



Generalized multiple change-point detection in the structure of multivariate, possibly high-dimensional, data sequences

Andreas Anastasiou¹ · Angelos Papanastasiou¹

Received: 1 June 2022 / Accepted: 24 May 2023 / Published online: 29 June 2023
© The Author(s) 2023

Abstract

The extensive emergence of big data techniques has led to an increasing interest in the development of change-point detection algorithms that can perform well in a multivariate, possibly high-dimensional setting. In the current paper, we propose a new method for the consistent estimation of the number and location of multiple generalized change-points in multivariate, possibly high-dimensional, noisy data sequences. The number of change-points is allowed to increase with the sample size and the dimensionality of the given data sequence. Having a number of univariate signals, which constitute the unknown multivariate signal, our algorithm can deal with general structural changes; we focus on changes in the mean vector of a multivariate piecewise-constant signal, as well as changes in the linear trend of any of the univariate component signals. Our proposed algorithm, labeled Multivariate Isolate–Detect (MID) allows for consistent change-point detection in the presence of frequent changes of possibly small magnitudes in a computationally fast way.

Keywords Change-point detection · High-dimensional setting · Piecewise-constant structure · Piecewise-linear structure

1 Introduction

Change-point detection algorithms have been actively developed and investigated from the scientific community. Their ability to segment data into smaller, homogeneous parts have helped researchers and practitioners to develop flexible statistical models that can adapt in non-stationary environments. Due to the natural data heterogeneity in many real problems, such algorithms have been applied in a wide range of application areas, such as bioinformatics (Picard et al. 2011; Hocking et al. 2013), cyber security (Siris and Papagalou 2004), or finance (Lavielle and Teyssiere 2007; Schröder and Fryzlewicz 2013). The advantages of detecting changes in the behaviour of data fall into two main categories; interpretation and forecasting. Interpretation comes naturally since the detected changes are usually connected with life events that took place near the estimation time. Associating the changes with such real-life phenomena can lead to a better under-

standing and quantification of the effect that these events had on the behaviour of the stochastic process. With respect to forecasting, the role of the final segment is important because it allows for a more accurate prediction of the future values of the data sequence at hand.

Based on whether we have full knowledge of the data to be analysed, change-point detection is split into two main categories; offline detection, where the data are already obtained, and online detection, in which the observations arrive sequentially at present. With respect to the dimensionality of the data, change-point detection can be further separated into algorithms that act only on univariate data and to those that are suitable for change-point detection in multivariate data sequences. In this paper, we focus on multivariate, possibly high-dimensional, offline settings; our aim is to estimate the number and locations of certain structural changes in the behaviour of given multivariate data. The model is

$$X_t = f_t + \epsilon_t, \quad t = 1, \dots, T, \quad (1)$$

where $X_t \in \mathbb{R}^{d \times 1}$ are the observed data and $f_t \in \mathbb{R}^{d \times 1}$ is the d -dimensional deterministic signal with structural changes at certain points. The signals that we treat in the current manuscript are those that changes appear in the mean structure or in the vector of the first order derivatives. The noise

✉ Andreas Anastasiou
anastasiou.andreas@ucy.ac.cy

Angelos Papanastasiou
aggelospapanastasiou905@gmail.com

¹ Department of Mathematics and Statistics, University of Cyprus, Nicosia, Cyprus

terms $\epsilon_t \in \mathbb{R}^{d \times 1}$ are random vectors with mean the zero vector and covariance matrix, Σ , which is positive definite and not necessarily diagonal. The true number, N , of the change-points, as well as their locations r_1, \dots, r_N , are unknown and our aim is to estimate them; N is free to grow with the sample size, T , and the dimensionality, d .

The initial purpose of change-point detection algorithms has been to detect a single change in the mean structure of a univariate signal under the setting of Gaussian noise, but much progress has since been made. Researchers have heavily focused on the detection of multiple change-points in the mean structure of a univariate data sequence. Towards this purpose, optimization-based methods have been developed, in which the estimated signal is chosen based on its fit to the data, penalized by a complexity rule. To solve the implied penalization problem, dynamic programming approaches, such as the Segment Neighborhood (SN) and Optimal Partitioning (OP) methods of Auger and Lawrence (1989) and Jackson et al. (2005), have been developed. Killeck et al. (2012) and Rigaiil (2015) introduce improvements over the classical OP and SN algorithms, respectively. In the context of regression problems, Frick et al. (2014) introduced the simultaneous multiscale change-point estimator (SMUCE) for change-point detection in exponential family regression. Apart from optimization based algorithms, a popular method in the literature is binary segmentation where changes are detected one at a time through an iterative binary splitting of the data. Recent variants of binary segmentation with improved performance are the Wild Binary Segmentation (WBS) of Fryzlewicz (2014) and its recently developed second version of Fryzlewicz (2020), the Narrowest-Over-Threshold (NOT) method of Baranowski et al. (2019), and the Seeded Binary Segmentation of Kovács et al. (2020). The Isolate–Detect (ID) algorithm has been developed in Anastasiou and Fryzlewicz (2022) to detect, one by one, structural changes in a data sequence. It is based on an isolation technique, which leads to very good accuracy on the estimated number and locations of the change-points, particularly in scenarios with many frequent change-points. Our proposed method is partly based on ID; therefore we elaborate on its important parts later. For a more thorough review of the literature on the detection of multiple change-points in the mean of univariate data sequences, see Cho and Kirch (2020) and Yu (2020). Apart from changes in the mean of a univariate data sequence, research has also been done for the detection of change-points under more complex scenarios, such as detection of changes in the slope for piecewise-linear models (Anastasiou and Fryzlewicz 2022; Baranowski et al. 2019; Fearnhead et al. 2019; Maeng and Fryzlewicz 2019; Tibshirani 2014) changes in the variance (Inclán and Tiao 1994), as well as for distributional changes under a non-parametric setting (Matteson and James 2014; Zou et al. 2014; Arlot et al. 2019).

Even though there is an extensive literature on change-point detection for univariate data sequences, the multivariate, possibly high-dimensional setting, which is the focus of this paper, has not been investigated in such degree. Working under the model in (1), Vert and Bleakley (2010) proposes a method for approximating the signal f_t as the solution of a convex optimization problem. In order to achieve this, the problem is first reformulated to a group LASSO one and then the group least-angle regression (LARS) procedure explained in Yuan and Lin (2006) is employed. Another interesting approach with very good behaviour has been introduced in Wang and Samworth (2018). The algorithm, called *inspect*, estimates the number and locations of the change-points in the mean structure of f_t as in (1). Firstly, *inspect* applies a cumulative sum (CUSUM) transformation to the original data matrix. Secondly, a projection direction of the transformed matrix is computed as its leading left singular vector, and, finally, a univariate change-point detection algorithm is applied to the projected series. It is among the many methods that employ CUSUM-type statistics for change-point detection in the multivariate setting. In general, methods that belong to this category, mainly either use CUSUM aggregations of the d component data sequences in order to test the obtained values against a threshold or to construct alternative test-statistics. For instance, Groen et al. (2013) examines the asymptotic behaviour of the maximum absolute and average CUSUM and gives finite-sample performance results. Focusing on testing for the existence of a change-point in the mean structure of the multivariate signal, Enikeeva and Harchaoui (2019) and Horváth and Hušková (2012) propose tests based on the ℓ_2 aggregation of the CUSUM statistics for each univariate component, while Jirak (2015) employs the ℓ_∞ aggregation of the aforementioned values. In Cho (2016), the Double CUSUM (DC) operator is introduced, which takes as input the ordered absolute CUSUM values of each individual component and performs a weighted ℓ_1 aggregation to construct the DC statistic which is then compared against a test criterion. Departing from the detection of changes in the mean structure of a multivariate signal, Cho and Fryzlewicz (2015) propose the Sparsified Binary Segmentation algorithm (SBS) for the detection of multiple change-points in the second-order structure of a multivariate data sequence. SBS is based on a first, “sparsifying” step which is used to exclude individual component data sequences from an ℓ_1 aggregation; a pre-defined threshold is used for the exclusion. The recent work of Anastasiou et al. (2022) introduces Cross-covariance isolate detect (CCID), which, motivated from the necessity of estimating changes in time-varying functional connectivity networks, detects multiple change-points in the second-order (cross-covariance or network) structure of multivariate, possibly high-dimensional time series. Ombao et al. (2005) investigates the application of smooth localized complex

exponentials (SLEX) waveforms, to the detection of changes in spectral characteristics of EEG data. Lavielle and Teysiere (2006) detects changes in the covariance structure of i.i.d. multivariate time series based on the minimization of a penalized Gaussian log likelihood, while Bücher et al. (2014) uses a test statistic based on sequential empirical copula processes to detect changes in the cross-covariance structure. For a survey of various offline change-point detection algorithms on multivariate time series see Truong et al. (2020).

In this paper, we propose a method called Multivariate Isolate–Detect (MID) for the consistent estimation of multiple change-points under the multivariate, possibly high-dimensional, structure of the model in (1). Our method builds on the foundations of the ID algorithm developed in Anastasiou and Fryzlewicz (2022); that is, we first isolate each true change-point within subintervals of the domain $[1, \dots, T]$ and then we proceed to detect them. Isolation enhances detection power, especially in frameworks with frequently occurring change-points. MID is explained in detail in Sect. 2; here, we only give a brief description of its important steps. The main idea is that for the observed data sequences $x_{t,j} \ t = 1, \dots, T, \ j = 1, \dots, d$, and with λ_T a positive constant, playing the role of an expansion step as in Anastasiou and Fryzlewicz (2022), our method first creates two ordered sets of $K = \lceil T/\lambda_T \rceil$ right- and left-expanding intervals. For $i = 1, \dots, K$, the i th right expanding interval is $R_i = [1, \min \{i\lambda_T, T\}]$ while the i th left-expanding interval is $L_i = [\max \{1, T - i\lambda_T + 1\}, T]$. We collect these intervals in the ordered set $S_{RL} = \{R_1, L_1, R_2, L_2, \dots, R_K, L_K\}$. The algorithm first acts on the interval $R_1 = [1, \lambda_T]$ by calculating, for every univariate component data sequence, the contrast function value for the Q possible candidates in this interval (details are given in Sect. 2). This process will return Q vectors $\mathbf{y}_j, j = 1, \dots, Q$ of length d each; for example, the elements of $\mathbf{y}_1 \in \mathbb{R}^d$ will be the contrast function values related to the first change-point candidate in R_1 , for each of the d component data sequences, the elements of $\mathbf{y}_2 \in \mathbb{R}^d$ will be the relevant values for the second candidate in R_1 , and so on. The next step is to apply to each \mathbf{y}_j a mean-dominant norm $L : \mathbb{R}^d \rightarrow \mathbb{R}$. To be more precise, since the contrast function values within each \mathbf{y}_j are all non-negative, in our case $L : (\mathbb{R}^d)^+ \rightarrow \mathbb{R}$. The definition of mean-dominant norms can be found in Sect. 2 of Carlstein (1988) and examples include

$$L_2 := L_2(\mathbf{y}_j) = \frac{1}{\sqrt{d}} \sqrt{\sum_{i=1}^d y_{j,i}^2}$$

$$L_\infty := L_\infty(\mathbf{y}_j) = \sup_{i=1, \dots, d} \{y_{j,i}\}. \tag{2}$$

Applying $L(\cdot)$ to each \mathbf{y}_j , will return a vector \mathbf{v} of length Q . We identify $\tilde{b}_{R_1} := \operatorname{argmax}_j \{v_j\}$. If $v_{\tilde{b}_{R_1}}$ exceeds a certain

threshold, denoted by $\zeta_{T,d}$, then \tilde{b}_{R_1} is taken as a change-point. If not, then the process tests the next interval in S_{RL} . Upon detection, the algorithm makes a new start from the end-point (respectively, start-point) of the right- (respectively, left-) expanding interval where the detection occurred. Upon correct choice of $\zeta_{T,d}$, MID ensures that we work on intervals with at most one change-point.

The rest of the paper is structured as follows. Section 2 describes our proposed methodology and its associated theory. The computational complexity of MID, useful variants of our algorithm and the choice of important parameter values are explained in Sect. 3. In Sect. 4, we perform a thorough simulation study to compare our algorithm with state-of-the-art methods. Furthermore, we show and discuss on the practical performance of MID in the cases of spatially dependent data (spatial independence does not need to be assumed for the theoretical results related to the consistency of the proposed method) and when the normality assumption for the error terms is violated. Sect. 5 illustrates the behaviour of MID on two examples of real data; the monthly percentage changes in the UK house price index over a period of twenty three years in twenty London Boroughs, and the daily number of new COVID-19 cases in the four constituent countries of the United Kingdom; England, Northern Ireland, Scotland, and Wales. In Sect. 6, we first discuss on how we can treat cases where the data exhibit temporal dependence, and then, we conclude the paper with general remarks and reflections on our proposed algorithm. The proof of Theorem 1 is given in the Appendix.

2 Methodology and theory

2.1 Methodology

We work under the model given in (1). The objective is to estimate both the number, N , and the locations r_1, \dots, r_N where the multivariate deterministic signal \mathbf{f}_t exhibits structural changes. We note that N can possibly grow with the sample size T and with the dimensionality d . In addition, a change-point does not necessarily appear in all univariate component signals; any level of sparsity is allowed. Before providing a full, step-by-step explanation of our algorithm, two simple examples are given to assist in ease of understanding. In Fig. 1, we have a three dimensional data sequence of length $T = 200$ with three change-points in the mean vector at locations $r_1 = 27, r_2 = 73$ and $r_3 = 165$. To be more precise,

$$f_{t,1} = \begin{cases} 0, & t = 1, \dots, 27 \\ 6, & t = 28, \dots, 165 \\ 0, & t = 166, \dots, 200 \end{cases},$$

Fig. 1 An example of a three dimensional data sequence that undergoes three changes in its mean structure at locations $r_1 = 27, r_2 = 73$ and $r_3 = 165$

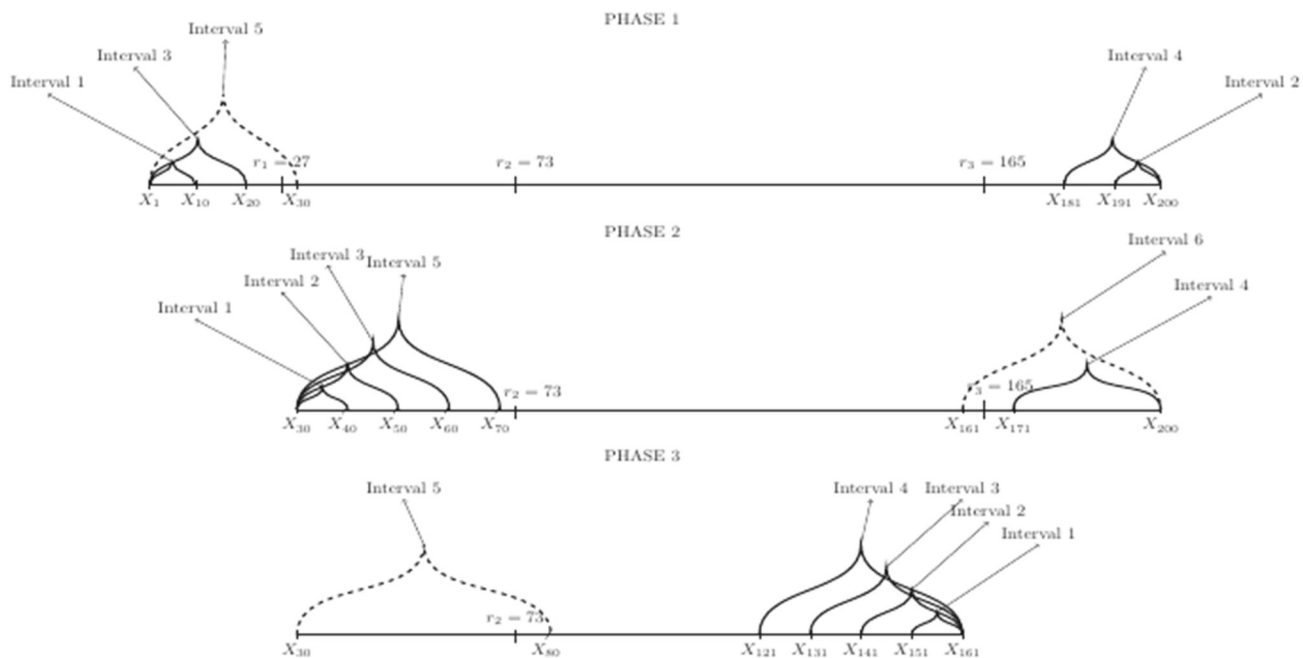
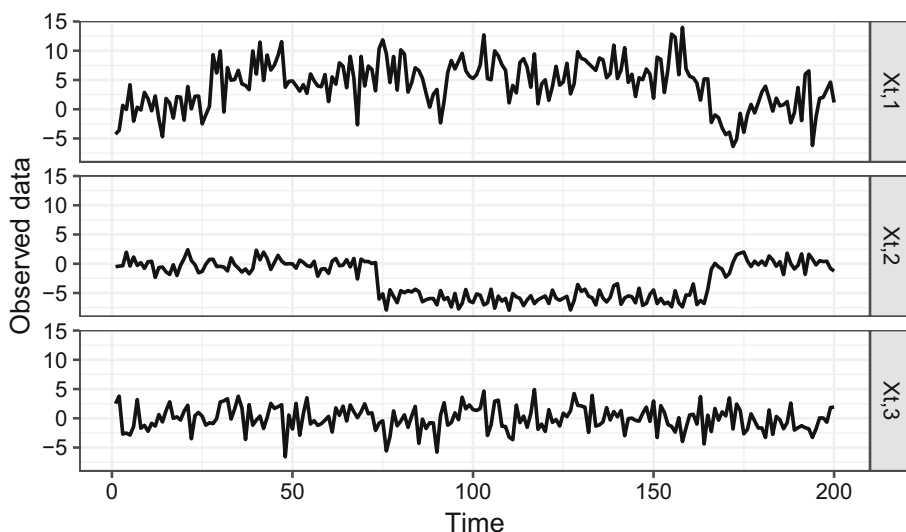


Fig. 2 An example with three change-points in the mean structure; $r_1 = 27, r_2 = 73$ and $r_3 = 165$. The dashed line is the interval in which the detection took place in each phase

$$f_{t,2} = \begin{cases} 0, & t = 1, \dots, 73 \\ -6, & t = 74, \dots, 165 \\ 0, & t = 166, \dots, 200 \end{cases}, \quad \Sigma_1 = \begin{pmatrix} 9 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 4 \end{pmatrix}.$$

$$f_{t,3} = 0, t = 1, \dots, 200. \tag{3}$$

From now on, $\mathcal{N}_d(\boldsymbol{\mu}, \Sigma)$ denotes the d -variate normal distribution with mean vector $\boldsymbol{\mu} \in \mathbb{R}^{d \times 1}$ and covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$. For the example, we take $\boldsymbol{\epsilon}_t \sim \mathcal{N}_3(\mathbf{0}, \Sigma_1)$, where

The component data sequences $X_{t,1}$ and $X_{t,2}$ share a common change-point at $t = 165$, while they also have their own change-points at $t = 27$ and $t = 73$, respectively. There are no change-points in $X_{t,3}$.

Let us denote by $r_0 = 0, r_{N+1} = T$. In this toy example, we take the expansion parameter $\lambda_T = 10$, while $\zeta_{T,d}$

is a well-chosen predefined threshold. The choice of the aforementioned parameters is discussed in detail in Sect. 3.3; special attention will be given to the dimensionality of the data sequence in order to make a robust threshold choice. Figure 2 shows the steps of MID until all change-points are detected. We will be referring to Phases 1, 2 and 3 involving five, six, and five intervals, respectively; these are clearly indicated in Fig. 2. These phases are only related to this specific example of three change-points; in cases with a different number of change-points we would have a different number of such phases. At the beginning of the detection process, we have that $s = 1$ and $e = T = 200$. As already mentioned in Sect. 1, the proposed algorithm acts both sequentially and interchangeably on subintervals of the full domain; this being $[1, \dots, 200]$ for this example. For a well-chosen threshold $\zeta_{T,d}$, then $r_1 = 27$ is the first change-point to be detected; this occurs in the interval $[1, 30]$ as shown in Phase 1 of Fig. 2. We now briefly explain how the detection occurred for r_1 . Let A be the 30×3 matrix, with each column being the first 30 observations (since we are working in the interval $[1, 30]$) of each of the three univariate data sequences. The next step is to compute the contrast function (in this specific case of piecewise-constant signals, the function is the absolute value of the widely used CUSUM statistic as given in (5)) values for each candidate point and for all three component data sequences. We end up with a matrix $B \in \mathbb{R}^{29 \times 3}$, with $B_{i,j}$ being the value of the contrast function for the i th data point of the j th data sequence when we work in the interval $[1, 30]$; the last point of the interval is not among the change-point candidates. Applying a mean-dominant norm to each row of B gives us a vector of length 29. Figure 3 (Detection 1), illustrates these values, when we employed the L_2 and the L_∞ mean-dominant norms as defined in (2). In Fig. 3, we see that for both employed norms, $t = 27$ has the highest value, which exceeds the predefined threshold value obtained as in Sect. 3.3. Therefore, $\hat{r}_1 = 27$ is assigned as the estimated location for r_1 . After the detection, s is updated as the end-point of the (right-expanding) interval where the detection occurred; therefore $s = 30$, and MID is, in Phase 2, applied to the interval $[30, 200]$. Then, $r_3 = 165$ gets detected at the sixth step of Phase 2 in the interval $[161, 200]$. After this second detection, MID proceeds to Phase 3, where it is applied to the interval $[30, 161]$ and r_2 gets isolated (for the first time) and detected in the interval $[30, 80]$ as shown in Figs. 2 and 3. In the end, MID is applied to the interval $[80, 161]$, where there will be no expanding interval that contains a point with an aggregated CUSUM value that surpasses the threshold $\zeta_{T,d}$; therefore, the process will terminate after scanning all the data.

In Fig. 4, we graphically provide an example of a three dimensional data sequence of length $T = 200$ with three change-points in the slope at locations $r_1 = 53, r_2 = 100$ and $r_3 = 124$. To be more precise,

$$f_{t,1} = \begin{cases} -t + 1, & t = 1, \dots, 53 \\ 2t - 158, & t = 54, \dots, 124 \\ -t + 214, & t = 125, \dots, 200 \end{cases},$$

$$f_{t,2} = \begin{cases} -t + 1, & t = 1, \dots, 100 \\ 2t - 299, & t = 101, \dots, 124 \\ -t + 73, & t = 125, \dots, 200 \end{cases},$$

$$f_{t,3} = t, t = 1, \dots, 200.$$

We take $\epsilon_t \sim \mathcal{N}_3(\mathbf{0}, \Sigma_2)$, with $\Sigma_2 = 49I_{3 \times 3}$, where for $k \in \mathbb{Z}^+$, the matrix $I_{k \times k}$ is the $k \times k$ identity matrix. The first two component data sequences have two change-points each; for $X_{t,1}$ at locations $t = 53$ and $t = 124$, while for $X_{t,2}$ at locations $t = 100$ and $t = 124$. There are no change-points in $X_{t,3}$.

After giving two examples of structures that MID can be employed to, our method can now be described in a general framework. Our proposed algorithm is based on the same isolation technique as that of the univariate change-point detection method ID and therefore, extensive details of how this isolation is achieved are avoided and can be found in Section 3.1 of Anastasiou and Fryzlewicz (2022). For the better understanding of MID, we provide its step-by-step outline through pseudocode, followed by a succinct narrative of the steps. For $K = \lceil T/\lambda_T \rceil$, let $c'_j = \min \{j\lambda_T, T\}$ and $c^l_j = \max \{1, T - j\lambda_T + 1\}$ for $j = 1, \dots, K$. For a generic interval $[s, e]$, define the sequences

$$R_{s,e} = [c'_{k_1}, c'_{k_1+1}, \dots, e], \quad L_{s,e} = [c^l_{k_2}, c^l_{k_2+1}, \dots, s], \tag{4}$$

where $k_1 := \operatorname{argmin}_{j \in \{1, 2, \dots, K\}} \{j\lambda_T > s\}$ and $k_2 := \operatorname{argmin}_{j \in \{1, 2, \dots, K\}} \{T - j\lambda_T + 1 < e\}$. We denote by

$$C_{s,e}^b(\mathbf{X}) = (C_{s,e}^b(\mathbf{X}^{(1)}), \dots, C_{s,e}^b(\mathbf{X}^{(d)}))$$

the vector in $(\mathbb{R}^+)^d$ that has the contrast function values for each one of the d component data sequences, $\mathbf{X}^{(i)}, i = 1, \dots, d$, at the location b when we work in the interval $[s, e]$. Then, denoting by $|A|$ the cardinality of any sequence A , by $A(j)$ its j th element, and for $L(\cdot)$ being any mean-dominant norm (as those in (2)) employed for the aggregation of the contrast function values, the pseudocode of the main function for the proposed algorithm is as below:

Pseudocode explaining the MID algorithm

```
function MID( $s, e, \lambda_T, \zeta_{T,d}, L$ )
  if  $e - s < 1$  then
    STOP
  else
    For  $j \in \{1, \dots, \lfloor R_{s,e} \rfloor\}$ , denote
```

$$[s_{2j-1}, e_{2j-1}] := [s, R_{s,e}(j)].$$

Fig. 3 CUSUM values in the relevant detection intervals explained in Fig. 2. The vertical dashed line indicates the time point with the highest value in the corresponding interval, while the horizontal dashed line is the optimal threshold value. On the left column you can see the results when the L_2 aggregation method was used, while for the right column we employed the L_∞ -based aggregation approach

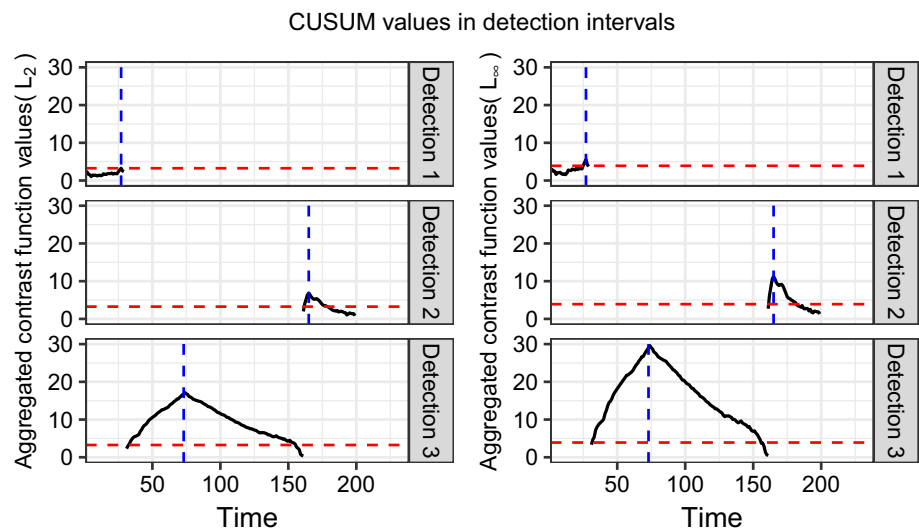
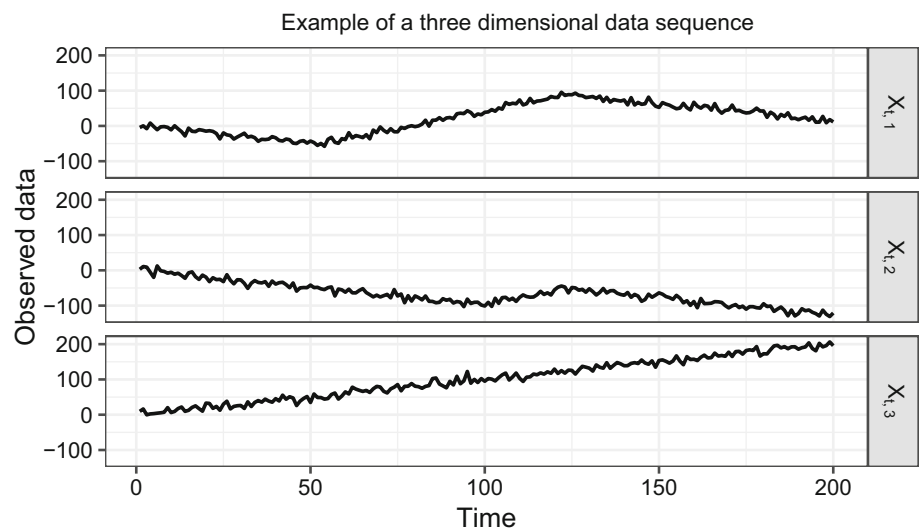


Fig. 4 An example of a three dimensional data sequence with piecewise-linear structure, that undergoes three changes in its first derivative at locations $r_1 = 53, r_2 = 100$ and $r_3 = 124$



For $j \in \{1, \dots, |L_{s,e}|\}$, denote

$$[s_{2j}, e_{2j}] := [L_{s,e}(j), e].$$

```

i = 1
(Main part)
b_{2i-1} := argmax_{t \in [s_{2i-1}, e_{2i-1}]} L(C_{s_{2i-1}, e_{2i-1}}^t(X))
if L(C_{s_{2i-1}, e_{2i-1}}^{b_{2i-1}}(X)) > \zeta_{T,d} then
    add b_{2i-1} to the set of estimated change-points.
    MID(e_{2i-1}, e, \lambda_T, \zeta_{T,d}, L)
else
    b_{2i} := argmax_{t \in [s_{2i}, e_{2i}]} L(C_{s_{2i}, e_{2i}}^t(X))
    if L(C_{s_{2i}, e_{2i}}^{b_{2i}}(X)) > \zeta_{T,d} then
        add b_{2i} to the set of estimated change-points.
        MID(s, s_{2i}, \lambda_T, \zeta_{T,d}, L)
    else
        i = i + 1
        if i \le \max\{|L_{s,e}|, |R_{s,e}|\} then
            Go back to (Main part) and repeat
        else
            STOP
        end if
    end if
end if
end if
    
```

```

end if
end function
    
```

A brief explanation of the pseudocode follows. With K already defined above, the intervals $[s_1, e_1], \dots, [s_{2K}, e_{2K}]$ are those used for the isolation step. Notice that in the odd intervals $[s_1, e_1], [s_3, e_3], \dots, [s_{2K-1}, e_{2K-1}]$ the start-point is fixed, unchanged, and equal to s , meaning that $s_1 = s_3 = \dots = s_{2K-1} = s$. In the even intervals $[s_2, e_2], [s_4, e_4], \dots, [s_{2K}, e_{2K}]$, it is the end-point that is kept fixed and equal to e , meaning that $e_2 = e_4 = \dots = e_{2K} = e$. The process will follow until there are intervals to check. The term “expanding intervals” that is used throughout the paper is due to this one-sided expansion (of magnitude λ_T) of the intervals. As shown in Fig. 2 for the specific example in Fig. 1, the pseudocode makes it also clear that, in general, MID is looking for change-points interchangeably in right- and left-expanding intervals which, with high probability, contain at most one change-point. In each of these

intervals, MID acts in the same way as shown in the main part of the pseudocode. The MID procedure is launched by the call $\mathbf{MID}(1, T, \lambda_T, \zeta_{T,d}, L)$.

2.2 Theory

We work under the setting in (1). Denoting by $r_0 = 0$ and $r_{N+1} = T$, the illustration scenarios are:

(S1) Changes in the mean structure: For $k = 1, \dots, N + 1$, $f_t = \mu_k \in \mathbb{R}^d$ for $t = r_{k-1} + 1, \dots, r_k$. In this case, the univariate component signals $f_{t,j}$, for $j = 1, \dots, d$ are piecewise-constant.

(S2) Changes in the first derivative: For $k = 1, \dots, N + 1$, $f_t = \mu_{1,k} + \mu_{2,k}t$ for $t = r_{k-1} + 1, \dots, r_k$, where $\mu_{1,k}$ and $\mu_{2,k}$ are vectors in \mathbb{R}^d . In addition, we require that for $j = 1, \dots, N$, the equality $\mu_{1,j} + \mu_{2,j}r_j = \mu_{1,j+1} + \mu_{2,j+1}r_j$ is satisfied. Under this framework, the change-points, r_k , satisfy that $f_{r_{k-1}} + f_{r_{k+1}} \neq 2f_{r_k}$. Therefore, the univariate component signals $f_{t,j}$, for $j = 1, \dots, d$ are continuous and piecewise-linear.

The aforementioned scenarios are only two specific illustration cases in which the proposed MID algorithm can be applied. Due to its change-point isolation step prior to detection, our algorithm can be applied in more complicated scenarios where each univariate signal could be, for example, piecewise-polynomial or piecewise-exponential. In Sects. 2.2.1 and 2.2.2, we provide the main theorems for the consistency of our method in accurately estimating the true number and the location of the change-points in Scenarios (S1) and (S2), respectively. The theoretical results presented in this section are for either the L_∞ or the L_2 mean-dominant norms discussed in the paper for the aggregation of the information from the component data sequences; under a similar proof strategy, results can be obtained for other mean-dominant norms as well.

2.2.1 Scenario (S1)

As already discussed, the first step of the detection process depends on an appropriately chosen contrast function, which, for every component data sequence, is applied to each change-point candidate. In Scenario (S1), the contrast function applied to the component data sequences $X_{t,j}, \forall j \in \{1, \dots, d\}$ is the absolute value of the widely used CUSUM statistic, with the latter being defined as

$$\tilde{X}_{s,e}^{b,j} = \sqrt{\frac{e-b}{n(b-s+1)}} \sum_{t=s}^b X_{t,j} - \sqrt{\frac{b-s+1}{n(e-b)}} \sum_{t=b+1}^e X_{t,j},$$

where $1 \leq s \leq b < e \leq T$ and $n = e - s + 1$. Before proceeding with the main theoretical result for the consistency

of our method, and with $L(\cdot)$ denoting any one of the mean-dominant norms in (2), allow us to introduce some more notation, as below.

$$\begin{aligned} \delta_T &:= \min_{j=1, \dots, N+1} |r_j - r_{j-1}|, \\ \Delta_j &:= |f_{r_{j+1}} - f_{r_j}|, \quad j = 1, \dots, N \\ \underline{f}_T &:= \inf_{j=1, \dots, N} \{L(\Delta_j)\}. \end{aligned} \tag{6}$$

The absolute value of the vector $\Delta_j \in \mathbb{R}^d$ in (6) is taken component-wise. For the consistency result with respect to the estimated number of change-points and their estimated locations obtained by MID, we work under the assumption (A1) as follows:

(A1) The quantities δ_T and \underline{f}_T , defined in (6), are connected by

$$\sqrt{\delta_T} \underline{f}_T \geq \underline{C} \sqrt{\log(Td^{1/4})}, \text{ for a large enough constant } \underline{C}.$$

The number of change-points, N , is allowed to grow with the sample size T and the dimensionality d . Theorem 1 provides the main theoretical result for Scenario (S1) when either L_∞ or L_2 are employed for the aggregation of the contrast function values. The proof is given in the appendix.

Theorem 1 *Let $\{X_t\}_{t=1, \dots, T}$ follow model (1) with f_t as in Scenario (S1) and $\epsilon_t \sim \mathcal{N}_d(\mathbf{0}, \Sigma)$, where $\Sigma \in \mathbb{R}^{d \times d}$ is positive definite. Let N and $r_j, j = 1, \dots, N$ be the number and locations of the change-points, while \hat{N} and $\hat{r}_j, j = 1, \dots, \hat{N}$ are their estimates sorted in increasing order. Assuming that (A1) holds, then, there exist positive constants C_1, C_2, C_3 , and C_4 , which do not depend on T or d , such that for $C_1 \sqrt{\log(Td^{1/4})} \leq \zeta_{T,d} < C_2 \sqrt{\delta_T} \underline{f}_T$, we have:*

For $L(\cdot) = L_\infty(\cdot)$,

$$\mathbb{P} \left(\hat{N} = N, V_\infty \leq C_3 \log \left(Td^{\frac{1}{4}} \right) \right) \geq 1 - \frac{C_4}{T}, \tag{7}$$

where $V_\infty := \max_{j=1, \dots, N} \left\{ |\hat{r}_j - r_j| \left(\Delta_j^{q_j} \right)^2 \right\}$ and $q_j := \operatorname{argmax}_{k=1, \dots, d} |\tilde{X}_{s_j, e_j}^{\hat{r}_j, k}|$, for $[s_j, e_j]$ being the interval where \hat{r}_j is obtained.

For $L(\cdot) = L_2(\cdot)$,

$$\mathbb{P} \left(\hat{N} = N, V_2 \leq C_3 \log \left(Td^{\frac{1}{4}} \right) \right) \geq 1 - \frac{C_4}{T}, \tag{8}$$

where $V_2 := \max_{j=1, \dots, N} \left\{ |\hat{r}_j - r_j| L_2^2(\Delta_j) \right\}$.

The lower bounds for the probabilities in (7) and (8) do not depend on the dimensionality d and their order is

$1 - \mathcal{O}(T^{-1})$. Furthermore, the rate of convergence of the estimated change-point locations does not depend on the minimum distance between two change-points, δ_T ; the aggregated jump magnitude, $\Delta_j^{q_j}$ for L_∞ and $L_2(\Delta_j)$ for L_2 , is the only quantity that affects the rate. We notice, though, that to be able to match the estimated change-point locations with the true ones, then δ_T should be larger than the distance between the estimated and the true change-point locations. Therefore, based on (7) and (8), we deduce that δ_T must be at least $\mathcal{O}(\log(Td^{1/4}))$. To put the obtained rate of convergence into perspective, it has already been proven in the literature (see, for example Chan and Walther 2013) that in the univariate case the smallest possible $\delta_T f_T^2$, which allows for the detection of changes in the mean of a data sequence, is $\mathcal{O}(\log T - \log(\log T))$. In our case, for $d = 1$, the $\mathcal{O}(\log T)$ rate is attained, which is near-optimal up to the rather negligible double logarithmic term. Moving now to the case where $d > 1$, we will compare the consistency results of MID to those obtained for two known recent procedures for multiple change-point detection in multivariate settings; the *inspect* algorithm of Wang and Samworth (2018) and the kernel change-point (KCP) algorithm of Arlot et al. (2019). The finite sample bounds like those in (7) and (8), imply a rate of convergence in an asymptotic setting. The comparison of the consistency results between the aforementioned methods is carried out following a known convention in the literature (see, for example Venkatraman (1992)). For \mathcal{P}_T being a class of distributions of $X \in \mathbb{R}^{d \times T}$, which is as in (1), we state that the procedure under consideration is consistent for \mathcal{P}_T with rate of convergence ρ_T if

$$\inf_{P \in \mathcal{P}_T} \mathbb{P}_P \left(\hat{N} = N, \max_{j=1, \dots, N} |r_j - \hat{r}_j| \leq T \rho_T \right) \xrightarrow{T \rightarrow \infty} 1. \tag{9}$$

For MID, Theorem 1 shows that the relevant rate of convergence is $\rho_T^{\text{MID}} = \mathcal{O}(T^{-1} \underline{f}_T^{-2} \log(Td^{1/4}))$. Theorem 2 of Wang and Samworth (2018) indicates that, apart from its dependence on the magnitude of the changes, the rate of convergence of the estimated change-point locations in *inspect* is also affected by the minimum distance between consecutive change-points; as already mentioned, this is not the case for MID. More specifically, employing our notation, Wang and Samworth (2018) show that, as long as $\delta_T \geq 14$, for *inspect* the rate of convergence of the estimated locations is $\rho_T^{\text{insp}} = \mathcal{O}(T^3 \log(Td) \vartheta^{-2} \delta_T^{-4})$, where ϑ is such that $\vartheta \leq \|\theta^{(i)}\|_2, \forall i \in \{1, \dots, N\}$, where $\theta^{(i)} \in \mathbb{R}^d$ is the vector of the change magnitudes associated to the i^{th} change-point. Corollary 3 of Wang and Samworth (2018) explains that if $\log(d) = \mathcal{O}(\log(T))$, $\vartheta \asymp T^{-\alpha}$ and $\delta_T \asymp T^{1-\beta}$, with $2\alpha + 5\beta < 1$, then *inspect* estimates all change-points with rate of convergence $\rho_T^{\text{insp}} = o(T^{-(1-2\alpha-4\beta)+\delta})$, for

any $\delta > 0$. Under the same scenario, and since $\underline{f}_T \leq \|\theta^{(i)}\|_2, \forall i \in \{1, \dots, N\}$, meaning that we can take $\underline{f}_T \asymp \vartheta$, our method estimates the change-points with rate of convergence $\rho_T^{\text{MID}} = o(T^{-(1-2\alpha)+\delta})$, which is an improvement over the rate attained by *inspect*. Consistency results for the KCP algorithm of Arlot et al. (2019) have been extensively studied in Garreau and Arlot (2018). More specifically, for $k : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ a positive semidefinite kernel, and with M a positive constant such that $k(X_t, X_t) \leq M^2 < \infty, \forall t \in \{1, \dots, T\}$, Theorem 3.1 of Garreau and Arlot (2018) shows that the change-points are estimated by KCP with rate of convergence $\rho_T^{\text{KCP}} = \mathcal{O}(T^{-1} \underline{f}_T^{-2} N M^2 \log(T))$, where N is the true number of change-points; the order of the lower bound for the probability is the same as in (7) and (8). To compare the consistency result of KCP with that of our method, we first highlight that an upper bound on the true number of change-points, N , is needed in order for KCP to be able to estimate the change-points; this is not the case in MID since there is no prior information on N which is allowed to grow with the sample size, T , and the dimensionality, d . Furthermore, in our understanding the positive constant M^2 appearing in the rate of convergence of KCP depends on the dimensionality d . For example in the special case of the linear kernel (see Arlot et al. (2019) for classical examples of kernels) $k^{\text{lin}}(x, y) = \langle x, y \rangle_{\mathbb{R}^d}$, for $x, y \in \mathbb{R}^d$, we deduce that $M^2 = \mathcal{O}(d)$, leading to $\rho_T^{\text{KCP}} = \mathcal{O}(NdT^{-1} \underline{f}_T^{-2} \log T)$, which is worse than ρ_T^{MID} if either d or N are allowed to diverge with T .

Assumption (A1) requires $\delta_T (\underline{f}_T)^2$ to be of order at least $\mathcal{O}(\log(Td^{1/4}))$. Combining this with the fact that the minimum distance, δ_T , between successive change-points is at least $\mathcal{O}(\log(Td^{1/4}))$, means that \underline{f}_T could decrease with T in cases where δ_T is of order higher than $\mathcal{O}(\log(Td^{1/4}))$.

With respect to the threshold parameter, $\zeta_{T,d}$, the rate of its lower bound is $\mathcal{O}(\sqrt{\log(Td^{1/4})})$; this will also be used in practice as the default rate. Therefore, we have that

$$\zeta_{T,d} = C \sqrt{\log(Td^{1/4})}, \tag{10}$$

where C is a positive constant. More details on the choice of C are given in Sect. 3.3.

2.2.2 Scenario (S2)

We are under the scenario where for any $j \in \{1, \dots, d\}$, the underlying signal $f_{t,j}, t = 1, \dots, T$ has a continuous and piecewise-linear structure as in Fig. 4. In this case, the contrast function applied to the component data sequences $X_{t,j}, \forall j \in \{1, \dots, d\}$ is

$$C_{s,e}^b(X_j) = |\langle X_j, \phi_{s,e}^b \rangle|, \tag{11}$$

where for $n = e - s + 1$ and

$$\alpha_{s,e}^b = \sqrt{\frac{6}{n(n^2 - 1)(1 + (e - b + 1)(b - s + 1) + (e - b)(b - s))}}$$

$$\beta_{s,e}^b = \sqrt{\frac{(e - b + 1)(e - b)}{(b - s + 1)(b - s)}}$$

we have that the contrast vector, $\phi_{s,e}^b(t)$, is equal to

$$\alpha_{s,e}^b \beta_{s,e}^b [(e + 2b - 3s + 2)t - (be + bs - 2s^2 + 2s)], \text{ for } t \in [s, b]$$

$$\frac{\alpha_{s,e}^b}{\beta_{s,e}^b} [(2e^2 + 2e - be - bs) - (3e - 2b - s + 2)t], \text{ for } t \in [b + 1, e]$$

0, otherwise.

For more details on how this vector is constructed for (S2), please see section B.2 in the online supplementary material of Baranowski et al. (2019). Under Scenario (S2), we denote by

$$\Delta_j := |f_{r_j-1} + f_{r_j+1} - 2f_{r_j}|, \quad j = 1, \dots, N$$

$$\underline{f}_T := \inf_{j=1, \dots, N} \{L(\Delta_j)\}, \tag{12}$$

whereas δ_T has the same expression as in (6). We are ready to proceed with the consistency result of MID when applied under Scenario (S2). We work under the assumption:

(A2) The quantities δ_T and \underline{f}_T are connected by $\delta_T^{3/2} \underline{f}_T \geq \underline{C}^* \sqrt{\log(Td^{1/4})}$, where \underline{C}^* is a large enough constant.

Theorem 2 provides the main theoretical result for Scenario (S2). The steps for its proof are similar as those employed in the proof of Theorem 1. Therefore, the proof is given in the supplementary material.

Theorem 2 *Let $\{X_t\}_{t=1, \dots, T}$ follow model (1) with f_t as in Scenario (S2) and $\epsilon_t \sim \mathcal{N}_d(\mathbf{0}, \Sigma)$, where $\Sigma \in \mathbb{R}^{d \times d}$ is positive definite. Let N and $r_j, j = 1, \dots, N$ be the number and locations of the change-points, while \hat{N} and $\hat{r}_j, j = 1, \dots, \hat{N}$ are their estimates sorted in increasing order. Assuming that (A2) holds, then, there exist positive constants C_1, C_2, C_3 , and C_4 , which do not depend on T or d , such that for $C_1 \sqrt{\log(Td^{1/4})} \leq \zeta_{T,d} < C_2 \delta_T^{3/2} \underline{f}_T$, we obtain that:*

For $L(\cdot) = L_\infty(\cdot)$,

$$\mathbb{P} \left(\hat{N} = N, \tilde{V}_\infty \leq C_3 \left(\log(Td^{1/4}) \right)^{1/3} \right) \geq 1 - \frac{C_4}{T}, \tag{13}$$

where $\tilde{V}_\infty := \max_{j=1, \dots, N} \left\{ |\hat{r}_j - r_j| \left(\Delta_j^{q_j} \right)^{2/3} \right\}$ and $q_j := \operatorname{argmax}_{k=1, \dots, d} \left\{ C_{s_j, e_j}^{\hat{r}_j}(\mathbf{X}_k) \right\}$, for $[s_j, e_j]$ denoting the interval where \hat{r}_j is obtained during MID.

For $L(\cdot) = L_2(\cdot)$,

$$\mathbb{P} \left(\hat{N} = N, \tilde{V}_2 \leq C_3 \left(\log(Td^{1/4}) \right)^{1/3} \right) \geq 1 - \frac{C_4}{T}, \tag{14}$$

where $\tilde{V}_2 := \max_{j=1, \dots, N} \left\{ |\hat{r}_j - r_j| \left(L_2(\Delta_j) \right)^{2/3} \right\}$.

Similarly as in (S1), the upper bounds for the probabilities in (13) and (14) do not depend on the dimensionality of the data sequence X_t ; their rate of convergence is $1 - \mathcal{O}(T^{-1})$. Furthermore, the rate of convergence of the estimated change-point locations depends only on the aggregated change magnitude, $\Delta_j^{q_j}$ for L_∞ and $L_2(\Delta_j)$ for L_2 . Under (S2), the change-points are estimated with rate of convergence $\mathcal{O} \left(T^{-1} f_T^{-2/3} \left(\log(Td^{1/4}) \right)^{1/3} \right)$. In the special case of $\underline{f}_T \asymp T^{-1}$, the rate of convergence becomes $\mathcal{O} \left(T^{-1/3} \left(\log(Td^{1/4}) \right)^{1/3} \right)$. In the univariate case, it is proven in Raimondo (1998) that the asymptotic minimax rate for the change-point detection problem under (S2) is $\mathcal{O}(T^{-1/3})$, which differs to the convergence rate obtained by MID only by the logarithmic factor.

As in Scenario (S1), the lower bound for the threshold, $\zeta_{T,d}$, is of order $\mathcal{O} \left(\sqrt{\log(Td^{1/4})} \right)$; this is the default rate that is used in practise in Sects. 4 and 5. Therefore,

$$\zeta_{T,d} = C^* \sqrt{\log(Td^{1/4})}, \tag{15}$$

where $C^* > 0$. More details on the choice of the values for C^* are given in Sect. 3.3.

3 Computational complexity and practicalities

3.1 Computational complexity

With $K = \lceil T/\lambda_T \rceil$, the total number of distinct intervals, \mathcal{I} , required in order to cover the whole data sequence is at most $2K$ (K intervals from each expanding direction). Choosing the expansion step, λ_T , small enough leads to isolation with high probability; more information on how to choose λ_T in order to obtain good accuracy performance while maintaining low computational cost can be found in Sect. 3.3. With δ_T the minimum distance between consecutive change-points, isolation is guaranteed as long as $\lambda_T < \delta_T$. We use this inequality, which leads to $K > \lceil T/\delta_T \rceil$ and therefore, in the worst case scenario $\mathcal{I} = 2K > 2\lceil T/\delta_T \rceil$.

The lower bound is of order $\mathcal{O}(T/\delta_T)$. For generic intervals $[s, e]$, $1 \leq s < e \leq T$ considered throughout the MID algorithm, the relevant contrast function values $C_{s,e}^b(\mathbf{X}_j)$ will be calculated $\forall b \in [s, e]$ and $\forall j \in \{1, \dots, d\}$ before aggregation takes place. In both scenarios studied in our paper, for fixed j , the cost of computing $C_{s,e}^b(\mathbf{X}_j)$ is $\mathcal{O}(e - s + 1)$, meaning that it is linear in time; see Section B.1 of the online supplement in Baranowski et al. (2019). Therefore, the calculation of the contrast function values for all the component data sequences has computational cost of order $\mathcal{O}(KdT)$. Applying now the mean-dominant norms, as in (2), to the obtained values has an order $\mathcal{O}(d)$ computational cost. Combining the complexities explained at each step of MID, we conclude that the total computational complexity of the algorithm is of order $\mathcal{O}(Kd^2T) = \mathcal{O}(d^2T^2\delta_T^{-1})$.

3.2 Mean-dominant norms

The proposed MID methodology is based on an aggregation step of the contrast function values obtained from each component data sequence using mean-dominant norms in the way these are defined in Section 2 of Carlstein (1988). Theorems 1 and 2 for the Scenarios (S1) and (S2), respectively, cover the theoretical behaviour of MID under both the L_∞ and L_2 mean-dominant norms given in (2). Therefore, in the remaining sections, the discussion on the choice of the parameter values as well as results on the practical performance of MID will be focused when our method is combined with the L_∞ or the L_2 mean-dominant norm. A data-adaptive variant of MID that chooses the most suitable mean-dominant norm to be used for the aggregation step will also be introduced.

3.3 Choice of parameter values

In order to choose the constants C in (10) and C^* in (15), we ran a large scale simulation study involving data sequences $\{\mathbf{X}_t\}_{t=1, \dots, T}$, for $T = 700, 1400$ following model (1), where $\mathbf{f}_t = \mathbf{0}$ while $\boldsymbol{\epsilon}_t$ are follow the d -variate standard normal distribution. Specifically, for each $d \in \{1, \dots, 50\}$ we generated 500 replicates and applied MID under scenarios (S1) and (S2) to each one of those replicates using various threshold constant values C and C^* , respectively, in order to estimate the number of change-points. For each dimension, in order to control the Type I error rate, α , of falsely detecting change-points, we chose the default constant to be the one that its number of times that successfully did not detect any change-points was closer to $(1 - \alpha) * 500$. For $d > 50$, we keep the threshold that gave the best results for the simulated 50-dimensional data sequence. Table 1 presents the results for $\alpha \in \{0.05, 0.1\}$ under Scenarios (S1) and (S2). From now on, the obtained values for C and C^* will be referred to as the default constants.

Regarding the expansion parameter, λ_T , we have already discussed that isolation of the change-points is guaranteed in theory, as long as we take $\lambda_T < \delta_T$. More specifically, in the proofs of Theorems 1 and 2, we show change-point isolation and detection for $\lambda_T \leq \delta_T/3$. However, any value of $\lambda_T \leq \delta_T/m$, for $m > 1$ suffices; the adjustments to be made in the proof in such a case, as well as why $m = 3$ leads to a more symmetric approach for the intervals where detection occurs, are explained in Remark 1 of the online supplement of Anastasiou and Fryzlewicz (2022). In practice, δ_T is unknown, and to ensure isolation, λ_T can be taken to be as small as equal to 1. If $\lambda_T > 1$, then isolation is guaranteed with high probability. Now, the computational cost of running our algorithm is inversely proportional to the size of λ_T because the smaller the λ_T the larger the number of the created intervals, which are examined for change-points. However, the computational speed of MID provides flexibility in the choice of the expansion parameter; in all practical examples in Sects. 4 and 5 we take $\lambda_T = 3$. Tables 2, 3 and 4 indicate the high accuracy of our method even in quite complex scenarios that exhibit low sparsity (see (16) for the relevant definition) and/or a large number of regularly occurring change-points.

3.4 Decision on the aggregation method

In Cho (2016), Cho and Fryzlewicz (2015), and Anastasiou et al. (2022), it has been explained that the L_∞ -based aggregation of the contrast functions for each component data sequence tends to exhibit a better behaviour (compared to the L_2 -based aggregation) in scenarios where the true change-points appear only in a small number of the component data sequences. In contrast, due to spuriously large contrast function values, the L_∞ approach could suffer in situations where the change-points are not sparse across the panel of the data sequences.

This difference in the behaviour between the L_2 and the L_∞ mean-dominant norms examined in our paper is what has motivated us to introduce an introductory step in MID, where the sparsity in a given d -dimensional data sequence, \mathbf{X}_t , is first estimated. For r_1, \dots, r_N being the N true change-points, allow us, for any $j \in \{1, \dots, N\}$, to define by $A_j \subseteq \{1, \dots, d\}$ the set of indices for the univariate component data sequences that contain the change-point r_j . Then, for \mathbf{X}_t , the sparsity is given by

$$sp = \max_{j=1, \dots, N} \{|A_j|\} / d, \quad (16)$$

where $|A_j|$ is the cardinality of the set A_j . It is straightforward that $sp \in [0, 1]$. Our proposed hybrid approach aims to data-adaptively decide on the aggregation method to be used and eliminates the necessity of choosing between the L_2 and

Table 1 The optimal values for the threshold constants, C and C^* , which control the Type I error rate α under the Scenarios (S1) and (S2), respectively, for $d = 1, \dots, 50$. Results are presented for the L_2 and the L_∞ norms

Mean-dominant norm	Results for the (S1) scenario			Results for the (S2) scenario		
	d	$C (\alpha = 0.05)$	$C (\alpha = 0.1)$	d	$C^* (\alpha = 0.05)$	$C^* (\alpha = 0.1)$
L_2	1	1.7	1.55	1	1.65	1.55
	2	1.25	1.25	2	1.25	1.2
	3	1.1	1.05	3	1.05	1.05
	4	1.05	0.95	4	0.95	0.95
	5	0.95	0.9	5	0.9	0.9
	6	0.9	0.9	6	0.9	0.85
	7	0.9	0.8	7	0.8	0.8
	8	0.8	0.8	8	0.8	0.75
	9	0.8	0.75	9–11	0.75	0.75
	10–13	0.75	0.75	12–16	0.7	0.7
	14	0.75	0.65	17–19	0.65	0.6
	15–20	0.7	0.65	20–22	0.6	0.6
	21–23	0.65	0.6	24–42	0.6	0.55
	24–39	0.6	0.6	43–50	0.55	0.55
	40–50	0.6	0.55			
L_∞	1	1.7	1.55	1	1.65	1.55
	2–3	1.75	1.7	2	1.7	1.6
	4–6	1.8	1.7	3	1.75	1.6
	7–13	1.85	1.75	4–5	1.75	1.65
	14–25	1.9	1.8	6–13	1.75	1.7
	26–28	1.9	1.85	14–25	1.8	1.75
	29–50	1.95	1.85	26–38	1.85	1.8
				39–50	1.9	1.85

L_∞ mean-dominant norms. The steps to achieve our purpose are as follows.

Step 1 We apply MID paired with the L_∞ aggregation rule in order to obtain $\hat{r}_1, \dots, \hat{r}_M$. It has already been explained that the L_∞ norm is preferable and provides very good results in cases with sparse change-points, while it tends to over-estimate the number of change-points (due to the contrast function taking spuriously large values) under the scenario of dense change-points. Therefore, $M \geq N$.

Step 2 We will now estimate the sparsity in the given data as defined in (16). With $\hat{r}_0 = 0$ and $\hat{r}_{M+1} = T$, we first collect the triplets $(\hat{r}_{m-1} + 1, \hat{r}_m, \hat{r}_{m+1}), \forall m \in \{1, \dots, M\}$. After this, $\forall i \in \{1, \dots, d\}$, we calculate $CS^{(i)}(\hat{r}_m) := C_{\hat{r}_{m-1}+1, \hat{r}_{m+1}}^{\hat{r}_m}(X_i)$, with $C_{s,e}^b(X_i)$ being the relevant contrast function (based on whether we are under Scenario (S1) of Sect. 2.2.1 or Scenario (S2) of Sect. 2.2.2) value for the point b , when we are in the interval $[s, e]$, for the univariate component data sequence X_i . The d contrast function values for each \hat{r}_m are collected in the sets

$$S_m = \{CS^{(1)}(\hat{r}_m), \dots, CS^{(d)}(\hat{r}_m)\}, \quad j = 1, \dots, M. \quad (17)$$

Step 3 For each $m = 1, \dots, M$, all the elements of S_m are tested against the relevant threshold value, ζ_T , for univariate change-point detection which is of the order $\mathcal{O}(\sqrt{\log T})$ in both scenarios of piecewise-constant and continuous piecewise-linear signals (representing scenarios (S1) and (S2), respectively, covered in this paper). In terms of the threshold constants, we employ those of Anastasiou and Fryzlewicz (2022). We denote by $\hat{sp}_m := \#\{i : CS^{(i)}(\hat{r}_m) > \zeta_T\} / d$. The estimated sparsity is then $\hat{sp} = \max_{m \in \{1, \dots, M\}} \{\hat{sp}_m\}$.

Step 4 For cases where $\hat{sp} \leq 0.4$, we accept the result from Step 1, where MID has been paired with the L_∞ mean-dominant norm, whereas if $\hat{sp} \geq 0.6$, then MID is paired with the L_2 mean-dominant norm. Extensive simulations have shown that there is no significant difference on the MID's practical performance with respect to accuracy (on both the estimated number of change-points and the estimated locations) when $\hat{sp} \in (0.4, 0.6)$. Therefore, MID could be paired with any of the aforementioned two norms and give very good results. For computational cost reasons, in such cases we accept the result of Step 1. From now on, we denote by

Table 2 Distribution of $\hat{N} - N$ over 100 simulated multivariate data sequences with 3 change-points under Scenario (S1) of Sect. 2.2.1. The signal strength, as defined in (18), is equal to 2 for each change-point. The average ARI, d_H , and computational times are also given

Method	sp	d	$\hat{N} - N$						ARI	d_H	Time (s)
			≤ -2	-1	0	1	2	3			
MID_{opt}	0.2	30	0	0	98	2	0	0	0.991	0.015	1.63
MIDPERL ₂	0.2	30	0	2	82	14	0	2	0.962	0.047	330.81
MIDPERL _∞	0.2	30	0	0	79	18	2	1	0.956	0.049	172.72
DC	0.2	30	0	0	93	5	2	0	0.993	0.042	19.84
INSPECT	0.2	30	0	0	95	4	1	0	0.992	0.030	1.08
KCP_Grid	0.2	30	88	10	2	0	0	0	0.422	1.615	34.29
KCP_Scree	0.2	30	0	3	97	0	0	0	0.990	0.008	34.29
SBS	0.2	30	0	0	95	5	0	0	0.986	0.022	3.56
MID_{opt}	0.2	100	0	0	97	2	1	0	0.981	0.028	4.21
MIDPERL ₂	0.2	100	0	4	85	7	4	0	0.971	0.042	1098.40
MIDPERL _∞	0.2	100	0	1	82	13	4	0	0.954	0.071	615.85
DC	0.2	100	0	0	94	6	0	0	0.993	0.029	32.30
INSPECT	0.2	100	0	0	96	3	1	0	0.991	0.025	4.18
KCP_Grid	0.2	100	92	8	0	0	0	0	0.438	1.458	41.94
KCP_Scree	0.2	100	0	4	96	0	0	0	0.988	0.009	41.94
SBS	0.2	100	1	4	81	14	0	0	0.942	0.149	6.38
MID_{opt}	0.5	30	0	0	98	1	1	0	0.985	0.020	1.68
MIDPERL ₂	0.5	30	0	1	89	9	1	0	0.972	0.056	331.91
MIDPERL _∞	0.5	30	0	1	82	15	2	0	0.959	0.061	161.49
DC	0.5	30	0	0	97	3	0	0	0.994	0.009	20.42
INSPECT	0.5	30	0	0	95	5	0	0	0.993	0.016	1.14
KCP_Grid	0.5	30	96	0	4	0	0	0	0.428	1.499	34.07
KCP_Scree	0.5	30	0	5	95	0	0	0	0.988	0.015	34.07
SBS	0.5	30	0	0	90	10	0	0	0.980	0.042	3.60
MID_{opt}	0.5	100	0	1	98	1	0	0	0.954	0.075	4.56
MIDPERL ₂	0.5	100	0	0	93	7	0	0	0.972	0.041	1069.54
MIDPERL _∞	0.5	100	0	0	93	7	0	0	0.973	0.042	751.34
DC	0.5	100	0	0	94	6	0	0	0.993	0.015	32.06
INSPECT	0.5	100	0	0	98	2	0	0	0.993	0.012	4.09
KCP_Grid	0.5	100	88	12	0	0	0	0	0.492	1.182	41.79
KCP_Scree	0.5	100	0	5	95	0	0	0	0.990	0.010	41.79
SBS	0.5	100	0	7	78	15	0	0	0.939	0.159	6.41
MID_{opt}	0.8	30	0	0	97	3	0	0	0.988	0.022	2.36
MIDPERL ₂	0.8	30	0	1	92	5	2	0	0.983	0.043	324.60
MIDPERL _∞	0.8	30	0	0	85	14	1	0	0.971	0.076	161.49
DC	0.8	30	0	0	96	3	1	0	0.993	0.023	20.60
INSPECT	0.8	30	0	0	93	6	1	0	0.991	0.023	1.17
KCP_Grid	0.8	30	88	10	2	0	0	0	0.462	1.380	33.53
KCP_Scree	0.8	30	0	10	90	0	0	0	0.989	0.10	33.53
SBS	0.8	30	0	2	94	6	0	0	0.980	0.035	3.68
MID_{opt}	0.8	100	0	4	91	5	0	0	0.933	0.133	5.01
MIDPERL ₂	0.8	100	0	0	94	4	1	1	0.976	0.036	1085.59
MIDPERL _∞	0.8	100	0	0	91	8	1	0	0.971	0.042	602.87
DC	0.8	100	0	0	96	3	1	0	0.993	0.014	32.26
INSPECT	0.8	100	0	0	95	5	0	0	0.990	0.023	3.72

Table 2 continued

Method	<i>sp</i>	<i>d</i>	$\hat{N} - N$						ARI	d_h	Time (s)
			≤ -2	-1	0	1	2	3			
KCP_Grid	0.8	100	82	18	0	0	0	0	0.511	0.142	41.99
KCP_Scree	0.8	100	0	4	96	0	0	0	0.989	0.009	41.99
SBS	0.8	100	0	7	83	10	0	0	0.926	0.160	6.47

MID_{opt} to be the data-adaptive, sparsity-based MID version explained in this section, where an aggregation method is chosen based on the estimated sparsity of the change-points in the given data.

3.5 A permutation-based approach

MID is a threshold-based algorithm because at each step, the largest aggregated contrast function value is tested against a predefined threshold in order to decide whether there is a change-point at the corresponding location. In Sect. 3.3 we have explained how the threshold constant is carefully chosen to control the Type I error taking into account the dimensionality of the data. However, misspecification of the threshold can possibly lead to the misestimation of the number of change-points. To solve such issues, we propose a variant of MID based on permutation.

The idea of a “data-adaptive threshold” through permutations or bootstrapping is not new. In Cho (2016), a bootstrap procedure is proposed, which is motivated by the representation theory developed for the Generalised Dynamic Factor Model, in order to approximate the quantiles of the developed double CUSUM test statistic under the null hypothesis of no change-points; the obtained quantiles are used as a test criterion for detection. In Cabrieto et al. (2018), a permutation-based approach is used to test the presence of correlation changes in multivariate time series. Under a univariate framework, in Antoch and Hušková (2001) a permutation scheme is proposed for deriving critical values for test statistics based on functionals of the partial sums $\sum_{i=1}^k (X_i - \bar{X}_n)$, $k = 1, \dots, n$, where X_i are the observed data and \bar{X}_n their mean value. The proposed scheme consists of three steps: a) compute the test statistic using the original data, b) construct the permutation distribution by computing the relevant test statistic on permuted versions of the data, and c) reject the null hypothesis of no change-points if the test statistic lies in the tails of the distribution.

We propose a variant that combines the isolation technique of MID with an extension of the permutation procedure used in Antoch and Hušková (2001) to the multivariate framework.

Although this permutation procedure tends to be computationally expensive, it has a straightforward implementation. As generally holds for permutation procedures, the test-statistic obtained from the original data is compared to those obtained when applying the same steps to several permuted versions of the data. For our proposed permutation scheme all steps remain the same as MID with the only difference being the way the algorithm chooses to accept or reject a change-point within each interval. To be more precise, suppose that, for a given data sequence $\{X_t\}_{t=1, \dots, T}$, the MID algorithm is at a step where it looks for a change-point in the interval $I = [s^*, e^*]$, where $1 \leq s^* < e^* \leq T$. As described in Sect. 2.1, MID returns a vector $v \in \mathbb{R}^J$, where J is the amount of all change-point candidates. The elements of v correspond to the aggregated contrast function values for each candidate point in I . The next step is to store $T_{I_{max}} = \max_{i \in \{1, \dots, J\}} \{v_i\}$ and repeat the following procedure a prespecified amount, denoted by K , of times:

1. Generate a random permutation from (s^*, \dots, e^*) .
2. Reorder the data according to the permutation.
3. Calculate and store the maximum aggregated contrast function value for each permutation.

The empirical distribution obtained from the maximum values is used to construct our test. More precisely, we identify $\hat{b}_I = \operatorname{argmax}_t \{v_t\}$ as a change-point if, for given $\alpha \in (0, 1)$, $T_{I_{max}} > q_{1-\alpha}$, where $q_{1-\alpha}$ is the 100(1 - α)% quantile of the empirical distribution.

The parameter α controls the probability of false detections. Small values of α will make it harder for the algorithm to reject the null hypothesis of no change-points, whereas large values can reduce the probability of a Type II error. For the simulations in Sect. 4, we take $\alpha = 0.01$. Regarding the parameter K , the simulation results provided in Antoch and Hušková (2001)—although for a univariate framework—suggest that the empirical quantiles get stabilised quickly. Therefore, for our simulations, we take $K = 1000$.

4 Numerical studies

4.1 Comparative simulation study

In this section, we investigate the performance of our method in Scenarios (S1) and (S2) covered in Sects. 2.2.1 and 2.2.2, respectively. Furthermore, in (S1), MID is compared with state-of-the-art methods through a comprehensive simulation study. The competitors are the Double Cusum (DC) method of Cho (2016), the Sparsified Binary Segmentation (SBS) algorithm of Cho and Fryzlewicz (2015), the INSPECT algorithm of Wang and Samworth (2018), and the KCP method of Arlot et al. (2019). DC and SBS are implemented in the *hdbinseg* R package. For INSPECT we have used the *InspectChangepoint* R package, while KCP is implemented in the *kcpRS* R package. SBS and INSPECT were called with their default arguments. For DC, the parameters used were the ones that gave the best results in the simulation study carried out in the relevant paper (Cho 2016). Regarding KCP, an upper bound, K_{\max} , for the true number, N , of change-points needs to be provided in order for the method to work. In the simulations carried out in this section, the true number of change-points is no more than 50; we take $K_{\max} = 100$. Furthermore, in the *kcpRS* R package, the KCP-based estimated number of change-points is selected either through a model selection procedure where the penalty constants are decided through an extensive grid search, or through a scree test. In our simulation results that follow, the grid-based procedure is denoted by *KCP_Grid*, while the scree-test-based estimation result is denoted by *KCP_Scree*. With respect to our algorithm, we give results for the data-adaptive, sparsity-based MID version of Sect. 3.4, denoted by MID_{opt} , where the threshold values are taken from Table 1 for $\alpha = 0.05$. In addition, results are presented for the permutation-based variants explained in Sect. 3.5; these are denoted by $MIDPERL_2$ and $MIDPERL_\infty$ for the L_2 and L_∞ mean-dominant norms, respectively.

Simulation setup We considered the settings where $T = 1500$, $d \in \{30, 100\}$ and $N \in \{3, 20, 50\}$. With the definition of sparsity as in (16), we took $sp \in \{0.2, 0.5, 0.8\}$, meaning that at least one true change-point appears in $d \times sp$ data sequences. In an attempt to increase the difficulty in the simulations the remaining true change-points appear in only few data sequences. The signal strength at the i th change-point is defined as

$$ss_i^{(S1)} := \|f_{r_i} - f_{r_i+1}\|_2 \tag{18}$$

in the case of Scenario (S1) and as

$$ss_i^{(S2)} := \|f_{r_i-1} + f_{r_i+1} - 2f_{r_i}\|_2, \tag{19}$$

with $\|\cdot\|_2$ being the known Euclidean norm. For our simulation study, for Scenario (S1), we take $ss_i^{(S1)} = s_1, \forall i \in \{1, \dots, N\}$ and we test the performance of all methods on two signal strength settings with $s_1 \in \{2, 2.5\}$. Regarding Scenario (S2), where changes in the vector of the first order derivatives are treated, we have $ss_i^{(S2)} = s_2, \forall i \in \{1, \dots, N\}$ and we test the performance of MID on two signal strength settings with $s_2 \in \{0.3, 0.5\}$. Standard Gaussian noise was added to the signals. In total, we tested the methods in 36 different setups covering a range of different multivariate sequences regarding dimensionality, sparsity, number of change-points, and signal strength.

We ran 100 replications for each setup. The frequency distribution of $\hat{N} - N$ is provided, while the accuracy of the estimated locations is evaluated through the Adjusted Rand Index (ARI) of the estimated segmentation against the true one (Hubert and Arabie 1985), and the scaled Hausdorff distance,

$$d_H = n_s^{-1} \max \left\{ \max_j \min_k |r_j - \hat{r}_k|, \max_k \min_j |r_j - \hat{r}_k| \right\},$$

where n_s is the length of the largest segment. The average computational times, in seconds, are also provided. The results, for MID and the competing methods under Scenario (S1) and for the signal strength, s_1 , being equal to 2, are given in Tables 2, 3, and 4; the results for $s_1 = 2.5$ are given in the online supplement. For each simulation setup, the method with the highest empirical frequency of $\hat{N} - N$ being equal to zero (or close to zero, depending on the example) and those within 5% off the highest are given in bold. For Scenario (S2) with $s_2 = 0.5$, the results are presented in Table 5, while for $s_2 = 0.3$ the results can be found in the online supplement.

As the tables show, MID performs extremely well in all setups in both (S1) and (S2). More specifically, for (S1) our method is either the best method overall or within 5% off the best method with respect to the estimated number of change-points in all signals. Furthermore, in all cases, MID attains very high values for the ARI, and quite small (in most cases it actually attains the smallest value) ones for the scaled Hausdorff distance; such results justify that apart from being extremely accurate in estimating the correct number of change-points, MID is also very accurate regarding the estimated change-point locations. We highlight that there seems to be a significant advantage in the performance of MID compared to the rest of the methods in cases with a large number of regularly occurring change-points; see, more specifically, Tables 3 and 4. Regarding the permutation-based variants of MID, both $MIDPERL_2$ and $MIDPERL_\infty$ show a very good behaviour in all different scenarios in terms of accuracy with respect to both the estimated number and the estimated change-point locations. Although these permutation-based variants do not require specification of the threshold, their

Table 3 Distribution of $\hat{N} - N$ over 100 simulated multivariate data sequences with 20 change-points under Scenario (S1) of Sect. 2.2.1. The signal strength, as defined in (18), is equal to 2 for each change-point. The average ARI, d_H , and computational times are also given

Method	sp	d	$\hat{N} - N$					ARI	d_h	Time (s)
			≤ -10	$(-10, -1)$	$[-1, 1]$	$(1, 10)$	≥ 10			
MID_{opt}	0.2	30	0	0	100	0	0	0.968	0.101	1.68
MIDPERL ₂	0.2	30	0	16	84	0	0	0.942	0.254	476.36
MIDPERL_∞	0.2	30	0	5	95	0	0	0.957	0.167	110.60
DC	0.2	30	0	60	40	0	0	0.877	0.849	16.73
INSPECT	0.2	30	0	0	64	36	0	0.948	0.246	0.99
KCP_Grid	0.2	30	100	0	0	0	0	0.093	9.128	33.17
KCP_Scree	0.2	30	64	2	30	0	4	0.382	5.411	33.17
SBS	0.2	30	79	21	0	0	0	0.495	1.887	3.21
MID_{opt}	0.2	100	0	0	100	0	0	0.932	0.761	4.27
MIDPERL ₂	0.2	100	0	5	95	0	0	0.953	0.241	1488.64
MIDPERL_∞	0.2	100	0	1	99	0	0	0.969	0.157	427.60
DC	0.2	100	0	56	44	0	0	0.875	0.833	26.99
INSPECT	0.2	100	0	2	81	17	0	0.933	0.306	3.12
KCP_Grid	0.2	100	100	0	0	0	0	0.093	9.208	41.81
KCP_Scree	0.2	100	100	0	0	0	0	0.120	7.974	41.81
SBS	0.2	100	38	62	0	0	0	0.542	1.532	5.68
MID_{opt}	0.5	30	0	0	100	0	0	0.939	0.608	1.61
MIDPERL ₂	0.5	30	0	1	99	0	0	0.952	0.183	461.75
MIDPERL_∞	0.5	30	0	0	99	1	0	0.956	0.163	103.31
DC	0.5	30	0	57	43	0	0	0.864	0.873	16.85
INSPECT	0.5	30	0	0	54	46	0	0.945	0.263	0.96
KCP_Grid	0.5	30	100	0	0	0	0	0.095	9.001	33.06
KCP_Scree	0.5	30	100	0	0	0	0	0.257	6.243	33.06
SBS	0.5	30	73	27	0	0	0	0.497	1.936	3.22
MID_{opt}	0.5	100	0	0	100	0	0	0.927	0.843	4.06
MIDPERL ₂	0.5	100	0	0	100	0	0	0.957	0.163	1471.95
MIDPERL_∞	0.5	100	0	0	100	0	0	0.946	0.151	450.34
DC	0.5	100	0	68	32	0	0	0.870	0.866	27.01
INSPECT	0.5	100	0	0	77	23	0	0.935	0.278	3.10
KCP_Grid	0.5	100	100	0	0	0	0	0.093	9.005	41.80
KCP_Scree	0.5	100	100	0	0	0	0	0.108	8.301	41.80
SBS	0.5	100	31	69	0	0	0	0.546	1.581	5.72
MID_{opt}	0.8	30	0	0	100	0	0	0.930	0.783	1.58
MIDPERL ₂	0.8	30	0	1	99	0	0	0.961	0.199	417.58
MIDPERL_∞	0.8	30	0	0	100	0	0	0.954	0.177	95.89
DC	0.8	30	0	62	38	0	0	0.873	0.835	16.96
INSPECT	0.8	30	0	0	61	39	0	0.943	0.257	1.01
KCP_Grid	0.8	30	100	0	0	0	0	0.097	9.137	32.27
KCP_Scree	0.8	30	50	0	42	0	8	0.483	4.207	32.27
SBS	0.8	30	78	22	0	0	0	0.497	1.846	3.24
MID_{opt}	0.8	100	0	0	100	0	0	0.927	0.875	4.09
MIDPERL ₂	0.8	100	0	0	100	0	0	0.958	0.072	1467.22
MIDPERL_∞	0.8	100	0	0	100	0	0	0.951	0.141	459.46

Table 3 continued

Method	sp	d	$\hat{N} - N$					ARI	d_h	Time (s)
			≤ -10	$(-10, -1)$	$[-1, 1]$	$(1, 10)$	≥ 10			
DC	0.8	100	0	63	37	0	0	0.880	0.876	26.93
INSPECT	0.8	100	0	0	80	20	0	0.931	0.257	3.12
KCP_Grid	0.8	100	100	0	0	0	0	0.092	9.140	41.78
KCP_Scree	0.8	100	100	0	0	0	0	0.102	8.677	41.78
SBS	0.8	100	42	58	0	0	0	0.537	1.576	5.70

Table 4 Distribution of $\hat{N} - N$ over 100 simulated multivariate data sequences with 50 change-points under Scenario (S1) of Sect. 2.2.1. The signal strength, as defined in (18), is equal to 2 for each change-point. The average ARI, d_H , and computational times are also given

Method	sp	d	$\hat{N} - N$					ARI	d_h	Time (s)
			< -20	$[-20, -6)$	$[-6, 6]$	$(6, 20]$	> 20			
MID_{opt}	0.2	30	0	6	94	0	0	0.869	0.584	2.79
MIDPERL₂	0.2	30	0	7	93	0	0	0.849	0.612	650.05
MIDPERL_∞	0.2	30	0	7	93	0	0	0.835	0.634	101.91
DC	0.2	30	96	4	0	0	0	0.547	1.291	17.21
INSPECT	0.2	30	0	5	95	0	0	0.844	0.547	1.03
KCP_Grid	0.2	30	100	0	0	0	0	0.039	15.570	32.43
KCP_Scree	0.2	30	65	0	5	26	4	0.324	8.616	32.43
SBS	0.2	30	100	0	0	0	0	0.225	3.662	3.31
MID_{opt}	0.2	100	0	14	86	0	0	0.862	0.588	5.93
MIDPERL₂	0.2	100	0	17	83	0	0	0.851	0.592	2188.56
MIDPERL_∞	0.2	100	0	15	85	0	0	0.847	0.585	530.94
DC	0.2	100	88	12	0	0	0	0.558	1.300	27.06
INSPECT	0.2	100	0	85	15	0	0	0.732	0.618	3.25
KCP_Grid	0.2	100	100	0	0	0	0	0.039	14.968	42.37
KCP_Scree	0.2	100	100	0	0	0	0	0.047	13.753	42.37
SBS	0.2	100	100	0	0	0	0	0.287	2.557	5.82
MID_{opt}	0.5	30	0	10	90	0	0	0.866	0.583	2.62
MIDPERL₂	0.5	30	0	11	89	0	0	0.861	0.591	576.55
MIDPERL_∞	0.5	30	0	9	91	0	0	0.879	0.542	85.13
DC	0.5	30	85	15	0	0	0	0.539	1.401	16.91
INSPECT	0.5	30	0	13	87	0	0	0.834	0.555	1.03
KCP_Grid	0.5	30	100	0	0	0	0	0.043	14.641	32.29
KCP_Scree	0.5	30	72	0	2	15	11	0.256	8.618	32.29
SBS	0.5	30	100	0	0	0	0	0.227	3.648	3.28
MID_{opt}	0.5	100	0	8	92	0	0	0.865	0.585	7.02
MIDPERL₂	0.5	100	0	5	95	0	0	0.873	0.538	2147.27
MIDPERL_∞	0.5	100	0	7	93	0	0	0.868	0.552	524.08
DC	0.5	100	0	92	8	0	0	0.566	1.241	34.90
INSPECT	0.5	100	0	90	10	0	0	0.723	0.614	3.74
KCP_Grid	0.5	100	100	0	0	0	0	0.038	14.939	51.25
KCP_Scree	0.5	100	88	0	0	0	12	0.105	12.443	51.25
SBS	0.5	100	100	0	0	0	0	0.283	2.567	7.08

Table 4 continued

Method	sp	d	$\hat{N} - N$					ARI	d_h	Time (s)
			< -20	$[-20, -6)$	$[-6, 6]$	$(6, 20]$	> 20			
MID_{opt}	0.8	30	0	9	91	0	0	0.865	0.585	2.17
MIDPERL₂	0.8	30	0	7	93	0	0	0.874	0.519	513.94
MIDPERL_∞	0.8	30	0	10	90	0	0	0.861	0.575	77.26
DC	0.8	30	93	7	0	0	0	0.539	1.366	16.67
INSPECT	0.8	30	0	7	92	1	0	0.839	0.544	1.03
KCP_Grid	0.8	30	100	0	0	0	0	0.039	15.598	32.15
KCP_Scree	0.8	30	78	0	0	20	2	0.207	10.733	32.15
SBS	0.8	30	100	0	0	0	0	0.221	3.938	3.25
MID_{opt}	0.8	100	0	10	90	0	0	0.865	0.584	6.76
MIDPERL₂	0.8	100	0	7	93	0	0	0.901	0.512	1723.12
MIDPERL_∞	0.8	100	0	12	88	0	0	0.853	0.580	511.36
DC	0.8	100	87	13	0	0	0	0.567	1.253	34.65
INSPECT	0.8	100	0	90	10	0	0	0.725	0.629	3.56
KCP_Grid	0.8	100	100	0	0	0	0	0.038	15.282	46.38
KCP_Scree	0.8	100	94	0	0	0	6	0.074	13.438	46.38
SBS	0.8	100	100	0	0	0	0	0.284	2.634	6.99

computational cost seems to be quite large (see Tables 2, 3, 4).

In regards to the competing methods, INSPECT has a very accurate behaviour in Table 2 which concerns the case of having three change-points. Even though the method’s behaviour is not very bad in all different combinations examined, Tables 3 and 4 show that INSPECT struggles to accurately estimate the change-points in cases where N is relatively high. DC’s, KCP_Scree’s, and SBS’ performances are also very good in those cases where we have three change-points, but they seem to underestimate (SBS more prominently) N in cases with more, regularly occurring change-points. KCP_Grid does not have a good performance in any of the scenarios tested.

With respect to Scenario (S2) as in Sect. 2.2.2, the results in Table 5 exhibit MID’s very strong performance in accurately estimating both the number, and the locations of the change-points. In conclusion, taking into account its low computational time, we can deduce that our proposed method is reliable and quick in accurately detecting change-points under various different (with respect to the number of change-points, the sparsity, the dimensionality of the data, the signal strength, and the structure of the changes) multivariate settings. R code and instructions to replicate the results in this section are available on <https://github.com/apapan08/Simulations-MID>.

4.2 Spatially dependent data

It has already been shown in the proofs of Theorems 1 and 2 that the covariance matrix, Σ , of the noise terms, ϵ_t , does not need to be diagonal. This means that MID allows for the detection of changes in data structures that exhibit spatial dependence. Apart from the theoretical justification, through the relevant proofs, of the aforementioned statement, in this section we investigate the practical performance of MID for data following two spatial dependence structures under both scenarios (S1) and (S2). In the first case, the covariance matrix is $\Sigma_{i,j}^{(1)} = 2^{-|i-j|}$, while in the second one, the spatial dependence is much stronger, with the covariance matrix being $\Sigma^{(2)} = \mathbf{1}_d \mathbf{1}_d^T$, where $\mathbf{1}_d = (1, 1, \dots, 1)^T \in \mathbb{R}^{d \times 1}$. With the notation as in Sect. 4.1, we take $T = 1500$, $d \in \{30, 100\}$, $sp \in \{0.2, 0.5\}$, $N \in \{3, 20\}$, and the signal strength for the scenarios (S1) and (S2) is taken to be equal to 2 and 0.5, respectively, for each one of the change-points. The results over 100 simulations are given in Table 6 for Scenario (S1) and in the online supplement for (S2).

Table 6 indicates the great practical behaviour of MID in such spatial dependence structures. More specifically, for the case of having 3 change-points, when the covariance matrix is $\Sigma^{(1)}$, MID succeeds in returning $\hat{N} = 3$ in at least 92 out of the 100 replications for all different scenarios regarding the dimensionality of the data sequence and the sparsity level employed. For the case where the covariance matrix is $\Sigma^{(2)}$, MID exhibits excellent behaviour in all settings apart from the one where $d = 100$ and $sp = 0.5$, in which our method accurately estimates the correct number

Table 5 Distribution of $\hat{N} - N$ over 100 simulated multivariate data sequences under Scenario (S2) of Sect. 2.2.2. We present the results of MID for different levels of sparsity and of dimensionality of the data sequence. The number of change-points is equal to 3, 20, or 50 and the signal strength, as defined in (19), is equal to 0.5. The average ARI, d_H , and computational times are also given

N	sp	d	$\hat{N} - N$					ARI	d_h	Time (s)
			-2	-1	0	1	2			
3	0.2	30	0	0	100	0	0	0.992	0.008	1.23
3	0.5	30	0	0	97	3	0	0.991	0.010	1.26
3	0.8	30	0	0	99	1	0	0.995	0.005	2.01
3	0.2	100	0	0	99	1	0	0.990	0.012	5.74
3	0.5	100	0	0	91	9	0	0.982	0.025	5.78
3	0.8	100	0	0	98	2	0	0.994	0.007	9.42

N	sp	d	$\hat{N} - N$					ARI	d_h	Time (s)
			≤ -10	$(-10, -1)$	$[-1, 1]$	$(1, 10)$	≥ 10			
20	0.2	30	0	0	100	0	0	0.967	0.055	1.99
20	0.5	30	0	0	99	1	0	0.965	0.071	2.04
20	0.8	30	0	0	100	0	0	0.968	0.047	3.84
20	0.2	100	0	0	100	0	0	0.959	0.071	5.75
20	0.5	100	0	0	99	1	0	0.953	0.088	6.16
20	0.8	100	0	0	100	0	0	0.969	0.053	9.26

N	sp	d	$\hat{N} - N$					ARI	d_h	Time (s)
			≤ -40	$(-40, -1)$	$[-1, 1]$	$(1, 40)$	≥ 40			
50	0.2	30	0	0	100	0	0	0.921	0.099	3.05
50	0.5	30	0	0	97	3	0	0.918	0.180	2.94
50	0.8	30	0	0	99	1	0	0.898	0.291	3.94
50	0.2	100	0	0	100	0	0	0.904	0.266	7.62
50	0.5	100	0	0	97	3	0	0.877	0.471	7.98
50	0.8	100	0	0	99	1	0	0.868	0.521	11.74

of change-points in 78% of the replications; however, this is still considered to be a good performance level. Moving now to the scenario of 20 change-points, we see in Table 6 that for either $\Sigma^{(1)}$ or $\Sigma^{(2)}$ and for any combination of d and sp , the estimated number of change-points by MID satisfies that $\hat{N} \in \{19, 20, 21\}$ in at least 98 out of the 100 replications. In addition to the outstanding behaviour of MID with respect to the accuracy on the estimated number of change-points, both the Adjusted Rand Index and the Hausdorff distance values indicate the method's accuracy regarding the estimated change-point locations in all settings studied.

4.3 Non-Gaussian noise

In this section, we investigate the practical performance of MID in cases where we deviate from the Gaussianity of the error terms, an assumption made in Theorems 1 and 2 in order to prove the consistency of our method in accurately estimating the true number and the location of the change-points. The simulation setup of Sect. 4.1 is repeated for MID; however, instead of standard Gaussian noise being added to each one of the d component signals, we now explore

the performance of our method in the following two noise structures:

$$\mathbf{S}_{\text{Unif}} : \epsilon_{t,j} \sim^{\text{iid}} \text{Unif}(-\sqrt{3}, \sqrt{3}),$$

$$\mathbf{S}_{\text{t}_8} : \epsilon_{t,j} \sim^{\text{iid}} \sqrt{6/8}t_8,$$

where $\text{Unif}(a, b)$ denotes the continuous Uniform distribution in the interval $[a, b]$, while t_v is the Student-t distribution with v degrees of freedom. The performance of MID in the case of S_{Unif} is very good and there are no significant deviations to the method's performance obtained under Gaussianity. However, preliminary simulations have shown that MID could lead to moderate overestimation on the number of change-points in the case of heavy-tailed distribution, such as the Student-t covered in the S_{t_8} scenario. Therefore, in such cases of heavy-tailed noise, we could take advantage of the Central Limit Theorem and pre-average the data over time in order to obtain a noise structure that is closer to Gaussianity; this is a technique also followed in Section 4.5 of Anastasiou and Fryzlewicz (2022). MID is then applied to the obtained, pre-averaged data in order to estimate the change-points. The results over 100 simulations for Scenario (S1) (changes in the mean structure) are given in Tables 7 and 8 for the cases of uniformly or Student-t distributed error terms, respectively.

Table 6 Distribution of $\hat{N} - N$ when MID was employed for change-point detection over 100 simulated multivariate data sequences under Scenario (S1), with the multivariate data exhibiting two settings of spatial dependence. We take $T = 1500, d \in \{30, 100\}, sp \in \{0.2, 0.5\}, N \in \{3, 20\}$ and the signal strength is taken to be equal to 2. The average ARI, d_H , and computational times are also given

Σ	N	sp	d	$\hat{N} - N$				ARI	d_H	Time (s)
				-1	0	1	≥ 2			
$\Sigma^{(1)}$	3	0.2	30	0	99	1	0	0.994	0.008	1.45
	3	0.2	100	0	92	5	3	0.985	0.038	4.53
	3	0.5	30	0	99	1	0	0.990	0.013	1.49
	3	0.5	100	0	95	3	2	0.973	0.045	4.86
	20	0.2	30	0	98	2	0	0.987	0.050	1.61
	20	0.2	100	49	50	1	0	0.957	0.479	4.76
	20	0.5	30	29	69	2	0	0.972	0.305	1.63
$\Sigma^{(2)}$	20	0.5	100	76	21	3	0	0.932	0.764	4.78
	3	0.2	30	0	98	2	0	0.996	0.005	1.27
	3	0.2	100	0	95	5	0	0.986	0.020	4.67
	3	0.5	30	0	98	2	0	0.988	0.018	1.25
	3	0.5	100	19	78	2	1	0.921	0.230	5.79
	20	0.2	30	0	99	0	1	0.982	0.053	1.35
	20	0.2	100	74	21	5	0	0.942	0.790	5.19
	20	0.5	30	79	21	0	0	0.952	0.714	1.39
	20	0.5	100	73	23	2	2	0.923	0.878	5.03

Similar results for Scenario (S2) (changes in the vector of the first order derivatives) are provided and discussed in the online supplement. From Tables 7 and 8 we can conclude that, with relatively minor overestimation in the case of Student-t noise when $d = 100$ and $N = 3$, MID maintains its strong practical performance under non-Gaussian noise and for various different scenarios involving the dimensionality of the given data, the sparsity level, and the true number of change-points.

5 Real data examples

5.1 UK House Price Index

In this section, the performance of our method is studied on monthly percentage changes in the UK house price index for all property types from January 2000 to January 2023 in twenty London Boroughs. The data are available from <http://landregistry.data.gov.uk/app/ukhpi> and they were last accessed in March 2023. We have a multivariate data sequence $X_t = (X_{t,1}, \dots, X_{t,20})^T$, with $t = 1, \dots, 276$. The boroughs used are: Barnet (Ba), Bexley (Be), Bromley (Br), Camden (Ca), Croydon (Cr), Ealing (Ea), Enfield (En), Greenwich (Gr), Hackney (Ha), Hammersmith and

Fulham (HF), Harrow (Har), Islington (Is), Kensington and Chelsea (KC), Lambeth (La), Merton (Me), Newham (Ne), Richmond upon Thames (RuT), Sutton (Su), Tower Hamlets (TH), and Wandsworth (Wa). Similar data have been investigated in Baranowski et al. (2019), Fryzlewicz (2014), Fryzlewicz (2020), and Anastasiou and Fryzlewicz (2022); however under a univariate setting. Figure 5 indicates the results when MID_{opt} of Sect. 3.4 was employed for the detection of changes under Scenario (S1) as explained in Sect. 2.2; with respect to the threshold constant, we used the optimal values as in Table 1 for $\alpha = 0.05$.

Our method identifies 27 change-points in the mean structure of the multivariate data sequence at hand. We have analysed the same data using the competing methods of Sect. 4. INSPECT detects 26 change-points at locations very similar to those detected by our method, DC detects one change-point at location 41, while SBS does not detect any change-points. We need to highlight, though, that in INSPECT, multiple change-points are estimated using a wild binary segmentation scheme, which, due to the randomness involved, does not necessarily detect the same change-points when it is employed more than once over the same data set.

5.2 The COVID-19 outbreak

The performance of our method is also investigated on data from the COVID-19 pandemic. In this case, we focus though on the detection of changes under Scenario (S2) as described in Sect. 2.2. The data under consideration consist of the daily number of new lab-confirmed COVID-19 cases in the four constituent countries of United Kingdom; England, Northern Ireland, Scotland, and Wales. The period under investigation is from 01/04/2020 until 30/04/2022; there are no data for Northern Ireland after the 15th of May, 2022. The data are available from <https://coronavirus.data.gov.uk> and they were last accessed on the 7th of March 2023. Based on the description, in this example we have a multivariate data sequence $X_t = (X_{t,1}, \dots, X_{t,4})^T$, with $t = 1, \dots, 760$. Due to the fact that the data are positive integer numbers, we perform the Anscombe transform, $a : \mathbb{N} \rightarrow \mathbb{R}$, with $a(x) = 2\sqrt{x + 3/8}$, to each $X_{t,j}$; this transform brings the distribution of the component data sequences closer to the Gaussian one with constant variance.

Our method has detected 32 change-points in the vector of the first partial derivatives (changes in the slope for the component univariate data sequences) which seem to capture the important movements in the data.

Table 7 Distribution of $\hat{N} - N$ when MID was employed for change-point detection over 100 simulated multivariate data sequences under Scenario (S1). The signal strength, as defined in (18), is equal to 2 for every change-point, and we present the results for different combinations involving the true number of change-points, the sparsity level, and the dimensionality of the data sequence at hand. The added noise at each component signal follows the $\text{Unif}(-\sqrt{3}, \sqrt{3})$ distribution. The average ARI, d_H , and computational times are also given

N	sp	d	$\hat{N} - N$					ARI	d_h	Time (s)
			-2	-1	0	1	2			
3	0.2	30	0	0	99	1	0	0.993	0.009	1.48
3	0.5	30	0	0	98	2	0	0.988	0.017	1.58
3	0.8	30	0	0	100	0	0	0.997	0.004	2.31
3	0.2	100	0	0	100	0	0	0.987	0.018	5.26
3	0.5	100	0	0	97	3	0	0.973	0.039	4.47
3	0.8	100	0	1	99	0	0	0.951	0.079	5.35

N	sp	d	$\hat{N} - N$					ARI	d_h	Time (s)
			≤ -10	$(-10, -2)$	$[-2, 2]$	$(2, 10)$	≥ 10			
20	0.2	30	0	0	100	0	0	0.987	0.048	1.41
20	0.5	30	0	0	100	0	0	0.968	0.369	1.49
20	0.8	30	0	0	100	0	0	0.944	0.709	1.62
20	0.2	100	0	0	100	0	0	0.947	0.644	4.87
20	0.5	100	0	0	100	0	0	0.925	0.869	5.10
20	0.8	100	0	0	100	0	0	0.922	0.858	4.74

Table 8 Distribution of $\hat{N} - N$ when MID was employed for change-point detection over 100 simulated multivariate data sequences under Scenario (S1). The signal strength, as defined in (18), is equal to 2 for every change-point, and we present the results for different combinations involving the true number of change-points, the sparsity level, and the dimensionality of the data sequence at hand. The added noise at each component signal follows the t_8 distribution and before applying MID we pre-average the data at every 5 observations. The average ARI, d_H , and computational times are also given

N	sp	d	$\hat{N} - N$					ARI	d_h	Time (s)
			-2	-1	0	1	≥ 2			
3	0.2	30	0	0	91	6	3	0.976	0.051	0.38
3	0.5	30	0	0	75	21	4	0.968	0.129	0.31
3	0.8	30	0	0	81	16	3	0.971	0.065	0.41
3	0.2	100	0	0	65	18	17	0.946	0.175	0.94
3	0.5	100	0	0	61	26	13	0.934	0.222	0.96
3	0.8	100	0	0	63	21	16	0.940	0.170	0.95

N	sp	d	$\hat{N} - N$					ARI	d_h	Time (s)
			≤ -10	$(-10, -2)$	$[-2, 2]$	$(2, 10)$	≥ 10			
20	0.2	30	0	0	100	0	0	0.917	0.097	0.61
20	0.5	30	0	0	99	1	0	0.910	0.156	0.46
20	0.8	30	0	0	99	1	0	0.901	0.330	0.47
20	0.2	100	0	0	96	4	0	0.902	0.302	1.56
20	0.5	100	0	0	100	0	0	0.885	0.562	1.34
20	0.8	100	0	0	100	0	0	0.883	0.680	1.41

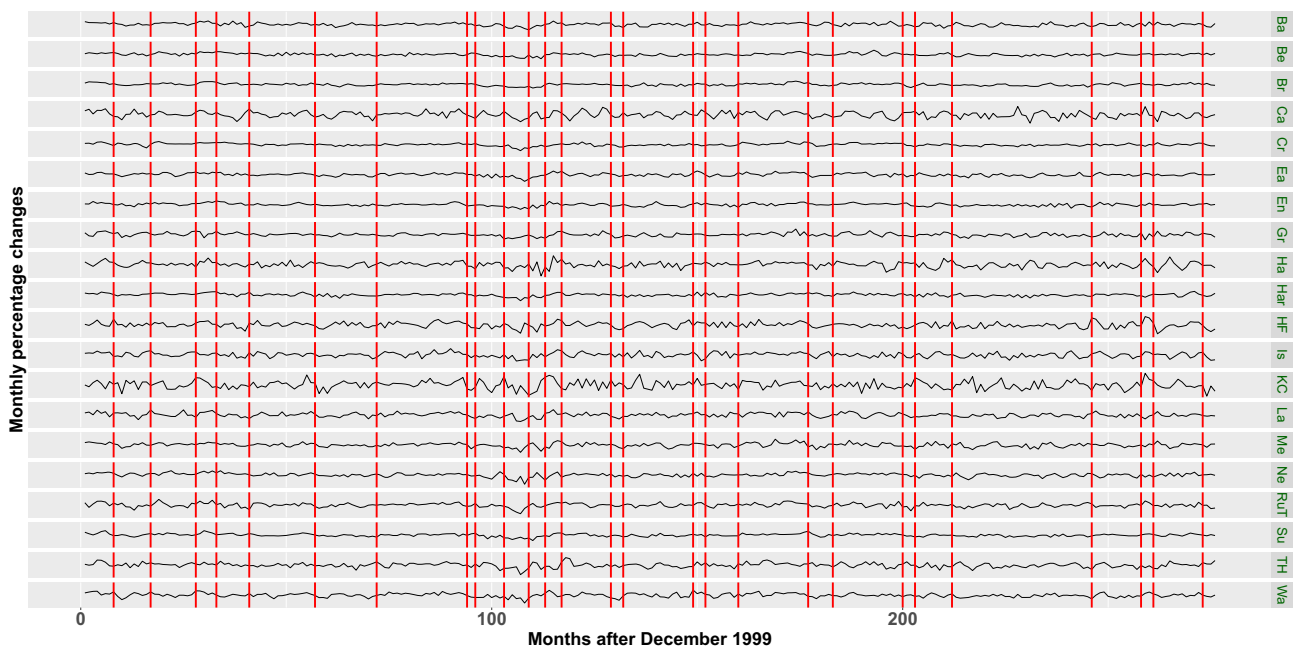


Fig. 5 The monthly percentage changes in the UK house price index for the twenty London boroughs under consideration. The estimated change-point locations when MID_{opt} was employed can be seen with red, vertical lines

6 Further discussion

6.1 Temporal dependence

In this section, we explore how our method can be employed for change-point detection in cases where the data exhibit temporal dependence. For $d = 1$, MID has already managed to earn independent praise regarding its robustness on moderate deviations from serial independence situations; see, for example, Fearnhead and Rigaiil (2020) where MID’s univariate analogue (the Isolate–Detect algorithm) is shown through an extensive simulation study to have very strong performance for a number of scenarios when either auto-correlated or heavy-tailed noise is added to the signals. Heuristically, such a good behaviour is expected because of the isolation aspect of MID, which ensures that the change-points will be detected one-by-one while they are still in intervals that contain no other change-points.

Practical ways that can be used in order to enhance MID’s performance in cases of very high serial correlation in the data, are the subsampling and pre-averaging techniques, which have already been described in detail in Section 5.1 of Anastasiou et al. (2022); here, we only give a brief idea of the steps involved. Starting with subsampling, the strategy is to choose a positive integer s and subsample every s data points from our original multivariate data sequence X_t , this will create s mutually exclusive (sub-sampled) data sequences; the autocorrelation in these new data sequences is expected to be lower than that in the original data. The

next step is to apply MID on each one of the s created data sequences. This will give back s different sets of estimated change-points. A majority voting rule is applied to these sets and we keep only those estimation values that appear at least a pre-decided amount of times, η , with $\eta \leq s$. Once the estimated change-point values, based on this majority voting rule, are extracted, the change-points are then transformed to represent the change-point locations with respect to the original data sequence. Regarding pre-averaging, which we have already applied in Sect. 4 in order to deal with heavy-tailed noise structures, the data are uniformly averaged over pre-specified, short time periods; this can significantly reduce autocorrelation. After the aforementioned pre-processing, MID is applied to the new, pre-averaged data.

6.2 Concluding reflections

In this paper, the MID methodology has been proposed for multiple generalized change-point detection in multivariate, possibly high-dimensional, data sequences, which could exhibit spatial dependence. Mean-dominant norms (see (2)) are employed for the aggregation of the information across the different components of the multivariate data sequence. The aggregated values for the relevant contrast function (depending on the structure of the changes) are collected and compared to a threshold value, $\zeta_{T,d}$. The rate of $\zeta_{T,d}$ with respect to both the length, T , of the data sequence and its dimensionality, d , has been proven to be $\mathcal{O}(\sqrt{\log(Td^{1/4})})$. The optimal multiplicative threshold constant, C , so that

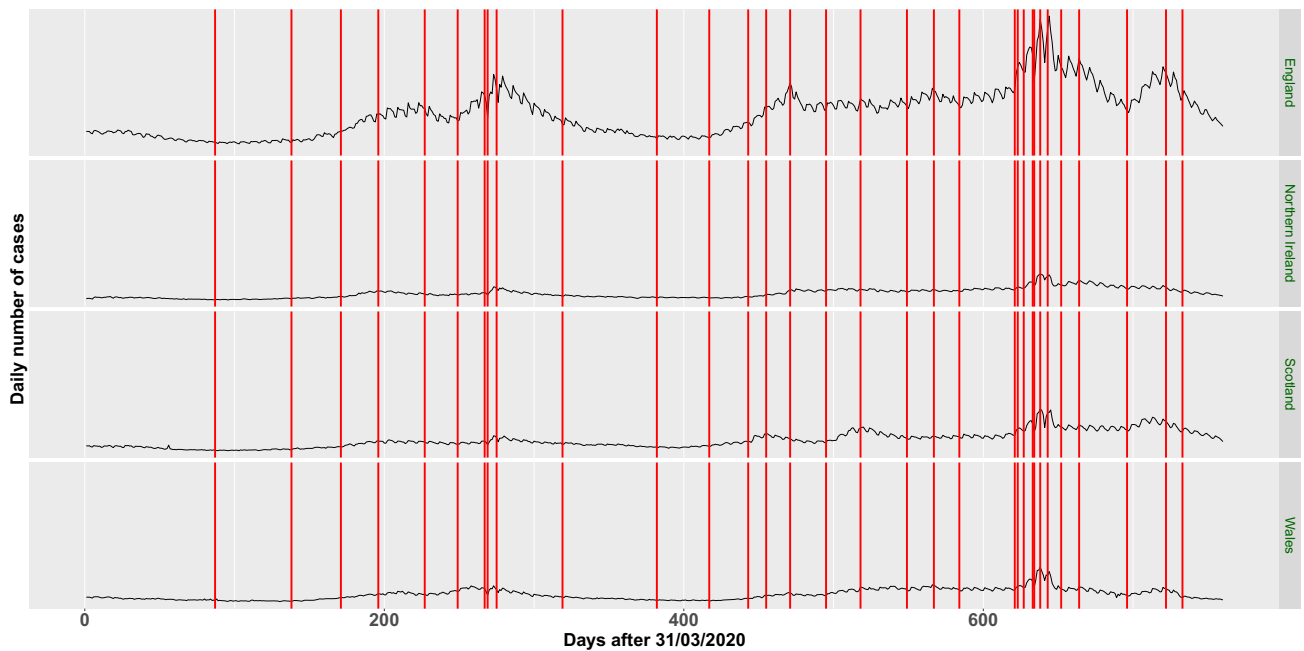


Fig. 6 The daily number of COVID-19 cases for England, Northern Ireland, Scotland, and Wales. The estimated change-point locations in the trend of each component data sequence when our proposed MID_{opt} method was employed can be seen with red, vertical lines

$\zeta_{T,d} = C\sqrt{\log(Td^{1/4})}$ has been derived through a large scale simulation study, with special attention in controlling the Type I error rate, α , of falsely detecting change-points for various values regarding the dimensionality of the data sequence.

Misspecification of the threshold is possible and could lead to misestimation of the underlying signal. To solve such issues, in Sect. 3.5, a permutation-based variant of our MID algorithm has been proposed with no threshold choice requirement. The algorithmic steps of MID remain the same; the difference lies in the way the method chooses to accept or reject a change-point within the interval under consideration, suppose this is denoted by $I = [s^*, e^*]$, where $1 \leq s^* < e^* \leq T$. In Sect. 2.1, we provide pseudocode for the better understanding of the method and it has been explained in detail that MID will first return a vector $\mathbf{v} \in \mathbb{R}^J$, where J is the amount of all change-point candidates. The elements of \mathbf{v} correspond to the aggregated contrast function values for each candidate point in the interval I . The next step is to store the maximum value of \mathbf{v} as obtained from the original data and compare it to the appropriate quantiles of the empirical distribution of the values obtained when applying the same steps to several permuted versions of the data; see the exact steps in Section 3.5. The proposed permutation-based variant of MID, though computationally expensive, performs very well in terms of accuracy with respect to the estimated number and locations of the change-points; see the results in Sect. 4.

The choice of the mean-dominant norm to be employed in the aggregation step of MID has already been discussed in the current paper, more specifically in Sect. 3.4. Our aim has been to provide a method that in practice would require minimal parameter choice to the user; towards this purpose, a data-adaptive variant, named MID_{opt} , of the method has been constructed. We first estimate the sparsity level in the given multivariate data; the steps are explained in detail in Sect. 3.4. Depending on the value of the estimated sparsity, MID_{opt} estimates the change-points employing either the L_∞ or the L_2 mean-dominant norm as defined in (2).

Through simulated and real-life data examples presented in Sects. 4 and 5, respectively, it has been shown that MID has very good performance in terms of accuracy and speed. Specifically, MID lies in the top 5% (in terms of the accurate estimation of the number and the location of the change-points) of the best methods when compared to the state-of-the-art competitors. In addition, MID_{opt} is a very quick detection method which in a few seconds can analyse signals with length in the range of thousands and dimensionality in the range of hundreds; this is carried out automatically with minimal decision making from the user on the aggregation method to be employed; more details on the computational complexity of our proposed algorithm can be found in Sect. 3.1.

MID has been proven to be a consistent method, with near-optimal rate, in accurately estimating the true number and the locations of the change-points. The consistency result also holds at the presence of spatial dependence between the

component data sequences; the very good practical performance of the proposed algorithm (without any modification) in scenarios of multivariate data sequences that exhibit spatial dependence can be found in Table 6. Regarding temporal dependence, the method of proof for the consistency result requires that the data are independent over time. In the univariate case, the method has already been shown (and has acquired independent praise on the matter) to be robust in the presence of auto-correlated data. In order to reduce the effect that, possibly strong, temporal dependence can have on our method’s accuracy in practice, in Sect. 6.1, we provide an explanation of two different approaches that can be followed prior to applying MID to the data; the first one is based on a sub-sampling scheme, while the second one requires to pre-average the given data.

No algorithm is perfect and we are now at the point to present limitations of MID regarding its practical behaviour. Firstly, the method can be slow in situations of long signals that do not exhibit any changes. The reason behind this behaviour is that, due to its expanding intervals characteristic, in such scenarios MID will keep testing for change-points in growing, overlapping intervals pushing the method to be slower than usual. In these settings one can eliminate the aforementioned MID weakness by splitting the data uniformly into smaller windows and then separately detect the change-points within each window. Even though MID has been shown to be quite robust in moderate deviations from Gaussianity (see, for example Table 7), the method seems to require some light data pre-processing in cases of heavy-tailed noise (see Table 8) or in the presence of temporal dependence.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s11222-023-10261-z>.

Acknowledgements We thank Piotr Fryzlewicz for constructive conversations. We also thank the Editor, the Associate Editor, and the two anonymous referees for their helpful comments which have lead to a significant improvement of the paper. This research occurred whilst Angelos Papanastasiou was employed as a Research Special Scientist at the University of Cyprus, supported by the start-up funding that Andreas Anastasiou has received from the University of Cyprus.

Author Contributions A.A. came up with the main idea of the proposed methodology, carried out the proofs of the theoretical results in the paper and the supplement, worked on the real-data analysis, reviewed the simulations, and has been regularly reviewing the manuscript. A.P. ran the simulations, came up with the permutation-based variant of the method, and drafted the manuscript. In addition, A.P. has created the GitHub repository and added all the R code which is useful for the replication of the results in the paper. Both authors have carefully reviewed the latest version of the manuscript.

Funding Partial financial support was received from the University of Cyprus (start-up grant of A.A).

Data Availability In Sect. 5, we provide the relevant links where all data and materials used in the paper can be obtained from.

Declarations

Conflict of interest The authors have no competing interests to declare that are relevant to the content of this article.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Code availability Code and instructions to replicate the results in the paper are available on <https://github.com/apapan08/Simulations-MID>.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

Appendix A Proof of Theorem 1

For $s \leq b < e$, we denote by $\tilde{X}_{s,e}^b := (\tilde{X}_{s,e}^{b,1}, \tilde{X}_{s,e}^{b,2}, \dots, \tilde{X}_{s,e}^{b,d})$, where $\tilde{X}_{s,e}^{b,j}$ is as in (5). Furthermore, $\tilde{f}_{s,e}^b := (\tilde{f}_{s,e}^{b,1}, \tilde{f}_{s,e}^{b,2}, \dots, \tilde{f}_{s,e}^{b,d})$, where $\tilde{f}_{s,e}^{b,j} = \sqrt{\frac{e-b}{n(b-s+1)}} \sum_{t=s}^b f_{t,j} - \sqrt{\frac{b-s+1}{n(e-b)}} \sum_{t=b+1}^e f_{t,j}$ is the value of the CUSUM statistic for the j th underlying component signal. Furthermore, for $L(\cdot)$ any mean-dominant norm as those in (2), and for $|\mathbf{x}|$ being the vector of the absolute value of the elements of an $\mathbf{x} \in \mathbb{R}^d$

$$C_{s,e}^b = L \left(\left| \tilde{X}_{s,e}^b \right| \right), \quad D_{s,e}^b = L \left(\left| \tilde{f}_{s,e}^b \right| \right),$$

$$A_T^* = \left\{ \max_{1 \leq s \leq b < e \leq T} L \left(\left| \tilde{X}_{s,e}^b - \tilde{f}_{s,e}^b \right| \right) \leq \sqrt{8 \log \left(T d^{\frac{1}{4}} \right)} \right\},$$

$$A_T = \left\{ \max_{1 \leq s \leq b < e \leq T} \left| C_{s,e}^b - D_{s,e}^b \right| \leq \sqrt{8 \log \left(T d^{\frac{1}{4}} \right)} \right\}. \tag{A1}$$

Proof STEP 1: For any mean-dominant norm $L(\cdot)$, using the mean dominance property as given in p.190 of Carlstein (1988), it holds that

$$0 \leq L_1(\mathbf{x}) \leq L(\mathbf{x}) \leq L_\infty(\mathbf{x}), \quad \forall \mathbf{x} \in (\mathbb{R}^d)^+.$$

Furthermore,

$$\begin{aligned} |C_{s,e}^b - D_{s,e}^b| &= \left| L \left(\left| \tilde{X}_{s,e}^b \right| \right) - L \left(\left| \tilde{f}_{s,e}^b \right| \right) \right| \\ &\leq \left| L \left(\left| \tilde{X}_{s,e}^b - \tilde{f}_{s,e}^b \right| \right) \right|. \end{aligned}$$

and it is straightforward to see that $\mathbb{P}(A_T) \geq \mathbb{P}(A_T^*)$. We will first show that

$$\mathbb{P}(A_T^*) \geq 1 - 1/(12\sqrt{\pi}T).$$

It suffices to show that $\mathbb{P}((A_T^*)^c) \leq 1/(12\sqrt{\pi}T)$. From the definition of our model in (1), we have that $\epsilon_t \sim \mathcal{N}_d(\mathbf{0}, \Sigma), \forall t \in \{1, \dots, T\}$. Denoting by $\sigma_j^2 := \Sigma_{j,j}, \forall j \in \{1, \dots, d\}$, it is straightforward that $\epsilon_{t,j} \sim \mathcal{N}(0, \sigma_j^2)$. The elements $\sigma_j, j = 1, \dots, d$, are of course, not assumed to be known; whenever they are unknown, then in the case of temporally independent errors with the univariate signal $f_{t,j}$ being piecewise-constant or piecewise linear (the scenarios covered in this work) then $\sigma_j, j = 1, \dots, d$, are estimated using the Median Absolute Deviation method explained in Hampel (1974). For simplicity, and without loss of generality, let $\sigma_j = 1, j = 1, \dots, d$. Then, from the equation of our model in (1), we have that for any $j \in \{1, \dots, d\}, \tilde{X}_{s,e}^{b,j} - \tilde{f}_{s,e}^{b,j} = \tilde{\epsilon}_{s,e}^{b,j}$, and simple calculations lead to $\tilde{\epsilon}_{s,e}^{t,j} \sim \mathcal{N}(0, 1)$. Therefore, for $\mathbf{Z} = (Z_1, Z_2, \dots, Z_d)$, with $Z_j \sim \mathcal{N}(0, 1)$, the Bonferroni inequality yields

$$\begin{aligned} \mathbb{P}((A_T^*)^c) &= \mathbb{P} \left(\max_{1 \leq s \leq b < e \leq T} L \left(\left| \tilde{X}_{s,e}^b - \tilde{f}_{s,e}^b \right| \right) > \sqrt{8 \log(Td^{\frac{1}{4}})} \right) \\ &= \mathbb{P} \left(\max_{1 \leq s \leq b < e \leq T} L \left(\left| \tilde{\epsilon}_{s,e}^b \right| \right) > \sqrt{8 \log(Td^{\frac{1}{4}})} \right) \\ &= \mathbb{P} \left(\max_{1 \leq s \leq b < e \leq T} L(|\mathbf{Z}|) > \sqrt{8 \log(Td^{\frac{1}{4}})} \right) \\ &\leq \mathbb{P} \left(\max_{1 \leq s \leq b < e \leq T} L_\infty(|\mathbf{Z}|) > \sqrt{8 \log(Td^{\frac{1}{4}})} \right) \\ &\leq \sum_{1 \leq s \leq b < e \leq T} \mathbb{P} \left(\max_{1 \leq j \leq d} |Z_j| > \sqrt{8 \log(Td^{\frac{1}{4}})} \right) \\ &\leq \frac{T^3}{6} \mathbb{P} \left(\max_{1 \leq j \leq d} |Z_j| > \sqrt{8 \log(Td^{\frac{1}{4}})} \right) \\ &\leq \frac{T^3}{6} \sum_{j=1}^d \mathbb{P} \left(|Z_j| > \sqrt{8 \log(Td^{\frac{1}{4}})} \right) \\ &= \frac{T^3}{3} \sum_{j=1}^d \mathbb{P} \left(Z_j > \sqrt{8 \log(Td^{\frac{1}{4}})} \right) \\ &= \frac{dT^3}{3} \frac{\phi \left(\sqrt{8 \log(Td^{\frac{1}{4}})} \right)}{\sqrt{8 \log(Td^{\frac{1}{4}})}} = \frac{1}{3T\sqrt{2\pi}\sqrt{8 \log(Td^{\frac{1}{4}})}} \\ &\leq \frac{1}{12\sqrt{\pi}T}, \end{aligned} \tag{A2}$$

where $\phi(x), x \in \mathbb{R}$ is the probability density function of the standard normal distribution. We conclude that $\mathbb{P}(A_T^*) \geq 1 - 1/(12\sqrt{\pi}T)$.

STEP 2: Let us denote by $\psi_{s,e}^b = (\psi_{s,e}^b(1), \dots, \psi_{s,e}^b(T))$ the contrast vector, where

$$\psi_{s,e}^b(t) = \begin{cases} \sqrt{\frac{e-b}{n(b-s+1)}}, & t = s, \dots, b, \\ -\sqrt{\frac{b-s+1}{n(e-b)}}, & t = b+1, \dots, e, \\ 0, & \text{otherwise,} \end{cases}$$

for $n = e - s + 1$. With $\mathbf{f}_k = (f_{1,k}, \dots, f_{T,k})$ and for $[s, e]$ being any interval that contains only one true change-point, namely r_j , allow us for ease of presentation to introduce the notation

$$A_{s,e}^b(k, r_j) := \frac{\langle \psi_{s,e}^b, \mathbf{f}_k \rangle - \psi_{s,e}^{r_j} \langle \mathbf{f}_k, \psi_{s,e}^{r_j} \rangle, \epsilon}{\left\| \psi_{s,e}^b \langle \mathbf{f}_k, \psi_{s,e}^b \rangle - \psi_{s,e}^{r_j} \langle \mathbf{f}_k, \psi_{s,e}^{r_j} \rangle \right\|_2}. \tag{A3}$$

In (A3) above, $k \in \{1, \dots, d\}$, while $\|\cdot\|_2$ is the Euclidean norm. Due to $\epsilon_{t,k}, t = 1, \dots, T$ following the standard normal distribution for all $k \in \{1, \dots, d\}$ (with the univariate error terms being independent over time, but possibly spatially dependent), then straightforward calculations lead to

$$A_{s,e}^b(k, r_j) \sim \mathcal{N}(0, 1), \forall k \in \{1, \dots, d\}. \tag{A4}$$

For

$$B_T = \left\{ \max_{\substack{j=1, \dots, N \\ k=1, \dots, d}} \max_{\substack{r_{j-1} < s \leq r_j \\ r_j < e \leq r_{j+1} \\ s \leq b < e}} |A_{s,e}^b(k, r_j)| \leq \sqrt{8 \log(Td^{\frac{1}{4}})} \right\}, \tag{A5}$$

we will show that $\mathbb{P}(B_T) \geq 1 - 1/(12\sqrt{\pi}T)$. Using (A4) and the Bonferroni inequality, then similar steps as in (A2) yield

$$\mathbb{P}(B_T^c) \leq \frac{1}{12\sqrt{\pi}T}. \tag{A6}$$

From (A2) and (A6) in STEPS 1 and 2, respectively, we conclude that $\mathbb{P}(A_T^* \cap B_T) \geq 1 - \frac{1}{6\sqrt{\pi}T}$.

STEP 3: This is the main part of our proof, where we explain in detail how to get the results in (7) and (8). For ease of understanding, we split this step into two smaller parts. From now on, we assume that A_T^* (and therefore also A_T) and B_T both hold. The constants we use are

$$c_1 = \sqrt{c_3} + \sqrt{8}, \quad c_2 = \frac{1}{\sqrt{6}} - \frac{2\sqrt{2}}{c}, \quad c_3 = 2(2\sqrt{2} + 4)^2,$$

where c is as in condition (A1).

STEP 3.1: In this step, we will show that there exists an interval where each change-point is isolated and detected.

For ease of presentation, we take $\lambda_T \leq \delta_T/3$; for more information on the general case of $\lambda_T \leq \delta_T/m$, for an $m > 1$, see Remark 1 in the online supplementary material of Anastasiou and Fryzlewicz (2022). Allow us now $\forall j \in \{1, \dots, N\}$, to define the intervals

$$I_j^R = \left[r_j + \frac{\delta_T}{3}, r_j + 2\frac{\delta_T}{3} \right)$$

$$I_j^L = \left(r_j - 2\frac{\delta_T}{3}, r_j - \frac{\delta_T}{3} \right]. \tag{A7}$$

It is apparent that in order for I_j^R and I_j^L to have at least one point, then we actually implicitly require that $\delta_T > 3$, which is the case for sufficiently large T . Since the length of the intervals in (A7) is equal to $\delta_T/3$ and $\lambda_T \leq \delta_T/3$, then MID ensures that for $K = \lceil T/\lambda_T \rceil$ and $k, m \in \{1, \dots, K\}$, there exists at least one $c_k^r = k\lambda_T$ and at least one $c_m^l = T - m\lambda_T + 1$ that are in I_j^R and I_j^L , respectively, $\forall j = 1, \dots, N$.

At the beginning of our algorithm, $s = 1, e = T$ (see, for example, Fig. 2) and depending on whether $r_1 \leq T - r_N$ then r_1 or r_N will get isolated in a right- or left-expanding interval, respectively. W.l.o.g., assume that $r_1 \leq T - r_N$ and, as already mentioned, our algorithm naturally ensures that $\exists k \in \{1, \dots, K\}$ such that $c_k^r \in I_1^R$. We highlight that there is no other change-point in $[1, c_k^r]$ apart from r_1 . With $C_{s,e}^b$ and $D_{s,e}^b$ as in (A1), we will show that for $\tilde{b} = \operatorname{argmax}_{1 \leq t < c_k^r} C_{1,c_k^r}^t$, then $C_{1,c_k^r}^{\tilde{b}} > \zeta_{T,d}$. Using (A1), we have that

$$C_{1,c_k^r}^{\tilde{b}} \geq C_{1,c_k^r}^{r_1} \geq D_{1,c_k^r}^{r_1} - \sqrt{8 \log \left(Td^{\frac{1}{4}} \right)}$$

$$= L \left(\left| \sqrt{\frac{r_1(c_k^r - r_1)}{c_k^r}} (f_{r_1} - f_{r_2}) \right| \right) - \sqrt{8 \log \left(Td^{\frac{1}{4}} \right)}$$

$$= \sqrt{\frac{r_1(c_k^r - r_1)}{c_k^r}} L(\Delta_1) - \sqrt{8 \log \left(Td^{\frac{1}{4}} \right)}$$

$$\geq \sqrt{\frac{\min \{c_k^r - r_1, r_1\}}{2}} L(\Delta_1) - \sqrt{8 \log \left(Td^{\frac{1}{4}} \right)}. \tag{A8}$$

with $\Delta_j, j = 1, 2, \dots, N$ defined in (6). For δ_T as in (6), we notice that $r_1 \geq \delta_T$. Furthermore, since $c_k^r \in I_1^R$ as in (A7), then $c_k^r - r_1 \geq \delta_T/3$, meaning that $\min \{c_k^r - r_1, r_1\} \geq \frac{\delta_T}{3}$. Therefore, using assumption (A1),

$$C_{1,c_k^r}^{\tilde{b}} \geq \sqrt{\frac{\delta_T}{6}} L(\Delta_1) - \sqrt{8 \log \left(Td^{\frac{1}{4}} \right)}$$

$$\geq \sqrt{\frac{\delta_T}{6}} \underline{f}_T - \sqrt{8 \log \left(Td^{\frac{1}{4}} \right)}$$

$$= \left(\frac{1}{\sqrt{6}} - \frac{\sqrt{8 \log \left(Td^{1/4} \right)}}{\sqrt{\delta_T} \underline{f}_T} \right) \sqrt{\delta_T} \underline{f}_T$$

$$\geq \left(\frac{1}{\sqrt{6}} - \frac{\sqrt{8 \log \left(Td^{1/4} \right)}}{C \sqrt{\log \left(Td^{1/4} \right)}} \right) \sqrt{\delta_T} \underline{f}_T$$

$$= \left(\frac{1}{\sqrt{6}} - \frac{2\sqrt{2}}{C} \right) \sqrt{\delta_T} \underline{f}_T = C_2 \sqrt{\delta_T} \underline{f}_T > \zeta_{T,d}. \tag{A9}$$

Therefore, there will be an interval of the form $[1, c_k^r]$, with $c_k^r > r_1$, such that $[1, c_k^r]$ contains only r_1 and $\max_{1 \leq b < c_k^r} \left\{ C_{1,c_k^r}^b \right\} > \zeta_{T,d}$.

STEP 3.2: Let us, for $k^* \in \{1, \dots, K\}$, to denote by $c_{k^*}^r \leq c_k^r$ the first right-expanding point where isolation and detection of r_1 as in STEP 3.1 occurs, and let $b_1 = \operatorname{argmax}_{1 \leq t < c_{k^*}^r} \left\{ C_{1,c_{k^*}^r}^t \right\}$ with $\left\{ C_{1,c_{k^*}^r}^{b_1} \right\} > \zeta_{T,d}$. In STEP 3.2, we will show that the estimated change-point is within a close neighbourhood of the true change-point location. We split our proof into two scenarios, depending on the mean-dominant norm used.

Scenario 1, $L(\cdot) = L_\infty(\cdot)$: For ease of presentation, allow us to denote by $q_1 := \operatorname{argmax}_{j=1, \dots, d} \left| \tilde{X}_{1,c_{k^*}^r}^{b_1,j} \right|$; this is the univariate component data sequence where the contrast function value got maximised at location b_1 . We will find $\gamma_T > 0$ such that for any $b^* \in \{1, \dots, c_{k^*}^r - 1\}$ with $|b^* - r_1| \left(\Delta_1^{q_1} \right)^2 > \gamma_T$, it holds that

$$\left(\tilde{X}_{1,c_{k^*}^r}^{r_1,q_1} \right)^2 > \left(\tilde{X}_{1,c_{k^*}^r}^{b^*,q_1} \right)^2. \tag{A10}$$

Proving (A10) and using the definition of b_1 we can conclude that $|b_1 - r_1| \left(\Delta_1^{q_1} \right)^2 \leq \gamma_T$. Now, since in our model, $X_{t,j} = f_{t,j} + \epsilon_{t,j}$, then (A10) can be expressed as

$$\left(\tilde{f}_{1,c_{k^*}^r}^{r_1,q_1} \right)^2 - \left(\tilde{f}_{1,c_{k^*}^r}^{b^*,q_1} \right)^2$$

$$> \left(\tilde{\epsilon}_{1,c_{k^*}^r}^{b^*,q_1} \right)^2 - \left(\tilde{\epsilon}_{1,c_{k^*}^r}^{r_1,q_1} \right)^2$$

$$+ 2 \left\langle \tilde{\psi}_{1,c_{k^*}^r}^{b^*} \langle f_{q_1}, \tilde{\psi}_{1,c_{k^*}^r}^{b^*} \rangle - \tilde{\psi}_{1,c_{k^*}^r}^{r_1} \langle f_{q_1}, \tilde{\psi}_{1,c_{k^*}^r}^{r_1} \rangle, \epsilon_{q_1} \right\rangle. \tag{A11}$$

W.l.o.g. assume that $b^* \geq r_1$ and a similar approach as below holds when $b^* < r_1$. Using Lemma 4 in the online supplementary material of Baranowski et al. (2019), gives for the left-hand side of the inequality in (A11) that

$$\left(\tilde{f}_{1,c_{k^*}^r}^{r_1,q_1} \right)^2 - \left(\tilde{f}_{1,c_{k^*}^r}^{b^*,q_1} \right)^2 = \frac{|b^* - r_1| r_1}{|b^* - r_1| + r_1} \left(\Delta_1^{q_1} \right)^2 := \Lambda. \tag{A12}$$

For the terms on the right-hand side of (A11), since we are working under A_T^* as in (A1), we obtain that

$$\left(\tilde{\epsilon}_{1,c_{k^*}^r}^{b^*,q_1} \right)^2 - \left(\tilde{\epsilon}_{1,c_{k^*}^r}^{r_1,q_1} \right)^2 \leq \left(\max_{s \leq b < e} \left\{ \max_{1 \leq j \leq d} \left| \tilde{\epsilon}_{s,e}^{b,j} \right| \right\} \right)^2$$

$$- \left(\tilde{\epsilon}_{1,c_{k^*}^r}^{r_1,q_1} \right)^2$$

$$\leq \left(\max_{s \leq b < e} \left\{ \max_{1 \leq j \leq d} \left| \tilde{\epsilon}_{s,e}^{b,j} \right| \right\} \right)^2$$

$$\leq 8 \log \left(Td^{\frac{1}{4}} \right),$$

while from (A5) and Lemma 4 from the online supplementary material of Baranowski et al. (2019),

$$\begin{aligned} & 2 \left\langle \psi_{1,c_{k^*}^r}^{b^*} \langle f_{q_1}, \psi_{1,c_{k^*}^r}^{b^*} \rangle - \psi_{1,c_{k^*}^r}^{r_1} \langle f_{q_1}, \psi_{1,c_{k^*}^r}^{r_1} \rangle, \epsilon_{q_1} \right\rangle \\ & \leq 2 \|\psi_{1,c_{k^*}^r}^{b^*} \langle f_{q_1}, \psi_{1,c_{k^*}^r}^{b^*} \rangle - \psi_{1,c_{k^*}^r}^{r_1} \langle f_{q_1}, \psi_{1,c_{k^*}^r}^{r_1} \rangle\|_2 \\ & \quad \times \sqrt{8 \log(Td^{\frac{1}{4}})} \\ & = 2\sqrt{8\Lambda \log(Td^{\frac{1}{4}})}. \end{aligned}$$

Therefore (A11) is satisfied if the stronger inequality $\Lambda > 8 \log(Td^{1/4}) + 2\sqrt{8\Lambda \log(Td^{1/4})}$ is satisfied, which has solution

$$\Lambda > (2\sqrt{2} + 4)^2 \log(Td^{\frac{1}{4}}).$$

From (A12) and since

$$\frac{|b^* - r_1| r_1}{|b^* - r_1| + r_1} \geq \frac{1}{2} \min\{|b^* - r_1|, r_1\},$$

we deduce that (A10) is implied by

$$\begin{aligned} \min\{|b^* - r_1|, r_1\} & > \frac{2(2\sqrt{2} + 4)^2 \log(Td^{\frac{1}{4}})}{(\Delta_1^{q_1})^2} \\ & = \frac{C_3 \log(Td^{\frac{1}{4}})}{(\Delta_1^{q_1})^2}. \end{aligned} \tag{A13}$$

However,

$$\min\{r_1, c_{k^*}^r - r_1\} > C_3 \frac{\log(Td^{\frac{1}{4}})}{(\Delta_1^{q_1})^2} \tag{A14}$$

and this is because if we assume that $\min\{r_1, c_{k^*}^r - r_1\} \leq C_3 \log(Td^{\frac{1}{4}}) / (\Delta_1^{q_1})^2$, then

$$\begin{aligned} C_{1,c_{k^*}^r}^{b_1} & = |\tilde{X}_{1,c_{k^*}^r}^{b_1,q_1}| = |\tilde{X}_{1,c_{k^*}^r}^{b_1,q_1} - \tilde{f}_{1,c_{k^*}^r}^{b_1,q_1} + \tilde{f}_{1,c_{k^*}^r}^{b_1,q_1}| \\ & \leq \sqrt{8 \log(Td^{\frac{1}{4}})} + |\tilde{f}_{1,c_{k^*}^r}^{b_1,q_1}| \leq \sqrt{8 \log(Td^{\frac{1}{4}})} + |\tilde{f}_{1,c_{k^*}^r}^{r_1,q_1}| \\ & = \sqrt{8 \log(Td^{\frac{1}{4}})} + \sqrt{\frac{r_1(c_{k^*}^r - r_1)}{c_{k^*}^r} \Delta_1^{q_1}} \\ & \leq \sqrt{8 \log(Td^{\frac{1}{4}})} + \sqrt{\min\{c_{k^*}^r - r_1, r_1\} \Delta_1^{q_1}} \\ & \leq (\sqrt{C_3} + \sqrt{8}) \sqrt{\log(Td^{\frac{1}{4}})} = C_1 \sqrt{\log(Td^{\frac{1}{4}})} \\ & \leq \zeta_{T,d}. \end{aligned}$$

This comes to a contradiction to $C_{1,c_{k^*}^r}^{b_1} > \zeta_{T,d}$, which has already been proven in (A9). Therefore, (A14) holds and (A13) is restricted to $|b^* - r_1| (\Delta_1^{q_1})^2 > C_3 \log T$, which implies (A10). Therefore, necessarily

$$|b_1 - r_1| (\Delta_1^{q_1})^2 \leq C_3 \log(Td^{\frac{1}{4}}). \tag{A15}$$

Scenario 2, $L(\cdot) = L_2(\cdot)$: We will find $\gamma_T > 0$ such that for any $b^* \in \{1, \dots, c_{k^*}^r - 1\}$ with $|b^* - r_1| L_2^2(\Delta_1) > \gamma_T$, it holds that

$$L_2^2(\tilde{X}_{1,c_{k^*}^r}^{r_1}) > L_2^2(\tilde{X}_{1,c_{k^*}^r}^{b^*}). \tag{A16}$$

Proving (A16) and using the definition of b_1 we can conclude that $|b_1 - r_1| L_2^2(\Delta_1) \leq \gamma_T$. Now, since in our model, $X_t = f_t + \epsilon_t$, then (A16) can be expressed as

$$\begin{aligned} & L_2^2(\tilde{f}_{1,c_{k^*}^r}^{r_1}) - L_2^2(\tilde{f}_{1,c_{k^*}^r}^{b^*}) \\ & > L_2^2(\tilde{\epsilon}_{1,c_{k^*}^r}^{b^*}) - L_2^2(\tilde{\epsilon}_{1,c_{k^*}^r}^{r_1}) \\ & + \frac{2}{d} \sum_{j=1}^d \left\langle \psi_{1,c_{k^*}^r}^{b^*} \langle f_j, \psi_{1,c_{k^*}^r}^{b^*} \rangle - \psi_{1,c_{k^*}^r}^{r_1} \langle f_j, \psi_{1,c_{k^*}^r}^{r_1} \rangle, \epsilon_j \right\rangle. \end{aligned} \tag{A17}$$

As in Scenario 1 earlier, w.l.o.g. assume that $b^* \geq r_1$ and a similar approach as below holds when $b^* < r_1$. Using Lemma 4 in the online supplementary material of Baranowski et al. (2019), gives for the left-hand side of the inequality in (A17) that

$$\begin{aligned} L_2^2(\tilde{f}_{1,c_{k^*}^r}^{r_1}) - L_2^2(\tilde{f}_{1,c_{k^*}^r}^{b^*}) & = \frac{|b^* - r_1| r_1}{|b^* - r_1| + r_1} L_2^2(\Delta_1) \\ & := \Lambda. \end{aligned} \tag{A18}$$

For the terms on the right-hand side of (A17), since we are working under A_7^* as in (A1), we obtain that

$$\begin{aligned} & L_2^2(\tilde{\epsilon}_{1,c_{k^*}^r}^{b^*}) - L_2^2(\tilde{\epsilon}_{1,c_{k^*}^r}^{r_1}) \\ & \leq \max_{s \leq b < e} \left\{ L_2^2(\tilde{\epsilon}_{1,c_{k^*}^r}^b) \right\} \\ & = \max_{s \leq b < e} \left\{ L_2^2(\tilde{X}_{1,c_{k^*}^r}^b - \tilde{f}_{1,c_{k^*}^r}^b) \right\} \\ & \leq 8 \log(Td^{\frac{1}{4}}). \end{aligned}$$

while from (A5) and Lemma 4 from the online supplementary material of Baranowski et al. (2019),

$$\begin{aligned} & \frac{2}{d} \sum_{j=1}^d \left\langle \psi_{1,c_{k^*}^r}^{b^*} \langle f_j, \psi_{1,c_{k^*}^r}^{b^*} \rangle - \psi_{1,c_{k^*}^r}^{r_1} \langle f_j, \psi_{1,c_{k^*}^r}^{r_1} \rangle, \epsilon_j \right\rangle \\ & \leq \frac{2}{d} \sum_{j=1}^d \|\psi_{1,c_{k^*}^r}^{b^*} \langle f_j, \psi_{1,c_{k^*}^r}^{b^*} \rangle - \psi_{1,c_{k^*}^r}^{r_1} \langle f_j, \psi_{1,c_{k^*}^r}^{r_1} \rangle\|_2 \end{aligned}$$

$$\begin{aligned}
 & \times \sqrt{8 \log \left(T d^{\frac{1}{4}}\right)} \\
 &= \frac{2}{d} \sqrt{8 \log \left(T d^{\frac{1}{4}}\right)} \sum_{j=1}^d \sqrt{\left(\tilde{f}_{1, c_{k^*}^r}^{r_1, j}\right)^2 - \left(\tilde{f}_{1, c_{k^*}^r}^{b^*, j}\right)^2} \\
 &= \frac{2}{d} \sqrt{8 \log \left(T d^{\frac{1}{4}}\right)} \sum_{j=1}^d \sqrt{\frac{\left|b^* - r_1\right| r_1}{\left|b^* - r_1\right| + r_1}} \left(\Delta_1^j\right)^2 \\
 &= \frac{2}{d} \sqrt{8 \log \left(T d^{\frac{1}{4}}\right)} \sqrt{\frac{\left|b^* - r_1\right| r_1}{\left|b^* - r_1\right| + r_1}} \sum_{j=1}^d \Delta_1^j \\
 &\leq 2 \sqrt{8 \log \left(T d^{\frac{1}{4}}\right)} \sqrt{\frac{\left|b^* - r_1\right| r_1}{\left|b^* - r_1\right| + r_1}} L_2\left(\Delta_1\right) \\
 &= 2 \sqrt{8 \Lambda \log \left(T d^{\frac{1}{4}}\right)},
 \end{aligned}$$

where for the last inequality, we used the mean-dominance norm property as given in p.190 of Carlstein (1988). Therefore (A17) is satisfied if the stronger inequality $\Lambda > 8 \log \left(T d^{1/4}\right) + 2 \sqrt{8 \Lambda \log \left(T d^{1/4}\right)}$ is satisfied, which has solution

$$\Lambda > \left(2 \sqrt{2} + 4\right)^2 \log \left(T d^{\frac{1}{4}}\right).$$

From (A18) and since

$$\frac{\left|b^* - r_1\right| r_1}{\left|b^* - r_1\right| + r_1} \geq \frac{1}{2} \min \left\{\left|b^* - r_1\right|, r_1\right\},$$

we deduce that (A16) is implied by

$$\begin{aligned}
 \min \left\{\left|b^* - r_1\right|, r_1\right\} &> \frac{2\left(2 \sqrt{2} + 4\right)^2 \log \left(T d^{\frac{1}{4}}\right)}{L_2^2\left(\Delta_1\right)} \\
 &= \frac{C_3 \log \left(T d^{\frac{1}{4}}\right)}{L_2^2\left(\Delta_1\right)}.
 \end{aligned} \tag{A19}$$

However,

$$\min \left\{r_1, c_{k^*}^r - r_1\right\} > C_3 \frac{\log \left(T d^{\frac{1}{4}}\right)}{L_2^2\left(\Delta_1\right)} \tag{A20}$$

and this is because if we assume that $\min \left\{r_1, c_{k^*}^r - r_1\right\} \leq C_3 \log \left(T d^{\frac{1}{4}}\right) / L_2^2\left(\Delta_1\right)$, then

$$\begin{aligned}
 C_{1, c_{k^*}^r}^{b_1} &= L_2\left(\left|\tilde{X}_{1, c_{k^*}^r}^{b_1}\right|\right) = L_2\left(\left|\tilde{X}_{1, c_{k^*}^r}^{b_1} - \tilde{f}_{1, c_{k^*}^r}^{b_1} + \tilde{f}_{1, c_{k^*}^r}^{b_1}\right|\right) \\
 &\leq \sqrt{8 \log \left(T d^{\frac{1}{4}}\right)} + L_2\left(\left|\tilde{f}_{1, c_{k^*}^r}^{b_1}\right|\right) \\
 &\leq \sqrt{8 \log \left(T d^{\frac{1}{4}}\right)} + L_2\left(\left|\tilde{f}_{1, c_{k^*}^r}^{r_1}\right|\right)
 \end{aligned}$$

$$\begin{aligned}
 &= \sqrt{8 \log \left(T d^{\frac{1}{4}}\right)} + \sqrt{\frac{r_1\left(c_{k^*}^r - r_1\right)}{c_{k^*}^r}} L_2\left(\Delta_1\right) \\
 &\leq \sqrt{8 \log \left(T d^{\frac{1}{4}}\right)} + \sqrt{\min \left\{c_{k^*}^r - r_1, r_1\right\}} L_2\left(\Delta_1\right) \\
 &\leq \left(\sqrt{C_3} + \sqrt{8}\right) \sqrt{\log \left(T d^{\frac{1}{4}}\right)} = C_1 \sqrt{\log \left(T d^{\frac{1}{4}}\right)} \\
 &\leq \zeta_{T, d}.
 \end{aligned}$$

This comes to a contradiction to $C_{1, c_{k^*}^r}^{b_1} > \zeta_{T, d}$, which has already been proven in (A9). Therefore, (A20) holds and (A19) is restricted to $\left|b^* - r_1\right| L_2^2\left(\Delta_1\right) > C_3 \log T$, which implies (A16). Therefore, necessarily

$$\left|b_1 - r_1\right| L_2^2\left(\Delta_1\right) \leq C_3 \log \left(T d^{\frac{1}{4}}\right). \tag{A21}$$

So far, for $\lambda_T \leq \delta_T / 3$, we have proven in STEPS 3.1 and 3.2 that working under the assumption that A_T^* (which implies that A_T also holds) and B_T hold, there will be an interval $\left[1, c_{k^*}^r\right]$, with $C_{1, c_{k^*}^r}^{b_1} > \zeta_{T, d}$, where $b_1 = \operatorname{argmax}_{1 \leq t < c_{k^*}^r} C_{1, c_{k^*}^r}^t$ is an estimation of r_1 that satisfies (A15) for $L(\cdot) = L_{\inf}(\cdot)$ and (A21) for $L(\cdot) = L_2(\cdot)$.

STEP 3.3: After detecting the first change-point, MID follows the same process as in STEPS 3.1 and 3.2, but now in the set $\left[c_{k^*}^r, T\right]$, which contains the change-points r_2, \dots, r_N . We do not check for possible change-points the interval $\left[b_1 + 1, c_{k^*}^r\right)$. However, this does not create any issues because:

- M.1 There is no change-point in $\left[b_1 + 1, c_{k^*}^r\right)$, apart from maybe the already detected r_1 ;
- M.2 $c_{k^*}^r$ is at a location which allows for detection of r_2 .

For M.1: We will split the explanation into two cases with respect to the location of b_1 .

Case 1: $b_1 < r_1 < c_{k^*}^r$. With q_1 as in Scenario 1 of STEP 3.2, we denote by

$$\alpha_1 := \begin{cases} \Delta_1^{q_1}, & \text{when } L(\cdot) = L_{\infty}(\cdot) \\ L_2\left(\Delta_1\right), & \text{when } L(\cdot) = L_2(\cdot) \end{cases},$$

using (A15) or (A21) (depending on the mean-dominant norm under consideration) with

$$\delta_T > 3 C_3 \frac{\log \left(T d^{\frac{1}{4}}\right)}{\alpha^2}, \tag{A22}$$

then since $c_k^r \in I_1^R$, we have that

$$\begin{aligned}
 c_{k^*}^r - b_1 &\leq c_k^r - b_1 = c_k^r - r_1 + r_1 - b_1 \\
 &< 2 \frac{\delta_T}{3} + r_1 - b_1
 \end{aligned}$$

$$\leq 2\frac{\delta_T}{3} + \frac{C_3 \log(Td^{1/4})}{\alpha^2} < \delta_T.$$

Since $r_2 - r_1 \geq \delta_T$ and r_1 is already in $[b_1 + 1, c_{k^*}^r)$, then there is no other change-point in $[b_1 + 1, c_{k^*}^r)$ apart from r_1 .

Case 2: $r_1 \leq b_1 < c_{k^*}^r$. By construction $c_{k^*}^r - r_1 < 2\delta_T/3$, which means that apart from r_1 there is no other change-point in $[r_1, c_{k^*}^r)$. With $r_1 \leq b_1$, then $[b_1 + 1, c_{k^*}^r)$ does not have any change-point.

Cases 1 and 2 above show that no matter the location of b_1 , there is no change-point in $[b_1 + 1, c_{k^*}^r)$ other than possibly the previously detected r_1 . Similarly to the approach in STEPS 3.1 and 3.2, our method applied now in $[s, e] = [c_{k^*}^r, T]$, will first isolate, and then detect, r_2 or r_N depending on whether $r_2 - c_{k^*}^r$ is smaller or larger than $T - r_N$. If $T - r_N < r_2 - c_{k^*}^r$ then r_N will get isolated first in a left-expanding interval and the procedure to show its detection is exactly the same as for the detection of r_1 in STEPS 3.1 and 3.2. Therefore, for the sake of showing M.2 let us assume that $r_2 - c_{k^*}^r \leq T - r_N$.

For M.2: By construction, there exists a right expanding point $c_{k_2}^r \in I_2^R$, with I_j^R defined in (A7). We will show that r_2 gets detected in $[c_{k_2}^r, c_{k_2^*}^r]$, for $k_2^* \leq k_2$ and its detection is $b_2 = \operatorname{argmax}_{c_{k^*}^r \leq t < c_{k_2^*}^r} C_{c_{k^*}^r, c_{k_2^*}^r}^t$, which, for $q_2 :=$

$$\operatorname{argmax}_{j=1,2,\dots,d} \left| \tilde{X}_{c_{k^*}^r, c_{k_2^*}^r}^{b_2, j} \right|, \text{ satisfies}$$

$$C_3 \log(Td^{1/4}) \geq \begin{cases} |b_2 - r_2| (\Delta_2^{q_2})^2, & \text{when } L(\cdot) = L_\infty(\cdot) \\ |b_2 - r_2| L_2^2(\mathbf{\Delta}_2), & \text{when } L(\cdot) = L_2(\cdot) \end{cases}$$

Following similar steps as in (A9), we have that

$$C_{c_{k^*}^r, c_{k_2^*}^r}^{b_2} > \zeta_{T,d}.$$

Now, with

$$\alpha_2 := \begin{cases} \Delta_2^{q_2}, & \text{when } L(\cdot) = L_\infty(\cdot) \\ L_2(\mathbf{\Delta}_2), & \text{when } L(\cdot) = L_2(\cdot) \end{cases},$$

it is straightforward to show, following the same process as in STEPS 3.1 and 3.2 of the proof, that $|b_2 - r_2| \alpha_2^2 \leq C_3 \log(Td^{1/4})$. Having detected r_2 , then our algorithm will proceed in the interval $[s, e] = [c_{k_2^*}^r, T]$ and all the change-points will get detected one by one since STEPS 3.1–3.3 will be applicable as long as there are undetected change-points in $[s, e]$.

Denoting by \hat{r}_j the estimation of r_j as we did in the statement of the theorem and for $[s^*, e^*]$ being the interval where the isolation and detection of r_j occurs (in the way that we explained in STEPS 3.1–3.3) allow us to denote by $q_j := \operatorname{argmax}_{m=1,2,\dots,d} \left| \tilde{X}_{s^*, e^*}^{\hat{r}_j, m} \right|$. Then, STEPS 3.1–3.3 have

shown that MID will detect all the change-points one by one and, $\forall j \in \{1, \dots, N\}$,

$$C_3 \log(Td^{1/4}) \geq \begin{cases} |\hat{r}_j - r_j| (\Delta_j^{q_j})^2, & \text{when } L(\cdot) = L_\infty(\cdot) \\ |\hat{r}_j - r_j| L_2^2(\mathbf{\Delta}_j), & \text{when } L(\cdot) = L_2(\cdot) \end{cases} \quad (\text{A23})$$

STEP 4: The arguments given in STEPS 1-3 hold in $A_T^* \cap B_T$. At the beginning of the algorithm, $s = 1, e = T$ and for $N \geq 1$, there exist $k_1 \in \{1, \dots, K\}$ such that $s_{k_1} = s, e_{k_1} \in I_1^R$ and $k_2 \in \{1, \dots, K\}$ such that $s_{k_2} \in I_N^L, e_{k_2} = e$. As in our previous steps, w.l.o.g. assume that $r_1 \leq T - r_N$ and r_1 gets isolated and detected first in an interval $[s, c_{k^*}^r]$, where $c_{k^*}^r$ is less than or equal to e_{k_1} . Then, $\hat{r}_1 = \operatorname{argmax}_{s \leq t < c_{k^*}^r} C_{s, c_{k^*}^r}^t$ is the estimated location for r_1 , which satisfies (A23). After this, the method continues in $[c_{k^*}^r, T]$ and keeps detecting all the change-points as explained in STEPS 3.1–3.3. There will not be any double detection issues because naturally, at each step of the algorithm, the new interval $[s, e]$ does not include any previously detected change-points. Once all the change-points have been detected one by one, then $[s, e]$ will contain no other change-points. Our method will of course keep checking for possible change-points in right- and left-expanding intervals denoted by $[s^*, e^*]$. However, MID will not detect anything in $[s^*, e^*]$ because $\forall b \in [s^*, e^*]$,

$$C_{s^*, e^*}^b \leq D_{s^*, e^*}^b + \sqrt{8 \log(Td^{1/4})} = \sqrt{8 \log(Td^{1/4})} < C_1 \sqrt{\log(Td^{1/4})} \leq \zeta_{T,d}.$$

After not detecting anything in all intervals of the above form, then the algorithm concludes that there are not any more change-points to be detected and stops. \square

References

Anastasiou, A., Fryzlewicz, P.: Detecting multiple generalized change-points by isolating single ones. *Metrika* **85**, 141–174 (2022)

Anastasiou, A., Cribben, I., Fryzlewicz, P.: Cross-covariance isolate detect: a new change-point method for estimating dynamic functional connectivity. *Med. Image Anal.* **75**, 102252 (2022)

Antoch, J., Hušková, M.: Permutation tests in change point analysis. *Stat. Probab. Lett.* **53**(1), 37–46 (2001)

Arlot, S., Celisse, A., Harchaoui, Z.: A kernel multiple change-point algorithm via model selection. *J. Mach. Learn. Res.* **20**, 1–56 (2019)

Auger, I.E., Lawrence, C.E.: Algorithms for the optimal identification of segment neighborhoods. *Bull. Math. Biol.* **51**, 39–54 (1989)

Baranowski, R., Chen, Y., Fryzlewicz, P.: Narrowest-over-threshold detection of multiple change points and change-point-like features. *J. R. Stat. Soc. Ser. B* **81**, 649–672 (2019)

- Bücher, A., Kojadinovic, I., Rohmer, T., Segers, J.: Detecting changes in cross-sectional dependence in multivariate time series. *J. Multivar. Anal.* **132**, 111–128 (2014)
- Cabrieto, J., Tuerlinckx, F., Kuppens, P., Hunyadi, B., Ceulemans, E.: Testing for the presence of correlation changes in a multivariate time series: a permutation based approach. *Sci. Rep.* **8**(1), 1–20 (2018)
- Carlstein, E.: Nonparametric change-point estimation. *Ann. Stat.* **16**, 188–197 (1988)
- Chan, H.P., Walther, G.: Detection with the scan and the average likelihood ratio. *Stat. Sin.* **23**, 409–428 (2013)
- Cho, H.: Change-point detection in panel data via double CUSUM statistic. *Electron. J. Stat.* **10**(2), 2000–2038 (2016)
- Cho, H., Fryzlewicz, P.: Multiple-change-point detection for high dimensional time series via sparsified binary segmentation. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 475–507 (2015)
- Cho, H., Kirch, C.: Data Segmentation Algorithms: Univariate Mean Change and Beyond. Preprint [arXiv:2012.12814](https://arxiv.org/abs/2012.12814) (2020)
- Enikeeva, F., Harchaoui, Z.: High-dimensional change-point detection under sparse alternatives. *Ann. Stat.* **47**(4), 2051–2079 (2019)
- Fearnhead, P., Rigaiil, G.: Relating and comparing methods for detecting changes in mean. *Stat.* **9**(1), e291 (2020)
- Fearnhead, P., Maidstone, R., Letchford, A.: Detecting changes in slope with an L_0 penalty. *J. Comput. Gr. Stat.* **28**, 265–275 (2019)
- Frick, K., Munk, A., Sieling, H.: Multiscale change point inference. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 495–580 (2014)
- Fryzlewicz, P.: Wild binary segmentation for multiple change-point detection. *Ann. Stat.* **42**, 2243–2281 (2014)
- Fryzlewicz, P.: Detecting possibly frequent change-points: Wild binary segmentation 2 and steepest-drop model selection. *J. Korean Stat. Soc.* **49**, 1027–1070 (2020)
- Garreau, D., Arlot, S.: Consistent change-point detection with kernels. *Electron. J. Stat.* **12**, 4440–4486 (2018)
- Groen, J.J., Kapetanios, G., Price, S.: Multivariate methods for monitoring structural change. *J. Appl. Econom.* **28**(2), 250–274 (2013)
- Hampel, F.R.: The influence curve and its role in robust estimation. *J. Am. Stat. Assoc.* **69**, 383–393 (1974)
- Hocking, T.D., Schleiermacher, G., Janoueix-Lerosey, I., Boeva, V., Cappo, J., Delattre, O., Bach, F., Vert, J.P.: Learning smoothing models of copy number profiles using breakpoint annotations. *BMC Bioinform* **14**(1), 1–15 (2013)
- Horváth, L., Hušková, M.: Change-point detection in panel data. *J. Time Ser. Anal.* **33**(4), 631–648 (2012)
- Hubert, L., Arabie, P.: Comparing partitions. *J. Classif.* **2**(1), 193–218 (1985)
- Inclán, C., Tiao, G.C.: Use of cumulative sums of squares for retrospective detection of changes of variance. *J. Am. Stat. Assoc.* **89**, 913–923 (1994)
- Jackson, B., Sargle, J.D., Barnes, D., Arabhi, S., Alt, A., Gioumousis, P., Gwin, E., Sangtrakulcharoen, P., Tan, L., Tsai, T.T.: An algorithm for optimal partitioning of data on an interval. *IEEE Signal Process. Lett.* **12**, 105–108 (2005)
- Jirak, M.: Uniform change point tests in high dimension. *Ann. Stat.* **43**(6), 2451–2483 (2015)
- Killick, R., Fearnhead, P., Eckley, I.A.: Optimal detection of change-points with a linear computational cost. *J. Am. Stat. Assoc.* **107**(500), 1590–1598 (2012)
- Kovács, S., Li, H., Bühlmann, P., Munk, A.: Seeded Binary Segmentation: A general methodology for fast and optimal change point detection. Preprint [arXiv:2002.06633](https://arxiv.org/abs/2002.06633) (2020)
- Lavielle, M., Teyssiere, G.: Detection of multiple change-points in multivariate time series. *Lith. Math. J.* **46**(3), 287–306 (2006)
- Lavielle, M., Teyssiere, G.: In: Long Memory in Economics, pp. 129–156. Springer (2007)
- Maeng, H., Fryzlewicz, P.: Detecting Linear Trend Changes and Point Anomalies in Data Sequences. Preprint [arXiv:1906.01939](https://arxiv.org/abs/1906.01939) (2019)
- Matteson, D.S., James, N.A.: A Nonparametric approach for multiple change point analysis of multivariate data. *J. Am. Stat. Assoc.* **109**, 334–345 (2014)
- Ombao, H., Von Sachs, R., Guo, W.: Slex analysis of multivariate nonstationary time series. *J. Am. Stat. Assoc.* **100**(470), 519–531 (2005)
- Picard, F., Lebarbier, E., Hoebeke, M., Rigaiil, G., Thiam, B., Robin, S.: Joint segmentation, calling, and normalization of multiple CGH profiles. *Biostatistics* **12**(3), 413–428 (2011)
- Raimondo, M.: Minimax estimation of sharp change points. *Ann. Stat.* **26**(4), 1379–1397 (1998)
- Rigaiil, G.: A pruned dynamic programming algorithm to recover the best segmentations with 1 to K_{max} change-points. *Journal de la Société Française de Statistique* **156**, 180–205 (2015)
- Schröder, A.L., Fryzlewicz, P.: Adaptive trend estimation in financial time series via multiscale change-point-induced basis recovery. *Stat. Interface* **6**(4), 449–461 (2013)
- Siris, V.A., Papagalou, F.: In: IEEE Global Telecommunications Conference, 2004. GLOBECOM'04., vol. 4, pp. 2050–2054. IEEE (2004)
- Tibshirani, R.J.: Adaptive piecewise polynomial estimation via trend filtering. *Ann. Stat.* **42**, 285–323 (2014)
- Truong, C., Oudre, L., Vayatis, N.: Selective review of offline change point detection methods. *Signal Process.* **167**, 107,299 (2020)
- Venkatraman, E.S.: Consistency results in multiple change-point problems. Doctoral Dissertation. Department of Statistics, Stanford University, Stanford (1992)
- Vert, J.P., Bleakley, K.: In: NIPS, pp. 2343–2351 (2010)
- Wang, T., Samworth, R.J.: High-dimensional changepoint estimation via sparse projection. *J. R. Stat. Soc. Ser. B* **80**, 57–83 (2018)
- Yu, Y.: A Review on Minimax Rates in Change Point Detection and Localisation. Preprint [arXiv:2011.01857](https://arxiv.org/abs/2011.01857) (2020)
- Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)* **68**(1), 49–67 (2006)
- Zou, C., Yin, G., Feng, L., Wang, Z.: Nonparametric maximum likelihood approach to multiple change-point problems. *Ann. Stat.* **42**, 970–1002 (2014)

Publisher’s Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.