More Powerful and General Selective Inference for Stepwise Feature Selection using the Homotopy Continuation Approach

Kazuya Sugiyama†, Vo Nguyen Le Duy††, Ichiro Takeuchi†∗

December 29, 2020

Abstract
Conditional Selective Inference (SI) has been actively studied as a new statistical inference framework for data-driven hypotheses. For example, conditional SI framework enables exact (non-asymptotic) inference on the features selected by stepwise feature selection (SFS) method. The basic idea of conditional SI is to make inference conditional on the selection event. The main limitation of existing conditional SI approach for SFS method is the loss of power due to over-conditioning for computational tractability. In this paper, we develop more powerful and general conditional SI method for SFS by resolving the over-conditioning issue by homotopy continuation approach. We conduct several experiments to demonstrate the effectiveness and efficiency of our proposed method.

1Nagoya Institute of Technology
2RIKEN
†Equal contribution
∗Correspondence to: Ichiro Takeuchi (takeuchi.ichiro@nitech.ac.jp)
1 Introduction

As machine learning (ML) is being applied to a greater variety of practical problems, ensuring the reliability of ML is recognized as becoming increasingly important. Among several potential approaches to reliable ML, conditional selective inference (SI) is recognized as a promising approach for evaluating the statistical reliability of data-driven hypotheses selected by ML methods. The basic idea of conditional SI is to make inference on a data-driven hypothesis conditional on the selection event that the hypothesis is selected by analyzing the data with the ML algorithm. Conditional SI has been actively studied especially in the context of feature selection. Notably, Lee et al. [1] and Tibshirani et al. [2] proposed conditional SI methods for exact conditional inference on selected features by using Lasso and stepwise feature selection (SFS), respectively. The basic idea of these conditional SI methods is to characterize the hypothesis selection event by a polytope, i.e., a set of linear inequalities, in the sample space. When a hypothesis selection event is characterized by a polytope, the authors in these studies developed practical methods for making inference on the selected hypotheses by deriving the exact (non-asymptotic) sampling distribution conditional on the polytope. In this paper, we call conditional SI based on a polytope polytope-based SI. These studies are regarded as significant advance in the field of statistical inference since traditional statistical inference cannot cope with hypotheses selected after observing the data.

Unfortunately, however, polytope-based SI has several limitations because it can be used only when the characterization of all the relevant selection events is represented by a polytope. In fact, in most of the existing polytope-based SI studies, extra-conditioning is required in order for the selection event to be characterized as a polytope. For example, in the SI for SFS method of Tibshirani et al. [2], the authors needed to consider conditioning not only on the selected features but also on extra event regarding the signs and the orders in the feature selection process. Such extra-conditioning leads to loss of power in the inference [3].
In this paper, we go beyond polytope-based SI and propose a novel SI method by using homotopy continuation for conditional inference on the features selected by SFS. We call the proposed method homotopy-based SI. Compared to the polytope-based SI for SFS method of [2], the proposed method is more powerful and more general. The basic idea of homotopy-based SI is to use the homotopy continuation approach to keep track of the hypothesis selection event when the data changes in the direction of the selected test statistic, which enables efficient identification of the subspace of the sample space in which the same hypothesis is selected. We demonstrate the effectiveness of the proposed homotopy-based SI method through intensive simulations and real data analyses.

Related Work A fundamental issue in current data analysis with more increasingly complicated datasets and algorithms is post-selection inference. Traditional statistical inference presumes that, before observing the data, we have decided on a statistical model and a statistical target for which we seek to conduct inference. Therefore, if we apply traditional statistical inference methods to hypotheses selected after observing the data, a selection bias is introduced and the inferential results are no longer valid. This issue has been extensively discussed in the context of post-feature-selection inference. In fact, even in commonly-used feature selection methods such as SFS, it has been difficult to correctly assess the statistical reliability of the selected features.

There have been several approaches suggested in the literature toward addressing this problem [4 5 6 7 8 9 10 11]. A particularly notable approach that has received considerable attention in the past few years is conditional selective inference introduced in the seminal paper by Lee et al. [1]. In their work, the authors showed that, for a feature selected by Lasso, the hypothesis selection event can be characterized as a polytope by conditioning on the selected set of features as well as extra events on their signs, and exact conditional inference can be conducted by exploiting the polyhedral selection event (polytope-based SI). Furthermore, Tibshirani et al. [2] showed that polytope-based SI is also applicable to SFS by conditioning on the set of selected features and by imposing
some extra conditions regarding the signs and the orders in the feature selection process. Conditional SI has been actively studied in the past few years and extended to various directions [12, 13, 14, 15, 16, 17, 18, 19, 2, 20, 21, 22, 23, 24].

Note that in conditional SIs, we typically prefer to condition on as little as possible [3], so that the resulting inference can be more powerful. Often, there are statistical reasons that make it necessary to condition [3], but other times the reasons are purely computational (e.g. conditioning on the signs and the orders in the SFS process). Namely, the main limitations of current polytope-based SI methods are that excessive conditioning is required to represent the selection event with a single polytope, and that selection events that cannot be represented with a polytope cannot be handled properly. In the seminal paper by Lee et al. [1], the authors already discussed the issue of over-conditioning, and explained how conditioning on signs can be omitted by an exhaustive enumeration of all possible signs and taking the union over the resulting polyhedra. However, such an exhaustive enumeration of exponentially increasing number of sign combination is feasible only when the number of selected features is fairly small. Several other approaches were proposed to circumvent the drawbacks and the restrictions of polytope-based SI. Loftus et al. [18] extended the polytope-based SI such that their method could handle selection events characterized by quadratic inequalities, but it inherits the over-conditioning issue in polytope-based SI. To improve the power, Tian et al. [14] proposed an approach to randomize the algorithm in order to condition on less. Terada et al. [25] proposed to use bootstrap re-sampling to characterize the selection event more generally. A drawback of these approaches is that additional randomness is introduced into the algorithm and/or the inference.

This study is motivated by Liu et al. [26] and Duy et al. [27]. The former studied Lasso SI for full-model parameters, while the latter extended the basic idea of the former so that it can be also applied to Lasso SI in more general settings including inference on partial-model parameters (exactly same problem setup as [1]). These two studies go beyond the polytope-based SI for more powerful Lasso SI without conditioning on the signs by carefully analyzing the
convex optimality conditions of the Lasso solutions. Unfortunately, in SI for SFS, we cannot use the optimality conditions of the solutions because SFS is not formulated as a convex optimization problem. Furthermore, in SI for SFS, we need to consider extra-conditioning not only on the signs but also on the order in the SFS process.

2 Problem Statement

We consider forward stepwise feature selection (SFS) method for regression problem. Let \( n \) be the number of instances and \( p \) be the number of original features. We denote the observed dataset as \((X, y^{\text{obs}})\) where \( X \in \mathbb{R}^{n \times p} \) is the design matrix and \( y^{\text{obs}} \in \mathbb{R}^n \) be the response vector. Following the problem setup in existing conditional SI literature such as [1] and [2], we assume that the observed response \( y^{\text{obs}} \) is a realization of the following random response vector

\[
y = (y_1, \ldots, y_n)^\top \sim \mathcal{N}(\mu, \Sigma),
\]

where \( \mu \in \mathbb{R}^n \) is the unknown mean vector, and \( \Sigma \in \mathbb{R}^{n \times n} \) is the covariance matrix which is known or estimable from independent data, and the design matrix \( X \) is assumed to be non-random. For notational simplicity, we assume that each column vector of \( X \) is normalized to have unit length.

Stepwise feature selection We consider the standard forward SFS method as studied in [2] — at each step of the SFS method, the feature which most improves the fit is newly added. When each feature has unit length, it is equivalent to the feature which is most correlated with the residual of the least-square regression model fitted with previously selected features. For a response vector \( y \in \mathbb{R}^n \) and a set of features \( M \subseteq [p] \), let \( r(y, X_M) \) be the residual vector obtained by regressing \( y \) onto \( X_M \) for a set of features \( M \), i.e.,

\[
r(y, X_M) = P_{X_M}^\perp y = (I_n - P_{X_M})y
\]
where $I_n$ is the $n$-by-$n$ identity matrix, $P_{X_M} = X_M(X_M^TX_M)^{-1}X_M^T$ and $P_{X_M}^\perp = I_n - X_M(X_M^TX_M)^{-1}X_M^T$. Let $K$ be the number of selected features by the SFS method\footnote{We discuss a situation where the number of selected features $K$ is selected by cross-validation in §3.4. In other parts of the paper, we assume that $K$ is determined before looking at the data.} When we need to clarify the fact that features are selected by applying $K$-step SFS method to a response vector $y \in \mathbb{R}^n$, we denote the set of selected features as $\mathcal{M}_K(y) \subseteq [p]$. Similarly, we denote the sequence of selected feature at each step as $j_k(y)$ for $k \in [K]$. In the SFS method, the selected feature at step $k \in [K]$ is defined as

$$j_k(y) = \arg\min_{j \in [p] \setminus \mathcal{M}_{k-1}(y)} \| r(y, X_{\mathcal{M}_{k-1}(y)} \cup \{j\}) \|_2^2$$

$$= \arg\max_{j \in [p] \setminus \mathcal{M}_{k-1}(y)} | x_j^T r(y, X_{\mathcal{M}_{k-1}(y)}) | , \quad (2)$$

where $\mathcal{M}_0(y) := \emptyset$ and $r(y, \mathcal{M}_0(y)) := y$.

**Statistical inference** In order to quantify the statistical significance of the relation between the selected features and the response, we consider the statistical test for each of the coefficient of the selected model parameters

$$\hat{\beta}_{\mathcal{M}_K(y)} = (X_{\mathcal{M}_K(y)}^TX_{\mathcal{M}_K(y)})^{-1}X_{\mathcal{M}_K(y)}^Ty,$$

where $X_{\mathcal{M}_K(y)}$ is the matrix with the set of columns corresponding to $\mathcal{M}_K(y)$.

Note that the $j^{\text{th}}$ coefficient is written as $\hat{\beta}_{\mathcal{M}_K(y),j} = \eta^T y$ by defining

$$\eta = X_{\mathcal{M}_K(y)} \left( X_{\mathcal{M}_K(y)}^TX_{\mathcal{M}_K(y)} \right)^{-1} e_j,$$

where $e_j \in \mathbb{R}^{\mathcal{M}_K(y)}$ is a unit vector whose $j^{\text{th}}$ element is 1 and 0 otherwise. Note that $\eta$ depends on $\mathcal{M}_K(y)$ and $j$, but we omit the dependence for notational simplicity. We consider the following statistical test for each of the coefficient $\beta_{\mathcal{M}_K(y),j} = \eta^T \mu$

$$H_0 : \beta_{\mathcal{M}_K(y),j} = 0 \quad \text{vs.} \quad H_1 : \beta_{\mathcal{M}_K(y),j} \neq 0,$$

where $\beta_{\mathcal{M}_K(y),j}$ is the $j^{\text{th}}$ element of the population least squares $\beta_{\mathcal{M}_K(y)} = P_{\mathcal{M}_K(y)} \mu$, i.e., the projection of $\mu$ onto the column space of $X_{\mathcal{M}_K(y)}$. 
Conditional Selective Inference (SI)  Since the target of the inference is selected by observing the data $y$, if we naively apply a traditional statistical test to the problem in (4) as if the inference target is pre-determined, the result is not valid (type I error cannot be controlled at the desired significance level) due to selection bias. To address the selection bias issue, we consider conditional selective inference (SI) framework introduced in [1] and [2]. In conditional SI framework, the inference is conducted based on the following conditional sampling distribution of the test-statistic:

$$\eta^\top y \mid \{M_K(y) = M_K(y^{obs}), q_K(y) = q_K(y^{obs})\},$$

where

$$q_K(y) = (I_n - c\eta^\top)y$$

is the nuisance parameters which is independent of the test statistic. The first condition $M_K(y) = M_K(y^{obs})$ in (5) indicates the event that the set of selected features obtained by applying the $K$-step SFS method to a random response vector $y$ is the same as the ones for $y^{obs}$. The second condition $q_K(y) = q_K(y^{obs})$ indicates the nuisance parameters for a random response vector $y$ is the same as the one for the observed vector $y^{obs}$.

To conduct the conditional inference for (5), the main task is to identify the conditional data space

$$\mathcal{Y} = \{y \in \mathbb{R}^n \mid M_K(y) = M_K(y^{obs}), q_K(y) = q_K(y^{obs})\}.$$ 

Once $\mathcal{Y}$ is identified, we can easily compute the pivotal quantity

$$F_{\eta^\top\eta, \Sigma\eta^\top\eta}^{Z_m,s^2} y \mid y \in \mathcal{Y},$$

where $F_{m,s^2}^{Z_m,s^2}$ is the c.d.f. of the truncated Normal distribution with mean $m$, variance $s^2$, and the truncation region $Z$. Later, we will explain how $Z$ is defined in (7) is defined. The pivotal quantity is crucial for calculating $p$-value

\[^2\text{The } q_K(y) \text{ corresponds to the component } z \text{ in the seminal paper (see [1], Sec 5, Eq 5.2 and Theorem 5.2).} \]
or obtaining confidence interval. Based on the pivotal quantity, we can obtain selective type I error or selective p-value in the form of

$$p_{K,j}^\text{selective} = 2 \min\{\pi_{K,j}, 1 - \pi_{K,j}\},$$

(8)

where $\pi_{K,j} = 1 - F_{Z}^{-1}(\eta^\top \eta | y)$. This p-value is valid in the sense that

$$\Pr_{H_0}(p_{K,j}^\text{selective} < \alpha) = \alpha, \forall \alpha \in [0, 1].$$

Furthermore, to obtain $1 - \alpha$ confidence interval for any $\alpha \in [0, 1]$, by inverting the pivotal quantity in (7), we can find the smallest and largest values of $\eta^\top \mu$ such that the value of pivotal quantity remains in the interval $[\alpha^2, 1 - \alpha^2]$.

**Characterization of the conditional data space $\mathcal{Y}$**

Using the second condition in (6), the data in $\mathcal{Y}$ is restricted to a line (see [24] and [3]). Therefore, the set $\mathcal{Y}$ can be re-written, using a scalar parameter $z \in \mathbb{R}$, as

$$\mathcal{Y} = \{ y(z) \in \mathbb{R}^n \mid y(z) = a + bz, z \in \mathcal{Z} \}$$

(9)

where $a = q_K(y_{\text{obs}})$, $b = \Sigma \eta (\eta^\top \Sigma \eta)^{-1}$, and

$$\mathcal{Z} = \{ z \in \mathbb{R} \mid M_K(y(z)) = M_K(y_{\text{obs}}) \}.$$  

(10)

Here, like $\eta$, $a$, $b$ and $\mathcal{Z}$ depend on $M_K(y)$ and $j$, but we omit the subscripts for notational simplicity. Now, let us consider a random variable $z \in \mathbb{R}$ and its observation $z_{\text{obs}} \in \mathbb{R}$ such that they respectively satisfy $y = a + bz$ and $y_{\text{obs}} = a + bz_{\text{obs}}$. The conditional inference in (5) is re-written as the problem of characterizing the sampling distribution of

$$z \mid z \in \mathcal{Z}.$$  

(11)

Since $z \sim \mathcal{N}(\eta^\top \mu, \eta^\top \Sigma \eta)$, $z \mid z \in \mathcal{Z}$ follows a truncated Normal distribution. Once the truncation region $\mathcal{Z}$ is identified, the pivotal quantity in (7) is obtained as $F_{\eta^\top \mu, \eta^\top \Sigma \eta}(z)$. Thus, the remaining task is reduced to the characterization of $\mathcal{Z}$. 

8
Figure 1: A schematic illustration of the proposed method. By applying $k$-step SFS algorithm on the observed data $y^{\text{obs}}$, we obtain a set of selected features. Then, we parametrize $y^{\text{obs}}$ with a scalar parameter $z$ in the dimension of test-statistic to identify the subspace $\mathcal{Y}$ whose data has the \textit{same set of selected features} as $y^{\text{obs}}$ has, regardless of the differences in signs and sequential orders. Finally, the valid statistical inference is conducted conditional on $\mathcal{Y}$. We introduce a homotopy continuation for efficiently characterizing the conditional data space $\mathcal{Y}$.

**Extra-conditioning in existing conditional SI methods** Unfortunately, it has been considered computationally infeasible to fully identify the truncation region $\mathcal{Z}$ in conditional SI for SFS method. Therefore, in the existing conditional SI studies such as [2], the authors circumvent the computational difficulty by \textit{overly conditioning} with \textit{extra-conditions}. Note that over-conditioning is not harmful for selective type-I error control, but it has been known that over-conditioning leads to the loss of power in conditional SI [3]. In fact, the decrease in the power due to over-conditioning is not unique issue for SFS in [2], but is a common major issue in many existing conditional SIs [1]. In the next section, we propose a method to overcome the difficulty by removing the extra-conditions for minimumly-conditioned SI for SFS.
3 Proposed Method

As we discussed in §2, to conduct the conditional SI, the truncation region \( Z \subseteq \mathbb{R} \) in (10) must be identified. To construct \( Z \), our idea is 1) computing \( M_K(y(z)) \) for all \( z \in \mathbb{R} \), and 2) identifying the set of intervals of \( z \in \mathbb{R} \) on which \( M_K(y(z)) = M_K(y_{\text{obs}}) \). However, it seems intractable to obtain \( M_K(y(z)) \) for infinitely many values of \( z \in \mathbb{R} \). To overcome the difficulty, we combine two approaches called \textit{extra-conditioning} and \textit{homotopy continuation}.

Figure 1 shows the schematic illustration of the proposed method. Our idea is motivated by regularization path of Lasso [28, 29], SVM [30] and other similar methods [31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41] in which the solution path along the regularization parameter can be computed by analyzing the KKT optimality conditions of the parametrized convex optimization problems. Although SFS cannot be formulated as a convex optimization problem, by introducing the notion of extra-conditioning, we note that conceptually similar approach as homotopy continuation can be used to keep track all possible changes of the selected features when the response vector \( y \) changes along the direction of the test-statistic.

3.1 Extra-Conditioning

First, we consider not only the feature selection event \( M_K(y) = M_K(y_{\text{obs}}) \) but also an \textit{extra} event \( E_K(y) = E_K(y_{\text{obs}}) \). In conditional SI for SFS method, the extra event consists of two types of information regarding \textit{orders} and \textit{signs} which are necessary for characterizing the process of SFS method. We use a symbol \( O \) for the former and \( S \) for the latter. Like the notation \( M_K(y) \), when we need to clarify the fact that they are \textit{selected} by applying \( K \)-step SFS method to a response vector \( y \in \mathbb{R}^n \), we denote them as \( O_K(y) \) and \( S_K(y) \), respectively. Thus, the extra event is denoted as \( E_K(y) = (O_K(y), S_K(y)) \).

The former \( O_K(y) \) contains the information on the order in which the \( K \) features are selected. Concretely, we define \( O_K(y) \) to be the selected permutation of the set of selected features \( M_K(y) = \{j_1, j_2, \ldots, j_K\} \). By combining
The latter $S_K(y)$ contains the information on the sign when each feature is entered to the model. Concretely, we define $S_K(y) \in \{-1, +1\}^K$ to be a binary vector with length $K$ whose $k^{th}$ element is defined as

$$S_{K,k}(y) = \text{sgn}(x_j^\top r(y, X_{M_{k-1}(y)}))$$

which indicates the correlation between the selected feature at step $k$ and the residual vector after $k - 1$ steps.

The following lemma tells that, by conditioning not only on the feature selection event but also on the extra events on sign and order, the over-conditioned truncation region can be simply represented as an interval in the line $\{y(z) \in \mathbb{R}^n \mid y(z) = a + bz, z \in \mathbb{R}\}$.

**Lemma 1.** For a response vector $y' \in \{y(z) \in \mathbb{R}^n \mid y(z) = a + bz, z \in \mathbb{R}\}$, the over-conditioned truncation region defined as

$$Z_{oc}(y') = \{z \in \mathbb{R} \mid M_K(y(z)) = M_K(y'), \mathcal{E}_K(y(z)) = \mathcal{E}_K(y')\}$$

is an interval in $\mathbb{R}$.

**Proof.** By conditioning on the feature selection event and extra event, the triplet $(M_K(y(z)), \mathcal{O}_K(y(z)), S_K(y(z)))$ is fixed as

$$M_K(y(z)) = M_K(y'), \mathcal{O}_K(y(z)) = \mathcal{O}_K(y'), S_K(y(z)) = S_K(y').$$

From the first two conditions,

$$\{M_K(y(z)) = M_K(y'), \mathcal{O}_K(y(z)) = \mathcal{O}_K(y')\}$$

$$\iff M_k(y(z)) = M_k(y') \forall k \in \{0\} \cup [K].$$

Therefore, we have

$$\{M_K(y(z)) = M_K(y'), \mathcal{O}_K(y(z)) = \mathcal{O}_K(y')\}$$

$$\iff \bigcap_{k=1}^K \bigcap_{j \in [p] \setminus M_{k-1}(y')} |x_j^\top r(y(z), X_{M_{k-1}(y')})| \geq \pm x_j^\top r(y(z), X_{M_{k-1}(y')}).$$

(13)
By further conditioning on the sign information, the condition is written as

\[
\{M_K(y(z)) = M_K(y'), O_K(y(z)) = O_K(y'), S_K(y(z)) = S_K(y')\}
\]

\[\Leftrightarrow \bigcap_{k=1}^{K} \bigcap_{j \in [p] \setminus \mathcal{M}_{k-1}(y')} S_{K,k}(y') x_{jk}^\top r(y(z), X_{\mathcal{M}_{k-1}(y')}) \geq \pm x_j^\top r(y(z), X_{\mathcal{M}_{k-1}(y')}).\]

By restricting on a line \(y(z) = a + bz, z \in \mathbb{R}\), the range of \(z\) is written as

\[
\max_{k \in [K], j \in [p] \setminus \mathcal{M}_{k-1}(y')} \frac{e_{(k,j)}}{d_{(k,j)}}, \min_{k \in [K], j \in [p] \setminus \mathcal{M}_{k-1}(y')} \frac{e_{(k,j)}}{d_{(k,j)}} \leq z \leq \min_{k \in [K], j \in [p] \setminus \mathcal{M}_{k-1}(y')} \frac{e_{(k,j)}}{d_{(k,j)}},
\]

where

\[
e_{(k,j)} = (x_j - x_{jk} S_{K,k}(y'))^\top P_{X_{\mathcal{M}_{k-1}(y')}}^\perp a,
\]

\[
d_{(k,j)} = (x_{jk} S_{K,k}(y') - x_j)^\top P_{X_{\mathcal{M}_{k-1}(y')}}^\perp b,
\]

for \(k \in [K], j \in [p] \setminus \mathcal{M}_{k-1}(y')\).

Note that the original conditional SI for SFS in [2] exactly considered only the over-conditioning case in the above lemma. The polytope-based SI is applicable to SFS method only with the extra conditioning \(E_K(y) = E_K(y_{\text{obs}})\). Although it is known that the over-conditioning leads to loss of power [3], it has been recognized to be computationally infeasible to remove the extra conditioning \(E_K(y) = E_K(y_{\text{obs}})\). In the following subsection, we overcome this computational difficulty by homotopy continuation approach.

### 3.2 Homotopy Continuation

Consider all possible triplets \((M_K(y(z)), O_K(y(z)), S_K(y(z)))\) for all possible response vectors \(y(z)\) in a line \(\{y(z) \in \mathbb{R}^n \mid y(z) = a + bz, z \in \mathbb{R}\}\). Clearly, the number of such triplets is finite. With a slight abuse of notation, we index each of the triplets by \(t = 1, 2, \ldots, T\) and denote it such as \((M_K^{(t)}, O_K^{(t)}, S_K^{(t)})\) where \(T\) is the number of the triplets. Lemma 1 indicates that each triplet corresponds to an interval in the line. Without loss of generality, we assume that the left most interval corresponds to \(t = 1\), the second left most interval corresponds to \(t = 2\),
and so on. Then, using an increasing sequence of real numbers $z_1 < z_2 < \cdots < z_T < z_{T+1}$, we can write these intervals as $[z_1, z_2], [z_2, z_3], \ldots, [z_T, z_{T+1}]$. In practice, we do not have to consider the entire line $z \in (-\infty, \infty)$, but it suffices to consider, e.g., $z \in [-20\sigma, 20\sigma]$ where $\sigma$ is the standard error of the test-statistic, since the probability mass in the out of the range is negligibly small. Thus, we set $z_{\min} = z_1 = -20\sigma$ and $z_{\max} = z_{T+1} = 20\sigma$ in our implementation and consider only the set of the intervals within the range.

Our simple idea is to compute all possible triplets $\{(M^{(t)}_K, O^{(t)}_K, S^{(t)}_K)\}_{t=1}^T$ by keeping track of the intervals $[z_1, z_2], [z_2, z_3], \ldots, [z_T, z_{T+1}]$ one by one, and then to compute the truncation region $Z$ by collecting the intervals in which the set of selected features $M^{(t)}_K$ is the same as the set of actually selected features from the observed data $M_K(y^{\text{obs}})$, i.e.,

$$Z = \bigcup_{t: M^{(t)}_K = M_K(y^{\text{obs}})} [z_t, z_{t+1}] = \bigcup_{t: M_K(y(z)) = M_K(y^{\text{obs}}) \text{ for } z \in [z_t, z_{t+1}]} [z_t, z_{t+1}].$$

We call $z_1, z_2, \ldots, z_{T+1}$ breakpoints. We start from applying the SFS method to the response vector $y(z_1)$ and obtain the first triplet $(M^{(1)}_K, O^{(1)}_K, S^{(1)}_K)$. Then, the next breakpoint $z_2$ is obtained by (14) with $(M_K(y'), O_K(y'), S_K(y')) = (M^{(1)}_K, O^{(1)}_K, S^{(1)}_K)$. Then, the second triplet $(M^{(2)}_K, O^{(2)}_K, S^{(2)}_K)$ is obtained by applying the SFS method to the response vector $y(z_2 + \Delta z)$ where $\Delta z$ is a small value such that $z_t + \Delta z < z_{t+1}$ for all $t \in [T]$. This process is repeated until the next breakpoint becomes greater than $z_{\max}$.

### 3.3 Algorithm

In this section, we present the detail of the proposed method. In Algorithm 1, we apply $K$-step SFS method to the observed dataset $(X, y^{\text{obs}})$, and obtain the set of selected features $M_K(y^{\text{obs}})$. Then, for each feature $j \in M_K(y^{\text{obs}})$, we first compute the direction of interest $\eta$ by (3). Next, for each feature $j \in M_K(y^{\text{obs}})$, we compute the truncation region $Z = \{z \in [z_{\min}, z_{\max}] | M_K(y(z)) = M_K(y^{\text{obs}})\}$ by the homotopy continuation. Note that $Z$ are different among different $j \in M_K(y^{\text{obs}})$ since the direction of interest $\eta$ depends
Algorithm 1 SFS\_conditional\_SI

\textbf{Input:} $X, y^{\text{obs}}, K, [z_{\min}, z_{\max}]$

1: $M_K(y^{\text{obs}}) \leftarrow$ Applying $K$-step SFS algorithm to $(X, y^{\text{obs}})$
2: \textbf{for} each selected feature $j \in M_K(y^{\text{obs}})$ \textbf{do}
3: \hspace{1em} Compute $\eta \leftarrow$ Equation (3)
4: \hspace{1em} Compute $a$ and $b \leftarrow$ Equation (9)
5: \hspace{1em} Truncation region $Z \leftarrow$ compute\_truncation\_region $(X, K, a, b, [z_{\min}, z_{\max}], M_K(y^{\text{obs}}))$
6: \hspace{1em} $p_{k,j}^{\text{selective}} \leftarrow$ Equation (6) (and/or selective confidence interval)
7: \textbf{end for}

\textbf{Output:} $\{p_{k,j}^{\text{selective}}\}_{j \in M_K(y^{\text{obs}})}$ (and/or selective confidence intervals)

Algorithm 2 compute\_truncation\_region

\textbf{Input:} $X, K, a, b, [z_{\min}, z_{\max}], M_{K}^{\text{obs}}$

1: Initialization: $t = 1, z_t = z_{\min}, Z = \emptyset$
2: \textbf{while} $z_t < z_{\max}$ \textbf{do}
3: \hspace{1em} $y(z_t + \Delta z) = a + b(z_t + \Delta z)$
4: \hspace{1em} $(M_K^{(t)}, O_K^{(t)}, S_K^{(t)}) \leftarrow$ Applying $K$-step SFS to $(X, y(z_t + \Delta z))$
5: \hspace{1em} $z_{t+1} \leftarrow$ Equation (14)
6: \hspace{1em} \textbf{if} $M_K^{(t)} = M_K(y^{\text{obs}})$ \textbf{then}
7: \hspace{2em} $Z \leftarrow Z \cup [z_t, z_{t+1}]$
8: \hspace{1em} \textbf{end if}
9: \hspace{1em} $t \leftarrow t + 1$
10: \textbf{end while}

\textbf{Output:} $Z = \{z \in [z_{\min}, z_{\max}] \mid M(y(z)) = M_K(y^{\text{obs}})\}$

on $j$. This task can be done by Algorithm 2. Finally, after having the truncation region $Z$, we can compute selective $p$-values and selective confidence intervals.

In Algorithm 2, multiple breakpoints are computed one by one. The algorithm is initialized at $z_t = z_{\min}, t = 1$. At each $z_t$, the task is to find the next breakpoint $z_{t+1}$, at which there is a change either in the set of selected features $M_K$ or orders $O_K$ or signs $S_K$. This step is repeated until $z_t > z_{\max}$.
3.4 Selecting the number of steps $K$ by cross-validation

In this section, we introduce a method for SI conditional also on the selection of the number of selected features $K$ via cross-validation. Consider selecting the number of steps $K$ in the SFS method from a given set of candidates $K = \{K_1, \ldots, K_L\}$ where $L$ is the number of candidates. Based on the cross-validation on the observed dataset $(X, y_{\text{obs}})$, suppose that $K(y_{\text{obs}}) \in K$ is selected as the best one. The test-statistic for the selected feature $j$ when applying the SFS method with $K(y_{\text{obs}})$ steps to $(X, y_{\text{obs}})$ is then defined as

$$\eta^\top y | \{M_{K(y_{\text{obs}})}(y) = M_{K(y_{\text{obs}})}(y_{\text{obs}}), V(y) = K(y_{\text{obs}}), q_{K(y_{\text{obs}})}(y) = q_{K(y_{\text{obs}})}(y_{\text{obs}})\}. \quad (15)$$

The conditional data space in (9) with the event of selecting $K$ is then written as

$$\mathcal{Y} = \{y(z) = a + bz \mid z \in Z_{\text{CV}}\}, \quad (16)$$

where $Z_{\text{CV}} = \{z \in \mathbb{R} \mid M_{K(y_{\text{obs}})}(y(z)) = M_{K(y_{\text{obs}})}(y_{\text{obs}}), V(y(z)) = K(y_{\text{obs}})\}$. The truncation region $Z_{\text{CV}}$ can be obtained by the intersection of the following two sets:

$$Z_1 = \{z \in \mathbb{R} \mid M_{K(y_{\text{obs}})}(y(z)) = M_{K(y_{\text{obs}})}(y_{\text{obs}})\}$$

$$Z_2 = \{z \in \mathbb{R} \mid V(y(z)) = K(y_{\text{obs}})\}.$$

Since the former $Z_1$ can be obtained by using the method described above, the remaining task is to identify the latter $Z_2$.

For notational simplicity, we consider the case where the dataset $(X, y)$ is divided into training and validation sets, and the latter is used for selecting $K(y_{\text{obs}})$. The following discussion can be easily extended to cross-validation scenario. Let us re-write

$$(X, y) = \{(X^{\text{tr}}, X^{\text{va}})\top \in \mathbb{R}^{n \times p}, (y^{\text{tr}}, y^{\text{va}})\top \in \mathbb{R}^n\}.$$ 

With a slight abuse of notation, for $K \in K$, let $M_K(y^{\text{tr}}(z))$ be the set of selected features by applying $K$-step SFS method to $(X^{\text{tr}}, y^{\text{tr}}(z))$. The validation error
is then defined as

\[ E_K(z) = \| y_{va}(z) - X_{\mathcal{M}_K(y^{tr}(z))}^{va} \hat{\beta}_K(z) \|_2^2, \]  

(17)

where \( \hat{\beta}_K(z) = \left( X_{\mathcal{M}_K(y^{tr}(z))}^{tr} X_{\mathcal{M}_K(y^{tr}(z))}^{tr} \right)^{-1} X_{\mathcal{M}_K(y^{tr}(z))}^{tr} y^{tr}(z) \). Then, we can write

\[ Z_2 = \{ z \in \mathbb{R} \mid E_K(y^{obs})(z) \leq E_K(z) \text{ for any } K \in \mathcal{K} \}. \]

Since the validation error \( E_K(z) \) in (17) is a piecewise-quadratic function of \( z \), we have a corresponding piecewise-quadratic function of \( z \) for each \( K \in \mathcal{K} \). The truncation region \( Z_2 \) can be identified by the intersection of the intervals of \( z \) in which the validation error \( E_K(y^{obs})(z) \) corresponding to \( K(y^{obs}) \) is minimum among a set of piecewise-quadratic functions for all the other \( K \in \mathcal{K} \).

Loftus [42] already discussed that it is possible to consider cross-validation event into conditional SI framework. However, his method is highly over-conditioned in the sense that in each single run of a SFS method in the entire process of cross-validation, extra-conditioning on orders and signs are required. Our method described above is minimumly-conditioned SI in the sense that our inference is conducted based exactly on the conditional sampling distribution of the test-statistic in (17) without any extra conditions.

4 Experiment

In the section, we demonstrate the performance of the proposed homotopy-based SI. We executed the experiment on Intel(R) Xeon(R) CPU E5-2687W v3 @ 3.10GHz. We show the false positive rates (FPRs), true positive rates (TPRs) and confidence intervals (CIs) for the following cases of conditional SI:

- **Active**: only conditioning on the active set.
- **ActiveSign**: conditioning on the active set and signs.
- **ActiveOrder**: conditioning on the active set and the sequential order that the feature enters the active set.
ActiveSignOrder: conditioning on the active set, signs, and sequential order, which is exactly same as the polytope-based SI in [2].

We also show the FPRs, TPRs, and CIs of data splitting (DS) method, which is the commonly used procedure for the purpose of selection bias correction. In this approach, the data is randomly divided in two halves — first half used for model selection and the other for inference.

We set the significance level $\alpha = 0.05$ for all the experiments. We generated $n$ outcomes as $y_i = x_i^T \beta + \varepsilon_i$, $i = 1, \ldots, n$, where $x_i \sim \mathcal{N}(0, I_p)$ in which $\varepsilon_i \sim \mathcal{N}(0, 1)$. We set $p = 5$, $K = 3$ for FPR and TPR experiments, and $\beta$ was respectively set to $[0, 0, 0, 0, 0]^T$ and $[0.25, 0.25, 0, 0, 0]^T$. We ran 100 trials and we repeated this experiments 20 times. For CI experiments, we set $n = 100$, $p = 10$, $K = 9$ and $\beta = [0.25, 0.25, 0.25, 0.25, 0, 0, 0, 0, 0]^T$.

The results of FPR and TPR are shown in Figure 2a and 2b. In all five cases, the FPRs are properly controlled under the significance level $\alpha$. Regarding the TPR comparison, it is obvious that Active has the highest power since we only condition on the selected features. Figure 2c shows the demonstration of CI for each selected feature. Note that we only show commonly selected features among all five cases of conditional inference. The lengths of CI obtained by Active are almost the shortest. We repeated this experiment 100 times and showed the dot plot of the lengths of the confidence intervals in Figure 2d. In summary, the CI results are consistent with the TPR results. In other words, the shortest length of CI in the case of Active indicates that Active has the highest power.

Especially, we also demonstrate the TPRs and the CIs between the cases when $K = 9$ is fixed and $K$ is selected from the set $\mathcal{K}_1 = \{3, 6, 9\}$, or $\mathcal{K}_2 = \{1, 2, \ldots, 10\}$ using 5-fold cross-validation. We set $p = 10$, only the first elements of $\beta$ was set to 0.25, and all the rest were set to 0. We show that the TPR tends to decrease when increasing the size of $\mathcal{K}$ in Figure 3. This is due to the fact that when we increase the size of $\mathcal{K}$, we have to condition on more information which leads to shorter truncation interval and results low TPR. The TPR results are consistent with the CI results shown in Figure 4 in which the length of CI is
longer when increasing the size of $K$.

Next, we demonstrate the computational efficiency of the proposed method. We show the result of comparing the computational time between the proposed method and the existing method using $n = 50$, $p = 10$ artificial dataset in Figure 5. For the existing studies, if we want to keep high statistical power, we have to consider a huge number of all the signs and order patterns $2^K \times K!$, which is unrealistic. With the proposed method, we are able to significantly reduce the computational cost while keeping high power.

We also show the violin plot of the actual number of interval of the test statistic $z$ that involves in the construction of truncated sampling distribution using $n = 250$, $p = 50$ artificial dataset in Figure 6. In this case, the number
of polytopes intersecting the line $z$ that we need to consider is larger than in the Lasso case because we have conditioned not only the signs but also the order to derive each polytope, but it is much smaller than $2^K \times K!$. When $K$ is large, because the ordering constraint is too strict, so the number of intervals of
Figure 5: The result of comparing the computational time between the proposed method and the existing method using $n = 50$, $p = 10$ artificial dataset.

Figure 6: The number of polytopes intersecting the line $z$ that we need to consider. The solid lines are shown the sample averages.

$z$ satisfying the condition including the order (ActiveOrder) becomes very small even after removing the sign constraint.
Figure 7: The computational time for using the proposed method on $n = 300$, $p = 300$ artificial dataset.

Table 1: The computational time for using the proposed method on the high dimensional real-world bioinformatics related datasets.

|          | $n$ | $p$   | $k$ | time (sec) |
|----------|-----|-------|-----|------------|
| Dataset 1| 89  | 5787  | 50  | 799.13     |
| Dataset 2| 76  | 5144  | 50  | 572.34     |
| Dataset 3| 133 | 5787  | 50  | 1709.49    |

We also show the computational time for using the proposed method on $n = 300$, $p = 300$ artificial dataset and the high dimensional real-world bioinformatics related datasets in Figure 7 and Table 1. The real-world datasets are available at [http://www.coepra.org/CoEPrA_regr.html](http://www.coepra.org/CoEPrA_regr.html).

5 Conclusion

In this paper, we proposed a more powerful and general conditional selective inference method for stepwise feature selection method. We resolve the over-conditioning issue in existing approach by introducing homotopy continuation
approach. The experimental results indicate that the proposed homotopy-based approach is more powerful and computationally efficient.

Acknowledgement

This work was partially supported by MEXT KAKENHI (20H00601, 16H06538), JST CREST (JPMJCR1502), RIKEN Center for Advanced Intelligence Project, and RIKEN Junior Research Associate Program.

References

[1] Jason D Lee, Dennis L Sun, Yuekai Sun, Jonathan E Taylor, et al. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(3):907–927, 2016.

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

[3] William Fithian, Dennis Sun, and Jonathan Taylor. Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*, 2014.

[4] Yoav Benjamini and Daniel Yekutieli. False discovery rate–adjusted multiple confidence intervals for selected parameters. *Journal of the American Statistical Association*, 100(469):71–81, 2005.

[5] Hannes Leeb and Benedikt M Pötscher. Model selection and inference: Facts and fiction. *Econometric Theory*, pages 21–59, 2005.

[6] Hannes Leeb, Benedikt M Pötscher, et al. Can one estimate the conditional distribution of post-model-selection estimators? *The Annals of Statistics*, 34(5):2554–2591, 2006.
[7] Yoav Benjamini, Ruth Heller, and Daniel Yekutieli. Selective inference in complex research. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 367(1906):4255–4271, 2009.

[8] Benedikt M Pötscher, Ulrike Schneider, et al. Confidence sets based on penalized maximum likelihood estimators in gaussian regression. *Electronic Journal of Statistics*, 4:334–360, 2010.

[9] Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, Linda Zhao, et al. Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837, 2013.

[10] Richard Lockhart, Jonathan Taylor, Ryan J Tibshirani, and Robert Tibshirani. A significance test for the lasso. *Annals of statistics*, 42(2):413, 2014.

[11] Jonathan Taylor, Richard Lockhart, Ryan J Tibshirani, and Robert Tibshirani. Post-selection adaptive inference for least angle regression and the lasso. *arXiv preprint arXiv:1401.3889*, 354, 2014.

[12] William Fithian, Jonathan Taylor, Robert Tibshirani, and Ryan Tibshirani. Selective sequential model selection. *arXiv preprint arXiv:1512.02565*, 2015.

[13] Yunjin Choi, Jonathan Taylor, Robert Tibshirani, et al. Selecting the number of principal components: Estimation of the true rank of a noisy matrix. *The Annals of Statistics*, 45(6):2590–2617, 2017.

[14] Xiaoqing Tian, Jonathan Taylor, et al. Selective inference with a randomized response. *The Annals of Statistics*, 46(2):679–710, 2018.

[15] Shuxiao Chen and Jacob Bien. Valid inference corrected for outlier removal. *Journal of Computational and Graphical Statistics*, pages 1–12, 2019.

[16] Sangwon Hyun, Kevin Lin, Max G’Sell, and Ryan J Tibshirani. Post-selection inference for changepoint detection algorithms with application to copy number variation data. *arXiv preprint arXiv:1812.03644*, 2018.
[17] Joshua R Loftus and Jonathan E Taylor. A significance test for forward stepwise model selection. *arXiv preprint arXiv:1405.3920*, 2014.

[18] Joshua R Loftus and Jonathan E Taylor. Selective inference in regression models with groups of variables. *arXiv preprint arXiv:1511.01478*, 2015.

[19] Snigdha Panigrahi, Jonathan Taylor, and Asaf Weinstein. Bayesian post-selection inference in the linear model. *arXiv preprint arXiv:1605.08824*, 28, 2016.

[20] Fan Yang, Rina Foygel Barber, Prateek Jain, and John Lafferty. Selective inference for group-sparse linear models. In *Advances in Neural Information Processing Systems*, pages 2469–2477, 2016.

[21] Shinya Suzumura, Kazuya Nakagawa, Yuta Umezu, Koji Tsuda, and Ichiro Takeuchi. Selective inference for sparse high-order interaction models. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3338–3347. JMLR. org, 2017.

[22] Kosuke Tanizaki, Noriaki Hashimoto, Yu Inatsu, Hidekata Hontani, and Ichiro Takeuchi. Computing valid p-values for image segmentation by selective inference. 2020.

[23] Vo Nguyen Le Duy, Hiroki Toda, Ryota Sugiyama, and Ichiro Takeuchi. Computing valid p-value for optimal changepoint by selective inference using dynamic programming. *arXiv preprint arXiv:2002.09132*, 2020.

[24] Vo Nguyen Le Duy, Shogo Iwazaki, and Ichiro Takeuchi. Quantifying statistical significance of neural network representation-driven hypotheses by selective inference. *arXiv preprint arXiv:2010.01823*, 2020.

[25] Yoshikazu Terada and Hidetoshi Shimodaira. Selective inference after variable selection via multiscale bootstrap. *arXiv preprint arXiv:1905.10573*, 2019.
[26] Keli Liu, Jelena Markovic, and Robert Tibshirani. More powerful post-selection inference, with application to the lasso. *arXiv preprint arXiv:1801.09037*, 2018.

[27] Vo Nguyen Le Duy and Ichiro Takeuchi. Parametric programming approach for powerful lasso selective inference without conditioning on signs. *arXiv preprint arXiv:2004.09749*, 2020.

[28] Michael R Osborne, Brett Presnell, and Berwin A Turlach. A new approach to variable selection in least squares problems. *IMA journal of numerical analysis*, 20(3):389–403, 2000.

[29] B. Efron and R. Tibshirani. Least angle regression. *Annals of Statistics*, 32(2):407–499, 2004.

[30] T. Hastie, S. Rosset, R. Tibshirani, and J. Zhu. The entire regularization path for the support vector machine. *Journal of Machine Learning Research*, 5:1391–415, 2004.

[31] S. Rosset and J. Zhu. Piecewise linear regularized solution paths. *Annals of Statistics*, 35:1012–1030, 2007.

[32] F. R. Bach, D. Heckerman, and E. Horvits. Considering cost asymmetry in learning classifiers. *Journal of Machine Learning Research*, 7:1713–41, 2006.

[33] K. Tsuda. Entire regularization paths for graph data. In *In Proc. of ICML 2007*, pages 919–925, 2007.

[34] G. Lee and C. Scott. The one class support vector machine solution path. In *Proc. of ICASSP 2007*, pages II521–II524, 2007.

[35] I. Takeuchi, K. Nomura, and T. Kanamori. Nonparametric conditional density estimation using piecewise-linear solution path of kernel quantile regression. *Neural Computation*, 21(2):539–559, 2009.
[36] Ichiro Takeuchi and Masashi Sugiyama. Target neighbor consistent feature weighting for nearest neighbor classification. In *Advances in neural information processing systems*, pages 576–584, 2011.

[37] M. Karasuyama and I. Takeuchi. Nonlinear regularization path for quadratic loss support vector machines. *IEEE Transactions on Neural Networks*, 22(10):1613–1625, 2010.

[38] T. Hocking, J. P. Vert, F. Bach, and A. Joulin. Clusterpath: an algorithm for clustering using convex fusion penalties. In *Proceedings of the 28th International Conference on Machine Learning*, pages 745–752, 2011.

[39] M. Karasuyama, N. Harada, M. Sugiyama, and I. Takeuchi. Multi-parametric solution-path algorithm for instance-weighted support vector machines. *Machine Learning*, 88(3):297–330, 2012.

[40] Kohei Ogawa, Motoki Imamura, Ichiro Takeuchi, and Masashi Sugiyama. Infinitesimal annealing for training semi-supervised support vector machines. In *International Conference on Machine Learning*, pages 897–905, 2013.

[41] Ichiro Takeuchi, Tatsuya Hongo, Masashi Sugiyama, and Shinichi Nakajima. Parametric task learning. *Advances in Neural Information Processing Systems*, 26:1358–1366, 2013.

[42] Joshua R Loftus. Selective inference after cross-validation. *arXiv preprint arXiv:1511.08866*, 2015.