The Annals of Statistics
2014, Vol. 42, No. 3, 970–1002
DOI: 10.1214/14-AOS1210

NONPARAMETRIC MAXIMUM LIKELIHOOD APPROACH TO MULTIPLE CHANGE-POINT PROBLEMS[1]

BY CHANGLIANG ZOU, GUOSHENG YIN, LONG FENG AND ZHAOJUN WANG

Nankai University, University of Hong Kong, Nankai University and Nankai University

In multiple change-point problems, different data segments often follow different distributions, for which the changes may occur in the mean, scale or the entire distribution from one segment to another. Without the need to know the number of change-points in advance, we propose a nonparametric maximum likelihood approach to detecting multiple change-points. Our method does not impose any parametric assumption on the underlying distributions of the data sequence, which is thus suitable for detection of any changes in the distributions. The number of change-points is determined by the Bayesian information criterion and the locations of the change-points can be estimated via the dynamic programming algorithm and the use of the intrinsic order structure of the likelihood function. Under some mild conditions, we show that the new method provides consistent estimation with an optimal rate. We also suggest a prescreening procedure to exclude most of the irrelevant points prior to the implementation of the nonparametric likelihood method. Simulation studies show that the proposed method has satisfactory performance of identifying multiple change-points in terms of estimation accuracy and computation time.

1. Introduction. The literature devoted to change-point models is vast, particularly in the areas of economics, genome research, quality control, and signal processing. When there are notable changes in a sequence of data, we can typically break the sequence into several data segments, so that the observations within each segment are relatively homogeneous. In the conventional change-point problems, the posited models for different data segments are often of the same structure but with different parameter values. However, the underlying distributions are typically unknown, and thus parametric methods potentially suffer from model misspecification. The least-squares fitting is the standard choice for the MCP, while its

Received November 2013; revised February 2014.

[1]Supported in part by the NNSF of China Grants 11131002, 11101306, 11371202, the RFDP of China Grant 20110031110002, the Foundation for the Author of National Excellent Doctoral Dissertation of PR China 201232, New Century Excellent Talents in University and also by a Grant (784010) from the Research grants Council of Hong Kong.

MSC2010 subject classifications. Primary 62G05; secondary 62G20.

Key words and phrases. BIC, change-point estimation, Cramér–von Mises statistic, dynamic programming, empirical distribution function, goodness-of-fit test.

performance often deteriorates when the error follows a heavy-tailed distribution or when the data contain outliers.

Without imposing any parametric modeling assumption, we consider the multiple change-point problem (MCP) based on independent data $\{X_i\}_{i=1}^n$, such that

$$X_i \sim F_k(x), \qquad \tau_{k-1} \le i \le \tau_k - 1, k = 1, \ldots, K_n + 1; i = 1, \ldots, n, \tag{1.1}$$

where K_n is the true number of change-points, τ_k's are the locations of these change-points with the convention of $\tau_0 = 1$ and $\tau_{K_n+1} = n + 1$, and F_k is the cumulative distribution function (C.D.F.) of segment k satisfying $F_k \neq F_{k+1}$. The number of change-points K_n is allowed to grow with the sample size n.

Although extensive research has been conducted to estimate the number of change-points K_n and the locations of these change-points τ_k's, most of the work assumes that F_k's belong to some-known parametric functional families or that they differ only in their locations (or scales). For a comprehensive coverage on single change-point problems ($K_n = 1$), see Csörgő and Horváth (1997). The standard approach to the MCP is based on least-squares or likelihood methods via a dynamic programming (DP) algorithm in conjunction with a selection procedure such as the Bayesian information criterion (BIC) for determining the number of change-points [Yao (1988); Yao and Au (1989); Chen and Gupta (1997); Bai and Perron (1998, 2003); Braun, Braun and Müller (2000); Hawkins (2001); Lavielle (2005)]. By reframing the MCP in a variable selection context, Harchaoui and Lévy-Leduc (2010) proposed a penalized least-squares criterion with a LASSO-type penalty [Tibshirani (1996)]. Chen and Zhang (2012) developed a graph-based approach to detecting change-points, which is applicable in high-dimensional data and non-Euclidean data. Other recent development in this area includes Rigaill (2010), Killick, Fearnhead and Eckley (2012) and Arlot, Celisse and Harchaoui (2012).

Our goal is to develop an efficient nonparametric procedure for the MCP in (1.1) without imposing any parametric structure on the F_k's; virtually any salient difference between two successive C.D.F.'s (say, F_k and F_{k+1}) would ensure detection of the change-point asymptotically. In the nonparametric context, most of the existing work focuses on the single change-point problem by using some seminorm on the difference between pre- and post-empirical distributions at the change-point [Darkhovskh (1976); Carlstein (1988); Dümbgen (1991)]. Guan (2004) studied a semiparametric change-point model based on the empirical likelihood, and applied the method to detect the change from a distribution to a weighted one. Zou et al. (2007) proposed another empirical likelihood approach without assuming any relationship between the two distributions. However, extending these methods to the MCP is not straightforward. Lee (1996) proposed to use the weighted empirical measure to detect two different nonparametric distributions over a window of observations and then run the window through the full data sequence to detect the number of change-points. Although the approach of Lee (1996) is simple and

easy to implement, our simulation studies show that even with elaborately chosen tuning parameters the estimates of the locations τ_k's as well as the number of change-points are not satisfactory. This may be partly due to the "local" nature of the running window, and thus the information in the data is not fully and efficiently utilized. Matteson and James (2014) proposed a new estimation method, ECP, under multivariate settings, which is based on hierarchical clustering by recursively using a single change-point estimation procedure.

Observing the connection between multiple change-points and goodness-of-fit tests, we propose a nonparametric maximum likelihood approach to the MCP. Our proposed nonparametric multiple change-point detection (NMCD) procedure can be regarded as a nonparametric counterpart of the classical least-squares MCP method [Yao (1988)]. Under some mild conditions, we demonstrate that the NMCD can achieve the optimal rate, $O_p(1)$, for the estimation of the change-points without any distributional assumptions. Due to the use of empirical distribution functions, technical arguments for controlling the supremum of the nonparametric likelihood function are nontrivial and are interesting in their own rights. As a matter of fact, some techniques regarding the empirical process have been nicely integrated with the MCP methodologies. In addition, our theoretical results are applicable to the situation with a diverging number of change-points, that is, when the number of change-points, K_n, grows as n goes to infinity. This substantially enlarges the scope of applicability of the proposed method, from a traditional fixed dimensionality to a more challenging high-dimensional setting.

In the proposed NMCD procedure, the number of change-points, K_n, is determined by the BIC. Given K_n, the DP algorithm utilizes the intrinsic order structure of the likelihood to recursively compute the maximizer of the objective function with a complexity of $O(K_n n^2)$. To exclude most of the irrelevant points, we also suggest an initial screening procedure so that the NMCD is implemented in a much lower-dimensional space. Compared with existing parametric and nonparametric approaches, the proposed NMCD has satisfactory performance of identifying multiple change-points in terms of estimation accuracy and computation time. It offers robust and effective detection capability regardless of whether the F_k's differ in the location, scale, or shape.

The remainder of the paper is organized as follows. In Section 2, we first describe how to recast the MCP in (1.1) into a maximization problem and then introduce our nonparametric likelihood method followed by its asymptotic properties. The algorithm and practical implementation are presented in Section 3. The numerical performance and comparisons with other existing methods are presented in Section 4. Section 5 contains a real data example to illustrate the application of our NMCD method. Several remarks draw the paper to its conclusion in Section 6. Technical proofs are provided in the Appendix, and the proof of a corollary and additional simulation results are given in the supplementary material [Zou et al. (2014)].

2. Nonparametric multiple change-point detection.

2.1. *NMCD method.* Assume that $Z_1, \ldots, Z_n$ are independent and identically distributed from F_0, and let $\widehat{F}_n$ denote the empirical C.D.F. of the sample, then $n\widehat{F}_n(u) \sim \text{Binomial}(n, F_0(u))$. If we regard the sample as binary data with the probability of success $\widehat{F}_n(u)$, this leads to the nonparametric maximum log-likelihood

$$n\{\widehat{F}_n(u)\log(\widehat{F}_n(u)) + (1-\widehat{F}_n(u))\log(1-\widehat{F}_n(u))\}.$$

In the context of (1.1), we can write the joint log-likelihood for a candidate set of change-points $(\tau_1' < \cdots < \tau_L')$ as

$$\begin{aligned}\mathcal{L}_u(\tau_1', \ldots, \tau_L') = \sum_{k=0}^{L} (\tau_{k+1}' - \tau_k')\{&\widehat{F}_{\tau_k'}^{\tau_{k+1}'}(u)\log(\widehat{F}_{\tau_k'}^{\tau_{k+1}'}(u)) \\ &+ (1-\widehat{F}_{\tau_k'}^{\tau_{k+1}'}(u))\log(1-\widehat{F}_{\tau_k'}^{\tau_{k+1}'}(u))\},\end{aligned} \tag{2.1}$$

where $\widehat{F}_{\tau_k'}^{\tau_{k+1}'}(u)$ is the empirical C.D.F. of the subsample $\{X_{\tau_k'}, \ldots, X_{\tau_{k+1}'-1}\}$ with $\tau_0' = 1$ and $\tau_{L+1}' = n+1$. To estimate the change-points $1 < \tau_1' < \cdots < \tau_L' \le n$, we can maximize (2.1) in an integrated form

$$R_n(\tau_1', \ldots, \tau_L') = \int_{-\infty}^{\infty} \mathcal{L}_u(\tau_1', \ldots, \tau_L')\, dw(u), \tag{2.2}$$

where $w(\cdot)$ is some positive weight function so that $R_n(\cdot)$ is finite, and the integral is used to combine all the information across u. The rationale of using (2.2) can be clearly seen from the behavior of its population counterpart. For simplicity, we assume that there exists only one change-point τ_1, and let $\tau_1/n \to q_1 \in (0,1)$ and $\tau_1'/n \to \theta \in (0,1)$. Through differentiation with respect to θ, it can be verified that the limiting function of $\mathcal{L}_u(\tau_1')/n$,

$$\begin{aligned}Q_u(\theta) = {}&\theta\{F_\theta^{(1)}(u)\log(F_\theta^{(1)}(u)) + (1-F_\theta^{(1)}(u))\log(1-F_\theta^{(1)}(u))\} \\ &+ (1-\theta)\{F_\theta^{(2)}(u)\log(F_\theta^{(2)}(u)) + (1-F_\theta^{(2)}(u))\log(1-F_\theta^{(2)}(u))\},\end{aligned}$$

increases as θ approaches q_1 from both sides, where

$$F_\theta^{(1)}(u) = \frac{\min(q_1,\theta)F_1(u) + \max(\theta-q_1,0)F_2(u)}{\min(q_1,\theta) + \max(\theta-q_1,0)} \quad \text{and}$$

$$F_\theta^{(2)}(u) = \frac{\max(q_1-\theta,0)F_1(u) + \min(1-\theta,1-q_1)F_2(u)}{\max(q_1-\theta,0) + \min(1-\theta,1-q_1)},$$

are the limits of $\widehat{F}_1^{\tau_1'}(u)$ and $\widehat{F}_{\tau_1'}^{n+1}(u)$, respectively. This implies that the function $\int_{-\infty}^{\infty} Q_u(\theta)\, dw(u)$ attains its local maximum at the true location of the change-point, q_1.

REMARK 1. The log-likelihood function (2.1) is essentially related to the two-sample goodness-of-fit (GOF) test statistic based on the nonparametric likelihood ratio [Einmahl and McKeague (2003); Zhang (2006)]. To see this, let $Z_1, \ldots, Z_n$ be independent, and suppose that $Z_1, \ldots, Z_{n_1}$ have a common continuous distribution function F_1, and $Z_{n_1+1}, \ldots, Z_n$ have F_2. We are interested in testing the null hypothesis H_0 that $F_1(u) = F_2(u)$ for all $u \in (-\infty, \infty)$ against H_1 that $F_1(u) \neq F_2(u)$ for some $u \in (-\infty, \infty)$. For each fixed $u \in (-\infty, \infty)$, a natural approach is to apply the likelihood ratio test,

$$G_u = n_1\left\{\widehat{F}_1^{n_1+1}(u)\log\left(\frac{\widehat{F}_1^{n_1+1}(u)}{\widehat{F}_n(u)}\right) + \bigl(1-\widehat{F}_1^{n_1+1}(u)\bigr)\log\left(\frac{1-\widehat{F}_1^{n_1+1}(u)}{1-\widehat{F}_n(u)}\right)\right\}$$
$$+ n_2\left\{\widehat{F}_{n_1+1}^{n+1}(u)\log\left(\frac{\widehat{F}_{n_1+1}^{n+1}(u)}{\widehat{F}_n(u)}\right) + \bigl(1-\widehat{F}_{n_1+1}^{n+1}(u)\bigr)\log\left(\frac{1-\widehat{F}_{n_1+1}^{n+1}(u)}{1-\widehat{F}_n(u)}\right)\right\},$$

where $\widehat{F}_n(u)$ corresponds to the C.D.F. of the pooled sample. By noting that $n_1\widehat{F}_1^{n_1+1}(u) + n_2\widehat{F}_{n_1+1}^{n+1}(u) = n\widehat{F}_n(u)$, G_u would be of the same form as (2.1) with $L = 1$ up to a constant which does not depend on the segmentation point n_1. Einmahl and McKeague (2003) considered using G_u to test whether there is at most one change-point.

In the two-sample GOF test, Zhang (2002, 2006) demonstrated that by choosing appropriate weight functions $w(u)$ we can produce new omnibus tests that are generally much more powerful than the conventional ones such as Kolmogorov–Smirnov, Cramér–von Mises and Anderson–Darling test statistics. If we take $dw(u) = \{\widehat{F}_n(u)(1-\widehat{F}_n(u))\}^{-1}\,d\widehat{F}_n(u)$, and also note that $\mathcal{L}_u$ is zero for $u \in (-\infty, X_{(1)})$ and $u \in (X_{(n)}, \infty)$ where $X_{(1)} < \cdots < X_{(n)}$ represent the order statistics, the objective function in (2.2) can be rewritten as

$$\begin{aligned} &R_n(\tau_1', \ldots, \tau_L') \\ &\quad= \int_{X_{(1)}}^{X_{(n)}} \mathcal{L}_u(\tau_1', \ldots, \tau_L')\{\widehat{F}_n(u)(1-\widehat{F}_n(u))\}^{-1}\,d\widehat{F}_n(u) \\ &\quad= n\sum_{k=0}^{L}\sum_{l=2}^{n-1}(\tau_{k+1}' - \tau_k')\frac{\widehat{F}_{kl}\log\widehat{F}_{kl} + (1-\widehat{F}_{kl})\log(1-\widehat{F}_{kl})}{l(n-l)}, \end{aligned} \tag{2.3}$$

where $\widehat{F}_{kl} = \widehat{F}_{\tau_k'}^{\tau_{k+1}'}(X_{(l)})$. As recommended by Zhang (2002), we take a common "continuity correction" by replacing $\widehat{F}_{kl}$ with $\widehat{F}_{kl} - 1/\{2(\tau_{k+1}' - \tau_k')\}$ for all k and l.

To determine L in the MCP, we observe that $Q_u(\theta)$ is a convex function with respect to θ, and thus

$$\max_{\tau_1'<\cdots<\tau_L'} R_n(\tau_1', \ldots, \tau_L') \le \max_{\tau_1'<\cdots<\tau_{L+1}'} R_n(\tau_1', \ldots, \tau_{L+1}'),$$

which means that the maximum log-likelihood $\max_{\tau_1'<\cdots<\tau_L'} R_n(\tau_1', \ldots, \tau_L')$ is a nondecreasing function in L. Hence, we can use Schwarz's Bayesian information criterion (BIC) to strike a balance between the likelihood and the number of change-points by incorporating a penalty for large L. More specifically, we identify the value of L by minimizing

$$\text{BIC}_L = -\max_{\tau_1'<\cdots<\tau_L'} R_n(\tau_1', \ldots, \tau_L') + L\zeta_n \tag{2.4}$$

and ζ_n is a proper sequence going to infinity. Yao (1988) used the BIC with $\zeta_n = \log n$ to select the number of change-points and showed its consistency in the least-squares framework. However, the traditional BIC tends to select a model with some spurious change-points. Detailed discussions on the choice of ζ_n and other tuning parameters are given in Section 3.2.

2.2. *Asymptotic theory.* In the context of change-point estimation, it is well known that the points around the true change-point cannot be distinguished asymptotically with a fixed change magnitude. In the least-squares fitting, the total variation with perfect segmentation is asymptotically equivalent to that with an estimate of the change-point in a neighborhood of the true change-point [Yao and Au (1989)]. For example, suppose that there is only one change-point τ with a change size δ, then we can only achieve $\delta^2|\hat{\tau}_{\text{MLE}} - \tau| = O_p(1)$ as $n \to \infty$, where $\hat{\tau}_{\text{MLE}}$ denotes the maximum likelihood estimator (MLE) of τ [see Chapter 1 of Csörgő and Horváth (1997)]. For single change-point nonparametric models, Darkhovskh (1976) obtained a rate of $o_p(n)$, Carlstein (1988) derived a rate of $O(n^\alpha)$ a.s. (almost surely) for any $\alpha > 1/2$, and Dümbgen (1991) achieved a rate of $O_p(1)$. The estimator in Lee (1996) is shown to be consistent a.s. and the differences between the estimated and true locations of change-points are of order $O(\log n)$ a.s.

Let $\mathcal{G}_n(L) = \{\hat{\tau}_1, \ldots, \hat{\tau}_L\}$ denote the set of estimates of the change-points using the proposed NMCD. The next theorem establishes the desirable property for the NMCD estimator when K_n is prespecified—$\mathcal{G}_n(K_n)$ is asymptotically close to the true change-point set. Let $C_{K_n}(\delta_n)$ contain all the sets in the δ_n-neighborhood of the true locations,

$$C_{K_n}(\delta_n) = \{(\tau_1', \ldots, \tau_{K_n}') : 1 < \tau_1' < \cdots < \tau_{K_n}' \le n, |\tau_s' - \tau_s| \le \delta_n \text{ for } 1 \le s \le K_n\},$$

where δ_n is some positive sequence. Denote $F_{k,\theta} = \theta F_k + (1-\theta)F_{k+1}$ for $0 < \theta < 1$. For $r = 1, \ldots, K_n$, define

$$\eta(u; F_r, F_{r,\theta}) = F_r(u)\log\left(\frac{F_r(u)}{F_{r,\theta}(u)}\right) + (1 - F_r(u))\log\left(\frac{1 - F_r(u)}{1 - F_{r,\theta}(u)}\right),$$

which is the Kullback–Leibler distance between two Bernoulli distributions with respective success probabilities $F_r(u)$ and $F_{r,\theta}(u)$. Hence, whenever $F_r(u) \neq$

$F_{r+1}(u)$, and accordingly $F_r(u) \neq F_{r,\theta}(u)$, $\eta(u; F_r, F_{r,\theta})$ is strictly larger than zero. Furthermore, for $r = 1, \ldots, K_n$, define

$$\eta_r(u) = \eta(u; F_r, F_{r,1/2}) + \eta(u; F_{r+1}, F_{r,1/2}).$$

To establish the consistency of the proposed NMCD, the following assumptions are imposed:

(A1) $F_1, \ldots, F_{K_n+1}$ are continuous and $F_k \neq F_{k+1}$ for $k = 1, \ldots, K_n$.

(A2) Let $\lambda_n = \min_{1 \le k \le K_n+1}(\tau_k - \tau_{k-1})$; $\lambda_n \to \infty$ as $n \to \infty$.

(A3) $\widehat{F}_n(u) \xrightarrow{\text{a.s.}} F(u)$ uniformly in u, where $F(u)$ is the C.D.F. of the pooled sample.

(A4) $\eta_{\min} \equiv \min_{1 \le r \le K_n} \int_0^1 \eta_r(u)/\{F(u)(1 - F(u))\}\, dF(u)$ is a positive constant.

Assumption (A1) is required in some exponential tail inequalities as detailed in the proof of Lemma 2, while the F_k's can be discrete or mixed distributions in practice. Assumption (A2) is a standard requirement for the theoretical development in the MCP, which allows the change-points to be asymptotically distinguishable. Assumption (A3) is a technical condition that is trivially satisfied by the Glivenko–Cantelli theorem when K_n is finite. Generally, it can be replaced by the conditions that $\lim_{n\to\infty} \sum_{k=1}^{K_n+1}(\tau_k - \tau_{k-1})/n F_k(u)$ exists and $\sum_{k=1}^{K_n+1}\{(\tau_k - \tau_{k-1})/n \sup_u |\widehat{F}^{\tau_k}_{\tau_{k-1}}(u) - F_k(u)|\}$ converges to 0 a.s. By the Dvoretzky–Kiefer–Wolfowitz inequality, the latter one holds if $\sum_{n=1}^{\infty} K_n \exp(-2\lambda_n \epsilon^2/K_n^2) < \infty$ for any $\epsilon > 0$. Assumption (A4) means that the smallest signal strength among all the changes is bounded away from zero.

We may consider relaxing $\lambda_n \to \infty$ in assumption (A2) by allowing $\eta_{\min} \to \infty$ as $n \to \infty$. It is intuitive that if two successive distributions are very different, then we do not need a very large λ_n to locate the change point. For the mean change problem, Niu and Zhang (2012) and Hao, Niu and Zhang (2013) revealed that in order to obtain the $O_p(1)$ consistency, a condition $\delta\lambda_n > 32 \log n$ is required, where δ is the minimal jump size at the change-points (similar to $\eta_{\min}$). In our nonparametric setting, such an extension warrants future investigation.

THEOREM 1. *Under assumptions* (A1)–(A4), *if* $K_n^3 (\log K_n)^2 (\log \delta_n)^2/\delta_n \to 0$ *and* $\delta_n/\lambda_n \to 0$, *then*

$$\Pr\{\mathcal{G}_n(K_n) \in C_{K_n}(\delta_n)\} \to 1 \qquad \text{as } n \to \infty.$$

Under the classical mean change-point model, Yao and Au (1989) studied the property of the least-squares estimator,

$$\arg\min_{\tau'_1 < \cdots < \tau'_{K_n}} \sum_{k=1}^{K_n+1} \sum_{i=\tau'_{k-1}}^{\tau'_k - 1} \{X_i - \hat{\mu}(\tau'_{k-1}, \tau'_k)\}^2, \tag{2.5}$$

where $\hat{\mu}(\tau'_{k-1}, \tau'_k)$ denotes the average of the observations $\{X_{\tau'_{k-1}}, \ldots, X_{\tau'_k-1}\}$. It is well known that the least-squares estimator is consistent with the optimal rate $O_p(1)$, when the number of change-points is known (and does not depend on n) and the change magnitudes are fixed; see Hao, Niu and Zhang (2013) and the references therein. Under a similar setting with $K_n \equiv K$, we can establish the same rate of $O_p(1)$ for our nonparametric approach.

COROLLARY 1. *Under assumptions* (A1), (A2) *and* (A4), $|\hat{\tau}_s - \tau_s| = O_p(1)$ *for* $s = 1, \ldots, K$.

The proof is similar to that of Theorem 1, which is provided in the supplementary material [Zou et al. (2014)]. With the knowledge of K, we can obtain an optimal rate of $O_p(1)$ without specifying the distributions, which is consistent with the single change-point case in Dümbgen (1991).

The next theorem establishes the consistency of the NMCD procedure with the BIC in (2.4). Let $\widehat{K}_n = \arg\min_{1 \le L \le \overline{K}_n} \mathrm{BIC}_L$, where $\overline{K}_n$ is an upper bound on the true number of change-points.

THEOREM 2. *Under assumptions* (A1)–(A4), $\lambda_n/(\overline{K}_n \zeta_n) \to \infty$, $\zeta_n = \overline{K}_n^3 (\log \overline{K}_n)^2 (\log n)^{2+c}$ *with any* $c > 0$, *then* $\Pr(\widehat{K}_n = K_n) \to 1$ *as* $n \to \infty$.

It is remarkable that in the conventional setting where $\overline{K}_n$ is bounded, we can use ζ_n of order $(\log n)^{2+c}$ instead of its least-squares counterpart $\log n$ in Yao (1988). In conjunction with Theorem 1, this result implies that $\Pr\{\mathcal{G}_n(\widehat{K}) \in C_K((\log n)^{2+c})\} \to 1$ with a fixed number of change-points.

3. Implementation of NMCD.

3.1. *Algorithm.* One important property of the proposed maximum likelihood approach is that (2.3) is separable. The optimum for splitting cases $1, \ldots, n$ into L segments conceptually consists of first finding the rightmost change-point $\hat{\tau}_L$, and then finding the remaining change-points from the fact that they constitute the optimum for splitting cases $1, \ldots, \hat{\tau}_L$ into $L - 1$ segments. This separability is called Bellman's "principle of optimality" [Bellman and Dreyfus (1962)]. Thus, (2.3) can be maximized via the DP algorithm and fitting such a nonparametric MCP model is straightforward and fast. The total computational complexity is $O(Ln^2)$ for a given L; see Hawkins (2001) and Bai and Perron (2003) for the pseudo-codes of the DP. Hawkins (2001) suggested using the DP on a grid of $m \ll n$ values. Harchaoui and Lévy-Leduc (2010) proposed using a LASSO-type penalized estimator to achieve a reduced version of the least-squares method. Niu and Zhang (2012) developed a screening and ranking algorithm to detect DNA copy number variations in the MCP framework.

Due to the DP's computational complexity in n^2, an optimal segmentation of a very long sequence could be computationally intensive; for example, DNA sequences nowadays are often extremely long [Fearnhead and Vasileiou (2009)]. To alleviate the computational burden, we introduce a preliminary screening step which can exclude most of the irrelevant points and, as a consequence, the NMCD is implemented in a much lower-dimensional space.

Screening algorithm.

(i) Choose an appropriate integer n_I which is the length of each subsequence of the data, and take the estimated change-point set $\mathcal{O} = \varnothing$.

(ii) Initialize $\gamma_i = 0$ for $i = 1, \ldots, n$; and for $i = n_I, \ldots, n - n_I$, update γ_i to be the Cramér–von Mises two-sample test statistic for the samples $\{X_{i-n_I+1}, \ldots, X_i\}$ and $\{X_{i+1}, \ldots, X_{i+n_I}\}$.

(iii) For $i = n_I, \ldots, n - n_I$, define $k = \arg\max_{i-n_I<j\le i+n_I} \gamma_j$. If $k = i$, update $\mathcal{O} = \mathcal{O} \cup \{i\}$.

Intuitively speaking, this screening step finds the most influential points that have the largest *local* jump sizes quantified by the Cramér–von Mises statistic, and thus helps to avoid including too many candidate points around the true change-point. As a result, we can obtain a candidate change-point set, $\mathcal{O}$, of which the cardinality, $|\mathcal{O}|$, is usually much smaller than n. Finally, we run the NMCD procedure within the set $\mathcal{O}$ using the DP algorithm to find the solution of

$$\arg\max_{\tau_1'<\cdots<\tau_L'\in\mathcal{O}} R_n(\tau_1', \ldots, \tau_L').$$

Apparently, the screening procedure is fast because it mainly requires calculating $n - 2n_I + 1$ Cramér–von Mises statistics. In contrast, Lee (1996) used a thresholding step to determine the number of change-points. The main difference between Lee (1996) and Niu and Zhang (2012) lies in the choice of the local test statistic; the former uses some seminorm of empirical distribution functions and the latter is based on the two-sample mean difference.

We next clarify how to choose n_I, which formally establishes the consistency of the screening procedure.

PROPOSITION 1. *Under assumptions* (A1)–(A2), *if* $n_I/\log n \to \infty$ *and* $n_I/\lambda_n^{1/2} \to 0$, *then we have* $\Pr\{\mathcal{O} \in H_{|\mathcal{O}|}(\log n)\} \to 1$, *where*

$$H_l(\delta_n) = \big\{(\tau_1', \ldots, \tau_l') : 1 < \tau_1' < \cdots < \tau_l' \le n, \text{ and for each } 1 \le r \le K_n \text{ there exists at least a } \tau_s' \text{ so that } |\tau_s' - \tau_r| \le \delta_n\big\}.$$

This result follows by verifying condition (A3) in Lee (1996); see Example II of Dümbgen (1991). With probability tending to one, the screening algorithm can at least include one δ_n-neighborhood of the true location set by choosing

an appropriate n_I. Given a candidate L, the computation of NMCD reduces to $O(L|\mathcal{O}|n)$, which is of order $O(\overline{K}_n^2|\mathcal{O}|n)$ in conjunction with the BIC. Both the R and FORTRAN codes for implementing the entire procedure are available from the authors upon request.

3.2. *Selection of tuning parameters.* We propose to take $dw(u) = \{\widehat{F}_n(u)(1 - \widehat{F}_n(u))\}^{-1} d\widehat{F}_n(u)$, which is found to be more powerful than simply using $dw(u) = d\widehat{F}_n(u)$. The function $\{\widehat{F}_n(u)(1 - \widehat{F}_n(u))\}^{-1}$ attains its minimum at $\widehat{F}_n(u) = 1/2$, that is when u is the median of the sample. Intuitively, when two successive distributions mainly differ in their centers, both choices of $dw(u)$ would be powerful because a large portion of observations are around the center. However, if the difference between two adjacent distributions lies in their tails, using $dw(u) = d\widehat{F}_n(u)$ may not work well because only very limited information is included in the integral of (2.2). In contrast, our weight would be larger for those more extreme observations (far way from the median).

To better understand this, we analyze the term $\eta_{\min}$, which reflects the detection ability to a large extent. Consider a special case

$$X_i \sim \begin{cases} U(0,1), & 1 \le i \le n/2, \\ U(1,2), & n/2+1 \le i \le n \end{cases}$$

and thus

$$\eta_1(u) = \left(u \log 2 + (1-u) \log \frac{1-u}{1-u/2}\right) I(0 < u < 1) + \left((u-1) \log \frac{2(u-1)}{u} + (2-u) \log \frac{2-u}{1-u/2}\right) I(1 < u < 2).$$

It is easy to check that $\eta_{\min} = \int_0^2 \eta_1(u)/\{F(u)(1 - F(u))\}\, dF(u)$ is unbounded, while the counterpart $\int_0^2 \eta_1(u)\, dF(u)$ is finite. Consequently, the NMCD procedure would be more powerful by using the weight $\{\widehat{F}_n(u)(1 - \widehat{F}_n(u))\}^{-1} d\widehat{F}_n(u)$.

Under the assumption that $\zeta_n = \overline{K}_n^3 (\log \overline{K}_n)^2 (\log n)^{2+c}$ with $c > 0$ and $\lambda_n/(\overline{K}_n \zeta_n) \to \infty$, we establish the consistency of the BIC in (2.4) for model selection. The choice of ζ_n depends on $\overline{K}_n$ and λ_n which are unknown. The value of $\overline{K}_n$ depends on the practical consideration of how many change-points are to be identified, while λ_n reflects the length of the smallest segment. For practical use, we take $\overline{K}_n$ to be fixed and recommend $\zeta_n = (\log n)^{2+c}/2$ with $c = 0.1$. A small value of c helps to prevent underfitting, as one is often reluctant to miss any important change-point. The performance of NMCD insensitive to the choice of $\overline{K}_n$, as long as $\overline{K}_n$ is not too small, which is also to avoid underfitting. We suggest $\overline{K}_n = |\mathcal{O}|$, that is, the cardinality of the candidate change-point set in the screening algorithm.

4. Simulation studies.

4.1. *Model setups*. To evaluate the finite-sample performance of the proposed NMCD procedure, we conduct extensive simulation studies, and also make comparisons with existing methods. We calculate the distance between the estimated set $\widehat{\mathcal{G}}_n$ and the true change-point set $\mathcal{C}_t$ [Boysen et al. (2009)],

$$\xi(\widehat{\mathcal{G}}_n \| \mathcal{C}_t) = \sup_{b \in \mathcal{C}_t} \inf_{a \in \widehat{\mathcal{G}}_n} |a - b| \quad \text{and} \quad \xi(\mathcal{C}_t \| \widehat{\mathcal{G}}_n) = \sup_{b \in \widehat{\mathcal{G}}_n} \inf_{a \in \mathcal{C}_t} |a - b|,$$

which quantify the over-segmentation error and the under-segmentation error, respectively. A desirable estimator should be able to balance both quantities. In addition, we consider the average Rand index [Fowlkes and Mallows (1983)], which measures the discrepancy of two sets from an average viewpoint.

Following model (I) introduced by Donoho and Johnstone (1995), we generate the Blocks datasets, which contains $K_n = 11$ change-points:

$$\begin{aligned}
\text{Model (I):}\ X_i &= \sum_{j=1}^{K_n} h_j J(nt_i - \tau_j) + \sigma \varepsilon_i, \qquad J(x) = \{1 + \operatorname{sgn}(x)\}/2, \\
\{\tau_j/n\} &= \{0.1, 0.13, 0.15, 0.23, 0.25, 0.40, 0.44, 0.65, 0.76, 0.78, 0.81\}, \\
\{h_j\} &= \{2.01, -2.51, 1.51, -2.01, 2.51, -2.11, 1.05, 2.16, \\
&\qquad\qquad -1.56, 2.56, -2.11\},
\end{aligned}$$

where there are n equally spaced covariates t_i in $[0, 1]$. Three error distributions for ε_i are considered: $N(0, 1)$, Student's t distribution with three degrees of freedom $t_{(3)}$, and the standardized (zero mean and unit variance) chi-squared distribution with one degree of freedom $\chi^2_{(1)}$. The Blocks datasets with $n = 1000$, as depicted in the top three plots of Figure A.1 in the supplementary material [Zou et al. (2014)], are generally considered difficult for multiple change-point estimation due to highly heterogeneous segment levels and lengths.

In a more complicated setting with both location and scale changes, we consider model (II) with $K_n = 4$:

$$\begin{aligned}
\text{Model (II):}\ X_i &= \sum_{j=1}^{K_n} h_j J(nt_i - \tau_j) + \sigma \varepsilon_i \prod_{j=1}^{\sum_{j=1}^{K_n} J(nt_i - \tau_j)} v_j, \\
\{h_j\} &= \{3, 0, -2, 0\}, \qquad \{\tau_j/n\} = \{0.20, 0.40, 0.65, 0.85\} \quad \text{and} \\
\{v_j\} &= \{1, 5, 1, 0.25\},
\end{aligned}$$

where all the other setups are the same as those of model (I). As shown by the bottom three plots in Figure A.1, there are two location changes and two scale changes.

In addition, we include a simulation study when the distributions differ in the skewness and kurtosis. In particular, we consider

$$\text{Model (III): } X_i \sim F_j(x), \qquad \tau_j/n = \{0.20, 0.50, 0.75\}, \qquad j = 1, 2, 3, 4,$$

where $F_1(x), \ldots, F_4(x)$ correspond to the standard normal, the standardized $\chi^2_{(3)}$ (with zero mean and unit variance), the standardized $\chi^2_{(1)}$, and the standard normal distribution, respectively. Because there is no mean or variance difference between the F_j's, as depicted in the left panel of Figure A.4, the estimation for such a change-point problem is rather difficult. All the simulation results are obtained with 1000 replications.

4.2. *Calibration of tuning parameters.* To study the sensitivity of the choice of ζ_n, Figure 1(a) shows the curves of $|\widehat{K}_n - K_n|$ versus the value of β with $\zeta_n = \beta(\log n)^{2.1}/2$ under model (I). Clearly, the estimation is reasonably well with a value of β around 1. For more adaptive model selection, a data-adaptive complexity penalty in Shen and Ye (2002) could be considered.

In the screening procedure, the choice of n_I needs to balance the computation and underfitting. By Proposition 1, $n_I \in (\log n, \lambda_n^{1/2})$, while λ_n is typically unknown. In practice, we recommend to choose $n_I = \lceil (\log n)^{3/2}/2 \rceil$, which is the smallest integer that is larger than $(\log n)^{3/2}/2$. Figure 1(b) shows the curves of under-segmentation errors versus the value of β with $n_I = \lceil \beta(\log n)^{3/2}/2 \rceil$ under model (I). In a neighborhood of $\beta = 1$, our method provides a reasonably effective reduction of the subset $\mathcal{O}$ and the performance is relatively stable. In general, we

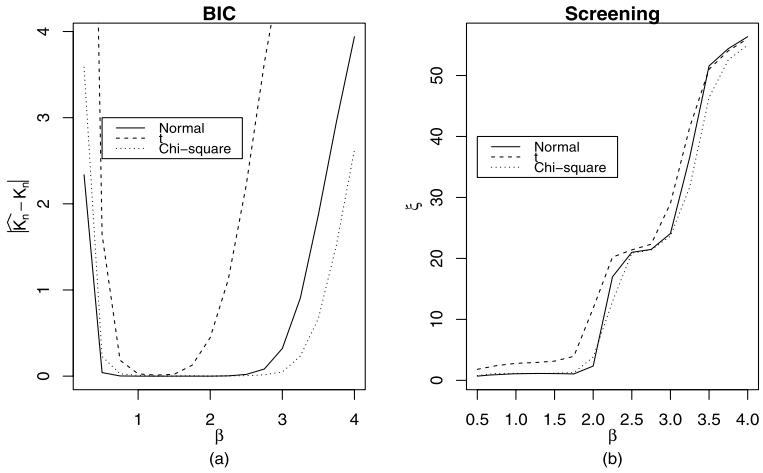

FIG. 1. *The performance of NMCD under model* (I) *with* $n = 1000$ *and* $\sigma = 0.5$ *when the tuning parameters vary*: (a) *the curves of* $|\widehat{K}_n - K_n|$ *versus the value of* β; (b) *the curves of* $\xi(\widehat{\mathcal{G}}_n \| \mathcal{C}_t)$ *versus the value of* β.

do not recommend a too large value of n_I so as to avoid underfitting. From the results shown in Section 4.6, the choice of $\zeta_n = (\log n)^{2.1}/2$ and $n_I = \lceil (\log n)^{3/2}/2 \rceil$ works also well when the number of change-points increases as the sample size increases.

4.3. *Comparison between NMCD and PL.* Firstly, under model (I) with location changes only, we make a comparison of NMCD with the parametric likelihood (PL) method which coincides with the classical least-squares method in (2.5) under the normality assumption [Yao (1988)]. We also consider a variant of NMCD by using $dw(u) = d\widehat{F}_n(u)$ (abbreviated as NMCD*). The comparison is conducted with and without knowing the true number of change-points K_n, respectively. Table 1 presents the average values of $\xi(\widehat{\mathcal{G}}_n \| \mathcal{C}_t)$ and $\xi(\mathcal{C}_t \| \widehat{\mathcal{G}}_n)$ for $n = 500$ and 1000 and $\sigma = 0.5$ when K_n is known to be 11. To gain more insight, we also present the standard deviations of the two distances in parentheses. Simulation results with other values of σ can be found in the supplementary material [Zou et al. (2014)].

As expected, the PL has superior efficiency for the case with normal errors, since the parametric model is correctly specified. The NMCD procedure also offers satisfactory performance and the differences in the two ξ values between NMCD and PL are extremely small, while both methods significantly outperform the NMCD* procedure. For the cases with $t_{(3)}$ and $\chi^2_{(1)}$ errors, the NMCD procedure almost uniformly outperforms the PL in terms of estimation accuracy of the locations. Not only are the distance values of $\xi(\widehat{\mathcal{G}}_n \| \mathcal{C}_t)$ and $\xi(\mathcal{C}_t \| \widehat{\mathcal{G}}_n)$ smaller, but the corresponding standard deviations are also much smaller using the NMCD.

TABLE 1

Comparison of the parametric likelihood (PL), NMCD, and NMCD methods when the number of change-points K_n is specified (known) under models* (I) *and* (II), *respectively. The standard deviations are given in parentheses*

			$\xi(\widehat{\mathcal{G}}_n \| \mathcal{C}_t)$			$\xi(\mathcal{C}_t \| \widehat{\mathcal{G}}_n)$		
Model	**Error**	n	**PL**	**NMCD**	**NMCD***	**PL**	**NMCD**	**NMCD***
(I)	$N(0,1)$	500	0.96 (1.19)	0.96 (1.14)	1.16 (1.15)	0.96 (1.19)	0.96 (1.14)	1.16 (1.15)
		1000	0.91 (1.15)	0.97 (1.16)	1.06 (1.21)	0.91 (1.15)	0.97 (1.16)	1.06 (1.21)
	$t_{(3)}$	500	13.6 (12.0)	3.77 (4.48)	3.86 (4.33)	14.3 (18.4)	3.95 (7.51)	3.97 (7.63)
		1000	20.2 (21.3)	2.58 (2.50)	2.90 (2.72)	21.9 (34.5)	2.56 (2.40)	2.90 (2.72)
	$\chi^2_{(1)}$	500	1.39 (2.91)	0.70 (0.80)	0.80 (1.22)	1.13 (1.57)	0.70 (0.80)	0.81 (1.41)
		1000	1.05 (2.15)	0.59 (0.77)	0.58 (0.71)	0.99 (1.38)	0.59 (0.77)	0.58 (0.71)
(II)	$N(0,1)$	500	1.59 (1.72)	2.35 (2.42)	3.34 (4.96)	1.59 (1.72)	2.35 (2.42)	3.34 (4.96)
		1000	1.58 (1.52)	2.68 (2.59)	2.74 (2.89)	1.58 (1.52)	2.68 (2.59)	2.74 (2.89)
	$t_{(3)}$	500	13.6 (25.8)	4.75 (6.87)	6.42 (8.84)	7.52 (10.2)	4.54 (5.19)	6.05 (6.42)
		1000	16.4 (40.2)	4.10 (3.88)	5.27 (7.20)	10.3 (18.0)	4.10 (3.88)	5.24 (6.85)
	$\chi^2_{(1)}$	500	6.36 (11.3)	1.57 (2.12)	1.65 (2.90)	5.88 (8.93)	1.57 (2.12)	1.65 (2.90)
		1000	4.80 (67.8)	1.17 (1.45)	1.49 (2.10)	4.80 (7.82)	1.17 (1.45)	1.49 (2.10)

TABLE 2

Comparison of the PL and NMCD methods when the number of change-points K_n is unknown (K_n is selected using the BIC) under models (I) *and* (II), *respectively. The standard deviations are given in parentheses*

			Parametric likelihood (PL)			**NMCD**		
Model	**Error**	n	$\xi(\widehat{\mathcal{G}}_n \| \mathcal{C}_t)$	$\xi(\mathcal{C}_t \| \widehat{\mathcal{G}}_n)$	$\lvert\widehat{K}_n - K_n\rvert$	$\xi(\widehat{\mathcal{G}}_n \| \mathcal{C}_t)$	$\xi(\mathcal{C}_t \| \widehat{\mathcal{G}}_n)$	$\lvert\widehat{K}_n - K_n\rvert$
(I)	$N(0,1)$	500	0.93 (1.08)	2.16 (6.57)	0.09 (0.31)	0.96 (1.34)	0.99 (1.05)	0.00 (0.04)
		1000	0.94 (1.14)	2.30 (10.3)	0.05 (0.25)	0.96 (1.25)	1.01 (1.25)	0.00 (0.04)
	$t_{(3)}$	500	2.91 (2.92)	39.0 (24.9)	6.05 (3.47)	3.34 (4.22)	8.64 (15.2)	0.36 (0.88)
		1000	2.94 (3.02)	95.2 (48.8)	9.70 (4.14)	2.54 (2.78)	10.0 (26.8)	0.36 (0.75)
	$\chi^2_{(1)}$	500	0.85 (0.99)	49.5 (23.6)	10.9 (4.69)	0.73 (0.95)	1.36 (5.59)	0.05 (0.28)
		1000	0.85 (1.05)	111 (46.2)	14.2 (4.06)	0.53 (0.69)	0.89 (4.28)	0.02 (0.20)
(II)	$N(0,1)$	500	1.66 (1.61)	2.22 (5.56)	0.04 (0.22)	2.28 (2.31)	4.45 (8.54)	0.13 (0.37)
		1000	1.69 (1.50)	1.71 (1.52)	0.01 (0.11)	2.19 (2.11)	3.93 (10.6)	0.06 (0.27)
	$t_{(3)}$	500	5.77 (6.57)	24.1 (20.0)	1.58 (1.56)	5.18 (6.18)	14.1 (16.5)	0.75 (1.01)
		1000	5.59 (6.26)	62.4 (41.3)	2.72 (2.21)	4.50 (4.44)	17.0 (28.4)	0.47 (0.87)
	$\chi^2_{(1)}$	500	5.03 (6.19)	43.1 (16.0)	4.71 (2.66)	1.67 (2.39)	7.27 (12.6)	0.43 (0.80)
		1000	5.00 (6.29)	91.1 (31.1)	6.22 (3.23)	1.26 (1.50)	9.45 (22.7)	0.28 (0.70)

Next, we consider the K_n unknown case, for which both the NMCD and PL procedures are implemented by setting $\overline{K}_n = 30$ and using the BIC to choose the number of change-points. The average values of the distances $\xi(\widehat{\mathcal{G}}_n \| \mathcal{C}_t)$ and $\xi(\mathcal{C}_t \| \widehat{\mathcal{G}}_n)$ are tabulated in Table 2. In addition, we also present the average values of $|\widehat{K}_n - K_n|$ with standard deviations in parentheses, which reflect the overall estimation accuracy of K_n. Clearly, the two methods have comparable performances under the normal error, while the proposed NMCD significantly outperforms PL in terms of $\xi(\mathcal{C}_t \| \widehat{\mathcal{G}}_n)$ and $|\widehat{K}_n - K_n|$ for the two nonnormal cases, because the efficiency of the BIC used in PL relies heavily on the parametric assumption. When we compare the results across Tables 1 and 2, the standard deviations for the distance measures increase from the K_n known to the K_n unknown cases, as estimating K_n further enlarges the variability.

We turn to the comparison between NMCD and PL under model (II) in which both location and scale changes are exhibited. In this situation, the standard least-squares method (2.5) does not work well because it is constructed for location changes only. To further allow for scale changes under the PL method, we consider

$$\arg\min_{\tau'_1 < \cdots < \tau_{K'_n}} \sum_{k=1}^{L} (\tau'_{k+1} - \tau'_k) \log \hat{\sigma}_k^2, \tag{4.1}$$

where $\hat{\sigma}_k^2 = (\tau'_{k+1} - \tau'_k)^{-1} \sum_{i=\tau'_{k-1}}^{\tau'_k - 1} \{X_i - \hat{\mu}(\tau'_{k-1}, \tau'_k)\}^2$, and the BIC is modified accordingly. The bottom panels of Tables 1 and 2 tabulate the values of $\xi(\widehat{\mathcal{G}}_n \| \mathcal{C}_t)$

and $\xi(\mathcal{C}_t \| \widehat{\mathcal{G}}_n)$ when K_n is specified in advance and estimated by using the BIC, respectively. Clearly, the NMCD method delivers a satisfactory detection performance for the normal case and performs much better than the PL method for the two nonnormal cases. Therefore, the conclusion remains that the PL method is generally sensitive to model specification, while the NMCD does not depend on any parametric modeling assumption and thus is much more robust.

4.4. *Comparisons of NMCD with other nonparametric methods*. We consider the methods of Lee (1996) and Matteson and James (2014), as they also do not make any assumptions regarding the nature of the changes. The NMCD is implemented with the initial nonparametric screening procedure, and K_n is selected by the BIC. In both our screening procedure and Lee's (1996) method, the window is set as $n_I = \lceil (\log n)^{3/2}/2 \rceil$, and the threshold value of the latter is chosen as $(\log n)^{3/4}$. The ECP method of Matteson and James (2014) is implemented using the "ecp" R package with the false alarm rate 0.05 and $\alpha = 1$.

Table 3 shows the comparison results based on $\xi(\widehat{\mathcal{G}}_n; \mathcal{C}_t) \equiv \xi(\widehat{\mathcal{G}}_n \| \mathcal{C}_t) + \xi(\mathcal{C}_t \| \widehat{\mathcal{G}}_n)$, $|\widehat{K}_n - K_n|$, and the Rand index under models (I)–(III) with $\sigma = 0.5$, respectively. Lee's (1996) method is unable to produce a reasonable estimate for K_n and the resulting models are much overfitted in all the cases, which indicates that its "local" nature incurs substantial loss of the information. Under model (I), the NMCD performs better than ECP for normal and $\chi^2_{(1)}$ errors, while the opposite is true for the t error distribution. Under model (II), the ECP also exhibits

TABLE 3
Comparison of NMCD, Lee's (1996) method and Matteson and James's (2014) ECP in terms of $\xi(\widehat{\mathcal{G}}_n; \mathcal{C}_t)$, *Rand and* $|\widehat{K}_n - K_n|$ *under models* (I)–(III) *with* $\sigma = 0.5$

			$\xi(\widehat{\mathcal{G}}_n; \mathcal{C}_t)$			Rand			$\lvert\widehat{K}_n - K_n\rvert$		
Model	**Error**	n	**Lee**	**ECP**	**NMCD**	**Lee**	**ECP**	**NMCD**	**Lee**	**ECP**	**NMCD**
(I)	$N(0,1)$	500	84.9	6.03	2.62	0.920	0.994	0.992	28.5	0.07	0.01
		1000	176	7.42	2.23	0.915	0.997	0.994	43.2	0.07	0.00
	$t_{(3)}$	500	86.7	4.95	8.94	0.920	0.995	0.988	27.5	0.06	0.22
		1000	177	7.29	7.63	0.914	0.997	0.993	42.9	0.08	0.02
	$\chi^2_{(1)}$	500	85.0	4.67	3.00	0.921	0.995	0.992	28.3	0.06	0.02
		1000	176	5.67	2.80	0.915	0.997	0.994	43.1	0.05	0.01
(II)	$N(0,1)$	500	69.0	17.6	14.4	0.832	0.980	0.980	33.8	0.06	0.11
		1000	140	17.5	14.4	0.830	0.990	0.987	51.3	0.07	0.03
	$t_{(3)}$	500	69.3	16.8	20.4	0.833	0.982	0.974	33.7	0.10	0.25
		1000	141	12.5	21.4	0.830	0.992	0.983	51.6	0.06	0.13
	$\chi^2_{(1)}$	500	67.9	8.25	10.5	0.833	0.989	0.983	34.0	0.05	0.12
		1000	139	10.2	12.6	0.830	0.994	0.987	51.2	0.07	0.09
(III)		500	120	394	78.2	0.822	0.446	0.894	35.4	1.73	0.53
		1000	243	452	43.9	0.818	0.714	0.965	52.8	1.22	0.19

certain advantage, especially for Student's t and $\chi^2_{(1)}$ error distributions. Both the NMCD and ECP methods significantly outperform that of Lee (1996) in models (I) and (II). Under model (III), both the methods of ECP and Lee (1996) appear not working well, while the NMCD still produces reasonable detection results. As the divergence measure used in the ECP is essentially similar to Euclidean distances, the ECP is expected to perform well when the distributions differ in the first two moments, which however is not the case for model (III). The advantages of NMCD are mainly due to the joint use of the nonparametric likelihood and the weight function $w(u) = \{\widehat{F}_n(u)(1 - \widehat{F}_n(u))\}^{-1} d\widehat{F}_n(u)$. Based on the empirical distribution functions, the nonparametric likelihood approach is capable of detecting various types of changes. In addition, the difference between two adjacent distributions under model (III) does not lie in their centers, and thus using our proposed $w(u)$ would provide certain improvement as discussed in Section 3.2. Due to the use of DP, our procedure is much faster than the ECP.

4.5. *Comparison of NMCD and LSTV.* Harchaoui and Lévy-Leduc (2010) proposed the least-squares total variation method (LSTV) to estimate the locations of multiple change-points. By reframing the MCP in a variable selection context, they use a penalized least-squares criterion with a LASSO-type penalty. The LSTV enjoys efficient computation using the least angle regression [Efron et al. (2004)], while it does not provide competitive performance relative to the classical least-squares method with the DP, even when the true number of change-points is known. To improve the performance, the so-called LSTV* was further developed by incorporating a reduced version of the DP. Roughly speaking, the LSTV plays essentially a similar role in the LSTV* as our screening procedure in the NMCD. We conduct comparisons between LSTV, LSTV* and NMCD under model (I) only as the former two methods are not effective for scale changes in model (II). The LSTV procedure is implemented until the cardinality of the active set is exactly $K_n = 11$, and both the NMCD and LSTV* procedures are implemented by setting $\overline{K}_n = 30$ and using the BIC to estimate the number of change-points.

The results in Table 4 show that the proposed NMCD and LSTV* substantially outperform LSTV in terms of both $\xi(\widehat{\mathcal{G}}_n \| \mathcal{C}_t)$ and $\xi(\mathcal{C}_t \| \widehat{\mathcal{G}}_n)$. Moreover, the NMCD performs uniformly better than LSTV*, which may be partly explained by the fact that the induced shrinkage of LASSO often results in significant bias toward zero for large regression coefficients [Fan and Li (2001)]. Consequently, the LSTV also suffers from such bias, which in turn may lead to unsatisfactory estimation of the locations τ_k's. In Table 4, we also report the average computation time of the NMCD and LSTV* methods using an Intel Core 2.2 MHz CPU. For a large sample size, NMCD is much faster.

4.6. *Performance of NMCD with a diverging number of change-points.* To examine the setting that the number of change-points increases with the sample size,

TABLE 4
Comparison of NMCD, LSTV and LSTV under model* (I)

		LSTV		LSTV*			NMCD		
n	σ	$\xi(\widehat{\mathcal{G}}_n\|\mathcal{C}_t)$	$\xi(\mathcal{C}_t\|\widehat{\mathcal{G}}_n)$	$\xi(\widehat{\mathcal{G}}_n\|\mathcal{C}_t)$	$\xi(\mathcal{C}_t\|\widehat{\mathcal{G}}_n)$	$\lvert\widehat{K}_n - K_n\rvert$	$\xi(\widehat{\mathcal{G}}_n\|\mathcal{C}_t)$	$\xi(\mathcal{C}_t\|\widehat{\mathcal{G}}_n)$	$\lvert\widehat{K}_n - K_n\rvert$
500	0.1	20.2	31.2	1.14	1.88	0.18	0.00	0.00	0.00
	0.25	23.4	29.4	1.08	2.05	0.18	0.07	0.07	0.00
	0.5	26.1	27.0	2.10	3.14	0.17	1.39	1.30	0.03
1000	0.1	43.1	60.2	2.82	2.21	0.15	0.00	0.00	0.00
	0.25	46.2	59.4	3.23	2.24	0.16	0.04	0.04	0.00
	0.5	48.4	51.0	4.45	2.49	0.17	1.20	1.20	0.01
Computation time per run (in seconds)				$n = 500 : 0.102$ $n = 1000 : 0.776$			$n = 500 : 0.054$ $n = 1000 : 0.24$		

we choose seven increasing sample sizes, $n = 1000$, 1500, 2000, 3000, 5000, 7500 and 10,000, under models (I) and (II), respectively. The number of change-points in model (I) is chosen as $K_n = \lceil 0.4n^{1/2} \rceil$, corresponding to the values of 13, 16, 18, 22, 29, 35 and 40. In each replication, we randomly generate the jump sizes h_j as follows: $h_{2k-1} = -1.5 + \nu_{2k-1}$ and $h_{2k} = 1.5 + \nu_{2k}$, $k = 1, \ldots, \lceil K_n/2 \rceil$, where $\nu_j \sim N(0, 0.2^2)$. In model (II), we take $K_n = \lceil 0.2n^{1/2} \rceil$, and we only consider the scale changes (i.e., $h_j = 0$ for all j) and the inflation (deflation) sizes v_j are chosen as: $v_{2k-1} = 1/(5 + \nu_{2k-1})$ and $v_{2k} = 5 + \nu_{2k}$, $k = 1, \ldots, \lceil K_n/2 \rceil$, where $\nu_j \sim N(0, 0.2^2)$. We take the error distributions to be $t_{(3)}$ and $\chi^2_{(1)}$ in models (I) and (II), respectively. We fix $\sigma = 0.5$, and generate $\{\tau_j/n\}_{j=1}^{K_n}$ from $U(0, 1)$. All the tuning parameters are the same as those in Section 4.3.

Figure 2 depicts the curves of $\xi(\widehat{\mathcal{G}}_n\|\mathcal{C}_t)$, $\xi(\mathcal{C}_t\|\widehat{\mathcal{G}}_n)$, and $100|\widehat{K}_n - K_n|$ versus the sample size, respectively. For both models, all the distance values are reasonably small and the three curves are generally stable. This demonstrates that the NMCD is able to deliver satisfactory detection performance with a diverging number of change-points. From all these numerical studies, we conclude that the proposed NMCD is a viable alternative approach to the MCP if we take into account its efficiency, computational speed, and robustness to error distributions and change patterns.

5. Example. For illustration, we apply the proposed NMCD procedure to identify changes in the isochore structure, which refers to the proportion of the G + C composition in the large-scale DNA bases rather than A or T [Oliver et al. (2004); Fearnhead and Vasileiou (2009)]. Such genetic information is important to understand the evolution of base composition, mutation and recombination rates. Figure 3 shows the G + C content in percentage of a chromosome sequence with long homogeneous genome regions characterized by well-defined mean G + C contents.

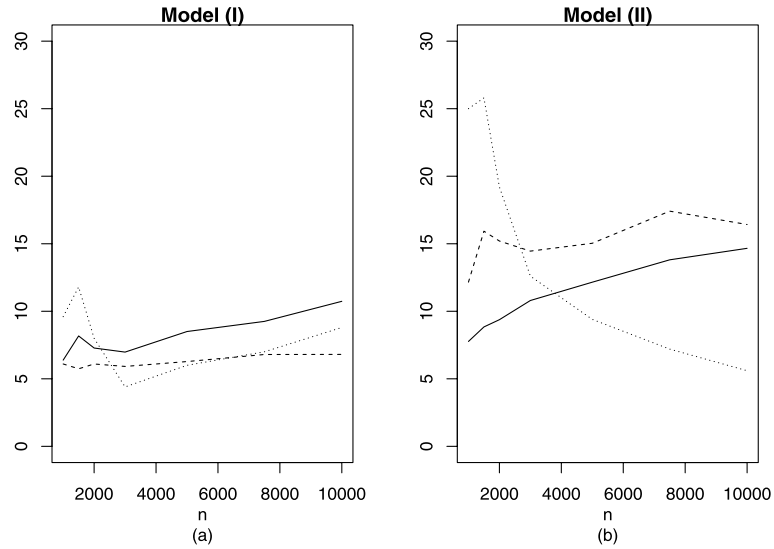

FIG. 2. *The performance of NMCD under models* (I) *and* (II) *when the number of change-points increases with the sample size*: *the solid*, *dashed and dotted lines represent* $\xi(\widehat{\mathcal{G}}_n \| \mathcal{C}_t)$, $\xi(\mathcal{C}_t \| \widehat{\mathcal{G}}_n)$, *and* $100|\widehat{K}_n - K_n|$ *versus the sample size*, *respectively.*

As the data sequence appears to be complicated without any obvious pattern and the sample size is large with $n = 8811$, identification of multiple change-points is very challenging. The data appear to contain quite a few outlying observations, and thus we expect that our nonparametric scheme would produce more robust detection results.

We take the upper bound for the number of change-points as $\overline{K}_n = 100$, and set $n_I = \lceil (\log n)^{3/2}/2 \rceil = 14$ and $\zeta_n = (\log n)^2/2 \approx 41$. After the initial screening procedure, 305 candidate points remain, which dramatically reduces the dimensionality of change-point detection. The BIC selection criterion further leads to the estimated number of change-points $\widehat{K}_n = 43$. The entire procedure is completed in 54 seconds using an Intel Core 2.2 MHz CPU. It can be seen from Figure 3 that the change-point estimates are generally reasonable based on the proposed NMCD procedure. It can detect some local and sharp features as well as those long unchanged data segments. For comparison, we also apply the LSTV* to the same dataset, and exhibit the result in Figure 3. The estimated number of change-points using LSTV* is $\widehat{K}_n = 26$. We can see that both methods perform well, and the line segments of the two methods are largely overlapping, except that the NMCD tends to detect relatively more picks or sharp changes. Some large changes could be overlooked by LSTV* due to the LASSO-type bias for large coefficients. This also explains that the number of change-points identified by the LSTV* is smaller than that of the proposed NMCD.

We performed the Shapiro–Wilk goodness-of-fit tests for normality on the 44 segments identified by NMCD and found that 34 tests are significant under the

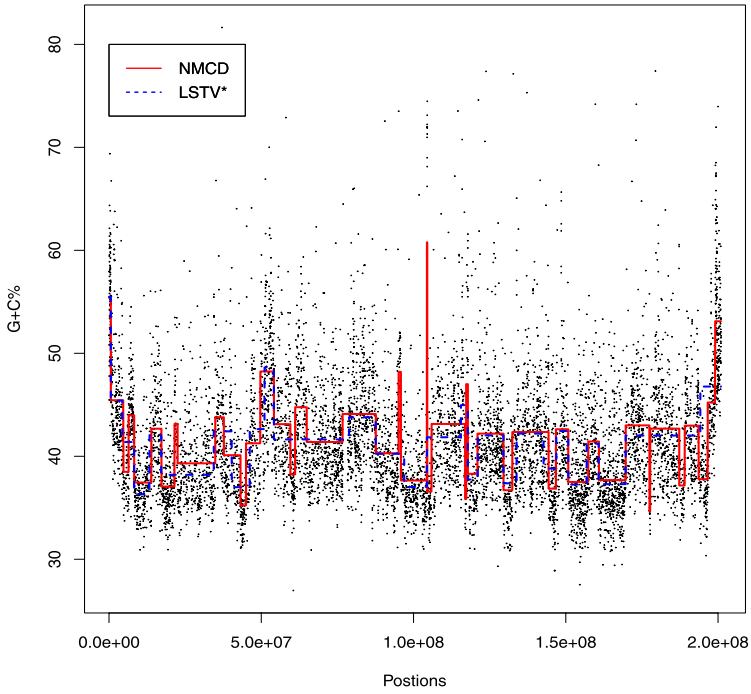

FIG. 3. *Illustration of a chromosome sequence with long homogeneous genome regions characterized by the mean $G + C$ contents, together with the estimated changepoints using the proposed NMCD and LSTV*, respectively. The red and blue solid lines represent the sample means in each segmentation estimated by NMCD and LSTV*, respectively.*

0.01 nominal level. As an example, Figure 4 shows the normal QQ-plot of the fifth segment, from which we can conclude that its distribution is far from normal. Furthermore, the density estimation of two consecutive segments (the 5th and 6th) shown in Figure 4 indicates that the two distributions differ not only in the location but also in the scale and shape. In light of these characteristics, our NMCD procedure is more desirable than those parametric methods which need to specify the mean or scale changes in advance.

6. Concluding remarks. In the MCP, we have proposed a nonparametric likelihood-based method for detection of multiple change-points. The consistency of the proposed NMCD procedure is established under mild conditions. The true number of change-points is assumed to be unknown, and the BIC is used to choose the number of change-points. To facilitate the implementation of NMCD, we suggest a DP algorithm in conjunction with a screening procedure, which has been shown to work well, particularly in large datasets. The computational scheme is fast and competitive with existing methods and, furthermore, numerical comparisons show that NMCD is able to strike a better balance for over- and under-

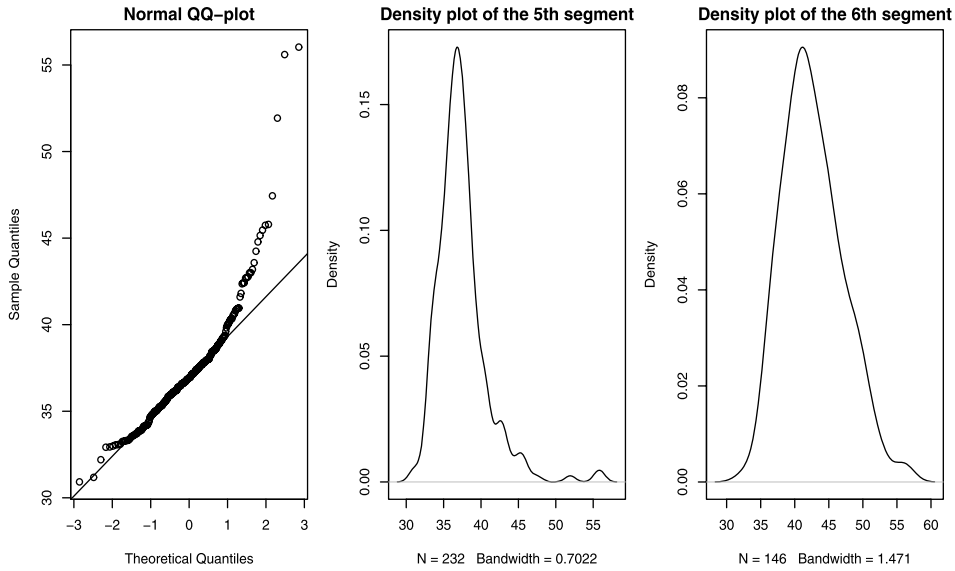

FIG. 4. *The first plot*: *normal QQ-plot of the* 5*th segment by using the NMCD*; *the second plot*: *density estimation of the* 5*th segment*; *the third plot*: *density estimation of the* 6*th segment*.

segmentation errors with nonnormal data and even has comparable performance with the parametric model under the correctly specified distributional assumption.

The proposed method is based on the assumption that there exists at least one change point. In practical applications, we need to use some tests within the nonparametric context to verify this assumption. The tests proposed by Einmahl and McKeague (2003) and Zou et al. (2007) are suited for this purpose. Our proposed NMCD is an omnibus method, and thus cannot diagnose whether a change occurs in the location, scale, or shape. To further determine which parameter changes, additional nonparametric tests need to be used as an auxiliary tool. Moreover, research is warranted to extend our method to other settings, such as the autocorrelated observations, multivariate cases [Matteson and James (2014)], and multiple structural changes in linear models [Bai and Perron (1998)].

APPENDIX

First of all, we present a lemma in Wellner (1978). Let $G_n(u)$ denote the empirical C.D.F. of a random sample of n uniform random variables on $(0,1)$, and define $\|G_n(u)/u\|_s^t \equiv \sup_{s\le u\le t}(G_n(u)/u)$ and $G_n^{-1}(u) = \inf\{s : G_n(s) \ge u\}$.

LEMMA 1. *For all* $\lambda \ge 0$ *and* $0 \le a \le 1$,

(i) $\Pr(\|G_n(u)/u\|_a^1 \ge \lambda) \le \exp\{-nah(\lambda)\}$,

(ii) $\Pr(\|u/G_n(u)\|_a^1 \geq \lambda) \leq \exp\{-nah(1/\lambda)\}$,
(iii) $\Pr(\|u/G_n^{-1}(u)\|_a^1 \geq \lambda) \leq \exp\{-naf(1/\lambda)\}$,
(iv) $\Pr(\|G_n^{-1}(u)/u\|_a^1 \geq \lambda) \leq \exp\{-naf(\lambda)\}$,
(v) $\Pr(|\|G_n(u)/u - 1\|_a^1| \geq \lambda) \leq 2\exp(-nah(1+\lambda))$,

where $h(x) = x(\log x - 1) + 1$ *and* $f(x) = x + \log(1/x) - 1$.

Before proceeding further, we state a key lemma, which allows us to control the supremum of the likelihood function.

LEMMA 2. *Suppose that assumptions* (A1)–(A2) *hold and* $K_n(\log \delta_n)/\delta_n \to 0$. *Let* $w_n \equiv C_\epsilon K_n(\log K_n)^2(\log(\delta_n K_n))^2$, *then*

$$\lim_{n\to\infty} K_n \Pr\Big\{ \sup_{\tau_{m-1}\leq k<l<\tau_{m-1}+\delta_n} \xi_m(k,l) \geq w_n \Big\} < \epsilon,$$

where

$$\xi_m(k,l) = n_{kl} \int_{X_{(1)}}^{X_{(n)}} \Big\{ \widehat{F}_k^l(u) \log\Big(\frac{\widehat{F}_k^l(u)}{F_m(u)}\Big) + (1-\widehat{F}_k^l(u)) \log\Big(\frac{1-\widehat{F}_k^l(u)}{1-F_m(u)}\Big) \Big\} \frac{d\widehat{F}_n(u)}{\widehat{F}_n(u)(1-\widehat{F}_n(u))},$$

$n_{kl} = l - k$ *and* C_ϵ *is given in the proof.*

PROOF. Without loss of generality, suppose that F_m is uniform on $[0,1]$ and $0 < X_1 < \cdots < X_n < 1$. Then we have

$$\xi_m(k,l) = n_{kl} \int_{X_{(1)}}^{X_{(n)}} H\big(\widehat{F}_k^l(u), u\big)\big\{\widehat{F}_n(u)\big(1-\widehat{F}_n(u)\big)\big\}^{-1} d\widehat{F}_n(u), \tag{A.1}$$

where

$$H(x,y) = x\log\Big(\frac{x}{y}\Big) + (1-x)\log\Big(\frac{1-x}{1-y}\Big).$$

By setting $a_n = 3h^{-1}(1+\alpha)\delta_n^{-1}\log(\delta_n K_n) \equiv D_\alpha \delta_n^{-1}\log(\delta_n K_n)$, $0 < \alpha < 1/2$, and noting that $h(1+\alpha) > 0$, we write

$$\begin{aligned}
&\xi_m(k,l) \\
&\quad = n_{kl}\Big(\int_{X_{(1)}}^{a_n} + \int_{a_n}^{1-a_n} + \int_{1-a_n}^{X_{(n)}}\Big) H\big(\widehat{F}_k^l(u), u\big)\big\{\widehat{F}_n(u)\big(1-\widehat{F}_n(u)\big)\big\}^{-1} d\widehat{F}_n(u) \\
&\quad \equiv \Delta_1 + \Delta_2 + \Delta_3.
\end{aligned}$$

First, we provide an upper bound for $K_n \Pr(\sup_{k,l} \Delta_1 \geq w_n/3)$, where $\Delta_1 \equiv \Delta_{11} + \Delta_{12}$ with

$$\Delta_{11} = n_{kl} \int_{X_{(1)}}^{a_n} \frac{\widehat{F}_k^l(u)}{u} \log\biggl(\frac{\widehat{F}_k^l(u)}{u}\biggr) \frac{u}{\widehat{F}_n(u)(1-\widehat{F}_n(u))}\, d\widehat{F}_n(u),$$

$$\Delta_{12} = n_{kl} \int_{X_{(1)}}^{a_n} \frac{1-\widehat{F}_k^l(u)}{1-u} \log\biggl(\frac{1-\widehat{F}_k^l(u)}{1-u}\biggr) \frac{1-u}{\widehat{F}_n(u)(1-\widehat{F}_n(u))}\, d\widehat{F}_n(u).$$

To show this, we choose $\lambda_{\epsilon 1}$ such that as $n \to \infty$,

$$\begin{aligned} &K_n \Pr\bigl(\bigl\|u/\widehat{F}_n(u)\bigr\|_{X_{(1)}}^1 > \log K_n \lambda_{\epsilon 1}\bigr) \\ &\quad\leq K_n \Pr\biggl(\biggl\| \frac{nu}{(\tau_m - \tau_{m-1})\widehat{F}_{\tau_{m-1}}^{\tau_m}(u)} \biggr\|_{X_{(1)}}^1 > \lambda_{\epsilon 1} \log K_n\biggr) \\ &\quad\leq n^{-1}(\tau_m - \tau_{m-1}) K_n^2 e \lambda_{\epsilon 1} \exp\bigl\{-n^{-1}(\tau_m - \tau_{m-1})\lambda_{\epsilon 1} \log K_n\bigr\} < \epsilon/12, \end{aligned}$$

based on assumption (A2) and the fact that

$$\Pr\bigl(\bigl\|u/G_n(u)\bigr\|_{X_{(1)}}^1 > \lambda\bigr) \leq \Pr\bigl(\bigl\|G_n^{-1}(u)/u\bigr\|_{1/n}^1 \geq \lambda\bigr) \leq e\lambda \exp\{-\lambda\}$$

by using Lemma 1(iv). Similarly,

$$K_n \Pr\bigl\{\bigl\|\widehat{F}_n^{-1}(u)/u\bigr\|_{1/n}^1 > \lambda_{\epsilon 1} \log K_n\bigr\} < \epsilon/12.$$

Also, we consider the event $A_m \equiv \bigcup_{k,l}\{\|\widehat{F}_k^l(u)/u\|_0^1 > \lambda_{\epsilon 2} K_n \delta_n / n_{kl}\}$, and thus

$$\begin{aligned} K_n \Pr(A_m) &= K_n \Pr\biggl(\bigcup_{k,l} \frac{n_{kl}}{\delta_n} \bigl\|\widehat{F}_k^l(u)/u\bigr\|_0^1 > \lambda_{\epsilon 2} K_n\biggr) \\ &\leq K_n \Pr\biggl(\bigcup_{k,l} \bigl\|\widehat{F}_{\tau_{m-1}}^{\tau_{m-1}+\delta_n}(u)/u\bigr\|_0^1 > \lambda_{\epsilon 2} K_n\biggr) \\ &= K_n \Pr\bigl(\bigl\|\widehat{F}_{\tau_{m-1}}^{\tau_{m-1}+\delta_n}(u)/u\bigr\|_0^1 > \lambda_{\epsilon 2} K_n\bigr) \leq e\lambda_{\epsilon 2}^{-1} < \epsilon/12 \end{aligned}$$

by choosing a proper $\lambda_{\epsilon 2}$. In parallel, let $B_m \equiv \bigcup_{k,l}\{\|(1-\widehat{F}_k^l(u))/(1-u)\|_0^1 > \lambda_{\epsilon 2} K_n \delta_n / n_{kl}\}$, and we have

$$K_n \Pr\bigl(\bigl\|(1-u)/\bigl(1-\widehat{F}_n(u)\bigr)\bigr\|_0^1 > \lambda_{\epsilon 1} \log K_n\bigr) < \epsilon/12$$

and $K_n \Pr(B_m) < e\lambda_{\epsilon 2}^{-1} < \epsilon/12$.

For the interaction of the events $\overline{A}_m$, $\|u/\widehat{F}_n(u)\|_0^1 \leq \lambda_{\epsilon 1} \log K_n$, and

$$\bigl\|\widehat{F}_n^{-1}(u)/u\bigr\|_{1/n}^1 \leq \lambda_{\epsilon 1} \log K_n,$$

we have

$$\begin{aligned}\Delta_{11} &= n_{kl}\int_{X_{(1)}}^{a_n} \frac{\widehat{F}_k^l(u)}{u}\log\Big(\frac{\widehat{F}_k^l(u)}{u}\Big)\frac{u}{\widehat{F}_n(u)}\frac{1}{(1-\widehat{F}_n(u))}\,d\widehat{F}_n(u)\\ &\le -n_{kl}\frac{K_n\delta_n}{n_{kl}}\lambda_{\epsilon 2}\log\Big(\frac{K_n\delta_n}{n_{kl}}\lambda_{\epsilon 2}\Big)\lambda_{\epsilon 1}\log K_n\log\big(1-\widehat{F}_n^{-1}(a_n)\big)\\ &\le -K_n\delta_n\lambda_{\epsilon 2}\log(K_n\delta_n\lambda_{\epsilon 2})\lambda_{\epsilon 1}\log K_n\log(1-\lambda_{\epsilon 1}a_n\log K_n)\\ &\le K_n(\log K_n)^2\delta_n a_n\lambda_{\epsilon 2}\lambda_{\epsilon_1}^2\log(\delta_n K_n)\big(1+o(1)\big)\end{aligned}$$

as $n\to\infty$. Consequently, as $n\to\infty$,

$$\begin{aligned}&K_n\Pr\Big(\sup_{k,l}\Delta_{11}\ge w_n/6\Big)\\ &\quad\le K_n\Pr(A_m)+K_n\Pr\big(\|\widehat{F}_n^{-1}(u)/u\|_{1/n}^1>\lambda_{\epsilon 1}\log K_n\big)\\ &\qquad+K_n\Pr\big(\|u/\widehat{F}_n(u)\|_{X_{(1)}}^1>\lambda_{\epsilon 1}\log K_n\big)+\delta_n^2K_n\Pr(\Delta_{11}\ge w_n/6)\\ &\quad\le\frac{1}{4}\epsilon+\delta_n^2K_n\Pr\big\{\lambda_{\epsilon 2}\lambda_{\epsilon_1}^2(\log\delta_nK_n)^2K_n(\log K_n)^2\big(1+o(1)\big)\ge w_n/6\big\}=\frac{1}{4}\epsilon,\end{aligned}$$

where the probability $\Pr\{(\log\delta_nK_n)^2K_n(\log K_n)^2\lambda_{\epsilon 2}\lambda_{\epsilon 1}^2(1+o(1))\ge w_n\}$ would be zero when n is sufficiently large, as long as $C_\epsilon>6D_\alpha\lambda_{\epsilon 2}\lambda_{\epsilon_1}^2$.

Similarly, we can show that $K_n\Pr(\sup_{k,l}\Delta_{12}\ge w_n/6)\le\epsilon/4$ as $n\to\infty$. Thus,

$$\Pr\Big(\sup_{k,l}\Delta_1\ge w_n/3\Big)\le\Pr\Big(\sup_{k,l}\Delta_{11}\ge w_n/6\Big)+\Pr\Big(\sup_{k,l}\Delta_{12}\ge w_n/6\Big)<\epsilon/2.$$

By symmetry, we immediately have

$$K_n\Pr\Big(\sup_{k,l}\Delta_3\ge w_n/3\Big)\le\frac{1}{2}\epsilon\qquad\text{as } n\to\infty.$$

Thus, it remains to give a bound of $K_n\Pr(\sup_{k,l}\Delta_2\ge w_n/3)$. Following similar argument in the proof of Theorem 3.1 of Jager and Wellner (2007), we can express $H(\widehat{F}_{kl}(u),u)$ as

$$H\big(\widehat{F}_n(u),u\big)=\frac{1}{2}\frac{(\widehat{F}_k^l(u)-u)^2}{\widehat{F}_{kl}^*(u)(1-\widehat{F}_{kl}^*(u))}$$

for $0<u<1$ where $|\widehat{F}_{kl}^*(u)-u|\le|\widehat{F}_k^l(u)-u|$. Then we rewrite Δ_2 as

$$\begin{aligned}\Delta_2&=\frac{1}{2}\int_{a_n}^{1-a_n}\frac{n_{kl}(\widehat{F}_k^l(u)-u)^2}{u(1-u)}\frac{u(1-u)}{\widehat{F}_{kl}^*(u)(1-\widehat{F}_{kl}^*(u))}\frac{d\widehat{F}_n(u)}{\widehat{F}_n(u)(1-\widehat{F}_n(u))}\\ &\le\frac{1}{2}\Big\|\frac{n_{kl}(\widehat{F}_k^l(u)-u)^2}{u(1-u)}\Big\|_{a_n}^{1-a_n}\Big\|\frac{u}{\widehat{F}_{kl}^*(u)}\Big\|_{a_n}^{1-a_n}\Big\|\frac{1-u}{1-\widehat{F}_{kl}^*(u)}\Big\|_{a_n}^{1-a_n}\\ &\quad\times\int_{a_n}^{1-a_n}\frac{d\widehat{F}_n(u)}{\widehat{F}_n(u)(1-\widehat{F}_n(u))}.\end{aligned}$$

Consider the event $C_m \equiv \bigcup_{k,l}\{|\|\widehat{F}_k^l(u)/u-1\|_{a_n}^{1-a_n}|>\alpha\}$ for some $0<\alpha<1$ and, by applying Lemma 1(v), we have

$$\begin{aligned} K_n \Pr(C_m) &\le \delta_n^2 K_n \Pr\bigl(\bigl|\bigl\|\widehat{F}_k^l(u)/u-1\bigr\|_{a_n}^{1-a_n}\bigr|>\alpha\bigr) \\ &\le 2\exp\bigl\{2\log(\delta_n K_n)-\delta_n a_n h(1+\alpha)\bigr\}\to 0. \end{aligned}$$

On the event $\overline{C}_m$ and $|\widehat{F}_{kl}^*(u)/u-1|<|\widehat{F}_{kl}(u)/u-1|<\alpha$, we have

$$\left\|\frac{u}{\widehat{F}_{kl}^*(u)}\right\|_{a_n}^{1-a_n}<\frac{1}{1-\alpha}.$$

Symmetrically, we also have

$$\left\|\frac{1-u}{1-\widehat{F}_{kl}^*(u)}\right\|_{a_n}^{1-a_n}<\frac{1}{1-\alpha}$$

on the event $\overline{D}_m$, where $D_m \equiv \bigcup_{k,l}\{|\|(1-\widehat{F}_k^l(u))/(1-u)-1\|_{a_n}^{1-a_n}|>\alpha\}$ occurs with the probability tending to zero. On the other hand, by using Lemma 1(v) again, it is easy to see that, for sufficiently large n,

$$\int_{a_n}^{1-a_n}\bigl\{\widehat{F}_n(u)\bigl(1-\widehat{F}_n(u)\bigr)\bigr\}^{-1}\,d\widehat{F}_n(u)\le -2\log a_n+C_\alpha\le 2\log(\delta_n K_n)\bigl(1+o_p(1)\bigr),$$

where the constant C_α depends on α.

Now, we consider the term $\|n_{kl}(\widehat{F}_k^l(u)-u)^2/\{u(1-u)\}\|_{a_n}^{1-a_n}$, and let $\varrho_n=(w_n/\log(\delta_n K_n))^{1/2}$. By taking $q(t)=\sqrt{t(1-t)}$ in Inequality 11.2.1 of Shorack and Wellner [(1986), page 446],

$$\begin{aligned} \Pr\left(\left\|\frac{n_{kl}(\widehat{F}_k^l(u)-u)^{\pm}}{\sqrt{u(1-u)}}\right\|_{a_n}^{1/2}\ge\varrho_n\right) &\le 6\int_{a_n}^{1/2}\frac{1}{t}\exp\left\{-\frac{1}{8}\gamma^{\pm}\varrho_n^2(1-t)\right\}dt \\ &\le 6\exp\left\{-\frac{1}{16}\gamma^{\pm}\varrho_n^2\right\}\log\delta_n\bigl(1+o(1)\bigr), \end{aligned}$$

where $\gamma^-=1$, $\gamma^+=\psi(\varrho_n/\sqrt{\delta_n a_n})$, and $\psi(x)=2h(1+x)/x^2$. By using the fact that $\psi(x)\sim 2(\log x)/x$ as $x\to\infty$ [Proposition 11.1.1 in Shorack and Wellner (1986)], $\gamma^+\sim\log(C_\epsilon K_n(\log K_n)^2)/(C_\epsilon^{1/2}K_n^{1/2}\log K_n)$ for sufficiently large C_ϵ. Consequently, we have

$$\begin{aligned} &K_n\Pr\left(\sup_{k,l}\left\|\frac{n_{kl}(\widehat{F}_k^l(u)-u)^2}{u(1-u)}\right\|_{a_n}^{1/2}\ge\frac{(1-\alpha)^2 w_n}{3\log(\delta_n K_n)}\right) \\ &\quad\le K_n\delta_n^2\Pr\left(\left\|\frac{n_{kl}(\widehat{F}_k^l(u)-u)^{\pm}}{\sqrt{u(1-u)}}\right\|_{a_n}^{1/2}\ge\varrho_n\right) \\ &\quad\le 12\exp\left(2\log(\delta_n K_n)-\frac{1}{16}\gamma^+\varrho_n^2\right)\log\delta_n\bigl(1+o(1)\bigr) \\ &\quad\to 0 \qquad \text{as } \delta_n\to\infty \end{aligned}$$

as long as C_ϵ is sufficiently large. By symmetry, we can also show that

$$K_n \Pr\Bigg(\sup_{k,l}\bigg\|\frac{n_{kl}(\widehat{F}_k^l(u)-u)^2}{u(1-u)}\bigg\|_{1/2}^{1-a_n} \geq \frac{(1-\alpha)^2 w_n}{3\log(\delta_n K_n)}\Bigg) \to 0.$$

Finally, we obtain as $n \to \infty$,

$$\begin{aligned} &K_n \Pr\Big(\sup_{k,l}\Delta_2 \geq w_n/3\Big) \\ &\quad \leq K_n \Pr(C_m) + K_n \Pr(D_m) \\ &\qquad + K_n \Pr\Bigg(\sup_{k,l}\bigg\|\frac{n_{kl}(\widehat{F}_k^l(u)-u)^2}{u(1-u)}\bigg\|_{a_n}^{1-a_n} \frac{1}{(1-\alpha)^2}\log(\delta_n K_n) \geq w_n/3\Bigg) \to 0, \end{aligned}$$

which completes the proof of this lemma. □

By Lemma 2, the next lemma follows immediately.

LEMMA 3. *Suppose that assumptions* (A1)–(A2) *hold and* $K_n(\log n)/n \to 0$. *Then*

$$\lim_{n\to\infty} K_n \Pr\Big\{\sup_{\tau_{m-1}\leq k<l<\tau_m} \xi_m(k,l) \geq u_n\Big\} < \epsilon,$$

where $u_n \equiv C_\epsilon K_n(\log K_n)^2(\log(nK_n))^2$ *with a sufficiently large* C_ϵ.

Let $\widetilde{O}_p(q_n; K_n)$ be a sequence of positive random variables Z_n if for any $\epsilon > 0$,

$$\lim_{n\to\infty} K_n \Pr(Z_n > C_\epsilon q_n) < \epsilon,$$

where C_ϵ is a constant depending only on ϵ.

LEMMA 4. *Suppose that assumptions* (A1)–(A2) *hold. For any* $L \geq 1$ *and* $\tau_s < \tau_1' < \cdots < \tau_L' < \tau_{s+1}$, *as* $n \to \infty$,

$$\begin{aligned} 0 &\leq R_n\big(\tau_s, \tau_1', \ldots, \tau_L', \tau_{s+1}\big) - R_n(\tau_s, \tau_{s+1}) \\ &= \widetilde{O}_p\big(L^2 K_n\big(\log(K_n L)\big)^2\big(\log(nK_n L)\big)^2; K_n\big). \end{aligned}$$

PROOF. By noting that $H(x,y)$ is a convex function, the left inequality is obvious. Without loss of generality, we assume $L = 1$, and for $L > 1$ the result follows by induction. By the fact that $(\tau_1' - \tau_s)\widehat{F}_{\tau_s}^{\tau_1'}(u) + (\tau_{s+1} - \tau_1')\widehat{F}_{\tau_1'}^{\tau_{s+1}}(u) = (\tau_{s+1} - \tau_s)\widehat{F}_{\tau_s}^{\tau_{s+1}}(u)$,

$$\begin{aligned} R_n\big(\tau_s, \tau_1', \tau_{s+1}\big) - R_n(\tau_s, \tau_{s+1}) &= \xi_s\big(\tau_s, \tau_1'\big) + \xi_s\big(\tau_1', \tau_{s+1}\big) - \xi_s(\tau_s, \tau_{s+1}) \\ &\leq \xi_s\big(\tau_s, \tau_1'\big) + \xi_s\big(\tau_1', \tau_{s+1}\big). \end{aligned}$$

Similarly, for any L, we have

$$\begin{aligned} &R_n(\tau_s, \tau'_1, \ldots, \tau'_L, \tau_{s+1}) - R_n(\tau_s, \tau_{s+1}) \\ &\quad \le \xi_s(\tau_s, \tau'_1) + \xi_s(\tau'_1, \tau'_2) + \cdots + \xi_s(\tau'_L, \tau_{s+1}). \end{aligned}$$

Thus, for any $\epsilon > 0$,

$$\begin{aligned} &\lim_{n\to\infty} K_n \Pr\{R_n(\tau_s, \tau'_1, \ldots, \tau'_L, \tau_{s+1}) - R_n(\tau_s, \tau_{s+1}) \\ &\qquad > C_\epsilon L^2 K_n (\log(K_n L))^2 (\log(nK_n L))^2\} \\ &\le \lim_{n\to\infty} K_n \Pr\{\xi_s(\tau_s, \tau'_1) + \cdots + \xi_s(\tau'_L, \tau_{s+1}) \\ &\qquad\qquad > C_\epsilon L^2 K_n (\log(K_n L))^2 (\log(nK_n L))^2\} \\ &\le L^{-1} \sum_{k=0}^{L} \lim_{n\to\infty} K_n L \Pr\{\xi_s(\tau'_k - \tau'_{k+1}) \\ &\qquad\qquad\qquad > C_\epsilon L^2 K_n (\log(K_n L))^2 (\log(nK_n L))^2\} \\ &< L^{-1}(L+1)\epsilon, \end{aligned}$$

where the last result follows immediately from Lemma 3. □

Next, we demonstrate that the global minimum of the BIC includes no less than K_n change-point estimators asymptotically.

PROPOSITION 2. *If assumptions* (A1)–(A4) *hold*, $\Pr\{\widehat{K}_n \ge K_n\} \to 1$.

PROOF. Define $\rho_n = \lambda_n/8$, and consider $0 < L < K_n$. Let

$$\begin{aligned} &B_r(L, \rho_n) \\ &\quad = \{(\tau'_1, \ldots, \tau'_L) : 1 < \tau'_1 < \cdots < \tau'_L \le n \text{ and } |\tau'_s - \tau_r| > \rho_n \text{ for } 1 \le s \le L\}, \end{aligned}$$

$r = 1, \ldots, K_n$. For $L < K_n$, $(\hat{\tau}_1, \ldots, \hat{\tau}_L)$ must belong to one $B_r(L, \rho_n)$. For every $(\tau'_1, \ldots, \tau'_L) \in B_r(L, \rho_n)$, we have

$$\begin{aligned} &R_n(\tau'_1, \ldots, \tau'_L) \\ &\quad \le R_n(\tau'_1, \ldots, \tau'_L, \tau_1, \ldots, \tau_{r-1}, \tau_r - \rho_n, \tau_r + \rho_n, \tau_{r+1}, \ldots, \tau_{K_n}) \end{aligned} \tag{A.2}$$

and the right-hand side of (A.2) can be expressed as $T_1 + \cdots + T_{K_n+2}$, where T_s ($s = 1, \ldots, r-1, r+2, \ldots, K_n+1$) is the sum of integrals involving the X_i's ($\tau_{s-1} \le i < \tau_s$); T_r is that involving the X_i's ($\tau_{r-1} \le i < \tau_r - \rho_n$); T_{r+1} is that involving the X_i's ($\tau_r + \rho_n \le i < \tau_{r+1}$); T_{K_n+2} is that involving the X_i's ($\tau_r - \rho_n \le i < \tau_r + \rho_n$). For $s = 1, \ldots, r-1, r+2, \ldots, K_n+1$, by Lemma 4, we have

$$\begin{aligned} R_n(\tau_{s-1}, \tau_s) \le T_s &\le R_n(\tau_{s-1}, \tau_s) + \widetilde{O}_p\big(L^2 K_n (\log(K_n L))^2 (\log(nK_n L))^2\big) \\ &= R_n(\tau_{s-1}, \tau_s) + \widetilde{O}_p(b_n; K_n), \end{aligned}$$

where $b_n = K_n^3(\log K_n)^2(\log n)^2$. Similarly, we have

$$T_r = R_n(\tau_{r-1}, \tau_r - \rho_n) + \widetilde{O}_p(b_n; K_n),$$
$$T_{r+1} = R_n(\tau_r + \rho_n, \tau_{r+1}) + \widetilde{O}_p(b_n; K_n)$$

and in addition,

$$\begin{aligned} T_{K_n+2} &= R_n(\tau_r - \rho_n, \tau_r + \rho_n) + R_n(\tau_r + \rho_n, \tau_{r+1}) \\ &= R_n(\tau_r - \rho_n, \tau_r) + R_n(\tau_r, \tau_r + \rho_n) + \Delta, \end{aligned}$$

where $\Delta \equiv R_n(\tau_r - \rho_n, \tau_r + \rho_n) - R_n(\tau_r - \rho_n, \tau_r) - R_n(\tau_r, \tau_r + \rho_n)$. Note that

$$\begin{aligned} \Delta &= 2\rho_n \int_{X_{(1)}}^{X_{(n)}} \big[\widehat{F}_{\tau_r-\rho_n}^{\tau_r+\rho_n}(u) \log\big(F_{r,1/2}(u)\big) \\ &\qquad + \big\{1 - \widehat{F}_{\tau_r-\rho_n}^{\tau_r+\rho_n}(u)\big\} \log\big(1 - F_{r,1/2}(u)\big)\big]\, dw(u) \\ &\quad - \rho_n \int_{X_{(1)}}^{X_{(n)}} \big[\widehat{F}_{\tau_r-\rho_n}^{\tau_r}(u) \log\big(F_r(u)\big) \\ &\qquad + \big\{1 - \widehat{F}_{\tau_r-\rho_n}^{\tau_r}(u)\big\} \log\big(1 - F_r(u)\big)\big]\, dw(u) \\ &\quad - \rho_n \int_{X_{(1)}}^{X_{(n)}} \big[\widehat{F}_{\tau_r}^{\tau_r+\rho_n}(u) \log\big(F_{r+1}(u)\big) \\ &\qquad + \big\{1 - \widehat{F}_{\tau_r}^{\tau_r+\rho_n}(u)\big\} \log\big(1 - F_{r+1}(u)\big)\big]\, dw(u) + \widetilde{O}_p(b_n; K_n) \\ &= -\rho_n \int_{X_{(1)}}^{X_{(n)}} \left[\widehat{F}_{\tau_r-\rho_n}^{\tau_r}(u) \log\left(\frac{F_r(u)}{F_{r,1/2}(u)}\right)\right. \\ &\qquad \left. + \big\{1 - \widehat{F}_{\tau_r-\rho_n}^{\tau_r}(u)\big\} \log\left(\frac{1 - F_r(u)}{1 - F_{r,1/2}(u)}\right)\right] dw(u) \\ &\quad - \rho_n \int_{X_{(1)}}^{X_{(n)}} \left[\widehat{F}_{\tau_r}^{\tau_r+\rho_n}(u) \log\left(\frac{F_{\tau_{r+1}}(u)}{F_{r,1/2}(u)}\right)\right. \\ &\qquad \left. + \big\{1 - \widehat{F}_{\tau_r}^{\tau_r+\rho_n}(u)\big\} \log\left(\frac{1 - F_{r+1}(u)}{1 - F_{r,1/2}(u)}\right)\right] dw(u) \\ &\quad + \widetilde{O}_p(b_n; K_n) \\ &\equiv -\widetilde{\Delta} + \widetilde{O}_p(b_n; K_n). \end{aligned}$$

Let $\widetilde{\Delta} = \widetilde{\Delta}_1 + \widetilde{\Delta}_2$, and then

$$\begin{aligned} \widetilde{\Delta}_1 \geq \rho_n \int_{X_{(1)}}^{X_{(n)}} &\left[\widehat{F}_{\tau_r-\rho_n}^{\tau_r}(u) \log\left(\frac{F_r(u)}{F_{r,1/2}(u)}\right)\right. \\ &\left. + \big\{1 - \widehat{F}_{\tau_r-\rho_n}^{\tau_r}(u)\big\} \log\left(\frac{1 - F_r(u)}{1 - F_{r,1/2}(u)}\right)\right] dw(u) \end{aligned}$$

$$= \rho_n \int_0^1 \Bigg[\widehat{F}^{\tau_r}_{\tau_r-\rho_n}(u) \log\Big(\frac{F_r(u)}{F_{r,1/2}(u)}\Big) + \big\{1 - \widehat{F}^{\tau_r}_{\tau_r-\rho_n}(u)\big\} \log\Big(\frac{1-F_r(u)}{1-F_{r,1/2}(u)}\Big)\Bigg] dw(u)$$

$$\equiv \widetilde{\Delta}_1'.$$

By assumption (A3), we have

$$\widetilde{\Delta}_1' = \rho_n \int_0^1 \Bigg[F_r(u) \log\Big(\frac{F_r(u)}{F_{r,1/2}(u)}\Big) + \big\{1 - F_r(u)\big\} \log\Big(\frac{1-F_r(u)}{1-F_{r,1/2}(u)}\Big)\Bigg] \times \frac{1}{F(u)(1-F(u))}\, dF(u)\big(1+o(1)\big), \qquad \text{a.s.}$$

Using the similar procedure, we can obtain the corresponding bound for $\widetilde{\Delta}_2$. As a result, as $n \to \infty$,

$$\begin{aligned}\widetilde{\Delta} \geq \rho_n \Bigg\{ & \int_0^1 \Bigg[F_r(u) \log\Big(\frac{F_r(u)}{F_{r,1/2}(u)}\Big) + \big\{1 - F_r(u)\big\} \log\Big(\frac{1-F_r(u)}{1-F_{r,1/2}(u)}\Big)\Bigg] \\ & \times \frac{1}{F(u)(1-F(u))}\, dF(u) \\ & + \int_0^1 \Bigg[F_{r+1}(u) \log\Big(\frac{F_{r+1}(u)}{F_{r,1/2}(u)}\Big) + \big\{1 - F_{r+1}(u)\big\} \log\Big(\frac{1-F_{r+1}(u)}{1-F_{r,1/2}(u)}\Big)\Bigg] \\ & \times \frac{1}{F(u)(1-F(u))}\, dF(u) \Bigg\} \\ \equiv\ & \rho_n S(F_r, F_{r+1}),\end{aligned}$$

in which the distance $S(F_r, F_{r+1})$ is strictly larger than zero.

Therefore,

$$\begin{aligned}& \max_{(\tau_1',\ldots,\tau_L') \in B_r(L,\rho_n)} R_n\big(\tau_1', \ldots, \tau_L'\big) \\ & \quad \leq \max_{(\tau_1',\ldots,\tau_L') \in B_r(L,\rho_n)} R_n\big(\tau_1', \ldots, \tau_L', \tau_1, \ldots, \tau_{r-1}, \tau_r - \rho_n, \tau_r + \rho_n, \\ & \qquad\qquad \tau_{r+1}, \ldots, \tau_{K_n}\big) \\ & \quad = \sum_{s \neq r, r+1}^{K_n+1} R_n(\tau_{s-1}, \tau_s) + R_n(\tau_{r-1}, \tau_r - \rho_n) + R_n(\tau_r - \rho_n, \tau_r) \\ & \qquad + R_n(\tau_r, \tau_r + \rho_n) + R_n(\tau_r + \rho_n, \tau_{r+1}) + \Delta + \widetilde{O}_p(b_n; K_n) \\ & \quad \leq R_n(\tau_1, \ldots, \tau_{K_n}) - \rho_n S(F_r, F_{r+1}) + \widetilde{O}_p(b_n; K_n).\end{aligned}$$

Let $\mathrm{BIC}_* = -R_n(\tau_1, \ldots, \tau_{K_n}) + K_n\zeta_n$, and for $L < K_n$, with probability tending to 1, we have

$$\mathrm{BIC}_L - \mathrm{BIC}_* \geq \rho_n S(F_r, F_{r+1}) - \widetilde{O}_p(b_n; K_n) - (K_n - L)\zeta_n$$

as $n \to \infty$. For any $\epsilon > 0$, we have, as $n \to \infty$,

$$\begin{aligned}\Pr(\widehat{K}_n < K_n) &= \Pr\left(\bigcup_{L=1}^{K_n-1} (\mathrm{BIC}_L < \mathrm{BIC}_*)\right) \leq \sum_{L=1}^{K_n-1} \Pr(\mathrm{BIC}_L < \mathrm{BIC}_*) \\ &\leq \sum_{L=1}^{K_n-1} \Pr\bigl(\widetilde{O}_p(b_n; K_n) > \rho_n S(F_r, F_{r+1}) - (K_n - L)\zeta_n\bigr) \\ &\leq K_n \Pr\bigl(\widetilde{O}_p(b_n; K_n) > b_n\bigr) < \epsilon.\end{aligned}$$

This completes the proof of this proposition. □

Let $\mathcal{Q}_L(\zeta_n)$ denote the set of global minimum of BIC with ζ_n and its cardinality is L.

PROPOSITION 3. *Suppose that assumptions* (A1)–(A4) *hold. For* $K_n \leq L \leq \overline{K}_n$ *and*

$$\Pr\left(\bigcup_{r=1}^{K_n} \{\mathcal{Q}_L(\zeta_n) \in D_r(L, \rho_n)\}\right) \to 0$$

as $n \to \infty$, *where*

$$\begin{aligned}&D_r(L, \rho_n) \\ &\quad = \bigl\{(\tau_1', \ldots, \tau_L') : 1 < \tau_1' < \cdots < \tau_L' \leq n \text{ and } |\tau_s' - \tau_r| > \rho_n \text{ for } 1 \leq s \leq L\bigr\}.\end{aligned}$$

PROOF. For every $(\tau_1', \ldots, \tau_L') \in D_r(L, \rho_n)$,

$$\begin{aligned}&R_n(\tau_1', \ldots, \tau_L') \\ &\quad \leq R_n(\tau_1', \ldots, \tau_L', \tau_1, \ldots, \tau_{r-1}, \tau_r - \delta_n, \tau_r + \delta_n, \tau_{r+1}, \ldots, \tau_{K_n})\end{aligned} \tag{A.3}$$

and the right-hand side of (A.3) can be expressed as $T_1 + \cdots + T_{K_n+2}$, where T_s $(s = 1, \ldots, r-1, r+2, \ldots, K_n+1)$ is the sum of squares involving the X_i's $(\tau_{s-1} \leq i < \tau_s)$; T_r is that involving the X_i's $(\tau_{r-1} \leq i < \tau_r - \rho_n)$; T_{r+1} is that involving the X_i's $(\tau_r + \rho_n \leq i < \tau_{r+1})$; T_{K_n+2} is that involving the X_i's $(\tau_r - \rho_n \leq i < \tau_r + \rho_n)$. Define $c_n = \overline{K}_n^3 (\log \overline{K}_n)^2 (\log(n\overline{K}_n))^2$. It can be further seen that uniformly in $(\tau_1', \ldots, \tau_L') \in D_r(L, \rho_n)$,

$$\begin{aligned}T_s &= R_n(\tau_{s-1}, \tau_s) + \widetilde{O}_p(c_n; \overline{K}_n), \qquad s = 1, \ldots, r-1, r+2, \ldots, K_n+1, \\ T_r &= R_n(\tau_{r-1}, \tau_r - \rho_n) + \widetilde{O}_p(c_n; \overline{K}_n), \\ T_{r+1} &= R_n(\tau_r + \rho_n, \tau_{r+1}) + \widetilde{O}_p(c_n; \overline{K}_n) \quad \text{and} \\ T_{K_n+2} &\leq R_n(\tau_r - \rho_n, \tau_r) + R_n(\tau_r, \tau_r + \rho_n) - \rho_n S(F_r, F_{r+1}) + \widetilde{O}_p(c_n; \overline{K}_n).\end{aligned}$$

These results imply that

$$\mathrm{BIC}_L - \mathrm{BIC}_* \geq \rho_n S(F_r, F_{r+1}) - \widetilde{O}_p(c_n; \overline{K}_n).$$

Thus, as $n \to \infty$,

$$\begin{aligned}\Pr\left(\bigcup_{r=1}^{K_n}\{\mathcal{Q}_L(\zeta_n) \in D_r(L, \rho_n)\}\right) &\leq \Pr\left(\bigcup_{r=1}^{K_n}(\mathrm{BIC}_L < \mathrm{BIC}_*)\right) \\ &\leq \Pr\left(\bigcup_{r=1}^{K_n}\{\rho_n S(F_r, F_{r+1}) < \widetilde{O}_p(c_n; \overline{K}_n)\}\right) \\ &\leq \overline{K}_n \Pr\bigl(\widetilde{O}_p(c_n; \overline{K}_n) > c_n\bigr) < \epsilon\end{aligned}$$

for any $\epsilon > 0$. Thus, the result follows. □

PROOF OF THEOREM 1. Define $d_n = K_n^3(\log K_n)^2(\log(\delta_n K_n))^2$. For every $(\tau_1', \ldots, \tau_{K_n}') \in D_r(K_n, \delta_n)$,

$$\begin{aligned}&\max_{(\tau_1', \ldots, \tau_{K_n}') \in D_r(K_n, \delta_n)} R_n\bigl(\tau_1', \ldots, \tau_{K_n}'\bigr) \\ &\quad\leq R_n\bigl(\tau_1', \ldots, \tau_{K_n}', \tau_1, \ldots, \tau_{r-1}, \tau_r - \delta_n, \tau_r + \delta_n, \tau_{r+1}, \ldots, \tau_{K_n}\bigr) \\ &\quad\leq R_n(\tau_1, \ldots, \tau_{K_n}) - \delta_n S(F_r, F_{r+1}) + \widetilde{O}_p(d_n; K_n)\end{aligned}$$

by Lemma 2. Thus, we know that

$$\max_{(\tau_1', \ldots, \tau_{K_n}') \in D_r(K_n, \delta_n)} R_n\bigl(\tau_1', \ldots, \tau_{K_n}'\bigr) < R_n(\tau_1, \ldots, \tau_{K_n})$$

with probability tending to one for each r. Consequently,

$$\begin{aligned}\Pr\bigl\{\mathcal{G}_n(K_n) \in C_{K_n}(\delta_n)\bigr\} &= 1 - \Pr\left\{\bigcup_r\bigl\{\mathcal{G}_n(K_n) \in D_r(K_n, \delta_n)\bigr\}\right\} \\ &\geq 1 - \sum_{r=1}^{K_n}\Pr\bigl\{\mathcal{G}_n(K_n) \in D_r(K_n, \delta_n)\bigr\} \to 1\end{aligned}$$

by the similar argument as that in Proposition 3. □

PROOF OF THEOREM 2. By Proposition 2, it suffices to show that $\Pr(\widehat{K}_n > K_n) \to 0$. This can be proved by contradiction. Let $E(L, \rho_n)$ be the complement of the union of $D_1(L, \rho_n), \ldots, D_{K_n}(L, \rho_n)$. As shown in Proposition 3, for $K_n < L < \overline{K}_n$ and every $(\tau_1', \ldots, \tau_L') \in E(L, \rho_n)$,

$$\begin{aligned}R_n\bigl(\tau_1', \ldots, \tau_L'\bigr) &\leq R_n\bigl(\tau_1', \ldots, \tau_L', \tau_1, \ldots, \tau_r, \tau_1 - \rho_n, \tau_{K_n} - \rho_n, \tau_1 + \rho_n, \tau_{K_n} + \rho_n\bigr) \\ &= R_n(\tau_1, \ldots, \tau_{K_n}) + \widetilde{O}_p(c_n; \overline{K}_n).\end{aligned}$$

Consequently, as $n \to \infty$,

$$\mathrm{BIC}_L - \mathrm{BIC}_* \geq (L - K_n)\zeta_n - \widetilde{O}_p(c_n; \overline{K}_n),$$

we obtain the result by the same argument as that in Proposition 2. □

Acknowledgments. The authors would like to thank Professor Runze Li, an Associate Editor, and three anonymous referees for their many insightful and constructive comments that have resulted in significant improvements in the article.

SUPPLEMENTARY MATERIAL

Supplement to "Nonparametric maximum likelihood approach to multiple change-point problems" (DOI: 10.1214/14-AOS1210SUPP; pdf). We provide technical details for the proof of Corollary 1, and additional simulation results.

REFERENCES

ARLOT, S., CELISSE, A. and HARCHAOUI, Z. (2012). Kernel change-point detection. Available at arXiv:1202.3878.

BAI, J. and PERRON, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica* **66** 47–78. MR1616121

BAI, J. and PERRON, P. (2003). Computation and analysis of multiple structural change models. *J. Appl. Econometrics* **18** 10–22.

BELLMAN, R. E. and DREYFUS, S. E. (1962). *Applied Dynamic Programming*. Princeton Univ. Press, Princeton, NJ. MR0140369

BOYSEN, L., KEMPE, A., LIEBSCHER, V., MUNK, A. and WITTICH, O. (2009). Consistencies and rates of convergence of jump-penalized least squares estimators. *Ann. Statist.* **37** 157–183. MR2488348

BRAUN, J. V., BRAUN, R. K. and MÜLLER, H.-G. (2000). Multiple changepoint fitting via quasilikelihood, with application to DNA sequence segmentation. *Biometrika* **87** 301–314. MR1782480

CARLSTEIN, E. (1988). Nonparametric change-point estimation. *Ann. Statist.* **16** 188–197. MR0924865

CHEN, J. and GUPTA, A. K. (1997). Testing and locating variance changepoints with application to stock prices. *J. Amer. Statist. Assoc.* **92** 739–747. MR1467863

CHEN, H. and ZHANG, N. R. (2012). Graph-based change-point detection. Available at arXiv:1209.1625v1.

CSÖRGŐ, M. and HORVÁTH, L. (1997). *Limit Theorems in Change-Point Analysis*. Wiley, Chichester. MR2743035

DARKHOVSKH, B. S. (1976). A nonparametric method for the a posteriori detection of the "disorder" time of a sequence of independent random variables. *Theory Probab. Appl.* **21** 178–183.

DONOHO, D. L. and JOHNSTONE, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.* **90** 1200–1224. MR1379464

DÜMBGEN, L. (1991). The asymptotic behavior of some nonparametric change-point estimators. *Ann. Statist.* **19** 1471–1495. MR1126333

EFRON, B., HASTIE, T., JOHNSTONE, I. and TIBSHIRANI, R. (2004). Least angle regression. *Ann. Statist.* **32** 407–499. MR2060166

EINMAHL, J. H. J. and MCKEAGUE, I. W. (2003). Empirical likelihood based hypothesis testing. *Bernoulli* **9** 267–290. MR1997030

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.* **96** 1348–1360. MR1946581

FEARNHEAD, P. and VASILEIOU, D. (2009). Bayesian analysis of isochores. *J. Amer. Statist. Assoc.* **104** 132–141. MR2663038

FOWLKES, E. B. and MALLOWS, C. L. (1983). A method for comparing two hierarchical clusterings. *J. Amer. Statist. Assoc.* **78** 553–569.

GUAN, Z. (2004). A semiparametric changepoint model. *Biometrika* **91** 849–862. MR2126037

HAO, N., NIU, Y. and ZHANG, H. (2013). Multiple change-point detection via a screening and ranking algorithm. *Statist. Sinica* **23** 1553–1572.

HARCHAOUI, Z. and LÉVY-LEDUC, C. (2010). Multiple change-point estimation with a total variation penalty. *J. Amer. Statist. Assoc.* **105** 1480–1493. MR2796565

HAWKINS, D. M. (2001). Fitting multiple change-point models to data. *Comput. Statist. Data Anal.* **37** 323–341. MR1856677

JAGER, L. and WELLNER, J. A. (2007). Goodness-of-fit tests via phi-divergences. *Ann. Statist.* **35** 2018–2053. MR2363962

KILLICK, R., FEARNHEAD, P. and ECKLEY, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *J. Amer. Statist. Assoc.* **107** 1590–1598. MR3036418

LAVIELLE, M. (2005). Using penalized contrasts for the change-points problems. *Signal Process.* **85** 1501–1510.

LEE, C.-B. (1996). Nonparametric multiple change-point estimators. *Statist. Probab. Lett.* **27** 295–304. MR1395582

MATTESON, D. S. and JAMES, N. A. (2014). A nonparametric approach for multiple change point analysis of multivariate data. *J. Amer. Statist. Assoc.* **109** 334–345. MR3180567

NIU, Y. S. and ZHANG, H. (2012). The screening and ranking algorithm to detect DNA copy number variations. *Ann. Appl. Stat.* **6** 1306–1326. MR3012531

OLIVER, J. L., CARPENA, P., HACKENBERG, M. and BERNAOLA-GALVÁN, P. (2004). IsoFinder: Computational prediction of isochores in genome sequences. *Nucleic Acids Res.* **32** W287–W292.

RIGAILL, G. (2010). Pruned dynamic programming for optimal multiple change-point detection. Available at arXiv:1004.0887.

SHEN, X. and YE, J. (2002). Adaptive model selection. *J. Amer. Statist. Assoc.* **97** 210–221. MR1947281

SHORACK, G. R. and WELLNER, J. A. (1986). *Empirical Processes with Applications to Statistics*. Wiley, New York. MR0838963

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* **58** 267–288. MR1379242

WELLNER, J. A. (1978). Limit theorems for the ratio of the empirical distribution function to the true distribution function. *Z. Wahrsch. Verw. Gebiete* **45** 73–88.

YAO, Y.-C. (1988). Estimating the number of change-points via Schwarz' criterion. *Statist. Probab. Lett.* **6** 181–189. MR0919373

YAO, Y.-C. and AU, S. T. (1989). Least-squares estimation of a step function. *Sankhyā Ser. A* **51** 370–381. MR1175613

ZHANG, J. (2002). Powerful goodness-of-fit tests based on the likelihood ratio. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 281–294. MR1904705

ZHANG, J. (2006). Powerful two-sample tests based on the likelihood ratio. *Technometrics* **48** 95–103. MR2236531

ZOU, C., YIN, G., FENG, L. and WANG, Z. Supplement to "Nonparametric maximum likelihood approach to multiple change-point problems." DOI:10.1214/14-AOS1210SUPP.

ZOU, C., LIU, Y., QIN, P. and WANG, Z. (2007). Empirical likelihood ratio test for the change-point problem. *Statist. Probab. Lett.* **77** 374–382. MR2339041

C. ZOU
Z. WANG
INSTITUTE OF STATISTICS
NANKAI UNIVERSITY
TIANJIN 300071
CHINA
E-MAIL: nk.chlzou@gmail.com
zjwang@nankai.edu.cn

G. YIN
DEPARTMENT OF STATISTICS
AND ACTUARIAL SCIENCE
UNIVERSITY OF HONG KONG
HONG KONG
E-MAIL: gyin@hku.hk

L. FENG
SCHOOL OF MATHEMATICAL SCIENCES
NANKAI UNIVERSITY
TIANJIN 300071
CHINA
E-MAIL: flnankai@126.com