Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: www.tandfonline.com/journals/uasa20

# Inference in High-Dimensional Online Changepoint Detection

Yudong Chen, Tengyao Wang & Richard J. Samworth

**To cite this article:** Yudong Chen, Tengyao Wang & Richard J. Samworth (26 May 2023): Inference in High-Dimensional Online Changepoint Detection, Journal of the American Statistical Association, DOI: 10.1080/01621459.2023.2199962

**To link to this article:** https://doi.org/10.1080/01621459.2023.2199962

© 2023 The Author(s). Published with license by Taylor & Francis Group, LLC.

View supplementary material

Published online: 26 May 2023.

Submit your article to this journal

Article views: 1594

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at https://www.tandfonline.com/action/journalInformation?journalCode=uasa20

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2023, VOL. 00, NO. 0, 1–12: Theory and Methods
https://doi.org/10.1080/01621459.2023.2199962

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS

Check for updates

# Inference in High-Dimensional Online Changepoint Detection

Yudong Chen[a,b], Tengyao Wang[b], and Richard J. Samworth[a]

[a]Statistical Laboratory, University of Cambridge, Cambridge, UK; [b]London School of Economics and Political Science, London, UK

**ABSTRACT**

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 produces an interval with guaranteed nominal coverage, and whose length is, with high probability, of the same order as the average detection delay, up to a logarithmic factor. The corresponding support estimate enjoys control of both false negatives and false positives. Simulations confirm the effectiveness of our methodology, and we also illustrate its applicability on the U.S. excess deaths data from 2017 to 2020. The supplementary material, which contains the proofs of our theoretical results, is available online.

**ARTICLE HISTORY**
Received November 2021
Accepted April 2023

**KEYWORDS**
Confidence interval; Sequential method; Sparsity; Support estimate

## 1. Introduction

The real-time monitoring of evolving processes has become a characteristic feature of 21st century life. Watches and defibrillators track health data, Covid-19 case numbers are reported on a daily basis and financial decisions are made continuously based on the latest market movements. Given that changes in the dynamics of such processes are frequently of great interest, it is unsurprising that the area of changepoint detection has undergone a renaissance over the last 5–10 years.

One of the features of modern datasets that has driven much of the recent research in changepoint analysis is high dimensionality, where we monitor many processes simultaneously, and seek to borrow strength across the different series to identify changepoints. The nature of series that are tracked in applications, as well as the desire to evade to the greatest extent possible the curse of dimensionality, means that it is commonly assumed that the signal of interest is relatively sparse, in the sense that only a small proportion of the constituent series undergo a change. Furthermore, the large majority of these works have focused on the retrospective (or *offline*) 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; Soh and Chandrasekaran 2017; Wang and Samworth 2018; Enikeeva and Harchaoui 2019; Padilla et al. 2021; Kaul et al. 2021; Liu, Gao, and Samworth 2021; Londschien, Kovács, and Bühlmann 2021; Rinaldo et al. 2021; Follain, Wang, and Samworth 2022). Nevertheless, the related problem where one observes data sequentially and seeks to declare changes as soon as possible after they have occurred, is nowadays receiving increasing attention (e.,g. Kirch and Stoehr 2019; Dette and Gösmann 2020; Gösmann et al. 2020; Chen, Wang, and Samworth 2022; Yu et al. 2021). Although the focus of our review here has been on recent developments, including finite-sample results in multivariate and high-dimensional settings, we also mention that changepoint analysis has a long history (e.g., Page 1954). Entry points to this classical literature include Csörgő and Horváth (1997) and Horváth and Rice (2014). For univariate data, sequential changepoint detection is also studied under the banner of statistical process control (Duncan 1952; Tartakovsky, Nikiforov, and Basseville 2014). In the field of high-dimensional statistical inference more generally, uncertainty quantification has become a major theme over the last decade, originating with influential work on the debiased Lasso in (generalized) 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).

The aim of this article 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 (e.g., Kaul et al. 2021) 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. By this we mean that the computational complexity for processing each new observation, as well as the storage requirements, should

**CONTACT** Richard J. Samworth r.samworth@statslab.cam.ac.uk Statistical Laboratory, University of Cambridge, Wilberforce Road, Cambridge, CB3 0WB, UK.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.
© 2023 The Author(s). Published with license by Taylor & Francis Group, LLC.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

depend only on the number of bits needed to represent the new data point observed.[1] The 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 $X_1,\dots,X_z \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ and $X_{z+1},X_{z+2},\dots \overset{\text{iid}}{\sim} \mathcal{N}(\theta,1)$. Without loss of generality, we assume that $\theta > 0$. Suppose that $\theta$ is known to be at least $b > 0$ and, for $n \in \mathbb{N}$, define *residual tail lengths*

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

In the case of a tie, we choose the smallest $h$ achieving the maximum. 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 $X_{n-h+1},\dots,X_n$, the quantity $t_{n,b}$ is the tail length for which the likelihood ratio statistic is maximized. 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 $\big[\max\{N-t_{N,b}-g(\alpha,b),0\},N\big]$, 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 2 for details, as well as Section 3.5 for further discussion.

The procedure for the inference tasks discussed above, which we call `ocd_CI` (short for **o**nline **c**hangepoint **d**etection **C**onfidence **I**ntervals), can be run in conjunction with any base sequential changepoint detection procedure. However, we recommend using the `ocd` algorithm introduced by Chen, Wang, and Samworth (2022), or 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. The corresponding inferential procedures inherit this same online property, thereby making them applicable even in very high-dimensional settings and where changes may be rare, so we may need to see many new data points before declaring a change.

In Section 3 we study the theoretical performance of the `ocd_CI` procedure. In particular, we prove in Theorem 1 that, whenever the base sequential detection procedure satisfies a patience and detection delay condition, the confidence interval has at least nominal coverage for a suitable choice of input parameters. Theorem 2 provides a corresponding guarantee on the length of the interval. In Section 3.3, we show that by using `ocd′` as the base procedure, the aforementioned patience and detection delay condition is indeed satisfied. As a result, the output confidence interval has guaranteed nominal coverage and the length of the interval is of the same order as the average detection delay for the base `ocd′` procedure, up to a poly-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, we prove that, with high probability, it is able both to recover the effective support of the signal (see Section 3.1 for a formal definition), and to avoid noise coordinates. We then broaden the scope of applicability of our methodology in Section 3.4 by relaxing our distributional assumptions to deal with sub-Gaussian or sub-exponential data. Finally, in Section 3.5, we introduce a modification of our algorithm that permits an arbitrarily loose lower bound $\beta > 0$ on the Euclidean norm of the vector of mean change to be employed, with only a logarithmic increase in the confidence interval length guarantee and the computational cost.

An attraction of our theoretical results is that we are able to handle arbitrary spatial (cross-sectional) dependence between the different coordinates of our data stream. On the other hand, two limitations of our analysis for practical use are that real data may exhibit both heavier than sub-exponential tails and temporal dependence. While a full theoretical analysis of the `ocd_CI` algorithm in these contexts appears to be challenging, we have made some practical suggestions regarding these issues in Sections 3.4 and 4.4, respectively.

Section 4 is devoted to a study of the numerical performance of our methodological proposals. Our simulations confirm that the `ocd_CI` methodology (with the `ocd` base procedure) 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. We further demonstrate the way in which naive application of offline methods may lead to poor performance in this problem. Moreover, in Section 4.4, we apply our methods to excess death data from the Covid-19 pandemic in the United States. Proofs, auxiliary results, extensions to sub-Gaussian and sub-exponential settings and additional simulation results are provided in the supplementary material.

We conclude this introduction with some notation used throughout the article. We write $\mathbb{N}_0$ for the set of all nonnegative integers. For $d \in \mathbb{N}$, we write $[d] := \{1,\dots,d\}$. 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 $\operatorname{sargmax}_{x\in S} f(x) := \min \operatorname{argmax}_{x\in S} f(x)$ and $\operatorname{largmax}_{x\in S} f(x) := \max \operatorname{argmax}_{x\in S} f(x)$ for the smallest and largest maximizers of $f$ in $S$, and define $\operatorname{sargmin}_{x\in S} f(x)$ and $\operatorname{largmin}_{x\in S} f(x)$ analogously. For a vector $v = \big(v^1,\dots,v^M\big)^\top \in$

[1] Here, we ignore the errors in rounding real numbers to machine precision; thus, we do not distinguish between continuous random variables and quantized versions where the data have been rounded to machine precision.

$\mathbb{R}^M$, we define $\|v\|_0 := \sum_{i=1}^M \mathbb{1}_{\{v^i \neq 0\}}$, $\|v\|_2 := \left\{\sum_{i=1}^M (v^i)^2\right\}^{1/2}$ and $\|v\|_\infty := \max_{i\in[M]} |v^i|$. In addition, for $j \in [M]$, we define $\|v^{-j}\|_2 := \left\{\sum_{i:i\neq j} (v^i)^2\right\}^{1/2}$. For a matrix $A = (A^{i,j}) \in \mathbb{R}^{d_1 \times d_2}$ and $j \in [d_2]$, we write $A^{\cdot,j} := \left(A^{1,j}, \ldots, A^{d_1,j}\right)^\top \in \mathbb{R}^{d_1}$ and $A^{-j,j} := \left(A^{1,j}, \ldots, A^{j-1,j}, A^{j+1,j} \ldots, A^{d_1,j}\right)^\top \in \mathbb{R}^{d_1-1}$. 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 $U \geq_{\text{st}} V$ or $V \leq_{\text{st}} U$ 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 $\min \emptyset := \infty$, $\max \emptyset := -\infty$.

## 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. Here and throughout, a stopping time is understood to be with respect to the natural filtration, so that the event $\{N = n\}$ belongs to the $\sigma$-algebra generated by $X_1, \ldots, X_n$. The focus of this work 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 $1 - \alpha$.

For $i \in \mathbb{N}$ and $j \in [p]$, let $X_i^j$ denote the $j$th coordinate of $X_i$. The `ocd_CI` algorithm relies on a lower bound $\beta > 0$ for the $\ell_2$-norm of the vector of mean change, sets of signed scales $\mathcal{B}$ and $\mathcal{B}_0$ defined in terms of $\beta$ and a base sequential changepoint detection procedure CP. As CP processes each new data vector, we update the matrix of residual tail lengths $(t^j_{n,b})_{j\in[p],b\in\mathcal{B}\cup\mathcal{B}_0}$ with $t^j_{n,b} := \text{sargmax}_{0\leq h\leq n} \sum_{i=n-h+1}^n (X_i^j - b/2)$, as well as corresponding tail partial sum vectors $\left(A^{\cdot,j}_{n,b}\right)_{j\in[p],b\in\mathcal{B}\cup\mathcal{B}_0}$, where $A^{j',j}_{n,b} := \sum_{i=n-t^j_{n,b}+1}^n X_i^{j'}$.

After the base procedure CP declares a change at a stopping time $N$, we identify an "anchor" coordinate $\hat{j} \in [p]$ and a signed anchor scale $\hat{b} \in \mathcal{B}$, where

$$(\hat{j}, \hat{b}) := \underset{(j,b)\in[p]\times\mathcal{B}}{\text{argmax}} \sum_{j'\in[p]\setminus\{j\}} \frac{\left(A^{j',j}_{N,b}\right)^2}{t^j_{N,b} \vee 1} \mathbb{1}_{\left\{|A^{j',j}_{N,b}| \geq a\sqrt{t^j_{N,b}\vee 1}\right\}}.$$

The intuition is that the anchor coordinate and signed anchor 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, 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 centered 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 have unit variance around this centering value. 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 $\theta^j$, for which $\left|E^{j,\hat{j}}_{N,\hat{b}}\right| - |b|(t^{\hat{j}}_{N,\hat{b}})^{1/2} \geq d_1$. In fact, such a signed scale, if it exists, can always be chosen to be from $\mathcal{B}_0$. As a convenient byproduct, the set of indices $j$ for which the latter inequality holds, which we denote as $\hat{\mathcal{S}}$, 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 $\left|E^{j,\hat{j}}_{N,\hat{b}}\right| - b(t^{\hat{j}}_{N,\hat{b}})^{1/2} \geq d_1$. 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 $\tilde{b}^j$; this can be regarded as a shrunken estimate of $\theta^j$, so plays the role of the lower bound $b$ from the univariate problem discussed in the introduction. Finally, then, our confidence interval is 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 $\tilde{b}^j$.

As a device to facilitate our theory, the `ocd_CI` algorithm allows the practitioner the possibility of observing a further $\ell$ 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 1 that the output 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 4 indicates that $\ell = 0$ provides confidence intervals of reasonable length in practice. Similarly, Theorem 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 this `ocd_CI` confidence interval construction is given in Algorithm 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 is equal to the sum of the corresponding quantities for the CP base procedure and $O\big(p^2 \log(ep)\big)$ regardless of the observation history. Thus, the `ocd_CI` method inherits the property of being an online algorithm, as discussed in the introduction, from any online CP base procedure.

A natural choice for the base online changepoint detection procedure CP is the ocd algorithm, or its variant ocd′, introduced by Chen, Wang, and Samworth (2022). Both are online algorithms, with computational complexity per new observation and storage requirements of $O\big(p^2\log(ep)\big)$. The ocd′ base procedure is considered for the theoretical analysis in Section 3 due to its known patience and detection delay guarantees, while we prefer ocd for numerical studies and practical use. For the reader's convenience, the ocd and ocd′ algorithms are provided as Algorithms S1 and S2, respectively in Section S3 of the supplementary materials.

**Algorithm 1:** Pseudo-code for the confidence interval construction algorithm ocd_CI

**Input:** $X_1, X_2, \ldots \in \mathbb{R}^p$ observed sequentially, $\beta > 0$, $a \geq 0$, an online changepoint detection procedure CP, $d_1, d_2 > 0$ and $\ell \in \mathbb{N}_0$

**Set:** $b_{\min} = \frac{\beta}{\sqrt{2^{\lfloor \log_2(2p)\rfloor}\log_2(2p)}}$, $\mathcal{B}_0 = \{\pm b_{\min}\}$, $\mathcal{B} = \big\{\pm 2^{m/2} b_{\min} : m = 1, \ldots, \lfloor \log_2(2p)\rfloor\big\}$, $n = 0$, $A_b = \mathbf{0} \in \mathbb{R}^{p\times p}$ 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 $X_n$ and update CP with $X_n$

  **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 $A_b^{\cdot,j} \leftarrow 0$

**until** CP *declares a change*;

Observe $\ell$ new data vectors $X_{n+1}, \ldots, X_{n+\ell}$

Set $E_b^{j',j} \leftarrow \frac{A_b^{j',j} + \sum_{i=n+1}^{n+\ell} X_i^{j'}}{\sqrt{(t_b^j+\ell)\vee 1}}$ for $j', j \in [p]$, $b \in \mathcal{B}\cup\mathcal{B}_0$

Compute $Q_b^j \leftarrow \sum_{j'\in[p]\setminus\{j\}} (E_b^{j',j})^2 \mathbb{1}_{\{|E_b^{j',j}|\geq a\}}$ for $j \in [p]$, $b\in\mathcal{B}$

$(\hat{j}, \hat{b}) \leftarrow \operatorname{argmax}_{(j,b)\in[p]\times\mathcal{B}} Q_b^j$

$\hat{\mathcal{S}} \leftarrow \Big\{j \in [p]\setminus\{\hat{j}\} : \big|E_{\hat{b}}^{j,\hat{j}}\big| - b_{\min}(t_{\hat{b}}^{\hat{j}} + \ell)^{1/2} \geq d_1\Big\}$

**for** $j \in \hat{\mathcal{S}}$ **do**

  $\tilde{b}^j \leftarrow \operatorname{sgn}\big(E_{\hat{b}}^{j,\hat{j}}\big)\max\Big\{b \in (\mathcal{B}\cup\mathcal{B}_0)\cap(0,\infty) : \big|E_{\hat{b}}^{j,\hat{j}}\big| - b(t_{\hat{b}}^{\hat{j}} + \ell)^{1/2} \geq d_1\Big\}$

**Output:** Confidence interval

$$\mathcal{C} = \Big[\max\Big\{n - \min_{j\in\hat{\mathcal{S}}}\Big\{t_{\tilde{b}^j}^j + \tfrac{d_2}{(\tilde{b}^j)^2}\Big\}, 0\Big\}, n\Big]$$

## 3. Theoretical Analysis

Throughout this section, we will assume that the sequential observations $X_1, X_2, \ldots$ are independent, and that for some unknown covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$ whose diagonal entries are all equal to 1, there exist $z \in \mathbb{N}_0$ and $\theta = (\theta^1, \ldots, \theta^p)^\top \neq 0$ for which $X_1, \ldots, X_z \sim \mathcal{N}_p(0,\Sigma)$ and $X_{z+1}, X_{z+2}, \ldots \sim \mathcal{N}_p(\theta,\Sigma)$. We let $\vartheta := \|\theta\|_2$, and write $\mathbb{P}_{z,\theta,\Sigma}$ for probabilities computed under this model, though in places we omit the subscripts for brevity. Define the *effective sparsity* of $\theta$, denoted $s(\theta)$, to be the smallest $s \in \big\{2^0, 2^1, \ldots, 2^{\lfloor\log_2(p)\rfloor}\big\}$ such that the corresponding *effective support* $\mathcal{S}(\theta) := \big\{j \in [p] : |\theta^j| \geq \|\theta\|_2/\sqrt{s\log_2(2p)}\big\}$ has cardinality at least $s(\theta)$. Thus, the sum of squares of coordinates in the effective support of $\theta$ has the same order of magnitude as $\|\theta\|_2^2$, up to logarithmic factors. Moreover, if at most $s$ components of $\theta$ are nonzero, then $s(\theta) \leq s$, and the equality is attained when, for example, all nonzero coordinates have the same magnitude.

For $r > 0$ and an online changepoint detection procedure CP characterized by an extended stopping time $N$, we define

$$g(r;N) := \sup_{z\in\mathbb{N}_0} \mathbb{P}_{z,\theta,\Sigma}(N > z + r). \tag{2}$$

### 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 whenever the base online changepoint detection procedure satisfies a patience and detection delay condition.

*Theorem 1.* Let $p \geq 2$ and fix $\alpha \in (0,1)$. Suppose that $\vartheta \geq \beta > 0$. Let CP be an online changepoint procedure characterized by an extended stopping time $N$ satisfying

$$\mathbb{P}_{z,\theta,\Sigma}(N \leq z) + g(r;N) + 4rp^2\log_2^2(4p)e^{-r\beta^2/(8s\log_2(2p))} \leq \frac{3}{4}\alpha \tag{3}$$

for some $r \geq 1$. Then, with inputs $(X_t)_{t\in\mathbb{N}}$, $\beta > 0$, $a \geq 0$, CP, $d_1 = \sqrt{\frac{5r\beta^2}{9s\log_2(2p)}}$, $d_2 = 4d_1^2$ and $\ell \geq 0$, the output confidence interval $\mathcal{C}$ from Algorithm 1 satisfies

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

As mentioned in Section 2, our coverage guarantee in Theorem 1 holds even with $\ell = 0$, that is, with no additional sampling. Condition (3) places a joint assumption on the base changepoint procedure CP and the parameter $r$, the latter of which appears in the inputs $d_1$ and $d_2$ of Algorithm 1. The first term on the left-hand side of (3) is the false alarm rate of the stopping time $N$ associated with CP. The second term can be interpreted as an upper bound on the probability of the detection delay of $N$ being larger than $r$, and in addition we also need $r$ to be at least of order $s/\beta^2$ up to logarithmic factors for the third term to be sufficiently small. See Section 3.3 for further discussion, where in particular we provide a choice of $r$ for which (3) holds with the ocd′ base procedure.

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

*Theorem 2.* Fix $\alpha \in (0,1)$. Assume that $\theta$ has an effective sparsity of $s := s(\theta) \geq 2$ and that $\vartheta \geq \beta > 0$. let CP be an online changepoint detection procedure characterized by an extended stopping time $N$ that satisfies (3) for some $r \geq 1$. Then there exists a universal constant $C > 0$ such that, with inputs $(X_t)_{t\in\mathbb{N}}$, $\beta > 0$, $a = C\sqrt{\log(rp/\alpha)}$, CP, $d_1 = \sqrt{\frac{5r\beta^2}{9s\log_2(2p)}}$, $d_2 = 4d_1^2$,

$\ell \geq 80r$, the length $L$ of the output confidence interval $\mathcal{C}$ satisfies

$$\mathbb{P}_{z,\theta,\Sigma}(L > 8r) \leq \alpha.$$

As mentioned following Theorem 1, we can take $r$ to be the maximum of an appropriate quantile of the detection delay distribution of CP and a quantity that is of order $s/\beta^2$ up to logarithmic factors. The main conclusion of Theorem 2 is that, with high probability, the length of the confidence interval is then of this same order $r$. Whenever the quantile of the detection delay distribution achieves the maximum above—which is the case, up to logarithmic factors, for the ocd′ base procedure (see Proposition 5)—we can conclude that with high probability, the length of the ocd_CI confidence interval is of the same order as this detection delay quantile (which is the best one could hope for). Note that the choices of inputs in Theorem 2 are identical to those in Theorem 1, except that we now ask for order $r$ additional observations after the changepoint declaration.

### 3.2. Support Recovery

Recall the definition of $\mathcal{S}(\theta)$ from the beginning of this section, and denote $\mathcal{S}_\beta(\theta) := \{j \in [p] : |\theta^j| \geq b_{\min}\}$, where $b_{\min}$, defined in Algorithm 1, is the smallest positive scale in $\mathcal{B} \cup \mathcal{B}_0$, We will suppress the dependence on $\theta$ of both these quantities in this section. Theorem 3 provides a support recovery guarantee for $\hat{\mathcal{S}}$, defined in Algorithm 1. 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 1 and 2.

*Theorem 3.* Let $p \geq 2$ and fix $\alpha \in (0,1)$. Suppose that $\vartheta \geq \beta > 0$. Let CP be an online changepoint detection procedure characterized by an extended stopping time $N$ that satisfies (3) for some $r \geq 1$.

(a) Then, with inputs $(X_t)_{t\in\mathbb{N}}$, $\beta > 0$, $a \geq 0$, CP, $d_1 = \sqrt{\frac{5r\beta^2}{9s\log_2(2p)}}$, $\ell \geq 0$, we have

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

(b) Now assume that $\theta$ has an effective sparsity of $s := s(\theta) \geq 2$. Then there exists a universal constant $C > 0$ such that, with inputs $(X_t)_{t\in\mathbb{N}}$, $\beta > 0$, $a = C\sqrt{\log(rp/\alpha)}$, CP, $d_1 = \sqrt{\frac{5r\beta^2}{9s\log_2(2p)}}$, $\ell \geq 80r$, we have

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

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 4 and in Section 3.3.

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. In fact, we can already establish this optimality by restricting the cross-sectional covariance matrix to be the identity matrix. Thus, given $\theta \in \mathbb{R}^p$ and $z \in \mathbb{N}_0$, we write $\mathbb{P}_{z,\theta}$ for a probability measure under which $(X_n)_{n\in\mathbb{N}}$ are independent with $X_n \sim \mathcal{N}_p(\theta \mathbb{1}_{\{n>z\}}, I_p)$. For $r > 0$ and $m \in [p] \cup \{0\}$, write

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

Define $\mathcal{T}$ to be the set of stopping times with respect to the natural filtration $(\mathcal{F}_n)_{n\in\mathbb{N}_0}$, and set

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

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 $\psi$ is said to be $\mathcal{F}_N$-measurable if for any $A \in 2^{[p]}$ and $n \in \mathbb{N}_0$, we have that $\psi^{-1}(A) \cap \{N = n\}$ is $\mathcal{F}_n$-measurable.

*Proposition 4.* For $r > 0$ and $m \geq 15$, we have

$$\inf_{N\in\mathcal{T}_{r,m}} \inf_{\psi\in\mathcal{J}_N} \sup_{z\in\mathbb{N}_0,\theta\in\Theta_{r,m}} \mathbb{E}_{z,\theta}\, d\big(\psi(X_1, X_2, \ldots), \mathrm{supp}(\theta)\big) \geq \frac{m}{32}.$$

This proposition considers any support estimation algorithm obtained from a stopping time in $\mathcal{T}_{r,m}$, 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. This result can be interpreted as an optimality guarantee for the support recovery property of the ocd_CI algorithm presented in Theorem 3(b), provided that the base procedure $N$ belongs to the class $\mathcal{T}_{r,m}$, and that $N$ and $r$ satisfy (3). See Section 3.3 for further discussion.

### 3.3. Using ocd′ as the Base Procedure

In this section, we provide a value of $r$ that suffices for condition (3) to hold when we take our base procedure to be ocd′. For the convenience of the reader, this algorithm is presented as Algorithm S2 in Section S3 of the supplementary materials, where we also provide interpretation to the input parameters $\tilde{a}$, $T^{\mathrm{diag}}$ and $T^{\mathrm{off}}$.

*Proposition 5.* Fix $\alpha \in (0,1)$ and $\gamma > 0$. Assume that $\theta$ has an effective sparsity of $s := s(\theta) \geq 2$, that $\vartheta \geq \beta > 0$ and that $z \leq 2\alpha\gamma$. Then with inputs $(X_t)_{t\in\mathbb{N}}$, $\beta > 0$, $\tilde{a} = \sqrt{2\log\{16p^2\gamma\log_2(2p)\}}$, $T^{\mathrm{diag}} = \log\{16p\gamma\log_2(4p)\}$ and $T^{\mathrm{off}} = 8\log\{16p\gamma\log_2(2p)\}$ in the ocd′ procedure, there exists a universal constant $C' > 0$ such that for all

$$r \geq \frac{C' s\log_2(2p)\log\{p\gamma\alpha^{-1}(\beta^{-2} \vee 1)\}}{\beta^2} + 2 =: r_1, \qquad (4)$$

the output $N$ satisfies (3).

By combining Proposition 5 with Theorems 1–3, respectively, we immediately arrive at the following corollary.

*Corollary 6.* Fix $\alpha \in (0,1)$, $\gamma > 0$. Assume that $\theta$ has an effective sparsity of $s := s(\theta) \geq 2$, that $\vartheta \geq \beta > 0$ and that $z \leq 2\alpha\gamma$. Let $(X_t)_{t\in\mathbb{N}}$, $\beta > 0$, $\tilde{a} = \sqrt{2\log\{16p^2\gamma\log_2(2p)\}}$, $T^{\mathrm{diag}} = \log\{16p\gamma\log_2(4p)\}$ and $T^{\mathrm{off}} = 8\log\{16p\gamma\log_2(2p)\}$ be the inputs of the ocd′ procedure. Then the following statements hold:

(a) With extra inputs $a \geq 0$, CP = ocd′, $d_1 = \sqrt{\frac{5r_1\beta^2}{9s\log_2(2p)}}$, $d_2 = 4d_1^2$ and $\ell \geq 0$ for Algorithm 1, the output confidence interval $\mathcal{C}$ and the support estimate $\hat{\mathcal{S}}$ satisfy $\mathbb{P}_{z,\theta,\Sigma}(z \in \mathcal{C}) \geq 1-\alpha$ and $\mathbb{P}_{z,\theta,\Sigma}(\hat{\mathcal{S}} \subseteq \mathcal{S}_\beta) \geq 1-\alpha$.

(b) There exists a universal constant $C > 0$ such that, with extra inputs $a = C\sqrt{\log(r_1p/\alpha)}$, CP = ocd′, $d_1 = \sqrt{\frac{5r_1\beta^2}{9s\log_2(2p)}}$, $d_2 = 4d_1^2$, $\ell \geq 80r_1$ for Algorithm 1, the length $L$ of the output confidence interval $\mathcal{C}$ and the support estimate satisfy $\mathbb{P}_{z,\theta,\Sigma}(L > 8r_1) \leq \alpha$ and $\mathbb{P}_{z,\theta,\Sigma}(\hat{\mathcal{S}} \cup \{\hat{j}\} \supseteq \mathcal{S}) \geq 1-\alpha$.

Corollary 6 reveals that, when ocd′ is used as the base procedure, the ocd_CI methodology provides guaranteed confidence interval coverage. Moreover, up to poly-logarithmic factors, with an additional $O\big(1 \vee (s/\beta^2)\big)$ post-change observations, the ocd_CI interval length is of the same order as the average detection delay. In terms of signal recovery, Corollary 6(b) shows that with high probability, ocd_CI 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 $s/\vartheta^2$ is bounded away from zero for simplicity of discussion (though see also Section 3.5 for discussion of the effect of the choice of $\beta$), Proposition S3 in the supplementary materials also reveals that the ocd′ base procedure belongs to $\mathcal{T}_{r,m}$ with $r$ of order $s/\vartheta^2$, up to logarithmic factors, and $m = |\{j : |\theta^j| \leq 1/(8\sqrt{r})\}|$. On the other hand, Proposition 4 shows that any such support estimation algorithm makes on average a nonvanishing 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 with base procedure ocd′ 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 Corollary 6(a)).

### 3.4. Relaxing the Gaussianity Assumption

It is natural to ask to what extent the theory of Sections 3.1–3.3 can be generalized beyond the Gaussian setting. The purpose of this section, then, is to describe how our earlier results can be modified to handle both sub-Gaussian and sub-exponential data. Recall that we say a random variable $Z$ with $\mathbb{E}Z = 0$ is *sub-Gaussian* with variance parameter $\sigma^2 > 0$ if $\mathbb{E}e^{\lambda Z} \leq e^{\sigma^2\lambda^2/2}$ for all $\lambda \in \mathbb{R}$, and is *sub-exponential* with variance parameter $\sigma^2 > 0$ and rate parameter $A > 0$ if $\mathbb{E}e^{\lambda Z} \leq e^{\sigma^2\lambda^2/2}$ for all $|\lambda| \leq A$.

We first consider the sub-Gaussian setting where $X_1,\ldots,X_z,X_{z+1}-\theta,X_{z+2}-\theta,\ldots$ are independent, each having sub-Gaussian components with variance parameter 1. Note that this data generating mechanism no longer requires all pre-change observations to be identically distributed, and likewise the post-change observations need not all have the same distribution. We assume that the base changepoint procedure, characterized by an extended stopping time $N$, satisfies a slightly strengthened version of (3), namely that

$$\mathbb{P}_{z,\theta,\Sigma}(N \leq z) + g(r;N) + 100rp^2\log_2^3(4p)(p\beta^{-2}\vee 1)e^{-r\beta^2/(8s\log_2(2p))} \leq \frac{3}{4}\alpha. \quad (5)$$

for some $r \geq 1$. Under (5), Theorems 1–3 hold with the same choices of input parameters. Moreover, the ocd′ base procedure satisfies the conclusion of Proposition 5, that is, there exists a universal constant $C' > 0$ such that (5) holds for $r \geq r_1 = r_1(C')$ in (4), provided that we use the modified input $\tilde{a} = \sqrt{2\log\{32p^2\gamma\log_2(2p)\}}$.

Generalizing these ideas further, now consider the model where $X_1,\ldots,X_z,X_{z+1}-\theta,X_{z+2}-\theta,\ldots$ are independent, each having sub-exponential components with variance parameter 1 and rate parameter $A > 0$. In this setting, provided the base procedure satisfies (5) for some $r \geq 1$ and $\vartheta \leq \sqrt{2A^2\log_2(2p)}$, Theorems 1–3 hold when we redefine $a := C\max\{\sqrt{\log(rp/\alpha)}, \frac{1}{A}\log(rp/\alpha)\}$ and $d_1 := \max\Big\{\sqrt{\frac{5r\beta^2}{9s\log_2(2p)}}, \frac{5r\beta^2}{9As\log_2(2p)}\Big\}$. Furthermore, with the modified input $\tilde{a} = \sqrt{2\log\{32p^2\gamma\log_2(2p)\}} \vee \frac{2}{A}\log\{32p^2\gamma\log_2(2p)\}$, the ocd′ base procedure satisfies the conclusion of Proposition 5 for

$$r \geq \frac{C's\log_2(2p)\log^2\{p\gamma\alpha^{-1}(\beta^{-2}\vee 1)\}\max\{1,A^{-2}\log(p\gamma)\}}{\beta^2} + 2,$$

where $C' > 0$ is a universal constant.

The claims made in the previous two paragraphs are justified in Section S4 of the supplementary materials. These results confirm the flexible scope of the ocd_CI methodology beyond the original Gaussian setting, at least as far as sub-exponential tails are concerned. Where data may exhibit heavier tails than this, clipping (truncation) and quantile transformations may represent viable ways to proceed, though further research is needed to confirm the theoretical validity of such approaches.

### 3.5. Confidence Interval Construction with Unknown Signal Size

In some settings, an experienced practitioner may have a reasonable idea of the magnitude $\vartheta$ of the $\ell_2$-norm of the vector of mean change that would be of interest to them, and this would facilitate a choice of a lower bound $\beta$ for $\vartheta$ in Algorithm 1. However, it is also worth considering the effect of this choice, and the extent to which its impact can be mitigated.

We first remark that the coverage probability guarantee for the ocd_CI interval in Corollary 6 remains valid for any (arbitrarily loose) lower bound $\beta$ on $\vartheta$. The only issue is in terms of power: if $\beta$ is chosen to be too small, then both the average detection delay and the high-probability bound on the length of the confidence interval may be inflated. In the remainder of this section, then, we describe a simple modification to Algorithm 1 that permits a loose lower bound $\beta$ to be employed that retains

coverage validity with only a logarithmic effect on the high-probability bound on the length of the confidence interval. The only other price we pay is that the computational cost increases as $\beta$ decreases, as we describe below.

Our change to Algorithm 1 is as follows: we replace the definition of $b_{\min}$ by setting

$$b_{\min} = \frac{\beta}{\sqrt{2^{\lfloor \log_2(2p) \rfloor} \log_2(2p)}} \wedge \frac{1}{\sqrt{2}},$$

set $M = \lceil 2\log_2(1/b_{\min}) \rceil$ and define $\mathcal{B} = \{\pm 2^{m/2} b_{\min} : m \in [M]\}$ in both the ocd′ base procedure and in Algorithm 1. The rest of algorithm remains as previously stated. Thus, if we choose a conservative (very small) $\beta$, then the effect of the modification is to increase the number of scales on which we search for a change, so that the largest element of $\mathcal{B}$ is of order 1. In order to state our theoretical results for this modified algorithm, first define $b_{\text{opt}} := \max\left\{b \in \mathcal{B} \cap (0,\infty) : b \le \frac{\vartheta}{\sqrt{s\log_2(2p)}}\right\}$, which satisfies $b_{\text{opt}} \ge \frac{\vartheta}{\sqrt{2s\log_2(2p)}} \wedge 1$. Under the same assumptions as in Proposition 5, and modifying the inputs to $(X_t)_{t\in\mathbb{N}}$, $\beta > 0$, $\tilde{a} = \sqrt{2\log(16p^2\gamma M)}$, $T^{\text{diag}} = \log\big(16p\gamma(M+1)\big)$ and $T^{\text{off}} = 8\log(16p\gamma M)$, it can be shown using very similar arguments to those in the proof of Proposition 5 that there exists a universal constant $C' > 0$ such that with $r \ge C' b_{\text{opt}}^{-2} \log\big(p\gamma\alpha^{-1}(\beta^{-2} \vee 1)\big) =: r_1$, the output $N$ satisfies

$$\mathbb{P}_{z,\theta,\Sigma}(N \le z) + g(r;N) + 4rp^2(M+1)^2 e^{-rb_{\text{opt}}^2/8} \le \frac{3}{4}\alpha.$$

With this in place, we can derive Corollary 7, which is the analog of Corollary 6 for our modified algorithm.

*Corollary 7.* Fix $\alpha \in (0,1), \gamma > 0$. Assume that $\theta$ has an effective sparsity of $s := s(\theta) \ge 2$, that $\vartheta \ge \beta > 0$ and that $z \le 2\alpha\gamma$. Let $(X_t)_{t\in\mathbb{N}}$, $\beta > 0$, $\tilde{a} = \sqrt{2\log(16p^2\gamma M)}$, $T^{\text{diag}} = \log\big(16p\gamma(M+1)\big)$ and $T^{\text{off}} = 8\log(16p\gamma M)$ be the inputs of the modified ocd′ procedure. Let $d_1 = \sqrt{5C'\log\big(p\gamma\alpha^{-1}(\beta^{-2}\vee 1)\big)/9}$ and $d_2 = 4d_1^2$. Then the following statements hold:

(a) With extra inputs CP = ocd′, $a \ge 0$, and $\ell \ge 0$ for the modified Algorithm 1, the output confidence interval $\mathcal{C}$ and the support estimate $\hat{\mathcal{S}}$ satisfy $\mathbb{P}_{z,\theta,\Sigma}(z \in \mathcal{C}) \ge 1-\alpha$ and $\mathbb{P}_{z,\theta,\Sigma}(\hat{\mathcal{S}} \subseteq \mathcal{S}_\beta) \ge 1-\alpha$.

(b) There exists a universal constant $C > 0$ such that, with extra inputs CP = modified ocd′, $a = C\sqrt{\log\big(p\gamma\alpha^{-1}(\beta^{-2}\vee 1)\big)}$, and $\ell \ge 80 r_1$ for the modified Algorithm 1, the length $L$ of the output confidence interval $\mathcal{C}$ and the support estimate satisfy

$$\mathbb{P}_{z,\theta,\Sigma}\left(L > 8C' \max\left\{\frac{2s\log_2(2p)}{\vartheta^2}, 1\right\} \log\big(p\gamma\alpha^{-1}(\beta^{-2}\vee 1)\big)\right) \le \alpha \tag{6}$$

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

The main difference between Corollaries 7 and 6 concerns the high-probability guarantees on the length of the confidence interval. Ignoring logarithmic factors, with high probability the length of the confidence interval in the modified algorithm is at most of order $(s/\vartheta^2) \vee 1$, whereas for the original algorithm it was of order $(s/\beta^2) \vee 1$. Thus, the modified algorithm has the significant advantage of enabling a conservative choice of $\beta$ with only a logarithmic effect on the length guarantee relative to an oracle procedure with knowledge of $\|\theta\|_2$. The computational complexity per new observation and the storage requirements of this modified algorithm are $O\Big(p^2\big(\log(ep) + \log(1/\beta)\big)\Big)$, so the order of magnitude is increased relative to the original algorithm only in an asymptotic regime where $\beta$ is small by comparison with $1/p^K$ for every $K > 0$. Moreover, the modified algorithm still does not require storage of historical data and the computational time per new observation after observing $n$ observations does not increase with $n$. Nevertheless, since the computational complexity now depends on $\beta$, the modified algorithm does not strictly satisfy our definition of an online algorithm given in introduction.

## 4. Numerical Studies

In this section, we study the empirical performance of the ocd_CI algorithm. Throughout this section, by default, the ocd_CI algorithm is used in conjunction with the recommended base online changepoint detection procedure CP = ocd.

### 4.1. Tuning Parameters

Chen, Wang, and Samworth (2022) found that the theoretical choices of thresholds $T^{\text{diag}}$ and $T^{\text{off}}$ 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 their Section 4.1. Likewise, as in Chen, Wang, and Samworth (2022), we take $a = \tilde{a} = \sqrt{2\log p}$ in our simulations.

For $d_1$ and $d_2$, as suggested by our theory, we take $d_2 = 4d_1^2$, and take $d_1$ to be of the form $d_1 = c\sqrt{\log(p/\alpha)}$. Here, we tune the parameter $c > 0$ through Monte Carlo simulation, as we now describe. We considered the parameter settings $p \in \{100, 500\}$, $s \in \{2, \lfloor\sqrt{p}\rfloor, p\}$, $\vartheta \in \{2, 1, 1/2\}$, $\Sigma = I_p$, $\alpha = 0.05$, $\beta \in \{2\vartheta, \vartheta, \vartheta/2\}$, $\gamma = 30{,}000$ and $z = 500$. Then, with $\theta$ generated as $\vartheta U$, 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. Figure 1 displays a subset of the results (the omitted curves were qualitatively similar). On this basis, 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 two paragraphs, in combination with Algorithms 1 and S1, 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 $\beta$, which represents a lower bound on the Euclidean norm of the vector of mean change. Fortunately, this description makes $\beta$ easily interpretable by practitioners. In cases where an informed

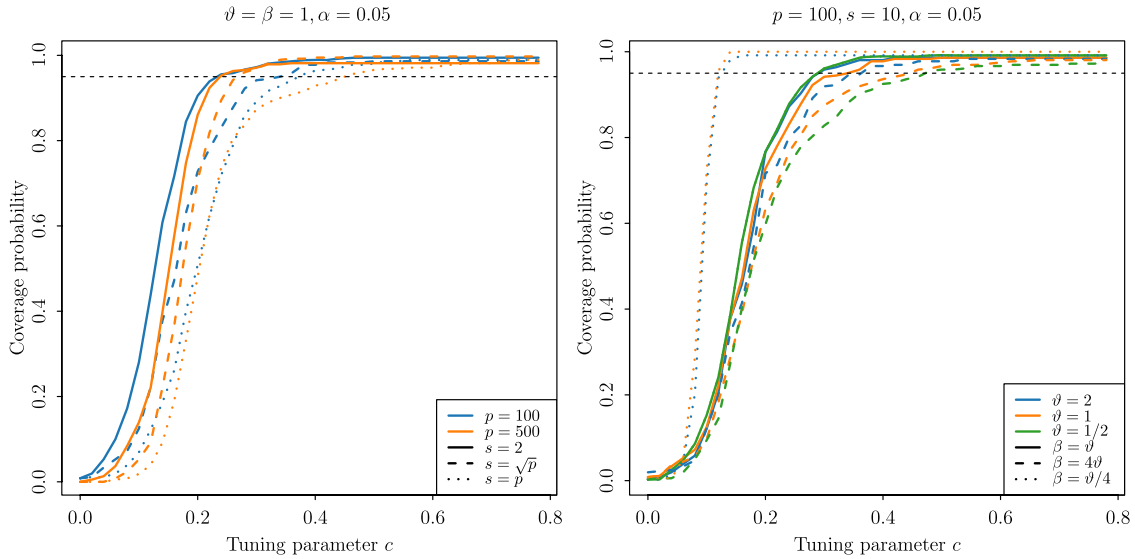


**Figure 1.** Coverage probabilities of the `ocd_CI` confidence interval as the parameter $c$, involved in the choice of tuning parameter $d_1$, varies.

default choice is not available, practitioners can make a conservative (very small) choice and use an increased grid of scales to with only a small inflation in the confidence interval length guarantee and computational cost; see Section 3.5.

### 4.2. Coverage Probability and Interval Length

In Table 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 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 $\beta$ 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 underestimate $\theta^j$ 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 $\beta$ 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 1 is to compare the average confidence interval lengths with the corresponding average detection delays. Corollary 6(b), as well as Chen, Wang, and Samworth (2022, Theorem 4), indicates that both of these quantities are of order $(s/\beta^2) \vee 1$, up to polylogarithmic factors in $p$ and $\gamma$, 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 2–3 times longer on average, and in all cases considered was less than seven times longer on average. Moreover, we can also observe that the confidence interval length increases with $s$ and decreases with $\beta$, as anticipated by our theory.

For comparison, we also present the corresponding coverage probabilities and average lengths of confidence intervals obtained using an offline procedure as described in the introduction. More precisely, after the `ocd` algorithm has declared a change, we treat the data up to the stopping time as an offline dataset, and apply the `inspect` algorithm (Wang and Samworth 2018), followed by the one-step refinement of Kaul et al. (2021), to construct an estimate, $\hat{z}^{\mathrm{KFJS}}$, of the changepoint location. As recommended by Kaul et al. (2021), we obtain an estimator $\hat{\vartheta}^{\mathrm{KFJS}}$ of $\vartheta$ using the $\ell_2$-norm of the soft-thresholded difference in empirical mean vectors before and after $\hat{z}^{\mathrm{KFJS}}$, with the soft-thresholding parameter chosen via the Bayesian Information Criterion. The final confidence interval is then of the form $[\hat{z}^{\mathrm{KFJS}} - q^{\alpha/2}/(\hat{\vartheta}^{\mathrm{KFJS}})^2, \hat{z}^{\mathrm{KFJS}} + q^{\alpha/2}/(\hat{\vartheta}^{\mathrm{KFJS}})^2]$, where $q^{\alpha/2}$ is the $1 - \alpha/2$ quantile of the distribution of the (almost surely unique) maximizer of a two-sided Brownian motion with a triangular drift as given by Kaul et al. (2021, Theorem 3.1). In particular, we have $q^{0.025} = 11.03$. The last two columns of Table 1 reveal that both the coverage probabilities and confidence interval lengths from this procedure are disappointing and not competitive with those of the `ocd_CI` algorithm. There are two main reasons for this: first, the nature of the online problem means that the changepoint is often located near the right-hand end of the dataset up to the stopping time; on the other hand, the theoretical guarantees of Kaul et al. (2021) are obtained under an asymptotic setting where the fraction of data either side of the change is bounded away from zero. Thus, the estimated changepoint from the one-step refinement is often quite poor. Moreover, the estimated magnitude of change, $\hat{\vartheta}^{\mathrm{KFJS}}$, is often a significant underestimate of $\vartheta$ due to the soft-thresholding operation, and this can lead to substantially inflated confidence interval lengths. We emphasize that the Kaul et al. (2021) procedure was not designed for use in this online setting, but we nevertheless present these results to illustrate the fact that the naive application of offline methods in sequential problems may fail badly.

While Table 1 covers the most basic setting for our methodology, our theory in Section 3 applies equally well to data with spatial dependence across different coordinates. To assess whether

**Table 1.** Estimated coverage and average length of the `ocd_CI` confidence interval and average detection delay over 2000 repetitions, with standard errors in brackets.

| | | | | `ocd_CI` | | | `Kaul_et_al` | |
|---|---|---|---|---|---|---|---|---|
| $p$ | $s$ | $\vartheta$ | $\beta$ | Delay | Coverage (%) | CI Length | Coverage (%) | CI Length |
| 100 | 2 | 2 | 4 | $9.8_{(0.1)}$ | $96.2_{(0.4)}$ | $20.1_{(0.7)}$ | $83.2_{(0.8)}$ | $732.7_{(9.6)}$ |
| 100 | 2 | 2 | 2 | $12.6_{(0.1)}$ | $97.0_{(0.4)}$ | $33.7_{(0.7)}$ | $82.5_{(0.8)}$ | $474.9_{(11.0)}$ |
| 100 | 2 | 2 | 1 | $14.1_{(0.1)}$ | $97.9_{(0.3)}$ | $80.8_{(1.0)}$ | $83.5_{(0.8)}$ | $341.4_{(10.4)}$ |
| 100 | 2 | 1 | 2 | $34.2_{(0.3)}$ | $95.8_{(0.4)}$ | $66.1_{(1.0)}$ | $76.6_{(0.9)}$ | $399.3_{(10.5)}$ |
| 100 | 2 | 1 | 1 | $44.2_{(0.3)}$ | $97.5_{(0.4)}$ | $122.0_{(1.4)}$ | $80.5_{(0.9)}$ | $123.8_{(6.5)}$ |
| 100 | 2 | 1 | 0.5 | $52.0_{(0.4)}$ | $97.4_{(0.4)}$ | $309.1_{(2.0)}$ | $81.5_{(0.9)}$ | $90.8_{(5.4)}$ |
| 100 | 10 | 2 | 4 | $14.7_{(0.1)}$ | $96.0_{(0.4)}$ | $32.5_{(0.8)}$ | $80.2_{(0.9)}$ | $636.4_{(10.5)}$ |
| 100 | 10 | 2 | 2 | $15.7_{(0.1)}$ | $97.4_{(0.4)}$ | $38.4_{(0.8)}$ | $77.4_{(0.9)}$ | $537.9_{(10.9)}$ |
| 100 | 10 | 2 | 1 | $15.9_{(0.1)}$ | $97.0_{(0.4)}$ | $80.2_{(1.1)}$ | $80.8_{(0.9)}$ | $542.4_{(10.9)}$ |
| 100 | 10 | 1 | 2 | $52.6_{(0.5)}$ | $96.2_{(0.4)}$ | $114.0_{(1.5)}$ | $75.8_{(1)}$ | $342.4_{(10.1)}$ |
| 100 | 10 | 1 | 1 | $56.9_{(0.4)}$ | $97.1_{(0.4)}$ | $142.5_{(1.8)}$ | $73.9_{(1)}$ | $262.6_{(9.1)}$ |
| 100 | 10 | 1 | 0.5 | $60.2_{(0.4)}$ | $98.2_{(0.3)}$ | $301.1_{(1.6)}$ | $75.9_{(1)}$ | $248.3_{(8.9)}$ |
| 100 | 100 | 2 | 4 | $27.2_{(0.2)}$ | $96.1_{(0.4)}$ | $77.6_{(0.9)}$ | $68.2_{(1.0)}$ | $533.9_{(10.7)}$ |
| 100 | 100 | 2 | 2 | $27.7_{(0.2)}$ | $96.0_{(0.4)}$ | $81.8_{(1.0)}$ | $71.3_{(1.0)}$ | $537.7_{(10.8)}$ |
| 100 | 100 | 2 | 1 | $28.2_{(0.2)}$ | $97.5_{(0.3)}$ | $99.4_{(1.3)}$ | $71.8_{(1.0)}$ | $556.0_{(10.7)}$ |
| 100 | 100 | 1 | 2 | $100.7_{(0.8)}$ | $94.7_{(0.5)}$ | $292.8_{(3.5)}$ | $87.7_{(0.7)}$ | $850.5_{(9.5)}$ |
| 100 | 100 | 1 | 1 | $100.5_{(0.9)}$ | $96.3_{(0.4)}$ | $296.0_{(3.4)}$ | $88.0_{(0.7)}$ | $863.7_{(9.3)}$ |
| 100 | 100 | 1 | 0.5 | $103.2_{(0.8)}$ | $97.3_{(0.4)}$ | $365.9_{(2.8)}$ | $89.3_{(0.7)}$ | $876.8_{(9.1)}$ |
| 500 | 2 | 2 | 4 | $11.3_{(0.1)}$ | $97.2_{(0.4)}$ | $23.1_{(0.7)}$ | $92.0_{(0.6)}$ | $958.7_{(4.3)}$ |
| 500 | 2 | 2 | 2 | $15.8_{(0.1)}$ | $97.7_{(0.3)}$ | $45.2_{(0.9)}$ | $83.3_{(0.8)}$ | $806.4_{(8.7)}$ |
| 500 | 2 | 2 | 1 | $17.7_{(0.1)}$ | $97.5_{(0.4)}$ | $117.3_{(1.0)}$ | $79.9_{(0.9)}$ | $624.9_{(10.7)}$ |
| 500 | 2 | 1 | 2 | $41.5_{(0.3)}$ | $97.3_{(0.4)}$ | $81.8_{(1.2)}$ | $80.0_{(0.9)}$ | $774.9_{(9.4)}$ |
| 500 | 2 | 1 | 1 | $55.0_{(0.4)}$ | $96.8_{(0.4)}$ | $168.9_{(1.5)}$ | $73.0_{(1)}$ | $275.9_{(9.4)}$ |
| 500 | 2 | 1 | 0.5 | $64.6_{(0.5)}$ | $98.1_{(0.3)}$ | $445.0_{(1.7)}$ | $75.6_{(1)}$ | $186.1_{(8.0)}$ |
| 500 | 22 | 2 | 4 | $23.6_{(0.2)}$ | $96.3_{(0.4)}$ | $55.4_{(1.0)}$ | $87.0_{(0.8)}$ | $884.9_{(7.3)}$ |
| 500 | 22 | 2 | 2 | $25.0_{(0.2)}$ | $97.0_{(0.4)}$ | $60.3_{(0.8)}$ | $85.5_{(0.8)}$ | $864.2_{(7.8)}$ |
| 500 | 22 | 2 | 1 | $25.5_{(0.2)}$ | $98.1_{(0.3)}$ | $119.7_{(0.8)}$ | $83.6_{(0.8)}$ | $823.0_{(8.6)}$ |
| 500 | 22 | 1 | 2 | $88.1_{(0.7)}$ | $97.0_{(0.4)}$ | $203.5_{(2.1)}$ | $77.2_{(0.9)}$ | $645.0_{(11.0)}$ |
| 500 | 22 | 1 | 1 | $91.9_{(0.6)}$ | $97.8_{(0.3)}$ | $229.7_{(2.2)}$ | $76.2_{(1)}$ | $562.8_{(11.1)}$ |
| 500 | 22 | 1 | 0.5 | $94.9_{(0.6)}$ | $98.3_{(0.3)}$ | $462.8_{(1.4)}$ | $75.5_{(1)}$ | $538.3_{(11.2)}$ |
| 500 | 500 | 2 | 4 | $79.8_{(0.6)}$ | $95.0_{(0.5)}$ | $238.9_{(2.7)}$ | $88.5_{(0.7)}$ | $913.0_{(8.0)}$ |
| 500 | 500 | 2 | 2 | $80.3_{(0.6)}$ | $95.8_{(0.4)}$ | $245.7_{(2.6)}$ | $90.3_{(0.7)}$ | $928.8_{(7.7)}$ |
| 500 | 500 | 2 | 1 | $80.9_{(0.6)}$ | $97.5_{(0.4)}$ | $250.2_{(2.5)}$ | $90.6_{(0.7)}$ | $928.3_{(7.7)}$ |
| 500 | 500 | 1 | 2 | $290.5_{(2.3)}$ | $94.5_{(0.5)}$ | $819.7_{(7.9)}$ | $95.2_{(0.5)}$ | $1189.4_{(7.3)}$ |
| 500 | 500 | 1 | 1 | $291.4_{(2.3)}$ | $95.2_{(0.5)}$ | $831.1_{(7.5)}$ | $94.3_{(0.5)}$ | $1204.9_{(7.0)}$ |
| 500 | 500 | 1 | 0.5 | $297.3_{(2.3)}$ | $98.1_{(0.3)}$ | $875.0_{(6.7)}$ | $94.6_{(0.5)}$ | $1207.4_{(6.8)}$ |

NOTE: Other parameters: $\gamma = 30{,}000$, $z = 1000$, $\Sigma = I_p$, $\alpha = 0.05$, $a = \tilde{a} = \sqrt{2\log p}$, $c = 0.5$, $d_1 = c\sqrt{\log(p/\alpha)}$, $d_2 = 4d_1^2$. For comparison, we also present the corresponding estimated coverage probabilities and average lengths of the procedure based on Kaul et al. (2021), as described in Section 4.2.

this theory carries over to empirical performance, Table S1 in the supplementary materials presents corresponding coverage probabilities and lengths for the `ocd_CI` procedure with the cross-sectional covariance matrix $\Sigma = (\Sigma_{jk})_{j,k\in[p]}$ taken to be Toeplitz with parameter $\rho \in \{0.5, 0.75\}$; in other words, $\Sigma_{jk} = \rho^{|j-k|}$. The results are again encouraging: the coverage remains perfectly satisfactory in all settings considered, and moreover, the lengths of the confidence intervals are very similar to those in Table 1.

### 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 2, 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 $p = 100$, $s \in \{5, 50\}$, $\vartheta \in \{1, 2\}$, $\Sigma = I_p$, 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 $a = \tilde{a} = \sqrt{2\log p}$, $\alpha = 0.05$, $d_1 = \sqrt{2\log(p/\alpha)}$, $\beta = \vartheta$, and, motivated by Corollary 6, took an additional $\ell = \lceil 2s\beta^{-2}\log_2(2p)\log p\rceil$ post-declaration observations in constructing the support estimates. The results reported in Table 2 provide empirical

**Table 2.** Estimated support recovery probabilities (with standard errors in brackets).

| $s$ | $\vartheta$ | 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 | $100.0_{(0.0)}$ | $100.0_{(0.0)}$ |

NOTE: Other parameters: $p = 100$, $\Sigma = I_p$, $\alpha = 0.05$, $a = \tilde{a} = \sqrt{2\log p}$, $d_1 = \sqrt{2\log(p/\alpha)}$, $\beta = \vartheta$, and with an additional $\ell = \lceil 2s\log_2(2p)\log(p)\beta^{-2}\rceil$ post-declaration observations.

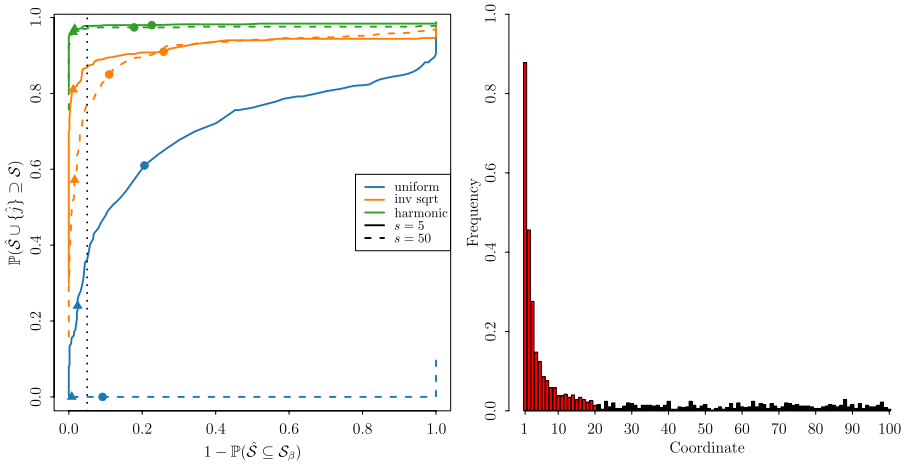


**Figure 2.** Support recovery properties of `ocd_CI`. In the left panel, we plot ROC curves for three different signal shapes and for sparsity levels $s \in \{5, 50\}$. The triangles and circles correspond to points on the curves with $d_1 = \sqrt{2\log(p/\alpha)}$ (with $\alpha = 0.05$), and $d_1 = \sqrt{2\log p}$, 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: $p = 100$, $\Sigma = I_p$, $\beta = \vartheta = 2$, $\ell = 0$, $a = \tilde{a} = \sqrt{2\log p}$.

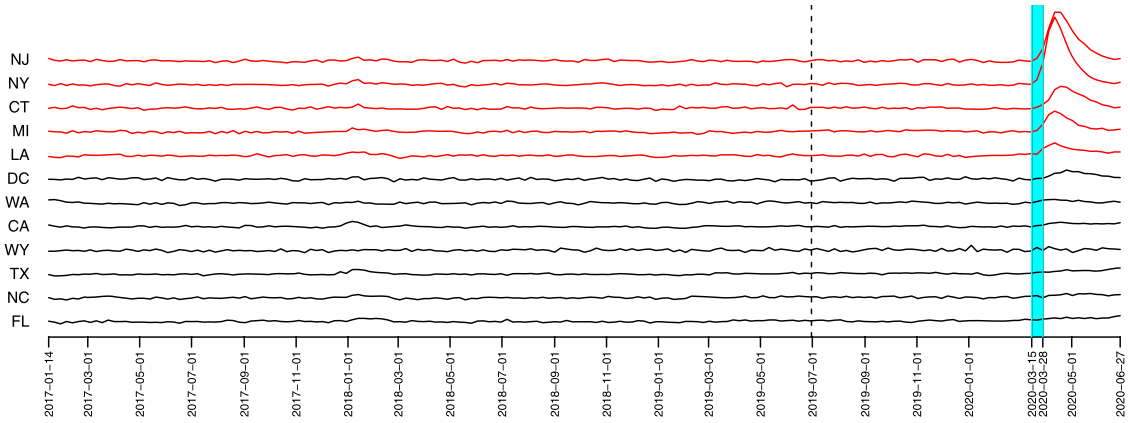


**Figure 3.** Standardized, transformed weekly excess death data from 12 states (including Washington, D.C.). The monitoring period starts from 30 June 2019 (dashed line). The data from the states in the support estimate are shown in red. The confidence interval [March 8, 2020, March 28, 2020] is shown in the light blue shaded region.

confirmation of the support recovery properties claimed in Corollary 6.

Finally in this section, we consider the extent to which the additional $\ell$ observations are necessary in practice to provide satisfactory support recovery. In the left panel of Figure 2, we plot Receiver Operating Characteristic (ROC) curves to study the estimated support recovery probabilities with $\ell = 0$ as a function of the input parameter $d_1$, which can be thought of as controlling the tradeoff 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 Corollary 6(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.

### 4.4. US Covid-19 Data Example

In this section, we apply `ocd_CI` to a dataset of weekly deaths in the United States between January 2017 and June 2020 (available at: *https://www.cdc.gov/nchs/nvss/vsrr/covid19/excess_deaths.htm*). The data up to 29 June 2019 are treated as our training data. An obvious discrepancy between underlying dynamics of these weekly deaths and the conditions assumed in our theory in Section 3 is temporal dependence, particularly induced by seasonal and weather effects. Although we can never hope to remove this dependence entirely, we seek to mitigate its impact by pre-processing the data as follows: for each of the 50 states, as well as Washington, D.C. ($p = 51$), we first estimate the "seasonal death curve," that is, 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 stabilize 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 standardize the transformed weekly excess deaths using the mean and standard deviation of the transformed data over the training period. The standardized, transformed data are plotted in Figure 3 for 12 states. When applying `ocd_CI` to these data, we take $a = \tilde{a} = \sqrt{2\log p}$, $T^{\text{diag}} = \log\{16p\gamma\log_2(4p)\}$, $T^{\text{off}} = 8\log\{16p\gamma\log_2(2p)\}$, $d_1 = 0.5\sqrt{\log(p/\alpha)}$ 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 standardizing as before, due to the lack of available data prior to 2017), it identifies a subtler change on the week ending January 6, 2018, with a confidence interval of [December 17, 2017, January 6, 2018]. This corresponds to a bad influenza season at the end of 2017 (see, *https://www.cdc.gov/flu/about/season/flu-season-2017-2018.htm*). Despite the natural interpretation of these findings, we recognize that the model in Section 3 under which we proved our theoretical results cannot capture the full complexity of the temporal dependence in this dataset even after our pre-processing transformations. A complete theoretical analysis of the performance of `ocd_CI` in time-dependent settings is challenging and beyond the scope of the current work; in practical applications, we advise careful modeling of this dependence to facilitate the construction of appropriate residuals for which the main effects of this dependence have been removed.

## Supplementary Materials

This contains the proofs of our main theoretical results and auxiliary results, as well as pseudocode for the `ocd` and `ocd′` base procedures, extensions of our results to sub-Gaussian and sub-exponential settings and additional simulation results.

## Acknowledgments

We are grateful to the Editor, the Associate Editor and two anonymous referees for their constructive feedback, which helped to improve the article.

## Disclosure Statement

The authors report there are no competing interests to declare.

## Funding

The research of the first and second authors was supported by Engineering and Physical Sciences Research Council (EPSRC) grant EP/T02772X/1 and that of the third author was supported by EPSRC grants EP/P031447/1 and EP/N031938/1, as well as European Research Council Advanced Grant 101019498.

## ORCID

Richard J. Samworth https://orcid.org/0000-0003-2426-4679

## References

Chan, H. P., and Walther, G. (2015), "Optimal Detection of Multi-Sample Aligned Sparse Signals," *Annals of Statistics*, 43, 1865–1895. [1]

Chen, Y., Wang, T., and Samworth, R. J. (2022), "High-Dimensional, Multiscale Online Changepoint Detection," *Journal of the Royal Statistical Society*, Series B, 84, 234–266. [1,2,4,7,8]

Cho, H. (2016), "Change-Point Detection in Panel Data via Double CUSUM Statistic," *Electronic Journal of Statistics*, 10, 2000–2038. [1]

Cho, H., and Fryzlewicz, P. (2015), "Multiple-Change-Point Detection for High Dimensional Time Series via Sparsified Binary Segmentation," *Journal of the Royal Statistical Society*, Series B, 77, 475–507. [1]

Csörgő, M., and Horváth, L. (1997), *Limit Theorems in Change-Point Analysis*, New York: Wiley. [1]

Dette, H., and Gösmann, J. (2020), "A Likelihood Ratio Approach to Sequential Change Point Detection for a General Class of Parameters," *Journal of the American Statistical Association*, 115, 1361–1377. [1]

Duncan, A. J. (1952), *Quality Control and Industrial Statistics*, Chicago: Richard D. Irwin Professional Publishing Inc. [1]

Enikeeva, F., and Harchaoui, Z. (2019), "High-Dimensional Change-Point Detection Under Sparse Alternatives," *Annals of Statistics*, 47, 2051–2079. [1]

Follain, B., Wang, T., and Samworth, R. J. (2022), "High-Dimensional Changepoint Estimation With Heterogeneous Missingness," *Journal of the Royal Statistical Society*, Series B, 84, 1023–1055. [1]

Gösmann, J., Stoehr, C., Heiny, J., and Dette, H. (2020), "Sequential Change Point Detection in High Dimensional Time Series," arXiv preprint, arxiv:2006.00636. [1]

Horváth, L., and Rice, G. (2014), "Extensions of Some Classical Methods in Change Point Analysis," *TEST*, 23, 219–255. [1]

Janková, J., and van de Geer, S. (2015), "Confidence Intervals for High-dimensional Inverse Covariance Estimation," *Electronic Journal of Statistics*, 9, 1205–1229. [1]

Javanmard, A., and Montanari, A. (2014), "Confidence Intervals and Hypothesis Testing for High-Dimensional Regression," *Journal of Machine Learning Research*, 15, 2869–2909. [1]

Jirak, M. (2015), "Uniform Change Point Tests in High Dimension," *Annals of Statistics*, 43, 2451–2483. [1]

Kaul, A., Fotopoulos, S. B., Jandhyala, V. K., and Safikhani, A. (2021), "Inference on the Change Point Under a High Dimensional Sparse Mean Shift," *Electronic Journal of Statistics*, 15, 71–134. [1,8,9]

Kirch, C., and Stoehr, C. (2019), "Sequential Change Point Tests Based on $U$-statistics," arXiv preprint, arxiv:1912.08580. [1]

Liu, H., Gao, C., and Samworth, R. J. (2021), "Minimax Rates in Sparse, High-dimensional Changepoint Detection," *Annals of Statistics*, 49, 1081–1112. [1]

Londschien, M., Kovács, S., and Bühlmann, P. (2021), "Change-Point Detection for Graphical Models in the Presence of Missing Values," *Journal of Computational and Graphical Statistics*, 30, 1–12. [1]

Padilla, O. H. M., Yu, Y., Wang, D., and Rinaldo, A. (2021), "Optimal Nonparametric Multivariate Change Point Detection and Localization," *IEEE Transactions on Information Theory*, 68, 1922–1944. [1]

Page, E. S. (1954), "Continuous Inspection Schemes," *Biometrika*, 41, 100–115. [1]

Rinaldo, A., Wang, D., Wen, Q., Willett, R., and Yu, Y. (2021), "Localizing Changes in High-dimensional Regression Models," *Proceedings of Machine Learning Research*, pp. 2089–2097. [1]

Soh, Y. S., and Chandrasekaran, V. (2017), "High-Dimensional Change-Point Estimation: Combining Filtering With Convex Optimization," *Applied and Computational Harmonic Analysis*, 43, 122–147. [1]

Tartakovsky, A., Nikiforov, I., and Basseville, M. (2014), *Sequential Analysis: Hypothesis Testing and Changepoint Detection*, London: Chapman and Hall. [1]

van de Geer, S., Bühlmann, P., Ritov, Y., and Dezeure, R. (2014), "On Asymptotically Optimal Confidence Regions and Tests for High-Dimensional Models," *Annals of Statistics*, 42, 1166–1202. [1]

Wang, T., and Samworth, R. J. (2018), "High Dimensional Change Point Estimation via Sparse Projection," *Journal of the Royal Statistical Society*, Series B, 80, 57–83. [1,8]

Yu, Y., Bradic, J., and Samworth, R. J. (2021), "Confidence Intervals for High-Dimensional Cox Models," *Statistica Sinica*, 31, 243–267. [1]

Yu, Y., Padilla, O. H. M., Wang, D., and Rinaldo, A. (2021), "Optimal Network Online Change Point Localisation," arXiv preprint, arxiv:2021.05477. [1]

Zhang, C.-H., and Zhang, S. S. (2014), "Confidence Intervals for Low Dimensional Parameters in High Dimensional Linear Models," *Journal of the Royal Statistical Society*, Series B, 76, 217–242. [1]