More Powerful Selective Inference for the Graph Fused Lasso

Yiqun Chen\textsuperscript{a}, Sean Jewell\textsuperscript{b,\dagger}, and Daniela Witten\textsuperscript{a,b}

\textsuperscript{a}Department of Biostatistics, University of Washington, Seattle, WA; \textsuperscript{b}Department of Statistics, University of Washington, Seattle, WA

\textbf{ABSTRACT}

The graph fused lasso—which includes as a special case the one-dimensional fused lasso—is widely used to reconstruct signals that are piecewise constant on a graph, meaning that nodes connected by an edge tend to have identical values. We consider testing for a difference in the means of two connected components estimated using the graph fused lasso. A naive procedure such as a z-test for a difference in means will not control the selective Type I error, since the hypothesis that we are testing is itself a function of the data. In this work, we propose a new test for this task that controls the selective Type I error, and conditions on less information than existing approaches, leading to substantially higher power. We illustrate our approach in simulation and on datasets of drug overdose death rates and teenage birth rates in the contiguous United States. Our approach yields more discoveries on both datasets. Supplementary materials for this article are available online.

\section{1. Introduction}

We consider a vector $Y \in \mathbb{R}^n$, assumed to be a noisy realization of a signal $\beta \in \mathbb{R}^n$,

$$Y_j = \beta_j + \epsilon_j, \quad \epsilon_j \sim_{iid} N(0, \sigma^2), \quad j = 1, \ldots, n,$$

(1)

with known variance $\sigma^2$. We assume that $\beta$ has some underlying structure of interest. For instance, $\beta$ might be \textit{sparse}, with few nonzero elements, or \textit{piecewise constant}, meaning that the elements of $\beta$ are ordered, and adjacent elements tend to take on equal values.

It is natural to estimate $\beta$ by solving the optimization problem

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^n} \left\{ \frac{1}{2} \|y - \beta\|_2^2 + \lambda \|D\beta\|_1 \right\},$$

(2)

where $D$ is an $m \times n$ penalty matrix that encodes the structure of $\beta$. Problem (2) is a special case of the generalized lasso

$$\min_{\beta \in \mathbb{R}^n} \left\{ \frac{1}{2} \|y - \beta\|_2^2 + \lambda \sum_{(j,f) \in E} |\beta_j - \beta_f| \right\},$$

(3)

where $G = (V,E)$ is an undirected graph, $V = \{1, \ldots, n\}$, and $(j, f) \in E$ indicates that the $j$th and $f$th vertices in the graph are connected by an edge (Tibshirani and Taylor 2011). For sufficiently large values of the nonnegative tuning parameter $\lambda$, we will have $\hat{\beta}_j = \hat{\beta}_f$ for some $(j, f) \in E$. We can segment $\hat{\beta}$ into \textit{connected components}—that is, sets of elements of $\hat{\beta}$ that are connected in the original graph and share a common value. We might then consider testing the null hypothesis that the true mean of $\beta$ is the same across two \textit{estimated} connected components, that is,

$$H_0 : \sum_{j \in \hat{C}_1} \beta_j / |\hat{C}_1| = \sum_{f \in \hat{C}_2} \beta_f / |\hat{C}_2| \quad \text{versus} \quad H_1 : \sum_{j \in \hat{C}_1} \beta_j / |\hat{C}_1| \neq \sum_{f \in \hat{C}_2} \beta_f / |\hat{C}_2|,$$

(4)

where $\hat{C}_1 \subseteq V$ and $\hat{C}_2 \subseteq V$ are connected components of $\hat{\beta}$, with cardinality $|\hat{C}_1|$ and $|\hat{C}_2|$, and $\hat{C}_1 \cap \hat{C}_2 = \emptyset$. This is

\textbf{ARTICLE HISTORY}

Received September 2021
Accepted June 2022

\textbf{KEYWORDS}

Change point detection; Hypothesis testing; Penalized regression; Piecewise constant

\textbf{CONTACT}

Yiqun Chen (yiqunc@uw.edu) \textsuperscript{a} Department of Biostatistics, University of Washington, Seattle, WA.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/riJCGS.

© 2022 American Statistical Association and Institute of Mathematical Statistics

\textbf{ARTICLE}{  \textsuperscript{\textcopyright}  2022 American Statistical Association and Institute of Mathematical Statistics
equivalent to testing \( H_0 : v^\top \beta = 0 \) versus \( H_1 : v^\top \beta \neq 0 \), where

\[ v_j = 1_{(j \notin \hat{C}_1)} / |\hat{C}_1| - 1_{(j \notin \hat{C}_2)} / |\hat{C}_2|, \quad j = 1, \ldots, n. \quad (5) \]

Here, \( H_0 \) is chosen based on the data, that is, we selected the contrast vector \( v \) in (5) because \( \hat{C}_1 \) and \( \hat{C}_2 \) are estimated connected components. We focus on developing a test of \( H_0 \) that controls the selective Type I error rate (Fithian, Sun, and Taylor 2014), that is, one for which the probability of rejecting \( H_0 \) at level \( \alpha \), given that \( H_0 \) holds and we decided to test \( H_0 \), is no greater than \( \alpha \):

\[ P_{H_0}(\text{reject } H_0 \text{ at level } \alpha \mid H_0 \text{ is tested}) \leq \alpha, \forall \alpha \in (0, 1). \quad (6) \]

It is not hard to see that a standard two-sample z-test of \( H_0 : v^\top \beta = 0 \), with p-value \( P_{H_0}(|v^\top Y| \geq |v^\top y|) \), fails to account for the fact that we decided to test \( H_0 \) after looking at the data, and therefore does not control the selective Type I error rate (6).

To address this problem, Hyun, G’Sell, and Tibshirani (2018) propose an elegant approach for testing \( H_0 : v^\top \beta = 0 \) that makes use of the selective inference framework developed by Lee et al. (2016), Fithian, Sun, and Taylor (2014), and Tibshirani et al. (2016). Their key insight is as follows: the set of \( v \) for which the inequality constraint in (7) is tight.

### 2. Background on the Generalized Lasso

In this section, we review the selective inference framework of Hyun, G’Sell, and Tibshirani (2018) for testing hypotheses based upon the generalized lasso estimator (2), which includes the graph fused lasso as a special case. Their framework relies on the dual path algorithm of Tibshirani and Taylor (2011) for solving (2). Thus, we begin with a very brief overview of that algorithm.

#### 2.1. The Dual Problem, and the Dual Path Algorithm

Tibshirani and Taylor (2011) develop an efficient path algorithm for solving the dual problem for (2), which takes the form

\[ \hat{u}(\lambda) = \arg\min_{u \in \mathbb{R}^m} \| y - D^\top u \|_2^2 \] subject to \[ \| u \|_\infty \leq \lambda, \] (7)

and is related to (2) through the identity \( \hat{\beta}(\lambda) = y - D^\top \hat{u}(\lambda) \), where the notation \( \hat{\beta}(\lambda) \) and \( \hat{u}(\lambda) \) makes explicit that \( \hat{\beta} \) and \( \hat{u} \) are functions of \( \lambda \). This dual path algorithm is detailed in Appendix A.1, supplementary materials. While the details of the algorithm are not important for the current paper, we briefly summarize the main idea. The algorithm begins with \( \lambda = \infty \), and then proceeds through a series of steps, corresponding to decreasing values of \( \lambda \). The kth step involves computing a boundary set \( B_k \subseteq \{m\} \), which consists of the subset of indices of the vector \( u \) for which the inequality constraint in (7) is tight. The signs of the elements of \( u \) associated with this boundary set, \( s_{B_k} \), are also computed. These quantities satisfy

\[ \hat{\beta}(\lambda) = \mathbb{P}_{\text{null}(D_{-B_k})}(y - \lambda \cdot D_{B_k}^\top s_{B_k}) \] (8)

for a range of \( \lambda \) values corresponding to the kth step (Tibshirani and Taylor 2011). In (8), \( D_{B_k} \) and \( D_{-B_k} \) correspond to the submatrices of \( D \) with rows in \( B_k \) and not in \( B_k \), respectively, and \( \mathbb{P}_{\text{null}(D_{-B_k})} \) is the projection matrix onto the null space of \( D_{-B_k} \). To summarize, (8) indicates that \( \hat{\beta}(\lambda) \) can be computed from \( (B_k, s_{B_k}) \), for an appropriate range of \( \lambda \) values.

The next proposition considers the special case of the graph fused lasso problem (3).

**Proposition 1.** Let \( B_k \) denote the boundary set that results from the kth step of the dual path algorithm for (3), and let \( \hat{\beta} \) denote
the solution to (3). Let $G_{\sim B_k}$ denote the subgraph of $G$ with edges in the boundary set $B_k$ removed, and let $C_1, \ldots, C_L$ denote the $L$ connected components of $G_{\sim B_k}$. Then, under (1), with probability $1$, $\hat{\beta}_j = \tilde{\beta}_j$ if and only if $j, j' \in C_l$ for some $l \in [L]$.

We close with a brief summary of the main points in this section:

- For the generalized lasso (2), $\hat{\beta}$ can be computed from $\lambda$ and $(B_k, B_{\sim B_k})$ from the dual path algorithm.
- In the special case of the graph fused lasso (3), the connected components of $\hat{\beta}$ can be computed from $B_k$ from the dual path algorithm.

2.2. Existing Work on Selective Inference for the Generalized Lasso

The main idea behind selective inference is as follows: when testing a null hypothesis that is a function of the data, to control the selective Type I error in the sense of (6), we must condition on the information used to construct that null hypothesis (Fithian, Sun, and Taylor 2014; Tibshirani et al. 2016; Lee et al. 2016). In particular, to test a null hypothesis of the form $H_0 : \nu^T \beta = 0$ where $\nu$ is a function of the data, we must condition on the information used to construct $\nu$.

In a recent elegant line of work, a number of authors have shown that the model selection events of several well-known model selection procedures, including the lasso (Lee et al. 2016), stepwise regression (Loftus and Taylor 2014; Tibshirani et al. 2014; Lee et al. 2016), and marginal screening (Reid, Taylor, and Tibshirani 2017), can be written as polyhedral constraints on $Y$. More precisely, conditioning on the selected model (and in some cases, additional information) is equivalent to conditioning on a polyhedral set $\{Y : AY \leq b\}$, where the matrix $A$ and the vector $b$ can be explicitly computed. Thus, we can test null hypotheses that are a function of the selected model by considering the null distribution of $Y$ truncated to a polyhedral set.

Recently, Hyun, G’Sell, and Tibshirani (2018) extended this line of work to develop an approach for selective inference for the generalized lasso (2). Their key insight is as follows:

The set of $Y$ that leads to a specified output for the first $K$ steps of the dual path algorithm for (2) is a polyhedron, i.e., $\{Y : AY \leq 0\}$, for a matrix $A$ that can be explicitly computed.

Proposition 2 details this result.

**Proposition 2** (Proposition 3.1 in Hyun, G’Sell, and Tibshirani (2018)). Consider solving (2) using the dual path algorithm for $y \in \mathbb{R}^n$. For the $k$th step, $k = 1, \ldots, K$, define

$$M_k(y) \equiv \{B_k(y), s_{B_k}(y), R_k(y), L_k(y)\}$$

where the boundary set $B_k(y)$ and the sign vector of the boundary set $s_{B_k}(y)$ are defined in Algorithm 1 (see Appendix A.1, supplementary materials), and $L_{k+1}(y) = \{i : i \in B_k(y), c_i < 0, d_i < 0\}$, for $a_i, c_i$, and $d_i$ specified in Algorithm 1.

Then the set $\{Y \in \mathbb{R}^n : \bigcap_{k=1}^K \{M_k(Y) = M_k(y)\}\}$ is of the form $\{Y : AY \leq 0\}$ for some matrix $A$ that can be constructed explicitly based on $M_1(y), \ldots, M_K(y)$.

Motivated by this result, Hyun, G’Sell, and Tibshirani (2018) proposed to test $H_0 : \nu^T \beta = 0$, where $\nu$ is a function of the generalized lasso estimator, via a $p$-value of the form

$$P_{\text{Hyun}} \equiv \mathbb{P}_{H_0}(\{\nu^T Y \geq |\nu^T y| \cap \bigcap_{k=1}^K \{M_k(Y) = M_k(y)\})$$

$$\Pi_\nu^+ Y = \Pi_\nu^+ y.$$

In (10), conditioning on $\Pi^+ \nu$ eliminates the nuisance parameter $\Pi^+ \nu \beta$; see sec. 3.1 of Fithian, Sun, and Taylor (2014). Now, under (1), the conditional distribution of $\nu^T Y$ is normal with mean zero and variance $\sigma^2||\nu||_2^2$, truncated to a set that can be characterized and efficiently computed using Proposition 2. This yields the $p$-value in (10). Furthermore, unlike the $z$-test based on the naive $p$-value $\mathbb{P}_{H_0}(\{\nu^T Y \geq |\nu^T y|\})$, a test that rejects $H_0$ when the $p$-value in (10) is less than some level $\alpha$ controls the selective Type I error rate, in the sense of (6).

We emphasize that the $p$-value in (10) conditions on the event $\bigcap_{k=1}^K \{M_k(Y) = M_k(y)\}$; that is, on all of the outputs of the first $K$ steps of the dual path algorithm (rather than simply the $k$th step). However, the contrast vector $\nu$ in $H_0 : \nu^T \beta = 0$ is constructed using only (at most) the output of the $k$th step in the dual path algorithm. In what follows, we will consider conditioning on much less information than (10). This will result in a test that controls the selective Type I error as in (6), and that has substantially higher power under the alternative.

3. Proposed Approach

3.1. What Should We Condition On?

To control the selective Type I error in (6), we must condition on the aspect of the data that led us to test the specific null hypothesis $H_0 : \nu^T \beta = 0$ (Fithian, Sun, and Taylor 2014; Hyun, G’Sell, and Tibshirani 2018).

If a data analyst wishes to choose the contrast vector $\nu$ in the null hypothesis $H_0 : \nu^T \beta = 0$ by inspecting the elements of $\hat{\beta}$ resulting from the $k$th step of the dual algorithm of the generalized lasso problem (2), then there is no reason to condition on $\bigcap_{k=1}^K \{M_k(Y) = M_k(y)\}$ (as was done by Hyun, G’Sell, and Tibshirani 2018), since the outputs of the first $K - 1$ steps of the dual path algorithm are not considered in constructing $\nu$. In fact, according to (8), which states that $B_K$ and $s_{B_K}$ uniquely determine $\hat{\beta}$, the data analyst need only condition on $B_K$ and $s_{B_K}$, rather than on $M_K = (B_K, s_{B_K}, R_K, L_K)$.

Furthermore, the data analyst might construct the contrast vector $\nu$ in $H_0 : \nu^T \beta = 0$ to take on a constant value within each connected component of $\hat{\beta}$, as in (4) and (5). Recall from Proposition 1 that the connected components of $\hat{\beta}$ are equivalent to the connected components of the subgraph $G_{\sim B_K}$. Therefore, it suffices to condition only on the connected components of the subgraph $G_{\sim B_K}$, or even on just the pair of connected components under investigation in (4).

What is the disadvantage of conditioning on $\bigcap_{k=1}^K \{M_k(Y) = M_k(y)\}$, as in Hyun, G’Sell, and Tibshirani...
(2018)? Conditioning on too much information leads to a loss of power (Fithian, Sun, and Taylor 2014; Lee et al. 2016; Liu, Markovic, and Tibshirani 2018; Jewell, Fearnhead, and Witten 2022). We wish to condition on less information to achieve higher power than Hyun, G’Sell, and Tibshirani (2018), while controlling the selective Type I error (6). Of course, this could result in computational challenges, as the conditioning sets described in the use cases above are not polyhedral, so the conditional distribution of $v^\top Y$ is no longer a normal truncated to an easily characterized set.

In what follows, we focus on the case where the data analyst constructs the contrast vector $v$ to take on a constant value in each connected component of $\hat{\beta}$, and consider a $p$-value of the form

$$ p_{\hat{C}_1, \hat{C}_2} = \mathbb{P}_{H_0}(v^\top Y \geq |v^\top y| | \hat{C}_1(y), \hat{C}_2(y) \in CC_K(Y), \Pi_{v^\top Y} = \Pi_{v^\top y}). $$

(11)

In (11), $\hat{C}_1(y)$ and $\hat{C}_2(y)$ are two connected components estimated from the data realization $y$ and used to construct the contrast vector $v$ in (5), and $CC_K(Y)$ is the set of connected components obtained from applying $K$ steps of the dual path algorithm for (3) to the random variable $Y$. Roughly speaking, this $p$-value answers the following question:

**Assuming that there is no difference between the population means of $\hat{C}_1$ and $\hat{C}_2$, then what’s the probability of observing such a large difference in the sample means of $\hat{C}_1$ and $\hat{C}_2$, given that these two connected components were estimated from the data?**

While our proposed $p$-value $p_{\hat{C}_1, \hat{C}_2}$ conditions on far less information than Hyun, G’Sell, and Tibshirani (2018), the recent proposal of Le Duy and Takeuchi (2021) takes an intermediate approach. They condition on the full boundary set $B_K$ at the $K$th step of the dual path algorithm (and thus, implicitly, on all of the connected components in $CC_K(Y)$), whereas we condition only on the two connected components of interest. See Appendix A.13, supplementary materials for further discussion and comparison.

### 3.2. Illustrative Example

We now demonstrate that conditioning on less information leads to increased power, using an example involving the graph fused lasso applied to a two-dimensional grid graph.

To begin, we constructed a graph composed of 64 nodes arranged in an $8 \times 8$ grid, such that each node is connected to its four closest (up, down, left, right) neighbors. We generated data on this grid according to (1), where $\beta$ has three piecewise constant segments, $C_1$, $C_2$, and $C_3$, with means of 3, 0, and $-3$, respectively. The true values of $\beta$, as well as the data generated from this model, are shown in Figures 1(a)–(b). On this particular dataset, $K = 13$ steps of the dual path algorithm for the graph fused lasso recovered the true connected components exactly.

For each pair of connected components, we then constructed a contrast vector $v$ as in (4), so that $H_0 : v^\top \beta = 0$ posits that the two components being tested have the same mean. We tested $H_0$ using the $p$-values $p_{\text{Hyun}}$ and $p_{\hat{C}_1, \hat{C}_2}$ given in (10) and (11), respectively. The $p$-values for all pairs of connected components

![Figure 1](image_url)

**Figure 1.** (a): We generated $\beta$ on an $8 \times 8$ grid. There are three true connected components, which take on values of $-3, 0,$ and $3$. (b): A noisy realization from the model $Y \sim N(\beta, I_{64})$. In this particular example, running 13 steps of the dual path algorithm for the graph fused lasso results in perfect recovery of the true connected components of $\beta$ (displayed in gray). (c): For each pair of estimated connected components, we tested the null hypothesis of equality in means using $p_{\text{Hyun}}$ in (10) and $p_{\hat{C}_1, \hat{C}_2}$ in (11). (d): The conditional null distributions of $v^\top Y$, where $v$ is chosen to test for a difference in means between $C_1$ and $C_2$, conditional on the conditioning sets in the definitions of $p_{\text{Hyun}}$ in (10) and $p_{\hat{C}_1, \hat{C}_2}$ in (11). In (d), the test statistic $|v^\top y| = 3.36$ is displayed as a dashed black line; this value is quite large relative to the null distribution of $p_{\hat{C}_1, \hat{C}_2}$, but modest relative to that of $p_{\text{Hyun}}$.**
are displayed in Figure 1(c). Because \( p_{\text{Hyun}} \) conditions on unnecessary information, the test based on \( p_{\text{Hyun}} \) has extremely low power and it cannot reject any \( H_0 \). By contrast, the test based on \( p_{\hat{C}_1, \hat{C}_2} \) has higher power. In Figure 1(d), we display the null distribution of \( v^\top Y \), conditional on the conditioning sets in (10) and (11).

### 3.3. Properties of \( p_{\hat{C}_1, \hat{C}_2} \)

The following result establishes key properties of \( p_{\hat{C}_1, \hat{C}_2} \) in (11).

**Proposition 3.** Suppose that \( Y \sim \mathcal{N}(\beta, \sigma^2 I_\nu) \). Define

\[
y'(\phi) = \Pi_+^\top y + \phi \cdot \frac{v}{||v||_2} = y + \left( \frac{\phi - v^\top y}{||v||_2^2} \right) v.
\]

(12)

Let \( \phi \sim \mathcal{N}(0, \sigma^2 ||v||_2^2) \). Then, under \( H_0 : v^\top \beta = 0 \),

\[
p_{\hat{C}_1, \hat{C}_2} = \mathbb{P} \left( |\phi| \geq |v^\top y|, \hat{C}_1(y), \hat{C}_2(y) \in CC_K(y'(\phi)) \right). \quad (13)
\]

Moreover, the test that rejects \( H_0 \) if \( p_{\hat{C}_1, \hat{C}_2} \leq \alpha \) controls the selective Type I error at level \( \alpha \).

Therefore, to compute the \( p \)-value in (11), it suffices to characterize the set

\[
S_{\hat{C}_1, \hat{C}_2} = \left\{ \phi \in \mathbb{R} : \hat{C}_1(y), \hat{C}_2(y) \in CC_K(y'(\phi)) \right\}. \quad (14)
\]

We can think of \( y'(\phi) \) in (12) as a perturbation of the data by a function of \( \phi \) along the direction defined by \( v \). Figure 2 illustrates this intuition in the toy example from Figure 1, in the context of a test for the difference in the means of \( C_1 \) and \( \hat{C}_2 \) (see Figure 2(a)). Panel (a) displays the observed data, for which \( v^\top y = 3.36 \), where \( v \) is defined in (5). In panel (b), we perturb the observed data to \( \phi = 0 \). Now the graph fused lasso with \( K = 13 \) no longer detects the three connected components. In panel (c), we perturb the observed data to \( \phi = -5 \); in this case, the graph fused lasso with \( K = 13 \) estimates all three connected components. Therefore, \( \phi = 3.36 \) and \( -5 \) are in the set \( \{ \phi \in \mathbb{R} : \hat{C}_1, \hat{C}_2 \in CC_3(y'(\phi)) \} \), but \( \phi = 0 \) is not. Panel (d) displays \( \{ \phi \in \mathbb{R} : \hat{C}_1, \hat{C}_2 \in CC_3(y'(\phi)) \} = (-\infty, -1.71) \cup (1.69, 12.3) \cup (36.3, \infty) \).

We now leverage ideas from Jewell, Fearnhead, and Witten (2022) to develop an efficient approach to compute the set (14). First, we characterize the set \( S_{\hat{C}_1, \hat{C}_2} \) in (14) in Proposition 4. Recall that \( M_k(y) = (B_k(y), s_B(y), R_k(y), L_k(y)) \) is the output of the \( k \)th step of Algorithm 1, supplementary materials. We first present a corollary of Proposition 2.

**Corollary 1.** The set \( \{ \phi \in \mathbb{R} : \bigcap_{k=1}^K \{ M_k(y'(\phi)) = M_k(y) \} \} \) is an interval.

**Proposition 4.** Let \( \mathcal{I} \) be the set of possible outputs of Algorithm 1, supplementary materials that yield \( \hat{C}_1 \) and \( \hat{C}_2 \) and can be obtained via a perturbation of \( y \) defined in (12), that is,

\[
\mathcal{I} \equiv \left\{ (m_1, \ldots, m_K) : \exists \alpha \in \mathbb{R} \text{ such that} \right\}
\]

\[
\hat{C}_1(y), \hat{C}_2(y) \in CC_K(y'(\alpha)), \bigcap_{k=1}^K \{ M_k(y'(\alpha)) = m_k \}. \quad (15)
\]

Then, there exists an index set \( \mathcal{J} \) and scalars \( a_{-1} < a_0 < a_1 < a_2 < \cdots \) such that
1. the set $\mathcal{S}_{\hat{C}_1, \hat{C}_2}$ in (14) is the union of $|\mathcal{J}|$ intervals:

$$
\mathcal{S}_{\hat{C}_1, \hat{C}_2} = \left\{ \phi \in \mathbb{R} : \hat{C}_1(y), \hat{C}_2(y) \in CC_K(y') \right\} = \bigcup_{i \in \mathcal{I}} \{a_i, a_{i+1}\},
$$

(16)

2. $|\mathcal{I}| = |\mathcal{J}|$ (i.e., the sets $\mathcal{I}$ and $\mathcal{J}$ have the same cardinality); and

3. $\forall i \in \mathcal{J}$, $\exists (m_1, \ldots, m_K) \in \mathcal{I}$ such that $\{a_i, a_{i+1}\} = \left\{ \phi \in \mathbb{R} : \bigcap_{k=1}^{K} \{M_k(y'(\phi)) = (m_1, \ldots, m_K)\} \right\}$.

In words, Proposition 4 states that the set $\mathcal{S}_{\hat{C}_1, \hat{C}_2}$ in (14) can be expressed as a union of intervals, each of which can be computed by applying Corollary 1 on a perturbation of $y$. Next, we use Proposition 4 to develop an efficient recipe to compute $\mathcal{S}_{\hat{C}_1, \hat{C}_2}$ by constructing the index set $\mathcal{J}$ and scalars $\cdots < a_{-2} < a_{-1} < a_0 < a_1 < a_2 < \cdots$. To begin, we run the first $K$ steps of the dual path algorithm on the data $y$. We then apply Corollary 1 to obtain the set $[a_0, a_1] = \left\{ \phi \in \mathbb{R} : \bigcap_{k=1}^{K} \{M_k(y'(\phi)) = M_k(y)\} \right\}$. By construction, $[a_0, a_1] \subseteq \mathcal{S}_{\hat{C}_1, \hat{C}_2}$, because $\hat{C}_1$ and $\hat{C}_2$ are connected components estimated from the data $y$. Therefore, we initialize the index set $\mathcal{J}$ as $\{0\}$. Then, for a small $\eta > 0$, we apply Corollary 1 to obtain the interval $\left\{ \phi \in \mathbb{R} : \bigcap_{k=1}^{K} \{M_k(y'(\phi)) = M_k(y'(a_1 + \eta))\} \right\}$. If the left endpoint of this interval does not equal $a_1$, then we must repeat with a smaller value of $\eta$ until we obtain an interval of the form $[a_1, a_2]$. We can then check whether $\hat{C}_1, \hat{C}_2 \in CC_K(y'(a_1 + \eta))$: if so, then $[a_1, a_2] \subseteq \mathcal{S}_{\hat{C}_1, \hat{C}_2}$ and we update $\mathcal{J}$ to include $[1]$. Otherwise, $\mathcal{J}$ remains unchanged. We continue in this vein, along the positive real line, until we reach an interval for which the right endpoint equals $\infty$.

Finally, we proceed along the negative real line: we apply Corollary 1 to compute the interval $[a_{-1}, a_0] = \left\{ \phi \in \mathbb{R} : \bigcap_{k=1}^{K} \{M_k(y'(\phi)) = M_k(y'(a_0 - \eta))\} \right\}$. If $\hat{C}_1, \hat{C}_2 \in CC_K(y'(a_0 - \eta))$, then $\mathcal{J}$ is set to $\mathcal{J} \cup \{-1\}$; otherwise, $\mathcal{J}$ remains unchanged. We iterate until the algorithm outputs an interval for which the left endpoint equals $-\infty$. Finally, $\mathcal{S}_{\hat{C}_1, \hat{C}_2} = \bigcup_{i \in \mathcal{J}} \{a_i, a_{i+1}\}$. The procedure is summarized in Algorithm 2 of Appendix A.3, supplementary materials.

In our implementation, we initialize with $\eta = 10^{-4}$, which proves to be an efficient choice in experiments in Section 5 (see details in Appendix A.3, supplementary materials). In principle, the running time of Algorithm 2 can be quite slow, and potentially even exponential in $K$. However, in practice, the runtime of Algorithm 2 is nowhere near the worst-case upper bound (see Appendix A.11, supplementary materials for a detailed empirical study of the timing complexity of Algorithm 2). In addition, in Proposition 5, we describe an “early stopping” rule that guarantees a conservative $p$-value and only requires running Algorithm 2 until we reach intervals containing $|u \uparrow y| + \delta$ and $-|u \downarrow y| - \delta$ for some $\delta > 0$, as opposed to $\infty$ and $-\infty$. Then, the set is appended with $(-\infty, -|u \downarrow y| - \delta]$ and $[|u \uparrow y| + \delta, \infty)$. This “early stopping” rule also applies to the extensions in Section 4. In practice, we suggest using $\delta = \max\{0, 10\sigma ||u||_2 - |u \downarrow y|\}$ to balance the inferential accuracy and computational efficiency (see details in Appendix A.6, supplementary materials).

**Proposition 5.** Provided that $\mathbb{P}(\phi \in S_{\hat{C}_1, \hat{C}_2}) > 0$, for any $\delta > 0$, we have that

$$
\mathbb{P}\left( |\phi| \geq |u \uparrow y| \right| \phi \in S_{\hat{C}_1, \hat{C}_2}) \geq \mathbb{P}\left( |\phi| \geq |u \uparrow y| \right| \phi \in S_{\hat{C}_1, \hat{C}_2}) \geq (\infty, -|u \downarrow y| - \delta] \cup S_{\hat{C}_1, \hat{C}_2} \cup [|u \uparrow y| + \delta, \infty).
$$

4. Extensions

4.1. Confidence Intervals for $u \uparrow \beta$

We now construct a $(1 - \alpha)$ confidence interval for $u \uparrow \beta$, the difference between the population means of two connected components $\hat{C}_1$ and $\hat{C}_2$ resulting from the graph fused lasso.

**Proposition 6.** Suppose that (1) holds, and let $\hat{C}_1$ and $\hat{C}_2$ be two connected components obtained from performing $K$ steps of the dual path algorithm for the graph fused lasso (3). For a given value of $\alpha \in (0, 1)$, define functions $\theta(t)$ and $\theta_u(t)$ such that

$$
F^{\mathcal{S}_{\hat{C}_1, \hat{C}_2}}(t) = \frac{1 - \alpha}{2}, \quad F^{\mathcal{S}_{\hat{C}_1, \hat{C}_2}}(t) = \frac{\alpha}{2},
$$

(18)

where $F^{\mathcal{S}_{\hat{C}_1, \hat{C}_2}}(t)$ is the cumulative distribution function of a $N(\mu, \sigma^2)$ random variable, truncated to the set $\mathcal{S}_{\hat{C}_1, \hat{C}_2}$ defined in (14). Then $[\theta(t \uparrow u), \theta_u(t \uparrow u)]$ has $(1 - \alpha)$ selective coverage (Fithian, Sun, and Taylor 2014; Lee et al. 2016; Tibshirani et al. 2016) for $u \uparrow \beta$, in the sense that

$$
\mathbb{P}\left( \theta(t \uparrow u), \theta_u(t \uparrow u) \mid \hat{C}_1, \hat{C}_2 \in CC_K(Y) \right) = 1 - \alpha.
$$

(19)

Computing $\theta(t)$ and $\theta_u(t)$ in (18) amounts to a root-finding problem, which can be solved using bisection (Chen and Bien 2020). A similar result is used to construct confidence intervals corresponding to $p_{\text{Hyun}}$ in Hyun, G’Sell, and Tibshirani (2018).

4.2. An Alternative Conditioning Set

The conditioning set for $p_{\hat{C}_1, \hat{C}_2}$ involves the connected components of the graph fused lasso solution after $K$ steps of the dual path algorithm. However, in practice, a data analyst might prefer a more “user-facing” choice of $K$, such as the value that yields $L$ connected components in the solution $\beta$.

For this reason, we now consider a slight modification of $p_{\hat{C}_1, \hat{C}_2}$.

$$
p_{\hat{C}_1, \hat{C}_2}^* = \mathbb{P}_{H_0}\left( |\phi \uparrow y| \geq |u \uparrow y| \right| \hat{C}_1(y), \hat{C}_2(y) \in CC(Y),
$$

$$
\Pi_{Y \uparrow} Y = \Pi_{Y \downarrow} Y),
$$

(20)

where the subscript $K$ on $CC$ has been dropped, indicating that the number of steps of the graph fused lasso algorithm is no longer fixed; instead, the function $CC$ now represents the graph fused lasso estimator tuned to yield exactly $L$ connected components. Thus, in $p_{\hat{C}_1, \hat{C}_2}^*$, we condition on datasets
for which $\hat{C}_1(y), \hat{C}_2(y)$ are among $L$ connected components estimated using the graph fused lasso. It is not hard to show that Proposition 4 and Algorithm 2, supplementary materials require only minor modifications to enable the computation of the p-values $p_{\hat{C}_1, \hat{C}_2}$; details are provided in Section A.8 of the Appendix, supplementary materials.

5. Simulation Study

We consider testing the null hypothesis $H_0 : \nu^T \beta = 0$ versus $H_1 : \nu^T \beta \neq 0$, where, unless otherwise stated, $\nu$ is defined in (5) for a randomly-chosen pair of estimated connected components $\hat{C}_1, \hat{C}_2$ of the solution to (3). We consider three p-values: $p_{\text{Hyun}}$ in (10), $p_{\hat{C}_1, \hat{C}_2}$ in (11), and the naive p-value

$$p_{\text{Naive}} \equiv \mathbb{P}_{H_0}(|\nu^T Y| \geq |\nu^T y|),$$

and compare the selective Type I error (6) and power of the tests that reject $H_0$ when these p-values are less than $\alpha = 0.05$.

In the simulations that follow, comparing the power of the tests requires a bit of care. Because the null hypothesis $H_0 : \nu^T \beta = 0$ involves the contrast vector $\nu$, which is a function of the data, the effect size $|\nu^T \beta|$ may differ across simulated datasets from the same data-generating distribution. Therefore, in what follows, we consider the power as a function of $|\nu^T \beta|$. Alternatively, we can separately assess the detection probability (i.e., the probability that $\hat{C}_1$ and $\hat{C}_2$ are true piecewise constant segments) and the “conditional power” (Gao, Bien, and Witten 2020; Hyun et al. 2021) (i.e., the probability of rejecting $H_0$, given that $\hat{C}_1, \hat{C}_2$ are true piecewise constant segments). Details are in Appendix A.9, supplementary materials.

5.1. One-Dimensional Fused Lasso

We first consider the special case of the graph fused lasso on a chain graph, in which the observations are ordered, and there is an edge between each pair of adjacent observations. This leads to the one-dimensional fused lasso problem (Tibshirani and Taylor 2011). We simulated from the “middle mutation” model of Hyun et al. (2021), where the signal contains two true changepoints of size $\delta$, and in turn, three connected components:

$$Y_j \sim \mathcal{N}(\beta_j, \sigma^2), \quad \beta_j = \delta \times 1_{(101 \leq j \leq 140)}, \quad j = 1, \ldots, 200.$$  

Figure 3(a) displays an example of this synthetic data with $\delta = 3$ and $\sigma = 1$.

5.1.1. Selective Type I Error Control under the Global Null

We simulated $y_1, \ldots, y_{200}$ according to (22) with $\delta = 0$ and $\sigma = 1$. Therefore, the null hypothesis $H_0 : \nu^T \beta = 0$ holds for all contrast vectors $\nu$ in (5), regardless of the pair of estimated connected components under consideration.

We solved (3) with $K = 2$ steps in the dual path algorithm, which yields exactly three estimated connected components by the properties of the one-dimensional fused lasso. Then, for each simulated dataset, we computed $p_{\hat{C}_1, \hat{C}_2}$ in (11), $p_{\text{Hyun}}$ in (10), and the naive p-value in (21).

Figure 3(b) displays the observed p-value quantiles versus Uniform(0,1) quantiles, aggregated over 1000 simulated datasets. We see that (i) the test based on the naive p-value in (21), which does not account for the fact that the connected components were estimated from the data, is anti-conservative; and (ii) tests based on $p_{\text{Hyun}}$ and $p_{\hat{C}_1, \hat{C}_2}$ control the selective Type I error (6).

5.1.2. Power as a Function of Effect Size

Next, we show that the test based on $p_{\hat{C}_1, \hat{C}_2}$ has higher power than that based on $p_{\text{Hyun}}$. We generated 1,500 datasets from (22) with $\sigma \in \{0.5, 1, 2\}$, for each of 10 evenly-spaced values of $\delta \in [0.5, 5]$. For every simulated dataset, we solved (3) with $K = 2$. We then rejected $H_0 : \nu^T \beta = 0$ if $p_{\text{Hyun}}$ or $p_{\hat{C}_1, \hat{C}_2}$ was less than $\alpha = 0.05$. Recalling that $\nu$ in (5) is a function of the data, and the effect size $|\nu^T \beta|$ will vary across simulated datasets drawn from an identical distribution, we created seven evenly-spaced bins of the observed values of $|\nu^T \beta|$, and then computed the proportion of simulated datasets for which we rejected $H_0$ within each bin.

![Figure 3](image-url)  

Figure 3. (a): One realization of $y$ generated according to (22) with $\delta = 3$ and $\sigma = 1$ (gray dots), along with the true signal $\beta$ (black curve). (b): When $\delta = 0$, tests based on both $p_{\text{Hyun}}$ in (10) and $p_{\hat{C}_1, \hat{C}_2}$ in (11) control the selective Type I error in the sense of (6). By contrast, the naive p-value in (21) leads to a test with inflated selective Type I error. (c): The power of the tests based on both $p_{\text{Hyun}}$ and $p_{\hat{C}_1, \hat{C}_2}$ increases as a function of $|\nu^T \beta|$. For a given bin of $|\nu^T \beta|$, the test based on $p_{\hat{C}_1, \hat{C}_2}$ has higher power than the test based on $p_{\text{Hyun}}$; the power of each test increases as $\sigma$ decreases.
Results are in Figure 3(c). The power of each test increases as the value of \(|v^T \beta|\) increases. For a given bin of \(|v^T \beta|\), the test based on \(p_{\hat{C}_1, \hat{C}_2}\) has higher power than the test based on \(p_{\text{Hyun}}\). For a given test and bin of \(|v^T \beta|\), a smaller value of \(\sigma\) results in higher power. As an alternative to binning, we can use regression splines to estimate the power as a smooth function of the effect size; see Appendix A.9, supplementary materials.

5.2. Two-Dimensional Fused Lasso

We consider the graph fused lasso on a grid graph, constructed by connecting each node to its four closest neighbors (up, down, left, right). This leads to the two-dimensional fused lasso problem, also known as total-variation denoising (Rudin, Osher, and Fatemi 1992; Tibshirani and Taylor 2011).

The signal \(\beta\) consists of with 64 observations arranged in an 8 \(\times\) 8 grid. It has three piecewise constant segments with means \(\delta, 0,\) and \(−\delta\), displayed in Figure 4(a):

\[
Y_j \sim \mathcal{N}(\beta_j, \sigma^2), \quad \beta_j = \delta \times 1_{(j \in C_1)} + (-\delta) \times 1_{(j \in C_2)},
\]

\[j = 1, \ldots, 64.\]  

5.2.1. Selective Type I Error Control under the Global Null

We simulated \(y_1, \ldots, y_{64}\) according to (23) with \(\delta = 0\) and \(\sigma = 1\). Thus, the null hypothesis \(H_0: v^T \beta = 0\) holds for any contrast vector \(v\) under consideration.

For each simulated dataset, we solved (3) with \(K = 15\) steps in the dual path algorithm, which typically yields between 2 and 4 estimated connected components. Then, provided that there was more than one connected component in the solution \(\hat{\beta}\), we computed \(p_{\text{Naive}}\) in (21), \(p_{\text{Hyun}}\) in (10), and \(p_{\hat{C}_1, \hat{C}_2}\) in (11). We rejected \(H_0\) if the \(p\)-values were less than \(\alpha = 0.05\).

Panel (b) of Figure 3 displays the observed \(p\)-values quantiles versus the Uniform(0, 1) quantiles, over 1,000 simulated datasets. As in Section 5.1.1, the tests based on both \(p_{\text{Hyun}}\) and \(p_{\hat{C}_1, \hat{C}_2}\) control the selective Type I error in (6), whereas the test based on \(p_{\text{Naive}}\) is anti-conservative.

5.2.2. Power as a Function of Effect Size

We generated data according to (23) with each of eight evenly-spaced values of \(\delta \in [0.5, 4]\) and \(\sigma \in [0.5, 1, 2]\). For each simulated dataset, we solved (3) with \(K = 15\) steps in the dual path algorithm. Provided that there were at least two estimated connected components, we then computed \(p_{\text{Hyun}}\) in (10) and \(p_{\hat{C}_1, \hat{C}_2}\) in (11), and rejected \(H_0\) if the \(p\)-values were less than 0.05.

In Figure 4(c), we display the proportion of simulated datasets for which we rejected \(H_0\) using the two tests, over seven evenly-spaced bins of \(|v^T \beta|\). For a given bin, the test based on \(p_{\hat{C}_1, \hat{C}_2}\) has substantially higher power than that based on \(p_{\text{Hyun}}\); the power of each test increases as \(\sigma\) decreases.

5.3. Allowing for Unknown Variance

Throughout this section, we have assumed that \(\sigma^2\) in (1) is known. In Appendix A.10, supplementary materials, we investigate the Type I error control and power of several variance estimators in simulations.

6. Data Applications

In this section, we apply our proposed \(p\)-value \(p_{\hat{C}_1, \hat{C}_2}\) to a dataset consisting of two measures: (a) drug overdose death rates (deaths per 100,000 persons), and (b) teenage birth rates (births per 1000 females aged 15–19), in each of the 48 contiguous states in the United States (Centers for Disease Control and Prevention 2020a, 2020b). In what follows, we consider the two measures after applying a log transformation. We can think of the data as noisy measurements of the true drug overdose death and teenage birth rate rates in each state, which are known to exhibit geographic trends (Ventura, Hamilton, and Matthews 2014; Amin et al. 2017; Schieber et al. 2019). Therefore, we solve the graph fused lasso in (3) with a custom graph that encodes the geography of the 48 states: each state is a node, and there is an edge between each contiguous pair of states. We then consider testing the equality of measures for pairs of estimated connected components.
For each pair of connected components, we computed three
$p$-values: $p_{\hat{C}_1, \hat{C}_2}$ in (11), $p_{\text{Hyun}}$ in (10), and $p_{\text{Naive}}$ in (21). We also computed
confidence intervals for $v^T \beta$, the difference between
population means of a pair of estimated connected components,
using $p_{\hat{C}_1, \hat{C}_2}$ and $p_{\text{Hyun}}$, as described in Section 4.1, along with
the naive confidence interval $[v^T y - z_{1-\alpha/2} \cdot \sigma ||v||_2, v^T y + z_{1-\alpha/2} \cdot \sigma ||v||_2]$, where $z_{\alpha}$ is the $\alpha$th quantile of the standard
normal distribution. For each $p$-value and confidence interval, we
used $\hat{\sigma}^2 = \frac{1}{48-\ell} \sum_{j=1}^{\ell} \sum_{j' \in \hat{C}_j} (y_j - (\sum_{j' \in \hat{C}_j} y_{j'}/|\hat{C}_j|))^2$ to
estimate $\sigma^2$ in (1), where $\hat{C}_1, \ldots, \hat{C}_4$ are the estimated connected
components. The average time required to test each hypothesis
$H_0 : v^T \beta = 0$ is 2 min.

6.1. Drug Overdose Death Rates in the Contiguous U.S. in 2018

Figure 5(a) displays the drug overdose death rate in a color
map. We solved (3) with $K = 30$ steps in the dual path
algorithm, which resulted in five connected components (see
Appendix A.12, supplementary materials for results with other
choices of $K$); the results are displayed in Figure 5(b). We have
estimated a constant drug overdose death rate in five geographical
regions, which we refer to as the Northeast ($\hat{C}_1$), Ohio ($\hat{C}_2$),
the West and Mountain region ($\hat{C}_3$), the Southeast ($\hat{C}_4$), and the
Midwest ($\hat{C}_5$). Among these regions, the Northeast and Midwest
have the highest estimated drug overdose rates.

We assess the equality of the means of each pair of connected
components using $p_{\text{Naive}}$ in (21), $p_{\text{Hyun}}$ in (10), and $p_{\hat{C}_1, \hat{C}_2}$ in
(11). The results are in Figure 5(c). The subset of pairs for
which $p_{\hat{C}_1, \hat{C}_2}$ is below 0.05 and $p_{\text{Hyun}}$ is not is displayed in bold.

For instance, the Northeast ($\hat{C}_1$) and the Southeast ($\hat{C}_4$) have a statistically significant difference in mean drug overdose death rates using the test based on $p_{\hat{C}_1, \hat{C}_2}$, but not using the test based on $p_{\text{Hyun}}$ at level $\alpha = 0.05$. Confidence intervals corresponding to these $p$-values are displayed in Figure 5(d). Intervals based on $p_{\text{Hyun}}$ are much wider than those based on $p_{\hat{C}_1, \hat{C}_2}$ across all 10 pairs of connected components. In addition, the confidence intervals based on $p_{\hat{C}_1, \hat{C}_2}$ are not much wider than those based on $p_{\text{Naive}}$, even though the latter do not have correct coverage for the true parameter $v^T \beta$.

6.2. Teenage Birth Rates in the Contiguous U.S. in 2018

Figure 6(a) displays the teenage birth rate in each of the 48
states. We solved the graph fused lasso with $K = 30$ steps of the
dual path algorithm, which results in five estimated connected
decomponents displayed in Figure 6(b); Appendix A.12, supple-
mentary materials contains additional results for $K = 20$. For
each pair of estimated connected components, we computed the
$p$-values $p_{\text{Naive}}, p_{\text{Hyun}},$ and $p_{\hat{C}_1, \hat{C}_2},$ along with the corresponding
confidence intervals for the difference in means. The results are displayed in Figures 6(c) and (d).

As in Section 6.1, at level $\alpha = 0.05$, the test based on $p_{\hat{C}_1, \hat{C}_2}$
makes more rejections than that based on $p_{\text{Hyun}}$. Additionally,
the confidence intervals based on $p_{\hat{C}_1, \hat{C}_2}$ are much narrower
than those based on $p_{\text{Hyun}}$; in some cases, the former are of
comparable length to those based on $p_{\text{Naive}}$.

![Figure 5](image_url)

**Figure 5.** (a) The observed drug overdose death rates (deaths per 100,000 persons) for the 48 contiguous U.S. states in the year 2018. (b): Applying the graph fused lasso to the drug overdose data results in five estimated connected components. (c): For each pair of estimated connected components, we computed $p_{\text{Naive}}$ in (21), $p_{\text{Hyun}}$ in (10), and $p_{\hat{C}_1, \hat{C}_2}$ in (11). For brevity, we use the notation $\hat{\beta}_j = \sum_{j' \in \hat{C}_j} \hat{\beta}_{j'}/|\hat{C}_j|$. (d): For each pair of estimated connected components, we constructed confidence intervals for the difference in means, corresponding to $p_{\text{Naive}}, p_{\text{Hyun}},$ and $p_{\hat{C}_1, \hat{C}_2}$. 

|   | $p_{\text{Naive}}$ | $p_{\text{Hyun}}$ | $p_{\hat{C}_1, \hat{C}_2}$ |
|---|------------------|------------------|------------------|
| $\beta_1$ | 0.78 | 0.78 | 0.45 |
| $\beta_1$ | $\beta_3$ | < 0.001 | < 0.001 | 0.890 |
| $\beta_1$ | $\beta_4$ | 0.003 | 0.024 | 0.820 |
| $\beta_1$ | $\beta_5$ | 0.12 | 0.49 | 0.68 |
| $\beta_1$ | $\beta_6$ | 0.007 | 0.64 | 0.58 |
| $\beta_1$ | $\beta_7$ | 0.10 | 0.70 | 0.65 |
| $\beta_1$ | $\beta_8$ | 0.29 | 0.78 | 0.52 |
| $\beta_1$ | $\beta_9$ | 0.03 | 0.21 | 0.36 |
| $\beta_1$ | $\beta_5$ | 0.003 | 0.039 | 0.650 |
| $\beta_1$ | $\beta_6$ | 0.36 | 0.24 | 0.60 |
7. Discussion

We have proposed a new procedure for testing the difference in the means of two connected components resulting from the graph fused lasso. Our approach conditions on less information than existing approaches, leading to substantially higher power while still controlling the selective Type I error.

Methods developed in this article are implemented in the R package GFLassoInference. Instructions on how to download and use this package can be found at https://yiqunchen.github.io/GFLassoInference. Code and files to reproduce the results in the article can be found at https://github.com/yiqunchen/GFLassoInference-experiment.

7.1. Incorporating the Selection of the Tuning Parameter

Throughout this article, we have chosen \( K \), the number of steps in the dual path algorithm for (3), without making use of the data. However, in practice, the tuning parameter \( K \) is often selected based on the data. For instance, we could choose the value of \( K \) that minimizes the modified Bayesian information criterion (Hyun, G’Sell, and Tibshirani 2018; Zhang and Siegmund 2007). We leave the details to future work.

7.2. Extension to other Generalized Lasso Problems

Ideas in this article apply beyond the setting of the piecewise constant model in (1) and the graph fused lasso estimator in (3). For instance, we can consider extending our proposal to the trend filtering problem, which postulates that the underlying signal is ordered and piecewise polynomial (Kim et al. 2009; Tibshirani 2014). Because trend filtering is a special case of (2) and can be solved using the dual path algorithm, an extension of the approach in Section 3 can be applied.

In addition, we can extend our proposal from an identity matrix in (2) to any design matrix \( X \in \mathbb{R}^{n \times q} \) with full column rank, that is, \( \hat{\beta} = \arg \min_{\beta \in \mathbb{R}^q} \{ \frac{1}{2} ||y - X\beta||_2^2 + \lambda ||D\beta||_1 \} \). Hyun, G’Sell, and Tibshirani (2018) showed that a p-value similar to (10) can be used in this case to test the hypothesis (4). Therefore, we can directly apply the computational insights in Section 3 to obtain a more powerful test.

We leave the details of outlined extensions, as well as comparisons to recent selective inference tools for trend filtering (e.g., Mehrizi and Chenouri 2021; Leiner et al. 2021), to future work.

7.3. Relaxes Assumptions in (1)

While the idea of conditioning on less information to improve the power of a selective inference procedure applies regardless of the distributions of the observations, the assumptions in model (1) are critical to the proof of Proposition 3, and therefore, the efficient computation of \( \bar{p}_{C_1,C_2} \). A line of recent work in selective inference has focused on relaxing these assumptions in high-dimensional linear modeling (Tibshirani et al. 2018; Tian and Taylor 2018; Charkhi and Claeskens 2018), and may be applicable to the generalized lasso. Alternatively, we can extend (1) to other exponential family distributions by leveraging the recent developments in generalized data carving (Rasines and Alastair Young 2021; Leiner et al. 2021; Schultheiss, Renaux, and Bühlmann 2021).
Supplementary Materials

The supplementary material contains proofs, technical details, and additional simulation results. [Online].

Acknowledgments

We thank the authors of Le Duy and Takeuchi (2021) for providing us with their software implementation.

Disclosure Statement

The authors report there are no competing interests to declare.

Funding

This work was partially supported by National Institutes of Health grants [R01EB026908, R01DA047869] and a Simons Investigator Award to D.W.

ORCID

Sean Jewell http://orcid.org/0000-0002-1843-5964

References

Amin, R., Deescare, J. Z., Hans, J., and Roussos-Ross, K. (2017), “Epidemiologic Surveillance of Teenage Birth Rates in the United States, 2006–2012,” Obstetrics and Gynecology, 129, 1068–1077. [584]
Arnold, T. B., and Tibshirani, R. J. (2016), “Efficient Implementations of the Generalized Lasso Dual Path Algorithm,” Journal of Computational and Graphical Statistics, 25, 1–27. [577]
Centers for Disease Control and Prevention. (2020a), “2018 Drug Overdose Death Rates,” available at https://www.cdc.gov/drugoverdose/data/statedeaths/drug-overdose-death-2018.html. [584]
——— (2020b), “2018 Teenage Birth Rates,” Available at https://www.cdc.gov/nchs/pressroom/sosmap/teenbirths/teenbirths.htm. [584]
Charkhi, A., and Claeskens, G. (2018), “Asymptotic Post-Selection Inference for the Akaike Information Criterion,” Biometrika, 105, 645–664. [586]
Chen, S., and Bien, J. (2020), “Valid Inference Corrected for Outlier Removal,” Journal of Computational and Graphical Statistics, 29, 323–334. [582]
Fithian, W., Sun, D., and Taylor, J. (2014), “Optimal Inference after Model Selection,” arXiv preprint arXiv:1410.2597. [578,579,580,582]
Friedman, J., Hastie, T., Hölting, H., and Tibshirani, R. (2007), “Pathwise Coordinate Optimization,” The Annals of Applied Statistics, 1, 302–332. [577]
Gao, L. L., Bien, J., and Witten, D. (2020), “Selective Inference for Hierarchical Clustering,” arXiv preprint arXiv:2012.02936. [583]
Harchaoui, Z., and Lévy-Leduc, C. (2010), “Multiple Change-Point Estimation with a Total Variation Penalty,” Journal of the American Statistical Association, 105, 1480–1493. [577]
Hastie, T., Tibshirani, R., and Wainwright, M. (2015), Statistical Learning with Sparsity: The Lasso and Generalizations, Boca Raton, FL: Chapman and Hall/CRC. [577]
Hyun, S., G’Sell, M., and Tibshirani, R. J. (2018), “Exact Post-selection Inference for the Generalized Lasso Path,” Electronic Journal of Statistics, 12, 1053–1097. [578,579,580,582,586]
Hyun, S., Lin, K. Z., G’Sell, M., and Tibshirani, R. J. (2021), “Post-Selection Inference for Change-point Detection Algorithms with Application to Copy Number Variation Data,” Biometrics, 77, 1037–1049. [583]
Jewell, S., Fearnhead, P., and Witten, D. (2022), “Testing for a Change in Mean after Change-point Detection,” Journal of the Royal Statistical Society, Series B (to appear). [578,580,581]
Johnson, N. A. (2013), “A Dynamic Programming Algorithm for the Fused Lasso and Δ-Segmentation,” Journal of Computational and Graphical Statistics, 22, 246–260. [577]
Kim, S.-J., Koh, K., Boyd, S., and Gorinevsky, D. (2009), “ℓ1 trend Filtering,” SIAM Review, 51, 339–360. [577,586]
Le Duy, V. N., and Takeuchi, I. (2021), “More Powerful Conditional Selective Inference for Generalized Lasso by Parametric Programming,” arXiv preprint arXiv:2105.04920. [578,580,587]
Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016), “Exact Post-selection Inference, with Application to the Lasso,” The Annals of Statistics, 44, 907–927. [578,579,580,582]
Leiner, J., Duan, B., Wasserman, L., and Ramdas, A. (2021), “Data Blurring: Sample Splitting a Single Sample,” arXiv preprint arXiv:2112.11079. [586]
Liu, K., Markovic, J., and Tibshirani, R. (2018), “More Powerful Post-Selection Inference, with Application to the Lasso,” arXiv preprint arXiv:1801.09037. [578,580]
Lofus, J. R., and Taylor, J. E. (2014), “A Significance Test for Forward Stepwise Model Selection,” arXiv preprint arXiv:1405.3920. [579]
Mehrizi, R. V., and Chenoury, S. (2021), “Valid Post-Detection Inference for Change Points Identified using Trend Filtering,” arXiv preprint arXiv:2104.12022. [586]
Ramdas, A., and Tibshirani, R. J. (2016), “Fast and Flexible ADMM Algorithms for Trend Filtering,” Journal of Computational and Graphical Statistics, 25, 839–858. [577]
Rasines, D. G., and Alastair Young, G. (2021), “Splitting Strategies for Post-selection Inference,” arXiv preprint arXiv:2102.02159. [586]
Reid, S., Taylor, J., and Tibshirani, R. (2017), “Post-Selection Point and Interval Estimation of Signal Sizes in Gaussian Samples,” Canadian Journal of Statistics, 45, 128–148. [579]
Rinaldo, A. (2009), “Properties and Refinements of the Fused Lasso,” The Annals of Statistics, 37, 2922–2952. [577]
Rudin, L. I., Osher, S., and Fatemi, E. (1992), “Nonlinear Total Variation Based Noise Removal Algorithms,” Physica D: Nonlinear Phenomena, 60, 259–268. [584]
Sadhanala, V., Wang, Y.-X., and Tibshirani, R. J. (2016), “Total Variation Classes Beyond 1d: Minimax Rates, and the Limitations of Linear Smoothers,” Advances in Neural Information Processing Systems, 29, 3513–3521. [577]
Schieber, L. Z., Guy, G. P., Seth, P., Young, R., Mattson, C. L., Milkos, C. A., and Schieber, R. A. (2019), “Trends and Patterns of Geographic Variation in Opioid Prescribing Practices by State, United States, 2006–2017,” JAMA Network Open, 2, e190665. [584]
Schultheiss, C., Renaux, C., and Bühlmann, P. (2021), “Multicarving for High-Dimensional Post-Selection Inference,” Electronic Journal of Statistics, 15, 1695–1742. [586]
Tian, X., and Taylor, J. (2018), “Selective Inference with a Randomized Response,” The Annals of Statistics, 46, 679–710. [586]
Tibshirani, R. (1996), “Regression Shrinkage and Selection via the Lasso,” Journal of the Royal Statistical Society, Series B, 58, 267–288. [577]
Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005), “Sparsity and Smoothness via the Fused Lasso,” Journal of the Royal Statistical Society, Series B, 67, 91–108. [577]
Tibshirani, R. J. (2014), “Adaptive Piecewise Polynomial Estimation via Trend Filtering,” The Annals of Statistics, 42, 285–323. [577,586]
Tibshirani, R. J., Rinaldo, A., Tibshirani, R., and Wasserman, L. (2018), “Uniform Asymptotic Inference and the Bootstrap After Model Selection,” The Annals of Statistics, 46, 1255–1287. [586]
Tibshirani, R. J., and Taylor, J. (2011), “The Solution Path of the Generalized Lasso,” The Annals of Statistics, 39, 1335–1371. [577,578,583,584]
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, 600–620. [578,579,582]
Ventura, S. J., Hamilton, B. E., and Matthews, T. J. (2014), “National and State Patterns of Teen Births in the United States, 1940–2013,” National Vital Statistics Reports, 63, 1–34. [584]
Xin, B., Kawahara, Y., Wang, Y., and Gao, W. (2014), “Efficient Generalized Fused Lasso and its Application to the Diagnosis of Alzheimer’s Disease,” in Twenty-Eighth AAAI Conference on Artificial Intelligence. [577]
Zhang, N. R., and Siegmund, D. O. (2007), “A Modified Bayes Information Criterion with Applications to the Analysis of Comparative Genomic Hybridization Data,” Biometrics, 63, 22–32. [586]
Zhu, Y. (2017), “An Augmented ADMM Algorithm with Application to the Generalized Lasso Problem,” Journal of Computational and Graphical Statistics, 26, 195–204. [577]