Valid Post-Detection Inference for Change Points Identified Using Trend Filtering

Reza Valiollahi Mehrizi *1 and Shojaeddin Chenouri1

1Department of Statistics and Actuarial Science, University of Waterloo, Canada

Abstract

There are many research works and methods about change point detection in the literature. However, there are only a few that provide inference for such change points after being estimated. This work mainly focuses on a statistical analysis of change points estimated by the PRUTF algorithm, which incorporates trend filtering to determine change points in piecewise polynomial signals. This paper develops a methodology to perform statistical inference, such as computing p-values and constructing confidence intervals in the newly developed post-selection inference framework. Our work concerns both cases of known and unknown error variance. As pointed out in the post-selection inference literature, the length of such confidence intervals are undesirably long. To resolve this shortcoming, we also provide two novel strategies, global post-detection and local post-detection which are based on the intrinsic properties of change points. We run our proposed methods on real as well as simulated data to evaluate their performances.

1 Introduction

Change point detection has achieved considerable attention in recent years and found applications in many fields such as finance and econometrics, Bai and Perron (2003), bioinformatics and genomics, Futschik et al. (2014), climatology, Pezzatti et al. (2013), and technology, Galceran et al. (2017). This problem was first studied by Page (1954, 1955) and since then, has been of interest to many scientists including statisticians.

Formally, change point detection seeks the locations of changes in the distribution of a signal for which an ordered sequence of observations is available. This paper considers change point detection in the mean of a distribution so that this underlying mean is a piecewise polynomial function. More specifically, let \( y = (y_1, \ldots, y_n) \) be a response vector following the model

\[
y_i = f_i + \varepsilon_i, \quad i = 1, \ldots, n,
\]

where \( f_i = f(i/n) \) is a deterministic and unknown signal with equally spaced input points over the interval \([0, 1]\). We assume that the error terms \( \varepsilon = (\varepsilon_1, \ldots, \varepsilon_n) \) are independent and

*Corresponding Author: rvalioll@uwaterloo.ca
have a Gaussian distribution with a mean of zero and finite variance $\sigma^2$ (either known or unknown). We suppose that signal $f$ undergoes $J_0$ unknown and distinct changes at locations $0 = \tau_0 < \tau_1 < \ldots < \tau_{J_0} < \tau_{J_0+1} = n$. Additionally, we allow the number of change points $J_0$ to grow with the sample size $n$. These change points divide the entire signal $f$ into $J_0 + 1$ sub-signals. We assume that each of these sub-signals are polynomial of order $r$ ($r = 0, 1, 2, \ldots$). We put no restriction over the signal at the locations of change points, i.e., the signal could be either continuous or discontinuous at change points. The focus in change point detection is on the estimation of the number and the positions of changes.

There is a vast and rich literature on change-point detection in statistical research. Depending on various aspects of the data at hand, many different techniques and algorithms exist in the literature. Greedy search techniques have long been applied to change point detection in various settings, in which change points are found one at a time. A traditional approach for the canonical change-point problem is the Binary Segmentation Vostrikova (1981). When the signal is piecewise constant, it is referred to as a canonical change-point problem in the literature. Recently refinement of Binary segmentation such as Circular Binary Segmentation Olshen et al. (2004), and Wild Binary Segmentation Fryzlewicz et al. (2014) have also been successfully applied. More greedy search methodologies adapted for various setting have recently been introduced: Li et al. (2016), Fryzlewicz et al. (2018), Baranowski et al. (2019), Anastasiou and Fryzlewicz (2019), and Fryzlewicz (2018). Some authors have considered the canonical change-point detection as a penalized least square problem, examples are Yao and Au (1989), Davies and Kovac (2001) Harchaoui and Lévy-Leduc (2010) and Qian and Jia (2016).

Although there is a huge body of work on canonical change point framework, a more general case in which signal $f$ is modelled as a piecewise polynomial has attracted less attention. A few works, mainly focused on a linear piecewise model, exist in the literature; see Bai (1997), Maidstone et al. (2017), Baranowski et al. (2019) and Anastasiou and Fryzlewicz (2019). The recent work of Mehrizi and Chenouri (2020) introduces an approach called PRUTF, which is designed for change point detection for piecewise polynomial signals. PRUTF exploits the trend filtering problem introduced in Tibshirani et al. (2014) and provides a novel algorithm to estimate change points.

After analysts perform change point detection/estimation, genuine interest is to make inferences about the uncertainty of the detected/estimated points. Many research works such as Eckley et al. (2011), Horváth and Rice (2014), and Chen et al. (2001) have discussed conducting such inferences. However, they have ignored the data-driven nature of the detected change points and regarded them as fixed locations. It is important to note that this data-dependent property of detected change points invalidates the inferences mentioned earlier. There is a new and rapidly growing framework that provides tools for conducting inference after a selection procedure. This type of inference is called post-selection inference and has been mainly developed for inference after variable selection in high-dimensional regression. See for example Berk et al. (2013); Lee et al. (2016); Tibshirani et al. (2016); Fithian et al. (2014). In this paper we adopt this body of work for inference after change point detection in piecewise polynomial signals.
Interest in post-selection inference research ignited after the work of Pötscher (1991). Later, in a sequence of articles, Leeb and Pötscher (2003, 2006, 2008) explored the estimation of the post-selection distribution. This subject has attracted much attention more recently as applications for model selection approaches have proliferated. Berk et al. (2013) performed valid and conservative selective inference by considering all possible procedures that produce a selected model. Lockhart et al. (2014) suggested an asymptotic test procedure to test whether a nonzero estimated coefficient in the lasso regression coincides with the true nonzero coefficient. Lee et al. (2016) developed an exact test based on truncated Gaussian distribution for the solution path of the lasso at a fixed value of the regularization parameter $\lambda$. Concurrently, a similar test is proposed by Tibshirani et al. (2016) for the lasso, LARs and forward stepwise regression with a fixed number of steps in the solution path. Fithian et al. (2014) studied the theoretical properties of post-selection inference and generalized the framework to the broad class of the exponential family of distributions. A number of extensions to different frameworks and applications are given in Loftus and Taylor (2014), Tian and Taylor (2017), Tibshirani et al. (2018), Hyun et al. (2018a), Liu et al. (2018), Bachoc et al. (2020).

In a post-selection inference setting, since data have been used to fit a model, a conditional inference is required to restrict the sample space to unused data. This conditional approach prevents an inference procedure from using data twice (once for selection and once for inference). As a result of this conditioning, the post-detection distribution changes from Gaussian to truncated Gaussian (Lee et al., 2016; Tibshirani et al., 2016), reflecting the restriction on the sample space. However, it turns out that this truncation is often very severe and leads statistical tests to lose power and their respective confidence intervals to become undesirably wide; see Section 6. These problems occur because most of the data are employed to select the model. Consequently, there is insufficient information to use for drawing conclusions after the selection procedure.

The application of post-selection inference in the change points context, referred to as post-detection inference, was first addressed in Hyun et al. (2018a). Their work, which is our main inspiration, applies the exact truncated Gaussian test for a given and selected number of steps in the solution path of generalized lasso. Similarly, Hyun et al. (2018b) study post-detection inference for some popular change point detection methodologies such as Binary Segmentation, Circular Binary Segmentation, and Wild Binary Segmentation.

In this paper, we study the problem of conducting valid statistical inferences for detected change points. For change point detection/estimation we use the Pattern Recovery Using Trend Filtering (PRUTF) algorithm introduced in Mehrizi and Chenouri (2020). At the core of our framework, an important result states that the set of change points detected by PRUTF constitutes a polyhedron (a convex cone). We apply the post-selection inference framework to compute valid post-detection p-values as well as confidence intervals.

1.1 Summary

We make the following contributions in this work.
• One fundamental aspect of our contribution is establishing that the set of change points identified using the PRUTF algorithm characterizes a polyhedron set in \( y \). This characterization allows us to use the post-selection inference methodology for conducting statistical inference after change point detection. We emphasize that, to the best of our knowledge, the inference procedures for the significance of detected change points developed in this paper are the first of their kind for piecewise polynomial signals. Mehrizi and Chenouri (2020) also developed a stopping criterion for selecting the number of required steps for the PRUTF algorithm, which uses the Gaussian bridge property of dual variables. We will show that this criterion also forms a polyhedron and, consequently, applies to the post-detection inference framework.

• The implication of post-detection inference for the Gaussian model in (1) leads us to propose two test statistics, one for known \( \sigma^2 \) and another for unknown \( \sigma^2 \). A significant feature of these test statistics is that their exact and finite sample distributions under the null hypothesis are \( U(0, 1) \). These test statistics allow us to conduct inference for the significance of change points estimated by PRUTF.

• We inspect the produced p-values and confidence intervals of detected change points using the two test statistics for both known and unknown noise variance cases. We show that the confidence intervals derived by these approaches are unacceptably wide. This also leads to a loss in the power of hypothesis tests. To resolve these shortcomings, we introduce two new methods of conditioning for post-detection inference. We call the first method *global post-detection* that focuses only on the target change point and makes inference regardless of the other change points. We call the second method *local post-detection* which takes into account detected locations that are the most relevant to the target change point.

• We conduct a comprehensive simulation study to investigate the performance of the proposed procedures in terms of power of post-detection tests and length of their confidence intervals. We also demonstrate applications of our algorithms to several real data sets.

The rest of this paper is structured as follows. We first provide a version of the PRUTF algorithm of Mehrizi and Chenouri (2020) along with a stopping criterion in Section 2. We also review conducting post-selection inference based on the polyhedron characterization of a selection procedure. In Section 3, we explain how to capture the representation of the corresponding polyhedron of the PRUTF algorithm. This representation also determines cases where the number of steps for the algorithm is adaptively selected using the stopping rule. Next, two different approaches are established in Section 4 to conduct valid post-detection inference when the error variance is assumed both known and unknown. In Section 6, we discuss the shortcomings of the post-selection inference, conditional on the entire selected model and its signs, such as reduction in the power of tests and wide confidence intervals. We then propose two strategies to resolve these shortcomings. Section 7 provides some numerical investigation, including real data analysis and a simulation study. We conclude the work with a discussion in Section 8.
2 Notations and Preliminaries

2.1 Notations

The following notations are used throughout the paper. For an $m \times n$ matrix $A$, a sub-matrix $A_I$ extracts the rows of the matrix $A$ that are in the set $I \subseteq \{1, \ldots, m\}$. Similarly, a sub-vector $a_I$ picks the components of a vector $a$ that are in $I$. Quantities $A_{-I}$ and $a_{-I}$ select rows and elements of $A$ and $a$ that are not in $I$, respectively. Additionally, for selecting $i$-th row of $A$, the notation $[A]_i$ and for its $(i,j)$-th element the notation $[A]_{ij}$ are used. Moreover, for a real number $x$, $\lfloor x \rfloor$ denotes the greatest integer less than or equal to $x$. Notation $\lceil x \rceil$ also gives the smallest integer greater or equal to $x$.

2.2 Solution Path of Trend Filtering

In this section, we briefly review the methodology introduced in Mehrizi and Chenouri (2020). Suppose a response vector $y$ and a penalty matrix $D^{(r+1)}$, to be defined, are given. The PRUTF estimate of change points in a piecewise polynomial signal $f$ of order $r$ is nonzero entries of $\hat{D}^{(r+1)} f$, where $\hat{f}$ is the solution of the following penalized least square problem

$$\min_{f \in \mathbb{R}^n} \frac{1}{2} \| y - f \|_2^2 + \lambda \| D^{(r+1)} f \|_1.$$  \hfill (2)

In this optimization problem known as trend filtering, $\lambda$ is the regularization parameter used to control the balance between sparsity and model fit, and $(n - r - 1) \times n$ matrix $D^{(r+1)}$ is the $(r+1)$-th discrete difference operator. The first order discrete difference operator matrix $D^{(1)}$ is defined as

$$D^{(1)} = \begin{pmatrix} -1 & 1 & 0 & \ldots & 0 & 0 \\ 0 & -1 & 1 & \ldots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \ldots & -1 & 1 \end{pmatrix}.$$  

For $r \geq 1$, the difference operator of order $r + 1$ can be recursively computed by $D^{(r+1)} = D^{(1)} \times D^{(r)}$, where $D^{(1)}$ here is the $(n - r - 2) \times (n - r - 1)$ version of the first difference operator. Henceforth, for simplicity, we suppress the superscript for $D^{(r+1)}$, i.e., $D = D^{(r+1)}$.

As a particular case of trend filtering ($r = 0$), one-dimensional fused lasso has been effectively used to detect change points and produce a piecewise constant estimate of a signal (Harchaoui and Lévy-Leduc, 2010; Rinaldo et al., 2009; Qian and Jia, 2016). Recently, Mehrizi and Chenouri (2020) applied trend filtering to detect change points in a piecewise polynomial signal. The method, called PRUTF, returns estimates of change points along the trend filtering solution path.

Problem (2) has a unique solution because $D$ is a full row rank matrix; see Ali et al. (2019) for more details. One way to estimate the signal $f$ is to solve its equivalent Lagrange dual problem. Following the argument of Tibshirani (2011) and duality conditions of Boyd and Vandenberghe (2004), the dual problem of trend filtering can be written as

$$\min_{u \in \mathbb{R}^m} \frac{1}{2} \| y - D^T u \|_2^2 \quad \text{subject to} \quad \| u \|_\infty \leq \lambda,$$  \hfill (3)
which is a quadratic problem constrained within a box with side length $2\lambda$. The connections between the primal and dual solutions are

$$\hat{\mathbf{u}}_\lambda = \lambda \hat{\gamma} \quad \text{and} \quad \hat{\mathbf{f}}_\lambda = \mathbf{y} - \mathbf{D}^T\hat{\mathbf{u}}_\lambda,$$

where $\hat{\gamma} \in \mathbb{R}^m$ is a subgradient of $\|\mathbf{x}\|_1$ computed at $\mathbf{x} = \mathbf{D}\hat{\mathbf{f}}_\lambda$, given by

$$\hat{\gamma}_i \in \begin{cases} 
{+1} & \text{if } |\mathbf{D}\hat{\mathbf{f}}_\lambda| > 0 \\
{-1} & \text{if } |\mathbf{D}\hat{\mathbf{f}}_\lambda| < 0 \\
{-1, +1} & \text{if } |\mathbf{D}\hat{\mathbf{f}}_\lambda| = 0.
\end{cases}$$

Tibshirani (2011) proposed an algorithm to compute the entire dual solution path along the regularization parameter $\lambda$. PRUTF modifies this algorithm by removing $r + 1$ rows of $\mathbf{D}$ once a coordinate joins the boundary set. This is in contrast with the original algorithm of Tibshirani (2011) in which only one row of $\mathbf{D}$ is removed. The modification in PRUTF makes the method well-suited for the discovery of change points; see (Mehrizi and Chenouri, 2020, Remark 3).

Below, we briefly review the PRUTF algorithm to identify change points along the solution path as the value of $\lambda$ descends. For more details, see Mehrizi and Chenouri (2020).

**Algorithm 1 (Dual Solution Path of Trend Filtering)**

*Define $r_b = \lfloor (r + 1)/2 \rfloor - 1$ and $r_a = \lfloor (r + 1)/2 \rfloor$, and also set $\bar{\tau}_0 = \emptyset$, the empty set.*

1. **At step $j = 1$, initialize the boundary set $\mathcal{B}_1 = \{\hat{\tau}_1 - r_b, \hat{\tau}_1 - r_b + 1, \ldots, \hat{\tau}_1\}$ and its associated sign vector $\hat{s}_{\mathcal{B}_1} = \{\hat{s}_1, \ldots, \hat{s}_1\}$, both with cardinality of $r_b + 1$, where $\hat{\tau}_1$ is obtained by

$$\hat{\tau}_1 = \arg\max_{i = 1, \ldots, m} |\hat{u}_i|,$$

and $\hat{s}_1 = \text{sign}(\hat{u}_1)$, where $\hat{u}_i$ is the $i$-th element of the vector $\hat{\mathbf{u}} = (\mathbf{D}\mathbf{D}^T)^{-1}\mathbf{D}\mathbf{y}$. In order to obtain the locations of change points, we maintain the set $\bar{\tau}_1 = \{\hat{\tau}_1\}$. We also record the first joining time $\lambda_1 = |\hat{u}_1|$ and keep track of the augmented boundary set $\mathcal{A}_1 = \{\hat{\tau}_1 - r_a, \ldots, \hat{\tau}_1 + r_a\}$ and its corresponding sign vector $\hat{s}_{\mathcal{A}_1} = \{\hat{s}_1, \ldots, \hat{s}_1\}$ of length $r + 1$.

2. **For step $j = 2, 3, \ldots ,**

(a) Obtain the pair $(\hat{\tau}_j^{\text{join}}, \hat{s}_j^{\text{join}})$ from

$$(\hat{\tau}_j^{\text{join}}, \hat{s}_j^{\text{join}}) = \arg\max_{i \in \mathcal{A}_{j-1}, \ s \in \{-1, 1\}} \frac{a_i}{s + b_i} \cdot \mathbb{1} \left\{ 0 \leq \frac{a_i}{s + b_i} \leq \lambda_{j-1} \right\},$$

and set the next joining time $\lambda_j^{\text{join}}$ as the value of $\frac{a_i}{s + b_i}$, for $i = \hat{\tau}_j^{\text{join}}$ and $s = \hat{s}_j^{\text{join}}$. The quantities $\mathbf{a}$ and $\mathbf{b}$ in the above equation are defined as

$$\mathbf{a} = (\mathbf{D}_{-\mathcal{A}} \mathbf{D}_{-\mathcal{A}}^T)^{-1} \mathbf{D}_{-\mathcal{A}} \mathbf{y} \quad \text{and} \quad \mathbf{b} = (\mathbf{D}_{-\mathcal{A}} \mathbf{D}_{-\mathcal{A}}^T)^{-1} \mathbf{D}_{-\mathcal{A}} \mathbf{D}_{-\mathcal{A}}^T \mathbf{s}_{\mathcal{A}}.$$  

(b) Obtain the pair $(\hat{\tau}_j^{\text{leave}}, \hat{s}_j^{\text{leave}})$ from

$$\left(\hat{\tau}_j^{\text{leave}}, \hat{s}_j^{\text{leave}}\right) = \arg\max_{i \in \mathcal{B}_{j-1}, \ s \in \{-1, 1\}} \frac{c_i}{d_i} \cdot \mathbb{1} \{ c_i \leq 0, \ d_i < 0 \},$$
and assign the next leaving time $\lambda_j^{\text{leave}}$ as the value of $\frac{\hat{u}_i}{d_i}$, for $i = \hat{s}_j^{\text{leave}}$ and $s = \hat{s}_j^{\text{leave}}$, where $c$ and $d$ are given by

$$c = \text{diag}(s_b)D_B(y - D^T_Aa) \quad \text{and} \quad d = \text{diag}(s_b)D_B(D^T_As_A - D^T_Ab).$$

(9)

(c) Let $\lambda_j = \max\{\lambda_j^{\text{join}}, \lambda_j^{\text{leave}}\}$, then the boundary set $B_j$ and its sign vector $\hat{s}_{B_j}$ are updated in the following fashion:

- Either append $\{\hat{\tau}_j^{\text{join}} - r_b, \hat{\tau}_j^{\text{join}} - r_b + 1, \ldots, \hat{\tau}_j^{\text{join}}\}$ and the corresponding signs $\{s_j^{\text{join}}, \ldots, s_j^{\text{join}}\}$ to $B_{j-1}$ and $\hat{s}_{B_{j-1}}$, respectively, provided that $\lambda_j = \lambda_j^{\text{join}}$. Also, add $\hat{\tau}_j^{\text{join}}$ to $\hat{s}_{j-1}$.
- Or remove $\{\hat{\tau}_j^{\text{leave}}, \hat{\tau}_j^{\text{leave}} + 1, \ldots, \hat{\tau}_j^{\text{leave}} + r_b\}$ and the corresponding signs $\{s_j^{\text{leave}}, \ldots, s_j^{\text{leave}}\}$ from $B_{j-1}$ and $\hat{s}_{B_{j-1}}$, respectively, provided that $\lambda_j = \lambda_j^{\text{leave}}$. Also, eliminate $\hat{\tau}_j^{\text{leave}}$ from $\hat{s}_{j-1}$.

In the same manner, the augmented boundary set, $A_j$ and its sign, $\hat{s}_{A_j}$ are formed by adding $\{\hat{\tau}_j^{\text{join}} - r_b, \ldots, \hat{\tau}_j^{\text{join}} + r_b\}$ and $\{s_j^{\text{join}}, \ldots, s_j^{\text{join}}\}$ to $A_{j-1}$ and $\hat{s}_{A_{j-1}}$ if $\lambda_j = \lambda_j^{\text{leave}}$ or, otherwise, by removing associated $\{\hat{\tau}_j^{\text{leave}}, \ldots, \hat{\tau}_j^{\text{leave}} + r\}$ and $\{s_j^{\text{leave}}, \ldots, s_j^{\text{leave}}\}$ from $A_{j-1}$ and $\hat{s}_{A_{j-1}}$. The dual solution is computed as $\hat{u}_{A_j}(\lambda) = a - \lambda b$ for interior coordinates and $\hat{u}_{A_j}(\lambda) = \lambda \hat{s}_{A_j}$ for boundary coordinates over $\lambda_j \leq \lambda \leq \lambda_{j-1}$.

3. Repeat Step 2 until $\lambda_j > 0$.

Along the dual solution path, the above algorithm keeps track of four sets: the set of dual change points denoted by $\hat{\tau}$, the set of change points signs denoted by $\hat{s}$, the set of coordinates that hits the boundary $\pm \lambda_j$, denoted by $B$ and the set of boundary coordinates plus $r_b$ coordinates right after each change point, denoted by $A$. The critical points $\lambda_j \geq \lambda_2 \geq \ldots \geq 0$ are called knots and indicate the values of the regularization parameter at which the boundary set $B$ changes. Notice that $\hat{\tau}$ derived along the above algorithm delivers locations of change points for the dual problem of trend filtering. In order to obtain locations of change points in the underlying signal (primal trend filtering) which is our main interest, we require to add $r_b$ to any element of $\hat{\tau}$, that is, $\{\hat{\tau}_1 + r_b, \hat{\tau}_2 + r_b, \ldots\}$.

### 2.3 Stopping Criterion

To prevent false discovery of change points by the PRUTF algorithm, we need a suitable criterion to end the search algorithm. Mehrizi and Chenouri (2020) designed a stopping rule based on the properties of dual variables within each block created by detected change points. More precisely, the authors have proven that the dual variables inside each block constitute a Gaussian bridge process. For more details, see Theorem 1 in Mehrizi and Chenouri (2020). It turns out that each entry of $\hat{u}^*_{A_j} = (D_{-A}D_{-A}^T)^{-1}D_{-A}y$ tends to stay close to the zero line, after locating all change points. Therefore, at each step of Algorithm 1, it is checked whether the maximum of $|\hat{u}^*_{A_j}|$ goes beyond a certain threshold. When there is no value to exceed this threshold, the
algorithm will stop. More precisely, for \( |A_{j_0}| \) as the cardinality of \( A_{j_0} \) and \( k = m - |A_{j_0}| \), the PRUTF algorithm terminates at the iteration \( j_0 \) if

\[
\max_{0 \leq t \leq 1} \left| \hat{u}^{st}_{-A_{j_0}} (\lfloor kt \rfloor) \right| \leq \sigma x_\alpha (k - r)^{(2r + 1)/2},
\]

where \( x_\alpha \) is derived from the equation

\[
\sum_{i=1}^{\infty} (-1)^{i+1} \exp \left( \frac{-2i^2 x_\alpha^2}{S_2^2(k)} \right) = \frac{\alpha}{2}. \tag{10}
\]

In the above equation, the quantity \( S_2^2(k) \) is the \( k \)-th diagonal element of the matrix \((D_{-\alpha}D_{-\alpha}^T)^{-1}\).

Remark 1 **Mehrizi and Chenouri (2020)** have established that the PRUTF algorithm is not consistent in pattern recovery when the true signal \( f \) contains staircase patterns. A staircase pattern refers to the phenomenon that the signs of two adjacent change points are in the same direction. This pattern causes PRUTF to identify many false change discoveries within segments with a staircase. Instead, they have proposed a modified algorithm called mPRUTF, which consistently recovers the true signal even in the presence of staircases.

2.4 Post-Selection Inference With Polyhedron Selection Procedures

In this section, we review some key concepts of post-selection inference in a linear regression setting. For details about a broader class of models, see Fithian et al. (2014). Suppose that the observations \( y \in \mathbb{R}^n \) follow a Gaussian regression model with either known or unknown variance. Also, let \( \mathcal{M} \) be a finite collection of all possible models \( \mathcal{M} \) obtained from a model selection procedure, that is, \( \mathcal{M} = \{M : M \subseteq \{1, \ldots, p\}\} \), for \( p \) to be the number of variables. The goal is, therefore, to carry out statistical inference for a selected model \( \widehat{M}(y) = M \). Since we adaptively choose \( \widehat{M}(y) \) using the data, for inference, it is natural to consider the conditional distribution given \( \widehat{M}(y) \). This conditional distribution is called post-selection distribution. For testing null hypothesis \( H_0 \), we seek to control the post-selection type-I error, defined as \( P_{H_0}(\text{Reject } H_0 \mid \widehat{M}(y) \in \mathcal{M}) \), at the nominal level \( \alpha \). By analogy to the classical theory, a post-selection confidence interval can then be built by inverting the associated post-selection hypothesis test.

Throughout this paper, we restrict our attention to a specific class of selection procedures known as polyhedron (affine) selection procedures. Again, suppose a variable selection approach picks a model \( \widehat{M}(y) \) from a finite collection of models \( \mathcal{M} \) using the data \( y \). This selection approach is called a polyhedron selection procedure if any \( \widehat{M}(y) \in \mathcal{M} \) can be characterized as a polyhedron set in respect to \( y \). In other words, any selected model \( \widehat{M}(y) \) chosen by the polyhedron selection procedure can be written in the form of

\[
\{ y : Ay \geq q \}, \tag{11}
\]

where the matrix \( A \in \mathbb{R}^{p \times n} \) and the vector \( q \in \mathbb{R}^p \) are dependent on \( y \). Observe that the inequality in (11) is interpreted componentwise. Selection approaches using the \( \ell_1 \)-penalized generalized linear model, including Lasso Lee et al. (2016); Tibshirani et al. (2016), LARS Taylor and Tibshirani (2018) and generalized lasso Hyun et al. (2018a), are examples of such polyhedron selection procedures.
3 Construction of Polyhedron

In this section, we describe how the change points detection procedure using PRUTF can be represented as a polyhedron. This representation enables us to state our post-detection inference conditional on a polyhedron. Suppose \( \{ \hat{\tau}_j, \hat{s}_j \} \) is the set of detected change points and their signs at step \( j \) of the PRUTF algorithm. Below, we describe that \( \{ \hat{\tau}_j = \hat{\tau}_j(y), \hat{s}_j = \hat{s}_j(y) \} \) as a function of \( y \) is indeed a polyhedron. Thus, formally, for a matrix \( A \) and vector \( q \),

\[
\left\{ y : \hat{\tau}_j = \hat{\tau}_j(y), \hat{s}_j = \hat{s}_j(y) \right\} = \left\{ y : Ay \geq q \right\}.
\]  

(12)

From now on, we call \( A \) the polyhedron matrix. Observe that, in general, \( \hat{\tau}_j \) does not necessarily contain \( j \) entries since (except in the case of \( r = 0 \)) the dual solution path of PRUTF can either add a change point to or remove it from the augmented boundary set at each step (Mehrizi and Chenouri, 2020).

3.1 Construction of Polyhedron Along the Solution Path

In the following, the construction process of the matrix \( A \) and vector \( q \) associated with the polyhedron along the dual solution path of PRUTF is presented. We present this construction in steps according to the steps of Algorithm 1.

1. For \( j = 1 \), the conditions for deriving the pair \( (\hat{\tau}_1, \hat{s}_1) \) can be rewritten as

\[
\hat{s}_1 \left[ (DD^T)^{-1} D \right]_{\hat{\tau}_1} y \geq \pm \left[ (DD^T)^{-1} D \right]_{\hat{\tau}_1} y,
\]  

(13)

for any \( i \neq \hat{\tau}_1 \). This implies that the polyhedron matrix after the first step, denoted by \( A_1 \), has \( 2(m - 1) \) rows, formed by \( \hat{s}_1 \left[ (DD^T)^{-1} D \right]_{\hat{\tau}_1} y \pm \left[ (DD^T)^{-1} D \right]_{\hat{\tau}_1} y \), for any \( i \neq \hat{\tau}_1 \).

2. For \( j = 1, 2, \ldots \), assume that \( A_j \) is the polyhedron matrix associated with \( \{ \hat{\tau}_j, \hat{s}_j \} \) at step \( j \) of Algorithm 1. In order to construct the polyhedron matrix for the set \( \{ \hat{\tau}_{j+1}, \hat{s}_{j+1} \} \), a number of rows will be appended to \( A_j \) according to Step 2 of Algorithm 1. Recall that part (a) of Step 2 corresponds to the specification of the joining coordinate and its sign. Therefore, specifying \( \hat{\tau}_{j+1} \) is equivalent to satisfying the conditions

\[
\text{sign}(a_i) \left[ (D_{-A_i}D_{-A_i}^T)^{-1} D_{-A_i} \right]_{\hat{\tau}_{j+1}} y \geq 0, \quad i \notin A_j, \]

\[
\frac{\left[ (D_{-A_j}D_{-A_j}^T)^{-1} D_{-A_j} \right]_{\hat{\tau}^\text{join}_{j+1}} y}{\hat{s}^\text{join}_{j+1}} + \frac{\left[ (D_{-A_j}D_{-A_j}^T)^{-1} D_{-A_j} \right]_{\hat{\tau}^\text{join}_{j+1}} D_{\hat{s}^\text{join}_{j+1}} S_{A_j}}{- \text{sign}(a_i) + \left[ (D_{-A_j}D_{-A_j}^T)^{-1} D_{-A_j} \right]_{\hat{\tau}^\text{join}_{j+1}} D_{\hat{s}^\text{join}_{j+1}} S_{A_j}} y \geq \frac{\left[ (D_{-A_j}D_{-A_j}^T)^{-1} D_{-A_j} \right]_{\hat{\tau}_{j+1}} y}{\hat{s}_{j+1}},
\]

(14)

for any \( i \notin A_j \cup \{ \hat{\tau}^\text{join}_{j+1} \} \). The vector \( a \) in the aforementioned equations is given in (7). The above inequalities will add \( 2(m - |A_j|) - 1 \) rows to the matrix \( A_j \).

Part (b) of Step 2 explains conditions for the determination of the leaving coordinate. First, rows corresponding to conditions \( c_i < 0 \) and \( d_i < 0 \) for \( i \in B_j \) are added to \( A_j \). To this end, we disregard those \( i \in B_j \) which \( d_i \geq 0 \) and partition the remaining entries into two groups. The first group called viable leaving coordinates includes those which \( c_i < 0 \)
and the second group as its complement. Denoting $L_{j+1}$ to be the collection of viable leaving coordinates, we consider

$$L_{j+1} = \left\{ i \in B_j : d_i < 0 \text{ and } c_i < 0 \right\},$$

with the complementary set

$$L_{j+1}^c = \left\{ i \in B_j : d_i < 0 \text{ and } c_i \geq 0 \right\}. \hspace{1cm} (15)$$

Since vector $\mathbf{d}$, defined in (9), does not depend on $\mathbf{y}$, the conditions $c_i < 0$ and $d_i < 0$ can be expressed as

$$\begin{bmatrix} \text{diag}(\mathbf{S}_j) \mathbf{D}_{B_j} \left( \mathbf{I} - \mathbf{D}^T_{-A_j} \left( \mathbf{D}_{-A_j} \mathbf{D}^T_{-A_j} \right)^{-1} \mathbf{D}_{-A_j} \right) \end{bmatrix} \mathbf{y} < 0, \text{ for } i \in L_{j+1},$$

$$\begin{bmatrix} \text{diag}(\mathbf{S}_j) \mathbf{D}_{B_j} \left( \mathbf{I} - \mathbf{D}^T_{-A_j} \left( \mathbf{D}_{-A_j} \mathbf{D}^T_{-A_j} \right)^{-1} \mathbf{D}_{-A_j} \right) \end{bmatrix} \mathbf{y} \geq 0, \text{ for } i \in L_{j+1}^c. \hspace{1cm} (16)$$

The above equations correspond to $|L_{j+1}| + |L_{j+1}^c|$ rows of $\mathbf{A}_{j+1}$. Second, the condition associated with the specification of the pair $(\hat{s}^\text{leave}_{j+1}, \hat{s}^\text{leave}_{j+1})$ must be added to the rows of $\mathbf{A}_j$. This condition can be captured for any $i \in L_{j+1} \setminus \{ \hat{s}^\text{leave}_{j+1} \}$ by the inequality

$$\begin{bmatrix} \text{diag}(\mathbf{S}_j) \mathbf{D}_{B_j} \left( \mathbf{I} - \mathbf{D}^T_{-A_j} \left( \mathbf{D}_{-A_j} \mathbf{D}^T_{-A_j} \right)^{-1} \mathbf{D}_{-A_j} \right) \end{bmatrix}_{\hat{s}^\text{leave}_{j+1}} \mathbf{y} \geq \begin{bmatrix} \text{diag}(\mathbf{S}_j) \mathbf{D}_{B_j} \left( \mathbf{I} - \mathbf{D}^T_{-A_j} \left( \mathbf{D}_{-A_j} \mathbf{D}^T_{-A_j} \right)^{-1} \mathbf{D}_{-A_j} \right) \end{bmatrix}_{\hat{s}^\text{leave}_{j+1}} \mathbf{D}^T_{A_j} \hat{s}_{A_j}, \hspace{1cm} (17)$$

which forms $|L_{j+1}| - 1$ rows of $\mathbf{A}_{j+1}$.

Lastly, the decision of adding or removing a coordinate in part (c) is required. The condition for the joining time is $\lambda^\text{join}_{j+1} \geq \lambda^\text{leave}_{j+1}$, which can be expressed as

$$\begin{bmatrix} \mathbf{D}_{-A_j} \mathbf{D}^T_{-A_j} \end{bmatrix}^{-1} \mathbf{D}_{-A_j} \mathbf{y}_{\hat{s}^\text{join}_{j+1}} + \begin{bmatrix} \mathbf{D}_{-A_j} \mathbf{D}^T_{-A_j} \end{bmatrix}^{-1} \mathbf{D}^T_{A_j} \hat{s}_{A_j} \geq \begin{bmatrix} \text{diag}(\mathbf{S}_j) \mathbf{D}_{B_j} \left( \mathbf{I} - \mathbf{D}^T_{-A_j} \left( \mathbf{D}_{-A_j} \mathbf{D}^T_{-A_j} \right)^{-1} \mathbf{D}_{-A_j} \right) \end{bmatrix}_{\hat{s}^\text{leave}_{j+1}} \mathbf{y}, \hspace{1cm} (18)$$

$$\begin{bmatrix} \text{diag}(\mathbf{S}_j) \mathbf{D}_{B_j} \left( \mathbf{I} - \mathbf{D}^T_{-A_j} \left( \mathbf{D}_{-A_j} \mathbf{D}^T_{-A_j} \right)^{-1} \mathbf{D}_{-A_j} \right) \end{bmatrix}_{\hat{s}^\text{join}_{j+1}} \mathbf{D}^T_{A_j} \hat{s}_{A_j}. \hspace{1cm} (19)$$

In the case of leaving time, which corresponds to $\lambda^\text{join}_{j+1} < \lambda^\text{leave}_{j+1}$, the sign of inequality in (19) will flip. Observe that the decision on whether to add or remove a coordinate will only add one row to $\mathbf{A}_j$. Also, it is important to note that $\mathbf{q} = \mathbf{0}$ in all the above steps.

### 3.2 Polyhedron Form After the Stopping Rule

In the preceding section, we described how to construct the polyhedron matrix $\mathbf{A}$ and vector $\mathbf{q}$ after running the PRUTF algorithm for a fixed number of steps. However, the PRUTF
algorithm, as discussed in Section 2.3, terminates using a stopping rule. This stopping rule is developed based on the stochastic term of the dual variables, $\hat{u}_{-\tau}^d(t) = \left[ (D_{-\tau} D_{-\tau}^T)^{-1} D_{-\tau} \right]_t y$, for $t \in \{1, \ldots, m\} \backslash A$. In the following, we will show that this stopping criterion can also be expressed as a polyhedron in $y$.

The stopping rule states that Algorithm 1 stops upon the time the inequality

$$\max_{0 \leq t \leq 1} \left| \hat{u}_{-\tau}^d (\left| kt \right|) \right| \leq \sigma x \left( k - r \right)^{(2r+1)/2}, \quad \text{for} \quad k = m - |A|, \quad (20)$$

is satisfied. In order to show that this stopping rule creates a polyhedron, note that the condition (20) can be rewritten as

$$\left[ (D_{-\tau} D_{-\tau}^T)^{-1} D_{-\tau} \right]_t y \geq -\sigma x \left( k - r \right)^{(2r+1)/2}, \quad \text{for} \quad t \notin A$$

$$-\left[ (D_{-\tau} D_{-\tau}^T)^{-1} D_{-\tau} \right]_t y \geq -\sigma x \left( k - r \right)^{(2r+1)/2}, \quad \text{for} \quad t \notin A. \quad (21)$$

These conditions append $2k$ rows to the matrix $A$ at each step. Additionally, the above conditions add $2k$ non-zero values $-\sigma x \left( k - r \right)^{(2r+1)/2}$ to $q$.

Remark 2 We have just shown that the event $\{ \hat{\tau} = \tau, \hat{s} = s \}$, for all fixed $\tau$ and $s$, constitutes a polyhedron of the form $\{ y : A \ y \geq q \}$. It would also be interesting to only characterize change point set $\{ \hat{\tau} = \tau \}$ in the form of a polyhedron. It turns out that $\{ \hat{\tau} = \tau \}$ can be represented as the union of such polyhedra. More precisely,

$$\left\{ y : \hat{\tau}(y) = \tau \right\} = \bigcup_s \left\{ y : \hat{\tau}(y) = \tau, \hat{s}(y) = s \right\} = \bigcup_s \left\{ y : A_s \ y \geq q_s \right\}, \quad (22)$$

where the union is over all sign vectors, $s \in \{-1, 1\}^{\text{row}(A)}$. Observe that the number of elements for this union is $2^{\text{row}(A)}$ which can grow fast and become intractable when $\text{row}(A)$ is moderately large.

4 Post-Detection Inference

Having detected change points using the PRUTF algorithm, an appealing follow-up goal would be performing statistical inference on the significance of the changes at these locations. In this section, we provide inference tools to apply after implementing the PRUTF algorithm using the post-selection inference framework. Given the vector of observations $y$, we assume that $\hat{\tau} = \{ \hat{\tau}_1, \ldots, \hat{\tau}_J \}$ is the set of ordered change points, $1 < \hat{\tau}_1 < \hat{\tau}_2 < \ldots < \hat{\tau}_J < n - r - 1$, detected using the PRUTF algorithm. Additionally, assume that $\hat{s} = \{ \hat{s}_1, \ldots, \hat{s}_J \}$ is the set of signs associated with $\hat{\tau}$. For notational convenience, we denote $\hat{\tau}_0 = 0$ and $\hat{\tau}_{J+1} = n - r$. Our focus is on conducting valid statistical inference for the significance of changes at the locations $\hat{\tau}_j$, $j = 1, \ldots, J$.

Hypothesis tests about the significance of the detected change points can be cast in a general linear form $H_0 : \eta^T \hat{f} = 0$, for some nonzero contrast vector $\eta \in \mathbb{R}^n$. Of particular interest in this paper is the hypothesis $H_0 : D_{\hat{\tau}_j} \ f = 0$, which tests the significance of change exactly at estimated location $\hat{\tau}_j$, where $D_{\hat{\tau}_j}$ is the $\hat{\tau}_j$-th row of the penalty matrix $D$. However, our inferential
Valid Post-Detection Inference

framework is not specific to the choice \( \eta^T = D \eta \). Many other types of contrast vectors are possible, as long as \( \eta \) is fixed by conditioning on the detection procedure. For example, the segment contrast, proposed in Hyun et al. (2018a), considers the difference between averages of two neighboring segments \( (\hat{\tau}_{j-1}, \hat{\tau}_j] \) and \( (\hat{\tau}_j, \hat{\tau}_{j+1}] \). Also, Jewell et al. (2019) have used a window contrast of size \( h \), which considers the difference between averages of \( h \) consecutive points just before the estimated change point, i.e., \( (\hat{\tau}_j - h, \hat{\tau}_j] \) and \( h \) consecutive points just after that, i.e., \( (\hat{\tau}_j, \hat{\tau}_j + h] \). The window contrast is suitable for checking whether a true change point exists near \( \hat{\tau}_j \). It is worth mentioning that the choice of \( \eta \) ultimately depends on the objective of the practitioner.

Data dependence nature of change point detection methodologies leads to a random change point set \( \hat{\tau} = \hat{\tau}(y) \). Associated with this randomly chosen \( \hat{\tau} \) is the vector \( \eta = \eta(\hat{\tau}) \) which, in turn, is a random object. This randomness invalidates classical theory for conducting statistical inference about \( \eta^T f \); see Berk et al. (2013) for a thorough discussion. In such cases, post-detection inference allows us to carry out our analysis. In particular, post-detection inference revolves around the conditional distribution of \( \eta^T y \) conditional on the selected change points. This conditioning makes \( \eta = \eta(\hat{\tau}) \) to become a fixed vector. Now, the goal is to test a hypothesis that controls the conditional type-I error rate at level \( \alpha \) as well as to build a conditional confidence interval \( I_\alpha \), such that

\[
\Pr \left( \eta^T f \in I_\alpha \mid \hat{\tau} = \tau \right) \geq 1 - \alpha,
\]

for all fixed \( \tau \).

In a change point detection setting, Hyun et al. (2018a,b) have exploited post-selection inference to compute \( p \)-values for the significance of change points found by fused lasso, Binary Segmentation, Wild Binary Segmentation, Circular Binary Segmentation. To boost the power of tests, Jewell et al. (2019) have suggested a post-detection approach which attempts to reduce the size of conditioning events. This approach covers change points detected in a piecewise constant model with identical and independent Gaussian random noises using a \( \ell_0 \)-penalization. In Duy et al. (2020), the authors have also considered post-selection inference for change points estimated by using a dynamic programming for a \( \ell_0 \)-penalization.

Conducting inference for \( \eta^T f \) in the post-detection framework requires the knowledge about the conditional distribution of \( \eta^T y \) given \( \{ \hat{\tau} = \tau \} \). As shown in Section 3, \( \{ \hat{\tau} = \tau, \hat{s} = s \} \) creates the polyhedron of the form \( \{ y : A y \geq q \} \), and \( \{ \hat{\tau} = \tau \} \) is a union of such polyhedra over all possible sign vectors \( s \). Therefore, it is easier to obtain the conditional distribution of \( \eta^T y \) given \( \{ \hat{\tau} = \tau, \hat{s} = s \} \), a single polyhedron. Observe that inferences that are valid conditional on this finer event will also be valid conditional on \( \{ \hat{\tau} = \tau \} \), (Lee et al., 2016). In what follows, we illustrate techniques to compute valid post-detection \( p \)-value for the hypothesis \( H_0 : \eta^T f = 0 \) and to construct a post-detection confidence interval for the parameter \( \eta^T f \).

Ignoring the post-detection framework for a moment, recall that when the change points are assumed fixed, the inference about \( \eta^T f \), depending on whether the error variance is known or unknown, is based on the normal or \( t \) distributions, respectively. More specifically, when \( \sigma^2 \) is known, the statistic \( Z = \eta^T \frac{y}{\sigma \| \eta \|} \) and when \( \sigma^2 \) is unknown, the statistic \( T = \eta^T \frac{y}{\hat{\sigma} \| \eta \|} \),
are employed to make inference for $\eta^T f$. We will essentially use the same statistics $Z$ and $T$ in post-detection inference and focus on determining their respective conditional distributions. We emphasize that these distributions are no longer the usual normal or $t$ distributions as they must be conditioned on the detected change points.

To define our proposed test statistics, again assume that the PRUTF algorithm has detected $J$ change points at locations $\hat{\tau} = \{\hat{\tau}_1, \ldots, \hat{\tau}_J\}$ with the corresponding signs $\hat{s} = \{\hat{s}_1, \ldots, \hat{s}_J\}$. These change points partition the entire signal $f$ into $J + 1$ segments, with each segment having its distinct polynomial signal of order $r$, namely $f_j, j = 0, \ldots, J$. Also, let $y_j$ denote the sub-vector of observations corresponding to the $j$-th segment; thus,

$$f^Ty = \sum_{j=0}^J f_j^Ty_j.$$  

Implementing the least square approach to estimate $f_j$ results in $\hat{f}_j = (X_j^TX_j)^{-1}X_j^Ty_j$, for $j = 0, \ldots, J$, where $X_j$ is defined as

$$X_j = \begin{pmatrix} 1 & \frac{\hat{\tau}_j}{n} & \left(\frac{\hat{\tau}_j}{n}\right)^2 & \cdots & \left(\frac{\hat{\tau}_j}{n}\right)^r \\ 1 & \frac{\hat{\tau}_{j+1}}{n} & \left(\frac{\hat{\tau}_{j+1}}{n}\right)^2 & \cdots & \left(\frac{\hat{\tau}_{j+1}}{n}\right)^r \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \frac{\hat{\tau}_J}{n} & \left(\frac{\hat{\tau}_J}{n}\right)^2 & \cdots & \left(\frac{\hat{\tau}_J}{n}\right)^r \end{pmatrix}.$$  

In words, $X_j$ is the design matrix of the $r$-th polynomial regression of $y_j$ on the indices of $j$-th segment, $\hat{\tau}_j, 1, \ldots, \hat{\tau}_{j+1}$. We also denote the projection matrix onto the column space of $X_j$ as $P_j = X_j(X_j^TX_j)^{-1}X_j^T$. Observe that $P_j f_j = f_j$, for $j = 0, \ldots, J$, or equivalently $P f = f$, where $P$ is a block diagonal matrix whose diagonal entries are the submatrices. With these notations, an unbiased estimator of the error variance, when $\sigma^2$ is unknown, is given by

$$\hat{\sigma}^2 = \frac{1}{d} \sum_{j=0}^J \| (I - P_j) y_j \|^2 = \frac{1}{d} \| (I - P) y \|^2,$$  

where $d = n - (J + 1)$.

According to the Gaussian model of (1), $y$ follows the exponential family of the form

$$y \sim \exp \left\{ \frac{1}{\sigma^2} f^Ty - \frac{1}{2\sigma^2} \|y\|^2 - \frac{\|f\|^2}{2\sigma^2} \right\}.$$  

Decomposing the data into the direction of $\eta$ and orthogonal to $\eta$, as well as using the fact that $P f = f$, yield

$$y \sim \exp \left\{ \frac{1}{\sigma^2} f^T(P_n + P - P_n)y - \frac{1}{2\sigma^2} \|y\|^2 - \frac{\|f\|^2}{2\sigma^2} \right\} = \exp \left\{ \left( \frac{1}{\sigma^2} \eta \right)^T \left( \frac{\eta^Ty}{\|\eta\|^2} \right) + \frac{1}{\sigma^2} f^T(P - P_n)y - \frac{1}{2\sigma^2} \|y\|^2 - \frac{\|f\|^2}{2\sigma^2} \right\},$$  

where $P_n$ is the orthogonal projection on the space spanned by $\eta \in \mathbb{R}^n$, defined as

$$P_n = \frac{\eta (\eta^T \eta)^{-1} \eta^T}{\|\eta\|^2}.$$  

13
Consider a multi-parameter exponential family in which the vector of natural parameters can be split into two sub-vectors: the target and the nuisance parameters. The conditional distribution, given sufficient statistics associated with nuisance parameters, depends only on the target parameters. It is known that this conditional distribution belongs to an exponential family with the same target parameters, the same sufficient statistics, but a different reference measure and the normalizing constant. This fact allows us to derive the conditional distribution of the statistics $Z$ and $T$, given the detected change points, their signs and the sufficient statistics of the nuisance parameters. Consequently, we can perform statistical inference for the target parameter $\eta^T f$ based on the derived distributions. We divide our presentation into two parts: known error variance and unknown error variance to perform such inferences.

4.1 Known Error Variance

When the error variance $\sigma^2$ is known, the statement (25) reveals that $\eta^T y$ and $V = (P - P_{\eta}) y$ are sufficient statistics for $\eta^T f$ and the nuisance parameters, respectively. As previously explained, the conditional distribution of $\eta^T y$ given $V$ eliminates nuisance parameters from our analysis. Therefore, when $\sigma^2$ is known, we base our analysis on the statistic

$$Z = \frac{\eta^T y}{\sigma \| \eta \|}. \quad (26)$$

We further seek to specify the conditional distribution of this statistic given $\{ V, Ay \geq q \}$. Recall that the polyhedron $\{ Ay \geq q \}$ is the substitute for $\{ \hat{\tau}, \hat{s} \}$, estimated using PRUTF. The following theorem illustrates that this conditional distribution is indeed a truncated normal distribution with an explicitly specified truncation set.

**Theorem 1** Suppose $y$ follows model (1), where $\sigma^2$ is assumed known. For a nonzero contrast vector $\eta$,

a) the conditional distribution of $Z$ in (26), given

$$\{ V, Ay \geq q \}, \quad (27)$$

is a normal distribution truncated to the interval $[\mathcal{V}_z^-, \mathcal{V}_z^+]$, provided $\mathcal{V}_z^0 \geq 0$. This distribution is denoted by $\text{TN}(\eta^T f, 1, [\mathcal{V}_z^-, \mathcal{V}_z^+])$. The truncation boundaries $\mathcal{V}_z^\pm = \mathcal{V}_z^\pm(V)$ and $\mathcal{V}_z^0 = \mathcal{V}_z^0(V)$ are given by

$$\mathcal{V}_z^- = \max_{\rho_i > 0} \frac{[q - AV]}{\sigma \rho_i}, \quad \mathcal{V}_z^+ = \min_{\rho_i < 0} \frac{[q - AV]}{\sigma \rho_i}, \quad \mathcal{V}_z^0 = \min_{\rho_i = 0} \frac{[AV - q]}{\sigma}, \quad (28)$$

where $\rho_i = [A \eta/\| \eta \|]$, for $i = 0, 1, \ldots, |\text{row}(A)|$. Observed that $|\text{row}(A)|$ denotes the number of rows in matrix $A$.

b) Moreover, let $\Phi^{[a, b]}(\cdot)$ be the cumulative distribution function of $\text{TN}(0, 1, [a, b])$, thus, under the null hypothesis $H_0 : \eta^T f = 0$, the conditional distribution of $1 - \Phi^{[\mathcal{V}_z^- \mathcal{V}_z^+]}(Z)$ given the event $\{ Ay \geq q \}$ is uniform on the unit interval $[0, 1]$, that is,

$$1 - \Phi^{[\mathcal{V}_z^- \mathcal{V}_z^+]}(Z) \left| \{ Ay \geq q \} \right. \sim U(0, 1). \quad (29)$$

We refer to this statistic as $\text{TN}$ statistic.
The proof is given in Appendix A.1.

Theorem 1 enables us to compute a post-detection \( p \)-value for the hypothesis \( H_0 : \eta^T f = 0 \) as well as to construct a post-detection confidence interval for \( \eta^T f \). These tasks can be carried out using the TN statistic in (29). In particular, for the two-sided hypothesis testing problem

\[
H_0 : \eta^T f = 0 \quad \text{vs} \quad H_1 : \eta^T f \neq 0, \tag{30}
\]

the value of \( 2 \min \{ 1 - \Phi^{[\hat{V}_z, V_z^+]}(Z), \Phi^{[\hat{V}_z, V_z^+]}(Z) \} \) serves as a valid post-detection \( p \)-value, since under the null hypothesis

\[
\Pr \left( 2 \min \{ 1 - \Phi^{[\hat{V}_z, V_z^+]}(Z), \Phi^{[\hat{V}_z, V_z^+]}(Z) \} \leq \alpha \| A y \geq q \right) = \alpha,
\]

for all \( 0 \leq \alpha \leq 1 \). To construct a two-sided post-detection confidence interval, define the confidence limits \( L_z(Z) \) and \( U_z(Z) \) such that

\[
1 - \Phi^{[\hat{V}_z, V_z^+]}(\frac{\eta^T y - L_z(Z)}{\sigma \| \eta \|}) = \frac{\alpha}{2}, \quad \text{and} \quad \Phi^{[\hat{V}_z, V_z^+]}(\frac{\eta^T y - U_z(Z)}{\sigma \| \eta \|}) = \frac{\alpha}{2}. \tag{31}
\]

These limits are well characterized since the survival function of TN(\( \mu, 1, [a, b] \)) monotonically increases with respect to \( \mu \). Hence, the interval \([L_z(Z), U_z(Z)]\) is a valid two-sided post-detection confidence interval at a \( 1 - \alpha \) level for \( \eta^T f \). This confidence interval can be interpreted in the following manner. If \( y \) is repeatedly drawn from model (1) and the PRUTF algorithm is run, among those cases in which \( \hat{\tau}_j \in D_{\hat{\tau}_j} \) is detected, the interval \([L_z(Z), U_z(Z)]\) contains the parameter \( \eta^T f \) with a relative frequency approaching \( 1 - \alpha \).

**Remark 3** The PRUTF algorithm estimates signs of change points in addition to their locations. We can incorporate this knowledge in forming the alternative hypothesis, that is, \( H_1 : \hat{s}_{j, \tau_j} > 0 \). This alternative hypothesis means that \( \hat{s}_{j, \tau_j} \) is a significant change point whose jump is in the direction of \( \hat{s}_{j, \tau_j} \). Note that this test is more powerful than its two-sided counterpart. Tibshirani et al. (2016) have provided a comparison between one-sided and two-sided tests for the significance of the selected variables using lasso. In general, suppose we are interested in testing the one-sided hypothesis \( H_0 : \eta^T f = 0 \) against \( H_1 : \eta^T f > 0 \). As with the two-sided hypotheses, \( 1 - \Phi^{[\hat{V}_z, V_z^+]}(Z) \) is a valid post-detection \( p \)-value for the one-sided test. Additionally, \([L_z(Z), \infty)\) is a one-sided post-detection confidence interval for \( \eta^T f \) where

\[
1 - \Phi^{[\hat{V}_z, V_z^+]}(\frac{\eta^T y - L_z(Z)}{\sigma \| \eta \|}) = \alpha.
\]

### 4.2 Unknown Error Variance

This section concerns post-detection inference after estimating change points, in the more realistic case, when the error variance \( \sigma^2 \) is unknown. The main methods proposed for post-selection inference such as in Lee et al. (2016), Tibshirani et al. (2016), and Taylor and Tibshirani (2018) proceed with a known \( \sigma^2 \), with the exception of Fithian et al. (2015). In the case of an unknown \( \sigma^2 \), we must further condition our inference on the sufficient statistic associated with the nuisance parameter \( \sigma^2 \).
According to (25), in the case of an unknown $\sigma^2$, the term $\eta^T f / \sigma^2$ is the sufficient statistic for the target parameter $\eta^T f / \sigma^2$ and $(V, \|y\|^2)$ is a joint sufficient statistic for the nuisance parameters. Since testing $\eta^T f / \sigma^2 = 0$ is equivalent to testing $\eta^T f = 0$, hence, we construct our analysis based on the statistic

$$T = \frac{\eta^T y}{\hat{\sigma} \|\eta\|},$$

where $\hat{\sigma}^2 = (d - 1) \left\| (I - P) y \right\|^2$, with $d = n - (J + 1)$. Notice that $\hat{\sigma}^2$ is simply a pooled estimate of the error variance $\sigma^2$ using $J + 1$ segments created by the detected change points $\hat{\tau} = \{\hat{\tau}_1, \ldots, \hat{\tau}_J\}$. The next step is to find the conditional distribution of $T$, given the sufficient statistics associated with the nuisance parameters and the polyhedron event identified by the detection procedure. Clearly, this statistic is distributed as a $t$ distribution constrained to the set $\{V, \|y\|^2, Ay \geq q\}$. We establish the corresponding distribution in the following theorem, whose proof is given in Appendix A.

**Theorem 2** Suppose $y$ follows model (1) and $\sigma^2$ is assumed unknown. For a nonzero contrast vector $\eta$,

a) the conditional distribution of $T$ given

$$\left\{V, \|y\|^2, Ay \geq q\right\},$$

is a location-scale $t$ distribution with mean $\eta^T f$, variance 1 and degrees of freedom $d = n - (J + 1)$, truncated to the interval $[V_T^- \cdot V_T^+, V_T^+ \cdot V_T^-]$, denoted by $Tt(\eta^T f, 1, d, [V_T^- \cdot V_T^+])$. The truncation boundaries $V_T^- = V_T^- (V, W)$ and $V_T^+ = V_T^+ (V, W)$ are given by

$$[V_T^- \cdot V_T^+] = \bigcap_{i=1}^{\text{row}(A)} \left\{ t \in \mathbb{R} : \left[ AV - q \right]_i t^2 + \left( 2 \frac{\eta^T (I - P) y}{\hat{\sigma} \|\eta\|} [AV - q] + W_i \hat{\sigma} \right) t + \left[ AV - q \right]_i d \geq 0 \right\},$$

where $W = \left\| (I - P + P_\eta) y \right\|^2$.

b) In addition, let $G_{[a,b]}(\cdot)$ denote the cumulative distribution function of $Tt(0, 1, d, [a, b])$, then under the null hypothesis

$$1 - G_{[V_T^- \cdot V_T^+]}(T) \left\{ Ay \geq q \right\} \sim U(0, 1).$$

We refer this statistic as the $Tt$ statistic.

Theorem 2 allows us to perform post-detection tests and construct post-detection confidence intervals for $\eta^T f$. In the same fashion as the truncated normal case, the quantity

$$2 \min \left\{ 1 - G_{[V_T^- \cdot V_T^+]}(T), G_{[V_T^- \cdot V_T^+]}(T) \right\}$$
is a post-detection $p$-value for the two-sided hypothesis problem (30). A post-detection confidence interval for $\eta^T f$, when $\sigma^2$ is unknown, is also given by $[L_\tau(T), U_\tau(T)]$, where

$$1 - G_{\hat{\sigma}^2}^{\{\hat{\nu}_2, \hat{\nu}_2^+\}} \left( \frac{\eta^T y - L_\tau(T)}{\hat{\sigma} \|\eta\|} \right) = \alpha/2,$$

and

$$G_{\hat{\sigma}^2}^{\{\hat{\nu}_2, \hat{\nu}_2^+\}} \left( \frac{\eta^T y - U_\tau(T)}{\hat{\sigma} \|\eta\|} \right) = \alpha/2. \tag{36}$$

The same technique explained in Remark 3 can be used to construct a one-sided confidence interval for $\eta^T f$.

Figure 1 demonstrates the distribution of TN and Tt statistics, stated in (29) and (35), by displaying their quantiles versus those of $U(0, 1)$. The figure also represents the distribution of the two statistics when the truncated normal and truncated t distributions in (29) and (35) are replaced with their untruncated counterparts. The figure certifies that the distribution of $Z$ and $T$ change from normal and t distributions to truncated normal and truncated t distributions, respectively, while accounting for the detection procedure.

**Remark 4** We would like to emphasize that the confidence intervals constructed by using (31) and (36) employ the knowledge of the conditional distributions given the estimated change points and their signs, $\{\hat{\tau} = \tau, \hat{s} = s\}$, and therefore, have to be interpreted conditionally. These confidence intervals can also be interpreted simultaneously after applying the Bonferroni correction to properly adjust for multiplicity. That is, we would compute confidence intervals at the level $1 - \frac{\alpha}{\hat{J}}$, where $\hat{J}$ is the number of change points estimated by mPRUTF.

One approach for constructing simultaneous confidence intervals in post-detection framework is to use the universal valid post-selection inference of Berk et al. (2013), by considering all possible detection procedures. More precisely, for any change point detection procedure $\hat{\tau} : \mathbf{y} \rightarrow \mathcal{T}$, where $\mathcal{T}$ is the collection of all possible change point sets, the goal of the universal
Valid Post-Detection Inference

post-selection inference, referred to as simultaneous PoSI, is to construct confidence intervals for $\eta_j^T f$. In this notation, $\eta_j$ denotes a nonzero contrast vector associated with $\hat{\tau}_j \in \hat{\tau}$. According to the simultaneous PoSI method, a valid confidence interval for $\eta_j^T f$ takes the form

$$\hat{I}_j(\alpha) = \left( \eta_j^T y \pm K_\alpha \sigma \| \eta_j \| \right),$$

where the constant $K_\alpha$ is derived such that

$$\Pr \left( \eta_j^T f \in \hat{I}_j(\alpha), \quad \forall \hat{\tau}_j \in \hat{\tau} \right) \geq 1 - \alpha, \quad \forall \hat{\tau} \in \mathcal{T}.$$

Note that the universal valid post-selection inference provides simultaneity both over the change points and over all detection procedures. This method has the advantage that it does not depend on the change point detection procedure. However, unless the number of possible change point sets is fairly small, this method is computationally challenging as the collection of all possible change point sets becomes intractable. Additionally, the confidence intervals derived from this method are unnecessarily wide as it disregards the knowledge of how the change points are estimated. To improve the universal valid post-detection inference, a method based on a simultaneous over selection (SoS) criterion was proposed by Zrnic and Jordan (2020). The method aims to construct confidence intervals for selected variables in a fixed stable model selection procedure in the regression context. We note that computing simultaneous confidence intervals is not among the goals of our inference and refer interested readers to Berk et al. (2013), Benjamini et al. (2019), Andrews et al. (2019) and Zrnic and Jordan (2020).

Remark 5 We have thus far explained how to carry out statistical inference for the detected change points after completing the PRUTF algorithm. In other words, we have computed post-selection $p$-values and confidence intervals conditional on the entire set of estimated change points $\hat{\tau}$. An alternative way is to perform statistical inferences for the change points in a sequential manner. In particular, suppose $\hat{\tau}_{j-1}$ and $\hat{s}_{j-1}$ are the vectors of detected change points and their corresponding signs at the iteration $j-1$ of the PRUTF algorithm. Additionally, suppose $\hat{\tau}_j$ is the detected change point at step $j$. In the sequential scheme, the interest is in making inference about the significance of change at the location $\hat{\tau}_j$, given $\{\hat{\tau}_j, \hat{s}_j\}$. Notice that the tools provided in Theorems 1 and 2 are still applicable in the sequential scheme, but the matrix $A$ and vector $q$ in the polyhedron representation must be adjusted accordingly. See Lockhart et al. (2014) and Tibshirani et al. (2016) which have pursued statistical inference in a sequential approach.

5 A Critique of Post-Detection Inference Methods

In post-detection inference, we lose part of the information in the data because it has already been used to estimate change points. As a result, the post-detection distributions change from normal or $t$ distributions to truncated normal or truncated $t$ distributions (Theorem 1 and Theorem 2). The performance of post-detection inference heavily depends on the amount of information in the data used for change point detection. Over-conditioning, which uses too
much information for the detection procedure and leaves little information for inference, leads to a significant loss in the power of tests and unacceptably wide confidence intervals; see Fithian et al. (2014), Lee et al. (2016) and Kivaranovic and Leeb (2020).

The over-conditioning issue in post-selection inference has motivated researchers to suggest various approaches that have higher left-over information for inference. One solution to the problem is data splitting or data carving in which the data is divided into two parts, one part for model selection and the other part for inference (Fithian et al., 2014). Another approach, put forward by Tian et al. (2018), is the idea of randomization, which selects the model based on not the actual dataset but a noise-perturbed version. Also, Liu et al. (2018) have suggested classifying the selected model into high-value and low-value submodels. The post-selection inference will then be conducted by only conditioning on the high-value submodel. Although these approaches improve the performance of post-selection inference, they also share some drawbacks; see Fithian et al. (2014).

The problem of wide length confidence intervals in post-selection inference has specifically been investigated by Kivaranovic and Leeb (2020) for models chosen by lasso. They have established that a confidence interval produced by the truncated Gaussian distribution with a finite truncation boundary (either the upper or lower bound) always has an infinite expected length. In the next theorem, we will show that the same property also holds for a truncated $t$ distribution. More precisely, let a truncation set $S$ to be the union of finitely many open intervals, where the intervals might be unbounded. In other words, $S$ can be represented in the form of $S = \bigcup_{i=1}^{k} (a_i, b_i)$, for a finite value $k$, and for some $a_1 < b_1 < a_2 < b_2 < \ldots < a_k < b_k$ in $\mathbb{R}$.

**Theorem 3**

a) Let $Z \sim \text{TN}(\mu, \sigma^2, S_Z)$, where the truncation set $S_Z$ is of the form $S_Z = \bigcup_{i=1}^{k} (a_i, b_i)$. Also, let $[L_Z(Z), U_Z(Z)]$ be a conditional confidence interval for $\mu$, given $S_Z$. Therefore, if the truncation set $S_Z$ is bounded, either from below ($-\infty < a_1$) or from above ($b_k < \infty$), then

$$E\left[U_Z(Z) - L_Z(Z)\right] = \infty. \quad (37)$$

b) Let $T \sim \text{Lt}(\mu, \sigma^2, d, S_T)$, where the truncation set $S_T$ is of the form $S_T = \bigcup_{i=1}^{m} (c_i, d_i)$. Also, let $[L_T(T), U_T(T)]$ be a conditional confidence interval for $\mu$, given $S_T$. Similar to part (a), if the truncation set $S_T$ is bounded, either from below ($-\infty < c_1$) or from above ($d_m < \infty$), then

$$E\left[U_T(T) - L_T(T)\right] = \infty. \quad (38)$$

The proof is provided in Appendix B.

Theorem 3 states that the truncation set is crucial in constructing a desirable confidence interval in post-selection inference. To figure out why post-selection confidence intervals are sometimes exceedingly wide, assume $T \sim \text{Lt}(\mu, \sigma^2, d, S_T)$. When the truncation set $S_T$ is
Valid Post-Detection Inference

\[ (a) \text{ Truncation set } S = [-3, 3]. \]

\[ (b) \text{ Truncation set } S = (-\infty, -3) \cup (3, \infty). \]

Figure 2: Lengths of confidence intervals for a truncated normal distribution and a truncated t distribution with a bounded truncation set (left panel) and an unbounded truncation set (right panel). The solid horizontal lines in both panels show the length of confidence intervals (at the level 0.95) for the usual (untruncated) normal and t distributions.

bounded, values of \( T \) which are close to the endpoints of \( \mathcal{S}_T \) leads to wide confidence intervals. This behaviour is because there are many values of \( \mu \) that would be consistent with the data (Lee et al., 2016; Kivaranovic and Leeb, 2020). On the other hand, when \( \mathcal{S}_T \) is unbounded, regardless of the values of \( T \), the interval length is always bounded. Similar arguments hold for a truncated normal distribution. Figure 2 displays the lengths of confidence intervals derived using truncated normal and truncated t distributions for a bounded and an unbounded truncation set. The left panel indicates that the length of confidence intervals for both truncated normal and truncated t distributions diverge as the values of \( z \) or \( t \) approach the edges of the truncation set. This observation certifies the results derived in Theorem 3. On the other hand, when the truncation set is unbounded, the right panel indicates that the length of confidence intervals for both distributions are bounded. Moreover, the interval length converges to the length of confidence interval obtained from an usual (untruncated) normal or t distributions as the values of the statistics diverge from the edges of the unbounded truncation sets. Additionally, for both cases of bounded and unbounded truncation set, the lengths of confidence intervals for the truncated t distribution are greater or equal than those of the truncated normal distribution.

To measure the quality of the confidence intervals built in (31) and (36), we inspect their corresponding truncation sets to determine whether they are unbounded. The next theorem verifies that the truncation boundaries provided in Theorems 1 and 2 are in fact bounded.

**Theorem 4** Consider the settings of Theorems 1 and 2. Also, let \( [V_{Z}^{-}(V), V_{Z}^{+}(V)] \) and \( [V_{T}^{-}(V, W), V_{T}^{+}(V, W)] \) be the truncation sets for the truncated normal and truncated t dis-
tributions, derived in (28) and (34), respectively. Then,

(a) either one or both truncation boundaries of (28) are finite, that is,

\[-\infty < V^-_Z(V) \quad \text{or} \quad V^+_Z(V) < \infty.\]

(b) either one or both truncation boundaries of (34) are finite, that is,

\[-\infty < V^-_T(V, W) \quad \text{or} \quad V^+_T(V, W) < \infty.\]

The proof is given in appendix C.

Remark 6 Theorem 4 states that the truncation sets derived in Theorems 2 and 3, for both truncated normal and t distributions, are bounded. Hence, from Theorem 4,

\[E\left[U_Z(Z) - L_Z(Z) \mid \{\hat{\tau}, \hat{s}\}\right] = \infty,\]

\[E\left[U_T(T) - L_T(T) \mid \{\hat{\tau}, \hat{s}\}\right] = \infty.\] (39)

In words, the average lengths of the confidence intervals for \(\eta^Tf\) with known \(\sigma^2\), given in (31), and with unknown \(\sigma^2\), given in (36), are infinite.

6 Optimal Post-Detection Inference

Motivated by the fact that post-detection confidence intervals might become extremely wide, we seek approaches to attain confidence intervals with narrower length properties. Given a target change point, we can essentially think of \(\hat{\tau}\) as containing two types of change points: those relevant to the discovery of the target change point and those irrelevant to it. The relevant change points are those change points that directly influence the detection of the target change point. In most cases, such points are located around the target change points. Therefore, one solution for improving the performance of post-detection inference is to allow only the relevant change points to participate in the analysis. In other words, the post-detection inference is carried out by only conditioning on a set containing relevant change points. We note that the set which includes these relevant change points is not necessarily a polyhedron. In the following, we will show that by narrowing down the conditional event to the set containing relevant change points, the resulting intervals becomes much shorter. This procedure will also lead to improved power for the associated hypothesis tests.

In the next two sections, we elaborate on two algorithms for conditioning on the relevant change point sets, leading to more powerful post-detection tests and narrower post-detection confidence intervals. We introduce two different setups. Section 6.1 involves post-detection inference by conditioning on the event that the given target change point is included in the estimated change point set. We refer to this setup as global post-detection because it verifies the significance of the target change point globally over the entire signal. In contrast, Section 6.2 involves statistical inference under the condition that the target change point and its adjacent change points are only included in the model. This setup is called local post-detection as it tests the significance of the target change points locally over its adjacent segments.
6.1 Global Post-Detection Inference

As explained before, global post-detection assumes that the target change point has been detected by the PRUTF algorithm and information about the rest of the change points is disregarded. In other words, for the given target change point \( \hat{\tau}_j \), we only assume that \( \hat{\tau}_j \in \hat{\tau} \). Now, the goal is to carry out post-detection inference by conditioning on the event \( \{ \hat{\tau}_j \in \hat{\tau} \} \).

Observe that this conditioning is different from conditioning on the entire change point set \( \tau \). We then attempt to compute a post-detection p-value for the hypothesis \( H_0 : D_j f = 0 \), as well as building a post-detection confidence interval for \( D_j f \). Depending on whether \( \sigma^2 \) is known or unknown, the post-detection tests are based on the statistics

\[
Z_{j}^{\text{glo}} = \frac{D_{y_j}^T y}{\sigma \|D_{y_j}^T\|} \quad \text{and} \quad T_{j}^{\text{glo}} = \frac{D_{y_j}^T y}{\bar{\sigma}_{j}^{(\text{glo})} \|D_{y_j}^T\|}
\]

which, under the null hypothesis, are distributed as the standard truncated normal and truncated \( t \) distributions, respectively. Given \( \{ \hat{\tau}_j \in \hat{\tau} \} \), we split \( y \) into two subvectors \( y_{1,j}^{\text{glo}} = (y_1, \ldots, y_{\hat{\tau}_j}) \) and \( y_{2,j}^{\text{glo}} = (y_{\hat{\tau}_j+1}, \ldots, y_n) \). The projection matrices \( P_{k,j}^{\text{glo}} \) associated with the subvectors \( y_{k,j}^{\text{glo}}, \; k = 1, 2 \), the block diagonal matrix \( B_{j}^{\text{glo}} = \text{diag}(P_{1,j}^{\text{glo}}, P_{2,j}^{\text{glo}}) \) and the sufficient statistic \( V_{j}^{\text{glo}} \) can be defined accordingly (see Section 4). The next theorem explains how to specify the truncation sets for the underlying distributions. We provide the proof of the theorem in Appendix D.

**Theorem 5** Consider the statistics \( Z_{j}^{\text{glo}} \) and \( T_{j}^{\text{glo}} \) as defined in (40).

(a) The conditional distribution of \( Z_{j}^{\text{glo}} \) given \( \{ \hat{\tau}_j \in \hat{\tau} \} \), under the null hypothesis \( H_0 : D_j f = 0 \), is the standard normal truncated to the set \( (\infty, V_{Z_j}^{-(\text{glo})}] \cup [V_{Z_j}^{+(\text{glo})}, \infty) \), where

\[
V_{Z_j}^{-(\text{glo})} = -\lambda_j D_{y_j}^T D_{y_j}^T \frac{D_{y_j}^T y}{\sigma \|D_{y_j}^T\|} \quad \text{and} \quad V_{Z_j}^{+(\text{glo})} = \frac{\lambda_j D_{y_j}^T D_{y_j}^T \hat{u}_{\hat{\tau}_j, -\hat{\tau}_j}}{\sigma \|D_{y_j}^T\|}.
\]

(b) The conditional distribution of \( T_{j}^{\text{glo}} \) given \( \{ \hat{\tau}_j \in \hat{\tau} \} \), under the null hypothesis, is \( t \) distribution with \( d_{j}^{(\text{glo})} = n - 2 \) degrees of freedom and truncated to the set \( (\infty, V_{T_j}^{-(\text{glo})}] \cup [V_{T_j}^{+(\text{glo})}, \infty) \), where

\[
V_{T_j}^{-(\text{glo})} = -\lambda_j D_{y_j}^T D_{y_j}^T \frac{D_{y_j}^T y}{\bar{\sigma}_{j}^{(\text{glo})} \|D_{y_j}^T\|} \quad \text{and} \quad V_{T_j}^{+(\text{glo})} = \frac{\lambda_j D_{y_j}^T D_{y_j}^T \hat{u}_{\hat{\tau}_j, -\hat{\tau}_j}}{\bar{\sigma}_{j}^{(\text{glo})} \|D_{y_j}^T\|},
\]

and \( \bar{\sigma}_{j}^{2(\text{glo})} = \| (I - P_{j}^{\text{glo}}) y \|^2 / d_{j}^{(\text{glo})} \).


We note that the TN and Tt statistics, provided in (29) and (35), respectively, can be constructed using the corresponding distributions in both parts (a) and (b) of Theorem 5. Therefore, we can apply the same procedure as in Section 4 to compute post-detection p-values and confidence intervals.

It is worth mentioning that \( \hat{\sigma}_j^2(\text{glo}) \) may not be a suitable estimator of \( \sigma^2 \). This is because global post-detection only assumes that \( \hat{\tau}_j \in \hat{\tau} \). Therefore, in the case that there exit other change points in \( \hat{\tau}_j^\text{glo} \), \( \hat{\sigma}_j^2(\text{glo}) \) is unable to accurately estimate the variation in the observations. As an alternative, we can apply Median Absolute Deviation (MAD) proposed by Hampel (1974), to robustly estimate \( \sigma^2 \). Also, see Baranowski et al. (2019) for more details about MAD estimation.

Let \( [L_j^\text{glo}, U_j^\text{glo}] \) and \( [L_j^\text{glo}, U_j^\text{glo}] \) be confidence intervals derived in the same manner as (31) and (36) using the pivotal quantities \( Z_j^\text{glo} \) and \( T_j^\text{glo} \). As the lower and upper bounds associated with their distributions are unbounded, we suspect that the underlying confidence intervals have finite expected lengths. In such a case, Kivaranovic and Leeb (2020) have provided an upper bound for the confidence intervals derived from a truncated normal distribution. In the following theorem, we will also give such an upper bound for confidence intervals using a truncated \( t \) distribution.

**Theorem 6** Let \( [L_j^\text{glo}, U_j^\text{glo}] \) and \( [L_j^\text{glo}, U_j^\text{glo}] \) be \((1 - \alpha)\)% confidence intervals derived from truncated normal and truncated \( t \) distributions given in Theorem 5.

(a) The length of confidence intervals derived from the truncated normal distribution is always upper bounded by

\[
U_j^\text{glo} - L_j^\text{glo} \overset{a.s.}{\leq} 2\sigma \Phi^{-1} \left( 1 - \frac{\alpha}{2} \right)^2 + V_j^\text{glo} - V_j^{-\text{glo}}.
\]

(b) For \( d^{(\text{glo})} \geq 3 \) and under the condition

\[
\alpha G_{d^{(\text{glo})}} \left( \frac{V_j^\text{glo} - V_j^{-\text{glo}}}{2\hat{\sigma}_j^{(\text{glo})}} \right) \geq G_{d^{(\text{glo})}} \left( \frac{1 - \alpha}{2} \right)
\]

the length of confidence intervals derived from the truncated \( t \) distribution is upper bounded by

\[
U_j^\text{glo} - L_j^\text{glo} \overset{a.s.}{\leq} 2\hat{\sigma}_j^{(\text{glo})} G_{d^{(\text{glo})}}^{-1} \left( 1 - \frac{\alpha}{2} \right)^2 + V_j^\text{glo} - V_j^{-\text{glo}}.
\]

A proof is provided in Appendix E.

### 6.2 Local Post-Detection Inference

Unlike in Section 6.1, here we are dealing with situations where the most relevant estimated change points to the target change point are involved in our inferential analysis. Obviously, the selection of \( \hat{\tau}_j \) as a change point relies on other estimated change points. Thus, conditioning only on \( \{\hat{\tau}_j \in \hat{\tau}\} \) is insufficient to decide whether \( \hat{\tau}_j \) is a meaningful change location. This fact indicates the need for knowledge about other change points. However, as previously discussed, conditioning our inference on the entire change point set yields an undesirable output such as...
wide confidence intervals. Note that the local nature of change point setting (Niu and Zhang, 2012) dictates that inferences for a change point should depend on its adjacent change points. For instance, only the immediate neighboring change points $\hat{\tau}_{j-1}$ and $\hat{\tau}_{j+1}$ play a role in the detection of the target change point $\hat{\tau}_j$, and the rest remain irrelevant. This fact has motivated us to develop the local post-detection algorithm.

Following the local property of change points, we suggest an idea for post-detection inference which leads to higher-powered tests and shorter confidence intervals. The idea is to condition the post-detection inference on the target change point as well as on its adjacent change points. In particular, for the given target change point $\hat{\tau}_j$, the goal is to test the hypothesis $H_0 : \mathbf{D}_j \mathbf{f} = 0$, given $\{(\hat{\tau}_{j-1}, \hat{\tau}_j, \hat{\tau}_{j+1}) \in \hat{\tau}\}$. In fact, this conditioning creates a new change point problem over the shorter subsignal $(f_{\hat{\tau}_j-1+1}, \ldots, f_{\hat{\tau}_j+1})$. We first define some notations for ease of exposition. Let $\mathbf{y}_{j}^{\text{loc}} = (y_{\hat{\tau}_j-1}, y_{\hat{\tau}_j})$ and $\mathbf{f}_{j}^{\text{loc}} = (f_{\hat{\tau}_j-1}, \ldots, f_{\hat{\tau}_j+1})$ denote the subvectors of $\mathbf{y}$ and $\mathbf{f}$ from $\hat{\tau}_{j-1}$ to $\hat{\tau}_{j+1}$. Also, let $\mathbf{D}_j$ be the $(\hat{\tau}_{j+1} - \hat{\tau}_{j-1} - 1) \times (\hat{\tau}_{j+1} - \hat{\tau}_{j-1})$ version of matrix $\mathbf{D}$. Applying these notations, the hypothesis of interest can be re-expressed as $H_0 : \mathbf{D}_j, \Delta_j, \mathbf{f}_{j}^{\text{loc}} = 0$, where $\Delta_j = \hat{\tau}_j - \hat{\tau}_{j-1}$ and $\mathbf{D}_j, \Delta_j$ is the $\Delta_j$-th row of $\mathbf{D}_j$. For inference, similar approach taken in the global post-detection can be applied, but this time using $\mathbf{y}_{j}^{\text{loc}}$ and $\mathbf{D}_j$. Depending on whether $\sigma^2$ is known or unknown, the post-detection tests are based on statistics

$$Z_{j}^{\text{loc}} = \frac{\mathbf{D}_j, \Delta_j, \mathbf{y}_{j}^{\text{loc}}}{\sigma \|\mathbf{D}_j, \Delta_j\|}, \quad T_{j}^{\text{loc}} = \frac{\mathbf{D}_j, \Delta_j, \mathbf{y}_{j}^{\text{loc}}}{\tilde{\sigma}_{j}^{\text{loc}} (\mathbf{D}_j, \Delta_j)}.$$  

(46)

We will obtain the distributions of these statistics in the next theorem whose proof is similar to that of Theorem 5 by replacing $\mathbf{y}, \mathbf{D}, \hat{\tau}_j, \mathbf{f}_{j}^{\text{loc}}$ and $\tilde{\sigma}_{j}^{2\text{(glo)}}$ with $\mathbf{y}_{j}^{\text{loc}}, \mathbf{D}_j, \Delta_j, \mathbf{f}_{j}^{\text{loc}}$ and $\tilde{\sigma}_{j}^{2\text{(loc)}}$.

**Theorem 7** Consider the statistics $Z_{j}^{\text{loc}}$ and $T_{j}^{\text{loc}}$ defined in (46).

(a) Given $\{(\hat{\tau}_{j-1}, \hat{\tau}_j, \hat{\tau}_{j+1}) \in \hat{\tau}\}$, the conditional distribution of $Z_{j}^{\text{loc}}$, under the null hypothesis $H_0 : \mathbf{D}_j \mathbf{f} = 0$, or equivalently $H_0 : \mathbf{D}_j, \Delta_j, \mathbf{f}_{j}^{\text{loc}} = 0$, is the standard normal truncated to $(-\infty, \mathcal{V}_{Z_{j}^{\text{loc}}}^{-}(\infty))$, where

$$\mathcal{V}_{Z_{j}^{\text{loc}}}^{-}(\infty) = \frac{-\lambda_j \mathbf{D}_j, \Delta_j, \mathbf{D}_j^T, \Delta_j + \mathbf{D}_j, \Delta_j, \mathbf{D}_j^T, -\Delta_j \hat{\tau}_j}{\sigma \|\mathbf{D}_j, \Delta_j\|}.$$  

$$\mathcal{V}_{Z_{j}^{\text{loc}}}^{+}(\infty) = \frac{\lambda_j \mathbf{D}_j, \Delta_j, \mathbf{D}_j^T, \Delta_j + \mathbf{D}_j, \Delta_j, \mathbf{D}_j^T, -\Delta_j \hat{\tau}_j}{\sigma \|\mathbf{D}_j, \Delta_j\|},$$  

(47)

(b) Given $\{(\hat{\tau}_{j-1}, \hat{\tau}_j, \hat{\tau}_{j+1}) \in \hat{\tau}\}$, the conditional distribution of $T_{j}^{\text{loc}}$, under the null hypothesis, is t-distribution with $d_{j}^{\text{loc}} = (\hat{\tau}_{j+1} - \hat{\tau}_{j-1}) - 2$ truncated to the set $(-\infty, \mathcal{V}_{T_{j}^{\text{loc}}}^{-}(\infty))$, where

$$\mathcal{V}_{T_{j}^{\text{loc}}}^{-}(\infty) = \frac{-\lambda_j \mathbf{D}_j, \Delta_j, \mathbf{D}_j^T, \Delta_j + \mathbf{D}_j, \Delta_j, \mathbf{D}_j^T, -\Delta_j \hat{\tau}_j}{\tilde{\sigma}_{j}^{\text{loc}} (\mathbf{D}_j, \Delta_j)}.$$  

$$\mathcal{V}_{T_{j}^{\text{loc}}}^{+}(\infty) = \frac{\lambda_j \mathbf{D}_j, \Delta_j, \mathbf{D}_j^T, \Delta_j + \mathbf{D}_j, \Delta_j, \mathbf{D}_j^T, -\Delta_j \hat{\tau}_j}{\tilde{\sigma}_{j}^{\text{loc}} (\mathbf{D}_j, \Delta_j)},$$  

(48)
\[
\text{and } \hat{\sigma}_j^2(\text{loc}) = \|(I - P_j^{\text{loc}}) y_j^{\text{loc}}\|^2 / d_j^{\text{loc}}.
\]

In the same manner as described in Section 4, the TN and Tt statistics can be employed to perform hypothesis testings and construct confidence intervals.

**Remark 7** From Theorem 7, it turns out that the lower and upper limits of truncation sets using local post-detection inference are infinite. Therefore, similar to global post-detection inference, the expected lengths of confidence intervals associated with such distributions are upper bounded. In particular, let \([L_{Zj}^{\text{loc}}, U_{Zj}^{\text{loc}}]\) and \([L_{Tj}^{\text{loc}}, U_{Tj}^{\text{loc}}]\) be \((1 - \alpha)\)% confidence intervals derived from truncated normal and truncated \(t\) distributions, respectively, provided in Theorem 7. Hence, the lengths of such intervals are upper bounded by

- when \(\sigma^2\) is known, the length of confidence intervals is always upper bounded by
  \[
  U_{Zj}^{\text{loc}} - L_{Zj}^{\text{loc}} \overset{a.s.}{\leq} 2 \sigma \Phi^{-1}\left(1 - \frac{\alpha}{2}\right) + \nu_{Zj}^{+\text{loc}} - \nu_{Zj}^{-\text{loc}}. \tag{49}
  \]
- when \(\sigma^2\) is unknown, for \(d^{\text{loc}} \geq 3\) and under the condition
  \[
  \alpha G_{d^{\text{loc}}} \left( - \frac{\nu_{Tj}^{+\text{loc}} - \nu_{Tj}^{-\text{loc}}}{2 \hat{\sigma}_j^{\text{loc}}} \right) \geq G_{d^{\text{loc}}} \left( G^{-1}_{d^{\text{loc}}} \left( \frac{\alpha}{2} \right) - \frac{\nu_{Tj}^{+\text{loc}} - \nu_{Tj}^{-\text{loc}}}{2 \hat{\sigma}_j^{\text{loc}}} \right), \tag{50}
  \]
  the length of confidence intervals is upper bounded by
  \[
  U_{Tj}^{\text{loc}} - L_{Tj}^{\text{loc}} \overset{a.s.}{\leq} 2 \hat{\sigma}_j^{\text{loc}} G^{-1}_{d^{\text{loc}}} \left(1 - \frac{\alpha}{2}\right) + \nu_{Tj}^{+\text{loc}} - \nu_{Tj}^{-\text{loc}}. \tag{51}
  \]

### 7 Numerical Studies

In this section, we investigate the performance of our proposed post-detection inference approaches for a piecewise constant signal \((r = 0)\) and a piecewise linear signal \((r = 1)\). We compare the performance of the approaches in terms of the empirical power for hypothesis testings and coverage probabilities for confidence intervals. We also apply the implementation of the proposed methods to three real datasets, which have been used to estimate change points using the PRUTF algorithm in Mehrizi and Chenouri (2020).

#### 7.1 Simulations

In order to investigate our proposed approaches for post-detection inference, we consider two signals: piecewise constant and piecewise linear. Suppose that \(f\) is a piecewise constant or linear signal of size \(n = 500\), with four change points, \(J_0 = 4\), at locations \(\tau = \{100, 200, 300, 400\}\).

For the piecewise constant signal, we consider a signal with the starting point 0 and the jump sizes \(\delta\). That is,

\[
 f_t^{\text{con}} = \begin{cases} 
 0 & \text{for } 1 \leq t \leq 100 \text{ or } 201 \leq t \leq 300 \text{ or } 401 \leq t \leq 500, \\
 \delta & \text{for } 101 \leq t \leq 200 \text{ or } 301 \leq t \leq 400,
\end{cases}
\]
where \( \delta \in \{1, 1.5, 2, \ldots, 4.5, 5\} \). Also, for the piecewise linear signal, we consider

\[
 f_{t}^{\text{lin}} = \begin{cases} 
 \delta (-0.5 + t) & \text{for } 1 \leq t \leq 100 \text{ or } 201 \leq t \leq 300 \text{ or } 401 \leq t \leq 500, \\
 \delta (0.5 - t) & \text{for } 101 \leq t \leq 200 \text{ or } 301 \leq t \leq 400.
\end{cases}
\]

We have generated a sample \( y \) from \( N(f, \sigma^2 I) \), with \( \sigma^2 = 1 \), for both piecewise constant and piecewise linear signals, and implemented PRUTF to the estimate change points. Next, we have applied our proposed approaches: polyhedron post-detection (poly), global post-detection

\[ (a) \text{ Piecewise constant with known } \sigma^2. \]

\[ (b) \text{ Piecewise constant with unknown } \sigma^2. \]

\[ (c) \text{ Piecewise linear with known } \sigma^2. \]

\[ (d) \text{ Piecewise linear with unknown } \sigma^2. \]

**Figure 3:** The empirical powers of the three proposed approaches, polyhedron, global and local post-detection inference. The results are provided for both cases when \( \sigma^2 \) is assumed known and unknown. The two top panels display the empirical powers for the piecewise constant signal. Also, the two bottom panels show the empirical powers for the piecewise linear signal. The solid orange lines in both panels (b) and (d) display the empirical powers of the global post-detection approach when \( \hat{\sigma}^2 (\text{glo}) \), an estimation of \( \sigma^2 \), is replaced with MAD estimation.
and local post-detection (local) inference procedures. Over \( N = 10000 \) repetitions, we have computed the empirical power of a change point hypothesis using the three methods poly, global and local, for both cases of known and unknown \( \sigma^2 \). Recall that the empirical power is defined as the ratio between the number of times the true change point is correctly detected and has \( p \)-value less than \( \alpha \), and the number of times the true change point is correctly detected, see Hyun et al. (2018b). Here, \( \alpha \) is set to 0.05. The results of the second change point, \( \tau_2 = 200 \), are displayed in Figure 3. For the piecewise constant signal, the empirical powers are computed over repetitions that contain the underlying change point. This is the spike contrast put forward by Hyun et al. (2018a). Whereas the window contrast, used in Jewell et al. (2019), with size, \( h = 15 \) is employed for the computation in the piecewise linear signal. This is because PRUTF mostly estimates change points near but not exactly at change points when the polyhedron degree \( r \) increases.

Panels (a) and (b) of Figure 3 provide the empirical powers for the piecewise constant signal with both known and unknown \( \sigma^2 \). When \( \sigma^2 \) is known, the local and global post-detection approaches outperform the polyhedron post-detection approach, as expected. In the case of unknown \( \sigma^2 \), the local post-detection approach again performs better than the polyhedron one. However, the global post-detection approach has a poor performance due to inaccurate \( \sigma^2 \) estimation, see Section 6.1. By replacing \( \hat{\sigma}^2_j^{(glo)} \) in Theorem 5 with the MAD estimation, the empirical power of the global post-detection approach has improved. The results for the piecewise linear signal are also plotted in panels (c) and (d) of Figure 3. Similar patterns are observed for this signal, as well.

Moreover, we have computed the coverage probability for confidence intervals, obtained using the polyhedron, global and local post-detection methods. The outputs are reported in Table 1.

| \( \delta \) | Poly | Global | Local | Poly | Global | Global (MAD) | Local |
|---|---|---|---|---|---|---|---|
| 2 | 0.9515 | 0.9527 | 0.9515 | 0.9515 | 0.9708 | 0.9527 | 0.9504 |
| 3 | 0.9554 | 0.9564 | 0.9554 | 0.9554 | 0.9817 | 0.9577 | 0.9564 |
| 4 | 0.9547 | 0.9586 | 0.9547 | 0.9605 | 0.9874 | 0.9601 | 0.9551 |
| 5 | 0.9543 | 0.9531 | 0.9543 | 0.9618 | 0.9889 | 0.9553 | 0.9545 |

| \( \delta \) | Poly | Global | Local | Poly | Global | Global (MAD) | Local |
|---|---|---|---|---|---|---|---|
| 2 | 0.9473 | 0.9531 | 0.9473 | 0.9473 | 0.9660 | 0.9450 | 0.9473 |
| 3 | 0.9550 | 0.9555 | 0.9550 | 0.9570 | 0.9737 | 0.9581 | 0.9555 |
| 4 | 0.9453 | 0.9491 | 0.9453 | 0.9491 | 0.9779 | 0.9514 | 0.9441 |
| 5 | 0.9431 | 0.9419 | 0.9431 | 0.9506 | 0.9786 | 0.9426 | 0.9431 |

Table 1: The coverage probabilities of confidence intervals obtained using polyhedron, global and local post-detection approaches. The results are reported for four values of \( \delta \in \{2, 3, 4, 5\} \).
Valid Post-Detection Inference

7.2 Real Data Analysis

For real datasets, Mehrizi and Chenouri (2020) have analyzed three datasets: UK House Price Index, the GISS Surface Temperature and COVID-19. In the following, we will apply our proposed post-detection inference approaches to evaluate the significance of change points estimated using PRUTF for these datasets.

Example 1 The UK House Price Index (HPI) is a National Statistic that shows changes in the value of residential properties in England, Scotland, Wales and Northern Ireland. The HPI measures the price changes of residential housing by calculating the price of completed house sale transactions as a percentage change from some specific start date. The UK HPI uses the hedonic regression model as a statistical approach to producing estimates of the change in house prices for each period. For more details, see https://landregistry.data.gov.uk/app/ukhpi. Many researchers, including Baranowski et al. (2019) and Anastasiou and Fryzlewicz (2019), have studied the UK HPI dataset by carrying out change point analysis. In the current study, we consider monthly percentage changes in the UK HPI at Tower Hamlets (an east borough of London) from January 1996 to November 2018.

The PRUTF algorithm has found five change points located at months: December 2002, April 2008 and August 2009, May 2012 and August 2015. See Mehrizi and Chenouri (2020) for more details. We have provided the results of post-detection inference for these change points in panel (a) of Figure 4. The plot displays 95\% post-detection confidence intervals for the change points. Based on the polyhedron and local post-detection results, we have concluded that all five estimated change points are significant as their corresponding intervals excluded zero.

Example 2 The Goddard Institute for Space Studies (GISS) monitors broad global changes around the world. The GISS Surface Temperature Analysis (GISTEMP) is an estimate of the global surface temperature changes (see https://www.giss.nasa.gov). In the analysis of GISTEMP data, the temperature anomalies are used rather than the actual temperatures. A temperature anomaly is a difference from an average or baseline temperature. The baseline temperature is typically computed by averaging thirty or more years of temperature data (1951 to 1980 in the current dataset). A positive anomaly indicates the observed temperature was warmer than the baseline, while a negative anomaly indicates the observed temperature was cooler than the baseline. For more details see Lenssen et al. (2019).

The GISTEMP dataset has been frequently explored in change point literature; for example, see James and Matteson (2015) and Baranowski et al. (2019). We have considered the monthly land-ocean temperature anomalies recorded from January 1880 to August 2019 (see https://data.giss.nasa.gov/gistemp). For this dataset, Mehrizi and Chenouri (2020) have identified six change points located at months: September 1899, February 1911, May 1929, April 1941, March 1960, October 1984.

We have applied our proposed post-detection inference approaches to assess the significance of these estimated change points by computing the associated confidence intervals. The results are provided in the right panel of Figure 4. The outputs obtained from both polyhedron and local
Figure 4: Valid post-detection confidence intervals for both the UK HPI dataset (left panel) and GISTEMP dataset (right panel). For each estimated change point, valid 95% post-detection confidence intervals, obtained from polyhedron, global using both pooled and MAD variances, and local post-detection inference approaches, are displayed.

Post-detection inference have confirmed the significance of all estimated change points except the one in September 1899.

Example 3 Since the initial outbreak of Coronavirus Disease (COVID-19) in Wuhan, China, in mid-November 2019, the virus has rapidly spread throughout the world. The pandemic has infected 167.25 million people and killed more than 3.46 (up to March 4, 2021) people, greatly stressing public health systems and adversely influencing global society and economies. Therefore, every country has attempted to slow down the transmission rate by various regional and national policies such as the declaration of national emergencies, quarantines and mass testing. Of vital interest to governments is understanding the pattern of the epidemic growth and assessing the effectiveness of policies undertaken. A scientist can investigate these matters by analyzing the sequence of infection data for COVID-19. Change point detection is one possible framework for studying the behaviour of COVID-19 infection curves and retrospectively investigating the effectiveness of interventions. By detecting the locations of alterations in the curves, change point analysis gives us insights into changes in the transmission rate or efficiency of interventions. It also enables us to raise warning signals if the disease pattern changes.

We have analyzed the logarithm of the cumulative daily number of confirmed cases using a piecewise linear model. The choice of the piecewise linear model is natural because the slope of each segment, estimated using the detected change points, indicates the growth rate of the coronavirus. Consequently, these slopes allow us to compare the virus growth rate among estimated segments and evaluate the effectiveness of undertaken strategies. Additionally, the linear trend
of the last segment can be used to predict the status of the pandemic for future dates.

Mehrizi and Chenouri (2020) have applied the mPRUTF algorithm to detect change points that have occurred in the transformed COVID-19 datasets from March 10, 2020 until April 30, 2021, for Australia, Canada, United Kingdom and the United States. We have applied our proposed post-detection inference approaches to these datasets. The 95\% post-detection confidence intervals for the identified change points are provided in Figure 5. For example, for Canada, based on the polyhedron and local post-detection inference approaches, the change points located on March 26, 2020; April 9, 2020; May 11, 2020; August 31, 2020 and January 12, 2021 are significant. For the United Kingdom, based on the polyhedron and local post-detection approaches, only the change point located on June 22, 2020 is insignificant. Moreover, the figure shows that the confidence intervals for March 18, 2021 in Canada and for February 23, 2021 in the United States derived from polyhedron post-detection method are very wide and skewed. These observations certify the results provided in Theorem 3.

8 Discussion

In this manuscript, we have attempted to quantify the uncertainty of change points estimated by the PRUTF algorithm, put forward by Mehrizi and Chenouri (2020). By applying trend filtering, PRUTF yields an efficient algorithm to detect change points in a piecewise polynomial signal. We have provided a post-detection inference framework to compute valid p-values for the significance of the estimated change points. We have also constructed confidence intervals for the magnitude of such estimated change points. These inferences have been executed by implementing conditional inference through three distinct conditional detection events, hence, giving polyhedron, global and local post-detection inference procedures. We have also shown that the global and local post-detection inferences lead to higher-powered tests by conditioning on much smaller detection events.

For future work in this area, there are several possible ideas. We would like to apply our proposed post-detection approaches to the extensively used change point detection algorithms such as binary segmentation, its variants and $\ell_0$-norm segmentation. Our focus in this work has been on the independent and identically distributed random noises to precisely highlight the main ideas. However, our proposed approaches are not confined only to these types of random noises. One possible extension would be to incorporate more complex random noises. For example, we can consider heavy-tailed distributions or even auto-correlated random noises. The extension of the proposed approaches in this work will allow us to provide inferential tools for a wide range of applications.
Figure 5: Valid post-detection confidence intervals for COVID-19 datasets. For each estimated change point, %95 post-detection confidence intervals obtained from polyhedron, global with both pooled and MAD variance estimations, and local post-detection inference approaches are displayed.

Appendix A

A.1 Proof of Theorem 1

a) Note that

\[ Ay = A(P_n + P - P_n) y = \left( \frac{A\eta}{\|\eta\|} \right) \frac{\eta^T y}{\|\eta\|} + AV, \]

which allows us to rewrite \( \{Ay \geq q\} \) as

\[ \{Ay \geq q\} = \left\{ \sigma \left[ A\eta/\|\eta\| \right] Z + [ AV - q ] \geq 0 \right\}. \]

This representation leads to the truncation set

\[ [\nu^- Z, \nu^+ Z] = \bigcap_{i=1}^{\text{row}(A)} \left\{ z \in \mathbb{R} : \sigma \left[ A\eta/\|\eta\| \right] z + [ AV - q ] \geq 0 \right\}. \]
Observe that solving the inequality in the above set with respect to $z$ depends on the sign of $\rho_i = \frac{[A\eta/\|\eta\|]_i}{\|\eta\|}$. Therefore,

$$
\left\{ z \in \mathbb{R} : \sigma \rho_i z + [AV-q]_i \geq 0, i = 1, \ldots, |\text{row}(A)| \right\} = \left\{ \begin{array}{ll}
z \geq \frac{[q-AV]}{\sigma \rho_i} & : \rho_i > 0, \\
z \leq \frac{[q-AV]}{\sigma \rho_i} & : \rho_i < 0, \\
0 \leq [AV-q]_i & : \rho_i = 0. 
\end{array} \right.
$$

The preceding statement leads to the interval $[V^{-z}, V^{+z}]$, provided $V^{-z}, V^{+z}$ are provided in (28). Given $V$, these truncation boundaries are fixed. Therefore, the distribution of $Z$ given $\{V, Ay \geq q\}$ is equivalent to the distribution of a normal distribution constrained to the interval $[V^{-z}, V^{+z}]$. Hence

$$Z \; \mid \; \{V, Ay \geq q\} \sim \text{TN}(\eta_T f, 1, [V^{-z}, V^{+z}]).$$

b) Applying the probability integral transform for (27) yields

$$1 - \Phi^{[V^{-z}, V^{+z}]} (Z) \; \mid \; \{V, Ay \geq q\} \sim U(0, 1).$$

For every $0 \leq u \leq 1$, by marginalizing the above statement over $V$, we have

$$\Pr \left( 1 - \Phi^{[V^{-z}, V^{+z}]} (Z) \leq u \; \mid \; Ay \geq q \right) = \int_V \Pr \left( 1 - \Phi^{[V^{-z}, V^{+z}]} (Z) \leq u \; \mid \; V, Ay \geq q \right) \Pr (V \; \mid \; Ay \geq q) \, dv = u \int_V \Pr (V \; \mid \; Ay \geq q) \, dv = u,$$

which establishes Equation (29).

### A Proof of Theorem 2

a) Let $W = \| (I - P + P_\eta) y \|^2$, since both projection matrices $P$ and $P_\eta$ are symmetric and idempotent, then

$$W = \| (I - P) y \|^2 + \left( \frac{\eta^T y}{\|\eta\|} \right)^2 + 2 \left( \frac{\eta^T (I - P) y}{\|\eta\|} \right) \left( \frac{\eta^T y}{\|\eta\|} \right)$$

$$= \hat{\sigma}^2 \left[ d + T^2 + 2 \left( \frac{\eta^T (I - P) y}{\hat{\sigma} \|\eta\|} \right) T \right].$$

From (A.52), the polyhedron set can be rewritten as

$$\{Ay \geq q\} = \left\{ \left( \frac{A\eta}{\|\eta\|} \right) \frac{\eta^T y}{\|\eta\|} + AV \geq q \right\}.$$
Multiplying both sides of the above inequality by \( W/\hat{\sigma}^2 \) yields the truncation set
\[
[V^-_T, V^+_T] = \bigcap_{i=1}^{\text{row}(A)} \left\{ t \in \mathbb{R} : \left( \frac{W_{\rho_i}}{\hat{\sigma}} \right) t + \left[ AV - q \right] , \left( d + t^2 + 2 \frac{\eta^T(I - P)y}{\hat{\sigma} \| \eta \|} \right) t \geq 0 \right\}
\]
\[
= \bigcap_{i=1}^{\text{row}(A)} \left\{ t \in \mathbb{R} : \left[ AV - q \right] , t^2 + \left( 2 \frac{\eta^T(I - P)y}{\hat{\sigma} \| \eta \|} \left[ AV - q \right] + \frac{W_{\rho_i}}{\hat{\sigma}} \right) t \right\}
\]
\[+ \left[ AV - q \right] , d \geq 0 \}. \tag{A.53}
\]

In the above intersection, for each \( i, i = 1, \ldots, \text{row}(A) \), the inequality can be solved explicitly. Note that \([V^-_T, V^+_T]\) is a bounded interval (see Theorem 4).

On the other hand, \( W \) can also be decomposed as
\[
W = \| (I - P + \rho_n) y \|^2 = \| y \|^2 - \| (P - \rho_n) y \|^2. \tag{A.54}
\]

Therefore, conditioning on \( \{ V, \| y \|^2 \} \) is equivalent to conditioning on \( \{ V, W \} \). This means that, given \( \{ V, \| y \|^2 \} \), the truncation boundaries \( V^-_T = V^-_T (V, W) \) and \( V^+_T = V^+_T (V, W) \) are fixed. Therefore, the distribution of \( T \) given \( \{ V, \| y \|^2 \} \) is equivalent to the distribution of a \( t \) distribution with \( d \) degrees of freedom, constrained to \([V^-_T, V^+_T]\).

Hence,
\[
T \bigg| \left\{ V, \| y \|^2, Ay \geq q \right\} \sim \text{Tr}(\eta^T f, 1, d, [V^-_T, V^+_T])
\]

b) Similar to part (b) of Theorem 1, by marginalization of (33) over \( \{ V, W \} \), we obtain the statement in (35).

### B Proof of Theorem 3

(a) The result of this part has been proved in Kivaranovic and Leeb (2020). Here, we give the proof in our notation. Lemma A.4 of the latter reference shows that for \( b_k < \infty \),
\[
\lim_{z \to b_k^-} (b_k - z) L_Z(z) = -\sigma^2 \log \left( 1 - \alpha/2 \right),
\]
\[
\lim_{z \to b_k^-} (b_k - z) U_Z(z) = -\sigma^2 \log \left( \alpha/2 \right).
\]

The above equations together leads to
\[
\lim_{z \to b_k^-} (b_k - z) \left[ U_Z(z) - L_Z(z) \right] = \sigma^2 \log \left( \frac{(2 - \alpha)/\alpha}{(2 - \alpha)/\alpha} \right)
\]

This limit states that there exits an \( \epsilon > 0 \), such that
\[
U_Z(z) - L_Z(z) \geq \frac{\sigma^2 \log \left( \frac{(2 - \alpha)/\alpha}{(2 - \alpha)/\alpha} \right)}{2 (b_k - z)} \quad \text{for any } z \in (b_k - \epsilon, b_k) \cap S_Z.
\]
Now, let \( f^* = \inf \left\{ f^S_{\mu, \sigma^2}(z) : z \in (b_k - \epsilon, b_k) \cap S_Z \right\} \). Clearly, \( f^* > 0 \) as \( f^S_{\mu, \sigma^2} \) is a probability density function. Therefore,

\[
E\left[ U_Z(Z) - L_Z(Z) \right| Z \in S_Z \right] = \int_{z \in S_Z} \left[ U_Z(z) - L_Z(z) \right] f^S_{\mu, \sigma^2}(z) \, dz \\
\geq \frac{\sigma^2 \log \left( \frac{(2 - \alpha)}{\alpha} \right)}{2} \int_{z \in (b_k - \epsilon, b_k) \cap s_Z} \frac{f^S_{\mu, \sigma^2}(z)}{b_k - z} \, dz \\
\geq \frac{\sigma^2 \log \left( \frac{(2 - \alpha)}{\alpha} \right)}{2} f^* \int_{z \in (b_k - \epsilon, b_k) \cap s_Z} \frac{1}{b_k - z} \, dz = \infty.
\]

This can be similarly shown for \( a_i > -\infty \).

(b) From the result in part (a), it suffices to show that

\[
E \left[ U_T(T) - L_T(T) \right] \geq E \left[ U_Z(Z) - L_Z(Z) \right].
\]

To this end, we show that a random variable \( T \sim T_t (\mu, \sigma^2, d, S_T) \) has a heavier tail than a random variable \( Z \sim TN (\mu, \sigma^2, S_T) \). It is known that a random variable \( Y \) has heavier tails than \( X \) if and only if

\[
\lim_{t \to \infty} \frac{1 - F_Y(t)}{1 - F_X(t)} = \infty.
\]

This equation is equivalent to

\[
\lim_{t \to \infty} \frac{f_Y(t)}{f_X(t)} = \infty.
\]

Now, let \( f^S_{\mu, \sigma^2}(\cdot) \) and \( g^S_{\mu, \sigma^2, d}(\cdot) \), denote the density function of \( Z \) and \( T \), respectively. Therefore,

\[
\lim_{x \to \infty} \frac{g^S_{\mu, \sigma^2, d}(x)}{f^S_{\mu, \sigma^2}(x)} = \lim_{y \to \infty} \frac{g^S_T(y)}{f^S_{0, 1}(y)} = \lim_{y \to \infty} \exp \left\{ \log g^S_T(y) - \log f^S_{0, 1}(y) \right\} \\
= \exp \left\{ \lim_{y \to \infty} -\frac{d + 1}{2} \log \left( 1 + \frac{y^2}{d} \right) + \frac{1}{2} \frac{y^2}{d} + C \right\} \\
= \exp \left\{ \lim_{y \to \infty} y^2 \left[ -\frac{d + 1}{2y^2} \log \left( 1 + \frac{y^2}{d} \right) + \frac{1}{2} \right] + C \right\},
\]

where in the first equality we use the change variable \( y = (x - \mu)/\sigma \). Since

\[
\lim_{y \to \infty} \frac{1}{y^2} \log \left( 1 + \frac{y^2}{d} \right) = 0,
\]

we have

\[
\lim_{x \to \infty} \frac{g^S_{\mu, \sigma^2, d}(x)}{f^S_{\mu, \sigma^2}(x)} = \infty.
\]
The preceding result proves that the truncated $t$ distribution has heavier tails than truncated normal.

Now, since $[L_Z(Z), U_Z(Z)]$ and $[L_T(T), U_T(T)]$ are confidence intervals at the same level for truncated normal and truncated $t$ distributions, respectively, and $U_Z(Z) - L_Z(Z)$ and $U_T(T) - L_T(T)$ are two non-negative random variables such that

$$U_T(T) - L_T(T) \overset{a.s.}{\geq} U_Z(Z) - L_Z(Z).$$

Hence,

$$E\left[U_T(T) - L_T(T)\right] \geq E\left[U_Z(Z) - L_Z(Z)\right].$$

The above equation and Equation (A.55) together lead to the statement in (38).

C Proof of Theorem 4

(a) From Equation (28), it can be viewed that $V_Z^-$ and $V_Z^+$ are unbounded if both sets $\{i : \rho_i < 0\}$ and $\{i : \rho_i > 0\}$ are empty, where $\rho_i = [A\eta/\|\eta\|]_i$, for $i = 1, \ldots, |\text{row}(A)|$. This condition implies $\rho_i = 0$ for all $i$ which in turn requires that $A\eta = 0$ for all $\eta \neq 0$. Hence every vector $\eta \neq 0$ is orthogonal to the row space of the matrix $A$, which contradicts with the fact that the conditioning set is a polyhedron.

(b) From (34), $V_T^-$ and $V_T^+$ are unbounded if the inequality

$$\left[AV - q\right]_i t^2 + \left(\frac{2\eta^T(I - P)y}{\sigma \|\eta\|} \left[AV - q\right]_i + \frac{W\rho_i}{\sigma}\right) t + \left[AV - q\right]_i d \geq 0 \quad (A.56)$$

always holds for any $i = 1, \ldots, |\text{row}(A)|$. This requires that the quadratic equation in (A.56) to have at most one real root, i.e.,

$$\left(\frac{2\eta^T(I - P)y}{\sigma \|\eta\|} \left[AV - q\right]_i + \frac{W\rho_i}{\sigma}\right)^2 - 4 d \left[AV - q\right]_i^2 \leq 0,$$

as well as the coefficient sign of $t^2$ must be positive, i.e., $[AV - q]_i > 0$. According to (28), this latter condition occurs if and only if $\rho_i = 0$, for all $i = 1, \ldots, |\text{row}(A)|$, which is impossible as shown in part (a).

D Proof of Theorem 5

(a) Since $s_{\hat{\tau}_j}$ is the sign of $[\hat{D}f]_{\hat{\tau}_j} = \hat{D}_{\hat{\tau}_j} \hat{f}$, hence, $s_{\hat{\tau}_j} \hat{D}_{\hat{\tau}_j} \hat{f} \geq 0$. From the primal-dual relationship in (4), we get

$$s_{\hat{\tau}_j} \hat{D}_{\hat{\tau}_j} (y - D^T \hat{\lambda}_j) \geq 0, \quad (A.57)$$

where $\lambda_j$ is the value of the regularization parameter associated with $\hat{\tau}_j$. Now, the goal is to find a range for $Z_{glo}^{j\text{th}}$ for which $\hat{\tau}_j$ is a change point. Recall that from KKT conditions of the
dual problem (3), if \( \hat{\tau}_j \) is a change point then \( \hat{\lambda}_j, \hat{\tau}_j = s_{\hat{\tau}_j} \lambda_j \). Applying the decomposition \( y = P_\eta y + (I - P_\eta) y \) into (A.57) and setting \( \hat{\lambda}_j, \hat{\tau}_j = s_{\hat{\tau}_j} \lambda_j \), we have

\[
s_{\hat{\tau}_j} D_{\hat{\tau}_j} \left( \frac{\eta}{\|\eta\|^2} \eta^T y + (I - P_\eta) y - s_{\hat{\tau}_j} \lambda_j D_{\hat{\tau}_j}^T \hat{\lambda}_{\hat{\tau}_j} \right).
\]

Hence, setting \( \eta = D_{\hat{\tau}_j}^T \) in the preceding equation, for \( s_{\hat{\tau}_j} = 1 \),

\[
D_{\hat{\tau}_j} y \geq \lambda_j D_{\hat{\tau}_j} D_{\hat{\tau}_j}^T + D_{\hat{\tau}_j} D_{\hat{\tau}_j}^T \hat{\lambda}_{\hat{\tau}_j},
\]

and, for \( s_{\hat{\tau}_j} = -1 \),

\[
D_{\hat{\tau}_j} y \leq -\lambda_j D_{\hat{\tau}_j} D_{\hat{\tau}_j}^T + D_{\hat{\tau}_j} D_{\hat{\tau}_j}^T \hat{\lambda}_{\hat{\tau}_j}.
\]

Therefore, given \( D_{\hat{\tau}_j}, \lambda_j, \hat{\lambda}_j, \) and \( \sigma^2 \), the quantity \( Z_{j}^{glo} = \frac{D_{\hat{\tau}_j} y}{\sigma \|D_{\hat{\tau}_j}^T\|} \) is restricted to regions \((-\infty, \mathcal{V}_{j}^{-(glo)}\)\] and \([\mathcal{V}_{j}^{+(glo)}, \infty)\) where \( \mathcal{V}_{j}^{-(glo)} \) and \( \mathcal{V}_{j}^{+(glo)} \) are given in (41).

(b) In the same fashion as part (a), given \( D_{\hat{\tau}_j}, \lambda_j \), and \( \hat{\lambda}_j, \) the quantity \( T_{j}^{glo} = \frac{D_{\hat{\tau}_j} y}{\sigma_j^{glo} \|D_{\hat{\tau}_j}^T\|} \) is restricted to the regions \((-\infty, \mathcal{V}_{j}^{-(glo)}\)\] and \([\mathcal{V}_{j}^{+(glo)}, \infty)\) where \( \mathcal{V}_{j}^{-(glo)} \) and \( \mathcal{V}_{j}^{+(glo)} \) are given in (42).

### E Proof of Theorem 6

We will prove the inequality for the truncated \( t \) distribution. A similar proof can be applied to the truncated normal distribution which is also provided in Kivaranovic and Leeb (2020).

Suppose \( T \sim t(\mu, \sigma^2, d) \) is a location-scale \( t \) distribution with cumulative distribution function \( G_{\mu, \sigma^2, d}(\cdot) \). For any \( 0 < \gamma < 1 \), define \( Q_\gamma(t) \) via

\[
G_{Q_\gamma, \sigma^2, d}(t) = \Pr_{Q_\gamma, \sigma^2, d}(T \leq t) = \gamma,
\]

or equivalently, \( Q_\gamma(t) = t - \sigma G_{d}^{-1}(\gamma) \). Thus, the cumulative distribution function of \( T \) truncated to \( S = \bigcup_{i=1}^{m} (c_i, d_i) \), is upper bounded by

\[
G_{S, Q_\gamma, \sigma^2, d}(t) = \frac{\Pr_{Q_\gamma, \sigma^2, d}(T \leq t \cap T \in S)}{\Pr_{Q_\gamma, \sigma^2, d}(T \in S)} \leq \frac{\gamma}{g},
\]

(A.60)
where \( g = \inf_\mu \Pr_{\mu, \sigma^2, d}(T \in S) \). On the other hand, from the inequality \( \Pr(A \cap B) \geq \Pr(A) + \Pr(B) - 1 \), we notice that

\[
G_{Q, \sigma^2, d}(t) = \Pr_{Q, \sigma^2, d}(T \leq t) + \Pr_{\mu, \sigma^2, d}(T \in S) - 1
\]

\[
\geq \inf_\mu \Pr_{\mu, \sigma^2, d}(T \leq t) + \Pr_{\mu, \sigma^2, d}(T \in S) - 1
\]

\[
= \frac{\gamma + \overline{g} - 1}{\overline{g}}, \tag{A.61}
\]

where \( \overline{g} = \sup_\mu \Pr_{\mu, \sigma^2, d}(T \in S) \). Equations (36) and (A.60) imply that

\[
G_{Q, \sigma^2, d}(T) \leq \alpha = G_{U(T), \sigma^2, d}(T).
\]

Because \( G_{\mu, \sigma^2, d}(t) \) is a decreasing function with respect to \( \mu \), hence,

\[
U(T) \leq Q_{\overline{g}}(T) = T - \sigma G^{-1}_d \left( \frac{\alpha}{2} \overline{g} \right).
\]

Similarly, from (36) and (A.61), we obtain

\[
L(T) \geq Q_{1 - \frac{\alpha}{2}}(T) = T - \sigma G^{-1}_d \left( 1 - \frac{\alpha}{2} \overline{g} \right).
\]

Therefore, from the last two equations, we have

\[
U(T) - L(T) \leq \sigma \left[ G^{-1}_d \left( 1 - \frac{\alpha}{2} \overline{g} \right) - G^{-1}_d \left( \frac{\alpha}{2} \overline{g} \right) \right]
\]

\[
= \sigma \left[ G^{-1}_d \left( 1 - \frac{\alpha}{2} \overline{g} \right) + G^{-1}_d \left( 1 - \frac{\alpha}{2} \overline{g} \right) \right]
\]

\[
\leq 2 \sigma G^{-1}_d \left( 1 - \frac{\alpha}{2} \overline{g} \right), \tag{A.62}
\]

Note that the last equality in the preceding statement is obtained from \( 1 - \frac{\alpha}{2} \overline{g} \leq 1 - \frac{\alpha}{2} g \). Now, consider the minimization problem

\[
\inf_\mu \Pr_{\mu, \sigma^2, d}(T < d_1 \text{ or } T > c_m),
\]

which minimizes at \( \frac{d_1 + c_m}{2} \), with the minimum value \( g_0 = 2G_d \left( -\frac{c_m - d_1}{2\sigma} \right) \). Observe that \( g_0 \leq \overline{g} \), and since \( G^{-1} \) is an increasing function, then

\[
G^{-1}_d \left( 1 - \frac{\alpha}{2} \overline{g} \right) \leq G^{-1}_d \left( 1 - \frac{\alpha}{2} g_0 \right).
\]

Lastly, to obtain the final upper bound in (45), it suffices to show

\[
G^{-1}_d \left( 1 - \frac{\alpha}{2} g_0 \right) \leq G^{-1}_d \left( 1 - \frac{\alpha}{2} \right) + \frac{c_m - d_1}{2\sigma}, \tag{A.63}
\]

37
or equivalently,

\[ G_d^{-1} \left( \frac{\alpha}{2} g_0 \right) \geq G_d^{-1} \left( \frac{\alpha}{2} \right) - \frac{c_m - d_i}{2\sigma}, \]  

(A.64)

Given the condition (44), we note that,

\[ \alpha G_d \left( - \frac{c_m - d_i}{2\sigma} \right) \geq G_d \left( G_d^{-1} \left( \alpha/2 \right) - \frac{c_m - d_i}{2\sigma} \right), \quad \text{for } d \geq 3. \]  

(A.65)

Hence, by plugging in the value of \( g_0 \), we have

\[ \frac{\alpha}{2} g_0 = \alpha G_d \left( - \frac{c_m - d_i}{2\sigma} \right) \geq G_d \left( G_d^{-1} \left( \alpha/2 \right) - \frac{c_m - d_i}{2\sigma} \right). \]

Applying the increasing function \( G_d^{-1}(\cdot) \) to both sides of the above inequality leads to (A.64), and in turn (A.63). Now, Equations (A.62) and (A.63) together yield

\[ U(T) - L(T) \overset{a.s.}{\leq} 2 \sigma G_d^{-1} \left( 1 - \frac{\alpha}{2} \right) + (c_m - d_i). \]

Finally, the upper bound associated with the truncated \( t \) distribution in (45) will be achieved by replacing \( L(T), \sigma, d_i, c_m \) and \( d \) with \( L^{glo}_{Tj}, U^{glo}_{Tj}, \tilde{\sigma}^{(glo)}_j, \nu^{(glo)}_j, \nu^{(glo)}_j \) and \( d^{(glo)}_j \), respectively.

To explore the condition (A.65), define

\[ h(x) = \frac{\alpha G_d(-x)}{G_d \left( G_d^{-1}(\alpha/2) - x \right)} - 1, \quad \text{for } d \geq 3. \]

The function \( h(x) \) has two roots, one at zero and one at a positive value \( x_0 \). Moreover, \( h(x) > 0 \), for \( x \in (0, x_0) \). Notice that the positive root \( x_0 \) diverges as degrees of freedom \( d \) increases.

More specifically, for large \( d \), the truncated \( t \) distribution behaves similar to normal one, and hence \( h(x) \) becomes an increasing function for all \( x > 0 \). See Figure 6, panel (a). Therefore, for large, the statement in (A.65) always holds. For small to moderate values of \( d \), (A.65) holds under the condition \( \frac{c_m - d_i}{2\sigma} \in (0, x_0] \). We have visualized the behaviour of lengths of confidence intervals and their upper bounds in Figure 6, panel (b), for various values of \( d \).

Bibliography

Ali, A., Tibshirani, R. J., et al. (2019). The generalized lasso problem and uniqueness. *Electronic Journal of Statistics*, 13(2):2307–2347.

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

Andrews, I., Kitagawa, T., and McCloskey, A. (2019). Inference on winners. Technical report, National Bureau of Economic Research.

Bachoc, F., Preinerstorfer, D., Steinberger, L., et al. (2020). Uniformly valid confidence intervals post-model-selection. *The Annals of Statistics*, 48(1):440–463.

Bai, J. (1997). Estimating multiple breaks one at a time. *Econometric theory*, 13(3):315–352.
Figure 6: Panel (a) displays the function $h(x)$ which, for moderate values of $d$, is a unimodal function with two roots, zero and a positive value $x_0$. Also, $h(x) \geq 0$ for $x \in [0, x_0]$. Panel (b) shows the lengths of confidence intervals for a truncated normal distribution and a truncated $t$ distribution, under various values of $d$. The solid horizontal lines in panel (b) shows the upper bounds given in Theorem 6, for both distributions.

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

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

Benjamini, Y., Hechtlinger, Y., and Stark, P. B. (2019). Confidence intervals for selected parameters. *arXiv preprint arXiv:1906.00505*.

Berk, R., Brown, L., Buja, A., Zhang, K., Zhao, L., et al. (2013). Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837.

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Chen, S. S., Donoho, D. L., and Saunders, M. A. (2001). Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159.

Davies, P. L. and Kovac, A. (2001). Local extremes, runs, strings and multiresolution. *Annals of Statistics*, pages 1–48.

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*.
Eckley, I. A., Fearnhead, P., and Killick, R. (2011). Analysis of changepoint models. *Bayesian Time Series Models*, pages 205–224.

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

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

Fryzlewicz, P. (2018). Detecting possibly frequent change-points: Wild binary segmentation 2 and steepest-drop model selection. *arXiv preprint arXiv:1812.06880*.

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

Fryzlewicz, P. et al. (2018). Tail-greedy bottom-up data decompositions and fast multiple change-point detection. *The Annals of Statistics*, 46(6B):3390–3421.

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

Galceran, E., Cunningham, A. G., Eustice, R. M., and Olson, E. (2017). Multipolicy decision-making for autonomous driving via changepoint-based behavior prediction: Theory and experiment. *Autonomous Robots*, 41(6):1367–1382.

Hampel, F. R. (1974). The influence curve and its role in robust estimation. *Journal of the American Statistical Association*, 69(346):383–393.

Harchaoui, Z. and Lévy-Leduc, C. (2010). Multiple change-point estimation with a total variation penalty. *Journal of the American Statistical Association*, 105(492):1480–1493.

Horváth, L. and Rice, G. (2014). Extensions of some classical methods in change point analysis. *Test*, 23(2):219–255.

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

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

James, N. A. and Matteson, D. S. (2015). Change points via probabilistically pruned objectives. *arXiv preprint arXiv:1505.04302*.

Jewell, S., Fearnhead, P., and Witten, D. (2019). Testing for a change in mean after changepoint detection. *arXiv preprint arXiv:1910.04291*.
Kivaranovic, D. and Leeb, H. (2020). On the length of post-model-selection confidence intervals conditional on polyhedral constraints. *Journal of the American Statistical Association*, (just-accepted):1–33.

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

Leeb, H. and Pötscher, B. M. (2003). The finite-sample distribution of post-model-selection estimators and uniform versus nonuniform approximations. *Econometric Theory*, 19(1):100–142.

Leeb, H. and Pötscher, B. M. (2006). Can one estimate the conditional distribution of post-model-selection estimators? *The Annals of Statistics*, 34(5):2554–2591.

Leeb, H. and Pötscher, B. M. (2008). Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory*, 24(2):338–376.

Lenssen, N. J., Schmidt, G. A., Hansen, J. E., Menne, M. J., Persin, A., Ruedy, R., and Zyss, D. (2019). Improvements in the uncertainty model in the goddard institute for space studies surface temperature (gistemp) analysis. *Journal of Geophysical Research: Atmospheres*.

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

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

Lockhart, R., Taylor, J., Tibshirani, R. J., and Tibshirani, R. (2014). A significance test for the lasso. *Annals of Statistics*, 42(2):413.

Loftus, J. R. and Taylor, J. E. (2014). A significance test for forward stepwise model selection. *arXiv preprint arXiv:1405.3920*.

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

Mehrizi, R. and Chenouri, S. (2020). Detection of change points in piecewise polynomial signals using trend filtering. *Electronic Journal of Statistics*, 9(1):1150–1172.

Niu, Y. S. and Zhang, H. (2012). The screening and ranking algorithm to detect dna copy number variations. *The annals of applied statistics*, 6(3):1306.

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

Page, E. (1955). A test for a change in a parameter occurring at an unknown point. *Biometrika*, 42(3/4):523–527.

Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, 41(1/2):100–115.
Pezzatti, G. B., Zumbrunnen, T., Bürgi, M., Ambrosetti, P., and Conedera, M. (2013). Fire regime shifts as a consequence of fire policy and socio-economic development: an analysis based on the change point approach. *Forest Policy and Economics*, 29:7–18.

Pötscher, B. M. (1991). Effects of model selection on inference. *Econometric Theory*, 7(2):163–185.

Qian, J. and Jia, J. (2016). On stepwise pattern recovery of the fused lasso. *Computational Statistics & Data Analysis*, 94:221–237.

Rinaldo, A. et al. (2009). Properties and refinements of the fused lasso. *The Annals of Statistics*, 37(5B):2922–2952.

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

Tian, X. and Taylor, J. (2017). Asymptotics of selective inference. *Scandinavian Journal of Statistics*, 44(2):480–499.

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

Tibshirani, R. J. (2011). *The solution path of the generalized lasso*. Stanford University.

Tibshirani, R. J. et al. (2014). Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics*, 42(1):285–323.

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

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

Vostrikova, L. Y. (1981). Detecting ”disorder” in multidimensional random processes. In *Doklady Akademii Nauk*, volume 259, pages 270–274. Russian Academy of Sciences.

Yao, Y.-C. and Au, S.-T. (1989). Least-squares estimation of a step function. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 370–381.

Zrnic, T. and Jordan, M. I. (2020). Post-selection inference via algorithmic stability. *arXiv preprint arXiv:2011.09462*.