High-dimensional Online Changepoint Detection

Yudong Chen

Statistical Laboratory
Department of Pure Mathematics and Mathematical Statistics
University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

Peterhouse

November 2021

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University of similar institution. It does not exceed the prescribed word limit for the Mathematics Degree Committee.

Chapters 2 and 3 are joint work with Tengyao Wang (University College London; London School of Economics and Political Science) and Richard Samworth (University of Cambridge). Chapter 2 has been published in the *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* as Chen, Wang and Samworth (2022). Chapter 3 has been submitted for publication as Chen, Wang and Samworth (2021).

Yudong Chen
November 2021

Abstract

High-dimensional Online Changepoint Detection

Yudong Chen

The problem of changepoint detection and estimation has a long history, dating back to at least Page (1954, 1955). As modern technological advances, data sets of unprecedented size can be collected at high frequency. This provides statisticians with new challenges in this field. In this thesis we study the online version of the changepoint detection problem in high-dimensional settings. In Chapter 1, we survey the field of changepoint detection. We focus, in particular, on the comparison between the offline problem and the online problem, the contrast between the univariate setting and the high-dimensional setting and inference problems associated with changepoints.

In Chapter 2, we introduce a novel method for high-dimensional, online changepoint detection in settings where a multivariate data stream may undergo a change in mean. The procedure works by performing Gaussian likelihood ratio tests against simple alternatives of different scales in each coordinate, and then aggregating test statistics across scales and coordinates. Our algorithm is online in the sense that both its storage requirements and worst-case computational complexity per new observation are independent of the number of previous observations. We prove that the patience, or average run length under the null, of our procedure is at least at the desired nominal level, and provide guarantees on its response delay under the alternative that depend on the sparsity of the vector of mean change. Our procedure shows excellent performance compared to existing methods in the numerical studies. Our algorithm is implemented in the `R` package `ocd`, and we also demonstrate its utility on a seismology data set.

In Chapter 3, we focus on the problem of inference for high-dimensional online changepoint detection. We propose a confidence interval for the changepoint location. The procedure first identifies coordinates with large signals and then combines univariate confidence intervals constructed from each of these coordinates. We prove that the confidence interval constructed has the desired coverage level and provide a guarantee on the length of the confidence interval. Our procedure also provides an estimate for the effective support of the signal as a byproduct. Simulations confirm the practical effectiveness of our proposal, and we also illustrate its applicability on both US excess deaths data from 2017–2020 and S&P 500 data from the 2007–2008 financial crisis.

Acknowledgements

First and foremost, I would like to thank my supervisors Professor Richard Samworth and Dr. Tengyao Wang for their guidance and support throughout my PhD. During my undergraduate studies, I did a summer research and wrote a Part III essay with them, and was fascinated by topics such as changepoints and data-perturbation techniques. These exciting experiences inspired me to pursue a PhD in the field of statistics. I have learned a lot not only about statistics but also about how to do academic research and how to become a better person from them in the last few years. I am extremely grateful for their generosity of time, expertise, guidance and feedback.

It has been a real pleasure to be a member of Richard's (extended) research group, which includes Tom Berrett, Tim Cannings, Yining Chen, Oliver Feng, Bertille Follain, Milana Gataric, Jana Janková, Ilmun Kim, Jing Lei, Anton Lundborg, Henry Reeve, Philip Thompson, Min Xu, Yachong Yang, Yoav Zemel and Ziwei Zhu. I have always enjoyed that ninety minutes every Tuesday afternoon when we get together (remotely) and discuss fascinating topics in statistics. I would also like to thank Idris Eckley, Paul Fearnhead and Hyeyoung Maeng for inviting me to attend and give a presentation at the StatScale seminar series and allowing me to know more people and their works in the changepoint research community.

I gratefully acknowledge the CCIMI for providing funding to my PhD, without which the work in this thesis would not have been possible. I would also like to thank my graduate advisor in the department Rajen Shah, as well as Saskia Murk Jansen and András Zsák for their support at Peterhouse.

Thanks also to my friends and colleagues in the CMS for making it a friendly and happy place to work in before the start of the COVID-19 pandemic, including Adam Goucher, Florian Pein, Ben Stokell, Xiaoyu Wang, Yuhao Wang and Wanlong Zheng. I am also deeply grateful to Daren Chen, Sheng Gao and Li Hua for their support and encouraging words in the last few years, and to Xuan Guo and Qiujia Li for forming a support bubble with me during the difficult and lonely lockdown times.

Finally, I would like to thank my family for all their love and encouragement over many many years, and especially during the last twenty months. It would not have been possible for me to be in today's position without them. Thank you!

Contents

Chapter 1

Introduction

Modern technology has not only allowed the collection of data sets of unprecedented size, but has also facilitated the real-time monitoring of many types of evolving processes of interest. Wearable health devices, astronomical survey telescopes, self-driving cars and transport network load-tracking systems are just a few examples of new technologies that collect large quantities of streaming data, and that provide new challenges and opportunities for statisticians.

Very often, a key feature of interest in the monitoring of a data stream is a *changepoint*; that is, a moment in time at which the data generating mechanism undergoes a change. Such times often represent events of interest, e.g. a change in heart function, and moreover, the accurate identification of changepoints often facilitates the decomposition of a data stream into stationary segments. Applications of changepoints include service attacks in Internet traffic monitoring (Peng, Leckie and Ramamohanarao, 2004), stock price movements in financial markets (Chen and Gupta, 1997), and blood oxygen level response change in functional Magnetic Resonance Imaging (fMRI) (Aston and Kirch, 2012).

Historically, it has tended to be univariate time series that have been monitored and studied, within the well-established field of statistical process control (e.g. Duncan, 1952; Page, 1954; Barnard, 1959; Oakland, 2007; Tartakovsky, Nikiforov and Basseville, 2014). More efficient algorithms for changepoint detection in univariate settings have been proposed and analysed in recent years (e.g. Fearnhead and Liu, 2007; Killick, Fearnhead and Eckley, 2012; Frick, Munk and Sieling, 2014; Fryzlewicz, 2014; Baranowski, Chen and Fryzlewicz, 2019; Wang, Yu and Rinaldo, 2020).

These days, however, it is frequently the case that many data processes are measured simultaneously. In the context of changepoint detection, this introduces the new challenge of borrowing strength across the different component series in an attempt to detect much smaller changes than would be possible through the observation of any individual series alone. The last 5-10 years have seen an increasing amount of works which study the changepoint problem under multivariate or high-dimensional settings. A large majority of these works have been focusing on the retrospective challenges of detecting and estimating changes after seeing all of the available data (e.g. Chan and Walther, 2015; Cho and Fryzlewicz, 2015; Jirak, 2015; Cho, 2016; Wang and Samworth, 2018; Enikeeva and Harchaoui, 2019; Kaul et al., 2021a; Liu, Gao and Samworth, 2021; Londschien, Kovács and Bühlmann, 2021; Padilla et al., 2021b; Rinaldo et al., 2021; Follain, Wang and Samworth, 2022).

Instead of working on the entire dataset in a retrospective way, one can also observe data sequentially and seek to declare changes as soon as possible after they have occurred. In fact, this often turns out to be a more natural approach in real world scenarios. For example, governments need to make public health decisions based on daily-reported COVID-19 case numbers and investors need to make trading decisions based on real-time market movements. This sequential/online approach to the changepoint problem is nowadays receiving increasing attention (e.g. Tartakovsky et al., 2006; Mei, 2010; Xie and Siegmund, 2013; Zou et al., 2015; Chan, 2017; Soh and Chandrasekaran, 2017; Kirch and Stoehr, 2019; Dette and Gösmann, 2020; Gösmann et al., 2020; Yu et al., 2020). Sequential changepoint detection has also been studied in the econometrics literature as well, where the problem is often referred to as that of monitoring structural breaks (Chu, Stinchcombe and White, 1996; Leisch, Hornik and Kuan, 2000; Zeileis et al., 2005).

In this thesis, we focus on a high-dimensional, online changepoint detection problem. We propose a novel method for this problem in Chapter 2. An *online detection procedure* is naturally sequential, but furthermore, the computational complexity for processing a new observation, as well as the storage requirements, can depend only on the number of bits needed to represent the new observation[1]. Importantly, they are not allowed to depend on the number of previously observed data points. This turns out to be a very stringent requirement, in the sense that finding online algorithms with good statistical performance is typically extremely challenging. Online algorithms must necessarily store only compact summaries of the historical observations, so the class of all possible procedures is severely restricted. In Chapter 3, we introduce and study two new inferential challenges associated with the sequential detection of change in a high-dimensional mean vector. First, we seek a confidence interval for the changepoint, and second, we estimate the set of indices of coordinates in which the mean changes. We propose an online algorithm that achieves these two goals.

In order for readers to better understand the context of this thesis, we provide a literature review on the entire field of changepoint detection in the rest of this chapter. We first survey the well-studied field of offline changepoint detection, both univariate and high-dimensional. We will mainly focus on the mean change problem. We then move on to introduce some basic concepts and performance measures in the sequential problem, as well as some well-known log-

[1] For the purpose of this definition, we ignore the errors in rounding real numbers to machine precision. Thus, when we later work with observations having Gaussian (or other absolutely continuous) distributions, we do not distinguish between these distributions and quantised versions where the data have been rounded to machine precision.

likelihood based control charts. These procedures are often useful when pre- and post-change parameters are known exactly. Finally, we provide a summary of some recent developments in the (high-dimensional) sequential/online changepoint detection problem.

Let $X_1, X_2, \ldots$ be a sequence of random vectors in $\mathbb{R}^p$. In most of this chapter, we will assume no temporal dependence, i.e. the data points are independent. Below, we assume that $F_0, F_1, \ldots$ are distributions with densities $f_0, f_1, \ldots$ correspondingly, with respect to some common base measure. For the single changepoint problem, in the offline setting, we have access to the entire data set up to time n and assume $X_1, \ldots, X_z \overset{\text{iid}}{\sim} F_0$ and $X_{z+1}, \ldots, X_n \overset{\text{iid}}{\sim} F_1$. A corresponding multiple changepoint model assumes that there exist ν (an unknown number) changepoints: $1 \leq z_1 < z_2 < \ldots < z_\nu \leq n-1$ such that $X_{z_j+1}, \ldots, X_{z_{j+1}} \overset{\text{iid}}{\sim} F_j$ for $0 \leq j \leq \nu$, where we define $z_0 = 0$ and $z_{\nu+1} = n$. In the sequential/online setting, we observe data points one at a time and assume $X_1, \ldots, X_z \overset{\text{iid}}{\sim} F_0$ and $X_{z+1}, X_{z+2}, \ldots \overset{\text{iid}}{\sim} F_1$. In later sections, we will consider simpler versions of these general formulations under different settings.

1.1 Offline changepoint problem

In the offline version of the changepoint problem, we have access to the entire dataset and work retrospectively to detect and estimate changes. We first define the cumulative sum (CUSUM) statistic. The original version of the CUSUM was introduced by Page (1954) in the field of quality control and sequential setting and is simply defined to be $S_n := \sum_{i=1}^n X_i$ (see Section 1.2.1). Now, let (s, t, e) be any triple satisfying $0 \leq s < t < e \leq n$. We define the (offline) CUSUM statistic:

$$\mathcal{T}_t^{s,e} = \mathcal{T}_t^{s,e}(X) := \sqrt{\frac{e-t}{(e-s)(t-s)}} \sum_{i=s+1}^{t} X_i - \sqrt{\frac{t-s}{(e-s)(e-t)}} \sum_{i=t+1}^{e} X_i, \tag{1.1}$$

where $X = (X_1, \ldots, X_n)^T$. Note that if $X_1, \ldots, X_n$ are all univariate normal random variables with equal variance, then $\max_{1 \leq t \leq n-1} \big|\mathcal{T}_t^{0,n}\big|^2$ is the generalised likelihood ratio statistic for testing the null of no mean change against the alternative that there exists a mean shift within these n data points.

1.1.1 Univariate multiple changepoint problem

We now consider the following simple univariate mean shift model

$$X_i = \sum_{j=0}^{\nu} \mu^{(j)} \mathbb{1}_{\{z_j < i \leq z_{j+1}\}} + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where $0 = z_0 < z_1 < z_2 < \ldots < z_\nu < z_{\nu+1} = n$ and $\varepsilon_1, \ldots, \varepsilon_n \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. The mean follows a piecewise-constant structure with $\nu + 1$ segments. This is consistent with the general form stated earlier. We also note that the normality assumption on the noise can be relaxed in many cases. The task is to estimate the number of changepoints ν, as well as their locations.

In this univariate multiple changepoint problem, one of the most popular methods is Binary Segmentation (BS). BS was used in cluster analysis as early as in Scott and Knott (1974) and was first seen used in random processes in Vostrikova (1981). The BS algorithm for the changepoint problem first searches for the time point $\hat{t}$ which maximises $|\mathcal{T}_t^{0,n}|$. If $|\mathcal{T}_{\hat{t}}^{0,n}|$ does not exceed a certain threshold, then no changepoint is detected and the procedure stops. Otherwise, we can then split the data into two segments and work recursively. BS enjoys a low computational complexity (typically) of order $O(n \log n)$.

Whilst BS is simple to implement and computationally fast, it has been shown to be statistically sub-optimal in some scenarios. In particular, Fryzlewicz (2014) showed that if we have two changes close together and the mean is the same before the first and after the second changepoint, then BS can have almost no power in detecting the changes in scenarios where other methods would have power close to one in detecting both. This is because BS only adds changes one at a time, and this can mean that it struggles in a situation where multiple changes are present but there is little improvement in fit that is possible from fitting a single change.

To overcome this issue, Fryzlewicz (2014) proposed a Wild Binary Segmentation (WBS) procedure. Instead of using the entire dataset $X_1, \dots, X_n$ to identify the first changepoint, we randomly choose a number of sub-intervals $[s_m, e_m]_{m \in \{1,\dots,R\}}$. For each $m \in \{1, \dots, R\}$, we can find a time point $\hat{t}_m$ that maximises $|\mathcal{T}_t^{s_m,e_m}|$. We compare $\max_{m \in \{1,\dots,R\}} |\mathcal{T}_{\hat{t}_m}^{s_m,e_m}|$ with a certain threshold, with the corresponding time point $\hat{t}_m$ as the first changepoint location if the threshold is exceeded. The domain is then split into two sub-intervals, one to the left of $\hat{t}_m$ and one to the right. The recursion continues by applying the previous steps to each of these two intervals.

With these randomly drawn intervals, it is hopeful that one interval will contain one changepoint only, and the changepoint location is well separated from the interval endpoints. Under such scenario, the CUSUM estimator works very well in picking up the changepoint location. These random intervals allow us to localise CUSUM statistics and therefore enable WBS to overcome the shortcomings of BS discussed earlier. WBS, up to today, remains one of the state-of-the-art univariate changepoint methods and is also instructive when we are dealing with multiple changepoints in high-dimensional settings. Recently, Fryzlewicz (2020) proposed the 'Wild Binary Segmentation 2' and 'Steepest Drop to Low Levels' (WBS2.SDLL) which improves upon the WBS procedure.

Based on the general idea of WBS, the Narrowest-Over-Threshold (NOT) method was proposed by Baranowski, Chen and Fryzlewicz (2019). In WBS, we consider all intervals whose CUSUM statistic is above a threshold, order these by the magnitude of the statistic, recursively add a new change with the highest value of the statistic and then remove subsequent intervals which overlap with this change. In NOT, we do the same but order the intervals based on the width of the interval instead, i.e. we keep all intervals whose CUSUM is greater than the threshold but then add a change based on the one from the narrowest interval first. In some changepoint problems (e.g. slope changes in a piecewise-linear signal), if we estimate a single changepoint when there is actually more than one, we can obtain a large value of the test statistic, but the estimated changepoint location may be far away from the true ones. The idea of NOT is that by focusing on narrow intervals, we can minimise the chance that we estimate a changepoint from a data segment containing more than one changepoint. NOT could also be extended beyond the simple piecewise constant model. Kovács et al. (2020) proposed Seeded Binary Segmentation (SeedBS), which uses a deterministic construction of intervals instead of random ones. Padilla et al. (2021a) generalised the BS/WBS approach to nonparametric settings by using the CUSUM Kolmogorov–Smirnov statistic.

An alternative to the CUSUM-based methods is to use the moving sum (MOSUM) statistic (Hušková and Slabý, 2001; Eichinger and Kirch, 2018). The MOSUM statistic is defined as:

$$M_t := \frac{1}{\sqrt{2G}}\left|\sum_{i=t+1}^{t+G} X_i - \sum_{i=t+1-G}^{t} X_i\right|, \qquad G \leq t \leq n-G, \tag{1.2}$$

with bandwidth $G = G(n)$. We can identify changepoints when the MOSUM statistic exceeds a certain threshold. The method requires a pre-specified bandwidth, though this could be circumvented by appropriately merging changepoint candidates obtained from a few automatically chosen bandwidths. We also remark that this MOSUM procedure could be deemed as a pseudo-sequential procedure, as the changepoints are estimated one by one.

Another popular approach to the univariate multiple changepoint problem is to maximise a penalised log-likelihood (equivalently, under a normality assumption, to minimise a penalised least squares criterion). The penalty term is to prevent overfitting. Yao and Au (1989) first used least squares to estimate the changepoint locations and showed that the estimate is consistent when ν is known. Yao (1988) used the BIC/SIC (Schwarz, 1978) to estimate an unknown ν. Since then, there has been a vast literature studying various forms of the penalty term in this optimisation problem (e.g. Lavielle and Moulines, 2000; Pan and Chen, 2006).

One major concern of such an approach is the computational complexity. For example, the Optimal Partitioning (OP) procedure (Jackson et al., 2005) is an exact search method that solves the optimisation problem mentioned in the last paragraph with the penalty term equal to the number of changes. Equivalently, we minimise

$$\beta m - 2\sum_{i=0}^{m} \ell(X_{\zeta_i+1}, \ldots, X_{\zeta_{i+1}-1}),$$

over all $m \in \{0, 1, \ldots, n-1\}$ and $0 = \zeta_0 < \zeta_1 < \ldots < \zeta_m < \zeta_{m+1} = n$ for each m, where $\ell(\cdot)$ denotes the maximum log-likelihood for data in a segment (maximising out the segment parameter). Let $F(k)$ be the minimised objective function on data $X_1, \ldots, X_k$. Then by considering the position of the last change, it can be shown that

$$F(k) = \min_{0 \leq k' < k} \big\{ \beta + F(k') - 2\ell(X_{k'+1}, \ldots, X_k) \big\}, \tag{1.3}$$

for $k \in \{1, \ldots, n\}$, with $F(0) = -\beta$. Thus, dynamic programming can be used to find the value of $F(k)$ and the changepoint locations within $X_1, \ldots, X_k$ for each $k \in \{1, \ldots, n\}$ recursively. The set of estimated changepoints up to the k-th observation can be computed via $cp(k) := \{cp(k^*(k)), k^*(k)\}$, where $k^*(k)$ is a minimiser of (1.3). Then the set $cp(n)$ is the final output for all estimated changepoint locations. The computational complexity of this procedure is $O(n^2)$. Hence, OP is significantly slower than BS (and its variants), especially when n is large.

Based on the Optimal Partitioning idea, Killick, Fearnhead and Eckley (2012) proposed the Pruned Exact Linear Time (PELT) procedure. At the k-th step of the iteration in dynamic programming above, instead of calculating k values in (1.3), we remove from consideration those candidate time points k' that can never be a minimiser. This extra pruning step can reduce the computational complexity to $O(n)$ in best-case scenarios, but the worst-case computational time remains $O(n^2)$ when no pruning occurs. Faster algorithms than PELT for maximising a penalised log-likelihood were later proposed by Rigaill (2015) and Maidstone et al. (2017).

Another state-of-the-art method is the Simultaneous Multiscale Change Point Estimator (SMUCE) proposed by Frick, Munk and Sieling (2014). SMUCE minimises the number of changepoints over all possible right continuous step (regression) functions subject to a log-likelihood ratio based multiscale statistic being below a certain threshold. The optimisation simultaneously produces an estimate of the number of changepoints and a way to estimate the changepoint locations as well. Additionally, SMUCE gives confidence bands for the mean (as a step function between 1 and n) and confidence intervals for the changepoint locations. We shall defer the detailed discussion of this method to Section 1.3. Since the above constrained optimisation problem can again be converted to an unconstrained problem with a penalty term, dynamic programming with pruning can also be used to reduce the computational complexity.

1.1.2 High-dimensional offline changepoint problem

Before discussing the high-dimensional problem, we first mention some prior works that extended the univariate methods to multivariate (but not necessarily high-dimensional) settings. Ombao, von Sachs and Guo (2005) utilised the Smooth Localised Complex Exponentials (SLEX) model originally developed in one-dimensional random process in multivariate time series. Kirch, Muhsal and Ombao (2015) extended the change point test statistics developed by Hušková, Prášková and Steinebach (2015) to the vector autoregressive (VAR) model.

When analysing their methods, most of these works in multivariate settings assume a fixed dimension p as the total number of observations n grows. In high-dimensional settings, however, we can have both p and n being large, and we may assume a sparse 'signal', as in sparse linear models (see Wainwright, 2019, Chapter 7 for detailed discussion). The natural sparsity assumption here is that the change vector is sparse, i.e. the mean change only occurs in a sparse subset of all coordinates. This sparsity assumption also arises naturally from applications such as stock price movements (Chen and Gupta, 1997) and chromosomal copy number abnormality in bioinformatics (Bleakley and Vert, 2011).

In this section, we consider the following model

$$X_i \sim \mathcal{N}_p(\mu_i, \sigma^2 I_p), \qquad 1 \le i \le n,$$

where there exist $0 = z_0 < z_1 < z_2 < \ldots < z_\nu < z_{\nu+1} = n$ such that $\mu_{z_i+1} = \ldots = \mu_{z_{i+1}} := \mu^{(i)}$ for $0 \le i \le \nu$. We will be focusing on the single changepoint problem (i.e. $\nu = 1$), unless otherwise stated, since we can combine a method developed for this model with a top-down approach such as the WBS to locate multiple changepoints in the dataset (e.g. Wang and Samworth, 2018). There are two tasks associated with this problem. One is to test the null hypothesis that there is no change. The other is to estimate the changepoint location, if there exists one. In practice, in many cases, taking the maximising time point t of the test statistics(s) can yield a good changepoint location estimate.

We denote the vector of change $\theta := \big(\mu_1^{(1)} - \mu_1^{(0)}, \ldots, \mu_p^{(1)} - \mu_p^{(0)}\big)$ and its magnitude $\vartheta := \|\theta\|_2$. We define the matrix $X := (X_1, X_2, \ldots, X_n) \in \mathbb{R}^{p\times n}$ and define its CUSUM matrix $\mathcal{T} = \mathcal{T}(X) \in \mathbb{R}^{p\times(n-1)}$ by:

$$[\mathcal{T}(X)]_{j,t} := \mathcal{T}_t^{0,n}(X_{j,\cdot}), \qquad j \in \{1, \ldots, p\} \text{ and } t \in \{1, \ldots, n-1\},$$

where $X_{j,\cdot}$ denotes the jth row of X and the right-hand side is defined in (1.1). Under the null, we have

$$(\mathcal{T}_{j,1}, \ldots, \mathcal{T}_{j,n-1}) \overset{d}{=} \left(\frac{\sigma B_t}{\sqrt{t(1-t)}}\right)_{t=\frac{1}{n},\ldots,\frac{n-1}{n}}, \tag{1.4}$$

where B_t is a standard Brownian bridge on $[0, 1]$, i.e. $B_t \overset{d}{=} W_t - tW_1$ for $t \in [0, 1]$, where $(W_t)_{t\ge 0}$ is a standard Brownian motion. Furthermore, when there is a change, if we apply the calculation of the CUSUM to the mean matrix of X instead, we find that this CUSUM matrix has rank 1 with leading left singular vector θ, the vector of change. These good properties make the CUSUM statistic arguably the most popular tool in the high-dimensional changepoint literature. The challenge, however, lies in finding an appropriate aggregation mechanism of the CUSUM statistic such that noise coordinates are left out.

Jirak (2015) constructed a test statistic by using a weighted ℓ_∞ aggregation, motivated by (1.4):

$$\Psi_\infty := \max_{1\le t\le n-1} \max_{1\le j\le p} \frac{\sqrt{t(n-t)}}{n} |\mathcal{T}_{j,t}|.$$

We reject the null hypothesis when Ψ_∞ exceeds a certain threshold. The asymptotic limiting distribution of Ψ_∞ under the null as $n, p \to \infty$ is derived in the paper. The changepoint location can also be estimated by finding the maximiser (in t) of a quantity similar to the test statistic. Though the method could be generalised to include multivariate ARMA and GARCH models, one major issue with a maximum statistic is that, for a fixed ϑ, the procedure can be inefficient when the change is evenly spread out across many coordinates. Yu and Chen (2020) recently used a Gaussian multiplier bootstrap to determine the threshold of another ℓ_∞-based CUSUM test statistic.

The ℓ_2 aggregation was studied by Zhang et al. (2010); Horváth and Hušková (2012) in multivariate settings without any sparsity assumption. Enikeeva and Harchaoui (2019) proposed a linear statistic and a scan statistic:

$$\Psi_{\text{linear}} := \frac{1}{H_1(p,\alpha_1)} \max_{1\le t\le n-1} \frac{\sum_{j=1}^p \mathcal{T}_{j,t}^2 - p}{\sqrt{2p}},$$
$$\Psi_{\text{scan}} := \max_{1\le s\le p} \frac{1}{H_2(s,p,\alpha_2)} \max_{1\le t\le n-1} \frac{\sum_{j=1}^s \mathcal{T}_{(j),t}^2 - s}{\sqrt{2s}},$$

where $|\mathcal{T}_{(1),t}| \ge |\mathcal{T}_{(2),t}| \ge \ldots \ge |\mathcal{T}_{(p),t}|$ and where $H_1(p,\alpha_1)$ and $H_2(s,p,\alpha_2)$ are thresholds that provide significance levels α_1 and α_2 for the linear statistic and the scan statistic respectively. We reject the null hypothesis if either Ψ_{linear} or Ψ_{scan} exceeds 1. The asymptotic regime in this work is $s \to \infty$, $s/p \to 0$ and $\frac{\log n}{s \log(p/s)} \to 0$ as $p \to \infty$. When the vector of change is dense or moderately sparse, the linear statistic will detect the change; when the vector of change is very sparse, the scan statistic will be more effective. The boundary between the two regimes is $s \asymp p^{1/2}$. This procedure has a vanishing testing error when

$$\frac{z_1(n-z_1)}{n}\vartheta^2 \ge \min\Big\{\sqrt{p\log p} + \sqrt{p \log\log n}, s\log(p/2)\Big\}. \tag{1.5}$$

Cho and Fryzlewicz (2015) used the following ℓ_1 aggregation of the CUSUM statistic:

$$\Psi_{\ell_1} := \max_{1\le t\le n-1} \sum_{j=1}^p |\mathcal{T}_{j,t}| \mathbb{1}_{\{|\mathcal{T}_{j,t}|\ge \pi_n\}},$$

with some threshold π_n, and combined this with the wild binary segmentation (WBS) to detect multiple changepoints. Cho (2016) extended this ℓ_1 aggregation method by providing a data-adaptive mechanism for choosing the threshold π_n and proposed the Double CUSUM method. The adaptive statistic is then able to detect both sparse and dense changes.

Another approach is to project the data matrix X along a good direction and reduce this to a univariate problem. The oracle projection direction is θ/ϑ. Thus, we seek a data-driven direction which is close to this oracle one. Recall that the vector of change θ is the leading singular vector of the CUSUM matrix of the mean matrix of X. If the sparsity s is known, then the s-sparse leading left singular vector of $\mathcal{T}$ is a consistent estimator of the oracle projection direction. However, computing this vector is a non-convex and NP-hard optimisation problem. Wang and Samworth (2018) proposed the Information Sparse Projection for Estimation of Changepoints algorithm (inspect), which uses a convex relaxation of this problem and computes the estimated oracle projection direction $\hat{v}$ as the leading singular vector of

$$\hat{M} \in \underset{M \in \mathcal{S}}{\operatorname{argmax}} \big\{ \operatorname{tr}(\mathcal{T}^\top M) - \lambda \|M\|_1 \big\},$$

with $\mathcal{S} := \{M \in \mathbb{R}^{p \times (n-1)} : \|M\|_* \leq 1\}$, where $\|M\|_1$ is the entrywise ℓ_1-norm and $\|M\|_*$ can be chosen to be either the nuclear norm or the entrywise l_2-norm of matrix M. The rate of convergence of the corresponding changepoint estimation can be shown to be minimax optimal up to a factor of $\log\log n$ under mild assumptions.

This sparse projection idea can also be generalised to exploit group sparsity structure (Cai and Wang, 2021) and handle heterogeneous missingness in high-dimensional changepoint problems (Follain, Wang and Samworth, 2022). For the latter problem, a new version of the CUSUM statistic that is suitable for missing data was introduced by Follain, Wang and Samworth (2022). Let $\Omega = (\omega_{j,t}) \in \{0,1\}^{p \times n}$ be the revelation matrix, with $\omega_{j,t} = 1$ if $X_{j,t}$ is observed, and $\omega_{j,t} = 0$ otherwise. The MissCUSUM matrix $\mathcal{T}^{\mathrm{Miss}} = \mathcal{T}^{\mathrm{Miss}}(X, \Omega) \in \mathbb{R}^{p \times (n-1)}$ is defined by:

$$[\mathcal{T}^{\mathrm{Miss}}(X,\Omega)]_{j,t} := \sqrt{\frac{\big(\sum_{r=1}^{t} \omega_{j,r}\big)\big(\sum_{r=t+1}^{n} \omega_{j,r}\big)}{\sum_{r=1}^{n} \omega_{j,r}}} \left(\frac{\sum_{r=1}^{t} (X \circ \Omega)_{j,r}}{\sum_{r=1}^{t} w_{j,r}} - \frac{\sum_{r=t+1}^{n} (X \circ \Omega)_{j,r}}{\sum_{r=t+1}^{n} w_{j,r}} \right),$$

for $j \in \{1, \ldots, p\}$ and $t \in \{1, \ldots, n-1\}$ such that $\sum_{r=1}^{t} w_{j,r} > 0$ and $\sum_{r=t+1}^{n} w_{j,r} > 0$ and 0 otherwise, where $X \circ \Omega$ denotes the Hadamard product of X and Ω. When X is fully observed, the MissCUSUM matrix coincides with the usual CUSUM matrix. Assume that Ω has a row-homogeneous distribution, i.e. there exists an observation rate vector $q = (q_1, \ldots, q_p)^\top \in (0,1]^p$ such that $\omega_{j,t} \sim \mathrm{Bern}(q_j)$, independently for all $j \in \{1, \ldots, p\}$ and $t \in \{1, \ldots, n\}$. Under this setting, the oracle projection direction is $\theta \circ \sqrt{q}$, where $\sqrt{q} := \big(q_1^{1/2}, \ldots, q_p^{1/2}\big)^T$. Thus, a similar optimisation problem

$$(\hat{v}, \hat{w}) \in \underset{\substack{v \in \mathbb{R}^p : \|v\|_2 \leq 1 \\ w \in \mathbb{R}^{n-1} : \|w\|_2 \leq 1}}{\operatorname{argmax}} \big\{ \operatorname{tr}(\mathcal{T}^\top v w^\top) - \lambda \|v\|_1 \big\},$$

can be solved to provide a good projection direction $\hat{v}$.

We list a couple more CUSUM-based methods below. Wang et al. (2022) constructed a test statistic based on a normalised two-sample U-statistic. Their approach does not address sparsity, though. A nonparametric extension of the CUSUM statistic in high-dimensional settings based on kernel density estimators was introduced by Padilla et al. (2021b). We now discuss some non CUSUM-based approaches. Kaul et al. (2021a) studied a plug-in least squares estimator, which achieves an optimal (without any logarithmic factor) rate of convergence of changepoint estimation on an integer scale ($O(\vartheta^{-2})$), but requires slightly stronger assumptions than those in Wang and Samworth (2018). Soh and Chandrasekaran (2017) used a filtering method (similar to the moving sum approach discussed in Section 1.1.1), combined with a denoising step which requires convex optimisation, to detect multiple changepoints. Their algorithm is applicable in the sequential settings, though the theoretical results are presented under the offline framework with a non-random sample size n.

Much of the theoretical analysis in previous literature has been devoted to quantifying the performance of changepoint location estimation. This is in contrast to the analysis of hypothesis tests which aim to detect whether a changepoint exists. Aston and Kirch (2018) introduced a concept of high-dimensional efficiency that allows the understanding of the detection power of different statistics. Enikeeva and Harchaoui (2019) derived the testing rate of their adaptive test statistic, as shown in (1.5). Liu, Gao and Samworth (2021) showed further that the minimax testing rate of a single changepoint is given by

$$\frac{z_1(n-z_1)}{n}\vartheta^2 \asymp \begin{cases} \sqrt{p\log\log(8n)} & \text{if } s \geq \sqrt{p\log\log(8n)} \\ \max\left\{ s\log\left(\frac{ep\log\log(8n)}{s^2}\right), \log\log(8n)\right\} & \text{if } s < \sqrt{p\log\log(8n)}. \end{cases}$$

This rate shows a phase transition when $s \asymp \sqrt{p\log\log(8n)}$. The authors also constructed an adaptive CUSUM-type testing procedure which achieves the minimax optimal testing rate.

We conclude this section by mentioning some works that focus on other changepoint models in high dimensions. Avanesov and Buzun (2018) and Wang, Yu and Rinaldo (2021b) studied the covariance structure change; Gibberd and Sandipan (2017), Kaul et al. (2021b) and Londschien, Kovács and Bühlmann (2021) focused on graphical models; Wang, Yu and Rinaldo (2021a) studied changes in sparse dynamic networks. Changepoint problems within high-dimensional regression models have also become more popular in recent years (Lee, Seo and Shin, 2016; Leonardi and Bühlmann, 2016; Kaul, Jandhyala and Fotopoulos, 2019; Rinaldo et al., 2021).

1.2 Sequential changepoint problem

Despite the rich literature on offline changepoint problems discussed in the previous section, it is the sequential[2] version of the problem that is arguably the more important for many applications: one would like to be able to detect a change as soon as possible after it has occurred. Entry points to this field include Lai (2001) and Tartakovsky, Nikiforov and Basseville (2014). Of course, one option here is to apply an offline method after seeing every new observation (or batch of observations). However, this is unlikely to be a successful strategy: not only is there a difficult and highly dependent multiple testing issue to handle when using the method repeatedly on datasets of increasing size (see also Chu, Stinchcombe and White (1996) for further discussion of this point), but moreover, the storage and running time costs may frequently be prohibitive.

A sequential changepoint procedure is an extended stopping time[3] N (with respect to the natural filtration) taking values in $\mathbb{N} \cup \{\infty\}$. Equivalently, we can think of it as a family of $\{0,1\}$-valued estimators $(\hat{H}_n)_{n=1}^{\infty}$, where $\hat{H}_n = \hat{H}_n(X_1, \ldots, X_n)$, and where the sequence is increasing in the sense that $\hat{H}_m(X_1, \ldots, X_m) \leq \hat{H}_n(X_1, \ldots, X_n)$ for $m \leq n$. Here, the correspondence arises from $\hat{H}_n = \mathbb{1}_{\{N \leq n\}}$ and $N = \inf\{n \in \mathbb{N} : \hat{H}_n = 1\}$, with the usual convention that $\inf \emptyset := \infty$.

1.2.1 Control charts

Prior to changepoint problems becoming popular in the second half of last century, statistical process control was already an important tool in manufacturing. Here, we briefly discuss some control charts, which are often used in process monitoring. Let Y_i be the score of the i-the sample (e.g. number of defectives). Assume $(Y_i)_{i\in\mathbb{N}}$ are independent and identically distributed with mean μ_0 and standard deviation σ_0. For simplicity, we assume μ_0 and σ_0 to be known.

Shewhart charts (Shewhart, 1931). A group size k is fixed. Let $Z_n := \sum_{i=(n-1)k+1}^{nk} Y_i/k$, the sample mean of the n-th group. The process is under control after nk observations if $\mu_0 - C\sigma_0/\sqrt{k} \leq Z_i \leq \mu_0 + C\sigma_0/\sqrt{k}$, and not in control otherwise, for some $C > 0$. The choice of group size k can be tricky here, as a large k can result in slow responses while a small k can trigger many undesired false alarms.

CUSUM charts (Page, 1954). Define $S_n := \sum_{i=1}^{n} Y_i$, with $S_0 := 0$. Action needs to be taken when $S_n - \min_{0\leq i\leq n} S_i$ exceeds a certain threshold c. One important result about CUSUM is that, equivalently, we can define $Z_n := \max\{Z_{n-1} + Y_n, 0\}$, with $Z_0 := 0$,

[2] Readers should take notice that in many previous works, the word *online* is used instead of *sequential*. To avoid confusion with the definition of an *online* procedure at the beginning of this chapter, we shall use the word *sequential* throughout this section.

[3] A random variable τ taking values in $\mathbb{N} \cup \{\infty\}$ is an *extended stopping time* with respect to the filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$, if $\{\tau = n\} \in \mathcal{F}_n$ for all $n \in \mathbb{N}$.

and declare when $Z_n > c$. We remark that in the above scheme, we only care about an upward deviation. A two-sided scheme can easily be constructed from two one-sided ones.

EWMA charts (Roberts, 1959). For a exponentially weighted moving chart (EWMA), we define recursively that $Z_n := (1-\lambda)Z_{n-1} + \lambda Y_n$, with $Z_0 := 0$, for some $\lambda \in (0,1]$. The process is in control if $\mu_0 - C\sigma_0\sqrt{\lambda/(2-\lambda)} \leq Z_n \leq \mu_0 + C\sigma_0\sqrt{\lambda/(2-\lambda)}$ when n is large. An EWMA chart incorporates all the information from previous data but puts more emphasis upon most recent observations. Note that an EWMA chart with $\lambda = 1$ coincides with a Shewhart chart with $k = 1$.

1.2.2 Classic sequential detection procedures and criteria

Many classic sequential detection procedures require knowing both pre-change and post-change densities f_0 and f_1. In this section, we use $\mathbb{P}_z$ to denote the joint distribution of $(X_n)_{n=1}^{\infty}$, where the change takes place at time z and $\mathbb{E}_z$ the expectation under this distribution. Note that $z = \infty$ corresponds to the case of no change.

We first discuss the minimax framework of sequential change point detection. Under this framework, the two most important performance measures of a detection procedure are patience and responsiveness. More specifically, the patience of a procedure N is its average run length (ARL) in the absence of change (under the null). We denote this quantity by $\mathrm{ARL}(N) := \mathbb{E}_{\infty}N$. Let $\mathcal{C}_\gamma$ be the class of detection procedures with ARL at least γ. The responsiveness of a procedure N is characterised by the essential supremum average detection delay (ESADD) or (worst-)worst-case response delay (Lorden, 1971):

$$\mathrm{ESADD}(N) := \sup_{z\in\mathbb{N}\cup\{0\}} \operatorname{ess\,sup} \mathbb{E}_z\big[(N-z)\vee 0 \;\big|\; X_1,\ldots,X_z\big].$$

The average detection delay here is maximised first over all possible pre-change observation sequences and then over all changepoint locations.

In the changepoint literature, Page's CUSUM chart/procedure introduced in the last section is often defined with Y_i being the log-likelihood ratio, i.e.

$$N_{\mathrm{Page}} = N_{\mathrm{Page}}(c) := \inf\left\{ n \in \mathbb{N} : \max_{1\leq k\leq n} \sum_{i=k}^{n} \log\frac{f_1(X_i)}{f_0(X_i)} \geq c \right\}, \tag{1.6}$$

with threshold $c > 0$. Let $Z_n := \max_{1\leq k\leq n}\sum_{i=k}^{n}\log\frac{f_1(X_i)}{f_0(X_i)}$ with $Z_0 := 0$ denote the tracked statistic. Page's procedure is the building block for many future works, so it is essential to understand the underlying idea. First, Z_n is the generalised likelihood ratio statistic for testing $H_0 : X_1,\ldots,X_n \sim f_0$ against $H_1 : \exists 0 \leq z < n$ such that $X_1,\ldots,X_z \sim f_0$ and $X_{z+1},\ldots,X_n \sim f_1$. This is exactly the test we are interested in at time n. Secondly, we consider the sequential testing task to determine whether $X_1, X_2, \ldots$ are from density f_0 or density f_1. The optimal test is the sequential probability ratio test (SPRT) (Wald, 1947; Wald and Wolfowitz, 1948), where we calculate $\sum_{i=1}^{n} \log(f_1(X_i)/f_0(X_i))$ at each time n. We use upper boundary $c > 0$ and lower boundary 0 here for our SPRT. Once the sum exceeds c, we declare that the sample is from f_1, and once the sum drops below 0, we declare that the sample is from f_0. Recall that we can update the tracked statistic in Page's procedure recursively via $Z_n := \max\{Z_{n-1} + \log(f_1(X_n)/f_0(X_n)), 0\}$. This corresponds to repeatedly using the SPRT. Once we have declared f_0, we then throw away everything up to this point and restart the SPRT from 0. This is ideal for our changepoint model, as we can eliminate a part of pre-change sample with every reset of the SPRT.

Lorden (1971) explored further the link between Page's procedure and the SPRT and showed that, with threshold $c = \log \gamma$, Page's procedure is asymptotically minimax optimal as $\gamma \to \infty$ and satisfies

$$\mathrm{ESADD}(N_{\mathrm{Page}}(\log \gamma)) \sim \inf_{N \in \mathcal{C}_\gamma} \mathrm{ESADD}(N) \sim \frac{\log \gamma}{D(f_1 \| f_0)},$$

where $D(f_1 \| f_0)$ denotes the Kullback–Leibler divergence from f_0 to f_1. Moustakides (1986) proved further that, non-asymptotically, Page's procedure is optimal for each $\gamma > 0$. An alternative proof of this optimality was given by Ritov (1990), where Page's procedure was viewed under a Bayesian perspective and an optimal strategy of a sequential stochastic game was considered.

The supremum conditional average detection delay (SCADD), proposed by Pollak (1985):

$$\mathrm{SCADD}(N) := \sup_{z \in \mathbb{N} \cup \{0\}} \mathbb{E}_z[N - z \mid N > z],$$

is a slightly less pessimistic responsiveness measure than Lorden's criterion. We shall discuss the optimal procedure under this criterion later in the section.

The Bayesian framework assumes that the changepoint location is random rather than fixed. Let $\pi = (\pi_k)_{k \in \mathbb{N} \cup \{0\}}$ be the prior distribution of the changepoint location and define $\mathbb{P}^\pi(A) := \sum_{z=0}^{\infty} \pi_z \mathbb{P}_z(A)$ for any measurable set $\mathcal{A}$ and $\mathbb{E}^\pi$ the expectation under this distribution. The corresponding patience measure of a procedure N under this framework is the probability of false alarm (PFA): $\mathrm{PFA}^\pi(N) := \mathbb{P}^\pi(N \leq z)$. Let $\mathcal{C}_\alpha^\pi$ be the clsss of detection procedures with PFA at most α. The responsiveness measure is the average detection delay (ADD): $\mathrm{ADD}^\pi(N) := \mathbb{E}^\pi[N - z \mid N > z]$. Note again that in the above expressions, both N and z are random.

When in the special case that the prior distribution π follows a geometric distribution with $\pi_k = p(1-p)^k$ for $k \in \mathbb{N} \cup \{0\}$, the optimal procedure was found by Shiryaev (1961, 1963) by computing the posterior probability of the changepoint location after each observation. More specifically, Shiryaev's procedure

$$N_{\mathrm{S}} = N_{\mathrm{S}}(c) := \inf\left\{n \in \mathbb{N} : \sum_{k=1}^{n}\prod_{i=k}^{n} \frac{f_1(X_i)}{(1-p)f_0(X_i)} \geq c\right\}$$

with threshold $c = c_\alpha$ such that $\mathrm{PFA}^\pi(N_{\mathrm{S}}(c_\alpha)) = \alpha$ minimises $\mathrm{ADD}^\pi(N)$ among all $N \in \mathcal{C}_\alpha^\pi$. As in Page's procedure, the tracked statistic in Shiryaev's procedure could also be calculated recursively. Let $Z_n := \sum_{k=1}^{n}\prod_{i=k}^{n} \frac{f_1(X_i)}{(1-p)f_0(X_i)}$. Then, we have $Z_n = (1 + Z_{n-1})\frac{f_1(X_n)}{(1-p)f_0(X_n)}$, with $Z_0 := 0$.

Taking $p = 0$ in the definition of the Shiryaev's procedure, we arrive at the Shiryaev–Roberts (SR) procedure (Roberts, 1966):

$$N_{\mathrm{SR}} = N_{\mathrm{SR}}(c) := \inf\left\{n \in \mathbb{N} : \sum_{k=1}^{n}\prod_{i=k}^{n} \frac{f_1(X_i)}{f_0(X_i)} \geq c\right\} = \inf\{n \in \mathbb{N} : Z_n \geq c\},$$

where here $Z_n := (1 + Z_{n-1})f_1(X_n)/f_0(X_n)$ with $Z_0 := 0$. The geometric prior discussed in the last paragraph now becomes an improper uniform prior on $\mathbb{N} \cup \{0\}$. The SR procedure is optimal (Pollak and Tartakovsky, 2009), subject to a constraint on $\mathrm{ARL}(N)$, in terms of another responsiveness measure, the integral average detection delay (IADD):

$$\mathrm{IADD}(N) := \frac{\sum_{z=0}^{\infty} \mathbb{E}_z\big[(N - z) \vee 0\big]}{\mathbb{E}_\infty N}.$$

This quantity could be understood as a limit of $\mathrm{ADD}^\pi(N)$ as $p \to 0$ in the family of geometric prior distributions. Since the definition of IADD does not explicitly contain Bayesian elements, the SR procedure could be viewed as a bridge between the Bayesian and minimax frameworks.

A few variants were proposed and analysed in order to understand the theoretical behaviour of the SR procedure more thoroughly. The SR-r procedure (Moustakides, Polunchenko and Tartakovsky, 2011) changed the initial value of the tracked statistic $(Z_n)_{n\in\mathbb{N}\cup\{0\}}$ to another deterministic value $Z_0 := r$, while the Shiryaev–Roberts–Pollak (SRP) procedure (Pollak, 1985) used a random initialisation with a quasi-stationary distribution. Pollak (1985) established the near optimality of the SRP procedure in the sense that

$$\mathrm{SCADD}(N_{\mathrm{SRP}}(c_\gamma)) - \inf_{N\in\mathcal{C}_\gamma} \mathrm{SCADD}(N) = o(1)$$

as $\gamma \to \infty$, where $N_{\mathrm{SRP}}(c_\gamma)$ is the SRP procedure with threshold c_γ chosen to satisfy $\mathrm{ARL}(N_{\mathrm{SRP}}) = \gamma$. The strict optimality under the SCADD criterion, however, is achieved by the SP-r procedure, with a particular choice of r (Polunchenko and Tartakovsky, 2010).

Another frequently used approach to tackle sequential changepoint detection problem is the window limited moving average scheme:

$$N_{\mathrm{MA}} = N_{\mathrm{MA}}(w, c) := \inf\left\{ n \in \mathbb{N} : n \geq w \text{ and } \sum_{i=n-w+1}^{n} \log \frac{f_1(X_i)}{f_0(X_i)} \geq c \right\}.$$

Lai (1995) showed that, by choosing threshold $c = \log \gamma$ and a window size m which satisfies $w \sim (\log \gamma)/D(f_1 \| f_0)$ and $\{w - (\log \gamma)/D(f_1 \| f_0)\}/\sqrt{\log \gamma} \to \infty$ as $\gamma \to \infty$, the procedure has the desired patience level $\mathrm{ARL}(N_{\mathrm{MA}}(w, c)) \geq \gamma$, and is asymptotically optimal: $\mathrm{ESADD}(N_{\mathrm{MA}}(w, c)) \sim (\log \gamma)/D(f_1 \| f_0)$ as $\gamma \to \infty$.

1.2.3 Unknown post-change parameter

In the last section, we have assumed that both the pre- and post-change densities to be known. However, this might be too restrictive. For simplicity, we now consider the case that the post-change density belongs to an exponential family $\{f_\theta(x) = \exp\{\theta x - K(\theta)\} f_0(x) : \theta \in \Theta \setminus \{0\}\}$, where f_0 is the known pre-change density.

One natural way to take into account a wide range of choices of the unknown parameter is to use the generalised likelihood ratio. We can conveniently modify (1.6) to form the GLR schemes:

$$N_{\mathrm{GLR}} = N_{\mathrm{GLR}}(\Theta_1, c) := \inf\left\{ n \in \mathbb{N} : \max_{1 \leq k \leq n} \sup_{\theta \in \Theta_1} \sum_{i=k}^{n} \log \frac{f_\theta(X_i)}{f_0(X_i)} \geq c \right\}.$$

We note that Θ_1 need not to be the same as the entire parameter space Θ. Lorden (1971) studies the case of Θ being an interval on the real line including 0. Then, by choosing $\Theta_1 = \{\theta : |\theta| \geq h_\gamma\}$ with $h_\gamma \sim \log^{-1} \gamma$ and a suitable threshold $c = c_\gamma$, the procedure (denoted as $N_{\mathrm{GLR,L}}$ for convenience) satisfies $\mathbb{E}_\infty N_{\mathrm{GLR,L}} \geq \gamma$ and as $\gamma \to \infty$

$$\mathrm{ESADD}_\theta(N_{\mathrm{GLR,L}}) \sim \frac{\log \gamma}{D(f_\theta \| f_0)}$$

for all $\theta \in \Theta \setminus \{0\}$. Instead of taking the maximum likelihood over Θ_1, an alternative way is to take the 'average' log likelihood. This leads to mixture detection procedures (Pollak and Siegmund, 1975):

$$N_{\mathrm{GLR}} = N_{\mathrm{GLR}}(\Theta_1, \Lambda, c) := \inf\left\{ n \in \mathbb{N} : \max_{1 \leq k \leq n} \int_{\Theta_1} \sum_{i=k}^{n} \log \frac{f_\theta(X_i)}{f_0(X_i)} \, d\Lambda(\theta) \geq c \right\},$$

where $\Lambda(\cdot)$ is a probability distribution on Θ_1. The optimality result about mixture detection procedures was established by Pollak (1978).

Although Lorden's GLR procedure could be implemented by tracking and updating only a small number of quantities (e.g. in univariate Gaussian case, this is $O(\log^3 \gamma)$), GLR schemes and mixture detection procedures, in general, are not computationally friendly.

1.2.4 An alternative sequential paradigm

A second sequential paradigm was introduced by Chu, Stinchcombe and White (1996). Instead of controlling the average run length under the null (no change), we aim to find procedures that have (asymptotic) control of the probability of type I error $\mathbb{P}_\infty(N < \infty)$ and have (asymptotic) power 1 when a change is present. This approach often requires a non-contaminated or stable historical dataset. Changes in both univariate mean structure and linear regression parameters have been studied extensively under this paradigm (Aue and Horváth, 2004; Horváth et al., 2004; Aue et al., 2006; Kirch, 2008). The type I error control requirement is very different from the ARL constraint introduced in Section 1.2.2. In fact, this paradigm puts a heavier penalty on any false alarm. Many classic procedures discussed in Section 1.2.2 can not be considered under this paradigm since they all have $\mathbb{P}_\infty(N < \infty) = 0$.

Yu et al. (2020) proposed a sequential detection procedure for the univariate mean change problem based on the offline CUSUM statistic (1.1), and provided its theoretical guarantees under both sequential paradigms. They showed that the procedure has the control of the probability of a type I error or the average run length under the null and achieves the minimax rate of detection delay, up to a logarithmic factor. However, unlike in Chu, Stinchcombe and White (1996), their proposed procedure does not require an initial stable dataset to work with. Their work provides a middle ground between the two paradigms. A faster algorithm implementing this procedure when the pre-change mean is unknown was proposed by Romano et al. (2022).

1.2.5 The multivariate and high-dimensional settings

There is a paucity of prior literature on multivariate/high-dimensional, sequential changepoint problems, though the field is gathering momentum in recent years. Tartakovsky et al. (2006), Mei (2010) and Zou et al. (2015) all proposed methods to the multivariate problem, where the dimension p is assumed to be fixed. For example, Mei (2010) considered using the statistic in Page's procedure (1.6) for each coordinate. Write $X_i := (X_{1,i}, \ldots, X_{p,i})^\top \in \mathbb{R}^p$ for $i \in \mathbb{N}$ and denote

$$Z_n^{j,\theta_j} := \max_{1 \le k \le n} \sum_{i=k}^{n} \log \frac{f_{\theta_j}(X_{j,i})}{f_0(X_{j,i})},$$

for $j \in \{1, \ldots, p\}$ and $\theta_j \in \mathbb{R}$. The two statistics considered in this paper are the sum statistic $Z_{\mathrm{sum}} := \sum_{j=1}^p Z_n^{j,\theta_j}$ and the maximum statistic $Z_{\max} := \max_{j \in \{1,\ldots,p\}} Z_n^{j,\theta_j}$. A change is declared when either statistic exceeds given thresholds. A dense change will be detected by the sum statistic while a sparse one will be picked up by the maximum statistic. The ARL and ESADD of the procedure using the sum statistic are also analysed in this work. However, these aforementioned works focused either on the case where both the pre- and post-change distributions are exactly known, or where, for each coordinate, both the sign and a lower bound on the magnitude of change, are known in advance. The amount of information about the post-change distribution required can be unrealistic.

A number of methods that involve scanning a moving window of fixed size for changes have also been proposed. Both Xie and Siegmund (2013) and Chan (2017) used mixture procedures to monitor multivariate data, and are similar in nature. Here, we summarise the contribution of Xie and Siegmund (2013) only. The basic idea is to reduce the changepoint problem to the following sequential testing task using the tail observations. Let $B_1, \ldots, B_p \overset{\mathrm{iid}}{\sim} \mathrm{Bernoulli}(p_0)$ for some known $p_0 \in [0,1]$. We test the null $(X_i)_{i\in\mathbb{N}}$ where $X_i \overset{\mathrm{iid}}{\sim} \mathcal{N}_p(0, I_p)$ against an alternative of a mixture distribution, where for each coordinate $j \in \{1,\ldots,p\}$, independently, $(X_{j,i})_{i\in\mathbb{N}}$ satisfies

$$X_{j,i} \mid B_j \overset{\mathrm{iid}}{\sim} \mathcal{N}\bigl(\mu_j \mathbb{1}_{\{B_j=1\}}, 1\bigr),$$

for some unknown $\mu_j \in \mathbb{R}$. The quantity p_0 is an estimate of the proportion of series affected by the change. The maximum likelihood estimate of the post-change mean is then used to complete the construction:

$$N_{\mathrm{XS}} := \inf\biggl\{n \in \mathbb{N} : \max_{1\le k\le n} \sum_{j=1}^{p} \log\Bigl(1 - p_0 + p_0 e^{\{\max(Z_{n,k,j},0)\}^2}\Bigr)\biggr\},$$

where $Z_{n,k,j} := k^{-1/2}\sum_{i=n-k+1}^{n} X_{j,i}$ for $n \in \mathbb{N}$, $k \in \{1,\ldots,n\}$ and $j \in \{1,\ldots,p\}$. A window limited version of the procedure could be used here by replacing $\max_{1\le k\le n}$ with $\max_{1\le k\le w}$ in the above expression, where w is the window size. This can reduce the memory requirements when implementing the procedure. As mentioned in Section 1.1.2, Soh and Chandrasekaran (2017) proposed an efficient multiple changepoint detection procedure for high-dimensional sparse signals. The procedure is constructed in a sequential manner. At each time step, we first calculate the sample mean using sample from a rolling window of size w. A denoising step which requires solving a convex optimisation problem is then applied before differencing. The output from the denoising step gives a better estimate of the true signal than the raw sample mean. These moving window methods can be effective when the signal-to-noise ratio is large enough that the change can be detected within the prescribed window, but may experience excessive response delay in other cases. Of course, the window size may be increased to compensate, but this correspondingly increases the computational complexity and storage requirements.

The high-dimensional problem has also been recently studied under the alternative sequential paradigm (Chu, Stinchcombe and White, 1996). Gösmann et al. (2020) and Gösmann, Kley and Dette (2021) used likelihood ratio based weighted CUSUM statistics and the ℓ_∞ aggregation to construct a sequential procedure for closed-end and open-end monitoring respectively. In the close-end scenario (i.e. there is a fixed endpoint for monitoring), the test statistic converges weakly to a Gumbel random variable under the null, as both the dimension p and the sample size of the initial stable dataset m tend to infinity.

1.3 Inference for changepoints

In this section, we review the existing works that address statistical inference for multiple changepoints. Many questions regarding uncertainty quantification can be asked. Here, we mainly focus on the literature whose aim is to provide confidence intervals/sets/statements about the locations of changepoints.

One approach is to provide confidence sets associated with Simultaneous Multiscale Changepoint Estimation (SMUCE), proposed by Frick, Munk and Sieling (2014). This work lays the foundation for future works in inferential aspects of multiple change points, where previous literature had been quite scarce. In this work, it is assumed that the observations follow

$$X_i \overset{\text{ind}}{\sim} F_{\phi(i/n)}, \qquad i = 1, \ldots, n,$$

where $(F_\theta)_{\theta \in \Theta}$ is a one-dimensional known exponential family and $\vartheta : [0,1] \to \Theta$ is a right continuous step function (also known as the regression function) with ν changepoints, where ν is unknown. SMUCE first estimates ν by minimising the 'number of changes' over all possible regression functions, subject to a log-likelihood ratio based multiscale statistic being below a certain threshold q. The minimal value $\hat{\nu}(q)$ gives the estimated number of change points and a confidence set (band) $\mathcal{C}(q)$ for the true regression function is given by the set of all functions that are optimisers of the above optimisation problem, i.e. regression functions that have $\hat{\nu}(q)$ many changepoints and satisfy the constraint. The SMUCE estimator $\hat{\vartheta}(q)$ for the regression function is then the constrained maximum likelihood estimator within the confidence set $\mathcal{C}(q)$. In addition, by inspecting the changepoint locations of the regression functions in $\mathcal{C}(q)$, we can produce simultaneous confidence intervals for changepoint locations.

The most crucial component of SMUCE is having a good estimate of the number of changepoints, as both the changepoint estimation and the uncertainty quantification rely heavily on it. Given a desired probability level α, the threshold $q = q_\alpha$ can be chosen to provide asymptotic control of the probability of overestimating the number of changepoints by α, while an exponential bound on the probability of underestimation is also derived. Based on these bounds, the confidence band for the regression function and the confidence intervals for changepoint locations have the desired (asymptotic) $(1-\alpha)$-coverage.

SMUCE can be used to study Gaussian observations with fixed and known (or estimated) variance σ^2 and piecewise constant mean. In fact, sharper and non-asymptotic theoretical results are possible under Gaussian settings. Pein, Sieling and Munk (2017) extended this methodology to handle the heterogeneous Gaussian change point model, while Dette, Eckle and Vetter (2020) proposed a similar procedure for dependent data.

Recall that by choosing a suitable threshold $q = q_\alpha$, SMUCE controls $\mathbb{P}(\hat{\nu}(q) > \nu)$ at level α asymptotically. This could also be viewed as a family-wise error rate (FWER) control. Though a corresponding theoretical result was provided on bounding the probability of underestimation, in practice when there are many changepoints or when the signal to noise ratio is small, the issue of underestimating ν could be magnified and affects the coverage of the confidence set and intervals. The problem of underestimation could also arise when the significance level α is small.

One way to overcome this is to use the false discovery rate (FDR) control instead. The rich history of FDR control in multiple testing dates back to Benjamini and Hochberg (1995). Here, we again work with Gaussian observations with fixed variance and piecewise constant mean and follow the framework described in Li, Munk and Sieling (2016). Consider the following multiple testing problem under the setting of changepoint detection:

$$H_{i,0} : i \text{ is not a changepoint, v.s. } H_{i,1} : i \text{ is a changepoint} \quad i = 1, \dots, n-1,$$

Let $\{\hat{z}_1, \dots, \hat{z}_{\hat{\nu}}\}$ denote all rejections/estimated changepoint locations. Intuitively, a rejection should be identified as a true discovery if it is 'close' to one of the true changepoints, and a false discovery otherwise. Several different notions are used to quantify closeness in literature. The first one is to use a uniform accuracy (Hao, Niu and Zhang, 2013; Cheng, He and Schwartzman, 2020). We say that $\hat{z}_i$ is a true discovery if it is within distance h of one of the true changepoints:

$$\min_{1 \le j \le \nu} |\hat{z}_i - z_j| \le h,$$

where h is a pre-specified threshold, and $z_1, \dots, z_\nu$ are true changepoints. This notion gives immediate guarantees on the accuracy of the estimated changepoint locations. However, choosing a good threshold h can be quite tricky in practice, as large values work well for long segments in between the changepoints, while small values are better for short segments. Another major drawback is that there could be many true discoveries corresponding to one single true changepoint, which almost certainly leads to an overestimation of the number of changepoints. An alternative notion is proposed by Li, Munk and Sieling (2016). We say that $\hat{z}_i$ is a true discovery if there exists a true changepoint within the interval

$$\left[\left\lceil \frac{\hat{z}_{i-1} + \hat{z}_i}{2} \right\rceil, \left\lceil \frac{\hat{z}_i + \hat{z}_{i+1}}{2} \right\rceil\right).$$

Since the intervals are disjoint for each i, it is guaranteed that there is at most one true discovery per true changepoint. This overcomes one drawback of using the uniform accuracy. The downside, however, is that we are sacrificing the accuracy of estimated changepoint locations – we allow a bigger gap (as big as almost $n/2$ in the most extreme case) between 'true discoveries' and true changepoints. Recall that FDR is defined to be the expectation of the false discovery proportion (FDP), where FDP equals the number of false discoveries divided by the total number of estimated changepoints (or 1 if zero estimated changepoint).

Based on the first notion of closeness, Hao, Niu and Zhang (2013) studied the Screening and Ranking algorithm (SaRa), first introduced by Niu and Zhang (2012) and proved the FDR control for the algorithm. Cheng, He and Schwartzman (2020) proposed a changepoint detection algorithm based on kernel smoothing and testing maxima and minima (dSTEM). This procedure also achieves asymptotic control of the FDR and power consistency. Using the alternative approach, Li, Munk and Sieling (2016) extended SMUCE and proposed a multiscale segmentation method (FDRSeg), which again controls the FDR. It is worth mentioning that though using the alternative approach, the FDRSeg algorithm still enjoys high accuracy in changepoint location estimation in the theoretical analysis.

In Section 1.1, we have introduced various offline multiple changepoint detection procedures and their corresponding test statistics. Many of these works include consistency results (when sample size n tends to infinity) for the number and locations of estimated changepoints; some also have finite sample bound results (e.g. Wang and Samworth, 2018). However, these results cannot be directly translated into confidence interval construction as the rate of convergence often involves unknown quantities such as minimum gap between changepoints and/or a lower bound on the change magnitude. In some cases, inferential tasks are still possible when the asymptotic or approximate distribution of the test statistics under the null hypothesis of no change can be derived, (e.g. Eichinger and Kirch, 2018; Fang, Li and Siegmund, 2020). For a given number of changepoints, using the relations between confidence intervals and hypothesis tests, we can construct joint asymptotic confidence regions for changepoint locations and means.

The moving sum (MOSUM) procedure introduced in Section 1.1 is another popular and easily implemented method in multiple changepoint detection. Moreover, MOSUM inherently contains some level of inferential arguments about the number and locations of the changepoints, in that if M_t defined in (1.2) exceeds the prescribed threshold, then with high probability there is a true changepoint within the interval $[t+1-G, t+G]$. As mentioned in the last paragraph, Eichinger and Kirch (2018) derived the asymptotic behaviour of the MOSUM statistics under the null. Cho and Kirch (2021) recently proposed a bootstrap procedure to construct confidence intervals for multiple changepoints and proved that these intervals attain asymptotic coverage. The method has its origins in Antoch, Hušková and Veraverbeke (1995) for the single changepoint setting.

Another set of articles works in the framework of post-selection inference, also known as selective inference (Hyun, G'Sell and Tibshirani, 2018; Duy et al., 2020; Hyun et al., 2021; Jewell, Fearnhead and Witten, 2022). Again, we consider univariate Gaussian observations of sample size n: $X = (X_1, \ldots, X_n)^\top \in \mathbb{R}^n$ with independent $X_i \sim \mathcal{N}(\mu_i, \sigma^2)$, where

$\mu = (\mu_1, \ldots, \mu_n)^\top$ is an unknown piecewise constant vector. One basic post-selection inference procedure adapted to our changepoint setting can be described as follows:

1. Given data X, use a base changepoint detection algorithm to detect k changepoints: $0 = \hat{z}_0 < \hat{z}_1 < \ldots < \hat{z}_k < \hat{z}_{k+1} = n$.

2. Determine contrast vectors $v_1, \ldots, v_k$ such that $v_j^\top x = \bar{x}_{(\hat{z}_j+1):\hat{z}_{j+1}} - \bar{x}_{(\hat{z}_{j-1}+1):\hat{z}_j}$ for any vector $x \in \mathbb{R}^n$. This is the sample mean difference between the segment to the right and to the left of the estimated changepoint.

3. For each j, test the null hypothesis $H_{0,j} : v_j^\top \mu = 0$ using a test statistic $T(X, v_j)$, which incorporates knowledge from both previous steps.

4. (Optional) Correct for multiple testing.

In the first step, k is usually the number of steps/iterations of an algorithm. The choice of k could be data-driven. Other contrast vectors are possible in step 2, such as a spike contrast. Extra information such as estimated directions of change is often included in the output of the changepoint detection algorithm and the contrast vectors as well. The core quantity in post-selection inference is the selective distribution

$$v^\top X \,\Big|\, \Big\{M(X) = M(x), q(X) = q(x)\Big\}$$

under the null, where $M(x)$ is the changepoint model selected by the base algorithm using observed data x and $q(X)$ is an extra vector of sufficient statistic of nuisance parameters included for tractability. The conditioning set above needs to be a polyhedron in order to apply the post-selection inference tools from Lee et al. (2016) and Tibshirani et al. (2016).

Hyun, G'Sell and Tibshirani (2018) studied using 1d fused lasso to estimate k changepoints. The framework can also be applied to trend filtering (Tibshirani, 2014) and graph clustering via graph fused lasso (Tibshirani and Taylor, 2011). Hyun et al. (2021) then studied a similar procedure, but considered other base multiple changepoint detection methods including Binary Segmentation (Vostrikova, 1981), Circular Binary Segmentation (Olshen et al., 2004) and Wild Binary Segmentation (Fryzlewicz, 2014). Jewell, Fearnhead and Witten (2022) considered using ℓ_0 penalisation for changepoint detection. The authors also gained power from theirs tests by conditioning on much less information than Hyun et al. (2021) and removing the polyhedron conditioning set requirement.

One drawback of the post-selection inference for multiple changepoints framework is that these methods usually focus on individual/local significance for each estimated changepoint, rather than a global statement. In addition, over-conditioning remains a big issue, as the resulting procedures will not be very powerful. Duy et al. (2020) and Jewell, Fearnhead and Witten (2022) discussed this issue in more detail and attempted to reduce the amount of conditioning.

Recent contributions to the multiple changepoint inference problem also include Fryzlewicz (2021a,b). The Narrowest Significance Pursuit (NSP) methodology proposed in Fryzlewicz (2021a) detects localised regions, each of which contains at least one changepoint, at a given global significance level. The robust variation of the algorithm (Fryzlewicz, 2021b) provides a better way to handle data with a more general heterogeneous noise structure.

Inference for multiple changepoint has also been studied from Bayesian approaches (e.g. Fearnhead, 2006; Nam, Aston and Johansen, 2012).

1.4 Notation

For the rest of this thesis, we will use the following notation. We write $\mathbb{N}_0$ for the set of all non-negative integers. For $d \in \mathbb{N}$, we write $[d] := \{1, \ldots, d\}$. Given $a, b \in \mathbb{R}$, we denote $a \vee b := \max(a, b)$ and $a \wedge b := \min(a, b)$. For a set S, we use $\mathbb{1}_S$ and $|S|$ to denote its indicator function and cardinality respectively. For a real-valued function f on a totally ordered set S, we write $\mathrm{sargmax}_{x \in S} f(x) := \min \mathrm{argmax}_{x \in S} f(x)$, the smallest maximiser of f in set S, and similarly we write $\mathrm{sargmin}_{x \in S} f(x) := \min \mathrm{argmin}_{x \in S} f(x)$, and $\mathrm{largmax}_{x \in S} f(x) := \max \mathrm{argmax}_{x \in S} f(x)$. For a vector $v = \bigl(v^1, \ldots, v^M\bigr)^\top \in \mathbb{R}^M$, we define $\|v\|_0 := \sum_{i=1}^M \mathbb{1}_{\{v^i \neq 0\}}$, $\|v\|_2 := \bigl\{\sum_{i=1}^M (v^i)^2\bigr\}^{1/2}$ and $\|v\|_\infty := \max_{i \in [M]} |v^i|$. In addition, for $j \in [M]$, we define $\|v^{-j}\|_2 := \bigl\{\sum_{i: i \neq j} (v^i)^2\bigr\}^{1/2}$. For a matrix $A = (A^{i,j}) \in \mathbb{R}^{d_1 \times d_2}$ and $j \in [d_2]$, we write $A^{\cdot,j} := \bigl(A^{1,j}, \ldots, A^{d_1,j}\bigr)^\top \in \mathbb{R}^{d_1}$ and $A^{-j,j} := \bigl(A^{1,j}, \ldots, A^{j-1,j}, A^{j+1,j} \ldots, A^{d_1,j}\bigr)^\top \in \mathbb{R}^{d_1 - 1}$. We use $\Phi(\cdot)$, $\bar{\Phi}(\cdot)$ and $\phi(\cdot)$ to denote the distribution function, survivor function and density function of the standard normal distribution respectively. For two real-valued random variables U and V, we write $U \geq_{\mathrm{st}} V$ or $V \leq_{\mathrm{st}} U$ if $\mathbb{P}(U \leq x) \leq \mathbb{P}(V \leq x)$ for all $x \in \mathbb{R}$. We adopt conventions that an empty sum is 0 and that $\min \emptyset := \infty$, $\max \emptyset := -\infty$.

Chapter 2

A high-dimensional, multiscale online changepoint detection procedure

2.1 Introduction

In this chapter, we are interested in developing algorithms for detecting changepoints in high-dimensional data that are observed sequentially. Moreover, we focus on *online algorithms*. Recall from Chapter 1 that an sequential procedure is online if the computational complexity for processing a new observation, as well as the storage requirements, depend only on the number of bits needed to represent the new observation.

To set the scene for our contributions, let $X_1, X_2, \ldots$ be a sequence of independent random vectors in $\mathbb{R}^p$. Assume that for some unknown, deterministic time $z \in \mathbb{N}_0$, the sequence is generated according to

$$X_1, \ldots, X_z \sim \mathcal{N}_p(\mu_-, I_p) \quad \text{and} \quad X_{z+1}, X_{z+2}, \ldots \sim \mathcal{N}_p(\mu_+, I_p), \tag{2.1}$$

for some $\mu_-, \mu_+ \in \mathbb{R}^p$. When $\mu_+ \neq \mu_-$, we say that there is a changepoint at time z. In many applications, such as in industrial quality control where the distribution of relevant properties of goods in a manufacturing process under regular conditions may be well understood, we may assume that the mean before the change is known (or at least can be estimated to high accuracy using historical data). However, the vector of change, $\theta := \mu_+ - \mu_-$, is typically unknown. Thus, for simplicity, we will work in the setting where $\mu_- = 0$ and $\mu_+ = \theta$. Let $\mathbb{P}_{z,\theta}$ denote the joint distribution of $(X_n)_{n=1}^{\infty}$ under (2.1) and $\mathbb{E}_{z,\theta}$ the expectation under this distribution. Note that when $\theta = 0$, the joint distribution of the data does not depend on z, and we therefore let $\mathbb{P}_0 = \mathbb{P}_{z,0}$ denote this joint distribution (with corresponding expectation $\mathbb{E}_0$). We will then say that the data is generated *under the null*. By contrast, if $\theta \neq 0$, we will say that the data is generated *under the alternative*, though we emphasise that in fact the alternative is composite, being indexed by $z \in \mathbb{N}_0$ and $\theta \in \mathbb{R}^p \setminus \{0\}$. In practice, in order for our procedure to have uniformly non-trivial power, it will be necessary to work with a subset of the alternative hypothesis parameter space that is well-separated from the null, in the sense that the ℓ_2-norm of the vector of mean change, $\vartheta := \|\theta\|_2$, is at least a known lower bound $\beta > 0$.

Recall from Section 1.2 that a sequential changepoint procedure is an extended stopping time (with respect to the natural filtration) taking values in $\mathbb{N} \cup \{\infty\}$. Then, following the concepts introduced in Section 1.2.2, we define the *patience* of a sequential changepoint procedure N to be $\mathbb{E}_0(N)$, and its *worst-case response delay* to be

$$\bar{\mathbb{E}}_\theta^{\mathrm{wc}}(N) := \sup_{z\in\mathbb{N}_0} \operatorname{ess\,sup} \mathbb{E}_{z,\theta}\bigl\{(N-z)\vee 0 \mid X_1,\ldots,X_z\bigr\}.$$

While controlling the worst-case response delay provides a very strong theoretical guarantee of the average detection delay of the procedure, even under the worst possible pre-change data sequence, obtaining a good bound for this quantity is often difficult. We therefore also consider the average-case response delay, or simply the *response delay* of a procedure N, defined as

$$\bar{\mathbb{E}}_\theta(N) := \sup_{z\in\mathbb{N}_0} \mathbb{E}_{z,\theta}\bigl\{(N-z)\vee 0\bigr\}.$$

We note that $\bar{\mathbb{E}}_\theta(N) \leq \bar{\mathbb{E}}_\theta^{\mathrm{wc}}(N)$. A good sequential changepoint procedure should have small worst- and average-case response delays, uniformly over the relevant class of alternatives $\{\mathbb{P}_{z,\theta} : (z,\theta) \in (\mathbb{N}\cup\{0\}) \times \mathbb{R}^p, \|\theta\|_2 \geq \beta\}$, subject to its patience being at least some suitably large, pre-determined $\gamma > 0$. Finally, as mentioned above, we are interested in sequential changepoint procedures that are online, so that the computational complexity per additional observation should be a function of p only.

Our main contribution in this work is to propose, in Section 2.2, a new algorithm called ocd (short for online changepoint detection), for high-dimensional, online changepoint detection in the above setting. The procedure works by performing likelihood ratio tests against simple alternatives of different scales in each coordinate, and then aggregating test statistics across scales and coordinates for changepoint detection. The ocd algorithm has worst-case computational complexity $O\bigl(p^2\log(ep)\bigr)$ per new observation, so satisfies our requirement for being an online algorithm. In fact, as we explain in Section 2.2.1, the algorithmic complexity is often even better than this. Moreover, as we illustrate in Section 2.4, it has extremely effective empirical performance. In terms of theoretical guarantees, it turns out to be more convenient to analyse a slight variant of our initial algorithm, which we refer to as ocd′. This has the same order of computational complexity per new observation as ocd, but enables us to ensure that whenever we are yet to declare that a change has occurred, only post-change observations contribute to the running test statistics. In practice, the original ocd algorithm also appears to have this property for typical pre-change sequences, and we argue heuristically that there is a sense in which it is more efficient than ocd′ by a factor of at most 2.

Our theoretical analysis in Section 2.3 initially considers separately versions of the `ocd`$'$ algorithm best tuned towards settings where the vector θ of change is dense, and where it is sparse in an appropriate sense. We then present results for a combined, adaptive procedure that seeks the best of both worlds. In all cases, the appropriate version of `ocd`$'$ has guaranteed patience, at least at the desired nominal level. In the (small-change) regime of primary interest, and when ϑ is of the same order as β, the response delay of `ocd`$'$ is of order at most $\sqrt{p}/\vartheta^2$ in the dense case, up to a poly-logarithmic factor; this can be improved to order s/ϑ^2, again up to a poly-logarithmic factor, when the effective sparsity of θ is $s < \sqrt{p}$.

Numerical results illustrate the performance of our `ocd` algorithm in Section 2.4. Proofs of our main results are given in Section 2.5. All the auxiliary lemmas and their proofs are provided in Section 2.6.

2.2 An online changepoint procedure

2.2.1 The `ocd` algorithm

In this section, we describe our online changepoint procedure, `ocd`, in more detail. As mentioned in the introduction, the procedure aggregates likelihood ratio test statistics against simple alternatives of different scales in different coordinates. For $i \in [n]$ and $j \in [p]$, we write X_i^j for the jth coordinate of X_i. Recall from Section 1.2.2 that if we want to test a null of $\mathcal{N}(0,1)$ against a simple post-change alternative distribution of $\mathcal{N}(b,1)$ for some $b \neq 0$ in coordinate $j \in [p]$, by Page (1954), the optimal online changepoint procedure is to declare that a change has occurred by time n when the test statistic

$$R_{n,b}^j := \max_{0 \leq h \leq n} \sum_{i=n-h+1}^{n} b(X_i^j - b/2) \tag{2.2}$$

exceeds a certain threshold. Note that $\sum_{i=n-h+1}^{n} b(X_i^j - b/2)$ can be viewed as the likelihood ratio test statistic between the null and this simple alternative using the tail sequence $X_{n-h+1}, \ldots, X_n$. Thus $R_{n,b}^j$ can be regarded as the most extreme of these likelihood ratio statistics, over all possible starting points for the tail sequence. Write

$$t_{n,b}^j := \operatorname*{sargmax}_{0 \leq h \leq n} \sum_{i=n-h+1}^{n} b(X_i^j - b/2) \tag{2.3}$$

for the length of the tail sequence in which the associated likelihood ratio statistic (in the jth coordinate) is maximised. One way to aggregate across the p coordinates would be to use $\sum_{j=1}^{p} R_{n,b}^j$ as a test statistic. However, this approach is not ideal for two reasons. Firstly, the exact distribution of the tail likelihood ratio statistic $R_{n,b}^j$ is hard to obtain, making it difficult to analyse the aggregated statistic under the null. More importantly, this aggregated statistic uses the same simple alternative $\mathcal{N}(b,1)$ in all coordinates, and so even after varying the magnitude of b, it is only effective against a very limited set of alternative distributions in $\{\mathbb{P}_{z,\theta} : z \in \mathbb{N}, \|\theta\|_2 \geq \beta\}$, namely those for which the change is of very similar magnitude in all coordinates. In order to overcome these problems, our procedure uses the coordinate-wise statistics $(R^j_{n,b} : j \in [p])$, which we call 'diagonal statistics', to detect changes that have a large proportion of their signal concentrated in one coordinate. To detect denser changes, for each $j \in [p]$, we also compute tail partial sums of length $t^j_{n,b}$ in all other coordinates $j' \neq j$, given by

$$A^{j',j}_{n,b} := \sum_{i=n-t^j_{n,b}+1}^{n} X^{j'}_i,$$

and aggregate them to form an 'off-diagonal statistic' anchored at coordinate j. Note that the number of summands in $A^{j',j}_{n,b}$ depends only on the observed data in the jth coordinate, and not on the data being aggregated in the j'th coordinate. These off-diagonal statistics are used to detect changes whose signal is not concentrated in a single coordinate. Intuitively, if a change has occurred and $\theta^j/b \geq 1$, then we can expect the tail length in coordinate j to be roughly of order $n-z$ for sufficiently large n, and this will ensure that the off-diagonal statistic anchored at coordinate j is close to the generalised likelihood ratio test statistic between the null and the composite alternative $\{\mathbb{P}_{z,\theta} : \|\theta\|_2 \neq 0\}$. If, in addition, a non-trivial proportion of the signal is contained in coordinates $[p] \setminus \{j\}$, then this statistic will be powerful for detecting the change.

The full description of the `ocd` procedure is given in Algorithm 2.1. Note that for notational simplicity, we have suppressed the time dependence of many variables as they are updated recursively in the algorithm. In the following, when necessary, we will make this dependence explicit by writing $A_{n,b}, t_{n,b}, Q_{n,b}, S^{\mathrm{diag}}_n$ and S^{off}_n for the relevant quantities at the end of the nth iteration of the repeat loop.

The algorithm takes inputs $X_1, X_2, \ldots \in \mathbb{R}^p$, observed sequentially, a known lower bound $\beta > 0$ for the ℓ_2-norm of the vector of mean change, a hard thresholding level $a \geq 0$ that can be chosen to detect dense or sparse signals, and two declaration thresholds $T^{\mathrm{diag}} > 0$ and $T^{\mathrm{off}} > 0$. We define sets of signed scales $\mathcal{B} := \Bigl\{\pm\frac{\beta}{\sqrt{2^\ell \log_2(2p)}} : \ell = 0, \ldots, \lfloor \log_2 p \rfloor\Bigr\}$ and $\mathcal{B}_0 := \Bigl\{\pm\frac{\beta}{\sqrt{2^{\lfloor \log_2 p \rfloor + 1}\log_2(2p)}}\Bigr\}$.

When a new observation X_n arrives, we first update

$$\begin{aligned} A^{\cdot,j}_{n,b} &:= A^{\cdot,j}_{n-1,b} + X_n \\ t^j_{n,b} &:= t^j_{n-1,b} + 1 \end{aligned}$$

for $(j,b) \in [p] \times \mathcal{B} \cup \mathcal{B}_0$. We then reset both $A^{\cdot,j}_{n,b}$ and $t^j_{n,b}$ to 0 if $bA^{j,j}_{n,b} - b^2 t^j_{n,b}/2 \leq 0$. By Lemma 2.10, $bA^{j,j}_{n,b} - b^2 t^j_{n,b}/2$ is equal to the quantity $R^j_{n,b}$ defined in (2.2) (we will also suppress its n dependence when it is clear from the context). Moreover, by Lemma 2.11, the two definitions of $t^j_{n,b}$ from Algorithm 2.1 and (2.3) coincide. In the algorithm, we allow b to range over the (signed) dyadic grid $\mathcal{B} \cup \mathcal{B}_0$, since the maximal signal strength in individual coordinates, $\|\theta\|_\infty$, can range from $\vartheta/\sqrt{p}$ to ϑ. In this way, the algorithm automatically adapts to different signal strengths in each coordinate. Here, the inclusion of $\mathcal{B}_0$ and the extra logarithmic factors in the denominators of elements of $\mathcal{B} \cup \mathcal{B}_0$ appear due to technical reasons in the theoretical analysis of the algorithm.

Algorithm 2.1 uses S^{diag} and S^{off} to aggregate diagonal and off-diagonal statistics respectively as mentioned above, and declares that a change has occurred as soon as either of these quantities exceeds its own pre-determined threshold. As mentioned previously, S^{diag} tracks the maximum of R^j_b over all scales b and coordinates j. Before introducing S^{off}, we first discuss the off-diagonal statistics Q^j_b in Algorithm 2.1, which are ℓ_2 aggregations of normalised tail sums $A^{j',j}_b/\sqrt{t^j_b \vee 1}$, each hard-thresholded at level a:

$$Q^j_{n,b} := \sum_{j' \in [p]: j' \neq j} \frac{(A^{j',j}_{n,b})^2}{t^j_{n,b} \vee 1} \mathbb{1}_{\left\{|A^{j',j}_{n,b}| \geq a\sqrt{t^j_{n,b}}\right\}}. \tag{2.4}$$

The hard thresholding level can be chosen to detect dense or sparse signals θ; in the sparse case a non-zero a facilitates an aggregation that aims to exclude coordinates with negligible change (thereby reducing the variance of the normalised tail sums). Finally, S^{off} is computed as the maximum of the Q^j_b over all anchoring coordinates $j \in [p]$ and scales $b \in \mathcal{B}$.

Although the off-diagonal statistics described in the previous paragraph are effective for detecting changes when the signal sparsity is known, it is desirable to the practitioner to have a combined procedure that adapts to the sparsity level. This may be computed straightforwardly by tracking S^{off} for $a = a^{\mathrm{dense}}$ and $a = a^{\mathrm{sparse}}$, as well as S^{diag}, and declaring a change when any of these three statistics exceeds a suitable threshold. Figure 2.1 illustrates the performance of this adaptive procedure, together with the time evolution of normalised versions of all three statistics tracked, in synthetic datasets both with and without a change. This adaptive procedure is analysed theoretically in Section 2.3.3 and empirically in Section 2.4.

The ocd procedure satisfies our definition of an online algorithm. Indeed, for each new observation X_n, ocd updates $t_{n,b} \in \mathbb{R}^p$ and $A_{n,b} \in \mathbb{R}^{p \times p}$ for $O\big(\log(ep)\big)$ different values of b. It then computes S^{diag}_n and S^{off}_n via $A_{n,b}$. These steps require $O\big(p^2 \log(ep)\big)$ operations. Moreover, the total storage used is $O\big(p^2 \log(ep)\big)$ throughout the algorithm.

In fact, the computational complexity of ocd can often be reduced, because typically $\mathcal{T} := \{t^j_b : j \in [p], b \in \mathcal{B}\}$ has cardinality much less than $p|\mathcal{B}|$ (which is the worst case, when all elements are distinct). Correspondingly, at each time step, we need only store the $p \times |\mathcal{T}|$ matrix $(B^{k,t})_{k \in [p], t \in \mathcal{T}}$ given by $B^{k,t^j_b} := A^{k,j}_b$, resulting in an improved per-iteration computational complexity and storage for ocd of $O(p|\mathcal{T}|)$. For simplicity of exposition, we have not presented this computational speed-up in Algorithm 2.1. We remark that Romano et al. (2022) recently provided some insights into the size of the set $\mathcal{T}$. Nevertheless we have

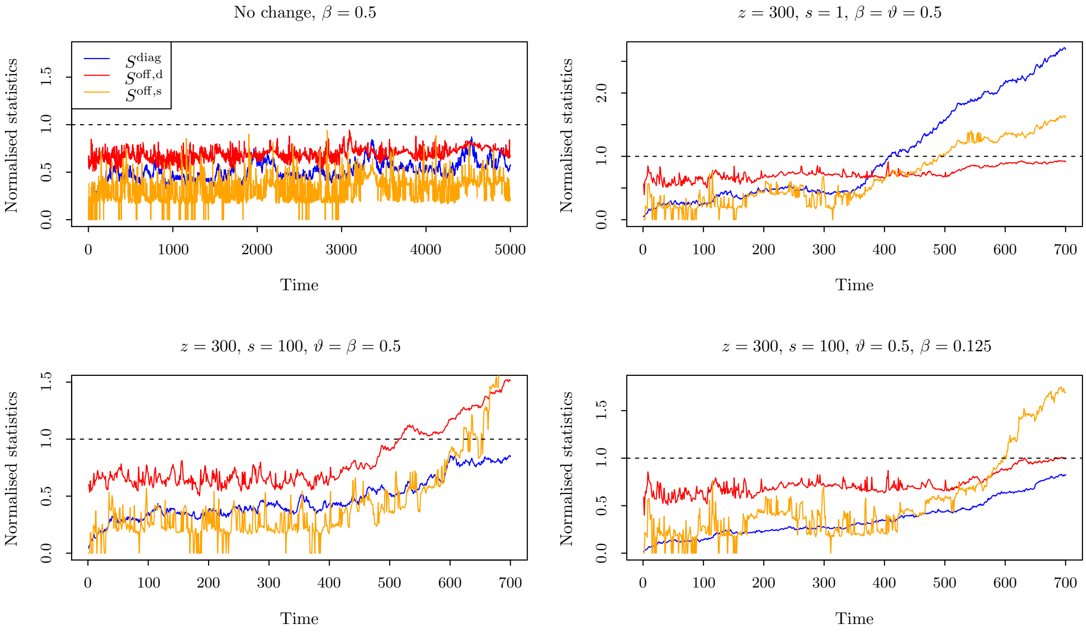

Fig. 2.1 Behaviour of the three normalised statistics in `ocd` under the null and under the alternative with different signal strength, sparsity level and assumed lower bound. A change is declared as soon as one of these three normalised statistics exceeds 1. The data were generated in the top-left panel according to $\mathbb{P}_0$, and, in the other panels, according to $\mathbb{P}_{z,\theta}$, with $p = 100$, $z = 300$ and $\theta = \vartheta U$, where U is uniformly distributed on the union of all s-sparse unit spheres in $\mathbb{R}^p$ (see Section 2.4.2 for a more detailed description).

implemented the algorithm in this form in the R package ocd (Chen, Wang and Samworth, 2020), and have found it to provide substantial computational savings in practice.

Algorithm 2.1: Pseudo-code of the ocd algorithm

```
Input: $X_1, X_2 \ldots \in \mathbb{R}^p$ observed sequentially, $\beta > 0$, $a \geq 0$, $T^{\mathrm{diag}} > 0$ and $T^{\mathrm{off}} > 0$
Set: $\mathcal{B} = \Big\{\pm\frac{\beta}{\sqrt{2^\ell \log_2(2p)}} : \ell = 0, \ldots, \lfloor \log_2 p \rfloor\Big\}$, $\mathcal{B}_0 = \Big\{\pm\frac{\beta}{\sqrt{2^{\lfloor \log_2 p \rfloor + 1}\log_2(2p)}}\Big\}$, $n = 0$,
    $A_b = \mathbf{0} \in \mathbb{R}^{p\times p}$ and $t_b = 0 \in \mathbb{R}^p$ for all $b \in \mathcal{B} \cup \mathcal{B}_0$
repeat
    $n \leftarrow n + 1$
    observe new data vector $X_n$
    for $(j, b) \in [p] \times (\mathcal{B} \cup \mathcal{B}_0)$ do
        $t_b^j \leftarrow t_b^j + 1$
        $A_b^{\cdot,j} \leftarrow A_b^{\cdot,j} + X_n$
        if $bA_b^{j,j} - b^2 t_b^j/2 \leq 0$ then
            $t_b^j \leftarrow 0$ and $A_b^{\cdot,j} \leftarrow 0$
        compute $Q_b^j \leftarrow \sum_{j' \in [p]: j' \neq j} \frac{(A_b^{j',j})^2}{t_b^j \vee 1} \mathbb{1}_{\{|A_b^{j',j}| \geq a\sqrt{t_b^j}\}}$
    $S^{\mathrm{diag}} \leftarrow \max_{(j,b)\in[p]\times(\mathcal{B}\cup\mathcal{B}_0)} \big(bA_b^{j,j} - b^2 t_b^j/2\big)$
    $S^{\mathrm{off}} \leftarrow \max_{(j,b)\in[p]\times\mathcal{B}} Q_b^j$
until $S^{\mathrm{diag}} \geq T^{\mathrm{diag}}$ or $S^{\mathrm{off}} \geq T^{\mathrm{off}}$;
Output: $N = n$
```

2.2.2 A slight variant of ocd

While the ocd algorithm performs very well numerically, it turns out to be easier theoretically to analyse a slight variant, which we call ocd′, and describe in Algorithm 2.2. Again, we have suppressed the time dependence n of many variables including $\tau_{n,b}, \tilde{\tau}_{n,b}, \Lambda_{n,b}$ and $\tilde{\Lambda}_{n,b}$ in the algorithm. The main difference between these two algorithms is that in ocd′, the off-diagonal statistics Q_b^j are computed using tail partial sums of length τ_b^j instead of t_b^j. These new tail partial sums are recorded in $\Lambda_b \in \mathbb{R}^{p\times p}$.

By Lemma 2.19, we always have

$$t_b^j/2 \leq \tau_b^j < 3t_b^j/4 \tag{2.5}$$

whenever $t_b^j \geq 2$. In this sense, the tail sample size used by ocd′ is smaller than that of ocd by a factor of at most 2. The benefit of using a shorter tail in ocd′ is that when n exceeds a known, deterministic threshold, we can be sure that whenever we have not declared that a change has occurred by time z, the tail partial sum consists exclusively of post-change observations. In practice, we observe that even in Algorithm 2.1, the tail lengths $t_{z,b}^j$ at the changepoint are generally very short for many coordinates, so the inclusion of a few pre-change observations in the tail partial sum calculation does not significantly affect the efficacy of the changepoint detection procedure. The practical performance of Algorithm 2.1 is statistically more efficient than Algorithm 2.2 in many settings by a factor of between 4/3 and 2, as suggested by (2.5). By construction, τ_b^j and $\Lambda_b^{\cdot,j}$ are computable online, through auxiliary variables $\tilde{\tau}_b^j$ and $\tilde{\Lambda}_b^{\cdot,j}$. Indeed, Algorithm 2.2 is also an online algorithm, with overall computational complexity per observation and storage remaining at $O\bigl(p^2\log(ep)\bigr)$ in the worst case; similar computational improvements to those mentioned for `ocd` at the end of Section 2.2.1 are also possible here.

Algorithm 2.2: Pseudo-code of the `ocd′` algorithm, a slight variant of `ocd`

```
Input: $X_1, X_2 \ldots \in \mathbb{R}^p$ observed sequentially, $\beta > 0$, $a \geq 0$, $T^{\mathrm{diag}} > 0$ and $T^{\mathrm{off}} > 0$.
Set: $\mathcal{B} = \Bigl\{\pm\frac{\beta}{\sqrt{2^\ell \log_2(2p)}} : \ell = 0, \ldots, \lfloor \log_2 p \rfloor\Bigr\}$, $\mathcal{B}_0 = \Bigl\{\pm\frac{\beta}{\sqrt{2^{\lfloor \log_2 p \rfloor + 1}\log_2(2p)}}\Bigr\}$, $n = 0$,
    $A_b = \Lambda_b = \tilde{\Lambda}_b = \mathbf{0} \in \mathbb{R}^{p\times p}$ and $t_b = \tau_b = \tilde{\tau}_b = 0 \in \mathbb{R}^p$ for all $b \in \mathcal{B} \cup \mathcal{B}_0$
repeat
    $n \leftarrow n + 1$
    observe new data vector $X_n$
    for $(j, b) \in [p] \times (\mathcal{B} \cup \mathcal{B}_0)$ do
        $t_b^j \leftarrow t_b^j + 1$  and  $A_b^{\cdot,j} \leftarrow A_b^{\cdot,j} + X_n$
        set $\delta = 0$ if $t_b^j$ is a power of 2 and $\delta = 1$ otherwise.
        $\tau_b^j \leftarrow \tau_b^j\delta + \tilde{\tau}_b^j(1-\delta) + 1$  and  $\Lambda_b^{\cdot,j} \leftarrow \Lambda_b^{\cdot,j}\delta + \tilde{\Lambda}_b^{\cdot,j}(1-\delta) + X_n$
        $\tilde{\tau}_b^j \leftarrow (\tilde{\tau}_b^j + 1)\delta$  and  $\tilde{\Lambda}_b^{\cdot,j} \leftarrow (\tilde{\Lambda}_b^{\cdot,j} + X_n)\delta$.
        if $bA_b^{j,j} - b^2 t_b^j/2 \leq 0$ then
            $t_b^j \leftarrow \tau_b^j \leftarrow \tilde{\tau}_b^j \leftarrow 0$
            $A_b^{\cdot,j} \leftarrow \Lambda_b^{\cdot,j} \leftarrow \tilde{\Lambda}_b^{\cdot,j} \leftarrow 0$
        compute $Q_b^j \leftarrow \sum_{j' \in [p]: j' \neq j} \frac{(\Lambda_b^{j',j})^2}{\tau_b^j \vee 1} \mathbb{1}_{\{|\Lambda_b^{j',j}| \geq a\sqrt{\tau_b^j}\}}$
    $S^{\mathrm{diag}} \leftarrow \max_{(j,b)\in[p]\times(\mathcal{B}\cup\mathcal{B}_0)}\bigl(bA_b^{j,j} - b^2 t_b^j/2\bigr)$
    $S^{\mathrm{off}} \leftarrow \max_{(j,b)\in[p]\times\mathcal{B}} Q_b^j$
until $S^{\mathrm{diag}} \geq T^{\mathrm{diag}}$ or $S^{\mathrm{off}} \geq T^{\mathrm{off}}$;
Output: $N = n$
```

2.3 Theoretical analysis

As mentioned in Section 2.2, the input a in Algorithms 2.1 and 2.2 allows users to detect changepoints of different sparsity levels. More precisely, for any $\theta \in \mathbb{R}^p$, we have by Lemma 2.18 that there exists a smallest $s(\theta) \in \{2^0, 2^1, \ldots, 2^{\lfloor \log_2 p \rfloor}\}$ such that the set

$$\mathcal{S}(\theta) := \biggl\{ j \in [p] : |\theta^j| \geq \frac{\|\theta\|_2}{\sqrt{s(\theta)\log_2(2p)}} \biggr\}$$

has cardinality at least $s(\theta)$. On the other hand, we also have $|\mathcal{S}(\theta)| \leq s(\theta)\log_2(2p)$. We call $s(\theta)$ the *effective sparsity* of the vector θ and $\mathcal{S}(\theta)$ its *effective support.* Intuitively, the sum of squares of coordinates in the effective support of θ has the same order of magnitude as $\|\theta\|_2^2$, up to logarithmic factors. Moreover, if θ is an s-sparse vector in the sense that $\|\theta\|_0 \leq s$, then $s(\theta) \leq s$, and the equality is attained when, for example, all non-zero coordinates have the same magnitude.

In this section, we initially analyse the theoretical performance of Algorithm 2.2 for two different choices of a in $S^{\mathrm{off}} = S^{\mathrm{off}}(a)$, namely $a = 0$ and $a = \sqrt{8\log(p-1)}$. We then present our combined, adaptive procedure and its performance guarantees. We note that the value of the universal constant C may vary from theorem to theorem in this section.

Define $N^{\mathrm{diag}} := \inf\{n : S_n^{\mathrm{diag}} \geq T^{\mathrm{diag}}\}$ and $N^{\mathrm{off}} = N^{\mathrm{off}}(a) := \inf\{n : S_n^{\mathrm{off}}(a) \geq T^{\mathrm{off}}\}$. Then the stopping time for our changepoint detection procedure is simply $N = N(a) = N^{\mathrm{diag}} \wedge N^{\mathrm{off}}(a)$.

2.3.1 Dense case

Here, we analyse the changepoint detection procedure $N = N(0)$, which, as we will see, is most suitable for detecting dense mean changes in the sense that $s(\theta) \geq \sqrt{p}$ (though we do not assume this in our theory). In this case, when $p \geq 2$ and conditionally on τ_b^j, the quantity Q_b^j follows a chi-squared distribution with $p-1$ degrees of freedom under the null, provided that τ_b^j is positive (When $p = 1$, we have that $Q_b^j = 0$ for all $j \in [p]$ and $b \in \mathcal{B}$, so $S^{\mathrm{off}} = 0$ and the off-diagonal statistic never triggers the declaration of a change. Similarly, if $p \geq 2$ but $\tau_{n,b}^j = 0$, then we also have $Q_{n,b}^j = 0$.). Motivated by the chi-squared tail bound of Laurent and Massart (2000, Lemma 1), we choose a threshold of the form

$$T^{\mathrm{off}} := p - 1 + \tilde{T}^{\mathrm{off}} + \sqrt{2(p-1)\tilde{T}^{\mathrm{off}}} =: \psi(\tilde{T}^{\mathrm{off}}), \tag{2.6}$$

say, for some $\tilde{T}^{\mathrm{off}} > 0$.

The following theorem provides control of the patience of `ocd′`.

Theorem 2.1. *Let $X_1, X_2, \ldots$ be generated according to $\mathbb{P}_0$. For any $\gamma \geq 1$, let $(X_t)_{t\in\mathbb{N}}$, $\beta > 0$, $a = 0$, $T^{\mathrm{diag}} = \log\{16p\gamma\log_2(4p)\}$ and $T^{\mathrm{off}} = \psi\bigl(\tilde{T}^{\mathrm{off}}\bigr)$ with $\tilde{T}^{\mathrm{off}} = 2\log\{16p\gamma\log_2(2p)\}\bigr)$ be the inputs of Algorithm 2.2, with corresponding output N. Then $\mathbb{E}_0(N) \geq \gamma$.*

We note that either of the two statistics S^{diag} and S^{off} may trigger a false alarm under the null. The two threshold levels T^{diag} and T^{off} are chosen so that $\mathbb{E}_0(N^{\mathrm{diag}})$ and $\mathbb{E}_0(N^{\mathrm{off}})$ have comparable upper bounds. We also remark that although Theorem 2.1 as stated only controls the expected value of N under the null, careful examination of the proof reveals that we can also control $\mathbb{P}_0(N \leq m)$ for every $m \in \mathbb{N}$. More precisely, from (2.16) and (2.17) in the proof, we can deduce that

$$\mathbb{P}_0(N \leq m) \leq \frac{m}{4\gamma}$$

for every $m \in \mathbb{N}$. The same bound holds for our other patience control results below, though we omit formal statements for brevity.

Our next result controls the response delay of $\texttt{ocd}'$ in both worst-case and average senses.

Theorem 2.2. *Assume that $X_1, X_2, \ldots$ are generated according to $\mathbb{P}_{z,\theta}$ for some z and θ such that $\|\theta\|_2 = \vartheta \geq \beta > 0$ and that θ has an effective sparsity of $s := s(\theta)$. Then there exists a universal constant $C > 0$, such that the output N from Algorithm 2.2, with inputs $(X_t)_{t\in\mathbb{N}}$, $\beta > 0$, $a = 0$, $T^{\mathrm{diag}} = \log\{16p\gamma\log_2(4p)\}$ and $T^{\mathrm{off}} = \psi\big(\tilde{T}^{\mathrm{off}}\big)$ with $\tilde{T}^{\mathrm{off}} = 2\log\{16p\gamma\log_2(2p)\}\big)$, satisfies*

$$\bar{\mathbb{E}}_\theta^{\mathrm{wc}}(N) \leq C\biggl\{\frac{\sqrt{p}\log(ep\gamma)}{\vartheta^2} \vee \frac{s\log(ep\gamma)\log(ep)}{\beta^2} \vee 1\biggr\}. \tag{2.7}$$

Furthermore, there exists $\beta_0(s) > 0$, depending only on s, such that for all $\beta \leq \beta_0(s)$, the output N satisfies

$$\bar{\mathbb{E}}_\theta(N) \leq C\biggl\{\frac{\sqrt{p}\log(ep\gamma)}{\vartheta^2} \vee \frac{\sqrt{s}\log(ep/\beta)\log(ep)}{\beta^2} \vee 1\biggr\}, \tag{2.8}$$

for $s \geq 2$, and

$$\bar{\mathbb{E}}_\theta(N) \leq C\biggl\{\frac{\log(ep\gamma)\log(ep)}{\beta\vartheta} \vee 1\biggr\}, \tag{2.9}$$

for $s = 1$.

We defer detailed discussion of our response delay bounds until after we have presented our adaptive procedure in Section 2.3.3.

2.3.2 Sparse case

We now assume that $p \geq 2$, and analyse the performance of $N = N\big(\sqrt{8\log(p-1)}\big)$; in other words, we choose $a = \sqrt{8\log(p-1)}$. This choice turns out to work particularly well when the vector of mean change is sparse in the sense that $s(\theta) \leq \sqrt{p}$, though again we do not assume this in our theory. The motivation for this choice of a comes from the fact that, for fixed b and j, we have $\Lambda_b^{j',j} \mid \tau_b^j \overset{\mathrm{iid}}{\sim} \mathcal{N}(0, \tau_b^j)$ for $j' \in [p] \setminus \{j\}$ under the null. Since a is the threshold level for $|\Lambda_b^{j',j}|/\sqrt{\tau_b^j}$, it is therefore natural to choose a to be of the same order as the maximum absolute value of $p-1$ independent and identically distributed $\mathcal{N}(0,1)$ random variables. The declaration threshold T^{off} is determined based on Lemma 2.20. Theorem 2.3 below shows that, in the sparse case, the patience of our procedure is also guaranteed to be at least at the nominal level $\gamma > 0$. In addition, as in the dense case, we can also control the response delay of $\texttt{ocd}'$ according to Theorem 2.4.

Theorem 2.3. *Let $X_1, X_2, \ldots$ be generated according to $\mathbb{P}_0$. For any $\gamma \geq 1$, let $(X_t)_{t\in\mathbb{N}}$, $\beta > 0$, $a = \sqrt{8\log(p-1)}$, $T^{\mathrm{diag}} = \log\{16p\gamma\log_2(4p)\}$ and $T^{\mathrm{off}} = 8\log\{16p\gamma\log_2(2p)\}$ be the inputs of Algorithm 2.2, with corresponding output N. Then $\mathbb{E}_0(N) \geq \gamma$.*

Theorem 2.4. *Assume that $X_1, X_2, \ldots$ are generated according to $\mathbb{P}_{z,\theta}$ for some z and θ such that $\|\theta\|_2 = \vartheta \geq \beta > 0$ and that θ has an effective sparsity of $s := s(\theta)$. Then there exists a universal constant $C > 0$, such that the output N from Algorithm 2.2, with inputs $(X_t)_{t\in\mathbb{N}}$, $\beta > 0$, $a = \sqrt{8\log(p-1)}$, $T^{\mathrm{diag}} = \log\{16p\gamma\log_2(4p)\}$ and $T^{\mathrm{off}} = 8\log\{16p\gamma\log_2(2p)\}$, satisfies*

$$\bar{\mathbb{E}}_\theta(N) \leq \bar{\mathbb{E}}_\theta^{\mathrm{wc}}(N) \leq C\biggl\{\frac{s\log(ep\gamma)\log(ep)}{\beta^2} \vee 1\biggr\}. \tag{2.10}$$

Comparing Theorems 2.2 and 2.4, we see that the thresholding induced by the non-zero choice of $a = \sqrt{8\log(p-1)}$ in Theorem 2.4 facilitates an improved dependence on the effective sparsity s in the bound on the response delay, whenever s is of smaller order than $\sqrt{p}$.

2.3.3 Adaptive procedure

To adapt to different sparsity levels s, we can run `ocd` (or `ocd′`) with two values of a simultaneously: we choose $a = a^{\mathrm{dense}} = 0$ to form the off-diagonal dense statistic $S^{\mathrm{off,d}} = S^{\mathrm{off}}(a^{\mathrm{dense}})$, and $a = a^{\mathrm{sparse}} = \sqrt{8\log(p-1)}$ to form the off-diagonal sparse statistic $S^{\mathrm{off,s}} = S^{\mathrm{off}}(a^{\mathrm{sparse}})$. We recall that the diagonal statistic S^{diag} does not depend on the choice of a. For clarity, we redefine the three stopping times here: $N^{\mathrm{diag}} := \inf\{n : S_n^{\mathrm{diag}} \geq T^{\mathrm{diag}}\}$, $N^{\mathrm{off,d}} := \inf\{n : S_n^{\mathrm{off,d}} \geq T^{\mathrm{off,d}}\}$ and $N^{\mathrm{off,s}} := \inf\{n : S_n^{\mathrm{off,s}} \geq T^{\mathrm{off,s}}\}$, for appropriately-chosen thresholds T^{diag}, $T^{\mathrm{off,d}}$ and $T^{\mathrm{off,s}}$. The output of this adaptive procedure is thus $N = N^{\mathrm{diag}} \wedge N^{\mathrm{off,d}} \wedge N^{\mathrm{off,s}}$.

The following results provide patience and response delay guarantees for this adaptive procedure.

Theorem 2.5. *Let $X_1, X_2, \ldots$ be generated according to $\mathbb{P}_0$. For any $\gamma \geq 1$, let $(X_t)_{t\in\mathbb{N}}$, $\beta > 0$, $T^{\mathrm{diag}} = \log\{24p\gamma\log_2(4p)\}$, $T^{\mathrm{off,d}} = \psi(\tilde{T}^{\mathrm{off,d}})$ with $\tilde{T}^{\mathrm{off,d}} = 2\log\{24p\gamma\log_2(2p)\}$ and $T^{\mathrm{off,s}} = 8\log\{24p\gamma\log_2(2p)\}$ be the inputs of the adaptive version of Algorithm 2.2, with corresponding output N. Then $\mathbb{E}_0(N) \geq \gamma$.*

Theorem 2.6. *Assume that $X_1, X_2, \ldots$ are generated according to $\mathbb{P}_{z,\theta}$ for some z and θ such that $\|\theta\|_2 = \vartheta \geq \beta > 0$ and that θ has an effective sparsity of $s := s(\theta)$. Then there exists a universal constant $C > 0$, such that the output N from the adaptive version of Algorithm 2.2, with inputs $(X_t)_{t\in\mathbb{N}}$, $\beta > 0$, $T^{\mathrm{diag}} = \log\{24p\gamma\log_2(4p)\}$, $T^{\mathrm{off,d}} = \psi(\tilde{T}^{\mathrm{off,d}})$ with $\tilde{T}^{\mathrm{off,d}} = 2\log\{24p\gamma\log_2(2p)\}$ and $T^{\mathrm{off,s}} = 8\log\{24p\gamma\log_2(2p)\}$, satisfies*

$$\bar{\mathbb{E}}_\theta^{\mathrm{wc}}(N) \leq C\biggl\{\frac{s\log(ep\gamma)\log(ep)}{\beta^2} \vee 1\biggr\}. \tag{2.11}$$

Furthermore, there exists $\beta_0(s) \in (0, 1/2]$, depending only on s, such that for all $\beta \leq \beta_0(s)$, the output N satisfies

$$\bar{\mathbb{E}}_\theta(N) \leq C\biggl\{\biggl(\frac{\sqrt{p}\log(ep\gamma)}{\vartheta^2} \vee \frac{\sqrt{s}\log(ep\beta^{-1})\log(ep)}{\beta^2}\biggr) \wedge \frac{s\log(ep\gamma)\log(ep)}{\beta^2}\biggr\}, \tag{2.12}$$

for $s \geq 2$*, and*

$$\bar{\mathbb{E}}_\theta(N) \leq \frac{C \log(ep\gamma) \log(ep)}{\beta^2}, \tag{2.13}$$

for $s = 1$*.*

Comparing these two results with the corresponding theorems in Sections 2.3.1 and 2.3.2, we see that by choosing slightly more conservative thresholds, the adaptive procedure retains the nominal patience control while (up to constant factors) achieving the best of both worlds in terms of its response delay guarantees under different sparsity regimes.

To better understand the worst-case and average-case response delay bounds in Theorem 2.6, it is helpful to assume that $\vartheta/C_1 \leq \beta \leq \vartheta \leq C_1$ and $\log(\gamma/\beta) \leq C_2 \log p$ for some $C_1, C_2 > 0$. Under these additional assumptions, the result of Theorem 2.6 takes the simpler form that for some $C > 0$, depending only on C_1 and C_2, we have

$$\bar{\mathbb{E}}_\theta^{\mathrm{wc}}(N) \leq \frac{Cs \log^2(ep)}{\vartheta^2} \quad \text{and} \quad \bar{\mathbb{E}}_\theta(N) \leq \frac{C(s \wedge p^{1/2}) \log^2(ep)}{\vartheta^2}.$$

In particular, the average-case response delay upper bound exhibits a phase transition when the effective sparsity level s is of order $\sqrt{p}$, which is the boundary between the sparse and dense cases. Similar sparsity-related elbow effects have been observed in the minimax rate for high-dimensional Gaussian mean testing (Collier, Comminges and Tsybakov, 2017) and the corresponding offline changepoint detection problem (Liu, Gao and Samworth, 2021). On the other hand, we note that quadratic dependence on ϑ in the denominator, and the logarithmic dependence on γ in the numerator, are known to be optimal in the case when $p = 1$ (Lorden, 1971, Theorem 3). The different dependencies on sparsity of the worst-case and average-case response delays for the dense, sparse and adaptive versions of ocd′ are illustrated in Figure 2.2.

2.3.4 Relaxation of assumptions

The setting we consider for our theoretical results, with independent Gaussian observations having identity covariance matrix, is convenient for facilitating a relatively clean presentation and to clarify the main ideas behind the ocd procedure. Nevertheless, it is of interest to consider more general data generating mechanisms, where these assumptions are relaxed. Focusing on the dense case for simplicity of exposition, the Gaussianity assumption ensures that our aggregated statistics have chi-squared distributions (under the null) or non-central chi-squared distributions (under the alternative), so we can apply existing sharp tail bounds. If, instead, our observations have sub-Gaussian distributions, then the corresponding statistics would have sub-Gamma distributions, in the terminology of Boucheron, Lugosi and Massart (2013), so Bernstein's inequality could be applied to give alternative bounds in this setting. Another place where we make use of the Gaussianity assumption is in comparing the trajectories of our

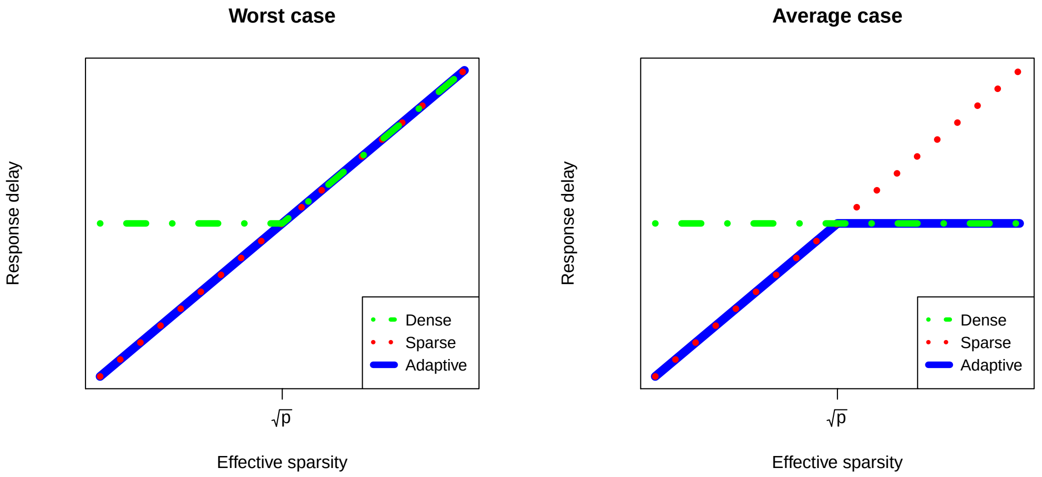

Fig. 2.2 Illustration of the dependencies on sparsity of the worst-case and average-case response delays for the dense, sparse and adaptive versions of ocd′, as given by Theorems 2.2, 2.4 and 2.6.

test statistics with a Brownian motion with drift (see, for instance, the proof of Lemma 2.16). Since we can view these trajectories as discrete Gaussian random walks, we can establish direct inequalities in this comparison. If we were to relax the Gaussianity, then we would need to rely on Donsker's invariance principle, or preferably its finite-sample version given by the Hungarian embedding (Komlós, Major and Tusnády, 1976).

In cases where the covariance matrix of the observations were unknown, it may be possible to estimate this using a training sample, known to come from the null hypothesis, and use this to pre-whiten the data. The form of the estimator to be used should be chosen to exploit any known dependence structure (e.g. banding, Toeplitz or tapering) between the different coordinates. Similar remarks apply when there is short-range serial (temporal) dependence between successive observations. In Section 2.4.4, we demonstrate one way of handling temporal dependence with real data, by studying the residuals of the fit of an autoregressive model.

2.4 Numerical studies

In this section, we study the empirical performance of the ocd algorithm and compare it with other online changepoint detection methods. Recall that the (adaptive) ocd algorithm declares a change when any of the three statistics S^{diag}, $S^{\mathrm{off,d}}$ and $S^{\mathrm{off,s}}$ exceeds their respective thresholds T^{diag}, $T^{\mathrm{off,d}}$ and $T^{\mathrm{off,s}}$. If a priori knowledge about the signal sparsity is available, it may be slightly preferable to use $N^{\mathrm{diag}} \wedge N^{\mathrm{off,d}}$ in the dense case, and $N^{\mathrm{diag}} \wedge N^{\mathrm{off,s}}$ in the sparse case, but for simplicity of exposition, we will focus on the adaptive version of our `ocd` procedure throughout the remainder of this section. While the threshold choices given in Theorem 2.5 guarantee that the patience of (adaptive) `ocd` will be at least at the nominal level, in practice, they may be conservative. We therefore describe a scheme for practical choice of thresholds in Section 2.4.1. Recall that, in order to form $S^{\mathrm{off,d}}$ and $S^{\mathrm{off,s}}$, two different entrywise hard thresholds for $A_b^{j',j}/\sqrt{t_b^j \vee 1}$ need to be specified. For $S^{\mathrm{off,d}}$, we choose $a = 0$ for both theoretical analysis and practical usage. For $S^{\mathrm{off,s}}$, the theoretical choice is $a = \sqrt{8\log(p-1)}$, but since this is also slightly conservative, the choice of $a = \sqrt{2\log p}$ is used in our practical implementation of the algorithm, and our numerical simulations below.

2.4.1 Practical choice of declaration thresholds

The purpose of this section is to introduce an alternative to using the theoretical thresholds T^{diag}, $T^{\mathrm{off,d}}$ and $T^{\mathrm{off,s}}$ provided by Theorem 2.5, namely to determine the thresholds through Monte Carlo simulation. The basic idea is that since the null distribution is known, we can simulate from it to determine the patience for any given choice of thresholds. A complicating issue is the fact that the choices of the three thresholds T^{diag}, $T^{\mathrm{off,d}}$ and $T^{\mathrm{off,s}}$ are related, so that we may be able to achieve the same patience by increasing T^{diag} and decreasing $T^{\mathrm{off,d}}$, for example. To handle this, we first argue that the renewal nature of the processes involved means that, at least for moderately large thresholds, the times to exceedence for each of the three statistics S^{diag}, $S^{\mathrm{off,d}}$ and $S^{\mathrm{off,s}}$ are approximately exponentially distributed. Evidence to support this is provided by Figure 2.3, where we present QQ-plots of $N^{\mathrm{diag}}/m(N^{\mathrm{diag}})$, $N^{\mathrm{off,d}}/m(N^{\mathrm{off,d}})$ and $N^{\mathrm{off,s}}/m(N^{\mathrm{off,s}})$, where the $m(N)$ statistics are empirical medians of the corresponding N statistics (divided by $\log 2$) over 200 repetitions.

We can therefore set an individual Monte Carlo threshold for S^{diag} as follows (the other two statistics can be handled in identical fashion): for $r \in [B]$, simulate $X_1^{(r)}, \ldots, X_\gamma^{(r)} \overset{\mathrm{iid}}{\sim} \mathcal{N}_p(0, I_p)$ and for each $n \in [\gamma]$, compute the diagonal statistic $S_n^{\mathrm{diag},(r)}$ on the rth sample. Now compute $V^{(r)} := \max_{1\leq n\leq \gamma} S_n^{\mathrm{diag},(r)}$, and take $\tilde{T}^{\mathrm{diag}}$ to be the $(1/e)$th quantile of $\{V^{(r)} : r \in [B]\}$. The rationale for the final step here is that if $\mathbb{P}_0(V^{(1)} < \tilde{T}^{\mathrm{diag}}) = 1/e$, then $\mathbb{P}_0(\tilde{N}^{\mathrm{diag}} > \gamma) = 1/e$, where $\tilde{N}^{\mathrm{diag}} := \min\{n : S_n^{\mathrm{diag}} \geq \tilde{T}^{\mathrm{diag}}\}$. Thus, under an exponential distribution for $\tilde{N}^{\mathrm{diag}}$, we have that $\tilde{N}^{\mathrm{diag}}$ has individual patience γ.

Having determined appropriate thresholds $\tilde{T}^{\mathrm{diag}}$, $\tilde{T}^{\mathrm{off,d}}$ and $\tilde{T}^{\mathrm{off,s}}$, we can then use similar ideas to set a suitable combined threshold T^{comb}. In particular, we also argue that $N^{\mathrm{diag}} \wedge N^{\mathrm{off,d}} \wedge N^{\mathrm{off,s}}$ has an approximate exponential distribution; see Figure 2.3 for supporting evidence. We therefore proceed as follows: for $r \in [B]$, simulate $\tilde{X}_1^{(r)}, \ldots, \tilde{X}_\gamma^{(r)} \overset{\mathrm{iid}}{\sim} \mathcal{N}_p(0, I_p)$ and use this new data to compute $\tilde{S}_n^{\mathrm{diag},(r)} := S_n^{\mathrm{diag},(r)}/\tilde{T}^{\mathrm{diag}}$, $\tilde{S}_n^{\mathrm{off,d},(r)} := S_n^{\mathrm{off,d},(r)}/\tilde{T}^{\mathrm{off,d}}$ and $\tilde{S}_n^{\mathrm{off,s},(r)} := S_n^{\mathrm{off,s},(r)}/\tilde{T}^{\mathrm{off,s}}$ for each $n \in [\gamma]$, and set $W^{(r)} := \max\bigl\{\tilde{S}_n^{\mathrm{diag},(r)} \vee \tilde{S}_n^{\mathrm{off,d},(r)} \vee \tilde{S}_n^{\mathrm{off,s},(r)} : n \in [\gamma]\bigr\}$ on the rth sample. Now take T^{comb} to be the $(1/e)$th quantile of $\{W^{(r)} : r \in [B]\}$. Similar to before, our reasoning here is that if $\mathbb{P}_0(W^{(1)} < T^{\mathrm{comb}}) = 1/e$, then $N^{\mathrm{diag}} := \min\bigl\{n : S_n^{\mathrm{diag}} \geq \tilde{T}^{\mathrm{diag}} T^{\mathrm{comb}}\bigr\}$, $N^{\mathrm{off,d}} := \min\bigl\{n : S_n^{\mathrm{off,d}} \geq \tilde{T}^{\mathrm{off,d}} T^{\mathrm{comb}}\bigr\}$ and $N^{\mathrm{off,s}} := \min\bigl\{n : S_n^{\mathrm{off,s}} \geq \tilde{T}^{\mathrm{off,s}} T^{\mathrm{comb}}\bigr\}$ satisfy

$$\mathbb{P}_0\Bigl(N^{\mathrm{diag}} \wedge N^{\mathrm{off,d}} \wedge N^{\mathrm{off,s}} > \gamma\Bigr) = 1/e.$$

Thus, under an exponential distribution for $N^{\mathrm{diag}} \wedge N^{\mathrm{off,d}} \wedge N^{\mathrm{off,s}}$, it again has the desired nominal patience. Our practical thresholds, therefore, are $T^{\mathrm{diag}} = \tilde{T}^{\mathrm{diag}} T^{\mathrm{comb}}$, $T^{\mathrm{off,d}} = \tilde{T}^{\mathrm{off,d}} T^{\mathrm{comb}}$ and $T^{\mathrm{off,s}} = \tilde{T}^{\mathrm{off,s}} T^{\mathrm{comb}}$ for S^{diag}, $S^{\mathrm{off,d}}$ and $S^{\mathrm{off,s}}$ respectively. Table 2.1 confirms that, with these choices of Monte Carlo thresholds, the patience of the adaptive `ocd` algorithm remains at approximately the desired nominal level.

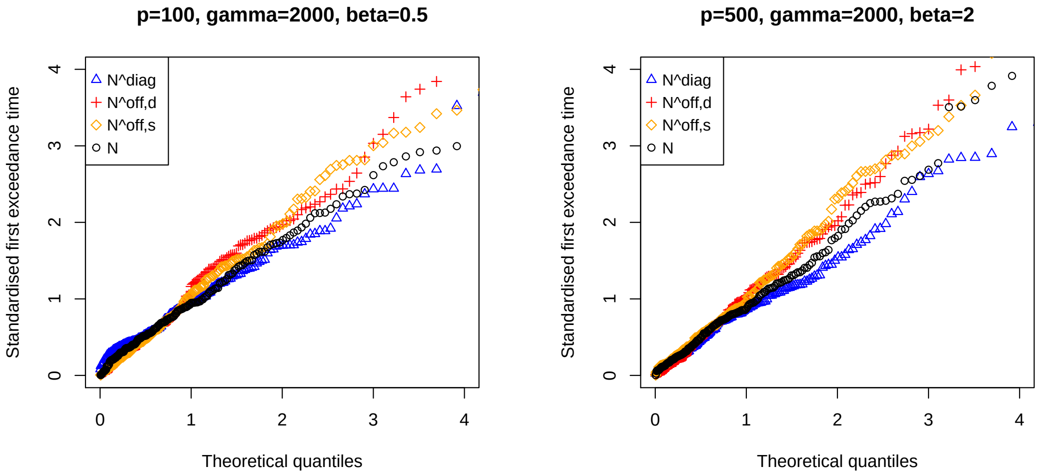

Fig. 2.3 QQ-plots of standardised versions of N^{diag}, $N^{\mathrm{off,d}}$ and $N^{\mathrm{off,s}}$, as well as $N = N^{\mathrm{diag}} \wedge N^{\mathrm{off,d}} \wedge N^{\mathrm{off,s}}$, against theoretical Exp(1) quantiles.

Table 2.1 Estimated run lengths under the null using the Monte Carlo thresholds described in Section 2.4.1 over 500 repetitions, with desired patience level $\gamma = 5000$. Algorithm is terminated after 20000 data points for each repetition. Each reported value is the average run length taken over the repetitions which have already declared prior to time 20000. For reference, $\mathbb{E}(X \mid X < 20000) \approx 4626.9$ when $X \sim \mathrm{Exp}(1/5000)$.

	$p = 100$	$p = 1000$
$\beta = 2$	4606.2	4480.8
$\beta = 1/2$	5291.5	4383.6

Table 2.2 Estimated response delays over 200 repetitions for N^{diag}, $N^{\mathrm{off,d}}$ and $N^{\mathrm{off,s}}$ and the response delay of the combined declaration time N for `ocd`, with the percentages of repetitions on which each statistics triggers the declaration first (or equal first) shown in parentheses. The quickest response in each setting is given in bold. Other parameters: $p = 100$, $\gamma = 5000$, $z = 0$ and $\theta = \vartheta U$, where the distribution of U is described in Section 2.4.2.

		$\beta = \vartheta$			
s	ϑ	N^{diag}	$N^{\mathrm{off,d}}$	$N^{\mathrm{off,s}}$	N
1	2	**11.5 (83.5)**	19.4 (1.5)	13.0 (35)	11.2
1	1	**40.6 (79.5)**	74.4 (1.5)	47.4 (19)	39.1
1	0.5	**136.3 (82)**	305.2 (1)	169.2 (17)	129.7
1	0.25	**455.4 (83)**	1124.5 (1)	635.0 (16)	433.6
10	2	20.1 (9.5)	19.2 (9.5)	**14.7 (88)**	14.3
10	1	69.7 (15.5)	72.6 (12)	**52.4 (73.5)**	50.4
10	0.5	240.4 (29.5)	308.0 (3)	**207.7 (68)**	197.1
10	0.25	**723.3 (56.5)**	1124.3 (6)	760.7 (37.5)	648.4
100	2	53.3 (0.5)	**19.7 (92)**	27.4 (10)	19.5
100	1	169.9 (2)	**75.2 (85)**	94.9 (14.5)	73.1
100	0.5	544.1 (9)	**300.6 (75.5)**	345.1 (15.5)	278.9
100	0.25	1493.6 (28.5)	**1206.0 (51.5)**	1420.2 (20)	1065.4

Table 2.3 Estimated response delays over 200 repetitions for N^{diag}, $N^{\mathrm{off,d}}$ and $N^{\mathrm{off,s}}$ and the response delay of the combined declaration time N for `ocd`. Settings where β is both over- and under-specified are given. The quickest response in each setting is given in bold. Other parameters: $p = 100$, $\gamma = 5000$, $z = 0$ and $\theta = \vartheta U$, where the distribution of U is described in Section 2.4.2.

		$\beta = 4\vartheta$				$\beta = \vartheta/4$			
s	ϑ	N^{diag}	$N^{\mathrm{off,d}}$	$N^{\mathrm{off,s}}$	N	N^{diag}	$N^{\mathrm{off,d}}$	$N^{\mathrm{off,s}}$	N
1	2	**7.7**	19.5	12.8	7.6	30.3	19.5	**12.6**	12.6
1	1	**27.8**	77.7	48.3	27.6	98.3	73.7	**45.2**	45.1
1	0.5	**92.9**	288.9	162.0	92.3	304.8	304.9	**171.8**	171.1
1	0.25	**351.7**	1148.7	657.2	342.8	746.7	1158.1	**614.0**	586.7
10	2	16.7	19.0	**14.9**	13.7	50.0	20.4	**15.1**	15.0
10	1	57.6	72.9	**51.2**	46.5	161.9	76.5	**54.7**	54.5
10	0.5	228.3	286.4	**201.0**	180.5	509.0	314.7	**203.6**	201.8
10	0.25	**739.3**	1175.1	787.9	645.1	1208.2	1189.6	**725.1**	715.9
100	2	59.2	**18.9**	25.3	18.7	110.8	**21.2**	27.2	20.5
100	1	213.9	**73.0**	92.4	71.0	347.4	**76.8**	95.5	74.2
100	0.5	696.5	**307.0**	385.0	284.8	1029.0	**310.2**	352.5	289.3
100	0.25	1811.5	**1218.1**	1327.4	967.1	2149.9	**1091.9**	1175.9	957.8

2.4.2 Numerical performance of `ocd`

In this section, we study the empirical performance of `ocd`. As shown in Figure 2.1, under the alternative, all three statistics S^{diag}, $S^{\mathrm{off,d}}$ and $S^{\mathrm{off,s}}$ in `ocd` can be the first to trigger a declaration that a mean change has occurred. We thus examine different settings under which each of these three statistics can respectively be the quickest to react to a change. Our simulations were run for $p = 100, s \in \{1, \lfloor p^{1/2} \rfloor, p\}, z \in \{0, 1000\}, \gamma = 5000, \vartheta \in \{2, 1, 0.5, 0.25\}, \beta \in \{\vartheta, 4\vartheta, \vartheta/4\}$. In all cases, θ was generated as ϑU, where U is uniformly distributed on the union of all s-sparse unit spheres in $\mathbb{R}^p$. By this, we mean that we first generate a uniformly random subset S of $[p]$ of cardinality s, then set $U := Z/\|Z\|_2$, where $Z = (Z^1, \ldots, Z^p)^\top$ has independent components satisfying $Z^j \sim \mathcal{N}(0,1)\mathbb{1}_{\{j \in S\}}$. Instead of terminating the `ocd` procedure once one of the three statistics declares a change (as we would in practice), we run the procedure until all three statistics have exceeded their respective thresholds. Tables 2.2 and 2.3 summarise the performance of the three statistics for $z = 0$. Simulation results for $z = 1000$ were similar, and are therefore not included here.

We first discuss the case when β is correctly specified (Table 2.2). When the sparsity s is small or moderate and ϑ is small, the diagonal statistic S^{diag} is likely to be the first to declare a change. The response delay of S^{diag} increases with s, which means that the off-diagonal sparse statistic $S^{\mathrm{off,s}}$ typically reacts quickest to a change when the s is moderate to large and ϑ is not too small. On the other hand, the stopping time $N^{\mathrm{off,d}}$, which is driven by the off-diagonal dense statistic, is not significantly affected by s (in agreement with our average-case bound in Theorem 2.2), and is usually the dominant statistic when the signal is dense. A further observation is that the three individual response delays, as well as the combined response delay, are all approximately proportional to ϑ^{-2}, a phenomenon which is supported by Theorem 2.6.

Table 2.3 presents corresponding results when β is both over- and under-specified. We note that both $N^{\mathrm{off,d}}$ and $N^{\mathrm{off,s}}$ are almost unaffected by either type of misspecification. For N^{diag}, a mild over-misspecification of β helps it to react faster, while an under-misspecification causes it to have increased response delay. However, since we can also observe that N^{diag} rarely declares first by a large margin, the performance of `ocd` is highly robust to misspecification of β, especially when s is not too small.

2.4.3 Comparison with other methods

We now compare our adaptive `ocd` algorithm with other online changepoint detection algorithms proposed in the literature, namely those of Mei (2010), Xie and Siegmund (2013) and Chan (2017). Since we were unable to find publicly-available implementations of any of these algorithms, we briefly describe below their methodology and the small adaptations that we made in order to allow them to be used in our settings.

Recall from Section 1.2.5 that Mei (2010) assumes knowledge of θ, and, on observing each new data point, aggregates likelihood ratio tests in each coordinate of the null $\mathcal{N}(0,1)$ against an alternative of $\mathcal{N}(\theta^j,1)$ in the jth coordinate. In our setting where we do not know θ and only assume that $\|\theta\|_2 \geq \beta$, we replace the original statistics $\sum_{j\in[p]} R^j_{n,\theta^j}$ and $\max_{j\in[p]} R^j_{n,\theta^j}$ with

$$\max\biggl\{\sum_{j=1}^p R^j_{n,\beta/\sqrt{p}}, \sum_{j=1}^p R^j_{n,-\beta/\sqrt{p}}\biggr\} \quad \text{and} \quad \max\Bigl\{\max_{j\in[p]} R^j_{n,\beta/\sqrt{p}}, \max_{j\in[p]} R^j_{n,-\beta/\sqrt{p}}\Bigr\}$$

respectively.

We now recall the algorithms of Xie and Siegmund (2013) and Chan (2017) from Chapter 1. Let $B^1,\ldots,B^p \overset{\mathrm{iid}}{\sim} \mathrm{Bernoulli}(p_0)$ for some known $p_0 \in [0,1]$. Both methods consider testing (applied to tail sequences for the purpose of changepoint detection) the null $(X_i)_{i\in\mathbb{N}}$ where $X_i \overset{\mathrm{iid}}{\sim} \mathcal{N}_p(0,I_p)$ against an alternative of a mixture distribution, where for each coordinate $j \in [p]$, independently, $(X_i^j)_{i\in\mathbb{N}}$ satisfies

$$X_i^j \mid B^j \overset{\mathrm{iid}}{\sim} \mathcal{N}\bigl(\mu^j \mathbb{1}_{\{B^j=1\}}, 1\bigr)$$

for some unknown $\mu^j \in \mathbb{R}$. Specifically, writing $Z^j_{n,r} := r^{-1/2}\sum_{i=n-r+1}^n X_i^j$ for $n \in \mathbb{N}$, $r \in [n]$ and $j \in [p]$, and with a pre-specified window size w, the test statistics are of the form

$$S^+_{\mathrm{XS,C}}(p_0,\lambda,\kappa,w) := \max_{r\in[w\wedge n]} \sum_{j=1}^p \log\biggl(1 - p_0 + \lambda p_0 e^{(Z^j_{n,r}\vee 0)^2/\kappa}\biggr),$$

where Xie and Siegmund (2013) take $(\lambda,\kappa,w) = (1,2,200)$ and Chan (2017) takes $(\lambda,\kappa,w) = (2\sqrt{2}-2,4,200)$. Since such a test statistic is only effective when $\sum_{j\in[p]}(\mu^j \vee 0)^2$ is large, we considered statistics of the form $S^+_{\mathrm{XS,C}}(p_0,\lambda,\kappa,w) \vee S^-_{\mathrm{XS,C}}(p_0,\lambda,\kappa,w)$, where $S^-_{\mathrm{XS,C}}(p_0,\lambda,\kappa,w)$ replaces the exponent $Z^j_{n,r} \vee 0$ with $Z^j_{n,r} \wedge 0$. An adaptive choice of p_0 is not provided by the authors, but the choices $p_0 \in \{0.1, 1/\sqrt{p}, 1\}$ have been considered; we found the choice $p_0 = 1/\sqrt{p}$ to be the most competitive overall, as seen in Table 2.4.

For each of the Mei (2010), Xie and Siegmund (2013) and Chan (2017) algorithms, we determined appropriate thresholds using Monte Carlo simulation, as suggested by the authors, and in a similar fashion to the way in which we set the `ocd` thresholds as described in Section 2.4.1. This guarantees that the algorithms have approximately the nominal patience, and so allows us to compare the methods by means of the response delay.

Table 2.5 displays the response delays for the `ocd` algorithm, as well as the alternative methods described above, for $p \in \{100, 2000\}$, $s \in \{5, \lfloor\sqrt{p}\rfloor, p\}$ and $\vartheta \in \{2, 1, 0.5, 0.25\}$. In fact, we also ran simulations for $p = 1000$, $s \in \{1, p/2\}$ and $\vartheta = 0.125$, but the results are qualitatively similar and are therefore omitted. Overall, the results reveal that `ocd` performs very well in comparison with existing methods, across a wide range of scenarios; in several

Table 2.4 Estimated response delay for the algorithms of Xie and Siegmund (2013) (`XS`) and Chan (2017) (`Chan`) over 200 repetitions, with $z = 0$, $\gamma = 5000$, $w = 200$, $p_0 \in \{0.1, 1/\sqrt{p}, 1\}$ and θ generated as described in Section 2.4.2. The smallest response delays are given in bold.

			`XS`			`Chan`		
p	s	ϑ	$p_0 = 0.1$	$p_0 = 1/\sqrt{p}$	$p_0 = 1$	$p_0 = 0.1$	$p_0 = 1/\sqrt{p}$	$p_0 = 1$
100	5	2	**13.1**	**13.1**	18.4	**11.9**	**11.9**	18.4
100	5	1	**47.3**	**47.3**	75.1	**42.0**	**42.0**	75.1
100	5	0.5	**194.3**	**194.3**	413.6	**163.7**	**163.7**	411.9
100	10	2	**15.2**	**15.2**	19.9	**14.5**	**14.5**	19.8
100	10	1	**52.9**	**52.9**	72.1	**51.5**	**51.5**	72.0
100	10	0.5	**255.8**	**255.8**	513.1	**245.6**	**245.6**	508.7
100	100	2	23.6	23.6	**22.9**	27.5	27.5	**22.9**
100	100	1	102.1	102.1	**84.1**	89.6	89.6	**84.1**
100	100	0.5	**526.8**	**526.8**	657.9	756.0	756.0	**647.5**
1000	5	2	23.0	**17.9**	49.0	17.2	**13.6**	52.3
1000	5	1	88.5	**67.9**	224.3	62.9	**52.4**	281.0
1000	5	0.5	781.5	**419.4**	2814.9	411.1	**236.5**	5627.1
1000	31	2	35.1	**31.3**	57.6	30.6	**29.8**	61.6
1000	31	1	133.5	**113.6**	297.8	113.3	**106.9**	364.7
1000	31	0.5	2129.1	**1692.7**	2678.6	1842.9	**1377.9**	5685.2
1000	1000	2	63.2	73.4	**59.2**	66.1	89.1	**65.0**
1000	1000	1	325.4	459.3	**296.4**	**355.4**	720.4	418.8
1000	1000	0.5	**2968.4**	3698.9	3355.8	**3090.8**	3846.6	6439.8
2000	5	2	30.2	**20.8**	65.0	20.7	**15.6**	67.6
2000	5	1	113.9	**79.5**	447.2	78.6	**59.5**	586.4
2000	5	0.5	1380.2	**607.7**	4333.9	570.2	**285.0**	6040.2
2000	44	2	45.0	**40.2**	75.3	38.8	**37.7**	79.1
2000	44	1	191.8	**149.1**	625.7	154.6	**145.0**	830.3
2000	44	0.5	3115.4	**2945.4**	4046.7	**2634.1**	2751.4	6066.1
2000	2000	2	89.3	103.2	**83.9**	93.2	136.7	**88.2**
2000	2000	1	**722.4**	1020.0	746.9	**765.9**	2074.7	967.0
2000	2000	0.5	**3326.7**	4669.3	4007.2	**3139.9**	4672.7	6197.0

cases it is by far the most responsive procedure, and in none of the settings considered is it outperformed by much. The Xie and Siegmund (2013) and Chan (2017) algorithms perform similarly to each other, and in most settings are both more competitive than the Mei (2010) method described above. We note that the performance of the Xie and Siegmund (2013) and Chan (2017) algorithms is relatively better when the signal-to-noise ratio ϑ is high; in these scenarios, the default window size $w = 200$ is large enough that sufficient evidence against the null can typically be accumulated within the moving window. For lower signal-to-noise ratios, this ceases to be the case, and from time $z + w$ onwards, the test statistic has the same marginal distribution (with no positive drift). This explains the relative deterioration in performance for those algorithms in the harder settings considered. As mentioned in the introduction, if the change in mean were known to be small, then the window size could be increased to compensate, but at additional computational expense; a further advantage of `ocd`, then, is that the computational time only depends on p (and not on β or other problem parameters). We remark that, in terms of the running time and CPU cost per new observation of the algorithm, the Mei (2010) algorithm is the fastest due to the simple nature of its aggregation technique. As discussed above, the computational time of the Xie and Siegmund (2013) and Chan (2017) algorithms depends on the choice of w, and is slower than that of the Mei (2010) algorithm when a large enough w is chosen such that the detection delays are relatively small. The computational time of our `ocd` algorithm can be slightly longer when the dimension p is quite large, but nonetheless the algorithm is able to process thousands of observations within a few seconds when $p = 1000$.

2.4.4 Real data example

We consider a seismic signal detection problem, using a dataset from the High Resolution Seismic Network, operated by the Berkeley Seismological Laboratory. Ground motion sensor measurements were recorded using geophones at a frequency of 250 Hz in three mutually perpendicular directions, at 13 stations near Parkfield, California for a total of 740 seconds from 2am on 23 December 2004. This dataset was also studied by Xie, Xie and Moustakides (2019), and was obtained from http://service.ncedc.org/fdsnws/dataselect/1/. To begin, we removed the linear trend in each coordinate and applied a 2–16 Hz bandpass filter to the data using the GISMO toolbox[1]; these are standard pre-processing steps in the seismology literature (e.g. Caudron et al., 2018; Xie, Xie and Moustakides, 2019). In order to reduce the effects of temporal dependence, we computed a root mean square amplitude envelope, downsampled to 16 Hz, and then extracted the residuals from the fit of an autoregressive model of order 1. The processed data are available as a built-in dataset in the `ocd` `R` package. The first four minutes of the series were used to estimate the baseline mean and variance for each sensor, and we plot the standardised data from 2:04am onwards in Figure 2.4. When

[1] Available at: http://geoscience-community-codes.github.io/GISMO/

Table 2.5 Estimated response delay for `ocd`, as well as the algorithms of Mei (2010) (`Mei`), Xie and Siegmund (2013) (`XS`) and Chan (2017) (`Chan`) over 200 repetitions, with $z = 0$, $\gamma = 5000$ and θ generated as described in Section 2.4.2. The smallest response delay is given in bold.

p	s	ϑ	`ocd`	`Mei`	`XS`	`Chan`
100	5	2	13.7	36.3	13.1	**11.9**
100	5	1	46.9	125.9	47.3	**42.0**
100	5	0.5	174.8	383.1	194.3	**163.7**
100	5	0.25	**583.5**	970.4	2147	1888.8
100	10	2	14.9	44.1	15.2	**14.5**
100	10	1	53.8	150.1	52.9	**51.5**
100	10	0.5	**194.4**	458.2	255.8	245.6
100	10	0.25	**629.7**	1171.3	2730.7	2484.9
100	100	2	**19.4**	72.7	23.6	27.5
100	100	1	**74.4**	268.3	89.6	102.1
100	100	0.5	**287.9**	834.9	526.8	756.0
100	100	0.25	**1005.8**	1912.9	3598.3	3406.6
2000	5	2	19.0	130.5	20.8	**15.6**
2000	5	1	67.3	316.7	79.5	**59.5**
2000	5	0.5	**247.3**	680.2	607.7	285.0
2000	5	0.25	**851.3**	1384.8	4459.2	3856.9
2000	44	2	**37.5**	247.7	40.2	37.7
2000	44	1	**136.0**	596.1	149.1	145.0
2000	44	0.5	**479.1**	1270.8	2945.5	2751.4
2000	44	0.25	**1584.2**	2428.8	4457.8	5049.7
2000	2000	2	**97.1**	949.9	103.2	136.7
2000	2000	1	**360.7**	2126.5	1020.0	2074.7
2000	2000	0.5	**1296.0**	3428.1	4669.3	4672.7
2000	2000	0.25	**3436.7**	4140.4	5063.7	5233.5

applying our ocd algorithm to this data, we specified the patience level to be $\gamma = 1.35 \times 10^6$, corresponding to a patience of one day, and $\beta = 150$. The ocd algorithm declared a change at 02:10:03.84, and was triggered by $S^{\mathrm{off,d}}$. According to the Northern California Earthquake Catalog[2], an earthquake of magnitude 1.47 Md hit near Atascadero, California (50 km away from Parkfield) at 02:09:54.01, so the delay was 9.8 seconds. It is known[3] that P waves, which are the primary preliminary wave and arrive first after an earthquake, travel at up to 6 km/s in the Earth's crust, which is consistent with this delay.

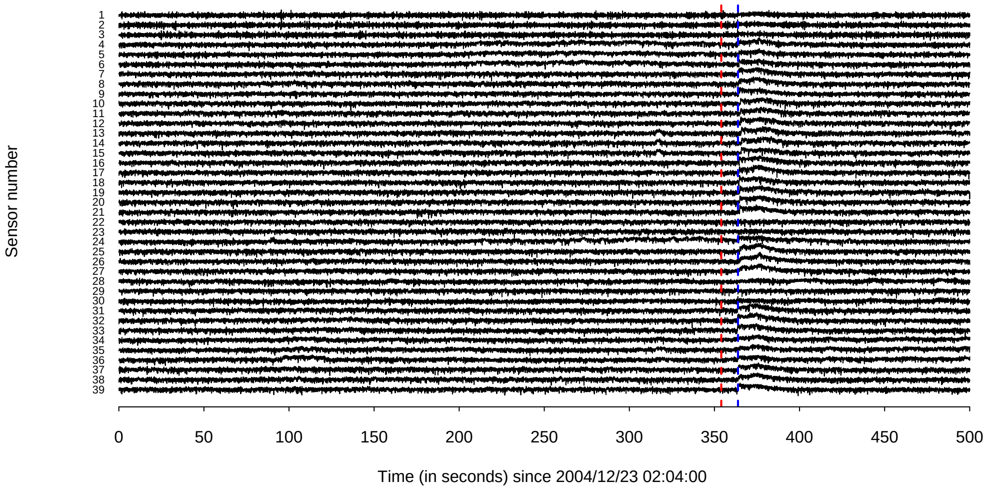

Fig. 2.4 Standardised, pre-processed earthquake data from 39 sensors. The time of the 1.47 Md earthquake is given by the vertical red dashed line, while time of ocd declaration of change is given as a blue dashed line.

2.5 Proofs of main results

2.5.1 Proofs from Section 2.3.1

Proof of Theorem 2.1. Define $m := \lfloor 2\gamma \rfloor$. It suffices to prove that (a) $\mathbb{P}_0(N^{\mathrm{off}} \leq m) \leq 1/4$ and (b) $\mathbb{P}_0(N^{\mathrm{diag}} \leq m) \leq 1/4$, since then we have

$$\begin{aligned}\mathbb{E}_0(N) = \mathbb{E}_0(N^{\mathrm{off}} \wedge N^{\mathrm{diag}}) &\geq 2\gamma \mathbb{P}_0(N^{\mathrm{off}} \wedge N^{\mathrm{diag}} > 2\gamma) \\ &\geq 2\gamma\bigl\{1 - \mathbb{P}_0(N^{\mathrm{off}} \leq m) - \mathbb{P}_0(N^{\mathrm{diag}} \leq m)\bigr\} \geq \gamma.\end{aligned}$$

[2] Available at: http://www.ncedc.org/ncedc/catalog-search.html.

[3] One source for this information is https://www.usgs.gov/natural-hazards/earthquake-hazards/science/seismographs-keeping-track-earthquakes.

We prove the two claims below.

(a) By (2.6) and a union bound, we have

$$\mathbb{P}_0(N^{\mathrm{off}} \leq m) \leq \sum_{\substack{n\in[m],j\in[p]\\ b\in\mathcal{B}}} \mathbb{P}_0\Big(Q^j_{n,b} \geq T^{\mathrm{off}}\Big) = \sum_{\substack{n\in[m],j\in[p]\\ b\in\mathcal{B}}} \mathbb{E}_0\Big[\mathbb{P}_0\Big(Q^j_{n,b} \geq T^{\mathrm{off}} \;\Big|\; \tau^j_{n,b}\Big)\Big]. \tag{2.14}$$

Recall that under the null, $\Lambda^{k,j}_b \mid \tau^j_b \overset{\mathrm{iid}}{\sim} \mathcal{N}(0, \tau^j_b)$ for all $b \in \mathcal{B}, j \in [p]$ and $k \in [p]\backslash\{j\}$, which implies that $Q^j_b \mid \tau^j_b \sim \chi^2_{p-1}\mathbb{1}_{\{\tau^j_b > 0\}}$. Thus, we have by Laurent and Massart (2000, Lemma 1) that for all $n \in [m], j \in [p]$ and $b \in \mathcal{B}$,

$$\mathbb{P}_0\Big(Q^j_{n,b} \geq T^{\mathrm{off}} \;\Big|\; \tau^j_{n,b}\Big) \leq e^{-\tilde{T}^{\mathrm{off}}/2}. \tag{2.15}$$

Combining (2.14) and (2.15), we have

$$\mathbb{P}_0(N^{\mathrm{off}} \leq m) \leq |\mathcal{B}|mpe^{-\tilde{T}^{\mathrm{off}}/2} \leq 1/4. \tag{2.16}$$

(b) For $j \in [p]$ and $b \in \mathcal{B} \cup \mathcal{B}_0$, denote $N^j_b := \inf\{n : R^j_{n,b} \geq T^{\mathrm{diag}}\}$, where $R^j_{n,b}$ is defined by (2.2). By Lemma 2.10, we have that $R^j_{n,b} = \{R^j_{n-1,b} + b(X^j_n - b/2)\} \vee 0$, and that this process is always non-negative. Then $N^{\mathrm{diag}} = \min\big\{N^j_b : j \in [p], b \in \mathcal{B} \cup \mathcal{B}_0\big\}$.

Now, fix some $j \in [p]$ and $b \in \mathcal{B} \cup \mathcal{B}_0$. Define $U_0 := 0$ and $U_h := \inf\{n > U_{h-1} : R^j_{n,b} \notin (0, T^{\mathrm{diag}})\}$ for $h \in \mathbb{N}$, and let $H := \inf\{h : R^j_{U_h,b} \geq T^{\mathrm{diag}}\}$. Then

$$N^j_b = U_H \geq H.$$

To study the distribution of H, consider the one-sided sequential probability ratio test of $H_{0,Z} : Z_1, Z_2, \ldots \overset{\mathrm{iid}}{\sim} \mathcal{N}(0,1)$ against $H_{1,Z} : Z_1, Z_2, \ldots \overset{\mathrm{iid}}{\sim} \mathcal{N}(b,1)$ with log-boundaries T^{diag} and $-\infty$. The associated stopping time for this test is

$$N_{\mathrm{os}} := \inf\bigg\{n \in \mathbb{N} : b\sum_{t=1}^{n}(Z_t - b/2) \geq T^{\mathrm{diag}}\bigg\}.$$

Since $(R^j_{n,b})_n$ is a Markov process that renews itself every time it hits 0, H follows a geometric distribution with success probability

$$\mathbb{P}_0(R^j_{U_1,b} \geq T^{\mathrm{diag}}) \leq \mathbb{P}_{H_{0,Z}}(N_{\mathrm{os}} < \infty) \leq e^{-T^{\mathrm{diag}}},$$

where the last inequality follows from Lemma 2.12. Consequently,

$$\mathbb{P}_0(N^j_b \leq m) \leq \mathbb{P}_0(H \leq m) \leq 1 - \Big(1 - e^{-T^{\mathrm{diag}}}\Big)^m.$$

As the above inequality holds for all $j \in [p]$ and $b \in \mathcal{B} \cup \mathcal{B}_0$, we have that

$$\begin{aligned}\mathbb{P}_0(N^{\mathrm{diag}} > m) &= \mathbb{P}_0\biggl(\bigcap_{j\in[p],b\in\mathcal{B}\cup\mathcal{B}_0}\{N_b^j > m\}\biggr) = \prod_{j\in[p]}\biggl\{1 - \mathbb{P}_0\biggl(\bigcup_{b\in\mathcal{B}\cup\mathcal{B}_0}\{N_b^j \le m\}\biggr)\biggr\} \\ &\ge \Bigl[1 - |\mathcal{B}\cup\mathcal{B}_0|\Bigl\{1 - \bigl(1 - e^{-T^{\mathrm{diag}}}\bigr)^m\Bigr\}\Bigr]^p \ge 1 - mp|\mathcal{B}\cup\mathcal{B}_0|e^{-T^{\mathrm{diag}}} \ge 3/4,\end{aligned} \tag{2.17}$$

as desired, where in the penultimate inequality, we twice used the fact that $(1-x)^\alpha \ge 1 - \alpha x$ for all $\alpha \ge 1$ and $x \in [0,1]$. □

The proof of Theorem 2.2 is quite involved. We first define some relevant quantities, and then state and prove some preliminary results. For $\theta \in \mathbb{R}^p$ with effective sparsity $s(\theta)$, there is at most one coordinate in θ of magnitude larger than $\vartheta/\sqrt{2}$, so there exists $b_* \in \bigl\{\beta/\sqrt{s(\theta)\log_2(2p)}, -\beta/\sqrt{s(\theta)\log_2(2p)}\bigr\} \subseteq \mathcal{B}$ such that

$$\mathcal{J} := \Bigl\{j \in [p] : \theta^j/b_* \ge 1 \text{ and } |\theta^j| \le \vartheta/\sqrt{2}\Bigr\} \tag{2.18}$$

has cardinality at least $s(\theta)/2$ (note that the condition $\theta^j/b_* \ge 1$ above ensures that $\{\theta^j : j \in \mathcal{J}\}$ all have the same sign as b_*). Both b_* and $\mathcal{J}$ can be chosen as functions of θ. Now, given any sequence $X_1, X_2, \ldots \in \mathbb{R}^p$ and $\theta \in \mathbb{R}^p$, define for any $\alpha \in (0,1]$ the function

$$q(\alpha) = q(\alpha; X_1, \ldots, X_z, \theta) := \inf\bigl\{y \in \mathbb{R} : \bigl|\{j \in \mathcal{J} : t_{z,b_*}^j \le y\}\bigr| \ge \alpha|\mathcal{J}|\bigr\}, \tag{2.19}$$

where t_{z,b_*}^j is obtained by running Algorithm 2.2 up to time z with $a = 0$ and $T^{\mathrm{diag}} = T^{\mathrm{off}} = \infty$. In other words, $q(\alpha)$ is the empirical α-quantile of the tail lengths $(t_{z,b_*}^j : j \in \mathcal{J})$ when we run the algorithm without declaring any change up to time z. Recall the definition of the function ψ in (2.6).

Proposition 2.7. *Assume that $X_1, X_2, \ldots$ are generated according to $\mathbb{P}_{z,\theta}$ for some z and θ such that $\|\theta\|_2 = \vartheta \ge \beta > 0$ and that θ has an effective sparsity of $s := s(\theta) \ge 2$. Then the output N from Algorithm 2.2, with input $(X_t)_{t\in\mathbb{N}}$, $\beta \in \mathbb{R}^p$, $a = 0$, $T^{\mathrm{diag}} \ge 1$ and $T^{\mathrm{off}} = \psi(\tilde{T}^{\mathrm{off}})$ for $\tilde{T}^{\mathrm{off}} \ge \log(ep)$, satisfies*

$$\mathbb{E}_{z,\theta}\bigl\{(N-z) \vee 0 \bigm| X_1, \ldots, X_z\bigr\} \le \frac{396\tilde{T}^{\mathrm{off}} + 65\sqrt{p\tilde{T}^{\mathrm{off}}}}{\vartheta^2} + \frac{24\log_2(2p)}{\alpha\beta^2} + 3q(\alpha) + 2, \tag{2.20}$$

for any $\alpha \in (0,1]$.

Proof. Since the bound in (2.20) is positive, we may, throughout the proof and for arbitrary $z \in \mathbb{N}$, restrict attention to realisations $X_1 = x_1, \ldots, X_z = x_z$ for which we have not declared a change by time z. In other words, we have $N > z$. This restriction also ensures that $q(\alpha)$

defined in (2.19) is now indeed the empirical α-quantile of the tail lengths $(t^j_{z,b^*} : j \in \mathcal{J})$ at the changepoint. Denote $\mathcal{J}_\alpha := \bigl\{j \in \mathcal{J} : t^j_{z,b_*} \le q(\alpha)\bigr\}$. Then we have $|\mathcal{J}_\alpha| \ge \alpha|\mathcal{J}| \ge \alpha s/2$.

We now fix some

$$r \ge \left\{ \frac{12\bigl(\tilde{T}^{\mathrm{off}} + \sqrt{2(p-1)\tilde{T}^{\mathrm{off}}}\bigr)}{\vartheta^2} \vee 3q(\alpha) \right\} + 2 =: r_0. \tag{2.21}$$

Note that $r_0 > 3q(\alpha) \ge 3t^j_{z,b_*}$ for all $j \in \mathcal{J}_\alpha$. For $j \in \mathcal{J}_\alpha$, we define the event

$$\Omega^j_r := \Bigl\{ t^j_{z+\lfloor r \rfloor, b_*} > 2\lfloor r \rfloor/3 \Bigr\}.$$

Under $\mathbb{P}_{z,\theta}$, conditional on $X_1 = x_1, \ldots, X_z = x_z$, we know that $X_{z+1}, X_{z+2}, \ldots \overset{\mathrm{iid}}{\sim} \mathcal{N}_p(\theta, I_p)$. Hence, by using Lemma 2.11 and applying Lemma 2.16(b) to $t^j_{z+\lfloor r \rfloor,b_*} \wedge \lfloor r \rfloor$ for $j \in \mathcal{J}_\alpha$, we obtain

$$\mathbb{P}_{z,\theta}\biggl(\bigcap_{j \in \mathcal{J}_\alpha} (\Omega^j_r)^c \biggm| X_1 = x_1, \ldots, X_z = x_z \biggr) \le \exp\bigl\{-|\mathcal{J}_\alpha| b_*^2 \lfloor r \rfloor/12\bigr\} \le \exp\{-\alpha s b_*^2 \lfloor r \rfloor/24\}. \tag{2.22}$$

We now work on the event Ω^j_r, for some $j \in \mathcal{J}_\alpha$. We note that (2.21) guarantees that $r \ge 2$, and thus $t^j_{z+\lfloor r \rfloor,b_*} \ge \lceil 2\lfloor r \rfloor/3 \rceil \ge 2$. Then, by Lemma 2.19 and the fact that $r_0 > 3t^j_{z,b_*}$, we have that

$$\frac{\lfloor r \rfloor}{3} < \frac{t^j_{z+\lfloor r \rfloor,b_*}}{2} \le \tau^j_{z+\lfloor r \rfloor,b_*} \le \frac{3t^j_{z+\lfloor r \rfloor,b_*}}{4} \le \frac{3\bigl(t^j_{z,b_*} + r\bigr)}{4} < r.$$

Hence we conclude that on the event Ω^j_r,

$$2/3 \le \lfloor r \rfloor/3 < \tau^j_{z+\lfloor r \rfloor,b_*} \le \lfloor r \rfloor. \tag{2.23}$$

Recall that $\Lambda^{\cdot,j}_{z+\lfloor r \rfloor,b_*} \in \mathbb{R}^p$ records the tail CUSUM statistics with tail length $\tau^j_{z+\lfloor r \rfloor,b_*}$. We observe by (2.23) that on Ω^j_r, only post-change observations are included in $\Lambda^{\cdot,j}_{z+\lfloor r \rfloor,b_*}$. Hence we have that on the event Ω^j_r,

$$\Lambda^{k,j}_{z+\lfloor r \rfloor,b_*} \bigm| \bigl\{\tau^j_{z+\lfloor r \rfloor,b_*}, X_1 = x_1, \ldots, X_z = x_z\bigr\} \overset{\mathrm{ind}}{\sim} \mathcal{N}\bigl(\theta^k \tau^j_{z+\lfloor r \rfloor,b_*}, \tau^j_{z+\lfloor r \rfloor,b_*}\bigr) \tag{2.24}$$

for $k \in [p]\backslash\{j\}$. Therefore, on the event Ω^j_r and conditional on $\tau^j_{z+\lfloor r \rfloor,b_*}, X_1 = x_1, \ldots, X_z = x_z$, the random variable $\frac{\|\Lambda_b^{-j,j}\|_2^2}{\tau^j_{z+\lfloor r \rfloor,b_*} \vee 1} = \frac{\|\Lambda_b^{-j,j}\|_2^2}{\tau^j_{z+\lfloor r \rfloor,b_*}}$ follows a non-central chi-squared distribution with $p-1$ degrees of freedom and noncentrality parameter $\|\theta^{-j}\|_2^2 \tau^j_{z+\lfloor r \rfloor,b_*}$. Since $j \in \mathcal{J}$ and $s \ge 2$, we observe, by (2.18) and (2.23) that $\|\theta^{-j}\|_2^2\tau^j_{z+\lfloor r\rfloor,b_*} \geq \vartheta^2\lfloor r\rfloor/6$ on Ω^j_r. Write

$$E^j_r := \biggl\{\frac{\|\Lambda^{-j,j}_{z+\lfloor r\rfloor,b_*}\|_2^2}{\tau^j_{z+\lfloor r\rfloor,b_*}\vee 1} < T^{\mathrm{off}}\biggr\}.$$

Then by Birgé (2001, Lemma 8.1), we have

$$\begin{aligned}\mathbb{P}_{z,\theta}\Bigl(E^j_r\cap\Omega^j_r \mid \tau^j_{z+\lfloor r\rfloor,b_*},X_1 = x_1,\ldots,X_z = x_z\Bigr)&\\ \leq \exp\Biggl\{-\frac{\Bigl(\vartheta^2\lfloor r\rfloor/6 - \tilde{T}^{\mathrm{off}} - \sqrt{2(p-1)\tilde{T}^{\mathrm{off}}}\Bigr)^2}{4\bigl(p-1+\vartheta^2\lfloor r\rfloor/3\bigr)}\Biggr\}&. \qquad (2.25)\end{aligned}$$

Combining (2.22) and (2.25), we deduce that

$$\begin{aligned}&\mathbb{P}_{z,\theta}\bigl(N > z+r \mid X_1 = x_1,\ldots,X_z = x_z\bigr) \leq \mathbb{P}_{z,\theta}\bigl(N > z+\lfloor r\rfloor \mid X_1 = x_1,\ldots,X_z = x_z\bigr)\\ &\leq \mathbb{P}_{z,\theta}\biggl(\bigcap_{j\in\mathcal{J}_\alpha}(\Omega^j_r)^c \biggm| X_1 = x_1,\ldots,X_z = x_z\biggr) + \sum_{j\in\mathcal{J}_\alpha}\mathbb{P}_{z,\theta}\Bigl(E^j_r\cap\Omega^j_r \Bigm| X_1 = x_1,\ldots,X_z = x_z\Bigr)\\ &\leq \exp\biggl\{-\frac{\alpha s b_*^2(r-1)}{24}\biggr\} + p\exp\Biggl\{-\frac{\Bigl(\vartheta^2(r-1)/6 - \tilde{T}^{\mathrm{off}} - \sqrt{2(p-1)\tilde{T}^{\mathrm{off}}}\Bigr)^2}{4\bigl(p-1+\vartheta^2(r-1)/3\bigr)}\Biggr\}\\ &\leq \exp\biggl\{-\frac{\alpha s b_*^2(r-1)}{24}\biggr\} + p\exp\biggl\{-\frac{\vartheta^4(r-1)^2}{576\bigl(p-1+\vartheta^2(r-1)/3\bigr)}\biggr\},\end{aligned}$$

where the last inequality uses (2.21). Therefore, we have

$$\begin{aligned}&\mathbb{E}_{z,\theta}\bigl\{(N-z)\vee 0 \mid X_1 = x_1,\ldots,X_z = x_z\bigr\}\\ &= \int_0^\infty \mathbb{P}_{z,\theta}\bigl(N > z+u \mid X_1 = x_1,\ldots,X_z = x_z\bigr)\,du\\ &\leq r_0 + \int_{r_0-1}^\infty \biggl[\exp\biggl\{-\frac{\alpha s b_*^2 u}{24}\biggr\} + p\exp\biggl\{-\frac{\vartheta^4 u^2}{576\bigl(p-1+\vartheta^2 u/3\bigr)}\biggr\}\biggr]\wedge 1\,du\\ &\leq r_0 + \frac{24}{\alpha s b_*^2} + \int_0^\infty\bigl(pe^{-\vartheta^2 u/384}\bigr)\wedge 1\,du + \int_0^\infty\Bigl(pe^{-\frac{\vartheta^4 u^2}{1152(p-1)}}\Bigr)\wedge 1\,du\\ &\leq r_0 + \frac{24}{\alpha s b_*^2} + \frac{384\log(ep)}{\vartheta^2} + \frac{24\sqrt{2(p-1)\log p}}{\vartheta^2} + \frac{12\sqrt{2\pi(p-1)}}{\vartheta^2}\\ &\leq r_0 + \frac{24}{\alpha s b_*^2} + \frac{384\log(ep)}{\vartheta^2} + \frac{48\sqrt{(p-1)\log(ep)}}{\vartheta^2},\end{aligned}$$

where the penultimate inequality follows from the fact that $1-\Phi(x) \leq \frac{1}{2}e^{-x^2/2}$ for $x \geq 0$. The desired bound (2.20) follows by substituting in the expressions for r_0 and b_*. □

The following two propositions control the residual tail length quantile term $q(\alpha)$ in (2.20) in the worst-case and average-case scenarios respectively.

Proposition 2.8. *Let $X_1, X_2, \ldots$, z, θ, s, a, p and N be defined as in Proposition 2.7. On the event $\{N > z\}$, we have*

$$q(1; X_1, \ldots, X_z, \theta) \leq \frac{8T^{\mathrm{diag}} s \log_2(2p)}{\beta^2}.$$

Proof. We will show the stronger result that on the event $\{N > z\}$, we have

$$t^j_{z,b} < \frac{8T^{\mathrm{diag}}}{b^2}$$

for all $b \in \mathcal{B}$ and $j \in [p]$. The desired result then follows immediately by taking $b = b_*$ and restricting to the subset $\mathcal{J} \subseteq [p]$.

Fix $b \in \mathcal{B}$ and $j \in [p]$. Recall from (2.2) and Lemma 2.10 the definition of $R^j_{n,b}$ and the recursive relation $R^j_{n,b} = \{R^j_{n-1,b} + b(X^j_n - b/2)\} \vee 0$. By the update procedure for $t^j_{n,b}$ in Algorithm 2.2 and Lemma 2.11, we have

$$R^j_{n,b} \begin{cases} = 0 & \text{when } n = z - t^j_{z,b}, \\ > 0 & \text{when } z - t^j_{z,b} < n \leq z. \end{cases} \tag{2.26}$$

We claim that

$$R^j_{n,b/2} \geq \frac{R^j_{n,b}}{2} + \frac{b^2(n - z + t^j_{z,b})}{8}, \tag{2.27}$$

for all $n \in \{z - t^j_{z,b}, \ldots, z\}$. To see this, the claim is true when $n = z - t^j_{z,b}$ since the right hand side of (2.27) is 0 by (2.26). Now, assume (2.27) is true for some $n = m - 1$. Then,

$$\begin{aligned} R^j_{m,b/2} \geq R^j_{m-1,b/2} + \frac{b}{2}\Big(X^j_m - \frac{b}{4}\Big) &\geq \frac{R^j_{m-1,b}}{2} + \frac{b^2(m - 1 - z + t^j_{z,b})}{8} + \frac{b}{2}\Big(X^j_m - \frac{b}{4}\Big) \\ &= \frac{R^j_{m,b}}{2} + \frac{b^2(m - z + t^j_{z,b})}{8}. \end{aligned}$$

This proves the claim by induction. In particular, on the event $\{N > z\}$, we have $T^{\mathrm{diag}} > R^j_{z,b/2} > b^2 t^j_{z,b}/8$ as desired. □

Proposition 2.9. *Let $X_1, X_2, \ldots$, z, θ, s, a, p and N be defined as in Proposition 2.7. There exists a universal constant C and $\beta_0(s) > 0$, depending only on s, such that for all $\beta < \beta_0(s)$, we have*

$$\mathbb{E}_{z,\theta}\{q(s^{-1/2}; X_1, \ldots, X_z, \theta)\} \leq \frac{C s^{1/2} \log(16 s^2 \beta^{-2} \log_2(2p)) \log_2(2p)}{\beta^2}.$$

Proof. Recall the definition of b_* in (2.18). We may assume, without loss of generality that $b_* = \beta/\sqrt{s \log_2(2p)}$ (the case $b_* = -\beta/\sqrt{s \log_2(2p)}$ can be proved in essentially the same way). We first prove the result for sufficiently large $s > s_0$. Recall that $t^j_{z,b_*} =$

$\mathrm{argmax}_{0\leq r\leq z}\sum_{i=z-r+1}^{z}(X_i^j - b_*/2)$. Define $Z_i := X_{z-i+1}$ for $i \in [z]$ and let $Z_{z+1}, Z_{z+2}, \ldots \overset{\mathrm{iid}}{\sim} \mathcal{N}_p(0, I_p)$ be independent from $Z_1, \ldots, Z_z$. For each $j \in [p]$, let

$$S_r^j := \sum_{i=1}^{r}\bigl(Z_i^j - b_*/2\bigr) \quad \text{and} \quad \tilde{S}_r^j := \sum_{i=1}^{r} Z_i^j$$

for $r \in \mathbb{N}$ and define $S_0^j := \tilde{S}_0^j := 0$. Writing $\xi_0^j := \mathrm{argmax}_{0\leq r\leq \Delta b_*^{-2}} S_r^j$, $\xi^j := \mathrm{argmax}_{r\in\mathbb{N}_0} S_r^j$, and $\tilde{\xi}_0^j := \mathrm{argmax}_{0\leq r\leq \Delta b_*^{-2}} \tilde{S}_r^j$, where $\Delta := 8\log(2s)$, we note that like t_{z,b_*}^j, these three maxima are also uniquely attained almost surely (see the proof of Lemma 2.16). By construction, we have for each $j \in [p]$ that

$$t_{z,b_*}^j = \underset{0\leq r\leq z}{\mathrm{argmax}} \sum_{i=z-r+1}^{z} (X_i^j - b_*/2) = \underset{0\leq r\leq z}{\mathrm{argmax}}\, S_r^j \leq \underset{r\in\mathbb{N}_0}{\mathrm{argmax}}\, S_r^j = \xi^j.$$

Writing $q_\xi(\alpha) := \inf\bigl\{y : |\{j \in \mathcal{J} : \xi^j \leq y\}| \geq \alpha|\mathcal{J}|\bigr\}$ as the empirical α-quantile of $(\xi^j : j \in \mathcal{J})$, it follows that $q(\alpha) \leq q_\xi(\alpha)$ and so it suffices to control $\mathbb{E}\{q_\xi(s^{-1/2})\}$ instead of $\mathbb{E}\{q(s^{-1/2})\}$. To this end, we observe that $\bigl\{16\Delta s^{-1/2} b_*^2 < \xi^j \leq \Delta b_*^{-2}\bigr\} \subseteq \bigl\{16\Delta s^{-1/2} b_*^{-2} < \xi_0^j \leq \Delta b_*^{-2}\bigr\}$ and $\tilde{\xi}_0^j \geq \xi_0^j$, and thus

$$\begin{aligned}\mathbb{P}\bigl(\xi^j \leq 16\Delta s^{-1/2} b_*^{-2}\bigr) &\geq \mathbb{P}\bigl(\xi_0^j \leq 16\Delta s^{-1/2} b_*^{-2}\bigr) - \mathbb{P}\bigl(\xi^j > \Delta b_*^{-2}\bigr) \\ &\geq \mathbb{P}\bigl(\tilde{\xi}_0^j \leq 16\Delta s^{-1/2} b_*^{-2}\bigr) - \mathbb{P}\bigl(\xi^j > \Delta b_*^{-2}\bigr). \end{aligned} \tag{2.28}$$

For the first term on the right hand side of (2.28), by Donsker's invariance principle and the continuity of the argmax map (see, e.g. van der Vaart and Wellner, 1996, Lemma 3.2.1 and Theorem 3.2.2), we have in the limit $\beta \searrow 0$ that $\Delta b_*^{-2} \to \infty$ and so

$$\frac{\tilde{\xi}_0^j}{\Delta b_*^{-2}} \xrightarrow{\mathrm{d}} \underset{t\in[0,1]}{\mathrm{argmax}}\, B_t,$$

where $(B_t)_{t\geq 0}$ denotes a standard Brownian motion. In particular, we can find $\beta_0(s) > 0$ depending only on s such that for $\beta \leq \beta_0(s)$ and $s > 256$, we have

$$\mathbb{P}\bigl(\tilde{\xi}_0^j \leq 16\Delta s^{-1/2} b_*^{-2}\bigr) \geq \frac{1}{2}\mathbb{P}\Bigl(\underset{t\in[0,1]}{\mathrm{argmax}}\, B_t \leq 16 s^{-1/2}\Bigr) = \frac{1}{\pi}\arcsin(4s^{-1/4}) \geq \frac{4s^{-1/4}}{\pi}. \tag{2.29}$$

where in the second step we used the arcsine law for Brownian motion (see, e.g. Mörters and Peres, 2010, Theorem 5.26), and in the final step we used the fact that $4s^{-1/4} < 1$.

For the second term on the right-hand side of (2.28), since $\Delta = 8\log(2s)$, for sufficiently large $s \geq s_0$ and sufficiently small $\beta \leq \beta_0(s)$, we have by Lemma 2.16(d) that

$$\mathbb{P}(\xi^j > \Delta b_*^{-2}) \leq 2e^{-\Delta/8} = s^{-1}. \tag{2.30}$$

Substituting (2.29) and (2.30) into (2.28), we have, for all $j \in \mathcal{J}$, that

$$\mathbb{P}\bigl(\xi^j \leq 16\Delta s^{-1/2} b_*^{-2}\bigr) \geq s^{-1/4}.$$

As a result, $\bigl|\{j \in \mathcal{J} : \xi^j \leq 16\Delta s^{-1/2} b_*^{-2}\}\bigr|$ is stochastically larger than $\mathrm{Bin}\bigl(|\mathcal{J}|, s^{-1/4}\bigr)$. Thus, for $s \geq s_0$, we have,

$$\mathbb{P}_{z,\theta}\bigl\{q_\xi(s^{-1/2}) > 16\Delta s^{-1/2} b_*^{-2}\bigr\} \leq \mathbb{P}\Bigl\{\mathrm{Bin}\bigl(|\mathcal{J}|, s^{-1/4}\bigr) \leq s^{-1/2}|\mathcal{J}|\Bigr\} \leq e^{-s^{1/2}/2},$$

where we have used Hoeffding's inequality and the fact that $|\mathcal{J}| \geq s/2$ in the last step. On the other hand, for sufficiently large $s \geq s_0$ and sufficiently small $\beta \leq \beta_0(s)$, we have,

$$\begin{aligned}
\mathbb{E}_{z,\theta}\Bigl\{q_\xi(s^{-1/2}) \Bigm| q_\xi(s^{-1/2}) > 16\Delta s^{-1/2} b_*^{-2}\Bigr\} &\leq \mathbb{E}_{z,\theta}\Bigl\{q_\xi(s^{-1/2}) \Bigm| q_\xi(s^{-1/2}) \geq \Delta b_*^{-2}\Bigr\} \\
\leq \mathbb{E}_{z,\theta}\Bigl\{q_\xi(1) \Bigm| q_\xi(|\mathcal{J}|^{-1}) \geq \Delta b_*^{-2}\Bigr\} &= \mathbb{E}_{z,\theta}\Bigl\{\max_{j\in\mathcal{J}} \xi^j \mid \min_{j\in\mathcal{J}} \xi^j \geq \Delta b_*^{-2}\Bigr\} \\
\leq \frac{61\bigl(\Delta + 4\log(2/b_*)\bigr)}{b_*^2},
\end{aligned}$$

where we have used Lemma 2.17(b) in the second inequality and Lemma 2.16(d) (with $\Delta/4$ taking the role of k and $b_*/2$ taking the role of b there) in the final inequality. As a result,

$$\begin{aligned}
\mathbb{E}_{z,\theta}\bigl\{q(s^{-1/2})\bigr\} \leq \mathbb{E}_{z,\theta}\bigl\{q_\xi(s^{-1/2})\bigr\} &\leq 16\Delta s^{-1/2} b_*^{-2} + 61 e^{-s^{1/2}/2}\bigl(\Delta + 4\log(2/b_*)\bigr) b_*^{-2} \\
&\leq \frac{C s^{1/2} \log(16 s^2 \beta^{-2} \log_2(2p)) \log_2(2p)}{\beta^2},
\end{aligned}$$

where we have used in the final step the fact that $e^{-s^{1/2}/2} \leq s^{-1/2}/100$ for sufficiently large s. This proves the desired result for $s \geq s_0$.

Finally, for $s \leq 256$, we have by Lemma 2.16(c) that, for $\beta < \sqrt{s}/2$,

$$\begin{aligned}
\mathbb{E}_{z,\theta}\bigl\{q(s^{-1/2})\bigr\} \leq \mathbb{E}_{z,\theta}\Bigl\{\max_{j\in\mathcal{J}} \xi^j\Bigr\} &\leq \frac{32 s \log(s^{3/2} \beta^{-1} \log_2^{1/2}(2p)) \log_2(2p)}{\beta^2} \\
&\leq \frac{C s^{1/2} \log(16 s^2 \beta^{-2} \log_2(2p)) \log_2(2p)}{\beta^2},
\end{aligned}$$

and the desired bound then follows. □

We are now in a position to prove Theorem 2.2.

Proof of Theorem 2.2. The proof proceeds with different arguments for the case $s \geq 2$ and the case $s = 1$.

Case 1: $s \geq 2$. Combining Propositions 2.7 (applied with $\alpha = 1$) and 2.8, we have

$$\bar{\mathbb{E}}_\theta^{\mathrm{wc}}(N) \leq \frac{396\tilde{T}^{\mathrm{off}} + 65\sqrt{p\tilde{T}^{\mathrm{off}}}}{\vartheta^2} + \frac{24\log_2(2p)}{\beta^2} + \frac{24T^{\mathrm{diag}}s\log_2(2p)}{\beta^2} + 2.$$

The desired bound (2.7) then follows by substituting in the expression for $\tilde{T}^{\mathrm{off}}$. On the other hand, combining Propositions 2.7 (applied with $\alpha = s^{-1/2}$) and 2.9, we have

$$\bar{\mathbb{E}}_\theta(N) \leq \frac{396\tilde{T}^{\mathrm{off}} + 65\sqrt{p\tilde{T}^{\mathrm{off}}}}{\vartheta^2} + \frac{24\sqrt{s}\log_2(2p)}{\beta^2} + \frac{3Cs^{1/2}\log(16s^2\beta^{-2}\log_2(2p))\log_2(2p)}{\beta^2} + 2,$$

which proves (2.8).

Case 2: $s = 1$. There exists $j_* \in [p]$ such that $|\theta^{j_*}| \geq \vartheta/\sqrt{\log_2(2p)}$, and recall from (2.18) that $b_* := \mathrm{sgn}(\theta^{j_*})\beta/\sqrt{\log_2(2p)} \in \mathcal{B}$. Note that $S_{n,1}^{\mathrm{diag}} = \max_{(j,b)\in[p]\times(\mathcal{B}\cup\mathcal{B}_0)} R_{n,b}^j \geq R_{n,b_*}^{j_*}$. We define $\bar{R}_n := \sum_{i=z+1}^{z+n} b_*(X_i^{j_*} - b_*/2)$ for $n \in \mathbb{N}_0$. Since $R_{z,b_*}^{j_*} \geq 0 = \bar{R}_0$ and $R_n - R_{n-1} = b_*(X_{z+n}^{j_*} - b_*/2) \leq R_{z+n,b_*}^{j_*} - R_{z+n-1,b_*}^{j_*}$, it follows by induction that $R_{z+n,b_*}^{j_*} \geq \bar{R}_n$ for all $n \in \mathbb{N}_0$. Then, for $n \geq \lceil 4T^{\mathrm{diag}}/(b_*\theta^{j_*})\rceil =: n_0$, we have

$$\begin{aligned}\mathbb{P}_{z,\theta}(N > z+n \mid X_1 = x_1, \ldots, X_z = x_z) &\leq \mathbb{P}_{z,\theta}\bigl(R_{z+n,b_*}^{j_*} \leq T^{\mathrm{diag}} \bigm| X_1 = x_1, \ldots, X_z = x_z\bigr) \\ &\leq \mathbb{P}_{z,\theta}\bigl(\bar{R}_n \leq T^{\mathrm{diag}}\bigr) = \Phi\biggl(-\frac{b_*n(\theta^{j_*} - b_*/2) - T^{\mathrm{diag}}}{n^{1/2}b_*}\biggr) \\ &\leq \frac{1}{2}\exp\biggl\{-\frac{(b_*n\theta^{j_*}/2 - T^{\mathrm{diag}})^2}{2nb_*^2}\biggr\} \leq \frac{1}{2}e^{-n(\theta^{j_*})^2/32}.\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}_{z,\theta}\bigl\{(N-z)\vee 0 \mid X_1 = x_1, \ldots, X_z = x_z\bigr\} &= \sum_{n=0}^{\infty}\mathbb{P}_{z,\theta}(N > z+n \mid X_1 = x_1, \ldots, X_z = x_z) \\ \leq n_0 + \frac{1}{2}\sum_{n=n_0}^{\infty} e^{-n(\theta^{j_*})^2/32} \leq n_0 + \frac{1}{2}\int_0^\infty e^{-u(\theta^{j_*})^2/32}\,du &\leq 1 + \frac{4T^{\mathrm{diag}}}{b_*\theta^{j_*}} + \frac{16}{(\theta^{j_*})^2}. \qquad (2.31)\end{aligned}$$

After substituting in the expressions for b_*, θ^{j_*} and T^{diag}, we see that

$$\bar{\mathbb{E}}_\theta(N) \leq \bar{\mathbb{E}}_\theta^{\mathrm{wc}}(N) \leq 1 + \frac{4\log(16p\gamma\log_2(4p))\log_2(2p)}{\beta\vartheta} + \frac{16\log_2(2p)}{\vartheta^2},$$

which proves both (2.7) and (2.9). □

2.5.2 Proofs from Sections 2.3.2 and 2.3.3

Proof of Theorem 2.3. It suffices to only prove $\mathbb{P}_0(N^{\mathrm{off}} \leq m) \leq 1/4$, since the remaining proof is identical to that of Theorem 2.1.

Since $\Lambda_b^{k,j} \mid \tau_b^j \overset{\text{iid}}{\sim} \mathcal{N}(0, \tau_b^j)$ for all $b \in \mathcal{B}, j \in [p]$ and $k \in [p]\backslash\{j\}$ under the null, by the fact that $T^{\text{off}} \geq 12$ and Lemma 2.20, we have

$$\mathbb{P}_0\big(Q_{n,b}^j \geq T^{\text{off}} \mid \tau_{n,b}^j\big) \leq \mathbb{P}_0\big(Q_{n,b}^j \geq 6 + T^{\text{off}}/2 \mid \tau_{n,b}^j\big) \leq \exp(-T^{\text{off}}/8).$$

Hence, it follows that

$$\mathbb{P}_0(N^{\text{off}} \leq m) \leq |\mathcal{B}|mpe^{-T^{\text{off}}/8} \leq 1/4, \tag{2.32}$$

as desired. □

Proof of Theorem 2.4. We note that the case $s = 1$ in the proof of Theorem 2.2 does not rely on the off-diagonal statistics. Hence (2.31) is still valid here with $a = \sqrt{8\log(p-1)}$ and the last expression in (2.31) again proves the desired bound (2.10). For the case $s \geq 2$, we follow exactly the proof of Proposition 2.7 until (2.24), with the only exception that we now fix, instead of (2.21),

$$r \geq \left\{\frac{24T^{\text{off}}\log_2(2p)}{\vartheta^2} \vee \frac{96s\log_2(2p)\log p}{\vartheta^2} \vee 3q(\alpha)\right\} + 2 =: \tilde{r}_0. \tag{2.33}$$

By the definition of the effective sparsity of θ, for a fixed $j \in \mathcal{J}_\alpha$,

$$\mathcal{L}^j := \left\{j' \in [p] : |\theta^{j'}| \geq \frac{\vartheta}{\sqrt{s\log_2(2p)}} \text{ and } j' \neq j\right\}$$

has cardinality at least $s-1$. On the event Ω_r^j, we have, by (2.23), that for all $k \in \mathcal{L}^j$

$$|\theta^k|\sqrt{\tau_{z+\lfloor r\rfloor,b_*}^j} \geq \sqrt{\frac{\vartheta^2\lfloor r\rfloor}{3s\log_2(2p)}} =: \tilde{a}_r.$$

We then observe, by (2.33), that

$$\tilde{a}_r \geq \sqrt{32\log p} > 2a. \tag{2.34}$$

Now, from (2.24) we have on the event Ω_r^j that, for all $k \in \mathcal{L}^j$,

$$\mathbb{P}_{z,\theta}\left(\Omega_r^j \cap \left\{|\Lambda_{z+\lfloor r\rfloor,b_*}^{k,j}| < \frac{1}{2}\tilde{a}_r\sqrt{\tau_{z+\lfloor r\rfloor,b_*}^j}\right\} \,\middle|\, \tau_{z+\lfloor r\rfloor,b_*}^j, X_1 = x_1, \ldots, X_z = x_z\right) \\ \leq \frac{1}{2}e^{-\tilde{a}_r^2/8} =: q_r.$$

We denote

$$U^j := \left|\left\{k \in \mathcal{L}^j : \left\{|\Lambda_{z+\lfloor r\rfloor,b_*}^{k,j}| < \frac{1}{2}\tilde{a}_r\sqrt{\tau_{z+\lfloor r\rfloor,b_*}^j}\right\}\right\}\right|.$$

Then, by the Chernoff–Hoeffding binomial tail bound (Hoeffding, 1963, Equation (2.1)), we have

$$\begin{aligned}
&\mathbb{P}_{z,\theta}\Bigl(\Omega_r^j \cap \bigl\{U^j \geq |\mathcal{L}^j|/2\bigr\} \Bigm| \tau^j_{z+\lfloor r\rfloor,b_*}, X_1 = x_1, \ldots, X_z = x_z\Bigr)\\
&\leq \exp\biggl\{-\frac{|\mathcal{L}^j|}{2}\log\biggl(\frac{1}{4q_r(1-q_r)}\biggr)\biggr\} \leq \exp\biggl\{|\mathcal{L}^j|\biggl(\frac{\log 2}{2} - \frac{\tilde{a}_r^2}{16}\biggr)\biggr\} \leq \exp\biggl\{-\frac{3|\mathcal{L}^j|\tilde{a}_r^2}{64}\biggr\}\\
&\leq \exp\biggl\{-\frac{\vartheta^2\lfloor r\rfloor}{128\log_2(2p)}\biggr\}, \qquad (2.35)
\end{aligned}$$

where the penultimate inequality follows from (2.34). Now, on the event $\Omega_r^j \cap \bigl\{U^j < |\mathcal{L}^j|/2\bigr\}$, we have

$$\begin{aligned}
&\sum_{j'\in[p]:j'\neq j} \frac{\bigl(\Lambda^{j',j}_{z+\lfloor r\rfloor,b_*}\bigr)^2}{\tau^j_{z+\lfloor r\rfloor,b_*} \vee 1} \mathbb{1}_{\bigl\{|\Lambda^{j',j}_{z+\lfloor r\rfloor,b_*}| \geq a\sqrt{\tau^j_{z+\lfloor r\rfloor,b_*}}\bigr\}}\\
&\geq \sum_{j'\in[p]:j'\neq j} \frac{\bigl(\Lambda^{j',j}_{z+\lfloor r\rfloor,b_*}\bigr)^2}{\tau^j_{z+\lfloor r\rfloor,b_*} \vee 1} \mathbb{1}_{\bigl\{|\Lambda^{j',j}_{z+\lfloor r\rfloor,b_*}| \geq \frac{\tilde{a}_r}{2}\sqrt{\tau^j_{z+\lfloor r\rfloor,b_*}}\bigr\}} \geq \frac{\tilde{a}_r^2}{4}\biggl\{|\mathcal{L}^j| - \biggl(\biggl\lceil\frac{|\mathcal{L}^j|}{2}\biggr\rceil - 1\biggr)\biggr\}\\
&= \frac{\tilde{a}_r^2}{4}\biggl\lceil\frac{|\mathcal{L}^j|+1}{2}\biggr\rceil \geq \frac{\vartheta^2\lfloor r\rfloor}{24\log_2(2p)} \geq T^{\mathrm{off}}, \qquad (2.36)
\end{aligned}$$

where the penultimate inequality uses the fact that $|\mathcal{L}^j| \geq s-1$ and the last inequality follows from (2.33). We now denote

$$\tilde{E}_r^j := \biggl\{\sum_{j'\in[p]:j'\neq j} \frac{\bigl(\Lambda^{j',j}_{z+\lfloor r\rfloor,b_*}\bigr)^2}{\tau^j_{z+\lfloor r\rfloor,b_*} \vee 1} \mathbb{1}_{\bigl\{|\Lambda^{j',j}_{z+\lfloor r\rfloor,b_*}| \geq a\sqrt{\tau^j_{z+\lfloor r\rfloor,b_*}}\bigr\}} < T^{\mathrm{off}}\biggr\}.$$

Combining (2.22), (2.35) and (2.36), we deduce that

$$\begin{aligned}
&\mathbb{P}_{z,\theta}\bigl(N > z + r \mid X_1 = x_1, \ldots, X_z = x_z\bigr) \leq \mathbb{P}_{z,\theta}\bigl(N > z + \lfloor r\rfloor \mid X_1 = x_1, \ldots, X_z = x_z\bigr)\\
&\quad\leq \mathbb{P}_{z,\theta}\biggl(\bigcap_{j\in\mathcal{J}_\alpha} (\Omega_r^j)^c \biggm| X_1 = x_1, \ldots, X_z = x_z\biggr) +\\
&\qquad\qquad\qquad \sum_{j\in\mathcal{J}_\alpha} \mathbb{P}_{z,\theta}\Bigl(\tilde{E}_r^j \cap \Omega_r^j \Bigm| X_1 = x_1, \ldots, X_z = x_z\Bigr)\\
&\quad\leq \mathbb{P}_{z,\theta}\biggl(\bigcap_{j\in\mathcal{J}_\alpha} (\Omega_r^j)^c \biggm| X_1 = x_1, \ldots, X_z = x_z\biggr) +\\
&\qquad\qquad\qquad \sum_{j\in\mathcal{J}_\alpha} \mathbb{P}_{z,\theta}\Bigl(\Omega_r^j \cap \bigl\{U^j \geq |\mathcal{L}^j|/2\bigr\} \Bigm| X_1 = x_1, \ldots, X_z = x_z\Bigr)\\
&\quad\leq \exp\biggl\{-\frac{\alpha s b_*^2(r-1)}{24}\biggr\} + p\exp\biggl\{-\frac{\vartheta^2(r-1)}{128\log_2(2p)}\biggr\}.
\end{aligned}$$

Therefore we have

$$\begin{aligned}
&\mathbb{E}_{z,\theta}\bigl\{(N-z)\vee 0 \bigm| X_1 = x_1,\ldots,X_z = x_z\bigr\} \\
&= \int_0^\infty \mathbb{P}_{z,\theta}\bigl(N > z+u \bigm| X_1 = x_1,\ldots,X_z = x_z\bigr)\,du \\
&\leq \tilde{r}_0 + \int_{\tilde{r}_0-1}^\infty \biggl[\exp\biggl\{-\frac{\alpha s b_*^2 u}{24}\biggr\} + p\exp\biggl\{-\frac{\vartheta^2 u}{128\log_2(2p)}\biggr\}\biggr] \wedge 1\,du \\
&\leq \tilde{r}_0 + \frac{24}{\alpha s b_*^2} + \int_0^\infty \Bigl(pe^{-\frac{\vartheta^2 u}{128\log_2(2p)}}\Bigr)\wedge 1\,du \leq \tilde{r}_0 + \frac{24}{\alpha s b_*^2} + \frac{128\log_2(2p)\log(ep)}{\vartheta^2} \\
&\leq \frac{24T^{\mathrm{off}}\log_2(2p) + 96s\log_2(2p)\log p}{\vartheta^2} + 3q(\alpha) + \frac{24\log_2(2p)}{\alpha\beta^2} + \frac{128\log_2(2p)\log(ep)}{\vartheta^2} + 2.
\end{aligned}$$

Combining this with Proposition 2.8 (applied with $\alpha = 1$), we have, by substituting in the expression for T^{off}, that

$$\bar{\mathbb{E}}_\theta(N) \leq \bar{\mathbb{E}}_\theta^{\mathrm{wc}}(N) \leq C\biggl\{\frac{s\log(ep\gamma)\log(ep)}{\beta^2} \vee 1\biggr\},$$

for some universal constant $C > 0$, as desired. □

Proof of Theorem 2.5. Let $T^{\mathrm{off,d}} = \psi(\tilde{T}^{\mathrm{off,d}})$. Then, similar to (2.16), (2.17) and (2.32), we have

$$\begin{aligned}
\mathbb{P}_0(N^{\mathrm{diag}} \leq m) &\leq mp|\mathcal{B}\cup\mathcal{B}_0|e^{-T^{\mathrm{diag}}} \leq 1/6, \\
\mathbb{P}_0(N^{\mathrm{off,d}} \leq m) &\leq mp|\mathcal{B}|e^{-\tilde{T}^{\mathrm{off,d}}/2} \leq 1/6, \\
\mathbb{P}_0(N^{\mathrm{off,s}} \leq m) &\leq mp|\mathcal{B}|e^{-T^{\mathrm{off,s}}/8} \leq 1/6.
\end{aligned}$$

and hence,

$$\begin{aligned}
\mathbb{E}_0(N) &= \mathbb{E}_0(N^{\mathrm{diag}} \wedge N^{\mathrm{off,d}} \wedge N^{\mathrm{off,s}}) \geq 2\gamma\mathbb{P}_0(N^{\mathrm{diag}} \wedge N^{\mathrm{off,d}} \wedge N^{\mathrm{off,s}} > 2\gamma) \\
&\geq 2\gamma\bigl\{1 - \mathbb{P}_0(N^{\mathrm{diag}} \leq m) - \mathbb{P}_0(N^{\mathrm{off,d}} \leq m) - \mathbb{P}_0(N^{\mathrm{off,s}} \leq m)\bigr\} \geq \gamma,
\end{aligned}$$

as desired. □

Proof of Theorem 2.6. We observe that

$$\begin{aligned}
\bar{\mathbb{E}}_\theta^{\mathrm{wc}}(N) &= \bar{\mathbb{E}}_\theta^{\mathrm{wc}}\bigl[(N^{\mathrm{diag}} \wedge N^{\mathrm{off,d}}) \wedge (N^{\mathrm{diag}} \wedge N^{\mathrm{off,s}})\bigr] \\
&\leq \bar{\mathbb{E}}_\theta^{\mathrm{wc}}\bigl[(N^{\mathrm{diag}} \wedge N^{\mathrm{off,d}})\bigr] \wedge \bar{\mathbb{E}}_\theta^{\mathrm{wc}}\bigl[(N^{\mathrm{diag}} \wedge N^{\mathrm{off,s}})\bigr],
\end{aligned}$$

and similarly for $\bar{\mathbb{E}}_\theta(N)$. The desired bounds (2.11), (2.12) and (2.13) are therefore direct consequences of Theorems 2.2 and 2.4 (note that the different constants in the thresholds only affect the value of the universal constant). □

2.6 Auxiliary results

Lemma 2.10. *For $n \in \mathbb{N}_0$, $b \in \mathcal{B} \cup \mathcal{B}_0$ and $j \in [p]$, we define $R^j_{n,b} := bA^{j,j}_{n,b} - b^2 t^j_{n,b}/2$, where $A_{n,b}$ and $t_{n,b}$ are taken from Algorithm 2.2 in the main text. Then*

$$R^j_{n,b} = \max_{0 \le h \le n} \sum_{i=n-h+1}^{n} b(X^j_i - b/2). \tag{2.37}$$

Proof. We prove the claim by induction on n. The base case $n = 0$ is true since, by definition, $R^j_{0,b} = 0$ and the sum on the right-hand side of (2.37) is empty. Assume (2.37) is true for $n = m - 1$. Then, by the update procedure in Algorithm 2.2 in the main text, we have

$$\begin{aligned} R^j_{m,b} &= \bigl\{R^j_{m-1,b} + b(X^j_m - b/2)\bigr\} \vee 0 = \biggl\{\max_{0 \le h \le m-1} \sum_{i=m-h}^{m-1} b(X^j_i - b/2) + b(X^j_m - b/2)\biggr\} \vee 0 \\ &= \biggl\{\max_{0 \le h \le m-1} \sum_{i=m-h}^{m} b(X^j_i - b/2)\biggr\} \vee 0 = \max_{0 \le h \le m} \sum_{i=m-h+1}^{m} b(X^j_i - b/2), \end{aligned}$$

and the desired result follows. □

Lemma 2.11. *For $n \in \mathbb{N}_0$, $b \in \mathcal{B} \cup \mathcal{B}_0$ and $j \in [p]$, let $t^j_{n,b}$ be defined as in Algorithm 2.2 in the main text and $R^j_{n,b}$ as in Lemma 2.10. Then*

$$t^j_{n,b} = \min\bigl\{0 \le i \le n : R^j_{n-i,b} = 0\bigr\} = \operatorname{sargmax}_{0 \le h \le n} \sum_{i=n-h+1}^{n} b(X^j_i - b/2). \tag{2.38}$$

Proof. We observe from the procedure in Algorithm 2.2 in the main text that $R^j_{n,b} = 0$ if and only if $t^j_{n,b} = 0$ and that $R^j_{n,b} > 0$ if and only if $t^j_{n,b} = t^j_{n-1,b} + 1$. Hence,

$$t^j_{n,b} = n - \max\bigl\{0 \le i \le n : R^j_{i,b} = 0\bigr\} = \min\bigl\{0 \le i \le n : R^j_{n-i,b} = 0\bigr\}.$$

We now prove that $t^j_{n,b} = \operatorname{sargmax}_{0 \le h \le n} \sum_{i=n-h+1}^{n} b(X^j_i - b/2)$ by induction on n. The base case $n = 0$ is true because $t^j_{n,b} = 0$, and the sum on the right-hand side of (2.38) is empty. Assume the claim is true for $n = m - 1$. Then, by the inductive hypothesis and Lemma 2.10,

$$\begin{aligned} t^j_{m,b} &= (t^j_{m-1,b} + 1)\mathbb{1}_{\{R^j_{m,b} > 0\}} = \biggl(\operatorname{sargmax}_{0 \le h \le m-1} \sum_{i=m-h}^{m-1} b(X^j_i - b/2) + 1\biggr)\mathbb{1}_{\{R^j_{m,b} > 0\}} \\ &= \biggl(\operatorname{sargmax}_{1 \le h \le m} \sum_{i=m-h+1}^{m} b(X^j_i - b/2)\biggr)\mathbb{1}_{\bigl\{\max_{0 \le h \le m} \sum_{i=m-h+1}^{m} b(X^j_i - b/2) > 0\bigr\}} \\ &= \operatorname{sargmax}_{0 \le h \le m} \sum_{i=m-h+1}^{m} b(X^j_i - b/2), \end{aligned}$$

and the desired result follows. □

For two distributions P_0 and P_1 on the same measurable space, the sequential probability ratio test of $H_0 : X_1, X_2, \ldots \overset{\text{iid}}{\sim} P_0$ against $H_1 : X_1, X_2, \ldots \overset{\text{iid}}{\sim} P_1$ with log-boundaries $a > 0$ and $b < 0$ is defined as the (extended) stopping time

$$N := \inf\biggl\{ n : \sum_{i=1}^{n} \log \frac{dP_1}{dP_0}(X_i) \notin (b, a) \biggr\},$$

together with the decision rule after stopping that accepts H_0 if $\sum_{i=1}^{N} \log\{(dP_1/dP_0)(X)\} \leq b$ and accepts H_1 if $\sum_{i=1}^{N} \log\{(dP_1/dP_0)(X)\} \geq a$.

Lemma 2.12. *Suppose N is the stopping time associated with the (one-sided) sequential probability ratio test of $H_0 : X_1, X_2, \ldots \overset{\text{iid}}{\sim} P_0$ against $H_1 : X_1, X_2, \ldots \overset{\text{iid}}{\sim} P_1$ with log-boundaries $a > 0$ and $b = -\infty$. Then*

$$\mathbb{P}_0(N < \infty) \leq e^{-a}.$$

Proof. Let $L_n := \prod_{i=1}^{n} (dP_1/dP_0)(X_i)$. On the event $\{N < \infty\}$, we have $L_N \geq e^a$. Therefore,

$$\mathbb{P}_0(N < \infty) = \sum_{n=1}^{\infty} \mathbb{P}_0(N = n) \leq e^{-a} \sum_{n=1}^{\infty} \mathbb{E}_0(L_n \mathbb{1}_{\{N=n\}}) = e^{-a} \sum_{n=0}^{\infty} \mathbb{P}_1(N = n) \leq e^{-a},$$

which proves the desired result. □

Lemma 2.13. *Assume that $X_1, X_2, \ldots$ are generated according to $\mathbb{P}_{z,\theta}$ for some z and θ such that $\|\theta\|_2 = \vartheta \geq \beta > 0$. Then the output N from Algorithm 2.2, with inputs $(X_t)_{t \in \mathbb{N}}$, $\beta > 0$, $a \geq \sqrt{8 \log(p-1)}$, $T^{\mathrm{diag}} = \log\{16 p \gamma \log_2(4p)\}$ and $T^{\mathrm{off}} = 8 \log\{16 p \gamma \log_2(2p)\}$, satisfies*

$$\mathbb{P}_{z,\theta}(N \leq z) \leq \frac{z}{4\gamma}.$$

Proof. This follows from (2.17) in the proof of Theorem 2.1 and (2.32) in the proof of Theorem 2.3. □

Lemma 2.14. *Let X, Y and Z be real-valued random variables. Assume that (X, Y) and Z are independent. Let $P_{Z|Z \leq Y}$ be the conditional distribution of Z given $Z \leq Y$. Then*

$$\mathbb{P}(X \geq Z \mid Y \geq Z) = \int_{\mathbb{R}} \mathbb{P}(X \geq u \mid Y \geq u) \, dP_{Z|Z \leq Y}(u).$$

Proof. Let P_Y and P_Z denote the marginal distribution of Y and Z respectively. Then, by the definition of $P_{Z|Z \leq Y}$, we have

$$P_{Z|Z \leq Y}\bigl([u, \infty)\bigr) = \mathbb{P}(Z \geq u \mid Z \leq Y) = \frac{\mathbb{P}(Y \geq Z \geq u)}{\mathbb{P}(Y \geq Z)} = \frac{\int_{\mathbb{R}^2} \mathbb{1}_{\{y \geq z \geq u\}} \, dP_Y(y) dP_Z(z)}{\mathbb{P}(Y \geq Z)}$$

$$= \frac{\int_u^\infty \mathbb{P}(Y \geq z)\, dP_Z(z)}{\mathbb{P}(Y \geq Z)},$$

where we have used the assumption that Y and Z are independent in the penultimate equality. Hence,

$$\begin{aligned}\int_{\mathbb{R}} \mathbb{P}(X \geq u \mid Y \geq u)\, dP_{Z|Z\leq Y}(u) &= \int_{\mathbb{R}} \mathbb{P}(X \geq u \mid Y \geq u) \frac{\mathbb{P}(Y \geq u)}{\mathbb{P}(Y \geq Z)}\, dP_Z(u) \\ &= \frac{\int_{\mathbb{R}} \mathbb{P}(X \geq u, Y \geq u)\, dP_Z(u)}{\mathbb{P}(Y \geq Z)} = \frac{\mathbb{P}(X \geq Z, Y \geq Z)}{\mathbb{P}(Y \geq Z)} \\ &= \mathbb{P}(X \geq Z \mid Y \geq Z),\end{aligned}$$

where we have used the assumption that (X, Y) and Z are independent in the penultimate equality. □

The proof of Lemma 2.16 below relies on the following result, due to Groeneboom (1989). It involves the Airy function Ai, defined for $x \in \mathbb{R}$ by

$$\mathrm{Ai}(x) := \frac{1}{\pi} \lim_{b\to\infty} \int_0^b \cos\Big(\frac{t^3}{3} + xt\Big)\, dt.$$

Lemma 2.15 (Corollary 3.4 of Groeneboom 1989). *Let $(W_t)_{t\in\mathbb{R}}$ be a two-sided standard Brownian motion and $Z := \mathrm{argmax}_{t\in\mathbb{R}}(W_t - t^2)$. Then Z has a density f_Z on $\mathbb{R}$ which is symmetric about zero, and which satisfies*

$$f_Z(z) = \frac{1}{2} \frac{4^{4/3}|z|}{\mathrm{Ai}'(\tilde{a}_1)} \exp\Big(-\frac{2}{3}|z|^3 + 2^{1/3}\tilde{a}_1|z|\Big)\{1 + o(1)\}$$

as $z \to \infty$, where $\tilde{a}_1 \approx -2.3381$ is the largest zero of the Airy function Ai *and where* $\mathrm{Ai}'(\tilde{a}_1) \approx 0.7022$.

In particular, there exists a universal constant $K \geq 1$ such that $f_Z(z) \geq ze^{-z^3}$ for $z \geq K^{1/3}$.

We collect in the following lemma some useful bounds on both the maximum and the argmax of a Brownian motion and a Gaussian random walk with a negative drift.

Lemma 2.16. *Fix $b > 0$, and let $(Z_t)_{t\geq 0}$ be given by $Z_t = W_t - bt$ for $t \geq 0$, where $(W_t)_{t\geq 0}$ is a standard Brownian motion. Define $\hat{M} := \sup_{t\geq 0} Z_t$ and $M := \sup_{r\in\mathbb{N}_0} Z_r$.*

(a) For any $a \geq 0$, we have

$$\frac{2\sqrt{ab}}{\sqrt{2\pi}(4ab + 1)} e^{-2ab} \leq \mathbb{P}(\hat{M} \geq a) \leq e^{-2ab}$$

and

$$\frac{3\sqrt{ab/2}}{\sqrt{2\pi}(9ab/2+1)}e^{-9ab/4}\mathbb{1}_{\{a\geq b\}}+\frac{2b}{\sqrt{2\pi}(4b^2+1)}e^{-2b^2}\mathbb{1}_{\{a<b\}}\leq\mathbb{P}(M\geq a)\leq e^{-2ab}.$$

(b) If $c \geq 0$ satisfies $bc \geq a \geq 0$, then

$$\mathbb{P}\Bigl(\sup_{r\in\mathbb{N}:r\geq c} Z_r \geq a\Bigr) \leq \mathbb{P}\Bigl(\sup_{t\geq c} Z_t \geq a\Bigr) \leq \exp\biggl\{\frac{-(bc+a)^2}{2c}\biggr\}.$$

Now let $\hat{\xi} := \mathrm{argmax}_{t\geq 0} Z_t$ and $\xi := \mathrm{argmax}_{r\in\mathbb{N}_0} Z_r$. Then $\hat{\xi}$ and ξ are both almost surely unique. Moreover, letting $\xi^1,\ldots,\xi^s$ denote independent copies of ξ, we have the following results:

(c) If $b \leq 1/2$, then

$$\mathbb{E}\Bigl(\max_{j\in[s]}\xi^j\Bigr) \leq \frac{8\log(s/b)}{b^2}.$$

(d) Taking $K \geq 1$ from Lemma 2.15, for all $k \geq K$ we have

$$e^{-2k} \leq \mathbb{P}(\hat{\xi} \geq kb^{-2}) \leq e^{-k/2}.$$

Moreover, for each $k \geq K$, there exists $b_0 > 0$, depending only on k, such that for all $b \leq b_0$ we have

$$\frac{1}{2}e^{-2k} \leq \mathbb{P}(\xi \geq kb^{-2}) \leq 2e^{-k/2} \tag{2.39}$$

and

$$\mathbb{E}\Bigl(\max_{j\in[s]}\xi^j \Bigm| \min_{j\in[s]}\xi^j \geq kb^{-2}\Bigr) \leq \frac{60}{b^2}\bigl\{k+\log(1/b)\bigr\}+sb^5.$$

Proof. (a) Since $M \leq \hat{M}$, we have

$$\mathbb{P}(M\geq a)\leq\mathbb{P}(\hat{M}\geq a)=\mathbb{P}\Bigl(\sup_{t\geq 0}(W_t-bt)\geq a\Bigr)=e^{-2ab},$$

where the calculation for the final equality can be found in, e.g. Siegmund (1986, Proposition 2.4 and Equation (2.5)). For the lower bounds, we note that

$$\mathbb{P}(\hat{M}\geq a)\geq\sup_{t\geq 0}\mathbb{P}(Z_t\geq a)=\sup_{t\geq 0}\Phi\biggl(-\frac{a+bt}{\sqrt{t}}\biggr)=\Phi\bigl(-2\sqrt{ab}\bigr).$$

Similarly, assuming without loss of generality that $a > 0$ (since otherwise the result is clear),

$$\mathbb{P}(M\geq a)\geq\sup_{r\in\mathbb{N}}\mathbb{P}(Z_r\geq a)=\sup_{r\in\mathbb{N}}\Phi\biggl(-\frac{a+br}{\sqrt{r}}\biggr)\geq\Phi\biggl(-\frac{a+br_0}{\sqrt{r_0}}\biggr),$$

where $r_0 = \lceil a/b \rceil \vee 1$. If $a \geq b$, then using the fact that the function $x \mapsto (a+bx)/\sqrt{x}$ is increasing on $[\sqrt{a/b}, \infty)$, we have

$$\frac{a+br_0}{\sqrt{r_0}} \leq \frac{a+b(a/b+1)}{\sqrt{a/b+1}} = 2\sqrt{b} \cdot \frac{a+b/2}{\sqrt{a+b}} \leq 2\sqrt{b}\biggl(a + \frac{b^2/4}{a+b}\biggr)^{1/2} \leq 3\sqrt{ab/2}.$$

On the other hand, if $a < b$, then

$$\frac{a+br_0}{\sqrt{r_0}} = a+b < 2b.$$

The desired results follow from the bound $\Phi(-x) \geq \frac{x}{\sqrt{2\pi}(x^2+1)} e^{-x^2/2}$ for all $x > 0$.

(b) By part (a), we have

$$\begin{aligned}
\mathbb{P}\Bigl(\sup_{r \in \mathbb{N}: r \geq c} Z_r \geq a\Bigr) &\leq \mathbb{P}\Bigl(\sup_{t \geq c} Z_t \geq a\Bigr) = \int_{-\infty}^{\infty} \mathbb{P}\Bigl(\sup_{t \geq c} Z_t \geq a \mid Z_c = x\Bigr) \frac{1}{\sqrt{2\pi c}} e^{-(x+bc)^2/(2c)}\,dx \\
&\leq \int_{-\infty}^{a} e^{-2(a-x)b} \frac{1}{\sqrt{2\pi c}} e^{-(x+bc)^2/(2c)}\,dx + \int_a^{\infty} \frac{1}{\sqrt{2\pi c}} e^{-(x+bc)^2/(2c)}\,dx \\
&= e^{-2ab}\Phi\Bigl(-\frac{bc-a}{\sqrt{c}}\Bigr) + \Phi\Bigl(-\frac{bc+a}{\sqrt{c}}\Bigr) \leq \exp\biggl\{\frac{-(bc+a)^2}{2c}\biggr\},
\end{aligned}$$

where in the final step we have used the fact that $bc \geq a$ and $\Phi(-x) \leq e^{-x^2/2}/2$ for $x \geq 0$.

To prove that ξ is almost surely unique, it suffices to note that

$$\mathbb{P}(\xi \text{ not unique}) \leq \mathbb{P}\biggl(\bigcup_{r_1, r_2 \in \mathbb{N}_0 : r_1 < r_2} \{Z_{r_1} = Z_{r_2}\}\biggr) \leq \sum_{r_1, r_2 \in \mathbb{N}_0 : r_1 < r_2} \mathbb{P}\Bigl(Z_{r_2} - Z_{r_1} = 0\Bigr) = 0,$$

since $Z_{r_2} - Z_{r_1} \sim \mathcal{N}\bigl(-b(r_2 - r_1), r_2 - r_1\bigr)$. To prove that $\hat{\xi}$ is almost surely unique, note that

$$\begin{aligned}
&\mathbb{P}(\hat{\xi} \text{ not unique}) \leq \mathbb{P}\biggl(\bigcup_{q_1, q_2 \in \mathbb{Q}: 0 < q_1 < q_2} \Bigl\{\max_{t \in [0, q_1]} Z_t = \max_{t \in [q_2, \infty)} Z_t\Bigr\}\biggr) \\
&\leq \sum_{q_1, q_2 \in \mathbb{Q}: 0 < q_1 < q_2} \mathbb{P}\Bigl(\max_{t \in [0, q_1]} Z_t = \max_{t \in [q_2, \infty)} Z_t\Bigr) \\
&= \sum_{q_1, q_2 \in \mathbb{Q}: 0 < q_1 < q_2} \mathbb{P}\biggl(\Bigl(\max_{t \in [q_2, \infty]} Z_t - Z_{q_2}\Bigr) = (Z_{q_2} - Z_{q_1}) - \Bigl(\max_{t \in [0, q_1]} Z_t - Z_{q_1}\Bigr)\biggr) = 0,
\end{aligned}$$

where we have used the Markov property of $(Z_t)_{t \geq 0}$ for the final equality.

(c) For any $x \in \mathbb{N}$, we have by two union bounds that

$$\mathbb{P}\Bigl(\max_{j \in [s]} \xi^j \geq x\Bigr) \leq s \sum_{r=x}^{\infty} \mathbb{P}(\xi = r) \leq s \sum_{r=x}^{\infty} \mathbb{P}(S_r \geq 0)$$

$$= s\sum_{r=x}^{\infty}\Phi(-b\sqrt{r}) \leq \frac{s}{2}\sum_{r=x}^{\infty} e^{-rb^2/2} = \frac{se^{-xb^2/2}}{2(1-e^{-b^2/2})}.$$

Now define $x_0 := \lceil 4b^{-2}\log(s/b)\rceil$. Then for $b \in (0, 1/2]$,

$$\begin{aligned}\mathbb{E}\Bigl(\max_{j\in[s]}\xi^j\Bigr) &= \sum_{x=1}^{\infty}\mathbb{P}\Bigl(\max_{j\in[s]}\xi^j \geq x\Bigr) \leq x_0 - 1 + \sum_{x=x_0}^{\infty}\frac{se^{-xb^2/2}}{2(1-e^{-b^2/2})} \\ &\leq \frac{4\log(s/b)}{b^2} + \frac{se^{-x_0b^2/2}}{2(1-e^{-b^2/2})^2} \leq \frac{4\log(s/b)}{b^2} + \frac{2}{b^2(1-1/16)^2} \\ &\leq \frac{8\log(s/b)}{b^2},\end{aligned}$$

where we have used the fact that $1 - e^{-x} \geq 15x/16$ for $x \in [0, 1/8]$.

(d) First note that $W_t - b^3t^2/k \leq W_t - bt$ for $t \geq kb^{-2}$ and $W_t - b^3t^2/k > W_t - bt$ for $t < kb^{-2}$. Thus, using the fact that $(W_t)_{t\geq 0} \overset{\mathrm{d}}{=} \bigl(a^{-1}W_{a^2t}\bigr)_{t\geq 0}$ for every $a > 0$, and taking $K \geq 1$ from Lemma 2.15, we have for $k \geq K$ that

$$\begin{aligned}\mathbb{P}(\hat{\xi} \geq kb^{-2}) &= \mathbb{P}\biggl(\operatorname*{argmax}_{t\geq 0}(W_t - bt) \geq \frac{k}{b^2}\biggr) \geq \mathbb{P}\biggl(\operatorname*{argmax}_{t\geq 0}\biggl(W_t - \frac{b^3t^2}{k}\biggr) \geq \frac{k}{b^2}\biggr) \\ &= \mathbb{P}\biggl(\operatorname*{argmax}_{t\geq 0}\biggl(\frac{W_{b^2k^{-2/3}t}}{bk^{-1/3}} - \frac{b^3t^2}{k}\biggr) \geq \frac{k}{b^2}\biggr) = \mathbb{P}\Bigl(\operatorname*{argmax}_{t\geq 0}(W_t - t^2) \geq k^{1/3}\Bigr) \\ &\geq \int_{k^{1/3}}^{\infty} 2z\exp(-z^3)\,dz \geq \int_{k^{1/3}}^{\infty}\biggl(\frac{3z}{2} + \frac{1}{2z^2}\biggr)\exp(-z^3)\,dz = \frac{e^{-k}}{2k^{1/3}} \geq e^{-2k},\end{aligned} \tag{2.40}$$

where the second inequality follows from Lemma 2.15. We also have, by part (b), that

$$\mathbb{P}(\hat{\xi} \geq kb^{-2}) = \mathbb{P}\Bigl(\operatorname*{argmax}_{t\geq 0}(W_t - bt) \geq kb^{-2}\Bigr) \leq \mathbb{P}\Bigl(\sup_{t\geq kb^{-2}}(W_t - bt) \geq 0\Bigr) \leq e^{-k/2}. \tag{2.41}$$

We now compute upper and lower bounds on the tail probabilities for ξ. By Donsker's invariance principle (Mörters and Peres, 2010, Theorem 5.22) and the continuity of the argmax map (e.g. van der Vaart and Wellner, 1996, Theorem 3.2.2), we have, as $b \to 0$, that

$$b^2\xi \overset{\mathrm{d}}{=} b^2\operatorname*{argmax}_{r\in\mathbb{N}_0}\biggl(\frac{W_{rb^2} - rb^2}{b}\biggr) \overset{\mathrm{d}}{=} \operatorname*{argmax}_{r\in b^2\mathbb{N}_0}(W_r - r) \xrightarrow{\mathrm{d}} \operatorname*{argmax}_{t\geq 0}(W_t - t).$$

Thus there exists $b_0 > 0$, depending only on k, such that for $b < b_0$, we have by (2.40) and (2.41) that

$$\mathbb{P}(\xi \geq kb^{-2}) \geq \frac{1}{2}\mathbb{P}\Bigl(\operatorname*{argmax}_{t\geq 0}(W_t - t) \geq k\Bigr) \geq \frac{1}{2}e^{-2k},$$

and that

$$\mathbb{P}(\xi \geq kb^{-2}) \leq 2\mathbb{P}\Bigl(\operatorname*{argmax}_{t\geq 0}(W_t - t) \geq k\Bigr) \leq 2e^{-k/2}.$$

We now move on to the final claim of Lemma 2.16(d). For $r \in \mathbb{N}_0$, we define $M_r := \max_{r'\in\{0,1,\ldots,r\}} Z_{r'}$ and let P_r denote the conditional distribution of $Z_r - M_r$ given that $\xi \geq r$. Note that $\{\xi \geq r\} = \{\max_{r'\in\mathbb{N}_0:r'\geq r} Z_{r'} \geq M_r\}$ up to a null set. Denote $x_0 := \lfloor 60\{k + \log(1/b)\}/b^2\rfloor$ and $c := \lceil kb^{-2}\rceil$. Without loss of generality, we may assume that $b_0 < 1/2$. Then for $b \leq b_0$, we have $c < x_0$, so

$$\begin{aligned}\mathbb{E}\Bigl(\max_{j\in[s]}\xi^j \Bigm| \min_{j\in[s]}\xi^j \geq c\Bigr) - x_0 &\leq \sum_{x=x_0}^{\infty}\mathbb{P}\Bigl(\max_{j\in[s]}\xi^j \geq x \Bigm| \min_{j\in[s]}\xi^j \geq c\Bigr)\\ &\leq s\sum_{x=x_0}^{\infty}\mathbb{P}\Bigl(\xi^1 \geq x \Bigm| \min_{j\in[s]}\xi^j \geq c\Bigr) = s\sum_{x=x_0}^{\infty}\frac{\mathbb{P}(\xi\geq x)\mathbb{P}(\xi\geq c)^{s-1}}{\mathbb{P}(\xi\geq c)^s}\\ &= s\sum_{x=x_0}^{\infty}\mathbb{P}(\xi \geq x \mid \xi \geq c).\end{aligned}$$

But, for every $x \in \mathbb{N}$ with $x \geq x_0$,

$$\begin{aligned}\mathbb{P}(\xi \geq x \mid \xi \geq c) &\leq \mathbb{P}\Bigl(\sup_{r\in\mathbb{N}:r\geq x} Z_r \geq M_c \Bigm| \sup_{r\in\mathbb{N}:r\geq c} Z_r \geq M_c\Bigr)\\ &= \mathbb{P}\Bigl(\sup_{r\in\mathbb{N}:r\geq x}(Z_r - Z_c) \geq M_c - Z_c \Bigm| \sup_{r\in\mathbb{N}:r\geq c}(Z_r - Z_c) \geq M_c - Z_c\Bigr)\\ &= \int_0^{\infty}\mathbb{P}\Bigl(\sup_{r\in\mathbb{N}:r\geq x}(Z_r - Z_c) \geq u \Bigm| \sup_{r\in\mathbb{N}:r\geq c}(Z_r - Z_c) \geq u\Bigr)\,dP_c(u)\\ &= \int_0^{\infty}\mathbb{P}\Bigl(\sup_{r\in\mathbb{N}:r\geq x-c} Z_r \geq u \Bigm| M \geq u\Bigr)\,dP_c(u), \qquad (2.42)\end{aligned}$$

where the second equality follows from Lemma 2.14 and the fact that $M_c - Z_c$ is independent of the sequence $(Z_r - Z_c)_{r\in\mathbb{N}:r\geq c}$. If $b(x-c)/4 \geq u \geq b$, then by Lemma 2.16(a) and (b) we have

$$\begin{aligned}\mathbb{P}\Bigl(\sup_{r\in\mathbb{N}:r\geq x-c} Z_r \geq u \Bigm| M \geq u\Bigr) &\leq \exp\biggl\{-\frac{(b(x-c)+u)^2}{2(x-c)}\biggr\}\cdot\frac{9ub/2+1}{3\sqrt{ub/(4\pi)}}e^{9ub/4}\\ &\leq e^{-b^2(x-c)/2+5bu/4}\biggl(3\sqrt{\pi ub} + \frac{2\sqrt{\pi}/3}{\sqrt{ub}}\biggr)\\ &\leq e^{-3b^2(x-c)/16}\biggl(\frac{3\sqrt{\pi}}{2}\sqrt{(x-c)b^2} + \frac{2\sqrt{\pi}}{3b}\biggr).\end{aligned}$$

Since the function $h \mapsto he^{-h^2/2}$ is decreasing for $h \geq 1$, we have that $3\sqrt{\pi(x-c)b^2/4} + 2\sqrt{\pi}/(3b) \leq 3e^{b^2(x-c)/16}/2$ for $x - c \geq 60b^{-2}\log(1/b)$, when $b \leq 1/2$. Thus,

$$\mathbb{P}\Big(\sup_{r \in \mathbb{N}: r \geq x-c} Z_r \geq u \,\Big|\, M \geq u\Big) \leq \frac{3}{2}e^{-b^2(x-c)/8}. \tag{2.43}$$

On the other hand, if $b > u$ (note that this implies $b(x-c) \geq u$), then by Lemma 2.16(a) and (b) we have that

$$\begin{aligned}\mathbb{P}\Big(\sup_{r \in \mathbb{N}: r \geq x-c} Z_r \geq u \,\Big|\, M \geq u\Big) &\leq \exp\Big\{-\frac{(b(x-c)+u)^2}{2(x-c)}\Big\} \cdot \frac{\sqrt{2\pi}(1+4b^2)}{2b}e^{2b^2} \\ &\leq e^{-b^2(x-c)/2+2b^2}\left(\frac{\sqrt{2\pi}}{2b} + 2\sqrt{2\pi}b\right) \\ &\leq \frac{\sqrt{2\pi}}{b}e^{-b^2(x-c)/4} \leq \frac{\sqrt{2\pi}}{2^{13/2}}e^{-b^2(x-c)/8},\end{aligned}$$

where we have used the fact that $x - c \geq 60b^{-2}\log(1/b) \geq 8$ in the final two bounds. Combining the above display with (2.43), we see that for $b(x-c)/4 \geq u$, we have

$$\mathbb{P}\Big(\max_{r \in \mathbb{N}: r \geq x-c} Z_r \geq u \,\Big|\, M \geq u\Big) \leq \frac{3}{2}e^{-b^2(x-c)/8}. \tag{2.44}$$

Thus, by reducing $b_0 > 0$ (still depending only on k) if necessary, we have for $b \leq b_0$ that

$$\begin{aligned}&\int_0^\infty \mathbb{P}\Big(\sup_{r \in \mathbb{N}: r \geq x-c} Z_r \geq u \,\Big|\, M \geq u\Big)\, dP_c(u) \\ &\quad\leq \int_0^{b(x-c)/4} \mathbb{P}\Big(\sup_{r \in \mathbb{N}: r \geq x-c} Z_r \geq u \,\Big|\, M \geq u\Big)\, dP_c(u) \\ &\qquad\qquad + \mathbb{P}\Big(M_c \geq \frac{b(x-c)}{8} \,\Big|\, \xi \geq c\Big) + \mathbb{P}\Big(Z_c \leq -\frac{b(x-c)}{8} \,\Big|\, \xi \geq c\Big) \\ &\quad\leq \frac{3}{2}e^{-b^2(x-c)/8} + 2e^{-b^2(x-c)/4}e^{2k} + 2\Phi\Big(-\frac{b(x-9c)}{8\sqrt{c}}\Big)e^{2k} \leq 5e^{-(b^2x-k)/8},\end{aligned}$$

where we have used (2.44), Lemma 2.16(a) and (2.39) in the penultimate inequality, and, in the final step, we have used the fact that $x \geq 60c$, the Gaussian tail bound $\Phi(-x) \leq \frac{1}{2}e^{-x^2/2}$ for $x \geq 0$ and the fact that

$$\frac{(x-9c)^2}{c} \geq \frac{(51/59)^2(x-c)^2}{c} \geq 59(51/59)^2(x-c) \geq 32(x-c).$$

Combining with (2.42), we conclude that

$$\mathbb{E}\Big(\max_{j \in [s]} \xi^j \,\Big|\, \min_{j \in [s]} \xi^j \geq kb^{-2}\Big) - x_0 \leq 5s\sum_{x=x_0}^{\infty} e^{-(b^2x-k)/8} \leq \frac{5se^{-15\log(1/b)/2-7}}{1-e^{-b^2/8}} \leq sb^5,$$

as desired, where we have used again the fact that $b \leq 1/2$ in the final inequality. □

Lemma 2.17. *(a) For any $n \in \mathbb{N}$, $0 < p \leq q < 1$ and $x \in \{0, 1, \dots, n\}$, we have*

$$\frac{\mathbb{P}\bigl(\mathrm{Bin}(n,p) \geq x\bigr)}{\mathbb{P}\bigl(\mathrm{Bin}(n,p/q) \geq x\bigr)} \leq \mathbb{P}\bigl(\mathrm{Bin}(n,q) \geq x\bigr). \tag{2.45}$$

(b) Let $W_1, \dots, W_n$ be independent and identically distributed, real-valued random variables, with corresponding order statistics $W_{(1)} \leq \ldots \leq W_{(n)}$. Then for every $s \geq t$ and every $m \in [n]$, we have that

$$\mathbb{P}(W_{(m)} \geq s | W_{(m)} \geq t) \leq \mathbb{P}(W_{(m)} \geq s | W_{(1)} \geq t).$$

In particular, $\mathbb{E}(W_{(m)} | W_{(m)} \geq t) \leq \mathbb{E}(W_{(m)} | W_{(1)} \geq t)$.

Proof. (a) Let $g(p)$ denote the left-hand side of (2.45). It suffices to prove that g is an increasing function on $(0, q]$. We may also assume that $x \geq 1$, because otherwise the result is clear. Now, let

$$h(p) := \mathbb{P}\bigl(\mathrm{Bin}(n,p) \geq x\bigr) = \sum_{r=x}^{n} \binom{n}{r} p^r (1-p)^{n-r}.$$

Then

$$\begin{aligned} h'(p) &= \sum_{r=x}^{n} \binom{n}{r} r p^{r-1} (1-p)^{n-r} - \sum_{r=x}^{n-1} \binom{n}{r} (n-r) p^r (1-p)^{n-r-1} \\ &= \sum_{r=x-1}^{n-1} \frac{n!}{r!(n-r-1)!} p^r (1-p)^{n-r-1} - \sum_{r=x}^{n-1} \frac{n!}{r!(n-r-1)!} p^r (1-p)^{n-r-1} \\ &= \frac{n!}{(x-1)!(n-x)!} p^{x-1} (1-p)^{n-x}. \end{aligned}$$

We can therefore compute

$$g'(p) = \frac{h(p/q) h'(p) - h(p) h'(p/q)/q}{h(p/q)^2},$$

and we note that

$$\begin{aligned} &h(p/q)h'(p) - h(p)h'(p/q)/q \\ &\quad = \frac{n!}{(x-1)!(n-x)!} p^{x-1} (1-p)^{n-x} \sum_{r=x}^{n} \binom{n}{r} \Bigl(\frac{p}{q}\Bigr)^r \Bigl(1 - \frac{p}{q}\Bigr)^{n-r} \\ &\qquad\qquad - \frac{n!}{(x-1)!(n-x)!} \frac{1}{q} \Bigl(\frac{p}{q}\Bigr)^{x-1} \Bigl(1 - \frac{p}{q}\Bigr)^{n-x} \sum_{r=x}^{n} \binom{n}{r} p^r (1-p)^{n-r} \\ &\quad = \frac{n! p^{x-1} (1-p)^{n-x} (1-p/q)^{n-x}}{q^x (x-1)!(n-x)!} \sum_{r=x}^{n} \binom{n}{r} p^r \biggl\{ \frac{1}{(q-p)^{r-x}} - \frac{1}{(1-p)^{r-x}} \biggr\} \geq 0, \end{aligned}$$

as required.

(b) Write F for the distribution function of W_1, and let $\bar{F} := 1 - F$. We also write $\bar{F}(x-) := \lim_{y \nearrow x} \bar{F}(x)$. For a Borel measurable set $A \subseteq \mathbb{R}$, let $N(A) := \sum_{i=1}^n \mathbb{1}_{\{W_i \in A\}}$. Then, for $s \geq t$,

$$\begin{aligned}\mathbb{P}(W_{(m)} \geq s | W_{(m)} \geq t) = \frac{\mathbb{P}(W_{(m)} \geq s)}{\mathbb{P}(W_{(m)} \geq t)} &= \frac{\mathbb{P}\bigl\{N\bigl([s,\infty)\bigr) \geq n-m+1\bigr\}}{\mathbb{P}\bigl\{N\bigl([t,\infty)\bigr) \geq n-m+1\bigr\}} \\ &= \frac{\mathbb{P}\bigl\{\mathrm{Bin}\bigl(n, \bar{F}(s-)\bigr) \geq n-m+1\bigr\}}{\mathbb{P}\bigl\{\mathrm{Bin}\bigl(n, \bar{F}(t-)\bigr) \geq n-m+1\bigr\}}.\end{aligned}$$

On the other hand,

$$\begin{aligned}\mathbb{P}(W_{(m)} \geq s | W_{(1)} \geq t) &= \frac{\mathbb{P}\bigl(W_{(m)} \geq s, W_{(1)} \geq t\bigr)}{\mathbb{P}(W_{(1)} \geq t)} \\ &= \frac{\mathbb{P}\bigl\{N\bigl((-\infty,t)\bigr) = 0, N\bigl([s,\infty)\bigr) \geq n-m+1\bigr\}}{\mathbb{P}\bigl\{N\bigl((-\infty,t)\bigr) = 0\bigr\}} \\ &= \frac{\sum_{r=n-m+1}^n \binom{n}{r} \bar{F}(s-)^r \bigl\{\bar{F}(t-) - \bar{F}(s-)\bigr\}^{n-r}}{\bar{F}(t-)^n} \\ &= \mathbb{P}\bigl\{\mathrm{Bin}\bigl(n, \bar{F}(s-)/\bar{F}(t-)\bigr) \geq n-m+1\bigr\}.\end{aligned}$$

The first conclusion therefore follows immediately from (a), and the second conclusion is an immediate consequence of the first. □

Lemma 2.18. *Let $v = (v_1, \ldots, v_p)^\top \in \mathbb{R}^p$ be a unit vector. There exists $\ell \in \{0, \ldots, \lfloor \log_2 p \rfloor\}$ such that*

$$\left| \left\{ j \in [p] : v_j^2 \geq \frac{1}{2^\ell \log_2(2p)} \right\} \right| \geq 2^\ell.$$

Proof. The case $p = 1$ is trivially true, so we may assume without loss of generality that $p \geq 2$. Let $L := \lfloor \log_2 p \rfloor$, $b_\ell := 2^{-\ell} \log_2^{-1}(2p)$ and $n_\ell := \bigl|\{j : v_j^2 \geq b_\ell\}\bigr|$ for $\ell \in \{0, \ldots, L\}$. Assume for a contradiction that $n_\ell < 2^\ell$ for all ℓ. Then by Fubini's theorem we have

$$\begin{aligned}\|v\|_2^2 &= \sum_{j=1}^p \int_{t=0}^1 \mathbb{1}_{\{v_j^2 \geq t\}} \, dt \leq n_0(1 - b_0) + \sum_{\ell=1}^L n_\ell(b_{\ell-1} - b_\ell) + pb_L \\ &\leq \sum_{\ell=1}^L (2^\ell - 1)(b_{\ell-1} - b_\ell) + pb_L = \sum_{\ell=0}^{L-1} 2^\ell b_\ell + (p - 2^L + 1)b_L \leq \frac{L+1}{\log_2(2p)} \leq 1.\end{aligned}$$

Note that the penultimate inequality is strict if $p + 1$ is not an integer power of 2 and the final inequality is strict if p is not an integer power of 2. Since $p \geq 2$, it cannot be the case that we have equality in both equalities, so $\|v\|_2^2 < 1$, which contradicts the fact that v is a unit vector. □

Lemma 2.19. *Define sequences $(a_n)_{n\in\mathbb{N}_0}$ and $(b_n)_{n\in\mathbb{N}_0}$ as follows: $a_0 := b_0 := 0$, $b_n := (b_{n-1}+1)\mathbb{1}_{\{n\notin\{2^\xi:\xi\in\mathbb{N}_0\}\}}$ and $a_n := (a_{n-1}+1)\mathbb{1}_{\{n\notin\{2^\xi:\xi\in\mathbb{N}_0\}\}} + (b_{n-1}+1)\mathbb{1}_{\{n\in\{2^\xi:\xi\in\mathbb{N}_0\}\}}$ for $n\in\mathbb{N}$. Then, we have*

$$n/2 \le a_n < 3n/4,$$

for all $n \ge 2$.

Proof. The two sequences $(a_n)_{n\in\mathbb{N}_0}$ and $(b_n)_{n\in\mathbb{N}_0}$ are tabulated below.

n	0	1	2	3	4	5	6	7	8	$\dots$	2^ξ	$2^\xi+1$	$\dots$	$2^{\xi+1}-1$	$\dots$
a_n	0	1	1	2	2	3	4	5	4	$\dots$	$2^{\xi-1}$	$2^{\xi-1}+1$	$\dots$	$3\cdot 2^{\xi-1}-1$	$\dots$
b_n	0	0	0	1	0	1	2	3	0	$\dots$	0	1	$\dots$	$2^\xi-1$	$\dots$

It is clear from the definition of $(b_n)_n$ that $b_{2^\xi+i} = i$ for $\xi\in\mathbb{N}_0$ and $0\le i\le 2^\xi-1$. Consequently, we have $a_{2^\xi} = b_{2^\xi-1}+1 = 2^{\xi-1}$ and $a_{2^\xi+i} = 2^{\xi-1}+i$ for $\xi\in\mathbb{N}$ and $1\le i\le 2^\xi-1$. Hence, we have

$$\frac{1}{2} = \frac{2^{\xi-1}}{2^\xi} \le \frac{a_{2^\xi+i}}{2^\xi+i} = \frac{2^{\xi-1}+i}{2^\xi+i} \le \frac{2^{\xi-1}+2^\xi-1}{2^\xi+2^\xi-1} < \frac{3}{4},$$

for all $\xi\in\mathbb{N}$ and $0\le i\le 2^\xi-1$ and the desired result follows. □

Lemma 2.20. *Let $Z_1,\ldots,Z_p \overset{\mathrm{iid}}{\sim} \mathcal{N}(0,1)$. Then for any $a>0$ and $x>0$, we have*

$$\mathbb{P}\biggl(\sum_{j=1}^p Z_j^2\mathbb{1}_{\{|Z_j|\ge a\}} \ge 6pe^{-a^2/8}+4x\biggr) \le e^{-x}.$$

Proof. This proof has some similarities with that of Lemma 17 of Liu, Gao and Samworth (2021). By a Chernoff bound, we have for any $u,\lambda>0$ that,

$$\mathbb{P}\biggl(\sum_{j=1}^p Z_j^2\mathbb{1}_{\{|Z_j|\ge a\}} \ge u\biggr) \le e^{-\lambda u}\bigl\{\mathbb{E}e^{\lambda Z_1^2\mathbb{1}_{\{|Z_j|\ge a\}}}\bigr\}^p. \tag{2.46}$$

We write $p(x) := (2\pi)^{-1/2}x^{-1/2}e^{-x/2}$ for the density of a χ_1^2 distribution. For $\lambda\in(0,1/4]$, we bound the moment generating function above as follows:

$$\begin{aligned}
\mathbb{E}e^{\lambda Z_1^2\mathbb{1}_{\{|Z_j|\ge a\}}} &= \int_{a^2}^\infty e^{\lambda x}p(x)\,dx \le 1+\int_{a^2}^\infty (e^{\lambda x}-1)p(x)\,dx = 1+\int_{a^2}^\infty \sum_{k=1}^\infty \frac{\lambda^k x^k}{k!}p(x)\,dx \\
&\le 1+\int_{a^2}^\infty \lambda x e^{\lambda x}p(x)\,dx \le 1+\frac{\lambda}{\sqrt{2\pi}}\int_{a^2}^\infty x^{1/2}e^{-x/4}\,dx \\
&= 1+\frac{4\lambda}{\sqrt{\pi}}\int_{a/\sqrt{2}}^\infty t^2e^{-t^2/2}\,dt = 1+\sqrt{\frac{8}{\pi}}\lambda a e^{-a^2/4}+4\sqrt{2}\lambda\Bigl\{1-\Phi\Bigl(\frac{a}{\sqrt{2}}\Bigr)\Bigr\} \\
&\le 1+\sqrt{\frac{8}{\pi}}\lambda a e^{-a^2/4}+2\sqrt{2}\lambda e^{-a^2/4} \le 1+\biggl(2\sqrt{\frac{8}{\pi}}e^{-1/2}+2\sqrt{2}\biggr)\lambda e^{-a^2/8} \\
&\le 1+5\lambda e^{-a^2/8},
\end{aligned}$$

where we use the fact that $xe^{-x^2/4} \leq 2e^{-1/2}e^{-x^2/8}$ for $x \in \mathbb{R}$ in the penultimate inequality. Hence, by substituting this bound into (2.46), we have for every $u > 0$, that

$$\mathbb{P}\biggl(\sum_{j=1}^{p} Z_j^2 \mathbb{1}_{\{|Z_j| \geq a\}} \geq u\biggr) \leq \exp\bigl\{-\lambda u + p\log(1 + 5\lambda e^{-a^2/8})\bigr\} \leq \exp\bigl(-\lambda u + 5p\lambda e^{-a^2/8}\bigr).$$

We set $u = 6pe^{-a^2/8} + 4x$. If $x \leq pe^{-a^2/8}/4$, choose $\lambda = p^{-1}xe^{a^2/8} \leq 1/4$; if $x > pe^{-a^2/8}/4$, choose $\lambda = 1/4$. In both cases, we have

$$\mathbb{P}\biggl(\sum_{j=1}^{p} Z_j^2 \mathbb{1}_{\{|Z_j| \geq a\}} \geq u\biggr) \leq e^{-x},$$

as required. □

Chapter 3

Inference in high-dimensional online changepoint detection

3.1 Introduction

In the field of high-dimensional statistical inference, uncertainty quantification has become a major theme over the last decade, originating with influential work on the debiased Lasso in (generalised) linear models (Javanmard and Montanari, 2014; van de Geer et al., 2014; Zhang and Zhang, 2014), and subsequently developed in other settings (e.g. Janková and van de Geer, 2015; Yu, Bradic and Samworth, 2021). Inference problems associated with multiple (offline) changepoints are also being studied in recent years. A number of approaches have become popular including simultaneous multiscale changepoint estimation (Frick, Munk and Sieling, 2014), false discovery rate control (e.g. Li, Munk and Sieling, 2016; Cheng, He and Schwartzman, 2020), post-selection inference (e.g. Hyun, G'Sell and Tibshirani, 2018; Jewell, Fearnhead and Witten, 2022) and narrowest significance pursuit (Fryzlewicz, 2021a).

The aim of this chapter is to propose methods to address two new inferential challenges associated with the high-dimensional, sequential detection of a sparse change in mean. The first is to provide a confidence interval for the location of the changepoint, while the second is to estimate the signal set of indices of coordinates that undergo the change. Despite the importance of uncertainty quantification and signal support recovery in changepoint applications, neither of these problems has previously been studied in the multivariate sequential changepoint detection literature, to the best of our knowledge. Of course, one option here would be to apply an offline confidence interval construction after a sequential procedure has declared a change. However, this would be to ignore the essential challenge of the sequential nature of the problem, whereby one wishes to avoid storing all historical data, to enable inference to be carried out in an online manner, see Chapter 1. This online requirement turns out to impose severe restrictions on the class of algorithms available to the practitioner, and lies at the heart of the difficulty of the problem.

To give a brief outline of our construction of a confidence interval with guaranteed $(1-\alpha)$-level coverage, consider for simplicity the univariate setting, where $(X_n)_{n\in\mathbb{N}}$ form a sequence of independent random variables with $X_1,\ldots,X_z \overset{\mathrm{iid}}{\sim} \mathcal{N}(0,1)$ and $X_{z+1},X_{z+2},\ldots \overset{\mathrm{iid}}{\sim} \mathcal{N}(\theta,1)$. Without loss of generality, we assume that $\theta > 0$. Suppose that θ is known to be at least $b > 0$ and, for $n \in \mathbb{N}$, let[1]

$$t_{n,b} := \operatorname*{argmax}_{0\leq h\leq n} \sum_{i=n-h+1}^{n} (X_i - b/2). \tag{3.1}$$

Since $\sum_{i=n-h+1}^{n}(X_i - b/2)$ can be viewed as the likelihood ratio statistic for testing the null of $\mathcal{N}(0,1)$ against the alternative of $\mathcal{N}(b,1)$ using $X_{n-h+1},\ldots,X_n$, the quantity $t_{n,b}$ is the tail length for which the likelihood ratio statistic is maximised. If N is the stopping time defining a good sequential changepoint detection procedure, then, intuitively, $N - t_{N,b}$ should be close to the true changepoint location z, and almost pivotal. This motivates the construction of a confidence interval of the form $\bigl[\max\bigl\{N - t_{N,b} - g(\alpha,b), 0\bigr\}, N\bigr]$, where we control the tail probability of the distribution of $N - t_{N,b}$ to choose $g(\alpha,b)$ so as to ensure the desired coverage. In the multivariate case, considerable care is required to handle the post-selection nature of the inferential problem, as well as to determine an appropriate left endpoint for the confidence interval. For this latter purpose, we only assume a lower bound on the Euclidean norm of the vector of mean change, and employ a delicate multivariate and multiscale aggregation scheme; see Section 3.2 for details.

In terms of the base sequential changepoint detection procedures, we focus on the `ocd` algorithm (short for **o**nline **c**hangepoint **d**etection) introduced in Section 2.2, as well as its variant `ocd`$'$, which provides guarantees on both the average and worst-case detection delays, subject to a guarantee on the *patience*, or average false alarm rate under the null hypothesis of no change. Crucially, these are both online algorithms. Our confidence intervals, which we correspondingly denote `ocd_CI` and `ocd_CI`$'$, inherit this same online property, thereby making them applicable even in very high-dimensional settings and where changes may be rare, so that we may need to see many new data points before declaring a change.

In Section 3.3 we study the theoretical performance of the `ocd_CI`$'$ procedure. In particular, we prove in Theorem 3.1 that, for a suitable choice of input parameters, the confidence interval has at least nominal coverage. Moreover, Theorem 3.2 ensures that, with high probability, its length is of the same order as the average detection delay for the base `ocd`$'$ procedure, up to a logarithmic factor. This is remarkable in view of the intrinsic challenge that the better such a changepoint detection procedure performs, the fewer post-change observations are available for inferential tasks.

A very useful byproduct of our `ocd_CI` methodology is that we obtain a natural estimate of the set of signal coordinates (i.e. those that undergo change). In Theorem 3.3, we prove

[1] In the case of a tie, we choose the smallest h achieving the maximum.

that, with high probability, it is able both to recover the effective support of the signal (see Section 3.3.1 for a formal definition), and avoids noise coordinates.

Section 3.4 is devoted to a study of the numerical performance of our methodological proposals. Our simulations confirm that the `ocd_CI` methodology attains the desired coverage level across a wide range of parameter settings, that the average confidence interval length is of comparable order to the average detection delay and that our support recovery guarantees are validated empirically. Moreover, in Section 3.4.4, we illustrate the practical utility of our methods by applying them to both excess death data during the COVID-19 pandemic in the US and S&P 500 data during the 2007–2008 financial crisis.

Proofs are given in Section 3.5, with auxiliary results deferred to Section 3.6. An `R` implementation of our methodology is available at github.com/yudongchen88/ocd_CI.

3.2 Confidence interval construction and support estimation methodology

In the multivariate sequential changepoint detection problem, we observe p-variate observations $X_1, X_2, \ldots$ in turn, and seek to report a stopping time N by which we believe a change has occurred. The focus here is on changes in the mean of the underlying process, and we denote the time of the changepoint by z. Moreover, since our primary interest is in high-dimensional settings, we will also seek to exploit sparsity in the vector of mean change. Given $\alpha \in (0,1)$, then, our primary goal is to construct a confidence interval $\mathcal{C} \equiv \mathcal{C}(X_1, \ldots, X_N, \alpha)$ with the property that $z \in \mathcal{C}$ with probability at least $1-\alpha$.

The algorithm takes inputs $X_1, X_2, \ldots \in \mathbb{R}^p$, observed sequentially, a known lower bound $\beta > 0$ for the ℓ_2-norm of the vector of mean change, a hard thresholding level $a \geq 0$ that can be chosen to detect dense or sparse signals, two changepoint declaration thresholds $T^{\mathrm{diag}} > 0$ and $T^{\mathrm{off}} > 0$, and two parameters d_1 and d_2 for confidence interval construction.

The first part of the algorithm is online changepoint detection, where we use the `ocd` algorithm (Algorithm 2.1). Recall that the `ocd` algorithm relies on the lower bound $\beta > 0$ and sets of signed scales $\mathcal{B}$ and $\mathcal{B}_0$ defined in terms of β (see Section 2.2).

From our perspective, the key aspects of this multiscale algorithm are that, in addition to returning a stopping time N as output, it produces a matrix of residual tail lengths $(t^j_{N,b})_{j\in[p], b\in\mathcal{B}\cup\mathcal{B}_0}$ with $t^j_{N,b} := \mathrm{sargmax}_{0\leq h\leq N} \sum_{i=N-h+1}^{N}(X^j_i - b/2)$ (similarly to (3.1)), an 'anchor' coordinate $\hat{j} \in [p]$, a signed anchor scale $\hat{b} \in \mathcal{B}$ and a tail partial sum vector $A^{\cdot,\hat{j}}_{N,\hat{b}} \in \mathbb{R}^p$ with jth component $A^{j,\hat{j}}_{N,\hat{b}} := \sum_{i=N-t^{\hat{j}}_{N,\hat{b}}+1}^{N} X^j_i$, where

$$(\hat{j}, \hat{b}) := \operatorname*{argmax}_{(j,b)\in[p]\times\mathcal{B}} Q^j_{N,b},$$

with $Q^{j}_{N,b}$ being the off-diagonal statistics defined in (2.4) at the time of changepoint declaration.

The intuition is that the anchor coordinate and signed scale are chosen so that the final $t^{\hat{j}}_{N,\hat{b}}$ observations provide the best evidence among all of the residual tail lengths against the null hypothesis of no change. Meanwhile, $A^{\cdot,\hat{j}}_{N,\hat{b}}$ aggregates the last $t^{\hat{j}}_{N,\hat{b}}$ observations in each coordinate, thereby providing a measure of the strength of this evidence against the null.

The main idea of our confidence interval construction is to seek to identify coordinates with large post-change signal. To this end, observe when $t^{\hat{j}}_{N,\hat{b}}$ is not too much larger than $N - z$, the quantity $E^{j,\hat{j}}_{N,\hat{b}} := A^{j,\hat{j}}_{N,\hat{b}}/(t^{\hat{j}}_{N,\hat{b}} \vee 1)^{1/2}$ should be centred close to $\theta^{j}(t^{\hat{j}}_{N,\hat{b}})^{1/2}$ for $j \in [p] \setminus \{\hat{j}\}$, with variance close to 1. Indeed, if $\hat{j}$, $\hat{b}$, N and $t^{\hat{j}}_{N,\hat{b}}$ were fixed, and if $0 < t^{\hat{j}}_{N,\hat{b}} \leq N - z$, then the former quantity would be normally distributed around this centering value, with unit variance. The random nature of these quantities, however, introduces a post-selection inference aspect to the problem. Nevertheless, by choosing an appropriate threshold value $d_1 > 0$, we can ensure that with high probability, when $j \neq \hat{j}$ is a noise coordinate, we have $|E^{j,\hat{j}}_{N,\hat{b}}| < d_1$, and when $j \neq \hat{j}$ is a coordinate with sufficiently large signal, there exists a signed scale $b \in (\mathcal{B} \cup \mathcal{B}_0) \cap [-|\theta^{j}|, |\theta^{j}|]$, having the same sign as θ^{j}, for which $\bigl|E^{j,\hat{j}}_{N,\hat{b}}\bigr| - |b|(t^{\hat{j}}_{N,\hat{b}})^{1/2} \geq d_1$. In fact, such a signed scale, if it exists, can always be chosen to be from $\mathcal{B}_0$. Thus we denote the set of indices j for which the latter inequality holds:

$$\hat{\mathcal{S}} := \Bigl\{ j \in [p] \setminus \{\hat{j}\} : \bigl|E^{j,\hat{j}}_{N,\hat{b}}\bigr| - b_{\min}(t^{\hat{j}}_{N,\hat{b}})^{1/2} \geq d_1 \Bigr\}.$$

Based on the discussions above, $\hat{\mathcal{S}}$, as a convenient byproduct of our algorithm, forms a natural estimate of the set of coordinates in which the mean change is large.

For each $j \in \hat{\mathcal{S}}$, there exists a largest scale $b \in (\mathcal{B} \cup \mathcal{B}_0) \cap (0, \infty)$ for which $\bigl|E^{j,\hat{j}}_{N,\hat{b}}\bigr| - b(t^{\hat{j}}_{N,\hat{b}})^{1/2} \geq d_1$. We denote the signed version of this quantity, where the sign is chosen to agree with that of $E^{j,\hat{j}}_{N,\hat{b}}$, by $\tilde{b}^{j}$:

$$\tilde{b}^{j} \leftarrow \operatorname{sgn}\bigl(E^{j,\hat{j}}_{N,\hat{b}}\bigr) \max\Bigl\{ b \in (\mathcal{B} \cup \mathcal{B}_0) \cap (0, \infty) : \bigl|E^{j,\hat{j}}_{N,\hat{b}}\bigr| - b(t^{\hat{j}}_{N,\hat{b}})^{1/2} \geq d_1 \Bigr\}.$$

This can be regarded as a shrunken estimate of θ^{j}, and therefore plays the role of the lower bound b from the univariate problem discussed in the introduction. Finally, then, our confidence interval can be constructed as the intersection over indices $j \in \hat{\mathcal{S}}$ of the confidence interval from the univariate problem in coordinate j, with signed scale $\tilde{b}^{j}$.

Pseudo-code for this `ocd_CI` confidence interval construction is given in Algorithm 3.1, where we suppress the n dependence on quantities that are updated at each time step. The computational complexity per new observation, as well as the storage requirements, of this algorithm are $O\big(p^2 \log(ep)\big)$ regardless of the observation history, so it satisfies the condition to be an online algorithm, as discussed in the introduction.

We now discuss a few technical details of the algorithm. The right endpoint of the confidence interval constructed in the `ocd_CI` algorithm is chosen to be the changepoint declaration time N. This is motivated by the fact that the probability of a false alarm (i.e. $N \leq z$) is small for a good changepoint detection procedure. However, when γ is very big, this may lead to an unbalanced confidence interval. In other words, the probability of the right endpoint of the confidence interval being smaller than z is tiny. To overcome this issue, we could try to control the conditional coverage probability $\mathbb{P}(L \leq z \mid N \geq z)$ instead, where L is the left endpoint of the confidence interval. However, we remark that $\mathbb{P}(L \leq z \mid N \geq z) \geq 1 - \alpha$ does not imply any guarantee on the unconditional coverage probability $\mathbb{P}(L \leq z \leq N)$. Another approach is to use a different right endpoint for the confidence interval, for example $N - \max_{j \in \hat{\mathcal{S}}} \Big\{ t^j_{N,\tilde{b}^j} + \frac{d_2'}{(\tilde{b}^j)^2} \Big\}$, which is somewhat similar to the current left endpoint. Extra care, though, is needed to guarantee that this is a valid and non-trivial interval, i.e. the left endpoint is smaller than the right endpoint. In this chapter, we will focus on the right endpoint being chosen to be N and not pursue the analysis of the aforementioned approaches.

The `ocd_CI` algorithm only works for $p \geq 2$, since otherwise the support estimate $\hat{\mathcal{S}}$ would be empty. The univariate problem has been discussed in Section 3.1 as a foundation for the `ocd_CI` algorithm. The case when there is only one 'strong' signal coordinate (see below in Section 3.3 for a formal definition of effective sparsity) is also interesting. The anchor coordinate will most likely be a noise coordinate, since its off-diagonal statistic needs to be the biggest, Then, as long as the corresponding tail length is not too bigger than $N - z$, that one signal coordinate will enter $\hat{\mathcal{S}}$ and provide a good confidence interval by itself.

In the formal definition of the support estimate $\hat{\mathcal{S}}$ in the algorithm, we have not included the anchor coordinate $\hat{j}$. One reason has been discussed in the previous paragraph. Another reason is that if we had included $\hat{j}$, it would introduce more complicated dependence between coordinates in the theoretical analysis. However, our support recovery result (see Theorem 3.3 below) covers both $\hat{\mathcal{S}}$ and $\hat{\mathcal{S}} \cup \{\hat{j}\}$ as support estimates.

3.2.1 A slight variant of the `ocd_CI` algorithm

While our experience is that `ocd_CI` performs very well empirically, in our theoretical analysis it turns out to be easier to study a slight variant of this algorithm, denoted `ocd_CI`$'$. There are two main differences between the algorithms. First, in `ocd_CI`, the base changepoint detection procedure is `ocd`, while in `ocd_CI`$'$, we use the `ocd`$'$ procedure (Algorithm 2.2) instead. This latter algorithm is designed to avoid difficulties caused by adversarial pre-change observations that may lead to lengthy response delays for the `ocd` procedure. In particular, for each $j \in [p]$ and $b \in \mathcal{B}$, instead of using the final $t^j_{n,b}$ observations at time n to construct test statistics

Algorithm 3.1: Pseudo-code for the confidence interval construction algorithm `ocd_CI`

```
Input: $X_1, X_2, \ldots \in \mathbb{R}^p$ observed sequentially, $\beta > 0$, $a \geq 0$, $T^{\mathrm{diag}}, T^{\mathrm{off}} > 0$,
       $d_1, d_2 > 0$
Set: $b_{\min} = \frac{\beta}{\sqrt{2^{\lfloor \log_2(2p) \rfloor} \log_2(2p)}}$, $\mathcal{B}_0 = \{\pm b_{\min}\}$,
     $\mathcal{B} = \big\{\pm 2^{\ell/2} b_{\min} : \ell = 1, \ldots, \lfloor \log_2(2p) \rfloor\big\}$, $n = 0$, $A_b = \mathbf{0} \in \mathbb{R}^{p \times p}$ and $t_b = 0 \in \mathbb{R}^p$
     for all $b \in \mathcal{B} \cup \mathcal{B}_0$
repeat
    $n \leftarrow n + 1$
    observe new data vector $X_n$
    for $(j, b) \in [p] \times (\mathcal{B} \cup \mathcal{B}_0)$ do
        $t_b^j \leftarrow t_b^j + 1$
        $A_b^{\cdot,j} \leftarrow A_b^{\cdot,j} + X_n$
        if $bA_b^{j,j} - b^2 t_b^j / 2 \leq 0$ then
            $t_b^j \leftarrow 0$ and $A_b^{\cdot,j} \leftarrow 0$
        $E_b^{\cdot,j} \leftarrow A_b^{\cdot,j} / \big(t_b^j \vee 1\big)^{1/2}$
        $Q_b^j \leftarrow \sum_{j' \in [p] \setminus \{j\}} (E_b^{j',j})^2 \mathbb{1}_{\{|E_b^{j',j}| \geq a\}}$
    $S^{\mathrm{diag}} \leftarrow \max_{(j,b) \in [p] \times (\mathcal{B} \cup \mathcal{B}_0)} \big(bA_b^{j,j} - b^2 t_b^j / 2\big)$
    $S^{\mathrm{off}} \leftarrow \max_{(j,b) \in [p] \times \mathcal{B}} Q_b^j$
until $S^{\mathrm{diag}} \geq T^{\mathrm{diag}}$ or $S^{\mathrm{off}} \geq T^{\mathrm{off}}$;
$(\hat{j}, \hat{b}) \leftarrow \mathrm{argmax}_{(j,b) \in [p] \times \mathcal{B}} Q_b^j$
$\hat{\mathcal{S}} \leftarrow \Big\{ j \in [p] \setminus \{\hat{j}\} : \big|E_{\hat{b}}^{j,\hat{j}}\big| - b_{\min} (t_{\hat{b}}^{\hat{j}})^{1/2} \geq d_1 \Big\}$
for $j \in \hat{\mathcal{S}}$ do
    $\tilde{b}^j \leftarrow \mathrm{sgn}\big(E_{\hat{b}}^{j,\hat{j}}\big) \max\Big\{ b \in (\mathcal{B} \cup \mathcal{B}_0) \cap (0, \infty) : \big|E_{\hat{b}}^{j,\hat{j}}\big| - b (t_{\hat{b}}^{\hat{j}})^{1/2} \geq d_1 \Big\}$
Output: Confidence interval $\mathcal{C} = \Big[ \max\Big\{ n - \min_{j \in \hat{\mathcal{S}}} \Big\{ t_{\tilde{b}^j}^j + \frac{d_2}{(\tilde{b}^j)^2} \Big\}, 0 \Big\}, n \Big]$
```

based on $A^{\cdot,j}_{n,b}$, the ocd′ procedure aggregates over a reduced number $\tau^j_{n,b}$ of observations to obtain test statistics based on $\Lambda^{\cdot,j}_{n,b}$, where $\tau^j_{n,b}$ is constructed in an online manner to lie in the interval $[t^j_{n,b}/2, 3t^j_{n,b}/4]$ for $t^j_{n,b} \geq 2$. Even though the reduced tail lengths may lead to a slight deterioration in empirical performance, provided no change has been declared by time z, they guarantee that from a later time of the form $z + O(b^{-2})$, the last $\tau^j_{n,b}$ observations consist entirely of post-change data.

Second, in ocd_CI′, we allow the practitioner to observe a further ℓ observations after the time of changepoint declaration, before constructing the confidence interval. The additional observations are used to determine the anchor coordinate $\hat{j}$ and scale $\hat{b}$, as well as the estimated support $\hat{\mathcal{S}}$ and the estimated scale $\tilde{b}^j$ for each $j \in \hat{\mathcal{S}}$. Thus, the extra sampling is used to guard against an unusually early changepoint declaration that leaves very few post-change observations for inference. Nevertheless, we will see in Theorem 3.1 below that the ocd_CI′ confidence interval has guaranteed nominal coverage even with $\ell = 0$, so that additional observations are only used to control the length of the interval. In fact, even for this latter aspect, the numerical evidence presented in Section 3.4 indicates that $\ell = 0$ provides confidence intervals of reasonable length in practice. Similarly, Theorem 3.3 ensures that with high probability, our support estimate $\hat{\mathcal{S}}$ contains no noise coordinates (i.e. has false positive control) even with $\ell = 0$, so that the extra sampling is only used to provide false negative control. Pseudo-code for the ocd_CI′ algorithm is given in Algorithm 3.2; its computational complexity per new observation, and storage requirements, remain $O\bigl(p^2 \log(ep)\bigr)$.

3.3 Theoretical analysis

Throughout this section, we will assume that the sequential observations $X_1, X_2, \ldots$ are independent, and that there exist $z \in \mathbb{N}_0$ and $\theta = (\theta^1, \ldots, \theta^p)^\top \neq 0$ for which $X_1, \ldots, X_z \sim \mathcal{N}_p(0, I_p)$ and $X_{z+1}, X_{z+2}, \ldots \sim \mathcal{N}_p(\theta, I_p)$. We let $\vartheta := \|\theta\|_2$, and write $\mathbb{P}_{z,\theta}$ for probabilities computed under this model, though in places we omit the subscripts for brevity. Recall from Section 2.3 that the *effective sparsity* of θ, denoted $s(\theta)$, is the smallest $s \in \bigl\{2^0, 2^1, \ldots, 2^{\lfloor \log_2(p) \rfloor}\bigr\}$ such that the corresponding *effective support* $\mathcal{S}(\theta) := \bigl\{j \in [p] : |\theta^j| \geq \|\theta\|_2/\sqrt{s \log_2(2p)}\bigr\}$ has cardinality at least $s(\theta)$. Thus, the sum of squares of coordinates in the effective support of θ has the same order of magnitude as $\|\theta\|_2^2$, up to logarithmic factors.

3.3.1 Coverage probability and length of the confidence interval

The following theorem shows that the confidence interval constructed in the ocd_CI′ algorithm has the desired coverage level.

Theorem 3.1. *Let $p \geq 2$. Fix $\alpha \in (0,1)$ and $\gamma \geq 1$ and assume that $z \leq 2\alpha\gamma$. Then there exist universal constants $C_1, C_2 > 0$, such that with inputs $(X_t)_{t\in\mathbb{N}}$, $0 < \beta \leq \vartheta$,*

Algorithm 3.2: Pseudo-code for the `ocd_CI′` algorithm, a slight variant of `ocd_CI`

```
Input: $X_1, X_2 \ldots \in \mathbb{R}^p$ observed sequentially, $\beta > 0$, $a \geq 0$, $T^{\mathrm{diag}}, T^{\mathrm{off}} > 0$, $d_1, d_2 > 0$
    and $\ell \in \mathbb{N}_0$
Set: $b_{\min} = \frac{\beta}{\sqrt{2^{\lfloor \log_2(2p) \rfloor} \log_2(2p)}}$, $\mathcal{B}_0 = \{\pm b_{\min}\}$,
    $\mathcal{B} = \bigl\{\pm 2^{\ell/2} b_{\min} : \ell = 1, \ldots, \lfloor \log_2(2p) \rfloor\bigr\}$, $n = 0$, $A_b = \Lambda_b = \tilde{\Lambda}_b = \mathbf{0} \in \mathbb{R}^{p \times p}$ and
    $t_b = \tau_b = \tilde{\tau}_b = 0 \in \mathbb{R}^p$ for all $b \in \mathcal{B} \cup \mathcal{B}_0$
repeat
    $n \leftarrow n + 1$
    observe new data vector $X_n$
    for $(j, b) \in [p] \times (\mathcal{B} \cup \mathcal{B}_0)$ do
        $t_b^j \leftarrow t_b^j + 1$  and  $A_b^{\cdot,j} \leftarrow A_b^{\cdot,j} + X_n$
        set $\delta = 0$ if $t_b^j$ is a power of 2 and $\delta = 1$ otherwise.
        $\tau_b^j \leftarrow \tau_b^j \delta + \tilde{\tau}_b^j (1 - \delta) + 1$  and  $\Lambda_b^{\cdot,j} \leftarrow \Lambda_b^{\cdot,j} \delta + \tilde{\Lambda}_b^{\cdot,j} (1 - \delta) + X_n$
        $\tilde{\tau}_b^j \leftarrow (\tilde{\tau}_b^j + 1)\delta$  and  $\tilde{\Lambda}_b^{\cdot,j} \leftarrow (\tilde{\Lambda}_b^{\cdot,j} + X_n)\delta$.
        if $bA_b^{j,j} - b^2 t_b^j / 2 \leq 0$ then
            $t_b^j \leftarrow \tau_b^j \leftarrow \tilde{\tau}_b^j \leftarrow 0$
            $A_b^{\cdot,j} \leftarrow \Lambda_b^{\cdot,j} \leftarrow \tilde{\Lambda}_b^{\cdot,j} \leftarrow 0$
        $E_b^{\cdot,j} \leftarrow \Lambda_b^{\cdot,j} / (\tau_b^j \vee 1)^{1/2}$
        $Q_b^j \leftarrow \sum_{j' \in [p] \setminus \{j\}} (E_b^{j',j})^2 \mathbb{1}_{\{|E_b^{j',j}| \geq a\}}$
    $S^{\mathrm{diag}} \leftarrow \max_{(j,b) \in [p] \times (\mathcal{B} \cup \mathcal{B}_0)} \bigl(bA_b^{j,j} - b^2 t_b^j / 2\bigr)$
    $S^{\mathrm{off}} \leftarrow \max_{(j,b) \in [p] \times \mathcal{B}} Q_b^j$
until $S^{\mathrm{diag}} \geq T^{\mathrm{diag}}$ or $S^{\mathrm{off}} \geq T^{\mathrm{off}}$;
Observe $\ell$ new data vectors $X_{n+1}, \ldots, X_{n+\ell}$
Set $\Xi_b^{j',j} \leftarrow \frac{\Lambda_b^{j',j} + \sum_{i=n+1}^{n+\ell} X_i^{j'}}{\sqrt{(\tau_b^j + \ell) \vee 1}}$ for $j', j \in [p]$, $b \in \mathcal{B} \cup \mathcal{B}_0$
Compute $\tilde{Q}_b^j \leftarrow \sum_{j' \in [p] \setminus \{j\}} (\Xi_b^{j',j})^2 \mathbb{1}_{\{|\Xi_b^{j',j}| \geq a\}}$ for $j \in [p]$, $b \in \mathcal{B}$
$(\hat{j}, \hat{b}) \leftarrow \mathrm{argmax}_{j \in [p], b \in \mathcal{B}} \tilde{Q}_b^j$
$\hat{\mathcal{S}} \leftarrow \bigl\{j \in [p] \setminus \{\hat{j}\} : |\Xi_{\hat{b}}^{j,\hat{j}}| - b_{\min}(\tau_{\hat{b}}^{\hat{j}} + \ell)^{1/2} \geq d_1\bigr\}$
for $j \in \hat{\mathcal{S}}$ do
    $\tilde{b}^j \leftarrow \mathrm{sgn}(\Xi_{\hat{b}}^{j,\hat{j}}) \max\bigl\{b \in (\mathcal{B} \cup \mathcal{B}_0) \cap (0, \infty) : |\Xi_{\hat{b}}^{j,\hat{j}}| - b(\tau_{\hat{b}}^{\hat{j}} + \ell)^{1/2} \geq d_1\bigr\}$
Output: Confidence interval $\mathcal{C} = \Bigl[\max\Bigl\{n - \min_{j \in \hat{\mathcal{S}}}\Bigl\{t_{\tilde{b}^j}^j + \frac{d_2}{(\tilde{b}^j)^2}\Bigr\}, 0\Bigr\}, n\Bigr]$
```

$a = C_1\sqrt{\log\{p\gamma(\beta^{-2} \vee 1)\alpha^{-1}\}}$, $T^{\mathrm{diag}} = \log\{16p\gamma\log_2(4p)\}$, $T^{\mathrm{off}} = 8\log\{16p\gamma\log_2(2p)\}$, $\ell \geq 0$, $d_1 = C_2 a$ and $d_2 = 4d_1^2$ in Algorithm 3.2, the output confidence interval $\mathcal{C}$ satisfies

$$\mathbb{P}_{z,\theta}(z \in \mathcal{C}) \geq 1 - \alpha.$$

As mentioned in Section 3.2.1, our coverage guarantee in Theorem 3.1 holds even with $\ell = 0$, i.e. with no additional sampling. The condition $z \leq 2\alpha\gamma$ ensures that the probability of a false alarm is at most $\alpha/2$, so that $\mathbb{P}_{z,\theta}(N \leq z) \leq \alpha/2$.

We now provide a guarantee on the length of the `ocd_CI′` confidence interval.

Theorem 3.2. *Assume that θ has an effective sparsity of $s := s(\theta) \geq 2$. Fix $\alpha \in (0,1)$ and $\gamma \geq 1$, and assume that $z \leq 2\alpha\gamma$. Then there exist universal constants $C_1, C_2, C_3, C_4 > 0$ such that, with inputs $(X_t)_{t\in\mathbb{N}}$, $0 < \beta \leq \vartheta$, $a = C_1\sqrt{\log\{p\gamma(\beta^{-2} \vee 1)\alpha^{-1}\}}$, $T^{\mathrm{diag}} = \log\{16p\gamma\log_2(4p)\}$, $T^{\mathrm{off}} = 8\log\{16p\gamma\log_2(2p)\}$, $d_1 = C_2 a$, $d_2 = 4d_1^2$ and $\ell \geq C_3\bigl(\frac{a^2 s\log_2(2p)}{\beta^2} + 1\bigr)$ in Algorithm 3.2, the length L of the output confidence interval $\mathcal{C}$ satisfies*

$$\mathbb{P}_{z,\theta}\biggl\{L > C_4\biggl(\frac{s\log_2(2p)\log\{p\gamma(\beta^{-2} \vee 1)\alpha^{-1}\}}{\beta^2} + 1\biggr)\biggr\} \leq \alpha.$$

The main conclusion of Theorem 3.2 is that, with high probability, the length of the confidence interval is of the same order, up to a logarithmic factor, as the average detection delay guarantee for the `ocd′` procedure (Theorem 2.4). Note that the choices of inputs in Theorem 3.2 are identical to those in Theorem 3.1, except that we now ask for some additional observations after the changepoint declaration, the number of which is of the same order of magnitude as the length of the interval.

3.3.2 Support recovery

Recall the definition of $\mathcal{S}(\theta)$ from the beginning of this section, and denote $\mathcal{S}_\beta(\theta) := \bigl\{j \in [p] : |\theta^j| \geq b_{\min}\bigr\}$, where $b_{\min}$, defined in Algorithm 3.2, is the smallest positive scale in $\mathcal{B} \cup \mathcal{B}_0$, We will suppress the dependence on θ of both these quantities in this subsection. Theorem 3.3 below provides a support recovery guarantee for $\hat{\mathcal{S}}$, defined in Algorithm 3.2. Since neither $\hat{\mathcal{S}}$ nor the anchor coordinate $\hat{j}$ defined in the algorithm depend on d_2, we omit its specification; the choices of other tuning parameters mimic those in Theorems 3.1 and 3.2.

Theorem 3.3. *Assume the conditions of Theorem 3.1.*

(a) There exist universal constants $C_1, C_2 > 0$, such that with inputs $(X_t)_{t\in\mathbb{N}}$, $0 < \beta \leq \vartheta$, $a = C_1\sqrt{\log\{p\gamma(\beta^{-2} \vee 1)\alpha^{-1}\}}$, $T^{\mathrm{diag}} = \log\{16p\gamma\log_2(4p)\}$, $T^{\mathrm{off}} = 8\log\{16p\gamma\log_2(2p)\}$, $\ell \geq 0$ and $d_1 = C_2 a$ in Algorithm 3.2, we have

$$\mathbb{P}_{z,\theta}(\hat{\mathcal{S}} \subseteq \mathcal{S}_\beta) \geq 1 - \alpha.$$

(b) Assume further that θ has effective sparsity $s := s(\theta) \geq 2$. There exist universal constants $C_1, C_2, C_3 > 0$ such that, with inputs $(X_t)_{t\in\mathbb{N}}$, $0 < \beta \leq \vartheta$, $a = C_1\sqrt{\log\{p\gamma(\beta^{-2} \vee 1)\alpha^{-1}\}}$, $T^{\mathrm{diag}} = \log\{16p\gamma\log_2(4p)\}$, $T^{\mathrm{off}} = 8\log\{16p\gamma\log_2(2p)\}$, $d_1 = C_2 a$ and $\ell \geq C_3\bigl(\frac{a^2 s\log_2(2p)}{\beta^2} + 1\bigr)$ in Algorithm 3.2, we have

$$\mathbb{P}_{z,\theta}(\hat{\mathcal{S}} \cup \{\hat{j}\} \supseteq \mathcal{S}) \geq 1 - \alpha.$$

Note that $\mathcal{S} \subseteq \mathcal{S}_\beta \subseteq \{j \in [p] : \theta^j \neq 0\}$. Thus, part (a) of the theorem reveals that with high probability, our support estimate $\hat{\mathcal{S}}$ does not contain any noise coordinates. Part (b) offers a complementary guarantee on the inclusion of all "big" signal coordinates, provided we augment our support estimate with the anchor coordinate $\hat{j}$. See also the further discussion of this result following Proposition 3.4 below.

We now turn our attention to the optimality of our support recovery algorithm, by establishing a complementary minimax lower bound on the performance of any support estimator. To this end, recall that for any given $\theta \in \mathbb{R}^p$ and $z \in \mathbb{N} \cup \{0\}$, we write $\mathbb{P}_{z,\theta}$ for a probability measure under which $(X_n)_{n\in\mathbb{N}}$ are independent with $X_n \sim \mathcal{N}_p(\theta\mathbb{1}_{\{n>z\}}, I_p)$. We further denote by $\mathbb{P}_{z,\theta}^{(n_0)}$ the restriction of $\mathbb{P}_{z,\theta}$ to the filtration $\mathcal{F}_{n_0} := \sigma(X_1, \ldots, X_{n_0})$. For $r > 0$ and $m \in [p] \cup \{0\}$, write

$$\Theta_{r,m} := \bigl\{\theta \in \mathbb{R}^p : |\{j \in [p] : |\theta^j| \leq 1/(8\sqrt{r})\}| \geq m\bigr\}.$$

Define $\mathcal{T}$ to be the set of stopping times with respect to the natural filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$, and set

$$\mathcal{T}_{r,m} := \biggl\{N \in \mathcal{T} : \sup_{z\in\mathbb{N}\cup\{0\},\theta\in\Theta_{r,m}} \mathbb{P}_{z,\theta}(N > z + r) \leq \frac{1}{4}\biggr\}.$$

Write $2^{[p]}$ for the power set of $[p]$, equipped with the symmetric difference metric $d : (A, B) \mapsto |(A \setminus B) \cup (B \setminus A)|$. For any stopping time N, denote

$$\mathcal{J}_N := \{\psi : (\mathbb{R}^p)^\infty \to 2^{[p]} : \psi \text{ is } \mathcal{F}_N\text{-measurable}\},$$

where we recall that ψ is said to be $\mathcal{F}_N$-measurable if for any $A \in 2^{[p]}$ and $n \in \mathbb{N}$, we have $\psi^{-1}(A) \cap \{N = n\}$ is $\mathcal{F}_n$-measurable.

Proposition 3.4. *For $r > 0$ and $m \geq 15$, we have*

$$\inf_{N\in\mathcal{T}_{r,m}} \inf_{\psi\in\mathcal{J}_N} \sup_{z\in\mathbb{N}\cup\{0\},\theta\in\Theta_{r,m}} \mathbb{E}_{z,\theta}\, d\bigl(\psi(X_1, X_2, \ldots), \mathrm{supp}(\theta)\bigr) \geq \frac{m}{32}.$$

Proposition 3.4 can be interpreted as an optimality guarantee for the support recovery property of the `ocd_CI′` algorithm presented in Theorem 3.3(b). To see this, recall from the definition of $\mathcal{S}$ and Theorem 3.3(b) that with high probability, the `ocd_CI′` algorithm with inputs as given in that result selects all signal coordinates whose magnitude exceeds $\vartheta/s^{1/2}$, up to logarithmic factors. Focusing on the case $\beta = \vartheta$ and where s/ϑ^2 bounded away from zero for simplicity of discussion, Proposition 3.7 also reveals that this version of the `ocd_CI`$'$ algorithm belongs to $\mathcal{T}_{r,m}$ with r of order s/ϑ^2, up to logarithmic factors, and $m = |\{j : |\theta^j| \leq 1/(8\sqrt{r})\}|$. Proposition 3.4 considers any support estimation algorithm obtained from a stopping time belonging to the same class, and we note that such a competing procedure is even allowed to store all data up to this stopping time, in contrast to our online algorithm. The result shows that any such support estimation algorithm makes on average a non-vanishing fraction of errors in distinguishing between noise coordinates and signals that are below the level $\vartheta/s^{1/2}$, again up to logarithmic factors. In other words, with high probability, the `ocd_CI`$'$ algorithm selects all signals that are strong enough (up to logarithmic factors) to be reliably detected, while at the same time including no noise coordinates (see Theorem 3.3(a)).

3.4 Numerical studies

In this section, we study the empirical performance of the `ocd_CI` algorithm. Recall that in `ocd_CI`, the off-diagonal statistics Q_b^j are computed using tail partial sums of length t_b^j and that we do not have any extra sampling beyond the time of declaration that a change has occurred.

3.4.1 Tuning parameters

Recall that in `ocd_CI`, the off-diagonal statistics Q_b^j are computed using tail partial sums of length t_b^j and that we do not have any extra sampling beyond the time of declaration that a change has occurred (i.e. $\ell = 0$ as in `ocd_CI`$'$). In Section 2.4, we found that the theoretical choice of thresholds T^{diag} and T^{off} for the `ocd` procedure were a little conservative, and therefore recommended determining these thresholds via Monte Carlo simulation; we replicate the method for choosing these thresholds described in Section 2.4.1. Likewise, as in Section 2.4, we take $a = \sqrt{2 \log p}$ in our simulations.

This leaves us with the choice of tuning parameters d_1 and d_2. As suggested by Theorems 3.1 and 3.2, we take $d_2 = 4d_1^2$. Finally, again as suggested by our theory, we take d_1 to be of the form $d_1 = c\sqrt{\log(p/\alpha)}$, and then tune the parameter $c > 0$ through Monte Carlo simulation, as we now describe. We considered the parameter settings $p \in \{100, 500\}$, $s \in \{2, \lfloor\sqrt{p}\rfloor, p\}$, $\vartheta \in \{2, 1, 1/2\}$, $\alpha = 0.05$, $\beta \in \{2\vartheta, \vartheta, \vartheta/2\}$, $\gamma = 30000$ and $z = 500$. Then, with θ generated as ϑU, where U is uniformly distributed on the union of all s-sparse unit spheres in $\mathbb{R}^p$ (independent of our data), we studied the coverage probabilities, estimated over 2000 repetitions as c varies, of the `ocd_CI` confidence interval for data generated according to the Gaussian model defined at the beginning of Section 3.3. Figure 3.1 displays a subset of the results (the omitted curves were qualitatively similar).

A large c results in a smaller support estimate set $\hat{\mathcal{S}}$ due to the threshold d_1 being higher; the output confidence interval is then longer due to both $\hat{\mathcal{S}}$ being smaller and d_2 being larger. From the left panel of Figure 3.1, we can see that, for a given c, the coverage probability decreases as the signal becomes denser. One reason for this is that we are using the 'sparse' version of the `ocd` algorithm. Another reason is that when the signal is denser and more spread out, more coordinates will enter $\hat{\mathcal{S}}$. This could lead to a more liberal confidence interval, as we are taking the minimum over all coorindates in $\hat{\mathcal{S}}$ in the final construction. The right panel indicates that an overspecification of β will likely yield a shorter and more liberal confidence interval, while an underestimation will result in a slightly more conservative one (see also below in Section 3.4.2). Therefore, a larger choice of c is needed to guarantee the nominal coverage in the case of a dense change or an overspecification of β. We recommend $c = 0.5$ as a safe choice across a wide range of data generating mechanisms, and we used this value of c throughout our confidence interval simulations.

The previous three paragraphs, in combination with Algorithm 3.1, provide the practical implementation of the `ocd_CI` algorithm that we use in our numerical studies and that we recommend for practitioners. The only quantity that remains for the practitioner to input (other than the data) is β, which represents a lower bound on the Euclidean norm of the vector of mean change. Fortunately, this description makes β easily interpretable by practitioners, and a sensible choice should typically be possible using domain knowledge.

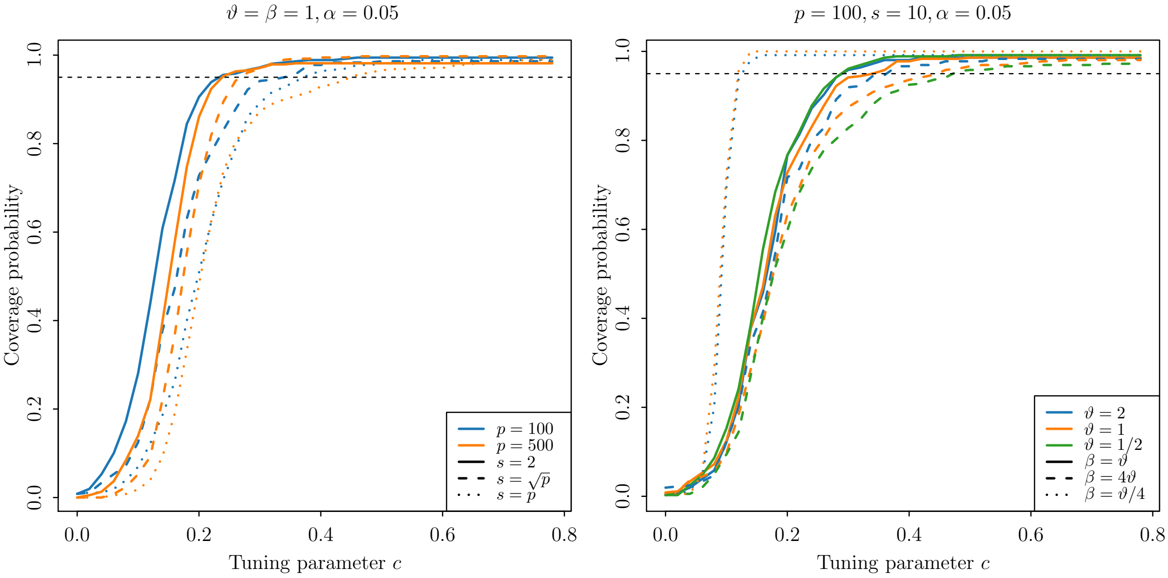

Fig. 3.1 Coverage probabilities of the `ocd_CI` confidence interval as the parameter c, involved in the choice of tuning parameter d_1, varies.

3.4.2 Coverage probability and interval length

In Table 3.1, we present the detection delay of the `ocd` procedure, as well as the coverage probabilities and average confidence interval lengths of the `ocd_CI` procedure, all estimated over 2000 repetitions, with the same set of parameter choices and data generating mechanism as in Section 3.4.1. From this table, we see that the coverage probabilities are at least at the nominal level (up to Monte Carlo error) across all settings considered. Underspecification of β means that the grid of scales that can be chosen for indices in $\hat{\mathcal{S}}$ is shifted downwards, and therefore increases the probability that $\tilde{b}^j$ will significantly underestimate θ^j for $j \in \hat{\mathcal{S}}$. In turn, this leads to a slight conservativeness for the coverage probability (and corresponding increased average confidence interval length). On the other hand, overspecification of β yields a shorter interval on average, though these were nevertheless able to retain the nominal coverage in all cases considered.

Another interesting feature of Table 3.1 is to compare the average confidence interval lengths with the corresponding average detection delays. Theorems 3.2 and 2.4 indicate that both of these quantities are of order $(s/\beta^2) \vee 1$, up to polylogarithmic factors in p and γ, but of course whenever the confidence interval includes the changepoint, its length must be at least as long as the detection delay. Nevertheless, in most settings, it is only 2 to 3 times longer on average, and in all cases considered was less than 7 times longer on average. Moreover, we can also observe that the confidence interval length increases with s and decreases with β, as anticipated by our theory.

From a practical perspective, it is also important to assess the robustness of the `ocd_CI` methodology to departures from the class of data generating mechanisms that underpins our theory in Section 3.3. In particular, in the remainder of this subsection, we seek to assess the impact of both spatial and temporal dependence on the performance of the `ocd_CI` algorithm, while in our analysis of S&P 500 data in Section 3.4.4, we also investigate the effect of heavy tails. Taking spatial dependence first, we consider a setting where the cross-sectional covariance matrix $\Sigma = (\Sigma_{jk})_{j,k\in[p]}$ for each observation is Toeplitz with parameter $\rho \in \{0.5, 0.75\}$; in other words, $\Sigma_{jk} = \rho^{|j-k|}$. Table 3.2 presents the results of applying the `ocd_CI` methodology in an unmodified way to this new data generating mechanism. Reassuringly, the coverage of the `ocd_CI` confidence intervals remains perfectly satisfactory in all settings considered, and moreover, the lengths of the confidence intervals are very similar to those in Table 3.1 where we have an identity cross-sectional covariance matrix.

Regarding temporal dependence, we consider an autoregressive AR(1) model for the sequentially-observed data with coordinate-dependent autoregressive parameters, so that $X_1^j = \epsilon_1^j$ and

$$X_t^j = \rho_j \cdot X_{t-1}^j + \sqrt{1-\rho_j^2} \cdot \epsilon_t^j$$

for $t = 2, \ldots, n$, with independent standard normal random variables $(\epsilon_t^j)_{t\in[n],j\in[p]}$. As with all (offline or online) changepoint procedures, the presence of temporal correlations with p new

p	s	ϑ	β	Detection Delay	Coverage (%)	CI Length
100	2	2	4	9.8$_{(0.1)}$	96.2$_{(0.4)}$	20.1$_{(0.7)}$
100	2	2	2	12.6$_{(0.1)}$	97.0$_{(0.4)}$	33.7$_{(0.7)}$
100	2	2	1	14.1$_{(0.1)}$	97.9$_{(0.3)}$	80.8$_{(1.0)}$
100	2	1	2	34.2$_{(0.3)}$	95.8$_{(0.4)}$	66.1$_{(1.0)}$
100	2	1	1	44.2$_{(0.3)}$	97.5$_{(0.4)}$	122.0$_{(1.4)}$
100	2	1	0.5	52.0$_{(0.4)}$	97.4$_{(0.4)}$	309.1$_{(2.0)}$
100	10	2	4	14.7$_{(0.1)}$	96.0$_{(0.4)}$	32.5$_{(0.8)}$
100	10	2	2	15.7$_{(0.1)}$	97.4$_{(0.4)}$	38.4$_{(0.8)}$
100	10	2	1	15.9$_{(0.1)}$	97.0$_{(0.4)}$	80.2$_{(1.1)}$
100	10	1	2	52.6$_{(0.5)}$	96.2$_{(0.4)}$	114.0$_{(1.5)}$
100	10	1	1	56.9$_{(0.4)}$	97.1$_{(0.4)}$	142.5$_{(1.8)}$
100	10	1	0.5	60.2$_{(0.4)}$	98.2$_{(0.3)}$	301.1$_{(1.6)}$
100	100	2	4	27.2$_{(0.2)}$	96.1$_{(0.4)}$	77.6$_{(0.9)}$
100	100	2	2	27.7$_{(0.2)}$	96.0$_{(0.4)}$	81.8$_{(1.0)}$
100	100	2	1	28.2$_{(0.2)}$	97.5$_{(0.3)}$	99.4$_{(1.3)}$
100	100	1	2	100.7$_{(0.8)}$	94.7$_{(0.5)}$	292.8$_{(3.5)}$
100	100	1	1	100.5$_{(0.9)}$	96.3$_{(0.4)}$	296.0$_{(3.4)}$
100	100	1	0.5	103.2$_{(0.8)}$	97.3$_{(0.4)}$	365.9$_{(2.8)}$
500	2	2	4	11.3$_{(0.1)}$	97.2$_{(0.4)}$	23.1$_{(0.7)}$
500	2	2	2	15.8$_{(0.1)}$	97.7$_{(0.3)}$	45.2$_{(0.9)}$
500	2	2	1	17.7$_{(0.1)}$	97.5$_{(0.4)}$	117.3$_{(1.0)}$
500	2	1	2	41.5$_{(0.3)}$	97.3$_{(0.4)}$	81.8$_{(1.2)}$
500	2	1	1	55.0$_{(0.4)}$	96.8$_{(0.4)}$	168.9$_{(1.5)}$
500	2	1	0.5	64.6$_{(0.5)}$	98.1$_{(0.3)}$	445.0$_{(1.7)}$
500	22	2	4	23.6$_{(0.2)}$	96.3$_{(0.4)}$	55.4$_{(1.0)}$
500	22	2	2	25.0$_{(0.2)}$	97.0$_{(0.4)}$	60.3$_{(0.8)}$
500	22	2	1	25.5$_{(0.2)}$	98.1$_{(0.3)}$	119.7$_{(0.8)}$
500	22	1	2	88.1$_{(0.7)}$	97.0$_{(0.4)}$	203.5$_{(2.1)}$
500	22	1	1	91.9$_{(0.6)}$	97.8$_{(0.3)}$	229.7$_{(2.2)}$
500	22	1	0.5	94.9$_{(0.6)}$	98.3$_{(0.3)}$	462.8$_{(1.4)}$
500	500	2	4	79.8$_{(0.6)}$	95.0$_{(0.5)}$	238.9$_{(2.7)}$
500	500	2	2	80.3$_{(0.6)}$	95.8$_{(0.4)}$	245.7$_{(2.6)}$
500	500	2	1	80.9$_{(0.6)}$	97.5$_{(0.4)}$	250.2$_{(2.5)}$
500	500	1	2	290.5$_{(2.3)}$	94.5$_{(0.5)}$	819.7$_{(7.9)}$
500	500	1	1	291.4$_{(2.3)}$	95.2$_{(0.5)}$	831.1$_{(7.5)}$
500	500	1	0.5	297.3$_{(2.3)}$	98.1$_{(0.3)}$	875.0$_{(6.7)}$

Table 3.1 Estimated coverage, average length of the `ocd_CI` confidence interval and average detection delay over 2000 repetitions, with standard errors in brackets. Other parameters: $\gamma = 30000$, $z = 1000$, $\alpha = 0.05$, $a = \sqrt{2\log p}$, $c = 0.5$, $d_1 = c\sqrt{\log(p/\alpha)}$, $d_2 = 4d_1^2$.

ρ	s	ϑ	Detection Delay	Coverage (%)	CI Length
0.5	2	2	$13.9_{(0.1)}$	$98.5_{(0.3)}$	$35.5_{(1.0)}$
0.5	2	1	$49.1_{(0.3)}$	$99.0_{(0.2)}$	$125.1_{(1.6)}$
0.5	2	0.5	$172.5_{(1.0)}$	$99.5_{(0.2)}$	$447.0_{(2.8)}$
0.5	10	2	$21.9_{(0.1)}$	$98.7_{(0.3)}$	$42.0_{(0.9)}$
0.5	10	1	$76.1_{(0.5)}$	$98.8_{(0.2)}$	$154.2_{(1.5)}$
0.5	10	0.5	$266.7_{(1.8)}$	$99.0_{(0.2)}$	$566.9_{(3.9)}$
0.5	100	2	$52.1_{(0.3)}$	$98.3_{(0.3)}$	$106.8_{(0.9)}$
0.5	100	1	$187.7_{(1.3)}$	$98.4_{(0.3)}$	$399.5_{(3.3)}$
0.5	100	0.5	$655.3_{(5.0)}$	$98.5_{(0.3)}$	$1366.2_{(10.1)}$
0.75	2	2	$13.9_{(0.1)}$	$96.9_{(0.4)}$	$51.1_{(2.6)}$
0.75	2	1	$47.9_{(0.3)}$	$96.8_{(0.4)}$	$146.0_{(3.3)}$
0.75	2	0.5	$171.5_{(1.1)}$	$97.7_{(0.3)}$	$463.8_{(4.2)}$
0.75	10	2	$21.8_{(0.2)}$	$96.4_{(0.4)}$	$48.6_{(1.7)}$
0.75	10	1	$75.3_{(0.5)}$	$96.7_{(0.4)}$	$165.0_{(2.7)}$
0.75	10	0.5	$266.3_{(1.9)}$	$96.0_{(0.4)}$	$558.9_{(4.5)}$
0.75	100	2	$50.9_{(0.3)}$	$96.8_{(0.4)}$	$106.8_{(1.2)}$
0.75	100	1	$184.8_{(1.4)}$	$95.6_{(0.5)}$	$401.8_{(3.8)}$
0.75	100	0.5	$647.3_{(5.4)}$	$94.6_{(0.5)}$	$1312.3_{(11.2)}$

Table 3.2 Spatial dependence. Estimated coverage and average length of the `ocd_CI` confidence interval and average detection delay over 2000 repetitions, with standard errors in brackets, under a Toeplitz cross-sectional covariance matrix Σ with entries $\Sigma_{jk} = \rho^{|j-k|}$ for $j, k \in [p]$. Other parameters: $p = 100$, $\beta = \vartheta$, $\gamma = 30000$, $z = 1000$, $\alpha = 0.05$, $a = \sqrt{2 \log p}$, $c = 0.5$, $d_1 = c\sqrt{\log(p/\alpha)}$, $d_2 = 4d_1^2$.

parameters in the model significantly increases the difficulty of the challenge. Therefore, in order to learn the autoregressive parameters, we also allow the algorithm access to a historical offline burn-in data set of length 1000 generated under the null. This was used initially to construct maximum likelihood estimates $\hat{\rho}_j$ of ρ_j, for $j \in [p]$. Next, we used the burn-in data again to compute estimates $\hat{\sigma}_j$ of the standard deviation of the auto-regressive residuals in the jth coordinate. In applying the `ocd_CI` algorithm subsequently, we pre-processed each new observation by replacing X_t^j with $(X_t^j - \hat{\rho}_j X_{t-1}^j)/\hat{\sigma}_j$ for $j = 2, 3, \ldots$. In the simulation results reported in Table 3.3 below, we generated $\rho_j \overset{\text{iid}}{\sim} U[-\rho_{\max}, \rho_{\max}]$ for $j \in [p]$, independent of all other sources of randomness, with $\rho_{\max} \in \{0.5, 0.75\}$. Again, we find that all of the coverage probabilities are satisfactory, and the confidence interval lengths in fact decrease slightly as the extent of the dependence increases (due to the declaration time typically being a little shorter).

$\rho_{\max}$	s	ϑ	Detection Delay	Coverage (%)	CI Length
0.5	2	2	$11.5_{(0.1)}$	$95.8_{(0.4)}$	$34.4_{(1.1)}$
0.5	2	1	$41.5_{(0.3)}$	$96.5_{(0.4)}$	$118.2_{(1.7)}$
0.5	2	0.5	$149.9_{(1.0)}$	$98.5_{(0.3)}$	$422.2_{(2.8)}$
0.5	10	2	$14.2_{(0.1)}$	$95.7_{(0.5)}$	$38.3_{(1.1)}$
0.5	10	1	$50.6_{(0.4)}$	$95.7_{(0.5)}$	$130.4_{(1.6)}$
0.5	10	0.5	$192.2_{(1.6)}$	$96.9_{(0.4)}$	$480.3_{(3.7)}$
0.5	100	2	$24.8_{(0.2)}$	$95.1_{(0.5)}$	$75.5_{(1.2)}$
0.5	100	1	$91.8_{(0.8)}$	$95.8_{(0.4)}$	$271.0_{(3.2)}$
0.5	100	0.5	$346.9_{(3.0)}$	$96.0_{(0.4)}$	$939.7_{(8.8)}$
0.75	2	2	$10.6_{(0.1)}$	$96.2_{(0.4)}$	$32.4_{(1.0)}$
0.75	2	1	$37.1_{(0.3)}$	$96.3_{(0.4)}$	$111.4_{(1.4)}$
0.75	2	0.5	$136.0_{(1.1)}$	$97.0_{(0.4)}$	$403.0_{(2.6)}$
0.75	10	2	$12.2_{(0.1)}$	$95.0_{(0.5)}$	$35.1_{(1.0)}$
0.75	10	1	$45.5_{(0.4)}$	$96.3_{(0.4)}$	$121.8_{(1.6)}$
0.75	10	0.5	$174.6_{(1.4)}$	$97.8_{(0.3)}$	$452.4_{(3.2)}$
0.75	100	2	$21.2_{(0.2)}$	$96.1_{(0.4)}$	$64.9_{(1.1)}$
0.75	100	1	$77.4_{(0.7)}$	$95.8_{(0.4)}$	$229.9_{(2.8)}$
0.75	100	0.5	$293.9_{(2.6)}$	$96.2_{(0.4)}$	$820.1_{(8.1)}$

Table 3.3 Temporal dependence. Estimated coverage and average length of the `ocd_CI` confidence interval and average detection delay over 2000 repetitions, with standard errors in brackets, under the AR(1) model with autoregressive parameters $\rho_j \overset{\text{iid}}{\sim} U[-\rho_{\max}, \rho_{\max}]$. Other parameters: $p = 100$, $\beta = \vartheta$, $\gamma = 30000$, $z = 1000$, $\alpha = 0.05$, $a = \sqrt{2 \log p}$, $c = 0.5$, $d_1 = c\sqrt{\log(p/\alpha)}$, $d_2 = 4d_1^2$.

Based on our numerical investigations, we make the following recommendations: for spatial dependence, we advocate applying the `ocd_CI` methodology in an unmodified fashion, and (at least when the dependence across coordinates is not too severe), our experience is that the performance of the algorithm is relatively unaffected. Temporal dependence, on the other hand, presents very significant challenges for changepoint algorithms, and careful modelling of this dependence is recommended in order to facilitate the construction of appropriate residuals for which the main effect of the dependence has been removed. Where an appropriate model for the temporal dependence structure is available, we have found that the `ocd_CI` algorithm again performs well.

3.4.3 Support recovery

We now turn our attention to the empirical support recovery properties of the quantity $\hat{\mathcal{S}}$ (in combination with the anchor coordinate $\hat{j}$) computed in the `ocd_CI` algorithm. In Table 3.4, we present the probabilities, estimated over 500 repetitions, that $\hat{\mathcal{S}} \subseteq \mathcal{S}_\beta$ and that $\hat{\mathcal{S}} \cup \{\hat{j}\} \supseteq \mathcal{S}$ for $p = 100$, $s \in \{5, 50\}$, $\vartheta \in \{1, 2\}$, and for three different signal shapes: in the uniform, inverse square root and harmonic cases, we took $\theta \propto (\mathbb{1}_{\{j \in [s]\}})_{j \in [p]}$, $\theta \propto (j^{-1/2}\mathbb{1}_{\{j \in [s]\}})_{j \in [p]}$ and $\theta \propto (j^{-1}\mathbb{1}_{\{j \in [s]\}})_{j \in [p]}$ respectively. As inputs to the algorithm, we set $a = \sqrt{2 \log p}$, $\alpha = 0.05$, $d_1 = \sqrt{2 \log(p/\alpha)}$, $\beta = \vartheta$, and, motivated by Theorem 3.3, took an additional $\ell = \lceil a^2 s \beta^{-2} \log_2(2p) \rceil$ post-declaration observations in constructing the support estimates. The results reported in Table 3.4 provide empirical confirmation of the support recovery properties claimed in Theorem 3.3.

s	ϑ	Signal Shape	$\hat{\mathcal{S}} \subseteq \mathcal{S}_\beta$ (%)	$\hat{\mathcal{S}} \cup \{\hat{j}\} \supseteq \mathcal{S}$ (%)
5	2	uniform	$99.8_{(0.2)}$	$97.6_{(0.7)}$
5	1	uniform	$100.0_{(0.0)}$	$97.6_{(0.7)}$
50	2	uniform	$100.0_{(0.0)}$	$95.6_{(0.9)}$
50	1	uniform	$100.0_{(0.0)}$	$97.8_{(0.7)}$
5	2	inv sqrt	$99.6_{(0.3)}$	$96.6_{(0.8)}$
5	1	inv sqrt	$100.0_{(0.0)}$	$98.8_{(0.5)}$
50	2	inv sqrt	$100.0_{(0.0)}$	$99.8_{(0.2)}$
50	1	inv sqrt	$100.0_{(0.0)}$	$100.0_{(0.0)}$
5	2	harmonic	$100.0_{(0.0)}$	$97.6_{(0.7)}$
5	1	harmonic	$99.6_{(0.3)}$	$97.8_{(0.7)}$
50	2	harmonic	$100.0_{(0.0)}$	$99.4_{(0.3)}$
50	1	harmonic	$100.0_{(0.0)}$	$100.0_{(0.0)}$

Table 3.4 Estimated support recovery probabilities (with standard errors in brackets). Other settings: $p = 100$, $a = \sqrt{2 \log p}$, $\alpha = 0.05$, $d_1 = \sqrt{2 \log(p/\alpha)}$, $\beta = \vartheta$, and with an additional $\ell = \lceil a^2 s \beta^{-2} \log_2(2p) \rceil$ post-declaration observations.

Finally in this section, we consider the extent to which the additional observations are necessary in practice to provide satisfactory support recovery. In the left panel of Figure 3.2, we plot Receiver Operating Characteristic (ROC) curves to study the estimated support recovery probabilities with $\ell = 0$ (i.e. no additional sampling) as a function of the input parameter d_1, which can be thought of as controlling the trade-off between $\mathbb{P}(\hat{\mathcal{S}} \cup \{\hat{j}\} \supseteq \mathcal{S})$ and $\mathbb{P}(\hat{\mathcal{S}} \subseteq \mathcal{S}_\beta)$. The fact that the triangles in this plot are all to the left of the dotted vertical line confirms the theoretical guarantee provided in Theorem 3.3(a), which holds with $d_1 = \sqrt{2\log(p/\alpha)}$, and even with $\ell = 0$); the less conservative choice $d_1 = \sqrt{2\log p}$, which roughly corresponds to an average of one noise coordinate included in $\hat{\mathcal{S}}$, allows us to capture a larger proportion of the signal. From this panel, we also see that additional sampling is needed to ensure that, with high probability, we recover all of the true signals. This is unsurprising: for instance, with a uniform signal shape and $s = 50$, it is very unlikely that all 50 signal coordinates will have accumulated such similar levels of evidence to appear in $\hat{\mathcal{S}} \cup \{\hat{j}\}$ by the time of declaration. The right panel confirms that, with an inverse square root signal shape, the probability that we capture each signal increases with the signal magnitude, and that even small signals tend to be selected with higher probability than noise coordinates.

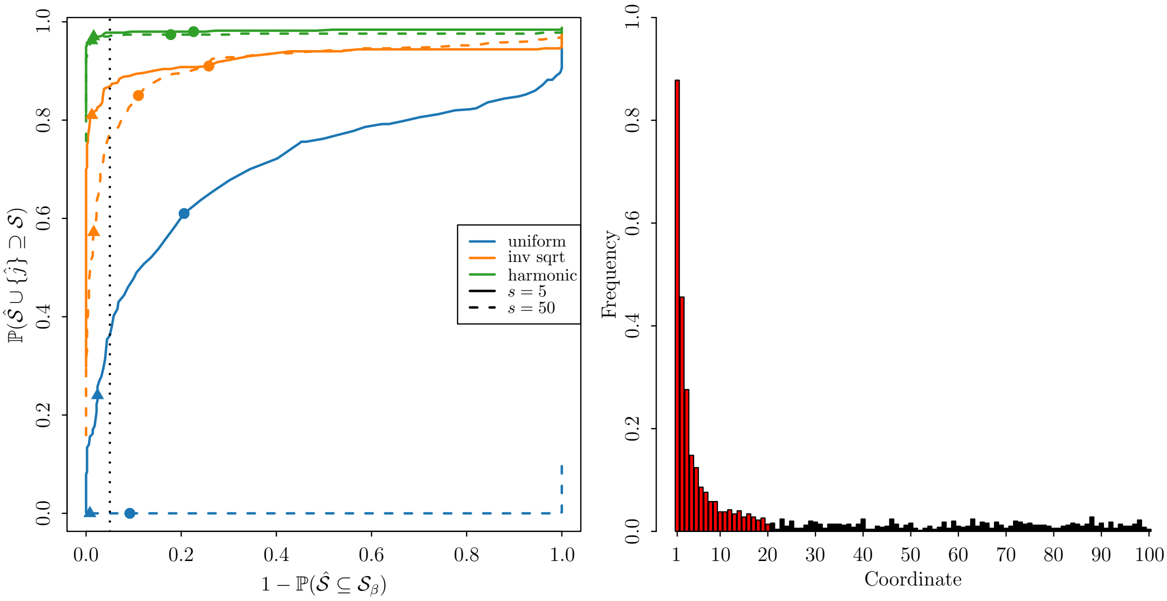

Fig. 3.2 Support recovery properties of `ocd_CI`. In the left panel, we plot ROC curves for three different signal shapes and for sparsity levels $s \in \{5, 50\}$. The triangles and circles correspond to points on the curves with $d_1 = \sqrt{2\log(p/\alpha)}$ (with $\alpha = 0.05$), and $d_1 = \sqrt{2\log p}$ respectively. The dotted vertical line corresponds to $\mathbb{P}(\hat{\mathcal{S}} \subseteq \mathcal{S}_\beta) = 1 - \alpha$. In the right panel, we plot the proportion of 500 repetitions for which each coordinate belongs to $\hat{\mathcal{S}} \cup \{\hat{j}\}$ with $d_1 = \sqrt{2\log p}$; here, the $s = 20$ signals have an inverse square root shape, and are plotted in red; noise coordinates are plotted in black. Other parameters for both panels: $p = 100$, $\beta = \vartheta = 2$, $\ell = 0$, $a = \sqrt{2\log p}$.

3.4.4 Real data examples

US COVID-19 data

We apply `ocd_CI` to a dataset of weekly deaths in the United States between January 2017 and June 2020[2]. The data up to 29 June 2019 are treated as our training data. For each of the 50 states, as well as Washington, D.C. ($p = 51$), we pre-process the data as follows. To remove the seasonal effect, we first estimate the 'seasonal death curve', i.e. the mean death numbers for each day of the year, for each state. The seasonal death curve is estimated by first splitting the weekly death numbers evenly across the seven relevant days, and then estimating the average number of deaths on each day of the year from these derived daily death numbers using a Gaussian kernel with a bandwidth of 20 days. As the death numbers follow an approximate Poisson distribution, we apply a square-root transformation to stabilise the variance; more precisely, the transformed weekly excess deaths are computed as the difference of the square roots of the weekly deaths and the predicted weekly deaths from the seasonal death curve. Finally, we standardise the transformed weekly excess deaths using the mean and standard deviation of the transformed data over the training period. The standardised, transformed data are plotted in Figure 3.3 for 12 states. When applying `ocd_CI` to these data, we take $a = \sqrt{2\log p}$, $T^{\mathrm{diag}} = \log\{16p\gamma\log_2(4p)\}$, $T^{\mathrm{off}} = 8\log\{16p\gamma\log_2(2p)\}$, $d_1 = 0.5\sqrt{\log(p/\alpha)}$ and $d_2 = 4d_1^2$, with $\alpha = 0.05$, $\beta = 50$ and $\gamma = 1000$. On the monitoring data (from 30 June 2019), the `ocd_CI` algorithm declares a change on the week ending 28 March 2020, and provides a confidence interval from the week ending 21 March 2020 to the week ending 28 March 2020. This coincides with the beginning of the first wave of COVID-19 deaths in the United States. The algorithm also identifies New York, New Jersey, Connecticut, Michigan and Louisiana as the estimated support of the change. Interestingly, if we run the `ocd_CI` procedure from the beginning of the training data period (while still standardising as before, due to the lack of available data prior to 2017), it identifies a subtler change on the week ending 6 January 2018, with a confidence interval of [17 December 2017, 6 January 2018]. This corresponds to a bad influenza season at the end of 2017[3].

S&P 500 data

We now use `ocd_CI` to study market movements leading up to the financial crisis of 2007–2008. We selected the $p = 254$ stocks that were both in the S&P 500 listing and were traded throughout the period from 1 January 2006 to 31 December 2007. The historical price data were downloaded from finance.yahoo.com using the `quantmod` `R` package (Ryan et al., 2020); a similar dataset was studied by Cai and Wang (2021). For each stock, we compute the daily logarithmic returns from the adjusted closing prices. We use the data from 2006 as the

[2] Available at: https://www.cdc.gov/nchs/nvss/vsrr/covid19/excess_deaths.htm.

[3] See https://www.cdc.gov/flu/about/season/flu-season-2017-2018.htm

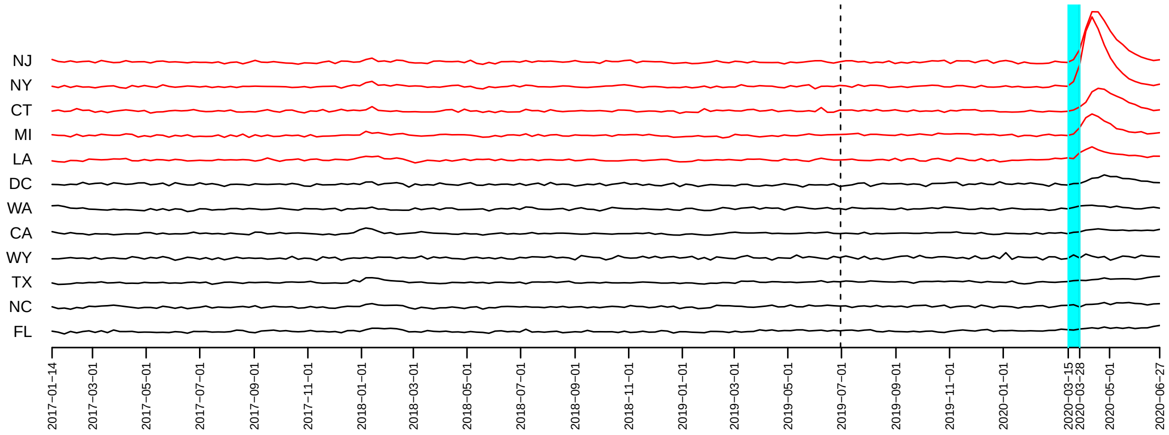

Fig. 3.3 Standardised, transformed weekly excess death data from 12 states (including Washington, D.C.). The monitoring period starts from 30 June 2019 (dashed line). The data from the states in the support estimate are shown in red. The confidence interval [8 March 2020, 28 March 2020] is shown in the light blue shaded region.

training data and standardise the entire data according to the mean and standard deviation over the training period.

A potential difficulty of applying the `ocd_CI` methodology directly to the raw data is the heavy tails that are a characteristic feature of financial return data. Nevertheless, simple transformations such as clipping (or trimming) have recently be shown to be extremely effective transformations for such data — see for instance Minsker (2018), Ke et al. (2019) and Zhu and Zhou (2021), as well as our discussion of robustness below. In our initial analysis, we clip the standardized data at $\pm\Phi^{-1}(0.999)$ as a pre-processing transformation.

When applying the `ocd_CI` procedure to this dataset, we used the same input parameters as in the previous example. So as to be able to use `ocd_CI` repeatedly to identify multiple changes, we also set a cool-down period of 10 trading days (i.e. the monitoring resets and restarts 10 trading days after a change is declared). This allows the market to recover from any loss (or gain) from the previous change so that the same market movement is not identified as more than one changepoint. The first four changes were declared on 27 February 2007, 24 May 2007, 24 July 2007 and 8 August 2007, with corresponding confidence intervals shown in Figure 3.4. This figure also depicts the relative sector impact of each change by showing the percentage of stocks in each sector (according to the Global Industry Classification Standard[4]) that belongs to the estimated support of a changepoint. In particular, the first and last identified changepoints are primarily associated with changes in Real Estate stocks; these correspond to an HSBC announcement indicating loan losses on subprime mortgages in February 2007, and American Home Mortgage Investment Corporation filing for bankruptcy in August 2007 respectively (Hausman and Johnston, 2014).

[4] See https://www.msci.com/our-solutions/indexes/gics

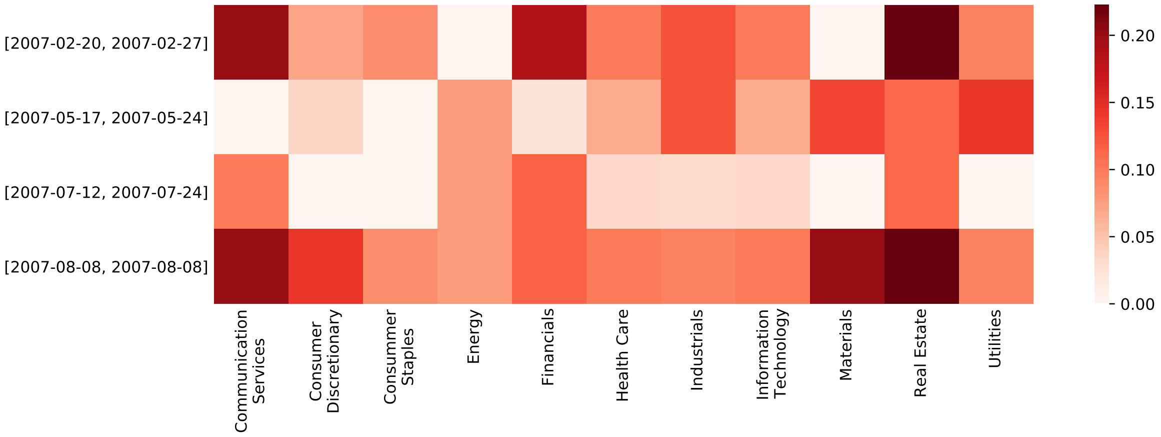

Fig. 3.4 Heatmap of sector impact of the four changepoints in the S&P 500 data identified by `ocd_CI`, measured as a proportion of the stocks in a sector that appear in the support estimate of the changepoint. The confidence intervals for each of the changepoints are given on the left.

To assess the robustness of our conclusions to our pre-processing transformation, we consider alternative clipping levels as well as a different normal transformation technique. More specifically, in addition to clipping the standardized data at $\pm\Phi^{-1}(0.999)$, we also consider clipping at $\pm\Phi^{-1}(1-\delta)$ with $\delta \in \{0.0005, 0.002, 0.004, 0.008\}$. The alternative transformation alluded to above involves applying the following two steps coordinatewise to our data. Given training data $X_1, \ldots, X_n$ with corresponding order statistics $X_{(1)} \leq \ldots \leq X_{(n)}$, we first define the piecewise linear function

$$\hat{F}(x) := \begin{cases} \frac{i}{n+1} + \frac{x - X_{(i)}}{(n+1)(X_{(i+1)} - X_{(i)})} & \text{if } x \in [X_{(i)}, X_{(i+1)}) \text{ with } i \in [n-1] \\ 1/(n+1) & \text{if } x < X_{(1)} \\ n/(n+1) & \text{if } x \geq X_{(n)}. \end{cases}$$

On our test data $X_{n+1}, \ldots, X_{n+m}$, we then compute $Z_i := \Phi^{-1}\bigl(\hat{F}(X_{n+i})\bigr)$ for $i \in [m]$. In Figure 3.5, we present the results of applying these different methods to our S&P 500 data. All of the methods declare four changepoints, and these are located around the same times. Small values of δ are less robust to heavy tails, but have greater power to detect changes, so declaration times tend to be slightly earlier, and the corresponding confidence intervals are somewhat shorter. Nevertheless, the figure is reassuring regarding the robustness of our pre-processing transformation.

3.5 Proofs of main results

Proof of Theorem 3.1. Fix $n > z, j \in [p]$, $b \in \mathcal{B}$ and $j' \in [p] \setminus \{j\}$. We assume, without loss of generality, that $\theta^{j'} \geq 0$. The case $\theta^{j'} < 0$ can be analysed similarly. Recall that $b_{\min}$,

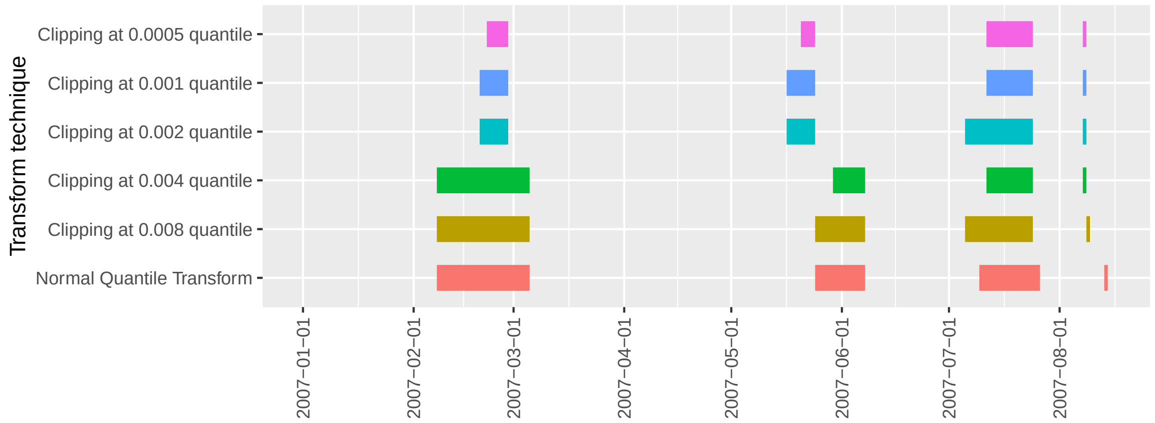

Fig. 3.5 Illustration of the robustness of `ocd_CI` methodology: the figure shows its application on the S&P 500 data with different clipping levels and an alternative normal transformation technique.

defined in Algorithm 3.2, is the smallest positive scale in $\mathcal{B} \cup \mathcal{B}_0$, and write $b^{j'}_{\mathrm{aux}} := \min\bigl\{b \in (\mathcal{B} \cup \mathcal{B}_0) \cap (0, \infty) : b \geq \theta^{j'}\bigr\}$. Then we have $\Lambda^{j',j}_{n,b} + \sum_{i=n+1}^{n+\ell} X^{j'}_i \mid \tau^j_{n,b} \sim \mathcal{N}\bigl(\theta^{j'} \min\{n + \ell - z, \tau^j_{n,b} + \ell\}, \tau^j_{n,b} + \ell\bigr)$. Thus, recalling the definition of $\hat{\mathcal{S}}$ and $\tilde{b}^{j'}$ from Algorithm 3.2, we have

$$\begin{aligned}
&\mathbb{P}\bigl(\{j' \in \hat{\mathcal{S}}\} \cap \{\tilde{b}^{j'} \notin (0, \theta^{j'})\} \cap \{N = n, \hat{\jmath} = j, \hat{b} = b\}\bigr) \\
&\quad = \mathbb{E}\Bigl\{\mathbb{P}\Bigl(\{j' \in \hat{\mathcal{S}}\} \cap \{\tilde{b}^{j'} \notin (-b_{\min}, b^{j'}_{\mathrm{aux}})\} \cap \{N = n, \hat{\jmath} = j, \hat{b} = b\} \Bigm| \tau^j_{n,b}\Bigr)\Bigr\} \\
&\quad \leq \mathbb{E}\biggl\{\mathbb{P}\biggl(\Lambda^{j',j}_{n,b} + \sum_{i=n+1}^{n+\ell} X^{j'}_i \geq b^{j'}_{\mathrm{aux}}(\tau^j_{n,b} + \ell) + d_1\bigl(\tau^j_{n,b} + \ell\bigr)^{1/2} \biggm| \tau^j_{n,b}\biggr)\biggr\} \\
&\qquad\quad + \mathbb{E}\biggl\{\mathbb{P}\biggl(\Lambda^{j',j}_{n,b} + \sum_{i=n+1}^{n+\ell} X^{j'}_i \leq -b_{\min}(\tau^j_{n,b} + \ell) - d_1\bigl(\tau^j_{n,b} + \ell\bigr)^{1/2} \biggm| \tau^j_{n,b}\biggr)\biggr\} \\
&\quad \leq \mathbb{E}\bigl\{\bar{\Phi}\bigl((b^{j'}_{\mathrm{aux}} - \theta^{j'})(\tau^j_{n,b} + \ell)^{1/2} + d_1\bigr)\bigr\} + \mathbb{E}\bigl\{\bar{\Phi}\bigl((b_{\min} + \theta^{j'})(\tau^j_{n,b} + \ell)^{1/2} + d_1\bigr)\bigr\} \\
&\quad \leq 2\bar{\Phi}(d_1). \qquad (3.2)
\end{aligned}$$

Moreover, by a similar argument to (3.15) in the proof of Proposition 3.5, for $b \in (0, \theta^{j'})$, we have

$$\mathbb{P}\bigl(n - t^{j'}_{n,b} - d_2/b^2 > z\bigr) \leq 2\bar{\Phi}\biggl(\frac{\sqrt{d_2}}{b}(\theta^{j'} - b/2)\biggr) \leq 2\bar{\Phi}\bigl(\sqrt{d_2}/2\bigr). \qquad (3.3)$$

Combining (3.2) and (3.3), we have

$$\begin{aligned}
&\mathbb{P}\Bigl(\{j' \in \hat{\mathcal{S}}\} \cap \{n - t^{j'}_{n,\tilde{b}^{j'}} - d_2/(\tilde{b}^{j'})^2 > z\} \cap \{N = n, \hat{\jmath} = j, \hat{b} = b\}\Bigr) \\
&\leq \mathbb{P}\Bigl(\{j' \in \hat{\mathcal{S}}\} \cap \{\tilde{b}^{j'} \notin (0, \theta^{j'})\} \cap \{N = n, \hat{\jmath} = j, \hat{b} = b\}\Bigr) + \sum_{b \in (\mathcal{B} \cup \mathcal{B}_0) \cap (0, \theta^j)} 2\bar{\Phi}\bigl(\sqrt{d_2}/2\bigr) \\
&\leq 2\bar{\Phi}(d_1) + 2\log_2(4p)\bar{\Phi}\bigl(\sqrt{d_2}/2\bigr).
\end{aligned}$$

Now, write

$$r_0 := \biggl(\frac{24T^{\mathrm{off}}}{\vartheta^2} \vee \frac{12(a^2 \vee 8\log 2)s}{\vartheta^2} \vee \frac{128(T^{\mathrm{diag}} + \log(8/\alpha))s}{\beta^2}\biggr)\log_2(2p) + 2. \tag{3.4}$$

By a union bound and Proposition 3.7, we have

$$\begin{aligned}
&\mathbb{P}\biggl(N - \min_{j\in\hat{\mathcal{S}}}\Bigl\{t^j_{N,\tilde{b}^j} + \frac{d_2}{(\tilde{b}^j)^2}\Bigr\} > z\biggr)\\
&\leq \mathbb{P}(N > z + r_0)\\
&\quad + \sum_{n=z+1}^{z+\lfloor r_0\rfloor}\sum_{j=1}^{p}\sum_{b\in\mathcal{B}}\sum_{j'=1}^{p}\mathbb{P}\biggl(\{j' \in \hat{\mathcal{S}}\} \cap \Bigl\{n - t^{j'}_{n,\tilde{b}^{j'}} - \frac{d_2}{(\tilde{b}^{j'})^2} > z\Bigr\} \cap \{N = n, \hat{\jmath} = j, \hat{b} = b\}\biggr)\\
&\leq \exp\biggl\{-\frac{\beta^2(r_0-1)}{48\log_2(2p)}\biggr\} + p\exp\biggl\{-\frac{\vartheta^2(r_0-1)}{128\log_2(2p)}\biggr\} + 4p^2\log_2^2(4p)r_0\bigl\{\bar{\Phi}(d_1) + \bar{\Phi}\bigl(\sqrt{d_2}/2\bigr)\bigr\}.
\end{aligned}$$

Therefore, for sufficiently large $C_1 > 0$ and $C_2 > 0$, the choice of d_1 and d_2 in the statement of the theorem ensures that

$$\mathbb{P}\biggl(N - \min_{j\in\hat{\mathcal{S}}}\Bigl\{t^j_{N,\tilde{b}^j} + \frac{d_2}{(\tilde{b}^j)^2}\Bigr\} > z\biggr) \leq \alpha/2.$$

Combining this with the fact that $\mathbb{P}(N \leq z) \leq z/(4\gamma) \leq \alpha/2$, which follows from Lemma 2.13 when $C_1 \geq \sqrt{8}$, we deduce the result. □

Proof of Theorem 3.2. Denote $\ell_0 := C_3\bigl(\frac{a^2 s\log_2(2p)}{\beta^2} + 1\bigr)$. Since the output of Algorithm 3.2 remains unchanged if we replace $(X^j_t : t \in \mathbb{N})$ by $(-X^j_t : t \in \mathbb{N})$ for any fixed j, we may assume without loss of generality that $\theta^1 \geq \theta^2 \geq \vartheta/\sqrt{s\log_2(2p)}$. For $j \in \{1,2\}$, we denote $b^j := \max\{b \in \mathcal{B}\cup\mathcal{B}_0 : b \leq \theta^j\}$. Since $\vartheta \geq \beta$ and $s \leq 2^{\lfloor\log_2(p)\rfloor}$, we have $b^1 \geq b^2 \geq \beta/\sqrt{s\log_2(2p)} \geq \sqrt{2}b_{\min}$. For $C_5 > 0$, let

$$r := \frac{C_5 a^2 s\log_2(2p)}{\beta^2} + 2, \quad u := \frac{\ell_0\beta^2}{80s\log_2(2p)} \quad \text{and} \quad \delta := \frac{a}{2\sqrt{r+\ell}}.$$

Now define the following events:

$$\begin{aligned}
\Omega_0 &:= \{z < N \leq z + r\}\\
\Omega_1 &:= \bigl\{t^j_{N,b} \leq N - z + ub^{-2} \text{ for all } j \in [p] \text{ and } b \in \mathcal{B}\cup\mathcal{B}_0\bigr\},\\
\Omega_2 &:= \biggl\{\biggl|\Lambda^{j',j}_{N,b} + \sum_{i=N+1}^{N+\ell} X^{j'}_i\biggr| < a\sqrt{\tau^j_{N,b} + \ell} \text{ for all } b \in \mathcal{B}\cup\mathcal{B}_0, j \in [p]\\
&\qquad\qquad \text{and all } j' \in [p]\setminus\{j\} \text{ with } \bigl|\theta^{j'}\bigr| \leq \delta\biggr\},\\
\Omega_3 &:= \bigl\{\tau^{\hat{\jmath}}_{N,\hat{b}} \leq N - z + \ell/20\bigr\}.
\end{aligned}$$

Finally, we denote event

$$\Omega_4 := \Omega_{4,1} \cup \Omega_{4,2},$$

with

$$\Omega_{4,1} := \bigl\{\hat{j} \neq 1, 1 \in \hat{\mathcal{S}}, \tilde{b}^1 \geq b^1/\sqrt{2}\bigr\}$$
$$\Omega_{4,2} := \bigl\{\hat{j} = 1, 2 \in \hat{\mathcal{S}}, \tilde{b}^2 \geq b^2/\sqrt{2}\bigr\}.$$

Henceforth, we will assume without loss of generality that $C_2 \geq 1/2$. Then, on the event $\bigcap_{k=0}^4 \Omega_k$, we have

$$\begin{aligned} L &= \min_{j \in \hat{\mathcal{S}}} \biggl\{ t^j_{N,\tilde{b}^j} + \frac{d_2}{(\tilde{b}^j)^2} \biggr\} \wedge N \leq N - z + \frac{2(u+d_2)}{(b^2)^2} \leq r + \frac{\ell_0}{40} + \frac{2sd_2 \log_2(2p)}{\beta^2} \\ &\leq C_1^2 \Bigl(C_5 + \frac{C_3}{40} + 8C_2^2 \Bigr) \biggl(\frac{s \log_2(2p) \log\{p\gamma(\beta^{-2} \vee 1)\alpha^{-1}\}}{\beta^2} + 1 \biggr). \end{aligned}$$

Let $C_4 := C_1^2(C_5 + \frac{C_3}{40} + 8C_2^2)$. Then

$$\mathbb{P}\biggl(L > C_4 \biggl(\frac{s \log_2(2p) \log\{p\gamma(\beta^{-2} \vee 1)\alpha^{-1}\}}{\beta^2} + 1 \biggr) \biggr) \leq \mathbb{P}\biggl(\bigcup_{k=0}^4 \Omega_k^{\mathrm{c}} \biggr). \tag{3.5}$$

By choosing $C_1 \geq \sqrt{8}$ and choosing C_5 to be a sufficiently large universal constant, we have $r \geq \bigl(\frac{24T^{\mathrm{off}}}{\vartheta^2} \vee \frac{12a^2 s}{\vartheta^2} \vee \frac{24T^{\mathrm{diag}} s}{\beta^2} \bigr) \log_2(2p) + 2$, so we may apply Proposition 3.7 and Lemma 2.13 to deduce that

$$\begin{aligned} \mathbb{P}(\Omega_0^{\mathrm{c}}) = \mathbb{P}(N > z + r) + \mathbb{P}(N \leq z) &\leq 2p \exp\biggl\{ -\frac{\beta^2(r-1)}{128 \log_2(2p)} \biggr\} + \frac{z}{4\gamma} \\ &\leq 2p \exp\biggl\{ -\frac{C_5 C_1^2 s \log\{p\gamma(\beta^{-2} \vee 1)\alpha^{-1}\}}{128} \biggr\} + \frac{z}{4\gamma}. \end{aligned} \tag{3.6}$$

On Ω_0, we have for any $j \in [p]$ and $b \in \mathcal{B} \cup \mathcal{B}_0$ that

$$\begin{aligned} t^j_{N,b} &= \operatorname*{sargmax}_{0 \leq h \leq N} \sum_{i=N-h+1}^{N} b(X_i^j - b/2) \leq \operatorname*{sargmax}_{N-z \leq h \leq N} \sum_{i=N-h+1}^{N} b(X_i^j - b/2) \\ &= N - z + \operatorname*{sargmax}_{0 \leq h \leq z} \sum_{i=z-h+1}^{z} b(X_i^j - b/2). \end{aligned}$$

Thus, by Lemma 3.6 (taking $\mu = -b/2$) and a union bound, we have

$$\mathbb{P}(\Omega_0 \cap \Omega_1^{\mathrm{c}}) \leq 2p \log_2(4p) e^{-u/8} = 2p \log_2(4p) \exp\biggl\{ -\frac{\ell_0 \beta^2}{640 s \log_2(2p)} \biggr\}. \tag{3.7}$$

Now observe that, for all $z < n \leq z + r, b \in \mathcal{B} \cup \mathcal{B}_0, j \in [p]$ and $j' \in [p] \setminus \{j\}$, we have

$$\Lambda^{j',j}_{n,b} + \sum_{i=n+1}^{n+\ell} X^{j'}_i \,\Big|\, \tau^j_{n.b} \sim \mathcal{N}\Big(\theta^{j'}\{\ell + \min\big(\tau^j_{n,b}, n - z\big)\}, \tau^j_{n,b} + \ell\Big).$$

Hence, when $\big|\theta^{j'}\big| \leq \delta$, we have that

$$\mathbb{P}\bigg(\bigg|\Lambda^{j',j}_{n,b} + \sum_{i=n+1}^{n+\ell} X^{j'}_i\bigg| \geq a\sqrt{\tau^j_{n,b} + \ell} \,\bigg|\, \tau^j_{n.b}\bigg) \leq \mathbb{P}(|Y_1| \geq a) \leq 2\mathbb{P}(Y_1 \geq a) \leq e^{-a^2/8},$$

where $Y_1 \sim \mathcal{N}(\delta\sqrt{n - z + \ell}, 1)$, and where the last inequality follows from the relation $a = 2\delta\sqrt{r + \ell}$. Thus, by a union bound, we have

$$\mathbb{P}(\Omega_0 \cap \Omega_2^{\mathrm{c}}) \leq 2rp^2 \log_2(4p) e^{-a^2/8} \leq 2rp^2 \log_2(4p)\{p\gamma(\beta^{-2} \vee 1)\alpha^{-1}\}^{-C_1^2/8}. \tag{3.8}$$

By Lemma 2.19, we have $t^j_{n,b}/2 \leq \tau^j_{n,b} \leq t^j_{n,b}$ for all $n \in \mathbb{N}_0$, $b \in \mathcal{B} \cup \mathcal{B}_0$ and $j \in [p]$. Moreover, when $C_3 \geq 80(C_5 \vee 2)$, we have $\ell_0 \geq 80r$. Define $b_* := \beta/\sqrt{s \log_2(2p)} \in \mathcal{B}$, so that $u = \ell_0 b_*^2/80$. We thus have for any $z < n \leq z + r, j \in [p]$ and $b \in \mathcal{B}$ that

$$\begin{aligned}
&\mathbb{P}\Big(\{N = n\} \cap \Omega_1 \cap \Omega_2 \cap \big\{\tilde{Q}^j_{n,b} \geq \tilde{Q}^j_{n,b_*}\big\} \cap \big\{\tau^j_{n,b} > n - z + \ell/20\big\} \,\Big|\, X^j_1, X^j_2, \dots\Big) \\
&\quad \leq \mathbb{P}\Bigg(\bigcup_{\substack{j' \in [p]\setminus\{j\}: \\ |\theta^{j'}| > \delta}} \{|\Xi^{j',j}_{n,b}| \geq |\Xi^{j',j}_{n,b_*}|\} \cap \{N = n\} \cap \Omega_1 \cap \Omega_2 \\
&\qquad\qquad\qquad\qquad\qquad\qquad \cap \big\{\tau^j_{n,b} > n - z + \ell/20\big\} \,\Bigg|\, X^j_1, X^j_2, \dots\Bigg) \\
&\quad \leq \sum_{\substack{j' \in [p]\setminus\{j\}: \\ |\theta^{j'}| > \delta}} \mathbb{P}\Big(\big\{|\Xi^{j',j}_{n,b}| \geq |\Xi^{j',j}_{n,b_*}|\big\} \cap \big\{n - z < \tau^j_{n,b_*} \leq n - z + \ell/80\big\} \\
&\qquad\qquad\qquad\qquad\qquad\qquad \cap \big\{\tau^j_{n,b} > n - z + \ell/20\big\} \,\Big|\, X^j_1, X^j_2, \dots\Big) \\
&\qquad + \sum_{\substack{j' \in [p]\setminus\{j\}: \\ |\theta^{j'}| > \delta}} \mathbb{P}\Big(\big\{|\Xi^{j',j}_{n,b}| \geq |\Xi^{j',j}_{n,b_*}|\big\} \cap \big\{\tau^j_{n,b_*} \leq n - z\big\} \cap \big\{\tau^j_{n,b} > n - z + \ell/20\big\} \,\Big|\, X^j_1, X^j_2, \dots\Big).
\end{aligned}$$

We apply Lemma 3.8(a) to each summand of the first term in the final expression above with $U = \sum_{i=n-\tau^j_{n,b_*}+1}^{z} X^{j'}_i$, $V = \sum_{i=n-\tau^j_{n,b}+1}^{n-\tau^j_{n,b_*}} X^{j'}_i$, $Y = \sum_{i=z+1}^{n+\ell} X^{j'}_i$, $\alpha = \theta^{j'}$, $\phi_1 = z - n + \tau^j_{n,b_*}$, $\phi_2 = z - n + \tau^j_{n,b}$, $\phi_3 = n - z + \ell$ and $\kappa = \ell/80$, and then apply Lemma 3.8(b) to each summand of the second term with $U = \sum_{i=n-\tau^j_{n,b}+1}^{z} X^{j'}_i$, $Y = \sum_{i=n-\tau^j_{n,b_*}+1}^{n+\ell} X^{j'}_i$, $Z = \sum_{i=z+1}^{n-\tau^j_{n,b_*}} X^{j'}_i$, $\alpha = \theta^{j'}$, $\phi_1 = z - n + \tau^j_{n,b}$, $\phi_3 = \ell + \tau^j_{n,b_*}$, $\phi_4 = n - z - \tau^j_{n,b_*}$ and $\kappa = \ell/80$. Then we have

$$\mathbb{P}\Big(\{N = n\} \cap \Omega_1 \cap \Omega_2 \cap \big\{\tilde{Q}^j_{n,b} \geq \tilde{Q}^j_{n,b_*}\big\} \cap \big\{\tau^j_{n,b} > n - z + \ell/20\big\} \,\Big|\, X^j_1, X^j_2, \dots\Big)$$

$$\le p\exp\biggl(-\frac{\ell\delta^2}{960}\biggr) = p\exp\biggl(-\frac{\ell a^2}{3840(r+\ell)}\biggr).$$

Observe that $\tilde{Q}^{\hat{j}}_{n,\hat{b}} \ge \tilde{Q}^{\hat{j}}_{n,b_*}$. Thus, by a union bound, we have

$$\begin{aligned}
&\mathbb{P}(\Omega_0 \cap \Omega_1 \cap \Omega_2 \cap \Omega_3^{\mathrm{c}}) \\
&\quad\le \sum_{j=1}^{p}\sum_{b\in\mathcal{B}}\sum_{n=z+1}^{z+\lfloor r\rfloor} \mathbb{P}\Bigl(\{N=n\} \cap \Omega_1 \cap \Omega_2 \cap \bigl\{\tilde{Q}^{j}_{n,b} \ge \tilde{Q}^{j}_{n,b_*}\bigr\} \cap \bigl\{\tau^{j}_{n,b} > n-z+\ell/20\bigr\}\Bigr) \\
&\quad\le 2rp^2\log_2(2p)\exp\biggl(-\frac{\ell a^2}{3840(r+\ell)}\biggr). \qquad (3.9)
\end{aligned}$$

When $C_3 \ge 144C_2^2 \vee 80C_5 \vee 160$, we have $\ell_0 \ge 80r$ and $d_1 \le \frac{\sqrt{\ell_0}\beta}{12\sqrt{s\log_2(2p)}} \le b^1\sqrt{\ell}/12$. Thus, on $\Omega_0 \cap \Omega_3$, we have

$$\tau^{\hat{j}}_{N,\hat{b}} \le N - z + \ell/20 \le r + \ell/20 \le \ell/16.$$

Hence, for any $z < n \le z + r$, $j \in [p] \setminus \{1\}$ and $b \in \mathcal{B}$, we have

$$\begin{aligned}
&\mathbb{P}\bigl(\Omega_3 \cap \{N=n, \hat{j}=j, \hat{b}=b\} \cap \Omega_{4,1}^{\mathrm{c}} \bigm| X_1^j, X_2^j, \ldots\bigr) \\
&\quad\le \mathbb{P}\biggl(\{\tau^{j}_{n,b} \le \ell/16\} \cap \Bigl\{\Xi^{1,j}_{n,b} - b^1\sqrt{\bigl(\tau^{j}_{n,b}+\ell\bigr)/2} < d_1\Bigr\} \biggm| X_1^j, X_2^j, \ldots\biggr) \le \frac{1}{2}e^{-d_1^2/2}.
\end{aligned}$$

Here, in the final bound, we have used the facts that $\Xi^{1,j}_{n,b} \mid \tau^{j}_{n,b} \sim \mathcal{N}\bigl(\theta^1 \min\{(n+\ell-z)(\tau^{j}_{n,b}+\ell)^{-1/2}, (\tau^{j}_{n,b}+\ell)^{1/2}\}, 1\bigr)$ and that

$$\begin{aligned}
&\theta^1 \min\bigl\{(n+\ell-z)(\tau^{j}_{n,b}+\ell)^{-1/2}, (\tau^{j}_{n,b}+\ell)^{1/2}\bigr\} - b^1\sqrt{\bigl(\tau^{j}_{n,b}+\ell\bigr)/2} \\
&\qquad\qquad \ge \frac{4\theta^1\sqrt{\ell}}{\sqrt{17}} - \frac{b^1\sqrt{17\ell}}{4\sqrt{2}} \ge \frac{b^1\sqrt{\ell}}{6} \ge 2d_1,
\end{aligned}$$

when $\tau^{j}_{n,b} \le \ell/16$, as well as the standard Gaussian tail bound used at the end of the proof of Lemma 3.6. By a similar argument, we also have for any $z < n \le z + r$ and $b \in \mathcal{B}$ that

$$\begin{aligned}
&\mathbb{P}\bigl(\Omega_3 \cap \{N=n, \hat{j}=1, \hat{b}=b\} \cap \Omega_{4,2}^{\mathrm{c}} \bigm| X_1^1, X_2^1, \ldots\bigr) \\
&\quad\le \mathbb{P}\biggl(\{\tau^{1}_{n,b} \le \ell/16\} \cap \Bigl\{\Xi^{2,1}_{n,b} - b^2\sqrt{\bigl(\tau^{1}_{n,b}+\ell\bigr)/2} < d_1\Bigr\} \biggm| X_1^1, X_2^1, \ldots\biggr) \le \frac{1}{2}e^{-d_1^2/2}.
\end{aligned}$$

Thus, by a union bound, we have

$$\begin{aligned}
\mathbb{P}(\Omega_0 \cap \Omega_3 \cap \Omega_4^{\mathrm{c}}) &= \mathbb{P}(\Omega_0 \cap \Omega_3 \cap \Omega_{4,1}^{\mathrm{c}} \cap \Omega_{4,2}^{\mathrm{c}}) \\
&\le \sum_{j=2}^{p}\sum_{b\in\mathcal{B}}\sum_{n=z+1}^{z+\lfloor r\rfloor} \mathbb{P}\bigl(\Omega_3 \cap \{N=n, \hat{j}=j, \hat{b}=b\} \cap \Omega_{4,1}^{\mathrm{c}}\bigr)
\end{aligned}$$

$$
\begin{aligned}
&+\sum_{b\in\mathcal{B}}\sum_{n=z+1}^{z+\lfloor r\rfloor}\mathbb{P}\bigl(\Omega_3\cap\{N=n,\hat{\jmath}=1,\hat{b}=b\}\cap\Omega_{4,2}^{\mathrm{c}}\bigr)\\
&\leq rp\log_2(2p)e^{-d_1^2/2}. \qquad (3.10)
\end{aligned}
$$

Hence by substituting (3.6), (3.7), (3.8), (3.9) and (3.10) into (3.5), we conclude that, by increasing the universal constant $C_1 > 0$ if necessary,

$$
\begin{aligned}
&\mathbb{P}\biggl(L > C_4\biggl(\frac{s\log_2(2p)\log\{p\gamma(\beta^{-2}\vee 1)\alpha^{-1}\}}{\beta^2}+1\biggr)\biggr)\\
&\quad\leq 2p\{p\gamma(\beta^{-2}\vee 1)\alpha^{-1}\}^{-C_1^2C_5 s/128}+\frac{z}{4\gamma}+2p\log_2(4p)\{p\gamma(\beta^{-2}\vee 1)\alpha^{-1}\}^{-C_1^2C_3/640}\\
&\qquad+2rp^2\log_2(4p)\{p\gamma(\beta^{-2}\vee 1)\alpha^{-1}\}^{-C_1^2/8}+2rp^2\log_2(2p)\{p\gamma(\beta^{-2}\vee 1)\alpha^{-1}\}^{-C_1^2/3888}\\
&\qquad+rp\log_2(2p)\{p\gamma(\beta^{-2}\vee 1)\alpha^{-1}\}^{-C_1^2C_2^2/2}\\
&\quad\leq\alpha,
\end{aligned}
$$

as required. □

Proof of Theorem 3.3. (a) For $j' \in \mathcal{S}_\beta^{\mathrm{c}}$, we have $|\theta^{j'}| < b_{\min}$, so the event $\{|\tilde{b}^{j'}| \leq |\theta^{j'}|\}$ is empty. Thus by (3.2), we have, for $n > z, j \in [p]$, $b \in \mathcal{B}$ and $j' \in \mathcal{S}_\beta^{\mathrm{c}}$, that

$$
\mathbb{P}_{z,\theta}\bigl(\{j'\in\hat{\mathcal{S}}\}\cap\{N=n,\hat{\jmath}=j,\hat{b}=b\}\bigr)\leq 2\bar{\Phi}(d_1).
$$

Hence, recalling the definition of r_0 from (3.4), by Lemma 2.13 applied with $C_1 \geq \sqrt{8}$, a union bound and Proposition 3.7, we have

$$
\begin{aligned}
\mathbb{P}_{z,\theta}(\hat{\mathcal{S}}\not\subseteq\mathcal{S}_\beta)&\leq\mathbb{P}_{z,\theta}(N\leq z)+\mathbb{P}_{z,\theta}(N>z+r_0)\\
&\qquad+\sum_{n=z+1}^{z+\lfloor r_0\rfloor}\sum_{j=1}^{p}\sum_{b\in\mathcal{B}}\sum_{j'\in\mathcal{S}_\beta^{\mathrm{c}}}\mathbb{P}_{z,\theta}\bigl(\{j'\in\hat{\mathcal{S}}\}\cap\{N=n,\hat{\jmath}=j,\hat{b}=b\}\bigr)\\
&\leq\frac{\alpha}{2}+\exp\biggl\{-\frac{\beta^2(r_0-1)}{48\log_2(2p)}\biggr\}+p\exp\biggl\{-\frac{\vartheta^2(r_0-1)}{128\log_2(2p)}\biggr\}+4p^2\log_2(2p)r_0\bar{\Phi}(d_1)\\
&\leq\alpha,
\end{aligned}
$$

where the final bound follows as in the proof of Theorem 3.1.

(b) We use the events $\Omega_0, \Omega_1, \Omega_2, \Omega_3$ defined in the proof of Theorem 3.2. Recall from the argument immediately below (3.9) that when $C_3 \geq 144C_2^2 \vee 80C_5 \vee 160$, we have $\tau_{N,\hat{b}}^{\hat{\jmath}} \leq \ell/16$ and $d_1 \leq \min_{j'\in\mathcal{S}}|\theta^{j'}|\sqrt{\ell}/12$ on $\Omega_0\cap\Omega_3$. Recall also the definition of $\Xi_{n,b}^{j',j}$ from Algorithm 3.2. Then, for any $z < n \leq z + r$, $j \in [p]$, $j' \in \mathcal{S}\setminus\{j\}$ and $b \in \mathcal{B}$, we have

$$
\mathbb{P}_{z,\theta}\bigl(\Omega_3\cap\{N=n,\hat{\jmath}=j,\hat{b}=b,j'\notin\hat{\mathcal{S}}\}\mid X_1^j,X_2^j,\ldots\bigr)
$$

$$\begin{aligned} &= \mathbb{P}_{z,\theta}\Bigl(\Omega_3 \cap \{N = n, \hat{j} = j, \hat{b} = b\} \cap \bigl\{|\Xi^{j',j}_{n,b}| < b_{\min}\sqrt{\tau^j_{n,b} + \ell} + d_1\bigr\} \mid X^j_1, X^j_2, \ldots\Bigr) \\ &\le \mathbb{P}_{z,\theta}\left(\{\tau^j_{n,b} \le \ell/16\} \cap \left\{|\Xi^{j',j}_{n,b}| - b_{\min}\sqrt{\tau^j_{n,b} + \ell} < d_1\right\} \,\middle|\, X^j_1, X^j_2, \ldots\right) \le \frac{1}{2}e^{-d_1^2/2}, \end{aligned} \tag{3.11}$$

where, in the final bound, we have used the facts that $\Xi^{j',j}_{n,b} \mid \tau^j_{n,b} \sim \mathcal{N}\bigl(\theta^{j'} \min\{(n+\ell-z)(\tau^j_{n,b}+\ell)^{-1/2}, (\tau^j_{n,b}+\ell)^{1/2}\}, 1\bigr)$ and that

$$\begin{aligned} |\theta^{j'}| \min\bigl\{(n+\ell-z)(\tau^j_{n,b}+\ell)^{-1/2}, (\tau^j_{n,b}+\ell)^{1/2}\bigr\} - b_{\min}\sqrt{\tau^j_{n,b}+\ell} \\ \ge \frac{4|\theta^{j'}|\sqrt{\ell}}{\sqrt{17}} - \frac{b_{\min}\sqrt{17\ell}}{4\sqrt{2}} \ge \frac{|\theta^{j'}|\sqrt{\ell}}{6} \ge 2d_1, \end{aligned}$$

when $\tau^j_{n,b} \le \ell/16$. Hence

$$\begin{aligned} &\mathbb{P}_{z,\theta}(\hat{\mathcal{S}} \cup \{\hat{j}\} \not\supseteq \mathcal{S}) \\ &\quad \le \mathbb{P}_{z,\theta}(\Omega_0^{\mathrm{c}}) + \mathbb{P}_{z,\theta}(\Omega_0 \cap \Omega_1^{\mathrm{c}}) + \mathbb{P}_{z,\theta}(\Omega_0 \cap \Omega_2^{\mathrm{c}}) + \mathbb{P}_{z,\theta}(\Omega_0 \cap \Omega_1 \cap \Omega_2 \cap \Omega_3^{\mathrm{c}}) \\ &\qquad + \sum_{n=z+1}^{z+\lfloor r \rfloor} \sum_{j=1}^{p} \sum_{b\in\mathcal{B}} \sum_{j'\in\mathcal{S}\setminus\{j\}} \mathbb{P}_{z,\theta}\bigl(\Omega_3 \cap \{N = n, \hat{j} = j, \hat{b} = b, j' \notin \hat{\mathcal{S}}\}\bigr) \\ &\quad \le 2p\{p\gamma(\beta^{-2} \vee 1)\alpha^{-1}\}^{-C_1^2 C_5 s/128} + \frac{z}{4\gamma} + 2p\log_2(4p)\{p\gamma(\beta^{-2} \vee 1)\alpha^{-1}\}^{-C_1^2 C_3/640} \\ &\qquad + 2rp^2\log_2(4p)\{p\gamma(\beta^{-2} \vee 1)\alpha^{-1}\}^{-C_1^2/8} + 2rp^2\log_2(2p)\{p\gamma(\beta^{-2} \vee 1)\alpha^{-1}\}^{-C_1^2/3888} \\ &\qquad + rp^2\log_2(2p)\{p\gamma(\beta^{-2} \vee 1)\alpha^{-1}\}^{-C_1^2 C_2^2/2} \le \alpha, \end{aligned}$$

where the penultimate inequality follows from (3.6), (3.7), (3.8), (3.9) and (3.11), and the last inequality follows by choosing the universal constant $C_1 > 0$ to be sufficiently large. □

Proof of Proposition 3.4. Fix $N \in \mathcal{T}_{r,m}$ and $\psi \in \mathcal{J}_N$. Denote

$$\tilde{\Theta} := \bigl\{\theta \in \mathbb{R}^p : \theta^j \in \{0, 1/(8\sqrt{r})\}, |\mathrm{supp}(\theta)| = m\bigr\},$$

and let $\tilde{\Theta}_{\mathrm{pa}} \subseteq \tilde{\Theta}$ be an $(m/4)$-packing set with respect to the symmetric difference metric defined above, i.e. for any $\theta, \tilde{\theta} \in \tilde{\Theta}_{\mathrm{pa}}$, we have $d\bigl(\mathrm{supp}(\theta), \mathrm{supp}(\tilde{\theta})\bigr) > m/4$. We also have $\mathrm{KL}\bigl(\mathbb{P}^{(z+r)}_{z,\theta}, P^{(z+r)}_{z,\tilde{\theta}}\bigr) = r\|\theta - \tilde{\theta}\|^2/2 \le m/64$.

Enumerate $\tilde{\Theta}_{\mathrm{pa}} = \bigl\{\theta_{(1)}, \theta_{(2)}, \ldots, \theta_{(|\tilde{\Theta}_{\mathrm{pa}}|)}\bigr\}$. Let $\phi^* := \mathrm{sargmin}_{\ell\in[|\tilde{\Theta}_{\mathrm{pa}}|]}\, d\bigl(\psi, \mathrm{supp}(\theta_{(\ell)})\bigr)$. Note that ϕ^* is also $\mathcal{F}_N$-measurable. Then for any $z \in \mathbb{N} \cup \{0\}$,

$$\sup_{\theta\in\Theta_{r,m}} \mathbb{E}_{z,\theta} d\bigl(\psi, \mathrm{supp}(\theta)\bigr) \ge \frac{m}{8|\tilde{\Theta}_{\mathrm{pa}}|} \sum_{\ell=1}^{|\tilde{\Theta}_{\mathrm{pa}}|} \mathbb{P}_{z,\theta_{(\ell)}}\Bigl(d\bigl(\psi, \mathrm{supp}(\theta_{(\ell)})\bigr) > \frac{m}{8}\Bigr)$$

$$
\begin{aligned}
&\geq \frac{m}{8|\tilde{\Theta}_{\mathrm{pa}}|}\sum_{\ell=1}^{|\tilde{\Theta}_{\mathrm{pa}}|}\mathbb{P}_{z,\theta_{(\ell)}}(\phi^* \neq \ell)\\
&= \frac{m}{8}\biggl\{1 - \frac{1}{|\tilde{\Theta}_{\mathrm{pa}}|}\sum_{\ell=1}^{|\tilde{\Theta}_{\mathrm{pa}}|}\mathbb{P}_{z,\theta_{(\ell)}}(\phi^* = \ell)\biggr\}\\
&\geq \frac{m}{8}\biggl\{\frac{3}{4} - \frac{1}{|\tilde{\Theta}_{\mathrm{pa}}|}\sum_{\ell=1}^{|\tilde{\Theta}_{\mathrm{pa}}|}\mathbb{P}_{z,\theta_{(\ell)}}(\phi^* = \ell, N \leq z+r)\biggr\}\\
&= \frac{m}{8}\biggl\{\frac{3}{4} - \frac{1}{|\tilde{\Theta}_{\mathrm{pa}}|}\sum_{\ell=1}^{|\tilde{\Theta}_{\mathrm{pa}}|}\mathbb{P}^{(z+r)}_{z,\theta_{(\ell)}}(\phi^* = \ell, N \leq z+r)\biggr\}. \qquad (3.12)
\end{aligned}
$$

Now set

$$
\tilde{\phi}^* := \begin{cases} \phi^* & \text{if } N \leq z+r \\ 1 & \text{if } N > z+r. \end{cases}
$$

Then $\tilde{\phi}^*$ is $\mathcal{F}_{z+r}$-measurable and by Fano's inequality (Yu, 1997, Lemma 3), we have

$$
\begin{aligned}
\frac{1}{|\tilde{\Theta}_{\mathrm{pa}}|}\sum_{\ell=1}^{|\tilde{\Theta}_{\mathrm{pa}}|}\mathbb{P}^{(z+r)}_{z,\theta_{(\ell)}}(\phi^* = \ell, N \leq z+r) &\leq \frac{1}{|\tilde{\Theta}_{\mathrm{pa}}|}\sum_{\ell=1}^{|\tilde{\Theta}_{\mathrm{pa}}|}\mathbb{P}^{(z+r)}_{z,\theta_{(\ell)}}(\tilde{\phi}^* = \ell)\\
&\leq \frac{\log 2 + |\tilde{\Theta}_{\mathrm{pa}}|^{-2}\sum_{j,k=1}^{|\tilde{\Theta}_{\mathrm{pa}}|}\mathrm{KL}\bigl(\mathbb{P}^{(z+r)}_{z,\theta_{(j)}}, P^{(z+r)}_{z,\theta_{(k)}}\bigr)}{\log|\tilde{\Theta}_{\mathrm{pa}}|}\\
&\leq \frac{\log 2 + m/64}{\log|\tilde{\Theta}_{\mathrm{pa}}|}. \qquad (3.13)
\end{aligned}
$$

By Massart (2007, Lemma 4.7), there exists an $(m/4)$-packing set with

$$
\log|\tilde{\Theta}_{\mathrm{pa}}| \geq m/8. \qquad (3.14)
$$

Combining (3.12), (3.13) and (3.14), we conclude that

$$
\begin{aligned}
\sup_{z\in\mathbb{N}\cup\{0\},\theta\in\Theta_{r,m}}\mathbb{E}_{z,\theta}\, d\bigl(\psi, \mathrm{supp}(\theta)\bigr) &\geq \frac{m}{8}\biggl(\frac{3}{4} - \frac{\log 2 + m/64}{m/8}\biggr)\\
&\geq \frac{m}{8}\biggl(\frac{3}{4} - \frac{8\log 2}{m} - \frac{1}{8}\biggr) \geq \frac{m}{32},
\end{aligned}
$$

where we have used the assumption that $m \geq 15$ in the final inequality. □

3.6 Auxiliary results

Proposition 3.5. *Let $X_1, X_2, \ldots$ be independent random variables with $X_1, \ldots, X_z \overset{\mathrm{iid}}{\sim} \mathcal{N}(0,1)$ and $X_{z+1}, X_{z+2}, \ldots \overset{\mathrm{iid}}{\sim} \mathcal{N}(\theta, 1)$. Assume that $0 < b \leq \theta$ and let $t_{n,b}$ be defined as in* (3.1) *for*

$n \in \mathbb{N}$. Then for any $\alpha \in (0,1)$, and any stopping time N satisfying $\mathbb{P}(N < z) \leq \alpha/2$, we have that the confidence interval

$$\mathcal{C}_0 := \biggl[N - t_{N,b} - \frac{4\{\Phi^{-1}(1-\alpha/4)\}^2}{b^2}, \, N\biggr]$$

satisfies $\mathbb{P}(z \in \mathcal{C}_0) \geq 1 - \alpha$.

Remark. *We could also replace $4\{\Phi^{-1}(1-\alpha/4)\}^2/b^2$ by $8\log(2/\alpha)/b^2$ in the confidence interval construction, if we apply the final bound from Lemma 3.6 to* (3.15)*.*

Proof. For $n \in \mathbb{N}$, define $R_{n,b} := \max\{R_{n-1,b} + b(X_n - b/2), 0\}$, with $R_{0,b} = 0$. By Lemma 2.11, we have $t_{N,b} = \min\{i : 0 \leq i \leq N, R_{N-i,b} = 0\} = \mathrm{sargmax}_{0 \leq h \leq N} \sum_{i=N-h+1}^{N} b(X_i - b/2)$. Let $U_{n,b} := \sum_{i=z+1}^{z+n}(X_i - b/2)$ for $n \in \mathbb{N}$, with $U_{0,b} := 0$. Then $R_{n+z,b} \geq bU_{n,b}$ for all $n \in \mathbb{N}$. Hence, for $y \in [0,\infty)$, we have

$$\mathbb{P}(N - t_{N,b} - y \geq z) \leq \mathbb{P}\Bigl(\inf_{n \in \mathbb{N}_0 : n \geq z+y} R_{n,b} = 0\Bigr) \leq \mathbb{P}\Bigl(\inf_{n \in \mathbb{N}_0 : n \geq y} U_{n,b} \leq 0\Bigr) \leq 2\bar{\Phi}\bigl(\sqrt{y}(\theta - b/2)\bigr), \tag{3.15}$$

where the last inequality follows from Lemma 3.6. Thus, if we choose $y = 4\{\Phi^{-1}(1-\alpha/4)\}^2/b^2$, then we are guaranteed that $\mathbb{P}(N - t_{N,b} - y > z) \leq \alpha/2$. Combining this with the assumption that $\mathbb{P}(N < z) \leq \alpha/2$, the desired result follows. □

Lemma 3.6. *Let $Y_1, Y_2, \ldots \overset{\mathrm{iid}}{\sim} \mathcal{N}(\mu, 1)$. Define $U_n := \sum_{i=1}^n Y_i$ for $n \in \mathbb{N}_0$, and let $\xi :=$ $\mathrm{sargmin}_{n \in \mathbb{N}_0} \mu U_n$. Then, for $y \in [0,\infty)$, we have*

$$\mathbb{P}(\xi \geq y) \leq \mathbb{P}\Bigl(\inf_{n \in \mathbb{N}_0 : n \geq y} \mu U_n \leq 0\Bigr) \leq 2\bar{\Phi}\bigl(\sqrt{y}|\mu|\bigr) \leq e^{-y\mu^2/2}.$$

Proof. The first inequality holds since $\mu U_\xi \leq \mu U_0 = 0$. For the second and third inequalities, we may assume without loss of generality that $\mu > 0$, since the result is clear when $\mu = 0$, and if $\mu < 0$ then the result will follow from the corresponding result with $\mu > 0$ by setting $Y_i' := -Y_i$ for $i \in \mathbb{N}$. Note that $(U_n - n\mu)_{n \in \mathbb{N}_0}$ is a standard Gaussian random walk starting at 0. Let $(B_t)_{t \in [0,\infty)}$ denote a standard Brownian motion starting at 0. Then, we have for any $y \in \mathbb{N}_0$ and $u > 0$ that

$$\mathbb{P}\biggl(\inf_{n \in \mathbb{N}_0 : n \geq y} U_n \leq 0 \biggm| U_y = u\biggr) \leq \mathbb{P}\biggl\{\inf_{t \in [y,\infty)} (B_t + t\mu) \leq 0 \biggm| B_y = u\biggr\} \leq e^{-2u\mu}, \tag{3.16}$$

where the final inequality follows from Siegmund (1986, Proposition 2.4 and Equation (2.5)). Thus, for $y \in [0,\infty)$, we have

$$\mathbb{P}\biggl(\inf_{n \in \mathbb{N}_0 : n \geq y} U_n \leq 0\biggr) = \mathbb{P}\bigl(U_{\lceil y \rceil} \leq 0\bigr) + \mathbb{E}\biggl\{\mathbb{P}\biggl(\inf_{n \in \mathbb{N}_0 : n \geq \lceil y \rceil} U_n \leq 0 \biggm| U_{\lceil y \rceil}\biggr)\mathbb{1}_{\{U_{\lceil y \rceil} > 0\}}\biggr\}$$

$$\begin{aligned}&\leq \bar{\Phi}\Big(\sqrt{\lceil y\rceil}\mu\Big) + \int_0^\infty \frac{1}{\sqrt{2\pi\lceil y\rceil}} e^{-\frac{(u-\lceil y\rceil\mu)^2}{2\lceil y\rceil}} e^{-2u\mu}\,du\\ &= 2\bar{\Phi}\Big(\sqrt{\lceil y\rceil}\mu\Big) \leq 2\bar{\Phi}\big(\sqrt{y}\mu\big) \leq e^{-y\mu^2/2}.\end{aligned}$$

where the first inequality follows from (3.16) and the fact that $U_{\lceil y\rceil} \sim \mathcal{N}(\lceil y\rceil\mu, \lceil y\rceil)$ and the last inequality follows from the standard normal distribution tail bound $\bar{\Phi}(x) \leq e^{-x^2/2}/2$ for $x \geq 0$. □

In the proposition below, we assume the Gaussian data generating mechanism given at the beginning of Section 3.3.

Proposition 3.7. *Assume that θ has an effective sparsity of $s := s(\theta) \geq 2$. Then, the right endpoint N of the interval output from the* `ocd_CI′` *algorithm, with inputs $(X_t)_{t\in\mathbb{N}}$, $0 < \beta \leq \vartheta$, $a > 0$, $T^{\mathrm{diag}} = \log\{16p\gamma\log_2(4p)\}$ and $T^{\mathrm{off}} = 8\log\{16p\gamma\log_2(2p)\}$, satisfies*

$$\mathbb{P}_{z,\theta}\big(N > z + r\big) \leq \exp\bigg\{-\frac{\beta^2(r-1)}{48\log_2(2p)}\bigg\} + p\exp\bigg\{-\frac{\vartheta^2(r-1)}{128\log_2(2p)}\bigg\},$$

for all $r \geq \Big\{\frac{24T^{\mathrm{off}}\log_2(2p)}{\vartheta^2} \vee \frac{12(a^2\vee 8\log 2)s\log_2(2p)}{\vartheta^2} \vee \frac{24T^{\mathrm{diag}}s\log_2(2p)}{\beta^2}\Big\} + 2$.

Proof. For $\theta \in \mathbb{R}^p$ with effective sparsity $s(\theta)$, there is at most one coordinate in θ of magnitude larger than $\vartheta/\sqrt{2}$, so there exists $b_* \in \big\{\beta/\sqrt{s(\theta)\log_2(2p)}, -\beta/\sqrt{s(\theta)\log_2(2p)}\big\} \subseteq \mathcal{B}$ such that

$$\mathcal{J} := \Big\{j \in [p] : \theta^j/b_* \geq 1 \text{ and } |\theta^j| \leq \vartheta/\sqrt{2}\Big\}$$

has cardinality at least $s(\theta)/2$. Note that the condition $\theta^j/b_* \geq 1$ above ensures that $\{\theta^j : j \in \mathcal{J}\}$ all have the same sign as b_*. By Proposition 2.8, we have on the event $\{N > z\}$ that

$$q(X_1,\ldots,X_z,\theta) := \max\big\{t^j_{z,b_*} : j \in \mathcal{J}\big\} \leq \frac{8T^{\mathrm{diag}}s\log_2(2p)}{\beta^2}. \tag{3.17}$$

We now fix

$$r \geq \bigg\{\frac{24T^{\mathrm{off}}\log_2(2p)}{\vartheta^2} \vee \frac{12(a^2\vee 8\log 2)s\log_2(2p)}{\vartheta^2} \vee \frac{24T^{\mathrm{diag}}s\log_2(2p)}{\beta^2}\bigg\} + 2 =: r_0. \tag{3.18}$$

For $j \in \mathcal{J}$, define the event

$$\Omega^j_r := \big\{t^j_{z+\lfloor r\rfloor,b_*} > 2\lfloor r\rfloor/3\big\}.$$

By applying Lemma 2.11 to $t^j_{z+\lfloor r\rfloor,b_*}$, we have for $j \in \mathcal{J}$ that

$$t^j_{z+\lfloor r\rfloor,b_*} = \operatorname*{sargmax}_{0\leq h\leq z+\lfloor r\rfloor} \sum_{i=z+\lfloor r\rfloor-h+1}^{z+\lfloor r\rfloor} b_*(X^j_i - b_*/2) \geq \operatorname*{sargmax}_{0\leq h\leq \lfloor r\rfloor} \sum_{i=z+\lfloor r\rfloor-h+1}^{z+\lfloor r\rfloor} b_*(X^j_i - b_*/2)$$

$$= \operatorname*{sargmax}_{0\le h\le\lfloor r\rfloor} \sum_{i=z+1}^{z+\lfloor r\rfloor-h} -b_*(X_i^j - b_*/2) = \lfloor r\rfloor - \operatorname*{largmax}_{0\le h\le\lfloor r\rfloor} \sum_{i=z+1}^{z+h} -b_*(X_i^j - b_*/2).$$

Recall that $X_{z+1}, X_{z+2}, \ldots \overset{\mathrm{iid}}{\sim} \mathcal{N}_p(\theta, I_p)$. Hence, by applying Lemma 2.16(b) with $a = 0, b = |b_*|/2$ and $c = \lfloor r\rfloor/3$, we have

$$\begin{aligned}
\mathbb{P}_{z,\theta}\biggl(\bigcap_{j\in\mathcal{J}} (\Omega_r^j)^{\mathrm{c}}\biggr) &= \prod_{j\in\mathcal{J}} \mathbb{P}_{z,\theta}\biggl(t^j_{z+\lfloor r\rfloor,b_*} \le \frac{2\lfloor r\rfloor}{3}\biggr) \\
&\le \prod_{j\in\mathcal{J}} \mathbb{P}_{z,\theta}\biggl(\operatorname*{largmax}_{0\le h\le\lfloor r\rfloor} \sum_{i=z+1}^{z+h} -b_*(X_i^j - b_*/2) \ge \frac{\lfloor r\rfloor}{3}\biggr) \\
&\le \prod_{j\in\mathcal{J}} \mathbb{P}_{z,\theta}\biggl(\sup_{h\ge\lfloor r\rfloor/3} \sum_{i=z+1}^{z+h} -\mathrm{sgn}(b_*)(X_i^j - b_*/2) \ge 0\biggr) \\
&\le \exp\bigl(-|\mathcal{J}|b_*^2\lfloor r\rfloor/24\bigr) \le \exp\bigl(-sb_*^2\lfloor r\rfloor/48\bigr). \qquad (3.19)
\end{aligned}$$

We now work on the event Ω_r^j, for some fixed $j \in \mathcal{J}$. We note that (3.18) guarantees that $r \ge 2$, and thus $t^j_{z+\lfloor r\rfloor,b_*} \ge \lceil 2\lfloor r\rfloor/3\rceil \ge 2$. Then, by (3.17) and (3.18), we have $r_0 > 3t^j_{z,b_*}$, and hence by Lemma 2.19,

$$\frac{\lfloor r\rfloor}{3} < \frac{t^j_{z+\lfloor r\rfloor,b_*}}{2} \le \tau^j_{z+\lfloor r\rfloor,b_*} \le \frac{3t^j_{z+\lfloor r\rfloor,b_*}}{4} \le \frac{3\bigl(t^j_{z,b_*} + r\bigr)}{4} < r.$$

We conclude that

$$2/3 \le \lfloor r\rfloor/3 < \tau^j_{z+\lfloor r\rfloor,b_*} \le \lfloor r\rfloor. \qquad (3.20)$$

Recall that $\Lambda^{\cdot,j}_{z+\lfloor r\rfloor,b_*} \in \mathbb{R}^p$ records the tail CUSUM statistics with tail length $\tau^j_{z+\lfloor r\rfloor,b_*}$. We observe by (3.20) that only post-change observations are included in $\Lambda^{\cdot,j}_{z+\lfloor r\rfloor,b_*}$. Hence we have that

$$\Lambda^{k,j}_{z+\lfloor r\rfloor,b_*} \mid \tau^j_{z+\lfloor r\rfloor,b_*} \overset{\mathrm{ind}}{\sim} \mathcal{N}\bigl(\theta^k\tau^j_{z+\lfloor r\rfloor,b_*}, \tau^j_{z+\lfloor r\rfloor,b_*}\bigr) \qquad (3.21)$$

for $k \in [p] \setminus \{j\}$. By the definition of the effective sparsity of θ, the set

$$\mathcal{L}^j := \biggl\{j' \in [p] : |\theta^{j'}| \ge \frac{\vartheta}{\sqrt{s\log_2(2p)}} \text{ and } j' \ne j\biggr\}$$

has cardinality at least $s - 1$. Hence, by (3.20), for all $k \in \mathcal{L}^j$,

$$|\theta^k|\sqrt{\tau^j_{z+\lfloor r\rfloor,b_*}} > \sqrt{\frac{\vartheta^2\lfloor r\rfloor}{3s\log_2(2p)}} =: \tilde{a}_r.$$

We then observe, from (3.18), that

$$\tilde{a}_r > 2\big(a \vee \sqrt{8\log 2}\big). \tag{3.22}$$

Hence, from (3.21), we have for all $k \in \mathcal{L}^j$ that

$$\mathbb{P}_{z,\theta}\biggl(\Omega^j_r \cap \biggl\{|\Lambda^{k,j}_{z+\lfloor r\rfloor,b_*}| < \frac{1}{2}\tilde{a}_r\sqrt{\tau^j_{z+\lfloor r\rfloor,b_*}}\biggr\}\,\biggm|\,\tau^j_{z+\lfloor r\rfloor,b_*}\biggr) \le \frac{1}{2}e^{-\tilde{a}_r^2/8} =: q_r.$$

We denote

$$U^j := \biggl|\biggl\{k \in \mathcal{L}^j : |\Lambda^{k,j}_{z+\lfloor r\rfloor,b_*}| < \frac{1}{2}\tilde{a}_r\sqrt{\tau^j_{z+\lfloor r\rfloor,b_*}}\biggr\}\biggr|.$$

Then, by the Chernoff–Hoeffding binomial tail bound (Hoeffding, 1963, Equation (2.1)), we have

$$\begin{aligned}\mathbb{P}_{z,\theta}\Bigl(\Omega^j_r \cap \bigl\{U^j \ge |\mathcal{L}^j|/2\bigr\}\,\bigm|\,\tau^j_{z+\lfloor r\rfloor,b_*}\Bigr) &\le \exp\biggl\{-\frac{|\mathcal{L}^j|}{2}\log\biggl(\frac{1}{4q_r(1-q_r)}\biggr)\biggr\}\\ &\le \exp\biggl\{-\frac{|\mathcal{L}^j|}{2}\biggl(\frac{\tilde{a}_r^2}{8}-\log 2\biggr)\biggr\} \le \exp\biggl(-\frac{3|\mathcal{L}^j|\tilde{a}_r^2}{64}\biggr) \le \exp\biggl\{-\frac{\vartheta^2\lfloor r\rfloor}{128\log_2(2p)}\biggr\},\end{aligned} \tag{3.23}$$

where the penultimate inequality follows from (3.22). Now, on the event $\Omega^j_r \cap \bigl\{U^j < |\mathcal{L}^j|/2\bigr\}$, we have

$$\begin{aligned}&\sum_{j'\in[p]:j'\neq j}\frac{\bigl(\Lambda^{j',j}_{z+\lfloor r\rfloor,b_*}\bigr)^2}{\tau^j_{z+\lfloor r\rfloor,b_*}\vee 1}\mathbb{1}_{\bigl\{|\Lambda^{j',j}_{z+\lfloor r\rfloor,b_*}|\ge a\sqrt{\tau^j_{z+\lfloor r\rfloor,b_*}}\bigr\}}\\ &\quad\ge \sum_{j'\in[p]:j'\neq j}\frac{\bigl(\Lambda^{j',j}_{z+\lfloor r\rfloor,b_*}\bigr)^2}{\tau^j_{z+\lfloor r\rfloor,b_*}\vee 1}\mathbb{1}_{\bigl\{|\Lambda^{j',j}_{z+\lfloor r\rfloor,b_*}|\ge \frac{1}{2}\tilde{a}_r\sqrt{\tau^j_{z+\lfloor r\rfloor,b_*}}\bigr\}} \ge \frac{\tilde{a}_r^2}{4}\biggl\{|\mathcal{L}^j| - \biggl(\biggl\lceil\frac{|\mathcal{L}^j|}{2}\biggr\rceil - 1\biggr)\biggr\}\\ &\quad= \frac{\tilde{a}_r^2}{4}\biggl\lceil\frac{|\mathcal{L}^j|+1}{2}\biggr\rceil \ge \frac{\vartheta^2\lfloor r\rfloor}{24\log_2(2p)} \ge T^{\mathrm{off}},\end{aligned} \tag{3.24}$$

where the penultimate inequality uses the fact that $|\mathcal{L}^j| \ge s-1$ and the last inequality follows from (3.18). We now denote

$$\tilde{E}^j_r := \biggl\{\sum_{j'\in[p]:j'\neq j}\frac{\bigl(\Lambda^{j',j}_{z+\lfloor r\rfloor,b_*}\bigr)^2}{\tau^j_{z+\lfloor r\rfloor,b_*}\vee 1}\mathbb{1}_{\bigl\{|\Lambda^{j',j}_{z+\lfloor r\rfloor,b_*}|\ge a\sqrt{\tau^j_{z+\lfloor r\rfloor,b_*}}\bigr\}} < T^{\mathrm{off}}\biggr\}.$$

Combining (3.19), (3.23) and (3.24), we deduce that

$$\begin{aligned}\mathbb{P}_{z,\theta}\bigl(N > z+r\bigr) \le \mathbb{P}_{z,\theta}\bigl(N > z+\lfloor r\rfloor\bigr) &\le \mathbb{P}_{z,\theta}\biggl(\bigcap_{j\in\mathcal{J}}(\Omega^j_r)^{\mathrm{c}}\biggr) + \sum_{j\in\mathcal{J}}\mathbb{P}_{z,\theta}\bigl(\tilde{E}^j_r \cap \Omega^j_r\bigr)\\ &\le \mathbb{P}_{z,\theta}\biggl(\bigcap_{j\in\mathcal{J}}(\Omega^j_r)^{\mathrm{c}}\biggr) + \sum_{j\in\mathcal{J}}\mathbb{P}_{z,\theta}\bigl(\Omega^j_r \cap \bigl\{U^j \ge |\mathcal{L}^j|/2\bigr\}\bigr)\end{aligned}$$

$$\leq \exp\biggl\{-\frac{sb_*^2(r-1)}{48}\biggr\} + p\exp\biggl\{-\frac{\vartheta^2(r-1)}{128\log_2(2p)}\biggr\},$$

as desired. □

Lemma 3.8. *Let* $U \sim \mathcal{N}(0,\phi_1)$*,* $V \sim \mathcal{N}(0,\phi_2-\phi_1)$*,* $Y \sim \mathcal{N}(\alpha\phi_3,\phi_3)$ *and* $Z \sim \mathcal{N}(\alpha\phi_4,\phi_4)$ *be independent random variables.*

(a) Assume that $\min\{\phi_2,\phi_3\}/4 \geq \kappa \geq \phi_1 \geq 0$ *for some* $\kappa > 0$*. Then*

$$\mathbb{P}\biggl(\frac{|U+V+Y|}{\sqrt{\phi_2+\phi_3}} \geq \frac{|U+Y|}{\sqrt{\phi_1+\phi_3}}\biggr) \leq \exp\biggl(-\frac{\kappa\alpha^2}{6}\biggr).$$

(b) Assume that $\min\{\phi_1,\phi_3\}/4 \geq \kappa \geq \phi_4 \geq 0$ *for some* $\kappa > 0$*. Then*

$$\mathbb{P}\biggl(\frac{|U+Y+Z|}{\sqrt{\phi_1+\phi_3+\phi_4}} \geq \frac{|Y|}{\sqrt{\phi_3}}\biggr) \leq \exp\biggl(-\frac{\kappa\alpha^2}{12}\biggr).$$

Proof. The case $\alpha = 0$ is trivial in both cases, so without loss of generality, we may assume $\alpha > 0$ throughout the rest of the proof.
(a) Let

$$W_1 := \bigl(\sqrt{\phi_2+\phi_3} - \sqrt{\phi_1+\phi_3}\bigr)(U+Y) - \sqrt{\phi_1+\phi_3}\,V,$$

so that

$$W_1 \sim \mathcal{N}\Bigl(\alpha\phi_3\bigl(\sqrt{\phi_2+\phi_3}-\sqrt{\phi_1+\phi_3}\bigr), \bigl\{\bigl(\sqrt{\phi_2+\phi_3}-\sqrt{\phi_1+\phi_3}\bigr)^2 + \phi_2 - \phi_1\bigr\}(\phi_1+\phi_3)\Bigr).$$

Hence, by the standard Gaussian tail bound used at the end of the proof of Lemma 3.6, we have

$$\mathbb{P}(W_1 \leq 0) \leq \frac{1}{2}e^{-\alpha^2/(2w_1)}, \tag{3.25}$$

where $w_1 := \frac{\phi_1+\phi_3}{\phi_3^2}\bigl(1 + \frac{\phi_2-\phi_1}{(\sqrt{\phi_2+\phi_3}-\sqrt{\phi_1+\phi_3})^2}\bigr)$. Then

$$\begin{aligned} w_1 &= \frac{\phi_1+\phi_3}{\phi_3^2}\left(1 + \frac{\bigl(\sqrt{\phi_2+\phi_3}+\sqrt{\phi_1+\phi_3}\bigr)^2}{\phi_2-\phi_1}\right) \\ &\leq \frac{5}{16\kappa}\left(1 + \frac{\bigl(\sqrt{8\kappa}+\sqrt{5\kappa}\bigr)^2}{3\kappa}\right) \leq \frac{3}{\kappa}, \end{aligned} \tag{3.26}$$

where the first inequality holds because w_1 is increasing in ϕ_1 and decreasing in both ϕ_2 and ϕ_3. Hence, using the fact that $-(U+V+Y) \leq_{\mathrm{st}} U+V+Y$, as well as (3.25) and (3.26), we have

$$\mathbb{P}\biggl(\frac{|U+V+Y|}{\sqrt{\phi_2+\phi_3}} \geq \frac{|U+Y|}{\sqrt{\phi_1+\phi_3}}\biggr) \leq \mathbb{P}\biggl(\biggl\{\frac{U+Y}{\sqrt{\phi_1+\phi_3}} \leq \frac{U+V+Y}{\sqrt{\phi_2+\phi_3}}\biggr\} \cap \{U+V+Y \geq 0\}\biggr)$$

$$
\begin{aligned}
&+\mathbb{P}\biggl(\biggl\{\frac{U+Y}{\sqrt{\phi_1+\phi_3}} \le -\frac{U+V+Y}{\sqrt{\phi_2+\phi_3}}\biggr\} \cap \{U+V+Y<0\}\biggr) \\
&\le 2\mathbb{P}\biggl(\frac{U+Y}{\sqrt{\phi_1+\phi_3}} \le \frac{U+V+Y}{\sqrt{\phi_2+\phi_3}}\biggr) \\
&= 2\mathbb{P}(W_1 \le 0) \le \exp\biggl(-\frac{\kappa\alpha^2}{6}\biggr),
\end{aligned}
$$

as required.

(b) Let

$$W_2 := \bigl(\sqrt{\phi_1+\phi_3+\phi_4} - \sqrt{\phi_3}\bigr)Y - \sqrt{\phi_3}(U+Z),$$

so that

$$W_2 \sim \mathcal{N}\Bigl(\alpha\phi_3\sqrt{\phi_1+\phi_3+\phi_4} - \alpha(\phi_3+\phi_4)\sqrt{\phi_3}, \bigl\{\bigl(\sqrt{\phi_1+\phi_3+\phi_4} - \sqrt{\phi_3}\bigr)^2 + \phi_1 + \phi_4\bigr\}\phi_3\Bigr).$$

Note that the assumption guarantees that $\mathbb{E}(W_2) > 0$. Hence, by the standard Gaussian tail bound used at the end of the proof of Lemma 3.6, we have

$$\mathbb{P}(W_2 \le 0) \le \frac{1}{2}e^{-\alpha^2/(2w_2)}, \tag{3.27}$$

where

$$w_2 := \frac{(\sqrt{\phi_1+\phi_3+\phi_4} - \sqrt{\phi_3})^2 + \phi_1 + \phi_4}{\bigl(\sqrt{\phi_3(\phi_1+\phi_3+\phi_4)} - \phi_3 - \phi_4\bigr)^2}.$$

Calculating the partial derivatives of w_2 with respect to ϕ_1, ϕ_3 and ϕ_4 and simplifying the expressions, we have

$$
\begin{aligned}
\frac{\partial w_2}{\partial \phi_1} &= \frac{(\phi_3+\phi_4)\sqrt{\phi_3} - (\phi_3+2\phi_4)\sqrt{\phi_1+\phi_3+\phi_4}}{\sqrt{\phi_1+\phi_3+\phi_4}\bigl(\sqrt{\phi_3(\phi_1+\phi_3+\phi_4)} - \phi_3 - \phi_4\bigr)^3} \le 0, \\
\frac{\partial w_2}{\partial \phi_3} &= \frac{-\bigl(\sqrt{\phi_1+\phi_3+\phi_4} - \sqrt{\phi_3}\bigr)^2\bigl[3\phi_1 + \phi_4 + \bigl(\sqrt{\phi_1+\phi_3+\phi_4} - \sqrt{\phi_3}\bigr)^2\bigr]}{2\sqrt{\phi_3(\phi_1+\phi_3+\phi_4)}\bigl(\sqrt{\phi_3(\phi_1+\phi_3+\phi_4)} - \phi_3 - \phi_4\bigr)^3} \le 0, \\
\frac{\partial w_2}{\partial \phi_4} &= \frac{2\phi_1\bigl(2\sqrt{\phi_1+\phi_3+\phi_4} - \sqrt{\phi_3}\bigr)^2 + 3(\phi_3+\phi_4)\bigl(\sqrt{\phi_1+\phi_3+\phi_4} - \sqrt{\phi_3}\bigr)^2}{2(\phi_1+\phi_3+\phi_4)\bigl(\sqrt{\phi_3(\phi_1+\phi_3+\phi_4)} - \phi_3 - \phi_4\bigr)^3} \\
&\qquad + \frac{(\phi_1+\phi_4)(\phi_3+\phi_4)}{2(\phi_1+\phi_3+\phi_4)\bigl(\sqrt{\phi_3(\phi_1+\phi_3+\phi_4)} - \phi_3 - \phi_4\bigr)^3} \ge 0.
\end{aligned}
$$

Thus w_2 is increasing in ϕ_4 and decreasing in both ϕ_1 and ϕ_3 and hence

$$w_2 \le \frac{6}{\kappa}. \tag{3.28}$$

Hence, using the fact that $-(U+Y+Z) \leq_{\mathrm{st}} U+Y+Z$, as well as (3.27) and (3.28), we have

$$\begin{aligned}
\mathbb{P}\biggl(\frac{|U+Y+Z|}{\sqrt{\phi_1+\phi_3+\phi_4}} \geq \frac{|Y|}{\sqrt{\phi_3}}\biggr) &\leq \mathbb{P}\biggl(\biggl\{\frac{Y}{\sqrt{\phi_3}} \leq \frac{U+Y+Z}{\sqrt{\phi_1+\phi_3+\phi_4}}\biggr\} \cap \{U+Y+Z \geq 0\}\biggr) \\
&\quad + \mathbb{P}\biggl(\biggl\{\frac{Y}{\sqrt{\phi_3}} \leq -\frac{U+Y+Z}{\sqrt{\phi_1+\phi_3+\phi_4}}\biggr\} \cap \{U+Y+Z < 0\}\biggr) \\
&\leq 2\mathbb{P}\biggl(\frac{Y}{\sqrt{\phi_3}} \leq \frac{U+Y+Z}{\sqrt{\phi_1+\phi_3+\phi_4}}\biggr) \\
&= 2\mathbb{P}(W_2 \leq 0) \leq \exp\biggl(-\frac{\kappa\alpha^2}{12}\biggr),
\end{aligned}$$

as required. □

Bibliography

Antoch, J., Hušková, M. and Veraverbeke, N. (1995) Change-point problem and bootstrap. *J. Nonparametr. Stat.*, **5**, 123–144.

Aston, J. A. D. and Kirch, C. (2012) Evaluating stationarity via change-point alternatives with applications to fMRI data. *Ann. Appl. Stat.*, **6**, 1906–1948.

Aston, J. A. D. and Kirch, C. (2018) High dimensional efficiency with applications to change point tests. *Electron. J. Stat.*, **12**, 1901–1947.

Aue, A. and Horváth, L. (2004) Delay time in sequential detection of change. *Stat. Probab. Lett.*, **67**, 221–231.

Aue, A., Horváth, L., Hušková, M. and Kokoszka, P. (2006) Change-point monitoring in linear models. *Econom. J.*, **9**, 373–403.

Avanesov, V. and Buzun, N. (2018) Change-point detection in high-dimensional covariance structure. *Electron. J. Stat.*, **12**, 3254–3294.

Baranowski, R., Chen, Y. and Fryzlewicz, P. (2019) Narrowest-Over-Threshold detection of multiple change points and change-point-like Features. *J. Roy. Statist. Soc., Ser. B*, **81**, 649–672.

Barnard, G. A. (1959) Control charts and stochastic processes. *J. Roy. Statist. Soc., Ser. B*, **21**, 239–271.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc., Ser. B*, **57**, 289–300.

Birgé, L. (2001) An alternative point of view on Lepski's method. *IMS Lecture Notes Monograph Series: State of the Art in Probability and Statistics (Leiden, 1999)*, **36**, 113–133.

Bleakley, K. and Vert, J.-P. (2011) The group fused Lasso for multiple change-point detection. hal–00602121.

Boucheron, S., Lugosi, G. and Massart, P. (2013) *Concentration Inequalities: A Nonasymptotic Theory of Independence.* Oxford University Press, Oxford.

Cai, H. and Wang, T. (2021) Estimation of high-dimensional change-points under a group sparsity structure. *arXiv preprint*, arxiv:2107.08724.

Caudron, R. S., C.and White, Green, R. G., Woods, J., Ágústsdóttir, T., Donaldson, C., Greenfield, T., Rivalta, E. and Brandsdóttir, B. (2018) Seismic amplitude ratio analysis of the 2014-15 Bárðarbunga-Holuhraun dike propagation and eruption. *J. Geophysical Res.: Solid Earth*, **123**, 264–276.

Chan, H. P. (2017) Optimal sequential detection in multi-stream data. *Ann. Statist.*, **45**, 2736–2763.

Chan, H. P. and Walther, G. (2015) Optimal detection of multi-sample aligned sparse signals. *Ann. Statist.*, **43**, 1865–1895.

Chen, J. and Gupta, A. K. (1997) Testing and locating variance changepoints with application to stock prices. *J. Amer. Statist. Assoc.*, **92**, 739–747.

Chen, Y., Wang, T. and Samworth, R. J. (2020) *ocd: high-dimensional, multiscale online changepoint detection.* R package, available at https://cran.r-project.org/package=ocd.

Chen, Y., Wang, T. and Samworth, R. J. (2021) Inference in high-dimensional online changepoint detection. *arXiv preprint*, arxiv:2111.01640.

Chen, Y., Wang, T. and Samworth, R. J. (2022) High-dimensional, multiscale online changepoint detection. *J. Roy. Statist. Soc., Ser. B*, **84**, 234–266.

Cheng, D., He, Z. and Schwartzman, A. (2020) Multiple testing of local extrema for detection of change points. *Electron. J. Stat.*, **14**, 3705–3729.

Cho, H. (2016) Change-point detection in panel data via double CUSUM statistic. *Electron. J. Stat.*, **10**, 2000–2038.

Cho, H. and Fryzlewicz, P. (2015) Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *J. Roy. Statist. Soc., Ser. B*, **77**, 475–507.

Cho, H. and Kirch, C. (2021) Bootstrap confidence intervals for multiple change points based on moving sum procedures. *arXiv preprint*, arxiv:2106.12844.

Chu, C.-S. J., Stinchcombe, M. and White, H. (1996) Monitoring structural change. *Econometrica*, **64**, 1045–1065.

Collier, O., Comminges, L. and Tsybakov, A. B. (2017) Minimax estimation of linear and quadratic functionals on sparsity classes. *Ann. Statist.*, **45**, 923–958.

Dette, H., Eckle, T. and Vetter, M. (2020) Multiscale change point detection for dependent data. *Scand. J. Stat.*, **47**, 1243–1274.

Dette, H. and Gösmann, J. (2020) A likelihood ratio approach to sequential change point detection for a general class of parameters. *J. Amer. Statist. Assoc.*, **115**, 1361–1377.

Duncan, A. J. (1952) *Quality Control and Industrial Statistics.* Richard D. Irwin Professional Publishing Inc., Chicago.

Duy, V. N. L., Toda, H., Sugiyama, R. and Takeuchi, I. (2020) Computing valid p-value for optimal changepoint by selective inference using dynaming programming. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, 11356–11367, NeurIPS.

Eichinger, B. and Kirch, C. (2018) A MOSUM procedure for the estimation of multiple random change points. *Bernoulli*, **24**, 526–564.

Enikeeva, F. and Harchaoui, Z. (2019) High-dimensional change-point detection under sparse alternatives. *Ann. Statist.*, **47**, 2051–2079.

Fang, X., Li, J. and Siegmund, D. (2020) Segmentation and estimation of change-point models: False positive control and confidence regions. *Ann. Statist.*, **48**, 1615–1647.

Fearnhead, P. (2006) Exact and efficient Bayesian inference for multiple changepoint problems. *Statist. Comput.*, **16**, 203–213.

Fearnhead, P. and Liu, Z. (2007) On-line inference for multiple changepoint problems. *J. Roy. Statist. Soc., Ser. B*, **69**, 589–605.

Follain, B., Wang, T. and Samworth, R. J. (2022) High-dimensional changepoint estimation with heterogeneous missingness. *J. Roy. Statist. Soc., Ser. B*, to appear.

Frick, K., Munk, A. and Sieling, H. (2014) Multiscale change point inference. *J. Roy. Statist. Soc., Ser. B*, **76**, 495–580.

Fryzlewicz, P. (2014) Wild binary segmentation for multiple change-point detection. *Ann. Statist.*, **42**, 2243–2281.

Fryzlewicz, P. (2020) Detecting possibly frequent change-points: Wild Binary Segmentation 2 and steepest-drop model selection. *J. Korean Statist. Soc.*, **49**, 1027–1070.

Fryzlewicz, P. (2021a) Narrowest Significance Pursuit: inference for multiple change-points in linear models. *arXiv preprint*, arxiv:2009.05431.

Fryzlewicz, P. (2021b) Robust Narrowest Significance Pursuit: inference for multiple change-points in the median. *arXiv preprint*, arxiv:2109.02487.

Gibberd, A. J. and Sandipan, R. (2017) Multiple changepoint estimation in high-dimensional Gaussian graphical models. *arXiv preprint*, arxiv:1712.05786.

Gösmann, J., Kley, T. and Dette, H. (2021) A new approach for open-end sequential change point monitoring. *J. Time Series Anal.*, **42**, 63–84.

Gösmann, J., Stoehr, C., Heiny, J. and Dette, H. (2020) Sequential change point detection in high dimensional time series. *arXiv preprint*, arxiv:2006.00636.

Groeneboom, P. (1989) Brownian motion with a parabolic drift and Airy functions. *Probab. Theory Related Fields*, **81**, 79–109.

Hao, N., Niu, Y. and Zhang, H. (2013) Multiple change-point detection via a screening and ranking algorithm. *Statist. Sinica*, **23**, 1553–1572.

Hausman, A. and Johnston, W. J. (2014) Timeline of a financial crisis: Introduction to the special issue. *J. Bus. Res.*, **67**, 2667–2670.

Hoeffding, W. (1963) Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.*, **58**, 13–30.

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

Horváth, L., Hušková, M., Kokoszka, P. and Steinebach, J. (2004) Monitoring changes in linear models. *J. Statist. Plann. Inference*, **126**, 225–251.

Hušková, M., Prášková, Z. and Steinebach, J. (2015) On the detection of changes in autoregressive time series I. Asymptotics,. *J. Statist. Plann. Inference*, **137**, 1243–1259.

Hušková, M. and Slabý, A. (2001) Permutation tests for multiple changes. *Kybernetika (Prague)*, **37**, 605–622.

Hyun, S., G'Sell, M. and Tibshirani, R. J. (2018) Exact post-selection inference for the generalized lasso path. *Electron. J. Stat.*, **12**, 1053–1097.

Hyun, S., Lin, K., G'Sell, M. and Tibshirani, R. J. (2021) Post-selection inference for change-point detection algorithms with application to copy number variation data. *Biometrics*, **77**, 1037–1049.

Jackson, B., Scargle, J., Barnes, D., Arabhi, S., Alt, A., Gioumousis, P., Gwin, E., Sangtrakulcharoen, P., Tan, L. and Tsai, T. T. (2005) An algorithm for optimal partitioning of data on an interval. *IEEE Signal Process. Lett.*, **12**, 105–108.

Janková, J. and van de Geer, S. (2015) Confidence intervals for high-dimensional inverse covariance estimation. *Electron. J. Stat.*, **9**, 1205–1229.

Javanmard, A. and Montanari, A. (2014) Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.*, **15**, 2869–2909.

Jewell, S., Fearnhead, P. and Witten, D. (2022) Testing for a change in mean after changepoint detection. *J. Roy. Statist. Soc., Ser. B*, to appear.

Jirak, M. (2015) Uniform change point tests in high dimension. *Ann. Statist.*, **43**, 2451–2483.

Kaul, A., Fotopoulos, S. B., Jandhyala, V. K. and Safikhani, A. (2021a) Inference on the change point under a high dimensional sparse mean shift. *Electron. J. Stat.*, **15**, 71–134.

Kaul, A., Jandhyala, V. K. and Fotopoulos, S. B. (2019) An efficient two step algorithm for high dimensional change point regression models without grid search. *J. Mach. Learn. Res.*, **20**, 1–40.

Kaul, A., Zhang, H., Tsampourakis, K. and Michailidis, G. (2021b) Inference on the change point for high dimensional dynamic graphical models. *arXiv preprint*, arxiv:2005.09711.

Ke, Y., Minsker, S., Ren, Z., Sun, Q. and Zhou, W.-X. (2019) User-friendly covariance estimation for heavy-tailed distributions. *Statist. Sci.*, **34**, 454–471.

Killick, R., Fearnhead, P. and Eckley, I. A. (2012) Optimal detection of changepoints with a linear computational cost. *J. Amer. Statist. Assoc.*, **107**, 1590–1598.

Kirch, C. (2008) Bootstrapping sequential change-point tests. *Sequential Anal.*, **27**, 330–349.

Kirch, C., Muhsal, B. and Ombao, H. (2015) Detection of changes in multivariate time series With application to EEG data. *J. Amer. Statist. Assoc.*, **110**, 1197–1216.

Kirch, C. and Stoehr, C. (2019) Sequential change point tests based on U-statistics. *arXiv preprint*, arxiv:1912.08580.

Komlós, J., Major, P. and Tusnády, G. (1976) An approximation of partial sums of independent RVs, and the sample DF. II. *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, **34**, 33–58.

Kovács, S., Li, H., Bühlmann, P. and Munk, A. (2020) Seeded Binary Segmentation: A general methodology for fast and optimal change point detection. *arXiv preprint*, arxiv:2002.06633.

Lai, T. L. (1995) Sequential changepoint detection in quality control and dynamical systems. *J. Roy. Statist. Soc., Ser. B*, **57**, 613–658.

Lai, T. L. (2001) Sequential analysis: some classical problems and new challenges. *Statist. Sinica*, **11**, 303–351.

Laurent, B. and Massart, P. (2000) Adaptive estimation of a quadratic functional by model selection. *Ann. Statist.*, **28**, 1302–1338.

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

Lee, J. D., Sun, D. L., Sun, Y. and Taylor, J. E. (2016) Exact post-selection inference, with application to the lasso. *Ann. Statist.*, **44**, 907–927.

Lee, S., Seo, M. H. and Shin, Y. (2016) The lasso for high dimensional regression with a possible change point. *J. Roy. Statist. Soc., Ser. B*, **78**, 193–210.

Leisch, F., Hornik, K. and Kuan, C.-M. (2000) Monitoring structural changes with the generalized fluctuation test. *Econometric Theory*, **16**, 835–854.

Leonardi, F. and Bühlmann, P. (2016) Computationally efficient change point detection for high-dimensional regression. *arXiv preprint*, arxiv:1601.03704.

Li, H., Munk, A. and Sieling, H. (2016) FDR-control in multiscale change-point segmentation. *Electron. J. Stat.*, **10**, 918–959.

Liu, H., Gao, C. and Samworth, R. J. (2021) Minimax rates in sparse, high-dimensional change point detection. *Ann. Statist.*, **49**, 1081–1112.

Londschien, M., Kovács, S. and Bühlmann, P. (2021) Change-point detection for graphical models in the presence of missing values. *J. Comput. Graph. Statist.*, **30**, 1–12.

Lorden, G. (1971) Procedures for reacting to a change in distribution. *Ann. Math. Statist.*, **42**, 1897–1908.

Maidstone, R., Hocking, T., Rigaill, G. and Fearnhead, P. (2017) On optimal multiple changepoint algorithms for large data. *Statist. Comput.*, **27**, 519–533.

Massart, P. (2007) *Concentration Inequalities and Model Selection: Ecole d'Eté de Probabilités de Saint-Flour XXXIII-2003.* Springer.

Mei, Y. (2010) Efficient scalable schemes for monitoring a large number of data streams. *Biometrika*, **97**, 419–433.

Minsker, S. (2018) Sub-Gaussian estimators of the mean of a random matrix with heavy-tailed entries. *Ann. Statist.*, **46**, 2871–2903.

Mörters, P. and Peres, Y. (2010) *Brownian Motion.* Cambridge University Press, Cambridge.

Moustakides, G. V. (1986) Optimal stopping times for detecting changes in distributions. *Ann. Statist.*, **14**, 1379–1387.

Moustakides, G. V., Polunchenko, A. S. and Tartakovsky, A. G. (2011) A numerical approach to performance analysis of quickest change-point detection procedures. *Statist. Sinica*, **21**, 571–596.

Nam, C. F. H., Aston, J. A. D. and Johansen, A. M. (2012) Quantifying the uncertainty in change points. *J. Time Series Anal.*, **33**, 807–823.

Niu, Y. S. and Zhang, H. (2012) The screening and ranking algorithm to detect DNA copy number variations. *Ann. Appl. Stat.*, **6**, 1306–1326.

Oakland, J. S. (2007) *Statistical Process Control.* Routledge, London.

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

Ombao, H. C., von Sachs, R. and Guo, W. (2005) SLEX analysis of multivariate nonstationary time series. *J. Amer. Statist. Assoc.*, **100**, 519–531.

Padilla, O. H. M., Yu, Y., Wang, D. and Rinaldo, A. (2021a) Optimal nonparametric change point analysis. *Electron. J. Stat.*, **15**, 1154–1201.

Padilla, O. H. M., Yu, Y., Wang, D. and Rinaldo, A. (2021b) Optimal nonparametric multivariate change point detection and localization. *IEEE Trans. Inform. Theory*, **68**, 1922–1944.

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

Page, E. S. (1955) A test for a change in a parameter occurring at an unknown point. *Biometrika*, **42**, 523–527.

Pan, J. and Chen, J. (2006) Application of modified information criterion to multiple change point. *J. Multivariate Anal.*, **97**, 2221–2241.

Pein, F., Sieling, H. and Munk, A. (2017) Heterogeneous change point inference. *J. Roy. Statist. Soc., Ser. B*, **79**, 1207–1227.

Peng, T., Leckie, C. and Ramamohanarao, K. (2004) Proactively detecting distributed denial of service attacks using source IP address monitoring. In *Networking 2004*, 771–782, Springer, Berlin.

Pollak, M. (1978) Optimality and almost optimality of mixture stopping rules. *Ann. Statist.*, **6**, 910–916.

Pollak, M. (1985) Optimal detection of a change in distribution. *Ann. Statist.*, **13**, 206–227.

Pollak, M. and Siegmund, D. (1975) Approximations to the expected sample size of certain sequential tests. *Ann. Statist.*, **3**, 1267–1282.

Pollak, M. and Tartakovsky, A. G. (2009) Optimality properties of the Shiryaev–Roberts procedure. *Statist. Sinica*, **19**, 1729–1739.

Polunchenko, A. S. and Tartakovsky, A. G. (2010) On optimality of the Shiryaev–Roberts procedure for detecting a change in distribution. *Ann. Statist.*, **38**, 3445–3457.

Rigaill, G. (2015) A pruned dynamic programming algorithm to recover the best segmentations with 1 to $K_{\max}$ change-points. *J. SFdS*, **156**, 180–205.

Rinaldo, A., Wang, D., Wen, Q., Willett, R. and Yu, Y. (2021) Localizing changes in high-dimensional regression models. In *International Conference on Artificial Intelligence and Statistics*, 2089–2097, PMLR.

Ritov, Y. (1990) Decision theoretic optimality of the cusum procedure. *Ann. Statist.*, **18**, 1464–1469.

Roberts, S. W. (1959) Control chart tests based on geometric moving averages. *Technometrics*, **1**, 239–250.

Roberts, S. W. (1966) A comparison of some control chart procedures. *Technometrics*, **8**, 411–430.

Romano, G., Eckley, I., Fearnhead, P. and Rigaill, G. (2022) Fast online changepoint detection via functional pruning CUSUM statistics. *arXiv preprint*, arxiv:2110.08205.

Ryan, J. A., Ulrich, J. M., Thielen, W., Teetor, P., Bronder, S. and Ulrich, M. J. M. (2020) *quantmod: quantitative financial modelling framework.* R package, available at https://cran.r-project.org/package=quantmod.

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

Scott, A. J. and Knott, M. (1974) A cluster analysis method for grouping means in the analysis of variance. *Biometrics*, **30**, 507–512.

Shewhart, W. A. (1931) *Economic Control of Quality of Manufactured Product.* Van Nostrand, New York.

Shiryaev, A. N. (1961) The problem of the most rapid detection of a disturbance in a stationary process. *Soviet Math. Dokl.*, **2**, 795–799.

Shiryaev, A. N. (1963) On optimum methods in quickest detection problems. *Theory Probab. Appl.*, **8**, 22–46.

Siegmund, D. (1986) Boundary crossing probabilities and statistical applications. *Ann. Statist.*, **14**, 361–404.

Soh, Y. S. and Chandrasekaran, V. (2017) High-dimensional change-point estimation: Combining filtering with convex optimization. *Appl. Comp. Harm. Anal.*, **43**, 122–147.

Tartakovsky, A., Nikiforov, I. and Basseville, M. (2014) *Sequential Analysis: Hypothesis testing and Changepoint Detection.* Chapman and Hall, London.

Tartakovsky, A. G., Rozovskii, B. L., Blažek, R. B. and Kim, H. (2006) Detection of intrusions in information systems by sequential change-point methods. *Statistical Methodology*, **3**, 252–293.

Tibshirani, R. J. (2014) Adaptive piecewise polynomial estimation via trend filtering. *Ann. Statist.*, **42**, 285–323.

Tibshirani, R. J. and Taylor, J. (2011) The solution path of the generalized lasso. *Ann. Statist.*, **39**, 1335–1371.

Tibshirani, R. J., Taylor, J., Lockhart, R. and Tibshirani, R. (2016) Exact post-selection inference for sequential regression procedures. *J. Amer. Statist. Assoc.*, **111**, 600–620.

van de Geer, S., Bühlmann, P., Ritov, Y. and Dezeure, R. (2014) On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.*, **42**, 1166–1202.

van der Vaart, A. W. and Wellner, J. A. (1996) *Weak Convergence and Empirical Processes.* Springer, New York.

Vostrikova, L. (1981) Detecting 'disorder' in multidimensional random processes. *Soviet Math. Dokl.*, **24**, 55–59.

Wainwright, M. (2019) *High-dimensional Statistics: A Non-asymptotic Viewpoint.* Cambridge University Press, Cambridge.

Wald, A. and Wolfowitz, J. (1948) Optimum character of the sequential probability ratio test. *Ann. Math. Statist.*, **19**, 326–339.

Wald, W. (1947) *Sequential Analysis.* John Wiley & Sons, Inc., New York.

Wang, D., Yu, Y. and Rinaldo, A. (2020) Univariate mean change point detection: penalization, CUSUM and optimality. *Electron. J. Stat.*, **14**, 1917–1961.

Wang, D., Yu, Y. and Rinaldo, A. (2021a) Optimal change point detection and localization in sparse dynamic networks. *Ann. Statist.*, **49**, 203–232.

Wang, D., Yu, Y. and Rinaldo, A. (2021b) Optimal covariance change point localization in high dimensions. *Bernoulli*, **27**, 554–575.

Wang, R., Zhu, C., Volgushev, S. and Shao, X. (2022) Inference for change points in high dimensional data via self-normalization. *Ann. Statist.*, **50**, 781–806.

Wang, T. and Samworth, R. J. (2018) High dimensional change point estimation via sparse projection. *J. Roy. Statist. Soc., Ser. B*, **80**, 57–83.

Xie, L., Xie, Y. and Moustakides, G. V. (2019) Asynchronous multi-sensor change-point detection for seismic tremors. In *IEEE International Symposium on Information Theory (ISIT)*, 787–791, IEEE.

Xie, Y. and Siegmund, D. (2013) Sequential multi-sensor change-point detection. *Ann. Statist.*, **41**, 670–692.

Yao, Y.-C. (1988) Estimating the number of change-points via Schwarz' criterion. *Stat. Probab. Lett.*, **6**, 181–189.

Yao, Y.-C. and Au, S. T. (1989) Least-squares estimation of a step function. *Sankhya A*, **51**, 370–381.

Yu, B. (1997) Assouad, Fano and Le Cam. In *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics* (Pollard, D., Torgersen, E. and L., Y. G., eds.), 423–435, Springer, New York.

Yu, M. and Chen, X. (2020) Finite sample change point inference and identification for high-dimensional mean vectors. *J. Roy. Statist. Soc., Ser. B*, **83**, 247–270.

Yu, Y., Bradic, J. and Samworth, R. J. (2021) Confidence intervals for high-dimensional Cox models. *Statist. Sinica*, **31**, 243–267.

Yu, Y., Padilla, O. H. M., Wang, D. and Rinaldo, A. (2020) A note on online change point detection. *arXiv preprint*, arxiv:2006.03283.

Zeileis, A., Leisch, F., Kleiber, C. and Hornik, K. (2005) Monitoring structural change in dynamic econometric models. *J. Appl. Econometrics*, **20**, 99–121.

Zhang, C.-H. and Zhang, S. S. (2014) Confidence intervals for low dimensional parameters in high dimensional linear models. *J. Roy. Statist. Soc., Ser. B*, **76**, 217–242.

Zhang, N. R., Siegmund, D. O., Ji, H. and Li, J. Z. (2010) Detecting simultaneous changepoints in multiple sequences. *Biometrika*, **97**, 631–645.

Zhu, Z. and Zhou, W. (2021) Taming heavy-tailed features by shrinkage. In *International Conference on Artificial Intelligence and Statistics*, 3268–3276, PMLR.

Zou, C., Wang, Z., Zi, X. and Jiang, W. (2015) An efficient online monitoring method for high-dimensional data streams. *Technometrics*, **57**, 374–387.