SUPPLEMENTARY MATERIAL FOR "CONSISTENT SELECTION OF THE NUMBER OF CHANGE-POINTS VIA SAMPLE-SPLITTING"

By Changliang Zou[*] Guanghui Wang[†], and Runze Li[‡]

Nankai University[*†] *and The Pennsylvania State University*[‡]

This supplementary material contains the proofs of all the technical lemmas and additional simulation results.

S.1. Proof of Lemma 3. First note that

$$\max_{1\le k_1<k_2\le N}(k_2-k_1)^{1/2}\|\bar{\breve{\mathbf{U}}}_{k_1,k_2}\| \ge \max_{1\le j\le N_d} d_N^{1/2}\|\bar{\breve{\mathbf{U}}}_{(j-1)d_N,jd_N}\|,$$

where d_N is chosen to be a sequence of integers diverging to infinity and $N_d=[N/d_N]$. Let $m_N=\epsilon(\log N)^{1/2}$ for a sufficiently small ϵ. We have

$$\begin{aligned}&\Pr\left\{\max_{1\le k_1<k_2\le N}(k_2-k_1)^{1/2}\|\bar{\breve{\mathbf{U}}}_{k_1,k_2}\|\le m_N\right\}\\ \le&\Pr\left\{\max_{1\le j\le N_d} d_N^{1/2}\|\bar{\breve{\mathbf{U}}}_{(j-1)d_N,jd_N}\|\le m_N\right\}\\ =&\prod_{1\le j\le N_d}\Pr\left\{d_N^{1/2}\|\bar{\breve{\mathbf{U}}}_{(j-1)d_N,jd_N}\|\le m_N\right\}\\ \equiv&\prod_{1\le j\le N_d}F_N^{(j)}(m_N).\end{aligned}$$

By the Berry-Esseen Theorem, we have, for each $1\le j\le N_d$, $\sup_x|F_N^{(j)}(x)-\Phi(x)|\le Cd_N^{-1/2}\mathrm{E}(\|\breve{\mathbf{U}}_1\|^3)$ for some constant C, where $\Phi(x)$ is the CDF of the standard normal distribution. Hence we conclude that

$$\begin{aligned}\prod_{1\le j\le N_d}F_N^{(j)}(m_N)&\le\left\{Cd_N^{-1/2}\mathrm{E}(\|\breve{\mathbf{U}}_1\|^3)+\Phi(m_N)\right\}^{N_d}\\ &\le\left\{1-\frac{m_N}{1+m_N^2}\phi(m_N)+C\mathrm{E}(\|\breve{\mathbf{U}}_1\|^3)d_N^{-1/2}\right\}^{N_d},\end{aligned}$$

[*]Supported by NNSF of China Grants (11690015 11622104 and 11431006), and NSF of Tianjin 18JCJQJC46000.

[†]Wang is the corresponding author.

[‡]Supported by NIH grants P50 DA039838, U19AI089672 and T32 LM012415, and a NSF grant DMS 1820702. Li is the corresponding author.

where the last inequality dues to the fact that $\Phi(x) < 1 - \frac{x}{1+x^2}\phi(x)$ for $x > 0$ with $\phi(x)$ being the PDF of the standard normal distribution. It can be easily verified that

$$\left\{1 - \frac{m_N}{1+m_N^2}\phi(m_N) + C\mathrm{E}(\|\breve{\mathbf{U}}_1\|^3)d_N^{-1/2}\right\}^{N_d} \to 0,$$

as $N \to \infty$, if d_N and ϵ are appropriately chosen; for instance, $d_N = [N^{2/3}]$ and $\epsilon < \sqrt{2/3}$. Hence the conclusion follows. □

S.2. Proof of Lemma 4. (i) Recall that $\rho_n = \underline{\lambda}_n/4$. If $L < K_n$, then there must exist some j $(1 \le j \le K_n)$ such that the interval $[\tau_j^* - \rho_n, \tau_j^* + \rho_n]$ contains no estimated change-points. We then introduce $K_n + 1$ intervals $\mathcal{I}_k^*$ for $k = 0, \ldots, K_n$, to wit, $\mathcal{I}_k^* = (\tau_k^*, \tau_{k+1}^*)$ for $k \notin \{j-1, j\}$, $\mathcal{I}_{j-1}^* = (\tau_{j-1}^*, \tau_j^* - \rho_n)$ and $\mathcal{I}_j^* = (\tau_j^* + \rho_n, \tau_{j+1}^*)$. Let $\#_{L,k}$ be the number of estimated change-points in the interval $\mathcal{I}_k^*$ (wlog, assume $\widehat{\mathcal{T}}_L \cap \mathcal{T}_{K_n}^* = \emptyset$), and thus we have $\sum_{k=0}^{K_n} \#_{L,k} = L$. Moreover, let $\widehat{\mathcal{T}}_{L,k}$ be the set of the corresponding estimated change-points in $\mathcal{I}_k^*$, and if $\#_{L,k} > 0$, we write $\widehat{\mathcal{T}}_{L,k} = (\widehat{\tau}_{L,k,1}, \ldots, \widehat{\tau}_{L,k,\#_{L,k}})$. Further denote $N_{L,k,l} = \widehat{\tau}_{L,k,l+1} - \widehat{\tau}_{L,k,l}$ for $l = 0, \ldots, \#_{L,k}$ with the convention of $\widehat{\tau}_{L,k,0} = a_k^*$ and $\widehat{\tau}_{L,k,\#_{L,k}+1} = b_k^*$, where a_k^*, b_k^* are the endpoints of interval $\mathcal{I}_k^*$.

Tedious but straightforward calculation yields that

$$\begin{aligned}
&\mathcal{S}_{\mathbf{E}}^2\{\widehat{\mathcal{T}}_L \cup \mathcal{T}_{K_n}^* \backslash \tau_j^* \cup (\tau_j^* - \rho_n, \tau_j^* + \rho_n)\} - \mathcal{S}_{\mathbf{E}}^2(\mathcal{T}_{K_n}^*) \\
&= -\sum_{k=0}^{K_n}\{\mathcal{R}_{\mathbf{E}}^2(a_k^*, b_k^*) - \mathcal{R}_{\mathbf{E}}^2(a_k^*, \widehat{\mathcal{T}}_{L,k}, b_k^*)\} \\
&\quad - \|\tilde{\mathbf{V}}_{\tau_{j-1}^*, \tau_j^*}^{\tau_j^* - \rho_n}\|_{\mathbf{W}_n}^2 - \|\tilde{\mathbf{V}}_{\tau_j^*, \tau_{j+1}^*}^{\tau_j^* + \rho_n}\|_{\mathbf{W}_n}^2 + \|\tilde{\mathbf{E}}_{\tau_j^* - \rho_n, \tau_j^* + \rho_n}^{\tau_j^*}\|_{\mathbf{W}_n}^2. \qquad \text{(S.1)}
\end{aligned}$$

It follows from (9) in Assumption 3 that

$$\begin{aligned}
\|\tilde{\mathbf{E}}_{\tau_j^* - \rho_n, \tau_j^* + \rho_n}^{\tau_j^*}\|_{\mathbf{W}_n}^2 &= \|\tilde{\boldsymbol{\theta}}_{\tau_j^* - \rho_n, \tau_j^* + \rho_n}^{\tau_j^*} + \tilde{\mathbf{V}}_{\tau_j^* - \rho_n, \tau_j^* + \rho_n}^{\tau_j^*}\|_{\mathbf{W}_n}^2 \\
&\ge \frac{\underline{\lambda}_n}{8}\underline{\omega}_n \min_{1 \le j \le K_n} \|\boldsymbol{\mu}_{j-1}^* - \boldsymbol{\mu}_j^*\|^2\{1 + o_p(1)\},
\end{aligned}$$

since

$$\begin{aligned}
\|\tilde{\boldsymbol{\theta}}_{\tau_j^* - \rho_n, \tau_j^* + \rho_n}^{\tau_j^*}\|_{\mathbf{W}_n}^2 &= \frac{\rho_n}{2}\|\boldsymbol{\mu}_{j-1}^* - \boldsymbol{\mu}_j^*\|_{\mathbf{W}_n}^2 \ge \frac{\underline{\lambda}_n}{8}\underline{\omega}_n\|\boldsymbol{\mu}_{j-1}^* - \boldsymbol{\mu}_j^*\|^2, \\
0 \le \|\tilde{\mathbf{V}}_{\tau_j^* - \rho_n, \tau_j^* + \rho_n}^{\tau_j^*}\|_{\mathbf{W}_n}^2 &\le \overline{\omega}_n\|\tilde{\mathbf{V}}_{\tau_j^* - \rho_n, \tau_j^* + \rho_n}^{\tau_j^*}\|^2 = O_p(\overline{\omega}_n\overline{\sigma}),
\end{aligned}$$

and

$$|(\tilde{\boldsymbol{\theta}}^{\tau_j^*}_{\tau_j^*-\rho_n,\tau_j^*+\rho_n})^\top \mathbf{W}_n \tilde{\mathbf{V}}^{\tau_j^*}_{\tau_j^*-\rho_n,\tau_j^*+\rho_n}| \le \|\tilde{\boldsymbol{\theta}}^{\tau_j^*}_{\tau_j^*-\rho_n,\tau_j^*+\rho_n}\|_{\mathbf{W}_n} \cdot \|\tilde{\mathbf{V}}^{\tau_j^*}_{\tau_j^*-\rho_n,\tau_j^*+\rho_n}\|_{\mathbf{W}_n}.$$

Similar, we can show $\|\tilde{\mathbf{V}}^{\tau_j^*-\rho_n}_{\tau_{j-1}^*,\tau_j^*}\|^2_{\mathbf{W}_n} = O_p(\overline{\omega}_n\overline{\sigma})$ and $\|\tilde{\mathbf{V}}^{\tau_j^*+\rho_n}_{\tau_j^*,\tau_{j+1}^*}\|^2_{\mathbf{W}_n} = O_p(\overline{\omega}_n\overline{\sigma})$.

For the first term of Eq. (S.1), we can show if $\#_{L,k} > 0$,

$$0 \le \mathcal{R}^2_{\mathbf{E}}(a_k^*, b_k^*) - \mathcal{R}^2_{\mathbf{E}}(a_k^*, \widehat{\mathcal{T}}_{L,k}, b_k^*) = \sum_{0\le l_1<l_2\le \#_{L,k}} \frac{N_{L,k,l_1}+N_{L,k,l_2}}{\tau_k^*-\tau_{k-1}^*}\|\tilde{\mathbf{V}}_{L,k,l_1,l_2}\|^2_{\mathbf{W}_n},$$

where

$$\tilde{\mathbf{V}}_{L,k,l_1,l_2} = \sqrt{\frac{N_{L,k,l_1}N_{L,k,l_2}}{N_{L,k,l_1}+N_{L,k,l_2}}}\Big(\bar{\mathbf{V}}_{\widehat{\tau}_{L,k,l_1},\widehat{\tau}_{L,k,l_1+1}} - \bar{\mathbf{V}}_{\widehat{\tau}_{L,k,l_2},\widehat{\tau}_{L,k,l_2+1}}\Big).$$

By Markov's inequality and by noting the fact that $\sum_{0\le l_1<l_2\le \#_{L,k}} \frac{N_{L,k,l_1}+N_{L,k,l_2}}{\tau_k^*-\tau_{k-1}^*} = \#_{L,k}$, we have, uniformly for any $\widehat{\mathcal{T}}_L$,

$$\begin{aligned}
0 &\le \sum_{k=0}^{K_n}\{\mathcal{R}^2_{\mathbf{E}}(a_k^*, b_k^*) - \mathcal{R}^2_{\mathbf{E}}(a_k^*, \widehat{\mathcal{T}}_{L,k}, b_k^*)\} \\
&\le \overline{\omega}_n \sum_{\substack{k=0,\dots,K_n \\ \#_{L,k}>0}} \sum_{0\le l_1<l_2\le \#_{L,k}} \frac{N_{L,k,l_1}+N_{L,k,l_2}}{\tau_k^*-\tau_{k-1}^*}\|\tilde{\mathbf{V}}_{L,k,l_1,l_2}\|^2 \\
&= O_p(K_n\overline{\omega}_n\overline{\sigma}).
\end{aligned}$$

By combining the above facts together, Lemma 4 (i) follows.

(ii) Obvious. (iii) A by-product of the proof of Lemma 4 (i). (iv) By assumption 2 and by using arguments similar to that in Yao and Au (1989), we can show that $\mathcal{S}^2_{\mathbf{E}}(\widehat{\mathcal{T}}_{K_n}) - \mathcal{S}^2_{\mathbf{E}}(\widehat{\mathcal{T}}_{K_n}\backslash\hat{\tau}_{K_n,i_j}\cup\tau_j^*) = o_p(\overline{\omega}_n\overline{\sigma}\log\log\overline{\lambda}_n)$ for some $j = 1,\dots,K_n$. Hence the conclusion follows. □

S.3. Proof of Lemma 5. (i) **Bounding the term $\mathcal{S}^2_{\mathbf{U}}(\mathcal{T}^*_{K_n}) - \mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_L \cup \mathcal{T}^*_{K_n})$:** If $L < K_n$, we let $\#_{L,j}$ be the number of estimated change-points in the interval (τ_j^*, τ_{j+1}^*) for $j = 0,\dots,K_n$ (wlog, assume $\widehat{\mathcal{T}}_L \cap \mathcal{T}^*_{K_n} = \emptyset$), and thus we have $\sum_{j=0}^{K_n}\#_{L,j} = L$. Moreover, let $\widehat{\mathcal{T}}_{L,j}$ be the set of the corresponding estimated change-points in (τ_j^*, τ_{j+1}^*), and if $\#_{L,j} > 0$, we write $\widehat{\mathcal{T}}_{L,j} = (\widehat{\tau}_{L,j,1},\dots,\widehat{\tau}_{L,j,\#_{L,j}})$. Further denote $N_{L,j,l} = \widehat{\tau}_{L,j,l+1} - \widehat{\tau}_{L,j,l}$ for $l = 0,\dots,\#_{L,j}$ with the convention of $\widehat{\tau}_{L,j,0} = \tau_j^*$ and $\widehat{\tau}_{L,j,\#_{L,j}+1} = \tau_{j+1}^*$.

It is straightforward to show that

$$0 \le \mathcal{S}^2_{\mathbf{U}}(\mathcal{T}^*_{K_n}) - \mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_L \cup \mathcal{T}^*_{K_n}) = \sum_{\substack{0\le j\le K_n\\ \#_{L,j}>0}} \sum_{0\le l_1<l_2\le \#_{L,j}} \frac{N_{L,j,l_1}+N_{L,j,l_2}}{\tau^*_j - \tau^*_{j-1}} \|\tilde{\mathbf{U}}_{L,j,l_1,l_2}\|^2_{\mathbf{W}_n},$$

$$\le \overline{\omega}_n \sum_{\substack{0\le j\le K_n\\ \#_{L,j}>0}} \sum_{k=0}^{\#_{L,j}} (\widehat{\tau}_{L,j,k+1} - \widehat{\tau}_{L,j,k}) \|\bar{\mathbf{U}}_{\widehat{\tau}_{L,j,k},\widehat{\tau}_{L,j,k+1}}\|^2,$$

where $\widehat{\tau}_{L,j,0} = \tau^*_j$ and $\widehat{\tau}_{L,j,\#_{L,j}+1} = \tau^*_{j+1}$. We observe that,

$$\sum_{\substack{0\le j\le K_n\\ \#_{L,j}>0}} \sum_{k=0}^{\#_{L,j}} (\widehat{\tau}_{L,j,k+1} - \widehat{\tau}_{L,j,k}) \|\bar{\mathbf{U}}_{\widehat{\tau}_{L,j,k},\widehat{\tau}_{L,j,k+1}}\|^2$$

$$\le \text{sum of several (at most } L) \max_{\tau^*_j < a < \tau^*_{j+1}} (\tau^*_{j+1} - a)\|\bar{\mathbf{U}}_{a,\tau^*_{j+1}}\|^2 s$$

$$+ \text{sum of several (at most } L) \max_{\tau^*_j < b < \tau^*_{j+1}} (b - \tau^*_j)\|\bar{\mathbf{U}}_{\tau^*_j,b}\|^2 s$$

$$+ \text{sum of several (at most } L-1) \max_{\tau^*_j < a < b < \tau^*_{j+1}} (b - a)\|\bar{\mathbf{U}}_{a,b}\|^2 s.$$

By Lemmas 1–2, we have

$$\max_{\tau^*_j < a < b < \tau^*_{j+1}} (b-a)\|\bar{\mathbf{U}}_{a,b}\|^2 \le \overline{\sigma} \max_{\tau^*_j < a < b < \tau^*_{j+1}} (b-a)\|\bar{\breve{\mathbf{U}}}_{a,b}\|^2 = O_p(\overline{\sigma}\overline{\lambda}_n^{2/m}),$$

$$\max_{\tau^*_j < a < \tau^*_{j+1}} (\tau^*_{j+1} - a)\|\bar{\mathbf{U}}_{a,\tau^*_{j+1}}\|^2 \le \overline{\sigma} \max_{\tau^*_j < a < \tau^*_{j+1}} (\tau^*_{j+1} - a)\|\bar{\breve{\mathbf{U}}}_{a,\tau^*_{j+1}}\|^2 = O_p(\overline{\sigma}\log\log\overline{\lambda}_n).$$

Hence we conclude that $\mathcal{S}^2_{\mathbf{U}}(\mathcal{T}^*_{K_n}) - \mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_L \cup \mathcal{T}^*_{K_n}) = O_p(K_n\overline{\omega}_n\overline{\sigma}\overline{\lambda}_n^{2/m})$ uniformly for any $\widehat{\mathcal{T}}_L$.

Bounding the term $\mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_L) - \mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_L \cup \mathcal{T}^*_{K_n})$: Let $\#_{L,l}$ be the number of true change-points in the interval $(\widehat{\tau}_{L,l}, \widehat{\tau}_{L,l+1})$ for $l = 0,\dots,L$ (wlog, assume $\widehat{\mathcal{T}}_L \cap \mathcal{T}^*_{K_n} = \emptyset$). Moreover, let $\mathcal{T}^*_{L,l}$ be the set of the corresponding true change-points in $(\widehat{\tau}_{L,l}, \widehat{\tau}_{L,l+1})$, and if $\#_{L,l} > 0$, we write $\mathcal{T}^*_{L,l} = (\tau^*_{L,l,1},\dots,\tau^*_{L,l,\#_{L,l}})$. Further denote $N_{L,l,j} = \tau^*_{L,l,j+1} - \tau^*_{L,l,j}$ for $j = 0,\dots,\#_{L,l}$ with the convention of $\tau^*_{L,l,0} = \widehat{\tau}_{L,l}$ and $\tau^*_{L,l,\#_{L,l}+1} = \widehat{\tau}_{L,l+1}$. Using arguments similar above, we can show that

$$
\begin{aligned}
0 &\le \mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_L) - \mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_L \cup \mathcal{T}^*_{K_n}) \\
&\le \overline{\omega}_n \sum_{\substack{0\le l\le L \\ \#_{L,l}>0}} \sum_{j=0}^{\#_{L,l}} (\tau^*_{L,l,j+1} - \tau^*_{L,l,j}) \|\bar{\mathbf{U}}_{\tau^*_{L,l,j},\tau^*_{L,l,j+1}}\|^2 \\
&\le \overline{\omega}_n \{\text{sum of several (at most } K_n+1)\ (\tau^*_{j+1}-\tau^*_j)\|\bar{\mathbf{U}}_{\tau^*_j,\tau^*_{j+1}}\|^2 s \\
&\quad + \text{sum of several (at most } K_n+1) \max_{\tau^*_j < a < \tau^*_{j+1}} (\tau^*_{j+1} - a)\|\bar{\mathbf{U}}_{a,\tau^*_{j+1}}\|^2 s \\
&\quad + \text{sum of several (at most } K_n+1) \max_{\tau^*_j < b < \tau^*_{j+1}} (b - \tau^*_j)\|\bar{\mathbf{U}}_{\tau^*_j,b}\|^2 s\} \\
&= O_p(K_n \overline{\omega}_n \overline{\sigma} \log\log \overline{\lambda}_n).
\end{aligned}
$$

(ii) To bound the term $\mathcal{S}^2_{\mathbf{U}}(\mathcal{T}^*_{K_n}) - \mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_{K_n} \cup \mathcal{T}^*_{K_n})$, we first introduce $\mathcal{T}^\delta_{2K_n} = \{\tau^*_j \pm \delta_{q,n},\ j = 1, \dots, K_n\}$. Then by using arguments similar to those when $L < K_n$, we have, uniformly for any $\widehat{\mathcal{T}}_{K_n}$,

$$
\begin{aligned}
&\mathcal{S}^2_{\mathbf{U}}(\mathcal{T}^*_{K_n}) - \mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_{K_n} \cup \mathcal{T}^*_{K_n}) \\
&= \{\mathcal{S}^2_{\mathbf{U}}(\mathcal{T}^*_{K_n}) - \mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_{K_n} \cup \mathcal{T}^*_{K_n} \cup \mathcal{T}^\delta_{2K_n})\} - \{\mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_{K_n} \cup \mathcal{T}^*_{K_n}) - \mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_{K_n} \cup \mathcal{T}^*_{K_n} \cup \mathcal{T}^\delta_{2K_n})\} \\
&= O_P(K_n \overline{\omega}_n \overline{\sigma}) + O_p(K_n \overline{\omega}_n \overline{\sigma} \log\log \delta_{q,n}).
\end{aligned}
$$

Bounding the term $\mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_{K_n}) - \mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_{K_n} \cup \mathcal{T}^*_{K_n})$ is much simpler than showing (iii) and thus the proof is omitted here.

(iii) Note that

$$
\begin{aligned}
&\mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_L) - \mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_L \cup \mathcal{T}^*_{K_n}) \\
&= \{\mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_L) - \mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_L \cup \mathcal{T}^*_{K_n} \cup \mathcal{T}^\delta_{2K_n})\} - \{\mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_L \cup \mathcal{T}^*_{K_n}) - \mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_L \cup \mathcal{T}^*_{K_n} \cup \mathcal{T}^\delta_{2K_n})\}.
\end{aligned}
$$

We first bound the term $\mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_L) - \mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_L \cup \mathcal{T}^*_{K_n} \cup \mathcal{T}^\delta_{2K_n})$. For each $l = 0, \dots, L$, we introduce $\mathcal{T}^{*\delta}_{L,l} = \{\tau \in \mathcal{T}^*_{K_n} \cup \mathcal{T}^\delta_{2K_n} : \widehat{\tau}_{L,l} < \tau < \widehat{\tau}_{L,l+1}\}$. Also write for short $D = \mathcal{S}^2_{\mathbf{U}}(\widehat{\tau}_{L,l}, \widehat{\tau}_{L,l+1}) - \mathcal{S}^2_{\mathbf{U}}(\widehat{\tau}_{L,l}, \mathcal{T}^{*\delta}_{L,l}, \widehat{\tau}_{L,l+1})$ and $\delta_n \equiv \delta_{q,n}$. If

(a) $\mathcal{T}^{*\delta}_{L,l} = (\tau^*_j, \tau^*_j + \delta_n, \tau^*_{j+1} - \delta_n, \tau^*_{j+1})$, we can similarly show $D = O_p(\overline{\omega}_n \overline{\sigma}) + O_p(\overline{\omega}_n \overline{\sigma} \log\log \delta_n)$ for any $\widehat{\tau}_{L,l}$ and $\widehat{\tau}_{L,l+1}$. The number of such terms is at most $K_n/2$.

(b) $\mathcal{T}^{*\delta}_{L,l} = (\tau^*_j, \tau^*_j + \delta_n, \tau^*_{j+1} - \delta_n)$, we also have $D = O_p(\overline{\omega}_n \overline{\sigma}) + O_p(\overline{\omega}_n \overline{\sigma} \log\log \delta_n)$ for any $\widehat{\tau}_{L,l}$ and $\widehat{\tau}_{L,l+1}$. The number of such terms is at most K_n.

(c) $\mathcal{T}^{*\delta}_{L,l} = (\tau^*_j, \tau^*_j + \delta_n)$, we have, for any $\widehat{\tau}_{L,l}$ and $\widehat{\tau}_{L,l+1}$,

$$
D \le O_p(\overline{\omega}_n \overline{\sigma}) + O_p(\overline{\omega}_n \overline{\sigma} \log\log \delta_n) + 2\overline{\omega}_n \frac{N_{\widehat{\tau}_{L,l},\tau^*_j+\delta_n}}{N_{\widehat{\tau}_{L,l},\widehat{\tau}_{L,l+1}}} N_{\tau^*_j+\delta_n,\widehat{\tau}_{L,l+1}} \|\bar{\mathbf{U}}_{\tau^*_j+\delta_n,\widehat{\tau}_{L,l+1}}\|^2.
$$

For any $\widehat{\tau}_{L,l}$ and $\widehat{\tau}_{L,l+1}$, denote $M = \widehat{\tau}_{L,l+1} - (\tau_j^* + \delta_n)$ for short. If $M = O_p(\delta_n)$, it is straightforward to show that

$$\frac{N_{\widehat{\tau}_{L,l},\tau_j^*+\delta_n}}{N_{\widehat{\tau}_{L,l},\widehat{\tau}_{L,l+1}}} N_{\tau_j^*+\delta_n,\widehat{\tau}_{L,l+1}} \|\bar{\mathbf{U}}_{\tau_j^*+\delta_n,\widehat{\tau}_{L,l+1}}\|^2 = O_p(\overline{\sigma} \log\log \delta_n).$$

On the other hand, if M/δ_n diverges to infinity with probability approaching one, then we have

$$\frac{N_{\widehat{\tau}_{L,l},\tau_j^*+\delta_n}}{N_{\widehat{\tau}_{L,l},\widehat{\tau}_{L,l+1}}} N_{\tau_j^*+\delta_n,\widehat{\tau}_{L,l+1}} \|\bar{\mathbf{U}}_{\tau_j^*+\delta_n,\widehat{\tau}_{L,l+1}}\|^2 = o_p(\overline{\sigma} \log\log \overline{\lambda}_n).$$

Moreover, the number of such terms is at most $q = L - K_n$.

(d) $\mathcal{T}_{L,l}^{*\delta} = \tau_j^*$, we also have $D = O_p(\overline{\omega}_n \overline{\sigma} \log\log \delta_n)$ for any $\widehat{\tau}_{L,l}$ and $\widehat{\tau}_{L,l+1}$. The number of such terms is at most K_n.

(e) other scenarios are considered, the problem is reduced to that under the above ones.

To summarize, we conclude, for any $\widehat{\mathcal{T}}_L$,

$$\begin{aligned}&\mathcal{S}_{\mathbf{U}}^2(\widehat{\mathcal{T}}_L) - \mathcal{S}_{\mathbf{U}}^2(\widehat{\mathcal{T}}_L \cup \mathcal{T}_{K_n}^* \cup \mathcal{T}_{2K_n}^{\delta})\\ &= \overline{\omega}_n \overline{\sigma}\{o_p(\log\log \underline{\lambda}_n) + O_p(K_n \log\log \delta_{q,n}) + O_p(K_n)\}.\end{aligned}$$

Similar, we can show $\mathcal{S}_{\mathbf{U}}^2(\widehat{\mathcal{T}}_L \cup \mathcal{T}_{K_n}^*) - \mathcal{S}_{\mathbf{U}}^2(\widehat{\mathcal{T}}_L \cup \mathcal{T}_{K_n}^* \cup \mathcal{T}_{2K_n}^{\delta}) = \overline{\omega}_n \overline{\sigma}\{o_p(\log\log \underline{\lambda}_n) + O_p(K_n \log\log \delta_{q,n}) + O_p(K_n)\}$. Hence Lemma 5 (iii) follows. □

S.4. Additional simulation results. Table S.1 reports the distribution of $\hat{K}_n - K_n$ together with its mean, standard deviation (SD) and mean-squared error (MSE) for the BIC and the COPSS in conjunction with various detection algorithms under Model I-CP(B) with Scenario (i). Except for the former four algorithms mentioned, we also consider the SaRa algorithm here, which involves a localizing window size h. We adopt Niu and Zhang (2012)'s recommendation $h = 0.75\lambda_n$ with λ_n being the minimum distance of two successive change-points while it is usually unknown in practice. The results of the SaRa under Model I-CP(A) is also listed below as Table S.2. From Tables S.1–S.2, again, we can see that most popular change detection approaches are sensitive to the specification of the penalization magnitude, and the COPSS is much more robust due to the introducing of the data-driven penalty. Figure S1 depicts the distribution of $\hat{K}_n - K_n$ under Model I-CP(B) with Scenarios (ii)–(iv), which reveals the robustness of the COPSS to different error distributions.

Table S.3 presents the distribution of $\hat{K}_n - K_n$ together with its mean, SD and MSE using the OP algorithm in conjunction with the COPSS, the traditional 5-fold and 10-fold CV criteria under Model I-CP(A) with Scenario (i), together with slight modifications of the latter two that preserve the time order. To use the traditional r-fold CV, we first randomly assign a fold index (from $1, \ldots, r$) for each observation, and then use all observations except those of index $k \in \{1, \ldots, r\}$, to train the change-point model, and those of index k to calculate the test error. Then we run over all possible fold indices for the validation. As to the modified r-fold CV in order to preserve the original time order, instead of randomly assigning a fold index (from $1, \ldots, r$) for each observation, we code a fold index k for the ith observation such that $i\%\%r = k$ where %% is the modulus operator. From this Table, we can see that the traditional or the modified multi-fold CV could also be consistent as reveled in Remark 4 in our paper; while our current parity splitting strategy could perform better sometime as it makes the training and validation sets the most similar.

As suggested by an anonymous referee, we also consider using our 2-fold splitting strategy with random assignments (RA). To be more specifically, for each pair $\{\mathcal{Z}_{2t-1}, \mathcal{Z}_{2t}\},\ t = 1, \ldots, T$, we randomly pick up one which goes to $\mathcal{Z}_O$, and the left goes to $\mathcal{Z}_E$. Tables S.4–S.5 present the distribution of $\hat{K}_n - K_n$ together with its mean, standard deviation (SD) and mean-squared error (MSE) by using our current 2-fold procedure and the one with RA under Model I-CP(A) with Scenario (i) the iid normal noises and Scenario (v) an AR(1) sequence with coefficient 0.3 and normal innovations. We can see from Table S.4 that our current 2-fold procedure is slightly better than the one with RA in most cases. With Scenario (v) (Table S.5), our current strategy yields two correlated data sets $\mathcal{Z}_O$ and $\mathcal{Z}_E$, which may affect the performance; while it seems that the random assignment strategy may mitigate the effect as it could reduce the variation. It is of interest to further investigate it in the future.

Figure S2 reports how the run time (in seconds) changes with the sample size $n = C_n \cdot 2048$ of both procedures under Model III-(A) with Scenario (i) for one replication using an Inter Xeon E5-2650v4 CPU. As we can see, our method is significantly faster and the advantage is more prominent as n increases.

At last, we consider an extension of our procedure to some non-stationary sequences (within each segment) with local structures, i.e., change-point detection in nonparametric regression; as mentioned in the Concluding Remarks in our paper. Consider here a simulated model, $Y_i = f(i/n) + 0.2\varepsilon_i,\ i =$

$1, \ldots, n$ with $n = 1000$, where

$$f(t) = \begin{cases} -3t + 2, & 0 \le t \le t_1^*, \\ -3t + 2.75 - \sin\left\{\frac{2\pi(t-t_1^*)}{t_2^*-t_1^*}\right\}, & t_1^* < t \le t_2^*, \\ t/2 + 3.5 - 7t_2^*/2, & t_2^* < t \le 1, \end{cases}$$

where $\varepsilon_i \overset{\text{iid}}{\sim} N(0,1)$. The goal is to identify the change-points t_1^* and t_2^*. Three cases for the choices of change-point locations are considered here: (i) $t_1^* = 0.3$ and $t_2^* = 0.7$ as used in Xia and Qiu (2015), (ii) $t_1^* = 0.45$ and $t_2^* = 0.55$, and (iii) $t_1^* = 0.25$ and $t_2^* = 0.5$. The sequence is naturally non-stationary with local structures in the sense that even in the same segment it has different pointwise mean. Xia and Qiu (2015) suggested a Jump Information Criterion (JIC) to choose the number of change-points (here 2), which may be slightly affected by the magnitude of their penalty. Here we apply our specialized 2-fold splitting strategy in the following way. For a given number of change-points, first we use the odd sample to train the model by the kernel-smoothing estimation method in Xia and Qiu (2015), then we calculate the test error using the even sample. Finally, we run over a sequence of candidate number of change-points and find the one which minimizes the test error. We also do the validation by altering the odd and even samples. The bandwidth in their method is chosen as $h = 0.3n^{-1/5}$ suggested by the authors, and in our CV implementation we use half of the h. The results are summarized in Table S.6. We report the distribution of $\hat{K}_n - K_n$ together with its mean, standard deviation (SD) and mean-squared error (MSE). For the JIC proposed by Xia and Qiu (2015), we consider three penalties, termed JIC-S, -M and -L (small, medium and large, respectively) as suggested by the authors. From Table S.6, we can see that our procedure performs reasonably well under all cases, and outperforms the JIC in cases (ii)–(iii). It offers us a data-driven way instead of selecting the best magnitude of the penalty, and hence could be much more robust.

References.

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

Xia, Z. and Qiu, P. (2015). Jump information criterion for statistical inference in estimating discontinuous curves. *Biometrika* **102** 397–408.

Yao, Y.-C. and Au, S. T. (1989). Least-squares estimation of a step function. *Sankhya A* **51** 370–381.

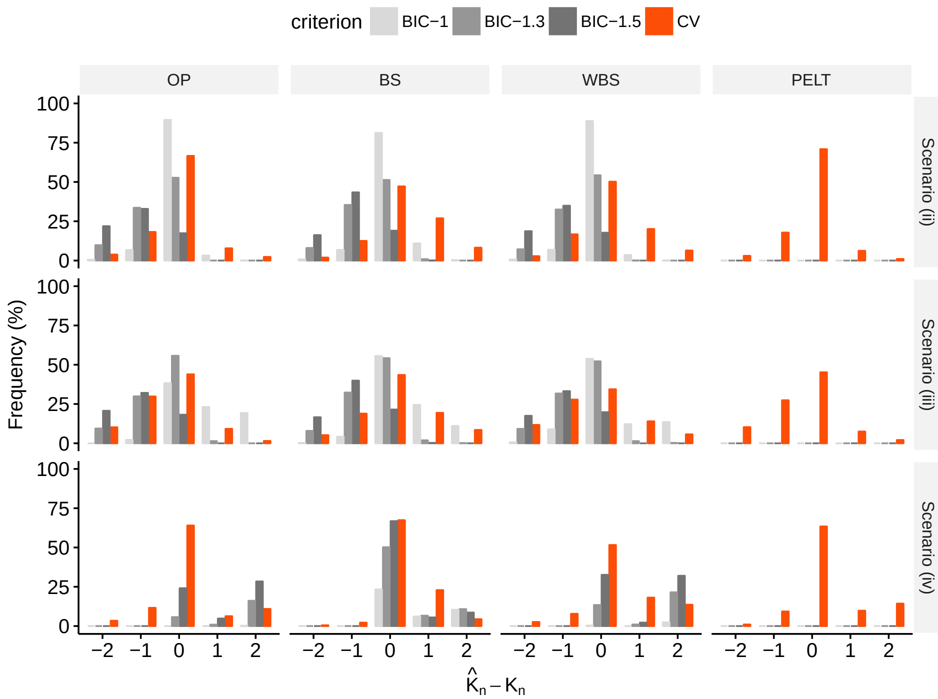

FIG S1. *Distribution of $\hat{K}_n - K_n$ for the BIC and our CV criteria in conjunction with the OP, BS, WBS and PELT algorithms under Scenarios (ii)–(iv) of Model I-CP(B).*

Changliang Zou, Guanghui Wang
Institute of Statistics and LPMC
Nankai University
Tianjin 300071, China
E-mail: nk.chlzou@gmail.com
ghwang.nk@gmail.com

Runze Li
Department of Statistics
and The Methodology Center
The Pennsylvania State University
University Park, PA 16802-2111, USA
E-mail: rzli@psu.edu

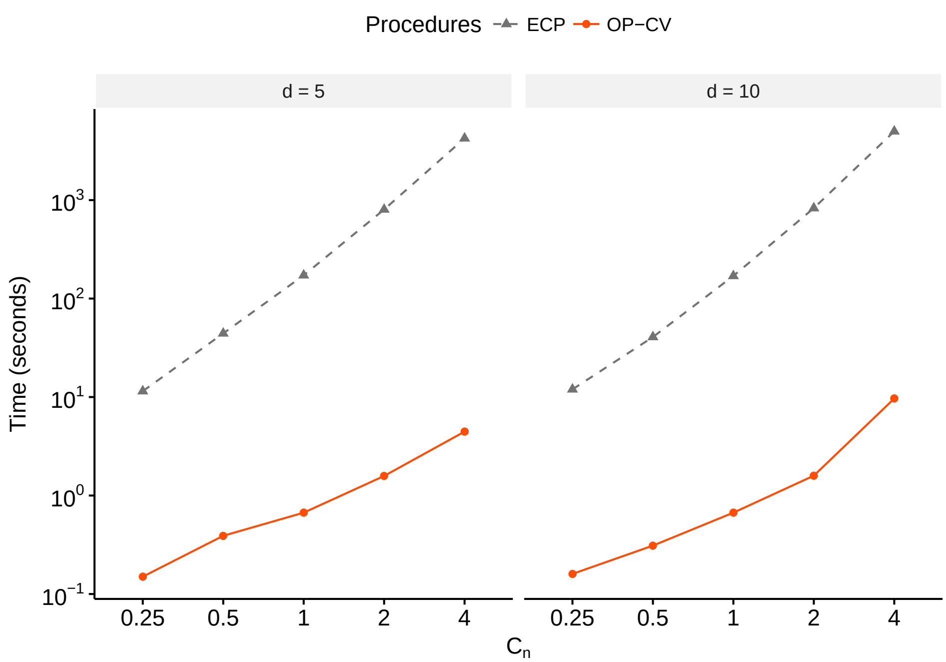

FIG S2. *Computation time of both detection procedures against the sample size* $n = C_n \cdot 2048$ *under Scenario (i) of Model III-CP(A) from one replication. The y-axis is in* $\log_{10}$*-scale.*

Table S.1
Distribution of $\hat{K}_n - K_n$ together with its mean, standard deviation (SD) and mean-squared error (MSE) using various detection algorithms under Model I. Scenario (i) and CP(B) are considered.

	$\hat{K}_n - K_n$									
Procedures	≤ -3	-2	-1	0	1	2	≥ 3	mean	SD	MSE
OP-BIC-1	0.0	0.4	5.5	89.8	4.1	0.2	0.0	0.11	0.35	0.12
OP-BIC-1.3	4.1	11.2	29.5	55.2	0.0	0.0	0.0	0.64	0.84	1.12
OP-BIC-1.5	29.8	21.3	31.6	17.3	0.0	0.0	0.0	1.71	1.23	4.43
OP-CV-O	0.5	5.1	20.2	59.6	10.2	2.6	1.8	0.54	0.95	0.91
OP-CV-E	1.4	5.2	18.7	62.4	8.3	3.2	0.8	0.51	0.92	0.87
OP-CV	0.6	4.9	18.9	66.3	8.0	1.0	0.3	0.41	0.74	0.59
BS-BIC-1	0.0	0.5	5.9	79.5	13.0	1.1	0.0	0.22	0.50	0.25
BS-BIC-1.3	4.2	10.6	31.2	52.7	1.3	0.0	0.0	0.66	0.85	1.13
BS-BIC-1.5	20.8	19.5	40.6	19.0	0.1	0.0	0.0	1.52	1.25	3.88
BS-CV-O	0.3	2.5	15.1	45.7	22.2	6.9	7.3	0.86	1.35	2.02
BS-CV-E	0.4	3.6	13.4	47.2	22.2	8.5	4.7	0.80	1.28	1.78
BS-CV	0.2	2.2	11.6	48.9	26.7	7.4	3.0	0.68	1.01	1.15
WBS-BIC-1	0.0	0.6	7.9	87.9	3.4	0.2	0.0	0.13	0.38	0.14
WBS-BIC-1.3	4.7	9.8	30.1	55.2	0.2	0.0	0.0	0.64	0.85	1.12
WBS-BIC-1.5	29.1	19.5	32.8	18.6	0.0	0.0	0.0	1.66	1.22	4.23
WBS-CV-O	0.3	4.0	17.7	45.0	18.7	8.3	6.0	0.86	1.36	1.95
WBS-CV-E	1.2	5.5	18.8	42.5	19.5	5.9	6.6	0.90	1.42	2.07
WBS-CV	0.3	3.5	17.1	52.3	18.6	5.5	2.7	0.64	1.02	1.06
SaRa-BIC-1	20.7	45.6	29.7	4.0	0.0	0.0	0.0	1.84	0.83	4.09
SaRa-BIC-1.3	26.0	42.5	27.7	3.8	0.0	0.0	0.0	1.94	0.88	4.53
SaRa-BIC-1.5	43.9	32.9	20.3	2.9	0.0	0.0	0.0	2.29	1.04	6.31
SaRa-CV-O	30.6	43.5	22.9	3.0	0.0	0.0	0.0	2.06	0.89	5.04
SaRa-CV-E	29.9	45.0	22.0	3.1	0.0	0.0	0.0	2.06	0.89	5.04
SaRa-CV	45.3	46.2	8.4	0.1	0.0	0.0	0.0	2.44	0.76	6.51
PELT-BIC-1	0.0	0.0	0.0	0.0	0.0	0.0	100.0	1768.38	24.10	3127733.81
PELT-BIC-1.3	0.0	0.0	0.0	0.0	0.0	0.0	100.0	1674.17	30.17	2803754.37
PELT-BIC-1.5	0.0	0.0	0.0	0.0	0.0	0.0	100.0	1594.74	34.79	2544408.08
PELT-CV-O	1.2	5.2	21.5	61.8	7.0	1.9	1.4	0.52	0.92	0.89
PELT-CV-E	2.5	4.4	19.2	64.2	6.6	2.3	0.8	0.50	0.91	0.87
PELT-CV	0.6	4.1	20.5	68.1	6.1	0.5	0.1	0.38	0.68	0.51

TABLE S.2
Distribution of $\hat{K}_n - K_n$ together with its mean, standard deviation (SD) and mean-squared error (MSE) using the SaRa algorithm under Model I. Scenario (i) and CP(A) are considered.

	$\hat{K}_n - K_n$									
Procedures	≤ -3	-2	-1	0	1	2	≥ 3	mean	SD	MSE
SaRa-BIC-1	2.4	25.4	43.7	23.8	3.6	0.9	0.2	1.08	0.92	1.77
SaRa-BIC-1.3	4.4	31.6	42.8	19.1	1.8	0.3	0.0	1.22	0.87	2.12
SaRa-BIC-1.5	10.0	38.2	39.1	12.6	0.1	0.0	0.0	1.47	0.87	2.92
SaRa-CV-O	5.0	15.5	20.9	18.9	12.7	8.3	18.7	1.97	2.74	7.91
SaRa-CV-E	2.5	11.2	22.1	20.5	14.5	9.8	19.4	1.85	2.57	7.23
SaRa-CV	0.2	8.3	21.3	24.8	16.7	10.6	18.1	1.63	2.28	5.91

TABLE S.3
Distribution of $\hat{K}_n - K_n$ together with its mean, standard deviation (SD) and mean-squared error (MSE) using the OP algorithm in conjunction with the COPSS, the traditional 5-fold and 10-fold CV criteria, together with the modifications of the latter two which preserve the time order (termed as "r-fold COPSS", $r = 5, 10$) under Model I. Scenario (i) and CP(A) are considered.

	$\hat{K}_n - K_n$										
SNR	Procedures	≤ -3	-2	-1	0	1	2	≥ 3	mean	SD	MSE
1	COPSS	0.0	0.0	24.8	66.2	7.5	1.3	0.2	0.35	0.61	0.39
	5-fold CV	0.0	0.0	8.9	75.3	11.2	2.4	2.2	0.33	0.78	0.62
	10-fold CV	0.0	0.0	5.2	75.8	12.7	4.1	2.2	0.35	0.81	0.71
	5-fold COPSS	0.0	0.0	6.3	80.5	10.1	1.9	1.2	0.25	0.67	0.46
	10-fold COPSS	0.0	0.0	4.3	75.6	13.6	3.7	2.8	0.35	0.81	0.72
2	COPSS	0.0	0.0	0.0	90.1	8.7	1.0	0.2	0.11	0.36	0.14
	5-fold CV	0.0	0.0	0.0	81.7	11.8	4.1	2.4	0.29	0.74	0.64
	10-fold CV	0.0	0.0	0.0	75.8	15.9	4.9	3.4	0.39	0.88	0.93
	5-fold COPSS	0.0	0.0	0.0	84.9	11.4	2.9	0.8	0.20	0.56	0.35
	10-fold COPSS	0.0	0.0	0.0	78.3	13.9	5.3	2.5	0.33	0.76	0.69
3	COPSS	0.0	0.0	0.0	89.7	8.5	1.6	0.2	0.12	0.39	0.17
	5-fold CV	0.0	0.0	0.0	79.0	13.4	5.2	2.4	0.33	0.80	0.75
	10-fold CV	0.0	0.0	0.0	75.0	16.6	5.2	3.2	0.39	0.84	0.86
	5-fold COPSS	0.0	0.0	0.0	82.0	13.4	3.3	1.3	0.25	0.62	0.44
	10-fold COPSS	0.0	0.0	0.0	76.7	14.8	5.4	3.1	0.38	0.88	0.91

Table S.4
Distribution of $\hat{K}_n - K_n$ together with its mean, standard deviation (SD) and mean-squared error (MSE) by using our current 2-fold strategy ("-CV") and the one with random assignment treatment ("-CV-RA") under Model I-CP(A) with Scenario (i).

	≤ -3	-2	-1	0	1	2	≥ 3	Mean	SD	MSE
OP-CV	0.0	0.0	22.6	68.3	7.5	1.2	0.4	0.34	0.61	0.39
OP-CV-RA	0.0	0.1	24.5	63.9	10.3	0.4	0.8	0.39	0.67	0.46
BS-CV	0.0	0.0	11.1	28.4	31.6	18.9	10.0	1.17	1.32	2.65
BS-CV-RA	0.0	0.2	8.8	27.4	31.3	21.3	11.0	1.23	1.30	2.77
WBS-CV	0.0	0.0	25.6	47.6	18.0	6.0	2.8	0.67	1.08	1.18
WBS-CV-RA	0.0	0.0	29.8	43.8	17.6	5.3	3.5	0.71	1.08	1.19
SaRa-CV	0.6	7.5	22.2	23.8	16.1	11.0	18.8	1.61	2.17	5.41
SaRa-CV-RA	0.8	7.4	22.5	27.0	15.1	9.5	17.7	1.54	2.19	5.33

Table S.5
Distribution of $\hat{K}_n - K_n$ together with its mean, standard deviation (SD) and mean-squared error (MSE) by using our current 2-fold strategy ("-CV") and the one with random assignment treatment ("-CV-RA") under Model I-CP(A) with Scenario (v).

	≤ -3	-2	-1	0	1	2	≥ 3	Mean	SD	MSE
OP-CV	0.0	0.0	4.6	20.8	15.5	9.7	49.4	3.36	3.15	20.58
OP-CV-RA	0.0	0.0	11.6	34.9	16.9	9.7	26.9	1.94	2.62	9.78
BS-CV	0.0	0.0	1.2	5.2	9.2	11.5	72.9	4.87	2.92	31.94
BS-CV-RA	0.0	0.0	2.9	10.8	16.7	14.0	55.6	3.50	2.78	19.57
WBS-CV	0.0	0.0	3.4	12.4	11.9	8.2	64.1	4.41	3.27	29.57
WBS-CV-RA	0.0	0.0	11.6	20.8	16.0	10.2	41.4	2.77	2.95	15.12
SaRa-CV	0.2	0.7	3.0	4.2	5.1	5.2	81.6	6.01	3.17	44.99
SaRa-CV-RA	0.1	2.0	5.1	8.8	7.2	7.8	69.0	4.83	3.39	33.05

Table S.6
Distribution of $\hat{K}_n - K_n$ together with its mean, standard deviation (SD) and mean-squared error (MSE) under the nonparametric settings.

	≤ -3	-2	-1	0	1	2	≥ 3	Mean	SD	MSE
					Case (i)					
JIC-S	0.0	0.0	0.0	100.0	0.0	0.0	0.0	0.00	0.00	0.00
JIC-M	0.0	0.0	0.1	99.9	0.0	0.0	0.0	0.00	0.03	0.00
JIC-L	0.0	0.0	61.8	38.2	0.0	0.0	0.0	0.62	0.49	0.62
CV	0.0	0.0	1.5	90.5	6.6	1.0	0.3	0.11	0.38	0.15
					Case (ii)					
JIC-S	0.0	0.0	100.0	0.0	0.0	0.0	0.0	1.00	0.00	1.00
JIC-M	0.0	0.0	100.0	0.0	0.0	0.0	0.0	1.00	0.00	1.00
JIC-L	0.0	0.0	100.0	0.0	0.0	0.0	0.0	1.00	0.00	1.00
CV	0.0	0.0	0.2	48.9	40.5	8.9	1.5	0.63	0.71	0.90
					Case (iii)					
JIC-S	0.0	0.0	4.1	65.9	30.0	0.0	0.0	0.34	0.52	0.34
JIC-M	0.0	0.0	26.0	68.7	5.3	0.0	0.0	0.31	0.52	0.31
JIC-L	0.0	0.0	97.8	2.2	0.0	0.0	0.0	0.98	0.15	0.98
CV	0.0	0.0	1.0	90.5	6.4	1.5	0.5	0.12	0.41	0.18