Optimal multiple change-point detection for high-dimensional data

Emmanuel Pilliat, Alexandra Carpentier, Nicolas Verzelen

Optimal multiple change-point detection for high-dimensional data

Emmanuel Pilliat[1], Alexandra Carpentier[2], and Nicolas Verzelen[3]

[1]Université de Montpellier, Montpellier, France
[2]Otto Von Guerike University, Magdeburg, Germany
[3]INRAE Montpellier, Montpellier, France

November 13, 2020

Abstract

This manuscript makes two contributions to the field of change-point detection. In a general change-point setting, we provide a generic algorithm for aggregating local homogeneity tests into an estimator of change-points in a time series. Interestingly, we establish that the error rates of the collection of test directly translate into detection properties of the change-point estimator. This generic scheme is then applied to the problem of possibly sparse multivariate mean change-point detection setting. When the noise is Gaussian, we derive minimax optimal rates that are adaptive to the unknown sparsity and to the distance between change-points. For sub-Gaussian noise, we introduce a variant that is optimal in almost all sparsity regimes.

1 Introduction

Change-point detection has a long history since seminal work of Wald [Wal45], Girshick and Rubin [GR52] and that lead to flourishing lines (see [NHZ16, TOV20] for recent surveys). Earlier work focused on the problems of detecting and localizing change-points in a univariate time series. Motivated by applications in genomics [OVLW04] and finance, there has been a recent trend in the literature towards the analysis of more complex time series for instance in a high-dimensional linear space [Jir15] or even belonging to a non-Euclidean space [CC+19].

In this work, we study high-dimensional time series whose mean may have change-points, some of them possibly arising on a small subset of the coordinates. See the introduction of [WS18] for an account of possible applications. In particular, we build a procedure which is able to detect and localize change-points under minimal assumptions on the height of these change-points. Along the way towards this optimal procedure, we define and analyze a scheme for general change-point problems that aggregates a collection of local tests into an estimator change-points. In this introduction, we first describe this generic scheme before turning to our results in high-dimensional sparse change-point detection.

1.1 General change-point setting

In its most general form, we consider a random sequence $Y = (y_1, y_2, \dots, y_n)$ in some measured space $\mathcal{Y}^n$ and, for $t = 1, \dots, n$, we write $\mathbb{P}_t$ for the marginal distribution of y_t. Consider a functional Γ mapping the probability distribution $\mathbb{P}_t$ to some space $\mathcal{V}$. Then, the purpose of change-point detection is to detect changes in the sequence $(\Gamma(\mathbb{P}_1), \Gamma(\mathbb{P}_2), \dots, \Gamma(\mathbb{P}_n))$ in $\mathcal{V}^n$ and to estimate the positions of these changes. This setting is very general and does not require that the random variables (y_t) are independent.

Let us shortly explain how this general framework encompasses most known settings of offline change-point detection. In the Gaussian mean univariate change-point setting, we have $\mathcal{Y} = \mathbb{R}$, the distributions $\mathbb{P}_t$ corresponds to the normal distribution with mean $\theta_t \in \mathbb{R}$ and variance σ^2 and $\Gamma(\mathbb{P}_t) = \theta_t$.

In the (heteroscedastic) mean univariate change-point problem, the distribution $\mathbb{P}_t$ is not necessarily Gaussian and, in particular, the variance of y_t is allowed to vary with t. Still, one is only interested in detecting variations of $\Gamma(\mathbb{P}_t) = \int x d\mathbb{P}_t = \mathbb{E}[y_t]$. By contrast, in the variance univariate change-point problems, one focuses on changes in the variance of y_t. This can be done by taking $\Gamma(\mathbb{P}_t) = \int x^2 d\mathbb{P}_t - [\int x d\mathbb{P}_t]^2 = \mathrm{Var}(y_t)$. If the statistician is interested in arbitrary changes in the distributions, then the functional Γ is simply the identity map.

As the problem of change-point detection is that of detecting changes in the sequence $(\Gamma(\mathbb{P}_1), \Gamma(\mathbb{P}_2), \ldots, \Gamma(\mathbb{P}_n))$, we consider an integer $0 \leq K \leq n-1$ and a vector of integers $\tau = (\tau_1, \ldots, \tau_K)$ satisfying $1 = \tau_0 < \tau_1 < \cdots < \tau_K < \tau_{K+1} = n+1$ such that $\Gamma(\mathbb{P}_t)$ is constant over each interval $[\tau_k, \tau_{k+1} - 1]$ and $\Gamma(\mathbb{P}_{\tau_k - 1}) \neq \Gamma(\mathbb{P}_{\tau_k})$. Hence, τ_k corresponds to the *position* of the k^{th} change-point. We shall often refer to τ_k as a *change-point*. Equipped with this notation, we are interested in building an estimator $\hat{\tau} = (\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{K}})$ of τ from the time series Y. $\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{K}}$ correspond to the *estimated change-points* of τ and $\hat{K}$ to the number of the estimated change-points.

Desirable Guarantees of an estimator. Before describing the generic scheme for estimating τ, let us first formalize our desireta for a good change-point procedure. With multiple change-points, the primary objectives are to detect most if not all change-points while estimating no (or at least very few) spurious change-points. We consider as in [VFLRB20] realistic settings where some change-points heights are possibly too small to be detected, so that we can only hope to detect the subset of *significant* change-points. Regarding spurious change-points, it is usually required that the number of change-points K is not overestimated by $\hat{\tau}$. The **(No Sp)** (see [VFLRB20]) property described below, is more precise, and requires the absence of *spurious* change-points.
(No Sp): An estimator $\hat{\tau}$ of size $\hat{K}$ of the change-points is said to detect no spurious change-points if, for all $1 \leq k \leq K$,

$$\left| \{\hat{\tau}_{k'}, \, 1 \leq k' \leq \hat{K}\} \cap \left[\tau_k - \frac{\tau_k - \tau_{k-1}}{2}, \tau_k + \frac{\tau_{k+1} - \tau_k}{2} \right] \right| \leq 1 \ ,$$

and

$$\{\hat{\tau}_{k'}, k' \leq \hat{K}\} \subset \left[\tau_1 - \frac{\tau_1 - 1}{2}, \tau_K + \frac{n + 1 - \tau_K}{2} \right] \ .$$

The second condition simply ensures that no change-point is estimated near the time boundaries of the time series. The first condition entails that, for each change-point τ_k there is at most one estimated change-point $\hat{\tau}_k$ in the interval $\left[\tau_k - \frac{\tau_k - \tau_{k-1}}{2}, \tau_k + \frac{\tau_{k+1} - \tau_k}{2}\right]$. In other words, it is required that, even on each sub-interval, the number of change-points is not overestimated.

Let us now turn to the **Detection** property. Since arbitrarily small change-points are allowed in our analysis, it is acknowledged that not all change-points are detectable. As a consequence, we consider a subset $\mathcal{K}^* \subset \{1, \ldots, K\}$ of change-point indices that correspond to *significant* change-points. Obviously, the significance of a particular change-point is relative to the change-point problem of consideration - data distribution, nature of change-points. As a primary example, we define the suitable notion of energy and significance of a change-point in the multivariate change-points setting studied in the next subsection. The second guarantee we aim for is to **Detect** all significant change-points. A change-point τ_k is said to be detected if there is at least one change-point estimated in the interval $\left[\tau_k - \frac{\tau_k - \tau_{k-1}}{2}, \tau_k + \frac{\tau_{k+1} - \tau_k}{2}\right]$.

A generic roadmap for change-point detection In this manuscript, our first contribution is a generic procedure for translating a collection of tests into an estimator $\widehat{\tau}$ of τ. For two positive integers (l, r), we consider the time interval $[l - r, l + r)$. Suppose we are given a collection $\mathcal{G}$ of (l, r), which therefore corresponds to a collection of time intervals. For each $(l, r) \in \mathcal{G}$, we are also given an homogeneity test $(T_{l,r})$ of the hypothesis $\{(\Gamma(\mathbb{P}_i)) \text{ is constant over the segment } [l - r, l + r)\}$ which is equivalent to the absence of change-point on the interval $(l - r, l + r)$. Given such a collections of homogeneity tests $(T_{l,r})$ with $(l, r) \in \mathcal{G}$, we build in this manuscript an estimator $\widehat{\tau}$ that satisfies the following properties: If the multiple testing procedure does not reject any true null hypothesis (no false positives), then $\widehat{\tau}$ does not estimate any spurious change-point, that is, it satisfies (**No Sp**). Besides,

consider any change-point τ_k that is detected by some test $T_{\bar{\tau}_k,\bar{r}_k}$. If the segment $[\bar{\tau}_k - \bar{r}_k, \bar{\tau}_k + \bar{r}_k)$ contains τ_k and is away from τ_{k-1} and τ_{k+1} - i.e. is contained in $\left[\tau_k - \frac{\tau_k - \tau_{k-1}}{2}, \tau_k + \frac{\tau_{k+1} - \tau_k}{2}\right]$ - then the estimator $\widehat{\tau}$ contains a point $\widehat{\tau}_l$ satisfying $|\widehat{\tau}_l - \tau_k| \leq \bar{r}_k - 1$. In other words, any change-point that is detected by the multiple testing procedure at some small scale $\bar{r}_k$ is also detected and reasonably well localized by our procedure $\widehat{\tau}$.

Thanks to this generic scheme, the construction of a change-point procedure boils down to building a suitable multiple testing procedure $(T_{l,r})$, $(l,r) \in \mathcal{G}$ whose family-wise error rate (FWER) is controlled, while being able to detect the most change-points.

Related Work and possible applications. Up to our knowledge, none of the available generic multiple change-point procedures (e.g. [KLBM20]) comes with guarantees in a general context.

Here, if we choose the test $(T_{l,r})$ to be two sample tests of the hypothesis $\{\Gamma(\mathbb{P}_{l-r}) = \ldots = \Gamma(\mathbb{P}_{l+r-1})\}$ versus $\{\Gamma(\mathbb{P}_{l-r}) = \ldots = \Gamma(\mathbb{P}_{l-1}) \neq \Gamma(\mathbb{P}_l) = \ldots = \Gamma(\mathbb{P}_{l+r-1})\}$, this allows us to build a generic multiple change-point procedure from a collection of two-sample tests. For instance, there has been a lot of interest for many high-dimension change-points problems where $y_i \in \mathbb{R}^p$. This includes changes in vector auto-regressive models [WYRW19], changes in the covariance matrix of y_i [WYR17], changes in the inverse covariance matrix of y_i [GR17, KLBM20]. All such change-point problems can be addressed through the construction and careful analysis of two-sample tests for auto-regressive models, covariance matrices, and inverse covariance matrices respectively. Similarly, one can build non-parametric change-points procedure as in [PYWR19] from two sample homogeneity tests or Kernel change-point procedures [ACH19, GA18] from kernel two-sample tests [GBR+12].

1.2 Sparse Multivariate Change-point Setting

Aside from the general case, we are mainly interested in multivariate mean change-point detection problem with sparse variations where one observes a time series $Y = (y_1, \ldots, y_n) \in \mathbb{R}^{p \times n}$ with unknown mean $\Theta = (\theta_1, \ldots, \theta_n) \in \mathbb{R}^{p \times n}$ so that we have the decomposition

$$y_t = \theta_t + \varepsilon_t \qquad t = 1, \ldots, n \ , \tag{1}$$

where the noise matrix $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ is made of independent and mean zero random $\mathbb{R}^p$. In this manuscript, we make two distributional assumptions on the noise. Either we suppose that all random vectors ε_i follow independent normal distribution with variance $\sigma^2 \mathbf{I}_p$ (see Section 3) or that the components of ε_i follow independent sub-Gaussian distributions with variance $\sigma^2 \mathbf{I}_p$ (see Section 4). In either case, we assume that σ^2 is known.

Here, we are interested in the the variations of the *mean* vector θ_t so that, relying on the formalism of the previous subsection, we have $\Gamma(\mathbb{P}_t) = \theta_t$. Considering the vector of change-points $\tau = (\tau_1, \ldots, \tau_K)$, we can define $K+1$ vectors $\mu_0, \ldots, \mu_K$ in $\mathbb{R}^p$ satisfying $\mu_k \neq \mu_{k+1}$ for all $k = 0, \ldots, K-1$ such that

$$\theta_t = \sum_{k=0}^{K} \mu_k \mathbf{1}_{\tau_k \leq t < \tau_{k+1}} \ .$$

Equivalently, μ_k is the constant mean of y over the interval $[\tau_k, \tau_{k+1} - 1]$. The difference $\mu_k - \mu_{k-1}$ in $\mathbb{R}^p$ measures the variation of Θ at the change-point τ_k and can possibly have many null coordinates. In this possibly sparse multi-dimensional setting, the significance of a change-point is measured through three quantities Δ_k, r_k, and s_k. First, the *height* Δ_k of the change-point τ_k is defined as the Euclidean norm of the signal difference. The *length* r_k of the change-point τ_k is the minimal distance from τ_k to another change-point, τ_{k-1} or τ_{k+1}. More precisely:

$$\Delta_k = \|\mu_k - \mu_{k-1}\| \ ; \qquad r_k = \min(\tau_{k+1} - \tau_k, \tau_k - \tau_{k-1}) \ . \tag{2}$$

As a simple example, Figure 1 depicts a one dimensional piece-wise constant sequence Θ with 3 change-points illustrating the setting presented above. In the univariate change-point literature (e.g. [Fry14, Fry18, CK19]) the height and the length of a change-point characterize the significance of a

change-point. In the multivariate setting, where the change-points can be sparse, meaning the number of non null coordinates of the vector $\mu_k - \mu_{k-1}$ is possibly small, one also considers the *sparsity* s_k of change-point τ_k, defined as

$$s_k = \|\mu_k - \mu_{k-1}\|_0 \quad , \tag{3}$$

where for any $v \in \mathbb{R}^p$, $\|v\|_0 = \sum_{1 \le i \le p} \mathbf{1}\{v_i \neq 0\}$.

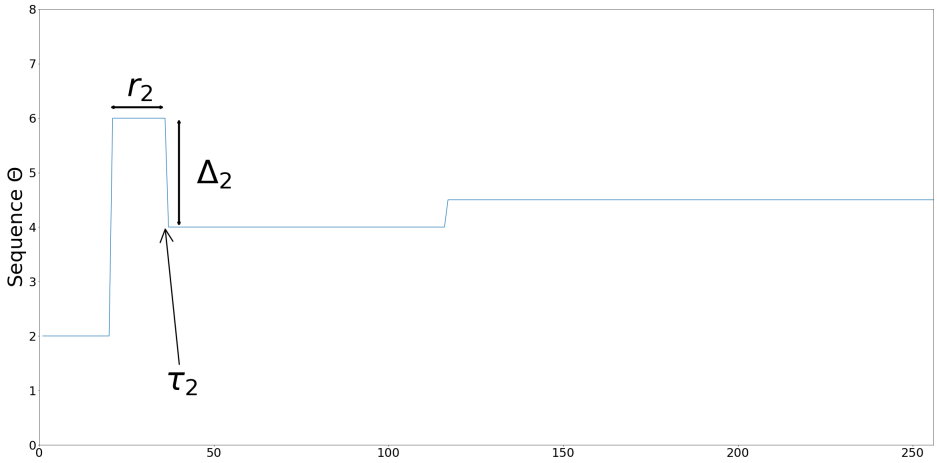

Figure 1: An example of a piece-wise constant sequence Θ with 3 change-points and $p = 1$.

Two-sample tests and CUSUM statistics Our objective is to detect and recover positions $(\tau_k)_{k \le K}$ under minimal conditions on the change-point height Δ_k, change-point length r_k and sparsity s_k. In view of the generic change-point procedure discussed in the previous subsection, this mainly boils down to building suitable tests of the assumptions $\{\Theta$ is constant over $[l-r, l+r)\}$ versus $\{\Theta$ is not constant on this segment$\}$. Following the literature on binary and wild binary segmentation, we consider the CUSUM statistic

$$\mathbf{C}_{l,r}(Y) = \sqrt{\frac{r}{2\sigma^2}} \left(\frac{1}{r} \sum_{i=l}^{l+r-1} y_i - \frac{1}{r} \sum_{t=l-r}^{l-1} y_i \right) \quad .$$

This statistic computes the normalized difference of empirical mean of y_i on $[l-r, l)$ and $[l, l+r)$. If the noise is Gaussian and if Θ is constant on $[l-r, l+r)$, then $\mathbf{C}_{l,r}(Y)$ simply follows a standard normal distribution. To simplify, consider a specific instance of our testing problem where we want to test whether $\{\Theta$ is constant over $[l-r, l+r)\}$ versus $\{\Theta$ contains exactly one change-point at l on the segment $[l-r, l+r)\}$. This corresponds to a two-sample mean testing problem, for which the CUSUM statistic $\mathbf{C}_{l,r}(Y)$ is a sufficient statistic if the noise is Gaussian. Then, given $\mathbf{C}_{l,r}(Y)$, one wants to test whether its expectation is 0 (no change-point on $[l-r, l+r)$) versus its expectation is non-zero but is s-sparse for some unknown s. This classical detection problem is well understood [DJ04] and it is well known that a combination of a χ^2-type test with a higher-criticism-type test is optimal. Here, the challenge stems from the fact that we do not want to perform a single such test, but a large collection of tests over a collection of $(l, r) \in \mathcal{G}$.

Our contribution As usual in the mean change-point literature, we consider $r_k \Delta_k^2$ as the energy of the change-point τ_k. Up to a possible factor in $[1/2, 1]$, $r_k \Delta_k^2$ is the square distance between Θ and its projection on the space of vectors Θ' with change-point at $(\tau_1, \ldots, \tau_{k-1}, \tau_{k+1}, \ldots, \tau_K)$ (see e.g. [VFLRB20] for a discussion in the univariate setting). In other words, the energy $r_k \Delta_k^2$ characterizes the significance of the change-point τ_k. In Section 3, we introduce a multi-scale change-point detection procedure detecting any change-point τ_k whose energy is higher (up to a numerical constant) than

$\sigma^2 s_k \log(1 + \frac{\sqrt{p}}{s_k}\sqrt{\log(n/r_k)}) + \sigma^2 \log(n/r_k)$. This result is valid for arbitrary length r_k and sparsity s_k, and does not require the knowledge of these two quantities. In summary, our procedure does not estimate any spurious change-point (**NoSp**) and **Detects** all the change-points whose energy is higher than the latter threshold. In Section 5, we establish that, as soon as the unknown number K of the change-points is larger than 1, the condition $\sigma^2 s_k \log(1 + \frac{\sqrt{p}}{s_k}\sqrt{\log(n/r_k)}) + \sigma^2 \log(n/r_k)$ on the energy is tight with respect to n, p, r_k and s_k, in the sense that no procedure achieving (**NoSp**) is able to detect with high probability a change-point whose energy is smaller (up to some constant) than the latter threshold. In Section 4, we consider the more general setting where the noise is L-sub-gaussian with known variance, and we establish a similar result to the Gaussian case up to a logarithmic loss in some regimes.

Related work For dense change-points ($s_k = p$) but with unknown covariance for the noise, Wang et al. [WVS19] (see also [WS20]) study the behavior of a procedure based on U-statistics of the CUSUM. Jirak [Jir15] and Yu and Chen [YC17] introduce binary segmentation procedures based on the l_∞ norm of the CUSUMs. Although those work explicitly characterize the asymptotic distribution of their test statistic and, for some of them, allow temporal dependencies in the data, the corresponding energy requirements for change-point detection are either not studied or turn out to be suboptimal. Recently, Liu et al. [LGS19] have characterized the optimal detection rate of a possibly sparse change-point in the specific case where there is at most one change-point. See also [EH19] for earlier results. Closest to our work, Wang and Samworth [WS18] have proposed the INSPECT method based on sparse projection to handle sparse change-points, but INSPECT provably detects the change-points under strong assumption on the energy; see Section 3 for a precise comparison.

In the univariate setting ($p = 1$), minimal energy requirements for change-point detection are well understood [FMS14, Fry18, WYR18, VFLRB20] and are nearly achieved by a wide range of procedures including penalized least-square and multi-scale tests methods.

2 A Generic algorithm for multiscale change-point detection on a grid

In this section, we study the problem of change-point detection in the general setting defined in Section 1.1. We propose an algorithm that aggregates a collection of homogeneity tests, performed at many positions, and for many scales, of our data. We prove that under some conditions on these tests - that have to be fulfilled on an event of high probability - the algorithm performs the change-point detection task.

2.1 Grid and multiscale statistics

Since our purpose is to translate a collection of local tests $T = (T_{l,r})_{(l,r)\in\mathcal{G}}$ indexed by a grid into a change-point procedure, we first need to formalize what we mean by a grid.

Henceforth, we call a grid $\mathcal{G}$ of $[n]$ a collection of locations and scales where a scale r is a positive integer smaller or equal to $\lfloor n/2 \rfloor$ and a location l is an integer between $r+1$ and $n-r$. This couple (l, r) refers to the segment $[l - r, l + r)$ centered at l and with radius r. Formally, $\mathcal{G}$ is therefore a subset of $J_n = \left\{(l, r) \ : \ r = 1, \dots, \left\lfloor \frac{n}{2} \right\rfloor \text{ and } l = r + 1, \dots, n - r\right\}$. Given a grid $\mathcal{G}$, we call $\mathcal{R}$ its collection of scales, that is $\mathcal{R} = \{r \ : \ \exists l \text{ s.t. } (l, r) \in \mathcal{G}\}$. For a scale $r \in \mathcal{R}$, $\mathcal{D}_r$ stands for the corresponding collection of locations, that is $\mathcal{D}_r = \{l \ : \ (l, r) \in \mathcal{G}\}$. We do not assume any condition on the grid $\mathcal{G}$ for now.

In the later section, we are mainly interested in two specific grids: the **complete** grid $\mathcal{G}_F = J_n$ and the **dyadic** grid $\mathcal{G}_D$ defined by $\mathcal{R} = \{1, 2, 4, \dots, 2^{\lfloor \log_2(n) \rfloor - 1}\}$, $\mathcal{D}_1 = [2, n]$, and

$$\mathcal{D}_r = \left\{ r + 1, 3r/2 + 1, 2r + 1, \dots, \left\lfloor \frac{2n}{r} \right\rfloor \frac{r}{2} + 1 \right\} \quad \text{for } r \in \mathcal{R} \setminus \{1\} \ . \tag{4}$$

See Figure 2 for a visual representation of the dyadic grid. At some points, we shall also mention a-adic grids $\mathcal{G}_a$. For any $a \in (0, 1)$, $\mathcal{G}_a$ is defined by $\mathcal{R} = \{1, \lfloor a^{-1} \rfloor, \lfloor a^{-2} \rfloor, \dots, \lfloor a^{1-\lfloor \log(n)/\log(a) \rfloor} \rfloor\}$ and $\mathcal{D}_r$ as in (4).

Given a fixed grid $\mathcal{G}$, a multiscale test is simply a multiple test $T = (T_{l,r})_{(l,r)\in\mathcal{G}}$ indexed by the elements of $\mathcal{G}$. It amounts to testing at all scales $r \in \mathcal{R}$ and all locations $l \in \mathcal{D}_r$ whether the distribution

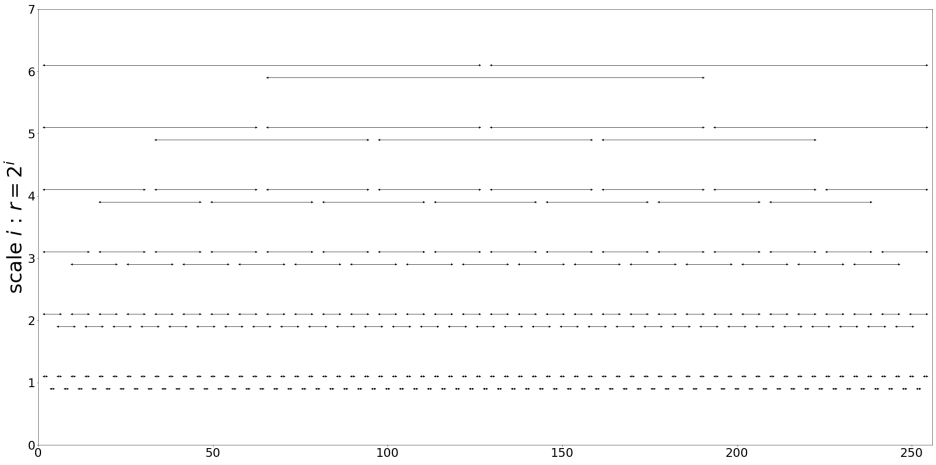

Figure 2: The dyadic grid is represented as follows : for each $r = 2^i$ and $l \in \mathcal{D}_r$, we draw the interval $[l - r + 1, l + r - 1]$ at position $(l, \log_2(r))$.

of (y_t) is constant over the segment $[l - r, l + r)$. Equivalently, $T_{l,r}$ tests whether there exists a change-point in $[l - r + 1, l + r - 1]$.

2.2 From a multiscale test to a change-point detection procedure

Our purpose is to introduce a generic procedure to translate a multiscale procedure into a vector of change-points. Intuitively, if, for some for some $(l, r) \in \mathcal{G}$, we have $T_{l,r} = 1$, then the distribution of y_t is possibly not constant over $[l - r, l + r)$ which entails that there is possibly at least one change-point in $[l - r + 1, l + r - 1]$. As a consequence, the multiscale procedure provides a collection $\mathcal{I}(T) = \{[l - r + 1, l + r - 1] \text{ s.t. } T_{l,r} = 1\}$ of intervals that tentatively contain at least one change-point.

If all these intervals were disjoint, then one simply would take $\widehat{\tau}$ as the sequence of centers of these intervals. Unfortunately, when two intervals $[l_1 - r_1 + 1, l_1 + r_1 - 1]$ and $[l_2 - r_2 + 1, l_2 + r_2 - 1]$ in $\mathcal{I}(T)$ have a non-empty intersection, one cannot necessarily decipher whether there is only one change-point in the intersection of both intervals or if each interval contains a specific change-point. Hence, our general objective is to transform the collection $\mathcal{I}(T)$ into a collection of non-intersecting intervals by either discarding or merging some intervals of $\mathcal{I}(T)$

More formally, we propose the following iterative procedures for building a collection of non-intersecting intervals. Start with $\mathcal{T}_0 = \mathcal{S}_0 = \varnothing$. For any scale $r \in \mathcal{R}$, we compute the collections $\mathcal{S}_r$ of intervals of scale r and the collection $\mathcal{T}_r$ of locations based on the following

$$\mathcal{T}_r = \left\{ l \in \mathcal{D}_r, \quad T_{l,r} = 1 \quad \text{and} \quad [l - r + 1, l + r - 1] \bigcap \Big(\bigcup_{r' < r,\ r' \in \mathcal{R}} \mathcal{S}_{r'} \Big) = \varnothing \ ; \right\}$$
$$\mathcal{S}_r = \bigcup_{l \in \mathcal{T}_r} [l - r + 1, l + r - 1] \ .$$

The sets $\mathcal{T}_1$ and $\mathcal{S}_1$ are made of all points l such that $T_{l,1} = 1$. More generally, $\mathcal{T}_r$ contains all locations l such that $T_{l,r} = 1$ and the corresponding intervals $[l - r + 1, l + r - 1]$ does not intersect with any of the detected interval at a smaller scale $r' < r$. The set $\mathcal{S}_r$ contains all intervals associated to $\mathcal{T}_r$.

One can easily check that $\mathcal{S} = \bigcup_r \mathcal{S}_r$ is a union of closed non-intersecting intervals. The collection $\mathcal{C} = \{\mathrm{C}_1, \ldots, \mathrm{C}_{\hat{K}}\}$ is the partition of $\mathcal{S}$ into connected components if it induces a partition of $\mathcal{S}$ and if, for all $1 \leq i < j \leq \hat{K}$, C_i is a closed segment and $\max \mathrm{C}_i < \min \mathrm{C}_j$. Finally, we estimate the vector of change-points $\widehat{\tau}$ by taking the center of each segment C_k. In other words, we take $\hat{\tau}_k := \frac{1}{2}(\min \mathrm{C}_k + \max \mathrm{C}_k)$ for any $1 \leq k \leq \widehat{K}$ - where $\hat{K}$ corresponds to the number of change-points detected by our $\widehat{\tau}$. The aggregation procedure is summarized in Algorithm 1 below.

Remark: If, for some $r \in \mathcal{R}$ and some $l_1 < l_2 \in \mathcal{D}_r$, we have $T_{l_1,r} = 1$, $T_{l_2,r} = 1$, and $l_1 + r - 1 \geq l_2 - r + 1$, then $\mathcal{S}_r$ contains the segment $[l_1 - r + 1, l_2 + r - 1]$. In other words, our aggregation procedure merges two intervals when both have the same scales.

Data: $y_t, t = 1 \dots n$ and local test statistics $(T_{l,r})_{(l,r)\in\mathcal{G}}$
Result: $(\hat{\tau}_k)_{k\leq\hat{K}}$
$\mathcal{T}_r, \mathcal{S}_r = \varnothing$ for all $r \in \mathcal{R}$ and $\mathcal{S} = \varnothing$;
for $r \in \mathcal{R}$ **do**
  for $l \in \mathcal{D}_r$ *s.t.* $T_{l,r} = 1$ **do**
    if $[l - r + 1, l + r - 1] \bigcap \mathcal{S} = \varnothing$ **then**
      $\mathcal{T}_r \leftarrow \mathcal{T}_r \cup \{l\}$;
      $\mathcal{S}_r \leftarrow \mathcal{S}_r \cup [l - r + 1, l + r - 1]$;
    end
  end
  $\mathcal{S} = \mathcal{S} \bigcup \mathcal{S}_r$;
end
Let $(\mathrm{C}_k)_{k=1,\dots,\hat{K}}$ be the connected components of $\mathcal{S}$ sorted in increasing order;
return $\left(\hat{\tau}_k = \frac{1}{2}(\min \mathrm{C}_k + \max \mathrm{C}_k)\right)_{k=1,\dots,\hat{K}}$

Algorithm 1: Merging procedure for Aggregation of multiscale tests

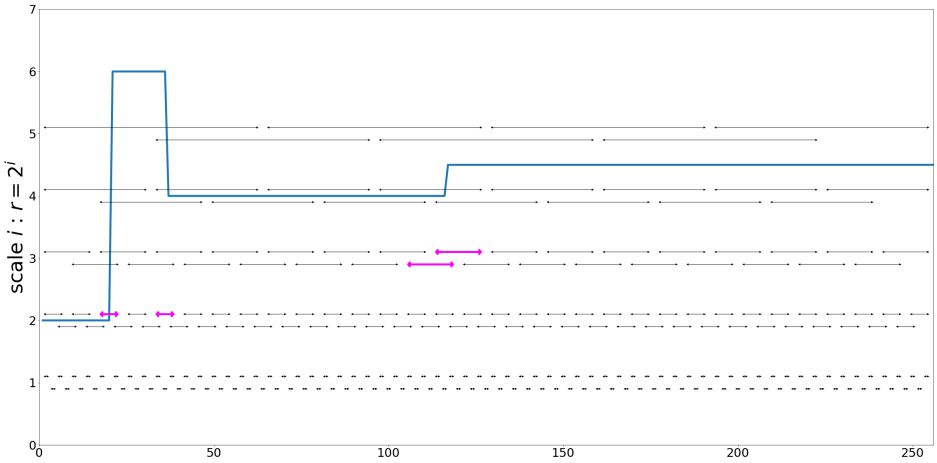

Figure 3: Example of our change-point estimation procedure with three change-points. The first two change-points have large height and are first detected at a small scale r (in magenta) while the third one is detected at a larger scale r.

Computational Complexity. A naive implementation of Algorithm 1 - and also of Algorithm 2 defined in Appendix - requires to compute all tests $T_{l,r}$ on the Grid, whereas the aggregation procedure is linear in the size of the Grid. As an example, the size of the complete grid J_n is of the order of $n^2/2$ whereas that of the dyadic grid is of the order of $3n$. One can speed up the full procedure by computing the statistics $T_{l,r}$ and aggregating on the fly by checking whether $[l-r+1, l+r-1]$ intersects $\mathcal{S}$ before evaluating $T_{l,r} = 1$.

2.3 General analysis

In this subsection, we provide an abstract theorem translating error controls of the multiple test procedure T in terms of properties of $\widehat{\tau}$. As explained in the introduction, the time series (y_t) may contain change-points that are too small to be detected. With this phenomenon in mind, we define a

subset $\mathcal{K}^* \subset [K]$ of indices corresponding to so-called significant change-points. For any such $k \in \mathcal{K}^*$, we introduce an element of the grid $(\bar{\tau}_k, \bar{r}_k) \in \mathcal{G}$ at which the statistics T is expected to detect τ_k.

We assume that the scales $\bar{r}_k$ and the location $\bar{\tau}_k$ of detection satisfy the two following conditions:

$$4(\bar{r}_k - 1) < r_k \quad \text{and} \quad |\bar{\tau}_k - \tau_k| \leq \bar{r}_k - 1. \tag{5}$$

Remark that the second condition is automatically satisfied if we take $\bar{\tau}_k$ as the best approximation of τ_k in $\mathcal{D}_{\bar{r}_k}$ and if we make the following assumption of approximation on the grid:

(App): At all scales $r \in \mathcal{R}$ and for any integer $l \in [r+1, n-r]$, there exists a location $l' \in \mathcal{D}_r$ such that

$$|l' - l| \leq r - 1 \ . \tag{6}$$

This property **(App)** entails that any point l can be approximated at distance $r-1$ by some location in $\mathcal{D}_r$. This also implies that each point $l \in [r+1, n-r]$ belongs to at least one segment $(l_1 - r, l_1 + r)$ where l_1 lies in $\mathcal{D}_r$. In practice, the a-adic grids $\mathcal{G}_a$ and the complete grid satisfy **(App)**.

Next, we introduce an event on the tests under which the change-point procedure performs well. In the following, we write $\mathcal{H}_0$, the collection of all possible $(l, r) \in J_n$ such that no change-point occurs in $[l - r + 1, l + r - 1]$, i.e. $\Gamma(\mathbb{P}_t)$ is constant on $[l - r, l + r)$. Equivalently, we have

$$(l, r) \in \mathcal{H}_0 \quad \text{iff} \quad (l - r, l + r) \cap \{\tau_k, k = 1, \dots, K\} = \varnothing \ . \tag{7}$$

For some elements of the grid $(\bar{\tau}_k, \bar{r}_k)$ satisfying (5), we say that the event $\mathcal{A}(T, \mathcal{K}^*, (\bar{\tau}_k, \bar{r}_k)_{k \in \mathcal{K}^*})$ holds if T satisfies the two following properties:

1. **(No False Positive)** $T_{l,r} = 0$ for all $(l, r) \in \mathcal{H}_0 \cap \mathcal{G}$
2. **(Significant change-point Detection)** for every $k \in \mathcal{K}^*$, we have $T_{\bar{\tau}_k, \bar{r}_k} = 1$.

The first property states that T performs no type I errors on the event $\mathcal{A}(T, \mathcal{K}^*, (\bar{\tau}_k, \bar{r}_k)_{k \in \mathcal{K}^*})$, whereas the second property states that all the significant change-points are detected on the event $\mathcal{A}(T, \mathcal{K}^*, (\bar{\tau}_k, \bar{r}_k)_{k \in \mathcal{K}^*})$.

Theorem 1. *The following holds for any grid $\mathcal{G}$, any local test statistic T, any non-negative integer K, any distribution with K change-points, any $\mathcal{K}^* \subset [K]$ and scales and locations $(\bar{\tau}_k, \bar{r}_k)_{k \in \mathcal{K}^*}$ in $\mathcal{G}$ satisfying Assumption (5). Under the event $\mathcal{A}(T, \mathcal{K}^*, (\bar{\tau}_k, \bar{r}_k)_{k \in \mathcal{K}^*})$, the estimated change-point vector $\hat{\tau}$ returned by Algorithm 1 satisfies the two following properties*

- **Significant change-points are detected and reasonably localized**: *for all $k \in \mathcal{K}^*$, there exists $k' \leq \hat{K}$ such that*

$$|\hat{\tau}_{k'} - \tau_k| \leq \bar{r}_k - 1 < \frac{r_k}{4} \ ,$$

- **No Spurious change-point is detected**: *for all $k \leq K$,*

$$\left| \{\hat{\tau}_{k'},\ 1 \leq k' \leq \hat{K}\} \cap \left[\tau_k - \frac{\tau_k - \tau_{k-1}}{2}, \tau_k + \frac{\tau_{k+1} - \tau_k}{2} \right] \right| \leq 1 \ .$$

and no spurious change-point is detected near the threshold of the time series

$$\{\hat{\tau}_{k'}, k' \leq \hat{K}\} \subset \left[\tau_1 - \frac{\tau_1 - 1}{2}, \tau_K + \frac{n + 1 - \tau_K}{2} \right] \ .$$

The first property states that so-called significant change-points $(\tau_k)_{k \in \mathcal{K}^*}$ are detected by the generic algorithm at the right scale. The no-spurious property guarantees that, around any true change-point τ_k, the procedure estimates at most one single change-point $\widehat{\tau}_l$.

Importantly, the theorem does not make any assumption on the non-significant change-points. In fact, change-points τ_k with $k \in [K] \setminus \mathcal{K}^*$ may or may not be detected. In general, we can only conclude from Theorem 1 that $|\mathcal{K}^*| \le \widehat{K} \le K$ on the event $\mathcal{A}(T, \mathcal{K}^*, (\bar{\tau}_k, \bar{r}_k)_{k\in\mathcal{K}^*})$.

Theorem 1 is abstract but its main virtue is to translate multiple testing properties into change-point estimation properties. For a specific problem such as multivariate mean change-point detection considered in the next section, the construction of a near optimal procedure boils down to introducing a collection of local test statistics, such that (a) change-points τ_k belong to $\mathcal{K}^*$ under suitable and minimal assumptions, (b) the scale $\bar{r}_k$ is the smallest possible, and (c) the event $\mathcal{A}(T, \mathcal{K}^*, (\bar{\tau}_k, \bar{r}_k)_{k\in\mathcal{K}^*})$ holds with high probability.

In the case where all the change-points are significant change-points, the result of Theorem 1 can be formulated as follows:

Corollary 1. *Assume that $\mathcal{K}^* = \{1, \dots, K\}$. The following holds for any grid $\mathcal{G}$, any local test statistic T, any non-negative integer K, any distribution with K change-points, any scales and locations $(\bar{\tau}_k, \bar{r}_k)_{k\in\mathcal{K}^*}$ in $\mathcal{G}$ satisfying Assumption (5). Under the event $\mathcal{A}(T, \mathcal{K}^*, (\bar{\tau}_k, \bar{r}_k)_{k\in\mathcal{K}^*})$, the estimated change-point vector $\hat{\tau}$ returned by Algorithm 1 satisfies the two following properties*

- $\hat{K} = K$,
- *for all $k \in K$, $|\hat{\tau}_k - \tau_k| < \bar{r}_k - 1 \le \frac{r_k}{4}$*

Let us respectively define the Hausdorff distance and the Wasserstein distance of two vectors $(u_1, \dots, u_K)$ and $(v_1, \dots, v_K)$ in $\mathbb{R}^K$ by $d_H(u, v) = \max_{k=1,\dots,K} |u_k - v_k|$ and $d_W(u, v) = \sum_{k=1,\dots,K} |u_k - v_k|$. Then, Corollary 1 implies that, if $\mathcal{K}^* = \{1, \dots, K\}$, then

$$d_H(\hat{\tau}, \tau) \le \max_{k=1,\dots,K} (\bar{r}_k - 1) \quad \text{and} \quad d_W(\hat{\tau}, \tau) \le \sum_{k=1,\dots,K} (\bar{r}_k - 1) \ .$$

3 Multivariate Gaussian change-point detection

We now turn to the multivariate change-point model introduced in Section 1.2. Throughout this section, we assume that the random vector ε_t are independently and identically distributed with $\varepsilon_t \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_p)$. Since we shall apply the general aggregation procedures introduced in the previous section, our main job here is to introduce a near-optimal testing procedure.

Fix some quantity $\delta \in (0, 1)$. At the end of the section, $1 - \delta$ will correspond to the probability of the event $\mathcal{A}(T, \mathcal{K}^*, (\bar{\tau}_k, \bar{r}_k)_{k\in\mathcal{K}^*})$ introduced in the previous section. Alternatively, one may interpret δ as an upper bound of the desired probability that the change-point detection procedure detects a spurious change-points. Recall that, for a change-point τ_k, s_k stands for the sparsity of the difference $\mu_{k+1} - \mu_k$. We say that the energy of a given change-point τ_k is c_0-high if

$$r_k \Delta_k^2 \ge c_0 \sigma^2 \left[s_k \log\left(1 + \frac{\sqrt{p}}{s_k} \sqrt{\log\left(\frac{n}{r_k \delta}\right)} \right) + \log\left(\frac{n}{r_k \delta}\right) \right] \ , \tag{8}$$

for some universal constant c_0 to be defined later. We show in this section that when c_0 is large enough, all high-energy change-points can be detected. Conversely, it is established in Section 5 that Condition (8) is (up to a multiplicative constant) optimal and cannot be weakened.

Let us now discuss the different regimes contained in Equation (8). In what follows, define

$$\psi^{(g)}_{n,r,s} := s \log\left(1 + \frac{\sqrt{p}}{s} \sqrt{\gamma_r} \right) + \gamma_r \ ; \qquad \gamma_r := \log\left(\frac{n}{r\delta}\right) \ ,$$

in order to alleviate notations. If $\gamma_r \ge p/2$, then $\psi^{(g)}_{n,r,s} \asymp \gamma_r$. This corresponds to the minimal energy condition for detection in the univariate case - i.e. when $p = 1$ - see [VFLRB20]. The condition $\gamma_r \ge p/2$

occurs when p is rather small and the scale r is much smaller than n. If $\gamma_r \leq p/2$, then

$$\psi^{(g)}_{n,r,s} \asymp \begin{cases} \gamma_r & \text{if } s \leq \frac{\gamma_r}{\log(p)-\log(\gamma_r)} \\ s\log\left(2\frac{p}{s^2}\gamma_r\right) & \text{if } \frac{\gamma_r}{\log(p)-\log(\gamma_r)} < s < \sqrt{p\gamma_r} \\ \sqrt{p\gamma_r} & \text{if } s \geq \sqrt{p\gamma_r} \ . \end{cases}$$

We define $\mathcal{K}^* \subset [K]$ as the subset of indices such that τ_k satisfies (8). For any $k \in \mathcal{K}^*$, we define r^*_k as the minimum radius r such that an inequality similar to (8) is satisfied for $r\Delta_k^2$, namely

$$r^*_k = \min\left\{ r \in \mathbb{R}^+ : \quad r\Delta_k^2 \geq c_0\sigma^2\left[s_k \log\left(1 + \frac{\sqrt{p}}{s_k}\sqrt{\log\left(\frac{n}{r\delta}\right)}\right) + \log\left(\frac{n}{r\delta}\right)\right]\right\} \ . \tag{9}$$

In the following, we introduce multi-scale tests for respectively dense and sparse change-points. In this section, we restrict our attention to the dyadic grid $\mathcal{G}_D = (\mathcal{R}, \mathcal{D})$ introduced in the previous section (see Equation (4)). To apply Theorem 1, a quantity $\bar{r}_k \in \mathcal{R}$ will be introduced in the proof of Corollary 2 which will be of the same order as $r^*_k \in \mathbb{R}^+$.

3.1 Dense change-points

We focus here on dense change-points for which s_k is possibly as large as p. Given $\kappa > 0$, we say that τ_k is a κ-*dense* high-energy change-point if

$$r_k\Delta_k^2 \geq \kappa\sigma^2\left(\sqrt{p\log\left(\frac{n}{r_k\delta}\right)} + \log\left(\frac{n}{r_k\delta}\right)\right) \ . \tag{10}$$

Note that the requirement (10) is analogous to (8) when $s_k \geq \sqrt{p\log(\frac{n}{r_k\delta})}$. For any κ-dense high-energy change-point, we define $\bar{r}^{(\mathrm{d})}_k \in \mathcal{R}$ as the minimum radius $r \in \mathcal{R}$ such that an inequality of the same type as (10) is satisfied for $r\Delta_k^2$,

$$\bar{r}^{(\mathrm{d})}_k = \min\left\{ r \in \mathcal{R} : 8r\Delta_k^2 \geq \kappa\sigma^2\left(\sqrt{p\log\left(\frac{n}{r\delta}\right)} + \log\left(\frac{n}{r\delta}\right)\right)\right\} \ .$$

Intuitively, $\bar{r}^{(\mathrm{d})}_k$ corresponds to the smallest scale such that τ_k is guaranteed to be detected. By definition, we have $4(\bar{r}^{(\mathrm{d})}_k - 1) \leq r_k$. Let $\bar{\tau}^{(d)}_k$ be the best approximation τ_k in the grid $\mathcal{D}_{\bar{r}^{(\mathrm{d})}_k}$ at scale $\bar{r}^{(\mathrm{d})}_k$. By definition of the dyadic grid, we have $|\bar{\tau}^{(d)}_k - \tau_k| \leq \bar{r}^{(\mathrm{d})}_k/4$.

For any positive integers $r \in [2; n]$ and $l \in [r+1, n+1-r]$, we define the statistic $\Psi^{(\mathrm{d})}_{l,r} := \left\|\mathbf{C}_{l,r}\right\|^2 - p$. If θ is constant over $[l-r, l+r)$, then the expectation of $\Psi^{(\mathrm{d})}_{l,r}$ is zero. Recall that the rescaled CUSUM statistic $\mathbf{C}_{l,r}$ depends on the noise level σ, and the statistic $\Psi^{(\mathrm{d})}_{l,r}$ therefore requires the knowledge of σ. To calibrate the corresponding test $T^{(\mathrm{d})}_{l,r}$ rejecting for large values of $\Psi^{(\mathrm{d})}_{l,r}$ we introduce

$$T^{(\mathrm{d})}_{l,r} := \mathbf{1}\left\{\Psi^{(\mathrm{d})}_{l,r} > x^{(\mathrm{d})}_r\right\} \ ; \qquad x^{(\mathrm{d})}_r := 4\left(\sqrt{p\log\left(\frac{2n}{r\delta}\right)} + \log\left(\frac{2n}{r\delta}\right)\right) \ .$$

Proposition 1. *There exists a universal constant $\kappa_{\mathrm{d}} > 0$ and an event $\xi^{(\mathrm{d})}$ of probability larger than $1 - 2\delta$ such that the two following properties hold on $\xi^{(\mathrm{d})}$:*

1. **(No False Positive)** $T^{(\mathrm{d})}_{l,r} = 0$ *for all* $(l, r) \in \mathcal{H}_0 \cap \mathcal{G}_D$.
2. **(Dense High-energy change-point Detection)** *for any κ_{d}-dense high-energy change-point τ_k, one has* $T^{(\mathrm{d})}_{\bar{\tau}^{(d)}_k, \bar{r}^{(\mathrm{d})}_k} = 1$.

The above proposition ensures that, on the event $\xi^{(\mathrm{d})}$, the collection of tests $T_{l,r}^{(\mathrm{d})}$ detects all dense high energy change-points at the scale $\bar{r}_k^{(\mathrm{d})}$ and has no false positives on the dyadic grid $\mathcal{G}_D$. If we plugged this collection of tests into the general multiple change-point procedure, then Theorem 1 would entail that all κ_{d}-dense high-energy change-points are discovered and localized and that $\widehat{\tau}$ does not detect any spurious change-point. In the next subsection, we introduce alternative tests that are tailored to sparse change-points and thereby allow to detect change-points that are not κ_{d}-dense high-energy but still satisfy the energy condition (8).

3.2 Sparse change-points

3.2.1 Energy condition

For a given $1 \le k \le K$ The change-point τ_k is a κ-*sparse* high-energy change-point if $s_k \le \sqrt{p \log\left(\frac{n}{r_k \delta}\right)}$ and

$$r_k \Delta_k^2 \ge \kappa \sigma^2 \left(s_k \log\left(\frac{p}{s_k^2} \log\left(\frac{n}{r_k \delta} \right) \right) + \log\left(\frac{n}{r_k \delta} \right) \right) \quad . \tag{11}$$

If τ_k is a κ-sparse high-energy change-point, we define $\bar{r}_k^{(\mathrm{s})}$ as the minimum scale such that an inequality similar to (11) is satisfied :

$$\bar{r}_k^{(\mathrm{s})} = \min \left\{ r \in \mathcal{R} : \quad 8 r \Delta_k^2 \ge \kappa \sigma^2 \left(s_k \log\left(\frac{p}{s_k^2} \log\left(\frac{n}{r \delta} \right) \right) + \log\left(\frac{n}{r \delta} \right) \right) \right\} \quad .$$

As in the dense case, we have $4(\bar{r}_k^{(\mathrm{s})} - 1) \le r_k$. Set $\bar{\tau}_k^{(s)}$ as the best approximation of τ_k in the grid $\mathcal{D}_{\bar{r}_k^{(\mathrm{s})}}$ at scale τ_k. By definition of the dyadic grid, we have $|\bar{\tau}_k^{(s)} - \tau_k| \le \bar{r}_k^{(\mathrm{s})}/4$. We introduce below two statistics for handling the testing problem.

3.2.2 Berk-Jones Test

The Berk-Jones test [MNS+16] is a variation of the Higher-Criticism test originally introduced in [DJ04] for signal detection. We decided to use the Berk-Jones test in this paper because of its intrinsic formulation in terms of the quantiles of a Bernoulli distribution, but the Higher-Criticism test would reach the same rates of detection within a constant factor. Given (l, r) in the grid $\mathcal{G}_D$, we first introduce $N_{x,l,r}$ as the number of coordinates of $\mathbf{C}_{l,r}$ that are larger than x in absolute value.

$$N_{x,l,r} = \sum_{i=1}^{p} \mathbf{1}_{|\mathbf{C}_{l,r,i}| > x} \tag{12}$$

If $(l, r) \in \mathcal{H}_0$, then the rescaled CUSUM statistic follows a standard normal distribution and $N_{x,l,r}$ therefore follows a Binomial distribution with parameters p and $2\overline{\Phi}(x)$. The Berk-Jones test amounts to rejecting the null, when at least one of the statistics $N_{x,l,r}$, for $x \in \mathbb{N}^*$, is significantly large. Next, we formalize what we mean by 'large'.

For any $u > 0$, any $q_0 \in [0, 1]$, and positive integer p_0, denote $\overline{Q}(u, p_0, q_0) = \mathbb{P}[\mathcal{B}(p_0, q_0) > u]$ the tail distribution function of a Binomial distribution with parameters p_0 and q_0. Given $\alpha \in [0, 1]$, we then write $\overline{Q}^{-1}(\alpha, q, p_0)$ for the corresponding quantile function,

$$\overline{Q}^{-1}(\alpha, p_0, q_0) = \inf_u \left[\mathbb{P}[\mathcal{B}(p_0, q_0) > u] \le \alpha \right] \quad .$$

Given a scale $x \in \mathcal{R}$ and a positive integer x, we define the weights

$$\alpha_{x,r}^{(BJ)} = \frac{6\delta r}{\pi^2 x^2 |\mathcal{D}_r| n} \quad . \tag{13}$$

This allows us to define the Berk-Jones statistic over $[l - r, l + r)$ as the test rejecting the null when at least one $N_{x,l,r}$ is large.

$$T_{l,r}^{(\mathrm{BJ})} = \max_{x \in \mathbb{N}^*} \mathbf{1} \left\{ N_{x,l,r} > \overline{Q}^{-1}(\alpha_{x,r}^{(BJ)}, p, 2\overline{\Phi}(x)) \right\} \quad . \tag{14}$$

Equivalently, $T_{l,r}^{(\mathrm{BJ})}$ is an aggregated test based on the statistics $N_{x,l,r}$ with weights $\alpha_{x,r}^{(BJ)}$. From the above remark and a union bound, we deduce that the probability that the collection of tests $\{T_{l,r}^{(\mathrm{BJ})}, (l,r) \in \mathcal{G}_D\}$ rejects a least one false positive is at most δ:

$$\mathbb{P}\left[\max_{(l,r)\in\mathcal{H}_0\cap\mathcal{G}_D} T_{l,r}^{(\mathrm{BJ})} = 1\right] \le \sum_{r\in\mathcal{R}}\sum_{l\in\mathcal{D}_r}\sum_{x\in\mathbb{N}^*} \alpha_{x,r}^{(BJ)} \le \sum_{r\in\mathcal{R}}\sum_{l\in\mathcal{D}_r} \frac{\delta r}{|\mathcal{D}_r| n} \le \sum_{r\in\mathcal{R}} \frac{\delta r}{n} \le \delta \ ,$$

where we recall that $(l,r) \in \mathcal{H}_0$ if and only if Θ is constant on $[l-r, l+r)$. Although one may think from the definition (14) that $T_{l,r}^{(\mathrm{BJ})}$ involves an infinite number of $N_{x,l,r}$, this is not the case. Indeed, $N_{x,l,r}$ is a non-increasing function of x whereas for all x such that $2p\overline{\Phi}(x) \le \alpha_{x,r}^{(BJ)}$, we have $\overline{Q}^{-1}(\alpha_{x,r}^{(BJ)}, p, 2\overline{\Phi}(x)) = 0$. Writing $x_{0,r}$ the smallest x such that $2p\overline{\Phi}(x) \le \alpha_{x,r}^{(BJ)}$ we derive

$$T_{l,r}^{(\mathrm{BJ})} = \max_{x=1,\dots,x_{0,r}} \mathbf{1}\left\{N_{x,l,r} > \overline{Q}^{-1}(\alpha_{x,r}, p, 2\overline{\Phi}(x))\right\} \ .$$

Since, for any $x > 0$, we have $\overline{\Phi}(x) \le e^{-x^2/2}$, one can deduce that $x_{0,r} \le c\sqrt{\log\left(\frac{np}{r\delta}\right)}$, for some numerical constant $c > 0$.

3.2.3 Partial norm statistics

The Berk Jones test is able to detect change-points τ_k for which there exists s such that the s largest squared coordinates of $\mu_k - \mu_{k-1}$ are larger than $C(\log(ep/s^2) + \log(n/r_k)/s)$ with a big enough constant C. However, it may happen that τ_k satisfies the energy condition (8) and that the s largest coordinates of $\mu_k - \mu_{k-1}$ are negligible with respect to $\log(n/r_k)/s$, mainly because $s \mapsto 1/s$ is not summable. To solve this issue, we introduce a second sparse statistic based on the partial sums. Let

$$\mathcal{Z} = \left\{1, 2, 2^2, \dots, 2^{\lfloor \log_2(p)\rfloor}\right\}$$

denote the dyadic set. Only the sparsities $s \in \mathcal{Z}$ will be analysed by the partial norm statistic. For any (l,r) in the Grid $\mathcal{G}_D$, we respectively write $\mathbf{C}_{l,r,(1)}, \mathbf{C}_{l,r,(2)}, \dots$ the reordered entries of $\mathbf{C}_{l,r}$ by decreasing absolute value, that is $|\mathbf{C}_{l,r,(1)}| \ge \cdots \ge |\mathbf{C}_{l,r,(p)}|$. Then, for $s \in \mathcal{Z}$, we define the partial CUSUM norm by

$$\Psi_{l,r,s}^{(p)} = \sum_{i=1}^{s} \left(\mathbf{C}_{l,r,(i)}\right)^2 \ . \tag{15}$$

Then, we define the test $T_{l,r}^{(p)}$ rejecting the null when at least one of the partial norms is large

$$x_{r,s}^{(p)} := x_{r,s}^{(p)}(\delta) = 4s\log\left(\frac{2ep}{s}\right) + 4\log\left(\frac{n}{r\delta}\right); \qquad T_{l,r}^{(p)} = \max_{s\in\mathcal{Z}} \mathbf{1}\left\{\Psi_{l,r,s}^{(p)} > x_{r,s}^{(p)}\right\} \ .$$

Finally, we define the sparse test by aggregating both the Berk-Jones test and the partial norm test. For any $(l,r) \in \mathcal{G}_D$, let $T_{l,r}^{(\mathrm{s})} = T_{l,r}^{(p)} \vee T_{l,r}^{(\mathrm{BJ})}$. The next proposition controls the error of this collection of tests.

Proposition 2. *There exists a universal constant $\kappa_{\mathrm{s}} > 0$ and an event $\xi^{(\mathrm{s})}$ of probability larger than $1 - 4\delta$ such that the two following properties hold on $\xi^{(\mathrm{s})}$:*

1. **(No False Positive)** *$T_{l,r}^{(\mathrm{s})} = 0$ for all $(l,r) \in \mathcal{H}_0 \cap \mathcal{G}_D$.*
2. **(Sparse High-energy change-point Detection)** *for any κ_{s}-sparse high-energy change-point τ_k, one has $T^{(\mathrm{s})}_{\bar{\tau}_k^{(s)}, \bar{r}_k^{(\mathrm{s})}} = 1$.*

Here we introduced two different statistics for the same sparse regime $s_k \le \sqrt{p \log\left(\frac{n}{r_k \delta}\right)}$ - the Berk Jones statistic and the partial sums statistic - essentially to solve a problem of integrability. We made this choice for the sake of simplicity, but we could have used a single test, as presented in [LGS19]

$$\Psi^{(LGS)}_{x,l,r} = \sum_{i=1}^{p} \left(\mathbf{C}^2_{l,r,i} - \mathbb{E}\left[Z | Z \ge x \right] \right) \mathbf{1}\{ \mathbf{C}^2_{l,r,i} \ge x \} \ ,$$

where Z follows a standard normal distribution $\mathcal{N}(0,1)$. This statistic gives the same type of result as the Berk Jones statistic when $\mu_k - \mu_{k-1}$ has large enough coordinates, and it is comparable to the partial sums statistic when its threshold x becomes low enough.

3.3 Consequences

To conclude this section, it suffices to observe that, for c_0 in (8), any c_0-high-energy change-point τ_k in the sense of (8) is either a $\frac{c_0}{2}$-dense or a $\frac{c_0}{2}$-sparse high-energy change-point. Hence, upon defining the test $T_{l,r} = T^{(\mathrm{d})}_{l,r} \vee T^{(\mathrm{s})}_{l,r}$ for $(l,r) \in \mathcal{G}_D$, we consider the change-point procedure $\widehat{\tau}$ defined in Algorithm 1. Gathering Theorem 1 with Proposition 1 and Proposition 2, we obtain the following.

Corollary 2. *There exists an universal constant $c_0 > 0$ such that the estimator $\widehat{\tau}$ based on $(T_{l,r})$ with $(l,r) \in \mathcal{G}_D$ satisfies the following properties with probability higher than $1 - 6\delta$*

- *$\widehat{\tau}$ does not detect any* **spurious** *change-points.*
- *c_0* **High-energy change-points are detected and well localized.** *for all k satisfying (8), there exists $l \le \hat{K}$ such that $|\hat{\tau}_l - \tau_k| < \frac{r^*_k}{2} \le \frac{r_k}{2}$, where r^*_k is defined by (9).*

If the change-points are of high-energy, that is $\mathcal{K}^* = \{1, \dots, K\}$, then Corollary 2 can be reformulated as follows:

Corollary 3. *Assume that for all $k = 1, \dots, K$, τ_k is a c_0-high-energy change-point (see (8)). If the constant c_0 is large enough, then the estimator $\widehat{\tau}$ based on $(T_{l,r})$ with $(l,r) \in \mathcal{G}_D$ satisfies the following properties with probability higher than $1 - 6\delta$:*

- *$\hat{K} = K$,*
- *for all $k \in K$, $|\hat{\tau}_k - \tau_k| < \frac{r^*_k}{2} \le \frac{r_k}{2}$*

In particular, one can respectively bounds the Hausdorffand the Wasserstein distance between $\hat{\tau}$ and τ by

$$d_H(\hat{\tau}, \tau) \le \max_{k=1,\dots,K} \frac{r^*_k}{2} \quad \text{and} \quad d_W(\hat{\tau}, \tau) \le \sum_{k=1,\dots,K} \frac{r^*_k}{2} \ , \tag{16}$$

with probability higher than $1 - \delta$.

In Section 5, we establish that the Condition (8) is (up to a multiplicative constant) is unimprovable and corresponds to the detection threshold for multivariate change-points.

Corollary 3 can be compared to the result of [WS18] on multivariate change-point detection in the multiple change-point setting. Using a method based on the CUSUM statistic and assuming that there are only high-energy change-points, they also obtain an upper bound on the energy necessary to detect the change-points. However, their result does not adapt to r_k, Δ_k, s_k, and their rate of detection is suboptimal in many regimes. Writing $r = \min_{k=1,\dots,K}(r_k)$, $\Delta = \min_{k=1,\dots,K}(\Delta_k)$ and $s = \max_{k=1,\dots,K}(s_k)$, Theorem 5 of [WS18] requires two conditions of the type $r\Delta^2 \gtrsim (\frac{n}{r})^4 \log(np)$ and $r\Delta^2 \gtrsim s\frac{n}{r}\log(np)$. Their rate of detection is thus suboptimal by a polynomial factor in n/r when r is of smaller order than n, and by a logarithmic factor $\log(np)$ instead of $\log(1 + \sqrt{p}/s \log(n/r)) + \frac{1}{s}\log(n/r)$ when r is of order n.

Comparison to the setting with only one change-point When the statistician knows that there is at most one change-point τ_1, then [LGS19] proved that it is possible to detect τ_1 if and only if $r_1\Delta_1^2 \gtrsim \sigma^2\left(s_1\log(1+\frac{1}{s_1}\sqrt{p\log\log 8n}) + \log\log 8n\right)$. The problem with only one change-point is more simple, and the paper uses similar statistics based on the CUSUM - a chi square statistics in the dense case and a thresholded sum of squared coordinates in the sparse case - to detect and localize τ_1. It turns out that the detection procedure of [LGS19] adapts to distance $r_1 = \max(\tau_1 - 1, n+1-\tau_1)$ of the change-point position to the boundary, and one could refine their result by stating that τ_1 is detectable if and only if $r_1\Delta_1^2 \gtrsim \sigma^2\left(s_1\log(1+\frac{1}{s_1}\sqrt{p\log\log(2n/r_1)}) + \log\log(2n/r_1)\right)$ which is more precise when r_1 is of the order of n. This refined result is related to our bounds in the mutiple change-point settings, but the rate is faster as they obtain a term in $\log\log(n/r_1)$ - instead of $\log(n/r_k)$ in our case. The reason for this faster rate is due to the relative simplicity of the problem with only one change-point. indeed, in single change-point detection, there is no need to look for change-points at all positions and scale at the same time, since scale and positions are related. This implies that it is possible to attain faster rates than in multiple change-point detection. The comparison between single and multiple change-point detection is thoroughly done in [VFLRB20] for univariate models.

4 Multi-scale change-point estimation with sub-Gaussian noise

In this subsection, we turn to the more general case of sub-Gaussian distributions [Ver18]. Given a random variable Z, define its ψ_2-norm by $\|Z\|_{\psi_2} = \inf\{x > 0,\ \mathbb{E}[\exp(Z^2/x^2)] \le 2\}$. Given $L > 0$, a mean zero real random variable is said to be L-sub-Gaussian if $\|Z\|_{\psi_2} \le L$. This implies in particular that, for all $x \ge 0$, one has $\mathbb{P}(|Z| \ge x) \le 2\exp(-x^2/L^2)$. For $p \ge 1$, a random vector $(X_1, \dots, X_p)$ taking values in $\mathbb{R}^p$ is said to be L-sub-Gaussian if its coordinates (X_i) are independent and L-sub-Gaussian, that is for every $i = 1, \dots, p$, one has $\|X_i\|_{\psi_2} \le L$.

Throughout this section, we assume that for $t = 1, \dots, n$, the random vectors ε_t are independent and L-sub-Gaussian, and that $\mathrm{Var}(\varepsilon_i) = \sigma^2\mathbf{I}_p$. As in the previous section, we apply the general aggregation procedures introduced in Section 2. As a consequence, our main task boils down to introducing a near-optimal multiple testing procedure indexed by a grid for detecting the existence of a change-point. Unlike the previous section, we shall rely on the complete grid $\mathcal{G}_F = J_n = \left\{(l,r) \ :\ r = 1, \dots, \left\lfloor \frac{n}{2} \right\rfloor \text{ and } l = r+1, \dots, n-r\right\}$ whose size is quadratic with respect to n. All the results in this section are still valid (but with different numerical constants) if we had kept the dyadic grid $\mathcal{G}_D$. Still, we use the complete grid here as a proof of concept that one can rely on the full collection of possible segments without deteriorating the rates. A detailed comparison between the complete and dyadic grids is made in Section 6.

In order to emphasize the common points with the previous section, we use the same notation $\mathcal{K}^*$ for the collection of high-energy change-points[1], $\bar{r}_k$ for the scales associated to the k-th change-points[2], Ψ for the statistics, T for the test and x for the thresholds although these quantities are slightly changed to handle the sub-Gaussian tail distribution. We follow the same scheme as for the Gaussian case and first introduce mutli-scale tests for dense change-points before turning to sparse change-points. As in the previous section, we consider some $\delta \in (0,1)$ corresponding to the type I error probability.

4.1 Dense change-points with sub-Gaussian noise

Recall that, for a change-point τ_k, s_k stands for the sparsity of the difference $\mu_{k+1} - \mu_k$. We focus here on dense change-points for which s_k is possibly as large as p. Given $\kappa > 0$, we say that τ_k is a κ-dense high-energy change-point if

$$r_k\Delta_k^2 \ge \kappa L^2\left(\sqrt{p\log\left(\frac{n}{r_k\delta}\right)} + \log\left(\frac{n}{r_k\delta}\right)\right) \quad . \tag{17}$$

This condition is very similar to its counterpart (10) for Gaussian noise. Still, we introduce it here for the sake of completeness. For $k \in \{1, \dots, K\}$ such that τ_k is a κ-dense high-energy change-point, we

[1] See Equation (20) as the energy condition is slightly different in the sub-Gaussian setting.

[2] Re-defined in Equation (21).

define $\bar{r}_k^{(\mathrm{d})}$ as the minimum length such that an inequality similar to (17) is satisfied :

$$\bar{r}_k^{(\mathrm{d})} = \min\left\{ r \in \mathbb{N}^* : \quad 4r\Delta_k^2 \geq \kappa L^2 \left(\sqrt{p \log\left(\frac{n}{r\delta}\right)} + \log\left(\frac{n}{r\delta}\right) \right) \right\} \quad .$$

As in the Gaussian case in Section 3, $\bar{r}_k^{(\mathrm{d})}$ corresponds to the smallest scale such that τ_k is guaranteed to be detected. For any κ-dense high-energy change-point, it holds that $4(\bar{r}_k^{(\mathrm{d})} - 1) < r_k$. For any positive integers $(l, r) \in \mathcal{G}_F$, we consider the same CUSUM-based statistic $\Psi_{l,r}^{(\mathrm{d})} := \left\| \mathbf{C}_{l,r} \right\|^2 - p$ as for Gaussian noise. Let $\bar{c}_{\mathrm{thresh}}^{(d)} > 0$ be a tuning parameter to be discussed later. To calibrate the corresponding multiple test procedures $(T_{l,r}^{(\mathrm{d})})$ with $(l, r) \in \mathcal{G}_F$ rejecting for large values of $\Psi_{l,r}^{(\mathrm{d})}$ we introduce

$$T_{l,r}^{(\mathrm{d})} := \mathbf{1}\left\{ \Psi_{l,r}^{(\mathrm{d})} > x_r^{(\mathrm{d})} \right\} \; ; \qquad x_r^{(\mathrm{d})} = \bar{c}_{\mathrm{thresh}}^{(d)} \frac{L^2}{\sigma^2} \left(\sqrt{p \log\left(\frac{n}{r\delta}\right)} + \log\left(\frac{n}{r\delta}\right) \right) \quad .$$

Proposition 3. *There exists a numerical constant $\bar{c}_{\mathrm{thresh}}^{(d)} > 0$ such that the following holds for any $\kappa_{\mathrm{d}} > 32\bar{c}_{\mathrm{thresh}}^{(d)}$. With probability higher than $1 - \delta$, the two following properties hold*

1. **(No False Positives)** *one has $T_{l,r}^{(\mathrm{d})} = 0$ for any $(l, r) \in \mathcal{G}_F \cap \mathcal{H}_0$.*
2. **(Dense high-energy change-points are detected)** *one has $T^{(\mathrm{d})}_{\tau_k, \bar{r}_k^{(\mathrm{d})}} = 1$ for any κ_{d}-dense high-energy change-point τ_k.*

In comparison to Proposition 1 in the previous section, there are two differences. First, we need to cope with SubGaussian distribution by applying the Hanson-Wright inequality. Most importantly, the grid $\mathcal{G}_F$ is much larger than $\mathcal{G}_D$ so that we cannot simply consider each test $T_{l,r}$ separately and simply apply a union bound as in the previous section. To handle the dependencies between the statistics $\Psi_{l,r}^{(\mathrm{d})}$, we have to apply a chaining argument. In fact, the thresholds $x_r^{(\mathrm{d})}$ are similar to their counterpart in the previous section, whereas the number $|\mathcal{G}_F|$ of tests is proportional to n^2. In principle, the benefit of using the full grid $\mathcal{G}_F$ is that $(\tau_k, \bar{r}_k^{(\mathrm{d})})$ belongs to $\mathcal{G}_F$ so that we can consider the CUSUM statistic based on a segment $[\tau_k - \bar{r}_k^{(\mathrm{d})}, \tau_k + \bar{r}_k^{(\mathrm{d})}]$ centered around the change-point τ_k. In contrast, $(\tau_k, \bar{r}_k^{(\mathrm{d})})$ does not necessarily belong to the dyadic grid $\mathcal{G}_D$ and we needed to consider its best approximation $(\bar{\tau}_k^{(d)}, \bar{r}_k^{(\mathrm{d})})$. The segment $[\bar{\tau}_k^{(d)} - \bar{r}_k^{(\mathrm{d})}, \bar{\tau}_k^{(d)} + \bar{r}_k^{(\mathrm{d})}]$ is therefore not centered on τ_k and the corresponding statistic $\Psi^{(\mathrm{d})}_{\bar{\tau}_k^{(d)}, \bar{r}_k^{(\mathrm{d})}}$ is in expectation smaller than $\Psi^{(\mathrm{d})}_{\tau_k, \bar{r}_k^{(\mathrm{d})}}$.

In summary, both the collections of dense tests $\Psi_{l,r}^{(\mathrm{d})}$ on $\mathcal{G}_D$ and $\mathcal{G}_F$ are able to detect change-points whose energy is (up to some multiplicative constants) higher than $L^2\Big[[p\log(\frac{n}{r_k\delta})]^{1/2} + \log(\frac{n}{r_k\delta})\Big]$.

4.2 Sparse change-points with sub-Gaussian noise

Unlike in the previous section, we do not know the exact distribution of the noise. As a consequence, the Berk-Jones test and more generally higher-criticism type tests cannot be applied to this setting. This is why we only rely on the partial norm statistic. Recall that $\mathcal{Z} = \left\{1, 2, 2^2, \ldots, 2^{\lfloor \log_2(p) \rfloor}\right\}$ stands for a dyadic set of sparsities. For $(l, r) \in \mathcal{G}_F$ and $s \in \mathcal{Z}$, we also recall that the partial CUSUM norm is defined as

$$\Psi_{l,r,s}^{(p)} = \sum_{i=1}^{s} \left(\mathbf{C}_{l,r,(i)} \right)^2 \quad .$$

Then, for any $(l, r) \in \mathcal{G}_F$, the test $T_{l,r}^{(p)}$ rejects the null when at least one of the partial norms is large

$$x_{r,s}^{(\mathrm{p})} = s + \bar{c}_{\mathrm{thresh}}^{(p)} \frac{L^2}{\sigma^2} \left[s \log\left(\frac{2ep}{s}\right) + \log\left(\frac{n}{r\delta}\right) \right]; \qquad T_{l,r}^{(p)} = \max_{s \in \mathcal{Z}} \mathbf{1}\left\{ \Psi_{l,r,s}^{(p)} > x_{r,s}^{(p)} \right\} \quad ,$$

where $\bar{c}^{(p)}_{\text{thresh}}$ is a tuning parameter in Proposition 4 below.

The partial norm test alone is not able to detect sparse high-energy change-points in the sense of (11) and we need to introduce a stronger condition on the energy. Given $\kappa > 0$, we say that a change-point τ_k is a κ-sparse high-energy change-point - in the sub-Gaussian setting - if $s_k \leq \sqrt{p \log\left(\frac{n}{r_k \delta}\right)}$ and

$$r_k \Delta_k^2 \geq \kappa L^2 \left[s_k \log\left(\frac{ep}{s_k}\right) + \log\left(\frac{n}{r_k \delta}\right)\right] \ . \tag{18}$$

Both Conditions (11) and (18) are compared at the end of the subsection. For a κ-sparse high-energy change-point τ_k, we define its scale $\bar{r}^{(\text{s})}_k$ by

$$\bar{r}^{(\text{s})}_k = \min\left\{ r \in \mathbb{N}^* : \quad 4r\Delta_k^2 \geq \kappa L^2 \left[s_k \log\left(\frac{ep}{s_k}\right) + \log\left(\frac{n}{r\delta}\right)\right]\right\} \ . \tag{19}$$

For any κ-sparse high-energy change-point, it holds that $4(\bar{r}^{(\text{s})}_k - 1) \leq r_k$.

Proposition 4. *There exists a numerical constant $\bar{c}^{(p)}_{\text{thresh}} > 0$ such that the following holds for any $\kappa_{\text{s}} > 32\bar{c}^{(p)}_{\text{thresh}}$. With probability higher than $1-\delta$, the two following properties hold*

1. **(No False Positives)** $T^{(\text{p})}_{l,r} = 0$ *for any* $(l,r) \in \mathcal{G}_F \cap \mathcal{H}_0$.
2. **(Sparse high-energy change-points are detected)** $T^{(\text{p})}_{\tau_k, \bar{r}^{(\text{s})}_k} = 1$ *for any κ_{s}-sparse high-energy change-point τ_k in the sense of* (18).

As for Proposition 3, the proof relies on a careful analysis of the joint distributions of the statistics $\Psi^{(p)}_{l,r,s}$ to handle the multiplicity of $\mathcal{G}_F$.

4.3 Consequences

Let $c_0 > 0$ be some constant that we will discuss later. A change-point τ_k is then said to be a c_0-high-energy change-points - in the sub-Gaussian setting - if

$$r_k \Delta_k^2 \geq c_0 L^2 \left[\left(\sqrt{p \log\left(\frac{n}{r_k\delta}\right)} \wedge \left(s_k \log\left(\frac{ep}{s_k}\right)\right)\right) + \log\left(\frac{n}{r_k\delta}\right)\right] \ . \tag{20}$$

We here re-introduce $\mathcal{K}^* \subset [K]$ as the subset of indices such that τ_k satisfies (20).

We gather both tests by considering, for any $(l,r) \in \mathcal{G}_F$, the test $T_{l,r} = T^{(\text{d})}_{l,r} \vee T^{(\text{p})}_{l,r}$ with tuning parameters $\bar{c}^{(d)}_{\text{thresh}}$ and $\bar{c}^{(p)}_{\text{thresh}}$ as in Propositions 3 and 4. Consider any $c_0 > 32(\bar{c}^{(d)}_{\text{thresh}} \vee \bar{c}^{(p)}_{\text{thresh}})$ and any c_0-high-energy change-point τ_k, which is either a c_0-sparse or a c_0-dense high-energy change-point. Defining

$$\bar{r}_k = \bar{r}^{(\text{d})}_k \wedge \bar{r}^{(\text{s})}_k, \tag{21}$$

we straightforwardly derive from Proposition 3 and Proposition 4 the following result.

Corollary 4. *There exists two numerical constants $\bar{c}^{(p)}_{\text{thresh}} > 0$ and $\bar{c}^{(d)}_{\text{thresh}} > 0$ such that the following holds. With probability higher than $1-\delta$, the tests $T_{l,r}$ with $(l,r) \in \mathcal{G}_F$ satisfy*

1. **(No False Positives)** $T_{l,r} = 0$ *for any* $(l,r) \in \mathcal{G}_F \cap \mathcal{H}_0$.
2. **(High-energy change-points are detected)** $T_{\tau_k, \bar{r}_k} = 1$ *for any c_0-high-energy change-point τ_k in the sense of* (20).

Then, it suffices to combine this multiple testing procedure with Algorithm 1 to get the change-point procedure $\widehat{\tau}$. Since, for a high-energy change-point in the sense of (20), we have $4(\bar{r}_k - 1) < r_k$, we are in position to apply Theorem 1.

Corollary 5. *There exist two numerical constant* $\bar{c}^{(p)}_{\text{thresh}} > 0$ *and* $\bar{c}^{(d)}_{\text{thresh}} > 0$ *such that the following holds. With probability higher than* $1-\delta$*, the estimator* $\widehat{\tau}$ *based on* $(T_{l,r})$ *with* $(l,r) \in \mathcal{G}_F$ *satisfies the two following properties*

- $\widehat{\tau}$ *does not detect any* **spurious** *change-points.*
- *Any* c_0*-high-energy change-point* τ_k *- in the sense of* (20) *- is* **detected** *and well localized, that is there exists* $l \le \hat{K}$ *such that*

$$|\hat{\tau}_l - \tau_k| \le \bar{r}_k - 1 \le \frac{r_k}{4} \ .$$

In the case where all change-points are c_0-high-energy change-points in the sense of (20), all change-points are detected, and something similar to Corollary 3 holds here, replacing $r^*_k/2$ by $\bar{r}_k - 1$. Also, both the Hausdorff distance and the Wasserstein distance, can be bounded as in Equation (16) - again, replacing $r^*_k/2$ by $\bar{r}_k - 1$.

Again, we could have obtained a similar result (but with different constants) using the dyadic grid $\mathcal{G}_D$ instead of $\mathcal{G}_F$. To conclude this section, let us compare the conditions (20) and (8). Define

$$\psi^{(sg)}_{n,r,s} = \sqrt{p\gamma_r} \wedge \left(s \log\left(\frac{ep}{s}\right)\right) + \gamma_r,$$

where we remind that $\gamma_r = \log\left(\frac{n}{r\delta}\right)$.

If $\gamma_r \ge p/2$, then $\psi^{(sg)}_{n,r,s} \asymp \gamma_r$. In low dimension, the energy threshold for multivariate change-point detection is the same as in the univariate setting, seev[VFLRB20]. If $\gamma_r \le p/2$, then

$$\psi^{(sg)}_{n,r,s} \asymp \begin{cases} \gamma_r & \text{if } s \le \frac{\gamma_r}{\log(p)-\log(\gamma_r)} \\ s\log\left(e\frac{p}{s}\right) & \text{if } \frac{\gamma_r}{\log(p)-\log(\gamma_r)} < s < \frac{\sqrt{p\gamma_r}}{\log(p)-\log(\gamma_r)} \\ \sqrt{p\gamma_r} & \text{if } s \ge \frac{\sqrt{p\gamma_r}}{\log(p)-\log(\gamma_r)} \end{cases}$$

As a consequence, $\psi^{(sg)}_{n,r,s}$ and $\psi^{(g)}_{n,r,s}$ are of the same order of magnitude for all s when $\gamma_r \ge p/2$. When $\log(n/r\delta) < p$, they are also of the same order of magnitude except when s is close but smaller than $\sqrt{p\gamma_r}$, for which the ratio $\psi^{(sg)}_{n,r,s}/\psi^{(g)}_{n,r,s}$ between these two quantities can be as large as $\log(p) - \log(\gamma_r)$. This gap corresponds to the regime where the test based on the Berk-Jones statistic defined in Equation (14), used in the Gaussian case, outperforms the test based on the partial CUSUM norm statistic defined in Equation (15).

5 Optimality of the result

In this section, we write for any $\Theta \in \mathbb{R}^{p\times n}$, the distribution of the time series $Y = (y_1, \dots, y_n)$ in the model (1) with Gaussian noise $\varepsilon_t \sim \mathcal{N}(0, \sigma^2 \mathbf{I}_p)$. In Section 3, we have established that any change-point satisfying the condition (8), that is

$$r_k \Delta_k^2 \ge c_0 \sigma^2 \left[s_k \log\left(1 + \frac{\sqrt{p}}{s_k}\sqrt{\log\left(\frac{n}{r_k\delta}\right)}\right) + \log\left(\frac{n}{r_k\delta}\right)\right] \ ,$$

is detected by our chane-point procedure. We show that this energy condition is unimprovable from a minimax point of view. More precisely, let us define, for any $u > 0$, the class $\bar{\mathcal{P}}_{n,p}(\sigma^2, u)$ as the set of mean parameters Θ satisfying for any change-point $1 \le k \le K$

$$r_k \Delta_k^2 \ge \frac{1}{2}\sigma^2 \left[s_k \log\left(1 + u\frac{\sqrt{p}}{s_k}\sqrt{\log\left(\frac{n}{r_k}\right)}\right) + u\log\left(\frac{n}{r_k}\right)\right] \ . \tag{22}$$

For u small enough, it turns out no change-point estimator is able to detect all change-points without estimating any spurious change-point with high probability on the full class $\bar{\mathcal{P}}_{n,p}(\sigma^2, u)$. Still, using this large class provides a somewhat pessimistic bounds. For instance, the most challenging distributions

in $\bar{\mathcal{P}}_{n,p}(\sigma^2, u)$ for the purpose of change-point detection satisfy $s_k = p$ and $r_k = 1$ (very close change-points). As a consequence, relying on the full collection $\bar{\mathcal{P}}_{n,p}(\sigma^2, u)$ is too pessimistic to assess the optimality of our results with respect to s_k and r_k. To adapt to these two quantities, we define, for any positive integers $1 \le r \le \lfloor n/2 \rfloor$ and any $1 \le s \le p$ the collection

$$\bar{\mathcal{P}}_{n,p}(\sigma^2, u, r, s) = \{\Theta \in \bar{\mathcal{P}}_{n,p}(\sigma^2, u) \ : \ \min_k r_k \ge r \text{ and } \max_k s_k \le s\} \ .$$

In the class $\bar{\mathcal{P}}_{n,p}(\sigma^2, u, r, s)$, all change-point have sparsities at most s and length at least r. Hence, $\bar{\mathcal{P}}_{n,p}(\sigma^2, u, r, s)$ becomes larger when s increases when r increases. Said differently, the worst-case difficulty of the problem increases with s and decreases with r. We are now able to state the following theorem which provides a lower bound for any estimator and which adapts to the values of s and r.

Theorem 2. *Let u be smaller than $1/8$. For any $\sigma > 0$ and any positive integers n, p, any length $1 \le r \le n/4$, and any sparsity $1 \le s \le p$, we have*

$$\inf_{\hat{\tau}} \sup_{\Theta \in \bar{\mathcal{P}}(r,s)} \mathbb{P}_\Theta(\hat{K} \ne K) \ge \frac{1}{4} \ ,$$

where for commodity we write $\bar{\mathcal{P}}(r, s) = \bar{\mathcal{P}}_{n,p}(\sigma^2, u, r, s)$ and the infimum is taken over all estimators $\hat{\tau}$ of the change-point vector τ.

Thus, for every distribution satisfying an assumption similar to (8) but with a smaller multiplicative constant factor, it holds that no change-point estimator can consistenly estimate the true number of change-points. To match the analysis of our change-point estimator, we can restate the above theorem in terms of spurious change-points or detection of change-points thereby echoing Corollary 3.

Corollary 6. *For any $\sigma > 0$ and any positive integers n, p, any length $1 \le r \le n/4$, any sparsity $1 \le s \le p$, and any estimator $\hat{\tau}$, there exists at least one problem characterised by $\Theta \in \bar{\mathcal{P}}_{n,p}(\sigma^2, u, r, s)$ such that the following holds. With $\mathbb{P}_\Theta$-probability larger than $1/4$, we have either that:*

- *at least a one* **spurious** *change-point is contained in $\hat{\tau}$,* **or**
- *at least a change-point τ_k with $1 \le k \le K$ is not* **detected**, *i.e. there is no change-point estimated in the interval $\left[\tau_k - \frac{\tau_k - \tau_{k-1}}{2}, \tau_k + \frac{\tau_{k+1} - \tau_k}{2}\right]$.*

This corollary is to be compared to Corollary 3 - indeed, the energy condition in Equation (22) differs from the one of Equation (8) only by a numerical multiplicative constant that does not depend on any parameter of the problem.

6 Discussion

6.1 Noise distribution for multivariate change-point detection

Comparison between Gaussian and sub-Gaussian rates. In this work, we have studied two types of noise distribution: Gaussian (Section 3) and general sub-Gaussian distributions (Section 4) without further knowledge on the distribution functions. Since the Gaussian setting is a specific instance of the sub-Gaussian setting, it is clear that the minimax lower bounds from Section 5 apply in both settings. As described in the previous subsection, the performances in the sub-Gaussian case almost match those in the Gaussian setting except for s_k slightly lower but close to $\sqrt{p \log(en/r_k)}$. Indeed, in that regime, Berk-Jones or Higher-Criticism type statistics heavily rely on the probability distribution function of the noise, which is not available in the general sub-Gaussian case. Still, we could slightly improve the sub-Gaussian rates if we further assume that the noise components are identically distributed with common CDF F.

- If F is known (know noise distribution), then one may adapt Berk-Jones test by replacing $\bar{\Phi}(x)$ in Equation (14) by $F(-x) + (1 - F(x))$. This would allow us to recover the exact same detection condition as in the Gaussian setting.
- If F is unknown and if there are not too many change-points, one could hope to estimate the quantiles of the CUSUM statistic at each scale r and plug them into a Berk-Jones statistics. This goes however beyond the scope of this paper.

Unknown variance or more general variance matrix We assumed throughout the manuscript that the variance σ^2 is known. Whereas the partial norm test only requires the knowledge of an upper bound on σ, the dense statistic $\Psi^{(d)}_{l,r}$ requires the exact knowledge of the variance. As soon as there are not too many change-points, it is possible to roughly estimate σ and therefore accommodate the partial norm test with an unknown variance. In contrast, the dense statistics needs to be replaced by a U-statistics. Consider any even positive integer r and define

$$\widetilde{\mathbf{C}}_{l,r}(Y) = \frac{\sqrt{r}}{2}\left(\frac{2}{r}\sum_{t=1}^{r/2} Y_{l-2(t-1)-1} - \frac{2}{r}\sum_{t=1}^{r/2} Y_{l+2(t-1)}\right), \quad \widetilde{\mathbf{C}}'_{l,r}(Y) = \frac{\sqrt{r}}{2}\left(\frac{2}{r}\sum_{t=1}^{r/2} Y_{l-2t} - \frac{2}{r}\sum_{t=1}^{r/2} Y_{l+2(t-1)+1}\right),$$

where $\widetilde{\mathbf{C}}_{l,r}(Y)$ and $\widetilde{\mathbf{C}}'_{l,r}(Y)$ are independent. If there is one change-point at position l and no other change-points in $(l-r, l+r)$, then these statistics are identically distributed and we consider $\widetilde{\Psi}^{''(\mathrm{d})}_{l,r} = \langle \widetilde{\mathbf{C}}_{l,r}(Y), \widetilde{\mathbf{C}}'_{l,r}(Y)\rangle$ whose expectation is null when there are no change-points in the segment. As a consequence, $\widetilde{\Psi}^{''(\mathrm{d})}_{l,r}$ does not require the knowledge of σ; only an upper bound of σ is required to calibrate the corresponding test. Such a U-statistics has already been introduced in [WVS19] and analyzed in an asymptotic setting. Unfortunately, since we can only consider even r, this precludes us to detecting change-points that are very close together with $r_k = 1$.

In the general case, where there is spatial covariance in the noise, that is $\mathrm{var}(\epsilon_i) = \Sigma$ for an unknown but general Σ, we can still use the same U-statistic described in the previous paragraph for the dense case. For the sparse case, one could use the supremum norm of the CUSUM statistics as in Jirak [Jir15] and Yu and Chen [YC17]. To calibrate those tests, we need to estimate both the Frobenius and the operator norm of Σ, which is doable as soon as there are not too many change-points.

6.2 Optimal Localization rates

In this work, we mainly considered the problem of **Detecting** change-points in the mean of a multi-dimensional mean. We provided tight conditions on the energy so that a change-point is detectable. When such a change-point τ_k is detected, Corollary 2 states that its position is estimated up to an error of $\overline{r}_k - 1$.

We did not address the optimality of this localization error. Corollary 6 implies that, when the energy is of the order of the detection threshold, one cannot hope to localize τ_k up to an error much smaller than $r_k \asymp \overline{r}_k$. However, when the energy is way above the threshold, it is not clear whether the error $\overline{r}_k - 1$ is optimal. From the definition of $\overline{r}_k$, one deduces that $\overline{r}_k$ is of the order of $1 + \sigma^2\Delta_k^{-2}\psi^{(g)}_{n,\overline{r}_k,s_k}$.

In the univariate setting ($p = 1$), [VFLRB20] has established that, above the detection threshold, a specific change-point position τ_k can be localized at the rate $\sigma^2\Delta_k^{-2}$. In the multivariate setting, the situation is certainly more tricky and there are certainly several localization regimes beyond the detection threshold. It is an interesting direction of research to pinpoint the exact localization rate between $\sigma^2\Delta_k^{-2}$ and $\sigma^2\Psi^{(g)}_{n,\overline{r}_k,s_k}\Delta_k^{-2}$. We leave this for future work.

6.3 On the choice of the grid

Let us wrap up the comparisons and open questions regarding the choice of the grids. As illustrated in Section 3, a simple dyadic Grid $\mathcal{G}_D$, or more generally an a-adic Grid is sufficient to achieve the optimal detection rates in Corollary 2. Such grids lead to fast procedures since the test aggregation step is linear in n and the control of the FWER of the testing procedures is easily done through an union bound.

Regarding the complete grid $\mathcal{G}_F$, the computational cost is quadratic in worst case and the control of the FWER in Section 4 is more involved as one has to take into account the dependences between the statistics. Still, one could hope that this grid could lead to a slightly more powerful procedure as argued in the beginning of Section 4.

6.4 On the generic algorithms

While Algorithm 1 is a reasonable proposal of a generic algorithm for aggregating multi-scale test statistics, variations of this algorithm could be used. For instance, Algorithm 2 defined in Appendix A also satisfies Theorem 1 and would also lead to the same detection bounds in both the Gaussian and sub-Gaussian settings as obtained with Algorithm 1.

These two algorithms are very related. Still Algorithm 1 is more conservative than Algorithm 2 since it merges all detection intervals at a given resolution while Algorithm 2 only keeps one interval at a given resolution when multiple intervals intersect - the one with smallest index t. While the minimax properties of both methods are comparable - at least up to a multiple constant - the choice of aggregation method will have an influence in practice on the outcome: Algorithm 1 will be slightly more stable, detect less change-points, and provide wider confidence interval around them, while Algorithm 2 will be slightly more sensitive to smaller changes, i.e. detect smaller change-points, will be more precise, and somewhat less stable.

Theorem 1 ensures that, if $T_{\overline{\tau}_k,\overline{r}_k} = 1$ with $(\overline{\tau}_k, \overline{r}_k)$ satisfying Assumption (5), then the change-point τ_k is detected. Inspecting the proof of Theorem 1, one easily checks that Assumption (5) is minimal for Algorithm 1 (and also for Algorithm 2). Still, one may wonder whether any generic algorithm has to require that $4(\overline{r}_k - 1) < r_k$ to detect the change-points or if there exists a generic algorithm where the constant 4 in the above condition can be improved.

6.5 Optimality of the generic algorithm in a broader context

Algorithm 1 aggregates homogeneity tests and provides theoretical guarantees on the event $\mathcal{A}(T, \mathcal{K}^*, (\bar{\tau}_k, \bar{r}_k)_{k\in\mathcal{K}^*})$ - i.e. the event where the outcomes of the tests are consistent - as stated in Theorem 1. In the possibly sparse high-dimensional mean change-point model, we introduced a suitable multiple testing procedure which, when combined with Algorithm 1, leads to a minimax optimal change-point detection procedure.

We described in Section 2 how to adapt this approach to other change-point problems such as kernel change-point or covariance change-point problems with homogeneity tests. One may then wonder whether this roadmap still leads to minimax optimal procedures for such problems. Consider the general setting from Section 1 where we are interested in detecting change-points in $(\Gamma(\mathbb{P}_t))_{t\in\{1,\dots,n\}}$. Upon endowing the space $\mathcal{V}$ with some distance d, we define, for any k,

$$\bar{\Delta}_k = d\left(\Gamma\left(\mathbb{P}_{\tau_k}\right), \Gamma\left(\mathbb{P}_{\tau_{k-1}}\right)\right) \quad ,$$

which corresponds to the change-point height. Then, one may wonder how large $\bar{\Delta}_k$ has to be - as a function of r_k - so that a change-point detection procedure achieving the no-spurious property (**NoSp**) with high probability is able to detect τ_k with high enough probability. In this discussion, we restrict our attention to independent observations, that is the random variables y_t are assumed to be independent and we consider the dyadic grid $\mathcal{G}_D$.

Fix $\alpha \in (0,1)$. At each scale $r \in \{1, 2, \dots, 2^{\lfloor \log_2(n) \rfloor - 1}\}$ and for each $l \in \mathcal{D}_r$, with $\mathcal{D}_r$ defined in (4), we consider the testing problem $H_{0,l,r} : \{\mathbb{P} :\ \Gamma(\mathbb{P}_{l-r}) = \dots = \Gamma(\mathbb{P}_{l+r-1})\}$ versus

$$H_{\rho,l,r} : \left\{ \mathbb{P} : \begin{array}{c} \Gamma(\mathbb{P}_{l-r}) = \dots = \Gamma(\mathbb{P}_{l-m-1}) \\ \Gamma(\mathbb{P}_{l-m}) = \dots = \Gamma(\mathbb{P}_{l+r-1}) \\ d(\Gamma(\mathbb{P}_{l-m-1}), \Gamma(\mathbb{P}_{l-m}) \geq \rho) \end{array} \text{ for some integer } m \in [-r/2, r/2] \right\}$$

This amounts to testing whether there is a single change-point near l of height at least ρ in the segment $(l-r, l+r)$. Given $\delta \in (0,1)$ and a test T we define the δ-separation distance of T by

$$\rho^*_{l,r}(T, \delta) = \inf\left\{ \rho :\ \sup_{\mathbb{P}\in H_{0,l,r}} \mathbb{P}(T = 1) \vee \sup_{\mathbb{P}\in H_{\rho,l,r}} \mathbb{P}(T = 0) \leq \delta \right\} .$$

This corresponds to the minimal change-point height that is detected by the test T. Then, the *minimax* separation distance $\rho^*_{l,r}(\delta)$ is simply $\inf_T \rho_{l,r}(T, \delta)$, i.e. the infimum over all tests T of the separation

distance. By translation invariance of the testing problem, note that $\rho^*_{l,r}(\delta)$ does not depend on l and is henceforth denoted $\rho^*_r(\delta)$.

For any (l,r), take any test $T_{l,r}$ (nearly)[3] achieving the minimax separation distance $\rho^*_r(\delta|\mathcal{D}_r|^{-1}\beta_r)$ with $\beta_r = 6\log_2^{-2}(n/r))\pi^{-2}$. Then, it follows from a simple union bound on the dyadic grid that, with probability higher than $1-\delta$, the collection of tests $T_{l,r}$, where (l,r) belongs to the dyadic grid, does not detect any false positive positive and detects any change-point τ_k such that $\overline{\Delta}_k$ is higher than $\rho^*_{\tilde{r}_k}(\delta|\mathcal{D}_{\tilde{r}_k}|^{-1}\beta_{\tilde{r}_k})$, where $\tilde{r}_k$ is the largest scale in $\mathcal{R}$ such that $4(\tilde{r}_k - 1) \le r_k$. As a consequence of Theorem 1, the corresponding detection procedure achieves, with probability higher than $1-\delta$, the properties (**NoSp**) and (**Detects**) any change-point satisfying the energy condition $\overline{\Delta}_k \ge \rho^*_{\tilde{r}_k}(r\delta\beta_r/2n)$.

Conversely, we believe that this energy condition is almost tight. Indeed, fix any even range $r \ge 2$. To simplify the discussion suppose that $n/(2r)$ is an integer. We consider a specific instance of the problem where the statistician knows that there are $n/(2r)-1$ evenly-spaced change-points respectively at $2r+1, 4r+1, \dots, n-2r+1$ that allow to reduce the change-point detection problem to $n/(2r)$ change-point detection problem in intervals $(l-r, l+r]$ for $l = r+1, 3r+1, 5r+1, \dots$. She further knows that, in each such segment, there exists at most one change-point that is situated in $[l-0.5r, l+0.5r]$, and if the change-point is present then its height is at least $\rho = \rho^*_r(\delta) - \zeta$ for ζ arbitrarily small. Since all $n/(2r)-1$ evenly-spaced change-points $2r+1, 4r+1, \dots, n-2r+1$ are known to the statistician, detecting all remaining change-points is equivalent to building an $n/(2r)$ multiple test of the hypotheses $H_{0,l,r}$ versus $H_{\rho,l,r}$ for $l = r+1, 3r+1, 5r+1, \dots$. If a change-point procedure achieves (**NoSp**) and **Detects** all change-points with radius at least $r/2$ and height at least ρ with probability at least $1-\delta$, then one is able, with probability uniformly higher than $1-\delta$, to simultaneously perform without error $n/(2r)$ independent tests $H_{0,l,r}$ versus $H_{\rho,l,r}$. Since any single test must endure an error with probability at least δ in the worst case, no collection of independents tests is able to endure less than $1-(1-\delta)^{n/(2r)}$. When n/r is large and $\delta < 2r/n$, the latter is of the order of $\delta 2r/n$. Based on this, we conjecture that no change-point procedure is able to achieve, with probability higher than $1-\delta$ the property (**NoSp**), and also to (**Detect**) all change-points with radius at least $r/2$ and height at least $\rho^*_r(2r\delta/n) - \zeta$ for $\zeta > 0$ arbitrarily small.

Comparing the performances of our procedure with the negative arguments that we just outlined, we see that aggregating optimal tests on a dyadic grid allows to detect change-points with (almost) uniform height higher $\rho^*_{\tilde{r}_k}(r_k\delta\beta_{r_k}/(2n))$ whereas, as explained above, we conjecture that a change-point τ_k can be detected only if $\bar{\Delta}_k \ge \rho^*_{r_k}(2r_k\delta/n)$. Since $\tilde{r}_k \ge (r_k/8) \vee 1$- since we considered the dyadic grid when constructing $\tilde{r}_k$ - the difference between these two bounds is mostly due to the term β_r which is of the order of $\log^2(n/r)$. Whereas it is possible to detect change-points at a given scale with a test of type I error probability $2r\delta/n$, our multi-scale procedure relies on a collection of single tests with type I error probability of the order of $r\delta/n/\log^2(n/r)$. This mild mismatch - that we introduce to deal with the multiplicity of scales - of order $\log^2(n/r)$ is harmless for the Gaussian mean-detection problem. Indeed, one may deduce from our analysis in Section 3 that $\rho^*_{r_k}(2r_k\delta/n)$ is of the same order as $\rho^*_{\tilde{r}_k}(\delta|\mathcal{D}_{\tilde{r}_k}|^{-1}\beta_{\tilde{r}_k})$.

In conclusion, one can build through Algorithm 1 an almost optimal change-point procedure in any model provided that we are given optimal homogeneity tests of the form $H_{0,l,r}$ versus $H_{\rho,l,r}$. This provides a universal reduction of the problem of change-point detection to the problem of homogeneity testing.

Acknowledgements. The work of A. Carpentier is partially supported by the Deutsche Forschungsgemeinschaft (DFG) Emmy Noether grant MuSyAD (CA 1488/1-1), by the DFG - 314838170, GRK 2297 MathCoRe, by the FG DFG, by the DFG CRC 1294 'Data Assimilation', Project A03, and by the UFA-DFH through the French-German Doktorandenkolleg CDFA 01-18 and by the UFA-DFH through the French-German Doktorandenkolleg CDFA 01-18 and by the SFI Sachsen-Anhalt for the project RE-BCI. A part of the work of E. Pilliat was supported by ENS Lyon.

[3]Since the minimax separation distance is defined as an infimum, it is not necessarily achieved by a test. Still, we can build a test whose separation distance is arbitrarily close to the optimal one. We neglect the additive error term for the purpose of the discussion.

A An alternative Algorithm

In Algorithm 2 below, we also introduce a variant of the procedure, where instead of merging relevant interesting intervals at the same scale, we only keep one of them. More precisely, we choose the convention of discarding the interval $[l-r+1, l+r-1]$ if there exists $l' < l$ such that $T_{l',r} = 1$ and $[l-r+1, l+r-1] \cap [l'-r+1, l'+r-1] \neq \varnothing$. Alternatively, we could have chosen to discard one of the intervals at random.

Data: $y_t, t = 1 \dots n$ and local test statistic $(T_{l,r})_{(l,r)\in\mathcal{G}}$
Result: $(\hat{\tau}_k)_{k\leq \hat{K}}$
$\mathcal{S} = \varnothing$ $\mathcal{T} = \varnothing$;
for $r \in \mathcal{R}$ **do**
  for $l \in \mathcal{D}_r$ *s.t.* $T_{l,r} = 1$ **do**
    if $[l-r+1, l+r-1] \cap \mathcal{S} = \varnothing$ **then**
      $\mathcal{S} \leftarrow \mathcal{S} \cup [l-r+1, l+r-1]$;
      $\mathcal{T} \leftarrow \mathcal{T} \cup \{l\}$;
    end
  end
end
return $\mathcal{T}$

Algorithm 2: Variant of Aggregation of multiscale tests

B Proofs

B.1 Proof of Theorem 1

Let $\Theta \in \mathbb{R}^{n\times p}$, T be a local test statistic, $\mathcal{K}^*$ be a set of indices of significant change-points and $(\bar{\tau}_k, \bar{r}_k)_{k\in\mathcal{K}^*}$ be elements of the grid $\mathcal{G}$ that satisfy (5). We assume that $\mathcal{A}(\Theta, T, \mathcal{K}^*, (\bar{\tau}_k, \bar{r}_k)_{k\in\mathcal{K}^*})$ holds, that is:

1. **(No False Positive)** $T_{l,r} = 0$ for all $(l,r) \in \mathcal{H}_0 \cap \mathcal{G}$, where $\mathcal{H}_0$ is defined by (7)
2. **(Significant change-point Detection)** for every $k \in \mathcal{K}^*$, we have $T_{\bar{\tau}_k, \bar{r}_k} = 1$.

for every $r \in \mathcal{R}$ define

$$\begin{aligned} \mathcal{T}_r^* &= \{l \in \mathcal{T}_r \ : \ \exists k \in \mathcal{K}^* \text{ s.t. } \tau_k \in [l-r+1, l+r-1]\}, \\ \mathcal{S}_r^* &= \bigcup_{l\in\mathcal{T}_r^*} [l-r+1, l+r-1]. \end{aligned}$$

In other words, for all $r \in \mathcal{R}$, $\mathcal{T}_r^*$ is the subset of $\mathcal{T}_r$ for which each interval of detection $[l-r+1, l+r-1]$ contains a significant change-point. The next proposition recursively analyzes the detection sets corresponding to significant change-points $(\mathcal{S}_r^*)_{r\geq 1}$. The first inclusion means that significant change-points which can be detected with a local statistic with radius smaller than r are detected before step r, while the second inclusion means that each connected component of $\bigcup_{r\in\mathcal{R}} \mathcal{S}_r^*$ is included in a close neighborhoods of some significant change-point τ_k, $k \in \mathcal{K}^*$.

Proposition 5. *For all $r \in \mathcal{R} \cup \{0\}$, we have the double inclusion*

$$\{\tau_k \ : \ k \in \mathcal{K}^* \text{ and } \bar{r}_k \leq r\} \subset \bigcup_{r'\leq r, r'\in\mathcal{R}} \mathcal{S}_{r'}^* \subset \bigcup_{k\in\mathcal{K}^*} [\tau_k - 2(\bar{r}_k - 1), \tau_k + 2(\bar{r}_k - 1)] \ . \tag{23}$$

the next proposition shows that for all step $r \in \mathcal{R}$, the subset of detection corresponding to non significant change-point is disjoint from $\bigcup_{r'\in\mathcal{R}} \mathcal{S}_{r'}^*$.

Proposition 6. *For all $r \in \mathcal{R}$, we have*

$$\bigcup_{l\in\mathcal{T}_r\setminus\mathcal{T}_r^*} [l-r+1, l+r-1] \cap \left(\bigcup_{r'\in\mathcal{R}} \mathcal{S}_{r'}^* \right) = \varnothing \ .$$

Recall that $(\mathrm{C}_k)_{k=1,\dots,\hat{K}}$ are defined as the connected component of $\bigcup_{r\in\mathcal{R}}\mathcal{S}_r$. To ease the notation, re-index (C_k) so that τ_k is the closest true change-point to $\hat{\tau}_k = \frac{\min \mathrm{C}_k + \max \mathrm{C}_k}{2}$. Since there is no false positive, $\tau_k \in \mathrm{C}_k$.
By Proposition 6, the two closed subset $\bigcup_{r\in\mathcal{R}}\bigcup_{l\in\mathcal{T}_r\setminus\mathcal{T}_r^*}[l-r+1, l+r-1]$ and $\bigcup_{r\in\mathcal{R}}\mathcal{S}_r^*$ are disjoints. For all $k\in\mathcal{K}^*$, it holds by Proposition 5 that $\tau_k \in \bigcup_{r\in\mathcal{R}}\mathcal{S}_r^*$, so that C_k is a connected components of $\bigcup_{r\in\mathcal{R}}\mathcal{S}_r^*$ containing the significant change-point τ_k. In particular, $\hat{K}\geq|\mathcal{K}^*|$. We have

- By Proposition 5, $\mathrm{C}_k \subset [\tau_k - 2(\bar{r}_k-1), \tau_k + 2(\bar{r}_k-1)]$ For every $k\in\mathcal{K}^*$. Thus
$$|\hat{\tau}_k - \tau_k| \leq (\bar{r}_k - 1) < \frac{r_k}{4}.$$
- For all $k\in[K]\setminus\mathcal{K}^*$, either τ_k does not belong to $\bigcup_{r\in\mathcal{R}}\mathcal{S}_r$ and it is simply not detected, or it is the closest true change-point to $\hat{\tau}_k = \frac{\min \mathrm{C}_k + \max \mathrm{C}_k}{2}$ so that
$$\hat{\tau}_k \in \left[\tau_k - \frac{\tau_k+\tau_{k-1}}{2}, \tau_k + \frac{\tau_k+\tau_{k+1}}{2}\right] .$$
In particular,
$$\{\hat{\tau}_{k'}, k'\leq\hat{K}\} \subset \left[\tau_1 - \frac{\tau_1-\tau_0}{2}, \tau_K + \frac{\tau_{K+1}-\tau_K}{2}\right] .$$
- Finally, if there exists two estimated change-points $\hat{\tau}_{k_1}, \hat{\tau}_{k_2}$ in $\left[\tau_k - \frac{\tau_k+\tau_{k-1}}{2}, \tau_k + \frac{\tau_k+\tau_{k+1}}{2}\right]$, then either C_{k_1} or C_{k_2} does not contain τ_k. Then Θ is constant on C_{k_1} or on C_{k_2} and we obtain a contradiction since there is no false positive.

This concludes the proof of Theorem 1.

Proof of Proposition 5. To prove the proposition, we do an induction on $r\in\mathcal{R}\cup\{0\}$. The case $r=0$ is trivial since by definition, $\mathcal{S}_0=\varnothing$. Let $r\in\mathcal{R}$ and assume that the double inclusion Proposition 5 holds for all $r'<r, r'\in\mathcal{R}\cup\{0\}$.

First inclusion: Let $k\in\mathcal{K}^*$ be such that $\bar{r}_k = r$ and assume that the corresponding significant change-point τ_k has not been detected before step r, that is $\tau_k \notin \bigcup_{r'<r}\mathcal{S}_{r'}^*$. Since $k\in\mathcal{K}^*$, this implies in particular that $\tau_k\notin\bigcup_{r'<r}\mathcal{S}_{r'}$. Let us show that $\tau_k\in\mathcal{S}_r$. To this end we prove that

$$[\bar{\tau}_k - r + 1, \bar{\tau}_k + r - 1] \cap \bigcup_{r'<r, r'\in\mathcal{R}} \mathcal{S}_{r'} = \varnothing \tag{24}$$

and

$$T_{\bar{\tau}_k, r} = 1, \tag{25}$$

which will be enough since $|\bar{\tau}_k - \tau_k| \leq \bar{r}_k - 1 = r-1$.

- **Proof of (24):** Assume for the sake of contradiction that there exists an integer z which belongs to $[\bar{\tau}_k - r + 1, \bar{\tau}_k + r - 1] \cap \bigcup_{\substack{r'<r\\ r'\in\mathcal{R}}} \mathcal{S}_{r'}$. There exists $r'<r$ such that $z\in\mathcal{S}_{r'}$ and $l(z)\in\mathcal{T}_{r'}$ such that $z\in[l(z)-r'+1, l(z)+r'-1]$. Since $\tau_k\notin\bigcup_{r'<r}\mathcal{S}_{r'}$, we have $\tau_k\notin[l(z)-r'+1, l(z)+r'-1]$. Moreover,
$$\begin{aligned}|l(z)-\tau_k| &\leq |l(z)-z| + |z-\bar{\tau}_k| + |\bar{\tau}_k - \tau_k| \\ &\leq (r'-1) + (r-1) + |\bar{\tau}_k - \tau_k| \\ &< r_k - r' \ ,\end{aligned}$$
Where the last inequality comes from the hypothesis $3(\bar{r}_k-1) + |\bar{\tau}_k - \tau_k| \leq r_k$ Consequently,
$$[l(z)-r', l(z)+r'] \subset [\tau_k - r_k, \tau_k + r_k)\setminus\{\tau_k\} \ ,$$
so that θ is constant on $[l(z)-r', l(z)+r')\cap\mathbb{N}$. Thus, $(l(z), r')\in\mathcal{H}_0$ and $l(z)\notin\mathcal{T}_{r'}$ since there is no false positive. This gives a contradiction and concludes the proof of (24).

- **Proof of (25):** This is simply a consequence of the fact that significant change-point are detected on the grid (See Item 2 in the definition of $\mathcal{A}$).

We have just shown that $\tau_k \in \mathcal{S}_r$ and hence $\tau_k \in \mathcal{S}_r^*$ so that the first inclusion holds at step r.

Second inclusion : Let x be an element of $\mathcal{S}_r^*$. There exists $l(x) \in \mathcal{T}_r^*$ such that $x \in [l(x) - r + 1, l(x) + r - 1]$. By definition of $\mathcal{T}_r^*$, there exists a significant change-point τ_k (i.e. such that $k \in \mathcal{K}^*$) belonging to $[l(x) - r + 1, l(x) + r - 1]$.

We necessarily have $\bar{r}_k \geq r$. Indeed, if $\bar{r}_k < r$, then by the induction hypothesis, $\tau_k \in \mathcal{S}_{r'}^*$ for some $r' < r$, which contradicts the fact that $\mathcal{S}_{r'}^*$ is disjoint from $[l(x) - r + 1, l(x) + r - 1] \subset \mathcal{S}_r^*$. Consequently,

$$\begin{aligned} |l(x) - \tau_k| + r - 1 &\leq 2r - 2 \\ &\leq 2(\bar{r}_k - 1) \end{aligned}$$

Thus

$$x \in [l(x) - r + 1, l(x) + r - 1] \subset [\tau_k - 2(\bar{r}_k - 1), \tau_k + 2(\bar{r}_k - 1)] \ .$$

We have just shown that $\mathcal{S}_r^* \subset \bigcup_{k \in \mathcal{K}^*} [\tau_k - 2(\bar{r}_k - 1), \tau_k + 2(\bar{r}_k - 1)]$.
Therefore, the proposition is verified at step r and the induction is proved. □

Proof of Proposition 6. Let $k \in \mathcal{K}^*$ and C_k be the detected connected component containing the significant change-point τ_k

$$\mathrm{C}_k = \bigcup_{r' \in \mathcal{R}} \mathcal{S}_{r'}^* \cap [\tau_k - 2(\bar{r}_k - 1), \tau_k + 2(\bar{r}_k - 1)] \ .$$

We know from Proposition 5 that C_k is a connected component of $\bigcup_{r' \in \mathcal{R}} \mathcal{S}_{r'}^*$ and we want to prove now that C_k does not overlap with $\bigcup_{l \in \mathcal{T}_r \setminus \mathcal{T}_r^*} [l - r + 1, l + r - 1]$ for some $r \in \mathcal{R}$. Let r_0 be such that C_k is the connected component of $\mathcal{S}_{r_0}$,

$$C_k \subset \mathcal{S}_{r_0}^* \ .$$

Such an r_0 exists and is unique since the sets $(\mathcal{S}_{r'}^*)$ are disjoints. We have from Proposition 5 that $\tau_k \in \bigcup_{r' \in \mathcal{R}, r' \leq \bar{r}_k} \mathcal{S}_{r'}^*$ so that

$$r_0 \leq \bar{r}_k \ .$$

Let $r \in \mathcal{R}$ and $l \in \mathcal{T}_r \setminus \mathcal{T}_r^*$ and assume without loss of generality that $l + r - 1 < \tau_k$. Since there is no false positive, $(l, r) \notin \mathcal{H}_0$ and there exists at least one true change-point in the interval of detection $[l - r + 1, l + r - 1]$. Denote $\tau_a, \ldots, \tau_b$ with $a \leq b$ the true change-points belonging to $[l - r + 1, l + r - 1]$. By definition of $\mathcal{T}_r \setminus \mathcal{T}_r^*$, $\tau_a, \ldots, \tau_b$ are not significant change-points, i.e. $a, a + 1, \ldots, b \notin \mathcal{K}^*$. We consider the two cases $r > \bar{r}_k$ and $r \leq \bar{r}_k$

- $\underline{r > \bar{r}_k}$: In that case, since the sets $(\mathcal{S}_{r'})$ are disjoints and $\mathrm{C}_k \subset \mathcal{S}_{r_0}^*$, we have $\mathrm{C}_k \cap [l - r + 1, l + r - 1] = \varnothing$.

- $\underline{r \leq \bar{r}_k}$: In that case, we have

  $$l + r - 1 \leq \tau_b + 2(r - 1) \leq \tau_b + 2(\bar{r}_k - 1) < \tau_k - 2(\bar{r}_k - 1) \ ,$$

  where we used the fact that $4(\bar{r}_k - 1) < r_k \leq \tau_k - \tau_b$. Since by Proposition 5 we have $\mathrm{C}_k \subset [l - r + 1, l + r - 1]$, we also have in that case $\mathrm{C}_k \cap [l - r + 1, l + r - 1] = \varnothing$.

This concludes the proof of the proposition. □

B.2 Proofs in the Gaussian setting

From now on, we use the following notation for all $(l, r) \in J_n$.

- For any $(v_1, \dots, v_n)$ with $v_t \in \mathbb{R}^p$, the left mean and right mean of v on $[l-r, l+r)$ are denoted by
$$\bar{v}_{l,+r} = \frac{1}{r}\sum_{t=l}^{l+r-1} v_t \quad \bar{v}_{l,-r} = \frac{1}{r}\sum_{t=l-r}^{l-1} v_t \ .$$
- The population term of the CUSUM statistic $\mathbf{C}_{l,r}$ is written
$$U_{l,r} = \sqrt{\frac{r}{2}}\left(\bar{\theta}_{l,+r} - \bar{\theta}_{l,-r}\right) \ .$$
- With these notation, we write $v_{l,+r,i}, v_{l,-r,i}, U_{l,r,i}$ for the i^{th} coordinate of the vector $v_{l,+r}, v_{l,-r}, U_{l,r}$.
- We define, for $1 \le s \le p$, the order statistics $U_{l,r,(s)}$ by $|U_{l,r,(1)}| \ge |U_{l,r,(2)}| \ge \dots |U_{l,r,(p)}|$.

B.2.1 Proof of Proposition 1

Step 0: Consequence of Equation (10) on the grid. Let $k \in \{1, \dots, K\}$ and assume that τ_k is a κ_{d}-dense high-energy change-point (see Equation (10)). We have that

$$\begin{aligned}\left\|U_{\bar{\tau}_k^{(d)},\bar{r}_k^{(\mathrm{d})}}\right\|^2 &\ge \frac{9}{16}\left\|U_{\tau_k,\bar{r}_k^{(\mathrm{d})}}\right\|^2 \\ &\ge \frac{9}{16\times 12}\kappa_{\mathrm{d}}\left(\sqrt{p\log\left(\frac{n}{\bar{r}_k^{(\mathrm{d})},\delta}\right)} + \log\left(\frac{n}{\bar{r}_k^{(\mathrm{d})},\delta}\right)\right),\end{aligned} \tag{26}$$

since by definition $\|\tau_k - \bar{\tau}_k^{(d)}\| \le \bar{r}_k^{(\mathrm{d})}/4$, so that $\|\overline{\theta}_{\bar{\tau}_k^{(d)},+\bar{r}_k^{(\mathrm{d})}} - \overline{\theta}_{\bar{\tau}_k^{(d)},-\bar{r}_k^{(\mathrm{d})}}\|^2 \ge \frac{9}{16}\|\overline{\theta}_{\tau_k,+\bar{r}_k^{(\mathrm{d})}} - \overline{\theta}_{\tau_k,-\bar{r}_k^{(\mathrm{d})}}\|^2$.

Step 1: Introduction of useful high probability events. Remark that

$$\frac{r}{2}\left[\left\|\overline{y}_{l,+r} - \overline{y}_{l,-r}\right\|^2 - \left\|\overline{\theta}_{l,-r} - \overline{\theta}_{l,+r}\right\|^2\right] - \sigma^2 p = r\langle\overline{\varepsilon}_{l,+r} - \overline{\varepsilon}_{l,-r}, \overline{\theta}_{l,+r} - \overline{\theta}_{l,-r}\rangle + \frac{r}{2}\left\|\overline{\varepsilon}_{l,+r} - \overline{\varepsilon}_{l,-r}\right\|^2 - \sigma^2 p \ .$$

The first term, written as

$$r\langle\overline{\varepsilon}_{l,+r} - \overline{\varepsilon}_{l,-r}, \overline{\theta}_{l,+r} - \overline{\theta}_{l,-r}\rangle \ ,$$

is a crossed term between the noise and the mean vector θ. Lemma 1 states that near the change-points and on the grid defined by the sets $\mathcal{R}, \mathcal{D}_r$, it is jointly controlled with high probability.

Lemma 1. *Let $1 \ge \delta > 0$. The event*

$$\begin{aligned}\xi_1^{(\mathrm{d})} = \bigcap_{k\in\{1,\dots,K\}} \Bigg\{ &\bar{r}_k^{(\mathrm{d})}\left|\langle\overline{\varepsilon}_{\bar{\tau}_k^{(d)},+\bar{r}_k^{(\mathrm{d})}} - \overline{\varepsilon}_{\bar{\tau}_k^{(d)},-\bar{r}_k^{(\mathrm{d})}}, \overline{\theta}_{\bar{\tau}_k^{(d)},+\bar{r}_k^{(\mathrm{d})}} - \overline{\theta}_{\bar{\tau}_k^{(d)},-\bar{r}_k^{(\mathrm{d})}}\rangle\right| \\ &\le \frac{1}{8}\bar{r}_k^{(\mathrm{d})}\left\|\overline{\theta}_{\bar{\tau}_k^{(d)},+\bar{r}_k^{(\mathrm{d})}} - \overline{\theta}_{\bar{\tau}_k^{(d)},-\bar{r}_k^{(\mathrm{d})}}\right\|^2 + 16\sigma^2\log\left(2\frac{n}{\bar{r}_k^{(\mathrm{d})}\delta}\right)\Bigg\} \ .\end{aligned}$$

holds with probability larger than $1-\delta$.

The second term, written as

$$\frac{r}{2}\left\|\overline{\varepsilon}_{l,+r} - \overline{\varepsilon}_{l,-r}\right\|^2 - \sigma^2 p \ ,$$

is a term of pure noise. Lemma 2 states that it is controlled jointly with high probability on the grid defined by the sets $\mathcal{R}, \mathcal{D}_r$.

Lemma 2. *Let $1 \geq \delta > 0$. The event*

$$\xi_2^{(\mathrm{d})} = \bigcap_{r\in\mathcal{R}} \bigcap_{l\in\mathcal{D}_r} \left\{ \left| \frac{r}{2} \left\| \overline{\varepsilon}_{l,+r} - \overline{\varepsilon}_{l,-r} \right\|^2 - \sigma^2 p \right| \leq 4\sigma^2 \left[\sqrt{p \log\left(2\frac{n}{r\delta}\right)} + \log\left(2\frac{n}{r\delta}\right) \right] \right\} \ ,$$

holds with probability larger than $1-\delta$.

Set now

$$\xi^{(\mathrm{d})} := \xi^{(\mathrm{d})} = \xi_1^{(\mathrm{d})} \cap \xi_2^{(\mathrm{d})} \ .$$

Note that

$$\mathbb{P}(\xi^{(\mathrm{d})}) \geq 1 - 2\delta \ .$$

Step 2: Study in the 'no change-point' situation. Consider $r \in \mathcal{R}, l \in \mathcal{D}_r$ such that $\{\tau_k, k \in \{1,\dots,K\}\} \cap [l-r, l+r) = \varnothing$. Note that since $\{\tau_k, k \in \{1,\dots,K\}\} \cap [l-r, l+r) = \varnothing$, we have $\overline{\theta}_{l,-r} = \overline{\theta}_{l,+r}$ so that

$$\frac{r}{2} \left\| \overline{\theta}_{l,-r} - \overline{\theta}_{l,+r} \right\|^2 = 0 \ ,$$

and

$$r\langle \overline{\varepsilon}_{l,+r} - \overline{\varepsilon}_{l,-r}, \overline{\theta}_{l,+r} - \overline{\theta}_{l,-r} \rangle = 0.$$

Moreover we have on $\xi^{(\mathrm{d})}$ that - see Lemma 2

$$\left| \frac{r}{2} \left\| \overline{\varepsilon}_{l,+r} - \overline{\varepsilon}_{l,-r} \right\|^2 - \sigma^2 p \right| \leq 4\sigma^2 \left[\sqrt{p \log\left(2\frac{n}{r\delta}\right)} + \log\left(2\frac{n}{r\delta}\right) \right] = \sigma^2 x_r^{(\mathrm{d})} \ .$$

And so

$$\Psi_{l,r}^{(\mathrm{d})} \leq x_r^{(\mathrm{d})} \ ,$$

so that

$$T_{l,r}^{(\mathrm{d})} = 0 \ ,$$

on $\xi^{(\mathrm{d})}$. This concludes the proof of the first part of the proposition.

Step 3: Study in the 'change-point' situation. Consider $k \in \{1,\dots,K\}$ τ_k is a κ_{d}-dense high-energy change-point - that is Equation(10) holds. We have from (26) that for κ_{d} large enough,

$$\begin{aligned} \frac{\overline{r}_k^{(\mathrm{d})}}{2} \left\| \overline{\theta}_{\overline{\tau}_k^{(d)}, -\overline{r}_k^{(\mathrm{d})}} - \overline{\theta}_{\overline{\tau}_k^{(d)}, +\overline{r}_k^{(\mathrm{d})}} \right\|^2 &\geq \frac{9}{16\times 12} \kappa_{\mathrm{d}} \sigma^2 \left(\sqrt{p \log\left(\frac{n}{\overline{r}_k^{(\mathrm{d})}, \delta}\right)} + \log\left(\frac{n}{\overline{r}_k^{(\mathrm{d})}, \delta}\right) \right) \\ &> 4\sigma^2 x_{\overline{r}_k^{(\mathrm{d})}}^{(\mathrm{d})} \ . \end{aligned}$$

So on $\xi^{(\mathrm{d})}$ this implies that - see Lemma 1

$$\overline{r}_k^{(\mathrm{d})} \left| \langle \overline{\varepsilon}_{\overline{\tau}_k^{(d)}, +\overline{r}_k^{(\mathrm{d})}} - \overline{\varepsilon}_{\overline{\tau}_k^{(d)}, -\overline{r}_k^{(\mathrm{d})}}, \overline{\theta}_{\overline{\tau}_k^{(d)}, +\overline{r}_k^{(\mathrm{d})}} - \overline{\theta}_{\overline{\tau}_k^{(d)}, -\overline{r}_k^{(\mathrm{d})}} \rangle \right| \leq \frac{\overline{r}_k^{(\mathrm{d})}}{4} \left\| \overline{\theta}_{\overline{\tau}_k^{(d)}, +\overline{r}_k^{(\mathrm{d})}} - \overline{\theta}_{\overline{\tau}_k^{(d)}, -\overline{r}_k^{(\mathrm{d})}} \right\|^2 .$$

Moreover we have on $\xi^{(\mathrm{d})}$ that - see Lemma 2

$$\left| \frac{\overline{r}^{(\mathrm{d})}}{2} \left\| \overline{\varepsilon}_{\overline{\tau}_k^{(d)}, +\overline{r}_k^{(\mathrm{d})}} - \overline{\varepsilon}_{\overline{\tau}_k^{(d)}, -\overline{r}_k^{(\mathrm{d})}} \right\|^2 - \sigma^2 p \right| \leq 4\sigma^2 \left[\sqrt{p \log\left(2\frac{n}{\overline{r}_k^{(\mathrm{d})}} \delta^{-1}\right)} + \log\left(2\frac{n}{\overline{r}_k^{(\mathrm{d})}} \delta^{-1}\right) \right] = \sigma^2 x_{\overline{r}_k^{(\mathrm{d})}}^{(\mathrm{d})} .$$

And so on $\xi^{(\mathrm{d})}$, combining the three previous displayed equations implies

$$\Psi^{(\mathrm{d})}_{\bar{\tau}_k^{(d)},\bar{r}_k^{(\mathrm{d})}} \geq \frac{\frac{\bar{r}_k^{(\mathrm{d})}}{2}\left\|\overline{\theta}_{\bar{\tau}_k^{(d)},+\bar{r}_k^{(\mathrm{d})}} - \overline{\theta}_{\bar{\tau}_k^{(d)},-\bar{r}_k^{(\mathrm{d})}}\right\|^2}{2\sigma^2} - x^{(\mathrm{d})}_{\bar{r}_k^{(\mathrm{d})}} > (2-1)x^{(\mathrm{d})}_{\bar{r}_k^{(\mathrm{d})}} = x^{(\mathrm{d})}_{\bar{r}_k^{(\mathrm{d})}} \ ,$$

so that

$$T^{(\mathrm{d})}_{\bar{\tau}_k^{(d)},\bar{r}_k^{(\mathrm{d})}} = 1 \ .$$

This concludes the proof of the second part of the proposition.

Proof of Lemma 1. Let $k \in \{1,\dots,K\}$. Since the vectors ε_t are i.i.d. and distributed as $\mathcal{N}(0,\sigma^2\mathbf{I}_p)$, it holds that

$$\bar{r}_k^{(\mathrm{d})}\langle \overline{\varepsilon}_{\bar{\tau}_k^{(d)},+\bar{r}_k^{(\mathrm{d})}} - \overline{\varepsilon}_{\bar{\tau}_k^{(d)},-\bar{r}_k^{(\mathrm{d})}}, \overline{\theta}_{\bar{\tau}_k^{(d)},+\bar{r}_k^{(\mathrm{d})}} - \overline{\theta}_{\bar{\tau}_k^{(d)},-\bar{r}_k^{(\mathrm{d})}}\rangle \sim \mathcal{N}\left(0, 2\bar{r}_k^{(\mathrm{d})}\sigma^2\left\|\overline{\theta}_{\bar{\tau}_k^{(d)},+\bar{r}_k^{(\mathrm{d})}} - \overline{\theta}_{\bar{\tau}_k^{(d)},-\bar{r}_k^{(\mathrm{d})}}\right\|^2\right) \ .$$

And so for $\delta_k > 0$, it holds with probability larger than $1-\delta_k$ it holds that

$$\begin{aligned}\bar{r}_k^{(\mathrm{d})}&\left|\langle \overline{\varepsilon}_{\bar{\tau}_k^{(d)},+\bar{r}_k^{(\mathrm{d})}} - \overline{\varepsilon}_{\bar{\tau}_k^{(d)},-\bar{r}_k^{(\mathrm{d})}}, \overline{\theta}_{\bar{\tau}_k^{(d)},+\bar{r}_k^{(\mathrm{d})}} - \overline{\theta}_{\bar{\tau}_k^{(d)},-\bar{r}_k^{(\mathrm{d})}}\rangle\right| \\ &\leq 2\sigma\left\|\overline{\theta}_{\bar{\tau}_k^{(d)},+\bar{r}_k^{(\mathrm{d})}} - \overline{\theta}_{\bar{\tau}_k^{(d)},-\bar{r}_k^{(\mathrm{d})}}\right\|\sqrt{\bar{r}_k^{(\mathrm{d})}\log(2\delta_k^{-1})} \ .\end{aligned}$$

Let us set $\delta_k = \frac{(\bar{r}_k^{(\mathrm{d})})^2\delta}{2n^2}$. Note that

$$\sum_{k\in\{1,\dots,K\}}\delta_k = \sum_{r\in R}\sum_{k\in\{1,\dots,K\}:\bar{r}_k^{(\mathrm{d})}=r}\frac{(\bar{r}_k^{(\mathrm{d})})^2\delta}{2n^2} \leq \sum_{r\in R}\sum_{l\in D_r}\frac{r^2\delta}{2n^2} \leq \sum_{r\in\mathcal{R}}\frac{r\delta}{2n} \leq \delta \ ,$$

since $r_k \geq \bar{r}_k^{(\mathrm{d})}$ and $|\mathcal{D}_r| \leq 2n/r$, and also by definition of $\mathcal{R}$ which implies $\sum_{r\in\mathcal{R}}\frac{r}{n} \leq 1$. And so if $\delta \leq 1$, then with probability larger than $1-\delta$, for any $k \in \{1,\dots,K\}$, we have

$$\begin{aligned}\bar{r}_k^{(\mathrm{d})}&\left|\langle \overline{\varepsilon}_{\bar{\tau}_k^{(d)},+\bar{r}_k^{(\mathrm{d})}} - \overline{\varepsilon}_{\bar{\tau}_k^{(d)},-\bar{r}_k^{(\mathrm{d})}}, \overline{\theta}_{\bar{\tau}_k^{(d)},+\bar{r}_k^{(\mathrm{d})}} - \overline{\theta}_{\bar{\tau}_k^{(d)},-\bar{r}_k^{(\mathrm{d})}}\rangle\right| \leq 2\sigma\left\|\overline{\theta}_{\bar{\tau}_k^{(d)},+\bar{r}_k^{(\mathrm{d})}} - \overline{\theta}_{\bar{\tau}_k^{(d)},-\bar{r}_k^{(\mathrm{d})}}\right\| \\ &\qquad\sqrt{2\bar{r}_k^{(\mathrm{d})}\log\left(2\frac{n}{\bar{r}_k^{(\mathrm{d})}}\delta^{-1}\right)} \ .\end{aligned}$$

This implies in particular that with probability larger than $1-\delta$, for any $k \in \{1,\dots,K\}$, we have

$$\begin{aligned}\bar{r}_k^{(\mathrm{d})}&\left|\langle \overline{\varepsilon}_{\bar{\tau}_k^{(d)},+\bar{r}_k^{(\mathrm{d})}} - \overline{\varepsilon}_{\bar{\tau}_k^{(d)},-\bar{r}_k^{(\mathrm{d})}}, \overline{\theta}_{\bar{\tau}_k^{(d)},+\bar{r}_k^{(\mathrm{d})}} - \overline{\theta}_{\bar{\tau}_k^{(d)},-\bar{r}_k^{(\mathrm{d})}}\rangle\right| \leq \frac{\bar{r}_k^{(\mathrm{d})}}{2}\frac{\left\|\overline{\theta}_{\bar{\tau}_k^{(d)},+\bar{r}_k^{(\mathrm{d})}} - \overline{\theta}_{\bar{\tau}_k^{(d)},-\bar{r}_k^{(\mathrm{d})}}\right\|^2}{4} \\ &\qquad + 16\sigma^2\log\left(2\frac{n}{\bar{r}_k^{(\mathrm{d})}}\delta^{-1}\right) \ .\end{aligned}$$

□

Proof of Lemma 2. Let $r \in \mathcal{R}$ and $l \in \mathcal{D}_r$. Since the vectors ε_t are i.i.d. and distributed as $\mathcal{N}(0,\sigma^2\mathbf{I}_p)$, it holds that

$$\frac{r}{2}\left\|\overline{\varepsilon}_{l,+r} - \overline{\varepsilon}_{l,-r}\right\|^2 \sim \sigma^2\chi^2_p,$$

which implies by properties of the χ^2_p distribution that for any $\delta_r > 0$ we have with probability larger than $1-\delta_r$

$$\left|\frac{r}{2}\left\|\overline{\varepsilon}_{l,+r} - \overline{\varepsilon}_{l,-r}\right\|^2 - \sigma^2 p\right| \leq 2\sigma^2\sqrt{p\log(2/\delta_r)} + 2\sigma^2\log(2/\delta_r) \ .$$

If we set, for $\delta > 0$, $\delta_r = \frac{r^2\delta}{2n^2}$, we have that with probability larger than $1 - \frac{r\delta}{n}$, that $\forall l \in D_r$

$$\left| \frac{r}{2} \left\| \overline{\varepsilon}_{l,+r} - \overline{\varepsilon}_{l,-r} \right\|^2 - \sigma^2 p \right| \leq 2\sigma^2 \sqrt{p \log(2/\delta_r)} + 2\sigma^2 \log(2/\delta_r) \ ,$$

since $|\mathcal{D}_r| \leq 2n/r$. And so with probability larger than $1 - \delta$, for all $r \in \mathcal{R}$ and $l \in \mathcal{D}_r$

$$\left| \frac{r}{2} \left\| \overline{\varepsilon}_{l,+r} - \overline{\varepsilon}_{l,-r} \right\|^2 - \sigma^2 p \right| \leq 2\sigma^2 \sqrt{p \log(2/\delta_r)} + 2\sigma^2 \log(2/\delta_r) \ ,$$

since $\sum_{r \in \mathcal{R}} \frac{r}{n} \leq 1$. And so finally for $\delta \leq 1$ and with probability larger than $1 - \delta$, for all $r \in \mathcal{R}$ and $l \in \mathcal{D}_r$

$$\left| \frac{r}{2} \left\| \overline{\varepsilon}_{l,+r} - \overline{\varepsilon}_{l,-r} \right\|^2 - \sigma^2 p \right| \leq 4\sigma^2 \left[\sqrt{p \log\left(2\frac{n}{r}\delta^{-1}\right)} + \log\left(2\frac{n}{r}\delta^{-1}\right) \right] \ .$$

This concludes the proof. □

B.2.2 Proof of Proposition 2

Step 1 : Analysis of the Berk Jones statistics We first define a threshold $x_{r,s}^{(\mathrm{BJ})}$ for the Berk Jones statistics for all $r, s \geq 1$

$$x_{r,s}^{(\mathrm{BJ})} = \min \left\{ x \geq 2 : \overline{\Phi}(x) \leq \frac{s^2}{28^2 p \log(2\alpha_{x,r}^{-1})} \right\} \ , \tag{27}$$

where we recall that $\alpha_{x,r}$ are the weights defined by (13):

$$\alpha_{x,r} = \frac{6\delta r}{\pi^2 x^2 |\mathcal{D}_r| n} \ .$$

Remark that $(x_{r,s}^{(\mathrm{BJ})})$ is nonincreasing with s and define for all $r \geq 1$

$$\bar{s}_r = \min \left\{ s \in \mathcal{Z} \ : \quad s \geq \frac{28}{3} \log\left(2\alpha_{x_{r,s}^{(\mathrm{BJ})},r}^{-1}\right) \right\} \ . \tag{28}$$

The second point of the following proposition ensures that if there exists $s \in \mathcal{Z}$ such that $U_{l,r,(s)} \geq t_s$ for some $s \geq \bar{s}_r$, for $(l,r) = (\bar{\tau}_k^{(s)}, \bar{r}_k^{(\mathrm{s})})$, then $T_{l,r}^{(\mathrm{BJ})} = 1$ with high probability. We recall that $|U_{l,r,(1)}| \geq \cdots \geq |U_{l,r,(p)}|$ are the sorted absolute values of the coordinate of $U_{l,r}$ and that $\mathcal{H}_0$ is defined by (7).

Proposition 7. *There exists an event $\xi^{(\mathrm{BJ})}$ of probability larger than $1 - 2\delta$ such that the following holds:*

- $T_{l,r}^{(\mathrm{BJ})} = 0$ *for any* $(l,r) \in \mathcal{H}_0 \cap G$.
- *For all $k \in [K]$, if there exists $s \in \mathcal{Z}$ such that $s \geq \bar{s}_{\bar{r}_k^{(\mathrm{s})}}$ and $U_{\bar{\tau}_k^{(s)},\bar{r}_k^{(\mathrm{s})},(s)} > x_{\bar{r}_k^{(\mathrm{s})},s}^{(\mathrm{BJ})}$, then $T_{\bar{\tau}_k^{(s)},\bar{r}_k^{(\mathrm{s})}}^{(\mathrm{BJ})} = 1$.*

Step 2 : Analysis of the max statistics Since it may happen that τ_k is a sparse high-energy change-point but there is no $s \geq \bar{s}_{\bar{r}_k^{(\mathrm{s})}}$ such that $U_{\bar{\tau}_k^{(s)},\bar{r}_k^{(\mathrm{s})},(s)} \geq x_{\bar{r}_k^{(\mathrm{s})},s}^{(\mathrm{BJ})}$, we use the following proposition on the partial norm test statistic $T_{l,r}^{(p)}$:

Proposition 8. *There exists an event $\xi^{(\mathrm{p})}$ of probability larger than $1 - 2\delta$ such that the following holds:*

- $T_{l,r}^{(p)} = 0$ *for any* $(l,r) \in \mathcal{H}_0 \cap G$.
- *for any $k \in \{1, \ldots, K\}$, if there exists $s \in \mathcal{Z}$ such that*

$$\sum_{s'=1}^{s} \left| U_{\bar{\tau}_k^{(s)},\bar{r}_k^{(\mathrm{s})},(s')} \right|^2 > 4x_{\bar{r}_k^{(\mathrm{s})},s}^{(p)} \ , \tag{29}$$

then $T_{\bar{\tau}_k^{(s)},\bar{r}_k^{(\mathrm{s})}}^{(p)} = 1$.

Step 3 : Combination of the two Statistics Let us return to the proof of Proposition 2. To conclude the proof, it suffices to show that if τ_k is a κ_s-sparse high-energy change-point - see (11) - for some large enough constant κ_s, then the result of one of the two preceding propositions holds. This is precisely what the following lemma shows.

Lemma 3. *There exists a constant κ_s such that if τ_k is a κ_s-sparse high-energy change-point, then one of the following propositions is true :*

- *There exists $s \in \mathcal{Z}$ such that $s > \bar{s}_{\bar{r}_k^{(s)}}$ and $\left|U_{\bar{\tau}_k^{(s)},\bar{r}_k^{(s)},(s)}\right| > x^{(\mathrm{BJ})}_{\bar{r}_k^{(s)},s}$.*
- *There exists $s \in \mathcal{Z}$ such that $s \le \bar{s}_{\bar{r}_k^{(s)}}$ and $\sum_{s'=1}^{s}\left|U_{\bar{\tau}_k^{(s)},\bar{r}_k^{(s)},(s')}\right|^2 > 4x^{(p)}_{\bar{r}_k^{(s)},s}$.*

Proof of Proposition 7. The first part of the proposition is a simple consequence of the definition together with an union bound.

$$\begin{aligned}\mathbb{P}\left[\max_{(l,r)\in\mathcal{H}_0} T^{(\mathrm{BJ})}_{l,r} = 1\right] &\le \sum_{r\in\mathcal{R}}\sum_{l\in\mathcal{D}_r}\sum_{x\in\mathbb{N}^*}\alpha^{(BJ)}_{x,r} \\ &\le \sum_{r\in\mathcal{R}}\sum_{l\in\mathcal{D}_r}\frac{\delta r}{|\mathcal{D}_r|n} \le \sum_{r\in\mathcal{R}}\frac{\delta r}{n} \le \delta.\end{aligned}$$

We focus on the second part of the proposition. To ease the reading, we introduce some notation

$$\gamma_{x,r} = \overline{Q}^{-1}[\alpha_{x,r}, p, 2\overline{\Phi}(x)] \; ; \quad \eta_{x,r,s} = \overline{Q}^{-1}[1-\alpha_{x,r}/2, p-s, 2\overline{\Phi}(x)] \; ;$$
$$\psi_{x,r,s}(u) = \overline{Q}^{-1}[1-\alpha_{x,r}/2, s, \overline{\Phi}(x-u)+\overline{\Phi}(x+u)] \; ,$$

for $x \ge 0$. In fact, $\gamma_{x,r}$ is the threshold of the statistics $N_{x,l,r}$. As for $\eta_{x,r,s}$, it stands for the contribution to $N_{x,l,r}$ of the $(p-s)$ coordinates i such that $\theta_{\cdot,i}$ is constant over $[l-r, l+r)$. Finally, $\psi_{x,r,s}(u)$ stands for the contribution to $N_{x,l,r}$ of the s coordinates i whose population CUSUM statistics $U_{l,r,i}$ is equal to u.

Lemma 4. *Consider any $r \in \mathcal{R}$ and $l \in \mathcal{D}_r$. If for some positive integers s and x we have*

$$\psi_{x,r,s}(|U_{l,r,(s)}|) > \gamma_{x,r} - \eta_{x,r,s} \; , \tag{30}$$

then $\mathbb{P}[T^{(\mathrm{BJ})}_{l,r} = 1] \ge 1 - \alpha_{x,r}$.

Denote $\mathcal{H}[\theta]$ the collection of (l,r) with $r \in \mathcal{R}$ and $l \in \mathcal{D}_r$ that satisfy Condition (30) for some s and some x. We easily deduce from the above Lemma together with an union bound that, with probability higher than $1-\delta$, $T^{(\mathrm{BJ})}_{l,r} = 1$ for all $(l,r) \in \mathcal{H}[\theta]$.

Let us now provide a more explicit characterisation of $\mathcal{H}[\theta]$ with the following Lemma.

Lemma 5. *For any $1 \le s \le p$ and $r \in \mathcal{R}$ define x_s by*

$$x_s := x^{(\mathrm{BJ})}_{r,s} = \min\left\{x \ge 2 : \overline{\Phi}(x) \le \frac{s^2}{28^2 p \log(2\alpha^{-1}_{x,r})}\right\} \; . \tag{31}$$

We have $\psi_{x_s,r,s}(t_s) > \gamma_{x_s,r} - \eta_{x_s,r,s}$ provided that

$$s \ge \frac{28}{3}\log(2\alpha^{-1}_{x_s,r}) \; . \tag{32}$$

Combining Lemma 5 and Lemma 4, we conclude the proof of the proposition. □

Proof of Lemma 4. Denote S any subset of size s, such that for any $j \in S$, $|U_{l,r,j}| \geq |U_{l,r,(s)}|$. Define

$$N^{(1)}_{x,l,r} = \sum_{i=1}^{p} \mathbf{1}_{i \notin S} \mathbf{1}_{|\mathbf{C}_{l,r,i}|>x}, \qquad N^{(2)}_{x,l,r} = \sum_{i=1}^{p} \mathbf{1}_{i \in S} \mathbf{1}_{|\mathbf{C}_{l,r,i}|>x}$$

Since, for any $x > 0$, the function $u \mapsto \overline{\Phi}(x+u) + \overline{\Phi}(x-u)$ is non-decreasing. As a consequence, the random variable $N^{(1)}_{x,l,r}$ is stochastically dominated by a Binomial distribution with parameters $(p-s, 2\overline{\Phi}(x))$. Besides, $N^{(2)}_{x,l,r}$ is stochastically dominated by a Binomial distribution with parameters $(s, \overline{\Phi}(x+|U_{l,r,(s)}|) + \overline{\Phi}(x-|U_{l,r,(s)}|))$. We obtain

$$\begin{aligned}
\mathbb{P}[T^{(BJ)}_{l,r} = 0] &\leq \mathbb{P}[N_{x,l,r} \leq \gamma_{x,r}] \leq \mathbb{P}[N^{(1)}_{x,l,r} < \eta_{x,r,s}] + \mathbb{P}[N^{(2)}_{x,l,r} \leq \gamma_{x,r} - \eta_{x,r,s}] \\
&\leq \frac{\alpha_{x,r}}{2} + 1 - \overline{Q}[\gamma_{x,r} - \eta_{x,r,s}, s, \overline{\Phi}(x-|U_{l,r,(s)}|) + \overline{\Phi}(x+|U_{l,r,(s)}|)] \\
&\leq \frac{\alpha_{x,r}}{2} + \frac{\alpha_{x,r}}{2} \leq \alpha_{x,r} \ .
\end{aligned}$$

□

Proof of Lemma 5. From Bernstein inequality, we deduce that, for any positive integers s and x,

$$\begin{aligned}
\gamma_{x,s} &\leq 2p\overline{\Phi}(x) + 2\sqrt{p\overline{\Phi}(x)\log(\alpha^{-1}_{x,r})} + \frac{2}{3}\log(\alpha^{-1}_{x,r}) \ ; \\
\eta_{x,r,s} &\geq 2(p-s)\overline{\Phi}(x) - 2\sqrt{p\overline{\Phi}(x)\log(2\alpha^{-1}_{x,r})} - \frac{2}{3}\log(2\alpha^{-1}_{x,r}) \ .
\end{aligned}$$

Hence, it follows that

$$\gamma_{x,s} - \eta_{x,r,s} \leq 2s\overline{\Phi}(x) + 4\sqrt{p\overline{\Phi}(x)\log(2\alpha^{-1}_{x,r})} + \frac{4}{3}\log(2\alpha^{-1}_{x,r}) \ .$$

For $u = x$, we have $\overline{\Phi}(x-u) + \overline{\Phi}(x+u) \geq \overline{\Phi}(0) = 1/2$ and we derive from Bernstein inequality that

$$\psi_{x,r,s}(t) \geq \frac{s}{2} - \sqrt{s\log(2\alpha^{-1}_{x,r})} - \frac{2}{3}\log(2\alpha^{-1}_{x,r}) \ .$$

As a ce, $\psi_{x,r,s}(t) > \gamma_{x,s} - \eta_{x,r,s}$ as long as

$$s(1 - 4\overline{\Phi}(x)) > 12\sqrt{p\overline{\Phi}(x)\log(2\alpha^{-1}_{x,r})} + \frac{12}{3}\log(2\alpha^{-1}_{x,r}) \ .$$

Provided that we take $x \geq 2$, the latter holds if

$$s \geq 14\sqrt{p\overline{\Phi}(x)\log(2\alpha^{-1}_{x,r})} + \frac{14}{3}\log(2\alpha^{-1}_{x,r}) \tag{33}$$

In view of the definition (31) of x_s, we have $14\sqrt{p\overline{\Phi}(x_s)\log(2\alpha^{-1}_{x_s,r})} \leq s/2$. Hence, under Condition (28), (33) holds and we conclude that $\psi_{x_s,r,s}(x_s) > \gamma_{x_s,s} - \eta_{x_s,r,s}$.

□

Proof of Proposition 8. The following lemma ensures that the max test returns 0 with high probability jointly at all positions where there is no change-point.

Lemma 6 (concentration of the pure noise for the second sparse statistic)**.** *If $1 \geq \delta > 0$, then the event*

$$\xi^{(\mathrm{p})}_1 = \left\{ \forall r \in \mathcal{R}, l \in \mathcal{D}_r, s \in \mathcal{Z} \quad \max_{S \in \bar{C}^s_p} \sum_{i \in S} \frac{r}{2\sigma^2} \left(\bar{\varepsilon}_{l,+r,i} - \bar{\varepsilon}_{l,-r,i}\right)^2 \leq x^{(p)}_{r,s} \right\} \ .$$

holds with probability higher than $1-\delta$.

We now state the following lemma, which ensures that the max test returns 1 with high probability jointly at relevant positions which are close to a change-point.

Lemma 7 (concentration on the change-points for the second sparse statistic)**.** *We write $\bar{\mathcal{K}}^*$ for the set of $k \in [K]$ such that*

- $s_k \leq \sqrt{p \log\left(\frac{n}{r_k \delta}\right)}$
- $\sum_{s'=1}^{s} \left| U_{\bar{\tau}_k^{(s)}, \bar{r}_k^{(s)}, (s')} \right|^2 \geq 4 x^{(p)}_{\bar{r}_k^{(s)}, s}$

If $1 \geq \delta > 0$, the event

$$\xi_2^{(\mathrm{p})} = \left\{ \forall k \in \bar{\mathcal{K}}^* \ : \exists s \in \mathcal{Z} \quad s.t. \ \Psi^{(p)}_{\bar{\tau}_k^{(s)}, \bar{r}_k^{(s)}, s} > x^{(p)}_{\bar{r}_k^{(s)}, s} \right\} ,$$

holds with probability higher than $1 - \delta$.

Lemmas 6 and 7 directly imply the result of the proposition. □

Proof of Lemma 6. Let $r \in \mathcal{R}, l \in \mathcal{D}_r, s \leq \bar{s}_r$ and $S \in \bar{C}_p^s$. Let $\delta > 0$, $\delta_{r,s} = \left(\frac{r}{n}\right)^2 \left(\frac{s}{2ep}\right)^s \delta$. Since $\sqrt{\frac{r}{2\sigma^2}} \left(\bar{\varepsilon}_{l,+r,i} - \bar{\varepsilon}_{l,-r,i}\right)$ follows a $\mathcal{N}(0,1)$ distribution for all l, r, i, we have by Bernstein's inequality that with probability larger than $1 - \delta_{r,s}$,

$$\begin{aligned}
\sum_{i \in S} \left(\bar{\varepsilon}_{l,+r,i} - \bar{\varepsilon}_{l,-r,i}\right)^2 &\leq s + 2\sqrt{s \log\left(\frac{1}{\delta_{r,s}}\right) + \log\left(\frac{1}{\delta_{r,s}}\right)} \\
&\leq 2\left(s + \log\left(\frac{1}{\delta_{r,s}}\right)\right) \\
&= 2\left(s + s\log\left(\frac{2ep}{s}\right) + \log\left(\frac{n^2}{r^2\delta}\right)\right) \\
&\leq 4\left(s \log\left(\frac{2ep}{s}\right) + \log\left(\frac{n}{r\delta}\right)\right) \ .
\end{aligned}$$

Since the number of such S is smaller than $\left(\frac{ep}{s}\right)^s$, a union bound gives

$$\begin{aligned}
\mathbb{P}\left(\xi_1^{(\mathrm{p})}\right) &\geq 1 - \sum_{r \in \mathcal{R}} \sum_{l \in \mathcal{D}_r} \sum_{s \in \mathcal{Z}} |\bar{C}_p^s| \left(\frac{s}{2ep}\right)^s \left(\frac{r}{n}\right)^2 \delta \\
&\geq 1 - \sum_{r \in \mathcal{R}} \sum_{l \in \mathcal{D}_r} \sum_{s \in \mathcal{Z}} \left(\frac{1}{2}\right)^s \left(\frac{r}{n}\right)^2 \delta \\
&\geq 1 - \delta \ ,
\end{aligned}$$

which yields the result. □

Proof of Lemma 7. Let $k \in \bar{\mathcal{K}}^*$, and $s \in \mathcal{Z}$ such that

$$\sum_{i=1}^{s} U^2_{\bar{\tau}_k^{(s)}, \bar{r}_k^{(s)}, (i)} > 4 x^{(p)}_{\bar{r}_k^{(s)}, s} \ . \tag{34}$$

To ease the reading, we write $(\tau, r) = (\bar{\tau}_k^{(s)}, \bar{r}_k^{(s)})$. Then on the event $\xi_1^{(\mathrm{p})}$ which holds with probability $1 - \delta$, we have

$$\begin{aligned}
\Psi^{(p)}_{\tau,r,s} &= \max_{S \in \bar{C}_p^s} \sum_{i \in S} \frac{r}{2\sigma^2} \left(\bar{\theta}_{\tau,+r,i} + \bar{\varepsilon}_{\tau,+r,i} - \bar{\theta}_{\tau,-r,i} - \bar{\varepsilon}_{\tau,-r,i}\right)^2 \\
&\geq \max_{S \in \bar{C}_p^s} \sum_{i \in S} \frac{1}{2} U^2_{\tau,r,i} - \frac{r}{2\sigma^2} \left(\bar{\varepsilon}_{\tau,+r,i} - \bar{\varepsilon}_{\tau,-r,i}\right)^2 \\
&> 2 x^{(p)}_{r,s} - x^{(p)}_{r,s} \\
&= x^{(p)}_{r,s} \ ,
\end{aligned}$$

where in the second inequality, we used the fact that $(a+b)^2 \geq \frac{1}{2}a^2 - b^2$ for all $a, b \in \mathbb{R}$. □

Proof of Lemma 3. First remark that there exists a large enough constant C such that for all $r, s \geq 1$,

$$\left(x_{r,s}^{(\mathrm{BJ})}\right)^2 \leq C \log\left(\frac{ep}{s^2}\log\left(\frac{n}{r\delta}\right)\right)$$
$$\bar{s}_r \leq C \log\left(\log\left(\frac{ep}{\bar{s}_r^2}\right)\frac{n}{r\delta}\right) \ ,$$

where we recall that $\bar{s}_r$ is defined by (28) and $x_{r,s}^{(\mathrm{BJ})}$ by (31). These two inequalies come from the fact that for all $t \geq 2$ and all $A > 0$, if $t \leq A + \log(t)$ then $t \leq 2A$. Assume that for all $s' = \bar{s}_{\bar{r}_k^{(\mathrm{s})}} + 1, \ldots, s_k$ we have $|U_{\bar{\tau}_k^{(s)}, \bar{r}_k^{(\mathrm{s})}, (s')}| < x_{\bar{r}_k^{(\mathrm{s})}, s'}^{(\mathrm{BJ})}$. To ease the notation, we write $\bar{s} = \bar{s}_{\bar{r}_k^{(\mathrm{s})}} \wedge s_k$ and in what follows we prove that $\sum_{s'=1}^{\bar{s}} |U_{\bar{\tau}_k^{(s)}, \bar{r}_k^{(\mathrm{s})}, (s')}|^2 > 4x_{\bar{r}_k^{(\mathrm{s})}, \bar{s}}^{(p)}$ when κ_{s} is a large enough constant. We have

$$\sum_{s'=\bar{s}_{\bar{r}_k^{(\mathrm{s})}}+1}^{s_k} U^2_{\bar{\tau}_k^{(s)}, \bar{r}_k^{(\mathrm{s})}, (s')} \leq C_1 \sum_{i=0}^{\lfloor \log(s_k) \rfloor} 2^i \log\left(\frac{ep}{2^{2i}}\log\left(\frac{n}{\bar{r}_k^{(\mathrm{s})}\delta}\right)\right)$$
$$\leq C_1 s_k \log\left(2e\log\left(\frac{n}{\bar{r}_k^{(\mathrm{s})}\delta}\right)\right) + C_1 \sum_{i=0}^{\lfloor \log(s_k) \rfloor} 2^i \log\left(\frac{p}{2^{2(i+1)}}\right) \ ,$$

for some universal constant C_1. To handle the second term remark that since $x \mapsto \log\left(\frac{p}{x^2}\right)$ is decreasing, we have

$$\sum_{i=0}^{\lfloor \log(s_k) \rfloor} 2^i \log\left(\frac{p}{2^{2(i+1)}}\right) \leq \int_1^{2s_k} \log\left(\frac{p}{x^2}\right) dx$$
$$= 2s_k \log\left(\frac{p}{(2s_k)^2}\right) + 2s_k - 1$$
$$\leq 2s_k \log\left(\frac{p}{s_k^2}\right) \ ,$$

and thus

$$\sum_{s'=\bar{s}_{\bar{r}_k^{(\mathrm{s})}}+1}^{s_k} U^2_{\bar{\tau}_k^{(s)}, \bar{r}_k^{(\mathrm{s})}, (s')} \leq 2C_1 s_k \log\left(2e\frac{p}{s_k^2}\log\left(\frac{n}{\bar{r}_k^{(\mathrm{s})}\delta}\right)\right),$$

which finally gives

$$\sum_{s'=1}^{\bar{s}} U^2_{\bar{\tau}_k^{(s)}, \bar{r}_k^{(\mathrm{s})}, (s')} \geq \frac{9}{16}\bar{r}_k^{(\mathrm{s})}\Delta_k^2 - 2C_1 s_k \log\left(\frac{2ep}{s_k^2}\log\left(\frac{n}{\bar{r}_k^{(\mathrm{s})}\delta}\right)\right)$$
$$\geq 4x_{\bar{r}_k^{(\mathrm{s})}, \bar{s}}^{(p)}.$$

In the first inequality we used the fact that

$$\left|\bar{\tau}_k^{(s)} - \tau_k\right| \leq \frac{1}{4}\bar{r}_k^{(\mathrm{s})},$$

so that for all i,

$$\left|\bar{\theta}_{\bar{\tau}_k^{(s)}, +\bar{r}_k^{(\mathrm{s})}, i} - \bar{\theta}_{\bar{\tau}_k^{(s)}, -\bar{r}_k^{(\mathrm{s})}, i}\right| = \frac{1}{\bar{r}_k^{(\mathrm{s})}}\left|\left(\bar{r}_k^{(\mathrm{s})} + \bar{\tau}_k^{(s)} - \tau_k\right)\mu_{k,i} - \left(\bar{r}_k^{(\mathrm{s})} - \bar{\tau}_k^{(s)} + \tau_k\right)\mu_{k-1,i}\right|$$
$$\geq \left(1 - \frac{\left|\bar{\tau}_k^{(s)} - \tau_k\right|}{\bar{r}_k^{(\mathrm{s})}}\right)\left|\mu_{k,i} - \mu_{k-1,i}\right|$$
$$> \frac{3}{4}\left|\mu_{k,i} - \mu_{k-1,i}\right| = \frac{3}{4}U_{k,i} \ .$$

In the second inequality, we used the fact that

- $8\bar{r}_k^{(\mathrm{s})}\Delta_k^2 \geq \kappa_{\mathrm{s}}\sigma^2\left(s_k \log\left(\frac{p}{s_k^2}\log\left(\frac{n}{\bar{r}_k^{(\mathrm{s})}\delta}\right)\right) + \log\left(\frac{n}{\bar{r}_k^{(\mathrm{s})}\delta}\right)\right)$ for a large enough constant κ_{s} (see (11)),
- $x \mapsto x\log\left(\frac{ep}{x^2}\right)$ is increasing for $x \leq p$, so that s_k can be replaced by $\bar{s}$,
- $\bar{s} \leq C\log\left(\log\left(\frac{ep}{\bar{s}^2}\right)\frac{n}{r\delta}\right)$.

This concludes the proof of the lemma. □

B.2.3 Proof of Corollary 2

Let $\xi^{(\mathrm{d})}$ and $\xi^{(\mathrm{s})}$ be two events such that Proposition 1 and Proposition 2 hold respectively with constants $\kappa_{\mathrm{d}}, \kappa_{\mathrm{s}}$ and with probability $1-2\delta$ and $1-4\delta$, and write $\xi = \xi^{(\mathrm{d})} \cap \xi^{(\mathrm{s})}$. From now on, we work on the event ξ, which holds with probability $1-6\delta$. Let us choose $c_0 \geq 2(\kappa_{\mathrm{d}} \vee \kappa_{\mathrm{s}})$ in (8). For all k such that τ_k is a c_0-high-energy change-point, define

$$(\bar{\tau}_k, \bar{r}_k) = \begin{cases} (\bar{\tau}_k^{(d)}, \bar{r}_k^{(\mathrm{d})}) \text{ if } s_k > \sqrt{p\log\left(\frac{n}{r_k\delta}\right)} \\ (\bar{\tau}_k^{(s)}, \bar{r}_k^{(\mathrm{s})}) \text{ if } s_k \leq \sqrt{p\log\left(\frac{n}{r_k\delta}\right)} \end{cases}.$$

$(\bar{r}_k, \bar{\tau}_k)$ is well defined. Indeed, If $s_k \leq \sqrt{p\log\left(\frac{n}{r_k\delta}\right)}$ then

$$s_k\log\left(1+\frac{\sqrt{p}}{s_k}\sqrt{\log\left(\frac{n}{r_k\delta}\right)}\right) + \log\left(\frac{n}{r_k\delta}\right) \geq \frac{1}{2}\left(s_k\log\left(\frac{p}{s_k^2}\log\left(\frac{n}{r_k\delta}\right)\right) + \log\left(\frac{n}{r_k\delta}\right)\right).$$

Now if $s_k \geq \sqrt{p\log\left(\frac{n}{r_k\delta}\right)}$ then using $\log(1+x) \geq \frac{x}{2}$ for $x \in [0,1]$ we have

$$s_k\log\left(1+\frac{\sqrt{p}}{s_k}\sqrt{\log\left(\frac{n}{r_k\delta}\right)}\right) + \log\left(\frac{n}{r_k\delta}\right) \geq \frac{1}{2}\left(\sqrt{p\log\left(\frac{n}{r_k\delta}\right)} + \log\left(\frac{n}{r_k\delta}\right)\right).$$

According to Theorem 1, it is sufficient to prove that the event $\mathcal{A}(\Theta, T, \mathcal{K}^*, (\bar{\tau}_k, \bar{r}_k)_{k\in\mathcal{K}^*})$ defined in Section 2.3 holds on ξ:

1. **(No False Positive):** for every $r \in \mathcal{R}$ and $l \in \mathcal{D}_r$, if Θ is constant on $[l-r, l+r)$ then
$$T_{l,r} = T_{l,r}^{(\mathrm{d})} \vee T_{l,r}^{(\mathrm{s})} = 0,$$
by Proposition 1 and Proposition 2.

2. **(high-energy change-point Detection):** for every k such that τ_k has c_0-high-energy, it holds by definition of $\bar{r}_k^{(\mathrm{d})}$ and $\bar{r}_k^{(\mathrm{s})}$ that
$$4(\bar{r}_k - 1) \leq r_k.$$
Moreover, $T_{\bar{\tau}_k,\bar{r}_k}^{(\mathrm{s})} = 1$ if $(\bar{\tau}_k, \bar{r}_k) = (\bar{\tau}_k^{(d)}, \bar{r}_k^{(\mathrm{d})})$ by Proposition 2 and $T_{\bar{\tau}_k,\bar{r}_k}^{(\mathrm{d})} = 1$ if $(\bar{\tau}_k, \bar{r}_k) = (\bar{\tau}_k^{(s)}, \bar{r}_k^{(\mathrm{s})})$ by Proposition 1.

Theorem 1 ensures that for all $k \in \{1, \dots, K\}$ such that τ_k is a c_0-high-energy change-point, there exists $k' \in \{1, \dots, \hat{K}\}$ such that

$$|\hat{\tau}_{k'} - \tau_k| \leq \bar{r}_k - 1.$$

It remains to show that

$$\bar{r}_k - 1 \leq \frac{r_k^*}{2},$$

where r_k^* is define by (9). Using $\log(1+x) \geq \frac{x}{2}$ for $x \in [0,1]$ and $\log(1+x) \geq \log(x)$ for $x \geq 1$ we have

$$8\bar{r}_k\Delta_k^2 \leq 4(\kappa_{\mathrm{d}} \vee \kappa_{\mathrm{s}})\left[s_k\log\left(1+\frac{\sqrt{p}}{s_k}\sqrt{\log\left(\frac{n}{\bar{r}_k\delta}\right)}\right) + \log\left(\frac{n}{\bar{r}_k\delta}\right)\right],$$

when $\bar{r}_k \geq 2$. Thus $2(\bar{r}_k - 1) \leq r_k^*$ for $c_0 \geq 2(\kappa_{\mathrm{d}} \vee \kappa_{\mathrm{s}})$. This concludes the proof of Corollary 2.

B.3 Proofs in the sub-Gaussian setting

We recall that in this section, we work on the complete grid $\mathcal{G}_F = J_n$ defined in Section 2.

B.3.1 Proof of Proposition 3

Step 1: Introduction of useful high probability events. We first introduce two events $\xi_1^{(d)}$ and $\xi_2^{(d)}$ on which the noise can be controlled. Remark that by a simple computation, the noise can be decomposed as follows :

$$\frac{r}{2}\left[\left\|\overline{y}_{l,+r}-\overline{y}_{l,-r}\right\|^2-\left\|\overline{\theta}_{l,-r}-\overline{\theta}_{l,+r}\right\|^2\right]-\sigma^2 p = r\langle\overline{\varepsilon}_{l,+r}-\overline{\varepsilon}_{l,-r},\overline{\theta}_{l,+r}-\overline{\theta}_{l,-r}\rangle+\frac{r}{2}\left\|\overline{\varepsilon}_{l,+r}-\overline{\varepsilon}_{l,-r}\right\|^2-\sigma^2 p \ .$$

The first term written as

$$r\langle\overline{\varepsilon}_{l,+r}-\overline{\varepsilon}_{l,-r},\overline{\theta}_{l,+r}-\overline{\theta}_{l,-r}\rangle$$

is a crossed term between the noise and the mean vector θ. Lemma 8 states that for l equal to a true change-point τ_k and r of order r_k^*, it is controlled on event $\xi_1^{(d)}$ with high probability.

Lemma 8 (concentration of the crossed terms)**.** *Assume that κ is a large enough universal constant. The event*

$$\xi_1^{(d)} = \Bigg\{\forall k\in\{1,\dots,K\} \text{ s.t. Equation (17) holds for } k,$$
$$\bar{r}_k^{(\mathrm{d})}\left|\langle\overline{\varepsilon}_{\tau_k,+\bar{r}_k^{(\mathrm{d})}}-\overline{\varepsilon}_{\tau_k,-\bar{r}_k^{(\mathrm{d})}},\overline{\theta}_{\tau_k,+\bar{r}_k^{(\mathrm{d})}}-\overline{\theta}_{\tau_k,-\bar{r}_k^{(\mathrm{d})}}\rangle\right|\le\frac{\bar{r}_k^{(\mathrm{d})}}{4}\left\|\overline{\theta}_{\tau_k,+\bar{r}_k^{(\mathrm{d})}}-\overline{\theta}_{\tau_k,-\bar{r}_k^{(\mathrm{d})}}\right\|^2\Bigg\}$$

holds with probability higher than $1-\delta$.

The second term written as

$$\frac{r}{2}\left\|\overline{\varepsilon}_{l,+r}-\overline{\varepsilon}_{l,-r}\right\|^2-\sigma^2 p \ ,$$

is a term of pure noise. Lemma 9 states that it is controlled on event $\xi_2^{(d)}$ with high probability.

Lemma 9 (concentration of the pure noise)**.** *There exists a constant $\bar{c}_{\mathrm{conc}}>0$ such that the event*

$$\xi_2^{(d)} = \left\{\forall(l,r)\in J_n,\ \left|\frac{r}{2}\left\|\overline{\varepsilon}_{l,+r}-\overline{\varepsilon}_{l,-r}\right\|^2-\sigma^2 p\right|\le\bar{c}_{\mathrm{conc}}L^2\left(\sqrt{p\log\left(\frac{n}{r\delta}\right)}+\log\left(\frac{n}{r\delta}\right)\right)\right\}$$

holds with probability higher than $1-2\delta$.

Set now

$$\xi^{(d)} := \xi_1^{(d)}\cap\xi_2^{(d)} \ .$$

Note that

$$\mathbb{P}(\xi^{(d)})\ge 1-3\delta \ .$$

Step 2: Study in the 'no change-point' situation. We remind that $\mathcal{H}_0$ stands for elements (l,r) such that there is no change-point in $[l-r,l+r)$ and that it is defined in (7). Consider $(l,r)\in J_n\cap\mathcal{H}_0$. Note that since $\{\tau_k,k\in\{1,\dots,K\}\}\cap[l-r,l+r)=\varnothing$, we have $\overline{\theta}_{l,-r}=\overline{\theta}_{l,+r}$ so that

$$\frac{r}{2}\left\|\overline{\theta}_{l,-r}-\overline{\theta}_{l,+r}\right\|^2=0 \ ,$$

and

$$r\langle \overline{\varepsilon}_{l,+r} - \overline{\varepsilon}_{l,-r}, \overline{\theta}_{l,+r} - \overline{\theta}_{l,-r} \rangle = 0 \ .$$

Moreover we have on $\xi^{(d)}$ that - see Lemma 9

$$\left| \frac{r}{2} \left\| \overline{\varepsilon}_{l,+r} - \overline{\varepsilon}_{l,-r} \right\|^2 - \sigma^2 p \right| \leq \bar{c}_{\text{conc}} L^2 \left(\sqrt{p \log\left(\frac{n}{r\delta}\right)} + \log\left(\frac{n}{r\delta}\right) \right) \leq \sigma^2 x_r^{(\text{d})},$$

for $\bar{c}_{\text{thresh}} \geq \bar{c}_{\text{conc}}$ - note that $\bar{c}_{\text{conc}} > 0$ is a universal constant. And so

$$\Psi_{l,r}^{(\text{d})} \leq x_r^{(\text{d})} \ ,$$

so that

$$T_{l,r}^{(\text{d})} = 0 \ ,$$

on $\xi^{(d)}$. This concludes the proof of the first part of the proposition.

Step 3: Study in the 'change-point' situation. Consider $k \in \{1, \dots, K\}$ such that τ_k is a κ-dense high energy change-point - see Equation (17). We have

$$\frac{\bar{r}_k^{(\text{d})}}{2} \left\| \overline{\theta}_{\tau_k, -\bar{r}_k^{(\text{d})}} - \overline{\theta}_{\tau_k, +\bar{r}_k^{(\text{d})}} \right\|^2 \geq \frac{\kappa}{8} L^2 \left(\sqrt{p \log\left(\frac{n}{\bar{r}_k^{(\text{d})}\delta}\right)} + \log\left(\frac{n}{\bar{r}_k^{(\text{d})}\delta}\right) \right).$$

So on $\xi^{(d)}$ choosing κ large enough implies that - see Lemma 8

$$\bar{r}_k^{(\text{d})} \left| \langle \overline{\varepsilon}_{\tau_k, +\bar{r}_k^{(\text{d})}} - \overline{\varepsilon}_{\tau_k, -\bar{r}_k^{(\text{d})}}, \overline{\theta}_{\tau_k, +\bar{r}_k^{(\text{d})}} - \overline{\theta}_{\tau_k, -\bar{r}_k^{(\text{d})}} \rangle \right| \leq \frac{\bar{r}_k^{(\text{d})}}{4} \left\| \overline{\theta}_{\tau_k, +\bar{r}_k^{(\text{d})}} - \overline{\theta}_{\tau_k, -\bar{r}_k^{(\text{d})}} \right\|^2 \ .$$

Moreover we have on $\xi^{(d)}$ that - see Lemma 9

$$\left| \frac{\bar{r}_k^{(\text{d})}}{2} \left\| \overline{\varepsilon}_{\tau_k, +\bar{r}_k^{(\text{d})}} - \overline{\varepsilon}_{\tau_k, -\bar{r}_k^{(\text{d})}} \right\|^2 - \sigma^2 p \right| \leq \bar{c}_{\text{conc}} L^2 \left(\sqrt{p \log\left(\frac{n}{\bar{r}_k^{(\text{d})}\delta}\right)} + \log\left(\frac{n}{\bar{r}_k^{(\text{d})}\delta}\right) \right) \leq \sigma^2 x_{\bar{r}_k^{(\text{d})}}^{(\text{d})} \ ,$$

for $\bar{c}_{\text{thresh}} \geq \bar{c}_{\text{conc}}$ - note that $\bar{c}_{\text{conc}} > 0$ is a universal constant. Thus on $\xi^{(d)}$, combining the three previous displayed equations implies

$$\begin{aligned} \Psi_{\tau_k, \bar{r}_k^{(\text{d})}}^{(\text{d})} &\geq \frac{\bar{r}_k^{(\text{d})}}{4\sigma^2} \left\| \overline{\theta}_{\tau_k, +\bar{r}_k^{(\text{d})}} - \overline{\theta}_{\tau_k, -\bar{r}_k^{(\text{d})}} \right\|^2 - x_{\bar{r}_k^{(\text{d})}}^{(\text{d})} \\ &\geq \left(\frac{c_0}{16} - \bar{c}_{\text{thresh}} \right) \frac{L^2}{\sigma^2} \left(\sqrt{p \log\left(\frac{n}{\bar{r}_k^{(\text{d})}\delta}\right)} + \log\left(\frac{n}{\bar{r}_k^{(\text{d})}\delta}\right) \right) > x_{\bar{r}_k^{(\text{d})}}^{(\text{d})} \ , \end{aligned}$$

since $\kappa > 32\bar{c}_{\text{thresh}}$. And so on $\xi^{(d)}$:

$$T_{\tau_k, \bar{r}_k^{(\text{d})}}^{(\text{d})} = 1 \ .$$

This concludes the proof of the second part of the proposition.

Proof of Lemma 8. Let k be in $\{1, \dots K\}$ and such that Equation (17) is satisfied. Remark that θ is constant on $[\tau_k - \bar{r}_k^{(\text{d})}, \tau_k)$ and is equal to μ_{k-1}, and is also constant on $[\tau_k, \tau_k + \bar{r}_k^{(\text{d})})$ and is equal to

μ_k. First, from the definition of the ψ_2-norm of a vector, there exists a universal constant $C > 0$ such that for all $k = 1 \dots K$,

$$\begin{aligned}
\left\| \bar{r}_k^{(\mathrm{d})} \langle \overline{\varepsilon}_{\tau_k, +\bar{r}_k^{(\mathrm{d})}} - \overline{\varepsilon}_{\tau_k, -\bar{r}_k^{(\mathrm{d})}}, \overline{\theta}_{\tau_k, +\bar{r}_k^{(\mathrm{d})}} - \overline{\theta}_{\tau_k, -\bar{r}_k^{(\mathrm{d})}} \rangle \right\|_{\psi_2} &\leq \bar{r}_k^{(\mathrm{d})} \left\| \overline{\varepsilon}_{\tau_k, +\bar{r}_k^{(\mathrm{d})}} - \overline{\varepsilon}_{\tau_k, -\bar{r}_k^{(\mathrm{d})}} \right\|_{\psi_2} |\mu_k - \mu_{k-1}| \\
&\leq C \sqrt{\bar{r}_k^{(\mathrm{d})}} \, \|\varepsilon_1\|_{\psi_2} \, |\mu_k - \mu_{k-1}| \\
&\leq C \sqrt{\bar{r}_k^{(\mathrm{d})}} L \, |\mu_k - \mu_{k-1}| \\
&\leq C L \sqrt{r_k \Delta_k^2} \ .
\end{aligned}$$

Thus by definition of sub-Gaussianity, for all $t > 0$,

$$\mathbb{P}\left(\bar{r}_k^{(\mathrm{d})} \left| \langle \overline{\varepsilon}_{\tau_k, +\bar{r}_k^{(\mathrm{d})}} - \overline{\varepsilon}_{\tau_k, -\bar{r}_k^{(\mathrm{d})}}, \overline{\theta}_{\tau_k, +\bar{r}_k^{(\mathrm{d})}} - \overline{\theta}_{\tau_k, -\bar{r}_k^{(\mathrm{d})}} \rangle \right| \geq t \right) \leq \exp\left(-c \frac{t^2}{L^2 r_k \Delta_k^2} \right) \ ,$$

for some constant $c > 0$. Finally we apply the concentration inequality to $t = \frac{r_k \Delta_k^2}{4}$ - remembering that τ_k is a κ-dense high energy change-point in the sense of Equation (17) - and sum over k to obtain a union bound over ξ_2^c :

$$\begin{aligned}
\mathbb{P}\left(\xi_2^c\right) &\leq \sum_{k=1}^{K} \mathbb{P}\left(r \left| \langle \overline{\varepsilon}_{\tau_k, +\bar{r}_k^{(\mathrm{d})}} - \overline{\varepsilon}_{\tau_k, -\bar{r}_k^{(\mathrm{d})}}, \overline{\theta}_{\tau_k, +\bar{r}_k^{(\mathrm{d})}} - \overline{\theta}_{\tau_k, -\bar{r}_k^{(\mathrm{d})}} \rangle \right| \geq \frac{r_k \Delta_k^2}{4} \right) \\
&\leq \sum_{k=1}^{K} \exp\left(-c \frac{r_k \Delta_k^2}{16 L^2} \right) \\
&\leq \sum_{k=1}^{K} \exp\left(-c' \kappa \log\left(\frac{n}{\bar{r}_k^{(\mathrm{d})}} \delta^{-1} \right) \right) && (c' = c/16) \\
&\leq \sum_{k=1}^{K} \left(\frac{\bar{r}_k^{(\mathrm{d})}}{n} \right)^{c'\kappa} \delta^{c'\kappa} \\
&\leq \delta \ ,
\end{aligned}$$

where the last inequality comes from the fact that $\sum_{k=1}^{K} \bar{r}_k^{(\mathrm{d})} \leq n$ and the fact that κ is chosen large enough so that $c'\kappa \geq 1$. □

Proof of Lemma 9. Remark first that by homogeneity, we can assume without loss of generality that $L = 1$. To provide a proof, we will use the Hanson-Wright inequality in high dimension, which is a way to control quadratic forms of the noise.

Lemma 10 (Hanson-Wright inequality in high dimension)**.** *Let $A = (a_{ij})$ be a $m \times m$ matrix and $\varepsilon_1, \dots, \varepsilon_m$ be sub-Gaussian vectors of dimension p with norm smaller than 1. Then*

$$\mathbb{P}\left(\left| \sum_{1 \leq i,j \leq m} a_{i,j} \langle \varepsilon_i, \varepsilon_j \rangle - \mathbb{E}\left[\sum_{1 \leq i,j \leq m} a_{i,j} \langle \varepsilon_i, \varepsilon_j \rangle \right] \right| \geq t \right) \leq 2 \exp\left(-c \min\left(\frac{t^2}{p \|A\|_F^2}, \frac{t}{\|A\|} \right) \right) \ ,$$

where c is an absolute constant, $\|A\|_F^2 = \sum_{i,j} a_{i,j}^2$ is the squared Frobenius norm of A and $\|A\|$ is the operator norm of A.

The proof of this lemma relies on the classical Hanson Wright inequality that is proved for example in [RV+13]. To prove the proposition, we will use a chaining argument. To this end, we let $(N_u)_{u \geq 0}$ be the following covering sets of J_n :

$$N_u = J_n \cap \{ i 2^{\kappa_1 - u}, i \in \mathbb{N} \}^2 \ ,$$

where we define $\kappa_1 = \lfloor \log_2(n) \rfloor$, and more generally $\kappa_r = \lfloor \log_2(n/r) \rfloor$ for $r = 1, \dots n$. Remark that the higher u is, the finer the covering set N_u is, and $N_{\kappa_1} = J_n$. For all $u \geq 0$, we define the projection map π_u from J_n to N_u by

$$\pi_u(l,r) = \underset{(\hat{l},\hat{r}) \in N_u}{\arg\min} \left(|\hat{l} - l| + |\hat{r} - r| \right) \ .$$

In the sequel, we will use the slight abuse of notation for (l,r) in J_n :

$$(l_u, r_u) = \pi_u(l,r) \ .$$

A useful lemma to control the distance between (l,r) and its projection (l_u, r_u) can be stated as follow.

Lemma 11. *For all $(l,r) \in J_n$ and $0 \leq u \leq \kappa_1$ such that $N_u \neq \varnothing$,*

$$|l_u - l| + |r_u - r| \leq 2\frac{n}{2^u} \ .$$

Let $(l,r) \in J_n$. From know on, we write $\varepsilon_{l,+r} = r\bar{\varepsilon}_{l,+r} = \sum_{t=l}^{l+r-1} \varepsilon_t$ and $\varepsilon_{l,-r} = r\bar{\varepsilon}_{l,+r}$. The chaining relation can be written as

$$\begin{aligned} \frac{r}{2} \left\| \bar{\varepsilon}_{l,+r} - \bar{\varepsilon}_{l,-r} \right\|^2 - \sigma^2 p = \frac{1}{2r} \left[\left\| \varepsilon_{l_{\kappa_r},+r_{\kappa_r}} - \varepsilon_{l_{\kappa_r},-r_{\kappa_r}} \right\|^2 - 2r_{\kappa_r}\sigma^2 p \right] \\ + \frac{1}{2r} \sum_{v=\kappa_r}^{\kappa_1} \left[\left\| \varepsilon_{l_{v+1},+r_{v+1}} - \varepsilon_{l_{v+1},-r_{v+1}} \right\|^2 - \left\| \varepsilon_{l_v,+r_v} - \varepsilon_{l_v,-r_v} \right\|^2 - 2(r_{v+1} - r_v)\sigma^2 p \right] \ . \end{aligned}$$

Remark that the chaining summation starts at scale $u = \kappa_r$ so that $\frac{n}{2^u} \asymp r$. The first term of the chaining is an approximation on the grid at level u of the term $\frac{r}{2} \left\| \bar{\varepsilon}_{l,+r} - \bar{\varepsilon}_{l,-r} \right\|^2 - \sigma^2 p$. The second term can be viewed as an error term, and we will show that it is of the same order as the first term. Since both terms are quadratic forms of the noise, we will need an upper bound on the norm of their corresponding matrix to apply the Hanson Wright inequality - see Lemma 10.

Lemma 12 (Control of the Frobenius norm)**.** *Let (l,r) be a fixed element of J_n. Let A and B be the corresponding matrix of the two following quadratic form :*

$$\varepsilon^T A \varepsilon = \left\| \varepsilon_{l,+r} - \varepsilon_{l,-r} \right\|^2 \quad \text{and} \quad \varepsilon^T B \varepsilon = \left\| \varepsilon_{l,+r} - \varepsilon_{l,-r} \right\|^2 - \left\| \varepsilon_{l',+r'} - \varepsilon_{l',-r'} \right\|^2 \ .$$

Then

$$\begin{aligned} \|A\|_F^2 &\leq 16r^2 \\ \|B\|_F^2 &\leq 24 \left(\left| l - l' \right| + \left| r - r' \right| \right) \left(r + r' + |l - l'| \right) \ . \end{aligned}$$

The following lemma aims at upper bounding the first term of the chaining relation with high probability.

Lemma 13. *There exists a constant C_N such that for all n, the event*

$$\xi_N^{(d)} = \bigcap_{u \geq 0} \bigcap_{\substack{(l,r) \in N_u \\ r \leq 3\frac{n}{2^u}}} \left\{ \left| \left\| \varepsilon_{l,+r} - \varepsilon_{l,-r} \right\|^2 - 2r\sigma^2 p \right| \leq C_N r \left(\sqrt{p \log\left(2^u \delta^{-1}\right)} + \log\left(2^u \delta^{-1}\right) \right) \right\} \ .$$

holds with probability higher than $1 - \delta$.

For $u = \kappa_r$, $(l_u, r_u) \in N_u$ Lemma 11 gives $r_u \leq r + 2\frac{n}{2^u} \leq 3\frac{n}{2^u}$. Consequently, on the event $\xi_N^{(d)}$, we obtain

$$\left| \frac{1}{2r} \left\| \varepsilon_{l_{\kappa_r},+r_{\kappa_r}} - \varepsilon_{l_{\kappa_r},-r_{\kappa_r}} \right\|^2 - \frac{r_{\kappa_r}}{r} \sigma^2 p \right| \leq C'_N \left(\sqrt{p \log\left(\frac{n}{r\delta}\right)} + \log\left(\frac{n}{r\delta}\right) \right) \ ,$$

for C'_N a large absolute constant. To upper bound the second term, we use the following lemma :

Lemma 14. *For all (l,r) and (l',r') in J_n, set*

$$\xi^{(d)}_{\Delta,v}(l,r,l',r') = \left\{ \left| \|\varepsilon_{l',+r'} - \varepsilon_{l',-r'}\|^2 - \|\varepsilon_{l,+r} - \varepsilon_{l,-r}\|^2 - 2(r'-r)\sigma^2 p \right| \le C_\Delta \sqrt{\frac{rn}{2^v}} \left(\sqrt{p \log\left(2^v \delta^{-1}\right)} + \log\left(2^v \delta^{-1}\right) \right) \right\} \ .$$

There exists a constant C_Δ such that for all n, The event

$$\xi^{(d)}_\Delta = \bigcap_{v \ge 0} \left\{ \xi^{(d)}_{\Delta,v}\left(l,r,l',r'\right) \text{ holds for all } ((l,r),(l',r')) \in N_v \times N_{v+1} \text{ s.t. } |l-l'| + |r-r'| \le 3\frac{n}{2^v} \right\} \ .$$

holds with probability higher than $1-\delta$.

For $v \ge \kappa_r$, $((l_v,r_v),(l_{v+1},r_{v+1})) \in N_v \times N_{v+1}$ and by Lemma 11,

$$\begin{aligned} |r_v - r_{v+1}| + |l_v - l_{v+1}| &\le |r_v - r| + |l_v - l| + |r - r_{v+1}| + |l - l_{v+1}| \\ &\le 3\frac{n}{2^v}. \end{aligned}$$

Therefore, on the event $\xi^{(d)}_\Delta$,

$$\begin{aligned} &\left| \frac{1}{2r} \sum_{v=\kappa_r}^{\kappa_1 - 1} \left[\left\| \varepsilon_{l_{v+1},+r_{v+1}} - \varepsilon_{l_{v+1},-r_{v+1}} \right\|^2 - \left\| \varepsilon_{l_v,+r_v} - \varepsilon_{l_v,-r_v} \right\|^2 - 2(r_{v+1} - r_v)\sigma^2 p \right] \right| \\ &\le C_\Delta \frac{1}{2r} \sum_{v=\kappa_r}^{\kappa_1 - 1} \sqrt{\frac{r_v n}{2^v}} \left(\sqrt{p \log\left(2^v \delta^{-1}\right)} + \log\left(2^v \delta^{-1}\right) \right) \\ &\le C'_\Delta \sum_{v' \ge 0} \frac{1}{2^{v'}} \left(\sqrt{p \log\left(\frac{n 2^{v'}}{r\delta}\right)} + \log\left(\frac{n 2^{v'}}{r\delta}\right) \right) \\ &\le C''_\Delta \left(\sqrt{p \log\left(\frac{n}{r\delta}\right)} + \log\left(\frac{n}{r\delta}\right) \right), \end{aligned}$$

where C'_Δ, C''_Δ are large absolute constants. Hence, letting $\bar{c}_{\text{conc}} = C'_N + C''_\Delta$ we obtain

$$\xi^{(d)}_N \cap \xi^{(d)}_\Delta \subset \xi^{(d)}_2 \ ,$$

which must be of probability higher than $1 - 2\delta$. □

Proof of Lemma 11. Since the mesh of the grid N_u is equal to $2^{\kappa_1 - u} \le \frac{n}{2^u}$, there exists $(\tilde{l},\tilde{r}) \in N_u$ such that

$$|l - \tilde{l}| \le \frac{n}{2^u} \quad \text{and} \quad |r - \tilde{r}| \le \frac{n}{2^u} \ .$$

□

Proof of Lemma 12. Let us write

$$\varepsilon^T A \varepsilon = \sum_{l-r \le i,j < l+r} a_{ij} \langle \varepsilon_i, \varepsilon_j \rangle \quad \text{and} \quad \varepsilon^T B \varepsilon = \sum_{m_1 \le i,j < m_2} b_{ij} \langle \varepsilon_i, \varepsilon_j \rangle,$$

where $m_1 = \min(l-r, l'-r')$, $m_2 = \max(l+r, l'+r')$. Remark that for all i,j in $[l-r, l+r)$, $a_{ij} \le 2$. This gives the first inequality.
For the second inequality, assume without loss of generality that $l \le l'$. As for the first inequality, $b_{ij} \le 2$ for all $i,j \in [m_1,m_2)$. Remark that b_{ij} can be non zero only if (i,j) is in one of the following cases :

1. i or j is in $[\min(l+r, l'+r'), \max(l+r, l'+r'))$
2. i or j is in $[\min(l-r, l'-r'), \max(l-r, l'-r'))$
3. i or j is in $[l, l')$.

Hence there is at most $(4(|l-l'|+|r-r'|)+2|l-l'|)(r+r'+|l-l'|)$ non zero b_{ij}, and we obtain the second inequality. □

Proof of Lemma 13. The probability of $(\xi_N^{(d)})^c$ can be written as :

$$\mathbb{P}\left((\xi_N^{(d)})^c\right) = \mathbb{P}\Bigg(\exists u \geq 0, \exists (l,r) \in N_u \ \text{ s.t. } \ r \leq 3\frac{n}{2^u} \text{ and } \\ \left|\left\|\varepsilon_{l,+r} - \varepsilon_{l,-r}\right\|^2 - 2r\sigma^2 p\right| \leq C_N r\left(\sqrt{p\log\left(2^u\delta^{-1}\right)} + \log\left(2^u\delta^{-1}\right)\right)\Bigg) \ .$$

First, fix $u \geq 0$ and $(l,r) \in N_u$ such that $r \leq 3\frac{n}{2^u}$.

Applying the first inequality of Lemma 12 and the Hanson-Wright inequality - see Lemma 10, we obtain for all $t \geq 0$

$$\mathbb{P}\left(\left|\left\|\varepsilon_{l,+r} - \varepsilon_{l,-r}\right\|^2 - 2r\sigma^2 p\right| \geq t\right) \leq 2\exp\left(-c\min\left(\frac{t^2}{pr^2}, \frac{t}{r}\right)\right) \ ,$$

where c is an absolute constant. Choosing

$$t = C_N r\left(\sqrt{p\log\left(2^u\delta^{-1}\right)} + \log\left(2^u\delta^{-1}\right)\right) \ ,$$

we obtain

$$\mathbb{P}\left(\left|\left\|\varepsilon_{l,+r} - \varepsilon_{l,-r}\right\|^2 - 2r\sigma^2 p\right| \geq C_N r\left(\sqrt{p\log\left(2^u\delta^{-1}\right)} + \log\left(2^u\delta^{-1}\right)\right)\right) \leq C\left(\frac{1}{2^u}\right)^{cC_N}\delta^{cC_N} \ ,$$

where c, C are absolute constants. Since the cardinal of N_u is upper bounded by 2^{2u+2}, A union bound on each N_u for each $u \geq 0$ gives :

$$\begin{aligned}\mathbb{P}\left((\xi_N^{(d)})^c\right) &\leq \sum_{u\geq 0} C\,|N_u|\left(\frac{1}{2^u}\right)^{cC_N}\delta^{cC_N} \\ &\leq \sum_{u\geq 0} 4C\left(\frac{1}{2^u}\right)^{2-cC_N}\delta^{cC_N} \ ,\end{aligned}$$

which is convergent. For C_N large enough, we obtain $\mathbb{P}\left(\xi_N^c\right) \leq 1-\delta$. □

Proof of Lemma 14.

$$\mathbb{P}\left((\xi_\Delta^{(d)})^c\right) = \mathbb{P}\left(\exists v \geq 0, \exists((l,r),(l',r')) \in N_v \times N_{v+1} \text{ s.t. } |l-l'|+|r-r'| \leq 4\frac{n}{2^v} \text{ and } (\xi_{\Delta,v}^{(d)}\left(l,r,l',r'\right))^c \text{ holds }\right) \ .$$

First fix $v \geq 0$ and $((l,r),(l',r')) \in N_v \times N_{v+1}$. Remark that by definition of N_v,

$$r \geq \frac{n}{2^{v+1}} \ .$$

Thus,

$$r + r' + |l-l'| \leq 2r + |l-l'| + |r-r'| \leq 10r \ .$$

Then by Lemma 12, letting B be the matrix such that $\varepsilon^T B\,\varepsilon = \left\|\varepsilon_{l',+r'} - \varepsilon_{l',-r'}\right\|^2 - \left\|\varepsilon_{l,+r} - \varepsilon_{l,-r}\right\|^2$, we obtain

$$\|B\|^2 \leq \|B\|_F^2 \leq 40r\frac{n}{2^v} \ .$$

Thus, by the Hanson Wright inequality - see Lemma 10,

$$\mathbb{P}\left(\left|\varepsilon^T B_u\,\varepsilon - \mathbb{E}\left[\varepsilon^T B_u\,\varepsilon\right]\right| \geq t\right) \leq 2\exp\left(-c\min\left(\frac{2^v}{pnr}t^2, \sqrt{\frac{2^v}{nr}}t\right)\right) \ .$$

From now on, we choose

$$t = C_\Delta \sqrt{\frac{rn}{2^v}} \left(\sqrt{p \log\left(2^v \delta^{-1}\right)} + \log\left(2^v \delta^{-1}\right) \right) \ .$$

There are at most 2^{4v+6} elements in $N_v \times N_{v+1}$. Therefore, a union bound on $v \geq 0$ and $N_v \times N_{v+1}$ gives

$$\begin{aligned}
\mathbb{P}\left((\xi_\Delta^{(d)})^c \right) &\leq \sum_{u \geq 0} 2 |N_v \times N_{v+1}| \, (2^v)^{-cC_\Delta} \, \delta^{cC_\Delta} \\
&\leq \sum_{u \geq 0} 2^7 \, (2^v)^{4 - cC_\Delta} \, \delta^{cC_\Delta} \\
&\leq C \delta^{cC_\Delta},
\end{aligned}$$

where the last inequality holds if C_Δ is large enough, for c, C universal constants. □

B.3.2 Proof of Proposition 4

Step 1: Introduction of useful high probability events. Let $s \leq p$ and consider $S \in \bar{C}_p^s$. In what follows and for an vector $u \in \mathbb{R}^p$, we write $u^{(S)}$ for the vector u restricted to the set S.

Remark that by a simple computation, the noise can be decomposed as follows :

$$\begin{aligned}
&\frac{r}{2} \left[\left\| \bar{y}_{l,+r}^{(S)} - \bar{y}_{l,-r}^{(S)} \right\|^2 - \left\| \bar{\theta}_{l,-r}^{(S)} - \bar{\theta}_{l,+r}^{(S)} \right\|^2 \right] - \sigma^2 s \\
= \ & r \langle \bar{\varepsilon}_{l,+r}^{(S)} - \bar{\varepsilon}_{l,-r}^{(S)}, \bar{\theta}_{l,+r}^{(S)} - \bar{\theta}_{l,-r}^{(S)} \rangle + \frac{r}{2} \left\| \bar{\varepsilon}_{l,+r}^{(S)} - \bar{\varepsilon}_{l,-r}^{(S)} \right\|^2 - \sigma^2 s \ .
\end{aligned}$$

The first term written as

$$r \langle \bar{\varepsilon}_{l,+r}^{(S)} - \bar{\varepsilon}_{l,-r}^{(S)}, \bar{\theta}_{l,+r}^{(S)} - \bar{\theta}_{l,-r}^{(S)} \rangle \ ,$$

is a crossed term between the noise and the mean vector θ. Lemma 8 states that for l equal to a true change-point τ_k, r of order r_k^*, and S being the corresponding support of the change-point, it is controlled on event $\xi_1^{(\mathrm{p})}$ with high probability.

Lemma 15. *For $k \in \{1, \dots, K\}$, let us write $S_k \subset \{1, \dots, K\}$ for the support of $\mu_k - \mu_{k-1}$. Assume that c_0 is a large enough universal constant. The event*

$$\xi_1^{(\mathrm{p})} := \xi_1^{(\mathrm{p})}(\delta) = \Bigg\{ \forall k \in \{1, \dots, K\} \text{ s.t. Equation (18) holds for } k,$$

$$\bar{r}_k^{(\mathrm{d})} \left| \langle \bar{\varepsilon}_{\tau_k, +\bar{r}_k^{(\mathrm{d})}}^{(S_k)} - \bar{\varepsilon}_{\tau_k, -\bar{r}_k^{(\mathrm{s})}}^{(S_k)}, \bar{\theta}_{\tau_k, +\bar{r}_k^{(\mathrm{s})}}^{(S_k)} - \bar{\theta}_{\tau_k, -\bar{r}_k^{(\mathrm{s})}}^{(S_k)} \rangle \right| \leq \frac{\bar{r}_k^{(\mathrm{d})}}{4} \left\| \bar{\theta}_{\tau_k, +\bar{r}_k^{(\mathrm{s})}} - \bar{\theta}_{\tau_k, -\bar{r}_k^{(\mathrm{s})}} \right\|^2 \Bigg\}$$

holds with probability higher than $1 - \delta$.

The proof of this lemma follows directly from the one of Lemma 8, restricting the term corresponding to change-point k to S_k - and diminishing the deviation by doing so.

The second term written as

$$\frac{r}{2} \left\| \bar{\varepsilon}_{l,+r}^{(S)} - \bar{\varepsilon}_{l,-r}^{(S)} \right\|^2 - \sigma^2 s$$

is a term of pure noise. Lemma 16 states that it is controlled on event $\xi_2^{(\mathrm{p})}(S)$ with high probability.

Lemma 16. *There exists a constant $\bar{c}_{\mathrm{conc}} > 0$ such that the event*

$$\begin{aligned}
\xi_2^{(\mathrm{p})}(S) := \xi_2^{(\mathrm{p})}(S, \delta) = \Bigg\{ & \forall (l, r) \in J_n, \ \left| \frac{r}{2} \left\| \bar{\varepsilon}_{l,+r}^{(S)} - \bar{\varepsilon}_{l,-r}^{(S)} \right\|^2 - \sigma^2 s \right| \\
& \leq \bar{c}_{\mathrm{conc}} L^2 \left(\sqrt{s \log\left(\frac{n}{r\delta} \right)} + \log\left(\frac{n}{r\delta} \right) \right) \Bigg\}
\end{aligned}$$

holds with probability higher than $1 - 2\delta$.

The proof of this lemma is exactly the same as the one of Lemma 9, restricting all vectors to S.

Set $\delta_s = \delta/(2^s\binom{p}{s})$. Lemma 16 implies that with probability larger than $1-2\delta$, $\forall (l,r) \in J_n$, $\forall S \subset \{1,\dots,p\}$

$$\left|\frac{r}{2}\left\|\bar{\varepsilon}^{(S)}_{l,+r} - \bar{\varepsilon}^{(S)}_{l,-r}\right\|^2 - \sigma^2 s\right| \leq \bar{c}_{\text{conc}} L^2 \left(\sqrt{s\log\left(\frac{n}{r\delta_s}\right)} + \log\left(\frac{n}{r\delta_s}\right)\right).$$

And so since $\binom{p}{s} \leq \left(\frac{ep}{s}\right)^s$, we have probability larger than $1-2\delta$, $\forall (l,r) \in J_n$, $\forall S \subset \{1,\dots,p\}$

$$\begin{aligned}\left|\frac{r}{2}\left\|\bar{\varepsilon}^{(S)}_{l,+r} - \bar{\varepsilon}^{(S)}_{l,-r}\right\|^2 - \sigma^2 s\right| &\leq \bar{c}_{\text{conc}} L^2 \left(\sqrt{s\log\left(\frac{n}{r\delta}\right) + s\log\left(\frac{2ep}{s}\right)} + \log\left(\frac{n}{r\delta}\right) + s\log\left(\frac{2ep}{s}\right)\right) \\ &\leq 4\bar{c}_{\text{conc}} L^2 \left(\log\left(\frac{n}{r\delta}\right) + s\log\left(\frac{2ep}{s}\right)\right).\end{aligned}$$

And so the event

$$\begin{aligned}\xi_2^{(\mathrm{p})} := \xi_2^{(\mathrm{p})}(\delta) = \Bigg\{ \forall (l,r) \in J_n, \forall S \subset \{1,\dots,p\}, \left|\frac{r}{2}\left\|\bar{\varepsilon}^{(S)}_{l,+r} - \bar{\varepsilon}^{(S)}_{l,-r}\right\|^2 - \sigma^2 s\right| \\ \leq 4\bar{c}_{\text{conc}} L^2 \left(\log\left(\frac{n}{r\delta}\right) + s\log\left(\frac{2ep}{s}\right)\right) \Bigg\}\end{aligned} \tag{35}$$

has probability larger than $1-2\delta$.

Set now

$$\xi^{(\mathrm{p})} := \xi_1^{(\mathrm{p})} \cap \xi_2^{(\mathrm{p})}.$$

Note that

$$\mathbb{P}(\xi^{(\mathrm{p})}) \geq 1 - 3\delta.$$

Step 2: Study in the 'no change-point' situation. Consider $(l,r) \in J_n$ such that $\{\tau_k, k \in \{1,\dots,K\}\} \cap [l-r, l+r) = \varnothing$, and $S \subset \{1,\dots,p\}$. Note that since $\{\tau_k, k \in \{1,\dots,K\}\} \cap [l-r, l+r) = \varnothing$, we have $\bar{\theta}^{(S)}_{l,-r} = \bar{\theta}^{(S)}_{l,+r}$ so that

$$\frac{r}{2}\left\|\bar{\theta}^{(S)}_{l,-r} - \bar{\theta}^{(S)}_{l,+r}\right\|^2 = 0,$$

and

$$r\langle \bar{\varepsilon}^{(S)}_{l,+r} - \bar{\varepsilon}^{(S)}_{l,-r}, \bar{\theta}^{(S)}_{l,+r} - \bar{\theta}^{(S)}_{l,-r}\rangle = 0.$$

Moreover we have on $\xi^{(\mathrm{p})}$ that - see Equation (35)

$$\left|\frac{r}{2}\left\|\bar{\varepsilon}^{(S)}_{l,+r} - \bar{\varepsilon}^{(S)}_{l,-r}\right\|^2 - \sigma^2 s\right| \leq 4\bar{c}_{\text{conc}} L^2 \left(\log\left(\frac{n}{r\delta}\right) + s\log\left(\frac{2ep}{s}\right)\right) \leq \sigma^2 x_r^{(\mathrm{p})},$$

for $\bar{c}_{\text{thresh}} \geq 4\bar{c}_{\text{conc}}$ - note that $\bar{c}_{\text{conc}} > 0$ is a universal constant. And so

$$\Psi^{(\mathrm{p})}_{l,r} \leq x_r^{(\mathrm{p})},$$

so that on $\xi^{(d)}$,

$$T^{(\mathrm{p})}_{l,r} = 0 \ .$$

This concludes the proof of the first part of the proposition.

Step 3: Study in the 'change-point' situation. Consider $k \in \{1,\dots,K\}$ such that τ_k is a κ-sparse high-energy change-point, - see Equation (18). Since S_k is the support of $\mu_k - \mu_{k-1}$ - and therefore of $\bar{\theta}_{\tau_k, -\bar{r}_k^{(\mathrm{s})}} - \bar{\theta}_{\tau_k, +\bar{r}_k^{(\mathrm{s})}}$ - we have

$$\frac{\bar{r}_k^{(\mathrm{s})}}{2}\left\|\bar{\theta}^{(S_k)}_{\tau_k, -\bar{r}_k^{(\mathrm{s})}} - \bar{\theta}^{(S_k)}_{\tau_k, +\bar{r}_k^{(\mathrm{s})}}\right\|^2 \geq \frac{\kappa}{8} L^2 \left(s_k \log\left(\frac{2ep}{s_k}\right) + \log\left(\frac{n}{\bar{r}_k^{(\mathrm{s})}\delta}\right)\right) \ .$$

So on $\xi^{(\mathrm{p})}$ this implies that - see Lemma 15

$$\bar{r}_k^{(\mathrm{d})} \left| \langle \bar{\varepsilon}^{(S_k)}_{\tau_k,+\bar{r}_k^{(\mathrm{s})}} - \bar{\varepsilon}^{(S_k)}_{\tau_k,-\bar{r}_k^{(\mathrm{s})}}, \bar{\theta}^{(S_k)}_{\tau_k,+\bar{r}_k^{(\mathrm{s})}} - \bar{\theta}^{(S_k)}_{\tau_k,-\bar{r}_k^{(\mathrm{s})}} \rangle \right| \leq \frac{\bar{r}_k^{(\mathrm{s})}}{4} \left\| \overline{\theta}_{\tau_k,+\bar{r}_k^{(\mathrm{s})}} - \overline{\theta}_{\tau_k,-\bar{r}_k^{(\mathrm{s})}} \right\|^2 .$$

Moreover we have on $\xi^{(\mathrm{p})}$ that - see Equation (35)

$$\left| \frac{\bar{r}_k^{(\mathrm{s})}}{2} \left\| \bar{\varepsilon}^{(S_k)}_{\tau_k,+\bar{r}_k^{(\mathrm{s})}} - \bar{\varepsilon}^{(S_k)}_{\tau_k,-\bar{r}_k^{(\mathrm{s})}} \right\|^2 - \sigma^2 s \right| \leq 4\bar{c}_{\mathrm{conc}} L^2 \left(\log\left(\frac{n}{\bar{r}_k^{(\mathrm{s})}\delta} \right) + 2s_k \log\left(\frac{2ep}{s_k} \right) \right) \leq \sigma^2 x^{(\mathrm{p})}_{\bar{r}_k^{(\mathrm{s})}} \quad ,$$

for $\bar{c}_{\mathrm{thresh}} \geq 4\bar{c}_{\mathrm{conc}}$ - note that $\bar{c}_{\mathrm{conc}} > 0$ is a universal constant. And so on $\xi^{(\mathrm{p})}$, combining the three previous displayed equations implies

$$\begin{aligned} \Psi^{(\mathrm{p})}_{\tau_k,\bar{r}_k^{(\mathrm{s})}} &\geq \frac{\bar{r}_k^{(\mathrm{d})}}{4\sigma^2} \left\| \bar{\theta}^{(S_k)}_{\tau_k,+\bar{r}_k^{(\mathrm{s})}} - \bar{\theta}^{(S_k)}_{\tau_k,-\bar{r}_k^{(\mathrm{s})}} \right\|^2 - x^{(\mathrm{p})}_{\bar{r}_k^{(\mathrm{s})}} \\ &\geq \left(\frac{\kappa}{16} - \bar{c}_{\mathrm{thresh}} \right) \frac{L^2}{\sigma^2} \left(\log\left(\frac{n}{\bar{r}_k^{(\mathrm{s})}\delta} \right) + s_k \log\left(\frac{2ep}{s_k} \right) \right) > x^{(\mathrm{p})}_{\bar{r}_k^{(\mathrm{s})}} \quad , \end{aligned}$$

since $\kappa > 32\bar{c}_{\mathrm{thresh}}$. And so on $\xi^{(\mathrm{p})}$

$$T^{(\mathrm{p})}_{\tau_k,\bar{r}_k^{(\mathrm{s})}} = 1.$$

This concludes the proof of the second part of the proposition.

B.3.3 Proof of Corollary 5

Let $\xi^{(\mathrm{d})}$ and $\xi^{(\mathrm{s})}$ be two events such that Proposition 3 and Proposition 4 both hold with probability $1-3\delta$, and write $\xi = \xi^{(\mathrm{d})} \cap \xi^{(\mathrm{p})}$. From now on, we work on the event ξ, which holds with probability $1-6\delta$. Define here simply $\bar{\tau}_k = \tau_k$. Note that by definition of $\bar{r}_k$ in the sub-Gaussian regime:

$$\bar{r}_k = \begin{cases} \bar{r}_k^{(\mathrm{d})} \text{ if } s_k \log\left(\frac{ep}{s_k}\right) > \sqrt{p \log\left(\frac{n}{r_k\delta}\right)} \\ \bar{r}_k^{(\mathrm{s})} \text{ if } s_k \log\left(\frac{ep}{s_k}\right) \leq \sqrt{p \log\left(\frac{n}{r_k\delta}\right)} \end{cases}$$

According to Theorem 1, it is sufficient to prove that $\mathcal{A}(\Theta, T, \mathcal{K}^*, (\bar{\tau}_k, \bar{r}_k)_{k\in\mathcal{K}^*})$ holds.

1. **(No False Positive):** $T_{l,r} = T^{(\mathrm{p})}_{l,r} \vee T^{(\mathrm{d})}_{l,r} = 0$ for any $(l,r) \in \mathcal{G}_F \cap \mathcal{H}_0$. by Proposition 3 and Proposition 4.

2. **(Significant change-point Detection):** for every $k \in \mathcal{K}^*$ (see (20)), we have by definition of $\bar{r}_k$:

   $$4(\bar{r}_k - 1) \leq r_k.$$

   Now if $s_k \log\left(\frac{ep}{s_k}\right) \geq \sqrt{p \log\left(\frac{n}{r_k\delta}\right)}$, we have $T^{(\mathrm{d})}_{\bar{\tau}_k,\bar{r}_k} = 1$ by Proposition 3, by definition of c_0, and for $\bar{c}^{(d)}_{\mathrm{thresh}}$ as in Proposition 3.

   If $s_k \log\left(\frac{ep}{s_k}\right) \leq \sqrt{p \log\left(\frac{n}{r_k\delta}\right)}$, we have $T^{(\mathrm{p})}_{\bar{\tau}_k,\bar{r}_k} = 1$ by Proposition 4, by definition of c_0, and for $\bar{c}^{(p)}_{\mathrm{thresh}}$ as in Proposition 4.

Theorem 1 ensures that for all $k \in \mathcal{K}^*$, there exists $k' \in \{1, \ldots, \hat{K}\}$ such that

$$|\hat{\tau}_{k'} - \tau_k| \leq \bar{r}_k - 1.$$

This concludes the proof since $4(\bar{r}_k - 1) \leq r_k$ for $k \in \mathcal{K}^*$.

B.4 Proof of Theorem 2

Let us fix $(r,s) \in [1, n/4] \times [1,p]$. Let Δ be such that

$$r\Delta^2 = \frac{1}{2}\sigma^2\left[s\log\left(1+u\frac{\sqrt{p}}{s}\sqrt{\log\left(\frac{n}{r}\right)}\right)+u\log\left(\frac{n}{r}\right)\right],$$

for some $u \leq \frac{1}{8}$.

In what follows, we consider any change-point detection method that outputs an estimator $\hat{\tau}$ of the change-points, associated to a number $\hat{K}$ of detected change-points, i.e. the length of $\hat{\tau}$. We also write $\mathbb{P}_\Theta$ for the distribution of the data when the mean parameter or the time series is fixed to a $n \times p$ matrix Θ, i.e. of $\Theta + \varepsilon$ where the noise entries $(\varepsilon_t)_j$ are i.i.d. and follow $\mathcal{N}(0,\sigma^2)$ as in Section 3. Also abusing slightly notations, we write $\mathbb{P}_0$ for the distribution of the data when the parameter is constant and equal to 0.

Consider also any prior π over the set of $n \times p$ matrices Θ such that the number of true change-points over the support of the prior is larger than 1 - i.e. the prior puts mass only on problems where more than one change-point occurs. Let $\bar{\mathbb{P}}_\pi$ be the corresponding distribution of the data, namely the distribution of the matrix of data when the mean parameter of the time series is the random matrix $\tilde{\Theta} \sim \pi$. Otherwise said, $\bar{\mathbb{P}}_\pi$ is the distribution of $\tilde{\Theta} + \varepsilon$ where $\tilde{\Theta} \sim \pi$.

We remind that in our setting K is the number of true change-points in a given problem - which would be either 0 under $\mathbb{P}_0$, or more than 1 under $\bar{\mathbb{P}}_\pi$. If the support of π_1 is included in $\mathcal{P}(r,s)$, then

$$\begin{aligned}\sup_{\Theta \in \mathcal{P}(r,s)} \mathbb{P}_\Theta(\hat{K} \neq K) &\geq \frac{1}{2}\left(\bar{\mathbb{P}}_\pi(\hat{K}=0) + \mathbb{P}_0(\hat{K}\neq 0)\right)\\ &\geq \frac{1}{2}(1 - d_{TV}(\bar{\mathbb{P}}_\pi, \mathbb{P}_0)),\end{aligned} \tag{36}$$

where d_{TV} is the total variation distance. From the Cauchy-Schwarz inequality, we have

$$d_{TV}(\bar{\mathbb{P}}_\pi, \mathbb{P}_0) \leq \frac{1}{2}\sqrt{\chi^2(\bar{\mathbb{P}}_\pi, \mathbb{P}_0)}, \tag{37}$$

where χ^2 is the divergence between probability distributions:

$$\chi^2(\bar{\mathbb{P}}_\pi, \mathbb{P}_0) = \mathbb{E}_{\mathbb{P}_0}\left[\left(\frac{\mathrm{d}\bar{\mathbb{P}}_\pi}{\mathrm{d}\,\mathbb{P}_0} - 1\right)^2\right].$$

By a simple computation that can be found for example in [Wu17]

$$\chi^2(\bar{\mathbb{P}}_\pi, \mathbb{P}_0) = \mathbb{E}_{\tilde{\Theta},\tilde{\Theta}}\left[e^{\frac{1}{\sigma^2}\langle\tilde{\Theta},\tilde{\Theta}\rangle}\right] - 1, \tag{38}$$

where $\tilde{\Theta}$ and $\tilde{\Theta}'$ are i.i.d. and distributed according to π, $\langle\tilde{\Theta},\tilde{\Theta}'\rangle = \mathrm{Tr}(\tilde{\Theta}'\tilde{\Theta}^T)$ is the standard scalar product, and $\mathbb{E}_{\tilde{\Theta},\tilde{\Theta}'}$ is the expectation according to $\tilde{\Theta}$ and $\tilde{\Theta}'$.

Let us consider the three following cases for the couple (r,s):

$$\begin{aligned}&\textbf{Case 1}: u\log\left(\frac{n}{r}\right) \leq s\log\left(1+u\frac{\sqrt{p}}{s}\sqrt{\log\left(\frac{n}{r}\right)}\right) \quad \text{and} \quad s \leq u\sqrt{p\log\left(\frac{n}{r}\right)},\\ &\textbf{Case 2}: u\log\left(\frac{n}{r}\right) \leq s\log\left(1+u\frac{\sqrt{p}}{s}\sqrt{\log\left(\frac{n}{r}\right)}\right) \quad \text{and} \quad s > u\sqrt{p\log\left(\frac{n}{r}\right)},\\ &\textbf{Case 3}: u\log\left(\frac{n}{r}\right) > s\log\left(1+u\frac{\sqrt{p}}{s}\sqrt{\log\left(\frac{n}{r}\right)}\right).\end{aligned}$$

Each case corresponds to the regime of detection of one of the three statistics. The first one corresponds to the Berk-Jones statistic, the second one to the dense statistic and the last one to the max statistic.

Case 1 : In that case, $r\Delta^2 \leq \sigma^2 s \log\left(4u\frac{p}{s^2}\log\left(\frac{n}{r}\right)\right)$. Let us define a probability distribution on the parameter $\Theta \in \mathcal{P}(r,s)$. For $\zeta = \left\lfloor \frac{n}{r} \right\rfloor - 1$ and $l \in \tilde{\mathcal{D}}_r = \{1, r+1, 2r+1, \dots \zeta r + 1\}$, define the column vector $v_l = \sum_{j=l}^{l+r-1} e_j$, where e_j is the j^{th} element of the canonical basis of $\mathbb{R}^n$. Let a be a random variable uniformly distributed in $\{x \in \{0,1\}^p, |x|_0 = s\}$ and ν be a random variable independent from a and uniformly distributed on $\{v_l \ : \ l \in \tilde{\mathcal{D}}_r\}$. Let

$$\tilde{\Theta}_{(1)} = \frac{\Delta}{\sqrt{s}} a\nu^T \in \mathbb{R}^{p\times n},$$

and π_1 be the distribution of the random variable $\tilde{\Theta}_{(1)}$, and $\bar{\mathbb{P}}_{\pi_1}$ be the corresponding distribution of the data.

Consider two independent copies $\tilde{\Theta}_{(1)}$ and $\tilde{\Theta}'_{(1)}$ that are distributed like π_1. The probability that $\tilde{\Theta}_{(1)}$ and $\tilde{\Theta}'_{(1)}$ have the same support is exactly $\frac{1}{\zeta+1}$. Hence, from Equation (38)

$$\chi^2(\bar{\mathbb{P}}_{\pi_1}, \mathbb{P}_0) = \frac{1}{\zeta+1}\left(\mathbb{E}_{a,a'}\left[e^{\frac{r\Delta^2}{s\sigma^2}\langle a,a'\rangle} - 1\right]\right) \ , \tag{39}$$

where a' is an independent copy of a, and $\mathbb{E}_{a,a'}$ is the expectation according to a, a'. Remark by symmetry that $\langle a, a'\rangle$ has the same law as $\sum_{i=1}^s a_i$. Hence

$$\mathbb{E}_{a,a'}\left[e^{\frac{r\Delta^2}{s\sigma^2}\langle a,a'\rangle}\right] = \mathbb{E}_a\left[e^{\frac{r\Delta^2}{s\sigma^2}\sum_{i=1}^s a_i}\right] \ ,$$

where $\mathbb{E}_a$ is the expectation according to a.

Remark that $(a_1, \dots, a_p)$ has the same distribution as a random sampling without replacement of the list of length p containing $(1, \dots, 1, 0, \dots, 0)$ - the list containing exactly s times the quantity 1 and otherwise only 0. The following lemma allows us to replace the variables a_i by independent Bernoulli random variables $Z_i \sim \mathcal{B}(s/p)$.

Lemma 17. *Let $c = (c_1, \dots, c_p) \in \mathbb{R}^p$. We associate to the list c two random sampling processes: (i) the sampling process without replacement $(X_i)_{i=1\dots s}$ of s elements uniformly on the list c and (ii) the sampling process with replacement $(Z_i)_{i=1\dots s}$ of s elements uniformly in the list. Then for any convex function f,*

$$\mathbb{E}\left[f\left(\sum_{i=1}^s X_i\right)\right] \leq \mathbb{E}\left[f\left(\sum_{i=1}^s Z_i\right)\right] \ .$$

The proof of this lemma can be found in [Hoe94]. Thus, if $(Z_i)_{i=1\dots s}$ is an i.i.d sequence of Bernoulli variables with parameter $\frac{s}{p}$ as described above, we obtain

$$\begin{aligned}
\chi^2(\bar{\mathbb{P}}_{\pi_1}, \mathbb{P}_0) &\leq \frac{1}{\zeta+1}\left(\mathbb{E}_Z\left[e^{\frac{r\Delta^2}{s\sigma^2}\sum_{i=1}^s Z_i}\right] - 1\right) \qquad (40)\\
&= \frac{1}{\zeta+1}\left[\left(\frac{s}{p}e^{\frac{r\Delta^2}{s\sigma^2}} + 1 - \frac{s}{p}\right)^s - 1\right] \leq \frac{1}{\zeta+1}\left[e^{\frac{s^2}{p}\left(e^{\frac{r\Delta^2}{s\sigma^2}}-1\right)} - 1\right]\\
&\leq 2\frac{r}{n}e^{\frac{s^2}{p}\left(e^{\log\left(4u^2\frac{p}{s^2}\log\left(\frac{n}{r}\right)\right)}\right)} \leq 2\left(\frac{r}{n}\right)^{1-4u^2} \leq 1 \ , \qquad (41)
\end{aligned}$$

where $\mathbb{E}_Z$ is the expectation according to the $(Z_i)_i$ and where in the last inequality we used $u \leq 1/3$ and $n \geq 4r$.

Case 2 : In that case, $r\Delta^2 \leq \sigma^2 u\sqrt{p\log\left(\frac{n}{r}\right)}$. Let $s_0 = \left\lceil u\sqrt{p\log\left(\frac{n}{r}\right)}\right\rceil$ and b be a random variable uniformly distributed in $\{x \in \{0,1\}^p, |x|_0 = s_0\}$ and ν be defined as in **Case 1**. Let

$$\tilde{\Theta}_{(2)} = \frac{\Delta}{\sqrt{p}} b\nu^T,$$

let π_2 be the distribution of $\tilde{\Theta}_{(2)}$ and $\bar{\mathbb{P}}_{\pi_2}$ be the associated probability distribution of the data. Doing the same reasoning and similar computations as for **Case 1**, see in particular the steps of Equations (39) and (40) - replacing s by s_0 and a by b - we have

$$\chi^2(\bar{\mathbb{P}}_{\pi_2}, \mathbb{P}_0) = \mathbb{E}_{\tilde{\Theta}_{(2)}, \tilde{\Theta}'_{(2)}}\left[e^{\frac{1}{\sigma^2}\langle \tilde{\Theta}_{(2)}, \tilde{\Theta}'_{(2)}\rangle}\right] - 1 = \frac{1}{\zeta+1}\mathbb{E}_{b,b'}\left[e^{\frac{r\Delta^2}{p\sigma^2}\langle b, b'\rangle} - 1\right] \leq \frac{1}{\zeta+1}\left[e^{\frac{s_0^2}{p}\left(e^{\frac{r\Delta^2}{s_0\sigma^2}}-1\right)} - 1\right]$$

$$\leq \frac{1}{\zeta+1}e^{2\frac{s_0 r\Delta^2}{p\sigma^2}} \leq 2\frac{r}{n}e^{4u\log\frac{n}{r}} = 2\left(\frac{r}{n}\right)^{1-4u} \leq 1 \ , \tag{42}$$

where $\mathbb{E}_{\tilde{\Theta}_{(2)}, \tilde{\Theta}'_{(2)}}$ is the expectation according to $\tilde{\Theta}_{(2)}, \tilde{\Theta}'_{(2)}$ (where $\tilde{\Theta}'_{(2)}$ is an independent copy of $\tilde{\Theta}_{(2)}$) and where $\mathbb{E}_{b,b'}$ is the expectation according to b, b' (where b' is an independent copy of b), and where in the last step we used $u \leq 1/8$ and $n \geq 4r$.

Case 3 : In that case, $r\Delta^2 \leq u\log\left(\frac{n}{r}\right)$. Let $c = (1, 0, 0, \ldots, 0)$ be the vector with 0 entries except the first one. Let ν be the random vector defined as in **Case 1**. Let

$$\tilde{\Theta}_{(3)} = \Delta c\nu^T,$$

and π_3 be the distribution of the random variable $\tilde{\Theta}_{(3)}$ - and $\bar{\mathbb{P}}_{\pi_3}$ be the associated probability distribution of the data. Doing the same reasoning as in **Case 1** - see in particular the step of Equation (39) - replacing a by c and s by 1 - for the prior π_3, we obtain

$$\chi^2(\bar{\mathbb{P}}_{\pi_3}, \mathbb{P}_0) = \mathbb{E}_{\tilde{\Theta}_{(3)}, \tilde{\Theta}'_{(3)}}\left[e^{\frac{1}{\sigma^2}\langle \tilde{\Theta}_{(3)}, \tilde{\Theta}'_{(3)}\rangle}\right] - 1 = \frac{1}{\zeta+1}e^{\frac{r\Delta^2}{\sigma^2}} \leq 2\frac{r}{n}e^{u\log\left(\frac{n}{r}\right)} \leq 2\left(\frac{r}{n}\right)^{1-u} \leq 1 \ , \tag{43}$$

where $\mathbb{E}_{\tilde{\Theta}_{(3)}, \tilde{\Theta}'_{(3)}}$ is the expectation according to $\tilde{\Theta}_{(3)}, \tilde{\Theta}'_{(3)}$ (where $\tilde{\Theta}'_{(3)}$ is an independent copy of $\tilde{\Theta}_{(3)}$) and where in the last step we used $n \geq 4r$ and $u \leq 1/2$.

Thus, in all cases - combining Equations (36) and (37) with Equations (41), (42) and (43) - we obtain in all three cases

$$\sup_{\Theta \epsilon \mathcal{P}(r,s)} \mathbb{P}_\Theta(\hat{K} \neq K) \geq \frac{1}{4} \ .$$

and this concludes the proof.

References

[ACH19] Sylvain Arlot, Alain Celisse, and Zaid Harchaoui. A kernel multiple change-point algorithm via model selection. *J. Mach. Learn. Res.*, 20:Paper No. 162, 56, 2019.

[CC+19] Lynna Chu, Hao Chen, et al. Asymptotic distribution-free change-point detection for multivariate and non-euclidean data. *The Annals of Statistics*, 47(1):382–414, 2019.

[CK19] Haeran Cho and Claudia Kirch. Localised pruning for data segmentation based on multiscale change point procedures. *arXiv preprint arXiv:1910.12486*, 2019.

[DJ04] David Donoho and Jiashun Jin. Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.*, 32(3):962–994, 2004.

[EH19] Farida Enikeeva and Zaid Harchaoui. High-dimensional change-point detection under sparse alternatives. *Ann. Statist.*, 47(4):2051–2079, 2019.

[FMS14] Klaus Frick, Axel Munk, and Hannes Sieling. Multiscale change point inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(3):495–580, 2014.

[Fry14] Piotr Fryzlewicz. Wild binary segmentation for multiple change-point detection. *The Annals of Statistics*, 42(6):2243–2281, 2014.

[Fry18] Piotr Fryzlewicz. Tail-greedy bottom-up data decompositions and fast multiple change-point detection. *The Annals of Statistics*, 46(6B):3390–3421, 2018.

[GA18] Damien Garreau and Sylvain Arlot. Consistent change-point detection with kernels. *Electron. J. Stat.*, 12(2):4440–4486, 2018.

[GBR+12] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

[GR52] Meyer A. Girshick and Herman Rubin. A bayes approach to a quality control model. *The Annals of Mathematical Statistics*, 23(1):114–125, 1952.

[GR17] Alex J Gibberd and Sandipan Roy. Multiple changepoint estimation in high-dimensional gaussian graphical models. *arXiv preprint arXiv:1712.05786*, 2017.

[Hoe94] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. In *The Collected Works of Wassily Hoeffding*, pages 409–426. Springer, 1994.

[Jir15] Moritz Jirak. Uniform change point tests in high dimension. *The Annals of Statistics*, 43(6):2451–2483, 2015.

[KLBM20] Solt Kovács, Housen Li, Peter Bühlmann, and Axel Munk. Seeded binary segmentation: A general methodology for fast and optimal change point detection. *arXiv preprint arXiv:2002.06633*, 2020.

[LGS19] Haoyang Liu, Chao Gao, and Richard J Samworth. Minimax rates in sparse, high-dimensional changepoint detection. *arXiv preprint arXiv:1907.10012*, 2019.

[MNS+16] Amit Moscovich, Boaz Nadler, Clifford Spiegelman, et al. On the exact berk-jones statistics and their p-value calculation. *Electronic Journal of Statistics*, 10(2):2329–2354, 2016.

[NHZ16] Yue S Niu, Ning Hao, and Heping Zhang. Multiple change-point detection: A selective overview. *Statistical Science*, 31(4):611–623, 2016.

[OVLW04] Adam B Olshen, ES Venkatraman, Robert Lucito, and Michael Wigler. Circular binary segmentation for the analysis of array-based dna copy number data. *Biostatistics*, 5(4):557–572, 2004.

[PYWR19] Oscar Hernan Madrid Padilla, Yi Yu, Daren Wang, and Alessandro Rinaldo. Optimal nonparametric multivariate change point detection and localization. *arXiv preprint arXiv:1910.13289*, 2019.

[RV+13] Mark Rudelson, Roman Vershynin, et al. Hanson-wright inequality and sub-gaussian concentration. *Electronic Communications in Probability*, 18, 2013.

[TOV20] Charles Truong, Laurent Oudre, and Nicolas Vayatis. Selective review of offline change point detection methods. *Signal Processing*, 167:107299, 2020.

[Ver18] Roman Vershynin. *High-Dimensional Probability*. 2018.

[VFLRB20] Nicolas Verzelen, Magalie Fromont, Matthieu Lerasle, and Patricia Reynaud-Bouret. Optimal Change-Point Detection and Localization. *arXiv preprint arXiv:2010.11470*, 2020.

[Wal45] Abraham Wald. Sequential tests of statistical hypotheses. *The annals of mathematical statistics*, 16(2):117–186, 1945.

[WS18] Tengyao Wang and Richard J. Samworth. High dimensional change point estimation via sparse projection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 80(1):57–83, 2018.

[WS20] Runmin Wang and Xiaofeng Shao. Dating the break in high-dimensional data. *arXiv preprint arXiv:2002.04115*, 2020.

[Wu17] Yihong Wu. Lecture notes for ece598yw: Information-theoretic methods for high-dimensional statistics, 2017.

[WVS19] Runmin Wang, Stanislav Volgushev, and Xiaofeng Shao. Inference for change points in high dimensional data. *arXiv preprint arXiv:1905.08446*, 2019.

[WYR17] Daren Wang, Yi Yu, and Alessandro Rinaldo. Optimal covariance change point localization in high dimension. *arXiv preprint arXiv:1712.09912*, 2017.

[WYR18] Daren Wang, Yi Yu, and Alessandro Rinaldo. Univariate mean change point detection: Penalization, cusum and optimality. *arXiv preprint arXiv:1810.09498*, 2018.

[WYRW19] Daren Wang, Yi Yu, Alessandro Rinaldo, and Rebecca Willett. Localizing changes in high-dimensional vector autoregressive processes. *arXiv preprint arXiv:1909.06359*, 2019.

[YC17] Mengjia Yu and Xiaohui Chen. Finite sample change point inference and identification for high-dimensional mean vectors. *arXiv preprint arXiv:1711.08747*, 2017.