Optimal Covariance Change Point Detection in High Dimension

Daren Wang[1], Yi Yu[2], and Alessandro Rinaldo[1]

[1]Department of Statistics and Data Science, Carnegie Mellon University
[2]School of Mathematics, University of Bristol

Abstract

We study the problem of change point detection for covariance matrices in high dimensions. We assume that $\{X_i\}_{i=1,\ldots,n}$ is a sequence of independent, centered p-dimensional sub-Gaussian random vectors is observed whose covariance matrices are piece-wise constant. Our task is to recover with high accuracy the number and locations the of change points, which are unknown. Our generic model setting allows for all the model parameters to change with n, including the dimension p, the minimal spacing between consecutive change points, the magnitude of smallest change and the maximal operator norm of the covariance matrices of the sample points. We introduce two procedures, one based on the binary segmentation algorithm (e.g. Vostrikova, 1981) and the other on its extension known as wild binary segmentation of Fryzlewicz (2014), and demonstrate that, under suitable conditions, both are able to consistently estimate the number and locations of change points. Our second algorithm, called Wild Binary Segmentation through Independent Projection (WBSIP) is shown to be minimax optimal in the in terms of all the relevant parameters. Our minimax analysis also reveals a phase transition effect based on our generic model setting. To the best of our knowledge, this type of results has not been established elsewhere in the change point detection literature.

Keywords: Change point detection; High dimensional covariance estimation; High dimensional covariance testing; Binary segmentation; Wild binary segmentation; Random projection; Minimax optimal.

1 Introduction

Change point detection in time series data has a long history and can be traced back to World War II, during which the change point detection was specifically demonstrated in a sequential fashion and heavily used in quality control. Wald (1945) was published shortly after the war and formally introduced sequential probability ratio test (SPRT), which consists of simple null and alternative hypotheses and is based on sequentially calculating the cumulative sum of log-likelihood ratios. Page (1954) relaxes the statistic to the cumulative sum of weights, which are not restricted to log-likelihood ratios, namely cumulative sum (CUSUM). Since then, a tremendous amount of efforts have been dedicated to this problem (see, e.g. James et al., 1987; Siegmund and Venkatraman, 1995) with applications ranging from clinical trials to education testing, signal processing and cyber security.

In the last few decades, with the advancement of technology, an increasing demand from emerging application areas, e.g. finance, genetics, neuroscience, climatology, has fostered the development of statistical theorties and methods for change point detection also in an *offline* fashion. Given a time series data set, instead of assuming it is stationary over the whole time course, it is more robust and realistic to assume that it is stationary only within time segments, and the underlying models may change at certain time points. Formally, we assume that we observed n variables $X_1 \ldots, X_n$ such that,

$$X_i \sim \begin{cases} F_1, & i = 1, \ldots, t_1 - 1, \\ F_2, & i = t_1, \ldots, t_2 - 1, \\ \ldots, & \\ F_K, & i = t_{K-1}, \ldots, n, \end{cases} \tag{1}$$

where $F_1, \ldots, F_K$, $K \geq 2$, are distribution functions, $F_{k-1} \neq F_k$, $k = 2, \ldots, K$, and $\{t_1, \ldots t_{K-1}\} \subset \{2, \ldots, n-1\}$ are unknown change point locations. Testing the existence of change points and estimating the locations of the change points are of primary interest.

The simplest and best studied scenario of model (1) is that $\{X_i = f_i + \varepsilon_i\}_{i=1}^n$ is a univariate time series, and $\{f_i\}_{i=1}^n$ is piecewise constant signal. Change point detection here means detecting changes in the mean of a univariate time series. An abundance of literature exists on this model and variants thereof. See Section 1.1 below for a literature review.

The properties of the covariance are also of key importance in statistics, and detecting the covariance changes in a time series data set also has a long history. See Section 1.1 below for some literature review. In this paper we investigate the problem of change point detection and localization for covariance matrices. We consider what is arguably the most basic setting for the problem in (1): we observed a sequence $(X_1 \ldots, X_n)$ of independent and centered random vectors in $\mathbb{R}^d$ wih respective covariance matrices $(\Sigma_1, \ldots, \Sigma_n)$ with that $\max_i \|\Sigma_i\|_{\mathrm{op}} \leq B$. We assume that $\Sigma_i = \Sigma_{i+1}$ for all coordinates i except for a few, which we refer to as the change points. We denote with Δ the minimal distance between two consecutive change points and with κ the magnitude of the smallest change, measured by the operator norm of the difference between the covariance matrices at two consecutive change points. The parameters p, Δ, B and κ completely characterize the difficulty of change point localization problem, which, for a given n, is increasing in p and B and decreasing in Δ and κ. In fact, for reason that will become clear later on, it will be convenient to aggregate the parameter B and κ into one parameter

$$\frac{\kappa}{B^2},$$

which effectively plays the role of the signal to noise ratio. Our goal to identify all the change points and to localize them, i.e. estimate their values accurately. In our analysis, we allow all the relevant parameters p, Δ and $\frac{\kappa}{B^2}$ to change with n and derive upper and lower bounds for the localization rate in terms of such parameters.

We propse two algorithms for covariance change point localization. The first one, called BSOP (Binary Segmentation in Operator Norm), is based on an adaptation to the covariance setting of the popular binary segmentaiton (BS) algorithm (e.g. Vostrikova, 1981) for change point detection of a univariate signal. The BSOP algorithm, described below in Algorithm 1, is rather simple to implement. Under appropriate assumptions we show in Theorem 1 that BSOP can consistently estimate all the change point, with a localization rate that is sub-optimal, especially, when the dimension p is allowed to be increasing with n. Our second algorithm, called WBSIP (Wilde Binary Segmentation through Independent Projections) is significantly more refined, as it combines data splitting, random projections and the wild binary segmentation (WBS) procedure of Fryzlewicz (2014); see Algorithm 3. For all its complexity, WBSIP yields much sharper localization rates than WBOP under a different, milder set of assumptions; see Theorem 2. We further obtain lower bounds on the localization rate and demonstrate that WBSIP is in fact minimax rate optimal. Interestingly, our analysis reveals a phase transition for the problem of covariance change point localization: if $\Delta = \Theta\left(B^4 p\kappa^{-2}\right)$ then no estimator will be able to estimate the location of the change point at a vanishing rate; see Lemma 3. On the other hand, if $\Delta = \Omega\left(B^4 p\kappa^{-2} \log n\right)$ then WBSIP will guarantee a localization rate of the order $B^4\kappa^{-2}\log n$, which, up to a $\log n$ term, is a minimax optimal rate; see Lemma 4. It appears that this type of phase transition is unique to our generic model setting. In fact, many authors (e.g. Aue et al., 2009; Avanesov and Buzun, 2016) have already studied the consistency of change point estimation in covariate structure. To the best of our knowledge, this phase transition based on the spacing parameter Δ and the dimension p has not been established elsewhere. Overall, our results provide a complete characterization of the problem of change point localization in the covariance setting described above.

1.1 Relevant and Related Literature

For the classical problem of change point detection and localization of the mean of a univariate time series, least squares estimation is a natural choice. For example, Yao and Au (1989) uses the least squares estimators to show that the distance between the estimated change points and the truth are within $O_p(1)$. Lavielle (1999) and Lavielle and Moulines (2000) consider penalized least-squares estimation and prove consistency in the presence of strong mixing and long-range dependence of the error terms ε_i's. Harchaoui and Lévy-Leduc (2010) consider the least squares criterion with a total variation.

Besides least squares estimation based methods, other attempts have also been made. For instance, Davis et al. (2006) propose a model based criterion on an autoregressive model. Wang (1995) detects the change points using fast discrete wavelet transform. Harchaoui and Lvy-Leduc (2010), Qian and Jia (2012), Rojas and Wahlberg (2014) and Lin et al. (2017) consider the fused lasso (Tibshirani et al., 2005) to penalise the differences between the signals at two consecutive points to detect the change points. Frick et al. (2014) introduce a multiscale statistic based procedure, namely simultaneous multiscale change point estimator (SMUCE), to detect change points in piecewise constant signals with errors from exponential families. More recently, Li et al. (2017) extends the multiscale method to a class of misspecified signal functions which are not necessarily piecewise constants. The method of Davies and Kovac (2001) can also be used to determine the location of change points in a piece-wise constant signal. Chan and Walther (2013) study the power of the likelihood ratio test statistics to detect the presence of change points.

Among all the methods, binary segmentation (BS) (e.g. Vostrikova, 1981) is 'arguably the most widely used change-point search method' (Killick et al., 2012). It goes through the whole time course and searches for a change point. If a change point is detected, then the whole time course is split into two, and the same procedure is conducted separately on the data sets before and after the detected change point. The procedure is carried on until no change point is detected, or the remaining time course consists of too few time points to continue the test. Venkatraman (1992) proves the consistency results of the BS method in univariate time series mean change point detection, with the number of change points allowed to increase with the number of time points.

It is worth to mention a variant of BS, namely wild binary segmentation (WBS), which is proposed in Fryzlewicz (2014) and which can be viewed as a flexible moving window techniques, or a hybrid of moving window and BS. Instead of starting with the whole data set and doing binary segmentation, WBS randomly draws a collection of intervals under certain conditions, conducts

BS on each interval, and return the one which has the most extreme criterion value among all the intervals. Compared to BS, under certain conditions, WBS is more preferable when multiple change points are present. In the univariate time series mean change point detection problem, Venkatraman (1992) shows that in order to achieve the estimating consistency using BS algorithm, the minimum gap between two consecutive change points should be at least of order $n^{1-\beta}$, where n is the number of time points, and $0 \leq \beta < 1/8$; as claimed in Fryzlewicz (2014), by using WBS algorithm, this rate can be reduced to $\log(n)$.

All the literature mentioned above tackles univariate time series models, however, in the big data era, data sets are now routinely more complex and often appear to be multi- or high-dimensional, i.e. $X_i \in \mathbb{R}^p$, where p is allowed to grow with the number of data points n. Horváth and Hušková (2012) propose a variant of the CUSUM statistic by summing up the square of the CUSUM statistic in each coordinate. Cho and Fryzlewicz (2012) transform a univariate non-stationary time series into multi-scale wavelet regime, and conduct BS at each scale in the wavelet context. Jirak (2015) allows p to tend to infinity together with n, by taking maxima statistics across panels coordinate-wise. Cho and Fryzlewicz (2015) propose sparsified binary segmentation method which aggregates the CUSUM statistics across the panel by adding those which exceed a certain threshold. Cho (2015) proposes the double CUSUM statistics which, at each time point, picks the coordinate which maximizes the CUSUM statistic, and *de facto* transfers the high-dimensional data to a univariate CUSUM statistics sequence. Aston and Kirch (2014) introduces the asymptotic concept of high-dimensional efficiency which quantifies the detection power of different statistics in this setting. Wang and Samworth (2016) study the problem of estimating the location of the change points of a multivariate piecewise-constant vector-valued function under appropriate sparsity assumptions on the number of changes.

As for change point detection in more general scenarios, the SPRT procedure (Wald, 1945) can be easily used for the variance change point detection. Based on a generalized likelihood ratio statistic, Baranowski et al. (2016) tackle a range of univariate time series change point scenarios, including the variance change situations, although theoretical results are missing. Picard (1985) proposes tests on the existence of change points in terms of spectrum and variance. Inclan and Tiao (1994) develop an iterative cumulative sums of squares algorithm to detect the variance changes. Gombay et al. (1996) propose some tests on detection of possible changes in the variance of independent observations and obtained the asymptotic properties under the non-change null hypothesis. Berkes et al. (2009), among others, extend the tests and corresponding results to linear processes, as well as ARCH and GARCH processes. Aue et al. (2009) considers the problem of variance change point detection in a multivariate time series model, allowing the observations to have m-dependent structures. Note that the consistency results in Aue et al. (2009) are in the asymptotic sense that the number of time points diverges and the dimension of the time series remains fixed. Aue et al. (2009) also require the existence of good estimators of the covariance and precision matrices, and the conditions thereof are left implicit. Barigozzi et al. (2016) deal with a factor model, which is potentially of high dimension $p/n = O(\log^2(n))$, and use the wavelet transforms to make the data possibly dependent across the timeline. Note that the model in Barigozzi et al. (2016) can be viewed as a specific covariance change point problem, where the additional structural assumption allows the dimensionality to go beyond the sample size.

As for the problem of hypothesis testing for high dimensional covariance matrices, which corresponds to the problem of change point detection, the literature is also abundant. In the cases where p fixed and $n \to \infty$, the likelihood ratio test (Anderson, 2003) has a $\chi^2_{p(p+1)/2}$ limiting distribution under $H_0 : \Sigma = I$. When both $n, p \to \infty$ and $p/n \to c \in (0, \infty)$, Johnstone (2001) extended Roy's largest root test (Roy, 1957) and derived the Tracy–Widom limit of its null distribution, to name but a few. In the case where $n, p \to \infty$ and $p/n \to \infty$, Birke and Dette (2005) derived the asymptotic null distribution of the Ledoit–Wolf test (Ledoit and Wolf, 2002). More recently, Cai and Ma (2013) studied the testing problem in the setting $p \to \infty$ from a minimax point of view and derived the testable region in terms of Frobenius norm is of order $\sqrt{p/n}$.

Organization of the paper

The rest of the paper is organized as follows. In Section 2, we propose two different methods for covariance change point detection problem. All three methods are shown to be consistent, but under different conditions, and the wild binary segmentation with random projection has the location error rate being $\log(n)$. In Section 3, we show the lower bound of the location error rate in the covariance change point detection problem is $\log(n)$, which implies that the wild binary segmentation with random projection is minimax optimal. Further discussion and future work directions can be found in Section 4.

Notation

For a vector $v \in \mathbb{R}^p$ and matrix $\Sigma \in \mathbb{R}^{p \times p}$, $\|v\|$ and $\|\Sigma\|_{\mathrm{op}} = \max_{\|v\|=1} |v^\top \Sigma v|$ indicate the Euclidean and the operator norm, respectively; for any integrable function $f(\cdot) : \mathbb{R} \mapsto \mathbb{R}$, denote $\|f\|_1 = \int_{x \in \mathbb{R}} |f(x)| dx$ as the ℓ_1-norm of $f(\cdot)$. For functions $f(\cdot)$ and $g(\cdot)$ satisfying $f(n)/g(n) \to a$, if $a \in (0, \infty)$, we denote by $f(n) = O(g(n))$.

2 Main Results

In this paper, we study the covariance change point detection in high dimension. To be specific, we consider a centered and independent time series $\{X_i\}_{i=1}^n \subset \mathbb{R}^{p \otimes n}$. Let $\{\eta_k\}_{k=1}^K \subset \{1, \dots, n\}$ be the collection of time points at which the covariance matrices of X_i's change. The model is formally summarized as follows with Orlicz norm $\|\cdot\|_{\psi_2}$, which is defined in Definition 2 in Appendix A.

Assumption 1. *Let $X_1, \dots, X_n \in \mathbb{R}^p$ be independent sub-Gaussian random vectors such that $E(X_i) = 0$, $E(X_i X_i') = \Sigma_i$ and $\|X_i\|_{\psi_2} \le B < \infty$ for all i. Let $\{\eta_k\}_{k=0}^{K+1} \subset \{0, \dots, n\}$ be a collection of change points, such that $\eta_0 = 0$ and $\eta_{K+1} = n$ and that*

$$\Sigma_{\eta_k+1} = \Sigma_{\eta_k+2} = \ldots = \Sigma_{\eta_{k+1}}, \text{ for any } k = 0, \dots, K.$$

Assume the jump size $\kappa = \kappa(n)$ and the spacing $\Delta = \Delta(n)$ satisfy that

$$\inf_{k=1,\dots,K+1} \{\eta_k - \eta_{k-1}\} \ge \Delta > 0,$$

and

$$\|\Sigma_{\eta_k} - \Sigma_{\eta_{k-1}}\|_{\mathrm{op}} = \kappa_k \ge \kappa > 0, \text{ for any } k = 1, \dots, K+1.$$

Remark 1. The parameter B. *The assumption that $\max_i \|X_i\|_{\psi_2} \le B$ is imposed in order to control the order of magnitude of the random fluctuations of the CUSUM covariance statistics, given below in Definition 1, around its means. See Lemma 5 in the Appendix for details. At the same* *time, the chain of inequalities* (2) *shows also that* $\max_i \|\Sigma_i\|_{\mathrm{op}} \leq 2B^2$*, so that the same assumptions amounts also to a uniform upper bound on the operator norm of the covariance matrices of the date points. In this regard,* B *may be reminiscent of the assumption that the signal be of bounded magnitude sometimes used in the litarature on mean change point detection: see, e.g. Fryzlewicz (2014, Assumption 3.1(ii)) and Venkatraman (1992, Condition (iii) on page 11). However, while in the mean change point detection problem such boundedness assumption can be in fact removed because the optimal solution is translation invariant (see, e.g., Wang and Samworth, 2016), this is not the case in the present setting. Indeed, a constant shift in the largest eigenvalue of the population covariance matrices also implies a change in the precision with which such matrices can be estimated. Thus, it is helpful to think of* B^2 *as some form of variance term.*

Remark 2. Relationship between κ **and** B**.** *The parameters* κ *and* B *are not variation independent, as they satisfy the inequality* $\kappa \leq B^2/4$*. In fact,*

$$\kappa \leq \max_{k=1}^{K} \|\Sigma_{\eta_k} - \Sigma_{\eta_{k-1}}\|_{\mathrm{op}} \leq 2 \max_{i=1}^{n} \|\Sigma_i\|_{\mathrm{op}} = 2 \max_{i=1}^{n} \sup_{v \in \mathcal{S}^{p-1}} \mathbb{E}\big[(v^\top X_i)^2\big] \leq 4 \max_{i=1}^{n} \|X_i\|_{\psi_2}^2 \leq 4B^2, \quad (2)$$

where the second-to-last inequality follows from Equation (16) *in Appendix A. For ease of readability, we will instead use the weaker bound* $\kappa \leq B^2$ *throughout. In fact, in our analysis we will quantify the combined effect of both* κ *and* B *with their ratio* $\frac{\kappa}{B^2}$*, which we will refer to as the signal-to-noise ratio. Larger values of such quantities lead to better performance of our algorithm. It is important to notice that the signal-to-noise ratio, and the task of change-point detection itself, remains invariant with respect to any multiplictive rescaling of the data by an arbitsraty non-zero constant.*

In our model Assumption 1, all the relevant parameters p, Δ, K, B and κ are allowed to be functions of the sample size n, although we do make this dependence explicit in our notation for ease of readability. This generic setting allows us to study the covariance change point problem with potential high dimensional data and with growing number of change points and decreasing jump sizes.

Motivated by the univariate CUSUM statistic for mean change point detection, we define the CUSUM statistic in the covariance context.

Definition 1 (Covariance CUSUM)**.** *For* $X_1, \ldots, X_n \in \mathbb{R}^p$*, a pair of integers* (s, e) *such that* $0 \leq s < e-1 < n$*, and any* $t \in \{s+1, \ldots, e-1\}$*, the covariance CUSUM statistic is defined as*

$$\widetilde{S}_t^{s,e} = \sqrt{\frac{e-t}{(e-s)(t-s)}} \sum_{i=s+1}^{t} X_i X_i^\top - \sqrt{\frac{t-s}{(e-s)(e-t)}} \sum_{i=t+1}^{e} X_i X_i^\top.$$

Its expected value is

$$\widetilde{\Sigma}_t^{s,e} = \sqrt{\frac{e-t}{(e-s)(t-s)}} \sum_{i=s+1}^{t} \Sigma_i - \sqrt{\frac{t-s}{(e-s)(e-t)}} \sum_{i=t+1}^{e} \Sigma_i.$$

In the rest of this section, we propose two algorithms to detect the covariance change points. Detailed algorithms and the main consistency results are presented in this section, with the proofs provided in the Appendices. The advantages of each algorithm will be discussed later in the section.

2.1 Consistency of the BSOP algorithm

We begin our study by analyzing the performance of a direct adaptation of the binary segmentation (BS) algorithm to the matrix setting based on the distance induced by the operator norm. The resulting algorithm, which we call BSOP, is given in Algorithm 1. The BSOP procedure works as follows: given any time interval (s, e), BSOP first computes the maximal operator norm of the covariance CUSUM statistics over the time points in $(s + \lceil p \log n \rceil, e - \lfloor p \log n \rfloor)$; if such maximal value exceeds a predetermined threshold τ, then BSOP will identify the location b of the maximum as a change point. The interval (s, e) is then split into two subintervals at b and the procedure is then iterated separately on each of the resulting sub-intervals (s, b) and (b, e) until an appropriate stopping condition is met.

The BSOP algorithm differs from the standard BS implementation in one aspect: the maximization of the of the operator norm of the CUSUM covariance operator is carried out only over the time points in (s, e) that are away by at least $p \log n$ from the endpoints of the interval. Such restriction is needed to obtain adequate tail bounds for the operator norm of the covariance CUSUM statistics $\tilde{S}_t^{s,e}$ given in Definition 1 and of the centered and weighted empirical covariance matrices. See Lemma 5 in Appendix A.

Algorithm 1 Binary Segmentation through Operator Norm. BSOP$((s, e), \tau)$

INPUT: $\{X_i\}_{i=s+1}^{e} \subset \mathbb{R}^{p \otimes (e-s)}$, $\tau > 0$.
  Initial FLAG $\leftarrow 0$,
  while $e - s > 2p\log(n) + 1$ and FLAG $= 0$ **do**
    $a \leftarrow \max_{\lceil s+p\log(n)\rceil \leq t \leq \lfloor e-p\log(n)\rfloor} \|\widetilde{S}_t^{s,e}\|_{\mathrm{op}}$
    if $a \leq \tau$ **then**
      FLAG $\leftarrow 1$
    else
      $b \leftarrow \arg\max_{\lceil s+p\log(n)\rceil \leq t \leq \lfloor e-p\log(n)\rfloor} \|\widetilde{S}_t^{s,e}\|_{\mathrm{op}}$
      add b to the collection of estimated change points
      BSOP$((s, b-1), \tau)$
      BSOP$((b, e), \tau)$
    end if
  end while
OUTPUT: The collection of estimated change points.

Remark 3 (Uniqueness of the maxima of the covariance CUSUM statistic.)**.** *The probability that the maximiser of* $\|\widetilde{S}_t^{s,e}\|_{\mathrm{op}}$ *over multiple time points is non unique is zero.*

To analyze the performance of the BSOP algorithm we will impose the following assumption, which is, for the most part, modeled after Assumption 3.2 in Fryzlewicz (2014).

Assumption 2. *For a sufficiently large constant* $C_\alpha > 0$ *and sufficient small constant* $c_\alpha, c_\kappa > 0$*, assume that* $\Delta\kappa B^{-2} \geq C_\alpha n^{\Theta}$*,* $p \leq c_\alpha n^{8\Theta-7}/\log(n)$*, where* $\Theta \in (7/8, 1]$*.*

When the parameters κ and B are fixed, the above assumption requires Δ, the minimal spacing between consecutive change point, to be of slightly smaller order than n, the size of the time series. This is precisely Assumption 3.3 in Fryzlewicz (2014) (see also Cho and Fryzlewicz, 2015). The fact that Δ cannot be too small compared to n in order for the BS algorithm to exhibit good performance is well known: see, e.g., Olshen et al. (2004). In Assumption 2, we require also the dimension p to be upper bounded by $\frac{n^{8\Theta-7}}{\log n}$, which means that p is allowed to diverge as $n \to \infty$.

Remark 4 (Generalizing Assumption 2)**.** *In Assumption 2 we impose certain constraints on the scaling of the quantities B, κ, Δ and p in relation to n that are captured by a single parameter Θ, whose admissible values lye in $(7/8, 1]$. The strict lower bound of $7/8$ on Θ is determined by the calculations outlined below on (5) and (6), which are needed to ensure the existence of a non-empty range of value for the input parameter τ to the BSOP algorithm. In fact, Assumption 2 may be generalized by by allowing for different types of scaling in n of the signal-to-noise ratio κB^{-2}, the minimal distance Δ between consecutive change points and the dimension p. In detail, we may require that $\kappa B^{-2} \succeq n^{\Theta_1}$, $\Delta \succeq n^{\Theta_2}$ and $p \log n \preceq n^{\Theta_3}$ for a given triplet of parameters $(\Theta_1, \Theta_2, \Theta_3)$ in an appropriate subset of $[0, 1]^3$. Such a generalization would then lead to consistency rates in n that depend on all these parameters simultaneously. However, the range of allowable values of $(\Theta_1, \Theta_2, \Theta_3)$ is not a product set due to non-trivial constraints among them. We will refrain from providing details and instead rely on the simpler formulation given in Assumption 2.*

Theorem 1 (Consistency of BSOP)**.** *Under Assumptions 1 and 2, let $\mathcal{B} = \{\hat{\eta}_k\}_{k=1}^{\widehat{K}}$ be the collection of the estimated change points from the* $\mathrm{BSOP}((0, n), \tau)$ *algorithm, where the parameter τ satisfies*

$$B^2\sqrt{p\log(n)} + 2\sqrt{\epsilon_n}B^2 < \tau < C_1\kappa\Delta(e-s)^{-1/2}, \tag{3}$$

for some constant $C_1 \in (0, 1)$ and where

$$\epsilon_n = C_2 B^2 \kappa^{-1} n^{5/2} \Delta^{-2} \sqrt{p\log(n)}$$

for some $C_2 > 0$. Then,

$$\mathbb{P}\Big(\widehat{K} = K \quad \text{and} \quad \max_{k=1,\ldots,K} |\eta_k - \hat{\eta}_k| \leq \epsilon_n\Big) \geq 1 - n^3 9^p 2n^{-cp}, \tag{4}$$

for some absolute constant $c > 0$.

Remark 5. *The condition (3) on the admissible values of the input parameter τ to the BSOP algorithm is well defined. Indeed, by Assumption 2, for all pairs (s, e) such that $e - s > 2p\log n$, we have that*

$$B^2\sqrt{p\log(n)} \leq B^2 c_\alpha n^{4\Theta-7/2} \leq B^2 c_\alpha n^{\Theta} n^{-1/2} \leq \frac{c_\alpha}{C_\alpha}\kappa\Delta n^{-1/2} \leq (1/8)\kappa\Delta(e-s)^{-1/2} \tag{5}$$

and

$$\begin{aligned} 2\sqrt{\epsilon_n}B^2 &= 2C_2^{1/2}B^3\kappa^{-1/2}n^{5/4}\Delta^{-1}(p\log(n))^{1/4} \leq (2C_2^{1/2}C_\alpha^{-1}c_\alpha^{1/4})B\kappa^{1/2}n^{5/4+\Theta-7/4} \\ &\leq (C_1/8)\kappa\Delta n^{-1/2}B^{-1}\kappa^{1/2} \leq (C_1/8)\kappa\Delta(e-s)^{-1/2}, \end{aligned} \tag{6}$$

where in the chain of inequalities we have used Assumption 2 repeatedly. It is also worth noting that the difference between the right hand side and the left hand side of (3) increase as Θ increases to 1. Finally we remark that in the proof of Theorem 1, we actually let $C_1 = 1/8$, but this is an arbitrary choice and it essentially depends on the constants C_α and c_α from Assumption 2.

Proof of Theorem 1. By induction, it suffices to consider any pair of integers s and e such that $(s,e) \subset (0,T)$ and satisfying

$$\eta_{r-1} \leq s \leq \eta_r \leq \ldots \leq \eta_{r+q} \leq e \leq \eta_{r+q+1}, \quad q \geq -1,$$
$$\max\{\min\{\eta_r - s, s - \eta_{r-1}\}, \min\{\eta_{r+q+1} - e, e - \eta_{r+q}\}\} \leq \epsilon_n,$$

where $q = -1$ indicates that there is no change point contained in (s,e). It follows that, for sufficiently small c_α and sufficiently large C_α,

$$\begin{aligned}\frac{\epsilon_n}{\Delta/4} &\leq \frac{C_2 B^2 \kappa^{-1} n^{5/2} \sqrt{p \log(n)} \Delta^{-2}}{\Delta/4} \\ &\leq 4C_2 B^2 \kappa^{-1} n^{5/2} \frac{c_\alpha^{1/2} n^{4\Theta - 7/2}}{C_\alpha^3 \kappa^{-3} B^6 n^{3\Theta}} \\ &\leq (4C_2 c_\alpha^{1/2} C_\alpha^{-3})(\kappa^2 B^{-4}) n^{\Theta-1} \\ &\leq (\kappa^2 B^{-4}) n^{\Theta-1} \\ &\leq 1\end{aligned}$$

where the second inequality stems from Assumption 2, the third inequality holds by choosing sufficiently small c_α and sufficiently large C_α and the last inequality follows from the fact that $\kappa \leq B^2$. Then, for any change point should be (s,e), it is either the case that

$$|\eta_p - s| \leq \epsilon_n,$$

or that

$$|\eta_p - s| \geq \Delta - \epsilon_n \geq \Delta - \Delta/4 = 3\Delta/4.$$

Similar considerations apply to the other endpoint e. As a consequence, the fact that $\min\{|\eta_p - e|, |\eta_p - s|\} \leq \epsilon_n$ implies that η_p is a detected change point found in the previous induction step, while if $\min\{\eta_p - s, \eta_p - e\} \geq 3\Delta/4$ we can conclude that $\eta_p \in (s,e)$ is an undetected change point.

In order to complete the induction step, it suffices to show that $\mathrm{BSOP}([s,e],\tau)$ (i) will not find any new change point in the interval (s,e) if there is none, or if all the change points in (s,e) have been already detected and (ii) will identify a location b such that $|\eta_p - b| \leq \epsilon_n$ if there exists at least one undetected change point in (s,e).

Set $\lambda = B^2 \sqrt{p \log(n)}$. Then, the event $\mathcal{A}_1(\{X_i\}_{i=1}^n, \lambda)$ defined in Equation (17) holds with probbability at least $1 - 2 \times 9^p n^3 n^{cp}$, for some universal constant $c > 0$. The proof will be completed in two steps.

Step 1. First we will show that on the event $\mathcal{A}_1(\{X_i\}_{i=1}^n, \lambda)$, $\mathrm{BSOP}((s,e),\tau)$ can consistently detect or reject the existence of undetected change points within (s,e).

Suppose there exists $\eta_p \in (s,e)$ such that $\min\{\eta_p - s, \eta_p - e\} \geq 3\Delta/4$. Set $\delta = p \log n$. Then $\delta \leq \frac{3}{32}\Delta$, since

$$p \log(n) \leq c_\alpha n^{8\Theta - 7} \leq c_\alpha n^{\Theta} \leq c_\alpha C_\alpha^{-1} \Delta B^{-2} \kappa^1 \leq 3\Delta/32,$$

where the last inequality follows from Assumption 2. With this choice of δ, we apply Lemma 20 in Appendix D (where we set $c_1 = 3/4$) and obtain that

$$\max_{t = \lceil s+\delta \rceil, \ldots, \lfloor e-\delta \rfloor} \|\widetilde{\Sigma}_t^{s,e}\|_{\mathrm{op}} \geq (3/8)\kappa\Delta(e-s)^{-1/2}.$$

On the event $\mathcal{A}_1(\{X_i\}_{i=1}^n, \lambda)$,

$$\max_{t=\lceil s+\delta\rceil,\ldots,\lfloor e-\delta\rfloor} \|\widetilde{S}_t^{s,e}\|_{\mathrm{op}} \geq \max_{t=\lceil s+\delta\rceil,\ldots,\lfloor e-\delta\rfloor} \|\widetilde{\Sigma}_t^{s,e}\|_{\mathrm{op}} - \lambda \geq (3/8)\kappa\Delta(e-s)^{-1/2} - \lambda \geq (1/8)\kappa\Delta(e-s)^{-1/2}, \tag{7}$$

where the last inequality follows from (5) (in the last step we have set $C_1 = 1/8$). If (3) holds, then, on the event $\mathcal{A}_1(\{X_i\}_{i=1}^n, \lambda)$, BSOP$((s,e),\tau)$ detects the existence of undetected change points if there are any.

Next, suppose there does not exist any undetected change point within (s,e). Then, one of the following cases must occur.

(a) There is no change point within (s,e);

(b) there exists only one change point η_r within (s,e) and $\min\{\eta_r - s, e - \eta_r\} \leq \epsilon_n$;

(c) there exist two change points η_r, η_{r+1} within (s,e) and that $\max\{\eta_r - s, e - \eta_{r+1}\} \leq \epsilon_n$.

Observe that if case (a) holds then, on the event $\mathcal{A}_1(\{X_i\}_{i=1}^n, \lambda)$, we have that

$$\max_{t=\lceil s+\delta\rceil,\ldots,\lfloor e-\delta\rfloor} \|\widetilde{S}_t^{s,e}\|_{\mathrm{op}} \leq \max_{t=\lceil s+\delta\rceil,\ldots,\lfloor e-\delta\rfloor} \|\widetilde{\Sigma}_t^{s,e}\|_{\mathrm{op}} + \lambda = \lambda < \tau,$$

where the last inequality follows from (3). If situation (c) holds, then on, the event $\mathcal{A}_1(\{X_i\}_{i=1}^n, \lambda)$, we have

$$\max_{t=\lceil s+\delta\rceil,\ldots,\lfloor e-\delta\rfloor} \|\widetilde{S}_t^{s,e}\|_{\mathrm{op}} \leq \max_{t=\lceil s+\delta\rceil,\ldots,\lfloor e-\delta\rfloor} \|\widetilde{\Sigma}_t^{s,e}\|_{\mathrm{op}} + \lambda \leq \max\{\|\widetilde{\Sigma}_{\eta_r}^{s,e}\|_{\mathrm{op}}, \|\widetilde{\Sigma}_{\eta_{r+1}}^{s,e}\|_{\mathrm{op}}\} + \lambda \leq 2\sqrt{\epsilon_n}B^2 + \lambda,$$

where the first inequality follows from $\mathcal{A}_1(\{X_i\}_{i=1}^n, \lambda)$, the second inequality from Lemma 19 and the third inequality from Lemma 23. (Both Lemmas are in Appendix D.2.) Case (b) can be handled in a similar manner. Thus, if (3) holds, then on the event $\mathcal{A}_1(\{X_i\}_{i=1}^n, \lambda)$, BSOP$((s,e),\tau)$ has no false positives when there are no undetected change points in (s,e).

Step 2. Assume now that there exists a change point $\eta_p \in (s,e)$ such that $\min\{\eta_p - s, \eta_p - e\} \geq 3\Delta/4$ and let

$$b \in \arg\max_{t=\lceil s+\delta\rceil,\ldots,\lfloor e-\delta\rfloor} \left\|\widetilde{S}_t^{s,e}\right\|_{\mathrm{op}}.$$

To complete the proof it suffices to show that $|b - \eta_k| \leq \epsilon_n$.

Let v be such that

$$v \in \arg\max_{\|u\|=1} |u^\top \widetilde{S}_b^{s,e} u|.$$

Consider the univariate time series $\{Y_i(v)\}_{i=1}^n$ and $\{f_i(v)\}_{i=1}^n$ defined in (53) and (54) in Appendix D.2. By Lemma 21 there, $b \in \arg\max_{s \leq t \leq e} |\widetilde{Y}_t(v)|$. Next, we wish to apply Corollary 10 to the time series $\{Y_i(v)\}_{i=s}^e$ and $\{f_i(v)\}_{i=s}^e$. Towards that end, we first need to ensure that the conditions required for that result to hold are verified. (Notice that in the statement of Corollary 10, the f_i's are assumed to be uniformly bounded by B_1, while in this proof the $f_i(v)$'s defined in (54) are assumed to be bounded by $2B^2$.) First, the collection of the change points of the time series $\{f_i(v)\}_{i=s+1}^e$ is a subset of $\{\eta_k\}_{k=0}^{K+1} \cap (s,e)$. The condition (26) and the inequality

$2\sqrt{\delta}B^2 \leq (3c_1/4)\kappa\Delta(e-s)^{-1/2}$ are straightforward consequences of Assumption 2, while (34) follows from the fact that

$$|\widetilde{f}_t^{s,e}(v) - \widetilde{Y}_t^{s,e}(v)| \leq \left\|\widetilde{S}_t^{s,e} - \widetilde{\Sigma}_t^{s,e}\right\|_{\text{op}} \leq \lambda.$$

Similarly, (33) stems from the relationships

$$\max_{t=\lceil s+\delta\rceil,\ldots,\lfloor e-\delta\rfloor} |\widetilde{Y}_t^{s,e}(v)| = \max_{t=\lceil s+\delta\rceil,\ldots,\lfloor e-\delta\rfloor} \|\widetilde{S}_t^{s,e}\|_{\text{op}} \geq \max_{t=\lceil s+\delta\rceil,\ldots,\lfloor e-\delta\rfloor} \|\widetilde{\Sigma}_t^{s,e}\|_{op} - \lambda \geq (1/8)\kappa\Delta(e-s)^{-1/2}$$

where the first inequality holds under the event $\mathcal{A}_1(\{X_i\}_{i=1}^n, \lambda)$ and the second inequality is due to (7) and Assumption 2. Thus, all the assumptions of Corollary 10 are met. An application of that result yields that there exists η_k, a change point of $\{f_i(v)\}_{i=s}^e$ satisfying (29), such that

$$|b - \eta_k| \leq C_2\lambda(e-s)^{5/2}\Delta^{-2}\kappa^{-1} \leq \epsilon_n.$$

The proof is complete by observing that (29) implies $\min\{\eta_k - s, \eta_k - e\} \geq 3\Delta/4$, as discussed in the argument before **Step 1**.

□

The proof of the theorem relies on a non-trivial extension of the arguments for proving consistency of the BS algorithm in one-dimensional mean change-point detection problems, as done in Venkatraman (1992) (see also Fryzlewicz, 2014). The main difficulty that prevents a direct application of those results is the fact the regions of monotonicity of the function $t \mapsto \left\|\widetilde{\Sigma}_t^{s,e}\right\|_{\text{op}}$ are hard to derive. Instead, for each par of integers $1 \leq s < e \leq n$ with $e - s > 2p\log n$, we study the one-dimensinal time series $\{(v^\top X_i)^2\}_{i=1,\ldots,n}$ of the squared coefficients of the projeciton of the data along a one-dimensional linear subspace spanned by a distinguished unit vector v, which we term the *shadow vector*. This is simply the leading singular vector of $\left\|\widetilde{\Sigma}_b^{s,e}\right\|_{\text{op}}$, where $b = \arg\max_{t\in(s+p\log n, e-p\log n)}\left\|\widetilde{\Sigma}_t^{s,e}\right\|_{\text{op}}$. As it turns out, with such a choice of the shadow vector, the local maxima of CUSUM statistic applied to the corresponding one-dimensional time series coincide with the local maxima of the time series of the values of the operator norm of the CUSUM covarance statistics. As a result, for the purpose of detecting local maxima of the CUSUM covariance statistic, it is enough and in fact much simpler to study the univariate times series of the squared projections onto the appropriate ghost vector. Of course, at each iteration of the BSOP algorithm a new ghost vector and a new univarite time series is obtained and a new local maximum is found. See Appendix D.2 for further comments on the uses and intepretation of the shadow vector.

Theorem 1 implies, that with high probability, the BSOP algorithm will identify all the change points and estimate their locations with an error that is bounded by

$$\epsilon_n \preceq \frac{B^2}{\kappa}\Delta^{-2}n^{5/2}\sqrt{p\log n}.$$

Noitce that, as expected, the performance of BSOP is decreasing in the inverse of the signal-to-noise ratio parmeter $\frac{\kappa}{B^2}$, the inverse of the minima distance Δ between change points and the dimension p. The above bound yields a family of rates of consistency for BSOP, depending on the scaling of each of the quantities involved in it. For example, in the simplest and most favorable scenario whereby B, κ and the dimension p are constants, the bound implies a rate for change point localization of the order

$$\frac{\epsilon_n}{n} \preceq n^{-2\Theta+3/2}\sqrt{\log n},$$

which is decreasing in the $\Theta \in (7/8, 1]$. In particular, when the number of change points is also kept constant, we have that $\Theta = 1$, yielding a localization rate of order $\sqrt{\frac{\log n}{n}}$.

As we will see in the next section, the dependence on the parameters B, κ and Δ is suboptimal. The advantage of BSOP over the rate-optimal algorithm we introduce next is that BSOP only requires one input parameter, the threshold value τ. Furthermore, when the spacing parameter Δ is comparable with n and the dimension p of the data grows slowly with respect with n, then BSOP can still deliver good consistency rates. Therefore, despite its suboptimality in general, BSOP is a simple and convenient algorithm which may serve as a competitive benchmark for other proceudres.

2.2 Consistency of the WBSIP algorithm

In this section we describe and analyze the performance of a new algorithm for covariance change-point detection, which we term WBSIP for Wild Binary Segmentation through Independent Projections. The WBSIP algorithm is a generalization of the wild binary segmentation or WBS procedure of Fryzlewicz (2014) for mean change-point detection and further exploits the properties shadow vectors. The WBSIP procedure begins by splitting the data into halves and by selecting at random a collection of M pairs of integers (s, e) such that $1 \leq s < e \leq n$ and $e - s > p\log n + 1$. In its second step, WBSIP computes, for each of the M random integer intervals previously generated, a shadow vector using one half od the data and its corresponding one-dimensional time series using the other half. The final step of the procedure is to apply the WBS algorithm over the resulting univariate time series. The details of the algorithm are given in Algorithm 2, which describes the computation of the shadow vectors by principal component methods, and Algorithm 3, which implements WBS to the resulting one dimensional time series.

Remark 6. *The idea of combining the WBS algorithm with sample splitting is due to Wang and Samworth (2016), who applied it to the problem of mean change-point detection in multivariate settings. Like us, Wang and Samworth (2016) employ one half of the sample to compute an optimal direction and then project the second half of the data along such direction to obtain a univariate time series which can then be handled using WBS. One of the main differences between our approach and that of Wang and Samworth (2016) is the fact that we do no require the shadow vectors to be consistent estimators of subspaces related the true covariance matrices. In particular, our analysis holds without any eigengap assumptions.*

In order to analyze the performance of the WBSIP procedure, we will impose the following assumption, which is significantly weaker then Assumption 2.

Assumption 3. *There exists a sufficiently large absolute constant $C > 0$ such that $\Delta\kappa^2 \geq Cp\log(n)B^4$.*

Remark 7. *We recall that all the parameters Δ, κ, p, B are allowed to depend on n. Since $\kappa \leq B^2$, and assuming without loss of generality that the constant C in the previous assumption is larger than 8, we further have that*

$$p\log(n) \leq \Delta\kappa^2 B^{-4}C^{-1} \leq \Delta/8,$$

which is used repeatedly below. In fact, in the proof we will set $C = 32\sqrt{2}$. This choice is of course arbitrary and is only made for convenience in carrying out the calculations below. We also recall that the ratio $B^4\kappa^{-2}$ is invariant of multiplicative scaling, i.e. if $X'_i = \alpha X_i$ for all i, then the corresponding ratio $B^4\kappa^{-2}$ stays the same.

Algorithm 2 Principal Component Estimation $\mathrm{PC}(\{X_i\}_{i=1}^n, \{(\alpha_m, \beta_m)\}_{m=1}^M)$

INPUT: $\{X_i\}_{i=1}^n$, $\{(\alpha_m, \beta_m)\}_{m=1}^M$
  for $m = 1, \ldots, M$ **do**
    if $\beta_m - \alpha_m > 2p\log(n) + 1$ **then**
      $d_m \leftarrow \arg\max_{\lceil \alpha_m + p\log(n)\rceil \le t \le \lfloor \beta_m - p\log(n)\rfloor} \|\widetilde{S}_t^{\alpha_m,\beta_m}\|_{\mathrm{op}}$
      $u_m \leftarrow \arg\max_{\|v\|_2=1} |v^\top \widetilde{S}_{d_m}^{\alpha_m,\beta_m} v|$
    else
      $u_m \leftarrow 0$
    end if
  end for
OUTPUT: $\{u_m\}_{m=1}^M$.

Algorithm 3 Wild Binary Segmentation through Independent Projection. $\mathrm{WBSIP}((s,e), \{(\alpha_m, \beta_m)\}_{m=1}^M, \tau, \delta)$

INPUT: Two independent samples $\{W_i\}_{i=1}^n$, $\{X_i\}_{i=1}^n$, τ, δ.
  $\{u_m\}_{m=1}^M = PC(\{W_i\}_{i=1}^n, \{(\alpha_m, \beta_m)\}_{m=1}^M)$
  for $i \in \{s, \ldots, e\}$ **do**
    for $m = 1, \ldots, M$ **do**
      $Y_i(u_m) \leftarrow \left(u_m^\top X_i\right)^2$
    end for
  end for
  for $m = 1, \ldots, M$ **do**
    $(s'_m, e'_m) \leftarrow [s,e] \cap [\alpha_m, \beta_m]$ and $(s_m, e_m) \leftarrow (\lceil s'_m + \delta\rceil, \lfloor e'_m - \delta\rfloor)$
    if $e_m - s_m \ge 2\log(n) + 1$ **then**
      $b_m \leftarrow \arg\max_{s_m + \log(n) \le t \le e_m - \log(n)} |\widetilde{Y}_t^{s_m,e_m}(u_m)|$
      $a_m \leftarrow |\widetilde{Y}_{b_m}^{s_m,e_m}(u_m)|$
    else
      $a_m \leftarrow -1$
    end if
  end for
  $m^* \leftarrow \arg\max_{m=1,\ldots,M} a_m$
  if $a_{m^*} > \tau$ **then**
    add b_{m^*} to the set of estimated change points
    $\mathrm{WBSRP}((s, b_{m*}), \{(\alpha_m, \beta_m)\}_{m=1}^M, \tau)$
    $\mathrm{WBSRP}\ ((b_{m*}+1, e), \{(\alpha_m, \beta_m)\}_{m=1}^M, \tau)$
  end if
OUTPUT: The set of estimated change points.

Similarly to the BSOP algoithm, WBSIP also applies a slight modification to the WBS algorithm as originally proposed in Fryzlewicz (2014). When computing the shadow vectors in Algorithm 2, the search for the optimal direction onto which to project the data is restricted, for any given candidate interval, only to the time points that are at least $p \log n$ away from the endpoints of the interval. As remarked in the previous section, this ensures good tail bounds for the operator norms of the matrices involved.

A second, more substantial, adaptation of WBS is used in Algorithm 3: when searching for candidate change points inside a given interval, the algorithm only consider time points that are δ-away from the endpoints of the interval, where δ is an upper bound on the localization error – the term ϵ_n in Theorem 2 below. The reasons for such restriction are somewhat subtle: once an estimated change point is found near a true change point, in its next iteration the algorithm can no longer look for change points in that proximity, since this will result, with high probability, in spurious detections. This phenomenon is due to the fact that the behavior of the CUSUM statistic of the projected data is not uniform around its local maxima. Therefore, after a true change point has been detected, the algorithm must scan only nearby regions of low signal-to-noise ratio – so that the probability of false positives can be adequately controlled with a proper choice of the thresholding parameter τ. Thus the need to stay away by at leat the localization error form the detected chaing points. A very similar condition is imposed in the main algorithm of Wang and Samworth (2016) and implicitly in Korkas and Fryzlewicz (2017). The value of δ is left as an input paramerer (as in Wang and Samworth, 2016), but any value between the localization error ϵ_n given in the statement of Theorem 2 and the minimal distance Δ between change points will do. In the proof of we Theorem 2 we set $\delta \leq 3\Delta/32$.

Theorem 2 (Consistency of WBSIP). *Let Assumptions 1 and 3 hold let* $\{(\alpha_m,\ \beta_m)\}_{m=1}^M \subset (0, T)^M$ *be a collection of intervals whose end points are drawn independently and uniformly from* $\{1, \ldots, T\}$ *and such that* $\max_{1\leq m\leq M}(\beta_m - \alpha_m) \leq C\Delta$ *for an absolute constant* $C > 0$*. Set*

$$\epsilon_n = C_1 B^4 \log(n)\kappa^{-2},$$

for a $C_1 > 0$*. Suppose there exist* $c_2, c_3 > 0$*, sufficiently small, such that the input parameters* τ *and* δ *satisfy*

$$\begin{aligned} &B^2\sqrt{\log(n)} < \tau < c_2\kappa\sqrt{\Delta}, \\ &\epsilon_n < \delta \leq c_3\Delta. \end{aligned} \tag{8}$$

Then the collection of the estimated change points $\mathcal{B} = \{\hat{\eta}_k\}_{k=1}^{\widehat{K}}$ *returned by WBSIP with input parameters of* $(0, n)$*,* $\{(\alpha_m, \beta_m)\}_{m=1}^M$*,* τ *and* δ *satisfies*

$$\begin{aligned} &\mathbb{P}\Big\{\widehat{K} = K; \quad \max_{k=1,\ldots,K} |\eta_k - \hat{\eta}_k| \leq \epsilon_n\Big\} \\ \geq &1 - 2n^2Mn^{-c} - n^3 9^p 2n^{-cp} - \exp\big(\log(n/\Delta) - M\Delta^2/(16n^2)\big) \end{aligned}$$

for some absolute constants $c > 0$*.*

Remark 8 (**The relationships among the constants in Theorem 2**). *The choice of the constant* C *is essentially arbitrary but will affect the choice of the constants* C_1*,* c_2 *and* c_3*, where* c_2 *and* c_3 *in particular have to be picked small enough. This dependence can be tracked in the proof but we refrain from giving further details.*

The above theorem implies that the WBSIP algorithm can estimate the change points perfectly well, with high probability, with a localization rate upper bounded by

$$\frac{\epsilon_n}{n} \preceq \frac{B^4}{\kappa^2}\frac{\log n}{n}.$$

The fact that the dimension p does not appear explicitly in the localization rate is an interesting, if not perhaps surprising, finding. Of course, the dimension does affect (negatively) the performance of the algorithm through Assumption 3: keeping n and Δ fixed, a larger value of p implies a larger value of $\frac{B^4}{\kappa^2}$ in order for that assumption to hold. In turn, this leads to a larger bound in Theorem 2. Furthermore, the dimension p appears in the probability of the event that WBSIP fails to localize all the change points. We remark that, for the different problem of high-dimensional mean change point detection, Wang and Samworth (2016) also obtained a localization rate independent of the dimension: see Theorem 3 there. In Section 3 below we will prove that Assumption 3 is in fact essentially necessary for any algorithm to produce a vanishing localization rate.

Proof. Since ϵ_n is the desired order of localization rate, by induction, it suffices to consider any generic $(s, e) \subset (0, T)$ that satisfies

$$\eta_{r-1} \leq s \leq \eta_r \leq \ldots \leq \eta_{r+q} \leq e \leq \eta_{r+q+1}, \quad q \geq -1,$$
$$\max\{\min\{\eta_r - s, s - \eta_{r-1}\}, \min\{\eta_{r+q+1} - e, e - \eta_{r+q}\}\} \leq \epsilon_n,$$

where $q = -1$ indicates that there is no change point contained in (s, e).

Note that under Assumption 3, $\epsilon_n \leq \Delta/4$; it, therefore, has to be the case that for any change point $\eta_p \in (0, T)$, either $|\eta_p - s| \leq \epsilon_n$ or $|\eta_p - s| \geq \Delta - \epsilon_n \geq 3\Delta/4$. This means that $\min\{|\eta_p - e|, |\eta_p - s|\} \leq \epsilon_n$ indicates that η_p is a detected change point in the previous induction step, even if $\eta_p \in (s, e)$. We refer to $\eta_p \in [s, e]$ as an undetected change point if $\min\{\eta_p - s, \eta_p - e\} \geq 3\Delta/4$.

In order to complete the induction step, it suffices to show that WBSIP$((s, e), \{(\alpha_m, \beta_m)\}_{m=1}^M, \tau, \delta)$ (i) will not detect any new change point in (s, e) if all the change points in that interval have been previous detected, and (ii) will find a point b in (s, e) (in fact, in $(s+\delta, e-\delta)$) such that $|\eta_p - b| \leq \epsilon_n$ if there exists at least one undetected change point in (s, e).

Let

$$\{u_m\}_{m=1}^M = PC(\{X_i\}_{i=1}^n, \{(\alpha_m, \beta_m)\}_{m=1}^M).$$

Since the intervals $\{(\alpha_m, \beta_m)\}_{m=1}^M$ are generated independently from $\{X_i\}_{i=1}^n \cup \{W_i\}_{i=1}^n$, the rest of the argument is made under $\mathcal{M}$ defined in Equation (20) of Appendix A, which has no effects on the distribution of $\{X_i\}_{i=1}^n \cup \{W_i\}_{i=1}^n$.

Step 1. Let $\lambda_1 = B^2\sqrt{p\log(n)}$. In this step, we are to show that, on the event $\mathcal{A}_1(\{W_i\}_{i=1}^n, \lambda_1)$ and for some $c_1' > 0$,

$$\sup_{1 \leq m \leq M} |u_m^\top(\Sigma_{\eta_k} - \Sigma_{\eta_{k-1}})u_m| \geq c_1'\|\Sigma_{\eta_k} - \Sigma_{\eta_{k-1}}\|_{\mathrm{op}} = c_1'\kappa_k \quad \text{for every} \quad k = 1, \ldots, K+1 \tag{9}$$

On the event $\mathcal{M}$, for any $\eta_k \in (0, n)$, without loss of generality, there exists $\alpha_k \in [\eta_k - 3\Delta/4, \eta_k - \Delta/2]$ and $\beta_k \in [\eta_k + \Delta/2, \eta_k + 3\Delta/4]$. Thus $[\alpha_k, \beta_k]$ contains only one change point η_k. Using Lemma 20 in Appendix D and the inequality $p\log(n) \leq \Delta/8$, we have that

$$\max_{t = \lceil \alpha_k + \delta \rceil, \ldots, \lfloor \beta_k - \delta \rfloor} \|\widetilde{\Sigma}_t^{\alpha_k, \beta_k}\|_{\mathrm{op}} = \|\widetilde{\Sigma}_{\eta_k}^{\alpha_k, \beta_k}\|_{\mathrm{op}} \geq (1/2)\|\Sigma_{\eta_k} - \Sigma_{\eta_{k-1}}\|_{\mathrm{op}}\sqrt{\Delta}. \tag{10}$$

Let $b_k \in \arg\max_{t=\lceil\alpha_k+\delta\rceil,\ldots,\lfloor\beta_k-\delta\rfloor} \|\widetilde{S}_t^{\alpha_k,\beta_k}\|_{\mathrm{op}}$, where $\widetilde{S}_t^{s,e}$ denote the covariance CUSUM statistics of $\{W_i\}_{i=s+1}^{e}$ at evaluated t. Since $\|\Sigma_{\eta_k} - \Sigma_{\eta_{k-1}}\|_{\mathrm{op}} = \kappa_k$, by definition,

$$\begin{aligned}
|u_k^\top \widetilde{\Sigma}_{b_k}^{\alpha_k,\beta_k} u_k| &\geq |u_k^\top \widetilde{S}_{b_k}^{\alpha_k,\beta_k} u_k| - \lambda_1 \\
&= \max_{t=\lceil\alpha_k+\delta\rceil,\ldots,\lfloor\beta_k-\delta\rfloor} \|\widetilde{S}_t^{\alpha_k,\beta_k}\|_{\mathrm{op}} - \lambda_1 \\
&\geq \max_{t=\lceil\alpha_k+\delta\rceil,\ldots,\lfloor\beta_k-\delta\rfloor} \|\widetilde{\Sigma}_t^{\alpha_k,\beta_k}\|_{\mathrm{op}} - 2\lambda_1 \\
&\geq (1/2)\|\Sigma_{\eta_k} - \Sigma_{\eta_{k-1}}\|_{\mathrm{op}}\sqrt{\Delta} - 2\lambda_1 \\
&\geq (1/4)\|\Sigma_{\eta_k} - \Sigma_{\eta_{k-1}}\|_{\mathrm{op}}\sqrt{\Delta}
\end{aligned}$$

where the first and second inequalities hold on the event $\mathcal{A}_1(\{W_i\}_{i=1}^n, \lambda_1)$, the third inequality follows from (10) and the last inequality from Assumption 3. Next, observe that

$$\widetilde{\Sigma}_t^{\alpha_k,\beta_k} = \begin{cases} \sqrt{\frac{t-\alpha_k}{(\beta_k-\alpha_k)(\beta_k-t)}}(\beta_k - \eta_k)(\Sigma_{\eta_k} - \Sigma_{\eta_{k-1}}), & t \leq \eta_k, \\ \sqrt{\frac{\beta_k-t}{(\beta_k-\alpha_k)(t-\alpha_k)}}(\eta_k - \alpha_k)(\Sigma_{\eta_k} - \Sigma_{\eta_{k-1}}), & t \geq \eta_k. \end{cases}$$

Using the above expression, for $b_k \geq \eta_k$, we have that

$$\begin{aligned}
(1/4)\|\Sigma_{\eta_k} - \Sigma_{\eta_{k-1}}\|_{\mathrm{op}}\sqrt{\Delta} &\leq \left|u_k^\top \widetilde{\Sigma}_{b_k}^{\alpha_k,\beta_k} u_k\right| \\
&= \sqrt{\frac{\beta_k - b_k}{(\beta_k - \alpha_k)(b_k - \alpha_k)}}(\eta_k - \alpha_k)\left|u_k^\top(\Sigma_{\eta_k} - \Sigma_{\eta_{k-1}})u_k\right| \\
&\leq \sqrt{\frac{(\beta_k - \eta_k)(\eta_k - \alpha_k)}{\beta_k - \alpha_k}}\left|u_k^\top(\Sigma_{\eta_k} - \Sigma_{\eta_{k-1}})u_k\right| \\
&\leq \sqrt{2\Delta}\left|u_k^\top(\Sigma_{\eta_k} - \Sigma_{\eta_{k-1}})u_k\right|.
\end{aligned}$$

Therefore (9) holds with $c_1' = 1/(2\sqrt{2})$. The case of $b_k < \eta_k$ follows from very similar calculations.

Step 2. In this step, we will show that WBSIP$((s,e), \{(\alpha_m, \beta_m)\}_{m=1}^M, \tau, \delta)$ will consistently detect or reject the existence of undetected change points within (s,e), provided that (9) holds and on the two events $\mathcal{B}_1(\{X_i\}_{i=1}^n, \{u_m\}_{m=1}^M, \lambda_2)$, where $\lambda_2 = B^2\sqrt{\log(n)}$, and $\mathcal{M}$, given in Equation (18) and Equation (20) in Appendix A, respectively.

Let a_m, b_m and $m*$ be defined as in WBSIP$((s,e), \{(\alpha_m, \beta_m)\}_{m=1}^M, \tau, \delta)$. Denote $Y_i(u_m) = (u_m^\top X_i)^2$ and $f_i(u_m) = u_m^\top \Sigma_i u_m$. Let $\widetilde{Y}_t^{s,e}(u_m)$ and $\widetilde{f}_t^{s,e}(u_m)$ be defined as in (53) and (54) of Appendix D.2 respectively.

Suppose there exists a change point $\eta_p \in (s,e)$ such that $\min\{\eta_p - s, e - \eta_p\} \geq 3\Delta/4$. Let $\delta \leq 3\Delta/32$. Then, on the event $\mathcal{M}$, there exists an interval (α_m, β_m) selected by WBSIP such that $\alpha_m \in [\eta_p - 3\Delta/4, \eta_p - \Delta/2]$ and $\beta_m \in [\eta_p + \Delta/2, \eta_p + 3\Delta/4]$.

Then $[s_m', e_m'] = [\alpha_m, \beta_m] \cap [s,e]$, and $[s_m, e_m] = [s_m' + \delta, e_m' - \delta]$ (see details of the WBSIP procedure in Algorithm 3). Moreover, we have that $\min\{\eta_p - s_m, e_m - \eta_p\} \geq (1/2)\Delta$. Thus, $[s_m, e_m]$ contains at most one change point of the time series $\{f_i(u_m)\}_{i=1}^n$. A similar calculation as the one shown in the proof of Lemma 20 gives that

$$\max_{\lceil s_m+\log(n)\rceil \leq t \leq \lfloor e_m - \log(n)\rfloor} |\widetilde{f}_t^{s_m,e_m}(u_m)| \geq (1/8)\sqrt{\Delta}|u_m^\top(\Sigma_{\eta_p} - \Sigma_{\eta_{p-1}})u_m|,$$

where $e_m - s_m \leq (3/2)\Delta$ is used in the last inequality. Therefore

$$\begin{aligned} a_m &= \max_{\lceil s_m + \log(n) \rceil \leq t \leq \lfloor e_m - \log(n) \rfloor} |\widetilde{Y}_t^{s_m, e_m}(u_m)| \\ &\geq \max_{\lceil s_m + \log(n) \rceil \leq t \leq \lfloor e_m - \log(n) \rfloor} |\widetilde{f}_t^{s_m, e_m}(u_m)| - \lambda_2 \\ &\geq (1/8)\sqrt{\Delta} |u_m'(\Sigma_{\eta_p} - \Sigma_{\eta_{p-1}}) u_m| - \lambda_2, \end{aligned}$$

where the first inequality holds on the event $\mathcal{B}_1(\{X_i\}_{i=1}^n, \{u_m\}_{m=1}^M, \lambda_2)$. Thus for any undetected change point η_p within (s, e), it holds that

$$a_{m*} = \sup_{1 \leq m \leq M} a_m \tag{11}$$

$$\geq \sup_{1 \leq m \leq M} (1/8)\sqrt{\Delta} |u_m'(\Sigma_p - \Sigma_{p-1}) u_m| - \lambda_2 \geq (c_1'/8)\kappa_p \sqrt{\Delta} - \lambda_2 \tag{12}$$

$$\geq (c_1'/16)\kappa_p \sqrt{\Delta} \tag{13}$$

where the second inequality follows from (9), and the last inequality from

$$\lambda_2 = B^2 \sqrt{\log(n)} \leq (c_1'/16)\kappa\sqrt{\Delta},$$

by choosing the constant C in Assumption 3 to be at least $\sqrt{32}$.

Then, WBSIP$((s, e), \{(\alpha_m, \beta_m)\}_{m=1}^M, \tau, \delta)$ correctly accepts the existence of undetected change points on the events (9), $\mathcal{B}_1(\{X_i\}_{i=1}^n, \{u_m\}_{m=1}^M, \lambda_2, \delta)$ and $\mathcal{M}$.

Suppose there does not exists any undetected change points within (s, e), then for any $(s_m', e_m') = (\alpha_m, \beta_m) \cap (s, e)$, one of the following situations must hold.

(a) There is no change point within (s_m', e_m');

(b) there exists only one change point η_r within (s_m, e_m) and $\min\{\eta_r - s_m', e_m' - \eta_r\} \leq \epsilon_n$; or

(c) there exist two change points η_r, η_{r+1} within (s_m, e_m) and $\max\{\eta_r - s_m', e_m' - \eta_{r+1}\} \leq \epsilon_n$.

Observe that if (a) holds, then, on the event $\mathcal{B}_1(\{X_i\}_{i=1}^n, \{u_m\}_{m=1}^M, \lambda_2)$ given in Equation (18) in Appendix A, for $(s_m, e_m) = (s_m' + \delta, e_m' - \delta)$, we have

$$\max_{\lceil s_m + \log(n) \rceil \leq t \leq \lfloor e_m - \log(n) \rfloor} |\widetilde{Y}_t^{s_m, e_m}(u_m)| \leq \max_{\lceil s_m + \log(n) \rceil \leq t \leq \lfloor e_m - \log(n) \rfloor} |\widetilde{f}_t^{s_m, e_m}(u_m)| + \lambda_2 = 0 + \lambda_2.$$

If (b) or (c) holds, then since $(s_m, e_m) = (s_m' + \delta, e_m' - \delta)$ and $\delta \leq \epsilon_n$, it must be the case that (s_m, e_m) does not contain any change points. This reduces to case (a). Therefore if (8) holds, then WBSIP$((s, e), \{(\alpha_m, \beta_m)\}_{m=1}^M, \tau, \delta)$ will always correctly reject the existence of undetected change points, on the event $\mathcal{B}_1(\{X_i\}_{i=1}^n, \{u_m\}_{m=1}^M, \lambda_2)$.

Step 3. Assume that there exists a change point $\eta_p \in (s, e)$ such that $\min\{\eta_p - s, \eta_p - e\} \geq 3\Delta/4$. Let a_m, b_m and $m*$ be defined as in WBSIP$((s, e), \{(\alpha_m, \beta_m)\}_{m=1}^M, \tau)$.

To complete the proof it suffices to show that, on the events $\mathcal{B}_1(\{X_i\}_{i=1}^n, \{u_m\}_{m=1}^M, \lambda_2)$ and $\mathcal{B}_2(\{X_i\}_{i=1}^n, \{u_m\}_{m=1}^M, \lambda_2)$ given in Equation (18) and Equation (19) respectively of Appendix A, there exists a change point $\eta_k \in [s_{m*}, e_{m*}]$ such that $\min\{\eta_k - s, \eta_k - e\} \geq 3\Delta/4$ and $|b_{m*} - \eta_k| \leq \epsilon_n$.

Consider the univariate time series $\{Y_i(u_{m*})\}_{i=1}^n$ and $\{f_i(u_{m*})\}_{i=1}^n$ defined in (53) and (54) of Appendix D.2. Since the collection of the change points of the time series $\{f_i(u_{m*})\}_{i=s_{m*}}^{e_{m*}}$ is a subset of that of $\{\eta_k\}_{k=0}^{K+1} \cap [s,e]$, we may apply Corollary 13 to the time series $\{Y_i(u_{m*})\}_{i=s_{m*}}^{e_{m*}}$ and $\{f_i(u_{m*})\}_{i=s_{m*}}^{e_{m*}}$. To that end, we will need to ensure that the assumptions of Corollary 13 are verified. Let $\delta' = \log(n)$ and $\lambda = \lambda_2$. Observe that (48) and (49) are straightforward consequences of Assumption 3, (46) and (47) follow from the definiiton of $\mathcal{B}_1(\{X_i\}_{i=1}^n, \{u_m\}_{m=1}^M, \lambda_2)$ and $\mathcal{B}_2(\{X_i\}_{i=1}^n, \{u_m\}_{m=1}^M, \lambda_2)$, and that (45) follows from (13).

Thus, all the conditions in Corollary 13 are met, and we therefore conclude that there exists a change point η_k, which is also a change point of $\{f_i(v)\}_{i=s_{m*}}^{e_{m*}}$, satisfying

$$\min\{e_{m*} - \eta_k, \eta_k - s_{m*}\} > \Delta/4 \tag{14}$$

and

$$|b_{m*} - \eta_k| \leq \max\{C_3\lambda_2^2\kappa^{-2}, \delta'\} \leq \epsilon_n,$$

where the last inequality holds because $\lambda_2^2\kappa^{-2} = B^4\log(n)\kappa^{-2} \geq \log(n)$, which is a consequence of the inequality $B^2 \geq \kappa$.

The proof is complete with the following two observations: i) The change points of $\{f_i(u_{m*})\}_{i=s}^{e}$ belong to $(s,e) \cap \{\eta_k\}_{k=1}^K$; and ii) Equation (14) and $(s_{m*}, e_{m*}) \subset (s,e)$ imply that

$$\min\{e - \eta_k, \eta_k - s\} > \Delta/4 > \epsilon_n.$$

As discussed in the argument before **Step 1**, this implies that η_k must be an undetected change point of $\{X_i\}_{i=1}^n$ in the covariance structure. □

3 Lower bounds

In this section, we provide lower bounds for the problem of chenge point estimation with high dimensional covariance matrices.

In Theorem 2 we showed that if the distribution of $\{X_i\}_{i=1}^n$ satisfies Assumption 1 and additionally the condition that $\Delta \geq CB^4\kappa^{-2}p\log(n)$ for sufficiently large C as given in Assumption 3, then the WBSIP algorithm can, with high probability, detect all the change points with a localization rate of the order

$$\frac{B^4}{\kappa^2}\frac{\log n}{n}. \tag{15}$$

Assumption 3 might seem a bit arbitrary at first glance. However, we will show in the next result that if the class of distribution of interest allows for a spacing parameter $\Delta = cB^4\kappa^{-2}p$ for some sufficiently small constant c, then it is *not* possible to estimate the location of the change point at a vanoishing rate. This result furher implies that the WBSIP algorithm is optimal in the sense of requiring the minimax scaling for the problem parameters.

To that effect, we will consider the following class of data generating distribution.

Assumption 4. *Let $X_1, \ldots, X_n \in \mathbb{R}^p$ be independent Gaussian random vectors such that $X_i \sim N_p(0, \Sigma_i)$ with $\|\Sigma_i\|_{\mathrm{op}} \leq 2\sigma$. Let $\{\eta_k\}_{k=0}^{K+1} \subset \{0, \ldots, n\}$ be a collection of change points, such that $\eta_0 = 0$ and $\eta_{K+1} = n$ and that*

$$\Sigma_{\eta_k+1} = \Sigma_{\eta_k+2} = \ldots = \Sigma_{\eta_{k+1}}, \text{ for any } k = 1, \ldots, K+1.$$

Assume there exists parameters $\kappa = \kappa(n)$ and $\Delta = \Delta(n)$ such that

$$\inf_{k=1,\ldots,K+1}\{\eta_k - \eta_{k-1}\} \geq \Delta > 0,$$
$$\|\Sigma_{\eta_k} - \Sigma_{\eta_{k-1}}\|_{\mathrm{op}} = \kappa_k \geq \kappa > 0, \text{ for any } k = 1, \ldots, K+1.$$

We will denote with $P^n_{\kappa,\Delta,\sigma}$ the joint distribution of the data when the above assumption is in effect. Notice that $P^n_{\kappa,\Delta,\sigma}$ satisfies Assumption 1 with $B = 8\sigma$. Below we prove that consistent estimation of the locations of the change point requires $\frac{\Delta\kappa^2}{p}B^4$ to diverge, as in Assumption 3. We recall that all the parameters of interest, Δ, κ, σ and p are allowed to change with n. The proof is based on a construction used in Cai and Ma (2013) to obtain minimax lower bounds for a class of hypothesis testing problems involving covariance matices.

Lemma 3. *Let $\{X_i\}_{i=1}^n$ be a time series satisfies Assumption 4 with only one change point and let $P^n_{\kappa,\Delta,\sigma}$ denote the corresponding joint distribution. Consider the class of distributions*

$$\mathcal{P} = \mathcal{P}(n) = \left\{ P^n_{\kappa,\Delta,\sigma} : \frac{\sigma^4 p}{33\kappa^2} \leq \Delta \leq n/3, \kappa \leq \sigma^2/4 \right\},$$

Then

$$\inf_{\hat{\eta}} \sup_{P\in\mathcal{P}} E_P(|\hat{\eta} - \eta|) \geq n/6.$$

Remark 9 (Phase transiton for the localization rate in covariance change point detection problem). *In light of Lemma 3 and Theorem 2, we conclude that, in the covariance change point detection problem, the solution undergoes a phase transition, which are able to characterize up to logarithmic factor (in n). Specifically,*

- *if $\Delta \geq CB^4 p \log(n)/\kappa^2$ for a sufficiently large constant $C > 0$, then it is possible to estimate the locations of the change points with a localization rate vanishing in n;*
- *on the other hand, if $\Delta = CB^4 p/\kappa^2$ for a sufficiently small constant $c > 0$, then the localization rate of any algorithm remains, at least in the worst case, bounded away from 0.*

To the best of our knowledge, this phase transition effect is new and unique to our settings.

We conclude this section by showing that the localization rate that we have obtained for the WBSIP algorithm, given above in (15), is, up to a logarithmic factor, minimax optimal.

Lemma 4. *Consider the class of distribution $\mathcal{Q} = \left\{ P^n_{\kappa,\Delta,\sigma} : \Delta < n/2, \kappa \leq \sigma^2/4 \right\}$. Then,*

$$\inf_{\hat{\eta}} \sup_{P\in\mathcal{Q}} E_P(|\hat{\eta} - \eta|) \geq c\sigma^4\kappa^{-2}.$$

4 Discussion

In this paper, we tackle the problem of change point detection for a time series of length n of independent p-dimensional random vectors with covariance matrices that are pirece-wise constant. We allow the dimension, as well as other parameters quantifying the difficulty of the problem, to change with n. We have devised two procedures based on existing algorithms for change point detection – binary segmentation and wild binary segmentation – and show that the localization rates they yield are consistent with those in the univariate time series mean change point detection problems. In particular we demonstrate the algorithm WBSIP, which applies wild binary segmentation to carefully chosen univariate projections of the data, produces a localization rate that is, up to a logarithmic factor, minimax optimal.

The model setting adopted in the paper allows for the dimension p to grow with n. However, in order for the localization rates of any procedure to vanish with n it must be the case that p is of smaller order than n. One possible future direction is to consider different high dimensional settings whereby p is permitted to grow even faster than n, with additional structural assumptions on the underlying covariance matrices. For instance, we may model the covariance matrices as spiked matrices with sparse leading eigenvectors.

A common, undesirable feature of the WBSIP algorithms is the fact that, for given interval (s, e), the search of the next change point is limited to points inside the interval that are δ away from the endpoints, where δ is an inout parameter that is larger than the localization error. Such restriction, which appears also in the algorithm of Wang and Samworth (2016) for mean change point localization of high-dimensional time series, is made in order to prevent the algorithms from returning spurious change points in the proximity of a true change point. The reason for such phenomenon is subtle, and is ultimately due to the fact that the rate of decay of the expected value of the covariance CUSUM statistics around the true change points is in general not uniform, as it depends on the magnitude of the change. A possible solution to such an issue – which appears to be unavoidable – would be to design adaptive algorithm yielding local rates, one for each change point. We will pursue this line of research in future work.

One key assumption used in this paper is the the time series of comprised of independent observations. This is of course a rather strong condition, which might not apply to many real-life problems. In order to handle time dependence, a natural approach is introduce mixing conditions (see, e.g. Wu, 2005) or assume the observations come from certain well-defined time series models. Further extensions that will be worth pursuing include the cases in which model is mis-specified or the observations are contaminated We leave these interesting extensions to future work but expect that many of the results derived in this manuscript will provide the theoretical underpinning for devising and studying more complicated algorithms.

Acknowledgments

We thank Dr Haeran Cho for her constructive comments.

References

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. 3rd ed. Wiley Series in Probability and Statistics, John Wiley & Sons.

Aston, J. A. D. and Kirch, C. (2014). Efficiency of change point tests in high dimensional settings. *arXiv preprint arXiv: 1409.1771*.

Aue, A., Hömann, S., Horváth, L. and Reimherr, M. (2009). Break detection in the covariance structure of multivariate nonlinear time series models. *The Annals of Statistics*, **37** 4046–4087.

Avanesov, V. and Buzun, N. (2016). Change-point detection in high-dimensional covariance structure. *arXiv preprint arXiv:1610.03783*.

Baranowski, R., Chen, Y. and Fryzlewicz, P. (2016). Narrowest-Over-Threshold detection of multiple change-points and change-point-like feature. *arXiv preprint arXiv: 1609.00293*.

Barigozzi, M., Cho, H. and Fryzlewicz, P. (2016). Simultaneous multiple change-point and factor analysis for high-dimensional time series. *arXiv preprint arXiv: 1612.06928*.

Berkes, I., Gabrys, R., Horváth, L. and Kokoszka, P. (2009). Detecting changes in the mean of functional observations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **71** 927–946.

Berthet, Q. and Rigollet, P. (2013). Optimal detection of sparse principal components in high dimension. *The Annals of Statistics*, **41** 1780–1815.

Birke, M. and Dette, H. (2005). A note on testing the covariance matrix for large dimension. *Statistics and Probability Letters*, **74** 281–289.

Cai, T. T. and Ma, Z. (2013). Optimal hypothesis testing for high dimensional covariance matrices. *Bernoulli*, **19** 2359–2388.

Chan, H. P. and Walther, G. (2013). Detection with the scan and the average likelihood ratio. *Statistica Sinica*, **1** 409–428.

Cho, H. (2015). Change-point detection in panel data via double cusum statistic. *Electronic Journal of Statistics* in press.

Cho, H. and Fryzlewicz, P. (2012). Multiscale and multilevel technique for consistent segmentation of nonstationary time series. *Statistica Sinica*, **22** 207–229.

Cho, H. and Fryzlewicz, P. (2015). Multiple change-point detection for high-dimensional time series via Sparsified Binary Segmentation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **77** 475–507.

Davies, P. L. and Kovac, A. (2001). Local extremes, runs, strings and multiresolution. *Ann. Statist.*, **29** 1–65.

Davis, R. A., Lee, T. C. M. and Rodriguez-Yam, G. A. (2006). Structural break estimation for nonstationary time series models. *Journal of the American Statistical Association*, **101** 223–239.

Frick, K., Munk, A. and Sieling, H. (2014). Multiscale change point inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **76** 495–580.

Fryzlewicz, P. (2014). Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, **42** 2243–2281.

Gombay, E., Horváth, L. and Hušková, M. (1996). Estimators and tests for change in variances. *Statistics and Risk Modeling*, **14** 145–160.

Harchaoui, Z. and Lévy-Leduc, C. (2010). Multiple change-point estimation with a total variation penalty. *Journal of American Statistical Association*, **105** 1480–1493.

Harchaoui, Z. and Lvy-Leduc, C. (2010). Multiple change-point estimation with a total variation penalty. *Journal of the American Statistical Association*, **105** 1480–1493.

Horváth, L. and Hušková, M. (2012). Change-point detection in panel data. *Journal of Time Series Analysis*, **33** 631–648.

Inclan, C. and Tiao, G. C. (1994). Use of cumulative sums of squares for retrospective detection of changes of variance. *Journal of the American Statistical Association*, **89** 913–923.

James, B., James, K. L. and Siegmund, D. (1987). Tests for a change-point. *Biometrika*, **74** 71–83.

Jirak, M. (2015). Uniform change point tests in high dimension. *The Annals of Statistics*, **43** 2451–2483.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal components analysis. *The Annals of Statistics*, **29** 295–327.

Killick, R., Fearnhead, P. and Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, **107** 1590–1598.

Korkas, K. K. and Fryzlewicz, P. (2017). Multiple change-point detection for non-stationary time series using wild binary segmentation. *Statistica Sinica*, **27** 287–311.

Lavielle, M. (1999). Detection of multiple changes in a sequence of dependent variables. *Stochastic Processes and their Applications*, **83** 79–102.

Lavielle, M. and Moulines, E. (2000). Least-squares estimation of an unknown number of shifts in a time series. *Journal of Time Series Analysis*, **21** 33–59.

Ledoit, O. and Wolf, M. (2002). Some hypothesis tests for the covariance in principal components analysis. *The Annals of Statistics*, **30** 1081–1102.

Li, H., Guo, Q. and Munk, A. (2017). Multiscale change-point segmentation: Beyond step functions. *arXiv preprint arXiv: 1708.03942*.

Lin, K., Sharpnack, J. L., Rinaldo, A. and Tibshirani, R. J. (2017). A sharp error analysis for the fused lasso, with application to approximate changepoint screening. In *Advances in Neural Information Processing Systems 30* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett, eds.). 6887–6896.

Olshen, A. B., Venkatraman, E., Lucito, R. and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics*, **5** 557–572.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, **41** 100–115.

Picard, D. (1985). Testing and estimating change-points in time series. *Advances in Applied Probability*, **17** 841–867.

Qian, J. and Jia, J. (2012). On pattern recovery of the fused lasso. Tech. rep. Https://arxiv.org/abs/1211.5194.

Rojas, C. R. and Wahlberg, B. (2014). On change point detection using the fused lasso method. Tech. rep. Https://arxiv.org/abs/1401.5408.

Roy, S. N. (1957). *Some Aspects of Multivariate Analysis*. New York: Wiley.

Siegmund, D. and Venkatraman, E. S. (1995). Using the generalized likelihood ratio statistic for sequential detection of a change-point. *The Annals of Statistics* 255–271.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67** 91–108.

Venkatraman, E. S. (1992). *Consistency results in multiple change-point problems*. Ph.D. thesis, Stanford University.

Vershynin, R. (2010). Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*.

Vostrikova, L. (1981). Detection of the disorder in multidimensional random-processes. *Doklady Akademii Nauk SSSR*, **259** 270–274.

Wald, A. (1945). Sequential tests of statistical hypotheses. *The Annals of Mathematical Statistics*, **16** 117–186.

Wang, T. and Samworth, R. J. (2016). High-dimensional changepoint estimation via sparse projection. *arXiv preprint arXiv:1606.06246*.

Wang, Y. (1995). Jump and sharp cusp detection by wavelets. *Biometrika*, **82** 385–397.

Wu, W. B. (2005). Nonlinear system theory: Another look at dependence. *Proceedings of the National Academy of Sciences of the United States of America*, **102** 14150–14154.

Yao, Y.-C. and Au, S.-T. (1989). Least-squares estimation of a stop function. *Sankhyā: The Indian Journal of Statistics, Series A* 370–381.

Yu, B. (1997). *Festschrift for Lucien Le Cam*, vol. 423, chap. Assouad, Fano, and Le Cam. Springer Science & Business Media, 435.

Appendices

A Probabilistic bounds

In this section, we give basic high-probability concentration bounds on the fluctuations of te covariance CUSUM statistics using the notions of sub-Gaussian and sub-Exponential random vectors. We also state some properties of the randomly selected intervals $\{s_m, e_m\}$ in the WBS algorithm, which hold with high probability.

We start by introducing the definitions of sub-Gaussian and sub-Exponential random variables through Orlicz norms. See, e.g., Vershynin (2010) for more details.

Definition 2. *(i) A random variable $X \in \mathbb{R}$ is sub-Gaussian if*

$$\|X\|_{\psi_2} := \sup_{k \geq 1} k^{-1/2} \{\mathbb{E}(|X|^k)\}^{1/k} < \infty.$$

A random vector $X \in \mathbb{R}^p$ is sub-Gaussian if

$$\|X\|_{\psi_2} := \sup_{v \in \mathcal{S}^{p-1}} \|v^\top X\|_{\psi_2} < \infty,$$

where $\mathcal{S}^{p-1}$ denote unit sphere in Eucliden norm in $\mathbb{R}^p$.

(ii) A random variable $Y \in \mathbb{R}$ is sub-Exponential if

$$\|Y\|_{\psi_1} = \sup_{k \geq 1} k^{-1} \mathbb{E}(|Y|^k)^{1/k} < \infty.$$

A random vector $Y \in \mathbb{R}^p$ is sub-Exponential if

$$\|Y\|_{\psi_1} = \sup_{v \in \mathcal{S}^{p-1}} \|v^\top Y\|_{\psi_1} < \infty.$$

We note that if $X \in \mathbb{R}$ is sub-Gaussian, then X^2 is sub-Exponential, due to the easily verifiable fact that

$$\|X\|_{\psi_2}^2 \leq \|X^2\|_{\psi_1} \leq 2\|X\|_{\psi_2}^2. \tag{16}$$

Recall the sample and the population versions of the covariance CUSUM statistics given in Definition 1. For $\lambda > 0$, we define the events, which depend on $\{X_i\}_{i=1}^n$.

$$\mathcal{A}_1(\{X_i\}_{i=1}^n, \lambda) = \left\{ \sup_{0 \leq s < t < e \leq n} \left\| \widetilde{S}_t^{s,e} - \widetilde{\Sigma}_t^{s,e} \right\|_{\mathrm{op}} \leq \lambda, \quad \min\{t-s, e-t\} \geq p \log(n) \right\} \tag{17}$$

and

$$\mathcal{A}_2(\{X_i\}_{i=1}^n, \lambda) = \left\{ \sup_{0 \leq s < e \leq n} \frac{\left\| \sum_{i=s+1}^e (X_i X_i^\top - \Sigma_i) \right\|_{\mathrm{op}}}{\sqrt{e-s}} \leq \lambda, \quad e - s \geq p \log(n) \right\}.$$

Next, for an arbitary collection $\{v_m\}_{m=1}^M$ of deterministic unit vectors in $\mathbb{R}^p$ we define the events

$$\mathcal{B}_1(\{X_i\}_{i=1}^n, \{v_m\}_{m=1}^M, \lambda) = \left\{ \sup_{1 \leq m \leq M} \sup_{0 \leq s < t < e \leq n} \left| v_m^\top (\widetilde{S}_t^{s,e} - \widetilde{\Sigma}_t^{s,e}) v_m \right| \leq \lambda, \quad \min\{t-s, e-t\} \geq \log(n) \right\} \tag{18}$$

and

$$\mathcal{B}_2(\{X_i\}_{i=1}^n, \{v_m\}_{m=1}^M, \lambda) = \left\{ \sup_{1 \leq m \leq M} \sup_{0 \leq s < e \leq n} \frac{\left| \sum_{i=s+1}^e v_m^\top (X_i X_i^\top - \Sigma_i) v_m \right|}{\sqrt{e-s}} \leq \lambda, \quad e - s \geq \log(n) \right\}. \tag{19}$$

Lemma 5. *Suppose $\{X_i\}_{i=1}^n \subset \mathbb{R}^p$ are i.i.d sub-Gaussian centered random vectors such that*

$$\sup_{1 \leq i \leq n} \|X_i\|_{\psi_2} \leq B.$$

There exists an absolute constant $c > 0$ such that,

$$\mathbb{P}(\mathcal{A}_1(\{X_i\}_{i=1}^n, B^2\sqrt{p\log(n)})) \geq 1 - 2 \times 9^p n^3 n^{-cp},$$
$$\mathbb{P}(\mathcal{A}_2(\{X_i\}_{i=1}^n, B^2\sqrt{p\log(n)})) \geq 1 - 2 \times 9^p n^2 n^{-cp},$$
$$\mathbb{P}(\mathcal{B}_1(\{X_i\}_{i=1}^n, \{v_m\}_{m=1}^M, B^2\sqrt{\log(n)})) \geq 1 - 2n^3 M n^{-c},$$
$$\mathbb{P}(\mathcal{B}_2(\{X_i\}_{i=1}^n, \{v_m\}_{m=1}^M, B^2\sqrt{\log(n)})) \geq 1 - 2n^2 M n^{-c},$$

for any set $\{v_m\}_{m=1}^M$ of deterministic unit vectors.

Proof. We first tackle $\mathcal{A}_1$. For any $v \in \mathbb{R}^p$ such that $\|v\|_2 = 1$, we can write

$$v^\top(\widetilde{S}_t^{s,e} - \widetilde{\Sigma}_t^{s,e})v = \sum_{i=s+1}^{e} a_i \left((v^\top X_i)^2 - \mathbb{E}((v^\top X_i)^2)\right) = \sum_{i=s+1}^{e} a_i Z_i,$$

where

$$a_i = \begin{cases} \sqrt{\frac{e-t}{(e-s)(t-s)}} & s+1 \leq i \leq t, \\ \sqrt{\frac{t-s}{(e-s)(e-t)}} & t+1 \leq i \leq e, \end{cases}$$

and $Z_i = (v^\top X_i)^2 - \mathbb{E}[(v^\top X_i)^2]$. Since $\min\{t-s, e-t\} \geq p\log(n)$, we further have that

$$\sum_{i=s+1}^{e} a_i^2 = 1, \quad \text{and} \quad \max_{s+1 \leq i \leq e} |a_i| \leq 1/\sqrt{p\log(n)}.$$

Thus by Proposition 5.16 in Vershynin (2010), for any $\epsilon \geq 0$,

$$\mathbb{P}\left(\left|\sum_{i=s+1}^{e} a_i Z_i\right| \geq \epsilon\right) \leq 2\exp\left(-c\min\left\{\frac{\epsilon^2}{K^2}, \frac{\epsilon\sqrt{p\log(n)}}{K}\right\}\right),$$

where $c > 0$ is an absolute constant and

$$K = \max_i \|Z_i\|_{\psi_1} \leq 2\|(v^\top X_i)^2\|_{\psi_1} \leq 4\|X_i\|_{\psi_2}^2 \leq 4B^2.$$

Therefore,

$$\mathbb{P}\Big(\Big|v^\top(\widetilde{S}_t^{s,e} - \widetilde{\Sigma}_t^{s,e})v\Big| \geq B^2\sqrt{p\log(n)}\Big) \leq 2n^{-cp}.$$

Let $\mathcal{N}_{1/4}$ be a minimal 1/4-net (with respect to the Eucidean norm) of the unit sphere in $\mathbb{R}^p$. Then, $\mathrm{card}(\mathcal{N}_{1/4}) \leq 9^p$ and, by a standard covering argument followed by a union bound, we arrive at the inequality

$$\mathbb{P}(\mathcal{A}_1(\{X_i\}_{i=1}^n, B^2\sqrt{p\log(n)})) \geq 1 - 2 \times 9^p n^3 n^{-cp},$$

for a universal constant $c > 0$. Following the same arguments we have,

$$\mathbb{P}(\mathcal{A}_2(\{X_i\}_{i=1}^n, B^2\sqrt{p\log(n)})) \geq 1 - 2 \times 9^p n^2 n^{-cp},$$
$$\mathbb{P}(\mathcal{B}_1(\{X_i\}_{i=1}^n, \{v_m\}_{m=1}^M, B^2\sqrt{\log(n)})) \geq 1 - 2n^3 M n^{-c},$$

and

$$\mathbb{P}(\mathcal{B}_2(\{X_i\}_{i=1}^n, \{v_m\}_{m=1}^M, B^2\sqrt{\log(n)})) \geq 1 - 2n^2 M n^{-c},$$

for some $c > 0$. □

Let $\{s_m\}_{m=1}^M, \{e_m\}_{m=1}^M$ be two sequences independently selected at random in $[s, e]$, and

$$\mathcal{M} = \bigcap_{k=1}^{K} \big\{ s_m \in \mathcal{S}_k, e_m \in \mathcal{E}_k, \text{ for some } m \in \{1, \ldots, M\} \big\}, \tag{20}$$

where $\mathcal{S}_k = [\eta_k - 3\Delta/4, \eta_k - \Delta/2]$ and $\mathcal{E}_k = [\eta_k + \Delta/2, \eta_k + 3\Delta/4]$, $k = 1, \ldots, K$. In the lemma below, we give a lower bound on the probability of $\mathcal{M}$. Under the scaling assumed in our setting, this bound approaches 1 as n grows.

Lemma 6. *For the event $\mathcal{M}$ defined in (20), we have*

$$\mathbb{P}(\mathcal{M}) \geq 1 - \exp\left(\log \frac{n}{\Delta} - M \frac{\Delta^2}{16n^2} \right).$$

Proof. Since the number of change points are bounded by n/Δ,

$$\mathbb{P}\big\{\mathcal{M}^c\big\} \leq \sum_{k=1}^{K} \prod_{m=1}^{M} \big(1 - P\big(s_m \in \mathcal{S}_k, e_m \in \mathcal{E}_k\big)\big) \leq K(1 - \Delta^2/(16n^2))^M \leq n/\Delta(1 - \Delta^2/(16n^2))^M.$$

□

B Properties of the univariate CUSUM statistics

In this section, we derive some important and useful properties of the univariate CUSUM statistic. Our results and proofs build upon the existing literature on univariate mean change point detection; see in partiucular, Venkatraman (1992) and Fryzlewicz (2014), whose notation will be used throughout. It is imprtnat however to note that we have made several non-trivial modifications of those arguments, and have made a special effort in keeping track of the changes in all the key parameters. This careful treatment eventually allows us to achieve tight upper bounds for the the localization rate implied by the WBSIP algorithm and which in turn have revealed a phase transition in the problem parameters (see Section 3). In particular, the results of this section can be used to sharpen existing analyses of the BS and WBS algorithms.

B.1 Results from Venkatraman (1992)

We start by introducing some notations for one dimensional change point detection and the corresponding CUSUM statistics. Let $\{Y_i\}_{i=1}^n, \{f_i\}_{i=1}^n \subset \mathbb{R}$ be two univariate,

sequences. We will make the following assumptions on.

Assumption 5 (Univariate mean change points). *Let $\{\eta_k\}_{k=0}^{K+1} \subset \{0, \ldots, n\}$, where $\eta_0 = 0$ and $\eta_{K+1} = n$, and*

$$f_{\eta_{k-1}+1} = f_{\eta_{k-1}+2} = \ldots = f_{\eta_k} \quad \textit{for all} \quad 1 \leq k \leq K+1,$$

Assume

$$\inf_{k=1,\ldots,K+1} \{\eta_k - \eta_{k-1}\} \geq \Delta = \Delta(n) > 0,$$
$$|f_{\eta_{k+1}} - f_{\eta_k}| := \kappa_k > 0, \; k = 1, \ldots, K,$$
$$\sup_{k=1,\ldots,K+1} |f_{\eta_k}| < B_1.$$

We also have the corresponding CUSUM statistics over any generic interval $[s,e] \subset [1,T]$ defined as

$$\widetilde{Y}_t^{s,e} = \sqrt{\frac{e-t}{(e-s)(t-s)}} \sum_{i=s+1}^{t} Y_i - \sqrt{\frac{t-s}{(e-s)(e-t)}} \sum_{i=t+1}^{e} Y_i,$$

$$\widetilde{f}_t^{s,e} = \sqrt{\frac{e-t}{(e-s)(t-s)}} \sum_{i=s+1}^{t} f_i - \sqrt{\frac{t-s}{(e-s)(e-t)}} \sum_{i=t+1}^{e} f_i.$$

Throughout this Appendix B, all of our results are proven by regarding $\{Y_i\}_{i=1}^T$ and $\{f_i\}_{i=1}^T$ as two deterministic sequences. We will frequently assume that $\widetilde{f}_t^{s,e}$ is a good approximation of $\widetilde{Y}_t^{s,e}$ in ways that we will specify through appropriate assumptions.

Observe that the function $\widetilde{f}_t^{s,e}$ is only well defined on $[s,e] \cap \mathbb{Z}$. Our first result, which is taken from Venkatraman (1992), shows that the there exists a continuous realization of the discrete function $\widetilde{f}_t^{s,e}$

Lemma 7. *Suppose $[s,e] \subset [1,T]$ satisfies*

$$\eta_{r-1} \leq s \leq \eta_r \leq \ldots \leq \eta_{r+q} \leq e \leq \eta_{r+q+1}, \quad q \geq 0.$$

Then there exists a continuous function $\widetilde{F}_t^{s,e} : [s,e] \to \mathbb{R}$ such that $\widetilde{F}_r^{s,e} = \widetilde{f}_r^{s,e}$ for every $r \in [s,e] \cap \mathbb{Z}$ with the following additional properties.

1. *$|\widetilde{F}_t^{s,e}|$ is maximized at the change points within $[s,e]$. In other words,*

$$\arg\max_{s \leq t \leq e} |\widetilde{F}_t^{s,e}| \cap \{\eta_r, \ldots, \eta_{r+q}\} \neq \emptyset.$$

2. *If $\widetilde{F}_t^{s,e} > 0$ for some $t \in (s,e)$, then $\widetilde{F}_t^{s,e}$ is either monotonic or decreases and then increases within each of the interval $[s,\eta_r], \ldots, [\eta_{r+q}, e]$.*

The proof of this lemma can be found in Lemma 2.2 and 2.3 of Venkatraman (1992). We remark that if $\widetilde{F}_t^{s,e} \leq 0$ for all $t \in (s,e)$, then it suffices to consider the time series $\{-f_i\}_{i=1}^T$ and a similar result as in the second part of Lemma 7 still holds. Throughout the entire section, we always view $\widetilde{f}_t^{s,e}$ as a continuous function and frequently invoke Lemma 7 as a basic property of the CUSUM statistics without further notice.

Our next lemma is an adaptation of a result first obtained by Venkatraman (1992), which quantifies characterizes how fast the CUSUM statistics decays around a good change point. An analogous result, derive using different arguments, can be found in Proposition 21 in Wang and Samworth (2016).

Lemma 8 (Venkatraman (1992) Lemma 2.6)**.** *Let $[s,e] \subset [1,T]$ be any generic interval. For some $c_1, c_2 > 0$ and $\lambda > 0$ such that*

$$\min\{\eta_k - s, e - \eta_k\} \geq c_1 \Delta \tag{21}$$

$$\widetilde{f}_{\eta_k}^{s,e} \geq c_2 \kappa \Delta (e-s)^{-1/2}, \tag{22}$$

suppose there exists sufficient small c_3 such that

$$\max_{s \leq t \leq e} |\widetilde{f}_t^{s,e}| - \widetilde{f}_{\eta_k}^{s,e} \leq 2\lambda \leq c_3 \kappa \Delta^3 (e-s)^{-5/2} \tag{23}$$

Then there exists an absolute constant $c > 0$ such that if the point $d \in [s, e]$ is such that $|d - \eta_k| \le c_1\Delta/16$, then

$$\widetilde{f}^{s,e}_{\eta_k} - \widetilde{f}^{s,e}_{d} > c\widetilde{f}^{s,e}_{\eta_k}|\eta_k - d|\Delta(e-s)^{-2}$$

Remark 10. *If $\widetilde{f}^{s,e}_{\eta_k} < 0$, and $d \in [s, e]$ is such that $|d - \eta_k| \le c_1/16$, then by considering the sequence $\{-f_i\}_{i=1}^n$, it holds that*

$$(-\widetilde{f}^{s,e}_{\eta_k}) - (-\widetilde{f}^{s,e}_{d}) > c(-\widetilde{f}^{s,e}_{\eta_k})|\eta_k - d|\Delta(e-s)^{-2}$$

Proof. Without loss of generality, assume that $d \ge \eta_k$. Following the argument of Venkatraman (1992) Lemma 2.6, it suffices to consider two cases: (1) $\eta_{k+1} > e$, and (2) $\eta_{k+1} \le e$.

Case 1. Let E_l be defined as in the case 1 in Venkatraman (1992) Lemma 2.6. There exists a $c > 0$ such that, for every $d \in [\eta_k, \eta_k + c_1\Delta/16]$, $\widetilde{f}^{s,e}_{\eta_k} - \widetilde{f}^{s,e}_{d}$ (which in the notation of Venkatraman (1992) is the term E_l) can be written as

$$\widetilde{f}^{s,e}_{\eta_k}|d-\eta_k|\frac{e-s}{\sqrt{e-\eta_k}\sqrt{\eta_k - s + (d-\eta_k)}\left(\sqrt{(\eta_k - s + (d-\eta_k))(e-\eta_k)} + \sqrt{(\eta_k - s)(e - \eta_k - (d-\eta_k))}\right)}.$$

Using the inequality $(e - s) \ge 2c_1\Delta$, the previous expression is lower bounded by

$$\ge c'|d-\eta_k|\widetilde{f}^{s,e}_{\eta_k}\Delta(e-s)^{-2}.$$

Case 2. Let $h = c_1\Delta/8$ and $l = d - \eta_k \le h/2$. Then, following closely the initial calculations for case 2 of Lemma 2.6 of Venkatraman (1992), we obtain that

$$\widetilde{f}^{s,e}_{\eta_k} - \widetilde{f}^{s,e}_{d} \ge E_{1l}(1 + E_{2l}) + E_{3l},$$

where

$$E_{1l} = \frac{\widetilde{f}^{s,e}_{\eta_k} l(h-l)}{\sqrt{(\eta_k - s + l)(e - \eta_k - l)}\left(\sqrt{(\eta_k - s + l)(e-\eta_k - l)} + \sqrt{(\eta_k - s)(e-\eta_k)}\right)},$$

$$E_{2l} = \frac{((e-\eta_k - h) - (\eta_k - s))((e-\eta_k-h) - (\eta_k - s) - l)}{\left(\sqrt{(\eta_k - s + l)(e-\eta_k - l)} + \sqrt{(\eta_k - s + h)(e - \eta_k - h)}\right)\left(\sqrt{(\eta_k - s)(e-\eta_k)} + \sqrt{(\eta_k - s + h)(e-\eta_k - h)}\right)}$$

and

$$E_{3l} = -\frac{(\widetilde{f}^{s,e}_{\eta_k + h} - \widetilde{f}^{s,e}_{\eta_k})l}{h}\sqrt{\frac{(\eta_k - s + h)(e - \eta_k - h)}{(\eta_k - s + l)(e-\eta_k - l)}}.$$

Since $h = c''\Delta$ and $l \le h/2$,

$$E_{1l} \ge c''\widetilde{f}^{s,e}_{\eta_k}|d - \eta|\Delta(e-s)^{-2}.$$

Observe that

$$\eta_k - s \le \eta_k - s + l \le \eta_k - s + h \le 9(\eta_k - s)/8, \quad e - \eta_k \ge e - \eta_k - l \ge e - \eta_k - h \ge 7(e-\eta_k)/8. \tag{24}$$

Thus

$$
\begin{aligned}
&E_{2l}\\
=&\frac{((e-\eta_k-h)-(\eta_k-s))^2+l(h+\eta_k-s)-l(e-\eta_k)}{\left(\sqrt{(\eta_k-s+l)(e-\eta_k-l)}+\sqrt{(\eta_k-s+h)(e-\eta_k-h)}\right)\left(\sqrt{(\eta_k-s)(e-\eta_k)}+\sqrt{(\eta_k-s+h)(e-\eta_k-h)}\right)}\\
\geq&\frac{-l(e-\eta_k)}{(\eta_k-s+h)(e-\eta_k-h)}\\
\geq&\frac{-l(e-\eta_k)}{(\eta_k-s)(7/8)(e-\eta_k)}\geq -1/2
\end{aligned}
$$

where (24) is used in the second inequality and the fact that $l \leq h/2 \leq c_1\Delta/16 \leq (\eta_k - s)/16$ is used in the last inequality. For E_{3l}, observe that

$$\widetilde{f}^{s,e}_{\eta_k+h} - \widetilde{f}^{s,e}_{\eta_k} \leq |\widetilde{f}^{s,e}_{\eta_k+h}| - \widetilde{f}^{s,e}_{\eta_k} \leq \max_{s\leq t\leq e}|\widetilde{f}^{s,e}_t| - \widetilde{f}^{s,e}_{\eta_k} \leq 2\lambda$$

and that (21) and $l/2 \leq h = c_1\Delta/8$ imply that

$$\eta_k - s \leq \eta_k - s + l \leq \eta_k - s + h \leq 9(\eta_k - s)/8 \quad \text{and} \quad e - \eta_k \geq e - \eta_k - l \geq e - \eta_k - h \geq 7(e - \eta_k)/8.$$

Therefore,

$$
\begin{aligned}
E_{3l} &\geq -\frac{2(d-\eta_k)\lambda}{c_1\Delta/8}\sqrt{\frac{(9/8)(\eta_k-s)(e-\eta_k)}{(\eta_k-s)(7/8)(e-\eta_k)}}\\
&\geq -\frac{32(d-\eta_k)\lambda}{c_1\Delta}\\
&\geq -(c''/4)\widetilde{f}^{s,e}_{\eta_k}(d-\eta_k)\Delta(e-s)^{-2},
\end{aligned}
$$

where the first inequality follows from (24) and the last inequality follows from (22) and (23), for a sufficiently small c_3. Thus,

$$\widetilde{f}^{s,e}_{\eta_k} - \widetilde{f}^{s,e}_{d} \geq E_{1l}(1 + E_{2l}) + E_{3l} \geq (c''/4)\widetilde{f}^{s,e}_{\eta_k}|\eta_k - d|\Delta(e-s)^{-2}.$$

□

The following proposition is a direct consequence of Lemma 8 and essentially characterized the localization error rate of the BS algorithm.

Proposition 9. *Consider any generic interval $(s, e) \subset (0, T)$ such that*

$$\eta_{r-1} \leq s \leq \eta_r \leq \ldots \leq \eta_{r+q} \leq e \leq \eta_{r+q+1}, \quad q \geq 0. \tag{25}$$

Let $b \in \arg\max_{s\leq t\leq e}|\widetilde{Y}^{s,e}_t|$. Suppose for some $c_1 > 0$ and $\kappa > 0$,

$$\max\{\min\{\eta_r - s, s - \eta_{r-1}\}, \min\{\eta_{r+q+1} - e, e - \eta_{r+q}\}\} = \epsilon_n,$$

where

$$\epsilon_n < \min\{(3c_1/8)^2\kappa^2\Delta^2(e-s)^{-1}B_1^{-2}, \Delta/4\}, \tag{26}$$

and

$$|\widetilde{Y}_b^{s,e}| \geq c_1 \kappa \Delta (e-s)^{-1/2}. \tag{27}$$

Assume also that there exists sufficient small $c_3 > 0$ such that

$$\sup_{s \leq t \leq e} |\widetilde{f}_t^{s,e} - \widetilde{Y}_t^{s,e}| = \lambda_1 < \min\{c_3 \kappa \Delta^3 (e-s)^{-5/2}, \quad (c_1/4)\kappa\Delta(e-s)^{-1/2}\}. \tag{28}$$

Then there exists a change point $\eta_k \in [s,e]$ and an absolute constant $C_1 > 0$ such that

$$\min\{e-\eta_k, \eta_k - s\} > (3c_1/8)^2 \kappa^2 \Delta^2 (e-s)^{-1} B_1^{-2} \tag{29}$$

$$|\eta_k - b| \leq C_1 \lambda_1 (e-s)^{5/2} \Delta^{-2} \kappa^{-1},$$

and

$$|\widetilde{f}_{\eta_k}^{s,e}| \geq |\widetilde{Y}_b^{s,e}| - \lambda_1 \geq \max_{s \leq t \leq e} |\widetilde{f}_t^{s,e}| - 2\lambda_1. \tag{30}$$

Proof. Observe that from (28),

$$\max_{s<t<e} |\widetilde{f}_t^{s,e}| \leq \max_{s<t<e} |\widetilde{Y}_t^{s,e}| + \lambda_1 \leq |\widetilde{Y}_b^{s,e}| + \lambda_1 \leq |\widetilde{f}_b^{s,e}| + 2\lambda_1. \tag{31}$$

Suppose $\eta_k \leq b \leq \eta_{k+1}$ for some $r-1 \leq k \leq r+q$. Observe that

$$|\widetilde{f}_b^{s,e}| \geq |\widetilde{Y}_b^{s,e}| - \lambda_1 > (3c_1/4)\kappa\Delta(e-s)^{-1/2} > 0,$$

where the second inequality follows from (27) and (28). It suffices to consider the case in which $\widetilde{f}_b^{s,e} > 0$, since, if $\widetilde{f}_b^{s,e} < 0$, then the same arguments can be applied to the time series $\{-f_i\}_{i=1}^n$. From Lemma 7, $\widetilde{f}_t^{s,e}$ is either monotonic or decreasing and then increasing on $[\eta_k, \eta_{k+1}]$. Thus

$$\max\{\widetilde{f}_{\eta_k}^{s,e}, \widetilde{f}_{\eta_{k+1}}^{s,e}\} \geq \widetilde{f}_b^{s,e}.$$

If $\widetilde{f}_t^{s,e}$ is locally decreasing at b, then $\widetilde{f}_{\eta_k}^{s,e} \geq \widetilde{f}_b^{s,e}$. Therefore

$$\widetilde{f}_{\eta_k}^{s,e} \geq \widetilde{f}_b^{s,e} > (3c_1/4)\kappa\Delta(e-s)^{-1/2}. \tag{32}$$

Step 1. We first show that (32) implies (29). For the sake of contradiction, suppose that $\min\{e-\eta_k, \eta_k - s\} \leq ((3c_1/8)\kappa\Delta(e-s)^{-1/2}B_1^{-1})^2$. Then

$$\begin{aligned}\widetilde{f}_{\eta_k}^{s,e} &\leq \sqrt{\frac{(e-\eta_k)(\eta_k-s)}{e-s}}B_1 + \sqrt{\frac{(e-\eta_k)(\eta_k-s)}{e-s}}B_1 \\ &\leq 2\sqrt{\min\{e-\eta_k, \eta_k-s\}}B_1 \leq (3c_1/4)\kappa\Delta(e-s)^{-1/2}.\end{aligned}$$

This is a contradiction to (32). Therefore (29) holds for η_k.

Step 2. We now apply Lemma 8, since (22) and (23) hold in virtue of (32) and (28), respectively. Thus, we will need to prove that

$$\min\{\eta_k - s, e - \eta_k\} \geq (3/4)\Delta.$$

For the sake of contradiction, assume that $\min\{\eta_k - s, e - \eta_k\} < (3/4)\Delta$. Suppose $\eta_k - s < (3/4)\Delta$. Since $\eta_k - \eta_{k-1} \geq \Delta$, one has $s - \eta_{k-1} \geq \Delta/4$. This also means that η_k is the first change point within $[s, e]$. Therefore $k = r$ in (25). By (26), $\min\{\eta_r - s, s - \eta_{r-1}\} \leq \epsilon < \Delta/4$. Since $s - \eta_{k-1} = s - \eta_{p-1} \geq \Delta/4$, it must be the case that

$$\eta_k - s = \eta_r - s \leq \epsilon \leq (3c_1/8)^2 \kappa^2 \Delta^2 (e-s)^{-1} B_1^{-2}.$$

This is a contradiction to (29). Therefore $\eta_k - s \geq (3/4)\Delta$. The argument of $e - \eta_k \geq (3/4)\Delta$ can be made analogously.

Step 3. By Lemma 8, if there exists ad and a sufficient large constant C_1 satisfying

$$d \in [\eta_k, \eta_k + C_1 \lambda_1 (e-s)^{5/2} \Delta^{-2} \kappa^{-1}],$$

then

$$\widetilde{f}^{s,e}_{\eta_k} - \widetilde{f}^{s,e}_{d} > c \widetilde{f}^{s,e}_{\eta_k} |\eta_k - d| \Delta (e-s)^{-2} \geq \lambda_1,$$

where the last inequality follows from (32) and (28).
For the sake of contradiction, suppose $b \geq d$. Then

$$\widetilde{f}^{s,e}_{b} \leq \widetilde{f}^{s,e}_{d} < \widetilde{f}^{s,e}_{\eta_k} - \lambda_1 \leq \max_{s<t<e} |\widetilde{f}^{s,e}_{t}| - 2\lambda_1,$$

where the first inequality follows from Lemma 7 which ensures that $\widetilde{f}^{s,e}_{t}$ is decreasing on $[\eta_p, b]$ and $d \in [\eta_p, b]$. This is a contradiction to (31). Thus $b \in [\eta_k, \eta_k + C_1 \lambda_1 (e-s)^{5/2} \Delta^{-2} \kappa^{-1}]$.
(30) follows from (31) and (32).
The argument for the case when $\widetilde{f}^{s,e}_{b}$ is locally increasing at b is similar and therefore we omit the details.

□

Corollary 10. *Let $\delta > 0$ be such that $2\sqrt{\delta} B_1 \leq (3c_1/4)\kappa\Delta(e-s)^{-1/2}$. Let $b' = \arg\max_{\lceil s+\delta \rceil \leq t \leq \lfloor e-\delta \rfloor} |\widetilde{Y}^{s,e}_{t}|$. Suppose all the assumption in Proposition 9 hold except that (27) and (28) are replaced by*

$$|\widetilde{Y}^{s,e}_{b'}| \geq c_1 \kappa \Delta (e-s)^{-1/2} \tag{33}$$

and

$$\sup_{\lceil s+\delta \rceil, \ldots, \lfloor e-\delta \rfloor} |\widetilde{f}^{s,e}_{t} - \widetilde{Y}^{s,e}_{t}| = \lambda_1 < \min\{c_3 \kappa \Delta^3 (e-s)^{-5/2}, \quad (c_1/4)\kappa\Delta(e-s)^{-1/2}\}, \tag{34}$$

respectively. Then, all the conclusions of Proposition 9 still hold for b'.

Proof. For any $t \in [s+\delta] \cup [e-\delta]$,

$$|\widetilde{f}^{s,e}_{t}| \leq 2\sqrt{\min\{e-t, t-s\}} B_1 \leq 2\sqrt{\delta} B_1 \leq (3c_1/4)\kappa\Delta(e-s)^{-1/2}.$$

Let η_k and b defined as in the proof of Proposition 9. Then by (32) $\eta_k, b \in \lceil s+\delta \rceil, \ldots, \lfloor e-\delta \rfloor$. Therefore $b = b'$ and all the conclusions of Proposition 9 still hold.

□

B.2 Results from Fryzlewicz (2014)

Below, we will derive some further properties of the CUSUM statistic using the ANOVA decomposition-type of arguments first introduced by Fryzlewicz (2014), which are particularly effective at separating the noise from the signal. Some of the results below are contained in Fryzlewicz (2014), but others requires different, subtle arguments. For completeness, we include all the proofs.

For a pair (s, e) of positive integers with $s < e$, let $\mathcal{W}_d^{s,e}$ be the two dimensional linear subspace of $\mathbb{R}^{e-s}$ spanned by the vectors

$$u_1 = (\underbrace{1, \ldots, 1}_{d-s}, \underbrace{0, \ldots, 0}_{e-d})^\top \qquad u_2 = (\underbrace{0, \ldots, 0}_{d-s}, \underbrace{1, \ldots, 1}_{e-d})^\top.$$

For clarity, we will use $\langle \ , \ \rangle$ to denote the inner product of two vectors in the Euclidean space.

Lemma 11. *For $x = (x_{s+1}, \ldots, x_e)^\top \in \mathbb{R}^{e-s}$, let $\mathcal{P}_d^{s,e}(x)$ be the projection of x onto $\mathcal{W}_d^{s,e}$.*

1. *The projection $\mathcal{P}_d^{s,e}(x)$ satisfies*

$$\mathcal{P}_d^{s,e}(x) = \frac{1}{e-s} \sum_{i=s+1}^{e} x_i + \langle x, \psi_d^{s,e} \rangle \psi_d^{s,e},$$

where $\langle \cdot, \cdot \rangle$ is the inner product in Euclidean space, and $\psi_d^{s,e} = ((\psi_d^{s,e})_s, \ldots, (\psi_d^{s,e})_{e-s})^\top$ with

$$(\psi_d^{s,e})_i = \begin{cases} \sqrt{\frac{e-d}{(e-s)(d-s)}}, & i = s+1, \ldots, d, \\ -\sqrt{\frac{d-s}{(e-s)(e-d)}}, & i = d+1, \ldots, e, \end{cases}$$

i.e. the i-th entry of $\mathcal{P}_d^{s,e}(x)$ satisfies

$$\mathcal{P}_d^{s,e}(x)_i = \begin{cases} \frac{1}{d-s} \sum_{j=s+1}^{d} x_j, & i = s+1, \ldots, d, \\ \frac{1}{e-d} \sum_{j=d+1}^{e} x_j, & i = d+1, \ldots, e. \end{cases}$$

2. *Let $\bar{x} = \frac{1}{e-s} \sum_{i=s+1}^{e} x_i$. Since $\langle \bar{x}, \psi_d^{s,e} \rangle = 0$,*

$$\|x - \mathcal{P}_d^{s,e}(x)\|^2 = \|x - \bar{x}\|^2 - \langle x, \psi_d^{s,e} \rangle^2. \tag{35}$$

Proof. The results hold following the fact that the projection matrix of subspace $\mathcal{W}_d^{s,e}$ is

$$P_{\mathcal{W}_d^{s,e}}^{s,e} = \begin{pmatrix} 1/(d-s) & \cdots & 1/(d-s) & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1/(d-s) & \cdots & 1/(d-s) & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1/(e-d) & \cdots & 1/(e-d) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & 1/(e-d) & \cdots & 1/(e-d) \end{pmatrix}.$$

□

For any pair $d_1, d_2 \in \{s+1, \ldots, e\}$ and $f \in \mathbb{R}^{e-s}$, the following two statements are equivalent:

$$\langle f, \psi_{d_1}^{s,e} \rangle^2 \le \langle f, \psi_{d_2}^{s,e} \rangle^2 \iff \|f - \mathcal{P}_{d_1}^{s,e}(f)\|^2 \ge \|f - \mathcal{P}_{d_2}^{s,e}(f)\|^2.$$

Lemma 12. *Assume Assumption 1. Let $[s_0, e_0]$ be an interval with $e_0 - s_0 \le C_R\Delta$ and containing at lest one change point η_r such that*

$$\eta_{r-1} \le s_0 \le \eta_r \le \ldots \le \eta_{r+q} \le e_0 \le \eta_{r+q+1}, \quad q \ge 0.$$

Suppose that $\min\{\eta_{p'} - s_0, e_0 - \eta_{p'}\} \ge \Delta/16$ for some p' and let $\kappa_{\max}^{s,e} = \max\{\kappa_p : \min\{\eta_p - s_0, e_0 - \eta_p\} \ge \Delta/16\}$. Consider any generic $[s,e] \subset [s_0, e_0]$, satisfying

$$\min\{\eta_r - s_0, e_0 - \eta_r\} \ge \Delta/16 \quad \textit{for all } \eta_r \in [s,e].$$

Let $b \in \arg\max_{s<t<e} |\widetilde{Y}_t^{s,e}|$. For some $c_1 > 0$, $\lambda > 0$ and $\delta > 0$, suppose that

$$|\widetilde{Y}_b^{s,e}| \ge c_1 \kappa_{\max}^{s,e} \sqrt{\Delta} \tag{36}$$

$$\sup_{s<t<e} |\widetilde{Y}_t^{s,e} - \widetilde{f}_t^{s,e}| \le \lambda$$

and

$$\sup_{s_1<t<e_1} \frac{1}{\sqrt{e_1 - s_1}} \left| \sum_{t=s_1+1}^{e_1} (Y_t - f_t) \right| \le \lambda \quad \textit{for every} \quad e_1 - s_1 \ge \delta. \tag{37}$$

If there exists a sufficient small $c_2 > 0$ such that

$$\lambda \le c_2 \kappa_{\max}^{s,e} \sqrt{\Delta} \quad \textit{and} \quad \delta \le c_2 \Delta, \tag{38}$$

then there exists a change point $\eta_k \in (s,e)$ such that

$$\min\{e - \eta_k, \eta_k - s\} > \Delta/4 \quad \textit{and} \quad |\eta_k - b| \le \min\{C_3 \lambda^2 \kappa_k^{-2}, \delta\}.$$

Proof. Without loss of generality, assume that $\widetilde{f}_b^{s,e} > 0$ and that $\widetilde{f}_t^{s,e}$ is locally decreasing at b. Observe that there has to be a change point $\eta_k \in [s,b]$, or otherwise $\widetilde{f}_b^{s,e} > 0$ implies that $\widetilde{f}_t^{s,e}$ is decreasing, as a consequence of Lemma 18.

Thus if $s \le \eta_k \le b \le e$, then

$$\widetilde{f}_{\eta_k}^{s,e} \ge \widetilde{f}_b^{s,e} \ge |\widetilde{Y}_b^{s,e}| - \lambda \ge c_1 \kappa_{\max}^{s,e} \sqrt{\Delta} - c_2 \kappa_{\max}^{s,e} \sqrt{\Delta} \ge (c_1/2) \kappa_{\max}^{s,e} \sqrt{\Delta}. \tag{39}$$

Observe that $e - s \le e_0 - s_0 \le C_R\Delta$ and that that (s,e) has to contain at least one change point or otherwise $|\widetilde{f}_{\eta_k}^{s,e}| = 0$ which contradicts (39).

Step 1. In this step, we are to show that $\min\{\eta_k - s, e - \eta_k\} \ge \min\{1, c_1^2\}\Delta/16$.

Suppose η_k is the only change point in (s,e). So $\min\{\eta_k - s, e - \eta_k\} \ge \min\{1, c_1^2\}\Delta/16$ must hold or otherwise it follows from Lemma 17, we have

$$|\widetilde{f}_{\eta_k}^{s,e}| < \frac{c_1}{4} \kappa_k \sqrt{\Delta} \le \frac{c_1}{2} \kappa_{\max}^{s,e} \sqrt{\Delta},$$

which contradicts (39).

Suppose (s,e) contains at least two change points. Then $\eta_k - s \le \min\{1, c_1^2\}\Delta/16$ implies that η_k is the first change point in $[s,e]$. Therefore

$$\begin{aligned} |\widetilde{f}^{s,e}_{\eta_k}| &\le \frac{1}{4}|\widetilde{f}^{s,e}_{\eta_{k+1}}| + 2\kappa_r\sqrt{\eta_r - s} \le \frac{1}{4}\max_{s<t<e}|\widetilde{f}^{s,e}_t| + \frac{c_1}{2}\kappa_r\sqrt{\Delta} \\ &\le \frac{1}{4}|\widetilde{Y}^{s,e}_b| + \lambda + \frac{c_1}{2}\kappa^{s,e}_{\max}\sqrt{\Delta} \le \frac{3}{4}|\widetilde{Y}^{s,e}_b| + \lambda < |\widetilde{Y}^{s,e}_b| - \lambda \end{aligned}$$

where the first inequality follows from Lemma 18, the fourth inequality follows from (36), and the last inequality holds when c_2 is sufficiently small. This contradicts (39).

Step 2. By Lemma 8 there exists d such that

$$d \in [\eta_k, \eta_k + \lambda\sqrt{\Delta}(\kappa^{s,e}_{\max})^{-1}]$$

and that $\widetilde{f}^{s,e}_{\eta_k} - \widetilde{f}^{s,e}_d > 2\lambda$. For the sake of contradiction, suppose $b \ge d$. Then

$$\widetilde{f}^{s,e}_b \le \widetilde{f}^{s,e}_d < \widetilde{f}^{s,e}_{\eta_k} - 2\lambda \le \max_{s<t<e}|\widetilde{f}^{s,e}_t| - 2\lambda \le \max_{s<t<e}|\widetilde{Y}^{s,e}_t| + \lambda - 2\lambda = |\widetilde{Y}^{s,e}_b| - \lambda,$$

where the first inequality follows from Lemma 7, which ensures that $\widetilde{f}^{s,e}_t$ is decreasing on $[\eta_p, b]$ and $d \in [\eta_p, b]$. This is a contradiction to (39). Thus $b \in [\eta_k, \eta_k + \lambda\sqrt{\Delta}(\kappa^{s,e}_{\max})^{-1}]$.

Step 3. Let $f^{s,e} = (f_{s+1}, \ldots, f_e)^\top \in \mathbb{R}^{(e-s)}$ and $Y^{s,e} = (Y_{s+1}, \ldots, Y_e)^\top \in \mathbb{R}^{(e-s)}$. By the definition of b, it holds that

$$\left\|Y^{s,e} - \mathcal{P}^{s,e}_b(Y^{s,e})\right\|^2 \le \left\|Y^{s,e} - \mathcal{P}^{s,e}_{\eta_k}(Y^{s,e})\right\|^2 \le \left\|Y^{s,e} - \mathcal{P}^{s,e}_{\eta_k}(f^{s,e})\right\|^2.$$

For the sake of contradiction, throughout the rest of this argument suppose that, for some sufficient large constant C_3 to be specified,

$$\eta_k + \max\{C_3\lambda^2\kappa_k^{-2}, \delta\} < b. \tag{40}$$

(This will of course imply that $\eta_k + \max\{C_3\lambda^2(\kappa^{s,e}_{\max})^{-2}, \delta\} < b$). We will show that this leads to the bound

$$\left\|Y^{s,e} - \mathcal{P}^{s,e}_b(Y^{s,e})\right\|^2 > \left\|Y^{s,e} - \mathcal{P}^{s,e}_{\eta_k}(f^{s,e})\right\|^2, \tag{41}$$

which is a contradiction. To derive (41) from (40), we note that $\min\{e-\eta_k, \eta_k - s\} \ge \min\{1, c_1^2\}\Delta/16$ and that $|b - \eta_k| \le \lambda\sqrt{\Delta}(\kappa^{s,e}_{\max})^{-1}$ implies that

$$\min\{e-b, b-s\} \ge \min\{1, c_1^2\}\Delta/16 - \lambda\sqrt{\Delta}(\kappa^{s,e}_{\max})^{-1} \ge \min\{1, c_1^2\}\Delta/32, \tag{42}$$

where the last inequality follows from (38) and holds for an appropriately small $c_2 > 0$.

Equation (41) is in turn implied by

$$2\langle \varepsilon^{s,e}, \mathcal{P}_b(Y^{s,e}) - \mathcal{P}_{\eta_k}(f^{(s,e)})\rangle < \|f^{s,e} - \mathcal{P}_b(f^{s,e})\|^2 - \|f^{s,e} - \mathcal{P}_{\eta_k}(f^{s,e})\|^2, \tag{43}$$

where $\varepsilon^{s,e} = Y^{s,e} - f^{s,e}$. By (35), the right hand side of (43) satisfied the relationships

$$\begin{aligned} \|f^{s,e} - \mathcal{P}_b(f^{s,e})\|^2 - \|f^{s,e} - \mathcal{P}_{\eta_k}(f^{s,e})\|^2 &= \langle f^{s,e}, \psi_{\eta_k}\rangle^2 - \langle f^{s,e}, \psi_b\rangle^2 \\ &= (\widetilde{f}^{s,e}_{\eta_p})^2 - (\widetilde{f}^{s,e}_b)^2 \\ &\ge (\widetilde{f}^{s,e}_{\eta_k} - \widetilde{f}^{s,e}_b)|\widetilde{f}^{s,e}_{\eta_k}| \\ &\ge c|d - \eta_k|(\widetilde{f}^{s,e}_{\eta_k})^2\Delta^{-1} \\ &\ge c'|d - \eta_k|(\kappa^{s,e}_{\max})^2, \end{aligned}$$

where Lemma 8 and (39) are used in the second and third inequalities. The left hand side of (43) can in turn be rewritten as

$$2\langle \varepsilon^{s,e}, \mathcal{P}_b(X^{s,e}) - \mathcal{P}_{\eta_k}(f^{s,e})\rangle = 2\langle \varepsilon^{s,e}, \mathcal{P}_b(X^{s,e}) - \mathcal{P}_b(f^{s,e})\rangle + 2\langle \varepsilon^{s,e}, \mathcal{P}_b(f^{s,e}) - \mathcal{P}_{\eta_k}(f^{s,e})\rangle. \quad (44)$$

The second term on the right hand side of the previous display can be decomposed as

$$\begin{aligned}\langle \varepsilon^{s,e}, \mathcal{P}_b(f^{s,e}) - \mathcal{P}_{\eta_k}(f^{s,e})\rangle &= \left(\sum_{i=s+1}^{\eta_k} + \sum_{i=\eta_k+1}^{b} + \sum_{i=b+1}^{e}\right) \varepsilon_i^{s,e} \left(\mathcal{P}_b(f^{s,e})_i - \mathcal{P}_{\eta_k}(f^{s,e})_i\right) \\ &= I + II + III.\end{aligned}$$

In order to bound the terms I, II and III, observe that, since $e - s \leq e_0 - s_0 \leq C_R\Delta$, the interval $[s, e]$ must contain at most $C_R + 1$ change points. Let

$$\eta_{r'-1} < s \leq \eta_{r'} \leq \ldots \leq \eta_{p'+q'} < e \leq \eta_{p'+q'+1}.$$

Then $p' + q' + 1 - r' \leq C_R + 1$.

Step 4. We can write

$$I = \sqrt{\eta_k - s}\left(\frac{1}{\sqrt{\eta_k - s}}\sum_{i=s+1}^{\eta_k} \varepsilon_i^{s,e}\right)\left(\frac{1}{b-s}\sum_{i=s+1}^{b} f_i - \frac{1}{\eta_k - s}\sum_{i=s+1}^{\eta_k} f_i\right).$$

Thus,

$$\begin{aligned}&\left|\frac{1}{b-s}\sum_{i=s+1}^{b} f_i - \frac{1}{\eta_k - s}\sum_{i=s+1}^{\eta_k} f_i\right| = \left|\frac{(\eta_k - s)(\sum_{i=s+1}^{\eta_k} f_i + \sum_{i=\eta_k+1}^{b} f_i) - (b-s)\sum_{i=s+1}^{\eta_k} f_i}{(b-s)(\eta_k - s)}\right| \\ =&\left|\frac{(\eta_k - b)\sum_{i=s+1}^{\eta_k} f_i + (\eta_k - s)\sum_{i=\eta_k+1}^{b} f_i)}{(b-s)(\eta_k - s)}\right| = \left|\frac{(\eta_k - b)\sum_{i=s+1}^{\eta_k} f_i + (\eta_k - s)(b - \eta_k) f_{\eta_k+1})}{(b-s)(\eta_k - s)}\right| \\ =&\frac{b-\eta_k}{b-s}\left|-\frac{1}{\eta_k - s}\sum_{i=s+1}^{\eta_k} f_i + f_{\eta_{k+1}}\right| \leq \frac{b-\eta_k}{b-s}(C_R+1)\kappa_{\max}^{s,e}\end{aligned}$$

where Lemma 16 is used in the last inequality. It follows from Equation (37) that

$$|I| \leq \sqrt{\eta_k - s}\lambda\frac{|b-\eta_k|}{b-s}(C_R+1)\kappa_{\max}^{s,e} \leq \frac{4\sqrt{2}}{\min\{1, c_1\}}|b - \eta_k|\Delta^{-1/2}\lambda(C_R+1)\kappa_{\max}^{s,e},$$

where (42) is used in the last inequality.

Step 5. For the second term II, we have that

$$\begin{aligned}|II| =& \left|\sqrt{b-\eta_k}\left(\frac{1}{\sqrt{b-\eta_k}}\sum_{i=\eta_k+1}^{d}\varepsilon_i^{s,e}\right)\left(\frac{1}{b-s}\sum_{i=s+1}^{b} f_i - \frac{1}{e-\eta_k}\sum_{i=\eta_k+1}^{e} f_i\right)\right| \\ \leq& \sqrt{b-\eta_k}\lambda\left(\left|f_{\eta_k} - f_{\eta_{k+1}}\right| + \left|\frac{1}{b-s}\sum_{i=s+1}^{b} f_i - f_{\eta_k}\right| + \left|\frac{1}{e-\eta_k}\sum_{i=\eta_k+1}^{e} f_i - f_{\eta_{k+1}}\right|\right) \\ \leq& \sqrt{b-\eta_k}(\kappa_{\max}^{s,e} + (C_R+1)\kappa_{\max}^{s,e} + (C_R+1)\kappa_{\max}^{s,e}),\end{aligned}$$

where the first inequality follows from (42) and (37), and the second inequality from Lemma 16.

Step 6. Finally, we have that

$$III = \sqrt{e-b}\left(\frac{1}{e-b}\sum_{i=b+1}^{e}\varepsilon_i^{s,e}\right)\left(\frac{1}{e-\eta_k}\sum_{i=\eta_k+1}^{e}f_i - \frac{1}{e-b}\sum_{i=b+1}^{e}f_i\right).$$

Therefore,

$$|III| \leq \sqrt{e-b}\lambda\frac{b-\eta_k}{e-b}(C_R+1)\kappa_{\max}^{s,e} \leq \frac{4\sqrt{2}}{\min\{1,c_1\}}|b-\eta_k|\Delta^{-1/2}\lambda(C_R+1)\kappa_{\max}^{s,e}.$$

Step 7. Using the first part of Lemma 11, the first term on the right hand side of (44) can be bounded as

$$\langle \varepsilon^{s,e}, \mathcal{P}_d(X^{s,e}) - \mathcal{P}_d(f^{s,e})\rangle \leq \lambda^2$$

. Thus (43) holds if

$$|b-\eta_k|(\kappa_{\max}^{s,e})^2 \geq C\max\left\{|b-\eta_k|\Delta^{-1/2}\lambda\kappa_{\max}^{s,e}, \quad \sqrt{b-\eta_k}\lambda\kappa_{\max}^{s,e}, \quad \lambda^2\right\}.$$

Since $\lambda \leq c_3\sqrt{\Delta}\kappa$, the first inequality holds. The second inequality follows from $|b-\eta_k| \geq C_3\lambda^2(\kappa_k)^{-2} \geq C_3\lambda^2(\kappa_{\max}^{s,e})^{-2}$, as assumed in (40). This completes the proof. ☐

Corollary 13. *Let $[s_0,e_0]$ be a generic interval satisfying $e_0 - s_0 \leq C_R\Delta$ and containing at lest one change point η_r such that*

$$\eta_{r-1} \leq s_0 \leq \eta_r \leq \ldots \leq \eta_{r+q} \leq e_0 \leq \eta_{r+q+1}, \quad q \geq 0.$$

Suppose $\min\{\eta_{p'}-s_0, e_0-\eta_{p'}\} \geq \Delta/16$ *for some p' and, let* $\kappa_{\max}^{s,e} = \max\{\kappa_p : \min\{\eta_p - s_0, e_0-\eta_p\} \geq \Delta/16\}$*. Consider a generic interval $(s,e) \subset (s_0,e_0)$, satisfying*

$$\min\{\eta_p - s_0, e_0 - \eta_p\} \geq \Delta/16 \quad \textit{for all } \eta_p \in [s,e].$$

Let $\delta' > 0$ be some constant and $b' = \arg\max_{\lceil s+\delta'\rceil,\ldots,\lfloor e-\delta'\rfloor}|\widetilde{Y}_t^{s,e}|$. Suppose in addition that, for some positive constants c_1 and c_2,

$$|\widetilde{Y}_{b'}^{s,e}| \geq c_1\kappa_{\max}^{s,e}\sqrt{\Delta}, \tag{45}$$

$$\sup_{t=\lceil s+\delta'\rceil,\ldots,\lfloor e-\delta'\rfloor}|\widetilde{Y}_t^{s,e} - \widetilde{f}_t^{s,e}| \leq \lambda, \tag{46}$$

$$\sup_{s_1\leq t\leq e_1}\frac{1}{\sqrt{e_1-s_1}}|\sum_{t=s_1+1}^{e_1}(Y_t-f_t)| \leq \lambda, \quad \textit{for every} \quad e_1-s_1\geq\delta', \tag{47}$$

$$\lambda \leq c_2\kappa_{\max}^{s,e}\sqrt{\Delta}, \tag{48}$$

and

$$\delta' \leq c_2\Delta. \tag{49}$$

Then there exists a change point $\eta_k \in [s,e]$ such that

$$\min\{e-\eta_k, \eta_k - s\} > \Delta/4$$
$$|\eta_k - b'| \leq \max\{C_3\lambda^2\kappa_k^{-2}, \delta'\}.$$

Proof. By the same proof of Corollary 10, if b is defined as in Lemma 12, then $b = b'$ if c_2 is sufficiently small. ☐

C Proofs of the Results from Section 3

Proof of Lemma 3. For any vector $u \in \mathbb{R}^p$, denote $\widetilde{\Sigma}_u = \sigma^2 I_p + \kappa uu'$. Observe that if $\kappa \le \sigma^2/4$, $\|\widetilde{\Sigma}_u\|_{\text{op}} = \sigma^2 + \kappa \le 2\sigma^2$.

Step 1.

Let $\widetilde{P}^n_{0,u}$ denote the joint distribution of independent random vectors $\{X_i\}_{i=1}^n$ in R^p such that

$$X_1, \dots, X_\Delta \overset{i.i.d.}{\sim} N_p(0, \widetilde{\Sigma}_u) \quad \text{and} \quad X_{\Delta+1}, \dots, X_n \overset{i.i.d.}{\sim} N_p(0, \sigma^2 I).$$

Similarly, let $\widetilde{P}^n_{1,u}$ denote the joint distribution of independent random vectors $\{X_i\}_{i=1}^n$ in R^p with

$$X_1, \dots, X_{n-\Delta} \overset{i.i.d.}{\sim} N_p(0, \sigma^2 I) \quad \text{and} \quad X_{n-\Delta+1}, \dots, X_n \overset{i.i.d.}{\sim} N_p(0, \widetilde{\Sigma}_u).$$

Let $\widetilde{P}^n_i = \frac{1}{2^p} \sum_{u \in \{\pm 1\}^p/\sqrt{p}} \widetilde{P}^n_{i,u}$. Let $\eta(P^n_{i,u})$ denote the location of the change point associated to the distribution $\widetilde{P}^n_{i,u}$. Then since $\eta(\widetilde{P}^n_{0,u}) = \Delta$ and $\eta(\widetilde{P}^n_{1,u}) = n - \Delta$ for any $u \in \{\pm 1\}^p/\sqrt{p}$, $|\eta(\widetilde{P}^n_{0,u}) - \eta(\widetilde{P}^n_{1,u})| \ge n/3$. By Le cam's lemma (see, e.g. Yu, 1997),

$$\inf_{\hat{\eta}} \sup_{P \in \mathcal{P}^n_{\widetilde{\kappa},\Delta,\sigma}} E_P(|\hat{\eta} - \eta|) \ge (n/3)(1 - d_{TV}(\widetilde{P}^n_0, \widetilde{P}^n_1)),$$

where $TV(\widetilde{P}^n_0, \widetilde{P}^n_1) = \frac{1}{2}\|\widetilde{P}^n_0 - \widetilde{P}^n_1\|_1$.

Let $\Sigma_u = I_p + \widetilde{\kappa} uu'$, where $\widetilde{\kappa} = \kappa/\sigma^2$. Observe that by assumption $\widetilde{\kappa} \le 1/4$. Denote with $P^n_{0,u}$ and $P^n_{1,u}$ the joint distributions of independent samples $\{X_i\}_{i=1}^n$ in R^p where

$$X_1, \dots, X_\Delta \overset{i.i.d.}{\sim} N_p(0, \Sigma_u) \quad \text{and} \quad X_{\Delta+1}, \dots, X_n \overset{i.i.d.}{\sim} N_p(0, I)$$

and

$$X_1, \dots, X_{n-\Delta} \overset{i.i.d.}{\sim} N_p(0, I) \quad \text{and} \quad X_{n-\Delta+1}, \dots, X_n \overset{ii.d.}{\sim} N_p(0, \Sigma_u),$$

respectively. Since total variation distance is invariant under rescaling of the covariance, then $\|\widetilde{P}^n_0 - \widetilde{P}^n_1\|_1 = \|P^n_0 - P^n_1\|_1$. Therefore

$$\inf_{\hat{\eta}} \sup_{P \in \mathcal{P}^n_{\widetilde{\kappa},\Delta,\sigma}} E_P(|\hat{\eta} - \eta|) \ge (n/3)(1 - d_{TV}(P^n_0, P^n_1)) \tag{50}$$

Step 2.

Let $x = (x_1, \dots, x_\Delta)$, $y = (x_{\Delta+1}, \dots, x_{n-\Delta})$ and $z = (x_{n-\Delta+1}, \dots, x_n)$. Let

- $f_0(x)$ denote the joint distribution of $X_1, \dots, X_\Delta \sim N_p(0, I)$ and $f_u(x)$ denote the joint distribution of $X_1, \dots, X_\Delta \overset{i.i.d.}{\sim} N_p(0, \Sigma_u)$;
- $g_0(x)$ denote the joint distribution of $X_{\Delta+1}, \dots, X_{n-\Delta} \overset{i.i.d.}{\sim} N_p(0, I)$;
- $h_0(x)$ denote the joint distribution of $X_{n-\Delta+1}, \dots, X_n \overset{i.i.d.}{\sim} N_p(0, I)$ and $h_u(x)$ denote the joint distribution of $X_{n-\Delta+1}, \dots, X_n \overset{i.i.d.}{\sim} N_p(0, \Sigma_u)$.

Then,

$$
\begin{aligned}
&\|P_0^n - P_1^n\|_1 \\
=&\int\int\int \left| \frac{1}{2^p} \sum_{u\in\{\pm 1\}^p/\sqrt{p}} f_u(x)g_0(y)h_0(z) - \frac{1}{2^p} \sum_{u\in\{\pm 1\}^p/\sqrt{p}} f_0(x)g_0(y)h_u(z) \right| dxdydz \\
=&\int g_0(y)dy \int\int \left| \frac{1}{2^p} \sum_{u\in\{\pm 1\}^p/\sqrt{p}} f_u(x)h_0(z) - \frac{1}{2^p} \sum_{u\in\{\pm 1\}^p/\sqrt{p}} f_0(x)h_u(z) \right| dxdz \\
=&\int\int \left| \frac{1}{2^p} \sum_{u\in\{\pm 1\}^p/\sqrt{p}} f_u(x)h_0(z) - \frac{1}{2^p} \sum_{u\in\{\pm 1\}^p/\sqrt{p}} f_0(x)h_u(z) \right| dxdz \\
\le&\int\int \left| \frac{1}{2^p} \sum_{u\in\{\pm 1\}^p/\sqrt{p}} f_u(x)h_0(z) - f_0(x)h_0(z) \right| + \left| f_0(x)g_0(z) - \frac{1}{2^p} \sum_{u\in\{\pm 1\}^p/\sqrt{p}} f_0(x)h_u(z) \right| dxdz \\
=&2\|P_0^\Delta - P_1^\Delta\|_1,
\end{aligned}
$$

where P_0^Δ is the joint distribution of $X_1, \dots, X_\Delta \overset{i.i.d.}{\sim} N_p(0, I)$ and $P_1^\Delta = \frac{1}{2^p}\sum_{u\in\{\pm 1\}^p/\sqrt{p}} P_{1,u}^\Delta$, where $P_{1,u}^\Delta$ is the joint distribution of $X_1, \dots, X_\Delta \overset{i.i.d.}{\sim} N_p(0, \Sigma_u)$. Thus (50) becomes

$$
\inf_{\hat\eta} \sup_{P\in\mathcal{P}^n_{\widetilde\kappa,\Delta,\sigma}} E_P(|\hat\eta - \eta|) \ge (n/3)(1 - \|P_0^\Delta - P_1^\Delta\|_1). \tag{51}
$$

Step 3.
To bound $2\|P_0^\Delta - P_1^\Delta\|_1$, let $P_0 = N_p(0, I_p)$ and $P_u = N_p(0, \Sigma_u)$. It is easy to see that

$$
\chi^2(P_1^\Delta, P_0^\Delta) = E_{P_0^\Delta}\left(\frac{dP_1^\Delta}{dP_0^\Delta} - 1\right)^2 = \frac{1}{4^p}\sum_{u,v\in\{\pm 1\}^p} E_{P_0^\Delta}\left(\frac{dP_u^\Delta}{dP_0^\Delta}\frac{dP_v^\Delta}{dP_0^\Delta}\right) - 1.
$$

For any $u, v \in \{\pm 1\}^p$,

$$
E_{P_0^\Delta}\left(\frac{dP_u^\Delta}{dP_0^\Delta}\frac{dP_v^\Delta}{dP_0^\Delta}\right) = \left(E_{P_0}\left(\frac{dP_u}{dP_0}\frac{dP_v}{dP_0}\right)\right)^\Delta = (1 - (\widetilde\kappa u'v)^2)^{-\Delta/2},
$$

where the last equality follows from Lemma 14. Denote U and V to be the p dimensional Rademacher variables with U being independent of V and $\varepsilon_p = (1'V/p)^2$. Thus

$$
\begin{aligned}
\chi^2(P_1^\Delta, P_0^\Delta) &= \frac{1}{4^p}\sum_{u,v\in\{\pm 1\}^p} (1 - (\widetilde\kappa u'v)^2)^{-\Delta/2} - 1 \\
&= E_{U,V}\left(\left(1 - (\widetilde\kappa U'V/p)^2\right)^{\Delta/2}\right) - 1 \\
&= E_V\left(\left(1 - (\widetilde\kappa 1'V/p)^2\right)^{\Delta/2}\right) - 1 \\
&\le E\left(\exp(\varepsilon_p \widetilde\kappa^2 \Delta)\right) - 1,
\end{aligned}
$$

where the inequality follows from inequality $(1-t)^{-\Delta/2} \leq \exp(\Delta t)$ for any $t \leq 1/2$ and that $\widetilde{\kappa} \leq 1/\sqrt{2}$. The Höffding's equality, applied to Rademacher variables, gives

$$P(\varepsilon_p \geq \lambda) \leq 2e^{-2p\lambda}. \tag{52}$$

Thus

$$\begin{aligned} E\left(\exp(\varepsilon_p \widetilde{\kappa}^2 \Delta)\right) &= \int_0^\infty P\left(\exp\left(\varepsilon_p \widetilde{\kappa}^2 \Delta \geq u\right)\right) du \\ &\leq 1 + \int_1^\infty P\left(\varepsilon \geq \log(u)/(\widetilde{\kappa}^2 \Delta)\right) du \\ &\leq 1 + \int_1^\infty 2\exp\left(-\frac{\log(u)2p}{\widetilde{\kappa}^2 \Delta}\right) du \\ &= 1 - \frac{2}{1 - \frac{2p}{\widetilde{\kappa}^2 \Delta}}, \end{aligned}$$

where the second inequality follows from (52), and the last equality holds if $\frac{2p}{\widetilde{\kappa}^2\Delta} > 1$. Thus if $\Delta = \frac{2p}{33\widetilde{\kappa}^2} = \frac{2p\sigma^4}{33\kappa^2}$, then, using the well-known fact that $\|P_0^\Delta - P_1^\Delta\|_1 \leq 2\sqrt{\chi^2(P_1^\Delta, P_0^\Delta)}$, we obtain the bounds

$$\|P_0^\Delta - P_1^\Delta\|_1 \leq 2\sqrt{\chi^2(P_1^\Delta, P_0^\Delta)} \leq 2\sqrt{\frac{2}{\frac{2p}{\widetilde{\kappa}^2\Delta} - 1}} = 1/2.$$

This and (51) give

$$\inf_{\hat{\eta}} \sup_{P \in \mathcal{P}} E_P(|\hat{\eta} - \eta|) \geq n/6.$$

□

Proof of Lemma 4. Let

$$\mathcal{Q} = \left\{P_{\kappa,\Delta}^n : \Delta < n/2\right\}.$$

Using the same argument as in Step 1 of the proof of Lemma 3, it suffices to take

$$\widetilde{\kappa} = \kappa/\sigma^2 \leq 1/4,$$

and consider $\Sigma_u = I_p + \widetilde{\kappa} u u'$.

Let u to be any unit vector in $\mathbb{R}^p$ and δ a positive number. Let P_0^n denote the joint distribution of independent samples $\{X_i\}_{i=1}^n$ in $\mathbb{R}^p$ where

$$X_1, \ldots, X_{\Delta-1} \overset{iid.}{\sim} N_p(0, I), \quad X_\Delta, \ldots, X_n \overset{iid.}{\sim} N_p(0, \Sigma_u)$$

and P_1^n the joint distribution of independent samples $\{X_i\}_{i=1}^n$ in $\mathbb{R}^p$ where

$$X_1, \ldots, X_{\Delta+\delta} \overset{iid.}{\sim} N_p(0, I), \quad X_{\Delta+\delta+1}, \ldots, X_n \overset{iid.}{\sim} N_p(0, \Sigma_u).$$

Let $\eta(P_i^n)$ denote the location of the change point of distribution $P_{i,u}^n$. Then since $\eta(P_0^n) = \Delta$ and $\eta(P_1^n) = \Delta + \delta$, thus $|\eta(P_0^n) - \eta(P_1^n)| \geq \delta$. By Le cam's lemma (Yu, 1997),

$$\inf_{\hat{\eta}} \sup_{P \in \mathcal{Q}} E_P(|\hat{\eta} - \eta|) \geq \delta(1 - d_{TV}(P_0^n, P_1^n)) = \delta(1 - \frac{1}{2}\|P_0^n - P_1^n\|_1)$$

Since P_0^n and P_1^n are identically distributed on $X_1, \dots X_\Delta - 1$ and on $X_{\Delta+\delta}, \dots, X_n$, $\|P_0^n - P_1^n\|_1 = \|P_0^\delta - P_1^\delta\|_1$, where P_0^δ is the joint distribution of $X_1, X_\delta \sim N_p(0, I)$ and P_1^δ is the joint distribution of $X_1, X_\delta \sim N_p(0, \Sigma_u)$. By Lemma 14,

$$\chi^2(P_1^\delta, P_0^\delta) = (1 - \widetilde{\kappa}^2)^{-\delta/2} - 1 \leq 4\widetilde{\kappa}^2\delta.$$

if $\kappa^2\delta/2 \leq 1/4$ and $\delta/2 \geq 2$. Thus by taking $\delta = \widetilde{\kappa}^{-2}/4 \geq 4$ we obtain that

$$\inf_{\hat{\eta}} \sup_{P \in \mathcal{Q}} E_P(|\hat{\eta} - \eta|) \geq \delta \left(1 - \frac{1}{2}\sqrt{\chi^2(P_1^\delta, P_0^\delta)}\right) \geq \widetilde{\kappa}^{-2}/32 = \sigma^4\kappa^{-2}/32.$$

□

Lemma 14. *Let* $P_0 = N_p(0, I_p)$ *and* $P_u = N_p(0, I_p + \kappa uu')$*. Then*

$$E_{P_0}\left(\frac{dP_u}{dP_0}\frac{dP_v}{dP_0}\right) = \left(1 - (\kappa u'v)^2\right)^{-1/2}.$$

Proof. See, e.g., lemma 5.1 in Berthet and Rigollet (2013). □

Lemma 15. *For* $t \geq 2, x \geq 0$*, if* $tx \leq 1/4$*,* $(1-x)^{-t} - 1 \leq 4tx$

Proof. There exists $s \in [0, x]$ such that

$$(1-x)^{-t} - 1 = tx + t(t+1)x^2(1-s)^{-t-2} \leq tx + 4t^2x^2(1-x)^{-t} \leq 2tx + 4t^2x^2((1-x)^{-t} - 1).$$

Thus

$$(1-x)^{-t} - 1 \leq \frac{2tx}{1 - 4t^2x^2} \leq 4tx$$

□

D Properties of the covariance CUSUM statistic

D.1 Properties of 1d CUSUM statistics

Lemma 16. *Suppose* $[s, e] \subset [1, T]$ *such that* $e - s \leq C_R\Delta$*, and that*

$$\eta_{r-1} \leq s \leq \eta_r \leq \ldots \leq \eta_{r+q} \leq e \leq \eta_{r+q+1}, \quad q \geq 0.$$

Denote

$$\kappa_{\max}^{s,e} = \max\{\eta_p - \eta_{p-1} : r \leq p \leq r+q\}.$$

Then for any $r - 1 \leq p \leq r + q$*,*

$$\left|\frac{1}{e-s}\sum_{i=s}^{e} f_i - f_{\eta_p}\right| \leq C_R\kappa_{\max}^{s,e}.$$

Proof. Since $e-s \leq C_R\Delta$, the interval $[s,e]$ contains at most C_R+1 change points. Observe that

$$\begin{aligned}
&\left|\frac{1}{e-s}\sum_{i=s}^{e} f_i - f_{\eta_p}\right| \\
=&\frac{1}{e-s}\left|\sum_{i=s}^{\eta_r}(f_{\eta_{r-1}}-f_{\eta_p}) + \sum_{i=\eta_r+1}^{\eta_{r+1}}(f_{\eta_r}-f_{\eta_p}) + \ldots + \sum_{i=\eta_{r+q}+1}^{e}(f_{\eta_{r+q}}-f_{\eta_p})\right| \\
\leq&\frac{1}{e-s}\sum_{i=s}^{\eta_r}|p-r|\kappa_{\max}^{s,e} + \sum_{i=\eta_r+1}^{\eta_{r+1}}|p-r-1|\kappa_{\max}^{s,e} + \ldots + \sum_{i=\eta_{r+q}+1}^{e}|p-r-q-1|\kappa_{\max}^{s,e} \\
\leq&\frac{1}{e-s}\sum_{i=s}^{e}(C_R+1)\kappa_{\max}^{s,e},
\end{aligned}$$

where $|p_1-p_2| \leq C_R+1$ for any $\eta_{p_1},\eta_{p_2} \in [s,e]$ is used in the last inequality. ◻

Lemma 17. *If η_p is the only change point in $[s,e]$, then*

$$|\widetilde{f}_{\eta_p}^{s,e}| = \sqrt{\frac{(\eta_p-s)(e-\eta_p)}{e-s}}\kappa_p \leq \sqrt{\min\{\eta_p-s, e-\eta_p\}}\kappa_p$$

Lemma 18. *Let $[s,e]$ contains two or more change points such that*

$$\eta_{r-1} \leq s \leq \eta_r \leq \ldots \leq \eta_{r+q} \leq e \leq \eta_{r+q+1}, \quad q \geq 1.$$

If

$$\eta_r - s \leq c_1^2\Delta$$

then

$$|\widetilde{f}_{\eta_r}^{s,e}| \leq c_1|\widetilde{f}_{\eta_{r+1}}^{s,e}| + 2\kappa_r\sqrt{\eta_r-s}.$$

This can be useful in testing when there are exactly two change points with

$$\eta_r - s \leq \lambda^2\kappa_r^{-2}, \quad e-\eta_{r+1} \leq \lambda^2\kappa_{r+1}^{-2}.$$

It is also useful to show $\eta_r - s \geq \Delta/4$ for some absolute constant c when

$$|\widetilde{f}_{\eta_r}^{s,e}| \geq \max_{s\leq t\leq e}|\widetilde{f}_t^{s,e}| - 2\lambda.$$

Proof. Consider the sequence $\{g_t\}_{t=s+1}^{e}$ be such that

$$g_t = \begin{cases} f_{\eta_{r+1}} & \text{if} \quad s+1 \leq t \leq \eta_r, \\ f_t & \text{if} \quad \eta_r+1 \leq t \leq e. \end{cases}$$

For any $t \geq \eta_r$,

$$\widetilde{f}_{\eta_r}^{s,e} - \widetilde{g}_{\eta_r}^{s,e} = \sqrt{\frac{(e-s)-t}{(e-s)(t-s)}}(\eta_r-s)(f_{\eta_{r+1}}-f_{\eta_r}) \leq \sqrt{\eta_r-s}\kappa_r$$

Thus

$$
\begin{aligned}
|\widetilde{f}^{s,e}_{\eta_r}| &\leq |\widetilde{g}^{s,e}_{\eta_r}| + \sqrt{\eta_r - s}\kappa_r \\
&\leq \sqrt{\frac{(\eta_r - s)(e - \eta_{r+1})}{(\eta_{r+1} - s)(e - \eta_r)}} |\widetilde{g}^{s,e}_{\eta_{r+1}}| + \sqrt{\eta_r - s}\kappa_r \\
&\leq \sqrt{\frac{c_1^2 \Delta}{\Delta}} |\widetilde{g}^{s,e}_{\eta_{r+1}}| + \sqrt{\eta_r - s}\kappa_r \\
&\leq c_1 |\widetilde{f}^{s,e}_{\eta_{r+1}}| + 2\sqrt{\eta_r - s}\kappa_r.
\end{aligned}
$$

where the first inequality follows from the observation that the first change point of g_t in $[s, e]$ is at η_{r+1}. □

D.2 Properties of the covariance CUSUM statistics

All of our consistency results heavily rely on the properties of population quantity of the CUSUM statistic. In the covariance change point detection problem, however, it is not trivial to analyze the properties of the function $t \mapsto \|\widetilde{\Sigma}^{s,e}_t\|_{\mathrm{op}}$ in the multiple change point case. For example, it is difficult to determine the regions of monotonicity of $\|\widetilde{\Sigma}^{s,e}_t\|_{\mathrm{op}}$ as a function of t as is done in Venkatraman (1992, Lemma 2.2). As a remedy, we introduce the concept of *shadow vector*, which is defined as a maximizer of the operator norm of the CUSUM statistics in all the following results. In this way, we turn the covariance change point detection problem into a mean change point detection problem.

For any $v \in \mathbb{R}^p$ with $\|v\| = 1$, let $Y_i(v) = (v^\top X_i)^2$ and $f_i(v) = v^\top \Sigma_i v$, $i = 1, \ldots, n$. Note that both $\{Y_i(v)\}$ and $\{f_i(v)\}$ are univariate sequences, we hence have the corresponding CUSUM statistics defined below

$$
\widetilde{Y}^{s,e}_t(v) = \sqrt{\frac{e-t}{(e-s)(t-s)}} \sum_{i=s+1}^{t} Y_i(v) - \sqrt{\frac{t-s}{(e-s)(e-t)}} \sum_{i=t+1}^{e} Y_i(v), \tag{53}
$$

$$
\widetilde{f}^{s,e}_t(v) = \sqrt{\frac{e-t}{(e-s)(t-s)}} \sum_{i=s+1}^{t} f_i(v) - \sqrt{\frac{t-s}{(e-s)(e-t)}} \sum_{i=t+1}^{e} f_i(v). \tag{54}
$$

The key rationale of the CUSUM based BS algorithm or any variants thereof being a powerful tool selecting the change points is that the population version of the CUSUM statistic achieving its maxima at the true change points. In Lemma 19, we show the same holds for the covariance CUSUM statistic.

Lemma 19. *Assume $(s, e) \cap \{\eta_k\}_{k=1}^K \neq \emptyset$ and Assumption 1. The quantity $\|\widetilde{\Sigma}^{s,e}_t\|_{\mathrm{op}}$ as a function of t achieves its maxima at the true change points, i.e.*

$$
\operatorname*{arg\,max}_{t=s+1,\ldots,e-1} \|\widetilde{\Sigma}^{s,e}_t\|_{\mathrm{op}} \cap \{\eta_k\}_{k=1}^K \neq \emptyset.
$$

Proof. For the sake of contradiction suppose that there exists $t^* \in (s, e) \setminus \{\eta_k\}_{k=1}^K$ such that

$$
t^* \in \operatorname*{arg\,max}_{t=s+1,\ldots,e-1} \|\widetilde{\Sigma}^{s,e}_t\|_{\mathrm{op}},
$$

and

$$\big\|\widetilde{\Sigma}^{s,e}_{t^*}\big\|_{\mathrm{op}} > \max_{k:\, \eta_k \in (s,e)} \big\|\widetilde{\Sigma}^{s,e}_{\eta_k}\big\|_{\mathrm{op}}.$$

Let $v \in \arg\max_{\|u\|=1} |u'\widetilde{\Sigma}^{s,e}_{t^*}u|$, and consider the sequence $\{f_i(v)\}_{i=1}^n = \{v'\Sigma_i v\}_{i=1}^n$. By the above display, we have

$$\big|\widetilde{f}^{s,e}_{t^*}(v)\big| = \big\|\widetilde{\Sigma}^{s,e}_{t^*}\big\|_{\mathrm{op}} > \max_{k:\, \eta_k \in (s,e)} \big\|\widetilde{\Sigma}^{s,e}_{\eta_k}\big\|_{\mathrm{op}} \geq \max_{k:\, \eta_k \in (s,e)} \big|\widetilde{f}_{\eta_k}(v)\big|, \tag{55}$$

where $\widetilde{f}^{s,e}_t(v)$ is defined in (54). It follows from Lemma 2.2 of Venkatraman (1992), the quantities $\big|\widetilde{f}^{s,e}_t(v)\big|$ is maximized at the change points of the time series $\{f_t(v)\}_{t=s+1}^e$. Note that the change points of the sequence $\{f_t(v)\}_{t=s+1}^e$ is a subset of $\{\eta_k\}_{k=1}^K$, this contradicts (55). □

Lemma 19 shows that the population version of the covariance CUSUM statistic is maximized at the true change points in terms of the operator norm. In Lemma 20 below, we give the lower bound of the maxima thereof. One can interpret it as the signal strength.

Lemma 20. *Under Assumption 1, let $0 \leq s < \eta_k < e \leq n$ be any interval satisfying*

$$\min\{\eta_k - s, e - \eta_k\} \geq c_1\Delta.$$

Then for any $0 < \delta < (c_1/8)\Delta$,

$$\max_{t=\lceil s+\delta\rceil,\ldots,\lfloor e-\delta\rfloor} \|\widetilde{\Sigma}^{s,e}_t\|_{\mathrm{op}} \geq (c_1/2)\kappa\Delta(e-s)^{-1/2}.$$

Proof. Let

$$v \in \arg\max_{\|u\|=1} \big|u^\top(\Sigma_{\eta_k} - \Sigma_{\eta_{k+1}})u\big|.$$

Therefore $\big\|\widetilde{\Sigma}^{s,e}_t\big\|_{\mathrm{op}} \geq \big|\widetilde{f}^{s,e}_t(v)\big|$. Since

$$\big|f_{\eta_k}(v) - f_{\eta_{k+1}}(v)\big| = \big|v^\top(\Sigma_{\eta_k} - \Sigma_{\eta_{k+1}})v\big| = \big\|\Sigma_{\eta_k} - \Sigma_{\eta_{k+1}}\big\|_{\mathrm{op}} \geq \kappa,$$

we have

$$\max_{t=\lceil s+\delta\rceil,\ldots,\lfloor e-\delta\rfloor} \|\widetilde{\Sigma}^{s,e}_t\|_{\mathrm{op}} \geq \max_{t=\lceil s+\delta\rceil,\ldots,\lfloor e-\delta\rfloor} |\widetilde{f}^{s,e}_t(u)| \geq (c_1/2)\kappa\Delta(e-s)^{-1/2},$$

where the second inequality follows from the same arguments in Venkatraman (1992, Lemma 2.4) by regarding $\widetilde{f}^{s,e}_t(u)$ as the CUSUM statistic for a univariate time series. □

In Lemma 21, we show that we can conform the problem of change point detection on covariance to the one on mean via shadow vectors. This translation allows us to convert a high dimensional covariance problem to a univariate mean problem.

Lemma 21. *Under the same assumptions as Lemma 20, let*

$$b \in \arg\max_{t=\lceil s+\delta\rceil,\ldots,\lfloor e-\delta\rfloor} \big\|\widetilde{S}^{s,e}_t\big\|_{\mathrm{op}},$$

and denote a shadow vector by

$$v \in \arg\max_{\|u\|=1} \big|u^\top \widetilde{S}^{s,e}_b u\big|. \tag{56}$$

Then

$$b \in \arg\max_{t=\lceil s+\delta\rceil,\ldots,\lfloor e-\delta\rfloor} \big|\widetilde{Y}^{s,e}_t(v)\big|.$$

Proof. It suffices to show that

$$\left|\widetilde{Y}_b^{s,e}(v)\right| = \left\|\widetilde{S}_b^{s,e}\right\|_{\mathrm{op}} \geq \max_{t=\lceil s+\delta\rceil,\ldots,\lfloor e-\delta\rfloor} \left\|\widetilde{S}_t^{s,e}\right\|_{\mathrm{op}} \geq \max_{t=\lceil s+\delta\rceil,\ldots,\lfloor e-\delta\rfloor} \left|\widetilde{Y}_t^{s,e}(v)\right|.$$

□

For the shadow vector v defined in (56), it is tempting to argue that we can estimate v and then use $\{Y_i(v)\}_{i=1}^n$ as the new sequence to derive an upper for the rate of localization. This ideal approach, however, suffers from two non-trivial obstacles.

- We don't have any guarantee that the estimation of v is consistent. This is because estimating the first principle direction of any sample matrix in general requires a non-vanish gap between the first and the second eigenvalues of the corresponding population matrix. Since b depends on the data, without more involved additional assumption, it is difficult to show $\widetilde{S}_b^{s,e}$ converges to a population quantity as $n \to \infty$.
- Suppose v given in (56) has a well defined population quantity, say $E(v)$. The estimation of v depends on $\{X_i\}_{i=1}^n$ and therefore $E(Y_i(v)) \neq f_i(E(v))$.

On one hand our knowledge of the population version of the shadow vector v is very limited; on the other hand, in Lemma 22 below we show that without estimating the population shadow vector, the CUSUM statistics and its corresponding population quantity are close enough and that the maximum of the CUSUM statistic is detectable.

Lemma 22. *Under the same assumptions as Lemma 20, assume*

$$\max_{t=\lceil s+\delta\rceil,\ldots,\lfloor e-\delta\rfloor} \left\|\widetilde{S}_t^{s,e} - \widetilde{\Sigma}_t^{s,e}\right\|_{\mathrm{op}} \leq \lambda,$$

where $\lambda_1 > 0$. For the shadow vector v defined in (56), then

$$\max_{t=\lceil s+\delta\rceil,\ldots,\lfloor e-\delta\rfloor} \left|\widetilde{Y}_t^{s,e}(v)\right| \geq (c_1/2)\kappa\Delta(e-s)^{-1/2} - \lambda.$$

Proof. Observe that for all compatible t and v,

$$\left|\widetilde{Y}_t^{s,e}(v) - \widetilde{f}_t^{s,e}(v)\right| = \left|v^\top\left(\widetilde{S}_t^{s,e} - \widetilde{\Sigma}_t^{s,e}\right)v\right| \leq \left\|\widetilde{S}_t^{s,e} - \widetilde{\Sigma}_t^{s,e}\right\|_{\mathrm{op}} \leq \lambda.$$

In addition, for the shadow vector v, we have

$$\begin{aligned}\max_{t=\lceil s+\delta\rceil,\ldots,\lfloor e-\delta\rfloor} \left|\widetilde{Y}_t^{s,e}(v)\right| &= \max_{t=\lceil s+\delta\rceil,\ldots,\lfloor e-\delta\rfloor} \left\|\widetilde{S}_t^{s,e}\right\|_{\mathrm{op}} \\ &\geq \max_{t=\lceil s+\delta\rceil,\ldots,\lfloor e-\delta\rfloor} \left\|\widetilde{\Sigma}_t^{s,e}\right\|_{\mathrm{op}} - \lambda_1 \geq (c_1/2)\kappa\Delta(e-s)^{-1/2} - \lambda,\end{aligned}$$

where the last inequality follows from Lemma 20. □

We finally prove a simple property of covariance CUSUM statistics.

Lemma 23. $\|\widetilde{\Sigma}_t^{s,e}\|_{\mathrm{op}} \leq 2\sqrt{\min\{e-t, t-s\}}B^2$ *for any $t \in (s,e)$.*

Proof. Observe that

$$\|\widetilde{\Sigma}_t^{s,e}\|_{\mathrm{op}} \leq 2\sqrt{\frac{(e-t)(s-t)}{e-s}}B^2 \leq 2\sqrt{\min\{e-t, t-s\}}B^2.$$

□