—	0.966	0.996	0
	$\hat{P}(\hat{\kappa} > k_0)$	—	—	0.506	—	—	
200	$\hat{P}(\hat{\kappa} < k_0)$	0.026	0.003	0	0.033	0.003	
	$\hat{P}(\hat{\kappa} = k_0)$	—	—	—	0	0	0
	$\hat{P}(\hat{\kappa} > k_0)$	0.023	0.116	0.353	—	1	1
200	$\hat{P}(\hat{\kappa} < k_0)$	—	—	—	0.983	0	0
					—		
	$\hat{P}(\hat{\kappa} = k_0)$	0.963	0.883	0.646	—		
	$\hat{P}(\hat{\kappa} > k_0)$	—	0	0	0.016		
200	$\hat{P}(\hat{\kappa} < k_0)$	0.013	—	—			
	$\hat{P}(\hat{\kappa} = k_0)$	—	—	—			
	$\hat{P}(\hat{\kappa} > k_0)$	—	—	—			

Table 5

Constrained model with $k_0 = 1$ and $x_i \sim N(0.5, 1)$ for $i = 1, \dots, n$. $E(Y|x) = 1 + x - 0.5(x - 0.5)^+$

n	Prob	$\sigma = 0.2$			$\sigma = 0.1$		
		BIC ₂	BIC ₃	MIC	BIC ₂	BIC ₃	MIC
30	$\hat{P}(\hat{\kappa} < k_0)$	0.100	0.290	—	0	—	—
	$\hat{P}(\hat{\kappa} = k_0)$	—	—	0.276	0.910	0.003	0.003
	$\hat{P}(\hat{\kappa} > k_0)$	0.823	0.706	0.720	0.090	—	—
		—	—	—	—	0.983	0.983
50	$\hat{P}(\hat{\kappa} < k_0)$	0.076	0.003	0.003	—	—	—
	$\hat{P}(\hat{\kappa} = k_0)$	—	—	—	0.013	0.013	0.013
	$\hat{P}(\hat{\kappa} > k_0)$	—	—	—	—	—	—
		—	—	—	—	—	—
150	$\hat{P}(\hat{\kappa} < k_0)$	—	0.080	—	0	0	0
	$\hat{P}(\hat{\kappa} = k_0)$	0.016	—	0.103	—	—	—
	$\hat{P}(\hat{\kappa} > k_0)$	—	0.916	—	0.936	0.986	0.993
		0.926	—	0.893	—	—	—
	$\hat{P}(\hat{\kappa} < k_0)$	—	0.003	—	0.063	0.013	0.006
	$\hat{P}(\hat{\kappa} = k_0)$	0.056	—	0.063	—	—	—
	$\hat{P}(\hat{\kappa} > k_0)$	—	—	—	—	—	—
		—	—	—	—	—	—

Table 6

Constrained model with $k_0 = 2$ and $x_i = i/n$ for $i = 1, \dots, n$. $E(Y|x) = 1 + x - 0.5(x - 0.5)^+ + (x - 0.7)^+$

n	Prob	$\sigma = 0.5$			$\sigma = 0.02$		
		BIC ₂	BIC ₃	MIC	BIC ₂	BIC ₃	MIC
30	$\hat{P}(\hat{\kappa} < k_0)$	—	0.860	—	0	—	—
	$\hat{P}(\hat{\kappa} = k_0)$	0.586	0.140	0.853	0.920	0.026	0.023
		—	0	—	0.080	—	0.960
	$\hat{P}(\hat{\kappa} > k_0)$	0.363		0.146		0.956	—
50		0.050		0		—	0.016
						0.016	
	$\hat{P}(\hat{\kappa} < k_0)$	—	—	0.890	0	—	0.020
	$\hat{P}(\hat{\kappa} = k_0)$	0.503	0.816	0.110	—	0.013	—
150		—	—	0	0.893	—	0.966
	$\hat{P}(\hat{\kappa} > k_0)$	0.433	0.183		—	0.963	—
		—	0		0.106	—	0.013
		0.063				0.023	
150	$\hat{P}(\hat{\kappa} < k_0)$	—	0.280	0.530	0	0	0
	$\hat{P}(\hat{\kappa} = k_0)$	0.093	—	0.470	—	—	1
		—	0.713	0	0.933	0.996	0
	$\hat{P}(\hat{\kappa} > k_0)$	0.863	—		—	—	
150		—	0.006		0.066	0.003	
		0.043					