Received: 23 September 2019 | Revised: 6 July 2020 | Accepted: 11 August 2020

DOI: 10.1111/biom.13422

Biometrics
A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY
WILEY

BIOMETRIC METHODOLOGY

Post-selection inference for changepoint detection algorithms with application to copy number variation data

Sangwon Hyun[1] | Kevin Z. Lin[2] | Max G'Sell[3] | Ryan J. Tibshirani[3]

[1] Department of Data Sciences and Operations, University of Southern California, Los Angeles, California, USA

[2] Department of Statistics, University of Pennsylvania, Philadelphia, Pennsylvania, USA

[3] Department of Statistics and Data Science, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA

Correspondence:
Sangwon Hyun, Department of Data Sciences and Operations, University of Southern California, Los Angeles, CA 90089, USA.
Email: sangwonh@usc.edu

Funding information
Division of Mathematical Sciences, Grant/Award Numbers: DMS-1554123, DMS-1613202

Abstract

Changepoint detection methods are used in many areas of science and engineering, for example, in the analysis of copy number variation data to detect abnormalities in copy numbers along the genome. Despite the broad array of available tools, methodology for quantifying our uncertainty in the strength (or the presence) of given changepoints *post-selection* are lacking. Post-selection inference offers a framework to fill this gap, but the most straightforward application of these methods results in low-powered hypothesis tests and leaves open several important questions about practical usability. In this work, we carefully tailor post-selection inference methods toward changepoint detection, focusing on copy number variation data. To accomplish this, we study commonly used changepoint algorithms: binary segmentation, as well as two of its most popular variants, wild and circular, and the fused lasso. We implement some of the latest developments in post-selection inference theory, mainly auxiliary randomization. This improves the power, which requires implementations of Markov chain Monte Carlo algorithms (importance sampling and hit-and-run sampling) to carry out our tests. We also provide recommendations for improving practical useability, detailed simulations, and example analyses on array comparative genomic hybridization as well as sequencing data.

KEYWORDS

comparative genomic hybridization analysis, changepoint detection, copy number variation, hypothesis tests, post-selection inference, segmentation algorithms

1 | INTRODUCTION

Changepoint detection algorithms identify changes in data distribution along a sequence of observations and are commonly used in copy number variation (CNV) analyses that detect deviations in the copy number in any region along the genome (Zhang, 2010). Specifically, in this article, we study the canonical changepoint model, where changes occur only in the mean. Let the vector $\boldsymbol{Y} = (Y_1, \dots, Y_n) \in \mathbb{R}^n$ be a data vector with independent Gaussian entries,

$$Y_i \sim \mathcal{N}(\theta_i, \sigma^2), \quad i = 1, \dots, n, \tag{1}$$

where the unknown mean vector $\boldsymbol{\theta} \in \mathbb{R}^n$ forms a piecewise constant sequence. That is, for locations $1 \leq b_1 < \cdots < b_t \leq n-1$,

$$\theta_{b_j+1} = \cdots = \theta_{b_{j+1}}, \quad j = 0, \dots, t,$$

where, for convenience, we write $b_0 = 0$ and $b_{t+1} = n$. We call $b_1, \dots, b_t$ *changepoint* locations of θ. Given such models, changepoint detection algorithms typically focus on estimating the number of changepoints t (possibly none), as well as the locations $b_1, \dots, b_t$, from a single vector $\boldsymbol{Y}$. This includes *binary segmentation* (BS) and its many variants common in the literature, such as *wild binary segmentation* (WBS; Fryzlewicz, 2014) and *circular binary segmentation* (CBS; Olshen *et al.*, 2004). Loosely speaking, these algorithms estimate the changepoint locations by scanning the entire vector Y and finding locations where the empirical means to the left and right segments are well separated. Despite the large number of theoretical results that formalize the point-wise estimation performance of these algorithms (see Lin *et al.*, 2017 and references within), there have been much fewer works that focus on computing valid p-values that quantify the significance of such changepoints. Having valid p-values can be greatly beneficial for filtering changepoints in an automated fashion, where only statistically significant changepoints are kept for potential downstream analyses. As we show later in this section, this methodological gap can be problematic for CNV analyses since naive hypothesis tests can inflate the Type-I error, leading to undesirable filtering procedures. Hence, in this article, we leverage the recent developments in post-selection inference (Tibshirani *et al.*, 2018) to develop an downstream algorithm to compute p-values within this framework for the aforementioned changepoint algorithms. By developing valid downstream inferential tools, we strengthen commonly used changepoint detection algorithms in CNV analyses by enabling a principled way to filter changepoints in an automated fashion.

We provide a high-level summary of the post-selection inference applied to changepoint model (1). This serves both as an overview of the post-selection framework and also highlights the algorithmic challenges we resolve in article. The machinery that we build off is developed in works like Fithian *et al.* (2014, 2015) and Tian and Taylor (2018), whose results we rely on.

Basic inference procedure. The basic inference procedure we consider is as follows.

(1) Given data $\boldsymbol{Y}$, apply a changepoint algorithm to detect some fixed number of changepoints k. Denote the estimated changepoint locations by $\hat{b}_1, \dots, \hat{b}_k$, and their respective changepoint directions (whether the estimated change in mean was positive or negative) by $\hat{d}_1, \dots, \hat{d}_k \in \{-1, 1\}$. Let $\boldsymbol{I}_1, \dots, \boldsymbol{I}_{k+1}$ denote the partition of $\{1, \dots, n\}$ formed by $\hat{\boldsymbol{b}}_{1:k}$. The specifics of the changepoint algorithms that we consider are given in Section 2.1.

(2) Form contrast vectors $\boldsymbol{v}_1, \dots, \boldsymbol{v}_k \in \mathbb{R}^n$, defined so that for an arbitrary $\boldsymbol{y} \in \mathbb{R}^n$,

$$\boldsymbol{v}_j^T \boldsymbol{y} = \hat{d}_j \cdot \left\{ \frac{1}{|\boldsymbol{I}_{j+1}|} \left(\sum_{i \in \boldsymbol{I}_{j+1}} \boldsymbol{y}_i \right) - \frac{1}{|\boldsymbol{I}_j|} \left(\sum_{i \in \boldsymbol{I}_j} \boldsymbol{y}_i \right) \right\}, \quad (2)$$

for $j = 1, \dots, k$, where $|\boldsymbol{I}_j|$ denotes the cardinality of the set $\boldsymbol{I}_j$. Hence, $\boldsymbol{v}_j^T \boldsymbol{Y}$ represents the difference between the sample means of segments to right and left of $\hat{b}_j$,

(3) For each $j = 1, \dots, k$, we test the hypothesis $H_0 : \boldsymbol{v}_j^T \theta = 0$ by rejecting for large values of a statistic $T(\boldsymbol{Y}, \boldsymbol{v}_j)$, which is computed based on knowledge of the changepoint algorithm that produced $\hat{\boldsymbol{b}}_{1:k}$ in Step 1, the desired contrast vector (2) formed in Step 2, and the value of σ^2. Each statistic yields a p-value under the null (assuming the model (1)). The details of $T(\boldsymbol{Y}, \boldsymbol{v}_j)$ are given in Sections 2.2 and 3.

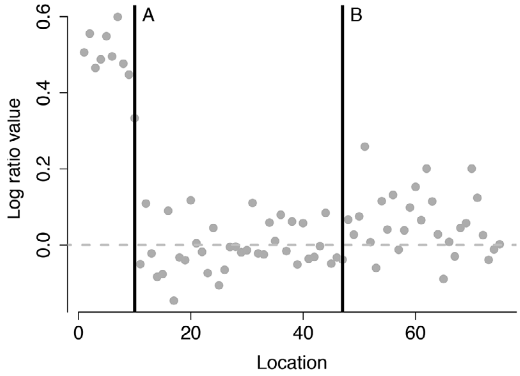
(4) Optionally, we can use the Bonferroni correction by multiplying the p-values by k, to account for multiplicity.

It is worth mentioning that several variants of this basic procedure are possible. For example, σ^2 can either be estimated from an alternative dataset or left unspecified; the number of changepoints k in Step 1 can be estimated from data; alternative contrast vectors to (2) in Step 2 may be used to measure more localized mean changes. We dedicate the following sections to highlight the novelties of our work within the simplified framework prescribed above and defer discussions of such variants to the Web Appendix. Additionally, though not covered in our article, the p-values from our tests can be inverted to form confidence intervals for population contrasts $\boldsymbol{v}_j^T \theta$ for $j = 1, \dots, k$ (Tibshirani *et al.*, 2018).

Contributions Our article has two primary contributions. First, we specialize the existing post-selection inference framework for CNV analyses in order to filter estimated changepoints and show their versatility on array-comparative genomic hybridization and sequencing data (Section 4). We prove results to facilitate concrete algorithms and provide extensive guidelines and variants that can be helpful for researchers. Second, we develop new methodologies to improve the power of our inferential tools and verify this improvement by simulation.

1.1 | CNV background and motivating analysis

To motivate the necessity for valid inferential tools in changepoint detection, we present a motivating CNV



Location	A	B
Karyotype	True	False
Classical t-test	0	5×10^{-4}
Saturated model test	0	0.050
Selected model test	0	0.027

FIGURE 1 (Left): aCGH data from the 14th chromosome of fibroblast cell line GM01750, from Snijders *et al.* (2001). The x-axis denotes the relative index of the genome position, and the y-axis denotes the measured $\log_2$ copy number ratio after a suitable preprocessing, with a dotted line denoting 0 for reference. The bold vertical lines denote the locations A and B from running WBS for two steps (corresponding to $\widehat{b}_1$ and $\widehat{b}_2$). (Right): The p-values using classical (naive) t-tests, saturated model tests, and selected model tests, at each location A and B. The ground truth is also given, as determined by karyotyping. The saturated model test used an estimated noise level σ^2 from the entire 23-chromosome dataset. The selected model test was performed in the unknown σ^2 setting. Specifically, we test the null hypothesis $H_0 : \boldsymbol{v}_j^T \boldsymbol{\theta} = 0$, for $j = 1, 2$, where the contrast vectors are as defined in (2). We see that both saturated and selected model tests correcting determine only location A to be associated with a true CNV

analysis on array comparative genomic hybridization (aCGH) data. Broadly speaking, CNV analyses investigate the structural variation of the genome where the total copy number of particular large regions in a chromosome deviate from two, the expected copy number—one from each parental cell. This type of variation has been implicated with tumor progression, and therefore many studies use CNV analyses to identify which regions of the genome to specifically investigate in future analyses (see works like Zhang, 2010). Toward this end, aCGH data is one of many types of data frequently collected to study CNV. This type of data is collected by microarrays, where the $\log_2$ copy number ratio between case and control cells along different regions along the genomes is measured by probes. However, since these measurements are often quite noisy, changepoint detection algorithms need to be used on aCGH data for effective detection of regions of CNV. The data in this motivating analysis originate from Snijders *et al.* (2001), which studies the CNV of fibroblast cell lines and is a common benchmark dataset for many changepoint detection algorithms, such as Olshen *et al.* (2004) and Duy *et al.* (2020).

Figure 1 illustrates how traditional inference tools can lead to invalid scientific conclusions and previews the results of the post-selection inferential tools we develop in this article. Here, we use a two-step WBS to estimate two locations $\widehat{b}_1$ and $\widehat{b}_2$ that segment the measured $\log_2$ ratios. Naively we might run t-tests for equality of means between the left and right neighboring segments of each estimated locations with the null hypothesis $H_0 : \boldsymbol{v}_j^T \boldsymbol{\theta} = 0$, for $j = 1, 2$ using the contrast vectors defined in (2). However, this would deem both changepoint locations as statistically significant, but an external karotyping dataset (from Snijders *et al.*, 2001) reveals that only one of the estimated locations is associated with a true CNV. At a high level, the t-test's erroneously small p-value arise since WBS detects specific changepoint locations where the empirical means to its left and right are well-separated, making the t-tests' null-hypotheses incorrect. To overcome this problem, we adapt the post-selection inference framework to the changepoint setting. As we will discuss in detail, our inferential tools can be done using either *saturated model* and a *selected model* on the mean vector $\boldsymbol{\theta}$. Moreover, using either such models leads to correctly deeming only one of the estimated changepoints as significant, as shown in Figure 1. We return to this dataset in Section 4.

We emphasize that while this motivating analysis is only for a single chromosome on one cell line, CNV analyses typically investigate the entire genome across many cell lines. Hence, an inflation in Type-I error can have profound effect in aggregate. Our inferential tools are well suited for these situations, as they can be used in an automated fashion as a way to filter out estimated changepoints in CNV analyses in a statistically principled fashion.

1.2 | Related work

The most common variant of post-selection inference is *sample splitting*, as discussed in Fithian *et al.* (2014). In our setting, this can be performed by dividing every odd and even index of $\boldsymbol{y}$ into two separate vectors, then applying BS or its variants on one vector, and using t-tests on the other. Since the data used to estimate the changepoints are independent of the data used for inference, this procedure yields valid p-values. However, as we will demonstrate empirically in Section 3, sample splitting reduces the test's power. Hence, in this article, we are interested in post-selection inferential tools that avoid sample splitting.

We mention additional works that use post-selection inference tools for changepoint detection. Most relevantly, Hyun *et al.* (2018) develop postinference tools for the generalized lasso (a generalization of the fused lasso), which can also be used for inference in the changepoint setting as well. However, those tools do not directly work for the variants of BS considered in this article and we additionally investigate methods to improve our tools' statistical power. Also, while writing this article, we became aware of the independent contributions in Duy *et al.* (2020), which appeared after the initial release of this article and develops variants that further improve our tools' statistical power. We remark that these notions of improved statistical power in this work are demonstrated mainly with empirical evidence. To the best of our knowledge, Azaïs *et al.* (2018)'s article is the only article that proves the power of post-selection inference methods, but its theoretical setting is not applicable for most changepoint analyses we encounter in practice. Aside from these articles, there is little focus on valid postdetection inference methods in changepoint analysis. On the other hand, there is a large literature on inference for *fixed* hypotheses in changepoint problems; we refer to works like Jandhyala *et al.* (2013).

2 | PRELIMINARIES

2.1 | Review: Changepoint algorithms

Below we describe the changepoint algorithms that we will study in this article. We will focus on formulations that run the algorithm for a given number of steps k. In what follows, we use the notation $\boldsymbol{y}_{a:b} = (y_a, y_{a+1}, \dots, y_b)$ and $\overline{\boldsymbol{y}}_{a:b} = (b-a+1)^{-1}\sum_{i=a}^{b} y_i$ for a vector $\boldsymbol{y}$. Similarly, for a set $\boldsymbol{I}$, $\overline{y}_I = |\boldsymbol{I}|^{-1}\sum_{i \in I} y_i$.

2.1.1 | Binary segmentation

Given a data vector $\boldsymbol{y} \in \mathbb{R}^n$, the k-step BS algorithm (see Fryzlewicz, 2014 and refers within) sequentially splits the data based on the cumulative sum (CUSUM) statistics, defined below. At a step $\ell = 1, \dots, k$, let $\widehat{\boldsymbol{b}}_{1:(\ell-1)}$ be the changepoints estimated so far, and let $\boldsymbol{I}_j$, $j = 1, \dots, \ell$ be the partition of $\{1, \dots, n\}$ induced by $\widehat{\boldsymbol{b}}_{1:(\ell-1)}$. Throughout this article, we use the convention that for $\ell = 1, \boldsymbol{I}_1 = \{1, \dots, n\}$. Intervals of length 1 are discarded. Let s_j and e_j be the start and end indices of $\boldsymbol{I}_j$. The next changepoint $\widehat{b}_\ell$ and maximizing interval $\widehat{j}_\ell$ are chosen to maximize the absolute CUSUM statistic,

$$\left\{\widehat{j}_\ell, \widehat{b}_\ell\right\} = \underset{\substack{j \in \{1,\dots,\ell-1\} \\ b \in \{s_j,\dots,e_j-1\}}}{\operatorname{argmax}} \left|\boldsymbol{g}^T_{(s_j,b,e_j)}\boldsymbol{y}\right|,$$

$$\text{where} \quad \boldsymbol{g}^T_{(s,b,e)}\boldsymbol{y} = \sqrt{\frac{1}{\frac{1}{|e-b|} + \frac{1}{|b+1-s|}}}\left(\overline{y}_{(b+1):e} - \overline{y}_{s:b}\right). \tag{3}$$

Additionally, the direction $\widehat{d}_\ell$ of the new changepoint is calculated by the sign of the maximizing absolute CUSUM statistic, $\widehat{d}_\ell = \text{sign}(\boldsymbol{g}^T_{(s_j,b_\ell,e_j)}\boldsymbol{y})$ for $j = \widehat{j}_{\ell+1}$.

2.1.2 | Wild binary segmentation

The k-step WBS algorithm (Fryzlewicz, 2014) is a modification of BS that calculates CUSUM statistics over randomly drawn segments of the data. Denote by $\boldsymbol{w} = \{\boldsymbol{w}_1, \dots, \boldsymbol{w}_B\} = \{(s_1, \dots, e_1), \dots, (s_B, \dots, e_B)\}$ a set of B uniformly randomly drawn intervals with $1 \le s_i < e_i \le n$, $i = 1, \dots, B$. At a step $\ell = 1, \dots, k$, let J_ℓ to be the index set of the intervals in $\boldsymbol{w}$ which do not intersect with the changepoints $\widehat{\boldsymbol{b}}_{1:(\ell-1)}$ estimated so far. The next changepoint $\widehat{b}_\ell$ and the maximizing interval $\widehat{j}_\ell$ are obtained by

$$\left\{\widehat{j}_\ell, \widehat{b}_\ell\right\} = \underset{\substack{j \in J_\ell \\ b \in \{s_j,\dots,e_j-1\}}}{\operatorname{argmax}} \left|\boldsymbol{g}^T_{(s_j,b,e_j)}\boldsymbol{y}\right|,$$

where $\boldsymbol{g}^T_{(s,b,e)}\boldsymbol{y}$ is defined in (3). Similar to BS, the direction of the changepoint $\widehat{d}_\ell$ is defined by the sign of the maximizing absolute CUSUM statistic. Fryzlewicz (2014) shows that the theoretical guarantees for WBS is strictly better than that for BS. However, while both can estimate the true changepoints asymptotically in a theoretic sense, both are prone to mistakes with finite data. This necessitates the need to develop valid inferential tools to prune the estimated changepoints.

We also consider CBS and the comparisons to the fused lasso (FL) in this article, but defer their discussions to Web Appendix C for brevity.

2.2 | Review: Post-selection inference

We briefly review post-selection inference as developed in Fithian *et al.* (2014) and related work, adapted for changepoint problems. For clarity, we notationally distinguish between a random vector $\boldsymbol{Y}$ distributed as in (1), and $\boldsymbol{y}_{\text{obs}}$, a single data vector we observe for changepoint analysis. When a changepoint algorithm—such as BS, WBS, or CBS —is applied to the data $\boldsymbol{y}_{\text{obs}}$, it selects a particular changepoint model $M(\boldsymbol{y}_{\text{obs}})$. The specific forms of such models are described in Section 3.1; for now, we may loosely think of $M(\boldsymbol{y}_{\text{obs}})$ as the changepoint locations and directions estimated by the algorithm on $\boldsymbol{y}_{\text{obs}}$, the data at hand. Post-selection inference revolves around the selective distribution, that is, the law of

$$\boldsymbol{v}^T\boldsymbol{Y} \mid (M(\boldsymbol{Y}) = M(\boldsymbol{y}_{\text{obs}}),\ q(\boldsymbol{Y}) = q(\boldsymbol{y}_{\text{obs}})), \tag{4}$$

under the null hypothesis $H_0 : \boldsymbol{v}^T\theta = 0$, for any vector $\boldsymbol{v}$ that is a measurable function of $M(\boldsymbol{y}_{\text{obs}})$, such as in (2). Here, $q(\boldsymbol{Y})$ is a vector of sufficient statistic of nuisance parameters that need to be conditioned on in order to tractably compute inferences based on (4). The explicit form of $q(\boldsymbol{Y})$ differs based on the assumptions imposed on θ under the null model. Broadly, there are two classes of null models we may study: saturated and selected models (Fithian *et al.*, 2014). As shown in the literature, computationally, in either null models, it is important for the selection event $\{\boldsymbol{y} : M(\boldsymbol{y}) = M(\boldsymbol{y}_{\text{obs}})\}$ be polyhedral. This is described in detail in Section 3.1, where we show that this holds for BS, WBS, and CBS.

2.2.1 | Saturated model

The *saturated model* assumes that $\boldsymbol{Y}$ is distributed as in (1) with known error variance σ^2 and assumes nothing about the mean vector θ. We set $q(\boldsymbol{Y}) = \Pi_{\boldsymbol{v}}^{\perp}\boldsymbol{Y}$, the projection of $\boldsymbol{Y}$ onto the hyperplane orthogonal to $\boldsymbol{v}$. The selective distribution (4) then becomes the law of

$$\boldsymbol{v}^T\boldsymbol{Y} \mid \left(M(\boldsymbol{Y}) = M(\boldsymbol{y}_{\text{obs}}),\ \Pi_{\boldsymbol{v}}^{\perp}\boldsymbol{Y} = \Pi_{\boldsymbol{v}}^{\perp}\boldsymbol{y}_{\text{obs}}\right). \tag{5}$$

2.2.2 | Selected model

The *selected model* again assumes that Y follows (1), but additionally assumes that the mean vector θ is piecewise constant with changepoints at the sorted estimated locations $\widehat{\boldsymbol{c}}_{1:k} = \widehat{\boldsymbol{c}}_{1:k}(\boldsymbol{y}_{\text{obs}})$, assuming we have run our changepoint algorithm for k steps. That is, letting s_j and e_j denote the start and end index of interval I_j, we assume

$$\theta_{s_j} = \cdots = \theta_{e_j}, \quad j \in \{1, \ldots, k+1\}.$$

Under this assumption, the law of $\boldsymbol{Y}$ becomes a $(k+1)$-parameter Gaussian distribution. Additionally, with the contrast vector $\boldsymbol{v}_j$ defined as in (2), for any fixed $j = 1, \ldots, k+1$, the quantity $\boldsymbol{v}_j^T\theta$ of interest is simply the difference between two of the parameters in this distribution. Specifically, let $\mathcal{I}_j = \{1, \ldots, k+1\}\backslash\{j, j+1\}$. Assuming σ^2 is known, the sufficient statistics $q(\boldsymbol{Y})$ are then the sample averages of the appropriate data segments, and the selective distribution (4) becomes the law of

$$\left(\overline{Y}_{\boldsymbol{I}_{j+1}} - \overline{Y}_{\boldsymbol{I}_j}\right) \Bigm| \Bigl(M(\boldsymbol{Y}) = M(\boldsymbol{y}_{\text{obs}}),\ \overline{Y}_{\boldsymbol{I}_j \cup \boldsymbol{I}_{j+1}} = \left(\overline{y}_{\text{obs}}\right)_{\boldsymbol{I}_j \cup \boldsymbol{I}_{j+1}},\ \overline{Y}_{\boldsymbol{I}_\ell} = \left(\overline{y}_{\text{obs}}\right)_{\boldsymbol{I}_\ell} \text{ for } \ell \in \mathcal{I}_j\Bigr). \tag{6}$$

The appeal of the selected model is that we can properly treat σ^2 as unknown; in this case, we must only additionally condition on the Euclidean norm of y_{obs} to account for this nuisance parameter, and the selective distribution (4) becomes the law of

$$\left(\overline{Y}_{\boldsymbol{I}_{j+1}} - \overline{Y}_{\boldsymbol{I}_j}\right) \Bigm| \Bigl(M(\boldsymbol{Y}) = M(\boldsymbol{y}_{\text{obs}}),\ \overline{Y}_{\boldsymbol{I}_j \cup \boldsymbol{I}_{j+1}} = \left(\overline{y}_{\text{obs}}\right)_{\boldsymbol{I}_j \cup \boldsymbol{I}_{j+1}},\ \overline{Y}_{\boldsymbol{I}_\ell} = \left(\overline{y}_{\text{obs}}\right)_{\boldsymbol{I}_\ell} \text{ for } \ell \in \mathcal{I}_j, \|\boldsymbol{Y}\|_2 = \|\boldsymbol{y}_{\text{obs}}\|_2\Bigr). \tag{7}$$

3 | INFERENCE FOR CHANGEPOINT ALGORITHMS

We describe our contributions that enable post-selection inference for changepoint analyses, beginning with the form of model selection events. We then describe computational details for saturated and selected model tests and auxiliary randomization.

3.1 | Polyhedral selection events

We show that, for each of the BS and WBS algorithms, there is a parameterization for their models such that event $\{\boldsymbol{y} : M(\boldsymbol{y}) = M(\boldsymbol{y}_{\text{obs}})\}$ is a polyhedron of the form $\{\boldsymbol{y} : \boldsymbol{\Gamma}\boldsymbol{y} \geq \boldsymbol{0}\}$, for a matrix $\boldsymbol{\Gamma} \in \mathbb{R}^{m\times n}$ that depends on $M(\boldsymbol{y}_{\text{obs}})$, where we interpret the inequality $\boldsymbol{\Gamma}\boldsymbol{y} \geq \boldsymbol{0}$ componentwise. Throughout the description of the polyhedra for each algorithm, we display the number of rows in $\boldsymbol{\Gamma}$ since it loosely denotes how "complex" each model selection event is. Overall, for a fixed k, the number of rows in the $\boldsymbol{\Gamma}$ matrix for BS is linear in n, and $O(Bp)$ for WBS using intervals of length p. This number can grow faster than linear in n if $B \geq n$, which is recommended in practice (Fryzlewicz, 2014). All the proofs are provided in Web Appendix H.

3.1.1 | Selection event for BS

We define the model for the k-step BS estimator as

$$M^{\mathrm{BS}}_{1:k}(\boldsymbol{y}_{\mathrm{obs}}) = \left\{\widehat{\boldsymbol{b}}_{1:k}(\boldsymbol{y}_{\mathrm{obs}}),\ \widehat{\boldsymbol{d}}_{1:k}(\boldsymbol{y}_{\mathrm{obs}})\right\},$$

where $\widehat{\boldsymbol{b}}_{1:k}(\boldsymbol{y}_{\mathrm{obs}})$ and $\widehat{\boldsymbol{d}}_{1:k}(\boldsymbol{y}_{\mathrm{obs}})$ are the changepoint locations and directions when the algorithm is run on $\boldsymbol{y}_{\mathrm{obs}}$, as described in Section 2.1.

Proposition 1. Given any fixed $k \geq 1$ and $\boldsymbol{b}_{1:k}, \boldsymbol{d}_{1:k}$, we can explicitly construct $\boldsymbol{\Gamma}$ where

$$\left\{\boldsymbol{y} : M^{\mathrm{BS}}_{1:k}(\boldsymbol{y}) = \{\boldsymbol{b}_{1:k}, \boldsymbol{d}_{1:k}\}\right\} = \{\boldsymbol{y} : \boldsymbol{\Gamma}\boldsymbol{y} \geq \boldsymbol{0}\},$$

where $\boldsymbol{\Gamma}$ has $2\sum_{\ell=1}^{k}(n-\ell-1)$ rows.

3.1.2 | Selection event for WBS

We define the model of the k-step WBS estimator as

$$M^{\mathrm{WBS}}_{1:k}(\boldsymbol{y}_{\mathrm{obs}}, \boldsymbol{w}) = \left\{\widehat{\boldsymbol{b}}_{1:k}(\boldsymbol{y}_{\mathrm{obs}}),\ \widehat{\boldsymbol{d}}_{1:k}(\boldsymbol{y}_{\mathrm{obs}}),\ \widehat{\boldsymbol{j}}_{1:k}(\boldsymbol{y}_{\mathrm{obs}})\right\},$$

where $\boldsymbol{w}$ is the set of B intervals that the algorithm uses, $\widehat{\boldsymbol{b}}_{1:k}(\boldsymbol{y}_{\mathrm{obs}})$ and $\widehat{\boldsymbol{d}}_{1:k}(\boldsymbol{y}_{\mathrm{obs}})$ are the changepoint locations and directions, and $\widehat{\boldsymbol{j}}_{1:k}(\boldsymbol{y}_{\mathrm{obs}})$ are the maximizing intervals. Note that unlike BS, the maximizing intervals $\widehat{\boldsymbol{j}}_{1:k}$ are part of WBS's model.

Proposition 2. Given any fixed $k \geq 1$ and $\{\boldsymbol{w}, \boldsymbol{b}_{1:k}, \boldsymbol{d}_{1:k}, \boldsymbol{j}_{1:k}\}$, we can explicitly construct $\boldsymbol{\Gamma}$ where

$$\left\{\boldsymbol{y} : M^{\mathrm{WBS}}_{1:k}(\boldsymbol{y}, \boldsymbol{w}) = \{\boldsymbol{b}_{1:k}, \boldsymbol{d}_{1:k}, \boldsymbol{j}_{1:k}\}\right\} = \left\{\boldsymbol{y} : \boldsymbol{\Gamma}\boldsymbol{y} \geq \boldsymbol{0}\right\}.$$

The number of rows in $\boldsymbol{\Gamma}$ will vary depending on the configuration of $\boldsymbol{w}$ and $\boldsymbol{b}_{1:k}$, but if each of the B intervals in w has length p, it will be at most $2\sum_{\ell=1}^{k}((B-\ell)\cdot(p-1)+(p-2))$.

In Web Appendix C, we additionally state the analogous results for the CBS model, as well as review the results for the FL model as derived in Hyun *et al.* (2018) for comparison.

3.2 | Computation of *p*-values

Given a precise description of the polyhedral selection event $\{\boldsymbol{y} : M(\boldsymbol{y}) = M(\boldsymbol{y}_{\mathrm{obs}})\}$, we can describe the methods to compute the p-value, that is the tail probability of the selective distributions described in Section 2.2. Without loss of generality, all of our descriptions will be specialized to testing the null hypothesis of $H_0 : \boldsymbol{v}^T\theta = 0$ against the one-sided alternative $H_1 : \boldsymbol{v}^T\theta > 0$. For saturated model tests, this exact calculation has been developed in previous works and we review it as it is relevant to our contributions on increasing its power. For selected model tests, we develop a new hit-and-run sampler. We emphasize, as stated in works like Fithian *et al.* (2014), the following methods provide p-values that are *exactly* uniformly distributed under the null hypothesis with respect to n, unlike those from t-tests.

3.2.1 | Saturated model tests: Exact formulae

As shown in Tibshirani *et al.* (2018) and related work, the saturated selective distribution (5) has a particularly computationally convenient distribution when $\boldsymbol{Y}$ is Gaussian and the model selection event $\{\boldsymbol{y} : M(\boldsymbol{y}) = M(\boldsymbol{y}_{\mathrm{obs}})\}$ is a polyhedral set in $\boldsymbol{y}$. In this case, the law of (5) is a *truncated Gaussian* (TG), whose truncation limits depend only on $\Pi^{\perp}_{\boldsymbol{v}}\boldsymbol{y}_{\mathrm{obs}}$ and can be computed explicitly. Its tail probability can be computed in closed form (without Monte Carlo sampling). That is, the probability that $\boldsymbol{v}^T\boldsymbol{Y} \geq \boldsymbol{v}^T\boldsymbol{y}_{\mathrm{obs}}$ under the law of (5) is exactly equal to

$$\left\{\Phi(\mathcal{V}_{\mathrm{up}}/\tau) - \Phi(\boldsymbol{v}^T\boldsymbol{y}_{\mathrm{obs}}/\tau)\right\}/\left\{\Phi(\mathcal{V}_{\mathrm{up}}/\tau) - \Phi(\mathcal{V}_{\mathrm{lo}}/\tau)\right\} \quad (8)$$

where $\Phi(\cdot)$ represents the standard Gaussian cumulative distribution function (CDF), $\tau = \sigma^2\|\boldsymbol{v}\|_2^2$, $\boldsymbol{\rho} = \boldsymbol{\Gamma}\boldsymbol{v}/\|\boldsymbol{v}\|_2^2$ and

$$\mathcal{V}_{\mathrm{lo}} = \boldsymbol{v}^T\boldsymbol{y}_{\mathrm{obs}} - \min_{j:\rho_j>0}(\boldsymbol{\Gamma}\boldsymbol{y}_{\mathrm{obs}})_j/\rho_j, \quad \text{and}$$
$$\mathcal{V}_{\mathrm{up}} = \boldsymbol{v}^T\boldsymbol{y}_{\mathrm{obs}} - \max_{j:\rho_j<0}(\boldsymbol{\Gamma}\boldsymbol{y}_{\mathrm{obs}})_j/\rho_j. \quad (9)$$

The statistic in (8) is commonly referred as the TG statistic. Since this statistic is a pivot and lies between [0,1], it is the p-value used for the saturated model test.

3.2.2 | Selected model tests: hit-and-run sampling

To compute the p-value for selected model tests, Fithian *et al.* (2015) proposed a hit-and-run strategy for sampling from the distribution for the known σ^2 setting (6). This was implemented by the authors, and we briefly review the details in Web Appendix D. For the unknown σ^2 setting, Fithian *et al.* (2014) developed an importance sampling strategy for sampling the distribution (7). However, we develop an alternative and intuitive hit-and-run strategy can be adapted to the unknown σ^2 setting and implement this as a new algorithm, explained next.

Given a changepoint $j \in \{1, \ldots, k\}$, observe that we can design a segment test contrast $\boldsymbol{v}$ where sampling from (7) is equivalent to sampling uniformly from the set

$$\Big\{\boldsymbol{v}^T\boldsymbol{Y} : M(\boldsymbol{Y}) = M(\boldsymbol{y}_{\text{obs}}),\ \|\boldsymbol{Y}\|_2 = \|\boldsymbol{y}_{\text{obs}}\|_2,\ \overline{Y}_{\boldsymbol{I}_j \cup \boldsymbol{I}_{j+1}} = \left(\overline{y}_{\text{obs}}\right)_{\boldsymbol{I}_j \cup \boldsymbol{I}_{j+1}},\ \overline{Y}_{\boldsymbol{I}_\ell} = \left(\overline{y}_{\text{obs}}\right)_{\boldsymbol{I}_\ell} \text{ for } \ell \in \mathcal{I}_j\Big\}. \quad (10)$$

Note that the above set no longer depends on $\boldsymbol{\theta}$ or σ^2. This is because we conditioned all the relevant sufficient statistics under the selected model. Our hit-and-run sampler then sequentially draws samples $\boldsymbol{v}^T\boldsymbol{Y}$ from the above set. For brevity, the explicit algorithm is deferred to Web Appendix D and leverages explicit formulas computing the intersection of two-dimensional circles with polytopes.

3.3 | Randomization and marginalization

We apply the ideas of auxiliary randomization in Tian and Taylor (2018) to improve the power of post-selection inference for changepoint algorithms. We investigate two specific forms of randomization—randomization over additive noise or over random intervals—specialized for saturated models. We note that similar randomization of selected model inferences is also possible but is doubly computationally burdensome.

3.3.1 | Marginalization over additive noise

Tian and Taylor (2018) shows that performing inference based on the selected model $M(\boldsymbol{y}_{\text{obs}} + \boldsymbol{w}_{\text{obs}})$ where $\boldsymbol{w}_{\text{obs}}$ is additive noise and then marginalizing over $\boldsymbol{W}$ leads to improved power. Here, $\boldsymbol{w}_{\text{obs}}$ is a realization of a random component $\boldsymbol{W}$ sampled from $\mathcal{N}(\boldsymbol{0}, \sigma^2_{\text{add}}\boldsymbol{I}_n)$, where $\sigma^2_{\text{add}} > 0$ is set by the user. Fithian *et al.* (2014) provide a mathematical framework for pursuing such randomization, stating that less conditioning results in an increase in Fisher information. For additive noise, the above model selection event is

$$\{\boldsymbol{y} : \boldsymbol{\Gamma}(\boldsymbol{y} + \boldsymbol{w}_{\text{obs}}) \geq \boldsymbol{0}\} = \{\boldsymbol{y} : \boldsymbol{\Gamma}\boldsymbol{y} \geq -\boldsymbol{\Gamma}\boldsymbol{w}_{\text{obs}}\}.$$

This suggests the following idea of using existing machinery to formulate the polyhedron formed by the model selection event based on perturbed data $\boldsymbol{y}_{\text{obs}} + \boldsymbol{w}_{\text{obs}}$.

Porting the ideas of Tian and Taylor (2018) to our setting, to test the one-sided null hypothesis $H_0 : \boldsymbol{v}^T\boldsymbol{\theta} = 0$, we want to compute the following tail probability of the marginalized selective distribution:

$$T(\boldsymbol{y}_{\text{obs}}, \boldsymbol{v}) = \mathbb{P}\left(\boldsymbol{v}^T\boldsymbol{Y} \geq \boldsymbol{v}^T\boldsymbol{y}_{\text{obs}} \,\middle|\, (M(\boldsymbol{Y} + \boldsymbol{W}) = M(\boldsymbol{y}_{\text{obs}} + \boldsymbol{W}),\ \Pi^{\perp}_{\boldsymbol{v}}\boldsymbol{Y} = \Pi^{\perp}_{\boldsymbol{v}}\boldsymbol{y}_{\text{obs}})\right). \quad (11)$$

It is hard to directly compute this. However, the formulas in (8) and (9) give us exact formulas to compute the non-marginalized tail probabilities,

$$T(\boldsymbol{y}_{\text{obs}}, \boldsymbol{v}, \boldsymbol{w}_{\text{obs}}) = \mathbb{P}\left(\boldsymbol{v}^T\boldsymbol{Y} \geq \boldsymbol{v}^T\boldsymbol{y}_{\text{obs}} \,\middle|\, (M(\boldsymbol{Y} + \boldsymbol{W}) = M(\boldsymbol{y}_{\text{obs}} + \boldsymbol{W}),\ \Pi^{\perp}_{\boldsymbol{v}}\boldsymbol{Y} = \Pi^{\perp}_{\boldsymbol{v}}\boldsymbol{y}_{\text{obs}}, = \boldsymbol{W} = \boldsymbol{w}_{\text{obs}})\right).$$

The following proposition shows that we can compute $T(\boldsymbol{y}_{\text{obs}}, \boldsymbol{v})$ by reweighting instances of $T(\boldsymbol{y}_{\text{obs}}, \boldsymbol{v}, \boldsymbol{w}_{\text{obs}})$ via importance sampling. Here, let $E_1 = \mathbb{1}[M(\boldsymbol{Y} + \boldsymbol{W}) = M(\boldsymbol{y}_{\text{obs}} + \boldsymbol{W})]$ and $E_2 = \mathbb{1}[\Pi^{\perp}_{\boldsymbol{v}}\boldsymbol{Y} = \Pi^{\perp}_{\boldsymbol{v}}\boldsymbol{y}_{\text{obs}}]$.

Proposition 3. Let Ω denote the support of the random component $\boldsymbol{W}$. If the distribution of $\boldsymbol{W}$ is independent of the random event E_2, *(11)* can be exactly computed as

$$T(\boldsymbol{y}_{\text{obs}}, \boldsymbol{v}) = \int_{\Omega} T(\boldsymbol{y}_{\text{obs}}, \boldsymbol{v}, \boldsymbol{w}_{\text{obs}}) \cdot a(\boldsymbol{w}_{\text{obs}})\, dP_{\boldsymbol{W}}(\boldsymbol{w}_{\text{obs}}) = \frac{\int_{\Omega} \Phi\left(\mathcal{V}_{up}/\tau\right) - \Phi\left(\boldsymbol{v}^T\boldsymbol{y}_{\text{obs}}/\tau\right) dP_{\boldsymbol{W}}(\boldsymbol{w}_{\text{obs}})}{\int_{\Omega} \Phi\left(\mathcal{V}_{up}/\tau\right) - \Phi(\mathcal{V}_{lo}/\tau)\, dP_{\boldsymbol{W}}(\boldsymbol{w}_{\text{obs}})}. \quad (12)$$

where the weighting factor is $a(\boldsymbol{w}_{\text{obs}}) = \mathbb{P}(\boldsymbol{W} = \boldsymbol{w}_{\text{obs}} | E_1, E_2)/\mathbb{P}(\boldsymbol{W} = \boldsymbol{w}_{\text{obs}})$.

The first equality in (12) demonstrates the reweighting of $T(\boldsymbol{y}_{\text{obs}}, \boldsymbol{v}, \boldsymbol{w}_{\text{obs}})$, but the second equality gives a sampling strategy where we approximate the integrals. Specifically, we sample T different instances of $\boldsymbol{w}_{\text{obs}}$ and compute $k(\boldsymbol{w}_{\text{obs}})$ and $g(\boldsymbol{w}_{\text{obs}})$, denoting the integrand of the last term's numerator and denominator in (12), respectively. Then the approximate for the tail probability (12) is the ratio between the summation of all the instances of $k(\boldsymbol{w}_{\text{obs}})$ and of all instances of $g(\boldsymbol{w}_{\text{obs}})$. (Observe that the calculations of $\mathcal{V}_{\text{up}}$ and $\mathcal{V}_{\text{lo}}$ now involve $\boldsymbol{w}_{\text{obs}}$.) For brevity, we defer the explicit algorithm to Web Appendix D.

3.3.2 | Marginalization over WBS intervals

In contrast to the above setting where $\boldsymbol{W}$ represents Gaussian noise, in WBS described in Section 2.1, W represents

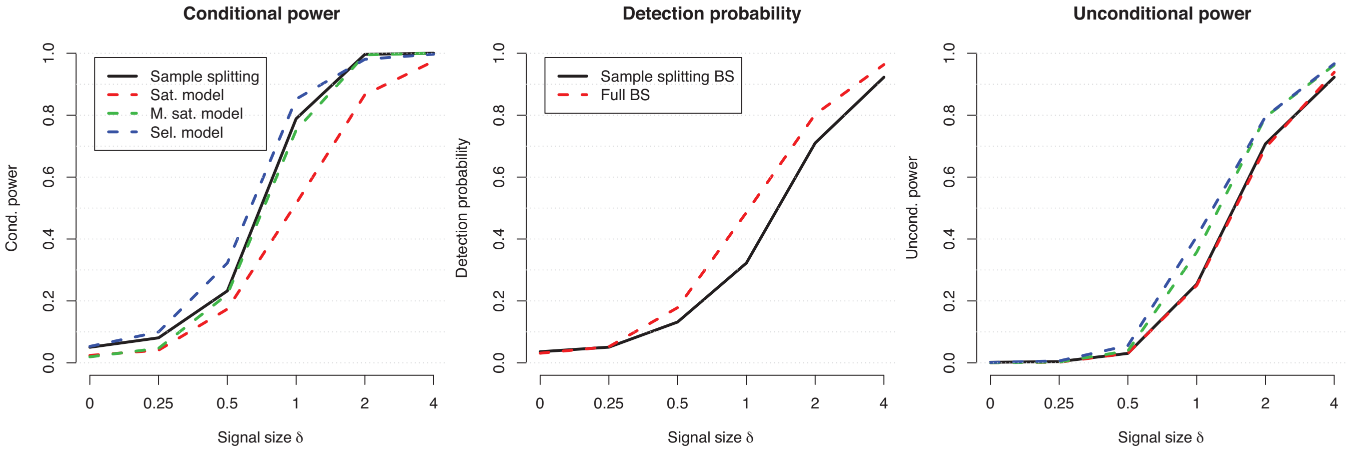


FIGURE 2 Simulation results displaying the empirical power for plain saturated model test (red dashed), additive noise marginalized saturated model test (green dashed), and selected model test with unknown σ^2 (blue dashed), and t-test for equality of mean after sample splitting (black solid), performed after a two-step BS. Here, we generate Gaussian data where θ has two true changepoints, and δ (the x-axis) parameterizes the signal (i.e., the difference between the piecewise constant segments). A larger δ means an easier simulation setting. We try more than 250 trials at each value of δ. (Left): Conditional power across all four methods, defined as the number of correctly detected and rejected changepoints instances divided by the number of corrected detected changepoints. (Middle): Detection probability for the BS, defined as the number of corrected detected changepoints divided by total number of trials. (Right): Unconditional power, defined by multiplying the conditional power curve and its relevant detection probability curve. Together, we see selected model tests and marginalized saturated model tests have higher unconditional power for larger δ than sample splitting and plain saturated model tests. More details behind the simulation are given in Web Appendix E. Note: this figure appears in color in the electronic version of this article, and any mention of color refers to that version

the set of B randomly drawn intervals. Observe that Proposition 3 still applies to this setting, where $M(\boldsymbol{y}_{\text{obs}} + \boldsymbol{w}_{\text{obs}})$ is now replaced with $M(\boldsymbol{y}_{\text{obs}}, \boldsymbol{w}_{\text{obs}})$, as described in Section 3.1. However, unlike the additive noise setting, the maximizing intervals $\hat{\boldsymbol{j}}_{1:k}$ in the model $M(\boldsymbol{y}_{\text{obs}}, \boldsymbol{w}_{\text{obs}})$ are embedded in the construction of the matrix $\boldsymbol{\Gamma}$ representing the polyhedra. This prevents a naive sampling of B new intervals. To overcome this, let $\{\boldsymbol{W}_{\hat{j}_1}, \dots, \boldsymbol{W}_{\hat{j}_k}\}$ be the maximizing intervals. We sample a new set of all other intervals, W_ℓ for $\ell \in \{1, \dots, B\} \backslash \{\hat{j}_1, \dots, \hat{j}_k\}$. Specifically, for each of such intervals $\boldsymbol{W}_\ell = (s_\ell, \dots, e_\ell)$, the indices s_ℓ and e_ℓ are sampled uniformly between 1 to n where $s_\ell < e_\ell$. After all $B - k$ intervals are resampled, a check is performed to ensure that $\{\boldsymbol{W}_{\hat{j}_1}, \dots, \boldsymbol{W}_{\hat{j}_k}\}$ are still the maximizing intervals when WBS is applied again to $\boldsymbol{y}_{\text{obs}}$. The full algorithm is also deferred to Web Appendix D.

3.4 | Simulation results

We provide a brief overview of simulation-based results that demonstrate the utility of our inferential tools, highlighting that our selected model tests and marginalized saturated model tests empirically higher power than plain saturated model tests and t-tests based on sample splitting, described in Section 1.2. In these simulations, we generate Gaussian data with θ having two changepoints with varying signal sizes and try different testing procedures based on two-step BS. We then compute the unconditional power of each test, defined as how often a test successfully detects the location of the true changepoint and then successfully rejects the null hypothesis. We make two observations from our results (Figure 2). First, the marginalized saturated model tests have substantially higher power over their plain counterparts, verifying the theoretical intuitions in Tian and Taylor (2018). Second, sample splitting leads to a lower unconditional power compared to our inferential tools primarily because sample splitting uses only half the data to detect changepoints, leading to a lower detection probability. Another competing method is the post-selection inference tools for the 1d fused lasso developed in Hyun *et al.* (2018). For brevity, we defer extensive simulations showing uniform null p-values, and those comparing the power among these methods, as well as additional details and discussions, to Web Appendix E.

4 | COPY NUMBER VARIATION APPLICATION

In this section, we apply our post-selection inference tools to study their empirical performance on both the aCGH data from Snijders *et al.* (2001) (introduced in Section 1) as well as sequencing data from Botton *et al.*

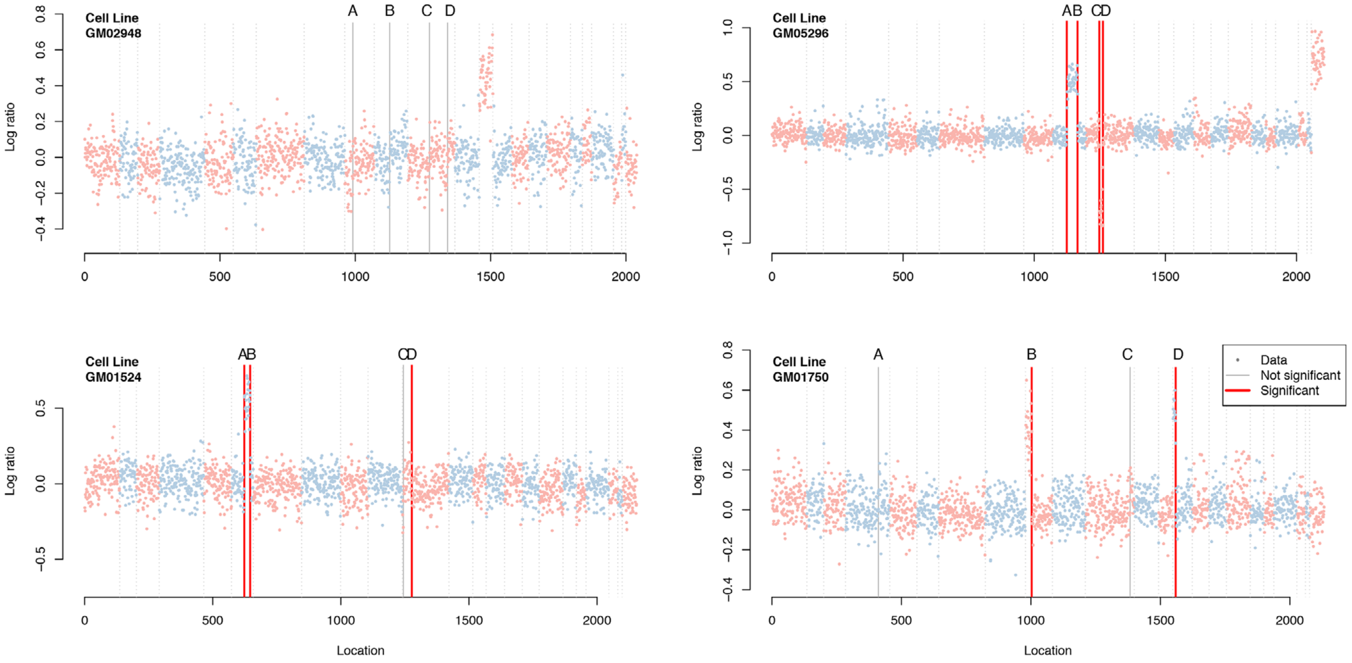


FIGURE 3 Pre-cut changepoint inference using saturated model tests for 4-step WBS marginalized over random intervals conducted on four cell lines across all 23 chromosomes, from Snijders *et al.* (2001). Data points are colored in two alternating tones, to visually depict the chromosomal boundaries. The x-axis denotes the relative index of the genome position while the y-axis denotes the measured $\log_2$ copy number. For each cell line, the letters A through D denote the estimated changepoints, $\hat{c}_1$ through $\hat{c}_4$ respectively. The bolded gray (or red) horizontal lines denote changepoints that were rejected (or not rejected) under the null hypothesis $H_0 : \boldsymbol{v}^T\boldsymbol{\theta} = 0$ at a Type-I error control level $\alpha = 0.05$ after Bonferroni-correction. (Top left): The analysis for the cell line GM02948 with no significant changepoints. This matches external karotyping results, marking a trisomy (i.e., increase in copy number) of the entire chromosome 13, meaning there are no CNVs within any chromosome. (Top right): The analysis for the cell line GM05296 with 4 significant changepoints, all at chromosome 10 and 11. This matches external karotyping results, marking a trisomy al chromosome 10 and a monosomy (i.e., decrease in copy number) at chromosome 11. (Bottom left): The analysis for the cell line GM01524 with 3 significant changepoints, at chromosome 6 and 9. The external karotyping results only reveal that the CNV at chromosome 6 is true (i.e., 6q15 to 6q25). (Bottom right): The analysis for the cell line GM01750 with 2 significant changepoints, at chromosome 9 and 14. This matches external karotyping results, marking trisomies at chromosome 9 and 14. Note: this figure appears in color in the electronic version of this article, and any mention of color refers to that version

(2013). Both datasets are specifically chosen since there are scientifically meaningful ways to quantify whether or not our inference tools were successful. Specifically, we first demonstrate in Section 4.1 that our inferential tools can be used for filtering changepoints by comparing our results applied on the aCGH data to external karotyping results. Then in Section 4.2, we show that we can adapt our inferential tools to handle heavy-tailed noise appearing in CNV analyses by demonstrating that p-values under the null hypothesis based on pseudo-real datasets are correctly distributed as uniform. Finally in Section 4.3, we show that while sequencing data (a newer technology based on counting DNA fragments) has technical variability that can differ dramatically from aCGH data (an older technology based on measuring light intensities), by preprocessing the sequencing data appropriately, our inferential tools shown to work for aCGH data can yield similar results on sequencing data. Together, these empirical results demonstrate that our inferential tools can be reliably used for automated filtering of changepoints for CNV analyses.

4.1 | Analysis on Snijders dataset

We first extend our analysis of the Snijders dataset from Section 1 by detecting and inferring about CNVs across the entire genome of different cell lines. These datasets are ideal because a ground truth—external karyotyping results—exists from the original study, which we compare our inferential results against. These cell lines originate from fibroblast cells and contain over 2000 probe measurements across all 23 chromosomes.

In our analysis, we use a four-step WBS and perform marginalized saturated model tests across the genome on each of the four cell lines (GM02948, GM05296, GM01524, and GM01750), shown in Figure 3. We precut at chromosome boundaries since the ordering of chromosomes 1–23

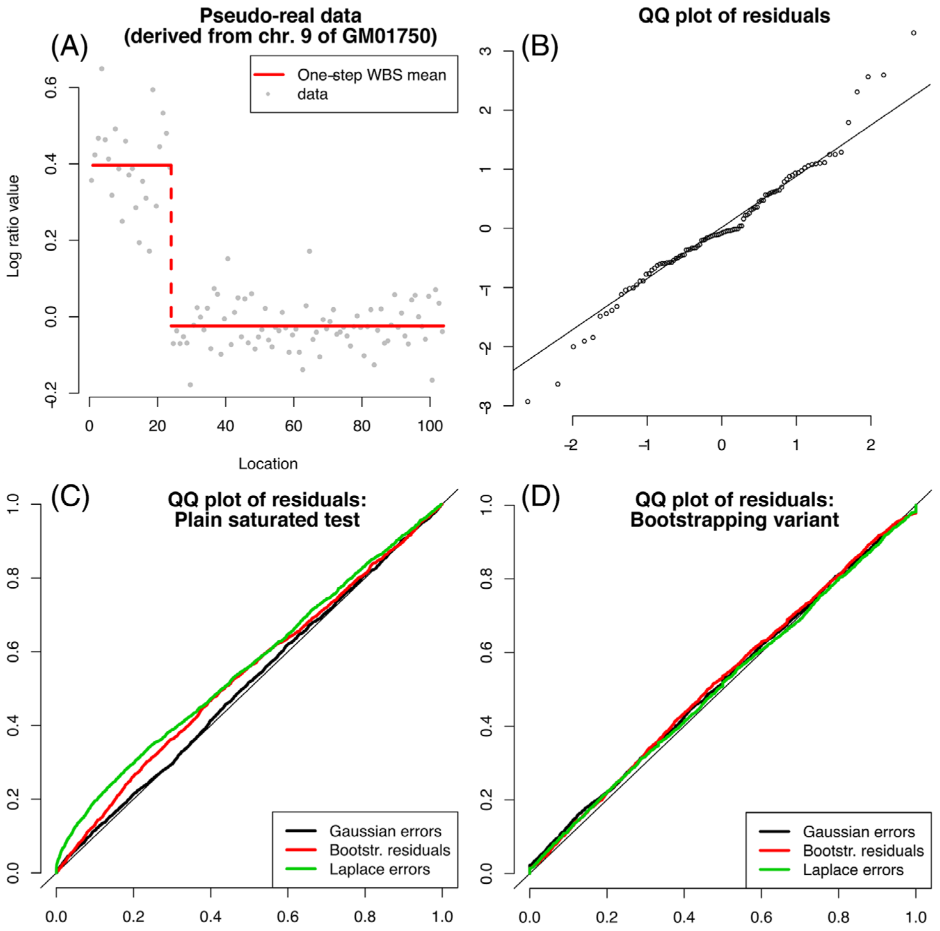


FIGURE 4 (A) Bootstrapped residuals added to the artificially constructed mean, generated from chromosome 9 in GM01750. (B): QQ plot of residuals. The remaining 2 panels show the p-values of saturated model tests under three different noise models, Gaussian (black), bootstrapped residuals (red) and Laplacian (green). (C): QQ-plot of p-values derived from plain saturated model tests. Exactly valid null p-values would follow the theoretical $U(0,1)$ distribution (i.e., valid Type-I control) while optimistic (superuniform) p-values would lie above the diagonal (i.e., invalid Type-I control). (D): QQ-plot of p-values derived from our modified bootstrap substitution method that involves bootstrapping $\boldsymbol{y} - \widehat{\theta}$ instead of $\boldsymbol{y} - \bar{\boldsymbol{y}}$. In Figures B-D, the X axis shows expected quantiles, and Y axis shows theoretical quantiles. Note: this figure appears in color in the electronic version of this article, and any mention of color refers to that version

is essentially arbitrary, meaning we initialize WBS with changepoints at the boundaries between chromosomes, but do not consider these as estimated changepoints. We test at a significance level $\alpha = 0.05$ after the Bonferroni correction. This is described in Web Appendix B. We provide additional details of our analyses in Web Appendix G.

Our results, shown in Figure 3, demonstrate that while WBS estimates changepoints at various locations across the genome, the significant changepoints determined by our inferential tools largely agree with external karotyping results. For example, in GM02948, external karotyping results show there is no CNV that occurs within a chromosome and, likewise, our inferential results filtered out all four estimated changepoints. Likewise, in GM05296 and GM01750, external karotyping match exactly with which changepoints were deemed as significant, and the remaining changepoints were filtered out. In GM01524, however, external karotyping results only validates the changepoints we estimated at chromosome 6. The analysis on these four cell lines demonstrates that our inferential tools can be used effectively to filter the changepoints across a wide range of scenarios.

4.2 | Follow-up analysis on Snijders dataset for impact of heavy tails

While our inferential tools are designed for Gaussian data, we demonstrate that they can be adapted to handle heavy-tailed data by bootstrapping the residuals instead of explicitly calculating the tail probabilities. For example, we can observe the heavy-tailed nature in the Snijders dataset by focusing on chromosome 9 in GM01750 (Figure 4). A QQ plot of the residuals $\boldsymbol{r}$ (computed by taking the

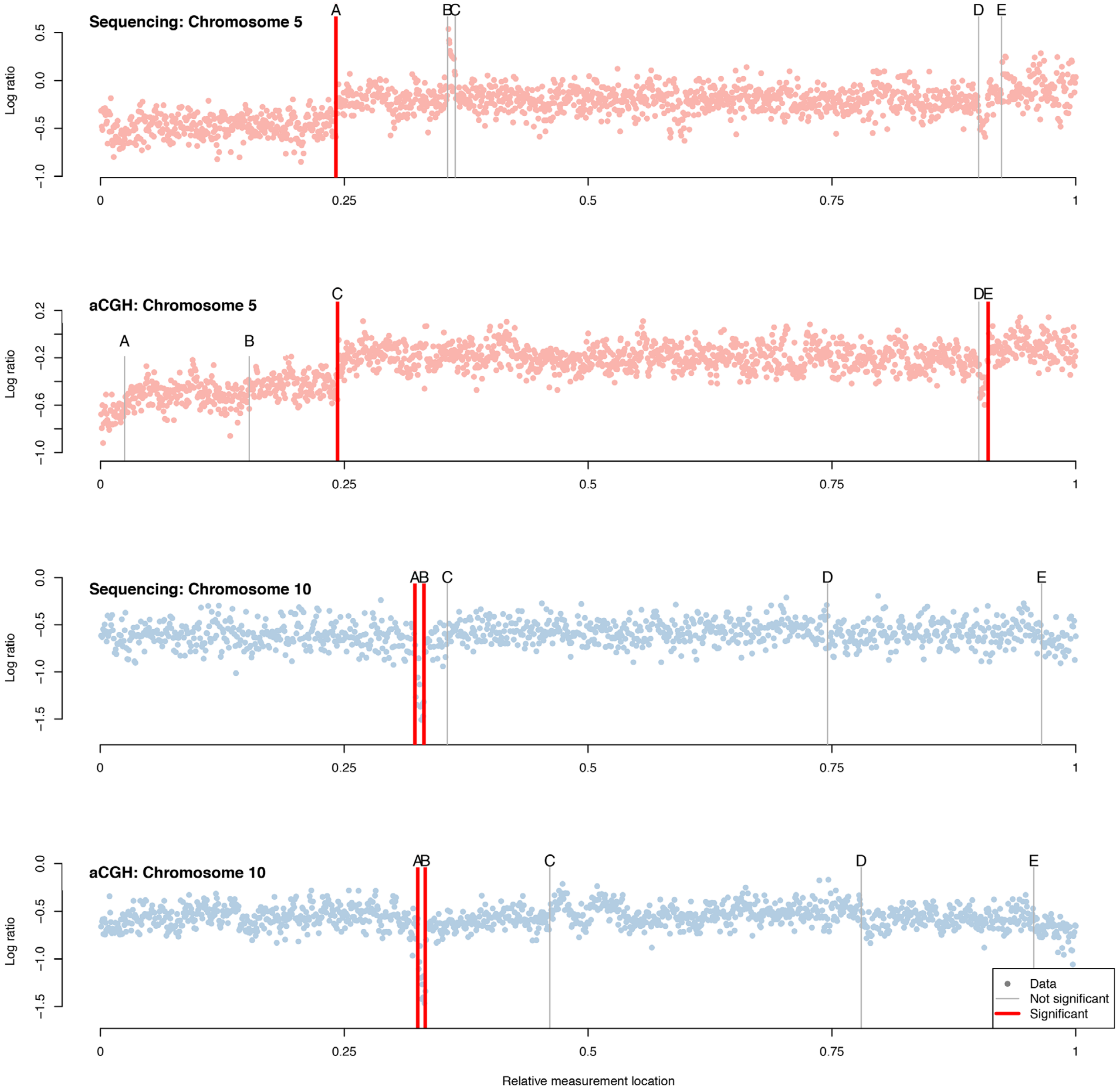


FIGURE 5 Changepoint inference using saturated model tests for five-step WBS marginalized over random intervals conducted over various chromosomes. In all the panels, the letters A through E denote the estimated changepoints, $\widehat{c}_1$ through $\widehat{c}_4$, respectively. The bolded gray (or red) horizontal lines denote changepoints that were rejected (or not rejected) under the null hypothesis. The x-axis denotes the relative index of the genome position (normalized to be between 0 and 1), while the y-axis denotes the measured $\log_2$ copy number. (Top two panels): Analyses for chromosome 5 using sequencing data on the top and aCGH data on the bottom, where a significant changepoint around location 0.25 is shown (A and C, respectively). However, in the aCGH analysis, changepoint E is deemed significant, which does not have a counterpart in the sequencing analysis. (Bottom two panels): Similar analysis but for chromosome 10, where changepoints A and B are marked significant between both data sources. These changepoints around location 0.3 is validated by FISH results (Talevich *et al.*, 2016). Note: This figure appears in color in the electronic version of this article, and any mention of color refers to that version

difference between the observation and the mean of its corresponding segment) suggests that the noise has heavier tails than a Gaussian (Figure 4(B)) and are closer in distribution to a Laplacian. Hence, we design a study based on pseduo-real datasets with heavy tails derived from this chromosome to investigate whether or not we can obtain uniform p-values under the null hypothesis in this setting.

To design this numerical study, we use the following procedure to generate pseudo-real datasets. Using the model $\boldsymbol{y} = \boldsymbol{\theta} + \boldsymbol{\epsilon}$, we add the noise variable ϵ in three different ways:

(1) Gaussian noise $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$,
(2) Laplace noise $\epsilon \sim \text{Laplace}(\mathbf{0}, \sigma/\sqrt{2} \cdot \boldsymbol{I}_n)$, and
(3) Bootstrapped residuals, $\epsilon = b(\boldsymbol{r})$, where $b(\cdot)$ samples the residuals $\boldsymbol{r}$ with replacement.

To simplify this study, we focus on the behavior of plain saturated model tests after a three-step BS across all three types of noises under the null hypothesis $H_0 : \boldsymbol{v}^T\theta = 0$.

The empirical distribution of the p-values under the null hypothesis under the three different noise models are shown in Figure 4(C). Specifically, while the plain saturated model test yield valid Type-I error control under Gaussian noise, its Type-I error control under Laplacian noise and bootstrapped residuals is slightly inflated. We introduce the following variant of our inferential tools to resolve this inflation.

Our variant is a modification of the *bootstrap substitution method* originally proposed in Tibshirani *et al.* (2018). Here, we approximate the law of $\boldsymbol{v}^T\boldsymbol{Y}$ under the null hypothesis $H_0 : \boldsymbol{v}^T\theta = 0$ with the bootstrapped distribution of $\boldsymbol{v}^T(\boldsymbol{Y} - \theta)$. Specifically, we consider bootstrapping the residuals, $\boldsymbol{r} = \boldsymbol{y} - \widehat{\theta}$, where $\widehat{\theta}$ is a piecewise constant estimate of θ. Here, we use a k-step BS model to estimate $\widehat{\theta}$, where we choose k using twofold cross validation from a twofold split of the data $\boldsymbol{y}$ into odd and even indices. Finally, we compute the p-value using Equation (8), but use the empirical CDF with respect to the bootstrapped values $\boldsymbol{v}^T\boldsymbol{r}$ instead of the Gaussian CDF $\Phi(\cdot)$. For our dataset, while this procedure is not valid in general and should be used with caution, these potential downsides do not seem to come to fruition in practice. Importantly, for our analysis derived from chromosome 9 in GM01750, the resulting p-values using this bootstrapped variant under any of the three noise models are convincingly uniform (Figure 4(D)). We provide results in Web Appendix G.

4.3 | Analysis on Botton dataset

While the above analysis were focused on aCGH data, we now show that our inferential tools can also be used in sequencing data used to study CNV if appropriately preprocessed. This is important to verify, as the technical noise of sequencing data without suitable preprocessing is vastly different from the technical noise of microarray data. Specifically, we investigate our inferential tools' performance on sequencing data collected in Botton *et al.* (2013), which was preprocessed using CNVkit (Talevich *et al.*, 2016), a recent codebase designed to analyze CNV from sequencing data. This sequencing data is derived from the C0902 melanoma cell line and has over 27,000 measurements across the entire genome. We choose this dataset in particular since we also have an external aCGH dataset of the same cell line, with comparable locations along the genome. Hence, we can see whether or not our inferential tools behave similarly between sequencing and aCGH data. We defer our preprocessing details performed by CNVkit to Web Appendix G.

The results, shown in Figure 5, demonstrate that while the estimated changepoints based on aCGH or sequencing data can differ dramatically, the significant changepoints determined by our inferential tools are largely the same. Specifically, we focus on chromosomes 5 and 10 and apply a five-step WBS followed by marginalized saturated model tests. In chromosome 5, the changepoint around location 0.25 is deemed significant in both aCGH and sequencing data, and many of the other estimated changepoints that disagree between both data sources are filtered out. In chromosome 10, the changepoints around location 0.3 is deemed significant in both aCGH and sequencing data, and in this case, Talevich *et al.* (2016) perform fluorescence in situ hybridization (FISH) experiments that additionally reveal there is a true CNV at this location. Overall, we note that since these two data sources are different, we should not expect the significant changepoints to exactly match in general. However, as we have demonstrated, our inferential tools filter changepoints in a way that is largely stable across multiple data sources.

5 | CONCLUSION

In this article, we have developed methods to perform valid post-selection inference for changepoint algorithms and applied them to CNV analyses. Through simulations, we have shown our inferential tools have higher power than competing methods. Through our applications on aCGH and sequencing data, we have shown that our inferential tools are beneficial for filtering changepoints. However, changepoint detection is a rapidly evolving field, and this article provides a blueprint on how to perform post-selection inference for newer methods.

ACKNOWLEDGMENTS

The authors used Pittsburgh Supercomputing Center resources (proposal/grant number: DMS180016P). Sangwon Hyun was supported by supported by NSF grants DMS-1554123 and DMS-1613202. Max G'Sell was supported by NSF grant DMS-1613202. Ryan Tibshirani was supported by NSF grant DMS-1554123. We also thank Nancy R. Zhang for helpful discussions and suggestions.

OPEN RESEARCH BADGES

This article has earned Open Data and Open Materials badges. Data and materials are available at https://osf.io/39wmv/.

DATA AVAILABILITY STATEMENT

The datasets in this article are obtainable from the following places: the Snijder analysis (Section 4.1) uses data directly from the GLAD Bioconductor package (https://www.bioconductor.org/packages/release/bioc/html/GLAD.html); for the Botton analysis (Section 4.3), the aCGH data can be retrieved from the CNVkit example GitHub package (https://github.com/etal/cnvkit-examples). The authors preprocessed the BAM files in the same package using the CNVkit software (https://github.com/etal/cnvkit) to obtain the sequencing data, using the steps outlined in the CNVkit example package.

ORCID

Sangwon Hyun https://orcid.org/0000-0003-0377-897X
Kevin Z. Lin https://orcid.org/0000-0002-1236-9847

REFERENCES

Azaïs, J.-M., De Castro, Y. and Mourareau, S. (2018) Power of the spacing test for least-angle regression. *Bernoulli*, 24, 465–492.

Botton, T., Yeh, I., Nelson, T., Vemula, S.S., Sparatta, A., Garrido, M.C., et al. (2013) Recurrent BRAF kinase fusions in melanocytic tumors offer an opportunity for targeted therapy. *Pigment Cell & Melanoma Research*, 26, 845–851.

Duy, V.N.L., Toda, H., Sugiyama, R. and Takeuchi, I. (2020) Computing valid p-value for optimal changepoint by selective inference using dynamic programming. arXiv preprint, arXiv:2002.09132.

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

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

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

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

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

Lin, K., Sharpnack, J.L., Rinaldo, A. and Tibshirani, R.J. (2017) A sharp error analysis for the fused lasso, with application to approximate changepoint screening. In *Advances in Neural Information Processing Systems*, pp. 6884–6893.

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

Snijders, A.M., Nowak, N., Segraves, R., Blackwood, S., Brown, N., Conroy, J., et al. (2001) Assembly of microarrays for genome-wide measurement of DNA copy number. *Nature Genetics*, 29, 263–264.

Talevich, E., Shain, A.H., Botton, T. and Bastian, B.C. (2016) CNVkit: Genome-wide copy number detection and visualization from targeted DNA sequencing. *PLoS Computational Biology*, 12, e1004873.

Tian, X. and Taylor, J. (2018) Selective inference with a randomized response. *Annals of Statistics*, 46, 619–710.

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

Zhang, N.R. (2010) DNA copy number profiling in normal and tumor genomes. In: Feng J., Fu W. and Sun F. (Eds) *Frontiers in Computational and Systems Biology*. London: Springer, vol 15, pp. 259–281.

SUPPORTING INFORMATION

Web Appendices, Tables, and Figures 6– 14 referenced in Sections 1–5, along with code and data used to generate all figures in this manuscript, are available with this paper at the Biometrics website on Wiley Online Library. The R software used for the numerical results in this paper can be found at https://github.com/sangwon-hyun/binseginf/. The scripts used to produce the numerical results in this paper are stored in an open source library (https://osf.io/39wmv/).

How to cite this article: Hyun S, Lin KZ, G'Sell M, Tibshirani RJ. Post-selection inference for changepoint detection algorithms with application to copy number variation data. *Biometrics*. 2021;1–13. https://doi.org/10.1111/biom.13422