Multiple Change Point Analysis: Fast Implementation and Strong Consistency

Jie Ding, *Student Member, IEEE*, Yu Xiang, *Member, IEEE*, Lu Shen, and Vahid Tarokh, *Fellow, IEEE*

***Abstract*—One of the main challenges in identifying structural changes in stochastic processes is to carry out analysis for time series with dependency structure in a computationally tractable way. Another challenge is that the number of true change points is usually unknown, requiring a suitable model selection criterion to arrive at informative conclusions. To address the first challenge, we model the data generating process as a segment-wise autoregression, which is composed of several segments (time epochs), each of which modeled by an autoregressive model. We propose a multi-window method that is both effective and efficient for discovering the structural changes. The proposed approach was motivated by transforming a segment-wise autoregression into a multivariate time series that is asymptotically segment-wise independent and identically distributed. To address the second challenge, we derive theoretical guarantees for (almost surely) selecting the true number of change points of segment-wise independent multivariate time series. Specifically, under mild assumptions, we show that a Bayesian information criterion like criterion gives a strongly consistent selection of the optimal number of change points, while an Akaike information criterion like criterion cannot. Finally, we demonstrate the theory and strength of the proposed algorithms by experiments on both synthetic- and real-world data, including the Eastern U.S. temperature data and the El Nino data. The experiment leads to some interesting discoveries about temporal variability of the summer-time temperature over the Eastern U.S., and about the most dominant factor of ocean influence on climate, which were also discovered by environmental scientists.**

***Index Terms*—Autoregression, change detection, information criteria, large deviation analysis, strong consistency, time series.**

I. Introduction

TIME series data usually exhibits occasional changes in their structure, such as network anomalies in complex IP networks [1], distributional changes in teletraffic models [2], sudden changes of volatility in stock markets due to financial crises [3], variations of an electroencephalogram (EEG) signal caused by mode changes in the brain [4], or environmental changes in various ecosystems [5], [6]. Change detection analysis tries to identify not only whether a time series is a concatenation of several segments, in which the neighboring ones are generated from different probability distributions, but also how many change points there are. There has been a vast amount of work in the filed of change point analysis. In the parametric settings, the likelihood function naturally plays a key role, for example in the cumulative sum [7], [8] and the generalized likelihood ratio [9] approaches. Various tests have been developed for tracking changes in time series statistics such as the mean [10], [11], the variance [12]–[14], the autocovariance function [10], [15], and the spectrum [16]. Nonparametric approaches usually rely on kernel density estimation. A widely used approach is to perform change detection by direct estimation of the ratio of probability densities [17]–[19] or using some dissimilarity measure in feature space [20], without estimating densities as an intermediary step. For practical implementations, bisection procedure and its extensions [21]–[24] have been widely studied. Exact search methods such as segment neighborhood [25] and optimal partitioning [26], [27] have also been widely applied. Other remarkable progress in change point discovery for dependent time series data have been made in [28]–[31]. More detailed references to the literature can be found in monographs and review papers such as [7], [32]–[34].

In this paper, we are focusing on the offline multiple change detection problem. As with any other statistical inference procedure, it is crucial to apply an appropriate model selection procedure in order to select the number of change points, whenever it is unknown. A common way is to apply the penalized approach, which selects the model dimension by minimizing the sum of goodness of fit and a penalty term. The three commonly used penalties are Akaike information criterion [35], [36], Bayesian information criterion (BIC) [37], and Hannan and Quinn information criterion (HQ) [38]. AIC is derived by minimizing the Kullback-Leibler divergence between the true distribution and the estimate of a candidate model, BIC is from a large sample approximation that aims at selecting a model of maximum posterior probability, HQ replaces the $\log N$ term in BIC by $c \log\log N (c > 1)$, where N is the sample size. In some parametric models where regularity conditions are met, such as autoregressive models, it has been rigorously proved that AIC produces an overfitting model with non-vanishing probability, while BIC or HQ selects the model that converges almost surely to the true model (if it is included in the candidate set). In addition, HQ was proved to be the smallest penalty term that guarantees strong consistency, i.e, almost sure convergence [38]. Though these three criteria have been used as

Manuscript received October 25, 2016; revised March 8, 2017 and May 11, 2017; accepted May 26, 2017. Date of publication June 5, 2017; date of current version June 28, 2017. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Joao Xavier. This work was supported by Defense Advanced Research Projects Agency under Grant W911NF-14-1-0508 and Grant N66001-15-C-4028. *(Corresponding author: Jie Ding.)*

The authors are with John A. Paulson School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138 USA (e-mail: jieding@fas.harvard.edu; yuxiang@seas.harvard.edu; lshen@fas.harvard.edu; vahid@seas.harvard.edu).

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org., provided by the author. The material includes software implementations (in MATLAB) and real data used in the paper. Contact jieding@fas.harvard.edu for further questions about this work.

Digital Object Identifier 10.1109/TSP.2017.2711558

general-purpose model selection rules in various statistical models, their validity in terms of asymptotic behavior need to be verified case by case, especially for parametric or semi-parametric models where regularity conditions may not hold. Examples include finite mixture models [39]–[43], and change point models considered in this paper. To the best of our knowledge, several remarkable works have studied the consistency in selecting the number of change points for the change detection problem, e.g., [44]–[47], but the theory of strong consistency for penalized method has not been well studied before.

A typical offline multiple change point analysis aims to solve the following problem. Given observations $y_1, \ldots, y_N \in \mathbb{R}^D$ and $M \in \mathbb{N}$, the goal is to find integers $0 < \ell_1 < \cdots < \ell_M < N$ that minimize the following sum of within-segment loss

$$e_M = \sum_{k=1}^{M+1} Loss(y_{\ell_{k-1}+1}, \ldots, y_{\ell_k}), \tag{1}$$

where $Loss(\cdot)$ is some selected loss function and by default $\ell_0 = 0, \ell_{M+1} = N$. Here, specified number of change points M is usually estimated using penalized approach. A simple and widely adopted loss function is the quadratic loss [46] defined by $Loss_q(x_{\ell_{j-1}+1}, \ldots, x_{\ell_j}) = \sum_{n=\ell_{j-1}+1}^{\ell_j} |x_n - \bar{x}|^2$, where $\bar{x}$ is the sample mean of $x_{\ell_{j-1}+1}, \ldots, x_{\ell_j}$, and $|\cdot|$ denotes the Euclidean norm of a vector. One reason for using the quadratic loss is that it enables efficient k-means type fast implementations. This is to be discussed later. Other commonly used loss functions include the negative log-likelihood associated with a specified parametric model [48], [49], or the cumulative sums [8], [50].

In this work, we investigate the following two directions in detecting structural changes in time series:

1) In Section II, we consider the formulation of change point analysis for a general stochastic process. The basic idea is to assume that the time series data consists of several segments each of which is generated from a finite order autoregressive process. For such dependent data, the loss function of each segment may be defined as the log-likelihood loss associated with an autoregressive model, and a standard change detection algorithm such as binary segmentation [22] is amenable to use with the loss function. However, the loss function depends on a particular parametric assumption of the autoregression noises, and it does not always support efficient algorithms to minimize e_M. In fact, even if the noises are assumed to be Gaussian, the loss function can lead to massive computations, as we shall discuss it in detail later. To obtain the change points in a robust and computationally efficient manner, we propose an alternative approach which casts the change detection problem for the original time series $\{y_n\}$ into that for segment-wise (asymptotically) independent and identically distributed (i.i.d.) multivariate data $\{x_n\}$. We can discover the change points of independent data more easily, and then use the results to infer the change points of the original time series.

2) In Section III, we show that change points for a segment-wise independent data $\{x_n\}$ can be discovered by minimizing (1) with quadratic loss function and appropriately designed penalized methods. Specifically, we investigate necessary and sufficient conditions under which the unknown true number of change points can be determined for sufficiently large sample size (almost surely).

Finally, we present experimental results to demonstrate the performance of the proposed method on both synthetic and real-world datasets. The real data experiments lead to interesting conclusions about temporal variability of the summer-time temperature over the Eastern US, and the most dominant factor of ocean influence on climate.

Notation and Abbreviation: Let $tr(\cdot)$, $(\cdot)^{\mathrm{T}}$, $\log$, $a.s.$, $i.o.$ respectively denote the trace of a square matrix, the transpose of a matrix or vector, natural logarithm, almost surely, and infinitely often. We write $h_N = \Theta(g_N)$ if $c < h_N/g_N < 1/c$ for some constant $c \neq 0$ for all sufficiently large N. Let $o(1)$ denote a deterministic sequence that converges to zero. We use $o_p(1)$ and $O_p(1)$ to respectively denote a sequence of random variables that converges in probability to zero and that is stochastically bounded. We write $\mathcal{G} \sim [\mu, V]$ if distribution $\mathcal{G}$ has mean μ and variance V. Let $\mathcal{N}(\mu, V)$ denote the multivariate normal distribution with mean μ and covariance matrix V. Throughout the paper, random variables and observed data are respectively represented by capital letters (e.g. Y_n) and small letters (e.g. y_n). Vectors are all column vectors. We use $\rightarrow_{a.s.}$, $\rightarrow_p$, $\rightarrow_d$ to respectively denote the almost sure, in probability, and in distribution convergence.

A generic change detection model assumes data to be the outcomes of a sequence of multi-dimensional real-valued random variables $\{Y_n : n = 1, \ldots, N\}$ that consists of of $M_0 + 1$ segments ($M_0 \in \mathbb{N} \cup \{0\}$), where each pair of neighboring segments have different data generating process. In this paper, Y_n's are sometimes substituted with X_n's in order to emphasize the independence of data. We denote the true segments by $\{Y_n : n = L_{k-1} + 1, \ldots, L_k\}$, $k = 1, \ldots, M_0 + 1$, where $L_1 < \cdots < L_{M_0}$ are referred to as the M_0 change points, and by default $L_0 = 0, L_{M_0+1} = N$. Let $N_k = L_k - L_{k-1}, k = 1, \ldots, M_0 + 1$ denote the size (length) of the kth segment. Clearly, $\sum_{k=1}^{M_0+1} N_k = N$. Throughout the paper, we use $\hat{M}$ to denote the estimated number of change points. Similarly, we represent the detected change points by $\hat{L}_k, k = 0, \ldots, \hat{M} + 1$, and segment sizes by $\hat{N}_k, k = 1, \ldots, \hat{M} + 1$.

II. Change Detection for Time Series With Dependency Structure

In this section, we consider a sequence of one-dimensional dependent data. The results can be readily extended to multi-dimensional data. We assume that the data is generated from the following *segment-wise autoregressive (AR) model*:

(M.1) The sequence $\{Y_n : n = 1, \ldots, N\}$ are one-dimensional and it consists of $M_0 + 1$ segments, each of which can be described by a linear autoregression. Additionally, the autoregressive coefficients in two neighboring segments are different (so that there are M_0 change points). In other words, for each $k = 1, \ldots, M_0 + 1$, we have $Y_n = \underline{Y}_n^{\mathrm{T}} \psi^{(k)} + \varepsilon_n^{(k)}$, $n = L_{k-1} + 1, \ldots, L_k$, where $\underline{Y}_n = [1, Y_{n-1}, \ldots, Y_{n-L}]^{\mathrm{T}}$ (for $L > 0$) or $\underline{Y}_n = 1$ (for $L = 0$), $\psi^{(k)} \in \mathbb{R}^{L+1}$ (referred to as AR filter of order L), $\psi^{(k)} \neq \psi^{(k+1)}$ for $k = 1, \ldots, M_0, Y_{1-L}, \ldots, Y_0$ have been used to denote initial values, $L_k - L_{k-1} = \Theta(N)$ for $k = 1, \ldots, M_0 + 1$, and $\varepsilon_n^{(k)}$ are zero mean independent noises which are identically distributed within each segment.

An autoregression of order $L \in \mathbb{N} \cup \{0\}$ is also denoted by AR(L). Note that we have assumed the same L in each segment of the data generating model for the simplicity of presentation

(clearly, any AR(ℓ) is necessarily AR(k) for $\ell < k < \infty$, so that we may let L be the maximum of all the AR orders from each segment). In the rest part of the paper, we assume that the order L is known as prior knowledge or from exploratory studies. Our goal is to identify the number of change points and their locations.

Before we proceed, it is worth mentioning that even though the above change detection model is semi-parametric since no assumption on how each AR model switches to another one was made, the change point analysis can serve as an exploratory study for more parametric settings. For example, the detected change points can be used to set up better initial values of Expectation-Maximization algorithm for complex parametric mixture models such as point process regression models [51] and multi-state autoregressive models [52].

It is natural to define the loss function based on

$$Loss_a(y_{\ell_{j-1}+1}, \ldots, y_{\ell_j}) = \sum_{n=\ell_{j-1}+1}^{\ell_j} (y_n - \underline{y}_n^{\mathrm{T}} \hat{\psi})^2 \qquad (2)$$

where $\hat{\psi}$ is estimated from $y_{\ell_{j-1}+1}, \ldots, y_{\ell_j}$ by Yule-Walker equation or least squares method. The above loss is interpreted as the cumulated prediction errors, or the rescaled negative log-likelihood associated with AR(L). The quadratic loss can be regarded as the special case when $L = 0$. We can find change points by minimizing the sum of within-segment loss in (1) using state-of-the-art algorithms such as binary segmentation [22]. However, an alternative idea is to turn the change detection of segment-wise autoregressive model into that of segment-wise Gaussian independent model.

A. Discussion of the Underlying Ideas

Here we are following model (M.1). We start by considering a single AR to simplify the explanation of ideas. Following that, we then consider the case of two or more AR's. Consider a sequence of N points that are generated from a single AR(L), i.e. $Y_n = \psi^{\mathrm{T}} \underline{Y}_n + \varepsilon_n$, where $\psi \in \mathbb{R}^{L+1}$, $\varepsilon_n \sim [0, \sigma^2]$. Suppose that the true change points of $\{Y_n : n = 1, \ldots, N\}$ are located at multiples of w, where $w > 2L$ is an integer, and the data are divided into N/w segments of size w. If each segment of data is used to estimate an AR(L) filter, we obtain N/w estimates of ψ, respectively denoted by $\hat{\psi}_1, \ldots, \hat{\psi}_{N/w}$. It has been well established that if $\hat{\psi}_i$ is estimated from either least squares or Yule Walker methods, $\sqrt{w}(\hat{\psi}_i - \psi)$ converges in distribution to $\mathcal{N}(0, \Gamma)$ as w goes to infinity, where Γ is a constant matrix depending only on ψ [53, Appendix 7.5]. Thus, $\hat{\psi}_i$ can be approximated by multivariate Gaussian random variables with mean ψ and variance Γ/w. The asymptotic independence of $\sqrt{w}(\hat{\psi}_i - \psi)$ are guaranteed by the following result.

Theorem 1: Suppose that $\{Y_1, \ldots, Y_N\}$ are generated from an autoregression with filter ψ. Let $\hat{\psi}_1 \in \mathcal{R}^{L+1}$ and $\hat{\psi}_2 \in \mathcal{R}^{L+1}$ respectively denote the estimated filters from $\{Y_1, \ldots, Y_{N_1}\}$ and $\{Y_{N_1+1}, \ldots, Y_N\}$ by least square methods, where N_1, $N_2 = N - N_1 \to \infty$ as $N \to \infty$. Assume that

(A.1) the distribution of ε_n has a nontrivial absolutely continuous component, and $E\varepsilon_n^4 < \infty$, $E[\max\{(\log|\varepsilon_n|), 0\}] < \infty$.

Then $\sqrt{N_1}(\hat{\psi}_1 - \psi)$ and $\sqrt{N_2}(\hat{\psi}_2 - \psi)$ converge to two Gaussian random variables that are independent.

Algorithm 1: change detection by multi-window method.

input $\{y_n \in \mathbb{R}, n = 1, \ldots, N\}$, L (lag order), $M_{\max}$ (the largest size of candidate models), $w_1 > \cdots > w_R$ (window sizes)

output $\widehat{cp} = \{\hat{I}_1, \ldots, \hat{I}_{\hat{M}}\}$ (ranges containing change points)

1: $s_n^{(0)} = 0, n = 1, \ldots, N$ (initialized score)
2: **for** $r = 1 \to R$ **do**
3: Let $N_r = N/w_r$. Estimate $\hat{\psi}_{n_r} \in \mathbb{R}^{L+1}$ from $\{Y_n : n = (n_r - 1)w_r + 1, \ldots, n_r w_r\}$, $n_r = 1, \ldots, N_r$.
4: Call Algo. 2 with input $\hat{\psi}_{n_r} : n_r = 1, \ldots, N_r$, $M_{\max}$, selected $f(N)$, $\beta(N)$, and obtain output $\hat{\ell}_1, \ldots, \hat{\ell}_{M_r}$
5: Define scores $s_n^{(r)} = s_n^{(r-1)} + \mathbf{1}_{n \in \bigcup_{k=1}^{M_r} I_k^{(r)}}$, $n = 1, \ldots, N$, where $I_k^{(r)} = [(\hat{\ell}_k - 1)w_r + 1, (\hat{\ell}_k - 1)w_r + 2, \ldots, (\hat{\ell}_k + 1)w_r]$, and $\mathbf{1}_{n \in A}$ equals one if n belongs to the set A and zero otherwise.
6: **end for**
7: Call Algo. 3 to obtain the peak ranges $\widehat{cp} = \{\hat{I}_1, \ldots, \hat{I}_{\hat{M}}\}$ ($\hat{M} \le M_{\max}$)

Assumption (A.1) is mild, as for instance, it is satisfied by the Gaussian distribution. Theorem 1 implies that if a data from the same autoregression is split into two (or more) parts, and each part gives an estimate of the true filter, then the estimators are asymptotically independent (up to a rescaling).

Now suppose that the stochastic process consists of two parts: the first N_1 points are generated from one AR(L) and the rest N_2 are from another AR(L). If a window size w that satisfies $2L < w < \min\{N_1, N_2\}$ is chosen, the estimated AR filters are approximately independent points in $\mathbb{R}^{L+1}$ and they contain a change point around the (N_1/w)th point. Here and afterwards, we assume that $N_1/w, N_2/w$ are integers. Extension to more general cases is straightforward. We therefore obtain a sequence of $(N_1 + N_2)/w$ 'independent' multivariate random variables, with a *change in mean* at around the (N_1/w)th point. We propose a multi-window (MW) change detection algorithm that chooses different w's and collect the information of the detected change points for each w in a proper way, in order to obtain a more accurate estimation of the change points of the stochastic process. From a computational point of view, starting from a large w also helps to reduce the cost, which is especially helpful in cases where massive time series data is involved.

B. Algorithmic Descriptions

Algo. 1 is a pseudo-code for MW method, followed by two subroutines: Algorithms 2 and 3. Illustrating experiments are provided in Section IV. Algo. 1 uses a sequence of R window sizes $w_1 > \cdots > w_R$ (discussed below) in order to capture any true segment of small size. For each w_r, the original data is turned into a sequence of $L + 1$ dimensional data that can be approximated as independent. By calling Algo. 2, we obtain a set of change points $\hat{\ell}_1, \ldots, \hat{\ell}_M$; By further mapping these change points back to the original scale $\{1, \ldots, N\}$, we obtain several short ranges (intervals) $I_k^{(r)}$ (each of size $2w_r$) that "probably" contain the desired change points. We repeat the

Algorithm 2: (generic) change detection by minimizing the sum of within-segment quadratic loss.

input $\{x_n \in \mathbb{R}^D, n = 1, \ldots, N\}$, $M = M_{\max} \in \mathbb{N}$ (the largest candidate number of change points), $f(N)$ (penalty term), $\beta(N)$ (minimal segment size)

output $\hat{M}, \hat{\ell}_1, \ldots, \hat{\ell}_{\hat{M}}$ (discovered change points).

1: **for** $k = 0 \to M_{\max}$ **do**
2: Define $\ell_0 = 0, \ell_{k+1} = N$; minimize $e_k = \sum_{j=1}^{k+1} Loss_q(x_{\ell_{j-1}+1}, \ldots, x_{\ell_j})$ over $\ell_j \in \mathbb{N}$ and record the optimum $\hat{\ell}_j : j = 1, \ldots, k$, $\hat{e}_k$
3: **if** size of the smallest segment $< \beta(N)$ **then**
4: Let $M = k - 1$; break the for loop
5: **end if**
6: **end for**
7: Choose $\hat{M} = \arg\min_{k=0,\ldots,M}(\hat{e}_k + kf(N))$, and $\{\hat{\ell}_j\}$ to be the solution to Step 2 under $k = \hat{M}$.

Algorithm 3: peak range selection.

input $s_n^{(r)}, n = 1, \ldots, N, r = 1, \ldots, R$ (recorded scores), $\tau \in \mathbb{N} \cup \{0\}$ (tolerance level)

output $\widehat{cp} = \{\hat{I}_1, \ldots, \hat{I}_{\hat{M}}\}$ (the output of Algo. 1)

1: **for** $r = R \to 1$ **do**
2: Let $S = \max_{n=1,\ldots,N}\{s_n^{(r)}\}$, and $H = \{n : s_n^{(r)} \geq S - \tau\}$. Arrange the elements of H in ascending order as $\{h_1, \ldots, h_p\}$
3: Initialize $M_r^* = 1, u_1 = 1$
4: **for** $i = 2 \to p - 1$ **do**
5: **if** $h_i < h_{i+1} - 1$ **then**
6: Let $v_{M_r^*} = h_i, u_{M_r^*+1} = h_{i+1}$
7: Let $M_r^* = M_r^* + 1$
8: **end if**
9: **end for**
10: Let $v_{M_r^*} = h_p$
11: From the above steps, we can rewrite $H = \cup_{m=1}^{M_r^*} J_m$, where each J_m is a peak range in the form of $J_m = [u_m + 1, u_m + 2, \ldots, v_m], m = 1, \ldots, M_r^*$, where the associated scores are at least $S - \tau$
12: **if** $M_r^* \leq M_{\max}$ **then**
13: Let $\widehat{cp} = H$, namely $\hat{M} = M_r^*$ and $\hat{I}_m = J_m$ for each $m = 1, \ldots, \hat{M}$; break the for loop
14: **end if**
15: **end for**

above procedure for different w_r, and combine the information in the following way: the detected ranges of change points from each window size are scored by one, the scores are aggregated, and the ranges with highest score or around the highest score (determined by the tolerance parameter τ) are finally selected. The output of the algorithm is $\hat{M}$ number of "peak" ranges that are most likely to contain the true change points.

In the descriptions of the algorithms, two of the important parameters are briefly explained here and the detailed explanations can be found in Section III. Parameter $\beta(N)$ is introduced for two purposes: for the technical convenience in deriving asymptotic results, and for faster implementation in practice. $\beta(N)$ must be selected such that $\lim_{N\to\infty} \beta(N) = \infty$. The penalty function is a linear function in the form of $kf(N)$, where $f(N)$ is referred to as the penalty term. The penalty terms $f(N) \propto 1$, $f(N) \propto \log\log N$, and $f(N) \propto \log N$ are referred to as the variants of AIC, HQ, and BIC, respectively.

Some detailed discussions of Algo. 1 and its subroutines are given below.

Algo. 1: In step 5, the score is introduced to facilitate the selection of the final $\hat{M}$ ranges. In particular, for each w_r, the detected ranges of change points from each window size are scored by one, and otherwise by zero. Then the scores are aggregated for all w_r, and the ranges with highest score or around the highest score (determined by τ) are finally selected.

Algo. 2: This subroutine is called by Algo. 1 at Step 4. Its input (X_n) is the sequence of estimated AR filters from each window (of size w_r) of the original time series (Y_n). It detects the number and locations of change points based on minimizing within-segment quadratic loss and applying penalized model selection approach (with $D = L + 1$). For clarity, we focus on Algo. 1 in this section, and defer detailed discussions of Algo. 2 to Section III.

Algo. 3: This subroutine is called by Algo. 1 at Step 7. It aims to select the narrow ranges with the highest scores, which are most likely to contain change points. Note that since $w_1 > w_2 > \cdots > w_R$, one needs to go from w_R to w_1 in order to pick the narrowest possible ranges. Steps 2–11 describe an algorithmic procedure to partition the set H (of high score) into disjoint intervals/ranges $[u_m + 1, \ldots, v_m]$. Its "for" loop (from R to 1) is a backward pruning procedure in order to ensure that $\hat{M} \leq M_{\max}$. The pruning was done by neglecting scores produced by the smallest window sizes, which are less reliable as the estimated AR filters from those windows have larger variances.

Window Sizes: Intuitively speaking, more reliable change detection results can be obtained by using multiple window sizes (instead of only one), since in practice we do not know what the true segment sizes are, and an inappropriately chosen w_r may be so large that a true segment is "missed". On the other hand, a small w_r leads to larger variance of AR filter estimates. A properly designed MW method strikes a tradeoff between estimation accuracy (since larger window sizes reduce variance of the estimated AR filters) and the resolution of the detected change points (since smaller window sizes produce narrower ranges). The specific choices of w_r's may depend on the desired accuracy and budget of computing time. In general, we suggest that $R = \Theta(\log N), w_1 = \Theta(N), w_R = \Theta(L)$, and $w_1 > \cdots > w_R$ is a geometric progression, based on our extensive synthetic data experiments. We provide the complexity analysis in Section IV.

Computing the Estimator $\hat{\psi}_{n_r}$: For a specified AR order L, $\hat{\psi}_{n_r}$ can be obtained either by least squares method, or by the Yule-Walker method (which requires slightly more data points, but supports fast computation by, e.g., the Levinson-Durbin recursion [54]).

Tolerance Parameter: The main purpose of introducing the tolerance parameter τ in step 8 of Algo. 1 is to ensure that the scoring produces fair comparisons among different ranges. Otherwise, small segments may be "missed" by some initial large window sizes. For example, suppose that $\tau = 0, w_1 = 200$ and there is only one true change point at $N_1 = 50$ in $N = 1000$ data points. Then in this scenario, it is harder to discover a change point from $N/w_1 = 5$ estimated filters.

It is worth mentioning that the output of MW method is a set of $\hat{M}$ narrow ranges instead of single points. In the cases where $\hat{M}$ exact change points are desired, we could use the results from Algo. 1 as starting point to further search optimal points within those ranges. In that sense, MW method can serve as a fast prescreening approach. In addition, the multiple windows can be implemented in parallel for massive time series, and it can be applied to independent data as well.

C. Discussion on the Implementation of Algorithm 2

Implementations of Algo. 2 based on popular methods such as binary segmentation [22], segment neighborhood [25], and optimal partitioning [26], [27] are possible. But since our loss function is quadratic, it is possible to have an algorithm that takes full advantage of this fact. We propose such a computationally efficient algorithm, which is analogous to but also differs from the usual k-means algorithm (in that each segment/cluster contains points with consecutive indices). It can then be regarded as an "ordered k-means" algorithm. The algorithm reduces the within-segment quadratic loss in each step by moving the change points based on the following result. Suppose that $\{X_n : n = 1, \ldots, N_1\}$ and $\{X_n : n = N_1 + 1, \ldots, N_1 + N_2\}$ are two segments. Consider the operation that shifts the change point from N_1 to $N_1 - t$ where $0 < t < N_1$: the two segments become $\{X_n : n = 1, \ldots, N_1 - t\}$ and $\{X_n : n = N_1 - t + 1, \ldots, N_1 + N_2\}$. The within-segment quadratic loss will be reduced after the operation if and only if

$$\frac{N_1 |\bar{X}_{0,N_1} - \bar{X}_{N_1-t,N_1}|^2}{N_1 - t} > \frac{N_2 |\bar{X}_{N_1,N_1+N_2} - \bar{X}_{N_1-t,N_1}|^2}{N_2 + t}, \tag{3}$$

where $\bar{X}_{n_1,n_2}$ denotes the sample mean of $\{X_n : n = n_1 + 1, \ldots, n_2\}$.

From the above argument, in order to decide whether a subsequence of data should be moved from one segment to its neighboring one, it only suffices to compute its mean and also the means of the original two segments. By iterative application of the result, a local optimum of step 2 in Algo. 2 could be achieved. Specifically, suppose that k is the specified number of change points and we start with initialized k segments (e.g. of equal sizes) located at $\ell_1, \ldots, \ell_k$. In each iteration, for $j = 1, \ldots, k$, we find the point $\hat{\ell}_j$ that minimizes $Loss_q(x_{\ell_{j-1}+1} + x_{\hat{\ell}_j}) + Loss_q(x_{\hat{\ell}_j+1}, \ldots, x_{\ell_{j+1}})$ and substitute ℓ_j with $\hat{\ell}_j$. Each iteration requires $\Theta(N)$ operations (multiplications and additions) by applying (3). Our synthetic data experiments indicate that this greedy algorithm converges very fast and reliably (usually achieves the global optimum). If the number of iterations is finite, then the overall computational cost is $\Theta(N)$. In our practical implementation, we specify a finite number of iterations and repeat with a few different random initial change points in order to enhance the possibility of obtaining the global optimum.

To analyze the complexity (in terms of the number of multiplications or additions), we assume that L and $M_{\max}$ are constants. By straightforward calculations, the complexity of estimating an AR filter from N data points by either Yule-Walker equation or least squares method is $\Theta(N)$. Thus, BS method requires the complexity $\Theta(k) + \Theta(N - k)$ to scan the kth point, resulting in $\Theta(N^2)$ operations to scan all the N points. Since the number of iterations is $M_{\max}$, and the complexity of the first iteration dominates that of the subsequent iterations, BS method has complexity $\Theta(N^2)$. On the other hand, for each window size w in MW method, the major computational cost is in Step 3 of Algo. 1 which is $(N/w)\Theta(w) = \Theta(N)$, and in Step 4 which is $f_{\text{inde}}(N/w)$, where $f_{\text{inde}}(n)$ represents the complexity of Algo. 2 given n data points. Therefore, the complexity of MW method can be expressed as $f_{\text{de}}(N) = \Theta(N) + f_{\text{inde}}(N/w_R)$ (w_R is the smallest window size). If Algo. 2 implements the method proposed above, then in accordance with the complexity analysis there, we obtain $f_{\text{inde}}(N/w_R) = \Theta(N/w_R)$ and $f_{\text{de}}(N) = \Theta(N + N/w_R) = \Theta(N)$. Moreover, if an analyst needs *exact* k change points that minimize the sum of within-segment loss defined in (2), he may carry out an exact search based upon the results from Algo. 1. Suppose that Algo. 1 outputs k intervals of size w_R, the additional computational cost is $\Theta(w_R^k N)$ which can be far below $\Theta(N^2)$ required by BS method.

III. Strong Consistency of Penalized Methods

Recall from Subsection II-A and II-B that the key idea of our methodology is to transform segment-wise autoregression into segment-wise asymptotically independent (multivariate) data. Fast implementation is then achieved, because 1) we can apply the efficient subroutine Algo. 2 (elaborated in Subsection II-C), and 2) we can implement the for loop in Algo. 1 in parallel (for different window sizes/resolutions).

In particular, the input of Algo. 2 is the sequence of estimated AR filters (of dimension $L + 1$) from each window of the original time series (Y_n). In other words, each window gives an estimate of the true filter, and then the estimators are asymptotically independent (up to a rescaling).

Since the performance (in terms of accurately discovering the locations and number of changes) of our MW method largely depends on the subroutine Algo. 2, we only focus on Algo. 2 in this section. Under assumption (M.1), the asymptotically independence motivates us to analyze a relaxed problem under assumption (M.2), where $\{X_n\}$ is a sequence of independent random variables. Note that even under assumption (M.2), the problem is still technically highly non-trivial. We make the following assumption on the input of Algo. 2, denoted by $\{X_n\}$ (with dimension $D = L + 1$).

(M.2) The sequence $\{X_n : n = 1, \ldots, N\}$ are D-dimensional ($D \in \mathbb{N}$) and independent random variables. Moreover, for each $k = 1, \ldots, M_0 + 1$, we have $N_k = \Theta(N)$, and $\{X_n : n = L_{k-1} + 1, \ldots, L_k\}$ are i.i.d. distributed according to $\mathcal{G}_k \sim [\mu_k, V_k]$. When $M_0 \geq 1$, $\mu_k \neq \mu_{k+1}$, $k = 1, \ldots, M_0$.

Algo. 2 discovers change points by minimizing the within-segment sum of quadratic loss e_k. Algo. 2 computes $\hat{e}_k$ for each candidate number of change points $k = \{0, \ldots, M\}$, where M is determined by the largest candidate number of segments $M_{\max}$ and minimal segment length $\beta(N)$. After that, the optimal number of change points $\hat{M}$ is estimated according to a penalized method. Further details are given below. In the remaining subsections, we shall show that when applied to a segment-wise independent data, Algo. 2 outputs $\hat{M}$ such that $\hat{M} \rightarrow_{a.s.} M_0$ as data size tends to infinity.

Parameter $\beta(N)$*:* It is introduced for two purposes: for the technical convenience in deriving asymptotic results, and for faster implementation in practice. $\beta(N)$ must be selected such

that $\lim_{N\to\infty}\beta(N)=\infty$. The rate of growth of $\beta(N)$ will be selected depending on the theoretical results we wish to prove.

Penalty Function: The common choice of penalty function is a linear function in the form of $kf(N)$, where $f(N)$ is referred to as the penalty term. For brevity, we consider the linear function in this paper, but the results can be applied to more general penalty functions. Three commonly used types of penalty terms are related to AIC, HQ, and BIC. In a parametric change detection problem, if there are k change points and p parameters in each segment, the total number of parameters to appear in AIC and BIC is $k+p(k+1)$. If the quadratic loss is treated as twice the negative log-likelihood of a Gaussian probability density function with variance equal to the identity matrix, the total number of parameters is $k+D(k+1)=k(D+1)+constant$. The penalty terms $f(N)\propto 1$, $f(N)\propto \log\log N$, and $f(N)\propto \log N$ are referred to as the variants of AIC, HQ, and BIC, respectively.

Strong Consistency: A penalized model selection approach is referred to be strongly consistent if $\hat{M}\to_{a.s.} M_0$ as data size tends to infinity. We may also say that $\hat{M}$ is strongly consistent.

A. Necessary Conditions for Strongly Consistent Model Selection

We start by examining the case when the true data generating process has no change point.

Theorem 2: Assume that the data generating model is given by (M.2) with $M_0\geq 0$. Then the smallest penalty term $f(N)$ that guarantees strong consistency of $\hat{M}$ in Algo. 2 is at least $\Theta(\log\log N)$.

If we additionally assume $M_0=0$ and $\beta(N)=\Theta(N)$, then there exists a constant $C>0$ such that $f(N)=C\log\log N$ guarantees strong consistency of $\hat{M}$.

Remark 1: Theorem 2 proves that the smallest penalty for strong consistency is $\Theta(\log\log N)$ (given by variants of HQ criterion). A by-product of its proof is a technical lemma (Lemma 1 in the Appendix) that implies that an AIC-like criterion (with constant penalty) always produces a non-vanishing overfitting probability. We recap this observation after Lemma 1. Interestingly, these observations are similar to those found for order selection of autoregressive models, even though an autoregressive model is purely parametric, and the proof in those cases require different technical approaches [38], [55].

Theorem 2 also proves that $f(N)=\Theta(\log\log N)$ is sufficient for strong consistency in the particular case $M_0=0$. The next theorem shows that the necessary condition is also sufficient for bounding the estimated number of change points when $M_0>0$.

We define

$$\bar{\Delta}_\mu=\max_{k=1,\ldots,M_0}\{|\mu_k-\mu_{k+1}|\},\ \underline{\Delta}_\mu=\min_{k=1,\ldots,M_0}\{|\mu_k-\mu_{k+1}|\}.$$

Theorem 3: Under the model assumption (M.2) with $M_0>0$, suppose that $\beta(N)=\Theta(N)$ and

(A.2) The largest candidate number of change points $M_{\max}$ is finite and $M_{\max}\geq M_0+3$,

(A.3) The true segment sizes satisfy $\beta(N)\leq N_k/4$, $N_k=\Theta(N)$ for $k=1,\ldots,M_0+1$. In addition, $f(N)=o(N)$.

Then there exists a positive constant C_0 such that whenever $f(N)\geq C_0\log\log N$, the estimated number $\hat{M}$ satisfies $M_0\leq \hat{M}\leq 2M_0$ for sufficiently large N almost surely, i.e.,

$$\mathrm{pr}\left\{\limsup_{N\to\infty}(\hat{M}<M_0)\cup(\hat{M}>2M_0)\right\}=0.$$

Moreover, the distances between the estimated change points and true ones satisfy

$$\limsup_{N\to\infty}\min_{k=1,\ldots,\hat{M}}\frac{|\hat{L}_k-L_j|}{2\beta(N)}\leq 1\quad (a.s.) \tag{4}$$

for each $j=1,\ldots,M_0$.

Remark 2: The requirement $M_{\max}\geq M_0+3$ (instead of $M_{\max}\geq M_0$) in (A.2) is for technical convenience in the proof of Theorem 3. Theorem 3 shows that $f(N)=\Theta(\log\log N)$ suffices to guarantee no underfitting. Although we cannot prove it avoids overfitting as well, we proved that the extent of overfitting is bounded (since $\hat{M}\leq 2M_0$ holds almost surely). In addition, Inequality (4) implies that each true change point is "almost" captured, since its nearest discovered change point is within distance $\beta(N)$, which can be chosen to be arbitrarily small compared with N (or each N_j). In the next subsection, we relax the assumption on $\beta(N)$ and obtain strongly consistent $\hat{M}$ by increasing the penalty to be BIC-like.

B. Sufficient Conditions for Strongly Consistent Model Selection

Suppose that $\{X_n: n=1,\ldots,N\}$ are i.i.d. sub-Gaussian random variables. Then there is some $c_0>0$ such that for every $a\in\mathbb{R}$,

$$\mathrm{pr}(|\bar{X}-E(X_1)|\geq a)\leq 2e^{-c_0 a^2 N} \tag{5}$$

where $\bar{X}$ denotes the sample mean. Assuming that X_n follows a sub-Gaussian distribution, it is possible to prove the strong consistency of $\hat{M}$. The assumption is for bounding the tail probability through the large deviation analysis [56]–[57].

Theorem 4: Under the model assumption (M.2) with $M_0\geq 0$, suppose that Assumptions (A.2), (A.3) in Theorem 3 hold and that

(A.4) $\mathcal{G}_k, k=1,\ldots,M_0+1$ are marginally sub-Gaussian. In other words, there exists a constant $c_0>0$ such that (5) holds for each marginal distribution of $\mathcal{G}_k$.

If

$$f(N)\geq 100\bar{\Delta}_\mu^2\eta^*(N), \tag{6}$$

where $\eta^*(N)=250Dc^{M_0-1}\log N/(c_0\underline{\Delta}_\mu^2)$, $c=4/(\sqrt{2}-1)^2$, then $\hat{M}$ is strongly consistent. Moreover,

$$\limsup_{N\to\infty}\left(\max_{k=1,\ldots,M_0}\frac{|\hat{L}_k-L_k|}{\eta^*(N)}\right)\leq 1\quad (a.s.)$$

Remark 3: Assumption (A.4) is satisfied by Gaussian, any bounded random variables, etc. By the conditions of Theorem 4, both the minimal distance and the minimal penalty required for strong consistency are no more than $\Theta(\log N)$. Note that we do not need the requirement $\beta(N)=\Theta(N)$. The constant term for $f(N)$ is proportional to the dimension D and the ratio $\bar{\Delta}_\mu^2/\underline{\Delta}_\mu^2$. Intuitively, higher dimension and larger variation in $|\mu_k-\mu_{k+1}|$ require stronger penalties. Besides this, it is interesting to observe that $f(N)$ depends on the ratio $\bar{\Delta}_\mu^2/\underline{\Delta}_\mu^2$ which is scale

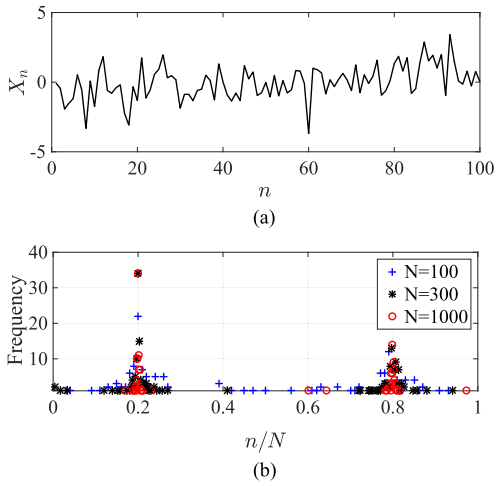

Fig. 1. (a) A sequence of independent data that contains two change points, and (b) the frequencies of discovered change points for each $N = 100, 300, 1000$.

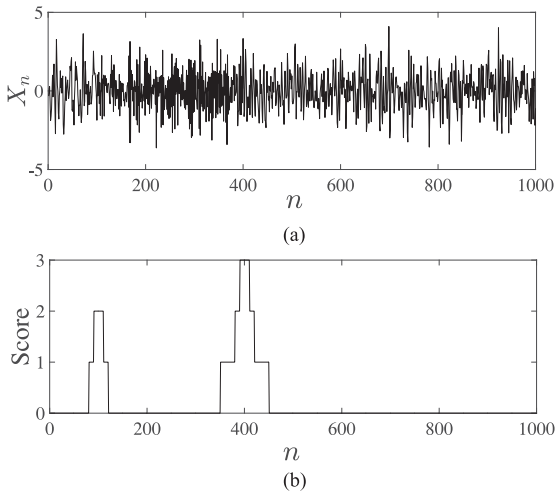

Fig. 2. (a) A time series that consists of three segments of various autoregressions, and (b) score plot for change detection.

invariant, while $\eta^*(N)$ only depends on the smallest distance between two neighboring distributions (in terms of the means).

IV. Experiments

In this section, we present experimental results to demonstrate the above theoretical results, and the advantages of MW method on both synthetic and real-world datasets.

A. Independent Data

In a synthetic data experiment, we generated data of two change points: $X_n \sim \mathcal{N}(\mu_1, \sigma^2)$, $n = 1, \ldots, 0.2N$, $X_n \sim \mathcal{N}(\mu_2, \sigma^2)$, $n = 0.2N + 1, \ldots, 0.8N$, $X_n \sim \mathcal{N}(\mu_3, \sigma^2)$, $n = 0.8N + 1, \ldots, N$. Let $[\mu_1, \mu_2, \mu_3, \sigma^2] = [-1, 0, 1, 1]$, $M_{\max} = 10$, $f(N) = 2 \log N$, $\beta(N) = \log \log N$. For illustration purpose, an example dataset with $N = 100$ is plotted in Fig. 1(a). For each $N = 100, 500, 1000$, we generate 100 independent datasets and summarize the detected change points (normalized by N) in Fig. 1(b). We also summarized the percentage frequencies of $\hat{M} < 2$, $\hat{M} = 2$, and $\hat{M} > 2$, respectively denoted by $f = (f_1, f_2, f_3)$. They are $f = (38, 60, 2)$ for $N = 100$, $f = (0, 89, 11)$ for $N = 300$, and $f = (0, 95, 5)$ for $N = 1000$. The results show that both the estimated number of and locations of change points become more and more accurate as the sample size grows.

B. Dependent Data

In a synthetic data experiment for dependent data, we generated data of two change points at $0.1N$ and $0.3N$. Data is generated from a zero mean autoregression in each of the three segments, and the associated AR filters are respectively $[\psi_1^{(1)}, \psi_2^{(1)}] = [0.8, -0.3]$, $[\psi_1^{(2)}, \psi_2^{(2)}] = [-0.5, 0.1]$, $[\psi_1^{(3)}, \psi_2^{(3)}] = [0.5, -0.5]$. Suppose that the noises are $\mathcal{N}(0, 1)$ and $M_{\max} = 5$, $f(N) = \log N$, $\tau = 1$. Fig. 2(a) illustrates one dataset with $N = 1000$. We set window sizes to be $[w_1, w_2, w_3, w_4] = [100, 50, 20, 10]$ and apply Algo. 1 to that dataset. The score is plotted in Fig. 2(b).

Next, we compare MW method with binary segmentation (BS) method (which is perhaps the most widely applied approach in the literature). The BS method first scans all the points and finds a single change point that minimizes the sum of within-segment loss defined in (2), and then extends to multiple change points discovery by iteratively repeating the method on different subsets of the series. This procedure is repeated until the maximal number of change points is reached or no more change point is detected.

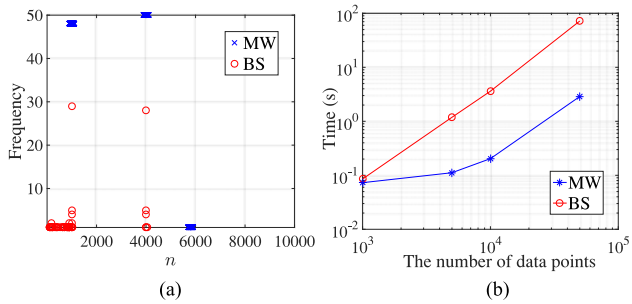

Fig. 3. (a) Frequencies of detected change points (or its ranges) by BS and MW methods, and (b) log-log plot of the computation time on multiple change points analysis.

To numerically compare the performance of MW and BS, we repeat the above experiment for 50 iterations. In each iteration, we generated three autoregressive filters of order $L = 2$ that are independent and uniformly distributed in the space of all stable AR(2) filters.[1] The change points are still $0.1N$ and $0.3N$. The number of points is $N = 10^4$. The discovered change points are plotted in Fig. 3(a). In order to compare the computational speed, we repeat the above experiment for each $N = [10^3, 5 \times 10^3, 10^4, 5 \times 10^4, 10^5]$. For the MW method, we use fixed number of windows $\{w_r\}_{r=1}^4 = N/10, N/20, N/50, N/100$ and tolerance parameter $\tau = 2$. We set the minimal length for BS method to be $10L$ (which is used to guarantee stability involved in matrix computations). For both methods, $M_{\max} = 4$. The comparison is plotted in Fig. 3(b). The average numbers of detected change points (with

[1]In general, for a stationary AR(L) processes with coefficients $\psi = [\psi_1, \ldots, \psi_L]$, ψ stays in a bounded subspace $S_L \subset \mathbb{R}^L$. For the purpose of fair comparison, in the experiment we draw AR filters that are uniformly distributed on S_L, using the technique proposed in [58].

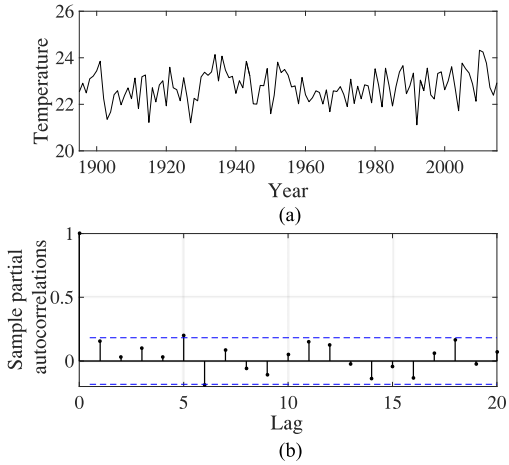

Fig. 4. (a) 1895-2015 summer-time temperature over the Eastern US (unit: °C), and (b) its sample partial autocorrelations.

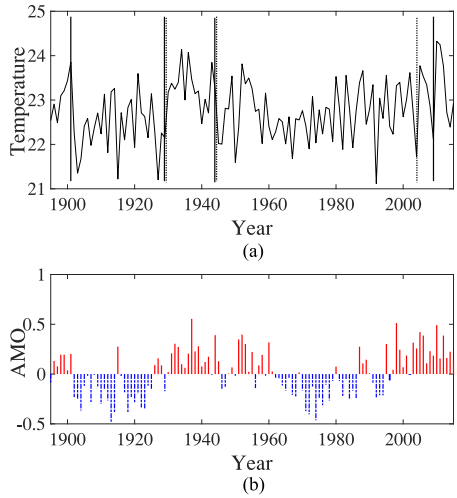

Fig. 5. (a) Detected change points of the Eastern US temperature, and (b) phase shifts of the AMO.

standard error inside the parenthesis) under each N are respectively $2.48(0.12)$, $1.98(0.04)$, $1.98(0.03)$, $1.98(0.04)$ for MW method, and $2.56(0.12)$, $3.2(0.11)$, $3.46(0.14)$, $3.7(0.14)$ for BS method. Here, if a discovered range has size no larger than twice the smallest window size and it contains a true change point, it is regarded as a successfully detected change point.

The simulation results shows that MW is more robust and computationally efficient than BS method. As was pointed out in the previous section, MW is robust because it looks into the data at different resolutions, thus reducing the risks of overfitting or underfitting (which the BS method suffers from).

C. Eastern US Temperature From 1895 to 2015

In this subsection, we investigate the temporal variability of the summer-time temperature over the Eastern US for 1895-2015 (121 points plotted in Fig. 4(a)) with our change detection algorithm. The temperature data is obtained from National Climatic Data Center (NCDC, http://www.ncdc.noaa.gov/) and averaged over the Eastern US (east of 100 °W). Fig. 4 shows the data and its sample partial autocorrelations, from which we recognize the data as independent. We choose $M_{\max} = 7$, and try a range of penalty terms $f(N) = j \log\log N, j = 1, ..., 5$. We start with $j = 1, 2$; the penalty is so small that it gives the maximally possible 7 change points. Then we increase $f(N)$ to $3 \log\log N$, and obtain 5 change points at years 1901, 1929, 1944, 2009, 2012 (marked in solid lines in Fig. 5(a)). If $f(N)$ is increased to $4 \log\log N$, the change points are the years 1929, 1944, 2004 (marked in dashed lines in Fig. 5(a)). If $f(N)$ is further increased to $j \log\log N, j \geq 5$, there is no change point detected. The segmentation of the time series of the Eastern US temperature over the past century matches the phase shift of the Atlantic Multi-decadal Oscillation (AMO), defined as the North Atlantic sea surface temperature after removing the long-term warming trend [59]. As seen from Fig. 5(b), since the early 20th century, there are warm phases from 1929 to 1960 and from 1990 to 2015, and cool phases from 1901 to 1929 and from 1965 to 1990, in synchrony with the segmentation of the Eastern US temperature time series defined by the change points. As the ocean has much larger heat capacity than the continent, this implies that the multi-decadal variability of Eastern US temperature is modulated by the AMO. The dynamic link between AMO and Eastern US climate has previously been reported. For example, based upon a global climate model, it was indicated in [59], [60] that the AMO plays an important role in driving the summer-time temperature in the Eastern US. This validates our conclusion derived from the change point detection algorithm.

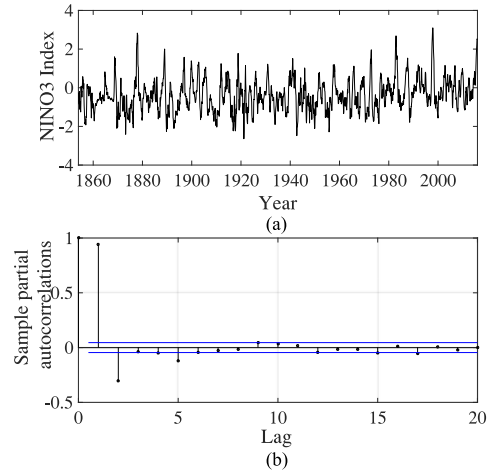

Fig. 6. (a) Monthly El Nino (Nino3) index from 1854 to 2015, and (b) its sample partial autocorrelations.

D. El Nino Data From 1854 to 2015

As the largest climate pattern, El Nino serves as the most dominant factor of oceanic influence on climate. The NINO3 index, defined as the area averaged sea surface temperature from 5 °S–5 °N and 150 °W–90 °W, is calculated from HadISST1 from 1854 to 2015 [61], as shown in Fig. 6(a) (with 1944 points). By looking at the partial autocorrelation of the complete dataset in Fig. 6(b), we tentatively set autoregression order $L = 2$ (in fact, we also experimented the cases $L = 3, 4, 5$ and the final results did not differ much). We apply Algo. 1 with window sizes $300, 250, 200, 150, 100, 50$, and $M_{\max} = \lfloor N/300 - 1 \rfloor = 5$ (where $\lfloor a \rfloor$ denotes the largest

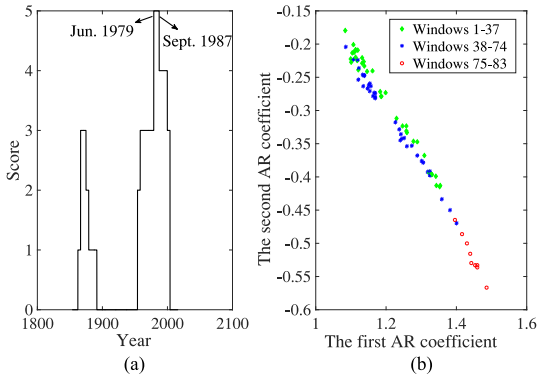

Fig. 7. (a) Score plot of El Nino data obtained from Algo. 1 which indicates the ranges of change points, and (b) the trace plot that illustrates how the coefficients of AR(2) vary with time.

integer that is no larger than a). We start with $f(N) = 2\log\log N$ and obtain the score plot as shown in Fig. 7(a). The plots show that the time period from June 1979 to September 1987 most likely contains one change point. We change the penalty to smaller or larger values, or use other window sizes, and found that the range is detected most of the time. In fact, we can trace how the AR coefficients change in Fig. 7(b), where each point is the AR coefficient estimated from a sliding window of size 300 and sliding step size 20. In other words, the windows are $\{X_1, \ldots, X_{300}\}, \ldots, \{X_{1641}, \ldots, X_{1941}\}$. The green diamond, blue star, and red circle indicate respectively the first 37 windows, the second 37 windows, and the last 9 windows. As illustrated from the plot, the red circles deviate nontrivially from other points, which means that the data has a structural change after 74 windows, and that time is exactly the year 1979. The shift of the Pacific Decadal Oscillation (PDO) from a long cold phase (1940–1978) to a warm phase (1979-present) is likely to explain why this year is unique in the past 150 years. The PDO can have a strong influence on the climate in the Northern hemisphere, including the drought frequency in the North America [62], ecosystem productivity [63], as well as the Bermuda High pressure system in Atlantic ocean [64] .

V. Conclusion

This work investigated the necessary and sufficient conditions under which a model selection criterion is strongly consistent. Our analysis is under the assumption of the independence of data, and for the quadratic loss function. Nevertheless, it appears that our proposed technical tools can be applied to studying richer data structures. Furthermore, we modeled a general stochastic process by segment-wise autoregressions, and proposed an effective and efficient multi-window technique for change detection. Generalization to other loss functions or procedures is possible and will be considered in future works.

Appendix

Let $S^{(k)}_{n_1:n_2} = \sum_{n=L_{k-1}+n_1+1}^{L_{k-1}+n_2}(X_n - \mu_k)$, and $S^{(k)} = S^{(k)}_{0:N_k}$. Let $S^{(k_1,k_2)}_{n_1:n_2} = S^{(k_1)}_{n_1:N_{k_1}} + S^{(k_1+1)} + \cdots + S^{(k_2-1)} + S^{(k_2)}_{0:n_2}$ for $k_1 < k_2$ and $S^{(k_1,k_2)}_{n_1:n_2} = S^{(k_1)}_{n_1:n_2}$ for $k_1 = k_2$.

We define the within-segment loss $Q^{(k)}_{n_1:n_2} = Loss_q(x_{L_{k-1}+n_1+1}, \ldots, x_{L_{k-1}+n_2})$, and the cross-segment loss

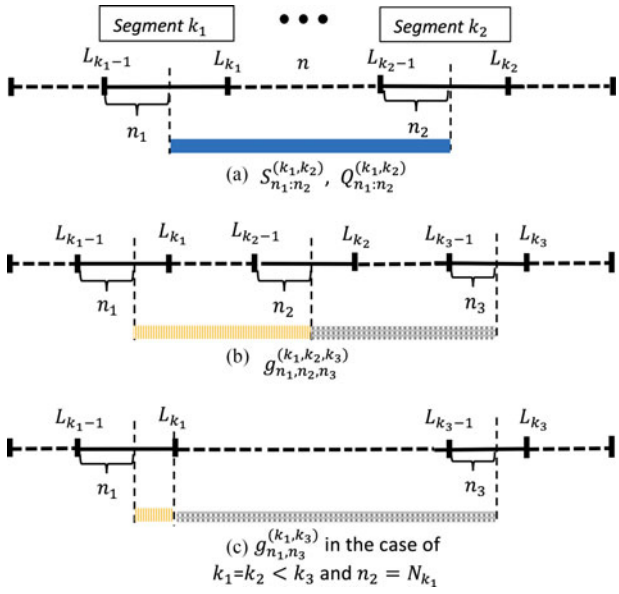

Fig. 8. Illustration of some frequently used notations in the proofs.

$Q^{(k_1,k_2)}_{n_1:n_2} = Loss_q(x_{L_{k_1-1}+n_1+1}, \ldots, x_{L_{k_2-1}+n_2})$. Let

$$g^{(k_1,k_2,k_3)}_{n_1,n_2,n_3} = Q^{(k_1,k_3)}_{n_1:n_3} - (Q^{(k_1,k_2)}_{n_1:n_2} + Q^{(k_2,k_3)}_{n_2:n_3}), \qquad (7)$$

referred to as the *decomposition gain*. In the case of $k_1 = k_2 < k_3, n_2 = N_{k_1}$, we denote $g^{(k_1,k_2,k_3)}_{n_1,n_2,n_3}$ by $g^{(k_1,k_3)}_{n_1,n_3}$; In the case of $k_1 = k_2 = k_3$, we denote $g^{(k_1,k_2,k_3)}_{n_1,n_2,n_3}$ by $g^{(k_1)}_{n_1,n_2,n_3}$. If $n_1 \geq n_2$ or $n_2 \geq n_3$ in the above definitions, the corresponding values are understood to be zeros. For each $d = 1, \ldots, D$, let $X_{n,d}$ and $S^{(k)}_{n_1:n_2,d}$ denote the dth component of X_n and $S^{(k)}_{n_1:n_2}$, respectively. Recall that the quadratic loss Q can also be interpreted as the negative log-likelihood of fitting a Gaussian distribution. Since the goodness of fit is improved by fitting two instead of one Gaussian, the decomposition gain g is always nonnegative. Figure 8 is an illustration of some frequently used notations used in the proofs.

Technical Lemmas: Lemma 1 proves the asymptotic distribution of a decomposition gain and its tight almost sure bound.

Lemma 1: Suppose that $\{X_n : n = 1, \ldots, N_1\}$ and $\{X_n : n = N_1 + 1, \ldots, N\}$ are independent random variables from the same distribution $\mathcal{G}$, with mean μ and variance V. Let $N_2 = N - N_1$. Assume that $N_1, N_2 \to \infty$ as $N \to \infty$, and N_1, N_2 depend only on N. Then $g^{(1,2)}_{0,N_2}$ (the decomposition gain) converges in distribution to $Z^{\mathrm{T}} V Z$, where $Z \in \mathcal{N}_D(0, I)$. Moreover,

$$\limsup_{N\to\infty} \frac{g^{(1,2)}_{0,N_2}}{\log\log(\min\{N_1, N_2\})} = C \quad (a.s.) \qquad (8)$$

for some positive constant $C \leq 8\, tr(V)$.

Proof: Let $p_1 \triangleq N_1/N, p_2 \triangleq 1 - p_1$. By direct calculation, we obtain

$$g^{(1,2)}_{0,N_2} = \left| \sqrt{\frac{p_2}{N_1}} \sum_{n=1}^{N_1} X_n - \sqrt{\frac{p_1}{N_2}} \sum_{n=N_1+1}^{N} X_n \right|^2. \qquad (9)$$

Let $Y_N^{(k)} \triangleq S_{0:N_k}^{(k)}/\sqrt{N_k}$, $k = 1, 2$. By the central limit theorem (CLT), $Y_N^{(1)}, Y_N^{(2)}$ converge in distribution to two independent $\mathcal{N}_D(0, V)$ random variables, respectively denoted by $Y^{(1)}, Y^{(2)}$. Therefore,

$$\sqrt{p_2}Y_N^{(1)} - \sqrt{p_1}Y_N^{(2)} = \sqrt{p_2}Y^{(1)} - \sqrt{p_1}Y^{(2)} + o_p(1)$$

converges in distribution to a random variable $W \sim \mathcal{N}_D(0, V)$. Let $W = V^{1/2}Z$, then $Z \sim \mathcal{N}_D(0, I)$. It follows that $g_{0,N_2}^{(1,2)}$ converges to $Z^{\mathrm{T}}VZ$. Furthermore, by the law of the iterated logarithm, for all $k = 1, 2$, $d = 1, \ldots, D$,

$$\limsup_{N_k \to \infty} Y_{N,d}^{(k)}/\sqrt{2V_{dd}\log\log N_k} = 1, \quad (a.s.), \tag{10}$$

where $Y_{N,d}^{(k)}$ and V_{dd} denote the dth entry of $Y_N^{(k)}$ and the (d, d)th entry of V, respectively. Note that $|\sqrt{p_2}Y_{N,1}^{(1)} - \sqrt{p_1}Y_{N,1}^{(2)}|^2 \le g_{0,N_2}^{(1,2)} \le \sum_{d=1}^{D}\{\sqrt{p_2}|Y_{N,d}^{(1)}| + \sqrt{p_1}|Y_{N,d}^{(2)}|\}^2$, where the second inequality follows from triangle inequality. We infer from (10) that for any fixed $\delta \in (0, 1)$,

$$\limsup_{N \to \infty} g_{0,N_2}^{(1,2)} \left[\sum_{d=1}^{D}\left(\sqrt{2p_2C_{d,1}} + \sqrt{2p_1C_{d,2}}\right)^2\right]^{-1} \le 1 \quad a.s., \tag{11}$$

$$g_{0,N_2}^{(1,2)} \ge \left(\sqrt{2p_2\delta C_{1,1}} + \sqrt{2p_1\delta C_{1,2}}\right)^2 \quad i.o. \tag{12}$$

for sufficiently large N. where $C_{d,1} \triangleq V_{dd}\log\log N_1$, $C_{d,2} \triangleq V_{dd}\log\log N_2$. From (12), it is easy to observe (with $\delta = 1/2$) that

$$g_{0,N_2}^{(1,2)} > V_{11}\log\log(\min\{N_1, N_2\}) \quad i.o. \tag{13}$$

It can be proved that for any $n_2 \ge n_1 \ge 34$,

$$\frac{n_1}{n_1 + n_2}\log\log(n_2) \le \frac{1}{2}\log\log(n_1). \tag{14}$$

It follows from (11) and (14) that

$$\limsup_{N \to \infty} g_{0,N_2}^{(1,2)} \left[8\,tr(V)\log\log(\min\{N_1, N_2\})\right]^{-1} \le 1 \quad (a.s.) \tag{15}$$

Furthermore, (13) and (15) imply the desired (8). ■

Remark 4: Lemma 1 implies that splitting a sequence of i.i.d. points into two halves increases the goodness of fit (measured by quadratic loss) by $O_p(1)$. Therefore, an AIC-like criterion (with constant penalty) always produces a non-vanishing overfitting probability.

Lemma 2 proves that there would be a large decomposition gain if the the identified change points are far away from the true change points. It is to be used in the proof of no underfitting.

Lemma 2: Under model assumption (M.2), for any $j \in \{1, \ldots, M_0\}$ and n_1, n_2 satisfying $N_j^{-1}n_1, N_{j+1}^{-1}n_2 \in [c^{-1}, 1]$, where $c > 1$ is some constant, we have

$$g_{N_j - n_1, n_2}^{(j,j+1)} > \frac{1}{3}|\mu_j - \mu_{j+1}|^2 \min\{n_1, n_2\} \tag{16}$$

for sufficiently large N almost surely.

Proof: From Equation (9) (note that its derivation does not require the two segments to have the same mean), we obtain

$$g_{N_j - n_1, n_2}^{(j,j+1)} = \left|\sqrt{\frac{n_2}{n}}Y^{(1)} - \sqrt{\frac{n_1}{n}}Y^{(2)} + \sqrt{\frac{n_1 n_2}{n}}(\mu_j - \mu_{j+1})\right|^2,$$

where $n = n_1 + n_2$, $Y^{(1)} = \sum_{i=L_j - n_1 + 1}^{L_j}(X_i - \mu_j)/\sqrt{n_1}$, $Y^{(2)} = \sum_{i=L_j+1}^{L_j+n_2}(X_i - \mu_{j+1})/\sqrt{n_2}$. By triangle inequality $g_{N_j - n_1, n_2}^{(j,j+1)} \ge (|B| - |A|)^2$, where

$$A = \sqrt{\frac{n_2}{n}}Y^{(1)} - \sqrt{\frac{n_1}{n}}Y^{(2)}, \quad B = \sqrt{\frac{n_1 n_2}{n}}(\mu_j - \mu_{j+1}).$$

By Strassen's invariance principle [66, Chapter 5], for each individual ω in a set of probability one, for each $d = 1, \ldots, D$

$$\limsup_{N_j \to \infty} \frac{\sum_{i=L_j - n_1(\omega)+1}^{L_j}(X_{i,d}(\omega) - \mu_{j,d})}{\sqrt{2V_{j,dd}N_j\log\log N_j}} \le 1,$$

and a similar inequality holds that replaces j by $j + 1$. This implies that

$$\frac{Y_d^{(1)}(\omega)}{\sqrt{2V_{j,dd}\log\log n_1}}, \frac{Y_d^{(2)}(\omega)}{\sqrt{2V_{j+1,dd}\log\log n_2}} \le \sqrt{c+1}$$

for sufficiently large N (thus N_j, N_{j+1}). For brevity, we have simplified $n_1(\omega), n_2(\omega)$ to n_1, n_2. Let $v_d \triangleq \max\{V_{j,dd}, V_{j+1,dd}\}$. From the above inequalities and (14),

$$|A|^2 < \sum_{d=1}^{D} 8(c+1)v_d\log\log(\min\{n_1, n_2\}) \tag{17}$$

for sufficiently large N almost surely. Then it follows from

$$|B| = \sqrt{\frac{n_1 n_2}{n}}|\mu_j - \mu_{j+1}| \ge \sqrt{\frac{\min\{n_1, n_2\}}{2}}|\mu_j - \mu_{j+1}|$$

that (16) holds. ■

Proof of Theorem 1: For the case $L = 0$, $\{Y_n : n = 1, \ldots, N\}$ are independent, and $\hat{\psi}_1 = \sum_{n=1}^{N_1} Y_n/N_1$, $\hat{\psi}_2 = \sum_{n=N-N_2+1}^{N} Y_n/N_2$. Thus, $\sqrt{N_1}(\hat{\psi}_1 - \psi)$ and $\sqrt{N_2}(\hat{\psi}_2 - \psi)$ converge to Gaussian random variables that are independent. It remains to prove for the case $L > 0$. Choose N_1', N_2' such that $N_1'/N_1, N_2'/N_2 \to 1, N_1 - N_1', N_2 - N_2' \to \infty$. Let $\hat{\psi}_1', \hat{\psi}_2' \in \mathcal{R}^{L+1}$ respectively denote the estimated filters from $\{X_1, \cdots, X_{N_1'}\}$ and $\{X_{N-N_2'+1}, \cdots, X_N\}$ using the least squares method. It is well known that $\sqrt{N_1'}(\hat{\psi}_1' - \psi)$, $\sqrt{N_2'}(\hat{\psi}_2' - \psi)$ respectively converge in distribution to $Z_1, Z_2 \sim \mathcal{N}(0, \sigma^2(\Gamma_L^*)^{-1})$, where $(\Gamma_L^* = \begin{smallmatrix} 1 & 0 \\ 0 & \Gamma_L \end{smallmatrix})$ and Γ_L is the covariance matrix of order L [53, Appendix 7.5]. Because X_n is strongly mixing under Assumption (A.1) [66], $\sqrt{N_1'}(\hat{\psi}_1' - \psi)$ and $\sqrt{N_2'}(\hat{\psi}_2' - \psi)$ are asymptotically independent. Thus, Z_1 and Z_2 are independent. To prove that $\sqrt{N_1}(\hat{\psi}_1 - \psi)$ and $\sqrt{N_2}(\hat{\psi}_2 - \psi)$ are asymptotically independent, by Slutsky' theorem, it remains to prove that

$$\sqrt{N_1}(\hat{\psi}_1 - \psi) = \sqrt{N_1'}(\hat{\psi}_1' - \psi) + o_p(1), \tag{18}$$

$$\sqrt{N_2}(\hat{\psi}_2 - \psi) = \sqrt{N_2'}(\hat{\psi}_2' - \psi) + o_p(1). \tag{19}$$

We prove the former equation since the latter one can be similarly proved. Let Z_1 denote the $(N_1 - L) \times L$ matrix whose (i,j)th element is $y_{N_1+1-i,N+1-i-j}$, and $W_1 \triangleq [y_{N_1}, \ldots, y_{L+1}]^{\mathrm{T}}$, $E_1 \triangleq [\varepsilon_{N_1}, \ldots, \varepsilon_{L+1}]^{\mathrm{T}}$. Since $\hat{\psi}_1$ is estimated from least squares method, it can be written in the matrix form $\hat{\psi}_1 = (Z_1^{\mathrm{T}} Z_1)^{-1} Z_1^{\mathrm{T}} W_1 = \psi_1 + (Z_1^{\mathrm{T}} Z_1)^{-1} Z_1^{\mathrm{T}} E_1$. We similarly define $Z_1^{'}, W_1^{'}, E_1^{'}$ by substituting N_1 with $N_1^{'}$, and write $\hat{\psi}_1^{'} = \psi_1 + \{(Z_1^{'})^{\mathrm{T}} Z_1^{'}\}^{-1} (Z_1^{'})^{\mathrm{T}} E_1^{'}$. Therefore,

$$\sqrt{N_1}(\hat{\psi}_1 - \psi_1) = \left(\frac{Z_1^{\mathrm{T}} Z_1}{N_1} \right)^{-1} \frac{Z_1^{\mathrm{T}} E_1}{\sqrt{N_1}}, \tag{20}$$

$$\sqrt{N_1}(\hat{\psi}_1^{'} - \psi_1) = \left\{ \frac{(Z_1^{'})^{\mathrm{T}} (Z_1^{'})}{N_1} \right\}^{-1} \frac{(Z_1^{'})^{\mathrm{T}} E_1^{'}}{\sqrt{N_1}}. \tag{21}$$

Let Γ_L denote the covariance matrix of order L. Recall that as $N \to \infty$,

$$\frac{Z_1^{\mathrm{T}} Z_1}{N_1}, \frac{(Z_1^{'})^{\mathrm{T}} Z_1^{'}}{N_1^{'}}, \frac{Z_1^{\mathrm{T}} Z_1 - (Z_1^{'})^{\mathrm{T}} Z_1^{'}}{N_1 - N_1^{'}} \to_p \Gamma_L^*$$

$$\frac{Z_1^{\mathrm{T}} E_1}{\sqrt{N_1}}, \frac{(Z_1^{'})^{\mathrm{T}} E_1^{'}}{\sqrt{N_1^{'}}}, \frac{Z_1^{\mathrm{T}} E_1 - (Z_1^{'})^{\mathrm{T}} E_1^{'}}{\sqrt{N_1 - N_1^{'}}} \to_d \mathcal{N}(0, \sigma^2 \Gamma_L^*) \tag{22}$$

due to the central limit theorem for martingale difference sequences [53, Appendix 7.5]. Therefore,

$$\frac{(Z_1^{'})^{\mathrm{T}} E_1^{'}}{\sqrt{N_1}} = \frac{Z_1^{\mathrm{T}} E_1 + \sqrt{N_1 - N_1^{'}} O_p(1)}{\sqrt{N_1}} = \frac{Z_1^{\mathrm{T}} E_1}{\sqrt{N_1}} + o_p(1),$$

and (20) further implies $\sqrt{N_1}(\hat{\psi}_1 - \psi_1) = O_p(1)$ and

$$\sqrt{N_1}(\hat{\psi}_1^{'} - \psi_1) = \left\{ \frac{(Z_1^{'})^{\mathrm{T}} (Z_1^{'})}{N_1} \right\}^{-1} \frac{Z_1^{\mathrm{T}} E_1}{\sqrt{N_1}} + o_p(1). \tag{23}$$

Straightforward calculations using (20) and (23) give

$$\sqrt{N_1}(\hat{\psi}_1 - \psi) - \sqrt{N_1^{'}}(\hat{\psi}_1^{'} - \psi) = \tau\, O_p(1) + o_p(1).$$

$$\text{where } \tau \triangleq \left(\frac{Z_1^{\mathrm{T}} Z_1}{N_1} \right)^{-1} - \left(\frac{(Z_1^{'})^{\mathrm{T}} (Z_1^{'})}{N_1} \right)^{-1}$$

It remains to prove that $\tau = o_p(1)$. In fact, from (22),

$$\tau = \frac{N_1}{N_1^{'}} \left\{ \frac{(Z_1^{'})^{\mathrm{T}} (Z_1^{'})}{N_1^{'}} \right\}^{-1} \frac{N_1^{'} - N_1}{N_1} \left\{ \frac{Z_1^{\mathrm{T}} Z_1 - (Z_1^{'})^{\mathrm{T}} (Z_1^{'})}{N_1 - N_1^{'}} \right\} \left(\frac{Z_1^{\mathrm{T}} Z_1}{N_1^{'}} \right)^{-1} = o_p(1).$$

Proof of Theorem 2: We first prove that $f(N)$ should be at least $\Theta(\log \log N)$ to ensure strong consistency for any $M_0 \geq 0$. We prove for the case $M_0 = 0$, and its proof can be straightforwardly extended to $M_0 > 0$. The event $\hat{M} = 0$ implies the event $Q_{0:N/2}^{(1)} + Q_{N/2,N}^{(1)} + f(N) \geq Q_{0:N}^{(1)}$. In other words, $g_{0,N/2,N}^{(1)} > f(N)$ implies the event $\hat{M} \neq 0$. By Lemma 1, there exists $C_1 > 0$ such that $g_{0,N/2,N}^{(1)} \geq C_1 \log \log N$ *i.o.* This implies that if $f(N) < C_1 \log \log N$, then $g_{0,N/2,N}^{(1)} > f(N)$ *i.o.* and thus $\hat{M} \neq M$ *i.o.*

Next, we prove that $\Theta(\log \log N)$ is sufficient for strong consistency when $M_0 = 0$. The event $\hat{M} > 0$ implies the event that there exist $0 < n_1 < n_2$ such that $g_{0,n_1,n_2}^{(1)} \geq f(N)$ and that $n_1, n_2 - n_1 \geq \beta(N) = \Theta(N)$. By a similar derivation to (17) in Lemma 2, we can show that for sufficiently large N

$$g_{0,n_1,n_2}^{(1)} < 8(c+1) tr(V_1) \log \log N \quad (a.s.) \tag{24}$$

where $c > 1$ is some constant. Thus, given that $f(N) = C_2 \log \log N$ for large enough $C_2 > 0$, $g_{0,n_1,n_2}^{(1)} < f(N)$ for sufficiently large N almost surely. This implies that $\hat{M} \to_{a.s.} 0$ as $N \to \infty$.

Proof of Theorem 3: We first prove that there is no underfitting, i.e. $\hat{M} \geq M_0$. It suffices to prove that for each ω from a set of probability one, there exists a positive integer N_ω such that for all $N > N_\omega$, $\hat{M} \neq m$ for each $m = 1, \ldots, M_0 - 1$. We prove the result by contradiction. Assume that $\hat{M} = m < M_0$. Then there exists at least one detected segment that consists of points from at least two neighboring segments, say the $(j-1)$th and jth, and that the numbers of points from the two segments are at least $N_{j-1}/2$ and $N_j/2$, respectively. Without loss of generality, we assume $N_1, \ldots, N_{M_0+1}$ to be even. In other words, the points $\{X_n : n = L_{j-1} - N_{j-1}/2 + 1, \ldots, L_{j-1} + N_j/2\}$ are contained in the kth detected segment for some $k = 1, \ldots, m+1$. Following the notation of Algo. 2, let $\hat{e}_m$ denote the minimal within-segment quadratic loss given m segments. We consider another configuration of change points: for the set of change points that give $\hat{e}_m$, keep all other segments except for the kth segment unchanged, and split the kth segment into four segments the middle two of which are $\{X_{L_{j-1}-N_{j-1}/4+1}, \ldots, X_{L_{j-1}}\}$ and $\{X_{L_{j-1}+1}, \ldots, X_{L_{j-1}+N_j/4}\}$. Then the number of segments will increase from m to $m+3$, and we obtain from Lemma 2 that for sufficiently large N, the increased within-segment quadratic loss is larger than $C_1 \min\{N_{j-1}, N_j\}$ almost surely, where the constant $C_1 = \underline{\Delta}_\mu^2/12$. Since $\hat{e}_{m+3}$ is the global minimum of the within-segment quadratic loss under $m+3$ change points, we obtain

$$\hat{e}_m - \hat{e}_{m+3} > C_1 \min\{N_{j-1}, N_j\} \quad (a.s.) \tag{25}$$

On the other hand, because $m + 3 \leq M_{\max}$ and the condition in step 3 of Algo. 2 is satisfied (since each new segment is at least $\min_{k=1,\ldots,M_0+1} N_k/4 \geq \beta(N)$ for sufficiently large N), $\hat{e}_{m+3}$ is a valid output of Algo. 2. Furthermore, the event $\hat{M} = m$ implies the event $\hat{e}_m - \hat{e}_{m+3} \leq 3f(N)$. In addition, $3f(N) < C_1 \min\{N_{j-1}, N_j\}$ for sufficiently large N due to Assumption (A.3). Thus, $\hat{e}_m - \hat{e}_{m+3} < C_1 \min\{N_{j-1}, N_j\}$ which contradicts the inequality in (25). Therefore, $\hat{M} \neq m$ for sufficiently large N almost surely. By similar reasoning we can prove Inequality (4).

Second, we prove the over-fitting part by contradiction. Assume that $\hat{M} = m > 2M_0$, by the pigeonhole principle there are two detected segments that are adjacent and that belong to the same true segment. Without loss of generality, suppose that $\{X_n : n = \tau + 1, \ldots, \tau + n_1\}$ and $\{X_n : n = \tau + n_1 + 1, \ldots, \tau + n_1 + n_2\}$ are from distribution $\mathcal{G}_k$. We consider the

configuration that merges the aforementioned two segments into one while keeping other segments unchanged. Since $n_1, n_2 \geq \beta(N) = \Theta(N)$, via a similar derivation of (17), it can be proved that for some constant $C_0 > 1$, $\hat{e}_{m-1} - \hat{e}_m < C_0 \log\log N$ for sufficiently large N almost surely. On the other hand, the event $\hat{M} = m$ implies that $\hat{e}_{m-1} - \hat{e}_m \geq f(N)$. Whenever $f(N) \geq C_0 \log\log(N)$, $\hat{e}_{m-1} - \hat{e}_m \geq C_0 \log\log(N)$ which is a contradiction to the previous inequality. Therefore, $\mathrm{pr}\{\limsup_{N\to\infty}(\hat{M} > 2M_0)\} \leq \mathrm{pr}\{\limsup_{N\to\infty}(\hat{e}_{m-1} - \hat{e}_m < C_0 \log\log N)\} = 0$.

Proof of Theorem 4: To prove Theorem 4, we need the following additional technical lemmas. The lemmas serve to enumerate various configurations of change points (events) that will not eventually happen given sufficiently large sample size. Loosely speaking, in those configurations, either "there exists a detected change point that is redundant" or "a true change point is too far away from all the detected change points".

Lemma 3 shows that if there are two neighboring segments that consist of points from the same underlying true segment, then Algo. 1 will almost surely merge them.

For each $k = 1, \ldots, M_0 + 1$, we define $P_k = \{L_{k-1} + 1, \ldots, L_k\}$, and use $\{X_n^{(k)} : n = 1, \ldots, N_k\}$ to represent the points in the kth true segment, namely $\{X_{L_{k-1}+1}, \ldots, X_{L_{k-1}+N_k}\}$.

Lemma 3: For each $k = 1, 2, \ldots, M_0 + 1$, let $E_{k,N}$ denote the event that Algo. 2 produces two neighboring segments that are both subsets of $\{X_n, n \in P_k\}$, the true kth segment. In other words, let $E_{k,N}$ denote the event "there exist $n_1, n_2, n_3 \in \mathbb{N}$ such that $0 \leq n_1 < n_2 < n_3 \leq N_k$, and $\{X_n^{(k)} : n_1 < n \leq n_2\}, \{X_n^{(k)} : n_2 < n \leq n_3\}$ are two detected segments". Assume that $f(N) \geq C \log N$ for some constant $C > 16D/c_0$. Then $\mathrm{pr}\big(\limsup_{N\to\infty} E_{k,N}\big) = 0$.

Proof: For brevity, we let $p_1 \triangleq (n_2 - n_1)/(n_3 - n_1)$, $p_2 \triangleq 1 - p_1$. Since $E_{k,N}$ implies the event that the loss of merging the two segments into one is larger than $f(N)$, we obtain from Equality (9) and the union bound that

$$\mathrm{pr}(E_{k,N}) \leq \sum_{\mathcal{I}} \mathrm{pr}\left\{\left|\frac{\sqrt{p_2}S^{(k)}_{n_1:n_2}}{\sqrt{n_2 - n_1}} - \frac{\sqrt{p_1}S^{(k)}_{n_2:n_3}}{\sqrt{n_3 - n_2}}\right|^2 > f(N)\right\}$$

where $I \triangleq \{(n_1, n_2, n_3) : 1 \leq n_1 < n_2 < n_3 \leq N_k\}$. For any tuple (n_1, n_2, n_3),

$$\begin{aligned}
&\mathrm{pr}\left\{\left|\sqrt{p_2}\frac{S^{(k)}_{n_1:n_2}}{\sqrt{n_2 - n_1}} - \sqrt{p_1}\frac{S^{(k)}_{n_2:n_3}}{\sqrt{n_3 - n_2}}\right|^2 > f(N)\right\} \\
&\quad\leq \mathrm{pr}\left\{\bigcup_{d=1}^{D}\left\{\left(\frac{\sqrt{p_2}S^{(k)}_{n_1:n_2,d}}{\sqrt{n_2 - n_1}} - \frac{\sqrt{p_1}S^{(k)}_{n_2:n_3,d}}{\sqrt{n_3 - n_2}}\right)^2 > \frac{f(N)}{D}\right\}\right\} \\
&\quad\leq \sum_{d=1}^{D}\mathrm{pr}\left\{\left(\frac{\sqrt{p_2}S^{(k)}_{n_1:n_2,d}}{\sqrt{n_2 - n_1}} - \frac{\sqrt{p_1}S^{(k)}_{n_2:n_3,d}}{\sqrt{n_3 - n_2}}\right)^2 > \frac{f(N)}{D}\right\}.
\end{aligned}$$

From triangular inequality and $p_1, p_2 < 1$, each term in the above summation is further upper bounded by

$$\begin{aligned}
&\sum_{(n',n'')=(n_1,n_2)\text{ or }(n_2,n_3)} \mathrm{pr}\left\{\left|\frac{S^{(k)}_{n':n'',d}}{n'' - n'}\right| > \frac{1}{2}\sqrt{\frac{f(N)}{D(n'' - n')}}\right\} \\
&< 2\exp\left\{-c_0(n'' - n')\frac{f(N)}{4D(n'' - n')}\right\} \qquad (26)
\end{aligned}$$

where the last inequality is due to Assumption (A.4). Taking and (26) into previous inequalities, we obtain

$$\mathrm{pr}(E_{k,N}) \leq N_k^3(2D)\exp\left\{-\frac{c_0 f(N)}{4D}\right\} \leq 2DN^{-C'}$$

for a constant $C' > 1$, where the last inequality follows from the assumption of Lemma 3. Therefore $\sum_{N=1}^{\infty}\mathrm{pr}(E_{k,N}) < \infty$ and by Borel-Cantelli lemma $\mathrm{pr}(\limsup_{N\to\infty} E_{k,N}) = 0$. ■

As a follow up result to Lemma 3, Lemma 4 shows that if there are at most η points from another true segment from one side involved, then Algo. 1 still merges them almost surely as long as η is small compared with the penalty increment $f(N)$.

Lemma 4: Suppose that $M_0 > 0$. For each $k = 1, 2, \ldots, M_0$, let $E_{k,N}$ denote the event that Algo. 2 produces two neighboring segments the first of which is a subset of $\{X_n, n \in P_k\}$ and the second of which consists of points from $\{X_n, n \in P_k\}$ and at most η points from $\{X_n, n \in P_{k+1}\}$, where $1 \leq \eta \leq N_{k+1}$. In other words, let $E_{k,N}$ denote the event "there exist $n_1, n_2, n_3 \in \mathbb{N}$ such that $0 \leq n_1 < n_2 < N_k, 1 \leq n_3 \leq \eta$, and $\{X_n^{(k)} : n = n_1 + 1, \ldots, n_2\}, \{X_n^{(k)} : n = n_2 + 1, \ldots, N_k\} \cup \{X_n^{(k+1)} : n = 1, \ldots, n_3\}$ are two detected segments". Assume that for some constant $C > 64D/c_0$,

$$f(N) \geq \max\{16|\mu_k - \mu_{k+1}|^2\eta,\ C\log N\}. \qquad (27)$$

Then $\mathrm{pr}\big(\limsup_{N\to\infty} E_{k,N}\big) = 0$.

Proof: The proof of Lemma 4 is similar to that of Lemma 3, so we only highlight the main differences.

Similar to the proof of Lemma 3, we obtain from (9) and the union bound that

$$\begin{aligned}
\mathrm{pr}(E_{k,N}) \leq \sum_{\substack{1\leq n_1<n_2<N_k \\ 1\leq n_3\leq\eta}} \mathrm{pr}\Bigg\{&\Bigg|\sqrt{\frac{N_k - n_2 + n_3}{N_k - n_1 + n_3}}\frac{S^{(k)}_{n_1:n_2}}{\sqrt{n_2 - n_1}} \\
&- \sqrt{\frac{n_2 - n_1}{N_k - n_1 + n_3}}\frac{S^{(k)}_{n_2:N_k} + S^{(k+1)}_{0:n_3}}{\sqrt{N_k - n_2 + n_3}} + C_{n_1,n_2,n_3}\Bigg|^2 > f(N)\Bigg\} \qquad (28)
\end{aligned}$$

where

$$C_{n_1,n_2,n_3} \triangleq \sqrt{\frac{n_2-n_1}{(N_k-n_1+n_3)(N_k-n_2+n_3)}} \times n_3(\mu_k-\mu_{k+1}) < \sqrt{n_3}|\mu_{k+1}-\mu_k| \le \sqrt{\eta}|\mu_{k+1}-\mu_k| = \frac{\sqrt{f(N)}}{4}.$$

From the above inequalities, and using the triangle inequality, we obtain

$$\mathrm{pr}(E_{k,N}) \le \sum_{\substack{1\le n_1<n_2<N_k \\ 1\le n_3\le\eta}} \mathrm{pr}\left\{\left|\frac{S^{(k)}_{n_1:n_2}}{\sqrt{n_2-n_1}}\right| + \left|\frac{S^{(k)}_{n_2:N_k}}{\sqrt{N_k-n_2}}\right| + \left|\frac{S^{(k+1)}_{0:n_3}}{\sqrt{n_3}}\right| > \frac{3\sqrt{f(N)}}{4}\right\}. \quad (29)$$

Using the union bound similar to (26), each term in the above summation can be upper bounded by

$$\sum_{k',n',n''} \mathrm{pr}\left\{\left|\frac{S^{(k')}_{n':n''}}{\sqrt{n''-n'}}\right| > \frac{\sqrt{f(N)}}{4}\right\} \le 6D\exp\left\{-\frac{c_0 f(N)}{16D}\right\}, \quad (30)$$

where the summation is taken over a tuple (k',n',n'') of three possible values: (k,n_1,n_2), (k,n_2,N_k), or $(k+1,0,n_3)$. Bringing (30) into (29), we obtain

$$\mathrm{pr}(E_{k,N}) \le 6DN^3\exp\left\{-\frac{c_0 f(N)}{16D}\right\} \le 6DN^{-C'}$$

for a constant $C'>1$, where the last inequality follows from (27). Therefore $\sum_{N=1}^{\infty}\mathrm{pr}(E_{k,N})<\infty$ and by Borel-Cantelli lemma $\mathrm{pr}(\limsup_{N\to\infty}E_{k,N})=0$. ■

Lemma 5 shows that if there are at most η points from another true segment from two sides involved, then Algo. 1 still merges them almost surely as long as η is small compared with the penalty increment $f(N)$.

Lemma 5: Suppose that $M_0>1$. For each $k=2,\ldots,M_0$ and $1\le\eta\le\min\{N_{k-1},N_{k+1}\}$, let $E_{k,N}$ denote the event "there exist n_1,n_2,n_3 such that $1\le n_1\le N_k, 1\le n_2\le\eta, 1\le n_3\le\eta$, and $\{X_n^{(k-1)}: n=N_{k-1}-n_3+1,\ldots,N_{k-1}\}\cup\{X_n^{(k)}: n=1,\ldots,n_1\}$, $\{X_n^{(k)}: n=n_1+1,\ldots,N_k\}\cup\{X_n^{(k+1)}: n=1,\ldots,n_2\}$ are two detected segments". Assume that $f(N)\ge\max\{100|\mu_{k-1}-\mu_k|^2\eta,\ 100|\mu_k-\mu_{k+1}|^2\eta, C\log N\}$ for some constant $C>100D/c_0$. Then we obtain the desired $\mathrm{pr}(\limsup_{N\to\infty}E_{k,N})=0$.

Proof: Let $n_1'=N_k-n_1$. The main difference with the proof of Lemma 4 is the treatment of the constant term, which is

$$const_{n_1,n_2,n_3} = \sqrt{\frac{(n_3+n_1)(n_1'+n_2)}{n_3+N_k+n_2}}\left(\frac{n_3}{n_3+n_1}(\mu_{k-1}-\mu_k) + \frac{n_2}{n_1'+n_2}(\mu_k-\mu_{k+1})\right)$$
$$\le \sqrt{\frac{n_3(n_1'+n_2)}{(n_3+n_1)(n_3+N_k+n_2)}}\sqrt{n_3}|\mu_{k-1}-\mu_k| + \sqrt{\frac{(n_3+n_1)n_2}{(n_3+N_k+n_2)(n_1'+n_2)}}\sqrt{n_2}|\mu_k-\mu_{k+1}|\Big)$$
$$\le 2\sqrt{\eta}\max\{|\mu_{k-1}-\mu_k|,|\mu_k-\mu_{k+1}|\} \le \frac{\sqrt{f(N)}}{5}.$$

The remaining proof is similar to that of Lemma 4. ■

Lemma 6 proves that with probability one, for large N there is no detected segment that consists of points from the same true segment while having a small size (compared with the penalty increment $f(N)$).

Lemma 6: Suppose that $M_0>0$. For each $k=2,\ldots,M_0+1$, let $E_{k,N}$ denote the event "there exist n_1,n_2,n_3 such that $1\le n_1<n_2\le\eta, 1\le s\le k-1, 1\le n_3\le N_{k-s}$ and $\{X_n^{(k-s)}: n=N_{k-s}-n_3+1,\ldots,N_{k-s}\}\cup\cdots\cup\{X_n^{(k)}: n=1,\ldots,n_1\}, \{X_n^{(k)}: n=n_1+1,\ldots,n_2\}$ are two detected segments" where $1\le\eta\le N_k$. Assume that

$$f(N)\ge\max\{(s+3)^2\bar{\Delta}_\mu^2\eta,\ C\log N\} \quad (31)$$

for some constant $C>4(s+3)^2D/c_0$. Then we obtain the desired $\mathrm{pr}\big(\limsup_{N\to\infty}E_{k,N}\big)=0$.

Proof: Let $p_1=(L_{k-1}-L_{k-s+1}+n_2+n_3)^{-1}(L_{k-1}-L_{k-s+1}+n_1+n_3), p_2=1-p_1$. Similar to Inequality (28) we obtain

$$\mathrm{pr}(E_{k,N}) \le \sum_{\substack{1\le n_1<n_2<\eta \\ 1\le n_3\le N_{k-s}}} \mathrm{pr}\left\{\left|\sqrt{p_2}\frac{S^{(k-s)}_{N_{k-s}-n_3:N_{k-s}}+\cdots+S^{(k)}_{0:n_1}}{\sqrt{L_{k-1}-L_{k-s+1}+n_1+n_3}} - \sqrt{p_1}\frac{S^{(k)}_{n_1:n_2}}{\sqrt{n_2-n_1}} + \sqrt{p_1(n_2-n_1)}(\mu^*-\mu_k)\right|^2 > f(N)\right\}$$

where

$$\mu^* = \frac{n_3\mu_{k-s}+\sum_{j=k-s+1}^{k-1}N_j\mu_j+n_1\mu_k}{n_3+\sum_{j=k-s+1}^{k-1}N_j+n_1}$$

The last term in the above summation is bounded by

$$\left|\sqrt{p_1(n_2-n_1)}(\mu^*-\mu_k)\right| \le \sqrt{n_2-n_1}|\mu^*-\mu_k| \le \sqrt{\eta}\bar{\Delta}_\mu \le \frac{\sqrt{f(N)}}{s+3}.$$

Following similar proof in Inequalities (29)–(30), we obtain

$$\mathrm{pr}(E_{k,N}) \le 2(s+3)DN^3\exp\left\{-\frac{c_0 f(N)}{(s+3)^2D}\right\},$$

which implies $\mathrm{pr}(\limsup_{N\to\infty} E_{k,N}) = 0$ under Condition (31) and Borel-Cantelli lemma. ■

Lemmas 7 and 8 show that with probability one, eventually each true change point can not be too far away from the detected change point nearest to it.

Lemma 7: Suppose that $M_0 > 0$. For each $k = 2, \ldots, M_0 + 1$, let $E_{k,N}$ denote the event "there exist n_1, n_2, n_3 such that $1 \le n_1 \le N_{k-1}, 1 \le n_2 < n_3 < N_k$, and $\{X_n^{(k-1)} : n = N_{k-1} - n_1 + 1, \ldots, N_{k-1}\} \cup \{X_n^{(k)} : n = 1, \ldots, n_2\}, \{X_n^{(k)} : n = n_2 + 1, \ldots, n_3\}$ are two detected segments", and let $A_{k,N}$ denote the event

$$\min\{n_1, n_2\} > q_k(N) \triangleq \frac{250D \log N}{c_0 |\mu_{k-1} - \mu_k|^2}.$$

Then $\mathrm{pr}\{\limsup_{N\to\infty}(A_{k,N} \cap E_{k,N})\} = 0$.

Lemma 8: Suppose that $M_0 > 1$ and η is an integer that satisfies $1 \le \eta \le N_{k-1}$. For each $k = 2, \ldots, M_0$, let $E_{k,N}$ denote the event "there exist n_1, n_2, n_3 such that $1 \le n_1 \le N_{k-1}, 1 \le n_2 \le N_k, 1 \le n_3 \le \eta$, and $\{X_n^{(k-1)} : n = N_{k-1} - n_1 + 1, \ldots, N_{k-1}\} \cup \{X_n^{(k)} : n = 1, \ldots, n_2\}, \{X_n^{(k)} : n = n_2 + 1, \ldots, N_k\} \cup \{X_n^{(k+1)} : n = 1, \ldots, n_3\}$ are two detected segments", and let $A_{k,N}$ denote the event

$$\min\{n_1, n_2\} \ge \max\left\{ \frac{4\eta}{(\sqrt{2}-1)^2} \frac{|\mu_k - \mu_{k+1}|^2}{|\mu_{k-1} - \mu_k|^2},\ 2q_k(N) \right\},$$

where $q_k(N) = 250D \log N/(c_0|\mu_{k-1} - \mu_k|^2)$. Then $\mathrm{pr}\{\limsup_{N\to\infty}(A_{k,N} \cap E_{k,N})\} = 0$.

Proof: We prove Lemma 8. The proof of Lemma 7 is similar. Consider the postulation that the two detected segments are $\{X_n^{(k-1)} : n = N_k - n_1 + 1, \ldots, N_k\}, \{X_n^{(k)} : n = 1, \ldots, N_k\} \cup \{X_n^{(k+1)} : n = 1, \ldots, n_3\}$ instead. Then the event $E_{k,N}$ implies that $\mathfrak{L}_1 \le \mathfrak{L}_2$, where

$$\mathfrak{L}_1 = (Q^{(k-1)}_{N_{k-1}-n_1:N_{k-1}} + Q^{(k)}_{0:n_2} + g^{(k-1,k)}_{N_{k-1}-n_1:n_2}) + Q^{(k,k+1)}_{n_2:n_3},$$

$$\mathfrak{L}_2 = Q^{(k-1)}_{N_{k-1}-n_1:N_{k-1}} + (Q^{(k)}_{0:n_2} + g^{(k,k,k+1)}_{0,n_2,n_3} + Q^{(k,k+1)}_{n_2:n_3}).$$

Thus, $g^{(k-1,k)}_{N_{k-1}-n_1:n_2} \le g^{(k,k,k+1)}_{0,n_2,n_3}$. Moreover, we can calculate

$$g^{(k,k+1)}_{n_2,n_3} = \left| \sqrt{p_2} \frac{S^{(k-1)}_{N_{k-1}-n_1:N_{k-1}}}{\sqrt{n_1}} - \sqrt{p_1} \frac{S^{(k)}_{0:n_2}}{\sqrt{n_2}} + C \right|^2,$$

$$g^{(k,k,k+1)}_{0,n_2,n_3} = \left| \sqrt{p'_2} \frac{S^{(k)}_{0:n_2}}{\sqrt{n_2}} - \sqrt{p'_1} \frac{S^{(k)}_{n_2:N_k} + S^{(k+1)}_{0:n_3}}{\sqrt{N_k - n_2 + n_3}} + C' \right|^2,$$

where we have let $p_1 = n_1/(n_1 + n_2)$, $p_2 = 1 - p_1$, $p'_1 = n_2/(N_k + n_3)$, $p'_2 = 1 - p'_1$, and

$$C = \sqrt{p_1 n_2}(\mu_{k-1} - \mu_k),\ C' = \sqrt{\frac{p'_1}{N_k - n_2 + n_3}} \times n_3(\mu_k - \mu_{k+1}).$$

Therefore, we obtain the following inequality $|C| \le |C'| +$

$$\left| \frac{S^{(k-1)}_{N_{k-1}-n_1:N_{k-1}}}{\sqrt{n_1}} \right| + 2\left| \frac{S^{(k)}_{0:n_2}}{\sqrt{n_2}} \right| + \left| \frac{S^{(k)}_{n_2:N_k}}{\sqrt{N_k - n_2}} \right| + \left| \frac{S^{(k+1)}_{0:n_3}}{\sqrt{n_3}} \right|.$$

Let $\bar{n} = \min\{n_1, n_2\}$. Since

$$|C| - |C'| \ge \sqrt{\frac{\bar{n}}{2}} |\mu_{k-1} - \mu_k| - \sqrt{n_3}|\mu_k - \mu_{k+1}| \ge \frac{1}{2}\sqrt{\bar{n}}|\mu_{k-1} - \mu_k| \ge \frac{1}{2}\sqrt{2q_k(N)}|\mu_{k-1} - \mu_k|$$

where the last two inequalities are under $A_{k,N}$, we obtain

$$\mathrm{pr}(A_{k,N} \cap E_{k,N}) \le \mathrm{pr}\left(\bigcup \left\{ \left| \frac{S^{(k-1)}_{N_{k-1}-n_1:N_{k-1}}}{\sqrt{n_1}} \right| + \left| \frac{S^{(k)}_{0:n_2}}{\sqrt{n_2}} \right| + 2\left| \frac{S^{(k)}_{n_2:N_k}}{\sqrt{N_k - n_2}} \right| + \left| \frac{S^{(k+1)}_{0:n_3}}{\sqrt{n_3}} \right| \ge \frac{\sqrt{2q_k(N)}}{2} |\mu_{k-1} - \mu_k| \right\} \right).$$

where the union is over $1 \le n_1 \le N_{k-1}, 1 \le n_2 \le N_k, 1 \le n_3 \le \eta$. Using similar techniques as in (26), we obtain $\mathrm{pr}(A_{k,N} \cap E_{k,N}) \le 10DN^{-2}$. Finally, it follows from Borel-Cantelli lemma that $\mathrm{pr}\{\limsup_{N\to\infty}(A_{k,N} \cap E_{k,N})\} = 0$. ■

Proof of Theorem 4 (Main Body): For the case $M_0 = 0$, Lemma 3 guarantees that there is not overfitting. Next, we prove for the case $M_0 > 0$. It has been proved in Theorem 3 that there is no underfitting for sufficiently large N almost surely. Note that in its proof, only Assumptions (A.2)–(A.3) were used. To prove the strong consistency, it remains to prove that there is no overfitting. To that end, we define the following sequence of M_0 ($M_0 > 0$) constants $\eta_k(N), k = 1, \ldots, M_0$:

$$\eta_k(N) = \max\left\{ \frac{4\eta_{k+1}(N)}{(\sqrt{2}-1)^2} \frac{|\mu_{k+1} - \mu_{k+2}|^2}{|\mu_k - \mu_{k+1}|^2},\ 2q_{k+1}(N) \right\},$$

$$\eta_{M_0}(N) = q_{M_0+1}(N)$$

where $q_k(N), k = 2, \ldots, M_0 + 1$ have been defined in Lemma 7. We prove in three steps sketched below:

Step 1) If Algo. 2 is applied to $\{X_n, n = 1, \ldots, N_1 + \eta_1\}$ where $0 \le \eta_1 \le \min\{N_2, \eta_1(N)\}$, then almost surely no change point is detected as $N \to \infty$. In other words, when the data consists of one true segment and at most $\eta_1(N)$ extra points from another segment at the end, there is no spurious discovery of change points.

Step 2) Suppose that $M_0 > 1$. If Algo. 2 is applied to $\{X_n : n = 1, \ldots, L_k + \eta_k\}$ where k, η_k are any integers such that $1 \le k \le M_0$ and $0 \le \eta_k \le \eta_k(N)$, then almost surely $k - 1$ change points are detected, and the largest deviation of each true change point with its nearest detected change point is no larger than $\eta_k(N)$. In other words, when the data consists of k true segments plus at most $\eta_k(N)$ points from the $(k + 1)$th true segment, the number of true change points $k - 1$ is correctly selected.

Step 3) Suppose that $M_0 > 1$. If Algo. 2 is applied to $X_{1:N}$, then almost surely M_0 change points are detected.

Before we prove each step, recall that $2q_k = 500D \log N/(c_0|\mu_{k-1} - \mu_k|^2)$ and $c = 4/(\sqrt{2}-1)^2$. By simple calculations, we obtain the identity in (32) shown at the top of the next page, for each $k = 1, \ldots, M_0 - 1$, where $\eta^*(N)$ is defined in Theorem 4.

Proof of Step 1): If there is at least one change point produced by Algo. 2, then its location (in terms of the subscript of X_n) belongs to either $\{1, \ldots, N_1\}$ or $\{N_1 + 1, \ldots, N_1 + \eta_1\}$.

$$\eta_k(N) = \max\left\{\bigcup_{\tilde{k}=k,\ldots,M_0-2}\left\{2q_{\tilde{k}+2}(N)\prod_{j=k}^{\tilde{k}}\left(c\frac{|\mu_{j+1}-\mu_{j+2}|^2}{|\mu_j-\mu_{j+1}|^2}\right)\right\}\cup\{2q_{k+1}(N)\}\cup\left\{\eta_{M_0}(N)\prod_{j=k}^{M_0-1}\left(c\frac{|\mu_{j+1}-\mu_{j+2}|^2}{|\mu_j-\mu_{j+1}|^2}\right)\right\}\right\}$$

$$= \frac{500D\log N}{c_0}\max\left\{\bigcup_{\tilde{k}=k,\ldots,M_0-2}\left\{\frac{c^{\tilde{k}-k+1}}{|\mu_k-\mu_{k+1}|^2}\right\}\cup\left\{\frac{1}{|\mu_k-\mu_{k+1}|^2}\right\}\cup\left\{\frac{c^{M_0-k}}{2|\mu_k-\mu_{k+1}|^2}\right\}\right\} \le \eta^*(N) \quad (32)$$

However, the former case will not happen i.o. due to Lemma 4 (Condition (27) with $E_{k,N}$); and the latter case will not happen i.o. due to Lemma 6 (Condition (31) with $s=1$). We note that Conditions (27) and (31) are guaranteed by Inequalities (6) and (32).

Proof of Step 2): Suppose that the last two change points discovered by Algo. 2 are denoted by y, z, i.e. $X_{y+1},\ldots,X_z$ and $X_{z+1},\ldots,X_{L_k+\eta_k}$ are the last two segments.

The case $k=1$ has been proved in Step 1). Assume that $k>1$ and the statement is true for each $\tilde{k}$ such that $1\le\tilde{k}<k$. We prove that the statement holds for $\tilde{k}=k$ as well. We consider the three possible events: z belongs to either $\{1,\ldots,L_{k-1}\}$, $\{L_{k-1}+1,\ldots,L_k\}$ or $\{L_k+1,\ldots,L_k+\eta_k(N)\}$, and prove that almost surely k change points are discovered given each event.

(E1) z belongs to $\{1,\ldots,L_{k-1}\}$. Then, by induction hypothesis, at most $k-2$ change points are discovered from $\{X_n : n=1,\ldots,z\}$. Thus, there are at most $k-1$ change points in total.

(E2) z belongs to $\{L_{k-1}+1,\ldots,L_k\}$. There are three possible events: (E2.1) $y\le L_{k-2}$; (E2.2) $L_{k-2}+1\le y\le L_{k-1}$ (E2.3) $L_{k-1}+1\le y<z$.

Given (E2.1), since the induction hypothesis guarantees that at most $k-3$ change points are discovered from $\{X_n : n=1,\ldots,y\}$, there are at most $k-1$ change points in total.

Given (E2.2), from Lemma 8 and the way $\eta_{k-1}(N)$ was constructed, we obtain $\min\{L_{k-1}-y,\ z-L_{k-1}\}\le\eta_{k-1}(N)$ for all sufficiently large N almost surely.

Consider the following two subevents of (E2.2). (E2.2.1) $1\le z-L_{k-1}\le\eta_{k-1}(N)$; by induction hypothesis at most $k-2$ change points are discovered from $\{X_n : n=1,\ldots,z\}$, so there are at most $k-1$ change points in total; (E2.2.2) $L_{k-1}-y\le\eta_{k-1}(N)$; this will not happen i.o. by using Lemma 5 (in which the condition in Lemma 5 is guaranteed by (6)).

The event (E2.3) will not happen i.o. by applying Lemma 4 (where Condition (27) is guaranteed by (6)).

(E3) z belongs to $\{L_k+1,\ldots,L_k+\eta_k(N)\}$. We consider four subevents: (E3.1) $y\le L_{k-2}$, (E3.2) $L_{k-2}+1\le y\le L_{k-1}$, (E3.3) $L_{k-1}+1\le y\le L_k$, and (E3.4) $L_k+1\le y<z$.

For the event (E3.1), induction hypothesis guarantees that at most $k-2$ change points are discovered from $\{X_n : n=1,\ldots,y\}$, so there are at most $k-1$ change points in total. Both the events (E3.2) and (E3.3) will not happen i.o. by application of Lemma 6 (with $s=1,2$, where Condition (31) is guaranteed by (6)). Applying Lemma 3 (with $E_{k+1,N}$), it can be seen that the event (E3.4) will not happen i.o.

Proof of Step 3): Step 3 can be regarded as a special type of step 2 with $k=M_0+1$, and its proof follows from the above proof for events (E1), (E2).

To complete the proof, it remains to prove that the largest deviation of each true change point with its nearest detected change point is less than $\eta^*(N)$. This can be proved in similar fashion as above.

Remark 5: In summary, the key part of the proof is Step 2) which is by induction on k, the number of underlying true segments (despite a small amount of extra points). The induction step is completed using the events (E1), (E2.1), (E2.2.1), and (E3.1) for each k. We note that the number of induction steps is finite, i.e. $k=1,\ldots,M_0$. Because of that, any (finite) union of events that will not infinitely often happen will not eventually happen too.

REFERENCES

[1] M. Thottan and C. Ji, "Anomaly detection in IP networks," *IEEE Trans. Signal Process.*, vol. 51, no. 8, pp. 2191–2204, Aug. 2003.

[2] R. Jana and S. Dey, "Change detection in teletraffic models," *IEEE Trans. Signal Process.*, vol. 48, no. 3, pp. 846–853, Mar. 2000.

[3] S. Hammoudeh and H. Li, "Sudden changes in volatility in emerging markets: The case of Gulf Arab stock markets," *Int. Rev. Financial Anal.*, vol. 17, no. 1, pp. 47–63, 2008.

[4] J. J. Vidal, "Real-time detection of brain events in EEG," *Proc. IEEE*, vol. 65, no. 5, pp. 633–641, May 1977.

[5] S. J. Hawkins, A. J. Southward, and M. J. Genner, "Detection of environmental change in a marine ecosystem–evidence from the western english channel," *Sci. Total Environ.*, vol. 310, no. 1, pp. 245–256, 2003.

[6] H. Huntington, T. Callaghan, S. Fox, and I. Krupnik, "Matching traditional and scientific observations to detect environmental change: A discussion on arctic terrestrial ecosystems," *Ambio*, vol. 13, pp. 18–23, 2004.

[7] M. Basseville and V. Nikiforov, *Detection of Abrupt Changes: Theory and Application*. Englewood Cliffs, NJ, USA: Prentice-Hall, 1993.

[8] E. Page, "Continuous inspection schemes," *Biometrika*, vol. 41, pp. 100–115, 1954.

[9] F. Gustafsson, "The marginalized likelihood ratio test for detecting abrupt changes," *IEEE Trans. Autom. Control*, vol. 41, no. 1, pp. 66–78, Jan. 1996.

[10] R. A. Davis, D. Huang, and Y.-C. Yao, "Testing for a change in the parameter values and order of an autoregressive model," *Ann. Stat.*, vol. 23, pp. 282–304, 1995.

[11] T. Vogelsang, "Testing for a shift in mean without having to estimate serial-correlation parameters," *J. Bus. Econ. Statist.*, vol. 16, pp. 73–80, 1998.

[12] C. Incln and G. C. Tiao, "Use of cumulative sums of squares for retrospective detection of change of variance," *J. Amer. Statist. Assoc.*, vol. 89, pp. 913–923, 1994.

[13] H. L. Gombay, E. and M. Huskova, "Estimators and tests for change in variances," *Statist. Decisions*, vol. 14, pp. 145–159, 1996.

[14] R. A. Davis, T. Lee, and G. A. Rodriguez-Yam, "Break detection for a class of nonlinear time series models," *J. Time Ser. Anal.*, vol. 29, no. 5, pp. 834–867, 2008.

[15] G. E. Berkes, I. and L. Horváth, "Testing for changes in the covariance structure of linear processes," *J. Statist. Planning Inference*, vol. 139, pp. 2044–2063, 2009.

[16] D. Picard, "Testing and estimating change-points in time series," *Adv. Appl. Probab.*, vol. 17, pp. 841–867, 1985.

[17] G. S. Fishman, *Monte Carlo: Concepts, Algorithms, and Applications*. Berlin, Germany: Springer-Verlag, 1996.

[18] J. Huang, A. Smola, A. Gretton, K. M. Borgwardt, and B. Scholkopf, "Correcting sample selection bias by unlabeled data," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 601–608.

[19] M. Sugiyama, T. Suzuki, S. Nakajima, H. Kashima, P. von Bünau, and M. Kawanabe, "Direct importance estimation for covariance shift adaptation," *Ann. Inst. Statist. Math.*, vol. 60, no. 4, pp. 699–746, 2008.
[20] F. Desobry, M. Davy, and C. Doncarli, "An online kernel change detection algorithm," *IEEE Trans. Signal Process.*, vol. 53, no. 8, pp. 2961–2974, Aug. 2005.
[21] L. Vostrikova, "Detection disorder in multidimensional random processes," *Soviet Math. Doklady*, vol. 24, pp. 55–59, 1981.
[22] A. Scott and M. Knott, "A cluster analysis method for grouping means in the analysis of variance," *Biometrics*, vol. 30, pp. 507–512, 1974.
[23] D. M. Hawkins, "Fitting multiple change-point models to data," *Comput. Statist. Data Anal.*, vol. 37, pp. 323–341, 2001.
[24] M. Lavielle and G. Teyssiére, "Detection of multiple change-points in multivariate time series," *Lithuanian Math. J.*, vol. 46, pp. 287–306, 2006.
[25] I. E. Auger and C. E. Lawrence, "Algorithms for the optimal identification of segment neighborhoods," *Bulletin Math. Biol.*, vol. 51, no. 1, pp. 39–54, 1989.
[26] Y.-C. Yao, "Estimation of a noisy discrete-time step function: Bayes and empirical Bayes approaches," *Ann. Stat.*, vol. 12, pp. 1434–1447, 1984.
[27] B. Jackson *et al.*, "An algorithm for optimal partitioning of data on an interval," *IEEE Signal Process. Lett.*, vol. 12, no. 2, pp. 105–108, Feb. 2005.
[28] R. A. Davis, T. C. M. Lee, and G. A. Rodriguez-Yam, "Structural break estimation for nonstationary time series models," *J. Amer. Statist. Assoc.*, vol. 101, no. 473, pp. 223–239, 2006.
[29] A. Khaleghi and D. Ryabko, "Locating changes in highly dependent data with unknown number of change points," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 3086–3094.
[30] A. Khaleghi and D. Ryabko, "Asymptotically consistent estimation of the number of change points in highly dependent time series," in *Proc. 31st Int. Conf. Mach. Learn.*, 2014, pp. 539–547.
[31] P. Preuss, R. Puchstein, and H. Dette, "Detection of multiple structural breaks in multivariate time series," *J. Amer. Statist. Assoc.*, vol. 110, no. 510, pp. 654–668, 2015.
[32] B. Brodsky and B. Darkhovsky, *Nonparametric Methods in Change-Point Problems*. Norwell, MA, USA: Kluwer, 1993.
[33] P. Perron, "Dealing with structural breaks," in *Palgrave Handbook of Econometrics: Econometric Theory*, vol. 1, K. Patterson and T. C. Mills, Eds. Basingstoke, U.K.: Palgrave Macmillan, 2006, pp. 278–352.
[34] V. Jandhyala, S. Fotopoulos, I. MacNeill, and P. Liu, "Inference for single and multiple change-points in time series," *J. Time Ser. Anal.*, vol. 34, pp. 423–446, 2013.
[35] H. Akaike, "Fitting autoregressive models for prediction," *Ann. Inst. Statist. Math.*, vol. 21, no. 1, pp. 243–247, 1969.
[36] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Selected Papers of Hirotugu Akaike*. Berlin, Germany: Springer-Verlag, 1998, pp. 199–213.
[37] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, 1978.
[38] E. J. Hannan and B. G. Quinn, "The determination of the order of an autoregression," *J. Roy. Statist. Soc. Ser. B*, vol. 41, no. 2, pp. 190–195, 1979.
[39] C. Keribin, "Consistent estimation of the order of mixture models," *Sankhyā, Indian J. Statist., Ser. A*, vol. 62, pp. 49–66, 2000.
[40] H. Chen, J. Chen, and J. D. Kalbfleisch, "A modified likelihood ratio test for homogeneity in finite mixture models," *J. Roy. Statist. Soc. Ser. B*, vol. 63, no. 1, pp. 19–29, 2001.
[41] H. Chen, J. Chen, and J. D. Kalbfleisch, "Testing for a finite mixture model with two components," *J. Roy. Statist. Soc. Ser. B*, vol. 66, no. 1, pp. 95–115, 2004.
[42] J. Chen and A. Khalili, "Order selection in finite mixture models with a nonsmooth penalty," *J. Amer. Statist. Assoc.*, vol. 103, no. 484, pp. 1674–1683, 2008.
[43] F. K. Hui, D. I. Warton, and S. D. Foster, "Order selection in finite mixture models: Complete or observed likelihood information criteria?" *Biometrika*, vol. 102, pp. 724–730, 2015.
[44] Y.-C. Yao, "Estimating the number of change-points via Schwarz' criterion," *Statist. Probab. Lett.*, vol. 6, no. 3, pp. 181–189, 1988.
[45] E. S. Venkatraman, "Consistency results in multiple change-point problems," Ph.D. dissertation, Dept. Statist., Stanford Univ., Stanford, CA, USA, 1992.
[46] G. Rigaill, "A pruned dynamic programming algorithm to recover the best segmentations with 1 to k_max change-points," *J. de la Société Française de Statistique*, vol. 156, no. 4, pp. 180–205, 2015.
[47] C. Du, C.-L. M. Kao, and S. C. Kou.,"Stepwise signal extraction via marginal likelihood," *J. Amer. Statist. Assoc.*, vol. 111, no. 513, pp. 314–330, 2016.
[48] L. Horváth, "The maximum likelihood method for testing changes in the parameters of normal observations," *Ann. Stat.*, vol. 21, pp. 671–680, 1993.
[49] J. Chen and A. K. Gupta, *Parametric Statistical Change Point Analysis: With Applications to Genetics, Medicine, and Finance*. Berlin, Germany: Springer-Verlag, 2011.
[50] C. Inclan and G. C. Tiao, "Use of cumulative sums of squares for retrospective detection of changes of variance," *J. Amer. Statist. Assoc.*, vol. 89, no. 427, pp. 913–923, 1994.
[51] A. Sheikhattar, J. Fritz, S. Shamma, and B. Babadi, "Recursive sparse point process regression with application to spectrotemporal receptive field plasticity analysis," *IEEE Trans. Signal Process.*, vol. 64, no. 8, pp. 2026–2039, Apr. 2016.
[52] J. Ding, M. Noshad, and V. Tarokh, "Learning the number of autoregressive mixtures in time series using the gap statistics," in *Proc. IEEE Int. Conf. Data Mining Workshop*, 2015, pp. 1441–1446.
[53] G. E. Box, G. M. Jenkins, and G. C. Reinsel, *Time Series Analysis: Forecasting and Control*, 4th ed., vol. 734. Hoboken, NJ, USA: Wiley, 2008.
[54] P. J. Brockwell and R. A. Davis, *Time Series: Theory and Methods*. Berlin, Germany: Springer-Verlag, 2013.
[55] R. Shibata, "Selection of the order of an autoregressive model by Akaike's information criterion," *Biometrika*, vol. 63, no. 1, pp. 117–126, 1976.
[56] P. Bühlmann and S. Van De Geer, *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Berlin, Germany: Springer, 2011.
[57] L. Boysen, A. Kempe, V. Liebscher, A. Munk, and O. Wittich, "Consistencies and rates of convergence of jump-penalized least squares estimators," *Ann. Stat.*, vol. 37, pp. 157–183, 2009.
[58] J. Ding, M. Noshad, and V. Tarokh, "Data-driven learning of the number of states in multi-state autoregressive models," in *Proc. 53rd Annu. Allerton Conf. Commun., Control, Comput.*, 2015, pp. 418–425.
[59] R. T. Sutton and D. L. Hodson, "Atlantic ocean forcing of North American and European summer climate," *Science*, vol. 309, no. 5731, pp. 115–118, 2005.
[60] R. T. Sutton and D. L. Hodson, "Climate response to basin-scale warming and cooling of the North Atlantic ocean," *J. Climate*, vol. 20, no. 5, pp. 891–907, 2007.
[61] N. Rayner *et al.*, "Global analyses of sea surface temperature, sea ice, and night marine air temperature since the late nineteenth century," *J. Geophys. Res., Atmospheres*, vol. 108, no. D14, pp. 1–37, 2003.
[62] G. J. McCabe, M. A. Palecki, and J. L. Betancourt, "Pacific and Atlantic ocean influences on multidecadal drought frequency in the United States," *Proc. Nat. Acad. Sci.*, vol. 101, no. 12, pp. 4136–4141, 2004.
[63] R. C. Francis, S. R. Hare, A. B. Hollowed, and W. S. Wooster, "Effects of interdecadal climate variability on the oceanic ecosystems of the NE Pacific," *Fisheries Oceanography*, vol. 7, no. 1, pp. 1–21, 1998.
[64] L. Li, W. Li, and Y. Kushnir, "Variation of the North Atlantic subtropical high western ridge and its implication to southeastern us summer precipitation," *Climate Dyn.*, vol. 39, no. 6, pp. 1401–1412, 2012.
[65] P. Billingsley, *Convergence of Probability Measures*. Hoboken, NJ, USA: Wiley, 2013.
[66] K. B. Athreya and S. G. Pantula, "A note on strong mixing of ARMA processes," *Statist. Probab. Lett.*, vol. 4, no. 4, pp. 187–190, 1986.

Jie Ding (S'12) has broad interests in signal processing, combinatorics, statistical inference, and machine learning.

Yu Xiang (M'10) is a postdoc working at Harvard University and his research interests include information theory, statistical signal processing, and computational biology.

Lu Shen studies atmospheric chemistry and climate dynamics at Harvard University.

Vahid Tarokh (F'09) is a professor in Applied Mathematics at Harvard University.