The Annals of Statistics
2020, Vol. 48, No. 1, 413–439
https://doi.org/10.1214/19-AOS1814

CONSISTENT SELECTION OF THE NUMBER OF CHANGE-POINTS VIA SAMPLE-SPLITTING

BY CHANGLIANG ZOU[1,*], GUANGHUI WANG[1,**] AND RUNZE LI[2]

[1]*School of Statistics and Data Science, Nankai University,* [*]*nk.chlzou@gmail.com;* [**]*ghwang.nk@gmail.com*
[2]*Department of Statistics and The Methodology Center, Pennsylvania State University, rzli@psu.edu*

In multiple change-point analysis, one of the major challenges is to estimate the number of change-points. Most existing approaches attempt to minimize a Schwarz information criterion which balances a term quantifying model fit with a penalization term accounting for model complexity that increases with the number of change-points and limits overfitting. However, different penalization terms are required to adapt to different contexts of multiple change-point problems and the optimal penalization magnitude usually varies from the model and error distribution. We propose a data-driven selection criterion that is applicable to most kinds of popular change-point detection methods, including binary segmentation and optimal partitioning algorithms. The key idea is to select the number of change-points that minimizes the squared prediction error, which measures the fit of a specified model for a new sample. We develop a cross-validation estimation scheme based on an order-preserved sample-splitting strategy, and establish its asymptotic selection consistency under some mild conditions. Effectiveness of the proposed selection criterion is demonstrated on a variety of numerical experiments and real-data examples.

1. Introduction. Change-point detection has received enormous attention due to the emergence of an increasing amount of temporal data. It is a process of detecting mean, variance, or distributional changes in time-ordered observations, and becomes an integrated part of modeling, estimation and inference. Comprehensive reviews of various existing approaches to the inference of multiple change-points (MCP) can be found, for instance, Chen and Gupta (2012) and Aue and Horváth (2013).

The determination of the number of change-points K in a dataset has been central to multiple change-point analysis for decades. It is often approached as a model selection problem, since K drives the model dimension. Bayesian information criterion (BIC, Schwarz (1978)) has become very popular in the change-point problems, for instance, see Yao (1988), Bai and Perron (1998), Braun, Braun and Müller (2000), Fryzlewicz (2014), Zou et al. (2014) and Wang, Zou and Yin (2018), and the asymptotic consistency of the resulting estimator of K has been established in particular contexts of interest. While the BIC is well grounded for general models, different BIC terms are required to adapt to different contexts of MCP problems, and more importantly, the optimal penalization magnitude usually varies from the model and error distribution (Hannart and Naveau (2012), Zhang and Siegmund (2007)). Several ad-hoc criteria for the change-point problem were also proposed, for instance, by Lavielle (2005) and Birgé and Massart (2001). Although these approaches could be visually useful in practice, their theoretical justification remains an open problem.

This article develops a new procedure that attempts to circumvent those limitations while improving the performance of existing criteria. Our strategy is to select the number of change-

Received November 2017; revised October 2018.
MSC2010 subject classifications. 62H12.
Key words and phrases. Cross-validation, dynamic programming, least-squares, model selection, multiple change-point model, selection consistency.

points that minimizes the squared prediction error, which measures the fit of a specified model for a new sample. A new estimation scheme is developed based on the sample splitting, selecting the estimated number of change-points yielding the smallest estimated squared prediction error. Specially, we divide the sample by the parity of the time order, being even or odd, resulting in a 2-fold cross-validation (CV) with order-preserved sample-splitting which is tailored for the change-point problem. The r-fold CV has been widely used to assess the quality of regression and classification models (Shao (1993), Yang (2007)), while analogous results for change-point problems seem rare. This may be because it is well recognized that under a parametric regression framework, the r-fold CV, which performs similar to the Akaike information criterion (AIC), tends to select the model with the optimal prediction performance (Zhang (1993)), while the BIC tends to identify the true sparse model well (Yang (2005)). Interestingly, asymptotic selection consistency of the proposed procedure can be established under some mild conditions, ensuring that the estimated number of change-points equals to the true one with probability tending to one. This may contradict with our intuition but can be understood by carefully examining the connection and difference between the linear regression and change-point problem; see Section 3.2 for details. The only related work we noticed is Arlot and Celisse (2011) which proposed to use a CV-based empirical risk instead of the commonly used least-squares loss function under a univariate mean change model with heterogeneity. However, no theoretical results and numerical evidences on the estimation of the number of change-points were provided.

Our selection criterion and its CV estimation are presented in Sections 2 and 3, respectively, using a unified parametric framework which includes classical univariate or multivariate location and scale problems, ordinary least-squares, generalized linear models, and many others as special cases, provided that the corresponding objective (likelihood or loss) function can be recast into their asymptotically equivalent least-squares problems. The proposed selection criterion makes minimum requirements on the change-point detection approach, and can be applied to almost all kinds of change detection algorithms, such as the local discrepancy based detection (Cao and Wu (2015), Niu and Zhang (2012)), binary segmentation and its variants (Fryzlewicz (2014)), and least-squares or likelihood methods via a dynamic programming algorithm (Bai and Perron (2003), Hawkins (2001), Yao (1988)). The proposed procedure could be also applicable for some other settings with minor modifications, including nonparametric models and correlated cases which are discussed in Section 3.3. In Section 4, numerical experiments indicate that the proposed criterion delivers superior performance in a variety of simulated and real examples. Section 5 concludes with some remarks, and theoretical proofs are delineated in the Appendix. Some technical details and additional numerical results are provided in the Supplementary Material (Zou, Wang and Li (2020)).

Notation. Let $\{\mathbf{x}, \mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a set of d-dimensional vectors and $\mathbf{M}$ be a positive definite matrix. Define the norm $\|\mathbf{x}\| = \sqrt{\mathbf{x}^\top \mathbf{x}}$ and $\|\mathbf{x}\|_{\mathbf{M}} = \sqrt{\mathbf{x}^\top \mathbf{M}\mathbf{x}}$. For any interval $(l, r]$ with $l \geq 0$ and $r \leq n$, denote $\bar{\mathbf{x}}_{l,r} = (r-l)^{-1} \sum_{i=l+1}^{r} \mathbf{x}_i$. Let $\mathcal{T}_L = (\tau_1, \ldots, \tau_L)$ be a set of L points such that $0 < \tau_1 < \cdots < \tau_L < n$. We introduce

$$\mathcal{S}^2_{\mathbf{x}}(\mathcal{T}_L; \mathbf{M}) = \sum_{l=0}^{L} \sum_{i=\tau_l+1}^{\tau_{l+1}} (\mathbf{x}_i - \bar{\mathbf{x}}_{\tau_l, \tau_{l+1}})^\top \mathbf{M} (\mathbf{x}_i - \bar{\mathbf{x}}_{\tau_l, \tau_{l+1}}),$$

where $\tau_0 = 0$ and $\tau_{L+1} = n$. Moreover, let $\tilde{\mathcal{T}}_{\tilde{L}} = (\tilde{\tau}_1, \ldots, \tilde{\tau}_{\tilde{L}})$ be another set of $\tilde{L}$ points such that $0 < \tilde{\tau}_1 < \cdots < \tilde{\tau}_{\tilde{L}} < n$ and we define $\mathcal{S}^2_{\mathbf{x}}(\mathcal{T}_L \cup \tilde{\mathcal{T}}_{\tilde{L}}; \mathbf{M}) = \mathcal{S}^2_{\mathbf{x}}(\mathrm{sort}(\mathcal{T}_L \cup \tilde{\mathcal{T}}_{\tilde{L}}); \mathbf{M})$, where $\mathrm{sort}(A)$ is the set of the sorted elements of A in ascending order. For a sequence $a_n > 0$, we denote $X_n \gtrsim a_n$ if there exists some constant $C > 0$ such that $X_n \geq C a_n$ for large enough n holds with probability approaching one.

2. A unified model and selection criterion.

2.1. *Model.* Suppose we have a sequence of independent data observations $\mathcal{Z} = \{\mathbf{Z}_1, \ldots, \mathbf{Z}_n\}$, collecting from the multiple change-point model,

$$\mathbf{Z}_i \sim m(\cdot \mid \boldsymbol{\beta}_j^*), \quad \tau_j^* < i \le \tau_{j+1}^*, \, j = 0, \ldots, K_n; \, i = 1, \ldots, n, \tag{1}$$

where K_n is the true number of change-points, τ_j^*'s are the locations of these change-points with the convention of $\tau_0^* = 0$ and $\tau_{K_n+1}^* = n$, $\boldsymbol{\beta}_j^*$ is a d-dimensional parameter vector of interest and $m(\cdot \mid \boldsymbol{\beta}_j^*)$ represents the model structure of the segment j satisfying $\boldsymbol{\beta}_j^* \ne \boldsymbol{\beta}_{j+1}^*$.

Denote $\mathcal{Z}_{\tau_j}^{\tau_{j+1}} = (\mathbf{Z}_{\tau_j+1}, \ldots, \mathbf{Z}_{\tau_{j+1}})$ and let $l(\boldsymbol{\beta}; \mathbf{Z}_i)$ be a plausible loss function for $\mathbf{Z}_i$ so that the minimizer of $\mathcal{L}(\boldsymbol{\beta}; \mathcal{Z}_{\tau_j}^{\tau_{j+1}}) = \sum_{i=\tau_j+1}^{\tau_{j+1}} l(\boldsymbol{\beta}; \mathbf{Z}_i)$, $\tilde{\boldsymbol{\beta}}(\mathcal{Z}_{\tau_j}^{\tau_{j+1}})$, is either a natural estimate of $\boldsymbol{\beta}_j^*$ or at least a good surrogate for $\boldsymbol{\beta}_j^*$ when $\tau_j = \tau_j^*$ for $j = 0, \ldots, K_n$. The number of change-points K_n is allowed to grow with the sample size n.

For example, we are frequently concerned with a univariate or multivariate mean change problem, that is, d-variate observations $\mathbf{X}_i$'s follow from

$$\mathbf{X}_i = \boldsymbol{\mu}_j^* + \boldsymbol{\varepsilon}_i, \quad \tau_j^* < i \le \tau_{j+1}^*, \, j = 0, \ldots, K_n; \, i = 1, \ldots, n, \tag{2}$$

where $\boldsymbol{\mu}_j^*$ is the true mean vector for the segment j and $\boldsymbol{\varepsilon}_i$ is a d-dimensional random vector with mean zero and a positive definite covariance matrix $\boldsymbol{\Sigma}$. By taking $\mathbf{Z}_i = \mathbf{X}_i$ and $\boldsymbol{\beta}_j^* = \boldsymbol{\mu}_j^*$, (2) is a special case of (1). The most popular $\mathcal{L}(\boldsymbol{\beta}; \cdot)$ may be the negative log-likelihood (up to constant factors) under normality or the so-called quadratic loss (Yao (1988))

$$\frac{1}{2} \sum_{i=\tau_j+1}^{\tau_{j+1}} \|\mathbf{X}_i - \boldsymbol{\beta}\|^2. \tag{3}$$

Consider another example of identifying structural break in linear regression. Let $\mathbf{Z}_i = (y_i, \mathbf{X}_i)$, where y_i's are the response observations and $\mathbf{X}_i$'s are the d-variate explanatory variables, and $\boldsymbol{\beta}_j^*$'s be the regression coefficients. The $\mathcal{L}(\boldsymbol{\beta}; \cdot)$ can be chosen as the conventional least-squares loss function (Bai and Perron (1998)) or some other robust loss function (Bai (1998)) in the form of

$$\sum_{i=\tau_j+1}^{\tau_{j+1}} \rho(y_i - \mathbf{X}_i^\top \boldsymbol{\beta}), \tag{4}$$

where $\rho(\cdot)$ is a pre-specified function.

2.2. *Criterion for measuring the goodness-of-fit.* Next, we introduce a simple yet effective criterion based on score functions which could avoid numerically obtaining many $\tilde{\boldsymbol{\beta}}(\mathcal{Z}_{\tau_j}^{\tau_{j+1}})$'s under the paradigm of loss function $\mathcal{L}(\boldsymbol{\beta}; \mathcal{Z}_{\tau_j}^{\tau_{j+1}})$. Note that very often $\mathrm{E}\{\mathbf{s}(\boldsymbol{\beta}_j^*; \mathbf{Z}_i)\} \approx \mathbf{0}$, $i \in (\tau_j^*, \tau_{j+1}^*]$, where $\mathbf{s}(\cdot; \cdot)$ is the first-order derivative of $l(\boldsymbol{\beta}; \mathbf{Z}_i)$ with respect to $\boldsymbol{\beta}$. Ideally, given a $\boldsymbol{\gamma}$, $\mathrm{E}\{\mathbf{s}(\boldsymbol{\gamma}; \mathbf{Z}_i)\} \ne \mathrm{E}\{\mathbf{s}(\boldsymbol{\gamma}; \mathbf{Z}_{i'})\}$ for $i \in (\tau_{j-1}^*, \tau_j^*]$ and $i' \in (\tau_j^*, \tau_{j+1}^*]$, which motivates us to consider a least-squares measure described below.

Given a candidate model, $\mathcal{M}_L$, which is specified by a set of change-points $\mathcal{T}_L = (\tau_1, \ldots, \tau_L)$ and the corresponding parameters $\boldsymbol{\mu}_j$'s that are approximations to $\mathrm{E}\{\mathbf{s}(\boldsymbol{\gamma}; \mathbf{Z}_i)\}$, $i \in (\tau_j, \tau_{j+1}]$, $j = 0, \ldots, L$, small values of

$$\mathcal{C}(\mathcal{M}_L; \mathcal{Z}) = \sum_{j=0}^{L} \sum_{i=\tau_j+1}^{\tau_{j+1}} \{\mathbf{s}(\boldsymbol{\gamma}; \mathbf{Z}_i) - \boldsymbol{\mu}_j\}^\top \mathbf{W}_n \{\mathbf{s}(\boldsymbol{\gamma}; \mathbf{Z}_i) - \boldsymbol{\mu}_j\} \tag{5}$$

may indicate a good fit of data. One can expect that inappropriate numbers of change-points may lead to a large value of $\mathcal{C}(\mathcal{M}_L; \mathcal{Z})$, where $\mathbf{W}_n$, possibly depending on $\boldsymbol{\gamma}$, serves as a rough scale estimator for standardization.

For instance, under the multivariate mean change model (2), $\mathbf{s}(\boldsymbol{\gamma}, \mathbf{Z}_i) = -(\mathbf{X}_i - \boldsymbol{\gamma})$ if the loss function (3) is used. Accordingly, the $\mathcal{C}(\mathcal{M}_L; \mathcal{Z})$ becomes

$$\sum_{j=0}^{L} \sum_{i=\tau_j+1}^{\tau_{j+1}} (\mathbf{X}_i - \boldsymbol{\gamma} + \boldsymbol{\mu}_j)^\top \mathbf{W}_n (\mathbf{X}_i - \boldsymbol{\gamma} + \boldsymbol{\mu}_j).$$

For another example, consider detecting the change in a univariate sequence where $\mathrm{E}(X_i) = \nu_j^*$ and $\mathrm{Var}(X_i) = \sigma^2 V(\nu_j^*)$ with some function $V(\cdot)$ for $\tau_j^* < i \le \tau_{j+1}^*$, $j = 0, \ldots, K_n$. Braun, Braun and Müller (2000) suggested using quasi-deviance as a goodness-of-fit criterion, in our notation, $l(\mu, x) = \int_\mu^x (x-t)/V(t)\,dt$. It can be easily checked that $\mathcal{C}(\mathcal{M}_L; \mathcal{Z}) = \sum_{j=0}^{L} \sum_{i=\tau_j+1}^{\tau_{j+1}} \{(X_i - \gamma)/V(\gamma) + \mu_j\}^2 W_n$.

The role of $\mathcal{C}(\mathcal{M}_L; \mathcal{Z})$ can be more clearly understood by further decomposing it into

$$\begin{aligned} \mathcal{C}(\mathcal{M}_L; \mathcal{Z}) &= \sum_{j=0}^{L} \sum_{i=\tau_j+1}^{\tau_{j+1}} \{\mathbf{s}(\boldsymbol{\gamma}; \mathbf{Z}_i) - \bar{\mathbf{s}}(\boldsymbol{\gamma}; \mathcal{Z}_{\tau_j}^{\tau_{j+1}})\}^\top \mathbf{W}_n \{\mathbf{s}(\boldsymbol{\gamma}; \mathbf{Z}_i) - \bar{\mathbf{s}}(\boldsymbol{\gamma}; \mathcal{Z}_{\tau_j}^{\tau_{j+1}})\} \\ &\quad + \sum_{j=0}^{L} n_j \{\bar{\mathbf{s}}(\boldsymbol{\gamma}; \mathcal{Z}_{\tau_j}^{\tau_{j+1}}) - \boldsymbol{\mu}_j\}^\top \mathbf{W}_n \{\bar{\mathbf{s}}(\boldsymbol{\gamma}; \mathcal{Z}_{\tau_j}^{\tau_{j+1}}) - \boldsymbol{\mu}_j\} \qquad (6) \\ &\equiv \mathcal{S}_{\mathbf{s}}^2(\mathcal{T}_L; \mathbf{W}_n) + \mathcal{D}(\mathcal{M}_L; \mathcal{Z}), \end{aligned}$$

where $\bar{\mathbf{s}}(\boldsymbol{\gamma}; \mathcal{Z}_{\tau_j}^{\tau_{j+1}}) = n_j^{-1} \sum_{i=\tau_j+1}^{\tau_{j+1}} \mathbf{s}(\boldsymbol{\gamma}; \mathbf{Z}_i)$ and $n_j = \tau_{j+1} - \tau_j$. By noting that $\mathcal{S}_{\mathbf{s}}^2(\mathcal{T}_L; \mathbf{W}_n) - \mathcal{S}_{\mathbf{s}}^2(\mathcal{T}_{K_n}; \mathbf{W}_n)$ could be quite large when $L < K_n$, $\mathcal{S}_{\mathbf{s}}^2(\mathcal{T}_L; \mathbf{W}_n)$ would help prevent the underfitting. On the other hand, when $L > K_n$, $\mathcal{S}_{\mathbf{s}}^2(\mathcal{T}_L; \mathbf{W}_n)$ does not decrease too much as L increases, but the term $\mathcal{D}(\mathcal{M}_L; \mathcal{Z})$ would dominate $\mathcal{D}(\mathcal{M}_{K_n}; \mathcal{Z})$ under certain conditions on $\boldsymbol{\mu}_j$'s. Thus, $\mathcal{C}(\mathcal{M}_L; \mathcal{Z})$ could be a useful measure to quantify the deviation from the true model. In practice, the candidate model $\mathcal{M}_L$ needs to be estimated based on the only available sample and thus a cross-validation based estimation procedure is developed.

3. Cross-validation for change-points.

3.1. *Algorithm.* In this section, we propose a new selection criterion based on the estimated $\mathcal{C}(\mathcal{M}_L; \cdot)$ through a special 2-fold cross-validation scheme. The key idea is to split the data into one training set $\mathcal{Z}_1$ and one validation set $\mathcal{Z}_2$, where the training set $\mathcal{Z}_1$ is used to construct a candidate model $\mathcal{M}_L$ via a given change detection algorithm, and $\mathcal{C}(\mathcal{M}_L; \mathcal{Z}_2)$ is estimated as the goodness-of-fit measured on the left-out validation set $\mathcal{Z}_2$. To further reduce the estimation variability due to splitting randomness, multiple data splittings can be performed (Zhang (1993)). However, since the change-point problem has an intrinsic order structure, randomly splitting may not be an ideal choice. A simple yet effective way is to use the parity splitting, that is, dividing the sample into

$$\mathcal{Z}_O = \{\mathcal{Z}_{2t-1}, t = 1, \ldots, T\} \quad \text{and} \quad \mathcal{Z}_E = \{\mathcal{Z}_{2t}, t = 1, \ldots, T\},$$

one of which is set as the training set and the other is used for validation, where we assume for convenience that $n = 2T$ is even. Using this splitting strategy, the original change-point structure can be preserved as much as possible and the difference between the training and validation samples is minimal. Our procedure is described as follows.

Suppose that a base change detection algorithm $\mathcal{A}(L; \mathcal{Z})$ and the largest possible number of change-points K_n^{U} are prespecified. The estimated number of change-points $\widehat{K}_n$ can be obtained via the following *Cross-validation with Order-Preserved Sample-Splitting (COPSS)* procedure.

Cross-validation with Order-Preserved Sample-Splitting (COPSS):

Step 1 (Initialization). Specify a proper $\boldsymbol{\gamma}$ and $\mathbf{W}_n$. Compute $\mathbf{s}(\boldsymbol{\gamma}; \mathbf{Z}_i)$ for $i = 1, \ldots, n$.

Step 2 (Training). Given L, obtain the set of change-points $\widehat{\mathcal{T}}_L^O = (\widehat{\tau}_{L,1}^O, \ldots, \widehat{\tau}_{L,L}^O)$ based on the $\mathcal{Z}_O$ using the detection algorithm $\mathcal{A}(L; \mathcal{Z}_O)$. Then compute $\bar{\mathbf{s}}(\boldsymbol{\gamma}; \mathcal{Z}_{\widehat{\tau}_{L,j}^O}^{\widehat{\tau}_{L,j+1}^O})$ for $j = 0, \ldots, L$ based on $\mathcal{Z}_O$. Denote the resulting change-point model as $\widehat{\mathcal{M}}_L^O$.

Step 3 (Validation). Compute $\mathcal{C}(\widehat{\mathcal{M}}_L^O; \mathcal{Z}_E)$ using (5) with $\boldsymbol{\mu}_j$'s replaced by $\bar{\mathbf{s}}(\boldsymbol{\gamma}; \mathcal{Z}_{\widehat{\tau}_{L,j}^O}^{\widehat{\tau}_{L,j+1}^O})$'s.

Step 4 (Cross-validation). Repeat Steps 2–3 by interchanging $\mathcal{Z}_O$ and $\mathcal{Z}_E$ and obtain $\mathcal{C}(\widehat{\mathcal{M}}_L^E; \mathcal{Z}_O)$.

Step 5 (Estimation). Set

$$\widehat{K}_n = \operatorname*{arg\,min}_{1 \le L \le K_n^{\mathrm{U}}} \{\mathcal{C}(\widehat{\mathcal{M}}_L^O; \mathcal{Z}_E) + \mathcal{C}(\widehat{\mathcal{M}}_L^E; \mathcal{Z}_O)\}$$

as the estimated number of change-points.

To better understand the mechanism of our proposed procedure, we consider the classical univariate mean change-point problem for illustration. Assume a univariate sequence of observations X_i's follow from (2). The popular BIC minimizes

$$S_{\mathrm{BIC}}(L) = \frac{n}{2} \log \left\{ n^{-1} \sum_{j=0}^{L} \sum_{i=\widehat{\tau}_{L,j}+1}^{\widehat{\tau}_{L,j+1}} (X_i - \bar{X}_{\widehat{\tau}_{L,j}, \widehat{\tau}_{L,j+1}})^2 \right\} + L\zeta_n, \tag{7}$$

where $(\widehat{\tau}_{L,1}, \ldots, \widehat{\tau}_{L,L})$ is obtained by $\mathcal{A}(L; \mathcal{Z})$ based on the whole sample. The second term of order ζ_n can be viewed as a penalty which is chosen to be slightly larger than the maximum variation level (no change) so that it can dominate the first term of $S_{\mathrm{BIC}}(L)$ under overfitting models with high probability and in this case it is usually chosen as of order $\log n$ (Yao (1988)).

Asymptotically speaking, as long as $\log n/\zeta_n \to c \in [0, \infty)$ and $\zeta_n/n \to 0$, the BIC estimator is consistent when the change magnitudes are fixed. However, the "optimal" penalty is always not easy to be determined since it may depend on the change magnitudes and error distributions. In contrast, it can be verified that taking the quadratic loss function (3), $\mathcal{C}(\widehat{\mathcal{M}}_L^O; \mathcal{Z}_E)$ is equivalent to (up to a scale constant)

$$\sum_{j=0}^{L} \sum_{i=\widehat{\tau}_{L,j}^O+1}^{\widehat{\tau}_{L,j+1}^O} (X_i^E - \bar{X}^E_{\widehat{\tau}_{L,j}^O, \widehat{\tau}_{L,j+1}^O})^2 + \sum_{j=0}^{L} n_j (\bar{X}^E_{\widehat{\tau}_{L,j}^O, \widehat{\tau}_{L,j+1}^O} - \bar{X}^O_{\widehat{\tau}_{L,j}^O, \widehat{\tau}_{L,j+1}^O})^2, \tag{8}$$

where the symbols with the superscripts "O" and "E" stand for the quantities based on the sample $\mathcal{Z}_O$ and $\mathcal{Z}_E$, respectively. In our CV-based procedure, the second term plays a similar role to the "ζ_n" term in the BIC, that is, avoiding overfitting (see Section 3.2 for theoretical discussion). As opposed to the BIC, the $\mathcal{C}(\widehat{\mathcal{M}}_L^O; \mathcal{Z}_E)$ can be viewed as a *data-driven* penalized loss function which greatly facilitates the determination of the number of change-points in practice. This data-driven feature benefits from the use of sample-splitting and thus certain efficiency loss would be incurred. Intuitively, using the summation of $\mathcal{C}(\widehat{\mathcal{M}}_L^E; \mathcal{Z}_O)$ and $\mathcal{C}(\widehat{\mathcal{M}}_L^O; \mathcal{Z}_E)$ may result in variance reduction that is verified by simulation in Section 4.

REMARK 1. The $\boldsymbol{\gamma}$ and $\mathbf{W}_n$ need to be specified to implement our algorithm. In many cases, such like the multivariate mean change-point model and change-point regression problem with least-squares loss function, it can be verified that the algorithm is invariant with $\boldsymbol{\gamma}$. In fact, our numerical results also reveal that the choice of $\boldsymbol{\gamma}$ is not crucial and thus we recommend using $\tilde{\boldsymbol{\beta}}_n \equiv \arg\min_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}; \mathcal{Z}_1^n)$ (assumed to exist) as $\boldsymbol{\gamma}$, when no preference is given. The performance of our procedure is not sensitive to $\mathbf{W}_n$ either, because the $\mathbf{W}_n$ plays only the role in standardizing the components of $\mathbf{s}(\boldsymbol{\gamma}; \mathbf{Z}_i)$ so that they are aggregated in a fair way. From the asymptotic analysis in Section 3.2 we can see that there is a minimal requirement for $\mathbf{W}_n$, and hence we can even simply choose it as the identity matrix $\mathbf{I}_d$ or the pooled sample covariance matrix based on $\{\mathbf{s}(\boldsymbol{\gamma}; \mathbf{Z}_1), \ldots, \mathbf{s}(\boldsymbol{\gamma}; \mathbf{Z}_n)\}$.

3.2. *Theoretical justification.* We now establish an asymptotic property regarding the selection consistency of the COPSS procedure. The consistency property ensures that the resulting estimated number of change-points equals to the true one with probability approaching one, when the change detection algorithm $\mathcal{A}(L; \mathcal{Z})$ performs reasonably well.

For ease of exposition, we introduce the following notation. Let $\mathcal{T}_{K_n}^* = (\tau_1^*, \ldots, \tau_{K_n}^*)$. Denote the minimal and maximal distance between change-points as $\underline{\lambda}_n = \min_{0 \le j \le K_n}(\tau_{j+1}^* - \tau_j^*)$ and $\overline{\lambda}_n = \max_{0 \le j \le K_n}(\tau_{j+1}^* - \tau_j^*)$, respectively, and the minimal signal strength as $\underline{\Delta}_n = \min_{1 \le j \le K_n} \|\boldsymbol{\mu}_{j-1}^* - \boldsymbol{\mu}_j^*\|^2$. Given $L \ge 1$, let $\widehat{\mathcal{T}}_L = (\widehat{\tau}_{L,1}, \ldots, \widehat{\tau}_{L,L})$ be the estimated change-points based on half of the data. For $j = 0, \ldots, K_n$, let $\boldsymbol{\mu}_j^* = \mathrm{E}\{\mathbf{s}(\boldsymbol{\gamma}; \mathbf{Z}_i)\}$, $\boldsymbol{\Sigma}_j^* = \mathrm{Cov}\{\mathbf{s}(\boldsymbol{\gamma}; \mathbf{Z}_i)\}$ and $\mathbf{U}_i = \mathbf{s}(\boldsymbol{\gamma}; \mathbf{Z}_i) - \boldsymbol{\mu}_j^*$, $i \in (\tau_j^*, \tau_{j+1}^*]$. Denote by $\overline{\sigma}(\boldsymbol{\Sigma}_j^*)$ and $\underline{\sigma}(\boldsymbol{\Sigma}_j^*)$ the maximum and minimum eigenvalues of $\boldsymbol{\Sigma}_j^*$ for $j = 0, \ldots, K_n$, and moreover let $\overline{\sigma} = \max\{\overline{\sigma}(\boldsymbol{\Sigma}_0^*), \ldots, \overline{\sigma}(\boldsymbol{\Sigma}_{K_n}^*)\}$ and $\underline{\sigma} = \min\{\underline{\sigma}(\boldsymbol{\Sigma}_0^*), \ldots, \underline{\sigma}(\boldsymbol{\Sigma}_{K_n}^*)\}$. Also, denote the maximum and minimum eigenvalues of $\mathbf{W}_n$ by $\overline{\omega}_n$ and $\underline{\omega}_n$, respectively.

ASSUMPTION 1 (Noises). The $\boldsymbol{\Sigma}_j^*$'s are positive-definite matrices and there exists a positive integer $m \ge 2$ such that $\mathrm{E}(\|\boldsymbol{\Sigma}_j^{*-1/2}\mathbf{U}_i\|^{2m}) < \infty$ for $i \in (\tau_j^*, \tau_{j+1}^*]$, $j = 0, \ldots, K_n$.

ASSUMPTION 2 (Detection Precision). If $q = L - K_n \ge 0$, then there exist $\widehat{\tau}_{L,i_1}, \ldots, \widehat{\tau}_{L,i_{K_n}}$ belonging to $\widehat{\mathcal{T}}_L$ such that $\max_{1 \le j \le K_n} |\widehat{\tau}_{L,i_j} - \tau_j^*| \le \delta_{q,n}$ holds with probability approaching one as $n \to \infty$, where $\delta_{q,n}$ is some positive sequence.

ASSUMPTION 3 (Minimum Signal). The jump sizes $\|\boldsymbol{\mu}_{j-1}^* - \boldsymbol{\mu}_j^*\|$'s satisfy

$$\frac{\underline{\lambda}_n \underline{\omega}_n \underline{\Delta}_n}{K_n \overline{\omega}_n \overline{\sigma} \overline{\lambda}_n^{2/m}} \to \infty \quad \text{as } n \to \infty. \tag{9}$$

REMARK 2. The moment conditions in Assumption 1 are used to control the supremum of the objective function and is commonly used in the literature, for example, Yao and Au (1989). Assumption 2 sets theoretical minimal requirements for the accuracy of the change-points detected by the algorithm $\mathcal{A}(L; \mathcal{Z})$ when $L \ge K_n$. This is reasonable because we cannot expect that our selection procedure would work well if all the estimated change-points are far away from the true ones. Under such circumstances, an appropriate algorithm generally results in a $\delta_{q,n}$-neighborhood of the true location set in the sense that for each true change-point there exists at least a point in the estimated set so that their distance is less than $\delta_{q,n}$ (Harchaoui and Lévy-Leduc (2010)). The condition on $\delta_{q,n}$ will be made in Theorem 1. We explicitly express the dependence of $\delta_{q,n}$ on q because for some algorithms different estimation accuracies may be achieved with different values of q; see the discussions below Theorem 1. The requirements on the smallest signal strength and distance between two

change-points are made in Assumption 3 so that the change-points are asymptotically identifiable. It can be further relaxed if (sub-)Gaussian noises are considered (Niu, Hao and Zhang (2016)). Under the conventional assumption that K_n does not depend on n, τ_j^*/n converges to a constant for each j and the change magnitudes are fixed, Assumption 3 is valid when $m > 2$.

It is worth noting that the conditions on the signal strength is made for $\|\boldsymbol{\mu}_j^* - \boldsymbol{\mu}_{j-1}^*\|$ rather than $\|\boldsymbol{\beta}_j^* - \boldsymbol{\beta}_{j-1}^*\|$ in Assumption 3. Simply speaking, an implicit assumption here is that the change in $\boldsymbol{\beta}$ would result in the change in $E\{\mathbf{s}(\boldsymbol{\gamma}; \mathbf{Z}_i)\}$ and consequently the segmentations with $\mathbf{s}(\boldsymbol{\gamma}; \mathbf{Z}_i)$ would be approximately equivalent to the original change-point model. In fact, $\|\boldsymbol{\mu}_j^* - \boldsymbol{\mu}_{j-1}^*\|$ is often an increasing function of $\|\boldsymbol{\beta}_j^* - \boldsymbol{\beta}_{j-1}^*\|$ for $j = 1, \ldots, K_n$. For example, under the classical multivariate mean change-point model (2), $\boldsymbol{\mu}_j^* - \boldsymbol{\mu}_{j-1}^* = \boldsymbol{\beta}_j^* - \boldsymbol{\beta}_{j-1}^*$. This is also true if the quasi-deviance is used (Braun, Braun and Müller (2000)). Also, for the regression problem with the loss function (4) being $\rho(x) = x^2/2$, we will have $\mathbf{s}(\boldsymbol{\gamma}; \mathbf{Z}_i) = -\mathbf{X}_i(y_i - \mathbf{X}_i^\top \boldsymbol{\gamma})$, and thus $\boldsymbol{\mu}_j^* - \boldsymbol{\mu}_{j-1}^* = E(\mathbf{X}_i\mathbf{X}_i^\top)(\boldsymbol{\beta}_j^* - \boldsymbol{\beta}_{j-1}^*)$.

THEOREM 1. *Suppose that Assumptions 1–3 hold. If there exist positive sequences* $\alpha_{q,n}$, $q = 0, 1, \ldots,$ *satisfying that* $K_n \log\log \delta_{q,n} = o(\alpha_{q,n})$, *and for* $L = K_n + q$ *with* $q \geq 1$,

$$\mathcal{S}_{\mathbf{U}}^2(\mathcal{T}_{K_n}^*; \mathbf{W}_n) - \mathcal{S}_{\mathbf{U}}^2(\widehat{\mathcal{T}}_L \cup \mathcal{T}_{K_n}^*; \mathbf{W}_n) \gtrsim \overline{\omega}_n \overline{\sigma} \alpha_{q,n}, \tag{10}$$

then our procedure is consistent in the sense that $\lim_{n\to\infty} \Pr(\widehat{K}_n = K_n) = 1$.

Intuitively speaking, the condition (10) implies that the reduction of total variation due to adding the points in $\widehat{\mathcal{T}}_L$ into the true set has a lower bound diverging to infinity. The mechanism of locating change-points is usually to search for a model from the candidate models so that the total variation is mostly reduced, and accordingly the condition (10) will be roughly satisfied for some $\alpha_{q,n}$.

The condition $K_n \log\log \delta_{q,n} = o(\alpha_{q,n})$ is quite mild and can be satisfied by many effective detection procedures. In particular, it holds for the binary segmentation (BS) method (Venkatraman (1992)) with $\alpha_{q,n} = \log\log \overline{\lambda}_n$, and for the optimal partitioning (OP) algorithm (Auger and Lawrence (1989)) and local discrepancy (LD) based algorithm (Niu and Zhang (2012)) with $\alpha_{q,n} = \eta_{q,n}$, where

$$\eta_{q,n} = \begin{cases} \log\log \overline{\lambda}_n & \text{if } q = 0, 1, \\ \log \overline{\lambda}_n & \text{if } q \geq 2. \end{cases}$$

For $q = 0$, say the number of change-point is correctly specified, it is well known that the change-point estimators obtained by the algorithms mentioned above are consistent with the optimal rate $O_p(1)$ in a parametric setting, when the change magnitudes are fixed; see Venkatraman (1992), Bai and Perron (1998), Hao, Niu and Zhang (2013) and the references therein. Thus, this condition holds if $K_n / \log\log \overline{\lambda}_n \to 0$. By Lemmas 4–5 given in the Appendix, we can verify that the case of $q > 0$ is also valid for those algorithms. This condition also restricts the relationship between the K_n and sample size n. Faster divergence rate of K_n may be possible, but more stringent conditions on the signal strength and tail probabilities of $\mathbf{U}_i$ would be required.

REMARK 3. It is interesting to examine the case with $\underline{\Delta}_n \to 0$. Generally, $\delta_{q,n} \gtrsim \underline{\Delta}_n^{-1}$ (Niu, Hao and Zhang (2016)), and thus $K_n \log\log \delta_{q,n} = o(\alpha_{q,n})$ would not hold if the minimal signal strength goes to zero in a polynomial rate for $q = 0, 1$. In such situations, the COPSS procedure is likely to yield an overfitting model but with only one redundant change-point since that condition may still hold for $q = 2$, at least for the LD or OP algorithm.

We show in the next theorem that (10) holds when $\widehat{\mathcal{T}}_L$ was obtained by the popular BS and OP algorithms. Specifically, for a given model size L, the OP algorithm finds the estimated change-point set by $\widehat{\mathcal{T}}_L = \arg\min_{\mathcal{M}_L} \mathcal{C}(\mathcal{M}_L; \mathcal{Z})$. For the BS algorithm, let $\widehat{\tau}_{1,1} = \arg\min_{\mathcal{M}_1} \mathcal{C}(\mathcal{M}_1; \mathcal{Z})$, and then for $2 \le l \le L$, one iteratively obtains

$$\widehat{\tau}_{l,l} = \underset{0\le k\le l-1}{\arg\min} \Big\{ \min_{\mathcal{M}_1} \mathcal{C}\big(\mathcal{M}_1; \mathcal{Z}_{\widehat{\tau}_{l,k}+1}^{\widehat{\tau}_{l,k+1}}\big) \Big\},$$

where $\widehat{\tau}_{l,k} = \widehat{\tau}_{l-1,k}$ for $k = 1, \ldots, l-1$ with $\widehat{\tau}_{l,0} = 0$ and $\widehat{\tau}_{l,l} = n$. The final estimated change-point set is $\widehat{\mathcal{T}}_L = (\widehat{\tau}_{L,1}, \ldots, \widehat{\tau}_{L,L})$. Note that the BS was typically used in conjunction with a thresholding criterion. Consequently, the estimated change-point number depends on the threshold and the procedure does not necessarily guarantee a model with any given size by choosing a suitable threshold. Hence, we modify it as above so that the algorithm is in a nested way. The BS will thereafter refer to this one which should not cause any confusion.

THEOREM 2. *If Assumptions 1–3 hold and* $\liminf_{n\to\infty}(\underline{\omega}_n\underline{\sigma})/(\overline{\omega}_n\overline{\sigma}) > 0$, *then the condition* (10) *is valid for the optimal partitioning and binary segmentation algorithms with* $\alpha_{q,n}$ *being* $\eta_{q,n}$ *and* $\log\log\overline{\lambda}_n$, *respectively, and accordingly the selection procedure is consistent.*

The proofs of Theorems 1–2 are given in the Appendix. The ideas of the proofs are similar to that of Nishii (1984). When a correct model is compared with an underfitting model, the first term of the criterion function $\mathcal{S}_{\mathbf{s}}^2(\widehat{\mathcal{T}}_L^O; \mathbf{W}_n)$ in (6), which measures the goodness-of-fit of the number and locations of the change-points obtained from the sample $\mathcal{Z}_O$ on the sample $\mathcal{Z}_E$, asymptotically dominates and the correct model is preferred; when comparing a simpler correct model with a more complex correct model, the second term of the criterion function, that is, the "penalty" term $\mathcal{D}(\widehat{\mathcal{M}}_L^O; \mathcal{Z}_E)$, asymptotically dominates and the simpler model is preferred. Hence, with probability approaching one, the CV criterion favors the true model over either an underfitting model or an overfitting model. We also want to point out that the condition (10) also holds for the LD algorithm such as the SaRa proposed by Niu and Zhang (2012), with $\alpha_{q,n} = \eta_{q,n}$, but more conditions on a sliding window size is needed.

In general, under a large-sample framework, in which the number of variables p is fixed and n goes to infinity, it has been pointed out in various models that the r-fold CV or the delete-k CV (if $\liminf_{n\to\infty}(n-k)/n > 0$) is not consistent (Shao (1993), Zhang (1993)). If the training sample size is negligible compared to n, then model consistency could be obtained. This has been confirmed theoretically by Shao (1993, 1997) for the variable selection problem in linear regression. It turns out that, when the goal is to identify the true model, the proportion of data used for evaluation in CV needs to be dominating in size. Using a very small proportion of the data for training is clearly not a good choice in our change-point problem, because the accuracy of change-point detection algorithms heavily relies on the sample size. On the other hand, under some high-dimensional or infinite-dimensional models, different consistency behaviors are noted (Bai, Fujikoshi and Choi (2017), Yang (2005)). In particular, Yang (2007) revealed interesting behaviors of CV: under some conditions, with an appropriate choice of data splitting ratio, CV is consistent when it is applied to compare between parametric and nonparametric methods or within nonparametric methods. These related findings shed light on understanding why the CV works in the MCP. In fact, if the candidate number of change-points is L, the cardinality of the collection of candidate models is diverging with n, say $\binom{n-1}{L}$, resulting in the validity of the condition (10). From the proof of Theorem 1, we can tell that $\mathcal{D}(\widehat{\mathcal{M}}_L^O; \mathcal{Z}_E)$ is approximately larger than $\mathcal{D}(\widehat{\mathcal{M}}_{K_n}^O; \mathcal{Z}_E)$ by an order of at least $\log\log\underline{\lambda}_n$ when $L > K_n$ and thus (10) holds, whereas $\mathcal{S}_{\mathbf{s}}^2(\widehat{\mathcal{T}}_{K_n}^O; \mathbf{W}_n) - \mathcal{S}_{\mathbf{s}}^2(\widehat{\mathcal{T}}_L^O; \mathbf{W}_n)$ is just $O_p(1)$ which would result in the favor of the true model.

However this is not the case in the classical regression problem where the number of variables is fixed and the cardinality of the collection of candidate models is accordingly fixed as $n \to \infty$.

For a clearer comparison, we consider the univariate sequence with $K = 0$, say no change-point. It can be verified that in this case the major term in $\mathcal{D}(\widehat{\mathcal{M}}^O_{K+1}; \mathcal{Z}_E) - \mathcal{D}(\widehat{\mathcal{M}}^O_K; \mathcal{Z}_E)$ is of the form

$$\max_{1 \le \tau \le T-1} \frac{\tau(T-\tau)}{T} \big(\bar{X}^O_{0,\tau} - \bar{X}^O_{\tau,T}\big)^2,$$

which is of $O_p(\log\log n)$ by the Darling–Erdős Theorem (Darling and Erdös (1956)). In contrast, in the regression problem with only one candidate covariate, $\mathcal{D}(\widehat{\mathcal{M}}^O_{K+1}; \mathcal{Z}_E)$ reflects only the difference between the two least-squares estimators obtained from $\mathcal{Z}_E$ and $\mathcal{Z}_O$, and thus $\mathcal{D}(\widehat{\mathcal{M}}^O_{K+1}; \mathcal{Z}_E) - \mathcal{D}(\widehat{\mathcal{M}}^O_{K+1}; \mathcal{Z}_E) = O_p(1)$; the CV will fail.

REMARK 4. From the proofs of Theorems 1–2, we can claim that our proposed procedure is also consistent if we use classical r-fold ($r > 2$) CV to replace the parity splitting. There is no general conclusion that the latter would outperform the commonly-used 5-fold or 10-fold CV. Intuitively speaking, a 5-fold CV would help to obtain a more accurate training model as we use 80% data, preventing the model from undefitting to certain degree. However, since the validation set with only 20% observations may not fully reflect the underlying change-point structure, overfitting would often be incurred if the sample size is not sufficiently large. Table S.3 in the Supplementary Material presents some results by using both the classical and a slightly modified order-preserved multi-fold CV. Though the 2-fold strategy in the COPSS procedure may not be always the optimal one, our numerical experience indicates that it is capable of providing balanced protection from the underfiting and overfitting because this splitting method makes the training and validation sets the most similar among all the choices of splitting. Considering its computational advantages, we would recommend it for practical use when there is little prior information about the data. A random assigning treatment in conjunction with our 2-fold splitting strategy as suggested by an anonymous referee could improve in some scenarios especially when the sequence has some systematical trend. More details can be found in the Supplementary Material.

3.3. *Extensions.*

3.3.1. *Modified CV procedure for the PELT.* The COPSS can be applied to most change detection algorithms which seek for all possible segmentations with the number of change-points $0 \le L \le K_n^{\mathrm{U}}$. In contrast, there are other efficient approaches such as the Pruned Exact Linear Time (PELT) Algorithm (Killick, Fearnhead and Eckley (2012)) which was designed for identifying multiple change-points by directly minimizing a "loss" plus "penalty" function over all possible numbers and locations of change-points. Consequently, the PELT outputs a single number of estimated change-points instead of running over all possible candidate models. The issue of specifying penalty terms for the PELT still remains open. The COPSS would be helpful, say we may choose a suitable penalty term which produces a relatively small squared prediction error over a sequence of penalization magnitudes. Although this procedure cannot go over all candidate models as it is uncontrollable to obtain a one-to-one correspondence from the model size to the penalization magnitude, it is able to considerably alleviate the dependence on the manual choice of penalty term. Some numerical evidence can be found in Section 4.

3.3.2. *A nonparametric setting*. Without imposing any parametric modeling assumption, consider the MCP based on independent data $\mathcal{Z} = \{X_i\}_{i=1}^n$, such that

$$X_i \sim F_j(x), \quad \tau_j^* < i \le \tau_{j+1}^*, j = 0, \ldots, K_n; i = 1, \ldots, n,$$

where F_j is the cumulative distribution (CDF) of the segment j satisfying $F_j \neq F_{j+1}$. Lee (1996) and Zou et al. (2014) discussed localization-based and global-loss-based detection algorithms using empirical CDF, respectively. Zou et al. (2014) and Haynes, Fearnhead and Eckley (2017) studied the estimation of K_n by the BIC. Following Zou et al. (2014), we may consider the criterion as

$$\mathcal{C}(\mathcal{M}_L; \mathcal{Z}) = \int_u \mathcal{L}_u(\mathcal{M}_L)\, dw(u), \tag{11}$$

where $\mathcal{L}_u$ is the negative joint nonparametric log-likelihood for each given candidate model $\mathcal{M}_L$, namely,

$$\begin{aligned}\mathcal{L}_u(\mathcal{M}_L) = -\sum_{j=0}^{L} (\tau_{j+1} - \tau_j)\{&\widehat{F}_{\tau_j}^{\tau_{j+1}}(u) \log\big(\tilde{F}_{\tau_j}^{\tau_{j+1}}(u)\big) \\ &+ \big(1 - \widehat{F}_{\tau_j}^{\tau_{j+1}}(u)\big) \log\big(1 - \tilde{F}_{\tau_j}^{\tau_{j+1}}(u)\big)\},\end{aligned}$$

$\widehat{F}_{\tau_j}^{\tau_{j+1}}(u)$ is the empirical CDF of the sample $\{X_{\tau_j+1}, \ldots, X_{\tau_{j+1}}\}$ and $w(u)$ is a nonnegative weight function. In this case, the model $\mathcal{M}_L$ is represented by a candidate set of change-points $(\tau_1 < \cdots < \tau_L)$ and the associated "pseudo" CDF $\tilde{F}_{\tau_j}^{\tau_{j+1}}(u)$ for $j = 0, \ldots, L$. Accordingly, $\mathcal{C}(\widehat{\mathcal{M}}_L^O; \mathcal{Z}_E)$ can be obtained by taking $\mathcal{A}(L; \mathcal{Z}_O)$ as the method proposed by Lee (1996), Zou et al. (2014) or its PELT version Haynes, Fearnhead and Eckley (2017).

3.3.3. *Cases when unknown correlations exist*. Though the asymptotic consistency of our proposed estimator is established under the assumption that $\mathbf{Z}_i$'s are independent, we may expect that the procedure is also applicable for dependent cases. The main difficulty lies in that the parity splitting would make $\mathcal{Z}_O$ and $\mathcal{Z}_E$ have similar error structures because the nearest observations are usually most correlated. By adapting the idea of moving block bootstrap for stationary series (Künsch (1989)), we suggest a pre-localizing procedure which is capable of alleviating the effect of autocorrelations to certain degree. Our first step is to locate the most influential points that have the largest local jump sizes quantified by certain measures.

LOCALIZING ALGORITHM.

Step 1. Choose an appropriate integer ω_n and take the change-point set as $\mathcal{O} = \varnothing$.

Step 2. Initialize $T_i = 0$ for $i = 1, \ldots, n$. For $i = \omega_n, \ldots, n - \omega_n$, update T_i to be a two-sample test statistic for the samples $\mathcal{Z}_{i-\omega_n}^i$ and $\mathcal{Z}_i^{i+\omega_n}$.

Step 3. For $i = \omega_n, \ldots, n - \omega_n$, if $i = \arg\max_{i-\omega_n < j \le i+\omega_n} |T_j|$, update $\mathcal{O} = \mathcal{O} \cup \{i\}$.

The ω_n is a sequence of sliding window lengths for which $\omega_n/n \to 0$. Properties of using local discrepancy measures to detect multiple change-points in univariate sequences have been widely studied; see, for example, Lee (1996), Jeng, Cai and Li (2010) and Niu and Zhang (2012). Unlike those works in which one specifies a threshold value to determine which are the true change-points in $\mathcal{O}$, the localizing algorithm aims only to help naturally split the data into many subsets.

Denote $\mathcal{O} = \{l_1, \ldots, l_{m-1}\}$, where $m = |\mathcal{O}| + 1$ and set $l_0 = 0$, $l_m = n + 1$. Intuitively, $\mathcal{O}$ provides an overfitting of the true model, say it at least includes a small neighborhood of the

true location set. Thus, the observations in each segment, divided by $\mathcal{O}$, have approximately the same parameters. This motives us to calculate the m estimated parameters $\tilde{\boldsymbol{\beta}}(\mathcal{Z}_{l_k}^{l_{k+1}})$ or average scores $\bar{\mathbf{s}}(\boldsymbol{\gamma}; \mathcal{Z}_{l_k}^{l_{k+1}})$, $k = 0, \ldots, m-1$ and for simplicity they are denoted as $\mathcal{X} = \{\mathbf{X}_1, \ldots, \mathbf{X}_m\}$. By this construction, the correlations among $\mathcal{X}$ are expected to be relatively weak. The original change-point problem is now re-framed into the change-points detection in this mean sequence of size m. Consequently, given a candidate model $\mathcal{M}_L$ specified by $(\tau_1, \ldots, \tau_L)$ and $(\boldsymbol{\mu}_0, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_L)$, the criterion $\mathcal{C}(\mathcal{M}_L; \mathcal{Z})$ can be defined as

$$\mathcal{C}(\mathcal{M}_L; \mathcal{X}) = \sum_{j=0}^{L} \sum_{i=\tau_j+1}^{\tau_{j+1}} N_i \left\| \mathbf{X}_i - \bar{\mathbf{X}}(\tau_j, \tau_{j+1}) \right\|^2 + \sum_{j=0}^{L} \sum_{i=\tau_j+1}^{\tau_{j+1}} N_i \left\| \boldsymbol{\mu}_j - \bar{\mathbf{X}}(\tau_j, \tau_{j+1}) \right\|^2,$$

where $N_i = l_{i+1} - l_i$, and $\bar{\mathbf{X}}(\tau_j, \tau_{j+1}) = \sum_{i=\tau_j+1}^{\tau_{j+1}} N_i \mathbf{X}_i / \sum_{i=\tau_j+1}^{\tau_{j+1}} N_i$ is the weighted sample mean vector of the segment $(\tau_j, \tau_{j+1}]$. The use of N_i distinguishes this objective function from standard least-squares function (8), because the sequence $\{\mathbf{X}_1, \ldots, \mathbf{X}_m\}$ is heterogeneous with the variance of $\mathbf{X}_i$ being approximately proportional to N_i. Then, the proposed CV procedure can be applied.

In this paper, we will use simulations to demonstrate the effectiveness of our proposed algorithm discussed in Sections 3.3.1–3.3.3 but theoretical investigation certainly warrants future research.

4. Numerical results. To evaluate the performance of our proposed COPSS procedure which utilizes a special CV criterion for identifying the number of change-points, we mainly compare with the BIC (or its variants by modifying the loss function and associated penalization term) on a range of simulated and real examples. The two criteria are in conjunction with a wide variety of change-point detection algorithms including OP algorithm (Bai and Perron (2003), Braun, Braun and Müller (2000), Zou et al. (2014)), BS method (Matteson and James (2014)) and its variant the wild binary segmentation (WBS) algorithm (Fryzlewicz (2014)), LD-based detection procedure, the SaRa, proposed by Niu and Zhang (2012), and the PELT (Haynes, Fearnhead and Eckley (2017), Killick, Fearnhead and Eckley (2012)). Several MCP models are considered, reflecting changes in different aspects such as the location, scale, distribution and regression relationship. The data can be univariate, multivariate or in linear model structure, either independent or correlated. Table 1 gives a short preview of all simulated models and the associated CV criteria we will use. For the BIC to be compared, we will either consider (7) with suitable penalty ζ_n tailored for a specific model in Table 1 or refer to the related literature and adopt the default formulation.

To further specify a MCP model in Table 1, we examine two kinds of generation mechanism of the number and locations of change-points (CP).

CP(A). Both the number and locations of change-points are fixed. We adopt the *blocks* setting which is widely used in the literature (Fryzlewicz (2014)). Specifically, $K_n = 11$ and $\mathcal{T}_{K_n}^*/n \approx (0.10, 0.13, 0.15, 0.23, 0.25, 0.40, 0.44, 0.65, 0.76, 0.78, 0.81)$.

CP(B). Both the number and locations of change-points can vary with the sample size n. We set $K_n = \lfloor (\log n)^{1.01} \rfloor$ with $\lfloor x \rfloor$ representing the largest integer not greater than x. The corresponding change-points are set as $\tau_j^* = j\lfloor n/(K_n+1) \rfloor + \text{Uniform}\{-a, a\}$ with $a = \lfloor n^{1/4} \rfloor$ for $j = 1, \ldots, K_n$, where Uniform$\{a, b\}$ with integers a, b denotes the discrete uniform distribution with support $\{a, a+1, \ldots, b\}$.

TABLE 1

Preview of simulated models and the associated CV criteria. Detailed generation of the change signal (such like θ_i's, σ_i's, $\boldsymbol{\theta}_i$'s, $(\alpha_i, \boldsymbol{\beta}_i^\top)$'s, $\boldsymbol{q}_i$'s and F_i's), together with other nuisance parameters (σ, p and q) are deferred in the specific context. The symbols with the superscripts "O" and "E" stand for the quantities based on the $\mathcal{Z}_O$ and $\mathcal{Z}_E$, respectively. For a given L, the change-points $\widehat{\tau}_{L,j}$'s are obtained on the basis of $\mathcal{Z}_O$ by certain change detection algorithm

No.	Model	$\mathcal{C}(\widehat{\mathcal{M}}_L^O; \mathcal{Z}_E)$
I	$X_i = \theta_i + \sigma\varepsilon_i$	$\sum_{j=0}^{L}\sum_{i=\widehat{\tau}_{L,j}^O+1}^{\widehat{\tau}_{L,j+1}^O}(X_i^E - \bar{X}^O_{\widehat{\tau}_{L,j}^O,\widehat{\tau}_{L,j+1}^O})^2$
II	$X_i = \sigma_i\varepsilon_i$	$\sum_{j=0}^{L}\sum_{i=\widehat{\tau}_{L,j}^O+1}^{\widehat{\tau}_{L,j+1}^O}(V_i^E - \bar{V}^O_{\widehat{\tau}_{L,j}^O,\widehat{\tau}_{L,j+1}^O})^2$, $V_i = \log X_i^2$
III	$\mathbf{X}_i = \boldsymbol{\theta}_i + \sigma\boldsymbol{\varepsilon}_i$	$\sum_{j=0}^{L}\sum_{i=\widehat{\tau}_{L,j}^O+1}^{\widehat{\tau}_{L,j+1}^O}\|\mathbf{X}_i^E - \bar{\mathbf{X}}^O_{\widehat{\tau}_{L,j}^O,\widehat{\tau}_{L,j+1}^O}\|^2_{\mathbf{W}_n}$, $\mathbf{W}_n = \operatorname{diag}^{-1}\{\operatorname{Cov}(\mathbf{X})\}$
IV	$Y_i = \alpha_i + \mathbf{X}_i^\top\boldsymbol{\beta}_i + \sigma\varepsilon_i$	$\sum_{j=0}^{L}\sum_{i=\widehat{\tau}_{L,j}^O+1}^{\widehat{\tau}_{L,j+1}^O}\|\mathbf{R}_i^E - \bar{\mathbf{R}}^O_{\widehat{\tau}_{L,j}^O,\widehat{\tau}_{L,j+1}^O}\|^2_{\mathbf{W}_n}$, $\mathbf{R}_i = (Y_i, Y_i\mathbf{X}_i^\top)^\top$, $\mathbf{W}_n = \operatorname{Cov}^{-1}([\mathbf{1}\dot{:}\mathbf{X}])$
V	$\mathbf{X}_i \sim \text{Multinomial}(n_0, \mathbf{q}_i)$	$\sum_{j=0}^{L}\sum_{i=\widehat{\tau}_{L,j}^O+1}^{\widehat{\tau}_{L,j+1}^O}\|\mathbf{X}_i^E - \bar{\mathbf{X}}^O_{\widehat{\tau}_{L,j}^O,\widehat{\tau}_{L,j+1}^O}\|^2_{\mathbf{W}_n}$, $\mathbf{W}_n = \operatorname{diag}^{-1}\{\operatorname{Cov}(\mathbf{X})\}$
VI	$X_i \sim F_i(\cdot)$	$-\int_u[\sum_{j=0}^{L}(\widehat{\tau}_{L,j+1}^O - \widehat{\tau}_{L,j}^O)\{\widehat{F_E}_{\widehat{\tau}_{L,j}^O}^{\widehat{\tau}_{L,j+1}^O}(u)\log(\widehat{F_O}_{\widehat{\tau}_{L,j}^O}^{\widehat{\tau}_{L,j+1}^O}(u)) + (1-\widehat{F_E}_{\widehat{\tau}_{L,j}^O}^{\widehat{\tau}_{L,j+1}^O}(u))\log(1-\widehat{F_O}_{\widehat{\tau}_{L,j}^O}^{\widehat{\tau}_{L,j+1}^O}(u))\}]\,dw(u)$
VII	$X_i = \theta_i + \sigma\varepsilon_i$, $\varepsilon_i s \sim \text{ARMA}(p,q)$	$\sum_{j=0}^{L}\sum_{i=\widehat{\tau}_{L,j}^O+1}^{\widehat{\tau}_{L,j+1}^O}N_i^E(S_i^E - \overline{S}^O_{\widehat{\tau}_{L,j}^O,\widehat{\tau}_{L,j+1}^O})^2$, $S_i = \bar{X}_{l_i,l_{i+1}}$, $\overline{S}_{i_1,i_2} = \sum_{i=i_1+1}^{i_2}N_iS_i/\sum_{i=i_1+1}^{i_2}N_i$, $l_1,\ldots,l_{m-1}$ local minimizers

We fix $K_n^{\mathrm{U}} = 20$ unless otherwise specified. For each example, 1000 replications is used to approximate the distribution of $\hat{K}_n - K_n$, where $\hat{K}_n$ is obtained by either the BIC or our proposed COPSS procedure in conjunction with the change-point detection algorithms under various examples specified in Table 1.

4.1. *Univariate examples*.

4.1.1. *Mean change-point model*. Detecting mean shift in a univariate time-series has been widely discussed in the literature. In this section, four commonly used detection algorithms, the OP, BS, WBS and PELT, are investigated. We consider the ready-made R-packages "wbs" and "changepoint," which implement the WBS and the PELT methods, respectively. We apply the conventional BIC, see (7), for comparison. As we mentioned earlier, the optimal penalization magnitude usually varies from the model and error distribution. To get a broader picture of the performance comparison, we choose the penalty term $\zeta_n = (\log n)^\alpha$ with $\alpha = 1, 1.3, 1.5$, as the order of magnitude $\log n$ has been shown to have superior performance when the noises are independently and identically distributed (i.i.d.) normal random variables (Fryzlewicz (2014)). To implement the PELT in conjunction with the newly proposed COPSS procedure, we follow the guidelines in Section 3.3.1 and consider a range of penalty values and choose the one yielding the minimum squared prediction

error. For the other three algorithms, we apply each in the training step, examining them one by one.

Model I-CP(A) is considered, where we set $n = 2048$. The signal function θ_i's are chosen as a piecewise constant function with $K_n = 11$ and the scale parameter σ is taken to be 7. Four scenarios of the error distribution are considered: (i) $\varepsilon_i \overset{\text{iid}}{\sim} N(0,1)$, (ii) $\varepsilon_i \overset{\text{iid}}{\sim} \sqrt{3}\,\text{Uniform}(-1,1)$, (iii) $\varepsilon_i \overset{\text{iid}}{\sim} \sin(2\pi i/n)/\sum_{j=1}^{n}\sin(2\pi j/n)\cdot N(0,1)$ and (iv) $\varepsilon_i \overset{\text{iid}}{\sim} 0.25t_3$, where $\text{Uniform}(a,b)$ is the continuous uniform distribution with support $[a,b]$ and t_ν is the Student's t-distribution with the degree of freedom ν.

Table 2 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(A) with Scenario (i). First of all, we observe that, in terms of the probability of correctly identifying the true number of change-points, the performance of the BIC could be seriously affected by different choices of penalization magnitude for every detection algorithm. The COPSS performs reasonably well with the OP or WBS algorithm, and has higher probability of correct identification than the BIC with $\alpha = 1.3$ or 1.5 with the OP. As we can expect, the BIC with the conventional choice of $\alpha = 1$ performs better than the COPSS under Scenario (i), that is, the normal error. This can be understood because the COPSS is in a data-driven nature; sacrificing certain estimation precision due to the use of sample-splitting. Especially under the CP(A), there are a few short segments

TABLE 2

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(A) are considered. Procedure using the BIC is named by the rule "Algorithm-BIC-α," where α is the tuning parameter appeared in the penalty; "Algorithm-CV" stands for an detection algorithm followed by the COPSS procedure; we also report the corresponding algorithm but with only a single $\mathcal{C}(\widehat{\mathcal{M}}_L^O; \mathcal{Z}_E)$- or $\mathcal{C}(\widehat{\mathcal{M}}_L^E; \mathcal{Z}_O)$-criterion, termed as "Algorithm-CV-O" and "Algorithm-CV-E," respectively*

	$\hat{K}_n - K_n$									
Procedures	≤ -3	-2	-1	0	1	2	≥ 3	Mean	SD	MSE
OP-BIC-1	0.0	0.0	3.0	93.5	3.1	0.4	0.0	0.07	0.28	0.08
OP-BIC-1.3	0.0	0.1	34.7	65.1	0.1	0.0	0.0	0.35	0.48	0.35
OP-BIC-1.5	0.3	5.7	75.4	18.6	0.0	0.0	0.0	0.88	0.49	1.01
OP-CV-O	0.0	0.5	25.4	59.8	10.9	2.0	1.4	0.46	0.79	0.63
OP-CV-E	0.1	0.2	24.9	59.7	10.6	3.1	1.4	0.47	0.80	0.65
OP-CV	0.0	0.0	24.8	66.2	7.5	1.3	0.2	0.35	0.61	0.39
BS-BIC-1	0.0	0.0	3.8	65.7	26.5	3.8	0.2	0.39	0.61	0.47
BS-BIC-1.3	0.0	0.2	39.1	53.2	7.0	0.5	0.0	0.47	0.62	0.49
BS-BIC-1.5	0.5	4.7	77.6	16.4	0.8	0.0	0.0	0.89	0.50	1.02
BS-CV-O	0.0	0.4	13.8	30.8	24.6	17.2	13.2	1.30	1.67	3.80
BS-CV-E	0.1	0.2	12.9	28.7	28.6	15.3	14.2	1.31	1.64	3.79
BS-CV	0.0	0.0	9.9	27.7	32.6	18.6	11.2	1.20	1.32	2.75
WBS-BIC-1	0.0	0.0	5.1	87.6	6.5	0.8	0.0	0.13	0.38	0.15
WBS-BIC-1.3	0.0	0.1	32.4	66.6	0.9	0.0	0.0	0.34	0.49	0.34
WBS-BIC-1.5	0.4	4.5	74.8	20.2	0.1	0.0	0.0	0.85	0.49	0.96
WBS-CV-O	0.0	0.3	26.9	44.9	15.5	6.9	5.5	0.78	1.28	1.68
WBS-CV-E	0.1	0.4	27.7	41.9	15.8	6.9	7.2	0.89	1.46	2.24
WBS-CV	0.0	0.1	25.6	48.1	17.4	6.0	2.8	0.65	1.00	1.02
PELT-CV-O	0.0	0.4	25.5	65.3	5.7	1.8	1.3	0.40	0.74	0.56
PELT-CV-E	0.0	0.5	25.7	65.5	5.9	1.9	0.5	0.38	0.67	0.47
PELT-CV	0.0	0.1	25.2	68.9	5.0	0.8	0.0	0.32	0.55	0.34

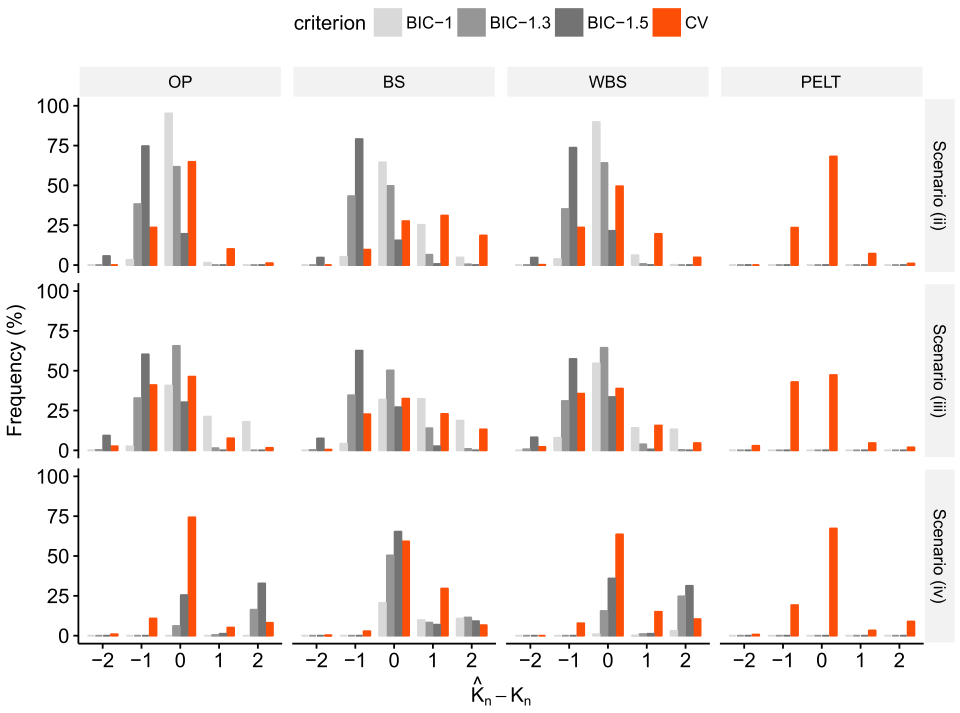

FIG. 1. *Distribution of $\hat{K}_n - K_n$ for the BIC and our CV criterion in conjunction with the OP, BS, WBS and PELT algorithms under Scenarios* (ii)–(iv) *of Model I-CP(A).*

whose length is only around 40. We only got 20 samples to fit the change-point models in such segments and thus may be inefficient. As a consequence, the probability of missing one change-point is a little high compared to the best BIC. Moreover, we found the PELT tends to overestimate the number of change-points; the values of $\hat{K}_n - K_n$ are almost all greater than 2 for all the ζ_n's and thus we omit those results in Table 2. This phenomenon has also been reported by Fryzlewicz (2014) in all of the examples he studied. Interestingly, using the COPSS procedure, this overfitting tendency disappeared and the probability of correct identification is even slightly higher than the OP with the COPSS.

The superiority of the BIC with $\alpha = 1$ does not always hold. Figure 1 depicts the distribution of $\hat{K}_n - K_n$ under Scenarios (ii)–(iv), which reveals that the performance of the BIC with $\alpha = 1$ is no longer the best and may be outperformed by the COPSS (corresponding to CV in Figure 1) for most cases. Now, we can find the BIC with $\alpha = 1$ performs the best under Scenario (ii), that is, light-tailed noises, as in Scenario (i); while the BIC with $\alpha = 1.3$ favors Scenario (iii), that is, the heterogeneous case; and finally the BIC with $\alpha = 1.5$ best suits Scenario (iv), that is, heavy-tailed noise. Consequently, the "oracle" penalty always differs from error to error and thus is not available when one has little knowledge about the data. In contrast, the COPSS is clearly more robust from Table 2 and Figure 1, benefiting from automatically adapting to the model and error distribution. Similar results under Model I-CP(B) are provided in the Supplementary Material, from which we can also conclude that the COPSS could achieve consistent estimation of K_n.

Table 2 also reports the results of the chosen algorithms followed by only a single $\mathcal{C}(\widehat{\mathcal{M}}_L^O; \mathcal{Z}_E)$- or $\mathcal{C}(\widehat{\mathcal{M}}_L^E; \mathcal{Z}_O)$-criterion, which reveals that our "crossed" training-validation procedure (the CV) indeed results in variance reduction.

4.1.2. *Variance change-point model.* Ideas of detecting changes in mean can be easily extended to the variance change-point problem (Chen and Gupta (1997)). To facilitate the comparison, we consider again the PELT method with penalty values specified as

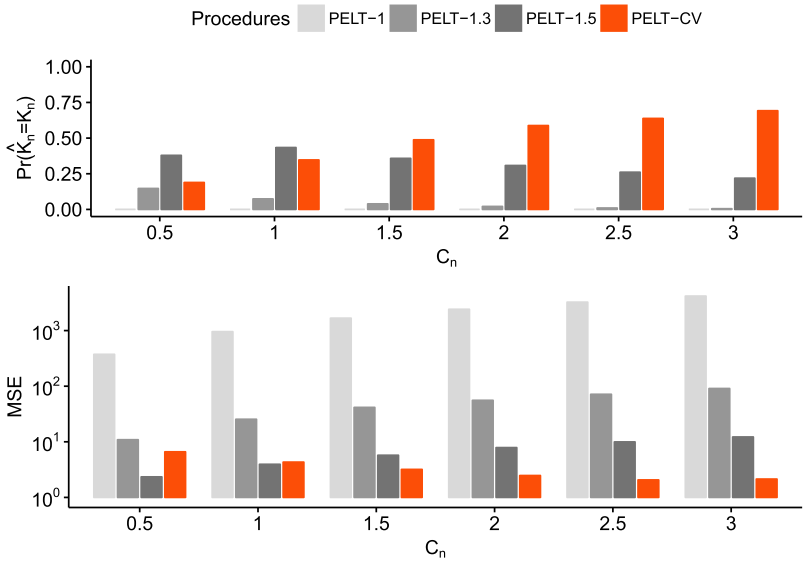

FIG. 2. *Probability of correct identification and the MSE of $\hat{K}_n - K_n$ against the sample size $n = C_n \cdot 2048$ for the PELT method and its CV implementation under Model II-CP(A).*

$\zeta_n = (\log n)^\alpha$, $\alpha = 1, 1.3, 1.5$, using the function "cpt.var()" in the R-package "changepoint" as discussed in Section 3.3.1, we search a range of penalty values and use the CV criterion in Table 1 to choose the best-fit model in order to implement the COPSS.

We take Model II with CP(A) as an illustration example, where we vary $n = C_n \cdot 2048$ over a range of values $C_n = 0.5, 1, 1.5, \ldots, 5$. The scale signal function σ_i's are chosen as a piecewise constant function with breaks at the $K_n = 11$ change-points and values between change-points 1, 0.25, 1, 5, 1, 0.25, 1, 5, 1, 0.25, 1, 5. The noises are independently generated as standardized t_5.

Figure 2 depicts the probability of correct identification and the MSE of $\hat{K}_n - K_n$ against the sample size n for the PELT and its CV implementation. Again, we observe that the performance of the PELT is sensitive to the penalization magnitude and unstable as the sample size varying. The detection ability of the PELT with $\alpha = 1.5$ appears better than our CV implementation when $n = 1024, 2048$ and exhibits a slightly increasing trend, but then drops significantly as n continues to increase. In contrast, our CV criterion presents a steady growth in the detection accuracy as more and more samples are gathered. In the meantime, the MSE of our procedure decreases fast.

4.2. *Multivariate examples.*

4.2.1. *Multivariate mean change-point model.* MCP problem for multivariate observations has gained more and more attention as well. In this section, we compare the COPSS in conjunction with the OP algorithm with a nonparametric method, ECP, proposed by Matteson and James (2014). The ECP method involves specifying the level at which to sequentially test if a proposed change point is statistically significant. In our simulation study, we use the default value 0.05 (see the R package "ecp").

Model III with CP(A) is used here, where we fix $n = 1024$ and 2048. For simplicity, each dimension of the signals $\boldsymbol{\theta}_i$'s are generated as the same as the signals θ_i's used in Model I-CP(A). Two scenarios for the error distribution are considered: (i) $\boldsymbol{\varepsilon}_i = (\boldsymbol{\varepsilon}_{i1}^\top, \boldsymbol{\varepsilon}_{i2}^\top)^\top$. $\boldsymbol{\varepsilon}_{i1} \overset{\text{iid}}{\sim} N_{d_1}(\mathbf{0}, \boldsymbol{\Sigma}_1)$ with $d_1 = \lfloor d/2 \rfloor$ and $\boldsymbol{\Sigma}_1 = (0.5^{|i-j|})$, $\boldsymbol{\varepsilon}_{i2} \overset{\text{iid}}{\sim} N_{d_2}(\mathbf{0}, \boldsymbol{\Sigma}_2)$ with $d_2 = d - d_1$

TABLE 3
Distribution of $\hat{K}_n - K_n$ with its MSE for the ECP procedure and the COPSS (labelled as OP-CV) in conjunction with the OP algorithm under Model III-CP(A)

			$d=5$						$d=10$					
			$\hat{K}_n - K_n$						$\hat{K}_n - K_n$					
Scenario	n	Procedure	-2	-1	0	1	2	MSE	-2	-1	0	1	2	MSE
(i)	1024	ECP	0.0	0.0	94.1	5.0	0.8	0.10	0.0	0.0	94.0	4.6	1.4	0.10
		OP-CV	0.3	10.8	84.5	3.3	1.1	0.20	4.2	36.9	53.1	4.0	1.2	0.69
	2048	ECP	0.0	0.0	94.7	3.3	1.7	0.13	0.0	0.0	94.3	3.4	2.3	0.13
		OP-CV	0.0	2.1	95.8	1.7	0.4	0.05	0.6	14.3	80.5	1.9	2.7	0.29
(ii)	1024	ECP	0.0	0.3	92.4	5.6	1.6	0.13	0.0	3.4	89.4	6.3	0.9	0.13
		OP-CV	0.0	7.9	87.4	4.4	0.3	0.14	0.0	28.1	69.4	2.5	0.0	0.31
	2048	ECP	0.0	0.0	92.9	5.6	1.3	0.13	0.0	0.0	91.3	6.5	2.0	0.16
		OP-CV	0.0	0.0	97.0	2.6	0.3	0.05	0.0	0.9	97.8	1.3	0.0	0.02

and $\mathbf{\Sigma}_2 = 0.3\mathbf{I}_{d_2} + 0.7\mathbf{1}_{d_2}\mathbf{1}_{d_2}^\top$, and $\boldsymbol{\varepsilon}_{i1}$ and $\boldsymbol{\varepsilon}_{i2}$ are independent, where $\mathbf{1}_d$ denotes the d-variate vector with all the components being one; (ii) $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \ldots, \varepsilon_{id})^\top$, where $\varepsilon_{i1}, \ldots, \varepsilon_{id_1} \overset{\text{iid}}{\sim} N(0,1)$, $\varepsilon_{i,d_1+1}, \ldots, \varepsilon_{id} \overset{\text{iid}}{\sim} 0.6t_5$. We set the dimension $d = 5, 10$ and adjust the scale parameter to $\sigma = 2.8\sqrt{d}$.

Table 3 presents the distribution of $\hat{K}_n - K_n$ with its MSE for the ECP procedure and the COPSS in conjunction with the OP algorithm under Scenarios (i)–(ii) with different configurations of (n, d). In terms of the probability of correct identification, the ECP performs quite robust and better than our approach when n is relatively small, while it is clear that the performance of the COPSS will significantly improve, even outperforms the ECP, when the sample size is doubled. In fact, the ECP can be also viewed as a "data-driven" procedure from the aspect of determining the number of change-points because it uses a permutation step to approximate the distribution of the test statistic. Hence, the ECP is more computationally extensive than the COPSS. Figure S2 in the Supplementary Material reports how the run-time (in seconds) changes with the sample size $n = C_n \cdot 2048$ of both procedures under Scenario (i) for one replication using an Inter Xeon E5-2650v4 CPU. Our method is significantly faster and the advantage is more prominent as n increases.

4.2.2. *Change-point in regression coefficients.* Another widely studied example is identifying structural breaks in regression model; see Bai and Perron (1998, 2003) for example. In this section, we perform the OP algorithm described by Bai and Perron (2003) in conjunction with their BIC and our CV criterion. For convenience, we will use the OP algorithm implemented in the R package "strucchange" for both criteria (for our CV criterion, this OP algorithm is used in the training step). For the BIC, we consider the conventional penalty "the number of parameters $\times \log n$."

We investigate Model IV with CP(B), where $n = 512, 1024$ and thus $K_n = 6, 7$ respectively. We consider the signal vector used in Model I-CP(A), that is, $\boldsymbol{\gamma} = (0, 14.64, -3.66, 7.32, -7.32, 10.98, -4.39, 3.29, 19.03, 7.68, 15.37)$ and let γ_{k-1} denote the kth element of $\boldsymbol{\gamma}$. We set $\alpha_i \equiv 0$ and $\boldsymbol{\beta}_i = \gamma_{\text{mod}(J_0+j,11)}$ for $\tau_j^* < i \le \tau_{j+1}^*$, $j = 0, \ldots, K_n$, where J_0 is an integer randomly sampling from $\{1, \ldots, 11\}$ and $\text{mod}(a, b)$ is the modulo operator. Hence, the signals is allowed to be random for each simulated replication. The covariate $\mathbf{X}_i$'s are generated as $\mathbf{X}_i \sim \sqrt{3}\sigma_{\mathbf{X}}\{\text{Uniform}(-1, 1) + \delta\}$ with $\delta = 0$ and 1 corresponding to the "Zero mean" and "Nonzero mean" situations, respectively, where $\sigma_X = 0.5\,\text{SD}(\boldsymbol{\beta}_i\text{s})$ and $\text{SD}(\{x_1, \ldots, x_n\})$

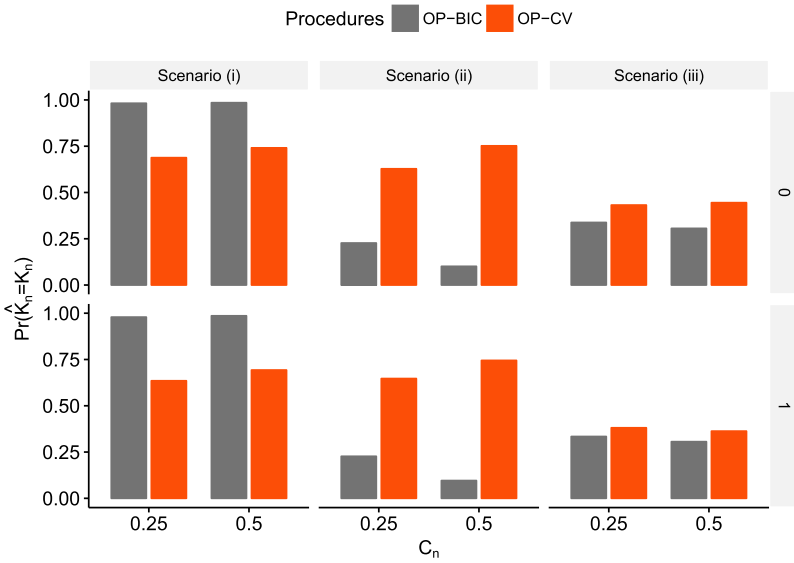

FIG. 3. *Probability of correct identification under Model IV-CP(B) when* $n = C_n \cdot 2048$.

denotes the sample standard deviation of $\{x_1, \ldots, x_n\}$. Three scenarios for the error distribution are considered here: (i) $\varepsilon_i \overset{\text{iid}}{\sim} N(0, 1)$, (ii) $\varepsilon_i \overset{\text{iid}}{\sim} t_3$, and (iii) an AR(1) sequence with coefficient 0.5 and $N(0, 1)$ innovations and the noises are standardized to have unit variance. Finally, the scale parameter σ is chosen such that $\mathrm{SD}(\{\mathbf{X}_i^\top \beta_i\}_{i=1}^n)/\mathrm{SD}(\sigma\varepsilon_i \mathrm{s}) = 3$ to control the signal-to-noise ratio.

Figure 3 depicts $\Pr(\hat{K}_n = K_n)$ for the BIC and our CV criterion under different scenarios. First, the performances of both procedures are not sensitive to the mean of response ($\delta = 0$ or 1). Second, the BIC with the default penalization magnitude performs very well with normal noises, while it is outperformed by the CV under Scenarios (ii)–(iii). This demonstrates that the order of the penalization magnitude $\log n$ may not be sufficient large to avoid overfitting under the heavy-tailed or correlated noises. Third, the detection accuracy of our procedure usually gets improved as the sample size increases.

4.2.3. *Changes in multinomial distributions*. In this section, we consider an example of MCP for multinomial distributions, where the variance of the observations depends on their mean. Braun, Braun and Müller (2000) embed this problem into a quasi-likelihood formulation and utilized the minimum deviance rule to fit the model. To determine the number of change-points, they also adopted the BIC with a penalty $\zeta_n = 0.5n^\alpha$. In particular, they considered the multinomial observations, that is, Model V in Table 1, and aimed to identify the breaks causing the changes in the probability vectors $\mathbf{q}_i$'s. They recommend using $\alpha = 0.23$ based on extensive simulations, which will be served as a benchmark for our comparison. For the COPSS, we adopt their algorithm in the training step, that is, given a candidate model size L, we obtain the estimated change-points by minimizing the corresponding quasi-deviance on the training samples.

Model V with CP(B) is used here, where we fix $n = 1000$ and vary n_0 over a range of values 40, 60, 800, 100, and the number of outcomes (i.e., the dimension of $\mathbf{q}_i$'s) takes value in 2, 4, 10. Under CP(B), $K_n = 7$ and the locations of change-points vary from replication to replication. We follow the mechanism in Braun, Braun and Müller (2000) to generate $\mathbf{q}_i$'s. For each replication, the initial mean vector $\mathbf{q} = (q_1, \ldots, q_d)^\top$ was obtained by normalizing a set of uniform deviates, that is, $q_k = U_k / \sum_{l=1}^d U_l$ for $k = 1, \ldots, d$, where $U_k \sim \mathrm{Uniform}(0, 1)$. Jumps were made on the logistic scale, and the resulting vectors are

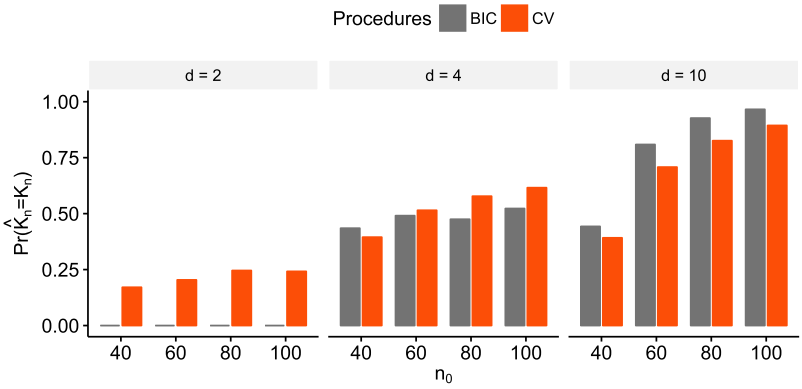

FIG. 4. *Probability of correct identification against the number of experiments n_0 under Model V-CP(B).*

normalized. To be specific, a new mean vector, say $\mathbf{q}'_k = (q'_1, \ldots, q'_d)^\top$, was obtained by normalizing expit(logit $q_k + U'_k$) for $k = 1, \ldots, d$, where $U'_k \sim \text{Uniform}(-J, J)$, logit is the logistic transform, and expit is its inverse. We specify the jump size $J = 1.2/\sqrt{d}$.

Figure 4 plots the probability $\Pr(\hat{K}_n = K_n)$ against the number of experiments n_0 for the BIC and our CV criterion under different number of outcomes, which again indicates that the BIC procedure is sensitive to the model variation but the performance of the COPSS (labelled as CV in Figure 4) in methodology is relatively stable.

4.3. *Extensions.*

4.3.1. *MCP for nonparametric models.* Here we consider the nonparametric MCP setting as described in Section 3.3.2. Zou et al. (2014) proposed a nonparametric maximum likelihood approach, NMCD, which used the BIC in conjunction with the OP algorithm to determine the number of change-points and they recommended using a penalty $\zeta_n = (\log n)^{2.01}/2$. Later, Haynes, Fearnhead and Eckley (2017) showed how the PELT can be applied to the NMCD and proposed the ED-PELT algorithm. The authors also pointed out that "the PELT requires a penalty to avoid under/over-fitting the model which can have a detrimental effect on the quality of the detected change-points." They then suggested using the CROPS algorithm (Haynes, Eckley and Fearnhead (2017)), which performs many PELTs for penalty values across a continuous range. In Haynes, Fearnhead and Eckley (2017), they used a "graphical" approach suggested by Lavielle (2005) in order to choose the best segmentation, which remains heuristic. In what follows, we show that the COPSS with the criterion (11) could be helpful in this case. Specifically, we apply the idea as illustrated in Section 3.3.1 to specify the optimal penalty values by running the ED-PELT over a range of candidate values, denoted as ED-PELT-CV. For comparison, we use the ED-PELT (the R-package "changpoint.np") with the penalty terms $\zeta_{n1} = 2\log n$ and $\zeta_{n2} = (\log n)^{2.01}/2$ as benchmarks (Haynes, Fearnhead and Eckley (2017)).

For Model VI, we consider a simple substitution by adopting similar settings in Model I. The change-points generation mechanism is taken as CP(A) with $K_n = 11$, and the sample size n is chosen to be $n = C_n \cdot 1000$ over a range of values $C_n = 1, \ldots, 10$. We further specify the signal function as what we used in Model I, and generate the noises as (i) independent normal or (ii) AR(1) sequence with coefficient 0.5 and $(\chi^2_1 - 1)/\sqrt{2}$ innovations. The scale parameter σ is specified so that $\text{SD}(\theta_i \text{s})/\text{SD}(\sigma\varepsilon_i \text{s}) = 1$.

Figure 5 depicts the quantity $\Pr(\hat{K}_n - K_n)$ against the sample size n for the ED-PELT-CV and the ED-PELT with two penalties. The ED-PELT with $2\log n$ penalty does not perform well as it appears to be too small to avoid underfitting. The penalty ζ_{n2} can provide accurate

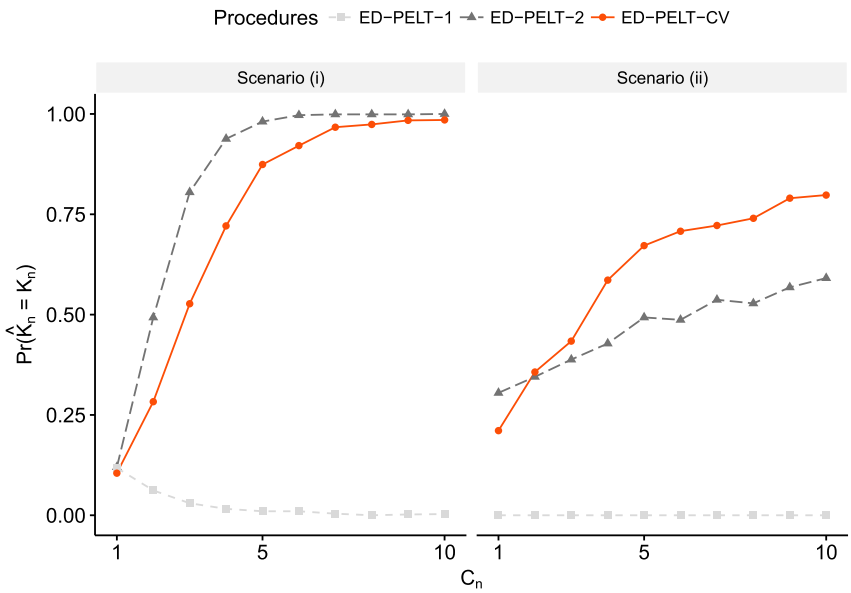

FIG. 5. *Probability of correct identification against the sample size n under Model VI-CP(A), where ED-PELT-1 and ED-PELT-2 stand for the ED-PELT with penalties ζ_{n1} and ζ_{n2} respectively.*

identification with the independence errors, but it is not an ideal one in the autoregressive case. The ED-PELT algorithm combined with the CV procedure performs reasonably well in most cases, and outperforms the benchmarks by a quite large margin when the independence assumption is violated, which again demonstrates its adaptiveness in practice.

4.3.2. *Changes for correlated sequences.* As a final simulation example, we investigate the performance of our modified CV criterion suggested in Section 3.3.3 for cases when unknown correlations exist. To implement the localizing algorithm, we consider the SaRa procedures, that is, using simple local two-sample mean test-statistics. The bandwidth h in SaRa is chosen as $h = \lfloor \log(n) \rfloor$. Once obtaining the set of the most influential points $\mathcal{O}$, we apply the OP algorithm in conjunction with our CV criterion (Table 1) to identify the number of change-points. We name the above procedure as "SaRa-OP-CV." As a benchmark, we also apply the SaRa with $h = \lfloor \log(n) \rfloor$ directly to identify the best model, whose size is determined by the BIC with the penalty $\zeta_n = \log n$. This procedure is named as "SaRa-BIC."

Model VII with CP(A) is considered here, where we vary the sample size $n = C_n \cdot 2048$ over a range of values $C_n = 5, 10, 15, 20$. The signal function θ_i is as the same as in Model I again, and the error ε is specified as ARMA(1, 1) with parameters (ϕ, φ) and innovations $\epsilon \sim N(0, \sigma_\epsilon^2)$. The scale parameter σ is specified such as $\mathrm{SD}(\theta_i \mathrm{s})/\mathrm{SD}(\sigma \varepsilon_i \mathrm{s}) = 1$. Four scenarios for the parameters $(\phi, \varphi, \sigma_\epsilon)$ are considered: (i) $(0.9, 0.5, 0.30)$, (ii) $(-0.9, 0.5, 0.74)$, (iii) $(-0.9, -0.5, 0.30)$ and (iv) $(0.9, -0.5, 0.74)$ such that $\mathrm{Var}(\varepsilon_i) \approx 1$.

Figure 6 presents the boxplot of $\hat{K}_n - K_n$ against the sample size $n = C_n \cdot 2048$ for the SaRa-BIC and the SaRa-OP-CV procedures under Scenarios (i)–(iv), from which we observe that the SaRa-BIC performs unstably and tends to overestimate the number of change-points except under Scenario (iii). In contrast, our SaRa-OP-CV procedure yields estimates fluctuating around the true number of change-points, and the variation significantly reduces as the sample size n increases.

4.4. *Real-data examples.* Here we revisit two examples appeared in the literature for illustration. The first dataset, FTSE100, is contained in the R package "changepoint," which

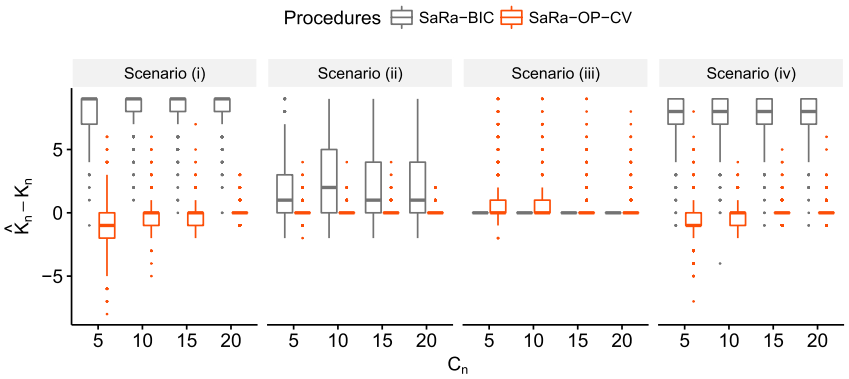

FIG. 6. *Boxplot of $\hat{K}_n - K_n$ against the sample size $n = C_n \cdot 2048$ for the SaRa-BIC and the SaRa-OP-CV procedures under Scenarios* (i)–(iv) *of Model VII.*

gives the daily returns of the UK FTSE 100 index from the Apr. 2, 1984 until Sep. 13, 2012. Our interest is to detect any changes in the variance. We first implement the PELT method with two penalty values $\zeta_{n1} = \log n$ and $\zeta_{n2} = 2\log n$ which are two default values in the "changepoint" package. The estimated number of change-points are, 80 and 32, respectively, which differ much. We then run the PELT over a sequence of penalty values combined with our CV procedure. By specifying the penalty yielding the minimum squared prediction error, we obtain an estimate of the number of change-points as 30, which is quite close to the estimate given under the penalty ζ_{n2}.

The second one is the example in Zou et al. (2014), where the authors considered detecting possible changes in the proportion of the G + C composition of a human chromosome sequence. The ED-PELT algorithm with a penalty $\zeta_n = (\log n)^{2.1}/2$ identifies 40 change-points, while the COPSS procedure detects 37 change-points. By further examining the prediction error in our CV criterion, we found the errors under the models with 37 and 40 candidate change-points are quite close. These two examples suggest that the COPSS is indeed able to provide a practical guide to determine the change-point number if no knowledge about the data is available.

5. Concluding remarks. Determination of the number of change-points is a long-standing problem. This paper proposes a CV-based procedure, COPSS, to select the number of change-points under a unified framework. Interestingly, the COPSS is shown to be consistent under mild conditions, and thus it could serve as a useful alternative to the classical BIC or ad-hoc graphical approaches in practice. We conclude the article with three remarks. First, our unified framework is developed using the score function. Though it is well recognized that in many cases the score- and likelihood- (loss-) based methods are approximately equivalent, the former may be sub-efficient especially when some nuisance parameters present. Thus, it is of interest to thoroughly compare the finite-sample performance of the proposed method with the likelihood-based method under some cases that the computation of $\tilde{\boldsymbol{\beta}}(\mathcal{Z}_{\tau_j}^{\tau_{j+1}})$'s is stable and fast. Second, our numerical results show that the CV procedure may also work well under large-dimensional or autocorrelated scenarios. Theoretical investigation is another interesting topic for future study. Third, though the COPSS procedure is developed under the parametric framework (1), some preliminary results given in the Supplementary Material show that it is also applicable for the nonparametric regression with multiple change-points (or called jump detection) (Loader (1996), Müller and Stadtmüller (1999)) in which the model is nonstationary within each segment (Wu and Zhao (2007)). Asymptotic studies on the consistency of the COPSS in such cases are desired.

APPENDIX: PROOFS

Let $\{\mathbf{x}, \mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and $\{\mathbf{y}, \mathbf{y}_1, \ldots, \mathbf{y}_n\}$ be two sets of d-dimensional vectors. Denote by $\mathcal{T}_L$ and $\tilde{\mathcal{T}}_{\tilde{L}}$ two sets of L and $\tilde{L}$ points, respectively, as defined at the end of Section 1. We introduce $\mathcal{S}_{\mathbf{xy}}(\mathcal{T}_L; \mathbf{M}) = \sum_{l=0}^{L} \mathcal{R}_{\mathbf{xy}}(\tau_l, \tau_{l+1}; \mathbf{M})$, where for each $l = 0, \ldots, L$

$$\mathcal{R}_{\mathbf{xy}}(\tau_l, \tau_{l+1}; \mathbf{M}) = \sum_{i=\tau_l+1}^{\tau_{l+1}} (\mathbf{x}_i - \bar{\mathbf{x}}_{\tau_l, \tau_{l+1}})^\top \mathbf{M} (\mathbf{y}_i - \bar{\mathbf{y}}_{\tau_l, \tau_{l+1}}).$$

By further introducing $\#_l$ more points in the sub-interval (τ_l, τ_{l+1}), say $\mathcal{T}_{L,l} = (\tau_{l,1}, \ldots, \tau_{l,\#_l})$, we extend the definition of $\mathcal{R}_{\mathbf{xy}}$ to $\mathcal{R}_{\mathbf{xy}}(\tau_l, \mathcal{T}_{L,l}, \tau_{l+1}; \mathbf{M}) = \sum_{k=0}^{\#_l} \mathcal{R}_{\mathbf{xy}}(\tau_{l,k}, \tau_{l,k+1}; \mathbf{M})$ with the convention of $\tau_{l,0} = \tau_l$ and $\tau_{l,\#_l+1} = \tau_{l+1}$. Moreover, define $\mathcal{S}_{\mathbf{xy}}(\mathcal{T}_L \cup \tilde{\mathcal{T}}_{\tilde{L}}; \mathbf{M}) = \mathcal{S}_{\mathbf{xy}}(\mathrm{sort}(\mathcal{T}_L \cup \tilde{\mathcal{T}}_{\tilde{L}}); \mathbf{M})$. Note that $\mathcal{S}^2_{\mathbf{x}} = \mathcal{S}_{\mathbf{xx}}$ and $\mathcal{R}^2_{\mathbf{x}} = \mathcal{R}_{\mathbf{xx}}$. Lastly, for any point $\tau \in (l, r)$, denote $\tilde{\mathbf{x}}^{\tau}_{l,r} = \sqrt{\frac{(\tau - l)(r - \tau)}{r - l}} (\bar{\mathbf{x}}_{l,\tau} - \bar{\mathbf{x}}_{\tau,r})$.

For notational convenience, we note that our estimation procedure can be reformulated as follows. Suppose we have two independent sets of d-dimensional observations $\{\mathbf{O}_1, \ldots, \mathbf{O}_n\}$ and $\{\mathbf{E}_1, \ldots, \mathbf{E}_n\}$ collected from the following multiple change-point model:

$$\mathbf{O}_i = \boldsymbol{\mu}^*_j + \boldsymbol{\Sigma}^{*1/2}_j \breve{\mathbf{U}}_i,$$

$$\mathbf{E}_i = \boldsymbol{\mu}^*_j + \boldsymbol{\Sigma}^{*1/2}_j \breve{\mathbf{V}}_i, \quad i = \tau^*_j + 1, \ldots, \tau^*_{j+1}, j = 0, \ldots, K_n,$$

where $\breve{\mathbf{U}}_1, \ldots, \breve{\mathbf{U}}_n, \breve{\mathbf{V}}_1, \ldots, \breve{\mathbf{V}}_n$ are independent standardized noises satisfying $\mathrm{E}(\breve{\mathbf{U}}_1) = \mathbf{0}$ and $\mathrm{Var}(\breve{\mathbf{U}}_1) = \mathbf{I}$, and $\breve{\mathbf{U}}_{\tau^*_j+1}, \ldots, \breve{\mathbf{U}}_{\tau^*_{j+1}}, \breve{\mathbf{V}}_{\tau^*_j+1}, \ldots, \breve{\mathbf{V}}_{\tau^*_{j+1}}$ are identically distributed for each $j = 0, \ldots, K_n$. Further let $\boldsymbol{\theta}_i = \boldsymbol{\mu}^*_j$, $\mathbf{U}_i = \boldsymbol{\Sigma}^{*1/2}_j \breve{\mathbf{U}}_i$ and $\mathbf{V}_i = \boldsymbol{\Sigma}^{*1/2}_j \breve{\mathbf{V}}_i$ for $i = \tau^*_j + 1, \ldots, \tau^*_{j+1}$, $j = 0, \ldots, K_n$. Given L, let $\widehat{\mathcal{T}}_L = (\widehat{\tau}_{L,1}, \ldots, \widehat{\tau}_{L,L})$ be the estimated change-points based on $\{\mathbf{O}_1, \ldots, \mathbf{O}_n\}$, the corresponding validation error on $\{\mathbf{E}_1, \ldots, \mathbf{E}_n\}$ can be formulated as

$$\begin{aligned}
\mathrm{Err}(L) &= \sum_{l=0}^{L} \sum_{i=\widehat{\tau}_{L,l}+1}^{\widehat{\tau}_{L,l+1}} (\mathbf{E}_i - \bar{\mathbf{O}}_{\widehat{\tau}_{L,l}, \widehat{\tau}_{L,l+1}})^\top \mathbf{W}_n (\mathbf{E}_i - \bar{\mathbf{O}}_{\widehat{\tau}_{L,l}, \widehat{\tau}_{L,l+1}}) \\
&= \mathcal{S}^2_{\mathbf{E}}(\widehat{\mathcal{T}}_L; \mathbf{W}_n) - \mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_L; \mathbf{W}_n) - \mathcal{S}^2_{\mathbf{V}}(\widehat{\mathcal{T}}_L; \mathbf{W}_n) + 2\mathcal{S}_{\mathbf{UV}}(\widehat{\mathcal{T}}_L; \mathbf{W}_n) \\
&\quad + \sum_{i=1}^{n} (\mathbf{U}_i - \mathbf{V}_i)^\top \mathbf{W}_n (\mathbf{U}_i - \mathbf{V}_i).
\end{aligned}$$

We will suppress the dependence on $\mathbf{W}_n$, which should not cause any confusion. To make the notation more readable, we let i index the observations, j the true change-points, and l the candidate change-points. If j or l has already been used in the former context, we choose k to be a substitution.

Before we present the proof of Theorem 1, we first state some useful lemmas.

LEMMA 1. *Suppose* $\breve{\mathbf{U}}, \breve{\mathbf{U}}_1, \ldots, \breve{\mathbf{U}}_N$ *are i.i.d. such that* $\mathrm{E}(\breve{\mathbf{U}}) = \mathbf{0}$. *If* $\mathrm{E}(\|\breve{\mathbf{U}}\|^{2m}) < \infty$ *for some positive integer* $m \geq 1$, *then as* $N \to \infty$,

$$\max_{0 \leq k_1 < k_2 \leq N} (k_2 - k_1) \|\bar{\breve{\mathbf{U}}}_{k_1, k_2}\|^2 = O_p(N^{2/m}).$$

LEMMA 2 (Multivariate Darling–Erdős theorem). *Suppose* $\breve{\mathbf{U}}, \breve{\mathbf{U}}_1, \ldots, \breve{\mathbf{U}}_N$ *are i.i.d. such that* $\mathrm{E}(\breve{\mathbf{U}}) = \mathbf{0}$ *and* $\mathrm{Var}(\breve{\mathbf{U}}) = \mathbf{I}$. *If* $\mathrm{E}(\|\breve{\mathbf{U}}\|^{2+\alpha}) < \infty$ *for some* $\alpha > 0$, *then*

$$\lim_{N \to \infty} \Pr\Big\{ a_N \max_{1 \leq k \leq N} k^{1/2} \|\bar{\breve{\mathbf{U}}}_{1,k}\| - b_{d,N} \leq t \Big\} = \exp\{-\exp(-t)\},$$

for all t, *where* $a_N = \sqrt{2\log\log N}$, $b_{d,N} = 2\log\log N + d/2\log\log\log N - \log\{\Gamma(d/2)\}$ *and* $\Gamma(\cdot)$ *is the Gamma function*.

LEMMA 3. *Suppose* $\breve{\mathbf{U}}, \breve{\mathbf{U}}_1, \ldots, \breve{\mathbf{U}}_N$ *are i.i.d. such that* $\mathrm{E}(\breve{\mathbf{U}}) = \mathbf{0}$ *and* $\mathrm{Var}(\breve{\mathbf{U}}) = \mathbf{I}$. *If* $\mathrm{E}(\|\breve{\mathbf{U}}\|^3) < \infty$, *then*

$$\max_{1\le k_1<k_2\le N} (k_2-k_1)^{1/2}\|\bar{\breve{\mathbf{U}}}_{k_1,k_2}\| \gtrsim \sqrt{\log N}.$$

Lemma 1 was obtained by Yao and Au (1989) under the univariate case, which can be easily extended to this multivariate version. Lemma 2 was obtained by Horváth (1993), which extends the one-dimensional Darling–Erdős theorem in Darling and Erdös (1956). As a corollary, we conclude that $\max_{1\le k\le N} k\|\bar{\breve{\mathbf{U}}}_{1,k}\|^2 = 2\log\log N\{1+o_p(1)\}$. Lemma 3 presents the lower bound for the terms in Lemma 1, whose proof is deferred in the Supplementary Material. We will repeatedly use the above facts in the proofs of the following lemmas and theorems. The proofs of Lemmas 4 and 5 are also given in the Supplementary Material.

LEMMA 4 (Variation on **E**). *Suppose Assumptions 1–3 hold.*

(i) *For any* $\widehat{\mathcal{T}}_L$ *with* $L < K_n$,

$$\mathcal{S}^2_{\mathbf{E}}\{\widehat{\mathcal{T}}_L \cup \mathcal{T}^*_{K_n}\setminus\tau^*_j \cup \{\tau^*_j - \rho_n\} \cup \{\tau^*_j + \rho_n\}\} - \mathcal{S}^2_{\mathbf{E}}(\mathcal{T}^*_{K_n})$$
$$\ge \frac{\underline{\lambda}_n}{8}\underline{\omega}_n \min_{1\le j\le K_n} \|\boldsymbol{\mu}^*_{j-1} - \boldsymbol{\mu}^*_j\|^2\{1+o_p(1)\},$$

where $\rho_n = \underline{\lambda}_n/4$.

(ii) *For any* $\widehat{\mathcal{T}}_L$ *with* $L \ge 0$, $\mathcal{S}^2_{\mathbf{E}}(\widehat{\mathcal{T}}_L) - \mathcal{S}^2_{\mathbf{E}}(\widehat{\mathcal{T}}_L \cup \mathcal{T}^*_{K_n}) \ge 0$.

(iii) *For any* $\widehat{\mathcal{T}}_L$ *with* $L \ge 0$, $\mathcal{S}^2_{\mathbf{E}}(\mathcal{T}^*_{K_n}) - \mathcal{S}^2_{\mathbf{E}}(\widehat{\mathcal{T}}_L \cup \mathcal{T}^*_{K_n}) = O_p(L\overline{\omega}_n\overline{\sigma})$.

(iv) *For any* $\widehat{\mathcal{T}}_{K_n}$, $\mathcal{S}^2_{\mathbf{E}}(\widehat{\mathcal{T}}_{K_n}) - \mathcal{S}^2_{\mathbf{E}}(\mathcal{T}^*_{K_n}) = o_p(\overline{\omega}_n\overline{\sigma}\log\log\overline{\lambda}_n)$.

LEMMA 5 (Variation on **U**). *Suppose Assumptions 1–3 hold.*

(i) *For any* $\widehat{\mathcal{T}}_L$ *with* $L < K_n$, $\mathcal{S}^2_{\mathbf{U}}(\mathcal{T}^*_{K_n}) - \mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_L \cup \mathcal{T}^*_{K_n}) = O_p(K_n\overline{\omega}_n\overline{\sigma}\overline{\lambda}_n^{-2/m})$ *and* $\mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_L) - \mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_L \cup \mathcal{T}^*_{K_n}) = O_p(K_n\overline{\omega}_n\overline{\sigma}\log\log\overline{\lambda}_n)$.

(ii) *For any* $\widehat{\mathcal{T}}_{K_n}$, $\mathcal{S}^2_{\mathbf{U}}(\mathcal{T}^*_{K_n}) - \mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_{K_n} \cup \mathcal{T}^*_{K_n}) = O_p(K_n\overline{\omega}_n\overline{\sigma}\log\log\delta_{0,n})$ *and* $\mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_{K_n}) - \mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_{K_n} \cup \mathcal{T}^*_{K_n}) = O_p(K_n\overline{\omega}_n\overline{\sigma}\log\log\delta_{0,n})$.

(iii) *For any* $\widehat{\mathcal{T}}_L$ *with* $L = K_n + q$ *and* $q \ge 1$, *then* $\mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_L) - \mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_L \cup \mathcal{T}^*_{K_n}) = \overline{\omega}_n\overline{\sigma}\{o_p(\log\log\overline{\lambda}_n) + O_p(K_n\log\log\delta_{q,n})\}$.

PROOF OF THEOREM 1. For any L, we observe that

$$\mathrm{Err}(L) - \mathrm{Err}(K_n) = \{\mathcal{S}^2_{\mathbf{E}}(\widehat{\mathcal{T}}_L) - \mathcal{S}^2_{\mathbf{E}}(\widehat{\mathcal{T}}_{K_n})\} + \{\mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_{K_n}) - \mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_L)\}$$
$$+ \{\mathcal{S}^2_{\mathbf{V}}(\widehat{\mathcal{T}}_{K_n}) - \mathcal{S}^2_{\mathbf{V}}(\widehat{\mathcal{T}}_L)\} + 2\{\mathcal{S}_{\mathbf{UV}}(\widehat{\mathcal{T}}_L) - \mathcal{S}_{\mathbf{UV}}(\widehat{\mathcal{T}}_{K_n})\}.$$

It suffices to show that for any $L \ne K_n$, $\Pr\{\mathrm{Err}(L) - \mathrm{Err}(K_n) > 0\} \to 1$ as $n \to \infty$. This can be revealed by demonstrating the following facts.

Fact (A). If $L < K_n$, then:

(a) $\mathcal{S}^2_{\mathbf{E}}(\widehat{\mathcal{T}}_L) - \mathcal{S}^2_{\mathbf{E}}(\widehat{\mathcal{T}}_{K_n}) \ge \underline{\lambda}_n/8\underline{\omega}_n \min_{1\le j\le K_n} \|\boldsymbol{\mu}^*_{j-1} - \boldsymbol{\mu}^*_j\|^2\{1+o_p(1)\}$;

(b) $\mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_{K_n}) - \mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_L) = O_p\big(K_n\overline{\omega}_n\overline{\sigma}\overline{\lambda}_n^{2/m}\big)$;

(c) $\mathcal{S}^2_{\mathbf{V}}(\widehat{\mathcal{T}}_{K_n}) - \mathcal{S}^2_{\mathbf{V}}(\widehat{\mathcal{T}}_L) = O_p(K_n\overline{\omega}_n\overline{\sigma})$;

(d) $\mathcal{S}_{\mathbf{UV}}(\widehat{\mathcal{T}}_L) - \mathcal{S}_{\mathbf{UV}}(\widehat{\mathcal{T}}_{K_n}) = O_p\big(K_n\overline{\omega}_n\overline{\sigma}\overline{\lambda}_n^{2/m}\big)$.

Fact (B). If $L = K_n + q$ with $q \geq 1$, then:

(a) $\mathcal{S}^2_{\mathbf{E}}(\widehat{\mathcal{T}}_L) - \mathcal{S}^2_{\mathbf{E}}(\widehat{\mathcal{T}}_{K_n}) = \text{a nonnegative term} + O_p(K_n\overline{\omega}_n\overline{\sigma}) + o_p(\overline{\omega}_n\overline{\sigma}\log\log\overline{\lambda}_n)$

(b) $\mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_{K_n}) - \mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_L) = \mathcal{S}^2_{\mathbf{U}}(\mathcal{T}^*_{K_n}) - \mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_L \cup \mathcal{T}^*_{K_n}) + o_p(\overline{\omega}_n\overline{\sigma}\alpha_{q,n})$;

(c) $\mathcal{S}^2_{\mathbf{V}}(\widehat{\mathcal{T}}_{K_n}) - \mathcal{S}^2_{\mathbf{V}}(\widehat{\mathcal{T}}_L) = O_p(K_n\overline{\omega}_n\overline{\sigma})$;

(d) $\mathcal{S}_{\mathbf{UV}}(\widehat{\mathcal{T}}_L) - \mathcal{S}_{\mathbf{UV}}(\widehat{\mathcal{T}}_{K_n}) = o_p\{\mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_L) - \mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_{K_n})\}$.

Verification of Fact (A). To show (a), consider the following identity:

$$\mathcal{S}^2_{\mathbf{E}}(\widehat{\mathcal{T}}_L) - \mathcal{S}^2_{\mathbf{E}}(\widehat{\mathcal{T}}_{K_n}) = \{\mathcal{S}^2_{\mathbf{E}}(\widehat{\mathcal{T}}_L) - \mathcal{S}^2_{\mathbf{E}}(\mathcal{T}^*_{K_n})\} - \{\mathcal{S}^2_{\mathbf{E}}(\widehat{\mathcal{T}}_{K_n}) - \mathcal{S}^2_{\mathbf{E}}(\mathcal{T}^*_{K_n})\}.$$

We observe that $\mathcal{S}^2_{\mathbf{E}}(\widehat{\mathcal{T}}_L) \geq \mathcal{S}^2_{\mathbf{E}}\{\widehat{\mathcal{T}}_L \cup \mathcal{T}^*_{K_n} \setminus \tau^*_j \cup \{\tau^*_j - \rho_n\} \cup \{\tau^*_j + \rho_n\}\}$. By Lemma 4(i), we have

$$\mathcal{S}^2_{\mathbf{E}}(\widehat{\mathcal{T}}_L) - \mathcal{S}^2_{\mathbf{E}}(\mathcal{T}^*_{K_n}) \geq \frac{\underline{\lambda}_n}{8}\underline{\omega}_n \min_{1\leq j\leq K_n} \|\boldsymbol{\mu}^*_{j-1} - \boldsymbol{\mu}^*_j\|^2\{1 + o_p(1)\}.$$

Then by Lemma 4(iv), (a) follows. (b) follows from Lemma 5(i)–(ii) that

$$\begin{aligned}\mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_L) - \mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_{K_n}) &= \{\mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_L) - \mathcal{S}^2_{\mathbf{U}}(\mathcal{T}^*_{K_n})\} - \{\mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_{K_n}) - \mathcal{S}^2_{\mathbf{U}}(\mathcal{T}^*_{K_n})\} \\ &= O_p\big(K_n\overline{\omega}_n\overline{\sigma}\overline{\lambda}_n^{2/m}\big).\end{aligned}$$

(c) can be obtained as a corollary of Lemma 4 and to verify (d), we just need to notice the following fact:

$$\begin{aligned}&\mathcal{S}_{\mathbf{UV}}(l, r) - \mathcal{S}_{\mathbf{UV}}(l, \tau_1, \ldots, \tau_L, r) \\ &\quad= \sum_{0\leq l_1<l_2\leq L} \frac{N_{\tau_{l_1},\tau_{l_1}+1} + N_{\tau_{l_2},\tau_{l_2}+1}}{r - l}\tilde{\mathbf{U}}^\top_{\tau_{l_1},\tau_{l_1}+1}\mathbf{W}_n\tilde{\mathbf{V}}_{\tau_{l_2},\tau_{l_2}+1}\end{aligned}$$

and $\tilde{\mathbf{U}}^\top\mathbf{W}_n\tilde{\mathbf{V}} \leq (\tilde{\mathbf{U}}^\top\mathbf{W}_n\tilde{\mathbf{U}} + \tilde{\mathbf{V}}^\top\mathbf{W}_n\tilde{\mathbf{V}})/2$.

Verification of Fact (B). By Lemma 4(ii)–(iv), (a) holds. By Lemma 5(ii)–(iii) and Assumption 2, (b) holds. (c) can also be obtained as a corollary of Lemma 4. To verify (d), first we can show that

$$\mathcal{S}_{\mathbf{UV}}(\widehat{\mathcal{T}}_{K_n}) - \mathcal{S}_{\mathbf{UV}}(\widehat{\mathcal{T}}_L) = \{\mathcal{S}_{\mathbf{UV}}(\mathcal{T}^*_{K_n}) - \mathcal{S}_{\mathbf{UV}}(\widehat{\mathcal{T}}_L \cup \mathcal{T}^*_{K_n})\} + o_p(\overline{\omega}_n\overline{\sigma}\alpha_{q,n}),$$

by using arguments similar to those in the verification of (b). By the assumption that $\mathcal{S}^2_{\mathbf{U}}(\mathcal{T}^*_{K_n}) - \mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_L \cup \mathcal{T}^*_{K_n}) \gtrsim \overline{\omega}_n\overline{\sigma}\alpha_{q,n}$, it suffices to show that

$$\mathcal{S}_{\mathbf{UV}}(\mathcal{T}^*_{K_n}) - \mathcal{S}_{\mathbf{UV}}(\widehat{\mathcal{T}}_L \cup \mathcal{T}^*_{K_n}) = o_p\{\mathcal{S}^2_{\mathbf{U}}(\mathcal{T}^*_{K_n}) - \mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_L \cup \mathcal{T}^*_{K_n})\}.$$

In fact, by the Cauchy–Schwarz inequality,

$$\begin{aligned}&|\mathcal{S}_{\mathbf{UV}}(\mathcal{T}^*_{K_n}) - \mathcal{S}_{\mathbf{UV}}(\widehat{\mathcal{T}}_L \cup \mathcal{T}^*_{K_n})| \\ &\quad\leq \{\mathcal{S}^2_{\mathbf{U}}(\mathcal{T}^*_{K_n}) - \mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_L \cup \mathcal{T}^*_{K_n})\}^{1/2}\{\mathcal{S}^2_{\mathbf{V}}(\mathcal{T}^*_{K_n}) - \mathcal{S}^2_{\mathbf{V}}(\widehat{\mathcal{T}}_L \cup \mathcal{T}^*_{K_n})\}^{1/2}.\end{aligned}$$

Hence the fact holds as $\{\mathcal{S}^2_{\mathbf{U}}(\mathcal{T}^*_{K_n}) - \mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_L \cup \mathcal{T}^*_{K_n})\}/(K_n\overline{\omega}_n\overline{\sigma}) \to \infty$.

Finally, according to *Facts (A)–(B)*, we have, with an overwhelming probability, $\widehat{K}_n = K_n$. □

PROOF OF THEOREM 2. First, assume $q \geq 2$. For each $j = 0, \ldots, K_n$, let τ_j and $\tau_{j'}$ be any points such that $\tau^*_j < \tau_j < \tau_{j'} < \tau^*_{j+1}$ and $\mathcal{T}^{(j)}_q = \{\tau_j\} \cup \{\tau_{j'}\} \cup \mathcal{T}_{q-2}$ where $\mathcal{T}_{q-2}$ is a set of $q-2$ points satisfying that each point is located outside the interval $[\tau^*_j, \tau^*_{j+1}]$. By the definition of OP algorithm, we observe $\mathcal{S}^2_{\mathbf{O}}(\widehat{\mathcal{T}}_L) \leq \min_{0\leq j\leq K_n} \min_{\tau^*_j<\tau_j<\tau_{j'}<\tau^*_{j+1}} \mathcal{S}^2_{\mathbf{O}}(\mathcal{T}^*_{K_n} \cup \mathcal{T}^{(j)}_q)$. We observe that

$$\begin{aligned}
\mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_L \cup \mathcal{T}^*_{K_n}) &= \mathcal{S}^2_{\mathbf{O}}(\widehat{\mathcal{T}}_L \cup \mathcal{T}^*_{K_n}) \leq \mathcal{S}^2_{\mathbf{O}}(\widehat{\mathcal{T}}_L) \\
&\leq \min_{0\leq j\leq K_n} \min_{\tau^*_j<\tau_j<\tau_{j'}<\tau^*_{j+1}} \mathcal{S}^2_{\mathbf{O}}(\mathcal{T}^*_{K_n} \cup \mathcal{T}^{(j)}_q) \\
&= \min_{0\leq j\leq K_n} \min_{\tau^*_j<\tau_j<\tau_{j'}<\tau^*_{j+1}} \mathcal{S}^2_{\mathbf{U}}(\mathcal{T}^*_{K_n} \cup \mathcal{T}^{(j)}_q).
\end{aligned}$$

Then, for any j and the corresponding any τ_j and $\tau_{j'}$,

$$\begin{aligned}
&\mathcal{S}^2_{\mathbf{U}}(\mathcal{T}^*_{K_n}) - \mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_L \cup \mathcal{T}^*_{K_n}) \\
&\geq \mathcal{S}^2_{\mathbf{U}}(\mathcal{T}^*_{K_n}) - \mathcal{S}^2_{\mathbf{U}}(\mathcal{T}^*_{K_n} \cup \mathcal{T}^{(j)}_q) \\
&\geq \underline{\omega}_n\underline{\sigma}\{\mathcal{R}^2_{\check{\mathbf{U}}}(\tau^*_j, \tau^*_{j+1}) - \mathcal{R}^2_{\check{\mathbf{U}}}(\tau^*_j, \tau_j, \tau_{j'}, \tau^*_{j+1})\} \\
&\geq \underline{\omega}_n\underline{\sigma}\{(\tau_{j'} - \tau_j)\|\bar{\check{\mathbf{U}}}_{\tau_j,\tau_{j'}}\|^2 - (\tau^*_{j+1} - \tau^*_j)\|\bar{\check{\mathbf{U}}}_{\tau^*_j,\tau^*_{j+1}}\|^2\}.
\end{aligned}$$

Hence, by Lemma 3, $\mathcal{S}^2_{\mathbf{U}}(\mathcal{T}^*_{K_n}) - \mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_L \cup \mathcal{T}^*_{K_n}) \gtrsim \underline{\omega}_n\underline{\sigma}\log(\tau^*_{j+1} - \tau^*_j)$ for any j. And by the assumption that $\liminf_{n\to\infty}(\underline{\omega}_n\underline{\sigma})/(\overline{\omega}_n\overline{\sigma}) > 0$, the conclusion follows. If $q = 1$, we can similarly show that

$$\begin{aligned}
\mathcal{S}^2_{\mathbf{U}}(\mathcal{T}^*_{K_n}) - \mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_L \cup \mathcal{T}^*_{K_n}) &\geq \underline{\omega}_n\underline{\sigma} \max_{0\leq j\leq K_n} \max_{\tau^*_j<\tau<\tau^*_{j+1}} \|\check{\tilde{\mathbf{U}}}^{\tau}_{\tau^*_j,\tau^*_{j+1}}\|^2 \\
&\gtrsim \overline{\omega}_n\overline{\sigma}\log\log\overline{\lambda}_n,
\end{aligned}$$

by using Lemma 2, which complete the proof.

For binary segmentation algorithm, the detection procedure is nested and thus

$$\mathcal{S}^2_{\mathbf{U}}(\mathcal{T}^*_{K_n}) - \mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_L \cup \mathcal{T}^*_{K_n}) \geq \mathcal{S}^2_{\mathbf{U}}(\mathcal{T}^*_{K_n}) - \mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_{K_n+1} \cup \mathcal{T}^*_{K_n}).$$

Again, we have

$$\mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_{K_n+1} \cup \mathcal{T}^*_{K_n}) = \mathcal{S}^2_{\mathbf{O}}(\widehat{\mathcal{T}}_{K_n+1} \cup \mathcal{T}^*_{K_n}) \leq \mathcal{S}^2_{\mathbf{O}}(\widehat{\mathcal{T}}_{K_n+1}).$$

For each $j = 0, \ldots, K_n$, let τ_j be any point such that $\tau^*_j < \tau_j < \tau^*_{j+1}$. By the construction of the algorithm, we know that

$$\mathcal{S}^2_{\mathbf{O}}(\widehat{\mathcal{T}}_{K_n+1}) \leq \min_{0\leq j\leq K_n} \min_{\tau^*_j<\tau_j<\tau^*_{j+1}} \mathcal{S}^2_{\mathbf{O}}(\tau_j \cup \widehat{\mathcal{T}}_{K_n}).$$

We can similarly show that

$$\begin{aligned}
\mathcal{S}^2_{\mathbf{O}}(\tau_j \cup \widehat{\mathcal{T}}_{K_n}) &= \mathcal{S}^2_{\mathbf{O}}(\tau_j \cup \widehat{\mathcal{T}}_{K_n} \cup \mathcal{T}^*_{K_n}) + o_p(\overline{\omega}_n\overline{\sigma}\log\log\overline{\lambda}_n) \\
&= \mathcal{S}^2_{\mathbf{U}}(\tau_j \cup \widehat{\mathcal{T}}_{K_n} \cup \mathcal{T}^*_{K_n}) + o_p(\overline{\omega}_n\overline{\sigma}\log\log\overline{\lambda}_n).
\end{aligned}$$

It follows that

$$\begin{aligned}
&\mathcal{S}^2_{\mathbf{U}}(\mathcal{T}^*_{K_n}) - \mathcal{S}^2_{\mathbf{U}}(\widehat{\mathcal{T}}_{K_n+1} \cup \mathcal{T}^*_{K_n}) \\
&\quad \geq \max_{0\leq j\leq K_n} \max_{\tau_j^*<\tau_j<\tau_{j+1}^*} \{\mathcal{S}^2_{\mathbf{U}}(\mathcal{T}^*_{K_n}) - \mathcal{S}^2_{\mathbf{U}}(\tau_j \cup \widehat{\mathcal{T}}_{K_n} \cup \mathcal{T}^*_{K_n})\} + o_p(\overline{\omega}_n\overline{\sigma}\log\log\overline{\lambda}_n) \\
&\quad \geq \max_{0\leq j\leq K_n} \max_{\tau_j^*<\tau_j<\tau_{j+1}^*} \{\mathcal{S}^2_{\mathbf{U}}(\mathcal{T}^*_{K_n}) - \mathcal{S}^2_{\mathbf{U}}(\tau_j \cup \mathcal{T}^*_{K_n})\} + o_p(\overline{\omega}_n\overline{\sigma}\log\log\overline{\lambda}_n) \\
&\quad \geq \underline{\omega}_n\underline{\sigma} \max_{0\leq j\leq K_n} \max_{\tau_j^*<\tau<\tau_{j+1}^*} \|\tilde{\mathbf{U}}^{\tau}_{\tau_j^*,\tau_{j+1}^*}\|^2 + o_p(\overline{\omega}_n\overline{\sigma}\log\log\overline{\lambda}_n).
\end{aligned}$$

□

Acknowledgments. The authors are grateful to the referees, Associate Editor and Editor for their insightful comments that have significantly improved the article.

The first author was supported by NNSF of China Grants 11690015, 11622104 and 11431006, and NSF of Tianjin Grant 18JCJQJC46000.

The third author was supported by NIH Grants P50 DA039838, U19AI089672 and T32 LM012415, and an NSF Grant DMS-1820702.

Guanghui Wang is the corresponding author.

SUPPLEMENTARY MATERIAL

Supplement to "Consistent selection of the number of change-points via sample-splitting" (DOI: 10.1214/19-AOS1814SUPP; .pdf). The Supplementary Material contains the proofs of all the technical lemmas and additional simulation results.

REFERENCES

ARLOT, S. and CELISSE, A. (2011). Segmentation of the mean of heteroscedastic data via cross-validation. *Stat. Comput.* **21** 613–632. MR2826696 https://doi.org/10.1007/s11222-010-9196-x

AUE, A. and HORVÁTH, L. (2013). Structural breaks in time series. *J. Time Series Anal.* **34** 1–16. MR3008012 https://doi.org/10.1111/j.1467-9892.2012.00819.x

AUGER, I. E. and LAWRENCE, C. E. (1989). Algorithms for the optimal identification of segment neighborhoods. *Bull. Math. Biol.* **51** 39–54. MR0978902 https://doi.org/10.1016/S0092-8240(89)80047-3

BAI, J. (1998). Estimation of multiple-regime regressions with least absolutes deviation. *J. Statist. Plann. Inference* **74** 103–134. MR1665123 https://doi.org/10.1016/S0378-3758(98)00082-2

BAI, Z., FUJIKOSHI, Y. and CHOI, K. P. (2017). Consistency of AIC and BIC in estimating the number of significant components in high-dimensional principal component analysis. *Ann. Statist.* **46** 1050–1076. MR3797996

BAI, J. and PERRON, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica* **66** 47–78. MR1616121 https://doi.org/10.2307/2998540

BAI, J. and PERRON, P. (2003). Computation and analysis of multiple structural change models. *J. Appl. Econometrics* **18** 1–22.

BIRGÉ, L. and MASSART, P. (2001). Gaussian model selection. *J. Eur. Math. Soc.* (*JEMS*) **3** 203–268. MR1848946 https://doi.org/10.1007/s100970100031

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 https://doi.org/10.1093/biomet/87.2.301

CAO, H. and WU, W. B. (2015). Changepoint estimation: Another look at multiple testing problems. *Biometrika* **102** 974–980. MR3431567 https://doi.org/10.1093/biomet/asv031

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 https://doi.org/10.2307/2965722

CHEN, J. and GUPTA, A. K. (2012). *Parametric Statistical Change Point Analysis*: *With Applications to Genetics*, *Medicine*, *and Finance*, 2nd ed. Birkhäuser/Springer, New York. MR3025631 https://doi.org/10.1007/978-0-8176-4801-5

DARLING, D. A. and ERDÖS, P. (1956). A limit theorem for the maximum of normalized sums of independent random variables. *Duke Math. J.* **23** 143–155. MR0074712

FRYZLEWICZ, P. (2014). Wild binary segmentation for multiple change-point detection. *Ann. Statist.* **42** 2243–2281. MR3269979 https://doi.org/10.1214/14-AOS1245

HANNART, A. and NAVEAU, P. (2012). An improved Bayesian information criterion for multiple change-point models. *Technometrics* **54** 256–268. MR2967976 https://doi.org/10.1080/00401706.2012.694780

HAO, N., NIU, Y. S. and ZHANG, H. (2013). Multiple change-point detection via a screening and ranking algorithm. *Statist. Sinica* **23** 1553–1572. MR3222810

HARCHAOUI, Z. and LÉVY-LEDUC, C. (2010). Multiple change-point estimation with a total variation penalty. *J. Amer. Statist. Assoc.* **105** 1480–1493. MR2796565 https://doi.org/10.1198/jasa.2010.tm09181

HAWKINS, D. M. (2001). Fitting multiple change-point models to data. *Comput. Statist. Data Anal.* **37** 323–341. MR1856677 https://doi.org/10.1016/S0167-9473(00)00068-2

HAYNES, K., ECKLEY, I. A. and FEARNHEAD, P. (2017). Computationally efficient changepoint detection for a range of penalties. *J. Comput. Graph. Statist.* **26** 134–143. MR3610414 https://doi.org/10.1080/10618600.2015.1116445

HAYNES, K., FEARNHEAD, P. and ECKLEY, I. A. (2017). A computationally efficient nonparametric approach for changepoint detection. *Stat. Comput.* **27** 1293–1305. MR3647098 https://doi.org/10.1007/s11222-016-9687-5

HORVÁTH, L. (1993). The maximum likelihood method for testing changes in the parameters of normal observations. *Ann. Statist.* **21** 671–680. MR1232511 https://doi.org/10.1214/aos/1176349143

JENG, X. J., CAI, T. T. and LI, H. (2010). Optimal sparse segment identification with application in copy number variation analysis. *J. Amer. Statist. Assoc.* **105** 1156–1166. MR2752611 https://doi.org/10.1198/jasa.2010.tm10083

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 https://doi.org/10.1080/01621459.2012.737745

KÜNSCH, H. R. (1989). The jackknife and the bootstrap for general stationary observations. *Ann. Statist.* **17** 1217–1241. MR1015147 https://doi.org/10.1214/aos/1176347265

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

LEE, C.-B. (1996). Nonparametric multiple change-point estimators. *Statist. Probab. Lett.* **27** 295–304. MR1395582 https://doi.org/10.1016/0167-7152(95)00089-5

LOADER, C. R. (1996). Change point estimation using nonparametric regression. *Ann. Statist.* **24** 1667–1678. MR1416655 https://doi.org/10.1214/aos/1032298290

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 https://doi.org/10.1080/01621459.2013.849605

MÜLLER, H.-G. and STADTMÜLLER, U. (1999). Discontinuous versus smooth regression. *Ann. Statist.* **27** 299–337. MR1701113 https://doi.org/10.1214/aos/1018031100

NISHII, R. (1984). Asymptotic properties of criteria for selection of variables in multiple regression. *Ann. Statist.* **12** 758–765. MR0740928 https://doi.org/10.1214/aos/1176346522

NIU, Y. S., HAO, N. and ZHANG, H. (2016). Multiple change-point detection: A selective overview. *Statist. Sci.* **31** 611–623. MR3598742 https://doi.org/10.1214/16-STS587

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 https://doi.org/10.1214/12-AOAS539

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014

SHAO, J. (1993). Linear model selection by cross-validation. *J. Amer. Statist. Assoc.* **88** 486–494. MR1224373

SHAO, J. (1997). An asymptotic theory for linear model selection. *Statist. Sinica* **7** 221–264. With comments and a rejoinder by the author. MR1466682

VENKATRAMAN, E. S. (1992). Consistency results in multiple change-point problems. Ph.D. thesis, Stanford Univ., ProQuest LLC, Ann Arbor, MI. MR2687536

WANG, G., ZOU, C. and YIN, G. (2018). Change-point detection in multinomial data with a large number of categories. *Ann. Statist.* **46** 2020–2044. MR3845009 https://doi.org/10.1214/17-AOS1610

WU, W. B. and ZHAO, Z. (2007). Inference of trends in time series. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **69** 391–410. MR2323759 https://doi.org/10.1111/j.1467-9868.2007.00594.x

YANG, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika* **92** 937–950. MR2234196 https://doi.org/10.1093/biomet/92.4.937

YANG, Y. (2007). Consistency of cross validation for comparing regression procedures. *Ann. Statist.* **35** 2450–2473. MR2382654 https://doi.org/10.1214/009053607000000514

YAO, Y.-C. (1988). Estimating the number of change-points via Schwarz' criterion. *Statist. Probab. Lett.* **6** 181–189. MR0919373 https://doi.org/10.1016/0167-7152(88)90118-6

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

ZHANG, P. (1993). Model selection via multifold cross validation. *Ann. Statist.* **21** 299–313. MR1212178 https://doi.org/10.1214/aos/1176349027

ZHANG, N. R. and SIEGMUND, D. O. (2007). A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics* **63** 22–32, 309. MR2345571 https://doi.org/10.1111/j.1541-0420.2006.00662.x

ZOU, C., WANG, G. and LI, R. (2020). Supplement to "Consistent selection of the number of change-points via sample-splitting." https://doi.org/10.1214/19-AOS1814SUPP.

ZOU, C., YIN, G., FENG, L. and WANG, Z. (2014). Nonparametric maximum likelihood approach to multiple change-point problems. *Ann. Statist.* **42** 970–1002. MR3210993 https://doi.org/10.1214/14-AOS1210