Statistics and Computing (2026) 36:128
https://doi.org/10.1007/s11222-026-10881-1

ORIGINAL PAPER

Post-selection inference for quantifying uncertainty in changes in variance

Rachel Carrington[1] · **Paul Fearnhead**[2]

Received: 24 May 2024 / Accepted: 8 April 2026

Abstract
Quantifying uncertainty in detected changepoints is an important problem. However it is challenging as the naive approach would use the data twice, first to detect the changes, and then to test them. This will bias the test, and can lead to anti-conservative p-values. One approach to avoid this is to use ideas from post-selection inference, which conditions on the information in the data used to choose which changes to test. As a result this produces valid p-values; that is, p-values that have a uniform distribution if there is no change. Currently such methods have been developed for detecting changes in mean only. This paper presents two approaches for constructing post-selection p-values for detecting changes in variance. These vary depending on the method used to detect the changes, but are general in terms of being applicable for a range of change-detection methods and a range of hypotheses that we may wish to test.

Keywords Changepoint detection · Breakpoint · Binary segmentation · Post-selection p-value

1 Introduction

The problem of detecting changes in time series has received a great deal of attention in recent years, with applications in finance (Schröder and Fryzlewicz 2013), quality control (Amiri et al. 2015), climate modelling (Reeves et al. 2007; Shi et al. 2022), genome sequencing (Muggeo and Adelfio 2011; Caron et al. 2012), neuroscience (Anastasiou et al. 2022), and epidemic modelling (Jiang et al. 2023), amongst many others. Various methods have been developed for detecting different types of change, for example changes in mean, variance, and slope. These are often developed based on a test for a single change of a specific type. To detect multiple changes, this test can be used recursively within algorithms such as binary segmentation (Scott and Knott 1974) and its variants such as wild binary segmentation (Fryzlewicz 2014), seeded binary segmentation (Kovács et al. 2023), and narrowest over threshold (Baranowski et al. 2019). Alternatively one can write down a likelihood for the data as a function of the number and location of changes and aim to estimate the changes by minimising a penalised version of this likelihood. For many models this optimisation can be performed exactly, see Killick et al. (2012); Maidstone et al. (2017). See Fearnhead and Rigaill (2020) and Shi et al. (2022) for recent reviews of changepoint methods.

Recently, increasing attention has been given to the problem of estimating the uncertainty associated with detected changepoints (Frick et al. 2014; Li et al. 2016; Chen et al. 2023). This is challenging due to the interplay of different elements of uncertainty, in terms of the number and location of the changepoints. Bayesian approaches (Barry and Hartigan 1993; Fearnhead 2006; Nam et al. 2012; Cappello and Padilla 2025) give a natural way of quantifying this uncertainty, but require specifying prior information with the results sensitive to these choices. Alternatively one can use the properties of test statistics under the assumption of no changepoints. For example, Fryzlewicz (2024) and Fryzlewicz (2024) use a probabilistic bound on the maximum of the test statistic for the presence of a change over all possible intervals of data to find the narrowest set of intervals which each must contain at least one changepoint with a certain probability. Similar bounds are used within the MOSUM method (Meier et al.

✉ Rachel Carrington
rjc204@bath.ac.uk

Paul Fearnhead
p.fearnhead@lancaster.ac.uk

[1] Department of Mathematical Sciences, University of Bath, Bath, UK

[2] School of Mathematical Sciences, Lancaster University, Lancaster, UK

Published online: 29 April 2026

2021) to produce asymptotic p-values for detected changepoints (Eichinger and Kirch 2018).

In this paper we consider post-selection inference methods (Bayarri and Berger 1999; Kuchibhotla et al. 2022). The idea is to construct a p-value for each detected changepoint, with these p-values being valid in finite samples, albeit under strong distributional assumptions for the data. This is a challenging problem in general as, if we use the data to estimate changepoint locations, naively re-using the same data to estimate the uncertainty associated with each changepoint will lead to biased results. The idea of post-selection inference is, when testing for the presence of a changepoint at some detected location τ, we condition on the information in the data that led to detection at τ when calculating the p-value. To date, almost all work in this area has focused on the univariate change in mean case, e.g. (Hyun et al. 2021; Jewell et al. 2022; Duy and Takeuchi 2022; Carrington and Fearnhead 2025), or the piecewise constant polynomial model (Mehrizi and Chenouri 2021). As far as we know, there are no current approaches for quantifying the evidence for detected changes in variance. In this paper we fill this gap by extending the post-selection inference ideas to cover the change in variance setting.

The post-selection inference methods for the change in mean problem use the fact that the natural test statistic for a change at a given location will have a normal distribution under the null hypothesis. To account for the fact that we are using the same data to calculate the test statistic as to detect changes – and thus to decide which locations to test – we need to condition this distribution on the fact that the tested location is a detected changepoint. Depending on precisely what information is conditioned on, this leads to a conditional null distribution that is constrained to lie on an interval (Hyun et al. 2021) or a union of intervals (Jewell et al. 2022) of the real line. Importantly it is straightforward to calculate this interval or intervals for changepoint detection methods that use the popular CUSUM statistic within a binary segmentation algorithm or penalised likelihood approach. One of the advantages of this approach is that it is general: it applies to a wide range of methods for detecting changes, and also to most natural null hypotheses that one may wish to test.

In this paper, we consider the two most common ways of detecting a change in variance. The first uses the idea that a change in variance in some time series, X_t say, will correspond to a change in mean in X_t^2. Thus we can shift the data so it has mean zero, and apply a method for detecting a change in mean to the square of this shifted data. The second method is where we estimate changepoint locations by directly maximizing the likelihood of the piecewise constant variance model. In both cases we can use the same test statistic for testing whether a detected change is real. This test statistic is equivalent to the likelihood-ratio test statistic for a change in variance in Gaussian data, and is the proportion of the sum of squares of the data in the region prior to the putative change. If we had used separate information, or independent data, to choose which putative change location to test then the null distribution of this test statistic is a Beta distribution with known parameters. For our setting we need to calculate the null distribution conditional on detecting the change, which leads to constraining the Beta distribution to a union of intervals of [0, 1].

To calculate the post-selection p-values requires us to calculate these intervals. For the first approach to detecting changes, this can be done analytically using the same ideas as for the change in mean setting. For the second approach, this is no longer possible so we show how Monte Carlo can be used to approximate the p-values (see Saha et al. 2025, for similar use of Monte Carlo in post-selection inference).

The rest of the paper is organized as follows. In Section 2 we define the model and set up the test we will use to detect changes in variance. Section 3 then describes the methods we will use to calculate p-values. Section 4 details some extensions to different scenarios, such as different null hypotheses and changepoint detection methods. Section 5 presents results of our methods on simulated data, and in Section 6 we implement our method on a financial data set. Section 7 then concludes with some discussion of our results and suggestions for further research.

2 Testing for a change in variance

We consider the univariate Gaussian model with fixed mean and piecewise constant variance:

$$X_t = \mu + \sigma_t \epsilon_t, \quad t = 1, \ldots, T, \tag{1}$$

where μ is assumed known, $\epsilon_t \sim N(0, 1)$, and $\sigma_{t+1} = \sigma_t$ except at K changepoints $\{\tau_1, \ldots, \tau_K\}$. We will assume, without loss of generality, that $\mu = 0$ (if this is not the case, we can just replace X_t with $X_t - \mu$).

Given the model in (1), suppose that we have a data set $\boldsymbol{X} = (X_1, \ldots, X_T)$ and we have obtained a set of estimated changepoint locations $\mathcal{M}(\boldsymbol{X}) = \{\hat{\tau}_1, \ldots, \hat{\tau}_K\}$, such that $0 = \hat{\tau}_0 < \hat{\tau}_1 < \ldots < \hat{\tau}_K < \hat{\tau}_{K+1} = T$. For a particular changepoint $\hat{\tau}$, we want to test whether there is indeed a changepoint at, or close to, $\hat{\tau}$.

First, we need to define our null hypothesis and choose a test statistic. Since we are generally interested in whether there is a change close to $\hat{\tau}$, rather than exactly at $\hat{\tau}$, it makes sense to set the null hypothesis to be that there is no changepoint within a certain window of $\hat{\tau}$. There are various ways of choosing such a window, which will generally be informed by our particular application: for example, we can use a fixed window length, or allow the window to be determined by the locations of neighbouring changepoint estimates. For now

we will assume that we have some fixed h such that the window is $(\hat{\tau}-h, \hat{\tau}+h)$, but note that the methods we describe are easily extended to other choices of H_0; we will discuss this in more detail in Section 4.1.

Hence, we have

$$H_0 \; : \; \sigma^2_{\hat{\tau}-h+1} = \ldots = \sigma^2_{\hat{\tau}} = \sigma^2_{\hat{\tau}+1} = \ldots = \sigma^2_{\hat{\tau}+h}, \tag{2}$$

with H_1 being that there is at least one inequality.

Let $C^2_{t_1:t_2} = \sum_{t=t_1}^{t_2} X_t^2$. A natural choice of test statistic is to consider what proportion of the total sum of squares within the region $(\hat{\tau}-h, \hat{\tau}+h)$ is accounted for by the data before the changepoint, i.e.

$$\phi = \frac{C^2_{(\hat{\tau}-h+1):\hat{\tau}}}{C^2_{(\hat{\tau}-h+1):(\hat{\tau}+h)}} = \frac{C^2_{(\hat{\tau}-h+1):\hat{\tau}}}{C^2_{(\hat{\tau}-h+1):\hat{\tau}} + C^2_{(\hat{\tau}+1):(\hat{\tau}+h)}}.$$

The terms in the last expression, $C^2_{(\hat{\tau}-h+1):\hat{\tau}}$ and $C^2_{(\hat{\tau}+1):(\hat{\tau}+h)}$, are independent gamma random variables with shape parameter $h/2$ and scale parameter $2\sigma^2$ under H_0. Thus, under H_0, $\phi \sim \text{Beta}\left(\frac{h}{2}, \frac{h}{2}\right)$. Under H_1, we would expect ϕ to be closer to 0 if the variance increases after $\hat{\tau}$, and closer to 1 if $\hat{\tau}$ is followed by a decrease in variance.

Our test statistic ϕ is equivalent to the F statistic, a standard test statistic used to test for a change in variance, in the sense that each of ϕ and F can be derived from the other, e.g. $F = \phi/(1-\phi)$. Hence, using either will lead to the same result. The formulation we define above is more convenient, as it allows us to write X_t^2 as a linear function of ϕ (see below), which will simplify the calculations in the next section.

Let ϕ_{obs} denote the observed value of ϕ. If $\hat{\tau}$ corresponds to an increase in variance, the one-sided p-value is $p = \Pr_{H_0}(\phi \leq \phi_{obs})$; otherwise, $p = \Pr_{H_0}(\phi \geq \phi_{obs})$. To calculate the two-sided p-value, we first find ϕ_* such that

$$\Pr_{H_0}(\phi \leq \phi_*) = 1 - \Pr_{H_0}(\phi \leq \phi_{obs}).$$

Then, letting $\phi_{lower} = \min\{\phi_{obs}, \phi_*\}$ and $\phi_{upper} = \max\{\phi_{obs}, \phi_*\}$, the two-sided p-value is

$$p = \Pr_{H_0}\left(\phi \leq \phi_{lower} \text{ or } \phi \geq \phi_{upper}\right). \tag{3}$$

We will assume for the remainder of this paper that we are using the two-sided p-value, i.e., we are not concerned about the direction of the change. However, our methods apply equally to the one-sided p-value case.

2.1 Post-selection inference

The p-value in (3) is valid if we use separate information, or independent data, to calculate it from that which we use to choose the location to test. However, using the same data to do both introduces bias. In particular, as demonstrated by Figure 1, when simulating under H_0 we tend to obtain p-values that are much smaller than they should be, so the Type I error rate is too high. In order to get valid p-values in this case, we need to calculate the probability of (3) conditional on the information used to choose to test for a change at $\hat{\tau}$: here, this is the fact that $\hat{\tau}$ is in the set of estimated changepoints. Letting $\mathcal{M}(\boldsymbol{X})$ denote the set of estimated changepoints $\{\hat{\tau}_1, \ldots, \hat{\tau}_K\}$ which we obtain when we apply a particular changepoint algorithm to $\boldsymbol{X}$, we therefore want to calculate the p-value in (3) conditional on the fact that $\hat{\tau} \in \mathcal{M}(\boldsymbol{X})$:

$$\Pr_{H_0}\left(\phi \leq \phi_{lower} \text{ or } \phi \geq \phi_{upper} \mid \hat{\tau} \in \mathcal{M}(\boldsymbol{X})\right).$$

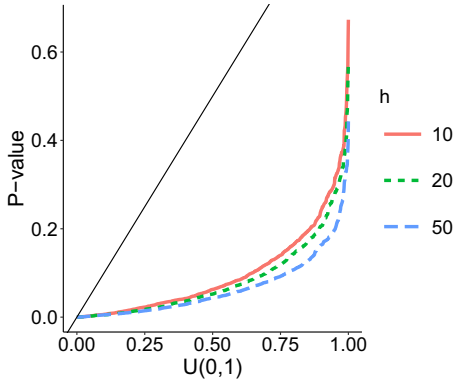

Fig. 1 QQ plot of p-values, calculated using (3), for data simulated under the null hypothesis of no change, where we selected change locations to test using binary segmentation. Different coloured lines correspond to different choices of h in the null hypothesis. Re-using the data results in p-values that are much smaller than they should be.

Our null hypothesis (2) is that the variance is constant inside the region $\{\hat{\tau}-h+1, \ldots, \hat{\tau}+h\}$. This does not specify the value that σ_t^2 takes for any t outside this window, nor the value of the fixed variance within the window. Therefore, in order for the test to be well-defined, i.e. have the same distribution for all data generating processes consistent with the null hypothesis and the model, we must condition on sufficient statistics for these parameters: the values of X_t outside of $\{\hat{\tau}-h+1, \ldots, \hat{\tau}+h\}$, and the sum of squares within the window: $C^2_{(\hat{\tau}-h+1):(\hat{\tau}+h)} = \sum_{t=\hat{\tau}-h+1}^{\hat{\tau}+h} X_t^2$.

The validity of post-selection p-values, in that they are uniformly distributed under the null conditional on the selection event, still holds if we condition on additional information (though this may lead to a reduction in power; Fithian et al. 2014; Jewell et al. 2022). In order to simplify the calculation of the p-values it is common to condition on data that is orthogonal to the test-statistic, ϕ, as we now describe.

To make explicit the dependence of $\boldsymbol{X}$ on ϕ, consider a re-parameterization of $\{X_{\hat{\tau}-h+1}, \ldots, X_{\hat{\tau}+h}\}$. Let $C_0^2 = C^2_{(\hat{\tau}-h+1):(\hat{\tau}+h)}$, the sum of the square of the data in the region being tested, and for $i = 1, \ldots, h-1$ let

$$W_i^l = \frac{C^2_{(\hat{\tau}-h+1):(\hat{\tau}-h+i)}}{C^2_{(\hat{\tau}-h+1):(\hat{\tau}-h+i+1)}}$$
$$W_i^r = \frac{C^2_{(\hat{\tau}+1):(\hat{\tau}+i)}}{C^2_{(\hat{\tau}+1):(\hat{\tau}+i+1)}} \quad (4)$$

The following results show how we can write $\{X^2_{\hat{\tau}-h+1}, \ldots, X^2_{\hat{\tau}+h}\}$ in terms of ϕ, C_0^2, and W_i^r and W_i^l for $i = 1, \ldots, h-1$, and that the latter set of variables are independent under the null hypothesis.

Lemma 1 *For $i = 1, \ldots, h$, we can write*

$$X^2_{\hat{\tau}-h+i} = \begin{cases} (1 - W^l_{i-1}) \left(\prod_{k=i}^{h-1} W^l_k\right) C_0^2 \phi & i = 1, \ldots, h-1 \\ (1 - W^l_{i-1}) C_0^2 \phi & i = h \end{cases}$$
$$X^2_{\hat{\tau}+i} = \begin{cases} (1 - W^r_{i-1}) \left(\prod_{k=i}^{h-1} W^r_k\right) C_0^2 (1-\phi) & i = 1, \ldots, h-1 \\ (1 - W^r_{i-1}) C_0^2 (1-\phi) & i = h, \end{cases} \quad (5)$$

where for convenience we set $W_0^l = W_0^r = 0$.

Lemma 2 *Under H_0, $W_1^l, \ldots, W_{h-1}^l, W_1^r, \ldots, W_{h-1}^r, \phi, C_0^2$ are all independent.*

Lemma 1 is easily verified. The proof of Lemma 2 is given in the supplementary material.

Therefore, conditioning on the dimensions of the data which are orthogonal to ϕ is equivalent to conditioning on the values of $W_1^l, \ldots, W_{h-1}^l$, $W_1^r, \ldots, W_{h-1}^r$, and C_0^2. We can see from (5) above that doing this while allowing only ϕ to change has the effect of multiplying $X_{\hat{\tau}-h+1,\ldots,X_{\hat{\tau}}}$ by $\sqrt{\frac{\phi}{\phi_{obs}}}$ and multiplying $X_{\hat{\tau}+1}, \ldots, X_{\hat{\tau}+h}$ by $\sqrt{\frac{1-\phi}{1-\phi_{obs}}}$. As we change ϕ we are re-scaling the data before and after $\hat{\tau}$ so that the proportion of the sum of squares of the data before and after $\hat{\tau}$ changes whilst the overall sum of squares of the data remains unchanged.

Hence, we can define the set of data that we would obtain as we vary ϕ but keep fixed all information orthogonal to ϕ:

$$X'_t(\phi) = \begin{cases} X_t & t \le \hat{\tau} - h \text{ or } t \ge \hat{\tau} + h + 1 \\ \sqrt{\frac{\phi}{\phi_{obs}}} X_t & \hat{\tau} - h + 1 \le t \le \hat{\tau} \\ \sqrt{\frac{1-\phi}{1-\phi_{obs}}} X_t & \hat{\tau} + 1 \le t \le \hat{\tau} + h. \end{cases} \quad (6)$$

Figure 2 illustrates this transformation.

Letting $\mathcal{S}$ denote the set $\mathcal{S} = \{\phi : \hat{\tau} \in \mathcal{M}(\boldsymbol{X}'(\phi))\}$ – the set of ϕ values such that the changepoint of interest, $\hat{\tau}$, is included as a change in the output of the changepoint algorithm – the post-selection p-value we want to calculate is

$$p = \Pr\left(\phi \le \phi_{lower} \text{ or } \phi \ge \phi_{upper} \mid \phi \in \mathcal{S}\right). \quad (7)$$

Since, under H_0, $\phi \sim \text{Beta}\left(\frac{h}{2}, \frac{h}{2}\right)$, the conditional distribution of $\phi \mid \phi \in \mathcal{S}$ is Beta $\left(\frac{h}{2}, \frac{h}{2}\right)$ truncated to $\mathcal{S}$, and calculating p depends on calculating $\mathcal{S}$. This is what we will consider in the next section.

3 Methods

3.1 CUSUM for change in variance

As a reminder, our model is

$$X_t = \sigma_t \epsilon_t, \quad t = 1, \ldots, T,$$

where $\epsilon_t \sim N(0, 1)$. A change in variance in $\boldsymbol{X}$ will correspond to a change in mean in $\boldsymbol{X}^2$. Therefore, one approach to detecting a change in variance in $\boldsymbol{X}$ is to look for a change in mean in $\boldsymbol{Y} = \boldsymbol{X}^2$, for example by maximizing the cumulative sum (CUSUM) statistic, which is defined on a interval $(s, e) = \{t : s \le t \le e\}$ as

$$G_{s,e}(t) = \left(\frac{(t-s+1)(e-t)}{e-s+1}\right)^{1/2} \left(\frac{1}{t-s+1}\sum_{j=s}^{t} Y_j - \frac{1}{e-t}\sum_{j=t+1}^{e} Y_j\right), \quad t = s, \ldots, e-1.$$

Figure 3 illustrates this process for a single changepoint. We will focus on using the CUSUM statistic with binary segmentation (Scott and Knott 1974); our approach to calculating p-values will also extend to variants such as wild binary segmentation, as well as the iterated cumulative sums of squares method of (Inclan and Tiao 1994), which uses a different statistic based on the cumulative sums of squares.

To detect whether there is a change in mean in an interval (s, e), where $1 \le s < e \le T$, we take

$$\hat{\tau} = \arg\max_t |G_{s,e}(t)|,$$

and determine that there is a changepoint at $\hat{\tau}$ if $|G_{s,e}(\hat{\tau})|$ is greater than some threshold. To detect multiple changepoints using binary segmentation, we start by maximizing $|G_{1,T}(t)|$ across the whole data set, then split the data at $\hat{\tau}$ and search for changepoints using $G_{s,e}(t)$ calculated on each segment. This process is repeated until no more changepoints are detected.

3.1.1 Calculating the post-selection *p*-value

To calculate the p-value in (7), we must compute $\mathcal{S} = \{\phi : \hat{\tau} \in \mathcal{M}(\boldsymbol{X}'(\phi))\}$ – in other words, we need to find the set of ϕ values such that our changepoint algorithm detects a changepoint at $\hat{\tau}$. Thus $\mathcal{S}$ is dependent on our choice of changepoint algorithm, in our case binary segmentation.

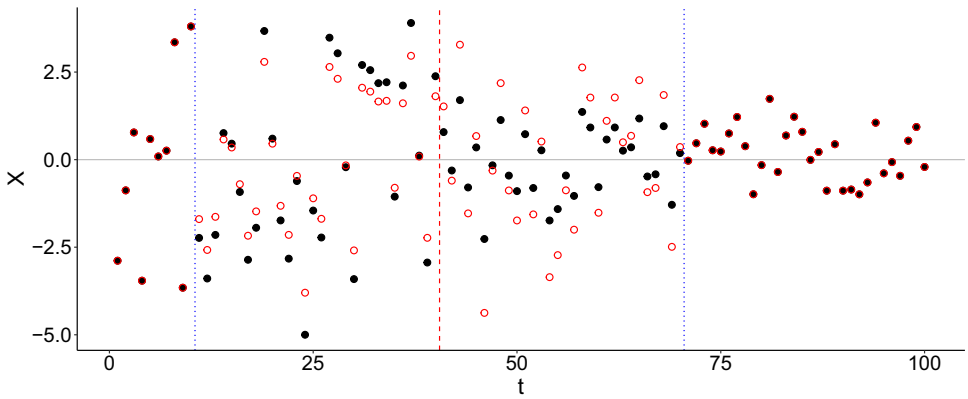

Fig. 2 An example of how the data is transformed when we apply the transformation $\boldsymbol{X}'(\phi)$. Here, the black points (filled circles) correspond to $\boldsymbol{X}$ and the red points (empty circles) to $\boldsymbol{X}'(0.5)$. Since $\phi_{obs} > 0.5$ in this case, the points before $\hat{\tau}$ (shown by the dashed red line) are pulled in towards 0, and the points after the changepoint move away from 0. The overall sum of squares in the region $\{\hat{\tau} - h + 1, \ldots, \hat{\tau} + h\}$ remains constant.

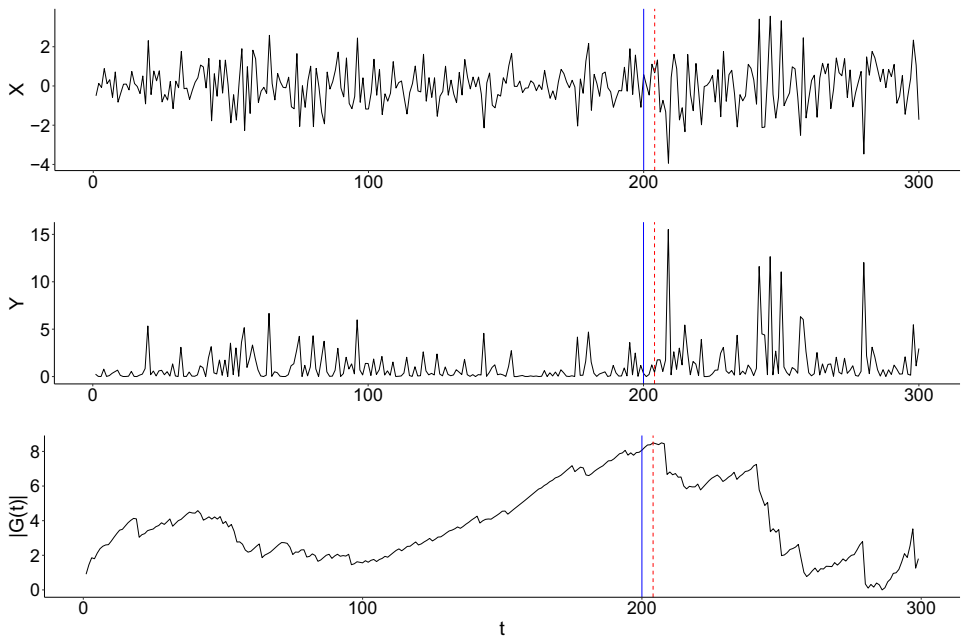

Fig. 3 A simulated data set with one change in variance. The data are simulated from $N(0, 1)$ for $t = 1, \ldots, 200$ and from $N(0, 1.7)$ for $t = 201, \ldots, 300$. The blue line shows the true changepoint location and the dashed red line the changepoint location estimated by maximizing $|G_{1,T}(t)|$.

To calculate $\mathcal{S}$ we can borrow ideas from methods used for the change in mean model. Hyun et al. (2021) showed that by conditioning on the set of changes $\mathcal{M}(\boldsymbol{X})$, the order in which the changes are found, and the direction of each change (positive or negative), we can write down a set of inequalities which must be satisfied by the CUSUM statistic. These are derived by considering each iteration of binary segmentation in turn, which is possible as conditioning on the order in which changes are found means we know the segments being tested at each iteration. For each iteration where we detect a change, the CUSUM statistics must be largest at the detected value and be above the threshold, while if no change is detected all the CUSUM statistics are below the threshold. We can write out inequalities for each of these possibilities.

For a given segment (s, e), in order to detect a change point at $\hat{\tau}$ we must have $|G_{s,e}(\hat{\tau})| \geq \lambda$, where λ is the changepoint detection threshold. This can be written as the linear equations, for $t = s, \ldots, e-1$ with $t \neq \hat{\tau}$:

$$-\hat{d} \cdot G_{s,e}(\hat{\tau}) > G_{s,e}(t), \quad \text{and} \quad -\hat{d} \cdot G_{s,e}(\hat{\tau}) > -G_{s,e}(t), \tag{8}$$

where $\hat{d} \in \{+1, -1\}$ is the direction of the change. For no changepoint to be detected in the interval (s, e), we must have, for $t = s, \ldots, e-1$,

$$G_{s,e}(t) < \lambda, \quad \text{and} \quad G_{s,e}(t) > -\lambda. \tag{9}$$

Hence, we can write down a set of linear inequalities in the data for each iteration of binary segmentation, which must all be satisfied for the particular set of changepoints, with the order and directions of change, to be found.

As described above, we can define perturbations of the data, $\mathbf{X}'(\phi)$, that correspond to changes in our test-statistic ϕ whilst fixing the information orthogonal to ϕ, see (6).

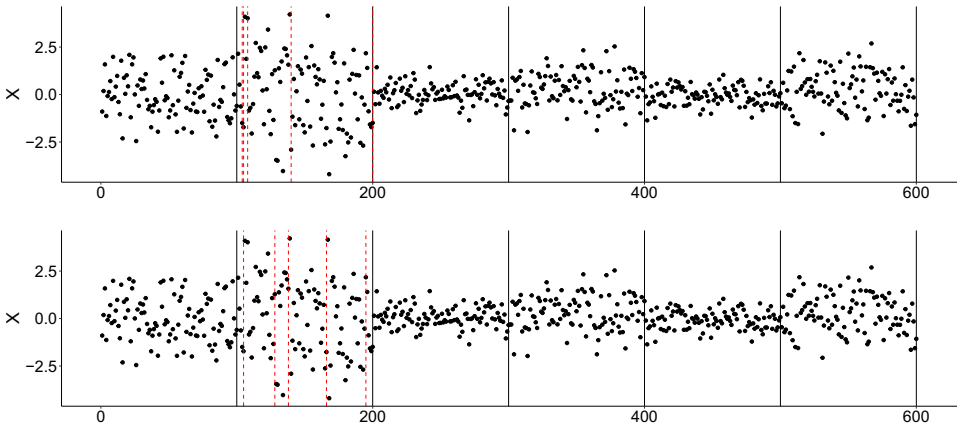

Fig. 4 Estimating changepoints using binary segmentation and wild binary segmentation with the CUSUM statistic for a simulated data set with multiple changes. The solid black lines correspond to true changes, the dashed red lines to estimated changepoints. The method over-estimates the number of changepoints within the region of highest variance, whilst missing the other changepoints

The CUSUM statistic is

$$G_{s,e}(t) = \left(\frac{(t-s+1)(e-t)}{e-s+1}\right)^{1/2} \cdot \left(\frac{1}{t-s+1}\sum_{j=s}^{t} Y_j - \frac{1}{e-t}\sum_{j=t+1}^{e} Y_j\right) = \left(\frac{(t-s+1)(e-t)}{e-s+1}\right)^{1/2} \cdot \left(\frac{1}{t-s+1}\sum_{j=s}^{t} X_j^2 - \frac{1}{e-t}\sum_{j=t+1}^{e} X_j^2\right).$$

If we replace the data $\mathbf{X}$ by $\mathbf{X}'(\phi)$, we get a set of CUSUM statistics that are functions of ϕ:

$$G_{s,e}(\phi, t) = \left(\frac{(t-s+1)(e-t)}{e-s+1}\right)^{1/2} \cdot \left(\frac{1}{t-s+1}\sum_{j=s}^{t} X'_j(\phi)^2 - \frac{1}{e-t}\sum_{j=t+1}^{e} X'_j(\phi)^2\right).$$

As $\mathbf{X}'(\phi)^2$ is linear in ϕ — see (6) — so is $G_{s,e}(\phi, t)$, and thus the inequalities (8) and (9) that, conditional on the order in which changes are found and the direction of each change, define the set $\mathcal{S}$ reduce to a set of linear inequalities which are simple to solve. Furthermore, Jewell et al. (2022) show how to extend this idea so we can calculate $\mathcal{S}$ without requiring us to condition on the order and direction of the changes. Hyun et al. (2018) and Jewell et al. (2022) also show how this idea can be extended to other changepoint detection algorithms.

3.2 Likelihood-ratio based detection of changes

Whilst post-selection inference is relatively straightforward when estimating changepoint locations using the CUSUM statistic, this can lead to poor detection of changes, as it tends to overestimate the number of changes in regions of higher variance, whilst failing to detect changes in lower-variance regions (see Figure 4).

We can achieve more accurate changepoint detection by maximizing the likelihood ratio statistic, which on an interval (s, e) is defined as

$$\Lambda_{s,e}(\tau) = (s-e+1)\log\sum_{t=s}^{e} X_t^2 - (\tau-s+1)\log\sum_{t=s}^{\tau} X_t^2 - (e-\tau)\log\sum_{t=\tau+1}^{e} X_t^2$$

To estimate a single changepoint, we calculate $\Lambda_{1,T}(\tau)$ for $\tau = 1, \ldots, T-1$, and take $\hat{\tau} = \arg\max_\tau \Lambda_{1,T}(\tau)$. For multiple changepoints, we can either use an iterative method such as binary segmentation, or a penalized likelihood method such as PELT (Killick et al. 2012). The approach we develop below will apply in both cases.

3.2.1 Post-selection inference

As a reminder, the post-selection p-value we want to calculate is

$$p = \Pr\left(\phi \le \phi_{lower} \text{ or } \phi \ge \phi_{upper} \mid \phi \in \mathcal{S}\right),$$

where $\mathcal{S} = \{\phi : \hat{\tau} \in \mathcal{M}(\boldsymbol{X}'(\phi))\}$. Previously, when detecting changes based on the CUSUM statistic, we were able to write down a set of linear inequalities in ϕ which defined the set $\mathcal{S}$. However, while we can write down inequalities in $\Lambda_{s,e}(\phi, \tau)$, $\Lambda_{s,e}(\phi, \tau)$ is no longer a linear function of ϕ (see the Supplementary Material for more details), and the inequalities cannot be resolved analytically. $\mathcal{S}$ will still consist of a union of intervals, however, as the inequalities are smooth functions of ϕ: Figure 5 illustrates this by showing a plot of $\Lambda(\tau, \phi)$ against ϕ for different values of τ. The region for which any particular τ value maximizes $\Lambda(\tau, \phi)$ consists of one or more intervals.

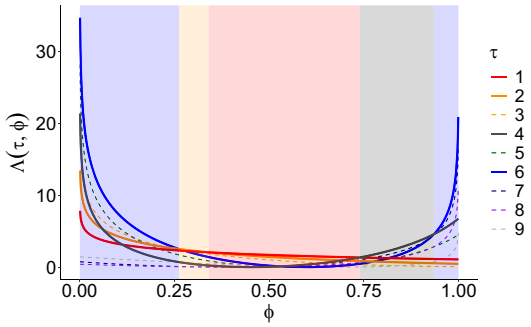

Fig. 5 Plot showing $\Lambda(\tau, \phi)$ against ϕ, for an example with $T = 10$. Each line on the plot corresponds to $\Lambda(\tau, \phi)$ for a different value of $\tau \in 1, \ldots, T-1$. Lines corresponding to values of τ that maximize $\Lambda(\tau, \phi)$ for some values of ϕ (here $\tau = 1, 2, 4, 6$) are shown in bold, with corresponding shaded regions showing for which values of ϕ each τ gives the maximum. We can hence see that for each $\tau \in \{1, 2, 4, 6\}$, the values of ϕ for which this τ maximizes $\Lambda(\tau, \phi)$ fall in an interval or union of intervals.

To consider how we can estimate p, let $q(\phi)$ denote the conditional distribution of $\phi \mid \phi \in S$:

$$q(\phi) = \frac{\pi(\phi)\, \mathbb{I}_{\phi \in \mathcal{S}}}{\int_0^1 \pi(\phi')\, \mathbb{I}_{\phi' \in \mathcal{S}}\, d\phi'}.$$

where $\pi(\phi)$ denotes the distribution function of the unconditional distribution of ϕ, i.e. Beta $\left(\frac{h}{2}, \frac{h}{2}\right)$, and $\mathbb{I}$ denotes the indicator function, such that $\mathbb{I}_{\phi \in \mathcal{S}} = 1$ if $\phi \in \mathcal{S}$ and 0 otherwise. Define $\mathcal{C} = \{\phi : \phi \leq \phi_{lower} \text{ or } \phi \geq \phi_{upper}\}$ to be the critical region of the test. Then the p-value can then be written as

$$p = \int_{\mathcal{C}} q(\phi) \mathrm{d}\phi. \tag{10}$$

Whilst $\mathcal{S}$, and hence $q(\phi)$, is unknown, we can evaluate $\mathbb{I}_{\phi \in \mathcal{S}}$ for any ϕ by re-running the changepoint algorithm on $\mathbf{X}'(\phi)$ and checking if $\hat{\tau} \in \mathcal{M}(\mathbf{X}'(\phi))$. This gives a number of different ways of approximating (10).

One approach involves using Monte Carlo to estimate the integral in (10). We can do this by sampling from q using rejection sampling. That is we sample $\phi_1, \ldots, \phi_N$ from $\pi(\phi)$, and accept ϕ_i as a draw from $q(\phi)$ if and only if $\mathbb{I}_{\phi_i \in \mathcal{S}} = 1$. This gives the following estimate

$$\hat{p} = \frac{\sum_{i=1}^N \mathbb{I}_{\phi_i \in \mathcal{S}} \cdot \mathbb{I}_{\phi_i \in \mathcal{C}}}{\sum_{i=1}^N \mathbb{I}_{\phi_i \in \mathcal{S}}}.$$

However, this estimator can require a large value of N as typically only a small proportion of samples will have $\mathbb{I}_{\phi \in \mathcal{S}} = 1$.

One way to reduce the variance of $\hat{p}$, without increasing the computation time, is to use stratified sampling. For $i = 1, \ldots, N$, we sample $z_i \sim U(\frac{i-1}{N}, \frac{i}{N})$. We then calculate ϕ_i to satisfy $\Pr(\phi < \phi_i) = z_i$ (see Section D.2 in the Supplementary Material). However, it is often the case that $\Pr(\phi \in \mathcal{S})$ is of the order of $\frac{1}{N}$ or smaller, in which case using stratified sampling is unlikely to improve the estimate $\hat{p}$.

Instead we consider using the information from $\mathbb{I}_{\phi^{(i)} \in \mathcal{S}}$, for $i = 1, \ldots, N$ to estimate $q(\phi)$ directly, and then use this estimate to approximate (10). Define $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_N)^\top$ and $\mathbf{I} = (\mathbb{I}_{\phi_1 \in \mathcal{S}}, \ldots, \mathbb{I}_{\phi_N \in \mathcal{S}})^\top$, to be, respectively, the vector of ϕ values and their associated indicator functions. A Bayesian approach would be to introduce a prior on the indicator function $\mathbb{I}_{\phi \in \mathcal{S}}$, as a function of ϕ, update these beliefs given $\boldsymbol{\phi}$ and $\mathbf{I}$. An estimate of $q(\phi)$ would then be

$$\hat{q}(\phi) \propto \mathrm{E}\left(\pi(\phi)\mathbb{I}_{\phi \in \mathcal{S}} | \boldsymbol{\phi}, \mathbf{I}\right) = \frac{\pi(\phi)\, \mathrm{E}\left(\mathbb{I}_{\phi \in \mathcal{S}} | \boldsymbol{\phi}, \mathbf{I}\right)}{\int_0^1 \pi(\phi')\, \mathrm{E}\left(\mathbb{I}_{\phi' \in \mathcal{S}} | \boldsymbol{\phi}, \mathbf{I}\right) \mathrm{d}\, \phi'}. \tag{11}$$

To do this, we model $\mathrm{E}(\mathbb{I}_{\phi \in \mathcal{S}})$ as the mean of a Gaussian process (GP) with some prior mean $\mu(\phi)$ and covariance kernel $K(\phi, \phi')$. Then, for any ϕ' we have

$$\mathrm{E}(\mathbb{I}_{\phi' \in \mathcal{S}} | \boldsymbol{\phi}, \mathbf{I}) = \mu(\phi') + \boldsymbol{K}(\boldsymbol{\phi}, \phi')^\top \boldsymbol{K}(\boldsymbol{\phi}, \boldsymbol{\phi})^{-1}(\mathbf{I} - \mu(\boldsymbol{\phi})), \tag{12}$$

where we use $\boldsymbol{K}(\boldsymbol{\phi})$ to be the vector whose jth entry is $K(\phi_j, \phi')$, $\boldsymbol{K}(\boldsymbol{\phi}, \boldsymbol{\phi})$ to denote the matrix whose (i, j)th entry is $K(\phi_i, \phi_j')$ and $\mu(\boldsymbol{\phi})$ is the vector whose jth entry is $\mu(\phi_j)$.

To implement this, we need to choose the mean and kernel functions. We choose an exponential kernel $\boldsymbol{K}(\phi_1, \phi_2) = e^{-\frac{1}{2l^2}|\phi_1 - \phi_2|}$, which is Markov. (This requires us to choose the value of l: we investigate this in Section D.3 of the Supplementary Material, where we find that at long as l is at least 1, the value we use does not have a noticeable effect on the results.) If we have evaluated $\mathbb{I}_{\phi \in \mathcal{S}}$ at $\phi_1 < \cdots < \phi_N$, and $\phi_i < \phi' < \phi_{i+1}$ the Markov property means that our Gaussian approximation to $p(\phi')$ will only depend on the values of $\mathbb{I}_{\phi_i \in \mathcal{S}}$ and $\mathbb{I}_{\phi_{i+1} \in \mathcal{S}}$. Furthermore, for such a kernel the mean of the Gaussian process will be an average of these values and the prior mean – and thus will always lie within [0, 1]. As we expect $\phi \in \mathcal{S}$ to be rare, we use the limit where $\mu(\phi) = 0$.

To implement the method we obtain $\phi_1, \ldots, \phi_N$ as a stratified sample from a uniform distribution on [0, 1]. We then estimate (10) using quadrature, where we replace $q(\phi)$ with $\hat{q}(\phi)$ as defined in (11) with the posterior expectation calculated using (12). An alternative to using quadrature to approximate (11) would be to use importance sampling with samples generated from (11).

Figure 6 shows an outline of this process. First, we sample $\phi_1, \ldots, \phi_N$. For each ϕ_i, we calculate $\boldsymbol{X}'(\phi_i)$ and apply the

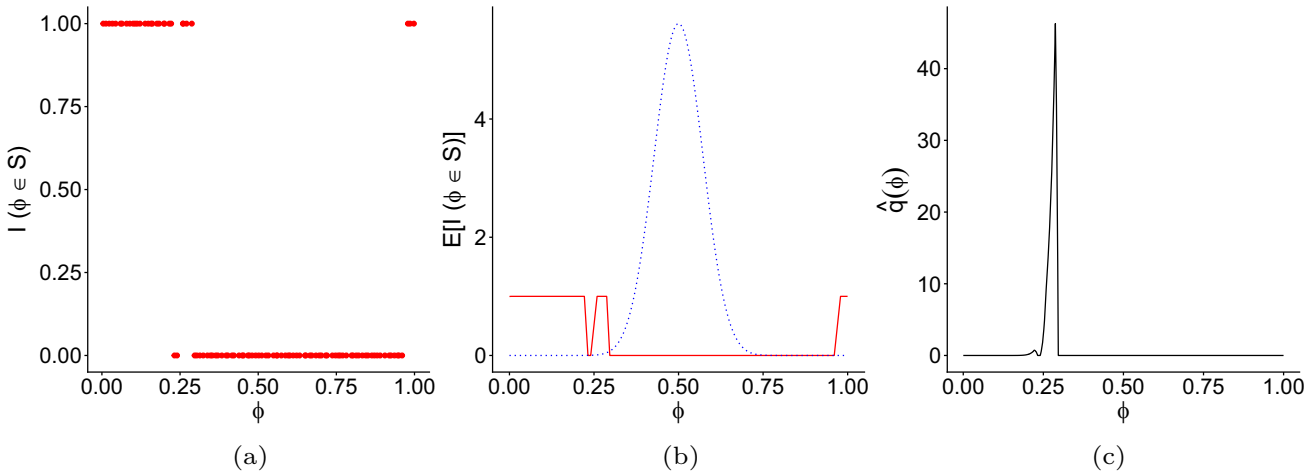

Fig. 6 Estimating $q(\phi)$ using a Gaussian process. Left: $\mathbb{I}_{\phi\in\mathcal{S}}$ plotted against ϕ for 100 samples. Middle: The blue line shows the posterior mean of the Gaussian process, which is an estimate of $I_{\phi\in S}$. The dotted red line shows the unconditional distribution $\pi(\phi)$. Right: Estimate of $\hat{q}(\phi)$, which is the normalized product of the GP mean and $\pi(\phi)$. The distribution is concentrated in the narrow region where both are substantially different from 0.

changepoint algorithm to evaluate $p(\phi_i) = \mathbb{I}_{\phi_i\in\mathcal{S}}$. Panel (a) shows a plot of $p(\phi)$ for our samples. Then we fit a Gaussian process: panel (b) shows the posterior mean as a solid blue line, with the prior $\pi(\phi)$ as a dashed red line. We then estimate $\hat{q}(\phi)$ as described above. This is shown in panel (c). In this case $\hat{q}(\phi)$ is very concentrated in a small region, as there is only a narrow range of ϕ values for which both $\pi(\phi)$ and the GP mean are different from 0. (Although $\pi(\phi)$ is not exactly 0 for $\phi \in (0, 1)$, for some subregions it is small enough as to be practically regarded as 0, and does not contribute to $\hat{q}(\phi)$.)

4 Extensions

4.1 Different null hypotheses

There are alternative choices for the null hypothesis we may want to test, in particular to deal with the possibility of detected changepoints being close together. For example, we could decide to truncate the region of interest if we find another changepoint within h of the changepoint being tested. Alternatively, if we denote our changepoint of interest as $\hat{\tau}_j$, we can take the null hypothesis that there is no changepoint within $\{\hat{\tau}_{j-1}+1, \ldots, \hat{\tau}_{j+1}-1\}$ (by convention setting $\hat{\tau}_0 = 0$ and $\hat{\tau}_{K+1} = T$, where K is the number of changepoints), or define the region we test to that closer to $\hat{\tau}_j$ than any other detected changepoint (see Jewell et al. 2022; Carrington and Fearnhead 2025).

In general, the methods we have described above will work in the same way for these different alternatives. However, in the case where the region of interest is determined by the locations of other estimated changepoints, it is necessary to redefine the conditioning set to take account of this. For example, if the window is determined by the locations of neighbouring changepoints, then we also need to condition on the fact that these changepoint locations are included in the set of changepoints $\mathcal{M}(\boldsymbol{X})$. This can be done, for example, by conditioning on the set of all estimated changepoints — i.e. $\mathcal{M}(\boldsymbol{X}'(\phi)) = \mathcal{M}(\boldsymbol{X}_{obs})$ — rather than just on the fact that $\hat{\tau} \in \mathcal{M}(\boldsymbol{X}'(\phi))$. To do this, for the CUSUM method we construct the set of intervals in the same way, but only choose the ones for which this new condition holds to be in $\mathcal{S}$. For the more general case in Section 3.2, it is straightforward to implement this.

4.2 Different changepoint detection methods

The method for post-selection inference for likelihood-ratio based detection of changes, described in Section 3.2, can be easily extended to the use of other binary segmentation methods (Fryzlewicz 2014, 2024; Kovács et al. 2023) or to penalised likelihood approaches (Killick et al. 2012; Fearnhead and Rigaill 2020). In each case we just need to use the same method for detecting changes to analyse the perturbed data sets as to evaluate the values of ϕ that are in the set $\mathcal{S}$.

The main computational cost is in re-running the changepoint detection algorithm on the perturbed data sets. Often one can reduce the computational cost through early stopping. If we are conditioning on $\hat{\tau} \in \mathcal{M}(\boldsymbol{X})$ then we can stop once this condition holds, whereas if we are conditioning on $\mathcal{M}(\boldsymbol{X})$ we can stop if ever we detect a change not in $\mathcal{M}(\boldsymbol{X})$ for our new data. Also we can often re-use calculations. For example, to implement methods such as wild binary segmen-

tation and narrowest over threshold we have to calculate the likelihood ratio statistic on a set of intervals $[s, e] \subseteq [1, T]$. The set of intervals is the same each time we implement the algorithm on $\boldsymbol{X}'(\phi)$. Since $\Lambda_{s,e}(t, \phi)$ is constant on intervals which do not overlap with the region of interest (i.e., $e \leq \hat{\tau} - h$ or $s \geq \hat{\tau} + h + 1$), for these intervals we only have to calculate $\Lambda_{s,e}(t)$ once. Therefore, at each sampling iteration we only have to re-calculate $\Lambda_{s,e}(t, \phi)$ on intervals which overlap with the region of interest.

4.3 Increasing power by conditioning on less information

So far, we have calculated p-values by conditioning on all aspects of the data except for the test statistic ϕ. However, as shown in (Carrington and Fearnhead 2025) for the change in mean case, it is possible to reduce this to conditioning on only the sufficient statistics for the parameters not specified under H_0, leaving $\boldsymbol{W} = \left(W_1^l, \ldots, W_{h-1}^l, W_1^r, \ldots, W_{h-1}^r\right)$ – as defined in (4) – as free parameters.

$$\Pr\left(\phi \leq \phi_{lower} \text{ or } \phi \geq \phi_{upper} \mid \hat{\tau} \in \mathcal{M}(\boldsymbol{X}'(\phi, \boldsymbol{W}))\right).$$

We can estimate this by Monte Carlo methods based on sampling $N_W - 1$ values of $\boldsymbol{W}$. More precisely, we set $\boldsymbol{W}^{(1)}$ to be the observed value of $\boldsymbol{W}$, and for $j = 2, \ldots, N_W$ sample $\boldsymbol{W}^{(j)}$ using the property that for $i = 1, \ldots, h-1$, W_i^l and W_i^r are independent Beta $\left(\frac{i}{2}, \frac{1}{2}\right)$ random variables under H_0. We then calculate

$$\hat{p} = \frac{\sum_{j=1}^{N_W} \Pr\left(\phi \geq \phi_{upper} \text{ or } \phi \leq \phi_{lower}, \hat{\tau} \in \mathcal{M}(\boldsymbol{X}), \mid \boldsymbol{W} = \boldsymbol{W}^{(j)}\right)}{\sum_{j=1}^{N_W} \Pr\left(\hat{\tau} \in \mathcal{M}(\boldsymbol{X}) \mid \boldsymbol{W} = \boldsymbol{W}^{(j)}\right)}.$$

As shown by (Carrington and Fearnhead 2025), using this method we obtain valid p-value estimates (i.e. the p-values follow a uniform distribution under H_0). If we have used a likelihood-ratio based detection method, we have to use estimates of $\Pr\left(\phi \geq \phi_{upper} \text{ or } \phi \leq \phi_{lower}, \hat{\tau} \in \mathcal{M}(\boldsymbol{X}), \mid \boldsymbol{W} = \boldsymbol{W}^{(j)}\right)$ and $\Pr\left(\hat{\tau} \in \mathcal{M}(\boldsymbol{X}) \mid \boldsymbol{W} = \boldsymbol{W}^{(j)}\right)$ using the methods in Section 3.2.1 Increasing N_W results in a more powerful test, albeit at an increased computational cost. We investigate applying this method on simulated data in Section 5.3.

5 Simulations

In this section we implement our methods on simulated data. All simulation results are based on 1000 replicates.

5.1 Detecting changes using CUSUM

Figure 7 shows QQ plots of p-values obtained by sampling from a model with $T = 200$. In (a) we sample from H_0; in (b)–(d) we sample from a model with a single change at $\tau = 100$, where in each case the variance before the changepoint is 1, and the variance after the changepoint is 0.25, 2, 4. We see that under H_0, $p \sim U(0, 1)$, and under H_1 the test has power to detect changes, which increases with the window size h and the size of the change.

To see what happens when we have multiple changes, Figure 8 shows similar plots where we simulate from a model with 4 changes, with $T = 200$ and $h = 10, 20$. In panels (a) and (b) we simulated from a model with 4 equally spaced changes, $\tau = 40, 80, 120, 160$. In (c) and (d), changepoints were chosen uniformly from $(h, T - h)$, with the constraint that the minimum distance between changepoints was at least h.

5.2 Detecting changes using the likelihood-ratio test

In Section 3.2.1, we discussed how we can estimate the p-value using a Gaussian process. To implement this we must choose the number of random samples N, and decide whether we will estimate the p-value directly from the posterior mean of the Gaussian process, or use importance sampling from the posterior distribution. We here investigate difference options for different simulated scenarios. In particular, we are interested in knowing what happens when the number of samples N is relatively small.

Figure 9 shows QQ plots of estimated p-values we obtain on simulated data for three different scenarios: (1) no change; (2) three changes at fixed equidistant locations, $\tau = \frac{T}{4}, \frac{T}{2}, \frac{3T}{4}$; (3) three changes at random locations, sampled uniformly from $\{h, \ldots, T - h\}$ and constrained to be at least h apart. In each case we set $T = 200$ and $h = 20$. For the cases with three changes, the variances on each segment were $(1, 4, 1, 0.25)$. For each simulated data set, we estimated changepoints using binary segmentation (Figure 9) and calculated an estimate of the p-value corresponding to the first detected changepoint using the posterior mean of the Gaussian process, and using a Gaussian process followed by importance sampling. Each line on the figure corresponds to a different proportion of samples (of $N = 50, 100, 200$ total) used to fir the Gaussian process, with the remainder used for importance sampling, to allow fair comparison between methods.

Under H_0, the p-values obtained from the posterior mean of the GP (i.e. the proportion of GP samples is 1) lie along the straight line implying they follow the distribution $U(0, 1)$, whilst using importance sampling the p-values are slightly conservative for small N, although still close to uniform. Similarly, when we sample data with changes, for small N we get slightly greater power just using the GP posterior mean. For $N = 200$ there seems to be very little difference between these methods. Overall, it appears that using the

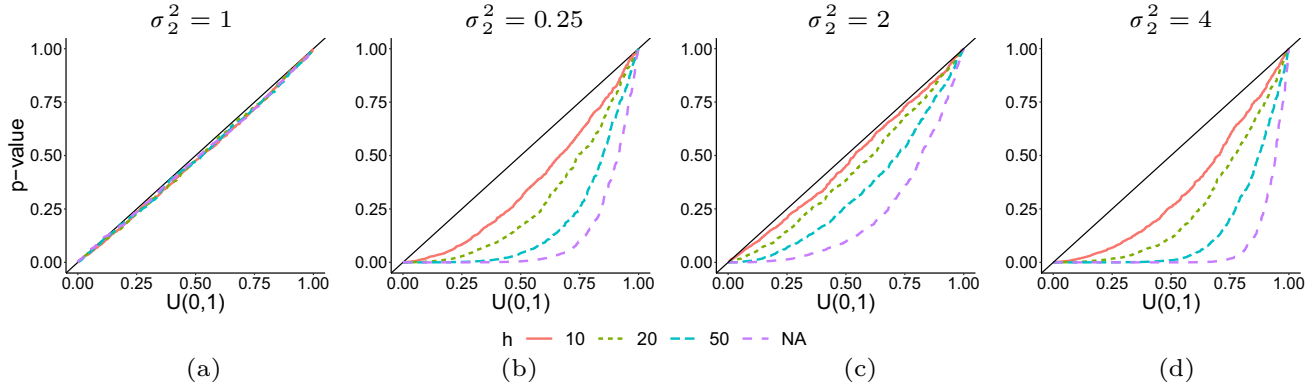

Fig. 7 QQ plots of p-values obtained using the CUSUM method. We set $T = 200$, and sampled data under H_0 (no change) and with a single changepoint sampled randomly from $1, \ldots, T$. The variance before the change was $\sigma_1^2 = 1$ in each case; the variance after the change is $\sigma^2 \in \{1, 0.25, 2, 4\}$. In each case we estimate a single changepoint using binary segmentation. We calculated p-values for different values of h, including the case where the window size is equal to the whole data set.

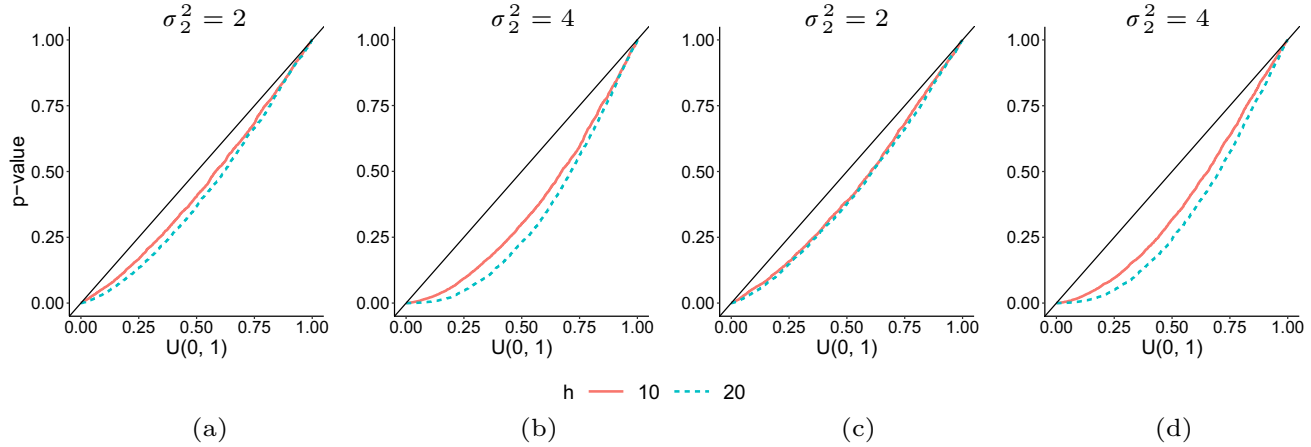

Fig. 8 QQ plots of p-values obtained using the CUSUM method, where we sample data from a model with $T = 200$ and 4 changepoints. In (a) and (b), the changes are equally spaced at $\tau = 40, 80, 120, 160$; in (c) and (d) they are random, sampled uniformly from $\{h, \ldots, T - h\}$ with the constraint that the distance between any two changes is at least h. The plots show sorted p-values for all estimated changepoints (up to 4 for each sampled data set).

posterior mean generally gives at least as good results as importance sampling, so this would be the preferred method.

How do we know when we have enough samples? We see that $N = 50$ seems to be sufficient in the above cases, as under H_0 the estimated p-values follow a uniform distribution, and under H_1 increasing N from 50 to 200 does not seem to lead to increased power. It is possible that we would need more samples if the number of data points T or the number of changes was larger.

5.3 Increasing power

Figure 10 shows p-values obtained when we estimated a single changepoint using the CUSUM statistic calculated p-values using the method described in Section 4.3. The number of samples (N_W) denotes the number of times we re-sample the additional parameters. We can see that under H_0 (left) we obtain valid p-values for any value of the number of samples N_W. When we sample from a model with a single change at $\frac{T}{2}$ (right), increasing N_W from 1 to 5 gives greater power against H_0, although increasing it beyond this does not seem to make much difference.

Figure 11 shows similar plots where we simulate from a model with a change at $\frac{T}{2}$, and use the likelihood ratio to estimate changepoints, and a Gaussian process with N_ϕ samples to estimate p-values. We see a similar pattern, where increasing the number of W samples initially leads to greater power.

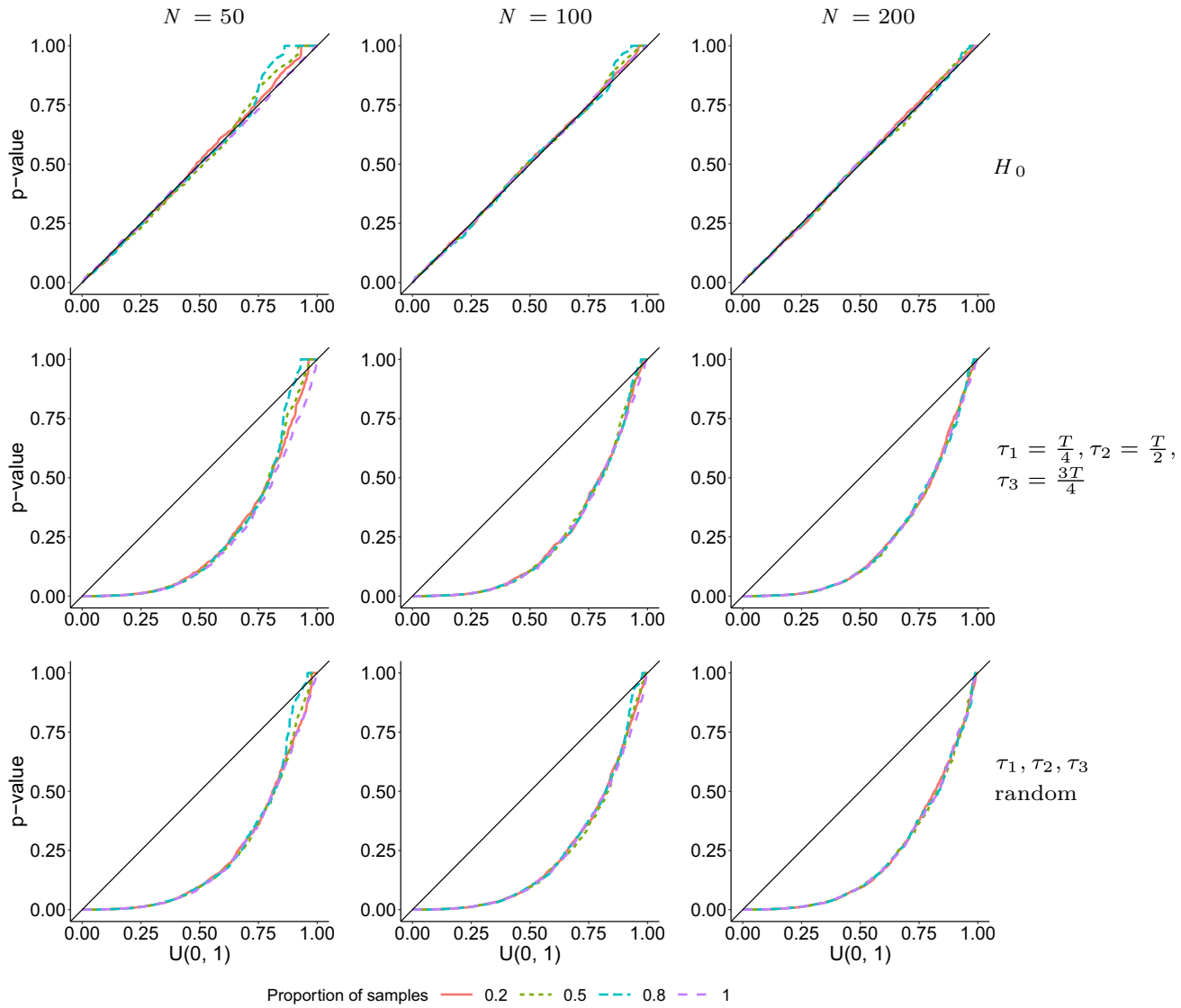

Fig. 9 QQ plots of p-values obtained using the likelihood ratio for three scenarios – no change, three equally spaced changes, and three random changes – using binary segmentation to estimate changepoints. Each row of the figure corresponds to a scenario, and columns correspond to the number of samples used. Each line corresponds to the proportion of samples that were used to fit the Gaussian process.

6 Application to financial data

To investigate how well our method works in a real-world setting, we applied our method to financial data (Figure 12). The data is log returns from S&P500 stock market data for 2264 days from July 2010 to October 2019[1]. We can see from Figure 12 that the data appears to have zero mean and several changes in variance. We estimated changepoints using three different methods: binary segmentation with the CUSUM statistic (top) and the likelihood ratio statistic (middle) and the penalized likelihood method PELT (Killick et al. 2012) (bottom), and calculated p-values for each estimated changepoint. We set thresholds so that in each case we estimated 11 changepoints, and set $h = 50$. For the cases where we had to estimate p-values, we used a Gaussian process with $N = 100$. On each plot estimated changepoints are shown by vertical lines, with red lines corresponding to estimated

[1] The data was downloaded from https://www.kaggle.com/datasets/mathisjander/s-and-p500-volatility-prediction-time-series-data

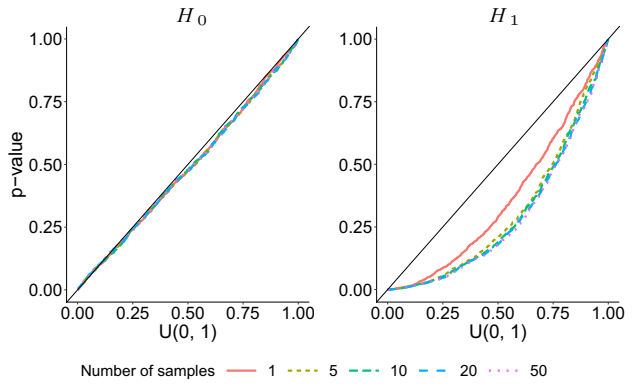

Fig. 10 QQ plots of p-values obtained when estimating a single changepoint using the CUSUM statistic, and calculating p-values using the method of 4.3.

changepoint locations whose p-values were smaller than 0.05 (after controlling for multiple testing using the Holm-Bonferroni method) and grey lines corresponding to p-values which were not significant. Using the likelihood ratio statistic generally seems to give better results than CUSUM, where there are more instances of multiple changepoints being estimated close together. However, these spurious changepoints generally have non-significant p-values associated with them, showing that our method for post-selection inference is effective at distinguishing true from false changepoint estimates.

7 Discussion

In this paper we have proposed a test for the piecewise constant change in variance model. We have developed methods for post-selection inference, which allow us to quantify uncertainty by computing p-values for estimated changes, and we have shown that our methods work on simulated and real data.

We have worked on the assumption that the data is i.i.d. Gaussian except for abrupt changes in variance. In practice, less restrictive models may be more appropriate in many cases, for example allowing the mean to change as well as the variance, or allowing autocorrelation between data points. Further work in this area could include extending or adapting our approach to apply to such models.

The strategies we have developed using Gaussian processes and importance sampling may be useful for other similar situations where we conduct post-selection inference but cannot characterise the selection event explicitly. More efficient implementations may be possible, for example using Bayesian optimisation (Frazier 2018) to adaptively choose ϕ values to consider based on some expected improvement in some appropriate measure of accuracy of the estimate of the p-value.

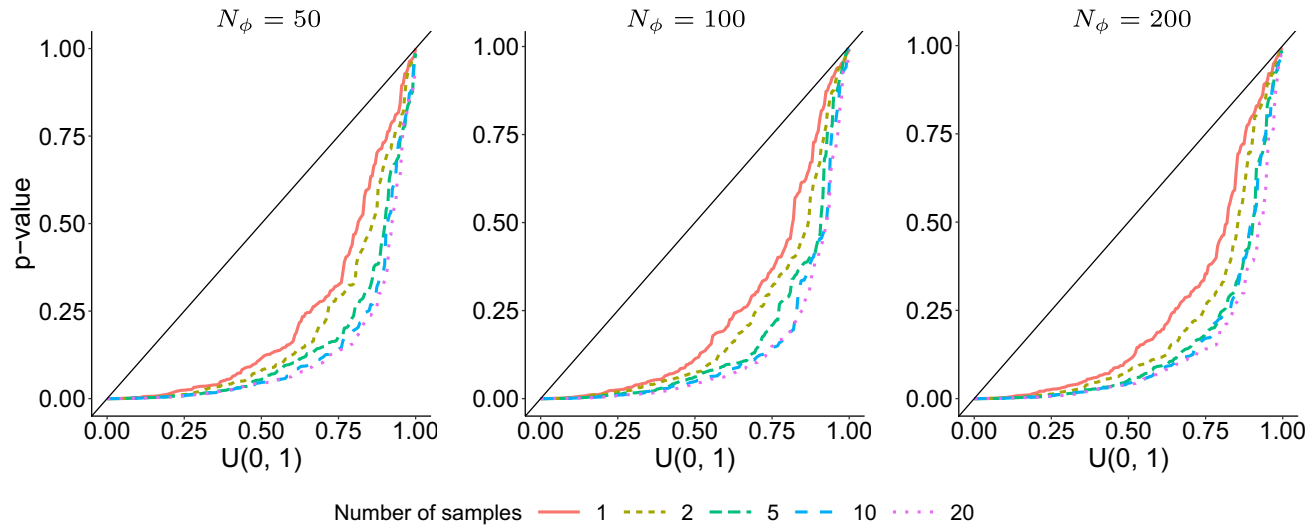

Fig. 11 P-values obtained sampling under H_1, and estimating a single changepoint using a Gaussian process, using the method of 4.3. N_ϕ is the number of ϕ samples for the Gaussian process.

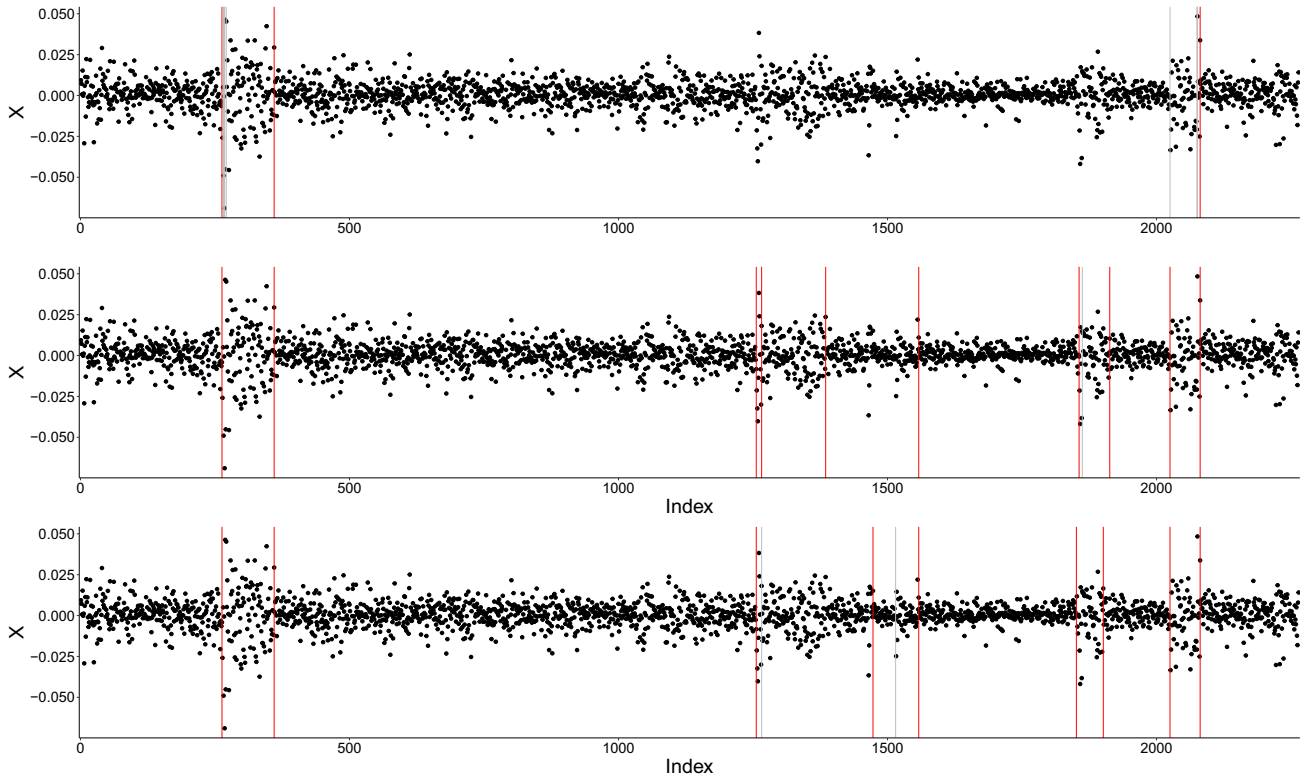

Fig. 12 Application of our method to financial log returns data. Changepoints are estimated using (i) CUSUM statistic with binary segmentation (top); (ii) likelihood ratio statistic with binary segmentation (middle); (iii) likelihood ratio statistic with PELT (Killick et al. 2012) (bottom). In each case we estimated 11 changepoints. Vertical lines indicate changepoints: those with a p-value smaller than 0.05 (after controlling for multiple testing using the Holm-Bonferroni method) are shown in red, and those with a p-value larger than 0.05 in grey.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11222-026-10881-1.

Acknowledgements This work is supported by EPSRC grant number EP/V053590/1.

Author contributions R.C. and P.F. contributed to the development of the methodology and to the writing of the manuscript. R.C. performed the simulations and data analysis and prepared the figures. All authors reviewed the manuscript.

Data availability No datasets were generated or analysed during the current study.

Declarations

Competing interests The authors declare no competing interests.

References

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

Amiri, A., Niaki, S.T.A., Moghadam, A.T.: A probabilistic artificial neural network-based procedure for variance change point estimation. Soft. Comput. **19**, 691–700 (2015)

Bayarri, M., Berger, J.O.: Quantifying surprise in the data and model verification. Bayes. Statistics **6**, 53–82 (1999)

Baranowski, R., Chen, Y., Fryzlewicz, P.: Narrowest-over-threshold detection of multiple change points and change-point-like features. J. Roy. Stat. Soc. B **81**(3), 649–672 (2019)

Barry, D., Hartigan, J.A.: A Bayesian analysis for change point problems. J. Am. Stat. Assoc. **88**(421), 309–319 (1993)

Caron, F., Doucet, A., Gottardo, R.: On-line changepoint detection and parameter estimation with application to genomic data. Stat. Comput. **22**, 579–595 (2012)

Carrington, R., Fearnhead, P.: Improving power by conditioning on less in post-selection inference for changepoints. Stat. Comput. **35**(1), 1–23 (2025)

Chen, Y.T., Jewell, S.W., Witten, D.M.: Quantifying uncertainty in spikes estimated from calcium imaging data. Biostatistics **24**(2), 481–501 (2023)
Cappello, L., Padilla, O.H.M.: Bayesian variance change point detection with credible sets. IEEE Transactions on Pattern Analysis and Machine Intelligence (2025)
Duy, V.N.L., Takeuchi, I.: More powerful conditional selective inference for generalized lasso by parametric programming. J. Mach. Learn. Res. **23**(1), 13544–13580 (2022)
Eichinger, B., Kirch, C.: A mosum procedure for the estimation of multiple random change points. Ber **24**, 526–564 (2018)
Fearnhead, P.: Exact and efficient Bayesian inference for multiple changepoint problems. Stat. Comput. **16**, 203–213 (2006)
Frick, K., Munk, A., Sieling, H.: Multiscale change point inference. J. Roy. Stat. Soc. B **76**(3), 495–580 (2014)
Fearnhead, P., Rigaill, G.: Relating and comparing methods for detecting changes in mean. Stat **9**(1), 291 (2020)
Frazier, P.I.: A tutorial on Bayesian optimization. arXiv preprint arXiv:1807.02811 (2018)
Fryzlewicz, P.: Wild binary segmentation for multiple change-point detection. Ann. Stat. **42**(6), 2243–2281 (2014)
Fryzlewicz, P.: Narrowest significance pursuit: inference for multiple change-points in linear models. Journal of the American Statistical Association **119**(546), 1633–1646 (2024)
Fryzlewicz, P.: Robust narrowest significance pursuit: Inference for multiple change-points in the median. Journal of Business & Economic Statistics **42**(4), 1389–1402 (2024)
Fithian, W., Sun, D., Taylor, J.: Optimal inference after model selection. arXiv preprint arXiv:1410.2597 (2014)
Hyun, S., G'Sell, M., Tibshirani, R.J.: Exact post-selection inference for the generalized lasso path. Electron. J. Stat. **12**(1), 1053–1097 (2018)
Hyun, S., Lin, K.Z., G'Sell, M., Tibshirani, R.J.: Post-selection inference for changepoint detection algorithms with application to copy number variation data. Biometrics **77**(3), 1037–1049 (2021)
Inclan, C., Tiao, G.C.: Use of cumulative sums of squares for retrospective detection of changes of variance. J. Am. Stat. Assoc. **89**(427), 913–923 (1994)
Jewell, S., Fearnhead, P., Witten, D.: Testing for a change in mean after changepoint detection. J. Roy. Stat. Soc. B **84**(4), 1082–1104 (2022)
Jiang, F., Zhao, Z., Shao, X.: Time series analysis of covid-19 infection curve: A change-point perspective. J. Econom **232**(1), 1–17 (2023)
Kovács, S., Bühlmann, P., Li, H., Munk, A.: Seeded binary segmentation: a general methodology for fast and optimal changepoint detection. Biometrika **110**(1), 249–256 (2023)
Killick, R., Fearnhead, P., Eckley, I.A.: Optimal detection of changepoints with a linear computational cost. J. Am. Stat. Assoc. **107**(500), 1590–1598 (2012)
Kuchibhotla, A.K., Kolassa, J.E., Kuffner, T.A.: Post-selection inference. Annu. Rev. Stat. Appl. **9**, 505–527 (2022)
Li, H., Munk, A., Sieling, H.: Fdr-control in multiscale change-point segmentation. Electron. J. Stat **10**, 918–959 (2016)
Muggeo, V.M., Adelfio, G.: Efficient change point detection for genomic sequences of continuous measurements. Bioinformatics **27**(2), 161–166 (2011)
Mehrizi, R.V., Chenouri, S.: Valid post-detection inference for change points identified using trend filtering. arXiv preprint arXiv:2104.12022 (2021)
Maidstone, R., Hocking, T., Rigaill, G., Fearnhead, P.: On optimal multiple changepoint algorithms for large data. Stat. Comput. **27**, 519–533 (2017)
Meier, A., Kirch, C., Cho, H.: mosum: A package for moving sums in change point analysis. J. Stat. Softw. **97**(8), 1–42 (2021)
Nam, C.F., Aston, J.A., Johansen, A.M.: Quantifying the uncertainty in change points. J. Time Ser. Anal. **33**(5), 807–823 (2012)
Reeves, J., Chen, J., Wang, X.L., Lund, R., Lu, Q.Q.: A review and comparison of changepoint detection techniques for climate data. J. Appl. Meteorol. Climatol. **46**(6), 900–915 (2007)
Shi, X., Beaulieu, C., Killick, R., Lund, R.: Changepoint detection: An analysis of the central England temperature series. J. Clim. **35**(19), 2729–2742 (2022)
Schröder, A.L., Fryzlewicz, P.: Adaptive trend estimation in financial time series via multiscale change-point-induced basis recovery. Statistics and Its Interface **6**, 449–461 (2013)
Shi, X., Gallagher, C., Lund, R., Killick, R.: A comparison of single and multiple changepoint techniques for time series data. Computational Statistics & Data Analysis **170**, 107433 (2022)
Scott, A.J., Knott, M.: A cluster analysis method for grouping means in the analysis of variance. Biometrics **30**, 507–512 (1974)
Saha, A., Witten, D., Bien, J.: Inferring independent sets of Gaussian variables after thresholding correlations. Journal of the American Statistical Association **120**(549), 370–381 (2025)