Random Matrices: Theory and Applications
Vol. 8, No. 4 (2019) 1950014 (35 pages)

DOI: 10.1142/S201032631950014X

Multiple change-points detection in high dimension

Yunlong Wang, Changliang Zou* and Zhaojun Wang
School of Statistics and Data Sciences, and LPMC
Nankai University, P. R. China
**nk.chlzou@gmail.com*

Guosheng Yin
Department of Statistics and Actuarial Science
The University of Hong Kong, Hong Kong

Received 8 August 2018
Accepted 10 December 2018
Published 30 January 2019

Change-point detection is an integral component of statistical modeling and estimation. For high-dimensional data, classical methods based on the Mahalanobis distance are typically inapplicable. We propose a novel testing statistic by combining a modified Euclidean distance and an extreme statistic, and its null distribution is asymptotically normal. The new method naturally strikes a balance between the detection abilities for both dense and sparse changes, which gives itself an edge to potentially outperform existing methods. Furthermore, the number of change-points is determined by a new Schwarz's information criterion together with a pre-screening procedure, and the locations of the change-points can be estimated via the dynamic programming algorithm in conjunction with the intrinsic order structure of the objective function. Under some mild conditions, we show that the new method provides consistent estimation with an almost optimal rate. Simulation studies show that the proposed method has satisfactory performance of identifying multiple change-points in terms of power and estimation accuracy, and two real data examples are used for illustration.

Keywords: Asymptotic normality; dynamic programming; feature screening; high-dimensional homogeneity test; large p, small n; sparse signals.

Mathematics Subject Classification 2010: 62H15, 62H12

1. Introduction

Change-point detection has received enormous attention due to emergence of an increasing amount of temporal data. It is a process of detecting mean, variance, or distributional changes in time-ordered observations, which becomes an integral part of modeling, estimation, and inference [3, 23, 30]. With rapid development in technology, high-dimensional data are often encountered in many fields, such

*Corresponding author.

as genomics, imaging analysis, signal processing, and e-commerce. Due to high dimensionality of such data, conventional methods for change-point detection may not work well.

In the multiple change-points problem with multivariate data, let $\mathbf{X}_i$ denote the p-dimensional observed data from subject i, and let $\mathbf{\Sigma}$ denote a $p \times p$ covariance matrix. We consider the mean-change model,

$$\mathbf{X}_i = \boldsymbol{\mu}_k + \mathbf{\Sigma}^{1/2}\boldsymbol{\varepsilon}_i, \quad \tau_{k-1}^* \le i \le \tau_k^* - 1, \;\; k = 1, \ldots, K+1; \quad i = 1, \ldots, n, \tag{1}$$

where $\boldsymbol{\varepsilon}_i$ is a p-dimensional random vector with mean zero and an identity covariance matrix, K is the true number of change-points, τ_k^*'s are the locations of the change-points with the convention of $\tau_0^* = 1$ and $\tau_{K+1}^* = n+1$, and $\boldsymbol{\mu}_k \neq \boldsymbol{\mu}_{k+1}$ (as long as one component of the $\boldsymbol{\mu}$ vectors is not equal). Moreover, let $\mathcal{J}_k$ be the set of variables for which changes occur at τ_k^*, say $\mathcal{J}_k = \{j : \mu_{jk} \neq \mu_{j,k+1}\}$, and $\mathcal{J} = \bigcup_{k=1}^{K} \mathcal{J}_k$, where μ_{jk} is the jth component of $\boldsymbol{\mu}_k$. Our goal is to test whether there is at least one change-point in the data, with $H_0 : K = 0$ versus $H_1 : K > 0$, and to further estimate the τ_k^*'s if the null hypothesis is rejected.

The standard procedure is to apply binary segmentation and perform multivariate homogeneity tests on two contiguous samples. Classical methods, such as Hotelling's T^2 test [34] or the likelihood ratio test [12], work well when p is relatively small in comparison with n. In modern applications where the number of variables is comparable to or even much larger than the sample size, classical test statistics are typically not well-defined and the asymptotic theories developed for a fixed dimension do not generally hold.

Except for functional data approaches [7], research in high-dimensional cases is rather limited. Bai [4] and Horváth and Hušková [22] considered the single change-point estimation and testing problem with panel data, which can be viewed as high-dimensional data $\mathbf{X}_i$ with a diagonal covariance matrix. The key feature of their proposals is to use the Euclidian norm, $\|\bar{\mathbf{X}}_{\tau-} - \bar{\mathbf{X}}_{\tau+}\|$, instead of the Mahalanobis norm, where $\bar{\mathbf{X}}_{\tau-}$ and $\bar{\mathbf{X}}_{\tau+}$ are the sample means before and after a candidate change-point τ, respectively. By using the max-norm of individual CUSUM statistics rather than the L_2-norm, Jirak [24] proposed to construct simultaneous confidence bands for dependent change-point tests. In general, the max-norm test is more suitable for sparse and strong signals, whereas the L_2-norm test is for dense but weak signals. Using a threshold aggregation approach in combination with the binary segmentation method [19], Cho and Fryzlewicz [11] studied the estimation of multiple global change-points. Despite being more effective than the traditional methods under sparse models, their method heavily relies upon the choice of the threshold value that unfortunately is not easy to determine. Enikeeva and Harchaoui [13] used adaptive Neyman's tests [14] to construct high-dimensional change-point tests, Aston and Kirch [2] proposed projection-based change-point tests, while the difficulty in estimating the optimal projection direction $\mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$ limits the applicability. Recently, high-dimensional change detection in a sequential context has been investigated; for example, see [28, 36, 38] and the references therein.

The contributions of our work are two-fold. First, we develop a novel testing procedure by combining the maximum and summation of p individual CUSUM statistics, which is capable of accommodating large p and sparse signals. Unlike most existing tests which resort to resampling procedures to characterize the null distribution, our test is based on the standard normal critical constants, which works well with typical values of n and p encountered in real applications. Our procedure is shown to be able to strike a balance between the detection abilities of L_2-norm and max-norm based procedures. Second, we form a cost function using the proposed test statistic when searching for multiple change-points. Traditional methods often minimize a penalized version of the cost, which, however, tend to yield an overfitting model in high-dimensional situations. We propose a new Schwarz' information criterion to determine the number of change-points, and the locations of the change-points can be further estimated via the dynamic programming algorithm or the pruned exact linear time (PELT) algorithm [25] in conjunction with the intrinsic order structure of the objective function. Due to the use of the new criterion, technical arguments for establishing estimation consistency are highly nontrivial and are of interest in their own rights. To exclude most of the irrelevant variables in the sparse signal cases, we also suggest an initial screening procedure so that the proposed method can be effectively implemented in a much lower dimensional space, resulting in a significant improvement on the estimation accuracy.

The remainder of this paper is organized as follows. In Sec. 2, we present the new test statistic and study its theoretical properties. In Sec. 3, we develop the estimation procedure for multiple change-points. Section 4 provides extensive simulation studies to compare our method with existing ones and two real data examples for illustration. Section 5 concludes with some remarks, and theoretical proofs are delineated in Appendix A. Some technical details and additional numerical results are provided in the Supplementary Material.

2. Test of Multiple Change-Points

2.1. *Test statistic*

To test the null hypothesis $H_0 : K = 0$, against the alternative in (1) with $K \geq 1$, we start with testing existence of one change-point [5], i.e. $K = 1$. In the asymptotic analysis, we let both p and n diverge to infinity. If τ is the true change-point ($\tau \leq n-1$), it is equivalent to testing whether the two groups, segmented by τ, have the same mean vector. Let X_{ji} denote the jth component of $\mathbf{X}_i$ and let σ_j^2 denote the jth diagonal element of $\mathbf{\Sigma}$. Based on the L_2-norm, an intuitive test statistic can be constructed as

$$L_\tau = \sum_{j=1}^{p} \frac{\tau(n-\tau)}{n\widehat{\sigma}_j^2}(\bar{X}_{j\tau-} - \bar{X}_{j\tau+})^2, \tag{2}$$

where $\bar{X}_{j\tau-} = \tau^{-1}\sum_{i=1}^{\tau} X_{ji}$, $\bar{X}_{j\tau+} = (n-\tau)^{-1}\sum_{i=\tau+1}^{n} X_{ji}$, and $\widehat{\sigma}_j^2$ is a consistent estimator of σ_j^2. Similar to the high-dimensional homogeneity test of two samples in

[18], L_τ naturally aggregates marginal information on individual changes. To test whether there exists a change-point, we run the test over all possible change-points and combine them in the form of

$$S_{n,p} = \sum_{\tau=1}^{n-1} L_\tau,$$

where the covariance matrix $\mathbf{\Sigma}$ is not incorporated, because the contamination bias in estimating $\mathbf{\Sigma}$ grows rapidly with p. In high-dimensional tests where p and n are comparable (e.g. [9, 10]), inclusion of the inverse of the estimated scatter matrix may not help to gain improvement. Similar phenomenon has also been revealed in high-dimensional classification problem, where the independence rule or "naive Bayes" classifier could greatly outperform the Fisher linear discrimination rule when p is much larger than n; see, for example, [29] and the references therein.

Remark 1. Conventionally, the maximum of L_τ, namely $M_{n,p} \stackrel{\text{def}}{=} \max_\tau L_\tau$, is often used as the change-point test statistic rather than $S_{n,p}$; e.g. see [12, 22]. However, it is recognized that the rate of convergence of the maximum statistic is slow; see [12, Sec. 1.3]. Consequently, the critical values from the asymptotic null distribution do not work well, and thus bootstrap resampling is generally required for the existing tests. In contrast, $S_{n,p}$ is asymptotically normal under some mild conditions and its power is at least comparable to that of $M_{n,p}$.

The proposed $S_{n,p}$ is essentially developed under the L_2-norm framework. To enhance power for sparse changes, thresholding and extreme-value tests may be considered [32]. It is widely recognized that those tests require either stringent conditions or bootstrap to derive the null distribution and they often suffer from size distortion due to the slow convergence. This drawback limits their applicability in the change-point detection problem because it is crucial to control false alarm rates effectively. Via connection with high-dimensional homogeneity tests, we suggest an enhanced test statistic by incorporating an extreme statistic,

$$\begin{aligned} T_{n,p} &= S_{n,p} + c_{n,p} I\left(\max_{\substack{1\le j\le p \\ \lceil an\rceil \le \tau \le \lceil bn\rceil}} \frac{\tau(n-\tau)}{n\widehat{\sigma}_j^2} (\bar{X}_{j\tau-} - \bar{X}_{j\tau+})^2 > h_{n,p} \right) \\ &\stackrel{\text{def}}{=} S_{n,p} + E_{n,p}, \end{aligned}$$

where $c_{n,p}$ is a large constant diverging to ∞ as $p \to \infty$, $0 < a < b < 1$ are fixed constants, for example, $a = 0.1$ and $b = 0.9$, and $\lceil x \rceil$ denotes the smallest integer not less than x. The sequence $h_{n,p}$ is chosen to be slightly larger than the maximum noise level, so that the enhancement term $E_{n,p}$ is zero under the null hypothesis with high probability, but diverges quickly under the alternative model with sparse signals. The statistic $T_{n,p}$ is basically in a similar spirit to the power-enhancement test statistic proposed by Fan *et al.* [16]. We use the trimmed maximization to avoid some technical difficulties as $n \to \infty$; see, for example, [12].

To make our test scalar-invariant, we need to estimate σ_j^2. Under the alternative hypothesis with mean changes, the pooled sample variance is not consistent. As a remedy, we suggest an estimator based on the moving ranges of neighboring samples,

$$\widehat{\sigma}_j^2 = \frac{1}{2(n-1)} \sum_{i=2}^{n} (X_{ji} - X_{j,i-1})^2,$$

which, as shown in Proposition 1, is always consistent under mild conditions.

2.2. *Null distribution*

For ease of exposition, the number of change-points, K, is assumed to be independent of n, and the change magnitude $\max_{1\le j\le p} \max_{1\le k\le K} |\mu_{j,\tau_k^*} - \mu_{j,\tau_k^*-1}|$ is bounded. To study the asymptotic behavior of $T_{n,p}$, we impose the following conditions.

(C1) The minimal distance between change-points, $\lambda_n = \min_{0\le k\le K}(\tau_{k+1}^* - \tau_k^*)$, diverges to infinity as $n \to \infty$.
(C2) It holds that $\mathrm{tr}(\mathbf{R}^4) = o(\mathrm{tr}^2(\mathbf{R}^2))$, where $\mathbf{R} \overset{\text{def}}{=} \mathbf{D}^{-1/2}\mathbf{\Sigma}\mathbf{D}^{-1/2}$ is the correlation matrix and $\mathbf{D}$ is the diagonal matrix of $\mathbf{\Sigma}$.
(C3) With some small constant $\nu > 0$, $p/n^{3-\nu} \to 0$.
(C4) The components of $\boldsymbol{\varepsilon}_i$ are independent sub-Gaussian variables, i.e. there exists a constant $\zeta > 0$ such that $\sup_{l\ge 1} l^{-1/2}(E|\varepsilon_{ji}|^l)^{1/l} \le \zeta$.

Remark 2. Condition (C1) allows the change-points to be asymptotically distinguishable, which is a standard requirement for the theoretical development with multiple change-points. Condition (C2) guarantees the asymptotic convergence of $S_{n,p}$, which is commonly used in high-dimensional tests (e.g. [17]). If all the eigenvalues of the correlation matrix $\mathbf{R}$ are bounded, (C2) is true for any p. If $\mathbf{R}$ contains many large entries, (C2) may not hold and neither does the asymptotic normality of $S_{n,p}$. Thus, asymptotic normality relies on the strength of dependency among the variables; certain sparseness on $\mathbf{R}$ is needed. Condition (C3) restricts the relationship between the dimension p and sample size n due to the use of the plug-in variance estimates, which is required for scalar transformation invariance. Condition (C4) is required in some exponential tail inequalities for obtaining the bound of the term $E_{n,p}$. In fact, the independency in (C4) can be replaced by exchangeability as in [6].

Theorem 1. *Suppose that conditions* (C2)–(C4) *hold. Under* H_0,

(i) *The expectation and variance of* $S_{n,p}$ *are respectively given by*

$$E(S_{n,p}) = (n+2)p + o(\sqrt{\mathrm{var}(S_{n,p})}),$$

$$\mathrm{var}(S_{n,p}) = \left[\frac{2\pi^2 - 18}{3} n^2 \,\mathrm{tr}(\mathbf{R}^2) + \frac{(15-\pi^2)n}{3}\{E(\boldsymbol{\varepsilon}^\top \widetilde{\mathbf{R}}\boldsymbol{\varepsilon})^2 - p^2\}\right]\{1 + o(1)\},$$

where $\widetilde{\mathbf{R}} = \mathbf{R}^{\top/2}\mathbf{R}^{1/2}$.

(ii) $\{S_{n,p} - E(S_{n,p})\}/\sqrt{\text{var}(S_{n,p})} \xrightarrow{\mathcal{L}} N(0,1)$.
(iii) *If* $h_{n,p}/\{\log(np)\} \to \infty$, *then*

$$\frac{T_{n,p} - E(S_{n,p})}{\sqrt{\text{var}(S_{n,p})}} \xrightarrow{\mathcal{L}} N(0,1).$$

Theorem 1 establishes the asymptotic normality of $T_{n,p}$ under the null hypothesis. The key step is that $S_{n,p} - E(S_{n,p})$ is asymptotically equivalent to a martingale difference sequence and consequently the assertion can be proved by applying the martingale central limit theorem; e.g. see [20, Corollary 3.1]. Although L_τ in (2) shares a similar form to the high-dimensional two-sample test statistic [10, 18], those results are not directly applicable for $T_{n,p}$, because the treatment on the summation of dependent statistics in L_τ is nontrivial. Theorem 1(iii) reveals that $T_{n,p}$ and $S_{n,p}$ would have the same asymptotic null behavior given an appropriate sequence of $h_{n,p}$.

High-dimensional change detection is challenging, because the estimation error increases fast as the dimension grows. It is extremely difficult, if not impossible, to construct a distribution-free multivariate change-point test in a strict and general sense [27]. Our proposed test statistic serves as an approximately distribution-free one as the central limit theorem is partially valid for the summation of all the components of random vectors. Hence, in this sense, global tests in high dimension may not be considered as a "curse" but a "bless".

Remark 3. Note that $S_{n,p}$ can be approximately viewed as a summation of $(n-1)p$ chi-squared variables with one degree of freedom, and thus its expectation is close to $(n-1)p$. Via an elaborated analysis of high-order expansion of $S_{n,p}$, Theorem 1(i) suggests a more accurate expectation, $(n+2)p$. This high-order approximation allows the dimensionality to grow in almost a cubic rate of the sample size. Also, in the two-sample test of [18], this rate is achieved through bias correction that involves the skewness and kurtosis of X_{ji}. From the proof in Appendix A, $S_{n,p}$ benefits from the additional summation over τ, resulting in a simple asymptotic expectation without other nuisance parameters.

In practical implementation, we need to find ratio consistent estimators for $\text{tr}(\mathbf{R}^2)$ and $E(\boldsymbol{\varepsilon}^\top \widetilde{\mathbf{R}} \boldsymbol{\varepsilon})^2$. To ensure consistency under both the null and alternative hypotheses, we adopt the procedure of moving ranges of neighboring samples. Let

$$\widehat{\mathbf{D}}_{(i_1,\ldots,i_m)} = \text{diag}\{\widehat{\sigma}^2_{1(i_1,\ldots,i_m)}, \ldots, \widehat{\sigma}^2_{p(i_1,\ldots,i_m)}\},$$

where $\widehat{\sigma}^2_{j(i_1,\ldots,i_m)}$ is constructed as $\widehat{\sigma}^2_j$ but excluding $X_{j,i_1}, \ldots, X_{j,i_m}$. We propose the estimates as

$$\widehat{E(\boldsymbol{\varepsilon}^\top \widetilde{\mathbf{R}} \boldsymbol{\varepsilon})^2} = \frac{1}{(n-2)} \sum_{i=2}^{n-1} \{(\mathbf{X}_i - \mathbf{X}_{i-1})^\top \widehat{\mathbf{D}}^{-1}_{(i-1,i,i+1)} (\mathbf{X}_i - \mathbf{X}_{i+1})\}^2 - 3\widehat{\text{tr}(\mathbf{R}^2)},$$

$$\widehat{\text{tr}(\mathbf{R}^2)} = \frac{1}{4(n-3)} \sum_{i=1}^{n-3} \{(\mathbf{X}_i - \mathbf{X}_{i+1})^\top \widehat{\mathbf{D}}^{-1}_{(i,i+1,i+2,i+3)} (\mathbf{X}_{i+2} - \mathbf{X}_{i+3})\}^2,$$

for which the ratio-consistent property holds; see Proposition 1. The "leave-two-out" and "leave-three-out" type estimators are used to remove certain terms that impose unnecessary demand on the dimensionality.

Proposition 1. *Under the change-point model* (1) *and conditions* (C1)–(C4),

$$\frac{\widehat{\mathrm{tr}(\mathbf{R}^2)}}{\mathrm{tr}(\mathbf{R}^2)} \xrightarrow{\mathcal{P}} 1 \quad \textit{and} \quad \frac{\widehat{E(\boldsymbol{\varepsilon}^\top \widetilde{\mathbf{R}} \boldsymbol{\varepsilon})^2}}{E(\boldsymbol{\varepsilon}^\top \widetilde{\mathbf{R}} \boldsymbol{\varepsilon})^2} \xrightarrow{\mathcal{P}} 1.$$

By Slutsky's theorem, we obtain that

$$\frac{T_{n,p} - (n+2)p}{\sqrt{\widehat{\mathrm{var}(S_{n,p})}}} \xrightarrow{\mathcal{L}} N(0,1),$$

as $p, n \to \infty$, where $\widehat{\mathrm{var}(S_{n,p})}$ is the plug-in estimator of $\mathrm{var}(S_{n,p})$. As a result, we reject H_0 at an α level of significance if the normalized $T_{n,p}$ exceeds z_α, where z_α is the upper αth quantile of the standard normal distribution.

2.3. *Consistency of the test*

We investigate the asymptotic behavior of our test under the alternative hypothesis, i.e. the change-point model in (1). Define

$$\Delta_{\boldsymbol{\mu}_S} = \sum_{j=1}^{p} \sum_{\tau=1}^{n-1} \frac{\tau(n-\tau)}{n\sigma_j^2} (\bar{\mu}_{j\tau-} - \bar{\mu}_{j\tau+})^2,$$

$$\Delta_{\boldsymbol{\mu}_E} = \max_{\substack{1 \le j \le p \\ 1 \le \tau \le n-1}} \frac{\tau(n-\tau)}{n\sigma_j^2} (\bar{\mu}_{j\tau-} - \bar{\mu}_{j\tau+})^2,$$

where $\bar{\mu}_{j\tau-} = \tau^{-1} \sum_{i=1}^{\tau} E(X_{ji})$ and $\bar{\mu}_{j\tau+} = (n-\tau)^{-1} \sum_{i=\tau+1}^{n} E(X_{ji})$. Thus, $\Delta_{\boldsymbol{\mu}_S}$ and $\Delta_{\boldsymbol{\mu}_E}$ are essentially the signals reflected through $S_{n,p}$ and $E_{n,p}$, respectively. Let $\lambda_{\max}(\mathbf{R})$ be the largest eigenvalue of $\mathbf{R} = (r_{ij})_{p \times p}$.

Theorem 2. *Suppose that conditions* (C1)–(C4) *hold, and if* $\Delta_{\boldsymbol{\mu}_S} / \max(n\lambda_{\max}(\mathbf{R}), p^\nu) \to \infty$ *with some* $\nu > 1$, *then*

$$\frac{T_{n,p} - (n+2)p}{\sqrt{\mathrm{var}(S_{n,p})}} = O_p\left(\frac{\Delta_{\boldsymbol{\mu}_S}}{\sqrt{\mathrm{var}(S_{n,p})}} \right) + \frac{c_{n,p}}{\sqrt{\mathrm{var}(S_{n,p})}} I(\Delta_{\boldsymbol{\mu}_E} > h_{n,p}) \{1 + o_p(1)\}. \tag{3}$$

This theorem demonstrates the consistency of our test under the alternatives of order larger than $O(np^{1/2})$ in terms of $\Delta_{\boldsymbol{\mu}_S}$, by noting that $\mathrm{var}(S_{n,p}) = O(n^2 p)$ if $\lambda_{\max}(\mathbf{R})$ is bounded. As shown in the Supplementary Material, $\Delta_{\boldsymbol{\mu}_S} = O(n^2 \delta_{n,p})$ under the assumption that $\lambda_n / n \to c < 1$, where $\delta_{n,p} = \min_{1 \le k \le K} \|\mathbf{D}^{-1/2}(\boldsymbol{\mu}_{k+1} - \boldsymbol{\mu}_k)\|^2$. Thus, our proposed test is consistent against the contiguous alternative with $n\delta_{n,p}/p^{1/2} \to \infty$. It also entails the rationale of using $E_{n,p}$ for power enhancement.

Suppose there exists only one change-point τ_1^*, and let $\tau_1^*/n \to q_1 \in (0,1)$. It can be verified that

$$\frac{\Delta_{\boldsymbol{\mu}_S}}{\sqrt{\mathrm{var}(S_{n,p})}} \approx nq_1(1-q_1)p^{-1/2}\sum_{j=1}^{p}\frac{(\mu_{j1}-\mu_{j2})^2}{\sigma_j^2},$$

$$\Delta_{\boldsymbol{\mu}_E} \approx nq_1(1-q_1)\max_{1\le j\le p}\frac{(\mu_{j1}-\mu_{j2})^2}{\sigma_j^2}.$$

On one hand, when the signal under the alternative is dense, say $\|\mathbf{D}^{-1/2}(\boldsymbol{\mu}_1-\boldsymbol{\mu}_2)\|$ grows faster with p but all the $\{|\mu_{j1}-\mu_{j2}|\}_{j=1}^p$ are not large compared with $h_{n,p}$, $T_{n,p}$ is as powerful as $S_{n,p}$. On the other hand, in sparse alternatives where $\delta_{n,p}$ may not grow fast with p but some of $\{|\mu_{j1}-\mu_{j,2}|\}_{j=1}^p$ are particularly large so that $\Delta_{\boldsymbol{\mu}_E} > h_{n,p}$, $T_{n,p}$ would be very powerful because $c_{n,p}/\sqrt{\mathrm{var}(S_{n,p})}$ dominates (2). In contrast, $S_{n,p}$ may not be powerful on its own by Theorem 2. The statistic $T_{n,p}$ gains strength by borrowing information from the max-norm and thus it is able to balance the detection abilities of the L_2-norm and max-norm based procedures to a certain extent.

3. Estimation of Multiple Change-Points

3.1. *Penalized cost function*

The test based on binary segmentation for detecting a single change-point in Sec. 2 can be applied recursively to detect multiple change-points. Due to its simplicity, binary segmentation is computationally efficient and roughly linear with the sample size, while it only provides an approximate solution and may lead to poor estimation of the number and locations of multiple change-points; see [19] and the references therein for variants of binary segmentation. Instead, we define an objective function based on segmentation and minimize its penalized version, which can be viewed as a global minimization procedure [5]. For a candidate set of L change-points, $\tau_1 < \cdots < \tau_L$, we introduce the objective cost function,

$$\mathcal{C}(\tau_1,\ldots,\tau_L) = \sum_{k=0}^{L-1}\sum_{i=\tau_k}^{\tau_{k+1}-1}\|\widehat{\mathbf{D}}^{-1/2}\{\mathbf{X}_i - \bar{\mathbf{X}}(\tau_k,\tau_{k+1})\}\|^2,$$

where $\tau_0 = 1$, $\tau_{L+1} = n+1$, and $\bar{\mathbf{X}}(\tau_k,\tau_{k+1}) = \sum_{i=\tau_k}^{\tau_{k+1}-1}\mathbf{X}_i/(\tau_{k+1}-\tau_k)$. The L change-points τ_l's can then be estimated by

$$(\widehat{\tau}_1,\ldots,\widehat{\tau}_L) = \underset{\tau_1<\cdots<\tau_L}{\arg\min}\,\mathcal{C}(\tau_1,\ldots,\tau_L),$$

which reduces to the procedure in [37] for $p=1$.

To determine L, we observe that $\mathcal{C}(\widehat{\tau}_1,\ldots,\widehat{\tau}_L)$ is a nonincreasing function in L. Hence we can use Schwarz' information criterion (SIC) to strike a balance between

the value of the objective function and the number of change-points by incorporating a penalty for large L [37]. We may consider minimizing

$$\mathcal{C}(\widehat{\tau}_1, \ldots, \widehat{\tau}_L) + L\xi_{n,p} \tag{4}$$

with respect to L, where $\xi_{n,p}$ is chosen to be slightly larger than the maximum noise level (no change-point) so that $\mathcal{C}(\widehat{\tau}_1, \ldots, \widehat{\tau}_L)$ is dominated by $L\xi_{n,p}$ under overfitting models with high probability. However, such an approach is not directly applicable in the high-dimensional setting. To gain more insight, we consider the variation of $\mathcal{C}(\tau_1^*)$ caused by adding a new change-point τ, i.e. $\mathcal{C}(\tau_1^*) - \mathcal{C}(\tau_1^*, \tau)$, when there is only one change-point τ_1^*. In the low-dimensional situation, the total variation reduced due to adding the new point is of the same order as the noise level. By contrast, $\mathcal{C}(\tau_1^*) - \mathcal{C}(\tau_1^*, \tau)$ is of order $p + O_p\{\sqrt{\mathrm{tr}(\mathbf{R}^2)}\}$ in our case. The noise term is, in fact, of the same order as the standard deviation of L_τ and vanishes compared to the expectation, p.

To circumvent the difficulty, we propose using $\tilde{\mathcal{C}}(\widehat{\tau}_1, \ldots, \widehat{\tau}_L) = \mathcal{C}(\widehat{\tau}_1, \ldots, \widehat{\tau}_L) + Lp$ instead of $\mathcal{C}(\widehat{\tau}_1, \ldots, \widehat{\tau}_L)$ in (4), so as to find the minimizer,

$$\widehat{K} = \underset{1 \le L \le \bar{K}}{\arg\min}\, \mathcal{C}(\widehat{\tau}_1, \ldots, \widehat{\tau}_L) + L(\xi_{n,p} + p), \tag{5}$$

where $\bar{K}$ is an upper bound on the true number of change-points. However, $\tilde{\mathcal{C}}(\widehat{\tau}_1, \ldots, \widehat{\tau}_L)$ is no longer a nonincreasing function in L, so that the standard technique to establish consistency in [37] is not applicable. Our theoretical derivations are very different and highly nontrivial, which require two more conditions as follows.

(C5) Assume $\lambda_n \delta_{n,p} / \{\mathrm{tr}(\mathbf{R}^2) \log^2 n\}^{1/2} \to \infty$.
(C6) $\lambda_{\max}(\mathbf{R}) = o\{\sqrt{\mathrm{tr}(\mathbf{R}^2)}\}$ and $p/n^2 \to 0$.

Condition (C5) sets a theoretical requirement for the smallest signal strength and distance between two change-points so that the change-points are asymptotically distinguishable. It is intuitive that if two successive means are very different, then we do not need a very large λ_n to locate the change-point.

Theorem 3. *Suppose that conditions* (C1)–(C6) *hold and* $\bar{K}$ *is bounded. If* $\lambda_n \delta_{n,p} / \xi_{n,p} \to \infty$, *then*

$$\Pr\left(\widehat{K} = K; \max_{1 \le k \le K} \min_{1 \le l \le \widehat{K}} |\tau_k^* - \widehat{\tau}_l| \le \Delta_{n,p} \right) \to 1,$$

provided that $\xi_{n,p} / \{\mathrm{tr}(\mathbf{R}^2) \log^2 n\}^{1/2} \to \infty$ *and* $\Delta_{n,p} \delta_{n,p} / \xi_{n,p} \to \infty$.

This theorem suggests that a larger $\delta_{n,p}$ would yield better detection results because $\Delta_{n,p}$ could be made smaller, as long as $\xi_{n,p}$ is chosen in the range of $(\{\mathrm{tr}(\mathbf{R}^2) \log^2 n\}^{1/2}, \lambda_n \delta_{n,p})$. This result can be shown with a concentration inequality for the quadratic form on the basis of an independent vector-valued sample, see [33]. As the concentration inequality is sharp, the rate of $\Delta_{n,p}$ given in this

theorem is "near-optimal" and cannot be improved beyond the degree of $\log^c n$ for $c > 1$. The optimal detection rate has been studied extensively in the literature; see [21] and the references therein. Niu *et al.* [30] discussed some commonly used technical conditions on signal strength in univariate multiple change-point detection problems; in the high-dimensional case, the signal strength could be defined as $\lambda_{n,p}\delta_{n,p}/\sqrt{\mathrm{tr}(\mathbf{R}^2)}$ which we require to be higher order of $\log(n)$, resulting in the consistency rate of $O_p(\sqrt{\mathrm{tr}(\mathbf{R}^2)\log(n)^2}/\delta_{n,p})$ with a suitable penalty $\xi_{n,p}$. To the best of our knowledge, Theorem 3 is the first theoretical investigation to the estimation accuracy of multiple change-points in high dimension.

The minimization problem in (5) can be solved via dynamic programming [5] in conjunction with a local-detection algorithm [31, 39] or the PELT algorithm [25]. The original dynamic programming requires to calculate all the costs $\mathcal{C}(\tau_1, \tau_1 + 1, \ldots, \tau_2)$ with $1 \le \tau_1 < \tau_2 \le n$ and thus the total complexity is as large as $O(n^2 p)$. The local-detection algorithm uses a local discrepancy statistic with moving window and searches for the most influential points that have the largest jump sizes, and then implements the dynamic programming in a relatively small searching space. Because the complexity of the screening procedure is generally linear in n and thus the overall complexity is about $O\{(n + m^2)p\}$, where m is the number of candidate points in the pre-selected set. The PELT has a pruning step so that many cost functions are not needed to evaluate. Killick *et al.* [25] have shown that the complexity of PELT could be linear in n under mild conditions and hence the minimization (5) would be accomplished with a $O(np)$ computational complexity.

3.2. *Pre-screening*

In many high-dimensional applications, it is often believed that only a subset of the p variables contribute to changes; that is, $|\mathcal{J}|$ is small compared to p. Change detection using all potential variables may cause difficulty in the interpretation and degrade the estimation performance due to the noise accumulation in estimating a large number of parameters. Enormous efforts have been devoted to developing effective high-dimensional variable selection or screening methods under sparsity [15]. To enhance the detection ability in such sparse cases, we introduce a preliminary screening procedure to reduce the original large-scale problem to a moderate scale and, as a consequence, the method proposed in Sec. 3.1 can be applied in a much lower dimensional space.

Define a screening index

$$\omega_j = \max_{1 \le \tau \le n-1} \frac{\tau(n-\tau)}{n\widehat{\sigma}_j^2}(\bar{X}_{j\tau-} - \bar{X}_{j\tau+})^2,$$

as a marginal utility to gauge the importance of the jth variable. When ω_j is large, the jth variable is more likely to have at least one change, and thus we select a set of variables with large ω_j,

$$\widehat{\mathcal{J}} = \{j : \omega_j \ge \{\log(np)\}^\kappa, \text{ for } 1 \le j \le p\},$$

where $\kappa > 1$ is a prespecified value. The consistency of this screening procedure can be formally established as follows.

Proposition 2. *Suppose that conditions* (C1) *and* (C4) *hold. If the change sizes satisfy*

$$\lambda_n^2 \min_{j\in\mathcal{J}} \min_{1\le k\le K} (\mu_{j,\tau_k-1} - \mu_{j\tau_k})^2/[n\sigma_j^2\{\log(np)\}^\kappa] \to \infty,$$

then $\Pr(\widehat{\mathcal{J}} = \mathcal{J}) \to 1$, *as* $n, p \to \infty$.

With probability tending to one, the screening algorithm can identify all the variables which exhibit changes. Finally, we find the solution of (5) by only considering the variables that belong to $\widehat{\mathcal{J}}$. Denote $\mathbf{A}_{(\mathcal{J})}$ as the square submatrix of a symmetric matrix $\mathbf{A}$ with the subset $\mathcal{J}$. Apparently, the result in Theorem 3 also holds for the post-screening change detection procedure, by replacing $\mathbf{R}$ with $\mathbf{R}_{(\mathcal{J})}$. The advantage of the pre-screening procedure is evident. Suppose that $\lambda_{\max}(\mathbf{R})$ is bounded, and note that $\mathrm{tr}(\mathbf{R}^2) = O(p)$, while $\mathrm{tr}(\mathbf{R}^2_{(\mathcal{J})}) = O(|\mathcal{J}|)$. Hence, the estimation accuracy (in terms of $\delta_{n,p}$) of the procedure after screening is improved to $(|\mathcal{J}|\log^2 n)^{1/2}/\delta_{n,p}$ compared to that without pre-screening, $(p\log^2 n)^{1/2}/\delta_{n,p}$.

3.3. *Selection of tuning parameters*

We provide some guidelines on the choices of the tuning parameters in our method. For the proposed test, we set $a = 0.1$ and $b = 0.9$ in $E_{n,p}$, and $c_{n,p} = 100\sqrt{\widehat{\mathrm{var}}(S_{n,p})}$, as the test performance is not affected by $c_{n,p}$, as long as it is large enough. In fact, we find that there is no difference in the performances of those $c_{n,p}$'s bigger than 10. However, the tuning parameter $h_{n,p}$ has effect on the balance between the false alarm rate and detection power. In general, small $h_{n,p}$ tends to increase the power while the size might be inflated; a large $h_{n,p}$ is able to control the size better but may results in some power loss, especially when the signal is dense but strong. We recommend $h_{n,p} = \{2\log(np)\}^{1.1}$ based on our theoretical results, while the performance of our test varies mildly around this $h_{n,p}$.

Choices of the threshold in the screening procedure would affect estimation accuracy. Theoretically, the upper bound for the maximum noise level is of $O_p\{\log(np)\}$ and we recommend to choose the threshold as $\{\log(np)\}^{1.01}$. In real applications, the signal strength may not satisfy the conditions required in Proposition 2 and then there will not exist one threshold that perfectly distinguishes the signals and noises. In such situation, a pre-specified $|\mathcal{J}|$, determined by some prior or expert knowledge, would be helpful.

The choice of $\xi_{n,p}$ is indeed subtle and depends on the application and so it is difficult to find a uniformly best $\xi_{n,p}$ for all situations encountered. A smaller choice potentially increases the estimated number of change-points with certain over-fitting risks, and vice versa. If it is important to find all change-points at some expense of potential false-positives, a smaller choice is recommended. As seen from Theorem 3, the lower bound for the penalty is of the form $p + \sqrt{\mathrm{tr}(\mathbf{R}^2)(\log n)^c}$ with

some constant $c > 2$. We fix $c = 2.2$ and turn to find a suitable c_0 in the SIC penalty $p + c_0\sqrt{\widehat{\mathrm{tr}(\mathbf{R}^2)}}(\log n)^{2.2}$. The results in Sec. 4.1 suggest that $c_0 = 2.5$ would be a reasonable choice.

4. Numerical Studies

To evaluate the performance of the proposed method, we conduct simulations by generating the error term $\mathbf{\Sigma}^{1/2}\boldsymbol{\varepsilon}_i \equiv (\epsilon_{1i}, \ldots, \epsilon_{pi})^\top$ in (1) from an ARMA model,

$$\epsilon_{ji} - \beta_1\epsilon_{j-1,i} = e_{ji} + \beta_2 e_{j-1,i},$$

where e_{ji} is the white noise with zero mean. We consider four correlation structures: $(\beta_1, \beta_2) = (0, 0)$, $(3/4, 0)$, $(0, 3/4)$ and $(1/2, 1/2)$, which correspond to the independent (IND), autoregressive (AR), moving-average (MA) and ARMA, respectively. Also, we study three distributions for the innovation e_{ji}: $N(0, 1)$, Student's t distribution with five degrees of freedom (t_5), and chi-squared distribution with three degrees of freedom (χ^2_3). All the simulation results are based on 1000 replications and the comparisons are made at the 0.05 significance level.

4.1. *Test size*

First, we compare the test size between our test (abbreviated as NEW) and those of Horváth and Hušková [22], Enikeeva and Harchaoui [13] and Jirak [24] (abbreviated as HH12, EH13 and J15, respectively). To obtain the critical constants for their tests, Horváth and Hušková [22] and Jirak [24] suggested using simulations under the standard multivariate normal distribution, and Enikeeva and Harchaoui [13] proposed a chi-squared approximation.

Table 1 presents the empirical sizes at a 5% nominal significance level under the correlation structures of IND and AR when $n = 200$ and 400, and those for MA and ARMA are reported in the Supplementary Material. The empirical sizes of our test approach the nominal level as n increases for all scenarios. In contrast, the test of Enikeeva and Harchaoui [13] encounters serious size distortion in many cases, which demonstrates that asymptotic-based cutoff values in the multi-level thresholding test may not always attain a desired false alarm rate. Although the simulation-based method works reasonably well for the test of Jirak [24], it is rather computationally intensive, especially when n and p are large. To gain more insight on the asymptotic normality of $T_{n,p}$, Fig. 1 displays the normal Q–Q plots under various settings, which indicate that our standardized test statistic generally follows the standard normal distribution.

4.2. *Power comparison*

For the power comparison of the four tests (NEW, EH13, J15 and HH12), we consider $n = 200$ and set the location of the change-point at $\tau = n/4$ or $n/2$. Without loss of generality, we fix $\boldsymbol{\mu}_1 = 0$ and examine the case of $\boldsymbol{\mu}_2$ with $p\rho$

Table 1. Comparison of empirical sizes (%) at a 5% significance level for the change-point test under scenarios of $(\beta_1, \beta_2) = (0, 0)$ or $(0.75, 0)$ and $n = 200$ or 400.

		$N(0,1)$				t_5				χ^2_3			
Correlation	p	NEW	EH13	J15	HH12	NEW	EH13	J15	HH12	NEW	EH13	J15	HH12
						$n = 200$							
IND	20	7.0	13.8	6.4	6.4	4.0	9.2	2.2	4.4	4.4	10.2	4.8	4.6
	50	6.8	10.2	4.6	6.0	4.8	5.2	2.6	3.4	5.6	6.8	3.8	4.8
	100	5.6	9.2	5.2	4.6	3.0	5.2	4.6	3.0	5.2	6.6	2.4	7.0
	500	4.4	8.0	5.6	4.0	2.6	5.6	3.6	4.6	3.6	4.8	4.2	4.4
	1000	2.2	34.2	5.0	5.2	2.0	22.2	4.2	4.4	2.2	20.6	5.4	3.6
AR	20	5.0	71.6	3.6	34.8	7.4	51.4	5.0	35.4	7.0	51.0	4.0	34.6
	50	7.4	73.0	3.6	40.2	6.2	54.0	3.4	38.4	6.0	52.6	3.4	40.2
	100	5.4	71.2	4.6	39.0	6.6	55.0	5.0	47.4	6.6	52.8	4.4	42.6
	500	7.8	72.2	3.8	46.0	6.0	55.2	4.8	45.0	3.8	51.6	3.2	43.2
	1000	6.0	90.6	5.6	47.2	4.6	70.8	3.6	43.8	4.2	77.2	4.2	48.0
						$n = 400$							
IND	20	6.8	20.0	6.0	5.2	4.6	10.0	3.6	4.0	5.0	9.0	4.8	3.2
	50	6.4	15.0	6.0	5.2	7.0	8.8	5.4	8.2	5.8	7.6	6.4	6.0
	100	4.8	9.8	5.4	5.4	5.4	5.4	6.4	4.4	4.6	6.4	6.0	5.8
	500	5.2	8.0	4.8	6.0	3.8	4.4	5.6	5.4	3.8	4.4	3.2	6.2
	1000	4.6	38.4	4.2	5.6	3.4	20.8	5.6	4.4	4.6	23.2	4.2	7.0
AR	20	7.0	77.0	3.6	36.2	6.4	54.0	2.4	37.0	6.0	50.4	4.0	35.2
	50	7.2	82.0	3.6	45.8	6.6	55.0	4.0	42.8	7.4	56.2	4.0	44.2
	100	6.4	77.6	4.6	44.2	6.0	54.4	4.6	43.4	5.0	56.6	4.8	45.0
	500	6.4	79.2	4.2	48.0	4.4	74.8	4.6	43.0	4.6	52.2	6.0	44.8
	1000	5.4	93.4	2.8	46.8	6.0	53.6	4.6	48.4	6.4	76.0	4.8	44.0

Note: NEW, EH13, J15 and HH12 represent our proposed test and those of Enikeeva and Harchaoui [13], Jirak [24] and Horváth and Hušková [22], respectively.

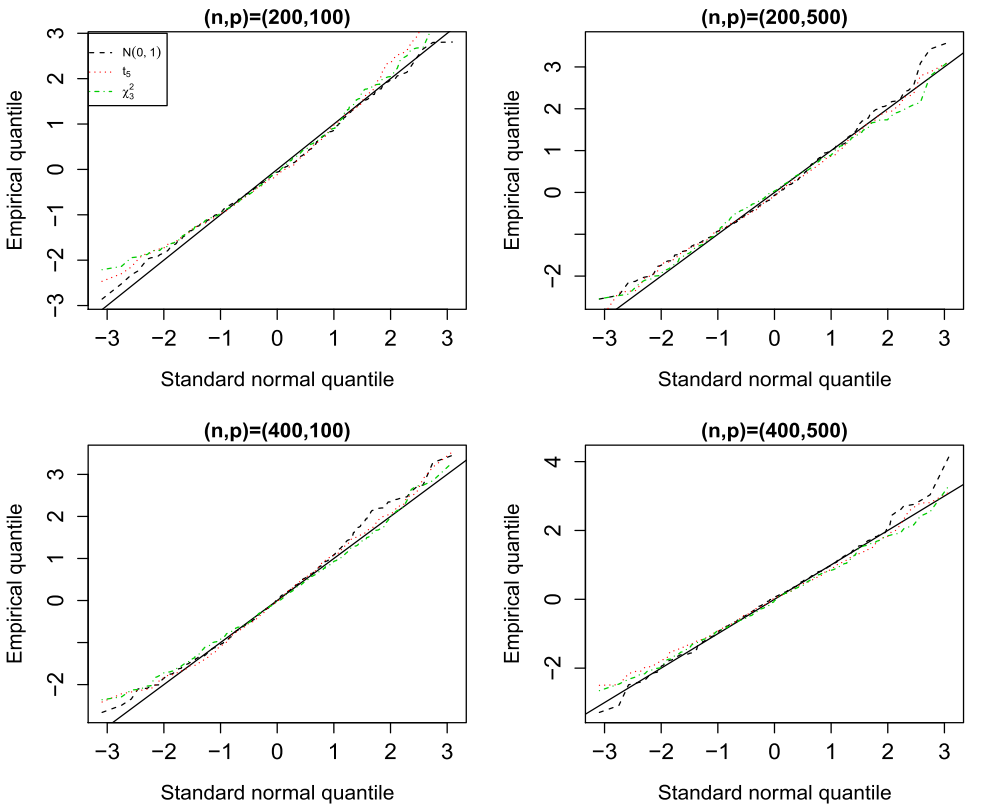


Fig. 1. Normal Q–Q plots of our test statistics under the independent structure.

components being δ and the others being zero. The locations of non-zero components are uniformly generated from $\{1, \ldots, p\}$. Figure 2 depicts the power curves of the four tests against δ, under different values of ρ, for $\tau = n/4$ and the normal error. We focus on the IND case, because EH13 and HH12 cannot control the size well under other correlation structures.

For the sparse cases with $\rho = 0.004$ and 0.01, our proposed test has comparable performance with J15, while HH12 exhibits the lowest power. In contrast, for the dense cases, such as $\rho = 0.05$ or $\rho = 0.1$, both EH13 and HH12 perform well (EH13 cannot preserve the test size), while J15 behaves poorly. The proposed test is inferior to HH12 but still possesses reasonably good detection ability and its power curve increases much more sharply than J15 as δ increases. Such findings are consistent with our theoretical analysis that the proposed test is capable of balancing the detection abilities between the sum- and max-type tests due to the use of power enhancement term $E_{n,p}$. The effect of the level of sparsity (characterized by ρ) is further investigated in Fig. 3, where the empirical power is plotted against ρ. To make power comparable among different configurations of the alternative, we fix

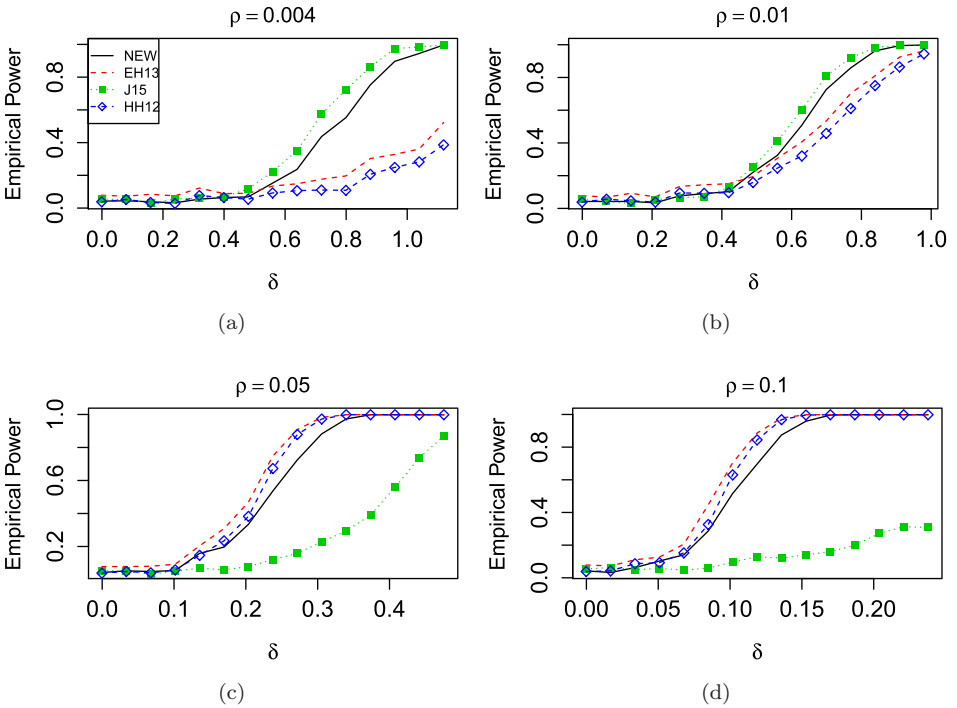


Fig. 2. Comparison of empirical power curves between the proposed test and those by Horváth and Hušková [22], Enikeeva and Harchaoui [13] and Jirak [24] with an increasing δ, under the independent structure, normal errors, and $(n, p) = (200, 500)$.

$\phi = n\rho\delta^2$ for each value of ρ. It is observed that the proposed test offers balanced protection of power over a range of values of ρ, where HH12 tends to be more powerful for larger values of ρ and J15 performs better with smaller ρ. Our method inherits advantages from both approaches and is always close to the best. It appears

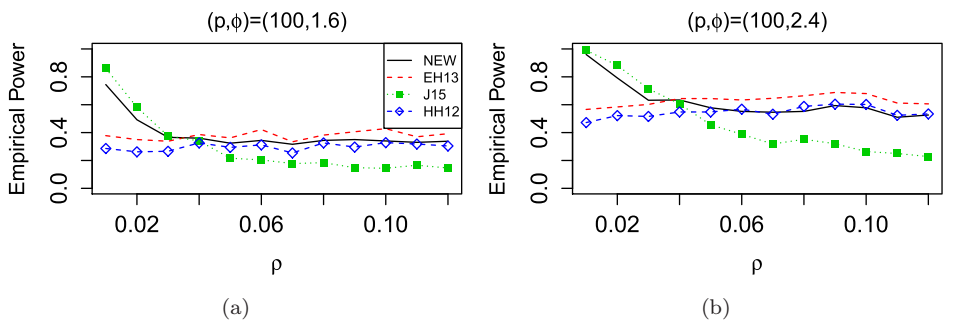
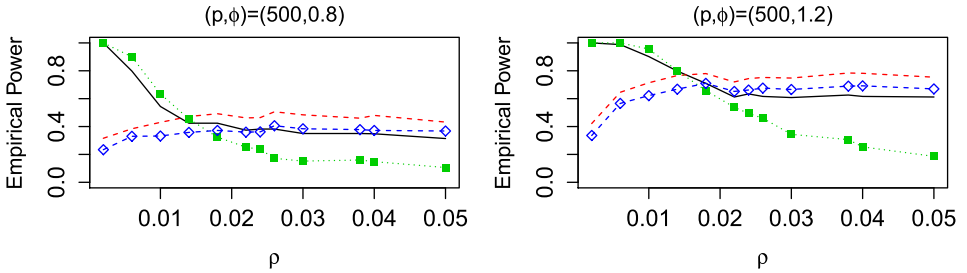


Fig. 3. Comparison of empirical power between the proposed test and those by Horváth and Hušková [22], Enikeeva and Harchaoui [13] and Jirak [24] with an increasing ρ, under the independent structure and normal errors.



(c) (d)

Fig. 3. (*Continued*)

that the EH13 test has comparable performance with ours, but its type I error rates are considerably inflated over the nominal level. Similar results under other settings can be found in the Supplementary Material.

4.3. *Estimation of change-points*

4.3.1. *Estimation of a single change-point*

Once the null hypothesis is rejected, we apply the proposed method to estimate the change-point. Table 2 summarizes the bias and standard deviation of the change-point estimate under the alternative models considered in Fig. 2. For comparison, the estimators of Jirak [24] and Horváth and Hušková [22] are included as well. In [24], the change-points are identified in each dimension independently, and thus the "average" bias is presented, say $\sum_{k=1}^{\widehat{K}} |\widehat{\tau}_k - \tau_k^*|/\widehat{K}$, where $\widehat{K}$ is the estimated number of change-points. Overall, the proposed estimator appears to be consistent under various configurations. Due to the use of the screening procedure, our method provides more robust estimation with different values of ρ.

Table 2. Simulation study on the consistency of our change-point detection procedure with the mean and the standard deviation (in parentheses) of $|\widehat{\tau}_1 - \tau_1^*|$'s under $n = 200$, $\tau_1 = n/4$, the independent structure and a normal error distribution.

		$p = 100$			$p = 500$		
ϕ	ρ	NEW	HH12	J15	NEW	HH12	J15
8	0.01	$0.53_{(0.87)}$	$7.26_{(10.9)}$	$2.99_{(5.98)}$	$0.04_{(0.18)}$	$1.31_{(2.56)}$	$1.84_{(2.08)}$
	0.05	$0.60_{(1.00)}$	$4.11_{(7.46)}$	$8.40_{(6.01)}$	$0.04_{(0.21)}$	$0.66_{(1.38)}$	$7.84_{(2.65)}$
	0.10	$1.09_{(1.75)}$	$3.25_{(5.40)}$	$12.10_{(8.31)}$	$0.12_{(0.48)}$	$0.59_{(1.43)}$	$12.20_{(4.99)}$
	0.50	$8.12_{(13.20)}$	$2.91_{(4.65)}$	$32.30_{(21.10)}$	$2.29_{(3.43)}$	$0.60_{(1.44)}$	$28.90_{(18.70)}$
16	0.01	$0.20_{(0.51)}$	$3.01_{(6.10)}$	$1.98_{(5.44)}$	$0.00_{(0.00)}$	$0.46_{(1.24)}$	$1.01_{(1.79)}$
	0.05	$0.26_{(0.59)}$	$1.74_{(2.94)}$	$4.53_{(3.15)}$	$0.00_{(0.00)}$	$0.23_{(0.61)}$	$4.31_{(1.39)}$
	0.10	$0.26_{(0.56)}$	$1.17_{(2.28)}$	$7.96_{(3.57)}$	$0.00_{(0.00)}$	$0.16_{(0.59)}$	$7.44_{(1.63)}$
	0.50	$1.32_{(1.98)}$	$1.02_{(2.00)}$	$22.10_{(14.8)}$	$0.22_{(0.52)}$	$0.14_{(0.52)}$	$21.80_{(10.4)}$

4.3.2. *Estimation of multiple change-points*

To assess our approach for detecting multiple change-points, we consider two different data generation processes. The first one is given by

$$\text{Model I}: \quad \{\mu_1, \ldots, \mu_5\} = \{0, -0.9, -0.3, 0.3, 0.9\},$$
$$\{\tau_1, \ldots, \tau_4\} = \{0.2n, 0.4n, 0.6n, 0.8n\}.$$

At each change-point τ_l, we randomly assign $p\rho$ components to be changed from μ_l to μ_{l+1}. The second data generation allows the signal strengths in different dimensions to be different,

$$\text{Model II}: \quad \{\mu_{j1}, \mu_{j2}, \mu_{j3}, \mu_{j4}\} = \{0, 2, 4, 6\}(C_j/[\rho p]),$$
$$\{\tau_1, \tau_2, \tau_3\} = \{0.2n, 0.5n, 0.8n\},$$

where $j \in \mathcal{J}$ and $\{C_1, \ldots, C_{[\rho p]}\}$ is a random permutation of $\{1, \ldots, [\rho p]\}$. We take $p = 500$, $n = 200$ and $\rho = 0.05$ or 0.01.

To evaluate the finite-sample performance, we consider the Hausdorff distance between the estimated change-point set and the true one (e.g. [8, 39]),

$$\text{OE} = \sup_{r=1,\ldots,K} \inf_{l=1,\ldots,\widehat{K}} |\widehat{\tau}_l - \tau_r^*| \quad \text{and} \quad \text{UE} = \sup_{l=1,\ldots,\widehat{K}} \inf_{r=1,\ldots,K} |\widehat{\tau}_l - \tau_r^*|,$$

which characterize the over- and under-segmentation errors, respectively. A desirable estimator should be able to balance both quantities. In addition, another measure on the number of change-points, $|\widehat{K} - K|$, is also considered. We present the mean and standard deviation of these three indices, $|\widehat{K} - K|$, OE and UE, based on 1000 replications.

The curves of the means of three quantities, $|\widehat{K} - K|$, OE and UE, versus the value of c_0 under Model I are presented in Fig. 4. To strike a balance among the three error measurements, we suggest that c_0 take a value around 2.5.

Tables 3 and 4 report the means and standard deviations of $|\widehat{K} - K|$, OE and UE under various configurations when $\rho = 0.05$ and 0.01, respectively. We take $(n, p) = (200, 500)$. In the more sparse case with $\rho = 0.01$, we amplify the signal by $\sqrt{5}$ so that the total signal strength, quantified by $\delta_{n,p}$, is similar for Tables 3 and 4. For comparison, we also include the ECP method of Matteson and James [27] and the thresholding method of Cho and Fryzlewicz [11], which are abbreviated as MJ14 and CF15, respectively. The MJ14 method is implemented using the "ecp" R package with the false alarm rate 0.05 and $\alpha = 1$. It can be observed that all the distance values using our method are reasonably small, and the performance is stable. The proposed method outperforms the two competitors, and the advantage becomes more prominent for $\rho = 0.01$. This demonstrates that the proposed global estimator, combined with the screening procedure, can deliver satisfactory detection performance in the presence of multiple change-points. Although it is not developed under the high-dimensional framework, the MJ14 method also provides reasonable estimation results in most cases.

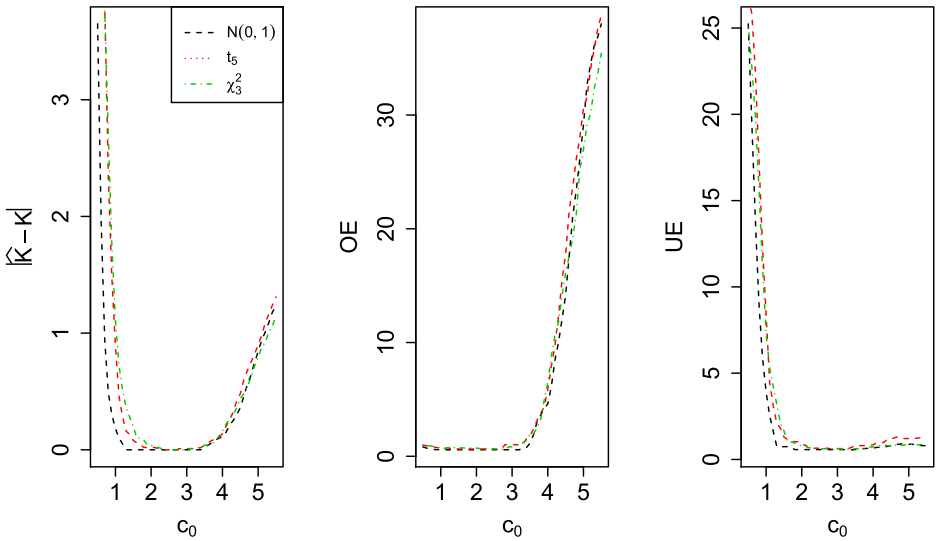


Fig. 4. Performance of the multiple change-points estimator under the first data generation process with $p = 500$, $n = 200$, $\rho = 0.05$, as the tuning parameter c_0 varies.

4.4. *Beyond the mean change*

As suggested by a referee, our method could be adapted to detect variance or distribution changes by using suitable transformation of the original data. For detecting the changes in variance, we transform the $\mathbf{X}_i$ by $\widetilde{\mathbf{X}}_i = (\mathbf{X}_i - \boldsymbol{\mu})^2$, where $\boldsymbol{\mu}$ is the known mean vector of $\mathbf{X}_i$ and can be replaced by some suitable estimator if it is unknown; for distribution-change model, denote $\widetilde{\mathbf{X}}_{ij} = (I(X_{ji} \in P_1), \ldots, I(X_{ji} \in P_k))^\top$, where $P_l, \ldots, P_k$s are some suitable non-overlapped partitions of $\mathbf{X}_i$'s domain, and construct $\widetilde{\mathbf{X}}_i = (\widetilde{\mathbf{X}}_{i1}^\top, \widetilde{\mathbf{X}}_{i2}^\top, \ldots, \widetilde{\mathbf{X}}_{ip}^\top)^\top$. Then the variance- or distribution-change model for $\mathbf{X}_i$ can be transformed to the mean-change model with $\widetilde{\mathbf{X}}_i$, and thus our method is still applicable.

We use the following simulation to show the performance of this method. The sample size is fixed as $n = 200, p = 100$. The variance change-point model considered is $\mathbf{X}_i = \sigma_1 \boldsymbol{\varepsilon}_i I(i \leq n/2 - 1) + \sigma_2 \boldsymbol{\varepsilon}_i I(i \geq n/2)$, with $\sigma_1^2 = 1$. For the distribution change-point model, $\mathbf{X}_i$'s are independently generated from the multivariate normal distribution $N_p(\mathbf{0}_p, \mathbf{I}_p)$ if $i < \tau_1^*$ and from multivariate t-distribution $T_p(\mathbf{0}_p, \mathbf{I}_p)$ with ν degrees of freedom if $i > \tau_1^*$. As a simple illustration, the partitions are chosen as $P_1 = (-\infty, -1.5], P_2 = (-1.5, 1.5], P_3 = (1.5, \infty)$.

The results are shown in Table 5. The proposed test satisfactorily controls the size and also has high power when the signal strength increases. In terms of estimation accuracy, our method performs well in the variance-change model but tends to under- or over-estimate when the signal is weak or strong under the distribution-change model. This is because the covariance matrix also changes at the change-point, and thus treating the covariance matrix as a constant may lead to loss of

Table 3. Performance evaluation on detection of multiple change-points with the standard deviations given in parentheses for $\rho = 0.05$.

			$\lvert\widehat{K} - K\rvert$			OE			UE		
Correlation	Model	Error	NEW	MJ14	CF15	NEW	MJ14	CF15	NEW	MJ14	CF15
IND	I	$N(0,1)$	$0.00_{(0.00)}$	$0.11_{(0.31)}$	$0.68_{(0.75)}$	$0.57_{(0.78)}$	$5.82_{(10.37)}$	$24.87_{(23.94)}$	$0.57_{(0.78)}$	$3.17_{(4.30)}$	$8.77_{(6.93)}$
		t_5	$0.05_{(0.23)}$	$0.01_{(0.30)}$	$0.99_{(0.81)}$	$0.63_{(0.80)}$	$4.78_{(9.27)}$	$24.15_{(27.00)}$	$1.41_{(4.94)}$	$3.03_{(4.00)}$	$12.53_{(8.01)}$
		χ^2_3	$0.04_{(0.18)}$	$0.10_{(0.32)}$	$1.16_{(1.10)}$	$0.73_{(1.04)}$	$4.77_{(9.14)}$	$20.34_{(23.99)}$	$1.20_{(3.80)}$	$2.99_{(3.87)}$	$13.47_{(8.53)}$
	II	$N(0,1)$	$0.01_{(0.07)}$	$0.07_{(0.25)}$	$0.42_{(0.60)}$	$0.01_{(0.07)}$	$0.02_{(0.12)}$	$2.17_{(1.87)}$	$0.19_{(2.55)}$	$1.43_{(5.59)}$	$8.68_{(9.58)}$
		t_5	$0.03_{(0.17)}$	$0.05_{(0.23)}$	$1.03_{(0.97)}$	$0.05_{(0.22)}$	$0.01_{(0.10)}$	$2.01_{(1.77)}$	$0.54_{(3.93)}$	$0.81_{(4.03)}$	$14.31_{(10.29)}$
		χ^2_3	$0.04_{(0.18)}$	$0.05_{(0.23)}$	$1.11_{(1.14)}$	$0.02_{(0.14)}$	$0.01_{(0.07)}$	$2.19_{(1.79)}$	$0.80_{(5.30)}$	$0.80_{(3.92)}$	$13.41_{(10.04)}$
ARMA	I	$N(0,1)$	$0.28_{(0.46)}$	$2.34_{(0.77)}$	$1.70_{(0.98)}$	$11.54_{(15.50)}$	$63.74_{(25.40)}$	$54.85_{(27.55)}$	$3.92_{(4.82)}$	$5.09_{(6.06)}$	$10.22_{(7.83)}$
		t_5	$0.33_{(0.47)}$	$2.38_{(0.79)}$	$1.78_{(0.90)}$	$12.17_{(15.31)}$	$66.35_{(26.60)}$	$58.21_{(26.01)}$	$4.04_{(5.51)}$	$5.52_{(5.85)}$	$11.16_{(8.36)}$
		χ^2_3	$0.22_{(0.42)}$	$2.25_{(0.77)}$	$1.65_{(0.91)}$	$9.48_{(13.12)}$	$60.36_{(24.03)}$	$55.86_{(25.54)}$	$3.88_{(4.01)}$	$6.08_{(5.72)}$	$11.75_{(8.08)}$
	II	$N(0,1)$	$0.02_{(0.14)}$	$0.03_{(0.17)}$	$0.27_{(0.47)}$	$0.31_{(0.59)}$	$1.60_{(4.60)}$	$5.07_{(8.87)}$	$0.96_{(4.68)}$	$1.85_{(4.10)}$	$7.94_{(8.29)}$
		t_5	$0.02_{(0.14)}$	$0.05_{(0.22)}$	$0.49_{(0.67)}$	$0.22_{(0.45)}$	$1.41_{(2.35)}$	$3.91_{(5.32)}$	$0.79_{(4.56)}$	$2.34_{(4.71)}$	$9.92_{(9.57)}$
		χ^2_3	$0.02_{(0.12)}$	$0.04_{(0.24)}$	$0.44_{(0.68)}$	$0.16_{(0.41)}$	$1.32_{(1.78)}$	$3.92_{(5.22)}$	$0.42_{(2.47)}$	$1.84_{(3.52)}$	$9.49_{(9.19)}$

Table 4. Performance evaluation on detection of multiple change-points with the standard deviations given in parentheses for $\rho = 0.01$.

			$\lvert\widehat{K} - K\rvert$			OE			UE		
Correlation	Model	Error	NEW	MJ14	CF15	NEW	MJ14	CF15	NEW	MJ14	CF15
IND	I	$N(0,1)$	$0.10_{(0.31)}$	$0.12_{(0.34)}$	$0.67_{(0.72)}$	$0.55_{(0.73)}$	$5.71_{(10.79)}$	$22.88_{(20.84)}$	$3.04_{(8.30)}$	$2.67_{(3.62)}$	$12.10_{(6.27)}$
		t_5	$0.12_{(0.37)}$	$0.11_{(0.33)}$	$1.07_{(0.94)}$	$0.54_{(0.84)}$	$4.75_{(9.68)}$	$24.81_{(24.29)}$	$2.24_{(7.05)}$	$2.77_{(3.92)}$	$14.69_{(7.29)}$
		χ^2_3	$0.13_{(0.39)}$	$0.12_{(0.35)}$	$1.05_{(0.98)}$	$0.54_{(0.74)}$	$3.76_{(6.96)}$	$26.15_{(25.29)}$	$2.84_{(8.29)}$	$3.61_{(4.70)}$	$15.20_{(7.22)}$
	II	$N(0,1)$	$0.10_{(0.31)}$	$0.06_{(0.24)}$	$0.49_{(0.66)}$	$0.00_{(0.00)}$	$0.01_{(0.07)}$	$0.65_{(8.22)}$	$2.63_{(8.68)}$	$1.20_{(4.91)}$	$14.83_{(8.61)}$
		t_5	$0.15_{(0.47)}$	$0.06_{(0.26)}$	$1.25_{(1.05)}$	$0.00_{(0.00)}$	$0.01_{(0.07)}$	$10.16_{(8.19)}$	$2.46_{(8.27)}$	$1.16_{(4.99)}$	$18.72_{(8.54)}$
		χ^2_3	$0.21_{(0.51)}$	$0.08_{(0.28)}$	$1.30_{(1.06)}$	$0.02_{(0.14)}$	$0.01_{(0.07)}$	$10.66_{(9.70)}$	$3.61_{(9.42)}$	$1.40_{(5.42)}$	$19.22_{(8.09)}$
ARMA	I	$N(0,1)$	$0.18_{(0.38)}$	$2.29_{(0.79)}$	$1.43_{(0.95)}$	$2.28_{(3.38)}$	$61.68_{(26.15)}$	$49.72_{(28.02)}$	$6.45_{(10.52)}$	$5.61_{(6.04)}$	$10.62_{(5.78)}$
		t_5	$0.09_{(0.32)}$	$2.37_{(0.76)}$	$1.34_{(0.93)}$	$1.89_{(1.97)}$	$64.88_{(27.39)}$	$48.89_{(29.43)}$	$2.82_{(5.25)}$	$5.74_{(5.89)}$	$12.18_{(6.53)}$
		χ^2_3	$0.16_{(0.41)}$	$2.27_{(0.79)}$	$1.21_{(0.94)}$	$2.31_{(2.37)}$	$62.86_{(27.18)}$	$46.64_{(29.52)}$	$4.83_{(8.43)}$	$5.55_{(6.12)}$	$13.11_{(6.80)}$
	II	$N(0,1)$	$0.18_{(0.45)}$	$0.06_{(0.25)}$	$0.47_{(0.62)}$	$0.07_{(0.27)}$	$0.77_{(1.54)}$	$15.59_{(15.05)}$	$4.62_{(11.14)}$	$1.60_{(4.09)}$	$15.10_{(8.05)}$
		t_5	$0.11_{(0.35)}$	$0.05_{(0.23)}$	$0.58_{(0.68)}$	$0.14_{(0.36)}$	$0.77_{(1.83)}$	$15.63_{(13.83)}$	$2.17_{(7.31)}$	$1.66_{(4.89)}$	$16.77_{(8.07)}$
		χ^2_3	$0.10_{(0.33)}$	$0.06_{(0.23)}$	$0.60_{(0.66)}$	$0.21_{(0.45)}$	$0.69_{(1.13)}$	$18.19_{(16.80)}$	$2.06_{(6.90)}$	$1.83_{(5.09)}$	$16.72_{(7.77)}$

Table 5. Performance evaluation on detection of variance- or distribution-change with the standard deviations given in parentheses.

Model	Size	Signal	Power	$\lvert\widehat{K}-K\rvert$	OE	UE
Variance	0.026	$\sigma_2^2 = 1.25$	0.735	$0.475_{(0.558)}$	$37.32_{(30.10)}$	$44.84_{(32.95)}$
		$\sigma_2^2 = 1.5$	1.000	$0.110_{(0.386)}$	$3.185_{(6.101)}$	$12.86_{(25.92)}$
Distribution	0.042	$\nu = 5$	0.362	$0.430_{(0.605)}$	$37.06_{(28.85)}$	$45.41_{(31.72)}$
		$\nu = 3$	0.878	$1.035_{(1.209)}$	$19.06_{(21.18)}$	$47.87_{(36.72)}$

information. Change-point detection with varying covariance matrix certainly warrants further research.

4.5. *Real-data applications*

4.5.1. *Changes in the stock returns*

We consider an illustrative example of 455 primary stocks which are the components of the S&P 500 index. The dataset recorded $n = 1295$ daily close price of each stock, from 1 January 2007 to 1 January 2012, which can be downloaded from https://beta.finance.yahoo.com/. It is known that a debt crisis or a policy change may affect many companies' stock returns. The goal of our analysis is to find the time points of critical events that would trigger change-points in the stock returns, by simultaneously examining all the constituent stocks. We consider the absolute log returns which can reflect the volatilities of the stocks.

As an illustration, Fig. 5 depicts the log returns of three stocks: the Exxon Mobil, Microsoft Corporation and Wells Fargo. We can observe that although the three plots show similar patterns at certain time points, the change-points in the three data series may not be the same in general. This can be further confirmed by Fig. 6(b), which presents the histogram of the change-points detected by using the "PELT" function in the R package "changepoint", when dealing with each stock sequence individually. Clearly, the detected change-points spread out across the five years and thus it is not easy to combine them to yield a reasonable common set of change-points.

We apply our method to all individual stock data and generate a common set of change-points. Using default tuning parameters, five change-points are identified: 1 October 2007; 26 November 2008; 4 February 2009; 8 July 2009; and 30 August 2011. Thirteen stocks are removed out by our screening procedure, because the stock market fluctuated drastically during those five years and it is expected that most series contain at least one change-point. The identified dates may correspond to five important events: the financial crisis caused by the subprime mortgage had begun to influence the market since August 2007; the US government started the first round of quantitative easing (QE) policy on 25 November 2008; General Motors submitted a detailed restructuring plan to the US Department of Treasury in February 2009, which increased the market instability; General Motors filed for

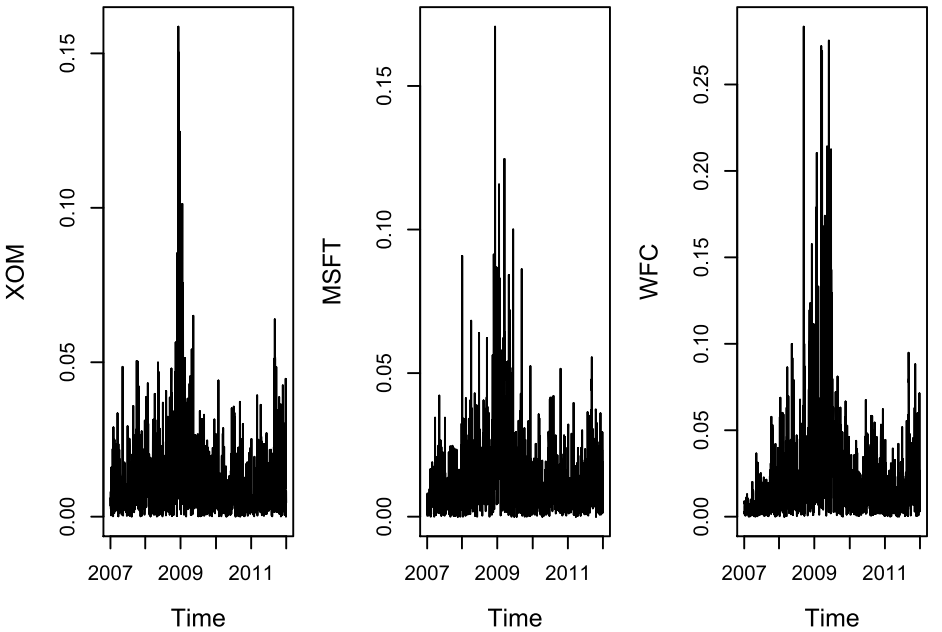


Fig. 5. The absolute log returns of the Exxon Mobil (XOM), Microsoft Corporation (MSFT) and Wells Fargo (WFC) from 1 January 2007 to 1 January 2012.

bankruptcy protection in June 2009, exacerbating the volatility of the stock market; and the second round of QE policy ended in June 2011, which indicated that the stock market was heading toward a new period.

A natural follow-up question is whether we can obtain similar results by simply applying univariate change-point detection methods to the S&P 500 index series. Figure 6(a) shows the detection results by applying the "PELT" function to the absolute log return series of the S&P 500 index. The red and green solid vertical lines indicate the segmentations of the univariate detection and our method, respectively. Although the two approaches exhibit similar patterns for the middle period of the series, the univariate method fails to find the change-point around 1 October 2007, and also leads to unreasonable detection at the end of the series; i.e. the two identified change-points are too close to each other (the last two red lines are largely overlapped). This can be partly explained by the fact that the S&P 500 index can be viewed a projection (linear combination) of all the stocks and thus using only the univariate detection incurs certain loss of the information.

4.5.2. *Changes in a series of fMRI images*

Functional magnetic resonance imaging (fMRI) has emerged as a noninvasive imaging technique that aims to localize functional brain areas in a living human brain.

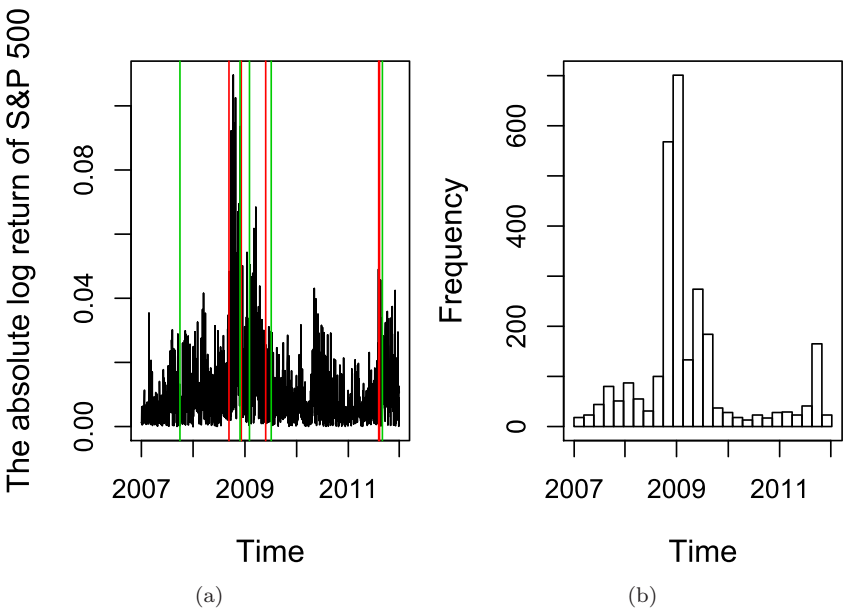
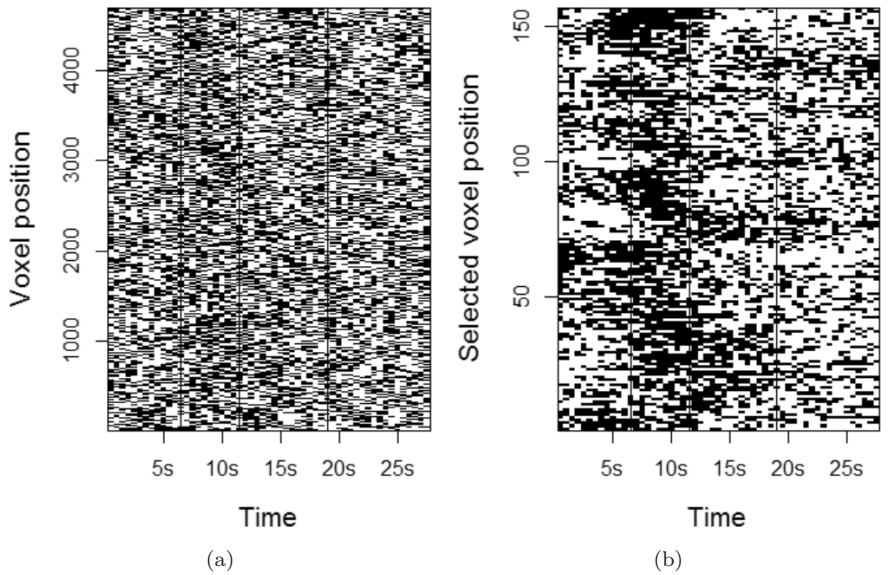


Fig. 6. (Color online) (a) The absolute log returns of the S&P 500 index; the change-points detected by our method and the PELT approach are shown by the green and red vertical lines, respectively. (b) The histogram of the change-points detected by the PELT approach with 455 univariate series.

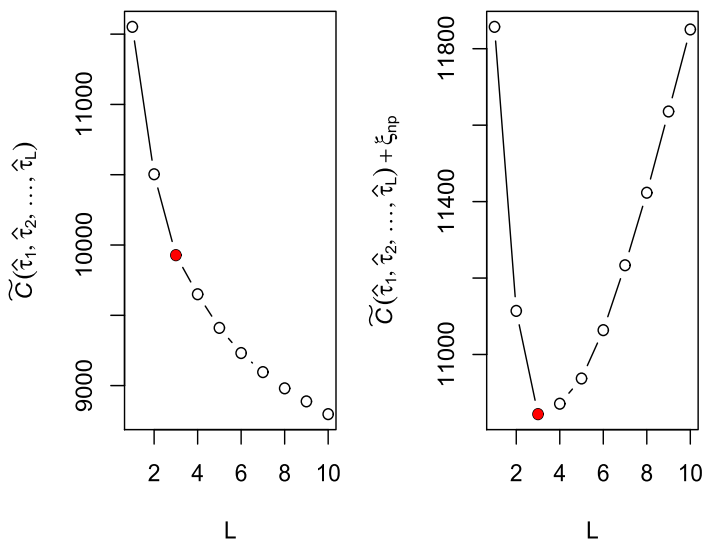
Change-point analysis appears to be a useful technique in fMRI [1], where different subjects react differently to stimuli such as stress or anxiety. In a brain functional study, subjects viewed a sentence and a picture sequentially, at 1852 ms and 1906 ms, and were then asked to press a button to indicate whether the sentence correctly described the picture. The dataset is available at http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-81/www/. The 3D picture ($64 \times 64 \times 8$) was taken every 500 ms and the experiment lasted 27 s; The sample size was $n = 54$, and each picture was transformed into a $p = 4634$ vector of voxels. As an illustration, we analyze the subject 04847 at the 37th experiment, who pressed the button at 4922 ms. We are interested in finding whether the mean of the measured noisy fMRI signal changed over time and estimating the locations of the change-points which reflected the times of brain state changed during the experiment. The information would be useful for studying the brain cognitive structure. The heatmaps of the fMRI observations for 54 time points are exhibited in Fig. 7(a). The data sequence appears to be complicated without any obvious pattern.

Based on our testing procedure, we obtain the test statistic, $S_{n,p} = 10.43$, which is highly significant compared with its standard normal null distribution. Following the rejection of the null hypothesis, we further estimate the change-points. After the initial screening procedure, 156 voxels remain, which dramatically reduces the dimensionality of change-point detection. The heatmaps of the fMRI observations of those 156 voxels are presented in Fig. 7(b). Figures 8(a) and 8(b) show the cost



(a) (b)

Fig. 7. Description of the fMRI data. (a) The heatmap of the measured noisy fMRI signals of 4634 voxels with 27 s (the brightness indicates the signal strength). (b) The heatmap of the measured noisy fMRI signals of the selected 156 voxels with 27 s.



(a) (b)

Fig. 8. (a) The cost function $\tilde{\mathcal{C}}(\widehat{\tau}_1, \ldots, \widehat{\tau}_L)$ versus the number of change-points. (b) The penalized cost function versus the number of change-points.

function $\tilde{\mathcal{C}}(\widehat{\tau}_1, \ldots, \widehat{\tau}_L)$ and its penalized version (5) versus the number of change-points, respectively. Figure 8(b) clearly suggests that the model with three change-points should fit the data the best; and the identified change-points are 6.5 s, 11.5 s and 19 s. These three change-points may correspond to the states that the subject started to think intensively after viewing the picture, began to relax after pressing the button, and recovered to the normal state, respectively. The first detected change-point 6.5 s is actually even after the time point 4922 ms when the object press the button. This further claims the existence of temporal delays (lags) in fMRI signals; see [1]. More interest may lie in estimating the distribution of the detected change-points when replicated observations are available. This segmentation result can also be verified by Fig. 8(a), where the rate of decline in the cost function changes more sharply at the point $L = 3$. Lavielle [26] suggested an intuitive method by first plotting the segmentation cost function versus the number of change-points and then finding the "elbow" in the plot, which would lead to the most suitable segmentation. The rationale underneath this method is that as more true change-points are detected, the cost function would continue to decrease, while at the same time we are likely to be detecting more false positives and thus the cost function starts to level off.

5. Concluding Remarks

Following our proposal, a natural question is whether it can handle ultra-high-dimensional cases with p increasing at an exponential rate of n. Unfortunately, it appears to be very difficult, if not impossible, when there is no sparse structure for all existing scale-invariant tests to correct bias terms. In general, it is not clear whether we can define a test statistic that is (at least) asymptotically effective without the sparsity assumption on the data structure. Our method offers some insight on how to construct tests and global multiple change-point estimators in the high-dimensional paradigm, and the results also serve as a fundamental step to move toward ultra-high dimensionality. In light of this discussion, the shrinkage estimation under sparse structures and other conditions may warrant further investigation [15].

Appendix A. Proofs of Theorems

In this appendix, we give succinct proofs of Theorems 1 and 3. Some technical arguments in Sec. 2 and the proofs of Lemmas A.2–A.5, Theorem 2, and Propositions 1 and 2 can be found in the Supplementary Material.

We first introduce a lemma which concerns concentration inequalities for the weighted sum of independent random variables, and the proof can be found in [33].

Lemma A.1. *Let* $\mathbf{X} = (X_1, \ldots, X_p)^\top$ *be a mean-zero random vector with independent components and satisfy the sub-Gaussian property* $\sup_{l \geq 1} l^{-1/2} \{E(|X_i|^l)\}^{1/l} \leq \zeta$ *for* $i = 1, \ldots, n$ *with some constant* $\zeta > 0$. *Let* $\mathbf{R}$ *be an* $p \times p$ *matrix. Then, for*

every $t \geq 0$,

$$\Pr(|\mathbf{X}^\top \mathbf{R}\mathbf{X} - E(\mathbf{X}^\top \mathbf{R}\mathbf{X})| > t) \leq \exp\left(-c\min\left\{\frac{t^2}{\zeta^4 \mathrm{tr}(\mathbf{R}^2)}, \frac{t}{\zeta^2 \lambda_{\max}}\right\}\right),$$

where $c > 0$ *is a constant and* $\lambda_{\max}$ *is the largest eigenvalue of* $\mathbf{R}$.

Define $\boldsymbol{\epsilon}_i = \mathbf{R}^{1/2}\boldsymbol{\varepsilon}_i$. The next lemma establishes the uniform convergence of $\widehat{\sigma}_j^2$ and the bound of the extreme statistic.

Lemma A.2. *Suppose condition* (C4) *holds. Under the change-point model* (1), *there exists a large constant* C_ζ *such that*

$$\Pr\left(\max_{1\leq j\leq p} |\widehat{\sigma}_j^2/\sigma_j^2 - 1| > \sqrt{C_\zeta \log(p)/n}\right) \to 0,$$

$$\Pr\left(\max_{1\leq j\leq p} \max_{1\leq \tau\leq n-1} \frac{\tau(n-\tau)}{n\sigma_j^2}(\bar{\epsilon}_{j\tau-} - \bar{\epsilon}_{j\tau+})^2 > C_\zeta \log(np)\right) \to 0,$$

as $n, p \to \infty$.

Proof of Theorem 1. Without loss of generality, we assume that $\mathbf{D} = \mathbf{I}$. Denote $\widetilde{\mathbf{R}} = \mathbf{R}^{\top/2}\mathbf{R}^{1/2}$ and remember that $\mathbf{R} = \mathbf{R}^{1/2}\mathbf{R}^{\top/2}$. In addition, denote the ith row and jth column component of $\mathbf{R}$ and $\widetilde{\mathbf{R}}$ as r_{ij} and $\widetilde{r}_{ij}$, respectively.

(i) Under the null hypothesis, there is no change-point, so we can replace the $\mathbf{X}_i$ in $S_{n,p}$ by $\boldsymbol{\epsilon}_i$. Remember that $S_{n,p} = \sum_{\tau=1}^{n-1} \tau(n-\tau)/n(\bar{\boldsymbol{\varepsilon}}_{\tau-} - \bar{\boldsymbol{\varepsilon}}_{\tau+})^\top \mathbf{R}^{\top/2}\widehat{\mathbf{D}}^{-1}\mathbf{R}^{1/2}(\bar{\boldsymbol{\varepsilon}}_{\tau-} - \bar{\boldsymbol{\varepsilon}}_{\tau+})$. By writing $\widehat{\mathbf{D}}^{-1} = (\widehat{\mathbf{D}}^{-1} - \mathbf{I}) + \mathbf{I}$, we make the decomposition $S_{n,p} = S_{n,p}^* + S_{n,p}^{**}$.

The flow chart of proving $E(S_{n,p}) = (n+2)p + o\{\sqrt{\mathrm{var}(S_{n,p})}\}$ is as follows. We first prove that $E(S_{n,p}^*) = (n-1)p$ and $E(S_{n,p}^{**}) = 3p + o\{\sqrt{\mathrm{var}(S_{n,p}^*)}\}$. Then $\mathrm{var}(S_{n,p}^{**}) = o\{\mathrm{var}(S_{n,p}^*)\}$ is proved. If the above two arguments are true, the proof is completed immediately.

By using the cyclic property of trace and exchanging the order of expectation and trace,

$$\begin{aligned} E(S_{n,p}^*) &= \sum_{\tau=1}^{n-1} \frac{\tau(n-\tau)}{n} \mathrm{tr}(E\{(\bar{\boldsymbol{\varepsilon}}_{\tau-} - \bar{\boldsymbol{\varepsilon}}_{\tau+})(\bar{\boldsymbol{\varepsilon}}_{\tau-} - \bar{\boldsymbol{\varepsilon}}_{\tau+})^\top\}\widetilde{\mathbf{R}}) \\ &= (n-1)\mathrm{tr}(\mathbf{R}) = (n-1)p. \end{aligned}$$

To derive $\mathrm{var}(S_{n,p}^*)$, we first write $S_{n,p}^*$ as

$$S_{n,p}^* = \sum_{l=1}^{n} b_l \boldsymbol{\varepsilon}_l^\top \widetilde{\mathbf{R}}\boldsymbol{\varepsilon}_l + \sum_{l=1}^{m-1}\sum_{m=2}^{n} a_{l,m}\boldsymbol{\varepsilon}_l^\top \widetilde{\mathbf{R}}\boldsymbol{\varepsilon}_m, \tag{A.1}$$

where the coefficients b_l and $a_{l,m}$ are given by $b_1 = b_n = \sum_{i=1}^{n-1} i^{-1} + n^{-1} - 1$ and $b_l = \sum_{i=1}^{l-1}(n-l)^{-1} + \sum_{i=l}^{n-1} i^{-1} + n^{-1} - 1$, $l = 2, \ldots, n-1$ and

$a_{l,m} = \sum_{i=1}^{l-1} 2(n-i)^{-1} + \sum_{i=m}^{n-1} 2i^{-1} + 2n^{-1} - 2$. If we approximate the Riemann sum with integral, it follows that

$$\int_0^1 2(-\log(1-s) - \log(s) + 1)ds = 2,$$

$$n^{-3}\sum_{l=1}^{n-1}\left(\sum_{m=l+1}^{n} a_{l,m}^2\right)^2 \approx 142, \quad \sum_{l<m}\left(\sum_{i=m+1}^{n} a_{l,i}a_{m,i}\right)^2 \le \left(\sum_{l=1}^{m-1}\sum_{m=2}^{n} a_{l,m}^2\right)^2.$$

By using the independence between $\boldsymbol{\varepsilon}_i$s, it is straightforward that

$$\begin{aligned}\mathrm{var}(S_{n,p}^*) &= \sum_{l=1}^{n} b_l^2(E\{(\boldsymbol{\varepsilon}_l^\top \widetilde{\mathbf{R}}\boldsymbol{\varepsilon}_l)^2\} - \{E(\boldsymbol{\varepsilon}_l^\top \widetilde{\mathbf{R}}\boldsymbol{\varepsilon}_l)\}^2) + \sum_{l=1}^{m-1}\sum_{m=2}^{n} a_{l,m}^2 E\{(\boldsymbol{\varepsilon}_l^\top \widetilde{\mathbf{R}}\boldsymbol{\varepsilon}_m)^2\} \\ &\approx \frac{(15-\pi^2)n}{3}(E\{(\boldsymbol{\varepsilon}_l^\top \widetilde{\mathbf{R}}\boldsymbol{\varepsilon}_l)^2\} - p^2) + \frac{(2\pi^2-18)}{n^2}\mathrm{tr}(\mathbf{R}^2),\end{aligned}$$

where the approximation is due to replacing Riemann sum by integral. Similar to (A.1), we can write $S_{n,p}^{**} = S_{n,p}^A + S_{n,p}^B$. By using the expansion, $\widehat{\mathbf{D}}^{-1} - \mathbf{I} = (\mathbf{I} - \widehat{\mathbf{D}}) + (\mathbf{I} - \widehat{\mathbf{D}})^2 + (\mathbf{I} - \widehat{\mathbf{D}})^3\widehat{\mathbf{D}}^{-1}$, we can further decompose $S_{n,p}^A$ (and $S_{n,p}^B$) into three parts $S_{n,p}^{A1}, S_{n,p}^{A2}$ and $S_{n,p}^{A3}$ ($S_{n,p}^{B1}, S_{n,p}^{B2}$ and $S_{n,p}^{B3}$), respectively. We first claim that $E(S_{n,p}^{A1} + S_{n,p}^{A2} + S_{n,p}^{B1} + S_{n,p}^{B2}) = 3p + o\{\sqrt{\mathrm{var}(S_{n,p}^*)}\}$. Write $S_{n,p}^{A1} + S_{n,p}^{A2} + S_{n,p}^{B1} + S_{n,p}^{B2} = \sum_{j=1}^{p}(\sum_{l=1}^{n} b_l\epsilon_{jl}^2 + \sum_{l=1}^{m-1}\sum_{m=2}^{n} a_{l,m}\epsilon_{jl}\epsilon_{jm})(2 - 3\widehat{\sigma}_j^2 + \widehat{\sigma}_j^4) \stackrel{\mathrm{def}}{=} 2A - 3B + C$. For the first term, it is obvious that $E(A) = p\sum_{l=1}^{n} b_l = (n-1)p$. Remember that $\widehat{\sigma}_j^2 = \sum_{i=1}^{n-1}(2n-2)^{-2}(\epsilon_{ji} - \epsilon_{j,i+1})^2$. By tedious calculation,

$$\begin{aligned}p^{-1}E(B) &= \sum_{l=1}^{n} b_l - \frac{1}{2(n-1)}\left(b_1 + b_n + 2\sum_{l=2}^{n-1} b_l\right) - \frac{1}{n-1}\sum_{l=1}^{n-1} a_{l,l+1} \\ &\quad + \frac{1}{2p(n-1)}\left(b_1 + b_n + 2\sum_{l=2}^{n-1} b_l\right)\sum_{j=1}^{p} E(\epsilon_j^4); \\ p^{-1}E(C) &= \sum_{l=1}^{n} b_l - \frac{1}{n-1}\left(b_1 + b_n + 2\sum_{l=2}^{n-1} b_l\right) - \frac{2}{n-1}\sum_{l=1}^{n-1} a_{l,l+1} \\ &\quad + \frac{1}{(n-1)p}\left(2b_1 + 2b_n + 3\sum_{l=2}^{n-1} b_l\right)\sum_{j=1}^{p} E(\epsilon_j^4) + o(n^{-\nu})\end{aligned}$$

for any $0 < \nu < 1$. Combining the above results, we have

$$E(2A - 3B + C) = \frac{p}{2(n-1)}\left(b_1 + b_n + 2\sum_{l=2}^{n-1} b_l\right) + \frac{p}{n-1}\sum_{k=2}^{n} a_{k-1,k} + o(pn^{-\nu}),$$

which converges to $3p + o(pn^{-\nu})$. By condition (C3), $p/n = o(\sqrt{\mathrm{var}(S_{n,p}^*)})$ and thus the claim is verified.

In the following we prove that $\mathrm{var}(S^{**}_{n,p}) = o(\mathrm{var}(S^{*}_{n,p}))$. We take $\mathrm{var}(S^{A1}_{n,p})$ as an example and the variances of the terms $S^{A2}_{n,p}$, $S^{B1}_{n,p}$ and $S^{B2}_{n,p}$ can be similarly proved to be $o(\mathrm{var}(S^{*}_{n,p}))$. Denote $\boldsymbol{\epsilon}_i = \mathbf{R}^{1/2}\boldsymbol{\varepsilon}_i$, $\widehat{\sigma}_j^2 - 1 = \sum_{i=1}^{2n-1} c_i\eta_{ji}$ with $c_1 = c_n = 1/(2n-2)$, $c_i = 1/(n-1)$ for $i = 2, \ldots, n-1$ and $c_i = -1/(n-1)$ for $i = n+1, \ldots, 2n-1$. Moreover, $\eta_{ji} = \epsilon_{ji}^2 - 1$ for $i = 1, \ldots, n$ and $\eta_{ji} = \epsilon_{j,i-1}\epsilon_{j,i-n+1}$ for $i = n+1, \ldots, 2n-1$. By calculation,

$$
\begin{aligned}
\mathrm{var}(S^{A1}_{n,p}) &= \sum_{s_1,s_2=1}^{p}\sum_{l_1,l_2=1}^{n}\sum_{o_1,o_2=1}^{2n-1} b_{l_1}b_{l_2}c_{o_1}c_{o_2}\mathrm{cov}(\epsilon^2_{s_1l_1}\eta_{s_1o_1}, \epsilon^2_{s_2l_2}\eta_{s_2o_2}),\\
&= \sum_{s_1,s_2=1}^{p}\sum_{l_1=l_2}^{n}\sum_{o_1=o_2\neq l_1}^{n} b_{l_1}b_{l_2}c_{o_1}c_{o_2}E(\epsilon^2_{s_1l_1}\epsilon^2_{s_2l_2})\mathrm{cov}(\epsilon^2_{s_1o_1}, \epsilon^2_{s_2o_2})\\
&\quad + \sum_{s_1,s_2=1}^{p}\sum_{l_1=l_2=1}^{n}\sum_{o_1=o_2=n+1}^{2n-1} b_{l_1}b_{l_2}c_{o_1}c_{o_2}E(\epsilon^2_{s_1l_1}\epsilon^2_{s_2l_2})r^2_{s_1s_2}\\
&\quad + 2\sum_{s_1,s_2=1}^{p}\sum_{l_1\neq l_2=1}^{n}\sum_{o_1=l_1,o_2=l_2}^{n} b_{l_1}b_{l_2}c_{o_1}c_{o_2}\mathrm{cov}(\epsilon^2_{s_1l_1}, \epsilon^2_{s_2o_1})\mathrm{cov}(\epsilon^2_{s_1l_2}, \epsilon^2_{s_2o_2})\\
&\quad + \sum_{s_1,s_2=1}^{p}\sum_{l_1=l_2=1}^{n}\sum_{o_1=l_1,o_2=l_2}^{n} b_{l_1}b_{l_2}c_{o_1}c_{o_2}\mathrm{cov}(\epsilon^2_{s_1l_1}\eta_{s_1o_1}, \epsilon^2_{s_2l_2}\eta_{s_2o_2})\\
&= \sum_{s_1,s_2=1}^{p} E(\epsilon^2_{s_1l_1}\epsilon^2_{s_2l_2})\mathrm{cov}(\epsilon^2_{s_1o_1}, \epsilon^2_{s_2o_2})O(1) + \sum_{s_1,s_2=1}^{p} E(\epsilon^2_{s_1l_1}\epsilon^2_{s_2l_2})r^2_{s_1s_2}O(1)\\
&\quad + \sum_{s_1,s_2=1}^{p} \mathrm{cov}(\epsilon^2_{s_1l_1}, \epsilon^2_{s_2o_1})\mathrm{cov}(\epsilon^2_{s_1l_2}, \epsilon^2_{s_2o_2})O(1) + O(p^2/n)\\
&= O(\mathrm{tr}(\mathbf{R}^2)) + O(p^2/n) = o(\mathrm{var}(S^{*}_{n,p})),
\end{aligned}
$$

where we use the fact that $\sum_{s_1,s_2=1}^{p}\{\mathrm{cov}(\epsilon^2_{s_1l}, \epsilon^2_{s_2l})\}^2 = O(\mathrm{tr}(\mathbf{R}^2))$ and the assumption $p/n^3 \to 0$. Thus we only need to prove that $S^{A3}_{n,p} + S^{B3}_{n,p} = o_p(\sqrt{\mathrm{var}(S^{*}_{n,p})})$. We have

$$
\begin{aligned}
S^{A3}_{n,p} + S^{B3}_{n,p} &= \sum_{j=1}^{p}\sum_{\tau=1}^{n-1}\frac{\tau(n-\tau)}{n}(\bar{\epsilon}_{j\tau-} - \bar{\epsilon}_{j\tau+})^2(1-\widehat{\sigma}_j^2)^3\widehat{\sigma}_j^{-2}\\
&\leq \sum_{j=1}^{p}\sum_{\tau=1}^{n-1}\frac{\tau(n-\tau)}{n}(\bar{\epsilon}_{j\tau-} - \bar{\epsilon}_{j\tau+})^2 \max_{1\leq j\leq p}(1-\widehat{\sigma}_j^2)^3 \max_{1\leq j\leq p}\widehat{\sigma}_j^{-2}\\
&= O_p(p\{\log(np)\}^{\kappa}/n^{1/2}) = o_p(\sqrt{\mathrm{var}(S^{*}_{n,p})}),
\end{aligned}
$$

where $\kappa > 3/2$ is a constant and the third equality is due to Lemma A.2. Thus (i) has been proved.

(ii) Define $\varsigma_l = b_l\{\boldsymbol{\varepsilon}_l^\top \widetilde{\mathbf{R}}\boldsymbol{\varepsilon}_l - \mathrm{tr}(\mathbf{R})\} + \sum_{l=1}^{m-1} a_{l,m}\boldsymbol{\varepsilon}_l^\top \widetilde{\mathbf{R}}\boldsymbol{\varepsilon}_l$. Then, $S_{n,p}^* - E(S_{n,p}^*) = \sum_{l=1}^n \varsigma_l$ is the sum of a martingale difference sequence. To prove the normality, it suffices to verify the following two convergence properties [20],

$$Z_1 = \sum_{l=1}^n E(\varsigma_l^2 \,|\, \mathcal{F}_{l-1})/\mathrm{var}(S_{n,p}^*) \xrightarrow{\mathcal{P}} 1, \tag{A.2}$$

$$Z_2 = \sum_{l=1}^n E(\varsigma_l^4)/\{\mathrm{var}(S_{n,p}^*)\}^2 \to 0, \tag{A.3}$$

where $\mathcal{F}_l$ is the σ-field constructed by $(\boldsymbol{\varepsilon}_1, \ldots, \boldsymbol{\varepsilon}_l)$. It is obvious that

$$Z_1 = \left\{\sum_{l=1}^n b_l^2 \mathrm{var}(\boldsymbol{\varepsilon}_l^\top \widetilde{\mathbf{R}}\boldsymbol{\varepsilon}_l) + \sum_{l_1=1}^{m-1}\sum_{l_2=1}^{m-1}\sum_{m=2}^{n} a_{l_1,m}a_{l_2,m}\boldsymbol{\varepsilon}_{l_1,m}^\top \widetilde{\mathbf{R}}^2\boldsymbol{\varepsilon}_{l_2,m}\right\} \Big/ \mathrm{var}(S_{n,p}^*),$$

with $E(Z_1) = 1$. By changing the sum orders, we have

$$\begin{aligned}
&\mathrm{var}(Z_1)\{\mathrm{var}(S_{n,p}^*)\}^2 \\
&\quad = \sum_{l=1}^{n-1}\left(\sum_{m=l+1}^n a_{l,m}^2\right)^2 \mathrm{var}(\boldsymbol{\varepsilon}_l^\top \widetilde{\mathbf{R}}^2\boldsymbol{\varepsilon}_l) + 4\sum_{l_1<l_2}\left(\sum_{m=l_2+1}^n a_{l_1,m}a_{l_2,m}\right)^2 \mathrm{tr}(\widetilde{\mathbf{R}}^4) \\
&\quad = O(n^3\{\mathrm{tr}(\mathbf{R}^2)\}^2 + n^4\mathrm{tr}(\mathbf{R}^4)) = o(\{\mathrm{var}(S_{n,p}^*)\}^2),
\end{aligned}$$

where we use the fact that $\sum_{l=1}^{n-1}(\sum_{m=l+1}^n a_{l,m}^2)^2 \approx n^3$, $\sum_{l<m}(\sum_{i=m+1}^n a_{l,i}a_{m,i})^2 \le (\sum_{l=1}^{m-1}\sum_{m=2}^n a_{l,m}^2)^2$. Thus (A.2) is verified.

For (A.3), denote $\widetilde{r}_{s_1s_2}^{(2)} = \sum_{s_3=1}^p \widetilde{r}_{s_1s_3}\widetilde{r}_{s_2s_3}$ as the s_1th row and s_2th column component of $\widetilde{\mathbf{R}}^2$, by using Cauchy inequality several times, we have

$$\begin{aligned}
&Z_2\{\mathrm{var}(S_{n,p}^*)\}^2 \\
&\quad \le 7\sum_{m=2}^n E\left\{\left(\sum_{l=1}^{m-1} a_{l,m}\boldsymbol{\varepsilon}_l^\top \widetilde{\mathbf{R}}\boldsymbol{\varepsilon}_m\right)^4\right\} + 5\sum_{l=1}^n b_l^4 E\{(\boldsymbol{\varepsilon}_l^\top \widetilde{\mathbf{R}}\boldsymbol{\varepsilon}_l - \mathrm{tr}(\mathbf{R}))^4\} \\
&\quad = 7\sum_{m=2}^n\sum_{l_1\ldots l_4}^{m-1}\sum_{s_1\ldots s_8}^p a_{l_1m}\cdots a_{l_4m}\widetilde{r}_{s_1s_2}\cdots\widetilde{r}_{s_7s_8}E(\varepsilon_{s_1l_1}\varepsilon_{s_2m}\cdots\varepsilon_{s_7l_4}\varepsilon_{s_8m}) \\
&\qquad + 5\sum_{l=1}^n\sum_{s_1\ldots s_8}^p b_l^4\widetilde{r}_{s_1s_2}\cdots\widetilde{r}_{s_7s_8}E(\{\varepsilon_{s_1l}\varepsilon_{s_2l} - I(s_1=s_2)\} \\
&\qquad \cdots\{\varepsilon_{s_7l}\varepsilon_{s_8l} - I(s_7=s_8)\}) \\
&\quad \le C\sum_{m=2}^n\sum_{l=1}^{m-1}\sum_{s_1\ldots s_8}^p a_{lm}^4\widetilde{r}_{s_1s_2}\cdots\widetilde{r}_{s_7s_8}E(\varepsilon_{s_1l}\varepsilon_{s_2m}\cdots\varepsilon_{s_7l}\varepsilon_{s_8m})
\end{aligned}$$

$$
\begin{aligned}
&+C\sum_{m=2}^{n}\sum_{l_1\neq l_2}^{m-1}\sum_{s_1\ldots s_8}^{p} a_{l_1m}\cdots a_{l_4m}\widetilde{r}_{s_1s_2}\\
&\cdots\widetilde{r}_{s_7s_8}E(\varepsilon_{s_1l_1}\varepsilon_{s_2m}\varepsilon_{s_3l_1}\varepsilon_{s_4m}\varepsilon_{s_5l_2}\varepsilon_{s_6m}\varepsilon_{s_7l_2}\varepsilon_{s_8m}\varepsilon_{s_8m})\\
&+C\sum_{l=1}^{n}\sum_{s=1}^{p} b_l^4\widetilde{r}_{ss}^4+C\sum_{l=1}^{n}\sum_{s_1\neq s_2}^{p} b_l^4\widetilde{r}_{s_1s_2}^4+C\sum_{l=1}^{n}\sum_{s_1\neq s_2}^{p} b_l^4\widetilde{r}_{s_1s_1}^2\widetilde{r}_{s_2s_2}^2\\
&+C\sum_{l=1}^{n}\sum_{s_1\neq s_2}^{p} b_l^4\widetilde{r}_{s_1s_1}\widetilde{r}_{s_1s_2}^2\widetilde{r}_{s_2s_2}+C\sum_{l=1}^{n}\sum_{s_1\neq s_2\neq s_3}^{p} b_l^4\widetilde{r}_{s_1s_1}\widetilde{r}_{s_1s_2}\widetilde{r}_{s_2s_3}\widetilde{r}_{s_3s_3}\\
&+C\sum_{l=1}^{n}\sum_{s_1\neq s_2\neq s_3}^{p} b_l^4\widetilde{r}_{s_1s_2}^2\widetilde{r}_{s_1s_3}\widetilde{r}_{s_2s_3}+C\sum_{l=1}^{n}\sum_{s_1\neq s_2\neq s_3\neq s_4}^{p} b_l^4\widetilde{r}_{s_1s_2}^2\widetilde{r}_{s_3s_4}^2\\
&+C\sum_{l=1}^{n}\sum_{s_1\neq s_2\neq s_3\neq s_4}^{p} b_l^4\widetilde{r}_{s_1s_2}\widetilde{r}_{s_2s_3}\widetilde{r}_{s_3s_4}\widetilde{r}_{s_1s_4}.
\end{aligned}
$$

By calculating the right-hand side of the above inequality,

$$
\begin{aligned}
Z_2\{\mathrm{var}(S_{n,p}^*)\}^2 &\le C'\sum_{m=2}^{n}\sum_{l=1}^{m-1} a_{lm}^4\sum_{s_1s_2}^{p}\widetilde{r}_{s_1s_2}^4+C'\sum_{m=2}^{n}\sum_{l=1}^{m-1} a_{lm}^4\sum_{s_1\neq s_2,s_3}^{p}\widetilde{r}_{s_1s_3}^2\widetilde{r}_{s_2s_3}^2\\
&\quad+C'\sum_{m=2}^{n}\sum_{l=1}^{m-1} a_{lm}^4\sum_{s_1\neq s_2,s_3\neq s_4}^{p}\widetilde{r}_{s_1s_3}^2\widetilde{r}_{s_2s_4}^2\\
&\quad+C'\sum_{m=2}^{n}\sum_{l_1\neq l_2}^{m-1} a_{l_1m}^2a_{l_2m}^2\sum_{s_1s_2s_3}^{p}\widetilde{r}_{s_1s_3}^2\widetilde{r}_{s_2s_3}^2\\
&\quad+C'\sum_{m=2}^{n}\sum_{l_1\neq l_2}^{m-1} a_{l_1m}^2a_{l_2m}^2\sum_{s_1s_2,s_3\neq s_4}^{p}\widetilde{r}_{s_1s_3}^2\widetilde{r}_{s_2s_4}^2\\
&\quad+C'\sum_{m=2}^{n}\sum_{l_1\neq l_2}^{m-1} a_{l_1m}^2a_{l_2m}^2\sum_{s_1s_2,s_3\neq s_4}^{p}\widetilde{r}_{s_1s_3}\widetilde{r}_{s_1s_4}\widetilde{r}_{s_2s_3}\widetilde{r}_{s_2s_4}\\
&\quad+C''\sum_{l=1}^{n} b_l^4\left\{\sum_{s_1s_2}^{p}\widetilde{r}_{s_1s_2}^2\right\}^2\\
&=O\left(\sum_{m=2}^{n}\sum_{l_1l_2}^{m-1} a_{l_1m}^2a_{l_2m}^2+\sum_{l=1}^{n} b_l^4p\right)\left\{\sum_{s_1s_2}^{p}\widetilde{r}_{s_1s_2}^2\right\}^2\\
&=O(n^3\{\mathrm{tr}(\widetilde{\mathbf{R}}^2)\}^2+np\mathrm{tr}(\widetilde{\mathbf{R}}^2))\\
&=o(\{\mathrm{var}(S_{n,p}^*)\}^2),
\end{aligned}
$$

where we use conditions (C2), (C3) and the fact that $\sum_{m=2}^{n}\sum_{l_1 l_2}^{m-1} a_{l_1 m}^2 a_{l_2 m}^2 \approx n^3$, $\sum_{l=1}^{n} b_l^4 \approx n$ and

$$\sum_{s_1,s_2,s_3\neq s_4}^{p} \widetilde{r}_{s_1 s_3}\widetilde{r}_{s_1 s_4}\widetilde{r}_{s_2 s_3}\widetilde{r}_{s_2 s_4} = \sum_{s_3\neq s_4}^{p}\{\widetilde{r}_{s_3 s_4}^{(2)}\}^2 \leq \sum_{s_3 s_4}^{p}\{\widetilde{r}_{s_3 s_4}^{(2)}\}^2 = \mathrm{tr}(\widetilde{\mathbf{R}}^4),$$

$$\sum_{s_1\neq s_2\neq s_3\neq s_4}^{p} \widetilde{r}_{s_1 s_3}\widetilde{r}_{s_1 s_4}\widetilde{r}_{s_2 s_3}\widetilde{r}_{s_2 s_4}$$

$$\leq \sum_{s_1\neq s_2\neq s_3\neq s_4}^{p}\{\widetilde{r}_{s_1 s_3}^2\widetilde{r}_{s_1 s_4}^2 + \widetilde{r}_{s_2 s_3}^2\widetilde{r}_{s_2 s_4}^2\}/2 \leq p\left\{\sum_{s_1 s_2}^{p}\widetilde{r}_{s_1 s_2}^2\right\}^2,$$

$$\sum_{s_1\neq s_2\neq s_3}^{p} \widetilde{r}_{s_1 s_2}^2\widetilde{r}_{s_1 s_3}\widetilde{r}_{s_2 s_3} \leq \sum_{s_1\neq s_2\neq s_3}(\widetilde{r}_{s_1 s_2}^2\widetilde{r}_{s_1 s_3}^2 + \widetilde{r}_{s_1 s_2}^2\widetilde{r}_{s_2 s_3}^2)/2 \leq \left\{\sum_{s_1 s_2}^{p}\widetilde{r}_{s_1 s_2}^2\right\}^2.$$

This completes the proof of (ii).

(iii) Note that, under the null hypothesis,

$$E_{n,p} = c_{n,p} I\left(\max_{1\leq j\leq p}\max_{1\leq\tau\leq n-1}\frac{\tau(n-\tau)}{n\sigma_j^2}(\bar{\epsilon}_{j\tau-} - \bar{\epsilon}_{j\tau+})^2\sigma_j^2\widehat{\sigma}_j^{-2} > h_{n,p}\right).$$

Therefore, by Lemmas A.2 and A.3,

$$\Pr(E_{n,p} > 0) \leq \Pr\left(\max_{1\leq j\leq p}\max_{1\leq\tau\leq n-1}\frac{\tau(n-\tau)}{n\sigma_j^2}(\bar{\epsilon}_{j\tau-} - \bar{\epsilon}_{j\tau+})^2 > h_{n,p}/2\right)$$

$$+\Pr\left(\max_{1\leq j\leq p}\sigma_j^2\widehat{\sigma}_j^{-2} > 2\right) \to 0,$$

when $h_{n,p}/\log(np) \to \infty$. By Slutsky's theorem, (iii) is proved. □

Before proceeding further, we claim two key lemmas, which allow us to control the upper bound of the total variation in $\mathcal{C}(\tau_1,\ldots,\tau_L)$ when adding a true change-point and the lower bound of that when adding an arbitrary point, respectively.

Lemma A.3. *Under conditions* (C1), (C4) *and* (C6), *we have*

$$\Pr\left(\max_{1\leq k<l<m\leq n}|\|\widehat{\mathbf{D}}^{-1/2}\mathbf{R}^{1/2}\bar{\boldsymbol{\varepsilon}}_{klm}\|^2 - \mathrm{tr}(\mathbf{R})| < \xi_{np}\right) \to 1,$$

as $\xi_{np}/\{\sqrt{\mathrm{tr}(\mathbf{R}^2)}\log(n)\} \to \infty$.

Lemma A.4. *For any p-dimension real vector* $\boldsymbol{\mu}$, *there exists a sufficiently large constant* $C > 0$ *so that*

$$\Pr\left(\min_{1\leq k<l<m\leq n}\|\widehat{\mathbf{D}}^{-1/2}\mathbf{R}^{1/2}(\bar{\boldsymbol{\varepsilon}}_{klm} - \boldsymbol{\mu})\|^2 - \mathrm{tr}(\mathbf{R}) \geq -C\sqrt{\mathrm{tr}(\mathbf{R}^2)}\log(n)\right) \to 1.$$

Proof of Theorem 3. We first prove the consistency of the change-point number estimator, i.e. $\Pr(\widehat{K} = K) \to 1$. When $\widehat{K} < K$, there exists at least one true change-point τ_m^* such that $\min_{1\le l\le \widehat{K}} |\widehat{\tau}_l - \tau_m^*| \ge \lambda_n/2$. Denote the true change-point set $\mathcal{B}^* = (\tau_1^*, \tau_2^*, \ldots, \tau_m^*)$ and the estimated change-point set $\widehat{\mathcal{B}} = (\widehat{\tau}_1, \widehat{\tau}_2, \ldots, \widehat{\tau}_{\widehat{K}})$. Let $\mathcal{B}_0 = \{\mathcal{B}^* \backslash \tau_m^*\} \cup \{\tau_m^* - \lambda_n/2, \tau_m^* + \lambda_n/2\}$ and $\mathcal{B}_1 = \mathcal{B}_0 \cup \widehat{\mathcal{B}}$. Without loss of generality, we assume that $\mathcal{B}_0 \cap \widehat{\mathcal{B}} = \emptyset$. Define $\mathcal{A}_k$, $k = 1, 2, \ldots, \widehat{K} + 1$ as

$$\mathcal{A}_k = \{\tau_l \,|\, \widehat{\tau}_{k-1} < \tau_l < \widehat{\tau}_k, \tau_l \in \mathcal{B}_0\},$$

with $\widehat{\tau}_0 = 1$ and $\widehat{\tau}_{\widehat{K}+1} = n + 1$. Rewrite $\mathcal{A}_k = \{\tau_{k,1}, \tau_{k,2}, \ldots, \tau_{k,l_k}\}$ and denote $l_k = |\mathcal{A}_k|$ as the cardinality of $\mathcal{A}_k$. By calculation,

$$\begin{aligned}
&\mathcal{C}(\widehat{\mathcal{B}}) - \mathcal{C}(\mathcal{B}_1) \\
&\quad = \sum_{k=1}^{\widehat{K}+1} \sum_{o=1}^{l_k} \frac{(\tau_{ko} - \tau_{k,o-1})(\widehat{\tau}_k - \tau_{ko})}{\widehat{\tau}_k - \tau_{k,o-1}} \|\widehat{\mathbf{D}}^{-1/2}(\bar{\mathbf{X}}_{\tau_{k,o-1},\tau_{ko}} - \bar{\mathbf{X}}_{\tau_{ko},\widehat{\tau}_k})\|^2 \\
&\quad = \sum_{k=1}^{\widehat{K}+1} \sum_{o=1}^{l_k} \frac{(\tau_{ko} - \tau_{k,o-1})(\widehat{\tau}_k - \tau_{ko})}{\widehat{\tau}_k - \tau_{k,o-1}} \|\widehat{\mathbf{D}}^{-1/2}\mathbf{R}^{1/2}(\bar{\boldsymbol{\varepsilon}}_{\tau_{k,o-1}\tau_{ko}\widehat{\tau}_k} - \bar{\boldsymbol{\mu}}_{\tau_{k,o-1}\tau_{ko}\widehat{\tau}_k})\|^2,
\end{aligned}$$

where $\tau_{k,0} = \widehat{\tau}_{k-1}$. Note that $\sum_{k=1}^{\widehat{K}+1} l_k = K + 1$. By Lemma A.4, with probability tending to 1,

$$\begin{aligned}
\mathcal{C}(\widehat{\mathcal{B}}) - \mathcal{C}(\mathcal{B}_1) &\ge (K+1)\mathrm{tr}(\mathbf{R}) - C\sqrt{\mathrm{tr}(\mathbf{R}^2)}\log(n) \\
&\ge (K+1)\mathrm{tr}(\mathbf{R}) + o_p(\xi_{np}).
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathcal{C}(\mathcal{B}_1) - \mathcal{C}(\mathcal{B}^*) &= \mathcal{C}(\mathcal{B}_1) - \mathcal{C}(\mathcal{B}_1 \cup \{\tau_m^*\}) + \mathcal{C}(\mathcal{B}_1 \cup \{\tau_m^*\}) - \mathcal{C}(\mathcal{B}^*) \\
&= \lambda_n \|\boldsymbol{\mu}_{m+1} - \boldsymbol{\mu}_m\|^2 (1/4 + o_p(1)) - \sum_{s=1}^{\widehat{K}+2} \bar{\boldsymbol{\varepsilon}}_{k_s l_s r_s}^\top \mathbf{R}^{\top/2} \widehat{\mathbf{D}}^{-1} \mathbf{R}^{1/2} \bar{\boldsymbol{\varepsilon}}_{k_s l_s r_s} \\
&\quad + \bar{\boldsymbol{\varepsilon}}_{k_{\widehat{K}+3} l_{\widehat{K}+3} r_{\widehat{K}+3}}^\top \mathbf{R}^{\top/2} \widehat{\mathbf{D}}^{-1} \mathbf{R}^{1/2} \bar{\boldsymbol{\varepsilon}}_{k_{\widehat{K}+3} l_{\widehat{K}+3} r_{\widehat{K}+3}},
\end{aligned}$$

where the ranges of k_s, l_s and r_s are all omitted, because by Lemma A.4, for $s = 1, \ldots, \widehat{K} + 3$, $\bar{\boldsymbol{\varepsilon}}_{k_s l_s r_s}^\top \mathbf{R}^{\top/2} \widehat{\mathbf{D}}^{-1} \mathbf{R}^{1/2} \bar{\boldsymbol{\varepsilon}}_{k_s l_s r_s} = \mathrm{tr}(\mathbf{R}) + o_p(\xi_{np})$. Accordingly,

$$\begin{aligned}
\mathrm{SIC}(\widehat{\mathcal{B}}) - \mathrm{SIC}(\mathcal{B}^*) &= \mathcal{C}(\widehat{\mathcal{B}}) - \mathcal{C}(\mathcal{B}^*) - (K - \widehat{K})(\mathrm{tr}(\mathbf{R}) + \xi_{np}) \\
&\ge \lambda_n \|\boldsymbol{\mu}_{m+1} - \boldsymbol{\mu}_m\|^2 (1/4 + o_p(1)) + O_p(\xi_{np})
\end{aligned}$$

by noticing the ratio consistency of $\widehat{\mathrm{tr}(\mathbf{R}^2)}$. By condition (C5), $\Pr(\mathrm{SIC}(\widehat{\mathcal{B}}) > \mathrm{SIC}(\mathcal{B}^*)) \to 1$, which implies that $\Pr(\widehat{K} \ge K) \to 1$.

Now, suppose that $\widehat{K} > K$. By Lemma A.4, $\mathcal{C}(\widehat{\mathcal{B}}) - \mathcal{C}(\widehat{\mathcal{B}} \cup \mathcal{B}^*) \geq K\text{tr}(\mathbf{R}) + o_p(\xi_{np})$, and by Lemma A.3, $\mathcal{C}(\widehat{\mathcal{B}} \cup \mathcal{B}^*) - \mathcal{C}(\mathcal{B}^*) = -\widehat{K}\text{tr}(\mathbf{R}) + o_p(\xi_{np})$. So,

$$\text{SIC}(\widehat{\mathcal{B}}) - \text{SIC}(\mathcal{B}^*) \geq (\widehat{K} - K)\xi_{np} + o_p(\xi_{np}) \to \infty.$$

This contradicts the definition of $\text{SIC}(\widehat{\mathcal{B}})$. Thus, we conclude that $\widehat{K} = K$ with probability tending to one.

If $\max_{1\leq k\leq K} \min_{1\leq l\leq \widehat{K}} |\tau_k^* - \widehat{\tau}_l| > \Delta_{np}$, there exists at least one true change-point τ_m^* such that $\min_{1\leq l\leq \widehat{K}} |\widehat{\tau}_l - \tau_m^*| > \Delta_{np}$. Redefine $\mathcal{B}_0 = \{\mathcal{B}^* \backslash \tau_m^*\} \cup \{\tau_m^* - \Delta_{np}/2, \tau_m^* + \Delta_{np}/2\}$ and use the same method as that in the proof of $\Pr(\widehat{K} > K) \to 1$, it is easy to prove that

$$\text{SIC}(\widehat{\mathcal{B}}) - \text{SIC}(\mathcal{B}^*) \geq \Delta_{n,p}\|\boldsymbol{\mu}_{m+1} - \boldsymbol{\mu}_m\|^2(1/4 + o_p(1)) + O_p(\xi_{np}).$$

Hence, as $\Delta_{np}\delta_{np}/\xi_{np} \to \infty$, $\Pr(\text{SIC}(\widehat{\mathcal{B}}) > \text{SIC}(\mathcal{B}^*)) \to 1$, which leads to the contradiction. This completes the proof of Theorem 3. □

Acknowledgments

The authors thank the editor, associate editor, and anonymous referees for many helpful comments that have resulted in significant improvements in the paper. This research was supported by NNSF of China Grants 1690015, 11622104, 11431006 and 11771332, and Excellent Youth Foundation of Tianjin Scientific Committee 18JCJQJC46000.

References

[1] J. A. D. Aston and C. Kirch, Evaluating stationarity via change-point alternatives with applications to FMRI data, *Ann. Appl. Statist.* **6** (2012) 1906–1948.
[2] J. A. D. Aston and C. Kirch, Change points in high dimensional settings, preprint (2014), arXiv:1409.1771.
[3] A. Aue and L. Horváth, Structural breaks in time series, *J. Time Series Anal.* **34** (2013) 1–16.
[4] J. Bai, Common breaks in means and variances for panel data, *J. Econ.* **157** (2010) 78–92.
[5] J. Bai and P. Perron, Estimating and testing linear models with multiple structural changes, *Econometrica* **70** (1998) 9–38.
[6] Z. Bai and H. Saranadasa, Effect of high dimension: By an example of a two sample problem, *Statist. Sin.* **6** (1996) 311–329.
[7] I. Berkes, R. Gabrys, L. Horváth and P. Kokoszka, Detecting changes in the mean of functional observations, *J. Roy. Statist. Soc.*: *Ser. B* (*Statist. Methodol.*) **71** (2009) 927–946.
[8] L. Boysen, A. Kempe, V. Liebscher, A. Munk and O. Wittich, Consistencies and rates of convergence of jump-penalized least squares estimators, *Ann. Statist.* **37** (2009) 157–183.
[9] H. Chen and T. Jiang, A study of two high-dimensional likelihood ratio tests under alternative hypotheses, *Random Matrices*: *Theory Appl.* **7** (2018) 1750016.
[10] S.-X. Chen and Y.-L. Qin, A two-sample test for high-dimensional data with applications to gene-set testing, *Ann. Statist.* **38** (2010) 808–835.

[11] H. Cho and P. Fryzlewicz, Multiple change-point detection for high dimensional time series via sparsified binary segmentation, *J. Roy. Statist. Soc.*: *Ser. B* (*Statist. Methodol.*) **77** (2015) 475–507.
[12] M. Csörgö and L. Horváth, *Limit Theorems in Change-point Analysis* (John Wiley & Sons, 1997).
[13] F. Enikeeva and Z. Harchaoui, High-dimensional change-point detection with sparse alternatives, preprint (2013), arXiv:1312.1900.
[14] J. Fan, Test of significance based on wavelet thresholding and Neyman's truncation, *J. Amer. Statist. Assoc.* **91** (1996) 674–688.
[15] J. Fan, F. Han and H. Liu, Challenges of big data analysis, *Natl. Sci. Rev.* **1** (2014) 293–314.
[16] J. Fan, Y. Liao and J. Yao, Power enhancement in high-dimensional cross-sectional tests, *Econometrica* **83** (2015) 1497–1541.
[17] L. Feng, C. Zou and Z. Wang, Multivariate-sign-based high-dimensional tests for the two-sample location problem, *J. Amer. Statist. Assoc.* **111** (2016) 721–735.
[18] L. Feng, C. Zou, Z. Wang and L. Zhu, Two-sample Behrens–Fisher problem for high-dimensional data, *Statist. Sin.* **25** (2015) 1297–1312.
[19] P. Fryzlewicz, Wild binary segmentation for multiple change-point detection, *Ann. Statist.* **42** (2014) 2243–2281.
[20] P. Hall and C. C. Heyde, *Martingale Limit Theory and its Applications* (Academic Press, 1980).
[21] N. Hao, Y. Niu and H. Zhang, Multiple change-point detection via a screening and ranking algorithm, *Statist. Sin.* **23** (2013) 1553–1572.
[22] L. Horváth and M. Hušková, Change-point detection in panel data, *J. Time Ser. Anal.* **33** (2012) 831–648.
[23] V. Jandhyala, S. Fotopoulos, I. MacNeill and P. Liu, Inference for single and multiple change-points in time series, *J. Time Ser. Anal.* **34** (2013) 423–446.
[24] M. Jirak, Uniform change point test in high dimension, *Ann. Statist.* **43** (2015) 2451–2483.
[25] R. Killick, P. Fearnhead and I. A. Eckley, Optimal detection of changepoints with a linear computational cost, *J. Amer. Statist. Assoc.* **107** (2012) 1590–1598.
[26] M. Lavielle, Using penalized contrasts for the change-point problem, *Signal Process.* **85** (2005) 1501–1510.
[27] D. S. Matteson and N. A. James, A nonparametric approach for multiple change point analysis of multivariate data, *J. Amer. Statisit. Assoc.* **109** (2014) 334–345.
[28] Y. Mei, Efficient scalable schemes for monitoring a large number of data streams, *Biometrika* **97** (2010) 419–433.
[29] Y. S. Niu, N. Hao and B. Dong, A new reduced-rank linear discriminant analysis method and its applications, *Statist. Sin.* **28** (2018) 189–202.
[30] Y. S. Niu, N. Hao and H. Zhang, Multiple change-point detection: A selective overview, *Statist. Sci.* **31** (2016) 611–623.
[31] Y. S. Niu and H. Zhang, The screening and ranking algorithm to detect DNA copy number variations, *Ann. Appl. Statist.* **6** (2012) 1306–1326.
[32] A. Onatski, Detection of weak signals in high-dimensional complex-valued data, *Random Matrices*: *Theory Appl.* **3** (2014) 1450001.
[33] M. Rudelson and R. Vershynin, Hanson–Wright inequality and sub-Gaussian concentration, *Electron. Commun. Probab.* **18** (2013) 1–9.
[34] M. S. Srivastava and K. J. Worsley, Likelihood ratio tests for a change in the multivariate normal mean, *J. Amer. Statist. Assoc.* **81** (1986) 199–204.

[35] E. S. Venkatraman, Consistency results in multiple change-point situations, Unpublished Ph.D. Thesis, Department of Statistics, Stanford University (1992).
[36] Y. Xie and D. Siegmund, Sequential multi-sensor change-point detection, Information Theory and Applications Workshop, IEEE (2013), pp. 670–692.
[37] Y. C. Yao, Estimating the number of change-points via Schwarz' criterion, *Statist. Probab. Lett.* **6** (1988) 181–189.
[38] C. Zou, Z. Wang, W. Jiang and X. Zi, An efficient online monitoring method for high-dimensional data streams, *Technometrics* **57** (2015) 374–387.
[39] C. Zou, G. Yin, L. Feng and Z. Wang, Nonparametric maximum likelihood approach to multiple change-point problems, *Ann. Statist.* **42** (2014) 970–1002.