Valid post-selection inference for Segmentation Methods with application to copy number variation data

Sangwon Hyun Kevin Lin Max G'Sell Ryan J. Tibshirani

Abstract

Changepoint inference is a relatively unexplored methodology area with several attractive applications in science and industry. In this paper, we extend existing post-selection inference methodology to various segmentation-based changepoint algorithms which give exact, finite-sample changepoint inference. These tools are directly useful for practitioners for interpreting and communicating results, and directly improves upon widely used changepoint detection tools. We characterize and implement the polyhedral selection space of these algorithms. Several extensions to the general post-selection tools such as randomization and different modeling assumptions are also explored. These extensions enable more powerful inferences. We make practical recommendations for modelling choices for the algorithm and inference, based on extensive simulations. In addition, we investigate the application of our proposed methods to real and pseudo-real array CGH data example, to demonstrate the inferential properties of our proposed tools. Lastly, we developed fast `R` software for implementing various selective inference approaches for common segmentation methods.

1 Introduction

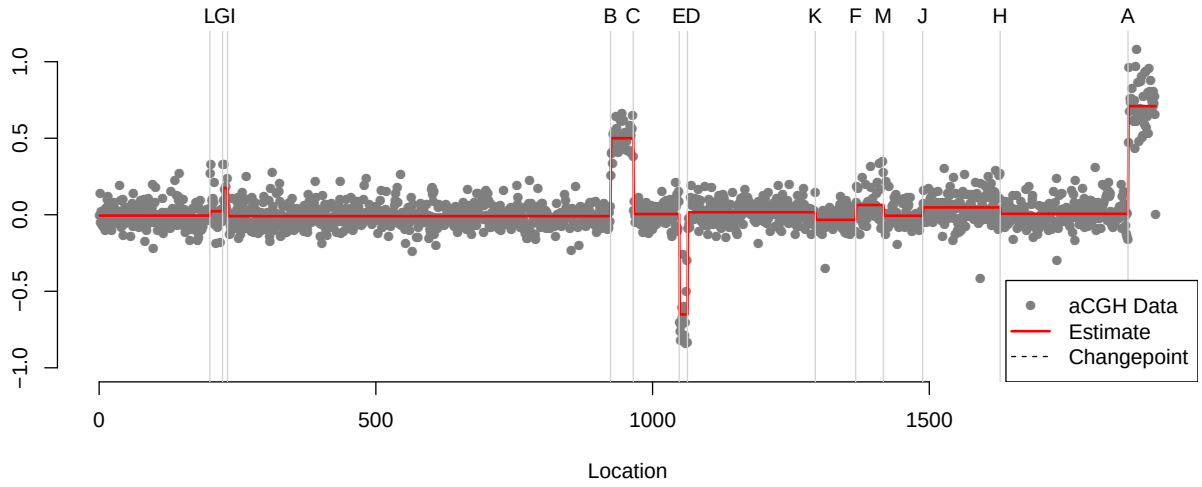
There are numerous scientific and industry data applications in which abrupt changes occur in the underlying structure of the data across some dimension. The study of methods to detect and test these changes from this type of data is called changepoint detection. Among changepoint detection algorithms, segmentation algorithms are a popular class of methods that can be applied to 1 dimensional data. Segmentation algorithms typically involve recursive splitting of the data at the most plausible location according to a split criterion. They are well studied in the literature, and hold the advantage of being intuitive to implement and communicate.

A single application of a segmentation algorithm to data gives a set of estimated changepoints. A valuable methodological addition is to be able to conduct statistical inference about these estimated changepoint locations – providing a level of confidence or plausibility of a detected jumps using hypothesis tests or confidence intervals. Approaches to carry out such inferences for common segmentation algorithms have not been developed in the literature. If classical tests of changepoint locations – t-tests or z-tests of means in data segments – are to be conducted *without* accounting for selection by the algorithm on the same data, inferential guarantees like type-I error do not hold.

Post-selection inference aims to solve one aspect of this problem, by explicitly conditioning on the event that application of the algorithm yielded the selected model, and conducting inferences under certain parametric assumptions. In this 'selective' distribution and under a suitable null, a hypothesis test can be conducted regarding linear constrasts describing interesting quantities of a changepoint model, directly formed from an algorithm's output. By a straightforward inversion of the p-values from these tests, valid CI's can also be produced.

Take the example (figure 1) of an observed array CGH (aCGH) dataset of fibroblast cell line GM05296, originally published in Snijders et al. (2001), of $n = 2011$ measurements from 23 chromosomes measured in a row. In aCGH data, regions of deviations from a baseline level of zero are scientifically interesting, and are often linked to genetically driven diseases. Detailed analysis on this dataset will be presented in a later section. Figure 1 shows the results from

carrying out our proposed tests to 13 changepoints after having applied binary segmentation with additive noise – a technique discussed in section 2.2. The model size 13 was chosen according to a data-dependent stopping rule. These test results were compared to two-sided naive Z-tests. We can see that the first five locations A through E are deemed significant against a Bonferonni-corrected cutoff of 0.05/13, while all others are deemed nonsignificant under the test. The Z-tests on the same locations conclude that twelve out of the thirteen locations are significant jumps, which is clearly an optimistic result.



Location	L	G	I	B	C	E	D	K	F	M	J	H	A
Naive	0.04	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Post-selection	0.40	0.26	0.19	0.00	0.00	0.00	0.00	0.01	0.06	0.95	0.97	0.91	0.00

Figure 1: *Wild binary segmentation inference applied to Array CGH data set of Coriell cell lines. Locations A through M were recovered in that order. Stop time was 13 according to two-rise BIC rule. By conducting post-selection segment tests in each location, A,B,C,D were significant changepoints (p-values were below the cutoff of $\alpha/13$). This is shown in the bottom row of the table. On the other hand, naive z-test inference of the same segments, ignoring the selection, shown in the top row of the table, deems all twelve locations except for location L as significant.)*

In this paper we extend the post-selection inference framework, first introduced in Tibshirani et al. (2016) and Lee et al. (2016), to enable inference on changepoint-related parametric quantities after selection by segmentation algorithms. In addition to the base case, we improve these tools by combining the ideas of randomization in Tian & Taylor (2015) and different null models suggested in Fithian et al. (2014) in order to enable more powerful inference. We also describe two Monte Carlo Markov Chain sampling strategies for implementing these extensions. Lastly, we demonstrate the application of our methods in simulation and on real data examples of several array CGH data.

Contributions of the paper

- We extend existing post-selection inference methodology to various segmentation-based changepoint algorithms.
- We characterize the polyhedral selection space of these algorithms.
- We make practical recommendations for modelling choices for the algorithm and inference, based on extensive simulations.
- We demonstrate application to real and pseudo-real array CGH data example, to demonstrate the inferential properties of our proposed tools.
- We developed `R` software for implementing various selective inference.

1.1 Notation

Whenever applicable, we will use colon notation $g_{a:b}$ for $\{g_a, \cdots, g_b\}$ and subscripted set notation for indices in A: $g_A = \{g_j : j \in A\}$. Similarly, $\bar{y}_{a:b} := \frac{1}{b+1-a}\sum_{i=a}^{b} y_i$ is the sample mean of the subvector $y_{a:b}$. We will also use the subscript notation as in P_{obs} to indicate the quantities based on the observed data, y_{obs}. The set of positive semi-definite matrices of size $n \times n$ is denoted by $\mathcal{S}_+^n$, and the $n \times n$ identity matrix as I_n. We will use the $\hat{\cdot}(y)$ notation in order to indicate specific fitted quantities. $\hat{b}_{1:k}(y) = (\hat{b}_1 y, \cdots \hat{b}_k(y))$ is k-length index set between 1 and $n-1$ of changepoints, and $\hat{s}_{1:k}(\cdot)$ denotes their directions; each element is equal to +1 for upward and -1 for downward changepoints.

1.2 Changepoint algorithms

Many algorithms have been developed in the literature for retrospective changepoint analysis. Segmentation algorithms are one of the most popular classes of algorithms for *multiple* changepoint detection in a single data stream or time series. Some examples are binary segmentation (Vostrikova 1981, Scott & Knott 1074), wild binary segmentation (Fryzlewicz 2014*a*) and circular binary segmentation (Olshen et al. 2004*b*). One significant advantage of binary segmentation is that the underlying mechanism is straightforward to understand and implement. Popular alternatives to segmentation methods are total variation denoising approaches, which minimize a likelihood function penalized by a total variation function. Many authors have studied the methodological and theoretical properties of total variation methods and the fused lasso (Rinaldo 2009, Harchaoui & Lvy-Leduc 2010). While not a focus of this paper, stochastic modeling like hidden Markov model based approaches has also been studied (Fridlyand et al. 2004). Recently, multiple stream segmentation has attracted interest and is studied by Fan & Mackey (2015). For 1d and regression model changepoint detection, Jandhyala et al. (2013), Horvath & Rice (2014), Aue & Horvath (2013), collectively provide nice survey of methods for changepoint estimation. In this paper, we focus on the most common segmentation algorithms, as well as the fused lasso.

Binary Segmentation The k-step binary segmentation (BS) algorithm takes a vector of data $y \in \mathbb{R}^n$ and sequentially splits the data according to break locations $\hat{b}_{(1:k)}(y) = (\hat{b}_1(y), \hat{b}_2(y), \cdots, \hat{b}_k(y))$ that produce the largest absolute cumulative sum (CUSUM) statistic $\tilde{y}^b_{s,e}$. Let s and e denote the starting and ending indices of a flat segment, and let b be the index of a proposed breakpoint. The CUSUM statistic is defined as:

$$\tilde{y}^b_{s,e} = \sqrt{\frac{1}{\frac{1}{|e-b|} + \frac{1}{|b+1-s|}}}(\bar{y}_{(b+1):e} - \bar{y}_{s:b}),$$

which is a variance-stabilized sample mean difference of two immediately adjacent segments to the left and right. Because all $\tilde{y}^b_{s,e}$ are of the same scale regardless of s, b or e, the comparison and maximization of the CUSUM statistic is still meaningful across different steps of the algorithm.

We now succinctly describe the algorithm. At the end of step i, denote by $0 = c_0 < c_1 < \cdots < c_{i+1} = n$ the sorted permutation of the detected changepoints $\hat{b}_{(1:i)}(y)$, and $c_0 = 0$ and $c_i = n$ for convenience. Also denote by $I_j = (c_{j-1} + 1) : c_j$ the data partitions made by $c_{1:j}$, for $j = 1, 2, \cdots, i+1$. Then, the next changepoint $\hat{b}_{i+1}(y)$ and the maximizing partition $\hat{j}_{i+1}(y)$ are obtained by the maximization of the absolute CUSUM statistic in the latest partitions of the data:

$$(\hat{j}_{i+1}(y), \hat{b}_{i+1}(y)) = \max_{\substack{j \in (1:(i+1)) \\ b \in I_j}} |\tilde{y}^b_{(c_{j-1}+1),c_j}|. \tag{1}$$

Additionally, the direction of the jump $\hat{s}_{i+1}(y)$ is calculated by the sign of the maximizing absolute CUSUM statistic $\hat{s}_{i+1}(y) = \text{sign}(\tilde{y}^{\hat{b}_{i+1}(y)}_{c_{j-1}+1,c_j})|_{j=\hat{j}_{i+1}(y)}$. To detect k changepoints, this procedure is repeated k times.

Wild binary segmentation Local, alternating jumps may go undetected when calculating the CUSUM statistic $\tilde{y}^b_{s,e}$ over longer segments. Motivated by this, wild binary segmentation (WBS) due to Fryzlewicz (2014*b*) modifies binary segmentation to calculate $\tilde{y}^b_{s,e}$ in randomly drawn subsegments of the data. Denote by $w = \{w_1, \cdots, w_B\} = \{(s_1 : e_1), \cdots, (s_B : e_B)\}$ a set of B randomly drawn intervals with endpoints between 1 and n.

Given data $y \in \mathbb{R}^n$, the k-step WBS algorithm proceeds as follows. Define $I_1 = 1 : B$ and I_i to be the subset of $(1 : B)$ such that none of $\hat{b}_{1:i}(y)$ are contained in w_j. Then, at step $i+1$, the next changepoint $\hat{b}_{i+1}(y)$ and the corresponding interval index $\hat{i}_{i+1}(y)$ are obtained by solving the following maximization problem:

$$(\hat{i}_{i+1}, \hat{b}_{i+1}) = \underset{\substack{j \in I_{I_i} \\ b \in w_j}}{\operatorname{argmax}} |\tilde{y}^b_{s_j,e_j}|. \tag{2}$$

As before, the direction of the jump $\hat{s}_{i+1}(y)$ is calculated by the sign of the maximizing absolute CUSUM statistic $\hat{s}_{i+1}(y) = \operatorname{sign}(\tilde{y}^{\hat{b}_{i+1}(y)}_{s_j,e_j})|_{j=\hat{i}_{i+1}(y)}$. This procedure is repeated for k steps.

Circular binary Segmentation Circular binary segmentation (CBS) due to Olshen et al. (2004*b*) specializes in detecting pairs of changepoints of alternating directions and of the same magnitude. In CBS, *pairs* of splits are made according to a modified CUSUM-type criterion which is the variance-stabilized difference in the sample means between a middle subsegment and the rest. The modified criterion is:

$$\tilde{\tilde{y}}^{a,b}_{s,e} = \sqrt{\frac{1}{\frac{1}{|b-a|} + \frac{1}{|e-s-b+a|}}}(\bar{y}_{(s+1):e} - \bar{y}_{(a+1):b}). \tag{3}$$

which, like $\tilde{y}$, has the same scale-invariance property. Denote by $\hat{b}^{\text{start}}_i(y)$ and $\hat{b}^{\text{end}}_i(y)$ the pair of changepoints at step i, and by $P_1, \cdots, P_{2i+1}$ the set of indices obtained by partitioning at the combined set of $2i$ changepoints $b^{\text{start}}_{1:i} \cup b^{\text{end}}_{1:i}$, similarly as in binary segmentation. (Specifically, $I_j = (c_{j-1} + 1) : c_j$ is defined in terms of the sorted permutation $0 = c_0 < c_1 < \cdots < c_{(2i+1)} = n$ of changepoints.) Then, at step $i + 1$, the next changepoint pair $\hat{b}^{\text{start}}_{i+1}$ and $\hat{b}^{\text{end}}_{i+1}$ and the maximizing partition $\hat{j}_{i+1}(y)$ are found by solving the following maximization problem:

$$(\hat{b}^{\text{start}}_{i+1}(y), \hat{b}^{\text{end}}_{i+1}(y), \hat{j}_{i+1}(y)) = \underset{\substack{j \in (1:(2i+1)) \\ a,b \in I_j}}{\operatorname{argmax}} |\tilde{\tilde{y}}^{a,b}_{c_{j-1}+1,c_j}|. \tag{4}$$

Additionally, the sign $\hat{s}_{i+1}$ is defined as the sign of the maximizing CUSUM statistic $\operatorname{sign}(\tilde{\tilde{y}}^{\hat{b}^{\text{start}}_{i+1}, b^{\text{end}}_{i+1}}_{c_{j-1}+1,c_j})|_{j=\hat{j}_{i+1}(y)}$ – the corresponding changepoint directions at $\hat{b}^{\text{start}}_{i+1}$ and $\hat{b}^{\text{end}}_{i+1}$ are ± 1 times this sign. Continue until k steps are completed.

Fused lasso The *1d fused lasso* due to (Tibshirani et al. 2005) – also known as *1d total variation denoising* in signal processing (Rudin et al. 1992) solves the least squares problem but with an ℓ_1 penalty on the successive differences:

$$\hat{\beta} = \min_{\beta \in \mathbb{R}^n} \|y - \beta\|_2^2 + \lambda \|D\beta\|_1, \quad D^{(1)} = \begin{bmatrix} -1 & 1 & 0 & \dots & 0 \\ 0 & -1 & 1 & \dots & 0 \\ \vdots & & \ddots & \ddots & \\ 0 & 0 & \dots & -1 & 1 \end{bmatrix}. \tag{5}$$

The generalized lasso dual path algorithm from Tibshirani & Taylor (2011) can be used to recover a piecewise linear path of $\hat{\beta}$ solving (5) over a range of $\lambda \in [0, \infty)$, with knot locations at $\lambda_1 \geqslant \cdots \geqslant 0$. The primal solution $\hat{\beta}$ is piecewise constant, with a single additional changepoint occurring in the solution at each knot. Hyun et al. (2017) develops post-selection inference for this algorithm. In the current paper, we extend this framework by incorporating additive-noise randomization and make comparisons to the post-selection inferences made with segmentation algorithms.

Recursive vs. sequential segmentation Segmentation algorithms in the literature are often defined to proceed recursively. In this paper, we take the variant of segmentation algorithms which sequentially splits the data instead of an ordinary recursion. For example a CUSUM statistic is sequentially maximized in *all* latest partitions to obtain a nested sequence of changepoint sets (models). This modification grants the advantage of being able to incorporate a data-dependent stopping rule which allows for valid post-selection inference, as will be discussed in Section 2.3. The unmodified, original recursive algorithm is also polyhedral, but requires a stopping rule in terms of a minimum threshold for the maximizing CUSUM each time a split occurs. There is no known polyhedral rule for choosing this threshold, which makes it less desirable than the fully data-dependent sequential variant.

1.3 Changepoint inference

1.3.1 Existing work

Inference methodology regarding changepoint models found in the literature focuses on the location or jump size of changepoints, or segment lengths, or goodness of fits among sequences of models. Some earlier representative works are from Hinkley (1970), Worsley (1986), Bai (1999). Jandhyala et al. (2013), Horvath & Rice (2014) collectively provide a nice survey of methods for changepoint estimation. There is also a body of Bayesian changepoint detection methods whose inference regarding changepoint locations can be calculated directly from posterior distributions (Yao 1984, Barry & Hartigan 1993, Yao 1993, Chib & Olin 1997, Tartakovsky & Moustakides 2010).

The most directly relevant tools for inference regarding changepoint locations is the simultaneous multiscale change-point estimator (SMUCE) estimator in Frick et al. (2014) and the generalized lasso post-selection inference tools in Hyun et al. (2017). The former produces confidence bands around the changepoint signal with exact simultaneous frequentist coverage. The latter enables the basic case post-selection inference tools after applying the generalized lasso path algorithm, of which a subcase is fused lasso. A simple transformation can be made to the SMUCE bands to make conservative confidence interval of linear contrasts of the mean – this was explored in Hyun et al. (2017) in comparison to fused lasso post-selection tests.

Another typical inference scheme in multiple changepoint detection is to conduct a sequence of likelihood ratio tests to admit changepoints until the first failure to reject. This scheme is used in the original binary segmentation methodology from Vostrikova (1981) as well as in Olshen et al. (2004*b*) for CBS. When each test is conducted at a prescribed level α, it is unclear what the aggregate control is among multiple tests. Bonferonni correction can be used for control of the global null, but only is powerful when a few strong signals exist. Otherwise each test is quite conservative, making it unsuitable when data is thought to have many changepoints. Most importantly, it is not clear whether the inference is always valid – each test is conducted in subsegments of the data chosen adaptively from results of subsequent tests.

1.3.2 Setup for changepoint inference framework

In the current paper, we discuss tools for inference after segmentation algorithms have been used to detect to changepoints. We give a brief setup of the methods in this paper. For the majority of the paper, we assume that the data $y \in \mathbb{R}^n$ is multivariate Gaussian around some mean $\theta \in \mathbb{R}^n$,

$$y \sim \mathcal{N}(\theta, \Sigma), \quad y, \theta \in \mathbb{R}^n, \quad \Sigma \in \mathcal{S}_+^n. \tag{6}$$

We can use a general positive semi-definite Σ, but for ease of application and explanation of the post-selection inference tools of Lee et al. (2016) and Tibshirani et al. (2016), we will assume that the covariance is $\Sigma = \sigma^2 I_n$ and can be summarized by a one dimensional noise parameter $\sigma^2 \in \mathbb{R}$.

We now desribe the entire procedure, focusing on binary segmentation for concreteness. First, a sequential binary segmentation algorithm is applied on the data on hand y_{obs} to recover a changepoint set $\hat{b}_{1:k}(y_{\text{obs}})$ and directions $\hat{d}_{1:k}(y_{\text{obs}})$, which jointly represent the outcomes of binary segmentation on y_{obs}.

From these, you can form a contrast vector $v \in \mathbb{R}^n$ whose linear contrast with the mean $v^T\theta$ is a meaningful parameter regarding detected changepoints. Now denote the selection space as $P_{\text{obs}} = \{z \in \mathbb{R}^n : \hat{b}(z) = \hat{b}(y_{\text{obs}}), \hat{d}(z) = \hat{d}(y_{\text{obs}})\}$. The two works Tibshirani et al. (2016) and Lee et al. (2016) make it possible to calculate a statistic $T(y_{\text{obs}})$ that serves as a p-value for the null hypothesis $H_0 : v^T\mu = 0$ with valid *selective* type-I error control,

$$P_{H_0}\left(T(y_{\text{obs}}) < \alpha | y \in P_{\text{obs}}\right) < \alpha, \tag{7}$$

for a preset significance level α. We can also invert these tests to obtain a *selective* confidence interval $C_{1-\alpha}$ that covers $v^T\mu$,

$$P\left(v^T\mu \in C_{1-\alpha} | y \in P_{\text{obs}}\right) = 1 - \alpha. \tag{8}$$

The overall inference procedure is as follows:

1. Observe data $y_{\text{obs}} \in \mathbb{R}^n$.
2. Obtain changepoint set $\hat{b}_{1:k}(y_{\text{obs}})$ and changepoint directions $\hat{d}_{1:k}(y_{\text{obs}})$.
3. Form contrast vectors $v_1, \cdots, v_{k_0} \in \mathbb{R}^n$ to use for inference. (The simplest case with no post-processing, $k_0 = k$, is to test all detected changepoints.)
4. Calculate $T(y_{\text{obs}}, v_i), i = 1, \cdots, k_0$ for p-values to the null hypotheses $H_0 : v_i^T\mu = 0$ against α/k_0, or compute $1 - \alpha/(2k_0)$ confidence interval covering $v_i^T\mu$.

In the last step, $T(y_{\text{obs}}, v_i)$ can be calculated under two different types of model assumptions – one that assumes a full dimensional mean vector θ but conditions on the rest of the structure, and one that assumes the latest underlying selected model. These are each called a saturated model and a selected model, whose details and implications are discussed in Section 2.

The current paper extends the saturated model post-selection inference framework to the changepoint detection setting, to use with segmentation algorithms. We characterize the selection space P_{obs} for several segmentation algorithms and devise sampling strategies for conducting inferences that incorporate randomization of Tian & Taylor (2015), and operate under the selected model. A useful contrast vector $v_i, i = 1, \cdots, k_0$ is an n-length vector onto which the mean θ can be projected to form the sample mean difference in two segments directly adjacent to a changepoint location of interest. Thus, $v_i^T\theta$ can represent a *shift* in mean level, from left to right, at a given changepoint location of interest. A careful choice of contrast v_i in step 3 formed after post-processing the detected changepoint set, or one that fully exploits the selection event, can result in more powerful inference – this will be discussed in Section 2.

2 Post selection inference for changepoint problems

2.1 General case

Here, we describe several ingredients for post-selection inference for changepoint inference, keeping the description general at this stage. Section 2.2 describes useful extensions to this general description, and 2.3 describes stopping rules and some other practical considerations when using these tools.

Linear test contrasts After having applied a changepoint algorithm for k steps to obtain changepoint locations $\hat{b}_{1:k}(y)$ and $\hat{d}_{1:k}(y)$ and after some post-processing (discussed in 2.2), the resulting "pruned" set of locations are $b_1, \cdots, b_{k_0}$ with corresponding directions $s_1, \cdots, s_{k_0}$. With this pruned set, one useful class of tests are *segment* tests, as coined in Hyun et al. (2017), which are about the difference in adjacent segment averages. Taking $c_{1:k_0}$ to be the increasing, sorted values of $b_1, \cdots, b_{k_0}$ with $c_0 = 0$ and $c_{k_0+1} = n$, a segment test contrast for testing location c_i is:

$$v_i^T\theta = d_i \left(\frac{1}{c_{i+1} - c_i} \sum_{i=c_i+1}^{c_{i+1}} \theta_i - \frac{1}{c_i - c_{i-1} + 1} \sum_{i=c_{i-1}+1}^{c_{i+1}} \theta_i \right) \tag{9}$$

which is the difference in the segment means immediately to the left and right of c_i, in the appropriate direction $d_i \in \{-1, 1\}$. We test the null hypothesis:

$$H_0 : v_i^T \theta = 0 \tag{10}$$

against the one sided alternative $H_0 : v_i^T \theta < 0$ or the two sided $H_0 : v_i^T \theta = 0$. If the sign s_i is not available, we can set $s_i = 1$ and conduct a two-sided test instead. When available, incorporating s_i and carrying out a one-sided test yields greater power.

Polyhedral selection spaces We can also formally characterize the selection space P_{obs} of $y \in \mathbb{R}^n$ that would result in the same segmentation selection event that of y_{obs} on hand. The following lemma states that the selection space P_{obs} for three segmentation algorithms can be represented as polyhedra:

Lemma 2.1. *In the following three cases, selection events are polyhedral sets in* $\mathbb{R}^n$*:*

Case 1. *the k-step binary segmentation selection event:*

$$P_{obs} = \{y : \hat{b}_{1:k}(y) = \hat{b}_{1:k}(y_{obs}), \hat{s}_{1:k}(y) = \hat{d}_{1:k}(y_{obs})\}$$

Case 2. *the k-step wild binary segmentation selection event*

$$P_{obs} = \{y : \hat{b}_{1:k}(y) = \hat{b}_{1:k}(y_{obs}), \hat{d}_{1:k}(y) = \hat{d}_{1:k}(y_{obs}), \hat{e}_{1:k}(y) = \hat{e}_{1:k}(y_{obs})\}, \tag{11}$$

Case 3. *and the k/2-step circular binary segmentation selection event*

$$P_{obs} = \{y : \hat{b}^{start}_{1:k}(y) = \hat{b}^{start}_{1:k}(y_{obs}), b^{end}_{1:k}(y) = \hat{b}^{end}_{1:k}(y_{obs}), \hat{d}_{1:k}(y) = \hat{d}_{1:k}(y_{obs})\}. \tag{12}$$

For the WBS selection event, $\hat{s}_i(y)$ and $\hat{e}_i(y)$ are the endpoints of the intervals in which the CUSUM statistic is maximized at step i, for $i = 1, \cdots, k$. For CBS, $\hat{b}_i^{start}(y)$ and $\hat{b}_i^{end}(y)$ denote the two breakpoints occurring at step i, and $\hat{s}_i(y)$ represents the direction ($+1$ for up-down, and -1 for down-up). The proofs for each algorithm are shown in appendix section B. Each proof follows the sequence of algorithm events and characterizes halfspaces for each event, inductively reasoning that the intersection of such spaces is the *exact* space of selection.

Selective distributions and p-values Having characterized P_{obs} and with a chosen test contrast vector v, the existing post-selection inference literature suggests carrying out inference out about $v^T\theta$ under a *selective* distribution instead of the naive, marginal distribution, in order to explicitly account for the previous selection based on the data. First consider the most basic selective distribution,

$$v^T Y \;\mid\; Y \in P_{\text{obs}}, \tag{13}$$

under which the contrast vector v is measurable, i.e. it is nonrandom assuming the same selection event as the one observed with y_{obs}.

In order to make the inference tractable, additional conditioning is needed beyond (13), which can be formulated in terms of two different null models, or null hypotheses. Closely following the ideas and terminology in Fithian et al. (2014, 2015), we first consider the *saturated* data model, which assumes a Gaussian model with full-dimensional mean θ but fixes a $n-1$ dimensional component in the parametrization by conditioning on $P_v^\perp Y$. Under the saturated model, additional conditioning on the $n-1$ dimensional orthogonal slice $P_v^\perp y$ is required to cover the nuisance parameter $P_v^\perp \theta$ to $v^T\theta$:

$$v^T Y \;\mid\; Y \in P_{\text{obs}},\; P_v^\perp Y = P_v^\perp y_{\text{obs}}. \tag{14}$$

In this distribution and under the null $H_0 : v^T\theta = 0$ of (10), we calculate the probability of $v^T Y$ exceeding its observed value $v^T y_{\text{obs}}$ to form a statistic $T(y, v)$ that serves as the p-value,

$$T(y, v) = \int \mathbb{1}(v^T y > v^T y_{\text{obs}})\, dP_{\mathbb{F}}, \tag{15}$$

This inference tool was developed in Lee et al. (2016) and Tibshirani et al. (2016), coined the *truncated Gaussian* (TG) statistic. The novelty of this work is in enabling the fast calculation of the p-value, by exploiting properties of a multivariate Gaussian such as independence of orthogonal components.

2.2 Modifications to the general case

Selected model inference Continue with the example of testing about a detected location c_i using a contrast vector $v_i \in \mathbb{R}^n$. Let $\tilde{C} = \{c_0, \ldots, c_{k_0+1}\} \backslash \{c_i\} = \{\tilde{c}_0, \ldots, \tilde{c}_{k_0}\}$, where $\tilde{c}_{0:k}$ are in sorted increasing order. Another null model we might operate under is that of a *selected model*, which assumes Gaussian data with a piecewise constant mean, i.e. θ lies in a smaller subspace in which entries are constant in each of the segments broken at locations in $\tilde{C}$,

$$\theta_{\tilde{c}_{j-1}+1} = \ldots = \theta_{\tilde{c}_j} = \mu_{k_0} \quad \text{for all } j \in \{1, \ldots, k_0\}. \tag{16}$$

The key idea shared by the two approaches – saturated and selected model testing – is in the handling of the nuisance parameters in their respective parametrizations, with respect to the pivotal. statistic. The statistic we use is the tail probability of $v^T y$ is a conditional pivotal (i.e. it does not depend on the unknown parametrization) only when the nuisance parameters are *covered*[1]. In a saturated model (where it is assumed σ is known), the nuisance is the $P_v^{\perp}\theta$. In a selected model (where σ is not assumed to be known), the nuisance parameters are all mean levels $\mu_i, i = 1, \cdots, k_0$ in each segment and the noise parameter σ, whose sufficient statistics are,

$$\{\bar{Y}_{(\tilde{c}_{j-1}+1):\tilde{c}_j} : j = 1, \cdots, k_0\}, \quad \|Y\|_2. \tag{17}$$

A closely related distinction can be made from the view point of the different null hypotheses (models). The saturated model null hypothesis of a full-dimensional mean is fairly general harder to reject than the selected model null. The tradeoff also comes in the form of stricter conditioning – a full $n-1$ dimensional slice in the $\mathbb{R}^n$ mean parameter – for tractable inference of $v^T\theta$. The selected null hypothesis asserts the correctness of a piecewise constant mean with all detected breaks excluding the one being tested. Because this is a more specific scenario than the saturated model null scenario, it may be easier to reject (i.e. powerful).

Hit-and-run sampler for selected model inference The null selective distribution of $Y \sim \mathcal{N}\left(\theta, \sigma^2 I\right)$ conditioned on

$$Y | \bar{Y}_{(\tilde{c}_{j-1}+1):\tilde{c}_j} = \bar{y}_{\text{obs},(\tilde{c}_{j-1}+1):\tilde{c}_j} \text{ for } j \in \{1, \ldots, k\}, \ \|Y\|_2 = \|y_{\text{obs}}\|_2, \text{ and } Y \in P_{\text{obs}}. \tag{18}$$

To form a p-value, we use a Monte-Carlo approach to simulate the null distribution of $v_i^T Y$ for $Y \sim \mathcal{N}\left(\theta, \sigma^2 I\right)$ conditioned on (18). That is, we first generate B samples $y^{(1)}, \ldots, y^{(B)} \in \mathbb{R}^n$ from this null distribution. Then, we calculate the p-value $T(y_{obs}, v_i)$ as in (15) for the test in (10) by calculating the empirical probability mass exceeding the value $v_i^T y_{\text{obs}}$ based on the sample $v_i^T y^{(1)}, \ldots, v_i^T y^{(B)}$.

To implement the Monte-Carlo sampler, we use a hit-and-run approach. We describe this approach briefly, with more details in APPENDIX. Observe that sampling $Y \sim \mathcal{N}\left(\theta, \sigma^2 I\right)$ conditioned on (18) is equivalent to sampling Y uniformly from the set

$$\left\{Y : \bar{Y}_{(\tilde{c}_{j-1}+1):\tilde{c}_j} = \bar{y}_{\text{obs},(\tilde{c}_{j-1}+1):\tilde{c}_j} \text{ for } j \in \{1, \ldots, k\}, \ \|Y\|_2 = \|y_{\text{obs}}\|_2, \text{ and } Y \in P_{\text{obs}}\right\}.$$

Hence, starting with $y^{(1)} = y_{\text{obs}}$, for iteration m, we uniformly select a random 2-dimensional slice of the high-dimensional sphere $\{Y : \bar{Y}_{(\tilde{c}_{j-1}+1):\tilde{c}_j} = \bar{y}_{\text{obs},(\tilde{c}_{j-1}+1):\tilde{c}_j}$ for $j \in \{1, \ldots, k\}$, $\|Y\|_2 = \|y_{\text{obs}}\|_2\}$ that passes through $y^{(m)}$. We can then explicitly sample $y^{(m+1)}$ uniformly from the region(s) of this 2-dimensional circle that lies within the polyhedra P_{obs}.

Randomization and Importance Sampling Tian & Taylor (2015) proposed the idea of randomization of the selective model distributions for improved numeric stability and for higher power. This is typically in the form of relaxing parts of the conditioning, or by choosing to condition on the model selection event after some controlled obfuscation. One example of this is *additive noise* randomization, which can be used with polyhedral sequential changepoint

[1] In an exponential family formulation, this can be done by conditioning on the sufficient statistics of the nuisance parameters.

algorithms – including the segmentation algorithms discussed in the current work and the 1d fused lasso path algorithm – described next.

First, another source of auxiliary variation is introduced – for additive noise randomization, an n-variate W distributed as $\mathcal{N}(0, \sigma^2_{\text{add}} I_n)$ – whose generating mechanism is under full control of the user. We draw a realization w_{obs} from W. Then inference about $v^T\theta$ can be carried out under the *randomized* selective distribution of (21),

$$Y|(Y+W) \in P_{\text{noisy}}, P_v^\perp Y = P_v^\perp y_{\text{obs}}, W = w_{\text{obs}}, \tag{19}$$

which conditions on the polyhedral space P_{noisy} which would lead to the same selection event based on the *obfuscated* data $Y + W$,

$$P_{\text{noisy}} = \{y : \hat{b}_{1:k}(Y+W) = \hat{b}_{1:k}(y_{\text{obs}} + w_{\text{obs}}), \hat{s}_{1:k}(Y+W) = \hat{s}_{1:k}(y_{\text{obs}} + w_{\text{obs}})\} \tag{20}$$

One valuable goal from randomization is to integrate out W so that (21) does not condition on the realization of the additive noise $W = w_{\text{obs}}$,

$$Y|(Y+W) \in P_{\text{noisy}}, P_v^\perp Y = P_v^\perp y_{\text{obs}}. \tag{21}$$

From this, we can calculate a functional to use for inference, such as the p-value in (15). Marginalizing out w leads to a strict increase in Fisher information, corresponding to an increase in power. This is verified through simulations.

We can use importance sampling to calculate a p-value under (21). Denote by $M_{\text{obs}} = (s_{1:k}, b_{1:k})$ the observed a k-step model. Also denote by $E_1 = \{Y|(Y+W) \in P_{\text{noisy}}\}$ the model selection event [2], and by $E_2 = \{P_v^\perp Y = P_v^\perp y_{\text{obs}}\}$ the $n-1$ dimensional orthogonal projection. Denote the k-step model $M_{\text{obs}} = (s_{1:k}, b_{1:k})$, the selection event $E_1 = \{\hat{M}(Y, W) = M_{\text{obs}}\}$, and the event of observing the orthogonal projection $E_2 = \{P_v^\perp Y = P_v^\perp y_{\text{obs}}\}$. The final p-value is calculated as,

$$P(v^T Y \geqslant v^T y_{\text{obs}} | E_1, E_2) = \int P(v^T Y \geqslant v^T y_{\text{obs}} | E_1, E_2, W = w) p_{W|E_1,E_2}(w) dw. \tag{22}$$

We calculate this using importance sampling, sampling from $W|E_2$ (assumed to be identical in distribution to W), and reweighting by

$$\frac{P(W|E_1, E_2)}{P(W|E_2)} = P(E_1|W = w, E_2)/P(E_1|E_2). \tag{23}$$

This importance sampling approach can be modified for WBS to relax conditioning on the randomly drawn intervals to improve power. Define the random variable $W = (W_1, \cdots, W_B), W_i = (W_i^1, W_i^2) \in \mathbb{R}^2$ to be the set of B endpoints for WBS intervals $(W_i^1 : W_i^2) \subseteq \{1, 2, \cdots, n\}$. Also adopt the notation of $W_A := \{W_i : i \in A\}$. We also define $E_0^{1:k}, E_1^{1:k}$ to replace E_0 and E_1 above. The set $E_0^{1:k} = \{W_{1:k} = w_{1:k}^{\text{obs}}\}$ characterize the realizations of the first k intervals. The set $E_1^{1:k}$ describes the sequence of k events $E_1^i, i = 1, \cdots$ in which, for each $i \in (1:k)$, breakpoint b_i in W_i maximizes the CUSUM $\tilde{y}^b_{W_i^1, W_i^2}$ out of all qualifying intervals. (Qualifying intervals at step $i > 1$ are ones that do not intersect with $b_{1:i-1}$, and for $i = 1$, all intervals $W = w_{\text{obs}}$ are qualified.) Then the above sampling scheme can be applied after replacing E_1 with $E_0^{1:k} \cap E_1^{1:k}$.

As with nonrandomized p-values, these p-values can be inverted to obtain confidence intervals that cover $v^T\theta$ after selection. Because calculating randomized TG p-values is computationally expensive than ordinary ones, it is recommended that the confidence interval endpoints are found using an efficient optimization than grid search, like binary search.

[2] Notice this space, or event, is random because W is random. It can be thought of as a wobbly polyhedron, shifted in a random direction $P_{\text{noisy}} W$.

2.3 Practicalities

For saturated model tests, the noise level σ is assumed to be known. In practice, it needs to be estimated. One example is to fit a low-bias or undersmoothed model, estimate the sample standard deviation from this relatively complex model, and then substitute in σ for the inference. In the case of copy number variation dat, there are typically long flat regions near either end of the data set. Parts of these regions could be excluded from changepoint detection and inference, and instead used to estimate the standard devation.

So far we have assumed that the number of segmentation steps k is fixed. Hyun et al. (2017) introduces a stopping rule based on information criteria (IC) wihich can be characterized as a polyhedral set and conditioned upon. First consider a sequence of nested *changepoint models* represented by increasing sets of changepoints $M_k = b_{1:k}$ (and $M_0 = \varnothing$) in the piecewise constant mean. The information criteria for this model is

$$J(M_k) = \|y - g_k(y)\| + p_n(k), \quad \hat{S}_k(y) = \text{sign}(J(M_k) - J(M_{k-1})), \tag{24}$$

where $g_i(y)$ is the projection of y onto a piecewise linear vector subspace with breaks at $b_{1:k}$. We will use $p_n(k) = \sigma^2 \cdot k \cdot \log(n)$, which resembles the Bayesian Information Criterion (BIC) for fixed changepoint models. The penalty term is proportional to the complexity of the changepoint model. Now define S_k, the sign of the difference between two steps k and $k+1$. The stoppping rule k is defined as

$$\hat{k}(y) = \min\{k : \hat{S}_k(y) = \hat{S}_{k+1}(y) = \cdots = \hat{S}_{k+q}(y) = 1\} \tag{25}$$

which is a local minimization of IC – the first time there are q consecutive rises in IC. As discussed in Hyun et al. (2017), $q = 2$ is a reasonable choice for the 1d changepoint detection. To carry out valid selective inference, we condition on the sequence $S_{1:k}(y)$, which is enough to determine $\hat{k}$.

Additional care is required when using this stopping rule in conjuction with randomized inference extension of Section 2.2. After introducing W, the conditioning event becomes $\{\hat{S}_{1:(k+q)}(Y, W) = S_{1:(k+q)}\}$, which is polyhedral in Y only when W is fixed. The importance sampling for obtaining saturated model p-values can be modified by intersecting the model selection events – E_1 in BS or $E_0 \cap E_1$ in WBS – with these new halfspaces. Then, during sampling, the expectations of $\mathbb{1}(v^T Y > v^T y)$ are calculated conditional on this new, smaller polyhedron.

3 Array CGH data changepoint inference

We now illustrate the application of this methodology to an array CGH dataset. Figure 1 shows an array CGH dataset of fibroblast cell line GM05296 originally published in Snijders et al. (2001). The data consists of a sequence of $n = 2011$ measurements from 23 chromosomes. This example appeared in the introduction; we describe the procedure in more detail here.

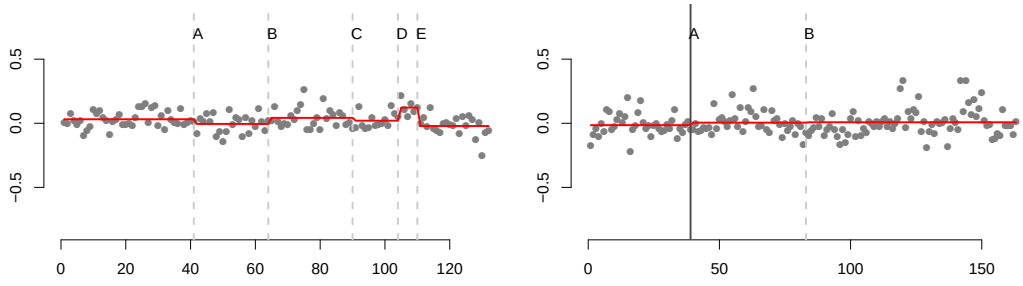
Prior to applying the methodology, we exclude three outlier points, as well as the first 200 points which are used to estimate the noise standard deviation ($\hat{\sigma} = 0.74$). We then apply additive noise binary segmentation with our BIC-based stopping rule ($q = 2$) resulting in $\hat{k} = 13$ steps. We construct segment test contrasts $\{v_i, i = 1, \cdots, 13\}$ for the detected changepoints $b_{1:k}$ as in (9) and conduct one-sided post-selection tests of the saturated null hypotheses $H_0 : v_i^T \theta = 0$. The corresponding post-selection p-values are shown in the bottom row of the table in Figure 1, and the naive Z-test p-values for the same hypotheses are shown in the top row. The first five locations A through E were deemed significant against a Bonferonni-corrected cutoff of 0.05/13, while all other changepoints were not deemed significant. The naive two-sided Z-tests on the same locations using the same contrasts $v_i, i = 1, \cdots, 13$ deemed twelve out of thirteen locations significant. The four significant jumps detected by post-selection tests occur in the two chromosomes that have known variation in copy number (Olshen et al. 2004*a*).

We also analyzed data from individual chromosomes, in both the GM05296 data and in another fibroblast cell line GM03563. Out of the 23 chromosomes in GM05296 (top two rows of Figure 2), only chromosome 10 and 11 have known gain and loss patterns. We apply a similar analysis steps as before to data from chromosomes 1, 4, 10 and 11, as analyzed in Olshen et al.

(2004*a*). We post-processed the chosen $\hat{k}$ changepoints using centroid-clustering of locations within 5 distance of each other; there was no need for post-processing the analysis in Figure 1. In each of chromosomes 10 and 11, we see that two detected jumps are deemed significant (after Bonferroni correction), in similar locations to the four significant locations from the larger dataset in Figure 1. These locations also match closely with those detected on the same data by Olshen et al. (2004*a*) who used the CBS algorithm combined with sequential likelihood ratio tests for stopping. There is one significant jump location (D) in chromosome 4 that appears to be a false discovery. Repeating this analysis in GM03563 fibroblast cell line data in chromosomes 1, 3, 9 and 11, we found no jumps in chromosome 1, and correctly found all jumps in the remaining three. As before, there is one jump (B) in chromosome 9 that may be spurious.

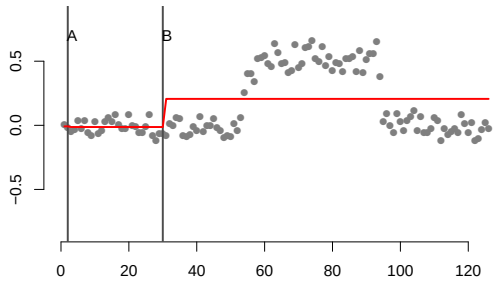
In the next section, we conduct simulation studies on synthetic and pseudo-real simulations to demonstrate the inferential properties of our proposed tools.

Coriell Cell Line GBM 05296

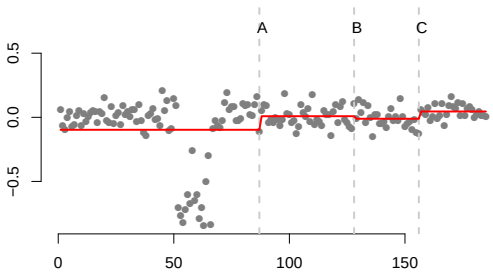


	A
location	129
pvalue	0.65

	A	B	C	D	E
location	47	-72	118	-126	141
p-values	0.78	0.74	0.10	0.01	0.17

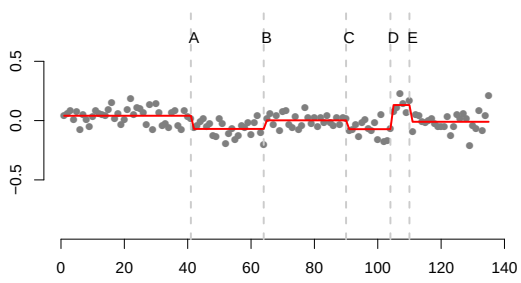


	A	B
location	53	-94
p-value	0.00	0.00

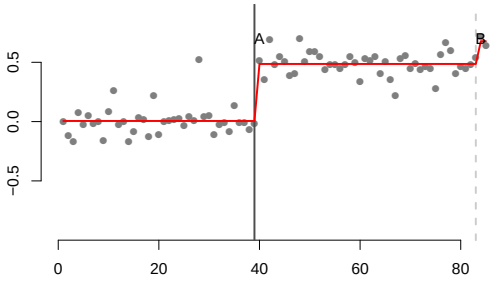


	A	B	C	D	E
location	-51	66	-86	-149	156
p-value	0.00	0.00	0.36	0.90	0.91

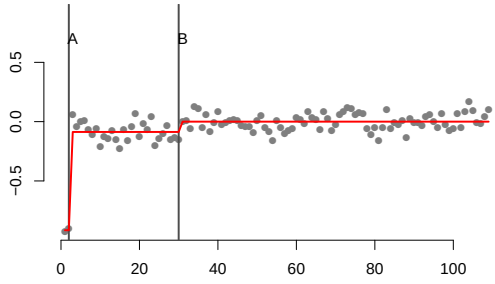
Coriell Cell Line GBM 03563



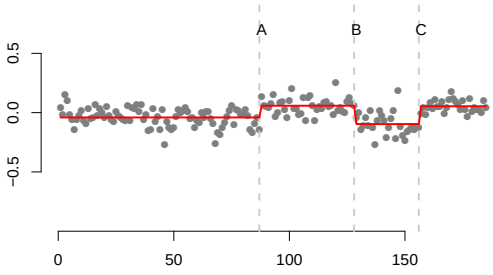
	A	B	C	D	E
locs	-41	64	-90	104	-110
pvs	0.14	0.65	0.68	0.45	0.50



	A	B
locs	39	83
pvs	0.00	0.95



	A	B
locs	2	30
pvs	0.00	0.00



	A	B	C
locs	87	-128	156
pvs	0.82	0.13	0.05

Figure 2: *Post-selection segment tests were applied to changepoint locations detected by WBS, in four chromosomes each from two observed array CGH datasets of fibroblast cell line GM05296 (top two rows) and of GM03536 (bottom two rows). The detected locations for which the post-selection tests were not significant vertical lines are shown in light-grey, dashed lines. The locations that were significant are shown in dark-grey, solid lines. The corresponding p-values are shown in the tables below each figure.*

4 Simulations

4.1 Synthetic simulations

In this section, we show simulation examples to demonstrate properties of the segmentation post-selection inference tools presented in the current paper. We will vary the signal size, denoted as δ, while generating data from a fixed noise level $\sigma = 1$. The main synthetic simulation setting consists of four alternating direction jumps in an otherwise constant mean:

$$y_i \sim \mathcal{N}(\theta_i, \sigma^2), \quad \theta_i = \begin{cases} \delta & \text{if } 41 \leqslant i \leqslant 80 \\ -2\delta & \text{if } 121 \leqslant i \leqslant 160 \\ 0 & \text{if } i \in (1:40) \cup (81:120) \cup 161:200 \end{cases} \tag{26}$$

The sample size $n = 200$ was chosen because it is in the scale of the data length in a typical array CGH dataset in a single chromosome. An example of this synthetic dataset can be seen in figure 3.

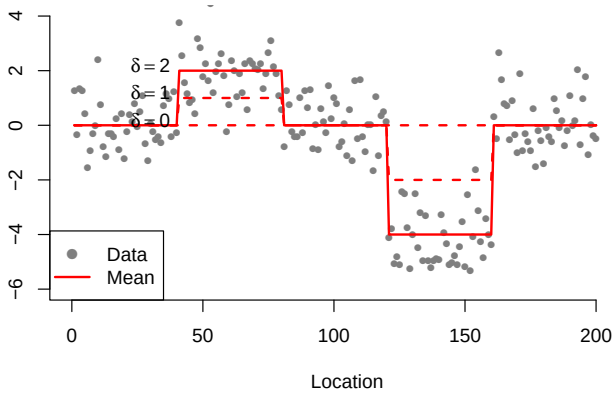


Figure 3: *Example realization for simulation* (26) *with* $\delta = 2$.

Type-I error control verification We consider simulations in the no-signal scenario ($\delta = 0$) to verify the expected Type-I error control of the proposed post-selection tests. The results are shown in Appendix A. For all of the proposed methods, the p-values are seen to be uniformly distributed under $\delta = 0$ at all steps in the model.

Power comparison by simulation Two signal size regimes are interesting. The no-signal regime $\delta = 0$ can be used to examine the validity (type I error control) of the inference; when there is no signal the null scenario $v^T\theta = 0$ is true so that Unif$(0, 1)$ distributed p-values are expected in simulation. This is shown in appendix A.

Simulations are carried out across a range of signal strengths δ to demonstrate and compare the power of the proposed tests. Because these tests are carried out only when a jump is selected, it is necessary to separate the effects of detection by a segmentation method, from test power. To that end, we define the following quantities:

$$\text{Detection} = \frac{\#\text{correctly detected}}{\#\text{simulations}} \tag{27}$$

$$\text{Unconditional power} = \frac{\#\text{correctly detected and rejected}}{\#\text{correctly detected}} \tag{28}$$

$$\text{Conditional power} = \frac{\#\text{rejections}}{\#\text{simulations}} \tag{29}$$

The overall power of an inference tool can only be assessed by examining the conditional and unconditional power in conjunction.

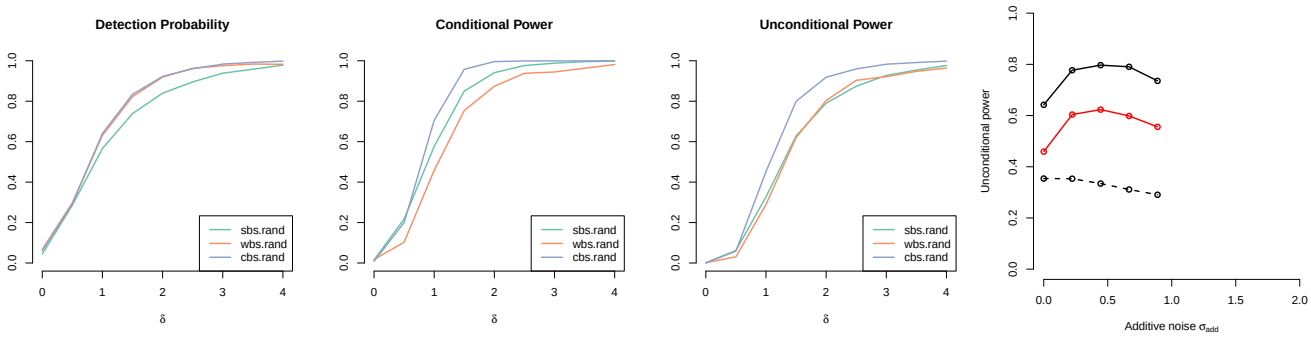


Figure 4: *Data was simulated from a four-jump mean as in* (26)*, over* $\delta \in (0, 4)$ *with* $n = 200$ *data points. Several four-step algorithms (WBS, SBS, CBS, FL) were applied, and post-selection segment test inference was conducted on the resulting four detected changepoints from each method. The detection proportions show out of all detected changepoints overall, the proportion of ones that were within* ± 2 *proximity of each true jump location* $(40, 80, 120, 160)$ *in the mean. The conditional power shows the proportion of p-values that are below* $\alpha/4 = 0.0125$*, out of the those that test the approximately correct detected changepoints. The unconditional power is the proportion of these p-values over all simulated p-values. For randomization,* $n_I = n = 200$ *for WBS was used, and an additive noise of* $\sigma_{add} = 0.2$ *was used for the rest. The rightmost plot shows the detection, unconditional and conditional power of noise-added BS inference, for simulations conducted across a range of* $\sigma_{add} = 0, 1, 2, 3, 4$*, for four-jump data with noise* $\sigma = 1$ *and signal size* $\delta = 1$*.*

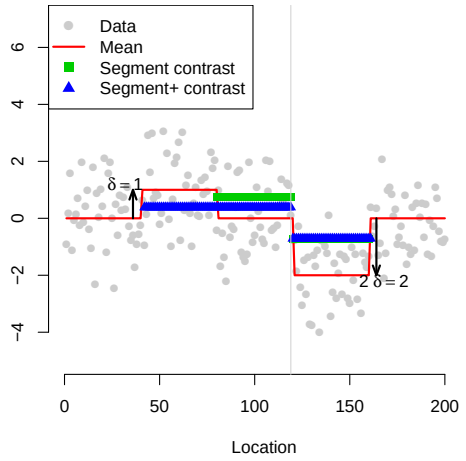
We examine the performance of four methods – binary segmentation (BS), 1 dimensional fused lasso (FL), wild binary segmentation (WBS) and circular binary segmentation (CBS). For BS, FL, and CBS, we use the randomized noise-added methods using Gaussian noise with standard deviation $\sigma_{\text{add}} = 0.2$. For WBS, we employ the randomization scheme as described in Section 2.2 and in Appendix C. The left panel of 4 shows the detection ability of three methods WBS and CBS and BS. Detection ability was calculated as the average fraction of changepoitns that were detected within three indices of the true locations. Power was calculated as in (27), but with approximate detection – within two locations of the correct changepoints in the mean – instead of exact detection.

First, we can notice that BS dominates FL in both detection ability and conditional power. FL often chooses closely neighboring points early in the path algorithm, and fails to capture all four points accurately. If we allow it to go $\hat{k}$ steps chosen by the two-rise IC rule, and also perform centroid clustering of the detected changpeoints prior to forming segment tests, it gains back some conditional power (middle panel). Both CBS and WBS dominate BS in detection ability. This is understandable, as CBS is designed to detect pairs of jumps in alternating directions, and WBS is designed for effective detection of local jumps. In conditional power, WBS is the weakest out of the three, which is likely because the conditioning (i.e. adaptivity to the data) is the strongest.

Improved test contrasts The segment test in (9) is practically and conceptually appealing because it is equivalent to a likelihood ratio test between two fixed models of piecewise constant mean which differ by the single tested changepoint b_i. However, in wild binary segmentation, we might be able to design a more powerful linear contrast by incorporating the endpoints of the intervals in which the CUSUM statistic was maximized. Coining this the 'segment+' test contrast, the exact form is:

$$v_i^T \theta = s_i \left(\frac{1}{W_i^2 - b_i} \sum_{i=b_i+1}^{W_i^2} \theta_i - \frac{1}{b_i - W_i^1} a \sum_{i=W_i^1}^{b_i} \theta_i \right), \tag{30}$$

where b_i is the i'th detected changepoint and $W_i = (W_i^1, W_i^2)$ is the couplet of interval endpoints in which the CUSUM was maximized. Using W_i in forming (30) is permissible because



δ	Segment+	Segment
Testing 120 ± 2		
0	0.000	0.006
0.5	0.463	0.272
1	0.858	0.756
2	0.959	0.952
3	1.000	0.980
4	1.000	0.990
Testing 80 ± 2		
0	0.011	0.022
0.5	0.079	0.067
1	0.229	0.325
2	0.766	0.794
3	0.869	0.908
4	0.970	0.949

Figure 5: *In a four-jump example with* $n = 200$ *as described in* (26) *with signal size* $\delta = 1$*, we can see that the power of the segment test is consistently higher than the original test. The first changepoint to be detected is typically around location* 120*. In this case, the power is dramatically increased. Power is defined as the proportion of tests that yield a p-value below a Bonferonni-corrected cutoff of* 0.05/4*.*

the endpoints of this winning interval is fixed with respect to the distribution conditioned on polyhedron (11). In the left panel of Figure 5, an example of a segment+ contrast is shown on a simulated data example from (26); the left region for the segment+ contrast is longer than that of the original segment contrast. In the top half of the table, we can see that the conditional power for this improved test statistic dominates wild binary segmentation consistently across signal sizes when testing 120 ± 2. This advantage disappears in further steps. The lower half of the table shows lower power for segment+ tests when conducted on locations 80 ± 2, which is typically detected in the third or fourth step. Because the endpoints used in segment+ are affected by segmentation from earlier steps, the endpoints cannot extend as long as was previously possible, and in fact are often shorter than is desirable, resulting in lower-power tests contrasts. This suggests that when testing changepoints detected in the first few steps of WBS, segment+ is a good substitute to the segment test to boost power (a more detailed analysis is provided in Appendix G).

Additive noise and power There is also an interesting tradeoff in power and the amount of additive noise used for randomization. As the additive noise level increases, the data used for fitting becomes more obfuscated, deteriorating the accuracy of detection. Hoever, the *conditional* power increases with σ_{add}. As a result, we expect an increase in *unconditional* power up to a point, followed by a decrease as segmentation fails. We can verify this from a simulation example. We consider a fixed signal-to-noise ratio regime of $\delta = 1$ (a weak signal) in the four-jump scenario of (26) with $n = 200$. We conduct post-selection segment tests on changepoint locations obtained from four-step WBS. Figure 6 shows conditional and unconditional power, and detection rate over a range of additive noise levels: $\sigma_{\text{add}} \in (0, 2)$. We see two effects of increasing additive noise – first, the detection ability uniformly decreases (dashed line), and both types of power have peak around roughly half ($\sigma_{\text{add}} = 0.5$) of the original noise.

4.2 Pseudo-real data application

From the same array CGH data as used in Figure 1, We create a pseudo-data simulation. We first fit a 1d fused lasso with an elastic net penalty, from which we extract locations of changepoints. These locations are then further pruned by post-processing to eliminate changepoints that are

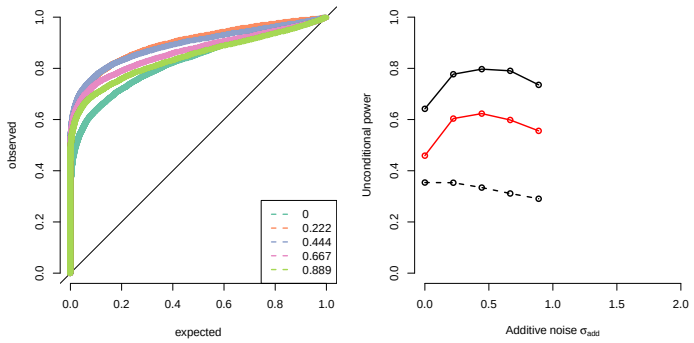


Figure 6: *In a four-jump example with* $n = 200$*, as described in* (26)*, the normal quantile-quantile plots (left) and powers (right) demonstrate the dependence of power on the amount of additive noise* σ_{add} *for randomized binary segmentation segment test p-values.*

adjacent. Taking regions in between these pruned points to be constant and the regions close to zero to be equal to zero, we create a flat mean $\hat{mu}$ of total length $n = 2012l$ with three regions of deviation from zero. The first 200 points are excluded for the estimation of noise.

(SH: need to /completely/ change this.) To this mean, we add bootstrapped versions of the residuals $\eta r = y - \hat{\mu}, \eta = 1$, and repeat the analysis. In addition to pure bootstrapped noise, we increase the difficulty of the problem by multiplying r by $\eta = 2, 3, 4$ prior to bootstrap. On these simulated datasets, noise-added SBS and WBS were run for a number of steps chosen by $q = 2$ rises in BIC, and segment tests were performed on the detected changepoints. Figure **??** shows the results of these simulations. We can see that both the conditional and unconditional power are very high for the original-scale ($\eta = 1$) case, and decreases as we increase the problem difficulty $\eta = 2, 3, 4$. We can see that both noise-added binary segmentation and circular binary segmentation perform well in terms of power.

5 Conclusions

We have described an approach to conduct post-selection inference on changepoints detected by common segmentaiton algorithms, using the same data for detection and testing. Segmentation algorithms are popular across many fields of application, including biology and economics, and have been well studied in the literature. The proposed approaches include adaptations of several recent developments in post-selection inference. For powerful randomized saturated model inference, we outlined an importance sampling strategy for calculating p-values from a selective distribution with relaxed conditioning – both over additive noise and also over intervals used in wild binary segmentation. For inference under a selected model, we demonstrated the procedure of conditioning on the sufficient statistic of a Gaussian changepoint model that assumes a piecewise constant mean, then outlined and implemented a hit-and-run Monte Carlo sampling scheme in order to calculate the p-values. We demonstrated the application in array CGH data, where we show that our methods effectively provide a statistical filter – in Figure 1, we show that the post-selection hypothesis tests after binary segmentation retain only the stronger five signals out of the thirteen detected. A pseudo-real simulation example on bootstrapped data also confirms this strength of our proposed tools. In addition , we demonstrate the detection probability and power over signal-to-noise ratios in a variety of simulations.

Future work in this area could improve the practical applicability of these methods. One useful extension would be to incorporate more complex and realistic noise models. For example, a noise model for CGH might incorporate some estimated spatial dependence. The impact of violations of model conditions – for example, using estimating noise levels for saturated models

– is also important item for future study. The selected model testing framework can also be extended to include other exponential family models for y_i. Post-selection inference may also be extended to multiple streams of copy number variation data from different subjects in order to more powerfully detect and make inferences about changepoint locations. Lastly, it may be possible to extend the framework of inference after cross validation (Loftus 2015) to post-selection changepoint inference.

References

Aue, A. & Horvath, L. (2013), 'Structural breaks in time series'.
URL: *http://dx.doi.org/10.1111/j.1467-9892.2012.00819.x*

Bai, J. (1999), 'Likelihood ratio tests for multiple structural changes', *Journal of Econometrics* **91**(2), 299–323.

Barry, D. & Hartigan, J. A. (1993), 'A Bayesian Analysis for Change Point Problems', *Journal of the American Statistical Association* **88**(421), 309–319.
URL: *https://doi.org/10.1080/01621459.1993.10594323*

Chib, S. & Olin, J. M. (1997), Estimation and comparison of multiple change-point models, number 86, pp. 221–241.

Fan, Z. & Mackey, L. (2015), 'An empirical bayesian analysis of simultaneous changepoints in multiple data sequences'.

Fithian, W., Sun, D. & Taylor, J. (2014), Optimal inference after model selection. arXv: 1410.2597.

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

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

Fridlyand, J., Snijders, A. M., Pinkel, D., Albertson, D. G. & Jain, A. N. (2004), 'Hidden markov models approach to the analysis of array cgh data', *Journal of Multivariate Analysis* **90**(1), 132 – 153. Special Issue on Multivariate Methods in Genomic Data Analysis.
URL: *http://www.sciencedirect.com/science/article/pii/S0047259X04000260*

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

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

Harchaoui, Z. & Lvy-Leduc, C. (2010), 'Multiple change-point estimation with a total variation penalty', *Journal of the American Statistical Association* **105**(492), 1480–1493.
URL: *https://doi.org/10.1198/jasa.2010.tm09181*

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

Horvath, L. & Rice, G. (2014), 'Extensions of some classical methods in change point analysis', *TEST* **23**(2), 219–255.

Hyun, S., G'sell, M. & Tibshirani, R. (2017), 'Exact post-selection inference for generalized lasso', *Electronic Journal of Statistics* .

Jandhyala, V., Fotopoulos, S., Macneill, I. & Liu, P. (2013), 'Inference for single and multiple change-points in time series', *Journal of Time Series Analysis* **34**(4), 423–446.

Lee, J., Sun, D., Sun, Y. & Taylor, J. (2016), 'Exact post-selection inference with application to the lasso', *Annals of Statistics* . To appear.

Loftus, J. R. (2015), 'Selective inference after cross-validation', *ArXiv e-prints* .

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

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

Rinaldo, A. (2009), 'Properties and refinements of the fused lasso', *Ann. Statist.* **37**(5B), 2922–2952.
URL: *https://doi.org/10.1214/08-AOS665*

Rudin, L. I., Osher, S. & Faterni, E. (1992), 'Nonlinear total variation based noise removal algorithms', *Physica D: Nonlinear Phenomena* **60**, 259–268.

Scott, A. & Knott, M. (1074), 'A cluster analysis method for grouping means in the analysis of variance', *Biometrics* **30**, 507–512.

Snijders, a. M., Nowak, N., Segraves, R., Blackwood, S., Brown, N., Conroy, J., Hamilton, G., Hindle, a. K., Huey, B., Kimura, K., Law, S., Myambo, K., Palmer, J., Ylstra, B., Yue, J. P., Gray, J. W., Jain, a. N., Pinkel, D. & Albertson, D. G. (2001), 'Assembly of microarrays for genome-wide measurement of DNA copy number.', *Nature genetics* **29**(3), 263–264.

Tartakovsky, A. G. & Moustakides, G. V. (2010), 'State-of-the-Art in Bayesian Changepoint Detection', *Sequential Analysis* **29**(2), 125–145.
URL: *https://doi.org/10.1080/07474941003740997*

Tian, X. & Taylor, J. E. (2015), 'Selective inference with a randomized response'.
URL: *http://arxiv.org/abs/1507.06739*

Tibshirani, R. J. & Taylor, J. (2011), 'The solution path of the generalized lasso', *Annals of Statistics* **39**(3), 1335–1371.

Tibshirani, R. J., Taylor, J., Lockhart, R., & Tibshirani, R. (2016), 'Exact post-selection inference for sequential regression procedures', *Journal of the American Statistical Association* . To appear.

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

Vostrikova, L. (1981), 'Detecting disorder in multidimensional random processes', *Soviet Math. Dokl.* **24**, 5559.

Worsley, K. J. (1986), 'Confidence-regions and tests for a change-point in a sequence of exponential family random-variables', *Biometrika* **73**(1), 91–104.

Yao, Q. (1993), 'Tests for change-points with epidemic alternatives', *Biometrika* **80**(1), 179–191.
URL: *http://www.jstor.org/stable/2336767*

Yao, Y.-C. (1984), 'Estimation of a noisy discrete-time step function: Bayes and empirical bayes approaches', *Ann. Statist.* **12**(4), 1434–1447.
URL: *https://doi.org/10.1214/aos/1176346802*

A Simulation examples

Here, we present more simulations to verify inferential properties. Having simulated Gaussian data from the four-jump mean as defined in (26) with $n = 20$, we verify the uniformity of p-values under two *null* scenarios in which the hypothesized linear contrasts after selection (by three methods – BS, noise-added BS, and WBS) is equal to zero. The first scenario is when the signal size δ is equal to zero. The p-values of *any* post-selection segment tests of changepoints detected from one or two algorithm steps, should be $U(0, 1)$ distributed. The second scenario is when the signal size is nonzero ($\delta > 0$) but the segment test contrast for changepoints from one or two step algorithm steps, $v^T\mu = 0$, is zero in mean. In this case, the resulting p-values should also be distributed as $U(0, 1)$. We can see in Figure 7 that all p-values that fall under these two null scenarios are indeed uniformly distributed.

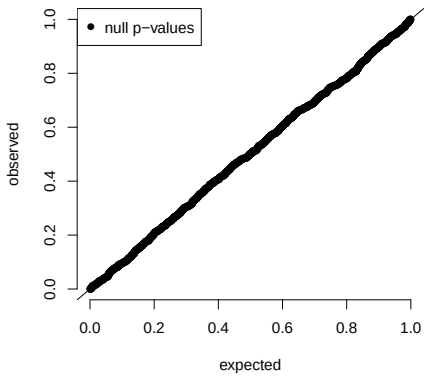


Figure 7: *Under various null scenarios, post-selection segment tests for simulated data generated around a four-jump mean (as in* (26)*) are shown to be distributed* $U(0, 1)$*. This verifies the Type-I error control property of our methods, under simulations (Placeholder figures, for now).*

B Proof of lemma 2.1

B.1 Binary segmentation

We describe the selection event for k-step binary segmentation where k is fixed a priori. Recalling the CUSUM statistic $y^b_{s,e}$ and related notation we will define a vector $w_{(s,b,e)} \in \mathbb{R}^n$ for any $s, b, e \in \{1, \dots, n\}$ where $s \leqslant b < e.$ such that

$$y^b_{s,e} = \sqrt{\frac{1}{\frac{1}{|e-b|} + \frac{1}{|b+1-s|}}}\left(\bar{y}_{(b+1):e} - \bar{y}_{s:b}\right) = w^T_{(s,b,e)}y.$$

Upon fitting the binary segmentation for a fixed k number of changepoints, recall that $b_1, \dots, b_k$ denote the set of changepoints, where b_i is the ith changepoint found in sequence. For any $i \in \{1, \dots, k\}$, denote $C_i = (c_0, c_1, \dots, c_{i-1}, c_i)$ where $0 = c_0 < \dots < c_i = n$, and $c_{1:(i-1)}$ are the sorted increasing permutation of $b_{1:(i-1)}$. Notice that C_i segments the indices $\{1, \dots, n\}$ into $i - 1$ partitions. We call these partitions $\mathcal{I}_i = \{I_1, \dots, I_i\}$, where $I_j = \{c_{j-1} + 1, \dots, c_j\}$ for $j \in \{1, \dots, i + 1\}$, and b_i lies in exactly one of these $i + 1$ partitions, which we will call P^*_i. Finally, we define (s_j, e_j) for $j \in \{1, \dots, i + 1\}$ so that $I_j = \{s_j, s_j + 1, \dots, e_j\}$. Similarly, we define (s^*_i, e^*_i) so that $I^*_i = \{s^*_i, s^*_i + 1, \dots, e^*_i\}$. $\mathcal{I}_j$ can be understood to represent the set of data partitions in which $y^b_{s,e}$ is maximized at step i.

We are ready to character the model selection event. Our model is characterized by

$$\mathcal{M} = \Big(\{b_1, \ldots, b_k\}, \{s_1, \ldots, s_k\}\Big),$$

where $(b_i, d_i) \in \{1, \ldots, n-1\} \times \{-1, 1\}$ is the index of the ith jump as well as the sign (i.e., direction) of the jump. The sign of the ith jump is 1 if

$$w^T_{(s_i^*, b_i, e_i^*)} y > 0,$$

and is -1 otherwise. Hence, we are interested in characterizing the set of y's such that applying the binary segmentation with for a fixed k number changepoints will result in the same model $\mathcal{M}$. We define our selection event recursively. For the first changepoint, $i = 1$, we have $2 \cdot (n-2)$ inequalities,

$$d_1 w^T_{(1,b_1,n)} y \geqslant w^T_{(1,j,n)} y, \quad \text{and} \quad d_1 w^T_{(1,b_1,n)} y \geqslant -w^T_{(1,j,n)} y, \quad \text{for } j \in \{1, \ldots, n\} \backslash \{b_1\}.$$

For the ith jump, for $i > 1$, we add an additional $2 \cdot (n - i - 2)$ inequalities. These can be understood as inequalities characterizing maximizations within I_i^*, and the rest. The inequalities pertaining to I_i^* are

$$d_i w^T_{(s_i^*, b_i, e_i^*)} y \geqslant w^T_{(s_i^*, j, e_i^*)} y, \quad \text{and} \quad d_i w^T_{(s_i^*, b_i, e_i^*)} y \geqslant -w^T_{(s_i^*, j, e_i^*)} y, \quad \text{for } j \in I_i^* \backslash \{b_i\}.$$

The inequalities not pertaining to I_i^* are

$$d_i w^T_{(s_i^*, b_i, e_i^*)} y \geqslant w^T_{(s_j, \ell, e_j)} y, \quad \text{and} \quad d_i w^T_{(s_i^*, b_i, e_i^*)} y \geqslant -w^T_{(s_j, \ell, e_j)} y, \quad \text{for } \ell \in I_j, \text{for } I_j \in \mathcal{P}_i \backslash \{I_i^*\}.$$

After forming all the inequalities, we note that all the inequalities are linear in y. That means we can construct a matrix Γ and a vector u such that our inequalities are succinctly represented by

$$\{y \ : \ \Gamma y \geqslant u\}.$$

(SH:Need to make this into an inductive reasoning.) (SH:Use different letter than w for the linear contrast vector.)

B.2 Wild binary segmentation

In addition to the changepoint $\hat{b}_{i,\text{obs}}$ defined in (2), define two other quantities:

$$s_{i,\text{obs}} = \text{sign}(\tilde{y}^{b_{i,\text{obs}}}_{s_{\hat{i}_i}, e_{\hat{i}_i}}) \tag{31}$$

and

$$i_{1,\text{obs}} = \underset{j \in (1:B)}{\text{argmax}} \left(\max_{b \in (s_j, e_j)} \tilde{y}^b_{s_j, e_j} \right) \tag{32}$$

which are respectively the sign of the maximized cusum statistic from (**??**), and the index of the interval in which the maximization occurs.

We proceed by induction. At step $i = 1$, it is straightforward to see that the triplet takes the values $(\hat{b}_1, \hat{i}_1, \hat{s}_1) = (b_{1,\text{obs}}, i_{1,\text{obs}}, s_{1,\text{obs}})$ if and only if the following hold:

$$\hat{s}_1 \tilde{y}^{\hat{b}_1}_{s_j, e_j} > 0 \tag{33}$$

and for all $b \in (s_j, e_j)$ for all $j \in (1:B) \backslash \{\hat{i}_1\}$ and $j \in \hat{i}_1$ and for all $b \in (s_j : e_j) \backslash \{\hat{b}_j\}$:

$$\hat{s}_1 \tilde{y}^{\hat{b}_1}_{s_{\hat{i}_1}, e_{\hat{i}_1}} \geqslant -\hat{s}_j \tilde{y}^b_{s_j, e_j} \tag{34}$$

$$\hat{s}_1 \tilde{y}^{\hat{b}_1}_{s_{\hat{i}_1}, e_{\hat{i}_1}} \geqslant \hat{s}_j \tilde{y}^b_{s_j, e_j}. \tag{35}$$

i.e. the sign of the maximizing CUSUM statistic, and the absolute value is bounded below and above by all others. At a general step $i > 1$, the triplet $(\hat{b}_i, \hat{i}_i, \hat{s}_i) = (b_{i,\text{obs}}, i_{i,\text{obs}}, s_{i,\text{obs}})$ is observed if and only if the following hold; the single inequality:

$$\hat{s}_i \tilde{y}^{\hat{b}_i}_{s_j, e_j} > 0, \tag{36}$$

and the following two inequalities for (j, b) such that $j \in (s_j : e_j)$ for all $j \in I_i \backslash \{\hat{i}_i\}$ and $(j, b) = (\hat{i}_i, b)$ for all $b \in (s_j : e_j) \backslash \{\hat{b}_j\}$:

$$\hat{s}_1 \tilde{y}^{\hat{b}_1}_{s_{\hat{i}_i}, e_{\hat{i}_i}} \geqslant -\hat{s}_j \tilde{y}^{b}_{s_j, e_j}$$
$$\hat{s}_1 \tilde{y}^{\hat{b}_1}_{s_{\hat{i}_i}, e_{\hat{i}_i}} \geqslant \hat{s}_j \tilde{y}^{b}_{s_j, e_j}.$$

□

B.3 Circular binary segmentation

First define the sign of the maximizing statistic:

$$s_{i,\text{obs}} = \text{sign}(\tilde{\tilde{y}}^{b_{i,\text{obs}}}_{s_{\hat{i}_i}, e_{\hat{i}_i}}) \tag{37}$$

We now proceed by induction. At step $i = 1$, it is straightforward to see that the triplet takes the values $(\hat{b}_1^{\text{start}}, \hat{b}_1^{\text{end}}, \hat{s}_1) = (b_{1,\text{obs}}^{\text{start}}, b_{1,\text{obs}}^{\text{end}}, s_{1,\text{obs}})$ if and only if the following holds:

$$\hat{s}_1 \tilde{\tilde{y}}^{\hat{b}_1^{\text{start}}, \hat{b}_1^{\text{end}}}_{s_j, e_j} > 0 \tag{38}$$

and the following two inequalities hold for all $1 \leqslant a < b \leqslant n$:

$$\hat{s}_1 \tilde{\tilde{y}}^{\hat{b}_1^{\text{start}}, \hat{b}_1^{\text{end}}}_{1,n} \geqslant -\hat{s}_j \tilde{\tilde{y}}^{a,b}_{1,n} \tag{39}$$
$$\hat{s}_1 \tilde{\tilde{y}}^{\hat{b}_1^{\text{start}}, \hat{b}_1^{\text{end}}}_{1,n} \geqslant \hat{s}_j \tilde{\tilde{y}}^{a,b}_{1:n} \tag{40}$$

At a general step $i > 1$, the triplet $(\hat{b}_i^{\text{start}}, \hat{b}_i^{\text{end}}, \hat{s}_i) = (b_{i,\text{obs}}^{\text{start}}, b_{i,\text{obs}}^{\text{end}}, s_{i,\text{obs}})$ is observed if and only if the following hold; the single inequality:

$$\hat{s}_i \tilde{y}^{\hat{b}_i}_{s_{\hat{i}_i}, e_{\hat{i}_i}} > 0, \tag{41}$$

and only if the following holds:

$$\hat{s}_{\hat{i}_i} \tilde{\tilde{y}}^{\hat{b}_1^{\text{start}}, \hat{b}_1^{\text{end}}}_{s_{\hat{i}_i}, e_{\hat{i}_i}} > 0 \tag{42}$$

and the following two inequalities hold for all (s, a, b, e) in the two cases (1) $s = \hat{s}_{\hat{i}_i}$ and $e = \hat{s}_{\hat{i}_i}$ and $a \neq \hat{b}_i^{\text{start}}$ and $b \neq \hat{b}_i^{\text{end}}$ and $a < b$, and (2) .

$$\hat{s}_{\hat{i}_i} \tilde{\tilde{y}}^{\hat{b}_1^{\text{start}}, \hat{b}_1^{\text{end}}}_{1,n} \geqslant -\tilde{\tilde{y}}^{a,b}_{s,e} \tag{43}$$
$$\hat{s}_{\hat{i}_i} \tilde{\tilde{y}}^{\hat{b}_1^{\text{start}}, \hat{b}_1^{\text{end}}}_{1,n} \geqslant \tilde{\tilde{y}}^{a,b}_{s,e} \tag{44}$$

□

C Sampling details for randomization

In addition to the randomness induced by the data $Y \sim F_y$, consider an external random component, $I \sim F_Y$, such that $F_Y \perp\!\!\!\perp F_I$ by design. For example, in wild binary segmentation (WBS), I can be thought of as randomly drawn endpoints (uniform in $\{1, \cdots, n\}$) to be used in the WBS algorithm. Another example is external additive noise to the original data. We outline the procedure for conducting randomized post-selection inference according to Tian & Taylor (2015).

Main derivation The end goal is to calculate a truncated Gaussian p-value without conditioning on the realization of the random component $I = i$:

$$h(t) = P(v^T Z \geqslant t | \underbrace{\hat{M}(Z, I) = M_{\text{obs}}}_{E_1}, \underbrace{P_v^\perp Z = P_v^\perp y_{\text{obs}}}_{E_2}) \tag{45}$$

First write this as an integral over the joint density of I and Z:

$$h(t) = \int_{z,i} \mathbb{1}(v^T Z \leqslant t) \underbrace{f_{I,Z|E_1,E_2}(i, z)}_{A} didz \tag{46}$$

Then the joint density A partitions into two components:

$$f_{I,Z|E_1,E_2}(i, z)dzdi = f_{Z|I=i,E_1,E_2}(z) f_{I|E_1,E_2}(i)dzdi \tag{47}$$

Using Bayes rule, we can write the latter probability mass function as:

$$f_{I|E_1,E_2}(i) = \frac{P(E_1|I = i, E_2) f_{I|E_2}(i)}{P(E_1|E_2)}, \tag{48}$$

where we have flipped E_1 and I while conditioning on E_2. With this, we can rewrite $h(t)$ as a double integral over I and Z

$$h(t) = \int \mathbb{1}(v^T Z \leqslant t) \cdot P(E_1|I = i, E_2) \frac{f_{I|E_2}(i)}{P(E_1|E_2)} \cdot f_{Z|I=i,E_1,E_2}(z)dzdi. \tag{49}$$

Now, rearranging (and writing the integral over i as a sum since it is discrete), we get:[3]

$$h(t) = \int_{\text{supp}(I)} \left[\int \mathbb{1}(v^T z \leqslant t) \cdot f_{Z|I=i,E_1,E_2}(z)dz \cdot P(E_1|I = i, E_2) \right] \frac{f_{I|E_2}(i)}{P(E_1|E_2)}, \tag{50}$$

Denoting by $g(i) = P(E_1|I = i, E_2)$ and $h_i(t) = \int \mathbb{1}(v^T Z \leqslant t) \cdot f_{Z|I=i,E_1,E_2}(z)\ dz \cdot g(i)$, (50) be rearranged to form:

$$h(t) = \int_{\text{supp}(I)} \left[\int \mathbb{1}(v^T Z \leqslant t) \cdot f_{Z|I=i,E_1,E_2}(z)\ dz \cdot g(i) \right] \frac{1}{P(E_1|E_2)} f_I(i) = \int_{\text{supp}(I)} h_i(t) \frac{1}{P(E_1|E_2)} f_I(i). \tag{51}$$

Now, substitute in the following:

$$P(E_1|E_2) = \int_{\text{supp}(I)} P(E_1|I = j, E_2) f_{I|E_2}(j)dj = \int_{\text{supp}(I)} g(j) f_{I|E_2}(j)dj. \tag{52}$$

so that $h(t)$ becomes:

$$h(t) = \int_{\text{supp}(I)} h_i(t) P_I(i) \cdot \frac{g(i)}{\int_{\text{supp}(I)} g(j) f_{I|E_2}(j)}. \tag{53}$$

The multiplier in the last part can be thought of as an importance weight by viewing it as a density ratio. This can be seen by applying Bayes rule to the numerator of (53), and rearranging:

$$\frac{g(i)}{\int_{\text{supp}(I)} g(i) P_I(i)} = \frac{P(E_1|E_2, I = i)}{P(E_1|E_2)} = \frac{P(I = i|E_1, E_2)}{P(I = i|E_2)} = \frac{P(I = i|E_1, E_2)}{P(I = i)} \tag{54}$$

From this, an importance sampling estimate of $h(t)$ can be deduced to be:

$$\hat{h}(t) = \sum_i h_i(t) P_I(i) \cdot \frac{g(i)}{\sum_j g(j) f_{I|E_2}(j)}. \tag{55}$$

The importance sampling scheme is as follows: instead of sampling from $I|E_1, E_2$, which is hard or impossible, sample from the easier reference distribution $I|E_2$, and applying the importance weight in (54). We additionally assume that I and E_2 are independent so that $I|E_2 \overset{d}{=} I$, making the reference distribution even simpler.

[3] We were able to bring three things – $P(E_1|E_2)$, $P(E_1|I = i, E_2)$ and $f_{I|E_2}(i) = p_I(i)$ – out of the integral with respect to $P_{Z|I=i,E_1,E_2}$, because they are constant with respect to $Z|I = i, E_1, E_2$; constant in the sense that it does not depend on a particular instance of it; just like $\int zP(Z \in A)dP_Z(z) = \int zdP_Z(z)P(Z \in A)$. This equality comes from our assume the distribution of I does not change after conditioning on E_2; this is debatable.

Sampling instructions The form in (55) can be simplified once more by noticing that $h_i(t) = k(i)/g(i)$, where $k(i) = F_{v^T\mu,\sigma^2\|v\|_2^2}(V^{up}) - F_{v^T\mu,\sigma^2\|v\|_2^2}(t)$ and cancelling $g(i)$ to get:

$$\hat{h}(t) = \frac{\sum_i k(i) \cdot P_I(i)}{\sum_j g(j) \cdot P_I(j)}, \tag{56}$$

giving a clear recipe for sampling:

1. Sample $I = i$ from the marginal distribution.
2. Calculate $\begin{cases} k(i) = F_{v^T\mu,\sigma^2\|v\|_2^2}(V^{up}) - F_{v^T\mu,\sigma^2\|v\|_2^2}(t) \\ g(i) = F_{v^T\mu,\sigma^2\|v\|_2^2}(V^{up}) - F_{v^T\mu,\sigma^2\|v\|_2^2}(V^{lo}) \end{cases}$.
3. Add each to collection $\{k(i)\}$ and $\{g(i)\}$

Then, the final p-value $\hat{h}(t)$ can be calculated as $\sum_i k(i)/\sum_i g(i)$.

Modifications for WBS Having fixed the drawn intervals $(i_1, \cdots, i_B)$, there are several options for conditioning:

$$P_1 = \left\{ y : \hat{M}(y, i) = \{+5\} \right\} \tag{57}$$

$$P_2 = \{ y : \text{1-step WBS applied to } y \text{ on intervals } i_1, \cdots, i_B \text{ resulted in } +5 \} \tag{58}$$

$$P_3 = \left\{ y : \begin{array}{l} \text{location 5 in } i_m = 3:8 \text{ had the largest CUSUM,} \\ \text{out of all other cusums.} \end{array} \right\} \tag{59}$$

A and B, which are equal, are a *union* of events – it is not specific about which interval the maximum occurs. For instance, it allows for the case that there is another interval $2:9$ whose maximizer is at 5.

So, we can take the conditioning statement in (45) to be

$$TG = P(v^T Z \leqslant t | \underbrace{\hat{M}(Z, I, i_{(\max)}) = M_{\text{obs}}}_{E_1}, \underbrace{P_v^{\perp} Z = P_v^{\perp} y_{\text{obs}}}_{E_2}, i_{(max)} = i_{\max}^{\text{obs}}) \tag{60}$$

where $i_{\max} \in I$, also a random variable, is the interval in which the maximum has occurred. The main modification is now to sample *other* $N-1$ random intervals whose maximum CUSUM statistics are smaller than what had occurred in the $i_{\max}$'th one, as the random component.

Continue here.

D Sampling details for randomized wild binary segmentation inference

The selection event (one that is a minimally unique characterizable) is the maximizations of the cusum statistics at the i'th step in a specific interval. Notating that 'winning' interval as the $\hat{G}_i(y_{\text{obs}})$'th one, the event is:

$$P_{\text{wbs}} = \{ y : \hat{b}_{1:k}(y) = \hat{b}_{1:k}(y_{\text{obs}}), \hat{s}_{1:k}(y) = \hat{s}_{1:k}(y), \hat{G}_{1:k}(y) = \hat{G}_{1:k}(y_{\text{obs}}) \} \tag{61}$$

This is still a polyhedral set.

Here the trick is to have all components in I, except for the actual interval in which the changepoint occurred, be allowed to be random. This minimal amount of conditioning still allows $E_0, E_1|E_2$ to be a *polyhedral* event, which enables closed form calculation of importance weights. A similar mechanism allows for importance sampling scheme after having fit larger number of steps of WBS. The details of sampling are deferred to the appendix.

E IC-based stopping rule with randomization

The IC-based stopping rule requires some special handling when combined with randomized post-selection infernece. For fused lasso or binary segmentation inference with additive noise randomization, recall our inference is based on a polyhedron that characterizes the selection event on noise-infused (fuzzy) data:

$$P_{noisy} = \{y : \Gamma_{\text{noisy}} y \geqslant u_{\text{noisy}}\},$$

and the event that we aim to condition on is:

$$\Gamma_{\text{noisy}}(Y + W) \geqslant u_{\text{noisy}} \iff \Gamma_{\text{noisy}} Y \geqslant u_{\text{noisy}} - \Gamma_{\text{noisy}} W.$$

The stopping time is to be calculated on the noise-added (fuzzed) data, so that we denote it $\Gamma^{ic}_{\text{noisy}}$ and u^{ic}_{noisy}. They simply enter the dataset as a

For wild binary segmentation, the IC minimization is with respect to nested changepoint models constructed from the changepoint locations $\hat{b}_{1:k}$ and signs $\hat{s}_{1:k}$. Let us say that k_{obs} is that observed stopping time. The randomized selective distribution is

$$Y|E_0, E_1, E_2, \hat{k}(y, \hat{b}) = k_{\text{obs}}$$

In the importance sampling involved in calculating the p-value function under this distribution requires $P(E_1|E_0, E_2)$ and $P(v^T Y > v^T y_{\text{obs}}|E_0, E_2)$. These are now modified to $P(E_1, \hat{k} = k_{\text{obs}}|E_0, E_2)$; essentially, replace E_1 with $E_1|\hat{k} = k_{\text{obs}}$ everywhere.

F Wild Binary Segmentation Number of intervals

We examine the effect of the choice of the number of intervals, n_I in wild binary segmentation inference. For a four-jump signal ($n = 50$), the detection ability of wild binary segmentation does not drastically differ after $n_I = n/2$, tapering off to a flat level of 0.75 at around $n_I = n$.

Fixing other simulation settings, conditional power does not notably change across different choices of n_I, as can be seen in 8

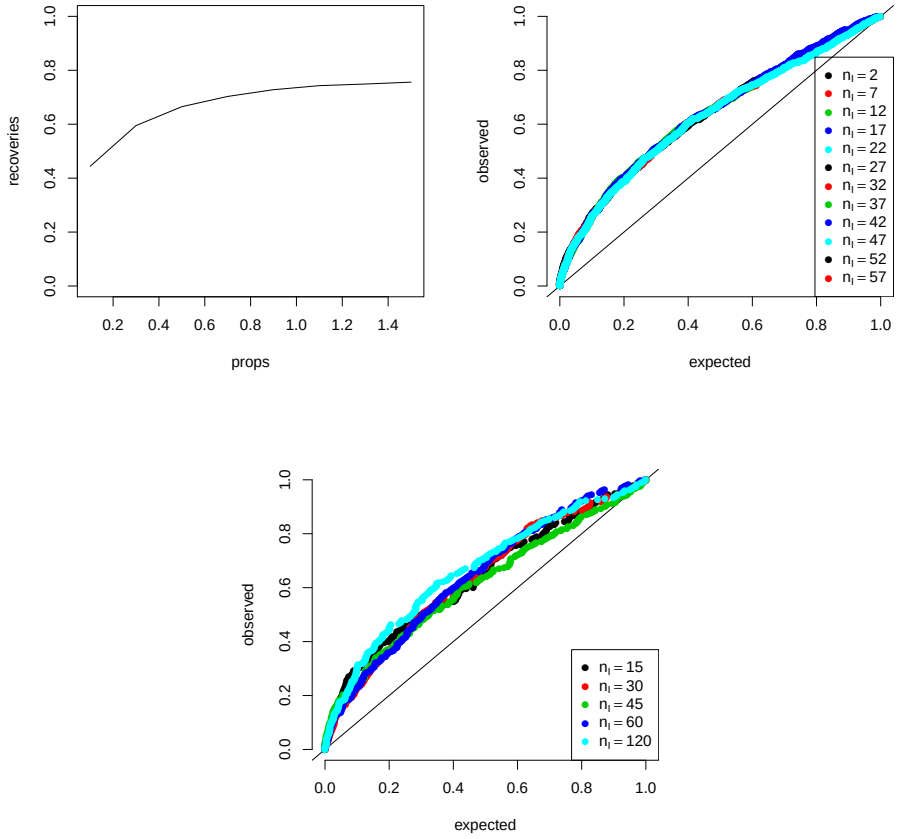


Figure 8: *(Top row) We generated data form a four-jump mean as in 26 with signal size $\delta = 1$, as shown in the left panel. Then, we calculated the average of the approximate correct detection proportion by a four-step fixed wild binary segmentation. Approximate correct recovery is defined to be within ± 1 viscinity of correct locations $n/5, \cdots, 4n/5$. This average detection proportion was calculated for n_I values ranging in $(0, 1.5n)$. (Bottom row) QQ plots of p-values for tests conducted about approximately correct locations, for various choices of n_I in the same range as above. The left shows p-value QQ plots for fixed inference, and right shows this for randomized inference. (SH: rerun everything with $n = 200$ with $\delta = 2$.)*

G Improved segment+ test for WBS

We continue the discussion of an improved segment test contrast – coined 'segment+' – that uses the winning intervals of wild binary segmentation events . We demonstrated that power can be noticeably improved when using segment+, in 5, in earlier steps. Figure 9 shows a detailed comparison of the test contrasts for some insight into why it is more powerful in earlier steps and why this trend is reversed in later steps. Take the four-jump example as described in (26) with $\delta = 1$.

Earlier steps (120 ± 5) We first isolate our attention to the four-step wild binary segmentation tests regarding location 120 ± 5, the third breakpoint location from the left. In the left panel, we can see that the segment+ test contrasts are more likely to use left endpoints that well extend to the left of 40 (the right end is similar on average). On the other hand, the segment+ test contrasts use left endpoints that are concentrated around 80. From the mean structure as

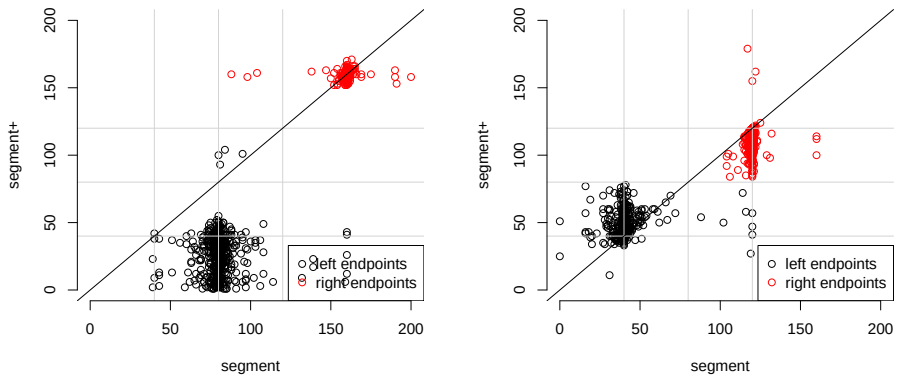


Figure 9: *Data was generated from a four-jump example in* (26) *with* $\delta = 1$. *After applying a four-step wild binary segmentation algorithm, corresponding segment and segment+ test contrasts endpoints are displayed in a scatterplot – each pair of points comes from the same simulation replicate. Two cases were considered – when (left)* 120 ± 5 *was tested, and when (right)* 80 ± 5 *was tested. (Left panel) In the former case, we can see that segment+ left endpoints often extend to the viscinity of location* 40*, while counteraprt segment test left endpoints are usually limited to the right of* 80*. (Right panel) In the latter case, the segment test endpoints are usually closer to* 40 *and* 80*, and segment+ contrasts are usually closer, making for shorter left and right segments.*

described in 26, we know that it is beneficial to use longer left segments when detecting a mean shift at location 80. As expected, the segment+ test has higher power.

Later steps (80 ± 5) When testing 80 ± 5, the second breakpoint location from the left, the trend reverses. Both contrasts use endpoints are usually within $(40, 120)$, and the segment+ contrasts have closer-by endpoints. This is understandable because coordinates in 80 ± 5 are detected later in the algorithm, so it is impossible for there to be larger segments, since the earlier segmentations limit this from happening. Because segment+ use strictly shorter segments, there is a loss of power that occurs

H Assessing model assumptions

(SH: This was originally kind of a stretch goal, but now I see how it is quite important for this paper. but I would want to show how close/far the data is from (i.i.d.) Gaussianity, and what effect it has on inference, and how to remedy it Start with: 'Data is never perfectly i.i.d. Gaussian. We discuss a departure from it.'

- How far is estimated noise from the true noise?
- As a diagnostic (after the BIC-stopped model), talk about (or simulate) residuals from this stopped model (optionally, KS test).
- Show QQ plots (of the left-out region)
- Produe 1 to 4 lag serial correlation tables.
- Compare the validity of WBS inference when you have 1-lag autocorrelated noise? Compare the following three in a table, in terms of the nominal rejection probability (i.e. proportion of p-values under $\alpha/4$) under simple, $lev = 0, 1, 2$ settings.
  - Use the naive i.i.d. assumption uner real noise
  - Use the naive i.i.d. assumption uner estimated noise

– Use an estimated Toeplitz Sigma matrix?