Testing for a Change in Mean After Changepoint Detection

Sean Jewell *
Department of Statistics, University of Washington
Paul Fearnhead
Department of Mathematics and Statistics, Lancaster University
Daniela Witten
Departments of Statistics and Biostatistics, University of Washington

Abstract

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, variants of binary segmentation, $\ell_0$ segmentation, or the fused lasso. Our setup allows us to condition on much smaller selection events than existing approaches, which yields higher powered tests. Our procedure leads to improved power in simulation and additional discoveries in a dataset of chromosomal guanine-cytosine content. Our new changepoint inference procedures are freely available in the R package ChangepointInference at https://jewellsean.github.io/changepoint-inference/.

Keywords: $\ell_0$ optimization, binary segmentation, fused lasso, selective inference

*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.
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 certain (unknown) timepoints, 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 $\tau$ — but also to conduct statistical inference on the estimated changepoints.

In this paper, we consider the most common changepoint model

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

and assume that $\mu_1, \ldots, \mu_T$ is piecewise constant, in the sense that $\mu_{\tau_j+1} = \mu_{\tau_j+2} = \ldots = \mu_{\tau_j+1}$, $\mu_{\tau_j+1} \neq \mu_{\tau_j+1+1}$, for $j = 0, \ldots, K - 1$, where $0 = \tau_0 < \tau_1 < \ldots < \tau_K < \tau_{K+1} = T$, and where $\tau_1, \ldots, \tau_K$ 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, and can be roughly divided into two classes. One class of proposals iteratively searches for one changepoint at a time (Vostrikova, 1981; Olshen et al., 2004; Fryzlewicz et al., 2014; Badagián et al., 2015; Anastasiou and Fryzlewicz, 2019); the canonical example of this approach is binary segmentation. Another class of proposals involves simultaneously estimating 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., 2018; Jewell et al., 2019); examples include $\ell_0$ segmentation and the fused lasso. We review these approaches in Section 2.

Despite the huge literature on estimating changepoint locations, the literature for performing inference on changepoints is less developed. One recent exception is the multiscale approach of Frick et al. (2014) which both 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 in situations where there are many changes, particularly when these changes 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”;

2
in extreme cases, this could include an estimated changepoint that is almost as far as $T/2$ observations from an actual changepoint.

In this paper, we consider testing the null hypothesis that there is no change in mean around an estimated changepoint. This is challenging since we must account for the estimation process when deriving the null distribution for a test statistic. A recent promising line of work was introduced by Hyun et al. (2016) and Hyun et al. (2018), who develop valid tests that appropriately control for the estimation process in the cases where changepoints are estimated with the generalized lasso and with binary segmentation, respectively. They leverage 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).

One disadvantage of the proposals of Hyun et al. (2016) and Hyun et al. (2018) is that, when defining $p$-values, they need to condition on much more information than is used to choose the null hypothesis that is tested. This is especially relevant since Fithian et al. (2014), Lee et al. (2016), and Liu et al. (2018) show that conditioning on extra information leads to a reduction in power. Our contribution is to implement selective inference while conditioning on less information, and moreover to illustrate that selective inference techniques can be used when changepoints are detected via $\ell_0$ segmentation. Our empirical results show that both 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 et al., 2014), and $\ell_0$ 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 $\ell_0$ segmentation. In Section 6, we examine the performance of our proposal, and compare it 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.

Our new changepoint inference procedures are freely available in the R package ChangepointInference. Code and data to produce all figures is available at https://jewellsean.github.io/changepoint-inference.
2 Background

2.1 Changepoint detection algorithms

2.1.1 Binary segmentation and its variants

The binary segmentation proposal of Vostrikova (1981) and its variants (Olshen et al., 2004; Fryzlewicz et al., 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

\[ g(s, \tau, e)^\top y := \sqrt{1 \over |e-\tau| + |\tau+1-s|} \left( \bar{y}(\tau+1) - \bar{y}_{s:e} \right), \]  

(2)

defined through a contrast \( g(s, \tau, e) \in \mathbb{R}^T \), which summarizes the evidence for a change at \( \tau \) in the data \( y_{s:e} := (y_s, \ldots, y_e) \) by the difference in the empirical mean of the data before and after \( \tau \) (normalized to have the same variance for all \( \tau \)).

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, \( \hat{\tau}_1 = \arg\max_{\tau \in [1:(T-1)]} \{|g(1, \tau, T)^\top y|\} \). 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 \( \hat{\tau}_2 = \arg\max_{\tau \in [1:(T-1)]\setminus \hat{\tau}_1} \left\{ |g(1, \tau, \hat{\tau}_1)^\top y|1_{1 \leq \tau < \hat{\tau}_1} + |g(\hat{\tau}_1, \tau, T)^\top y|1_{\hat{\tau}_1 < \tau < T} \right\} \).

We continue in this manner until a stopping criterion is met. Variants of this procedure have been proposed to improve performance (Olshen et al., 2004; Fryzlewicz et al., 2014).

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.

\[
\min_{0 = t_0 < \ldots < \tau_K < T_{K+1} = T} \left\{ \frac{1}{2} \sum_{k=0}^{K} \sum_{t = \hat{\tau}_k+1}^{\hat{\tau}_{k+1}} (y_t - u_k)^2 + \lambda K \right\}. 
\]  

(3)

This is equivalent to solving an \( \ell_0 \) penalized regression problem

\[
\min_{\mu \in \mathbb{R}^T} \left\{ \frac{1}{2} \sum_{t=1}^{T} (y_t - \mu_t)^2 + \lambda \sum_{t=2}^{T} 1_{(\mu_t \neq \mu_{t-1})} \right\}, 
\]  

(4)

4
in the sense that the vector \( \hat{\mu} \) that solves (4) is piecewise constant with breakpoints at \( \hat{\tau}_1, \ldots, \hat{\tau}_K \), where \( \hat{\tau}_1, \ldots, \hat{\tau}_K \) are the changepoints that solve (3). Here, \( \lambda \) is a tuning parameter that specifies the improvement in fit to the data needed to add an additional changepoint.

Replacing the \( \ell_0 \) penalty in (4) with an \( \ell_1 \) penalty leads to the well-studied trend filtering or fused lasso optimization problem (Rudin et al., 1992; Tibshirani et al., 2005),

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

(5)

### 2.2 Existing methods for inference on changepoints

Suppose we wish to quantify the evidence for a set of estimated changepoints \( \hat{\tau}_1, \ldots, \hat{\tau}_K \). We could naively apply a standard z-test for the difference in mean around each estimated changepoint. However, this procedure is fundamentally flawed since the same data is used to estimate the changepoints and thus to choose the hypothesis tests that we perform. Therefore, 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. (2016) and Hyun et al. (2018) 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. (2018) propose to test whether there is a change in mean around \( \hat{\tau}_j \). They construct the test statistic \( \hat{d}_j \nu^\top Y \), where \( \hat{d}_j \) is the sign of the estimated change in mean at \( \hat{\tau}_j \), and \( \nu \) is a \( T \)-vector of contrasts, defined as

\[
\nu_t = \begin{cases} 
0 & \text{if } t \leq \hat{\tau}_{j-1} \text{ or } t > \hat{\tau}_{j+1}, \\
\frac{1}{\hat{\tau}_j - \hat{\tau}_{j-1}} & \text{if } \hat{\tau}_{j-1} < t \leq \hat{\tau}_j, \\
-\frac{1}{\hat{\tau}_{j+1} - \hat{\tau}_j} & \text{if } \hat{\tau}_j < t \leq \hat{\tau}_{j+1}, 
\end{cases}
\]  

(6)

and consider the null hypothesis \( H_0 : \hat{d}_j \nu^\top \mu = 0 \) versus the one-sided alternative \( H_1 : \hat{d}_j \nu^\top \mu > 0 \). Since both \( \hat{d}_j \) and \( \nu \) are functions of the estimated changepoints themselves, 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., \( \mathcal{M}(y) = \{ \hat{\tau}_1, \ldots, \hat{\tau}_K \} \). Then, it is tempting to define the p-value as

\[
\Pr_{H_0} \left( \hat{d}_j \nu^\top Y \geq \hat{d}_j \nu^\top y \ | \ \mathcal{M}(Y) = \mathcal{M}(y) \right).
\]  

(7)

However, (7) 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 \( \{ y : Ay \leq b \} \) for a matrix \( A \) and vector \( b \). Thus, in the case of binary segmentation, Hyun et al. (2018) condition on three additional quantities: (i) the order in which the estimated changepoints enter the model, \( O(Y) = O(y) \); (ii) the sign of the change in mean due to the estimated changepoints, \( \Delta(Y) = \Delta(y) = \{ \hat{d}_1, \ldots, \hat{d}_K \} \); (iii) \( \Pi_\nu^Y = \Pi_\nu^y \), where \( \Pi_\nu^1 = I - \nu \nu^T / \| \nu \|^2 \) is the orthogonal projection matrix onto the subspace that is orthogonal to \( \nu \). Conditions (i) and (ii) ensure that the conditioning set is polyhedral, whereas condition (iii) ensures that the test statistic is a pivot. This leads to the \( p \)-value

\[
\Pr_{H_0} \left( \hat{d}_j \nu^T Y \geq \hat{d}_j \nu^T y \mid M(Y) = M(y), O(Y) = O(y), \Delta(Y) = \Delta(y), \Pi_\nu^1 Y = \Pi_\nu^1 y \right), \tag{8}
\]

which can be easily computed because the conditional distribution of \( \hat{d}_j \nu^T Y \) is a Gaussian truncated to an interval, which is computationally tractable. For slightly different conditioning sets, Hyun et al. (2018) show similar results for variants of binary segmentation and for the fused lasso.

Importantly, Hyun et al. (2018) choose the conditioning set in (8) for computational reasons: there is no clear statistical motivation for conditioning on \( O(Y) = O(y) \) 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 \( M(Y) = M(y) \). In fact, Fithian et al. (2014) argue that when conducting selective inference, it is better to condition on a smaller set, 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 reduce 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. (2018). This means that we do not need to use the full conditioning set in (8), and, in turn, leads to higher-powered tests. Additionally, since we do not need to condition on \( \Delta(Y) = \Delta(y) \), we are able to consider two-sided tests of

\[
H_0 : \nu^T \mu = 0 \text{ versus } H_1 : \nu^T \mu \neq 0, \tag{9}
\]

rather than the one-sided tests considered by Hyun et al. (2018).

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 et al., 2001; Barber et al., 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. By suitable choice of \( \nu \) in (9), we can test whether there is a change in mean near \( \hat{\tau}_j \), where near can be defined appropriately for a given application. It is possible that application of, for example, knockoffs (Barber et al., 2015; Candes et al., 2018) would enable us to control the false discovery rate for the null hypotheses that the changes in mean are precisely at \( \hat{\tau}_1, \ldots, \hat{\tau}_K \). However, our experience with such methods is that they have almost no power to detect small to moderate 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 (9) 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 \( \Pi_{\nu}^+ Y = \Pi_{\nu}^+ y \) (which is essentially the test proposed by Hyun et al. (2018)). In panel b) we condition on just \( \mathcal{M}(Y) = \mathcal{M}(y) \) and \( \Pi_{\nu}^+ Y = \Pi_{\nu}^+ y \). 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 a smaller set 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 smaller sets than previously possible.

3 Two new tests with smaller conditioning sets

In this section, we consider testing a null hypothesis of the form (9) using a much smaller conditioning set than used by Hyun et al. (2018). 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 \( \nu \) 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 (9) for \( \nu \) 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, \( \mathcal{M}(y) = \{\hat{\tau}_1, \ldots, \hat{\tau}_K\} \). Thus, we define the \( p \)-value

\[
p \equiv \Pr_{H_0}(|\nu^\top Y| \geq |\nu^\top y| \mid \mathcal{M}(Y) = \mathcal{M}(y), \Pi_\nu Y = \Pi_\nu y).
\]  

(10)

As in Hyun et al. (2018), we condition on \( \Pi_\nu Y = \Pi_\nu y \) for technical reasons; see Fithian et al. (2014) for additional discussion. Roughly speaking, (10) asks: “Out of all data sets yielding this particular set of changepoints, what is the probability, under the null that there is no changepoint at this location, that the difference in mean between the segments on either side of \( \hat{\tau}_j \) is as large as what is observed?”

Our next result reveals that computing (10) involves a univariate truncated normal distribution.

**Theorem 1** The \( p \)-value in (10) is equal to

\[
p = \Pr(|\phi| \geq |\nu^\top y| \mid \mathcal{M}(y'(\phi)) = \mathcal{M}(y)),
\]

(11)

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

\[
y'(\phi) = y - \frac{\nu \nu^\top y}{\|\nu\|_2^2} + \frac{\nu \phi}{\|\nu\|_2^2}.
\]

(12)

In light of Theorem 1, to evaluate (10) we must simply characterize the set

\[
S = \{\phi : \mathcal{M}(y'(\phi)) = \mathcal{M}(y)\};
\]

(13)

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 \( \ell_0 \) segmentation, respectively. We discuss the fused lasso in Section D 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 that there is no change in mean in a window \( h > 0 \) around the \( j \)th estimated changepoint,

\[
H_0 : \mu_{\hat{\tau}_j-h+1} = \ldots = \mu_{\hat{\tau}_j} = \ldots = \mu_{\hat{\tau}_j+h}.
\]

(14)

This is a special case of (9) for \( \nu \) defined as

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

(15)
When considering this null hypothesis, it makes sense to condition only on the $j$th estimated changepoint, leading to a $p$-value defined as

$$ p \equiv \Pr_{H_0}(|\nu^T Y| \geq |\nu^T y| \mid \hat{\tau}_j \in \mathcal{M}(Y), \Pi_{\nu}^+ Y = \Pi_{\nu}^+ y), \quad (16) $$

where once again, we condition on $\Pi_{\nu}^+ Y = \Pi_{\nu}^+ y$ for technical reasons. Roughly speaking, (16) asks: “Out of all data sets yielding a changepoint at $\hat{\tau}_j$, what is the probability, under the null that there is no changepoint at this location, that the difference in mean within a fixed window of $\hat{\tau}_j$ is as large as what is observed?”

The $p$-values in (16) and (10) are calculated for slightly different null hypotheses: the null for (16) is that there is no changepoint within a distance $h$ of the estimated changepoint, $\hat{\tau}_j$. By contrast, (10) tests for no change in mean between the estimated changepoints immediately before and after $\hat{\tau}_j$. Furthermore, (16) conditions on less information. We believe that in many applications, the null hypothesis assumed by (16) 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 (16) could also be applied to test for the null hypothesis assumed by (10), while conditioning on less information. We further investigate these issues in Sections 6 and 8.1.

Theorem 1 can be extended to show that (16) is equal to

$$ p = \Pr(|\phi| \geq |\nu^T y| \mid \hat{\tau}_j \in \mathcal{M}(y'(\phi))), \quad (17) $$

where $\phi \sim N(0, \|\nu\|^2\sigma^2)$, and where $y'(\phi)$ was defined in (12). Thus, computing the $p$-value requires characterizing the set

$$ S = \{\phi : \hat{\tau}_j \in \mathcal{M}(y'(\phi))\}; \quad (18) $$

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 $S$ can be efficiently characterized for binary and $\ell_0$ segmentation. We discuss the fused lasso in Section D of the Supplementary Materials.

### 3.3 Intuition for $y'(\phi)$ and $S$

To gain intuition for $y'(\phi)$ in (12), we consider $\nu$ defined in (6). We see that

$$ y'_j(\phi) \equiv \begin{cases} y_t & \text{if } t \leq \hat{\tau}_{j-1} \text{ or } t > \hat{\tau}_{j+1}, \\ y_t + \frac{\phi - \nu^T y}{1 + \frac{\hat{\tau}_{j+1} - \hat{\tau}_j}{\hat{\tau}_{j+1} - \hat{\tau}_j}} & \text{if } \hat{\tau}_{j-1} < t \leq \hat{\tau}_j, \\ y_t - \frac{\phi - \nu^T y}{1 + \frac{\hat{\tau}_{j+1} - \hat{\tau}_j}{\hat{\tau}_{j+1} - \hat{\tau}_{j-1}}} & \text{if } \hat{\tau}_j < t \leq \hat{\tau}_{j+1}. \end{cases} \quad (19) $$
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 \( \phi \) around \( \hat{\tau}_j \). In other words, we can view \( y'_t(\phi) \) 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, \( S = \{ \phi : M(y'_t(\phi)) = M(y) \} \) is the set of such perturbations that do not affect the set of estimated changepoints.

Figure 2 illustrates the intuition behind \( y'_t(\phi) \) in a simulated example 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 \( S = \{ \phi : M(y'_t(\phi)) = M(y) \} \), but \( 0 \) is not.

In Sections 4 and 5, and in Section D of the Supplementary Materials, we develop procedures to characterize \( S \) in the cases of binary segmentation, \( \ell_0 \) segmentation, and the fused lasso, respectively. Here, the procedure from Section 4 gives \( S = \{ \phi : M(y'_t(\phi)) = M(y) \} = (\infty, -0.2) \cup (0.2, \infty) \); see panel d) of Figure 2.

4 Efficient characterization of (13) and (18) for binary segmentation

We now turn our attention to computing the set (13) for k-step binary segmentation; (18) is detailed in Section B.3 of the Supplementary Materials. Extensions to variants of binary segmentation proposed in Olshen et al. (2004) and Fryzlewicz et al. (2014) are straightforward and, for brevity, are not included.

We begin by paraphrasing Proposition 1 of Hyun et al. (2018).

Proposition 1 (Proposition 1 of Hyun et al. (2018)) 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 \( \{ y : \Gamma y \leq 0 \} \) for a \( k(2T - k - 3) \times T \) matrix \( \Gamma \), which is a function of the estimated changepoints, orders, and signs.

We will now make use of this result in a new proposition. Recall from Section 2.2 that \( M(y), O(y), \) and \( \Delta(y) \) are defined as the set of estimated changepoints, orders, and signs.

Proposition 2 The set \( \{ \phi : M(y'_t(\phi)) = m, O(y'_t(\phi)) = o, \Delta(y'_t(\phi)) = d \} \) is an interval. Furthermore, the set \( S \) defined in (13) can be written as the union of such intervals,

\[
S = \{ \phi : M(y'_t(\phi)) = M(y) \} = \bigcup_{i=-N}^{N'} (a_i, a_{i+1}),
\]
where \( N' + N + 1 \) is the number of elements in the set
\[
I := \{(o, d) : \exists \alpha \in \mathbb{R} \text{ such that } o = O(y'(\alpha)), d = \Delta(y'(\alpha)), M(y) = \mathcal{M}(y'(\alpha))\}. \tag{21}
\]
That is, \( I \) is the set of possible orders and signs of the changepoints that can be obtained via a perturbation of \( y \) that yields changepoints \( \mathcal{M}(y) \).

Importantly, the set \( I \) has far fewer than \( 2^k k! \) elements, which is the total number of possible orders and signs for the \( k \) changepoints. The reason for the unconventional indexing in Proposition 2 will soon become apparent. Proposition 3 guarantees that Proposition 2 is of practical use.

**Proposition 3** \( \bigcup_{i=-N}^{N'} (a_i, a_{i+1}) \) defined in (20) can be efficiently computed.

Proposition 3 follows from a simple argument that we outline here. 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 the first statement in Proposition 2 with \( m = \mathcal{M}(y) \), \( o = \mathcal{O}(y) \), and \( d = \Delta(y) \) to identify the interval \([a_0, a_1]\). By construction, \([a_0, a_1] \subseteq \mathcal{S}\).

Next, for some small positive value of \( \eta \), we apply the first statement in Proposition 2 with \( m = \mathcal{M}(y'(a_1 + \eta)) \), \( o = \mathcal{O}(y'(a_1 + \eta)) \), and \( d = \Delta(y'(a_1 + \eta)) \) to identify the interval \([a_1, a_2]\). (If the left endpoint of this interval does not equal \( a_1 \), then we must repeat with a smaller value of \( \eta \).) We then check whether \( \mathcal{M}(y'(a_1 + \eta)) = \mathcal{M}(y) \); if so, then \([a_1, a_2] \subseteq \mathcal{S} \), and if not, then \([a_1, a_2] \not\subseteq \mathcal{S}\). Next, we apply the first statement of Proposition 2 with \( m = \mathcal{M}(y'(a_2 + \eta)) \), \( o = \mathcal{O}(y'(a_2 + \eta)) \), and \( d = \Delta(y'(a_2 + \eta)) \) to identify the interval \([a_2, a_3]\). We then determine whether \([a_2, a_3] \subseteq \mathcal{S} \). We continue in this way until we reach an interval containing \( \infty \). We then repeat this process in the other direction, applying the first statement of Proposition 2 with \( m = \mathcal{M}(y'(a_0 - \eta)) \), \( o = \mathcal{O}(y'(a_0 - \eta)) \), and \( d = \Delta(y'(a_0 - \eta)) \), and determining whether the resulting interval \([a_{-1}, a_0]\) belongs to \( \mathcal{S} \), until eventually we arrive at an interval containing \(-\infty\).

Proposition 4 shows that this procedure can be stopped early in order to substantially reduce computational costs, and obtain conservative \( p \)-values.

**Proposition 4** Let \( \tilde{\mathcal{S}} \) be defined as the set
\[
\tilde{\mathcal{S}} = (-\infty, a_{-r}) \cup \left\{ \bigcup_{i=-r}^{r'} (a_i, a_{i+1}) \right\} \cup (a_{r'+1}, \infty),
\]
for some \( r < N \) and \( r' < N' \), and for \( a_{-r} \leq -|\nu^\top y| \) and \( a_{r'+1} \geq |\nu^\top y| \). Then the \( p \)-value obtained by conditioning on \( \{\phi \in \tilde{\mathcal{S}}\} \) is greater than the \( p \)-value obtained by conditioning on \( \{\phi \in \mathcal{S}\} \), i.e.,
\[
Pr(|\phi| \geq |\nu^\top y| \mid \phi \in \tilde{\mathcal{S}}) \geq Pr(|\phi| \geq |\nu^\top y| \mid \phi \in \mathcal{S}).
\]
Section B of the Supplementary Materials contains proofs of Propositions 2 and 4. In that section, we also show that Propositions 2 and 3 can be easily modified to characterize (18). Section D of the Supplementary Materials contains a straightforward modification of this procedure to characterize (13) and (18) in the case of the fused lasso.

5 Efficient characterization of (13) and (18) for $\ell_0$ segmentation

In this section, we develop efficient algorithms to analytically characterize (13) for the $\ell_0$ segmentation problem (4) with a fixed value of $\lambda$; Section C.2 of the Supplementary Materials considers $\mathcal{S}$ in the case of (18). In particular, we wish to determine all values of $\phi$ such that $\phi \in \mathcal{S}$, without checking each value of $\phi$ individually. Roughly speaking, we show that it is possible to write (13) in terms of the cost to segment the perturbed data $y'(\phi)$. To compute the necessary cost functions, we derive recursions that look similar to those in Rigaill (2015) and Maidstone et al. (2017). However these recursions are for functions of two variables, rather than one, which requires fundamentally different techniques to avoid a computational cost that increases exponentially in $h$.

Let $\hat{K}$ denote the number of estimated changepoints resulting from $\ell_0$ segmentation (4) on the original data $y$ with fixed tuning parameter value $\lambda$, 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 $\hat{\tau}_{\hat{K}+1} \equiv T$. Recall the definition of $y'(\phi)$ in (12) and the definition of $\nu$ in (6). For a given value of $\phi$, $\mathcal{M}(y'(\phi)) = \mathcal{M}(y)$ if and only if the cost of $\ell_0$ segmentation of the data $y'(\phi)$ with the changepoints restricted to occur at $\hat{\tau}_1, \ldots, \hat{\tau}_{\hat{K}}$,

$$C(\phi) = \min_{u_0, u_1, \ldots, u_{\hat{K}}} \left\{ \frac{1}{2} \sum_{k=0}^{\hat{K}} \sum_{t=\hat{\tau}_k+1}^{\hat{\tau}_{k+1}} (y'_t(\phi) - u_k)^2 + \lambda \hat{K} \right\}, \quad (22)$$

is no greater than the cost of $\ell_0$ segmentation of $y'(\phi)$,

$$C'(\phi) = \min_{0=\tau_0 < \tau_1 < \ldots < \tau_{\hat{K}} < \tau_{\hat{K}+1} = T} \left\{ \frac{1}{2} \sum_{k=0}^{\hat{K}} \sum_{t=\tau_k+1}^{\tau_{k+1}} (y'_t(\phi) - u_k)^2 + \lambda \hat{K} \right\}. \quad (23)$$

In other words, $\mathcal{S} = \{ \phi : C(\phi) \leq C'(\phi) \}$. The following result will prove useful.

**Proposition 5** $C(\phi) = C'(\phi')$ for all $\phi$ and $\phi'$.

Proposition 5 follows from the fact that from (12) and (6), $y'(\phi)$ is equal to $y_t$ for $t \leq \hat{\tau}_{j-1}$ or $t > \hat{\tau}_{j+1}$, adds a constant that depends on $\phi$ to all data points for $\hat{\tau}_{j-1} < t \leq \hat{\tau}_j$, and subtracts a constant from all data points for $\hat{\tau}_j < t \leq \hat{\tau}_{j+1}$. Therefore, by inspection of (22), $C(\phi)$ does not depend on the value of $\phi$. 
Applying Proposition 5, we see that \( S = \{ \phi : C(\nu^\top y) \leq C'(\phi) \} \). Furthermore, 
\( C(\nu^\top y) \) is easy to calculate, by inspection of (22) (recall from (12) that \( y'(\nu^\top y) = y \)).
Hence, we simply need an efficient way to calculate \( C'(\phi) \), i.e., to perform \( \ell_0 \) segmentation on the perturbed data. In the interest of computational tractability, we need a single procedure that works for all values of \( \phi \) simultaneously, rather than (for instance) having to repeat the procedure for values of \( \phi \) on a fine grid.

We note that \( C'(\phi) \) can be decomposed into the cost of segmenting the data \( y'(\phi) \) with a changepoint at \( \hat{\tau}_j \),
\[
C'_{\hat{\tau}_j}(\phi) = \min_u \left\{ \text{Cost}(y'_{1:\hat{\tau}_j}(\phi); u) \right\} + \min_{u'} \left\{ \text{Cost}(y'_{(\hat{\tau}_j+1):T}(\phi); u') \right\} + \lambda, \tag{24}
\]
and the cost of segmenting the data \( y'(\phi) \) without a changepoint at \( \hat{\tau}_j \),
\[
C'_{-\hat{\tau}_j}(\phi) = \min_u \left\{ \text{Cost}(y'_{1:\hat{\tau}_j}(\phi); u) + \text{Cost}(y'_{(\hat{\tau}_j+1):T}(\phi); u) \right\}, \tag{25}
\]
where \( \text{Cost}(y'_{1:\hat{\tau}_j}(\phi); u) \) is the cost of segmenting \( y'_{1:\hat{\tau}_j}(\phi) \) with \( \mu_{\hat{\tau}_j} = u \). Combining (24) and (25), we have
\[
C'(\phi) = \min \left\{ C'_{\hat{\tau}_j}(\phi), C'_{-\hat{\tau}_j}(\phi) \right\}. \tag{26}
\]
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, \( \phi \), and the mean at the \( \hat{\tau}_j \)th timepoint, \( u \). A similar approach can be used to compute \( \text{Cost}(y'_{(\hat{\tau}_j+1):T}(\phi); u) \).

5.1 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 \( u \),
\[
\text{Cost}(y_{1:s}; u) = \min \left\{ \text{Cost}(y_{1:(s-1)}; u), \min_{u'} \left\{ \text{Cost}(y_{1:(s-1)}; u') \right\} + \lambda \right\} + \frac{1}{2} (y_s - u)^2. \tag{27}
\]
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 appears to be intractable due to the fact that, naively, Cost \((y_{1:s}; u)\) needs to be updated for each value of \(u \in \mathbb{R}\). However, surprisingly, Rigaill (2015) and Maitstone 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 (27) can be reused in the calculation of Cost \((y'_{1:\hat{\tau}_j} (\phi); u)\). In particular, we note that from (12) and (6), \(y_s' (\phi) = y_s\) for all \(s \notin \{ \hat{\tau}_{j-1} + 1, \ldots, \hat{\tau}_j + 1 \}\), and therefore, Cost \((y'_{1:\hat{\tau}_j-1} (\phi); u)\) = Cost \((y_{1:\hat{\tau}_j-1}; u)\). As a result, we only require a new algorithm to efficiently compute Cost \((y'_{1:(\hat{\tau}_{j-1}+1)} (\phi); u)\), \ldots, Cost \((y'_{1:\hat{\tau}_j} (\phi); u)\). However, since these cost functions are piecewise quadratic of two variables, developing functional pruning recursions similar to the one-dimensional recursions of (27) is fundamentally more difficult. Nonetheless, in Theorem 2 we show that Cost \((y'_{1:s} (\phi); u)\) for \(s = \hat{\tau}_{j-1} + 1, \ldots, \hat{\tau}_j\) is the pointwise minimum over a set \(\mathcal{C}_s\) that can be efficiently computed.

**Theorem 2** For \(\hat{\tau}_{j-1} < s \leq \hat{\tau}_j\),

\[
\text{Cost}(y'_{1:s}(\phi); u) = \min_{f \in \mathcal{C}_s} f(u, \phi),
\]

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

Theorem 2 is proven in Section C.1 of the Supplementary Materials. Timing results are in Section C.4 of the Supplementary Materials.

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

Recall from (26) that \(C'(\phi)\) is the minimum of \(C'_{\hat{\tau}_j}(\phi)\) and \(C'_{\hat{\tau}_j-1}(\phi)\), in (24) and (25), respectively. We now show how to compute \(C'_{\hat{\tau}_j}(\phi)\).

We use Theorem 2 to build the set \(\mathcal{C}_{\hat{\tau}_j}\). Additionally, we define \(\tilde{\mathcal{C}}_{\hat{\tau}_j+1} = \{\text{Cost}(y_{T:(\hat{\tau}_j+1)}; u)\}\), and build \(\tilde{\mathcal{C}}_{\hat{\tau}_j+1}, \ldots, \tilde{\mathcal{C}}_{\hat{\tau}_j+1}\) such that Cost \((y'_{1:(\hat{\tau}_j+1)} (\phi); u)\) = \(\min_{f \in \tilde{\mathcal{C}}_{\hat{\tau}_j+1}} f(u, \phi)\), using a modified version of Theorem 2 that accounts for the reversal of the timepoints.

Then, because Cost \((y'_{1:\hat{\tau}_j} (\phi); u)\) = \(\min_{f \in \mathcal{C}_{\hat{\tau}_j}} f(u, \phi)\) and Cost \((y'_{T:(\hat{\tau}_j+1)} (\phi); u)\) = \(\min_{f \in \tilde{\mathcal{C}}_{\hat{\tau}_j+1}} f(u, \phi)\), we have from (24) that

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

\[
= \min_{f \in \mathcal{C}_{\hat{\tau}_j}} \left\{ \min_u \{ f(u, \phi) \} \right\} + \min_{f \in \tilde{\mathcal{C}}_{\hat{\tau}_j+1}} \left\{ \min_{u'} \{ f(u', \phi) \} \right\} + \lambda. \tag{29}
\]
Since \( f(u, \phi) \) is piecewise quadratic in \( u \) and \( \phi \) (Theorem 2), we see that \( \min_u \{ f(u, \phi) \} \) is piecewise quadratic in \( \phi \). Therefore, the operation \( \min_{f \in \mathcal{C}_{\hat{\tau}_j}} \{ \min_u \{ f(u, \phi) \} \} \) can be efficiently performed using ideas from Rigaill (2015) and Maidstone et al. (2017), and in turn \( C'_{\hat{\tau}_j}(\phi) \) can be efficiently computed. Recall from Theorem 2 that the set \( \mathcal{C}_{\hat{\tau}_j} \) 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 \( \mathcal{C}_{\hat{\tau}_j} \), followed by performing the operation \( \min_u \{ f(u, \phi) \} \) a total of \( \hat{\tau}_j - \hat{\tau}_{j-1} + 1 \) times. We can similarly obtain the piecewise quadratic function \( C'_{\hat{\tau}_j}(\phi) \) of \( \phi \). Therefore, we can analytically compute \( C'(\phi) \).

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 vector \( \mu \in \mathbb{R}^{2000} \) has \( K = 50 \) changepoints, with absolute difference in mean \( \delta = |\mu_{r_j+1} - \mu_{r_j}| \), for \( \delta \in \{0.5, 1.0, 1.5, 2.0, 2.5, 3.0, 3.5, 4.0\} \). Panel a) of Figure 3 depicts a realization with \( \delta = 3 \).

We compare four different procedures for testing for a change in mean at an estimated changepoint, \( H_0 : \nu^T \mu = 0 \):

**Approach 1.** Conditioning on the estimated changepoints, order, and signs, \( \{ \phi : \mathcal{M}(y'(\phi)) = \mathcal{M}(y), \mathcal{O}(y'(\phi)) = \mathcal{O}(y), \Delta(y'(\phi)) = \Delta(y) \} \), for binary segmentation;

**Approach 2.** Conditioning on all of the estimated changepoints, \( \{ \phi : \mathcal{M}(y'(\phi)) = \mathcal{M}(y) \} \), for binary segmentation;

**Approach 3.** Conditioning on the \( j \)th estimated changepoint, \( \{ \phi : \hat{\tau}_j \in \mathcal{M}(y'(\phi)) \} \), for binary segmentation;

**Approach 4.** Conditioning on the \( j \)th estimated changepoint, \( \{ \phi : \hat{\tau}_j \in \mathcal{M}(y'(\phi)) \} \), for \( \ell_0 \) segmentation.

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 \( \ell_0 \) segmentation estimate the same (or very similar) number of changepoints\(^1\). We also assume that the underlying noise variance \( (\sigma^2 = 1) \) is known. In what follows, all results reported are averaged over 100 replicate data sets. Unless stated otherwise, we take the window size for testing (14) to be \( h = 50 \). In Approaches 1–3, we approximate the set \( \mathcal{S} \) with \( \hat{\mathcal{S}} \) as described in Proposition 4; we take \( |a_{-r}| = |a_{r+1}| = \max(10\sigma||\nu||_2, ||\nu^T y||) \).

In practice, model selection techniques can be used to estimate \( K \) (Yao, 1988; Lebarbier, 2005; Arlot et al., 2012). Similarly, one can estimate the noise variance \( \sigma^2 \) based on

---

\(^1\)On a given data set, there may not exist a value of \( \lambda \) such that \( \ell_0 \) penalization yields precisely 50 estimated changepoints.
the data \( y \) (Birgé and Massart, 2001; Lebarbier, 2005). Of course, the \( p \)-values in (11) and (17) do not account for these data-driven estimates.

In Section E of the Supplementary Materials, we present timing results for estimating changepoints as well as computing \( p \)-values using Approaches 1–4. Our algorithms are very efficient: for series of length \( T = 1000 \), estimating changepoints requires less than 0.02 seconds; calculating \( p \)-values requires less than 15 seconds in the case of Approach 4, and less than 150 seconds in the case of Approaches 1–3.

6.2 Type I error control under a global null

We take \( \delta = 0 \), so that \( \mu_1 = \ldots = \mu_{2000} \), and examine the \( p \)-values obtained from each of the four procedures for testing \( H_0 : \nu^\top \mu = 0 \) in Section 6.1. Panel b) of Figure 3 displays quantile-quantile plots of the observed \( p \)-value quantiles versus theoretical Unif\([0, 1]\) quantiles. The plots indicate Type I error control.

6.3 Increases in power due to smaller conditioning sets

Next, we illustrate that power increases as the size of the conditioning event decreases, by considering Approaches 1–3 from Section 6.1. Each approach uses binary segmentation; the only difference is in the size of the conditioning sets.

On a given dataset, we define the empirical power as the ratio between the number of true changepoints for which the nearest estimated changepoint has a \( p \)-value less than \( \alpha \) and is within \( \pm m \) timepoints, and the number of true changepoints,

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

Here, \( j(i) = \arg \min_{1 \leq l \leq K} |\tau_l - \hat{\tau}_l| \). Panel c) of Figure 3 shows the empirical power for each of the four approaches with \( \alpha = 0.05 \) and \( m = 2 \). As the size of the conditioning set decreases, the power increases substantially: the power increases by up to 15\% when we condition on \( \{ \phi : M(y'(\phi)) = M(y) \} \) instead of \( \{ \phi : \mathcal{M}(y'(\phi)) = \mathcal{M}(y) \} \), and it increases by another 20\% when we condition on \( \{ \phi : \hat{\tau}_j \in M(y'(\phi)) \} \) instead of \( \{ \phi : \mathcal{M}(y'(\phi)) = \mathcal{M}(y) \} \).

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 5.** Apply binary segmentation on the odd timepoints to estimate changepoints. Then apply a standard \( z \)-test of \( H_0 : \nu^\top \mu = 0 \) on the even timepoints;
Approach 6. Apply $\ell_0$ segmentation to the odd timepoints to estimate changepoints. Then apply a standard $z$-test of $H_0 : \nu^T \mu = 0$ on the even timepoints.

Because sample splitting involves estimating the changepoints on half of the data and testing for a change in mean using the other half of the data, the $p$-value resulting from a standard $z$-test for a change in mean is valid, but is conditional on the set of timepoints used to estimate the changepoints (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,

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

Figure 4 displays the power and detection probability for Approaches 1–6, where $\alpha = 0.05$ and $m = 2$. In panel a), we see that Approach 4 (which estimates changepoints via $\ell_0$ segmentation, and then conditions on only the $j$th estimated changepoint) yields the highest power, especially for larger values of $\delta$. In panel b), we observe that $\ell_0$ segmentation vastly outperforms binary segmentation in terms of its ability to detect true changepoints.

Additionally, Figure 4 illustrates the benefit of the inferential framework developed in this paper over naive sample-splitting approaches. Sample splitting is limited in its ability to detect changepoints, since only half of the data is used to estimate changepoints.

6.5 Assessment of different window sizes for testing (14)

The results in Figure 4 suggest that conditioning on just $\hat{\tau}_j \in \mathcal{M}(y'(\phi))$ as in (17) yields the greatest power to detect a difference in means around $\hat{\tau}_j$. However, this requires pre-specifying the window size in (14). We now address this possible weakness. For window sizes $h \in \{1, 30, 50\}$, we assess the performance of Approaches 3 and 4 from Section 6.1 in panel c) of Figure 4. We observe that, provided $h$ is large enough, the window size has little effect on the power.

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.

We estimate changepoints using 20-step binary segmentation, and $\ell_0$ segmentation using the penalty $\lambda = 2\hat{\sigma}^2 \log T$, which yields 20 estimated 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. We see that the number of discoveries (estimated changepoints whose p-value is less than 0.05) increases as the size of the conditioning set decreases. In Approach 1 we make 11 discoveries, in Approach 2 we make 13, and in Approaches 3 and 4 we make 15 discoveries.

8 Discussion

In this paper, we show that testing for a change in mean around an estimated changepoint simply requires characterizing the set $S$, defined in either (13) or (18). We introduce the necessary computational tools to do this for three popular changepoint detection algorithms. Importantly, since our approach does not rely on the polyhedral lemma of Lee et al. (2016), the conditioning sets that we use are much smaller than those in earlier work and lead to higher-powered tests. We now discuss a few extensions of our work.

8.1 Smaller conditioning sets for (10)

Similarly to Liu et al. (2018), we note that no special properties of the conditioning set were used in the proof of Theorem 1. For instance, instead of conditioning on the full set of changepoints as is done in Section 3.1, we could have instead conditioned on the $j$th estimated changepoint and its immediate neighbors. This would yield a p-value of the form $p = \Pr(\{\phi \geq \nu^T y \mid \{\hat{\tau}_{j-1}, \hat{\tau}_j, \hat{\tau}_{j+1}\} \subset M(y'(\phi))\}$, and requires only a minor modification to the algorithms in Sections 4 and 5 and in the Supplemental Materials.

For some conditioning sets and changepoint detection algorithms, it might be difficult to characterize $S$. In this case, it is still possible to approximate $S$ by testing whether or not $\phi \in S$ for a fine grid of $\phi$ values; this approach is also suggested by Liu et al. (2018).

8.2 Confidence intervals for the change in mean

To construct confidence intervals for the change in mean, we first define $H_0(c) : \nu^T \mu = c$. We note that since $C(\phi) = \{c \mid \Pr_{H_0(c)}(\{\phi \geq \nu^T y \mid \phi \in S\} \geq \alpha\}$ satisfies $\Pr(\nu^T \mu \in C(\phi) \mid \phi \in S) \geq 1 - \alpha$, the set $C(\phi)$ is a $100(1 - \alpha)$% confidence interval for $\nu^T \mu$. Importantly, we can efficiently calculate $C(\phi)$ since the set $S$ is unchanged as we vary $c$; only the mean of the null distribution for $\nu^T Y$ changes.

8.3 P-values for spikes obtained from calcium imaging data

The ideas in this paper apply beyond the change-in-mean model (1). In particular, the ideas in Section 3 only require conditioning on the sufficient statistics of $\nu^T Y$.

For example, we can apply these ideas to analyze 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 fired, $s_t > 0$:

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

In this model, scientific interest lies in determining the precise timepoints of the spikes. Jewell and Witten (2018) and Jewell et al. (2019) estimate the spikes by solving

$$\minimize_{c_1,\ldots,c_T} \left\{ \frac{1}{2} \sum_{t=1}^{T} (y_t - c_t)^2 + \lambda \sum_{t=2}^{T} 1_{(c_t - \gamma c_{t-1} \geq 0)} \right\},$$  

which is closely related to the $\ell_0$ segmentation problem (4) in Section 2.1.2. The framework from Section 3 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 $\nu$. Furthermore, the algorithms developed in Section 5 can be modified to efficiently characterize the selective distribution; we leave the details to future work.

**Acknowledgments**

We thank Zaid Harchaoui and Ali Shojaie for helpful conversations, and Jacob Bien for thoughtful suggestions based on an early version of this paper.

**References**

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

Arlot, S., Celisse, A., and Harchaoui, Z. (2012). A kernel multiple change-point algorithm via model selection. *arXiv preprint arXiv:1202.3878*.

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. and Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica*, pages 47–78.
Bai, J. and Perron, P. (2003). Computation and analysis of multiple structural change models. *Journal of Applied Econometrics*, 18(1):1–22.

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

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 (Methodological)*, 57(1):289–300.

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

Birgé, L. and Massart, P. (2001). A generalized Cp criterion for gaussian model selection.

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.

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.

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.

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. (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. et al. (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 *NIPS*, 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.

Hocking, T. D., Rigail, G., Fearnhead, P., and Bourque, G. (2018). Generalized functional pruning optimal partitioning (gfpop) for constrained changepoint detection in genomic data. *arXiv preprint arXiv:1810.00117*.

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.

Hyun, S., G’Sell, M., and Tibshirani, R. J. (2016). Exact post-selection inference for changepoint detection and other generalized lasso problems. *arXiv preprint arXiv:1606.03552*.

Hyun, S., Lin, K., G’Sell, M., and Tibshirani, R. J. (2018). Post-selection inference for changepoint detection algorithms with application to copy number variation data. *arXiv preprint arXiv:1812.03644*.

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.

Jewell, S. and Witten, D. (2018). Exact spike train inference via ℓ₀ optimization. *The Annals of Applied Statistics*, 12(4):2457.

Jewell, S. W., Hocking, T. D., Fearnhead, P., and Witten, D. M. (2019). Fast nonconvex deconvolution of calcium imaging data. *Biostatistics*. kxy083.

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.

Lebarbier, ´E. (2005). Detecting multiple change-points in the mean of gaussian process by model selection. *Signal Processing*, 85(4):717–736.

Lee, J. D., Sun, D. L., Sun, Y., Taylor, J. E., et al. (2016). Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927.
Li, H., Munk, A., Sieling, H., et al. (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 change point detection for genomic sequences of continuous measurements. Bioinformatics, 27(2):161–166.

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 $K_{\text{max}}$ change-points. Journal de la Société Française de Statistique, 156(4):180–205.

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.

Tian, X., Taylor, J., et al. (2018). Selective inference with a randomized response. The Annals of Statistics, 46(2):679–710.

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., 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.
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*.

Yao, Y.-C. (1988). Estimating the number of change-points via Schwarz’ criterion. *Statistics & Probability Letters*, 6(3):181–189.

Figure 1: The power of a test of (9) 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 (9) 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. (2018)). In panel (b), we conditioned on the much smaller set \( \mathcal{M}(Y) = \mathcal{M}(y) \) and \( \Pi_{\nu}^\perp Y = \Pi_{\nu}^\perp y \).
Figure 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 $\phi$ 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.
Figure 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) The panels display quantile-quantile plots comparing sample $p$-value quantiles under (1) with $\mu_1 = \ldots = \mu_{2000}$ versus theoretical quantiles of the Unif(0, 1) distribution, for the four approaches listed in Section 6.1. c) Empirical power, averaged over 100 replicates, is displayed for Approaches 1–3 defined in Section 6.1, each of which results from testing $H_0: \nu^T \mu = 0$ for changepoints estimated using binary segmentation with different conditioning sets. Various values of $\delta$, the true change in mean due to a changepoint, are shown on the $x$-axis. Power increases as the size of the conditioning set decreases.
Figure 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 5–6, which are described in Section 6.4. b) Detection probability for binary segmentation and $\ell_0$ segmentation using all of the data, as well as half of the data. c) Power of Approaches 3 and 4 from Section 6.1 for testing the null hypothesis (14), for three values of the window size $h$. 
Figure 5: The number of discoveries depends on the size of the conditioning set. 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). Changepoints are estimated via 20-step binary segmentation, and $\ell_0$ segmentation with tuning parameter $\lambda = 2\sigma^2 \log(2000) \approx 5.5$. 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.
Supplementary Materials:
Testing for a Change in Mean After Changepoint Detection

Sean Jewell ∗
Department of Statistics, University of Washington
Paul Fearnhead
Department of Mathematics and Statistics, Lancaster University
Daniela Witten
Departments of Statistics and Biostatistics, University of Washington

A Proof of Theorem 1

To characterize (10), we note that \( Y \) decomposes as

\[
Y = (I - \Pi_\nu^\perp)Y + \Pi_\nu^\perp Y, \tag{A1}
\]

where \( \Pi_\nu^\perp = I - \frac{\nu \nu^\top}{\|\nu\|^2} \). Then (10) becomes

\[
p = \Pr_{H_0} (|\nu^\top Y| \geq |\nu^\top y| \mid \mathcal{M}(Y) = \mathcal{M}(y), \Pi_\nu^\perp Y = \Pi_\nu^\perp y) \tag{A2}
= \Pr_{H_0} (|\nu^\top Y| \geq |\nu^\top y| \mid \mathcal{M}((I - \Pi_\nu^\perp)Y + \Pi_\nu^\perp y) = \mathcal{M}(y), \Pi_\nu^\perp Y = \Pi_\nu^\perp y) \tag{A3}
= \Pr_{H_0} (|\nu^\top Y| \geq |\nu^\top y| \mid \mathcal{M}((I - \Pi_\nu^\perp)Y + \Pi_\nu^\perp y) = \mathcal{M}(y)). \tag{A4}
\]

∗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.
Here, (A2) is our definition of a p-value (10), and (A3) follows from (A1) and the fact that \( \Pi_\nu^\bot Y = \Pi_\nu^\bot y \). Finally, (A4) follows from the fact that \( Y \) is Gaussian (see (1)) and so \( \nu^\top Y \) and \( \Pi_\nu^\bot Y \) are independent.

Moreover, we note that (1) implies that \( \nu^\top Y \sim N(\nu^\top \mu, \|\nu\|^2 \sigma^2) \), and that under the null hypothesis (9), \( \nu^\top Y \sim N(0, \|\nu\|^2 \sigma^2) \). We now define \( \phi = \nu^\top Y \); thus under the null hypothesis, \( \phi \sim N(0, \|\nu\|^2 \sigma^2) \). Recall that \( y' = y - \nu \nu^\top y \|\nu\|^2 + \nu \phi \|\nu\|^2 \).

Therefore, \( p = \Pr (|\phi| \geq |\nu^\top y\| \mid M(y') = M(y)) \). (A6)

### B Details related to Section 4

#### B.1 Proof of Proposition 2

To prove the first statement in Proposition 2, we note from Proposition 1 that the set of data that yields changepoints \( m \), orders \( o \), and signs \( d \) is of the form \( \{y : \Gamma y \leq 0\} \). Therefore, the set of \( \phi \) that yields \( M(y') = m \), \( O(y') = o \), and \( \Delta(y') = d \) is of the form \( \{\phi : \Gamma y'(\phi) \leq 0\} \). Since \( \Gamma y'(\phi) \leq 0 \) represents \( k(2T - k - 3) \) linear inequalities in \( \phi \), the set \( \{\phi : \Gamma y'(\phi) \leq 0\} \) is an interval.

The second statement in Proposition 2 follows from the fact that

\[
S = \bigcup_{o \in O, d \in D} \{\phi : M(y'(\phi)) = M(y), O(y'(\phi)) = o, \Delta(y'(\phi)) = d\} \tag{B7}
\]

\[
= \bigcup_{(o', d') \in \mathcal{I}} \{\phi : M(y'(\phi)) = M(y), O(y'(\phi)) = o', \Delta(y'(\phi)) = d'\} \tag{B8}
\]

\[
= \bigcup_{i=-N}^{N'} (a_i, a_{i+1}) \tag{B9}
\]

where \( O \) is the set of cardinality \( k! \) containing all possible orders of the \( k \) changepoints, and \( D := \{-1, +1\}^k \) is the set of possible signs. Recall that \( N' + N + 1 = |\mathcal{I}| \) for \( \mathcal{I} \) defined in (21).

A key insight of (B7)-(B9) is that (B7) is the union over \( 2^k k! \) intervals. By contrast, (B9) is a union over \( N' + N + 1 = |\mathcal{I}| \) intervals which in practice is much smaller than \( 2^k k! \).
B.2 Proof of Proposition 4

To prove Proposition 4, recall that \( S = \bigcup_{i=-N}^{N'} (a_i, a_{i+1}) \), as described in Section 4, where \( a_{-N} = -\infty \) and \( a_{N'} = \infty \). Also recall that \( \tilde{S} = (-\infty, a_{-r}) \cup \left( \bigcup_{i=-r}^{r'} (a_i, a_{i+1}) \right) \cup (a_{r'+1}, \infty) \), for some \( a_{-r} \leq -|\nu^T y| \) and \( a_{r'+1} \geq |\nu^T y| \). Since \( \{ |\phi| \geq |\nu^T y| \} \cap S = \{ \phi \in \tilde{S} \setminus S \} \), we have

\[
\Pr(\{ |\phi| \geq |\nu^T y| \} \mid \phi \in \tilde{S}) = \frac{\Pr(\{ |\phi| \geq |\nu^T y| \} \cap \{ \phi \in S \})}{\Pr(\phi \in \tilde{S})} = \frac{\Pr(\{ |\phi| \geq |\nu^T y| \} \cap \{ \phi \in S \}) + \Pr(\{ |\phi| \geq |\nu^T y| \} \cap \{ \phi \in \tilde{S} \setminus S \})}{\Pr(\phi \in \tilde{S}) + \Pr(\phi \in \tilde{S} \setminus S)} \geq \Pr(\{ |\phi| \geq |\nu^T y| \} \mid \phi \in S).
\]

B.3 Characterization of (18)

In this section, we show that we can characterize the set \( S \equiv \{ \phi : \hat{\tau}_j \in M(y'(\phi)) \} \) for changepoints estimated via binary segmentation. Our approach is very similar to that of Section 4. In the following two propositions, Propositions B1 and B2, we modify Propositions 2 and 3 for the case of \( S \) defined in (18).

**Proposition B1** The set \( \{ \phi : M(y'(\phi)) = m, O(y'(\phi)) = o, \Delta(y'(\phi)) = d \} \) is an interval. Furthermore, the set \( S \) defined in (18) can be written as the union of intervals,

\[
S = \left\{ \phi : \hat{\tau}_j \in M(y'(\phi)) \right\} = \bigcup_{i=-N}^{N'} (a_i, a_{i+1}), \quad (B10)
\]

where \( N' + N + 1 \) is the number of elements in the set

\[
\mathcal{I} := \{ (o, d) : \exists \alpha \in \mathbb{R} \text{ such that } o = O(y'(\alpha)), d = \Delta(y'(\alpha)), \hat{\tau}_j \in M(y'(\alpha)) \}. \quad (B11)
\]

\( \mathcal{I} \) is the set of possible orders and signs of the changepoints that can be obtained via a perturbation of \( y \) that yields a changepoint at \( \hat{\tau}_j \).

**Proposition B2** \( \bigcup_{i=-N}^{N'} (a_i, a_{i+1}) \) defined in (B10) can be efficiently computed.
We outline the proof for Proposition B2 here. We first run k-step binary segmentation on the data \( y \) in order to obtain estimated changepoints \( \mathcal{M}(y) \), orders \( \mathcal{O}(y) \), and signs \( \Delta(y) \). We then apply the first statement in Proposition B1 to obtain an interval \([a_0, a_1] \subset \mathcal{S}\). Next, for some small positive value of \( \eta \), we apply the first statement of Proposition B1 with \( m = \mathcal{M}(y'(a_1 + \eta)) \), \( o = \mathcal{O}(y'(a_1 + \eta)) \), and \( d = \Delta(y'(a_1 + \eta)) \) to identify the interval \([a_1, a_2] \). We then check whether \( \hat{\tau}_j \in \mathcal{M}(y'(a_1 + \eta)) \); if so, then \([a_1, a_2] \subset \mathcal{S} \), and if not, then \([a_1, a_2] \not\subset \mathcal{S} \). We continue in this vein, much as we did in Section 4, to obtain the full set \( \mathcal{S} \).

In fact, when characterizing the set \( \mathcal{S} = \{ \phi : \hat{\tau}_j \in \mathcal{M}(y'(\phi)) \} \), this procedure can be sped up. We first define the interval in \( \phi \) such that \( j \)-step binary segmentation yields the estimated changepoints \( m \), orders \( o \), and signs \( d \)

\[
\{ \phi : \mathcal{M}_j(y'(\phi)) = m, \mathcal{O}_j(y'(\phi)) = o, \Delta_j(y'(\phi)) = d \},
\]

where the subscripts indicate that we have used \( j \)-step binary segmentation as opposed to \( k \)-step binary segmentation.

Now, recall that \( \hat{\tau}_j \) is the \( j \)th estimated changepoint resulting from binary segmentation on the data \( y \). Suppose that \( j < k \). We first run \( j \)-step binary segmentation on \( y \) in order to obtain estimated changepoints \( \mathcal{M}_j(y) \), orders \( \mathcal{O}_j(y) \), and signs \( \Delta_j(y) \). Then we can identify an interval \([a_0, a_1] \subset \mathcal{S} \) by applying (B12) with \( m = \mathcal{M}_j(y) \), \( o = \mathcal{O}_j(y) \), and \( d = \Delta_j(y) \). This leads to substantial computational speed-ups if \( j \ll k \). Next, suppose that \( \hat{\tau}_j \) is the \( l \)th estimated changepoint resulting from \( k \)-step binary segmentation applied to \( y'(a_1 + \eta) \), for \( l < k \). Once again, we can identify an interval \([a_1, a_2] \subset \mathcal{S} \) by applying (B12) with \( m = \mathcal{M}_l(y'(a_1 + \eta)) \), \( o = \mathcal{O}_l(y'(a_1 + \eta)) \), and \( d = \Delta_l(y'(a_1 + \eta)) \). By contrast, if \( \hat{\tau}_j \notin \mathcal{M}_k(y'(a_1 + \eta)) \) or if \( \hat{\tau}_j \) is the \( k \)th estimated changepoint on the data \( y'(a_1 + \eta) \), then we must identify intervals using the first statement of Proposition B1.

## C Details related to Section 5

### C.1 Proof of Theorem 2

To compute \( \text{Cost}(y'_{1:s}(\phi); u) \) for \( s \in \{ \hat{\tau}_{j-1} + 1, \ldots, \hat{\tau}_{j+1} \} \), we will introduce a set of functions \( \mathcal{C}_s \); each function in the set will correspond to a possible configuration for the changepoints preceding the \( s \)th timepoint. Then, \( \text{Cost}(y'_{1:s}(\phi); u) = \min_{f \in \mathcal{C}_s} f(u, \phi) \). Importantly, we will construct the set \( \mathcal{C}_s \) in such a way that its size grows linearly, rather than exponentially, in the size of the set of values that \( s \) can take.
To begin, we let $C_{\tau_{j-1}} = \{\text{Cost}(y_{1:\tau_{j-1}}; u)\}$ be a set containing a single function, $\text{Cost}(y_{1:\tau_{j-1}}; u)$, which can be obtained by applying (27) for $s = 1, \ldots, \tau_{j-1}$. To obtain the set $C_{\tau_{j-1}+1}$, we must update $C_{\tau_{j-1}}$ to allow for the following two possibilities:

1. **There is no changepoint at the $(\hat{\tau}_{j-1})th$ timepoint.** In this case, the cost is
   $$\text{Cost}(y_{1:\tau_{j-1}}; u) + \frac{1}{2}(y_{\tau_{j-1}+1}(\phi) - u)^2.$$

2. **There is a changepoint at the $(\hat{\tau}_{j-1})th$ timepoint.** This incurs a penalty of $\lambda$, and leads to a cost of
   $$\min_{u'} \{\text{Cost}(y_{1:\tau_{j-1}}; u')\} + \frac{1}{2}(y_{\tau_{j-1}+1}(\phi) - u)^2 + \lambda.$$  

   Therefore, $\text{Cost}(y_{1:((\hat{\tau}_{j-1}+1))}(\phi); u) = \min_{f \in C_{\tau_{j-1}+1}} f(u, \phi)$, where
   $$C_{\tau_{j-1}+1} = \left\{\text{Cost}(y_{1:\tau_{j-1}}; u) + \frac{1}{2}(y_{\tau_{j-1}+1}(\phi) - u)^2, \min_{u'} \{\text{Cost}(y_{1:\tau_{j-1}}; u')\} + \frac{1}{2}(y_{\tau_{j-1}+1}(\phi) - u)^2 + \lambda\right\}.$$  

   Continuing on to the next timepoint, we can see that $\text{Cost}(y_{1:((\hat{\tau}_{j-1}+2))}(\phi); u) = \min_{f \in C_{\tau_{j-1}+2}} f(u, \phi)$, where
   $$C_{\tau_{j-1}+2} = \left\{\text{Cost}(y_{1:\tau_{j-1}}; u) + \frac{1}{2}(y_{\tau_{j-1}+1}(\phi) - u)^2 + \frac{1}{2}(y_{\tau_{j-1}+2}(\phi) - u)^2, \right. \tag{C13}$$
   $$\min_{u'} \{\text{Cost}(y_{1:\tau_{j-1}}; u')\} + \frac{1}{2}(y_{\tau_{j-1}+1}(\phi) - u)^2 + \frac{1}{2}(y_{\tau_{j-1}+2}(\phi) - u)^2 + \lambda, \tag{C14}$$
   $$\min_{u''} \left\{\text{Cost}(y_{1:\tau_{j-1}}; u'') + \frac{1}{2}(y_{\tau_{j-1}+1}(\phi) - u'')^2\right\} + \frac{1}{2}(y_{\tau_{j-1}+2}(\phi) - u)^2 + \lambda, \tag{C15}$$
   $$\min_{u'} \{\text{Cost}(y_{1:\tau_{j-1}}; u')\} + \min_{u''} \left\{\frac{1}{2}(y_{\tau_{j-1}+1}(\phi) - u'')^2 + \lambda\right\} + \frac{1}{2}(y_{\tau_{j-1}+2}(\phi) - u)^2 + \lambda. \tag{C16}$$

Here, (C13) corresponds to no changepoint at either $\hat{\tau}_{j-1}$ or $\hat{\tau}_{j-1}+1$, (C14) corresponds to a changepoint at $\hat{\tau}_{j-1}$, (C15) corresponds to a changepoint at $\hat{\tau}_{j-1}+1$, and (C16) corresponds to changepoints at $\hat{\tau}_{j-1}$ and $\hat{\tau}_{j-1}+1$. We could continue along this vein to create the sets
\( \mathcal{C}_{\tilde{\tau}_{j-1}+3}, \ldots, \mathcal{C}_{\tilde{\tau}_j} \), but the number of functions in the sets would scale exponentially, making computations intractable. Instead, we notice that we really care about the minimum of the functions in each set, as a function of \( u \) and \( \phi \); furthermore, since (C15) and (C16) are of the form \( h(\phi) + \frac{1}{2}(y'_{\tilde{\tau}_{j-1}+2}(\phi) - u)^2 + \lambda \), their minimum takes the form

\[
\min \left\{ \min_{u''} \left\{ \text{Cost}(y_{1:\tilde{\tau}_j}; u'') + \frac{1}{2}(y'_{\tilde{\tau}_{j-1}+1}(\phi) - u'')^2 \right\}, \min_{u'} \left\{ \text{Cost}(y_{1:\tilde{\tau}_{j-1}}; u') + \min_{u''} \left\{ \frac{1}{2}(y'_{\tilde{\tau}_{j-1}+1}(\phi) - u'')^2 + \lambda \right\} \right\} \right\} + \frac{1}{2}(y'_{\tilde{\tau}_{j-1}+2}(\phi) - u)^2 + \lambda. \tag{C17}
\]

Thus, it is not necessary for us to keep track of (C15) and (C16); we can just keep track of (C17) instead. Using this insight, as \( s \) increases by one, the set \( \mathcal{C}_s \) will increase by just one function, rather than increasing exponentially. Importantly, (C17) is a piecewise quadratic function of \( \phi \), plus a quadratic function of \( \phi \) and \( u \); therefore, it can be efficiently calculated and stored using ideas from Rigaill (2015) and Maidstone et al. (2017).

We now summarize the overall procedure. For \( s = \tilde{\tau}_{j-1} + 1, \ldots, \tilde{\tau}_j \), we update the set \( \mathcal{C}_s \) as

\[
\mathcal{C}_s = \left\{ f(u, \phi) + \frac{1}{2}(y'_{\phi}(\phi) - u)^2 : f \in \mathcal{C}_{s-1} \cup \{ h_s(\phi) \} \right\}, \tag{C18}
\]

where

\[
h_s(\phi) = \min_{f \in \mathcal{C}_{s-1}} \min_{u'} f(u', \phi) + \lambda. \tag{C19}
\]

Furthermore, from (C18)–(C19), the size of the set \( \mathcal{C}_s \) increases by one as \( s \) increases by one. Therefore, computing \( \text{Cost}(y'_{1:T}(\phi); u) \) requires \( 1 + 2 + \ldots + (\tilde{\tau}_j - \tilde{\tau}_{j-1}) = \mathcal{O}((\tilde{\tau}_j - \tilde{\tau}_{j-1})^2) \) operations in the case of (13).

### C.2 Characterization of (18)

In this section, we show that we can characterize the set \( \mathcal{S} \equiv \{ \phi : \tilde{\tau}_j \in \mathcal{M}(y'_{\phi}) \} \) for changepoints estimated via \( \ell_0 \) segmentation. For \( \mathcal{S} \) defined in (18), \( \phi \in \mathcal{S} \) if and only if the cost of segmenting \( y'_{1:T}(\phi) \) with a changepoint at \( \tilde{\tau}_j \),

\[
\tilde{C}(\phi) = \min_u \left\{ \text{Cost}(y'_{1:\tilde{\tau}_j}(\phi); u) \right\} + \min_u \left\{ \text{Cost}(y'_{T:(\tilde{\tau}_j+1)}(\phi); u) \right\} + \lambda, \tag{C20}
\]

is no greater than the cost of segmenting \( y'_{1:T}(\phi) \) with no changepoint at \( \tilde{\tau}_j \),

\[
\tilde{C}'(\phi) = \min_u \left\{ \text{Cost}(y'_{1:\tilde{\tau}_j}(\phi); u) + \text{Cost}(y'_{T:(\tilde{\tau}_j+1)}(\phi); u) \right\}, \tag{C21}
\]

...
where \( \text{Cost}(y_{1:s}; u) \) is defined in (27). Therefore, \( S = \{ \phi : \hat{\tau}_j \in \mathcal{M}(y' (\phi)) \} = \{ \phi : \hat{C}(\phi) \leq \hat{C'}(\phi) \} \). We note that (C20) and (C21) are identical to (24) and (25) defined in Section 5, except here the contrast \( \nu \) is defined in (15), whereas in Section 5 it is defined in (6). Therefore, we can compute \( S \) using a slightly modified version of the procedure of Section 5. Section C.3 of the Supplementary Materials illustrates the details on a small example.

We also note that computing \( \text{Cost}(y'_{1: \hat{\tau}_j} (\phi); u) \) requires \( 1 + 2 + \ldots + h = \mathcal{O}(h^2) \) operations in the case of (18). Timing results are presented in Section C.4 of the Supplementary Materials.

### C.3 An illustration of the procedure of Section C.2

To better grasp the procedure described in Section C.2 of the Supplementary Materials to characterize the set \( S = \{ \phi : \hat{\tau}_j \in \mathcal{M}(y' (\phi)) \} \) in (18) for \( \ell_0 \) segmentation, in this section we work through a simple example. Suppose we observe \( y = [1, 1, 1, 2, 2, 2] \), and estimate a changepoint at \( \hat{\tau} = 3 \) by solving (4) with \( \lambda = \frac{1}{2} \).

In this example, we take \( h = 2 \), and use the simplified perturbation model

\[
y'_t (\phi) = \begin{cases} 
y_t & t = 1, 6, \\
y_t + \phi & t = 2, 3, \\
y_t - \phi & t = 4, 5. 
\end{cases} \tag{C22}
\]

We wish to ultimately compute \( \mathcal{E}_3 \), so we begin with \( \mathcal{E}_1 = \{ \text{Cost}(y_1; u) \} \),

\[
\text{Cost}(y_1; u) = \frac{1}{2} (1 - u)^2,
\]

and repeatedly use (C18) and (C19) to obtain \( \mathcal{E}_2 \) from \( \mathcal{E}_1 \) and \( \mathcal{E}_3 \) from \( \mathcal{E}_2 \).

\( \mathcal{E}_2 \) contains two functions: the first function represents the cost of segmenting \([1, 1 + \phi]\) with zero changepoints and where the mean \( \mu_2 = u \); the second function represents the cost of segmenting \([1, 1 + \phi]\) with a changepoint at timepoint 1, and where the mean \( \mu_2 = u \). By (C18), this is simply

\[
\mathcal{E}_2 = \left\{ \frac{1}{2} (1 - u)^2 + \frac{1}{2} (1 + \phi - u)^2, h_2(u, \phi) + \frac{1}{2} (1 + \phi - u)^2 \right\},
\]

where

\[
h_2(u, \phi) = \min_{u'} \text{Cost}(y_1; u') + \lambda = \min_{u'} \frac{1}{2} (1 - u')^2 + \frac{1}{2} = \frac{1}{2}.
\]

7
More explicitly,

\[ C_2 = \left\{ \frac{1}{2}(1 - u)^2 + \frac{1}{2}(y'_2(\phi) - u)^2, \frac{1}{2} + \frac{1}{2}(y'_2(\phi) - u)^2 \right\} \]

\[ = \left\{ u^2 - 2u - u\phi + \frac{1}{2}\phi^2 + \phi + 1, \frac{1}{2}u^2 - u - u\phi + \frac{1}{2}\phi^2 + \phi + 1 \right\}. \]

To compute \( C_3 \), we first calculate the minimum (corresponding to a changepoint at time-point 2)

\[ h_3(u, \phi) = \min_{f \in \mathcal{C}_2} \min_{u'} f(u', \phi) + \lambda = \begin{cases} 1 & \phi < -\sqrt{2} \\ \frac{1}{4}\phi^2 + \frac{1}{2} & -\sqrt{2} \leq \phi \leq \sqrt{2} \\ 1 & \phi > \sqrt{2} \end{cases}, \]

and add the perturbed data point, \( 1 + \phi \), to obtain \( C_3 = \{ q_1(u, \phi), q_2(u, \phi), q_3(u, \phi) \} \), where

\[ q_1(u, \phi) = 1.5u^2 - 3u - 2u\phi + \phi^2 + 2\phi + 1.5, \]
\[ q_2(u, \phi) = u^2 - 2u - 2u\phi + \phi^2 + 2\phi + 1.5, \]
\[ q_3(u, \phi) = \begin{cases} 0.5u^2 - u - u\phi + 0.5\phi^2 + \phi + 1.5 & \phi < -\sqrt{2} \\ 0.5u^2 - u - u\phi + 0.75\phi^2 + \phi + 1 & -\sqrt{2} \leq \phi \leq \sqrt{2} \\ 0.5u^2 - u - u\phi + 0.5\phi^2 + \phi + 1.5 & \phi > \sqrt{2} \end{cases}. \]

For any \( u \) and \( \phi \), the optimal cost of segmenting \( y'_{1:3}(\phi) \) is given as \( \text{Cost}(y'_{1:3}(\phi); u) = \min_{f \in \mathcal{C}_3} f(u, \phi) \).

Applying similar steps in the reverse direction from timepoint 6 to timepoint 4, gives

\[ \text{Cost}(y'_{6:4}(\phi); u) = \min\{ f_1(u, \phi), f_2(u, \phi), f_3(u, \phi) \}, \]

where

\[ f_1(u, \phi) = 1.5u^2 - 6u + 2u\phi + \phi^2 - 4\phi + 6, \]
\[ f_2(u, \phi) = u^2 - 4u + 2u\phi + \phi^2 - 4\phi + 4.5, \] and
\[ f_3(u, \phi) = \begin{cases} 0.5u^2 - 2u + u\phi + 0.5\phi^2 - 2\phi + 3 & \phi < -\sqrt{2} \\ 0.5u^2 - 2u + u\phi + 0.75\phi^2 - 2\phi + 2.5 & -\sqrt{2} \leq \phi \leq \sqrt{2} \\ 0.5u^2 - 2u + u\phi + 0.5\phi^2 - 2\phi + 3 & \phi > \sqrt{2} \end{cases}. \]
\( \tilde{C}(\phi) \) and \( \tilde{C}'(\phi) \), defined in (C20) and (C21), are calculated as
\[
\tilde{C}(\phi) = \min_u \text{Cost}(y'_{1:3}(\phi); u) + \min_u \text{Cost}(y'_{6:4}(\phi); u) + \lambda = \begin{cases} 
\frac{3}{2} & \phi < -\sqrt{\frac{3}{2}} \\
\frac{2}{3}\phi^2 + \frac{1}{2} & -\sqrt{\frac{3}{2}} \leq \phi \leq \sqrt{\frac{3}{2}} \\
\frac{3}{2} & \phi > \sqrt{\frac{3}{2}} 
\end{cases},
\]
and
\[
\tilde{C}'(\phi) = \min_u \{ \text{Cost}(y'_{1:3}(\phi); u) + \text{Cost}(y'_{6:4}(\phi); u) \} = \begin{cases} 
\phi^2 - \phi + 2.25 & \phi < -1.41421 \\
1.5\phi^2 - \phi + 1.25 & -1.41421 \leq \phi \leq -1 \\
1.625\phi^2 - 1.25\phi + 0.875 & -1 \leq \phi \leq -0.1547 \\
2\phi^2 - 2\phi + 0.75 & -0.1547 \leq \phi \leq 1.76619 \\
1.375\phi^2 + 1.375\phi + 2.25 & 1.76619 \leq \phi \leq 1.89681 \\
\phi^2 - \phi + 2.25 & \phi > 1.89681 
\end{cases}.
\]

To determine \( S \), we recall from Section C.2 that \( S = \{ \phi : \tilde{C}(\phi) \leq \tilde{C}'(\phi) \} \). Therefore, we take the minimum
\[
\min \{ \tilde{C}(\phi), \tilde{C}'(\phi) \} = \begin{cases} 
1.5 & \phi < -1.22474 \quad \text{Minimizer: } \tilde{C}(\phi) \\
\frac{2}{3}\phi + \frac{1}{2} & -1.22474 \leq \phi \leq 0.13763 \quad \text{Minimizer: } \tilde{C}(\phi) \\
2\phi^2 - 2\phi + 0.75 & 0.13763 \leq \phi \leq 1.29057 \quad \text{Minimizer: } \tilde{C}'(\phi) \\
1.5 & \phi > 1.29057 \quad \text{Minimizer: } \tilde{C}(\phi) 
\end{cases}
\]
and for each point \( \phi \) track whether \( \tilde{C}(\phi) \) or \( \tilde{C}'(\phi) \) minimized the objective. Therefore, \( S = (-\infty, 0.13763] \cup [1.29057, \infty) \). Figure S1 shows \( \tilde{C}(\phi) \) and \( \tilde{C}'(\phi) \).

### C.4 Timing results for computing the set \( S \) defined in (18)

In this section, we investigate the claim of Section C.2 of the Supplementary Materials, that computing the set \( S \) defined in (18) in the case of \( \ell_0 \) segmentation requires \( O(h^2) \) computations, where \( h \) is the window size that appears in (14).

Figure S2 displays the average running time over 50 replicate datasets as a function of the window size, \( h \), on a simulated dataset of 2000 timepoints, which contains a single changepoint at the 1000th timepoint. We see that the running time is, in fact, approximately quadratic in the window size.
D Efficient analytical characterization of (13) and (18) for the fused lasso

The fused lasso problem (5) can be reformulated as the regression problem

$$\min_{\beta \in \mathbb{R}^T} \left\{ ||y - X\beta||_2^2 + \lambda||\beta||_1 \right\},$$

(D23)

for a $T \times T$ matrix $X$ whose $j$th row contains $j$ ones followed by $T-j$ zeros. (5) and (D23) are equivalent in the sense that $\hat{\beta}_t = \hat{\mu}_t - \hat{\mu}_{t-1}$ for $t = 2, \ldots, T$ and $\hat{\beta}_1 = \hat{\mu}_1$.

Lee et al. (2016) show that the set of $y$ for which the lasso (D23) results in a given set of selected variables and signs can be written as the polyhedral set $\{ y : Ay \leq b \}$ for a $T \times T$ matrix $A$ and a $T$-vector $b$. $A$ and $b$ have explicit formulas depending only on the selected variables and coefficient signs. Therefore, Lee et al. (2016) are able to compute p-values for the null hypothesis that the estimated coefficients are zero conditional on the selected variables, the signs of the estimated coefficients, and nuisance parameters.

To avoid conditioning on the signs of the estimated coefficients, we slightly modify the arguments outlined in Section 4. In the following propositions, Propositions D3 and D4, we modify Propositions 2 and 3 for $S = \{ \phi : \text{supp}(\hat{\beta}(y'(\phi))) = \text{supp}(\hat{\beta}(y)) \}$, where $\text{supp}(\hat{\beta}(y))$ denotes the set of selected variables obtained from solving (D23) with data $y$.

**Proposition D3** The set $\{ \phi : \text{supp}(\hat{\beta}(y'(\phi))) = m, \text{sign}(\hat{\beta}(y'(\phi))) = d \}$ is an interval. Furthermore, the set $S = \{ \phi : \text{supp}(\hat{\beta}(y'(\phi))) = \text{supp}(\hat{\beta}(y)) \}$ can be written as the union of intervals,

$$S = \{ \phi : \text{supp}(\hat{\beta}(y'(\phi))) = \text{supp}(\hat{\beta}(y)) \} = \bigcup_{i=-N}^{N'} (a_i, a_{i+1}),$$

(D24)

where $N' + N + 1$ is the number of elements in the set

$$\mathcal{I} := \left\{ d : \exists \alpha \in \mathbb{R} \text{ such that } d = \text{sign}(\hat{\beta}(y'(\alpha))), \text{supp}(\hat{\beta}(y)) = \text{supp}(\hat{\beta}(y'(\alpha))) \right\}. \quad (D25)$$

$\mathcal{I}$ is the set of possible coefficient signs that can be obtained via a perturbation of $y$ that yields the same non-zero coefficients as $\hat{\beta}(y)$.

**Proposition D4** $\bigcup_{i=-N}^{N'} (a_i, a_{i+1})$ defined in (D24) can be efficiently computed.

Now, we outline the proof for Proposition D4. We first solve (D23) on the data $y$ in order to obtain $\text{supp}(\hat{\beta}(y))$ and $\text{sign}(\hat{\beta}(y))$. We then apply the first statement in Proposition D3
to obtain an interval \([a_0, a_1] \subset S\). Next, for some small positive value of \(\eta\), we apply the first statement of Proposition D3 with \(m = \text{supp}(\hat{\beta}(y'(a_1 + \eta)))\) and \(d = \text{sign}(\hat{\beta}(y'(a_1 + \eta)))\) to identify the interval \([a_1, a_2]\). We then check whether \(\text{supp}(\hat{\beta}(y')) = \text{supp}(\hat{\beta}(y'(a_1 + \eta)))\); if so, then \([a_1, a_2] \subset S\), and if not, then \([a_1, a_2] \not\subset S\). We continue in this vein, much as we did in Section 4, to obtain the full set \(S\).

### D.1 Generalized lasso

In this section, we show that we can use the tools from Section 4 to characterize the selection event of the generalized lasso. In Section D of the Supplementary Materials we rewrote the fused lasso problem (5) in terms of a lasso (regression) problem (D23), which allowed us to develop a simple procedure to characterize \(S\). The generalized lasso (Tibshirani and Taylor, 2011) is the solution to the optimization problem

\[
\begin{align*}
\min_{\beta \in \mathbb{R}^T} \left\{ ||y - \beta||_2^2 + \lambda||D\beta||_1 \right\},
\end{align*}
\]

where \(D\) is a matrix whose rows encode our beliefs about the underlying structure in the data. For general \(D\), (D26) cannot be rewritten in the form of (D23), and so existing machinery for selective inference for the lasso cannot be applied. Nonetheless, by also conditioning on the order that variables enter the model, Hyun et al. (2016) show that the selection event of the generalized lasso is polyhedral. Therefore, an extension of the ideas in Section D could be applied in order to conduct selective inference using a smaller conditioning set.

### E Timing results for estimating changepoints and computing \(p\)-values

In this section, we present timing results for estimating changepoints and computing \(p\)-values. Figure S3 displays the running time, computed on a MacBook Pro with a 2.5 GHz Intel Core i7 processor, for estimating changepoints and calculating \(p\)-values for Approaches 1–4 defined in Section 6.1. We take \(\lambda = \log(T)\) for \(\ell_0\) segmentation and use \(\max(\hat{K}, 1)\)-step binary segmentation for \(\hat{K}\) equal to the number of estimated changepoints from \(\ell_0\) segmentation. Fifty replicate datasets are simulated according to model (1) with \(\sigma^2 = 1\), and with \(K = 10[\log_{10}(T)]\) changepoints sampled without replacement from the set \(\{1, \ldots, T\}\). At each changepoint, the absolute difference in mean is \(|\mu_{\tau_j+1} - \mu_{\tau_j}| = 1.5\). Our implementations of Approaches 1–3 approximate the set \(S\) with \(\tilde{S}\) as described in Proposition 4; we take \(|a-r| = |a_r' + 1| = \max(10\sigma||\nu||_2, |\nu^Ty|)\).
Estimating changepoints with binary and $\ell_0$ segmentation is very fast (under 0.06 seconds for all series lengths considered). On the other hand, inference is much more costly for all approaches. In particular, we note that Approach 4 is almost an order of magnitude faster than Approaches 1–3 for longer series lengths. We note that Approach 3 can be sped up using the idea presented in Section B.3.

References

Hyun, S., G’Sell, M., and Tibshirani, R. J. (2016). Exact post-selection inference for changepoint detection and other generalized lasso problems. *arXiv preprint arXiv:1606.03552*.

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

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.

Rigaill, G. (2015). A pruned dynamic programming algorithm to recover the best segmentations with 1 to $K_{\text{max}}$ change-points. *Journal de la Société Française de Statistique*, 156(4):180–205.

Tibshirani, R. J. and Taylor, J. (2011). The solution path of the generalized lasso. *The Annals of Statistics*, 39(3):1335–1371.
Figure S1: Optimal cost of segmenting $y'(\phi)$ as a function of $\phi$, in the example in Section C.3 of the Supplementary Materials. $\tilde{C}(\phi)$ is the optimal cost of segmenting $y'(\phi)$ as a function of $\phi$ given that there is a changepoint at $\hat{\tau} = 3$ (red). $\tilde{C}'(\phi)$ is the optimal cost of segmenting $y'(\phi)$ given that there is no changepoint at $\hat{\tau} = 3$ (blue).
Figure S2: Average time, in seconds, to compute the set $\mathcal{S}$ in (18), as a function of the window size $h$ on 50 replicated datasets. Both axes are displayed on the log scale. The function time $= e^{-3.3h^2}$ (red) is displayed for reference. Details are provided in Section C.4 of the Supplementary Materials.
Figure S3: Computational cost of Approaches 1–4 defined in Section 6.1. 50 replicate datasets are simulated according to model (1) with $\sigma^2 = 1$ and with $K = 10\lfloor\log_{10}(T)\rfloor$ changepoints sampled without replacement from \{1, ..., T\}. At each changepoint the absolute difference in mean, $|\mu_{\tau_j+1} - \mu_{\tau_j}|$, is 1.5. Details are provided in Section E of the Supplementary Materials.