Testing for a Change in Mean After Changepoint Detection

Sean Jewell

Department of Statistics, University of Washington, Seattle, USA.

Paul Fearnhead

Department of Mathematics and Statistics, Lancaster University, Lancaster, UK.

Daniela Witten

Departments of Statistics and Biostatistics, University of Washington, Seattle, USA.

Summary. While many methods are available to detect structural changes in a time series, few procedures are available to quantify the uncertainty of these estimates post-detection. In this work, we fill this gap by proposing a new framework to test the null hypothesis that there is no change in mean around an estimated changepoint. We further show that it is possible to efficiently carry out this framework in the case of changepoints estimated by binary segmentation and its variants, ℓ_0 segmentation, or the fused lasso. Our setup allows us to condition on much less information than existing approaches, which yields higher powered tests. We apply our proposals in a simulation study and on a dataset of chromosomal guanine-cytosine content. These approaches are freely available in the `R` package `ChangepointInference` at `https://jewellsean.github.io/changepoint-inference/`.

Keywords: ℓ_0 optimization, binary segmentation, fused lasso, selective inference

1. Introduction

Detecting structural changes in a time series is a fundamental problem in statistics, with a variety of applications (Bai and Perron, 1998, 2003; Muggeo and Adelfio, 2010; Schröder and Fryzlewicz, 2013; Futschik et al., 2014; Xiao et al., 2019; Harchaoui and Lévy-Leduc, 2007; Hotz et al., 2013). A structural change refers to the phenomenon that at a certain (unknown) timepoint τ, the law of the data may change: that is, observations $y_1, \ldots, y_T$ are heterogeneous, in the sense that $y_1, \ldots, y_\tau \sim F$, whereas $y_{\tau+1}, \ldots, y_T \sim G$, for distribution functions $F \neq G$. In the presence of possible structural changes, it is of interest not only to estimate the times at which these changes occur — that is, the value of τ — but also to conduct statistical inference on the estimated changepoints.

In this paper, we consider the most common changepoint model,

$$Y_t = \mu_t + \epsilon_t, \quad \epsilon_t \overset{\text{iid}}{\sim} \mathrm{N}(0, \sigma^2), \quad t = 1, \ldots, T, \tag{1}$$

and assume that $\mu_1, \ldots, \mu_T$ is piecewise constant, in the sense that $\mu_{\tau_j+1} = \mu_{\tau_j+2} = \ldots = \mu_{\tau_{j+1}} \neq \mu_{\tau_{j+1}+1}$, for $j = 0, \ldots, K-1$, and $\mu_{\tau_K+1} = \mu_{\tau_K+2} = \ldots = \mu_{\tau_{K+1}}$. Here $0 = \tau_0 < \tau_1 < \ldots < \tau_K < \tau_{K+1} = T$, and $\tau_1, \ldots, \tau_K$ represent the true changepoints. Changepoint detection refers to the task of estimating the changepoint locations $\tau_1, \ldots, \tau_K$, and possibly the number of changepoints K. A huge number of proposals for this task have been made in the literature; see Truong et al. (2020) and Fearnhead and

Rigaill (2020) for a comprehensive review. These proposals can be roughly divided into two classes. One class iteratively searches for one changepoint at a time (Vostrikova, 1981; Olshen et al., 2004; Fryzlewicz, 2014; Badagián et al., 2015; Anastasiou and Fryzlewicz, 2019); the canonical example of this approach is binary segmentation. Another class of proposals simultaneously estimates all changepoints by solving a single optimization problem (Auger and Lawrence, 1989; Jackson et al., 2005; Tibshirani et al., 2005; Niu and Zhang, 2012; Killick et al., 2012; Haynes et al., 2017; Maidstone et al., 2017; Jewell and Witten, 2018; Fearnhead et al., 2019; Hocking et al., 2020; Jewell et al., 2020); examples include ℓ_0 segmentation and the fused lasso. We review these approaches in Section 2. Although not a focus of our work, changepoint estimation and inference have also been studied from a Bayesian perspective (Fearnhead, 2006; Nam et al., 2012; Ruanaidh and Fitzgerald, 2012).

In the single changepoint setting, estimation and inference for the location of the changepoint have been studied in the asymptotic (Hinkley, 1970; Yao, 1987; James et al., 1987; Bai, 1994) and non-asymptotic (Enikeeva and Harchaoui, 2019) settings. These approaches are typically extended to the multiple changepoint case by repeated application of tests for a single changepoint to sliding subsets of the data.

In the multiple changepoint setting, the multiscale approach of Frick et al. (2014) estimates the changepoints and provides confidence intervals for the changepoint locations and the unknown mean. However, this approach aims to control the probability of falsely detecting a change, and can lose power when there are many changes, particularly when they are hard to detect. Similarly, Ma and Yau (2016) produce asymptotically valid confidence intervals, but assume an asymptotic regime where all of the changepoints are detected with probability tending to one; this regime is unrealistic in many settings.

To overcome these issues, Li et al. (2016) develop a multiscale procedure that controls the false discovery rate of detections. But their method uses a very weak definition of a "true changepoint". In extreme cases, this could include an estimated changepoint that is almost as far as $T/2$ observations from an actual changepoint.

Non-parametric approaches to estimate multiple changepoints, such as moving-sum or scan statistics, have also been proposed (Bauer and Hackl, 1980; Hušková, 1990; Chu et al., 1995). Eichinger and Kirch (2018) recently showed consistency for the number and locations of changepoints, and established rates of convergence for moving-sum statistics.

Despite the huge literature on estimation and inference in changepoint detection problems, there remains a gap between the procedures used by practitioners to estimate changepoints and the statistical tools to assess the uncertainty of these estimates:

(a) Much of the theory for changepoint detection, especially in the multiple changepoint setting, focuses on specialized estimation procedures that are designed to facilitate inference. Therefore, these results are not directly applicable to the procedures commonly used by data analysts to estimate changepoints in practice.
(b) Classical techniques to test for a single changepoint give (mostly) asymptotic results, which involve complicated limiting distributions that do not directly apply to the multiple changepoint setting.
(c) Earlier works (mostly) provide confidence statements for the location of the changepoint. However, downstream analyses often rely on the size of the shift in mean, and not its precise location.

To address these limitations, we consider testing the null hypothesis that there is no change in mean around an estimated changepoint. Our interest lies not in determining whether there is a change in mean at a precise location, but rather, whether there is a change in mean nearby. This is a challenging task, since we must account for the fact that the changepoint was estimated from the data — and thus that the null hypothesis was chosen using the data — when deriving the null distribution for a test statistic. A recent promising line of work was introduced by Hyun et al. (2018) and Hyun et al. (2021), who develop valid tests for a change in mean associated with changepoints estimated with the generalized lasso or binary segmentation, respectively. Their work leverages recent results for selective inference in the regression setting (Fithian et al., 2014, 2015; Tibshirani et al., 2016; Lee et al., 2016; Tian et al., 2018). In greater detail, they compute the probability of observing such a large change in mean associated with an estimated changepoint, conditional on the fact that the changepoint was estimated from the data, as well as some additional quantities required for computational tractability. However, the fact that they condition on much more information than is used to choose the null hypothesis that is tested leads to a substantial reduction in power, as pointed out by Fithian et al. (2014), Lee et al. (2016), and Liu et al. (2018).

In this paper, we consider testing for a change in mean associated with an estimated changepoint, while conditioning on far less information than Hyun et al. (2018) and Hyun et al. (2021). In effect, we conduct local conditioning, as opposed to the global conditioning needed in Hyun et al. (2021). Moreover, we develop a test for a change in mean associated with changepoints detected via ℓ_0 segmentation, rather than only fused lasso and binary segmentation. Both of these advances lead to more powerful procedures for testing for the presence of changepoints. We develop this framework in detail for the change-in-mean model, but the general ideas can be applied more widely.

The rest of this paper is organized as follows. In Section 2, we review the relevant literature on changepoint detection and inference. In Section 3, we introduce a framework for inference in changepoint detection problems, which allows us to test for a change in mean associated with a changepoint estimated on the same dataset. In Sections 4 and 5, we develop efficient algorithms that allow us to instantiate this framework in the special cases of binary segmentation (Vostrikova, 1981) and its variants (Olshen et al., 2004; Fryzlewicz, 2014), and ℓ_0 segmentation (Killick et al., 2012; Maidstone et al., 2017); the case of the fused lasso (Tibshirani et al., 2016) is straightforward and addressed in the Supplementary Materials. Our framework is an improvement over the existing approaches for inference on the changepoints estimated using binary segmentation and its variants and the fused lasso; it is completely new in the case of ℓ_0 segmentation. After a preprint of this work appeared (Jewell et al., 2019), another research group developed a less efficient dynamic programming approach to assess the uncertainty in changepoints estimated from ℓ_0 segmentation (Duy et al., 2020). In Section 6, we present a comparison to some recent proposals from the literature in a simulation study. In Section 7, we show that our procedure leads to additional discoveries versus existing methods on a dataset of chromosomal guanine-cytosine (G-C) content. Extensions are in Section 8, and some additional details are deferred to the Supplementary Materials.

The `R` package `ChangepointInference`, along with code and data to reproduce all figures, can be found at `https://jewellsean.github.io/changepoint-inference`.

2. Background

2.1. Changepoint detection algorithms

2.1.1. Binary segmentation and its variants

Binary segmentation (Vostrikova, 1981) and its variants (Olshen et al., 2004; Fryzlewicz, 2014) search for changepoints by solving a sequence of local optimization problems. For the change-in-mean problem, these use the cumulative sum (CUSUM) statistic

$$g_{(s,\tau,e)}^\top y := \sqrt{\frac{1}{\frac{1}{|e-\tau|}+\frac{1}{|\tau+1-s|}}}(\bar{y}_{(\tau+1):e} - \bar{y}_{s:\tau}), \tag{2}$$

defined through a contrast $g_{(s,\tau,e)} \in \mathbb{R}^T$, which summarizes the evidence for a change at τ in the data $y_{s:e} := (y_s, \ldots, y_e)$ by the difference in the empirical mean of the data before and after τ (normalized to have the same variance for all τ). In (2), the notation $\bar{y}_{a:b}$ represents the sample mean of $(y_a, \ldots, y_b)$.

In binary segmentation (Vostrikova, 1981), the set of estimated changepoints is simply the set of local CUSUM maximizers: the first estimated changepoint maximizes the CUSUM statistic over all possible locations, $\hat{\tau}_1 = \underset{\tau\in[1:(T-1)]}{\text{argmax}} \left\{|g_{(1,\tau,T)}^\top y|\right\}$. Subsequent changepoints are estimated at the location that maximizes the CUSUM statistic when we consider regions of the data between previously estimated changepoints. For example, the second estimated changepoint is $\hat{\tau}_2 = \underset{\tau\in[1:(T-1)]\setminus\hat{\tau}_1}{\text{argmax}} \left\{|g_{(1,\tau,\hat{\tau}_1)}^\top y|1_{(1\le\tau<\hat{\tau}_1)} + |g_{(\hat{\tau}_1,\tau,T)}^\top y|1_{(\hat{\tau}_1<\tau<T)}\right\}$. We continue in this manner until a stopping criterion is met.

2.1.2. Simultaneous estimation of changepoints

As an alternative to sequentially estimating changepoints, we can simultaneously estimate all changepoints by minimizing a penalized cost that trades off fit to the data against the number of changepoints (Killick et al., 2012; Maidstone et al., 2017), i.e.

$$\underset{\substack{0=\tau_0<\tau_1<\cdots<\tau_K<\tau_{K+1}=T,\\ u_0,u_1,\ldots,u_K,K}}{\text{minimize}} \left\{\frac{1}{2}\sum_{k=0}^{K}\sum_{t=\tau_k+1}^{\tau_{k+1}}(y_t-u_k)^2 + \lambda K\right\}. \tag{3}$$

This is equivalent to solving an ℓ_0 penalized regression problem

$$\underset{\mu\in\mathbb{R}^T}{\text{minimize}} \left\{\frac{1}{2}\sum_{t=1}^{T}(y_t-\mu_t)^2 + \lambda\sum_{t=1}^{T-1}1_{(\mu_t\neq\mu_{t+1})}\right\}, \tag{4}$$

in the sense that the vector $\hat{\mu}$ that solves (4) satisfies $\{t : \hat{\mu}_t \neq \hat{\mu}_{t+1}\} = \{\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{K}}\}$, where $\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{K}}$ are the changepoints that solve (3). The tuning parameter λ specifies the improvement in fit to the data needed to add an additional changepoint.

Replacing the ℓ_0 penalty in (4) with an ℓ_1 penalty leads to the well-studied trend filtering or fused lasso optimization problem (Rudin et al., 1992; Tibshirani et al., 2005),

$$\underset{\mu\in\mathbb{R}^T}{\text{minimize}} \left\{\frac{1}{2}\sum_{t=1}^{T}(y_t-\mu_t)^2 + \lambda\sum_{t=1}^{T-1}|\mu_t-\mu_{t+1}|\right\}. \tag{5}$$

2.2. Existing methods for inference on changepoints post-detection

Suppose that we estimate some changepoints $\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{K}}$, and then wish to quantify the evidence for these estimated changepoints. We might naively apply a standard z-test for the difference in mean around each estimated changepoint. However, this approach is problematic, because it entails using the same data for testing that was used to estimate the changepoints, and thus to select the hypotheses to be tested. In particular, the z-statistic is not normally distributed under the null hypothesis. In the linear regression setting, Tibshirani et al. (2016) and Lee et al. (2016) have shown that it is possible to select and test hypotheses based on the same set of data, provided that we condition on the output of the hypothesis selection procedure.

Hyun et al. (2018) and Hyun et al. (2021) extend these ideas to the changepoint detection setting. For each changepoint $\hat{\tau}_j$ estimated using either binary segmentation, its variants, or the fused lasso, Hyun et al. (2021) propose to test whether there is a change in mean around $\hat{\tau}_j$. They construct the test statistic $\hat{d}_j \nu_j^\top Y$, where $\hat{d}_j$ is the sign of the estimated change in mean at $\hat{\tau}_j$, and ν_j is a T-vector of contrasts, defined as

$$\nu_{j,t} = \begin{cases} 0 & \text{if } t \leq \hat{\tau}_{j-1} \text{ or } t > \hat{\tau}_{j+1}, \\ \frac{1}{\hat{\tau}_j - \hat{\tau}_{j-1}} & \text{if } \hat{\tau}_{j-1} < t \leq \hat{\tau}_j, \\ -\frac{1}{\hat{\tau}_{j+1} - \hat{\tau}_j} & \text{if } \hat{\tau}_j < t \leq \hat{\tau}_{j+1}. \end{cases} \tag{6}$$

They consider the null hypothesis $H_0 : \hat{d}_j \nu_j^\top \mu = 0$ versus the one-sided alternative $H_1 : \hat{d}_j \nu_j^\top \mu > 0$. Since both $\hat{d}_j$ and ν_j are functions of the estimated changepoints, it is clear that valid inference requires somehow conditioning on the estimation process, in the spirit of Tibshirani et al. (2016) and Lee et al. (2016). Define $\mathcal{M}(y)$ to be the set of changepoints estimated from the data y, i.e., $\mathcal{M}(y) = \{\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{K}}\}$. Then, it is tempting to define the p-value as $\Pr_{H_0}\left(\hat{d}_j \nu_j^\top Y \geq \hat{d}_j \nu_j^\top y \mid \mathcal{M}(Y) = \mathcal{M}(y)\right)$. However, this p-value is not immediately amenable to the selective inference framework proposed by Tibshirani et al. (2016) and Lee et al. (2016), which requires that the conditioning set be polyhedral; i.e., the conditioning set can be written as $\{Y : \boldsymbol{A}Y \leq b\}$ for a matrix $\boldsymbol{A}$ and vector b. Thus, in the case of binary segmentation, Hyun et al. (2021) condition on three additional quantities: (i) the order in which the estimated changepoints enter the model, $\mathcal{O}(Y) = \mathcal{O}(y)$; (ii) the sign of the change in mean due to the estimated changepoints, $\Delta(Y) = \Delta(y) = (\hat{d}_1, \ldots, \hat{d}_{\hat{K}})$; (iii) $\Pi^\perp_{\nu_j} Y = \Pi^\perp_{\nu_j} y$, where $\Pi^\perp_{\nu_j} = I - \nu_j \nu_j^\top / ||\nu_j||_2^2$ is the orthogonal projection matrix onto the subspace that is orthogonal to ν_j. Conditioning on (i) and (ii) ensures that the conditioning set is polyhedral, whereas conditioning on (iii) ensures that the test statistic is a pivot. This leads to the p-value

$$\Pr_{H_0}\left(\hat{d}_j \nu_j^\top Y \geq \hat{d}_j \nu_j^\top y \mid \mathcal{M}(Y) = \mathcal{M}(y), \mathcal{O}(Y) = \mathcal{O}(y), \Delta(Y) = \Delta(y), \Pi^\perp_{\nu_j} Y = \Pi^\perp_{\nu_j} y\right), \tag{7}$$

which can be easily computed because the conditional distribution of $\hat{d}_j \nu_j^\top Y$ is a Gaussian truncated to an interval. For slightly different conditioning sets, Hyun et al. (2021) show similar results for variants of binary segmentation and for the fused lasso.

Importantly, Hyun et al. (2021) choose the conditioning set in (7) for computational reasons: there is no clear statistical motivation for conditioning on $\mathcal{O}(Y) = \mathcal{O}(y)$ and

$\Delta(Y) = \Delta(y)$. Furthermore, it might be possible to account for the fact that changepoints are estimated from the data without conditioning on the full set $\mathcal{M}(Y) = \mathcal{M}(y)$. In fact, Fithian et al. (2014) argue that when conducting selective inference, it is better to condition on less information, i.e. to condition on Y being in a larger set of possible data, since conditioning on more information reduces the Fisher information that remains in the conditional distribution of the data.

For this reason, in the regression setting, some recent proposals seek to increase the size of the conditioning set. Lee et al. (2016) propose to condition on just the selected model, rather than on the selected model and the corresponding coefficient signs, by considering all possible configurations of the signs of the estimated coefficients. Unfortunately, this comes at a significant computational cost. Continuing in this vein, Liu et al. (2018) partition the selected variables into high value and low value subsets, and then condition on the former and the variable of interest.

In this paper, we develop new insights that allow us to test the null hypothesis that there is no change in mean at an estimated changepoint, without restriction to the polyhedral conditioning sets pursued by Hyun et al. (2021). Because we do not need to use the full conditioning set in (7), we obtain higher-powered tests. Additionally, since we avoid conditioning on $\Delta(Y) = \Delta(y)$, we can consider two-sided tests of

$$H_0 : \nu^\top \mu = 0 \text{ versus } H_1 : \nu^\top \mu \neq 0, \tag{8}$$

rather than the one-sided tests considered by Hyun et al. (2021). In (8), and for the remainder of this paper, we suppress the j subscript on ν_j for notational convenience. Thus, the vector ν should be interpreted as shorthand for ν_j.

It is natural to ask whether we can avoid the complications of selective inference and use alternative approaches that control the false discovery rate (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001; Barber and Candès, 2015; Candes et al., 2018). However, these alternatives are not suitable for the changepoint setting in the following sense. Often we do not want to know if a true changepoint is *exactly* at $\hat{\tau}_j$, but rather whether there is a true changepoint *near* $\hat{\tau}_j$; that is, we are willing to accept small estimation errors in the location of a changepoint. With a suitable choice of ν in (8), we can test whether there is a change in mean near $\hat{\tau}_j$, where *near* can be defined appropriately for a given application. By contrast, while knockoffs (Barber and Candès, 2015) or a related approach could likely be used to test for a change in mean at a precise location, in our experience such approaches tend to have almost no power to detect modest changes in the mean, due to the large uncertainty in the precise location of the change.

2.3. Toy example illustrating the cost of conditioning

In this section, we demonstrate that the power of a test of (8) critically depends on the size of the conditioning set. In Figure 1, we consider two choices for the conditioning set. In panel a), we condition on $\mathcal{M}(Y) = \mathcal{M}(y), \mathcal{O}(Y) = \mathcal{O}(y), \Delta(Y) = \Delta(y)$, and $\Pi_\nu^\perp Y = \Pi_\nu^\perp y$: this is essentially the test proposed by Hyun et al. (2021). In panel b) we condition on just $\mathcal{M}(Y) = \mathcal{M}(y)$ and $\Pi_\nu^\perp Y = \Pi_\nu^\perp y$. Observed data (grey points) are simulated according to (1) with the true underlying mean displayed in blue. 19-step binary segmentation is used to estimate changepoints, which are displayed as vertical

lines, and are colored based on whether the associated p-value is less than 0.05 (blue) or greater than 0.05 (red). In this example, *conditioning on less information allows us to reject the null hypothesis when it is false more often (i.e., we obtain five additional true positives), without inflating the number of false positives.*

With this toy example in mind, we turn to our proposal in the following section. It does not require polyhedral conditioning sets, and thus allows us to condition on much less information than previously possible.

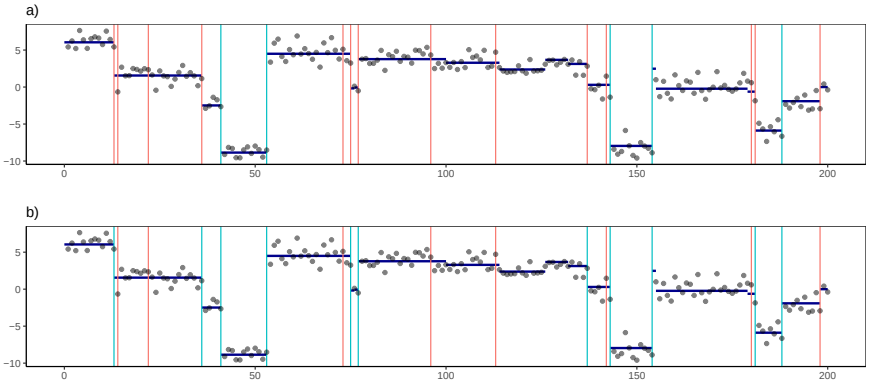

Fig. 1. The power of a test of (8) critically depends on the size of the conditioning set. Observations (displayed in grey) were simulated from (1) with $\sigma = 1$ and $\mu_1, \ldots, \mu_T$ displayed in dark blue. Our proposed test of (8) was conducted for each of the changepoints estimated via 19-step binary segmentation. Estimated changepoints for which the p-value is less than 0.05 are displayed in blue, and the remaining estimated changepoints are displayed in red. In panel (a), we conducted our proposed test by conditioning on $\mathcal{M}(Y) = \mathcal{M}(y), \mathcal{O}(Y) = \mathcal{O}(y), \Delta(Y) = \Delta(y)$, and $\Pi_\nu^\perp Y = \Pi_\nu^\perp y$ (this is essentially the proposal of Hyun et al. (2021)). In panel (b), we conditioned on the much larger set $\mathcal{M}(Y) = \mathcal{M}(y)$ and $\Pi_\nu^\perp Y = \Pi_\nu^\perp y$.

3. Two new tests with larger conditioning sets

In this section, we consider testing a null hypothesis of the form (8) using a much larger conditioning set than used by Hyun et al. (2021). Our approach is similar in spirit to the "general recipe" proposed in Section 6 of Liu et al. (2018). We consider two possible forms of the contrast vector ν in Sections 3.1 and 3.2.

3.1. A test of no change in mean between neighboring changepoints

We first consider testing the null hypothesis (8) for ν defined in (6). In order to account for the fact that we estimated the changepoints, it is natural to condition on all of the estimated changepoints, $\mathcal{M}(y) = \{\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{K}}\}$. Thus, we define the p-value

$$p \equiv \Pr_{H_0}\left(|\nu^\top Y| \geq |\nu^\top y| \mid \mathcal{M}(Y) = \mathcal{M}(y), \Pi_\nu^\perp Y = \Pi_\nu^\perp y\right). \tag{9}$$

As in Hyun et al. (2021), we condition on $\Pi_\nu^\perp Y = \Pi_\nu^\perp y$ for technical reasons; see Fithian et al. (2014) for additional discussion. Roughly speaking, (9) asks: "Out of all data sets

yielding this particular set of changepoints, what is the probability, under the null that there is no true change in mean at this location, of observing such a large difference in mean between the segments on either side of $\hat{\tau}_j$?" Our next result reveals that computing (9) involves a univariate truncated normal distribution. Related results appear in Tibshirani et al. (2016), Lee et al. (2016), and Liu et al. (2018).

THEOREM 1. *The p-value in* (9) *is equal to*

$$p = Pr\Big(|\phi| \geq |\nu^\top y| \mid \mathcal{M}(y'(\phi)) = \mathcal{M}(y)\Big), \tag{10}$$

where $\phi \sim N(0, \|\nu\|^2\sigma^2)$ *and where*

$$y'(\phi) = y - \frac{\nu\nu^\top y}{||\nu||_2^2} + \frac{\nu\phi}{||\nu||_2^2}. \tag{11}$$

In light of Theorem 1, to evaluate (9) we must simply characterize the set

$$\mathcal{S} = \{\phi : \mathcal{M}(y'(\phi)) = \mathcal{M}(y)\}; \tag{12}$$

as we will see in Section 3.3, this is the set of perturbations of y that result in no change to the estimated changepoints. In Sections 4 and 5, we do exactly this in the case of binary and ℓ_0 segmentation, respectively. We discuss the fused lasso in Section D.1 of the Supplementary Materials.

3.2. A test of no change in mean within a fixed window size

We now consider testing the null hypothesis (8) with ν given by

$$\nu_t = \begin{cases} 0 & \text{if } t \leq \hat{\tau}_j - h \text{ or } t > \hat{\tau}_j + h, \\ \frac{1}{h} & \text{if } \hat{\tau}_j - h < t \leq \hat{\tau}_j, \\ -\frac{1}{h} & \text{if } \hat{\tau}_j < t \leq \hat{\tau}_j + h. \end{cases} \tag{13}$$

Thus, we are testing whether the mean in a window to the left of the jth estimated changepoint equals the mean in a window to the right of the jth estimated changepoint, for a fixed window size $h > 0$. When considering this null hypothesis, it makes sense to condition only on the jth estimated changepoint, leading to a p-value defined as

$$p \equiv \mathrm{Pr}_{H_0}\left(|\nu^\top Y| \geq |\nu^\top y| \mid \hat{\tau}_j \in \mathcal{M}(Y), \Pi_\nu^\perp Y = \Pi_\nu^\perp y\right), \tag{14}$$

where once again, we condition on $\Pi_\nu^\perp Y = \Pi_\nu^\perp y$ for technical reasons. Roughly speaking, (14) asks: "Out of all data sets yielding a changepoint at $\hat{\tau}_j$, what is the probability, under the null that there is no true change in mean at this location, of observing such a large difference in mean between the windows of size h on either side of $\hat{\tau}_j$?"

The p-values in (14) and (9) are calculated for slightly different null hypotheses: the null for (14) is that there is no changepoint within a distance h of the estimated changepoint, $\hat{\tau}_j$. By contrast, (9) tests for no change in mean between the estimated changepoints immediately before and after $\hat{\tau}_j$. Furthermore, (14) conditions on less information. We believe that in many applications, the null hypothesis assumed by (14)

is more natural and informative, since it allows a practitioner to specify how accurately they want to detect changepoint locations, and it avoids rejecting the null due to changes that are arbitrarily far away from $\hat{\tau}_j$. Moreover, the ability to condition on less information intuitively should lead to higher power. If required, the ideas used to calculate (14) could also be applied to test for the null hypothesis assumed by (9), while conditioning on less information. We further investigate these issues in Sections 6 and 8.1.

Theorem 1 can be extended to show that (14) is equal to

$$p = \text{Pr}\left(|\phi| \geq |\nu^\top y| \mid \hat{\tau}_j \in \mathcal{M}(y'(\phi))\right), \tag{15}$$

where $\phi \sim N(0, \|\nu\|^2\sigma^2)$, and where $y'(\phi)$ was defined in (11). Thus, computing the p-value requires characterizing the set

$$\mathcal{S} = \{\phi : \hat{\tau}_j \in \mathcal{M}(y'(\phi))\}; \tag{16}$$

this is the set of perturbations of y that result in estimating a changepoint at $\hat{\tau}_j$.

We show in Sections 4 and 5 that $\mathcal{S}$ can be efficiently characterized for binary and ℓ_0 segmentation. We discuss the fused lasso in Section D.1 of the Supplementary Materials.

3.3. *Intuition for $y'(\phi)$ and $\mathcal{S}$*

To gain intuition for $y'(\phi)$ in (11), we consider ν defined in (6) (similar results apply for ν defined in (13)). We see that

$$y'_t(\phi) \equiv \begin{cases} y_t & \text{if } t \leq \hat{\tau}_{j-1} \text{ or } t > \hat{\tau}_{j+1}, \\ y_t + \frac{\phi - \nu^\top y}{1 + \frac{\hat{\tau}_j - \hat{\tau}_{j-1}}{\hat{\tau}_{j+1} - \hat{\tau}_j}} & \text{if } \hat{\tau}_{j-1} < t \leq \hat{\tau}_j, \\ y_t - \frac{\phi - \nu^\top y}{1 + \frac{\hat{\tau}_{j+1} - \hat{\tau}_j}{\hat{\tau}_j - \hat{\tau}_{j-1}}} & \text{if } \hat{\tau}_j < t \leq \hat{\tau}_{j+1}. \end{cases} \tag{17}$$

Thus, $y'_t(\phi)$ is equal to y_t for $t \leq \hat{\tau}_{j-1}$ or $t > \hat{\tau}_{j+1}$, and otherwise equals the observed data perturbed by a function of ϕ around $\hat{\tau}_j$. In other words, we can view $y'(\phi)$ as a perturbation of the observed data y by a quantity proportional to $\phi - \nu^\top y$, within some window of $\hat{\tau}_j$. Furthermore, $\mathcal{S} = \{\phi : \mathcal{M}(y'(\phi)) = \mathcal{M}(y)\}$ is the set of such perturbations that do not affect the set of estimated changepoints.

Figure 2 illustrates the intuition behind $y'(\phi)$ in a simulated example of length 200 with a change in mean at the 100th position, and where $\phi = \nu^\top y = -1$. In panel a), the observed data are displayed. Here, 1-step binary segmentation estimates $\hat{\tau}_1 = 100$. In panel b), the observed data are perturbed using $\phi = 0$ so that 1-step binary segmentation no longer estimates a changepoint at the 100th position. Conversely, in panel c), the data are perturbed using $\phi = -2$ to exaggerate the change at timepoint 100; 1-step binary segmentation again estimates a changepoint at the 100th position. Hence, for 1-step binary segmentation, -1 and -2 are in $\mathcal{S} = \{\phi : \mathcal{M}(y'(\phi)) = \mathcal{M}(y)\}$, but 0 is not. The procedure from Section 4 for efficiently characterizing $\mathcal{S}$ gives $\mathcal{S} = \{\phi : \mathcal{M}(y'(\phi)) = \mathcal{M}(y)\} = (-\infty, -0.2] \cup [0.2, \infty)$; see panel d) of Figure 2.

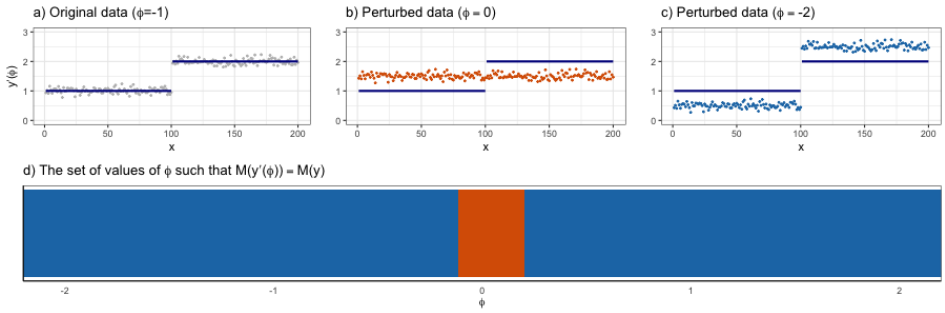

Fig. 2. a) A simulated dataset with $\phi = \nu^\top y = -1$ is displayed in grey, and the true underlying mean is shown in blue. b) The perturbed dataset $y'(\phi)$ is shown, with $\phi = \nu^\top y = 0$. The perturbed dataset does not have a change in mean at the 100th timepoint, and so 1-step binary segmentation does not detect a changepoint at that position. c) The perturbed dataset $y'(\phi)$ is shown, with $\phi = \nu^\top y = -2$. There is now a very pronounced change in mean at the 100th timepoint, and so 1-step binary segmentation does detect a changepoint at that position. d) Values of ϕ for which $\mathcal{M}(y'(\phi)) = \mathcal{M}(y)$ are shown in blue, and those for which $\mathcal{M}(y'(\phi)) \neq \mathcal{M}(y)$ are shown in red, for $\mathcal{M}$ given by 1-step binary segmentation.

4. Efficient characterization of (12) and (16) for binary segmentation and its variants

We now turn our attention to computing the set (12) for k-step binary segmentation; (16) is detailed in Section B.4 of the Supplementary Materials. We begin by paraphrasing Proposition 1 of Hyun et al. (2021).

PROPOSITION 1 (PROPOSITION 1 OF HYUN ET AL., (2021)). *The set of y for which k-step binary segmentation yields a given set of estimated changepoints, orders, and signs is polyhedral, and takes the form $\{y : \mathbf{\Gamma} y \leq 0\}$ for a $k(2T-k-3) \times T$ matrix $\mathbf{\Gamma}$, which is a function of the estimated changepoints, orders, and signs.*

Recall from Section 2.2 that $\mathcal{M}(y)$, $\mathcal{O}(y)$, and $\Delta(y)$ are the locations, orders, and signs of the changepoints estimated from k-step binary segmentation applied to data y. We first present a corollary of Proposition 1.

COROLLARY 1. *The set $\mathcal{S}$ in (12) is the union of sets of the form $\{\phi : \mathcal{M}(y'(\phi)) = m, \mathcal{O}(y'(\phi)) = o, \Delta(y'(\phi)) = d\}$, each of which is an interval. Furthermore, each of these intervals can be computed analytically using the matrix $\mathbf{\Gamma}$ in Proposition 1.*

This leads to the following result.

PROPOSITION 2. *Let $\mathcal{I}$ denote the set of orders and signs of the changepoints that can be obtained via a perturbation of y that yields changepoints $\mathcal{M}(y)$: that is,*

$$\mathcal{I} := \left\{(o,d) : \exists \phi \in \mathbb{R} \text{ such that } \mathcal{M}(y'(\phi)) = \mathcal{M}(y), \mathcal{O}(y'(\phi)) = o, \Delta(y'(\phi)) = d\right\}. \quad (18)$$

Then, there exists an index set $\mathcal{J}$ *and scalars* $\ldots, a_{-3}, a_{-2}, a_{-1}, a_0, a_1, a_2, a_3, \ldots$ *such that* $|\mathcal{J}| = |\mathcal{I}|$*, and the set* $\mathcal{S}$ *in* (12) *is the union of* $|\mathcal{J}|$ *intervals: that is,*

$$\mathcal{S} = \{\phi : \mathcal{M}(y'(\phi)) = \mathcal{M}(y)\} = \bigcup_{i \in \mathcal{J}} [a_i, a_{i+1}]. \tag{19}$$

In Proposition 2, $|\mathcal{J}|$ and $|\mathcal{I}|$ denote the cardinality of the sets $\mathcal{J}$ and $\mathcal{I}$, respectively. Importantly, $|\mathcal{J}| = |\mathcal{I}| \ll 2^k k!$, which is the total number of possible orders and signs for the k changepoints. To simplify notation in (19), we have used the convention that if $a_i = -\infty$ then $[a_i, a_{i+1}]$ should be interpreted as $(a_i, a_{i+1}]$, and similarly if $a_{i+1} = \infty$ then $[a_i, a_{i+1}]$ should be interpreted as $[a_i, a_{i+1})$.

Proposition 3 guarantees that Proposition 2 is of practical use.

PROPOSITION 3. $\bigcup_{i \in \mathcal{J}} [a_i, a_{i+1}]$ *defined in* (19) *can be efficiently computed.*

Proposition 3 follows from a simple argument; see Algorithm 1 of the Supplementary Materials for additional details. We first run k-step binary segmentation on the data y to obtain estimated changepoints $\mathcal{M}(y)$, orders $\mathcal{O}(y)$, and signs $\Delta(y)$. We then apply Corollary 1 to obtain the interval $[a_0, a_1] = \{\phi : \mathcal{M}(y'(\phi)) = \mathcal{M}(y), \mathcal{O}(y'(\phi)) = \mathcal{O}(y), \Delta(y'(\phi)) = \Delta(y)\}$. By construction, $[a_0, a_1] \subset \mathcal{S}$. The set $\mathcal{J}$ indexes the intervals comprising the set $\mathcal{S}$; therefore, we set $\mathcal{J} = \{0\}$.

Next, for some small $\eta > 0$, we apply Corollary 1 with $m = \mathcal{M}(y'(a_1 + \eta))$, $o = \mathcal{O}(y'(a_1 + \eta))$, and $d = \Delta(y'(a_1 + \eta))$ to obtain the interval $[a_1, a_2] = \{\phi : \mathcal{M}(y'(\phi)) = m, \mathcal{O}(y'(\phi)) = o, \Delta(y'(\phi)) = d\}$. (If the left endpoint of this interval does not equal a_1, then we must use a smaller value of η.) We then check whether $\mathcal{M}(y'(a_1 + \eta)) = \mathcal{M}(y)$. If so, then $[a_1, a_2] \subset \mathcal{S}$ and we set $\mathcal{J}$ equal to $\mathcal{J} \cup \{1\}$; if not, then $\mathcal{J}$ remains unchanged. Next, we apply Corollary 1 with $m = \mathcal{M}(y'(a_2 + \eta))$, $o = \mathcal{O}(y'(a_2 + \eta))$, and $d = \Delta(y'(a_2 + \eta))$ to obtain the interval $[a_2, a_3]$. We then determine whether $[a_2, a_3] \subset \mathcal{S}$; if so, then we set $\mathcal{J}$ equal to $\mathcal{J} \cup \{2\}$, and if not, then $\mathcal{J}$ remains unchanged. We continue in this way until we reach an interval containing ∞. We then repeat this process in the other direction, applying Corollary 1 with $m = \mathcal{M}(y'(a_0 - \eta))$, $o = \mathcal{O}(y'(a_0 - \eta))$, and $d = \Delta(y'(a_0 - \eta))$, determining whether the resulting interval $[a_{-1}, a_0]$ belongs to $\mathcal{S}$, and updating $\mathcal{J}$ accordingly. We continue until we arrive at an interval containing $-\infty$.

Proposition 4 shows that this procedure can be stopped early in order to obtain conservative p-values, while substantially reducing computational costs.

PROPOSITION 4. *Let* $\tilde{\mathcal{S}}$ *be defined as the set*

$$\tilde{\mathcal{S}} = (-\infty, a_{-r}] \cup \left(\bigcup_{i \in \mathcal{J} \cap \{-r, \ldots, r'\}} [a_i, a_{i+1}] \right) \cup [a_{r'+1}, \infty),$$

for some r *and* r' *such that* $a_{-r} \leq -|\nu^\top y|$ *and* $a_{r'+1} \geq |\nu^\top y|$*. Then the p-value obtained by conditioning on* $\{\phi \in \tilde{\mathcal{S}}\}$ *exceeds the p-value obtained by conditioning on* $\{\phi \in \mathcal{S}\}$*:*

$$Pr(|\phi| \geq |\nu^\top y| \mid \phi \in \tilde{\mathcal{S}}) \geq Pr(|\phi| \geq |\nu^\top y| \mid \phi \in \mathcal{S}).$$

Section B of the Supplementary Materials contains proofs of Corollary 1 and Propositions 2 and 4. In that section, we also show that Propositions 2 and 3 can be easily modified to characterize (16). Section D.1 of the Supplementary Materials contains a straightforward modification of this procedure to characterize (12) and (16) in the case of the fused lasso.

It turns out that all of the ideas developed in this section for binary segmentation can be directly applied to the circular binary segmentation proposal of Olshen et al. (2004) and the wild binary segmentation proposal of Fryzlewicz (2014). In particular, it is shown in the Supplementary Materials of Hyun et al. (2021) that a result almost identical to Proposition 1 holds for these two variants of binary segmentation, for a different matrix $\mathbf{\Gamma}$. This means that Propositions 2–4 follow directly.

We have assumed that k, the number of steps of binary segmentation, is pre-specified. Hyun et al. (2021) showed that a stopping rule based on the Bayesian information criterion yields a polyhedral conditioning set. Hence, we could extend the ideas in this section to select k adaptively. However, as shown by Hyun et al. (2021), this approach requires conditioning on additional information, and thereby results in a loss of power.

5. Efficient characterization of (12) and (16) for ℓ_0 segmentation

In this section, we develop an efficient algorithm to analytically characterize $\mathcal{S}$ in (12) for the ℓ_0 segmentation problem (4) with a fixed value of λ; Section C.2 of the Supplementary Materials considers $\mathcal{S}$ in (16). Recall that in the context of $\mathcal{S}$ in (12), $y'(\phi)$ is defined in (11) and ν is defined in (6).

Roughly speaking, we show that it is possible to write (12) in terms of the cost to segment the perturbed data $y'(\phi)$. To compute the necessary cost functions, we derive recursions similar to those in Rigaill (2015) and Maidstone et al. (2017). However, these recursions involve functions of two variables, rather than one. Consequently, fundamentally different techniques are required for efficient computation.

5.1. Recharacterizing $\mathcal{S}$ in (12) in terms of $C(\phi)$ and $C'(\phi)$

Let $\hat{K}$ denote the number of estimated changepoints resulting from ℓ_0 segmentation (4) on the data y with fixed tuning parameter value λ, and let $\hat{\tau}_1 < \ldots < \hat{\tau}_{\hat{K}}$ denote the positions of those estimated changepoints; for notational convenience, let $\hat{\tau}_0 \equiv 0$ and $\hat{\tau}_{\hat{K}+1} \equiv T$. For a given value of ϕ, $\mathcal{M}(y'(\phi)) = \mathcal{M}(y)$ if and only if the cost of ℓ_0 segmentation of the data $y'(\phi)$ with the changepoints restricted to occur at $\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{K}}$,

$$C(\phi) = \min_{u_0, u_1, \ldots, u_{\hat{K}}} \left\{ \frac{1}{2} \sum_{k=0}^{\hat{K}} \sum_{t=\hat{\tau}_k+1}^{\hat{\tau}_k} (y'_t(\phi) - u_k)^2 + \lambda \hat{K} \right\}, \tag{20}$$

is no greater than the cost of ℓ_0 segmentation of $y'(\phi)$,

$$C'(\phi) = \min_{\substack{0=\tau_0<\tau_1<\cdots<\tau_K<\tau_{K+1}=T, \\ u_0, u_1, \ldots, u_K, K}} \left\{ \frac{1}{2} \sum_{k=0}^{K} \sum_{t=\tau_k+1}^{\tau_{k+1}} (y'_t(\phi) - u_k)^2 + \lambda K \right\}. \tag{21}$$

In other words,

$$\mathcal{S} = \{\phi : C(\phi) \leq C'(\phi)\}. \tag{22}$$

The following result follows from the fact that (11) and (6) imply that for all $j = 0, \ldots, \hat{K}$, there exists a constant c_j such that $y'_t(\phi) = y_t + c_j$ for all $t = \hat{\tau}_j + 1, \ldots, \hat{\tau}_{j+1}$.

PROPOSITION 5. *$C(\phi)$ is a constant function of ϕ. That is, $C(\phi) = C(\phi')$ for all ϕ and ϕ'.*

Proposition 5 implies that $C(\phi)$ is easy to calculate: we just compute it for a single value of ϕ, e.g. $\phi = \nu^T y$. Hence, to characterize $\mathcal{S}$ using (22), it remains to calculate $C'(\phi)$, i.e., to perform ℓ_0 segmentation on $y'(\phi)$. In the interest of computational tractability, we need a single procedure that works for all values of ϕ simultaneously, rather than (for instance) having to repeat the procedure for values of ϕ on a fine grid.

Let $\text{Cost}(y'_{1:\hat{\tau}_j}(\phi); u)$ be the cost of segmenting $y'_{1:\hat{\tau}_j}(\phi)$ with $\mu_{\hat{\tau}_j} = u$. Then $C'(\phi)$ can be decomposed into the cost of segmenting the data $y'(\phi)$ with a changepoint at $\hat{\tau}_j$,

$$C'_{\hat{\tau}_j}(\phi) = \min_u \left\{ \text{Cost}(y'_{1:\hat{\tau}_j}(\phi); u) \right\} + \min_{u'} \left\{ \text{Cost}(y'_{T:(\hat{\tau}_j+1)}(\phi); u') \right\} + \lambda, \tag{23}$$

and the cost of segmenting the data $y'(\phi)$ without a changepoint at $\hat{\tau}_j$,

$$C'_{\neg\hat{\tau}_j}(\phi) = \min_u \left\{ \text{Cost}(y'_{1:\hat{\tau}_j}(\phi); u) + \text{Cost}(y'_{T:(\hat{\tau}_j+1)}(\phi); u) \right\}. \tag{24}$$

Combining (23) and (24), we have

$$C'(\phi) = \min \left\{ C'_{\hat{\tau}_j}(\phi), C'_{\neg\hat{\tau}_j}(\phi) \right\}. \tag{25}$$

Next, we will show that it is possible to analytically calculate $\text{Cost}(y'_{1:\hat{\tau}_j}(\phi); u)$ as a function of the perturbation, ϕ, and the mean at the $\hat{\tau}_j$th timepoint, u. A similar approach can be used to compute $\text{Cost}(y'_{T:(\hat{\tau}_j+1)}(\phi); u)$.

5.2. *Analytic computation of* $\text{Cost}(y'_{1:\hat{\tau}_j}(\phi); u)$

We first note that $\text{Cost}(y_{1:s}; u)$, the cost of segmenting $y_{1:s}$ with $\mu_s = u$, can be efficiently computed (Rigaill, 2015; Maidstone et al., 2017). The cost at the first timepoint is simply $\text{Cost}(y_1; u) = \frac{1}{2}(y_1 - u)^2$. For any $s > 1$ and for all u,

$$\text{Cost}(y_{1:s}; u) = \min \left\{ \text{Cost}(y_{1:(s-1)}; u), \min_{u'} \left\{ \text{Cost}(y_{1:(s-1)}; u') \right\} + \lambda \right\} + \frac{1}{2}(y_s - u)^2. \tag{26}$$

For each u, this recursion encapsulates two possibilities: (i) there is no changepoint at the $(s-1)$st timepoint, and the optimal cost is equal to the previous cost plus the cost of a new data point, $\text{Cost}(y_{1:(s-1)}; u) + \frac{1}{2}(y_s - u)^2$; (ii) there is a changepoint at the $(s-1)$st timepoint, and the optimal cost is equal to the optimal cost of segmenting up to $s-1$ plus the penalty for adding a changepoint at $s-1$ plus the cost of a new data point, $\min_{u'} \left\{ \text{Cost}(y_{1:(s-1)}; u') \right\} + \lambda + \frac{1}{2}(y_s - u)^2$. The resulting cost functions $\text{Cost}(y_1; u), \ldots, \text{Cost}(y_{1:T}; u)$ can be used to determine the exact solution to (4).

At first blush, the recursion in (26) appears to be intractable due to the fact that, naively, $\text{Cost}(y_{1:s}; u)$ needs to updated for each value of $u \in \mathbb{R}$. However, Rigaill (2015) and Maidstone et al. (2017) show that these updates can be performed by efficiently manipulating piecewise quadratic functions of u, without needing to explicitly consider individual values of u, using a procedure that they call *functional pruning*.

It turns out that many of the computations made in the recursion (26) can be reused in the calculation of $\text{Cost}(y'_{1:\hat{\tau}_j}(\phi); u)$. In particular, we note that from (11) and (6), $y'_s(\phi) = y_s$ for all $s \notin \{\hat{\tau}_{j-1}+1, \ldots, \hat{\tau}_{j+1}\}$, and therefore, $\text{Cost}(y'_{1:\hat{\tau}_{j-1}}(\phi); u) = \text{Cost}(y_{1:\hat{\tau}_{j-1}}; u)$. As a result, we only require a new algorithm to efficiently compute $\text{Cost}(y'_{1:(\hat{\tau}_{j-1}+1)}(\phi); u), \ldots, \text{Cost}(y'_{1:\hat{\tau}_j}(\phi); u)$. We now show that for $s = \hat{\tau}_{j-1}+1, \ldots, \hat{\tau}_j$, $\text{Cost}(y'_{1:s}(\phi); u)$ is the pointwise minimum over a set $\mathfrak{C}_s$ of piecewise quadratic functions of u and ϕ that can be efficiently computed.

THEOREM 2. *For $\hat{\tau}_{j-1} < s \leq \hat{\tau}_j$,*

$$\text{Cost}(y'_{1:s}(\phi); u) = \min_{f \in \mathfrak{C}_s} f(u, \phi), \tag{27}$$

where $\{f(u,\phi)\}_{f\in\mathfrak{C}_s}$ is a collection of $s - \hat{\tau}_{j-1} + 1$ piecewise quadratic functions of u and ϕ constructed recursively from $\hat{\tau}_{j-1}+1$ to s, and where $\mathfrak{C}_{\hat{\tau}_{j-1}} = \{\text{Cost}(y_{1:\hat{\tau}_{j-1}}; u)\}$. Furthermore, the set $\mathfrak{C}_{\hat{\tau}_j}$ can be computed in $\mathcal{O}((\hat{\tau}_j - \hat{\tau}_{j-1})^2)$ operations.

Section C.1 of the Supplementary Materials contains a proof of Theorem 2.

5.3. Computing $C'(\phi)$ based on $\text{Cost}(y'_{1:\hat{\tau}_j}(\phi); u)$ and $\text{Cost}(y'_{T:(\hat{\tau}_j+1)}(\phi); u)$

Recall from (25) that $C'(\phi)$ is the minimum of $C'_{\hat{\tau}_j}(\phi)$ and $C'_{\neg\hat{\tau}_j}(\phi)$, in (23) and (24), respectively. We now show how to compute $C'_{\hat{\tau}_j}(\phi)$.

We apply Theorem 2 to build the set $\mathfrak{C}_{\hat{\tau}_j}$, and recall from (27) that $\text{Cost}(y'_{1:\hat{\tau}_j}(\phi); u) = \min_{f\in\mathfrak{C}_{\hat{\tau}_j}} f(u,\phi)$. Additionally, we define $\tilde{\mathfrak{C}}_{\hat{\tau}_{j+1}+1} = \{\text{Cost}(y_{T:(\hat{\tau}_{j+1}+1)}; u)\}$, and build $\tilde{\mathfrak{C}}_{\hat{\tau}_{j+1}}, \ldots, \tilde{\mathfrak{C}}_{\hat{\tau}_j+1}$ such that $\text{Cost}(y'_{T:(\hat{\tau}_j+1)}(\phi); u) = \min_{f\in\tilde{\mathfrak{C}}_{\hat{\tau}_{j+1}}} f(u,\phi)$, using a modification of Theorem 2 that accounts for the reversal of the timepoints. Plugging into (23),

$$C'_{\hat{\tau}_j}(\phi) = \min_u \left\{ \min_{f\in\mathfrak{C}_{\hat{\tau}_j}} \{f(u,\phi)\} \right\} + \min_{u'} \left\{ \min_{f\in\tilde{\mathfrak{C}}_{\hat{\tau}_{j+1}}} \{f(u',\phi)\} \right\} + \lambda \tag{28}$$

$$= \min_{f\in\mathfrak{C}_{\hat{\tau}_j}} \left\{ \min_u \{f(u,\phi)\} \right\} + \min_{f\in\tilde{\mathfrak{C}}_{\hat{\tau}_{j+1}}} \left\{ \min_{u'} \{f(u',\phi)\} \right\} + \lambda. \tag{29}$$

Since $f(u,\phi)$ is piecewise quadratic in u and ϕ (Theorem 2), we see that $\min_u \{f(u,\phi)\}$ is piecewise quadratic in ϕ. Therefore, $\min_{f\in\mathfrak{C}_{\hat{\tau}_j}} \left\{ \min_u \{f(u,\phi)\} \right\}$ and $\min_{f\in\tilde{\mathfrak{C}}_{\hat{\tau}_{j+1}}} \left\{ \min_u \{f(u,\phi)\} \right\}$ can be efficiently performed using ideas from Rigaill (2015) and Maidstone et al. (2017), which allow for efficient manipulations of piecewise quadratic functions of a single variable. This means that $C'_{\hat{\tau}_j}(\phi)$ can be efficiently computed. Recall from Theorem 2 that

the set $\mathfrak{C}_{\hat{\tau}_j}$ contains $\hat{\tau}_j - \hat{\tau}_{j-1} + 1$ functions and can be computed in $\mathcal{O}((\hat{\tau}_j - \hat{\tau}_{j-1})^2)$ operations. Therefore, computing $C'_{\hat{\tau}_j}(\phi)$ requires $\mathcal{O}((\hat{\tau}_j - \hat{\tau}_{j-1})^2)$ operations to compute $\mathfrak{C}_{\hat{\tau}_j}$ and $\mathcal{O}((\hat{\tau}_{j+1} - \hat{\tau}_j)^2)$ operations to compute $\tilde{\mathfrak{C}}_{\hat{\tau}_j+1}$, followed by performing the operation $\min_u\{f(u,\phi)\}$ a total of $\mathcal{O}(\hat{\tau}_{j+1} - \hat{\tau}_{j-1})$ times. We can similarly obtain the piecewise quadratic function $C'_{-\hat{\tau}_j}(\phi)$ of ϕ. Therefore, we can analytically compute $C'(\phi)$.

Finally, recall from (22) that $\mathcal{S} = \{\phi : C(\phi) \leq C'(\phi)\}$. Since we have efficiently characterized both $C(\phi)$ and $C'(\phi)$, our characterization of $\mathcal{S}$ is complete.

6. Experiments

6.1. Simulation set-up and methods for comparison

We simulate $y_1, \ldots, y_{2000}$ according to (1) with $\sigma^2 = 1$. The mean vector $\mu \in \mathbb{R}^{2000}$ is piecewise constant with 50 changepoints. After each even-numbered changepoint the mean equals 0, and after each odd-numbered changepoint it equals δ, for $\delta \in \{0, 0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0\}$. The $K = 50$ changepoints are sampled without replacement from $\{1, 2, \ldots, 1999\}$. Panel a) of Figure 3 depicts a realization with $\delta = 3$.

We compare four tests of a change in mean at an estimated changepoint:

Approach 1. For the jth changepoint estimated by binary segmentation, test $H_0 : \nu^\top \mu = 0$ using ν in (6). Condition on the locations, orders, and signs of all of the estimated changepoints from binary segmentation. This is closely related to Hyun et al. (2021)'s proposal.

Approach 2. For the jth changepoint estimated by binary segmentation, test $H_0 : \nu^\top \mu = 0$ using ν in (6). Condition on the locations of all of the estimated changepoints from binary segmentation.

Approach 3. For the jth changepoint estimated by binary segmentation, test $H_0 : \nu^\top \mu = 0$ using ν in (13). Condition only on the location of the jth estimated changepoint from binary segmentation.

Approach 4. For the jth changepoint estimated by ℓ_0 segmentation, test $H_0 : \nu^\top \mu = 0$ using ν in (13). Condition only on the location of the jth estimated changepoint from ℓ_0 segmentation.

Unless stated otherwise, we take $h = 50$ in (13) for Approaches 3–4. As our aim is to compare the power of Approaches 1–4, we assume the true number of changepoints ($K = 50$) is known, so that both binary segmentation and ℓ_0 segmentation estimate the same number of changepoints†. We also assume that the underlying noise variance ($\sigma^2 = 1$) is known; see Section 8.3 for a more detailed discussion. All results are averaged over 100 replicate data sets with μ fixed.

In Section E of the Supplementary Materials, we present timing results for estimating changepoints as well as computing p-values using Approaches 1–4. Surprisingly, Approach 4 is even faster than Approaches 1–3: in our C++ implementation, the former takes only 15 seconds when $T = 1000$. Approaches 1–3 take longer because calculating $\mathcal{S}$ in the case of binary segmentation requires manipulating a large set of linear equations.

†On a given data set, there may not exist a λ such that ℓ_0 segmentation yields precisely 50 estimated changepoints. In this case, we select λ to give approximately 50 estimated changepoints.

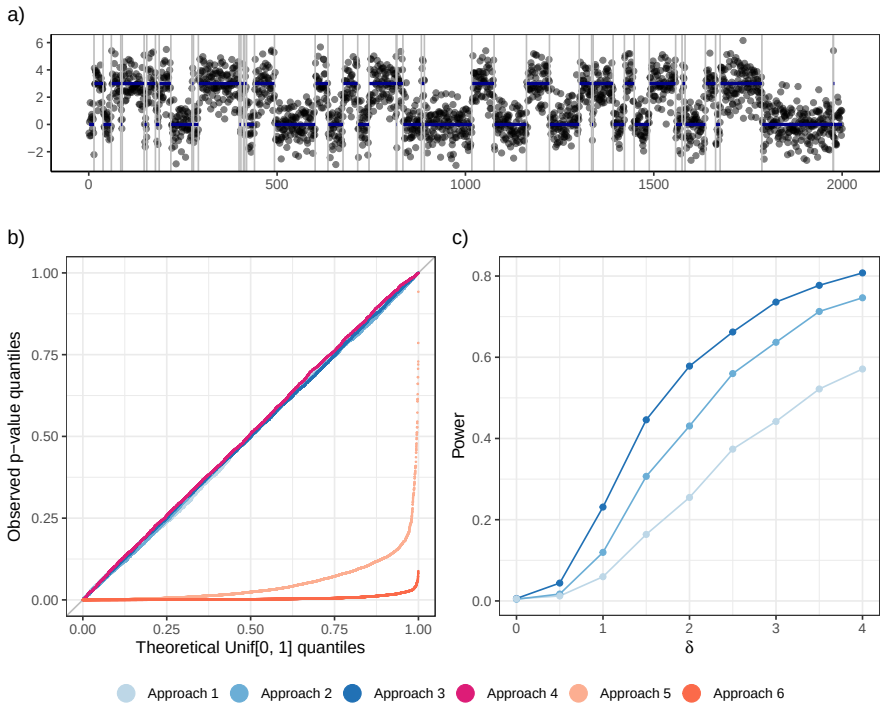

Fig. 3. a) The grey points represent a realization from the mean model (1), with true change in mean due to a changepoint $\delta = 3$. The mean $\mu_1, \ldots, \mu_T$ is shown as a blue line, and the changepoints are shown as grey vertical lines. b) Quantile-quantile plot comparing sample p-value quantiles under (1) with $\mu_1 = \ldots = \mu_{2000}$ versus theoretical quantiles of the $\mathrm{Unif}(0, 1)$ distribution, for Approaches 1–4 in Section 6.1, and Approaches 5–6 in Section 6.2. c) Empirical power, averaged over 100 replicates, is displayed for Approaches 1–3 defined in Section 6.1, each of which results from testing $H_0 : \nu^\top \mu = 0$ for changepoints estimated using binary segmentation with different conditioning sets. Various values of δ, the true change in mean due to a changepoint, are shown on the x-axis. Power increases with the size of the conditioning set.

6.2. Type I error control under a global null

We take $\delta = 0$, so that $\mu_1 = \ldots = \mu_{2000}$, and consider testing $H_0 : \nu^\top \mu = 0$ using Approaches 1–4, as well as the following two approaches that rely on a standard z-test:

> *Approach 5.* For the jth changepoint estimated by binary segmentation, test $H_0 : \nu^\top \mu = 0$ using ν in (6), without conditioning.
>
> *Approach 6.* For the jth changepoint estimated by ℓ_0 segmentation, test $H_0 : \nu^\top \mu = 0$ using ν in (6), without conditioning.

These two approaches do not account for the fact that the changepoints were estimated from the data. Panel b) of Figure 3 displays quantile-quantile plots of the observed p-value quantiles versus theoretical $\mathrm{Unif}[0, 1]$ quantiles. The plots indicate that Approaches 1–4 control the Type 1 error, whereas Approaches 5–6 do not.

6.3. Increases in power due to conditioning on less information

Next, we illustrate that the power increases as the size of the conditioning set increases, by considering Approaches 1–3 from Section 6.1. Each approach uses binary segmentation, though with different conditioning sets.

On a given dataset, we define the empirical power as the fraction of true changepoints for which the nearest estimated changepoint has a p-value below α and is within $\pm m$ timepoints,

$$\widehat{\text{Power}} := \frac{\sum_{i=1}^{K} 1_{(|\tau_i - \hat{\tau}_{j(i)}| \leq m \text{ and } p_{j(i)} \leq \alpha)}}{K}. \tag{30}$$

Here, $j(i) = \text{argmin}_{1 \leq l \leq K} |\tau_i - \hat{\tau}_l|$. Panel c) of Figure 3 shows the empirical power for the three approaches with $\alpha = 0.05$ and $m = 2$. As the size of the conditioning set increases, from $\{\phi : \mathcal{M}(y'(\phi)) = \mathcal{M}(y), \mathcal{O}(y'(\phi)) = \mathcal{O}(y), \Delta(y'(\phi)) = \Delta(y)\}$ to $\{\phi : \mathcal{M}(y'(\phi)) = \mathcal{M}(y)\}$ to $\{\phi : \hat{\tau}_j \in \mathcal{M}(y'(\phi))\}$, the power increases substantially.

6.4. Power and detection probability

We now compare the performances of Approaches 1–4, defined in Section 6.1, as well as two additional approaches that are based on *sample splitting* (Cox, 1975):

> *Approach 7.* Apply binary segmentation to the odd timepoints. For the jth estimated changepoint, test $H_0 : \nu^\top \mu = 0$ on the even timepoints, with ν in (6), without conditioning.
> *Approach 8.* Apply ℓ_0 segmentation to the odd timepoints. For the jth estimated changepoint, test $H_0 : \nu^\top \mu = 0$ on the even timepoints, with ν in (6), without conditioning.

Because we estimate and test the changepoints on two separate halves of the data, we can apply a standard z-test in Approaches 7 and 8 (Fithian et al., 2014).

In addition to calculating the empirical power (30) for each approach, we also consider each approach's ability to detect the true changepoints. This is defined as the fraction of true changepoints for which there is an estimated changepoint within $\pm m$ timepoints,

$$\widehat{\text{Detection probability}} := \frac{\sum_{i=1}^{K} 1_{(\min_{1 \leq l \leq K} |\tau_i - \hat{\tau}_l| \leq m)}}{K}. \tag{31}$$

Panels b) and c) of Figure 4 display the power and detection probability for Approaches 1–4 and 7–8, with $\alpha = 0.05$ and $m = 2$. Approach 4 (which makes use of ℓ_0 segmentation, and conditions only on the jth estimated changepoint) performs the best, in terms of both power and detection probability, especially as δ increases. Figure 4 also illustrates the benefit of the inferential framework developed in this paper over naive sample-splitting approaches. Sample splitting has limited ability to detect changepoints, since only half of the data is used to estimate changepoints.

6.5. Assessment of different window sizes for testing $H_0 : \nu^\top \mu = 0$ for ν in (13)

Figure 4 suggests that Approaches 3 and 4 from Section 6.1 have high power. However, they require pre-specifying the window size h in (13). We now address this possible

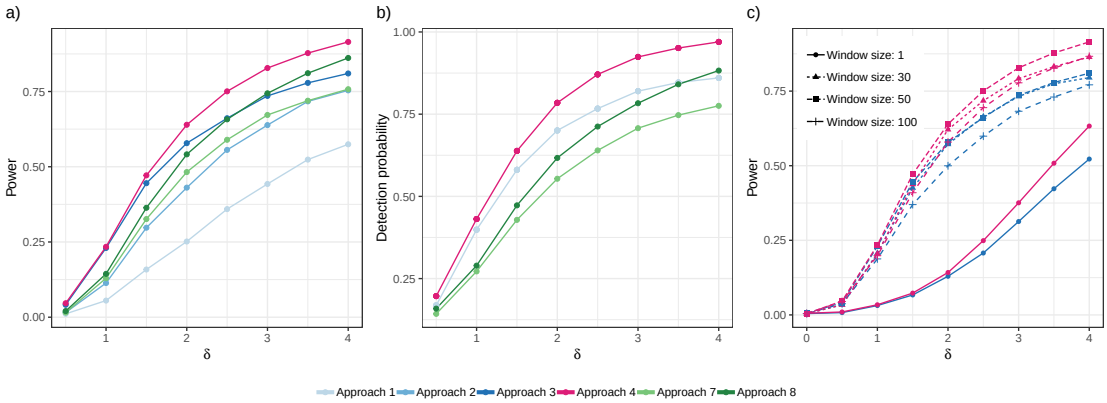

Fig. 4. Empirical power and detection probability for different changepoint estimation and inference procedures. a) Power for Approaches 1–4, which are described in Section 6.1, as well as Approaches 7–8, which are described in Section 6.4. b) Detection probability for binary segmentation and ℓ_0 segmentation using all of the data, as well as half of the data. In this panel, the curve shown for Approach 1 applies to Approaches 1-3 since Approaches 1-3 use binary segmentation. c) Power of Approaches 3 and 4 for testing $H_0 : \nu^\top \mu = 0$ for ν in (13), for three values of the window size h.

weakness. In Figure 4c), we assess the performance of Approaches 3 and 4 with $h \in \{1, 30, 50, 100\}$. In this particular setting, we see that our test has high power when h takes on a value of around 50. However, in general, the optimal value of h will depend on the true positions of changepoints. In Section 8.1 we discuss an extension that does not rely on a fixed window size h.

7. Real data example

We now consider guanine-cytosine (G-C) content on a 2Mb window of human chromosome one, binned so that $T = 2000$. Data was originally accessed from the National Center for Biotechnology Information, and is available via the `R` package `changepoint` (Killick and Eckley, 2014). We used a consistent estimator of σ described in Section 8.3 to scale the data and calculate p-values.

We estimate changepoints using k-step binary segmentation, where $k = 38$ is chosen based on the modified Bayesian information criterion (Zhang and Siegmund, 2007) implemented in the `changepoint` package. To facilitate comparisons, we then fit ℓ_0 segmentation using a value of λ that yields 38 changepoints. Figure 5 displays the estimated changepoints from these two methods, along with an indication of whether Approaches 1–4 from Section 6.1 resulted in a p-value below 0.05. The number of discoveries (estimated changepoints whose p-value is less than 0.05) is substantially greater using Approaches 2–4 than using Approach 1, which conditions on far more information. Approach 1 results in 15 discoveries, versus 26, 25, and 27 in Approaches 2, 3, and 4, respectively. These p-values can be adjusted for multiple testing using ideas from e.g. Benjamini and Hochberg (1995), Storey (2002), and Dudoit and Van Der Laan (2007).

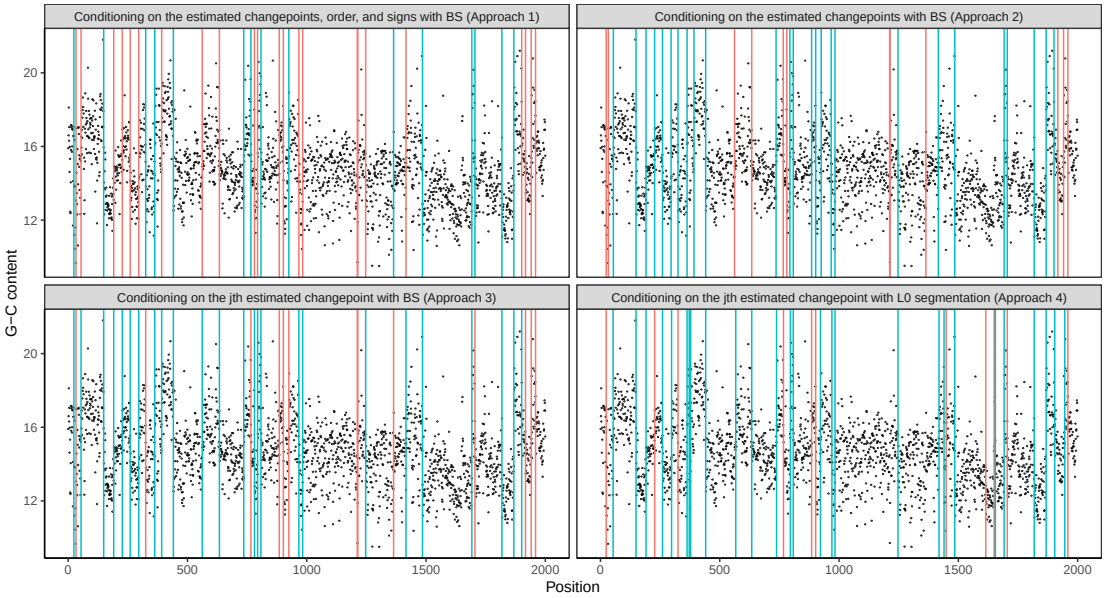

Fig. 5. The number of discoveries depends on the size of the conditioning set. Each panel displays scaled G-C content on a 2Mb window of human chromosome one. The G-C content is binned leading to $T = 2000$ (displayed in black). Estimated changepoints from Approaches 1–4 from Section 6.1 (organized by panel) for which the p-value is less than 0.05 are displayed in blue; the remaining estimated changepoints are displayed in red.

8. Discussion

8.1. Larger conditioning sets for testing (9) *with* ν *in* (6)

No special properties of the conditioning set were used to prove Theorem 1. Thus, instead of conditioning on the full set of changepoints as in Section 3.1, we could have instead conditioned on the jth estimated changepoint and its immediate neighbors. This would yield the p-value $p = \Pr\left(|\phi| \geq |\nu^\top y| \mid \{\hat{\tau}_{j-1}, \hat{\tau}_j, \hat{\tau}_{j+1}\} \subseteq \mathcal{M}(y'(\phi))\right)$. Characterizing the set $\mathcal{S} = \{\phi : \{\hat{\tau}_{j-1}, \hat{\tau}_j, \hat{\tau}_{j+1}\} \subseteq \mathcal{M}(y'(\phi))\}$ would require only minor modifications to the algorithms in Sections 4 and 5 and the Supplementary Materials.

8.2. Extensions to related problems

The ideas in this paper apply beyond the change-in-mean model (1). For instance, they can be applied to the analysis of data from calcium imaging, a recent technology for recording neuronal activity *in vivo* (Dombeck et al., 2007). A number of authors (Vogelstein et al., 2010; Friedrich et al., 2017) have assumed that the observed fluorescence trace for a neuron, y_t, is a noisy version of the underlying calcium concentration, c_t, which decays exponentially with a rate $\gamma < 1$, except when there is an instantaneous increase in the calcium because the neuron has spiked, $s_t > 0$:

$$Y_t = c_t + \epsilon_t, \quad \epsilon_t \overset{\text{iid}}{\sim} N(0, \sigma^2), \quad c_t = \gamma c_{t-1} + s_t.$$

In this model, scientific interest lies in determining the precise timepoints of the spikes, i.e. the set $\{t : s_t > 0\}$. Jewell and Witten (2018) and Jewell et al. (2020) estimate this quantity by solving a variant of the ℓ_0 segmentation problem (4) in Section 2.1.2. The framework from Section 3, and the algorithms from Section 5, can be used to test the null hypothesis that there is no increase in the calcium concentration around a spike, $H_0 : \nu^\top c = 0$, for a suitably chosen contrast ν. Details are in Chen et al. (2021).

It is natural to wonder whether these ideas can be extended to the change-in-slope proposals of Fearnhead et al. (2019) and Baranowski et al. (2019). Extending the ideas in Section 5 to the former is quite challenging, since the continuity constraint in the optimization problem induces dependence across segments that complicate the development of computationally-feasible recursions. By contrast, the latter is closely related to binary segmentation, and so an extension of the approach in Section 4 can be applied.

8.3. Additional extensions

Relaxing assumptions in (1) The model (1) assumes that the error terms are Gaussian, independent, and identically distributed. These assumptions are critical to the proof of Theorem 1, as they guarantee that $\nu^\top Y$ and $\Pi_\nu^\perp Y$ are independent. However, recent work in selective inference has focused on relaxing these assumptions (Tian et al., 2018; Tibshirani et al., 2018; Taylor and Tibshirani, 2018), and may be applicable here.

Estimation of the error variance in (1) Throughout this paper, we have assumed that the error variance in (1) is known. However, if it is unknown, then we can plug in any consistent estimator of σ in evaluating the p-values in (9) and (14). Then, under $H_0 : \nu^\top \mu = 0$, the resulting p-values will converge in distribution to a Unif$[0, 1]$ distribution, i.e. they will have asymptotic Type 1 error control. In Section F of the Supplementary Materials, we present the results of a simulation study using a simple consistent estimator of σ obtained by taking the median absolute deviation of the first differences of $y_1, \ldots, y_T$ and dividing by $\sqrt{2}\Phi^{-1}(3/4)$. We see that this approach leads to adequate Type 1 error control, as well as substantial power under the alternative.

Confidence intervals The conditional distribution of $\nu^\top Y$ can be used to develop a confidence interval for $\nu^\top \mu$ that has correct selective coverage; see, e.g., Lee et al. (2016).

Acknowledgments

Sean Jewell received funding from the Natural Sciences and Engineering Research Council of Canada. This work was partially supported by Engineering and Physical Sciences Research Council Grant EP/N031938/1 to Paul Fearnhead, and NSF CAREER DMS-1252624, NIH grants DP5OD009145, R01DA047869, and R01EB026908, and a Simons Investigator Award in Mathematical Modeling of Living Systems to Daniela Witten.

We thank Zaid Harchaoui and Ali Shojaie for helpful conversations, and Jacob Bien and three anonymous reviewers for suggestions that improved the quality of this paper.

References

Anastasiou, A. and Fryzlewicz, P. (2019). Detecting multiple generalized change-points by isolating single ones. *arXiv preprint arXiv:1901.10852*.

Auger, I. E. and Lawrence, C. E. (1989). Algorithms for the optimal identification of segment neighborhoods. *Bulletin of Mathematical Biology*, 51(1):39–54.

Badagián, A. L., Kaiser, R., and Peña, D. (2015). Time series segmentation procedures to detect, locate and estimate change-points. In *Empirical Economic and Financial Research*, pages 45–59. Springer.

Bai, J. (1994). Least squares estimation of a shift in linear processes. *Journal of Time Series Analysis*, 15(5):453–472.

Bai, J. and Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica*, 66(1):47–78.

Bai, J. and Perron, P. (2003). Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, 18(1):1–22.

Baranowski, R., Chen, Y., and Fryzlewicz, P. (2019). Narrowest-over-threshold detection of multiple change points and change-point-like features. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 81(3):649–672.

Barber, R. F. and Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085.

Bauer, P. and Hackl, P. (1980). An extension of the MOSUM technique for quality control. *Technometrics*, 22(1):1–7.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodological)*, 57(1):289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188.

Candes, E., Fan, Y., Janson, L., and Lv, J. (2018). Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 80(3):551–577.

Chen, Y. T., Jewell, S. W., and Witten, D. M. (2021). Quantifying uncertainty in spikes estimated from calcium imaging data. *Biostatistics*.

Chu, C.-S. J., Hornik, K., and Kaun, C.-M. (1995). MOSUM tests for parameter constancy. *Biometrika*, 82(3):603–617.

Cox, D. R. (1975). A note on data-splitting for the evaluation of significance levels. *Biometrika*, 62(2):441–444.

Dombeck, D. A., Khabbaz, A. N., Collman, F., Adelman, T. L., and Tank, D. W. (2007). Imaging large-scale neural activity with cellular resolution in awake, mobile mice. *Neuron*, 56(1):43–57.

Dudoit, S. and Van Der Laan, M. J. (2007). *Multiple testing procedures with applications to genomics*. Springer Science & Business Media.

Duy, V. N. L., Toda, H., Sugiyama, R., and Takeuchi, I. (2020). Computing valid p-value for optimal changepoint by selective inference using dynamic programming. *Advances in Neural Information Processing Systems*, 33:11356–11367.

Eichinger, B. and Kirch, C. (2018). A MOSUM procedure for the estimation of multiple random change points. *Bernoulli*, 24(1):526–564.

Enikeeva, F. and Harchaoui, Z. (2019). High-dimensional change-point detection under sparse alternatives. *The Annals of Statistics*, 47(4):2051–2079.

Fearnhead, P. (2006). Exact and efficient Bayesian inference for multiple changepoint problems. *Statistics and Computing*, 16(2):203–213.

Fearnhead, P., Maidstone, R., and Letchford, A. (2019). Detecting changes in slope with an L_0 penalty. *Journal of Computational and Graphical Statistics*, 28(2):265–275.

Fearnhead, P. and Rigaill, G. (2020). Relating and comparing methods for detecting changes in mean. *Stat*, 9(1):e291.

Fithian, W., Sun, D., and Taylor, J. (2014). Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*.

Fithian, W., Taylor, J., Tibshirani, R., and Tibshirani, R. J. (2015). Selective sequential model selection. *arXiv preprint arXiv:1512.02565*.

Frick, K., Munk, A., and Sieling, H. (2014). Multiscale change point inference. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(3):495–580.

Friedrich, J., Zhou, P., and Paninski, L. (2017). Fast online deconvolution of calcium imaging data. *PLoS Computational Biology*, 13(3):e1005423.

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

Futschik, A., Hotz, T., Munk, A., and Sieling, H. (2014). Multiscale DNA partitioning: statistical evidence for segments. *Bioinformatics*, 30(16):2255–2262.

Harchaoui, Z. and Lévy-Leduc, C. (2007). Catching change-points with lasso. In *Advances in Neural Information Processing Systems*, volume 617, page 624.

Haynes, K., Fearnhead, P., and Eckley, I. A. (2017). A computationally efficient nonparametric approach for changepoint detection. *Statistics and Computing*, 27(5):1293–1305.

Hinkley, D. V. (1970). Inference about the change-point in a sequence of random variables. *Biometrika*, 57(1):1–17.

Hocking, T. D., Rigaill, G., Fearnhead, P., and Bourque, G. (2020). Constrained dynamic programming and supervised penalty learning algorithms for peak detection in genomic data. *Journal of Machine Learning Research*, 21(87):1–40.

Hotz, T., Schütte, O. M., Sieling, H., Polupanow, T., Diederichsen, U., Steinem, C., and Munk, A. (2013). Idealizing ion channel recordings by a jump segmentation multiresolution filter. *IEEE transactions on NanoBioscience*, 12(4):376–386.

Hušková, M. (1990). Asymptotics for robust MOSUM. *Commentationes Mathematicae Universitatis Carolinae*, 31(2):345–356.

Hyun, S., G'Sell, M., and Tibshirani, R. J. (2018). Exact post-selection inference for the generalized lasso path. *Electronic Journal of Statistics*, 12(1):1053–1097.

Hyun, S., Lin, K. Z., G'Sell, M., and Tibshirani, R. J. (2021). Post-selection inference for changepoint detection algorithms with application to copy number variation data. *Biometrics*, 77(3):1037–1049.

Jackson, B., Scargle, J. D., Barnes, D., Arabhi, S., Alt, A., Gioumousis, P., Gwin, E., Sangtrakulcharoen, P., Tan, L., and Tsai, T. T. (2005). An algorithm for optimal partitioning of data on an interval. *IEEE Signal Processing Letters*, 12(2):105–108.

James, B., James, K. L., and Siegmund, D. (1987). Tests for a change-point. *Biometrika*, 74(1):71–83.

Jewell, S., Fearnhead, P., and Witten, D. (2019). Testing for a change in mean after changepoint detection. *arXiv preprint arXiv:1910.04291*.

Jewell, S. and Witten, D. (2018). Exact spike train inference via ℓ_0 optimization. *The Annals of Applied Statistics*, 12(4):2457–2482.

Jewell, S. W., Hocking, T. D., Fearnhead, P., and Witten, D. M. (2020). Fast nonconvex deconvolution of calcium imaging data. *Biostatistics*, 21(4):709–726.

Killick, R. and Eckley, I. (2014). changepoint: An R package for changepoint analysis. *Journal of Statistical Software*, 58(3):1–19.

Killick, R., Fearnhead, P., and Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500):1590–1598.

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927.

Li, H., Munk, A., and Sieling, H. (2016). FDR-control in multiscale change-point segmentation. *Electronic Journal of Statistics*, 10(1):918–959.

Liu, K., Markovic, J., and Tibshirani, R. (2018). More powerful post-selection inference, with application to the lasso. *arXiv preprint arXiv:1801.09037*.

Ma, T. F. and Yau, C. Y. (2016). A pairwise likelihood-based approach for changepoint detection in multivariate time series models. *Biometrika*, 103(2):409–421.

Maidstone, R., Hocking, T., Rigaill, G., and Fearnhead, P. (2017). On optimal multiple changepoint algorithms for large data. *Statistics and Computing*, 27(2):519–533.

Muggeo, V. M. and Adelfio, G. (2010). Efficient changepoint detection for genomic sequences of continuous measurements. *Bioinformatics*, 27(2):161–166.

Nam, C. F., Aston, J. A., and Johansen, A. M. (2012). Quantifying the uncertainty in change points. *Journal of Time Series Analysis*, 33(5):807–823.

Niu, Y. S. and Zhang, H. (2012). The screening and ranking algorithm to detect DNA copy number variations. *The Annals of Applied Statistics*, 6(3):1306.

Olshen, A. B., Venkatraman, E., Lucito, R., and Wigler, M. (2004). Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics*, 5(4):557–572.

Rigaill, G. (2015). 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(4):180–205.

Ruanaidh, J. J. O. and Fitzgerald, W. J. (2012). *Numerical Bayesian methods applied to signal processing*. Springer Science & Business Media.

Rudin, L. I., Osher, S., and Fatemi, E. (1992). Nonlinear total variation based noise removal algorithms. *Physica D: Nonlinear Phenomena*, 60(1-4):259–268.

Schröder, A. L. and Fryzlewicz, P. (2013). Adaptive trend estimation in financial time series via multiscale change-point-induced basis recovery. *Statistics and Its Interface*, 4(6):449–461.

Storey, J. D. (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 64(3):479–498.

Taylor, J. and Tibshirani, R. (2018). Post-selection inference for ℓ_1-penalized likelihood models. *Canadian Journal of Statistics*, 46(1):41–61.

Tian, X., Loftus, J. R., and Taylor, J. E. (2018). Selective inference with unknown variance via the square-root lasso. *Biometrika*, 105(4):755–768.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108.

Tibshirani, R. J., Rinaldo, A., Tibshirani, R., and Wasserman, L. (2018). Uniform asymptotic inference and the bootstrap after model selection. *Annals of Statistics*, 46(3):1255–1287.

Tibshirani, R. J., Taylor, J., Lockhart, R., and Tibshirani, R. (2016). Exact post-selection inference for sequential regression procedures. *Journal of the American Statistical Association*, 111(514):600–620.

Truong, C., Oudre, L., and Vayatis, N. (2020). Selective review of offline change point detection methods. *Signal Processing*, 167:107299.

Vogelstein, J. T., Packer, A. M., Machado, T. A., Sippy, T., Babadi, B., Yuste, R., and Paninski, L. (2010). Fast nonnegative deconvolution for spike train inference from population calcium imaging. *Journal of Neurophysiology*, 104(6):3691–3704.

Vostrikova, L. (1981). Detection of the disorder in multidimensional random-processes. *Doklady Akademii Nauk SSSR*, 259(2):270–274.

Xiao, F., Luo, X., Hao, N., Niu, Y. S., Xiao, X., Cai, G., Amos, C. I., and Zhang, H. (2019). An accurate and powerful method for copy number variation detection. *Bioinformatics*, 35(17):2891–2898.

Yao, Y.-C. (1987). Approximating the distribution of the maximum likelihood estimate of the change-point in a sequence of independent random variables. *The Annals of Statistics*, 15(3):1321–1328.

Zhang, N. R. and Siegmund, D. O. (2007). A modified Bayes information criterion with applications to the analysis of comparative genomic hybridization data. *Biometrics*, 63(1):22–32.