Selective Inference for High-order Interaction Features Selected in a Stepwise Manner

SHINYA SUZUMURA^{1,a)} KAZUYA NAKAGAWA^{1,b)} YUTA UMEZU^{1,c)} KOJI TSUDA^{2,3,d)}
 ICHIRO TAKEUCHI^{1,3,e)}

Received: August 13, 2020, Accepted: November 19, 2020

Abstract: In this paper, we study a stepwise feature selection algorithm for a high-order interaction model and we propose a new statistical inference for selected high-order interaction features. Feature selection and statistical inference for high-order interaction features are challenging tasks because the possible number of those interactions is extremely large. Our main contribution is to extend recently developed selective inference framework to high-order interaction model by developing a pruning technique for searching over tree which represents high-order interaction features. We demonstrate the effectiveness of the proposed approach by applying it to several synthetic problems and an HIV drug resistance prediction problem.

Keywords: stepwise feature selection, high-order interaction model, selective inference, tree pruning algorithm

1. Introduction

In this paper, we consider a stepwise feature selection algorithm for a high-order interaction model that has r -th order interactions of multiple features, and we propose a new statistical inference for selected high-order interaction features. Feature selection and statistical inference for high-order interaction features are important tasks. For example, in a biomedical study, co-occurrence of multiple mutations in multiple genes may have a significant influence on a response to a drug even if occurrence of single mutation in each of these genes has no influence [1], [2]. In high-order interaction model, a difficulty of these tasks is that there is a huge number of possible combinations of multiple features. If one has a dataset with d original features and takes into account interactions up to order r , the model has $D = \sum_{\rho=1}^r \binom{d}{\rho}$ features (e.g., for $d = 10,000$, $r = 5$, $D > 10^{17}$). Feature selection and statistical inference in such an extremely high-dimensional model are challenging both computationally and statistically.

A common approach to high-dimensional modeling is to consider a sparse model, i.e., a model only with a selected subset of features. In the past two decades, considerable amount of studies have been done on sparse modeling and feature selection in high-dimensional models. In these studies, a variety of

feature selection algorithms such as *marginal screening* [3], *orthogonal matching pursuit* [4], LASSO [5], and their various extensions have been developed. On the other hand, statistical inference for sparse models (hypothesis testing or confidence interval computation of the fitted coefficients) have not been deeply studied until recently. After the seminal work by Ref. [6], significant progress has been made on statistical inference for sparse linear models [6], [7], [8], [9], [10], [11], [12], [13], [14], and these approaches are sometimes called *Selective Inference* or *Post-Selection Inference*. The basic idea of selective inference is to consider the sampling distribution of test statistic after feature selection, where the sampling distribution is a conditional distribution which is characterized by selected and unselected features. By considering such a distribution, we can eliminate the *selection bias* [15] that arises in feature selection. In the following paragraph, we briefly discuss selection bias in testing selected features.

A simple illustration of selection bias is depicted in **Fig. 1**. In this example, we generated several images whose color was randomly generated based on the standard normal distribution $N(0, 1^2)$. By using a selection algorithm, we selected a region which is composed of 3×3 pixels, and we considered a naive hypothesis testing that the expected average color of pixels on the selected region is zero or not (the detail will be described later). Although each color of pixels was randomly generated, the color level on the selected region tends to be higher, and *type I error* could not be controlled under the desired significance level. The detail of the hypothesis testing is shown as follows. Let τ_j be the average color of pixels on the j -th selected region in an image. Since each color of pixels was randomly generated by using $N(0, 1^2)$, one expects that the sampling distribution of value τ_j is

¹ Nagoya Institute of Technology, Nagoya, Aichi 466–8555, Japan

² University of Tokyo, Kashiwa, Chiba 277–8568, Japan

³ RIKEN Center for Advanced Intelligence Project, Chuo, Tokyo 103–0027, Japan

a) suzumura.mllab.nit@gmail.com

b) nakagawa.k.mllab.nit@gmail.com

c) umezu.yuta@nagasaki-u.ac.jp

d) tsuda@k.u-tokyo.ac.jp

e) takeuchi.ichiro@nitech.ac.jp

also $N(0, 1^2)$. The classical z-test $H_{j,0} : \mathbb{E}[\tau_j] = 0, s \sim N(0, 1^2)$ is totally valid when j is completely randomly selected. On the other hand, $\mathbb{E}[\tau_j] \neq 0$ when j is selected based on observations of generated image. It is because the region would NOT be selected when the average value τ_j is relatively smaller than the others. This indicates that the distribution of value τ_j after selection is no longer normal distribution $N(0, 1^2)$, but rather a conditional distribution which is conditioned by the selected and unselected regions. To remove the bias, it is necessary to analyze the sampling space of the test statistic after selection.

Our Contributions

Our main contribution is to develop a selective inference procedure for selected features in high-order interaction model. In high-order interaction model, there is a serious computational problem because the sampling distribution depends not only on selected features but also on an extremely large number of unselected features. For circumventing the computational issue, we consider a tree structure among features and derive a novel pruning condition that enables us to efficiently identify a set of features which have no effect on the sampling distribution.

1.1 Outline

Here is the outline of this paper. Section 2 presents problem formulation, illustrative example, formal description of selective inference, and a brief review of recent selective inference literature. Section 3 describes our main contribution, where we develop a method that enables selective inference in extremely high-dimensional settings. Section 4 discusses extensions for computational efficiency and statistical power of selective inference. Section 5 covers numerical experiments for demonstrating the ad-

vantage of selective inference framework. Section 6 concludes the paper. The early version of this paper is presented in a conference [16].

2. Preliminaries

2.1 Notations

We use the following notations in the remainder. For any natural numbers n and d , we denote the set of indices as $[n] = \{1, \dots, n\}$, and we denote a vector and a matrix as $\mathbf{y} \in \mathbb{R}^n$ and $\mathbf{Z} \in \mathbb{R}^{n \times d}$, respectively. Furthermore, the i -th row and j -th column vectors of matrix $\mathbf{Z} \in \mathbb{R}^{n \times d}$ are written as $\mathbf{z}_{i:} \in \mathbb{R}^{1 \times d}$, $\mathbf{z}_{:j} \in \mathbb{R}^{n \times 1}$ respectively, and (i, j) -th element of matrix $\mathbf{Z}$ is written as z_{ij} . For a set of indices $S \subseteq [d]$, we denote a sub-matrix as $\mathbf{Z}_S \in \mathbb{R}^{n \times |S|}$ whose column vectors are composed of S columns of the matrix $\mathbf{Z}$.

2.2 Problem Statement for High-order Interaction Model

Let $\mathbf{z}_{i:} = [z_{i1}, \dots, z_{id}] \in [0, 1]^{1 \times d}$ be d -dimensional original features with i -th sample index. We consider the following high-order interaction model up to r -th order

$$f(\mathbf{z}_{i:}) = \sum_{j_1 \in [d]} \alpha_{j_1} z_{ij_1} + \sum_{\substack{(j_1, j_2) \in [d] \times [d] \\ j_1 \neq j_2}} \alpha_{j_1, j_2} z_{ij_1} z_{ij_2} + \dots + \sum_{\substack{(j_1, \dots, j_r) \in [d]^r \\ j_1 \neq \dots \neq j_r}} \alpha_{j_1, \dots, j_r} z_{ij_1} \dots z_{ij_r}, \quad (1)$$

where α_s are the coefficients, and the number of coefficients is $D = \sum_{j=1}^r \binom{d}{j}$. For notation simplicity, we write the high-order interaction model (1) as the following linear model

$$f(\mathbf{x}_i; \boldsymbol{\beta}) = \mathbf{x}_i \boldsymbol{\beta} = \beta_1 x_{i1} + \dots + \beta_D x_{iD}, \quad (2)$$

where $\beta_1, \dots, \beta_D$ are D coefficients corresponding to $\alpha_{j_1}, \dots, \alpha_{j_1, \dots, j_r}$ in Eq. (1), and $\mathbf{x}_i$ is D -dimensional high-order interaction features which is computed by multiplying original features. Although the bias term in Eq. (2) can be considered by appending an unit vector $\mathbf{x}_{:0} = \mathbf{1}$ and a coefficient β_0 , we omit the explicit notation.

Our goal is to select high-order interaction features which are highly associated with the n -sample continuous responses $\mathbf{y} = (y_1, \dots, y_n)^\top \in \mathbb{R}^n$, and provide the statistical inference of the association for each of the selected features. The original training set is denoted as $(\mathbf{Z}, \mathbf{y}) \in [0, 1]^{n \times d} \times \mathbb{R}^n$, while the expanded training set is written as $(\mathbf{X}, \mathbf{y}) \in [0, 1]^{n \times D} \times \mathbb{R}^n$. The short description of symbols is summarized in Table 1.

2.3 Sparse High-order Interaction Model

Since the number of all features D in the high-order interaction

Fig. 1 A simple demonstration of selection bias. In this example, we randomly generated $5 \times 5, \dots, 100 \times 100$ images whose color was generated based on normal distribution $N(0, 1^2)$. In each image, we selected the region (which is composed of 3×3 pixels) whose average value is the highest compared with the other regions in an image. We observed that the color of selected region looks “significantly higher” when it is selected from larger image. This indicates that the selection bias depends on both selected and unselected regions. The bottom plot shows the frequencies of the false positive findings for various sizes of images, in which we applied a naive hypothesis testing where the null hypothesis is that the average value on the selected region is equals to 0, with desired 5% significant level.

Table 1 The short description of symbols.

Symbol	Description
n	The number of samples
d	The number of original features
D	The number of high-order interaction features
r	The maximum order of interactions
$\mathbf{Z} \in [0, 1]^{n \times d}$	Original design matrix
$\mathbf{X} \in [0, 1]^{n \times D}$	Design matrix for high-order interaction features
$\mathbf{y} \in \mathbb{R}^n$	Response vector

model is enormous, we consider a sparse model by using a forward greedy feature selection algorithm called Orthogonal Matching Pursuit (OMP) [4]. Let $S^{(t)} \subseteq [D]$ be the set of indices for selected features at feature selection step $t \geq 1$. In OMP, each feature is selected by

$$\begin{aligned} S^{(t)} &= \{j^{(t)} \cup S^{(t-1)}\}, \\ j^{(t)} &= \arg \max_{j \in [D] \setminus S^{(t-1)}} |\mathbf{x}_{:,j}^\top \mathbf{r}^{(t)}|, \end{aligned} \quad (3)$$

where $S^{(0)} = \emptyset$ and $\mathbf{r}^{(1)} = \mathbf{y}$. The (residual) vector $\mathbf{r}^{(t)}$ is

$$\mathbf{r}^{(t)} = \mathbf{y} - X_{S^{(t-1)}} \hat{\boldsymbol{\beta}}_{S^{(t-1)}}, \quad (4)$$

and $\hat{\boldsymbol{\beta}}_{S^{(t-1)}}$ is the least square estimator $\hat{\boldsymbol{\beta}}_{S^{(t-1)}} = X_{S^{(t-1)}}^+ \mathbf{y}$ where the super script + means pseudo inverse. The feature selection step t is sequentially incremented.

In this paper, we do not care about the normalization of $\mathbf{x}_{:,j}$ for each $j \in [D]$, because high-order interaction features tend to be very sparse when original features are defined in $[0, 1]$. Moreover, we do not care about the centralization of features but it can be considered when the residual vector is centralized as $\mathbf{1}^\top \mathbf{r}^{(t)} = 0$ because the inner product will be shift invariant $\mathbf{x}_{:,j}^\top \mathbf{r}^{(t)} = (\mathbf{x}_{:,j} + c\mathbf{1})^\top \mathbf{r}^{(t)}$ for any scalar parameter c .

2.4 Selective Inference for Sparse Linear Model

In this section, we review statistical inference after feature selection, which is based on selective inference recently developed by Ref. [6]. To formalize the problem, we consider the following response which follows the multivariate normal distribution

$$\mathbf{y} \sim N(\boldsymbol{\mu}, \sigma^2 I_n), \quad (5)$$

where $\boldsymbol{\mu}$ is unknown mean, σ^2 is a known variance, and I_n is n sample identity matrix. We model the expected value of response as $\boldsymbol{\mu} = X_{S^{(t)}} \boldsymbol{\beta}_{S^{(t)}}$, where $X_{S^{(t)}}$ is the fixed design matrix at feature selection step t , and $\boldsymbol{\beta}_{S^{(t)}}$ is the best linear predictor, that is

$$\boldsymbol{\beta}_{S^{(t)}} = \arg \min_{\boldsymbol{\beta}_{S^{(t)}} \in \mathbb{R}^D} \|\mathbf{y} - X_{S^{(t)}} \boldsymbol{\beta}_{S^{(t)}}\|_2^2 = X_{S^{(t)}}^+ \mathbf{y}.$$

Lee et al. [6] suggested to consider a family of intervals $\{C_{S^{(t)},j}\}_{j \in S^{(t)}}$ that have conditional $(1 - \alpha)$ coverage:

$$\Pr(\boldsymbol{\beta}_{S^{(t)},j} \in C_{S^{(t)},j} | S^{(t)}) \geq 1 - \alpha. \quad (6)$$

If Eq. (6) holds, then, FCR (false coverage-statement rate) [17] and pFCR (positive false coverage-statement rate) [18] can be controlled (the proof is shown at Lemma 2.1 of Ref. [6]) as

$$\begin{aligned} \text{FCR} &= \mathbb{E} \left[\frac{|\{j \in S^{(t)} : \boldsymbol{\beta}_{S^{(t)},j} \notin C_{S^{(t)},j}\}|}{|S^{(t)}|}; |S^{(t)}| > 0 \right] \\ &\leq \alpha, \\ \text{pFCR} &= \mathbb{E} \left[\frac{|\{j \in S^{(t)} : \boldsymbol{\beta}_{S^{(t)},j} \notin C_{S^{(t)},j}\}|}{|S^{(t)}|} \middle| |S^{(t)}| > 0 \right] \\ &\leq \alpha. \end{aligned}$$

In this paper, our interest is to construct the conditional intervals shown in Eq. (6) for each of the selected high-order interaction features. Our interest is not in evaluating the correctness of

the selected model and we do not discuss how to terminate the feature selection step t in OMP. Lee et al. [6] showed that valid confidence intervals can be computed even if the selected model is wrong, but the requirement is the normality of error as shown in Eq. (5). On the other hand, the selective inference for determining the feature selection step t is studied in Ref. [8], where the interest is to check whether the selected model is correct or not.

2.4.1 Basic Idea of Selective Inference

Selective inference is developed for two stage methods, where a subset of features is *selected* in the first stage, and *inferences* for the selected features are made in the second stage. A key finding by Ref. [6] is that, if the first selection stage is described as a *linear selection event*, then exact (non-asymptotic) statistical inference for the fitted coefficients conditional on the selection event can be done.

2.4.2 Feature Selection Stage

Suppose that, in the first feature selection stage, a subset of features $S^{(t)} = \{j^{(1)}, \dots, j^{(t)}\} \subseteq [D]$ are selected, where $j^{(t)}$ is defined by Eq. (3). The selective inference framework in Ref. [6] can be applied to feature selection algorithms whose selection process can be characterized by a set of linear inequalities in the form of $A\mathbf{y} \leq \mathbf{b}$ with a certain matrix A and a certain vector $\mathbf{b}$ that do not depend on $\mathbf{y}$. This type of selection event is called a *linear selection event*. In the selective inference framework, inferences are made conditional on the selection event, it means that, in the case of a linear selection event, we only care about the cases where $\mathbf{y}$ is observed in a polytope $\text{Pol}(S^{(t)}) = \{\mathbf{y} \in \mathbb{R}^n \mid A\mathbf{y} \leq \mathbf{b}\}$. In Refs. [6], [19] and [20], some feature selection procedures are shown to be linear selection events.

As pointed out in Refs. [19], [20] the selection process of OMP is a linear selection event, i.e., characterized by a set of linear constraints. Let $S^{(t)}$ and $\bar{S}^{(t)}$ be the set of indices for selected and unselected features, respectively at feature selection step $t \in [k]$ for a fixed natural number k . It indicates that $S^{(1)} \subseteq S^{(2)} \subseteq \dots \subseteq S^{(k)}$ and $\bar{S}^{(1)} \supseteq \bar{S}^{(2)} \supseteq \dots \supseteq \bar{S}^{(k)}$. Moreover, let $j^{(t)}$ be the index of t -th selected feature, which is defined by Eq. (3). The selection event of OMP is rewritten by the following constraints

$$\forall j' \in \bar{S}^{(t)}, t \in [k], |\mathbf{x}_{:,j^{(t)}}^\top \mathbf{r}^{(t)}| \geq |\mathbf{x}_{:,j'}^\top \mathbf{r}^{(t)}|. \quad (7)$$

These constraints are rephrased by

$$\begin{aligned} &\forall j' \in \bar{S}^{(t)}, t \in [k], \\ &\text{if } \mathbf{x}_{:,j^{(t)}}^\top \mathbf{r}^{(t)} > 0, \text{ then} \\ &\quad (-\mathbf{x}_{:,j^{(t)}} - \mathbf{x}_{:,j'})^\top \mathbf{r}^{(t)} \leq 0, \\ &\quad (-\mathbf{x}_{:,j^{(t)}} + \mathbf{x}_{:,j'})^\top \mathbf{r}^{(t)} \leq 0, \\ &\text{if } \mathbf{x}_{:,j^{(t)}}^\top \mathbf{r}^{(t)} \leq 0, \text{ then} \\ &\quad (\mathbf{x}_{:,j^{(t)}} - \mathbf{x}_{:,j'})^\top \mathbf{r}^{(t)} \leq 0, \\ &\quad (\mathbf{x}_{:,j^{(t)}} + \mathbf{x}_{:,j'})^\top \mathbf{r}^{(t)} \leq 0. \end{aligned}$$

By using $s_{j^{(t)}} = \text{sign}(\mathbf{x}_{:,j^{(t)}}^\top \mathbf{r}^{(t)})$, these constraints are summarized as

$$\begin{aligned} &\forall j' \in \bar{S}^{(t)}, t \in [k], \\ &\quad (-s_{j^{(t)}} \mathbf{x}_{:,j^{(t)}} - \mathbf{x}_{:,j'})^\top \mathbf{r}^{(t)} \leq 0, \end{aligned}$$

$$\begin{aligned} (-s_{j^{(t)}} \mathbf{x}_{:j^{(t)}} + \mathbf{x}_{:j'})^\top \mathbf{r}^{(t)} &\leq 0, \\ -s_{j^{(t)}} \mathbf{x}_{:j^{(t)}}^\top \mathbf{r}^{(t)} &\leq 0, \end{aligned}$$

or equivalently

$$\begin{aligned} \forall(j', \xi) \in \bar{S}^{(t)} \times \{-1, 1, 0\}, t \in [k], \\ (-s_{j^{(t)}} \mathbf{x}_{:j^{(t)}} + \xi \mathbf{x}_{:j'})^\top \mathbf{r}^{(t)} &\leq 0. \end{aligned}$$

Since the residual vector (4) is $\mathbf{r}^{(t)} = (I_n - X_{S^{(t-1)}} X_{S^{(t-1)}}^\top) \mathbf{y}$, the constraints are rephrased by the set of linear inequalities in the form of $A\mathbf{y} \leq \mathbf{b}$ as

$$\begin{aligned} \forall(j', \xi) \in \bar{S}^{(t)} \times \{-1, 1, 0\}, t \in [k], \\ (-s_{j^{(t)}} \mathbf{x}_{:j^{(t)}} + \xi \mathbf{x}_{:j'})^\top \Gamma^{(t)} \mathbf{y} &\leq 0, \end{aligned} \quad (8)$$

where $\Gamma^{(t)} = (I_n - X_{S^{(t-1)}} X_{S^{(t-1)}}^\top)$ and $\Gamma^{(1)} = I_n$.

2.4.3 Statistical Inference Stage

In the remainder, we use the notation $S = S^{(k)}$ with a certain fixed natural number k . If we consider the case where S is NOT selected from the data, i.e., independent of $\mathbf{y}$, then the sampling distribution of each fitted coefficient $\hat{\beta}_{S,j} = (X_S^+ \mathbf{y})_j$ is

$$\hat{\beta}_{S,j} \sim N(\beta_{S,j}, \sigma_{S,j}^2), \text{ where } \sigma_{S,j}^2 = \sigma^2 (X_S^\top X_S)^{-1}_{jj}. \quad (9)$$

If we define $\ell_{\alpha/2}^{(j)}$ and $u_{\alpha/2}^{(j)}$ to be the lower and upper $\alpha/2$ points of the sampling distribution in Eq. (9), then the *type I error* at level α is controlled as

$$\Pr(\beta_{S,j} \notin [\ell_{\alpha/2}^{(j)}, u_{\alpha/2}^{(j)}]) \leq \alpha. \quad (10)$$

On the other hand, after feature selection, we would like to control the following *selective type I error*

$$\begin{aligned} \Pr(\beta_{S,j} \notin [\ell_{\alpha/2}^{(S,j)}, u_{\alpha/2}^{(S,j)}] \mid S) \\ = \Pr(\beta_{S,j} \notin [\ell_{\alpha/2}^{(S,j)}, u_{\alpha/2}^{(S,j)}] \mid \mathbf{y} \in \text{Pol}(S)) \leq \alpha, \end{aligned} \quad (11)$$

where the selection event is written as $\mathbf{y} \in \text{Pol}(S)$ in the case of a linear selection event. Reference [6] derived how to compute these confidence intervals as formally stated in the following lemma.

Lemma 1. Let $F_{\mu, \sigma^2}^{[L, U]}$ be the CDF of a truncated normal distribution with the mean μ , the variance σ^2 , and the truncation interval $[L, U]$, i.e.,

$$F_{\mu, \sigma^2}^{[L, U]}(x) = \frac{\Phi((x - \mu)/\sigma) - \Phi((L - \mu)/\sigma)}{\Phi((U - \mu)/\sigma) - \Phi((L - \mu)/\sigma)},$$

where Φ is the CDF of the standard normal distribution. Furthermore, let the truncation points be

$$L(S, j) = \hat{\beta}_{S,j} + \theta_L (X_S^\top X_S)^{-1}_{jj}, \quad (12a)$$

where $\theta_L = \min_{\theta \in \mathbb{R}} \theta$ s.t. $\mathbf{y} + \theta (X_S^+)^T \mathbf{e}_j \in \text{Pol}(S)$,

$$U(S, j) = \hat{\beta}_{S,j} + \theta_U (X_S^\top X_S)^{-1}_{jj}, \quad (12b)$$

where $\theta_U = \max_{\theta \in \mathbb{R}} \theta$ s.t. $\mathbf{y} + \theta (X_S^+)^T \mathbf{e}_j \in \text{Pol}(S)$,

where $\mathbf{e}_j$ is an unit vector, j -th element of which is one. If $\ell_{\alpha/2}^{(S,j)}$ and $u_{\alpha/2}^{(S,j)}$ are defined such that

$$F_{\ell_{\alpha/2}^{(S,j)}, \sigma_{S,j}^2}^{[L(S,j), U(S,j)]}(\hat{\beta}_{S,j}) = 1 - \alpha/2, \quad (13a)$$

Fig. 2 An illustration of polyhedral lemma. The polyhedron represents the selection event and truncation points can be computed by optimizing θ along with the direction $\boldsymbol{\eta} = (X_S^+)^T \mathbf{e}_j$.

$$F_{\ell_{\alpha/2}^{(S,j)}, \sigma_{S,j}^2}^{[L(S,j), U(S,j)]}(\hat{\beta}_{S,j}) = \alpha/2, \quad (13b)$$

then, the interval $[\ell_{\alpha/2}^{(S,j)}, u_{\alpha/2}^{(S,j)}]$ satisfies the conditional coverage property, i.e.,

$$\Pr(\beta_{S,j} \notin [\ell_{\alpha/2}^{(S,j)}, u_{\alpha/2}^{(S,j)}] \mid S) = \alpha.$$

The proof of Lemma 1 is presented in Appendix A.1.1 although it is easily proved by using the results in Ref. [6]. Lemma 1 indicates that the sampling distribution of each fitted coefficient $\hat{\beta}_{S,j} = (X_S^+ \mathbf{y})_j$ is a truncated normal distribution when the selection event is a linear selection event $\{A\mathbf{y} \leq \mathbf{b}\}$. **Figure 2** schematically illustrates that truncation points can be computed by optimizing θ along with the direction $\boldsymbol{\eta} = (X_S^+)^T \mathbf{e}_j$ since each fitted coefficient $\hat{\beta}_{S,j}$ is $\boldsymbol{\eta}^\top \mathbf{y}$.

Unfortunately, we cannot directly apply this selective inference framework to the high-order interaction model because the polytope $\text{Pol}(S)$ is characterized by extremely large number of linear inequalities, and the optimization problems in Eq. (12) are hard to solve.

3. Computational Tricks by Using Tree-based Feature Representation

The OMP feature selection (3) and the computation of truncation points (12) are computationally infeasible because the number of high-order interaction features D is extremely huge. For circumventing this issue, we introduce a tree-based representation of high-order interaction features shown in **Fig. 3** and consider the following computational pruning techniques in both feature selection and statistical inference stages.

3.1 Pruning Technique in Feature Selection Stage

Since a high-order interaction feature is the product of multiple original features whose value are defined in $[0, 1] \in \mathbb{R}$, the following property holds:

$$0 \leq x_{i\tilde{j}} \leq x_{ij}, \forall(i, \tilde{j}) \in [n] \times \text{Des}(j), \quad (14)$$

where $\text{Des}(j)$ is the set of indicates for descendant features of j -th parent feature. By using this, for any residual vector $\mathbf{r}^{(t)}$, we obtain

Fig. 3 (a) An example of high-order interaction features represented by a tree structure with a fixed sample index i , where the number of original features $d = 3$ and the maximum interaction order $r = 2$. In the tree, each node can be efficiently computed by using a tree (or graph) search technique such as DFS (depth first search). This tree originally has a lattice structure, e.g., the node $z_{i1}z_{i2}$ has two edges from z_{i1} and z_{i2} , but we remove the edges for simplicity. (b) Another tree-based feature representation by using high-order interaction features.

$$|\mathbf{x}_{:j}^{\top} \mathbf{r}^{(t)}| \leq \max \left\{ \sum_{i: r_i^{(t)} > 0} x_{ij} r_i^{(t)}, - \sum_{i: r_i^{(t)} < 0} x_{ij} r_i^{(t)} \right\}. \quad (15)$$

When we search over the tree, if the right-hand side of Eq. (15) is smaller than the current largest correlation in Eq. (3), then, we can quit searching over its descendant nodes $\tilde{j} \in \text{Des}(j)$. This pruning technique has been widely used in item-set or sub-graph mining, e.g., Refs. [21], [22], [23], [24].

3.2 Proposed Pruning Technique in Statistical Inference Stage

In the following lemma, we show the solutions of the optimization problems (12).

Lemma 2. For $\boldsymbol{\eta} = (\mathbf{X}_S^+)^{\top} \mathbf{e}_j$, the solutions of the optimization problems in Eq. (12) are respectively written as

$$\theta_L = \max_{t \in [k]} \max_{\substack{(j', \xi) \in \bar{S}^{(t)} \times \{-1, 1, 0\}, \\ (\xi \mathbf{x}_{:j'} - s_{j^{(0)}} \mathbf{x}_{:j^{(0)}})^{\top} \Gamma^{(t)} \boldsymbol{\eta} < 0}} \frac{(s_{j^{(0)}} \mathbf{x}_{:j^{(0)}} - \xi \mathbf{x}_{:j'})^{\top} \Gamma^{(t)} \boldsymbol{\eta}}{(\xi \mathbf{x}_{:j'} - s_{j^{(0)}} \mathbf{x}_{:j^{(0)}})^{\top} \Gamma^{(t)} \boldsymbol{\eta}} \leq 0, \quad (16a)$$

$$\theta_U = \min_{t \in [k]} \min_{\substack{(j', \xi) \in \bar{S}^{(t)} \times \{-1, 1, 0\}, \\ (\xi \mathbf{x}_{:j'} - s_{j^{(0)}} \mathbf{x}_{:j^{(0)}})^{\top} \Gamma^{(t)} \boldsymbol{\eta} > 0}} \frac{(s_{j^{(0)}} \mathbf{x}_{:j^{(0)}} - \xi \mathbf{x}_{:j'})^{\top} \Gamma^{(t)} \boldsymbol{\eta}}{(\xi \mathbf{x}_{:j'} - s_{j^{(0)}} \mathbf{x}_{:j^{(0)}})^{\top} \Gamma^{(t)} \boldsymbol{\eta}} \geq 0. \quad (16b)$$

where $\bar{S}^{(t)}$, $s_{j^{(0)}}$ and $\Gamma^{(t)}$ are appeared in Eq. (8).

The proof of Lemma 2 is presented in Appendix A.1.2. Since the inner maximization and inner minimization of Eq. (16) are difficult to compute, we present an efficient computational trick for those inner optimization problems, which is our main contribution. Our basic idea for addressing the computational difficulty is to note that most of the inequalities actually do not affect the results of the selective inference, and a large portion of them can be identified by exploiting the anti-monotonicity properties defined in the tree structure.

We consider a tree which consists of a set of nodes corresponding to each of the unselected features $j' \in \bar{S}^{(t)}$. We define $\text{Des}(j')$ be the set of descendant nodes of j' -th node in the tree. When we search over the tree, we introduce a novel pruning strategy by

Fig. 4 An example of our pruning technique behavior. In the tree, we consider the situation that the feature x_{i2} is selected at step $t = 1$ and the sets of indices are $S^{(1)} = \{2\}$, $\bar{S}^{(1)} = \{1, 3, 4, 5, 6, 7\}$. Then, we consider the tree corresponding to the unselected features $j' \in \bar{S}^{(1)}$. In this example, we update the current optimal solutions $\hat{\theta}_L$ and $\hat{\theta}_U$ by using the feature $x_{ij'}$ with the $j' = 1$. Moreover, the pruning conditions are satisfied at the $j' = 1$, then one can skip searching over its subtree which has no effect on the selective inference.

deriving a condition such that, if the j' -th node in the tree satisfies certain conditions, then all the $\tilde{j} \in \text{Des}(j')$ are guaranteed to be irrelevant to the selective inference results because they do not affect the optimal solutions in Eq. (16).

Theorem 3. Consider solving the inner maximization problem in Eq. (16a), and let $\hat{\theta}_L$ be the current optimal solution, i.e., we know that the optimal θ_L is at least no greater than $\hat{\theta}_L$. For any $j' \in \bar{S}^{(t)}$ and all $\xi \in \{-1, 1, 0\}$, if either of the following conditions

$$\begin{aligned} & \sum_{i: \xi(\Gamma^{(t)} \boldsymbol{\eta})_i < 0} x_{ij'} \xi (\Gamma^{(t)} \boldsymbol{\eta})_i - s_{j^{(0)}} \mathbf{x}_{:j^{(0)}}^{\top} \Gamma^{(t)} \boldsymbol{\eta} \geq 0, \text{ or} \\ & \frac{\max\{0, s_{j^{(0)}} \mathbf{x}_{:j^{(0)}}^{\top} \Gamma^{(t)} \boldsymbol{\eta} - \sum_{i: \xi(\Gamma^{(t)} \boldsymbol{\eta})_i > 0} x_{ij'} \xi (\Gamma^{(t)} \boldsymbol{\eta})_i\}}{\min\{0, \sum_{i: \xi(\Gamma^{(t)} \boldsymbol{\eta})_i < 0} x_{ij'} \xi (\Gamma^{(t)} \boldsymbol{\eta})_i - s_{j^{(0)}} \mathbf{x}_{:j^{(0)}}^{\top} \Gamma^{(t)} \boldsymbol{\eta}\}} \leq \hat{\theta}_L \end{aligned} \quad (17)$$

are satisfied, then its descendant nodes $\tilde{j} \in \text{Des}(j')$ do not affect the solution of Eq. (16a).

Similarly, for any $j' \in \bar{S}^{(t)}$ and all $\xi \in \{-1, 1, 0\}$, if either of the following conditions

$$\begin{aligned} & \sum_{i: \xi(\Gamma^{(t)} \boldsymbol{\eta})_i > 0} x_{ij'} \xi (\Gamma^{(t)} \boldsymbol{\eta})_i - s_{j^{(0)}} \mathbf{x}_{:j^{(0)}}^{\top} \Gamma^{(t)} \boldsymbol{\eta} \leq 0, \text{ or} \\ & \frac{\max\{0, s_{j^{(0)}} \mathbf{x}_{:j^{(0)}}^{\top} \Gamma^{(t)} \boldsymbol{\eta} - \sum_{i: \xi(\Gamma^{(t)} \boldsymbol{\eta})_i > 0} x_{ij'} \xi (\Gamma^{(t)} \boldsymbol{\eta})_i\}}{\max\{0, \sum_{i: \xi(\Gamma^{(t)} \boldsymbol{\eta})_i > 0} x_{ij'} \xi (\Gamma^{(t)} \boldsymbol{\eta})_i - s_{j^{(0)}} \mathbf{x}_{:j^{(0)}}^{\top} \Gamma^{(t)} \boldsymbol{\eta}\}} \geq \hat{\theta}_U \end{aligned} \quad (18)$$

are satisfied, then its descendant nodes $\tilde{j} \in \text{Des}(j')$ do not affect the solution of Eq. (16b).

The proof of Theorem 3 is presented in Appendix A.1.3. **Figure 4** shows that all the conditions in Theorem 3 can be checked at the j' -th node in each tree, and if the conditions are satisfied as the j' -th node, then one can skip searching over its subtree. It allows us to perform selective inference even if the number of constraints that defines the selection event is extremely large. A brief statement of selective inference with our proposed pruning technique is shown in Algorithm 1.

4. Extensions

In this section, we improve computational efficiency and sta-

Algorithm 1 Selective inference with our proposed pruning technique.

Input: Dataset $(\mathbf{Z}, \mathbf{y})$, the number of selected features k , the maximum interaction order r , desired significance level α .

Output: The set of selected features $S^{(k)}$, and confidence intervals for the selected features.

- 1: Consider the r th-order interaction features which are represented by the tree in Fig. 3.
- 2: **for** $t = 1, \dots, k$ **do**
- 3: Select a feature and update the current set $S^{(t)}$ by using the pruning condition (15) with OMP in Eq. (3).
- 4: **end for**
- 5: **for** $j = 1, \dots, k$ **do**
- 6: For $\boldsymbol{\eta} = (X_{S^{(t)}}^+)^T \mathbf{e}_j$, compute θ_L, θ_U in (16) by using our proposed pruning conditions (17) and (18).
- 7: Compute the confidence interval $[\ell_{\alpha/2}^{(S^{(t)}, j)}, u_{\alpha/2}^{(S^{(t)}, j)}]$ which satisfies Eq. (13).
- 8: **end for**

tistical power of selective inference. The drawback of selective inference is that one has to consider the number of considerable constraints as shown in Eq. (8). This is not desirable in terms of both computational efficiency and statistical power – according to Ref. [20], “a greater degree of conditioning will generally lead to less powerful tests and wider intervals,” and this fact is also studied in Ref. [25]. Moreover, the statistical power of selective inference is generally less powerful when the number of selected features is increased, because the degree of conditioning is linearly increased with the number of selected features.

In order to circumvent such a over conditioning, we consider the following confidence interval instead of Eq. (6): for any $t \in [k]$ with a fixed natural number k ,

$$\Pr(\beta_{S^{(t)}, t} \in C_{S^{(t)}, t} | S^{(t)}) \geq 1 - \alpha. \quad (19)$$

In this inequality, we consider incrementally selected k model, and we only interested in the coefficient of t -th selected feature in the t -th model. The advantage of this approach is that the coefficient $\beta_{S^{(t)}, t}$ is independent of $t + 1, \dots, k$ selected features. In other words, we do NOT need to consider the selection events that $t + 1, \dots, k$ features are selected, and then the degree of conditioning is strictly smaller than fully conditioned (8). Although a hypothesis testing for incrementally selected model is introduced in Refs. [8], [26] and the null hypothesis is called *incremental null*, our contribution is to utilize (19) for reducing the degree of conditioning.

Simulation Study for FCR Control

Here, we checked whether selective inference can control FCR or not when we consider the confidence interval (19). We generated the following synthetic response as $y_i = \mu_i + \epsilon_i, \mu_i = -z_{i1} + 2z_{i2} - 1.5z_{i3}, \epsilon_i \sim N(0, 0.5^2)$, and the element of design matrix $z_{ij} \in \{0, 1\}$ was randomly generated by

$$z_{ij} = \begin{cases} 0 & \text{if } u \leq \zeta, \\ 1 & \text{otherwise,} \end{cases} \quad (20)$$

where u was a random variable generated by the uniform distribution $\text{Unif}(0, 1)$, and ζ is the sparsity of design matrix. We fixed

Fig. 5 FCRs for various parameters: the number of samples n and selected features k . We fixed the number of original features d and sparsity of design matrix ζ as 100, 95%, respectively.

Table 2 The short description of approaches.

Symbol	Description
SI	Our selective inference approach.
SI _{inc}	SI with the incrementally constructed interval (19).
OLS	z-testing using ordinary least squares in Eq. (9).
split	Data-splitting approach.

the number of original features d and the sparsity ζ as 100, 95%, respectively. FCR was estimated by using v/k where v is the number of false rejections and k is the number of selected features, e.g., if two features z_{i1}, z_{i5} are selected with $k = 2$ and each of the intervals does not cover zero, then the number of false rejections v is one (because z_{i5} is not in the true model) and $\text{FCR} = 1/2$. We randomly generated the synthetic data over 2,000 times and averaged FCR over those independent simulations. **Figure 5** showed FCRs for various parameters with significance level was 0.05. FCR could be controlled for any the number of selected features $k \in \{1, \dots, 10\}$ even if the number of features in the true model was three. In particular, FCR in the case of greater n and smaller k was nearly zero since the selected features were in the true model in most cases and false rejections were almost none.

5. Experiments

5.1 Experiments on Synthetic Data

We compared our selective inference approach with non-adjusted z-testing using Eq. (9) and data-splitting approach [7] on synthetic data. In data-splitting, we splitted the data into two subsets, and used one for feature selection and another for z-testing. The performance of data-splitting is basically weak both in selection and inference stages because only a part of the available data is used in each stage. We wrote the short description of each approach in **Table 2**.

First, we generated the synthetic data whose response was $y_i = \epsilon_i, \epsilon_i \sim N(0, \sigma^2)$, this means that all coefficients are 0. The element of original design matrix $\mathbf{Z}$ was randomly generated by using Eq. (20). Since there were several hyper-parameters (σ, ζ , etc.), we used the default settings of parameters as shown in **Table 3**, but those parameters were changed depending on the experiment.

FCR Controlling

Here, we checked whether FCR can be properly controlled or not. FCR was estimated by using v/k where v is the number of false rejections and k is the number of selected features. We ran-

Table 3 Default setting of parameters.

Parameter	Value	Description
n	100	The number of samples.
d	100	The number of original features.
k	5	The number of selected features.
r	3	The maximum order of interactions.
σ	0.5	Noise level for response.
ζ	95%	Sparsity of original design matrix.
α	5%	Desired significance level.

Fig. 6 FCRs for various parameters. Our approaches and `split` could control FCRs under the desired significance level, while OLS could not control them.

domly generated the synthetic data over 2,000 times and averaged FCR over those independent simulations. **Figure 6** showed that our approaches and `split` could control FCRs for any parameters under the desired significance level, while OLS could not control them. (a) The selection bias was more intense if the number of samples n is increased. We conjecture that the non-adjusted confidence intervals might not cover 0 because a greater n will leads tighter confidence interval. (b) The selection bias critically depends on the number of original features. (c) The selection bias was smaller if the number of selected features k is increased. We conjecture that redundant features make the absolute value of fitted coefficients be smaller, and then the confidence intervals tend to cover 0 if k is excessively increased.

Statistical Power Comparison

Second, we evaluated the statistical power of the inference. We replaced the response as $y_i = -z_{i1} + 2z_{i2}z_{i3} - 3z_{i4}z_{i5}z_{i6} + \epsilon_i$, $\epsilon_i \sim N(0, \sigma^2)$, and we set the sparsity ζ of the original vectors $z_{:1}, \dots, z_{:6}$ to 80%. We computed TPR (true positive rate) of the approaches. We defined $\text{TPR} = \text{TP}/3$ where TP is the number of true positives, e.g., TP is 3 if the confidence intervals of coefficients corresponding to z_{i1} , $z_{i2}z_{i3}$ and $z_{i4}z_{i5}z_{i6}$ do not cover 0. We randomly generated the synthetic data over 2,000 times and averaged TPRs over those independent simulations. **Figure 7** showed that `SI_inc` quite outperformed data-splitting, while the power of SI decreases as the number of selected features increases. This indicates that the degree of conditioning critically depends on k in SI, which leads to wider intervals.

Computational Efficiency and Sensitivity

Lastly, we evaluated the computational efficiency and sensitiv-

Fig. 7 TPRs for various parameters. (a) and (b) `SI_inc` quite outperformed data-splitting. (c) the power of SI decreases as the number of selected features increases. (d) FPRs did not depend on the maximum interaction order r if $r \geq 4$; however, we observed that FPRs were slightly decreased depending on r when the number of samples n was very small.

Fig. 8 Computational times for various parameters: the number of original features $d \in \{500, 1000\}$, the maximum interaction order $r \in \{3, 5, 7, 9, \infty\}$, the sparsity of original design matrix $\zeta \in \{99.6\%, 99.5\%, 99.4\%\}$, the number of selected features $k \in \{5, 10\}$.

ity of our approaches. We used the same synthetic data in the previous paragraph (statistical power comparison), but we changed the number of original features d and the sparsity of original design matrix ζ . We randomly generated the synthetic data over 100 times and averaged computational time over those independent simulations. The machine spec was Intel(R) Core(TM) i7-3517U CPU 1.90 GHz with 4 GByte memory, and we implemented our approach as a single thread application. **Figure 8** showed that the computational cost strongly depends on three parameters: (i) the number of original features d , (ii) the sparsity ζ , and (iii) the

Fig. 9 The confidence intervals of fitted coefficients for $k = 30$ selected features with Bonferroni-adjusted significance level $\alpha = 0.05/k$.

maximum interaction order r when d is large and ζ is lower. SI and SI_{inc} were almost the same if the number of selected features k is small.

5.2 Application to HIV Drug Resistance Data

We applied our selective inference approaches to HIV-1 sequence data obtained from Stanford HIV Drug Resistance Database [27]. The goal here was to find statistically significant high-order interaction features that are highly associated with the drug resistances. The detail of the features is that, each feature means a mutation that is defined as amino acid differences from HIV-1 subtype B consensus sequence. If a position of the sequence is mutated, then the feature corresponding to the position is 1 otherwise 0. We used three datasets for antiretroviral drugs: Delavirdine (DLV), Tenofovir (TDF), and Atazanavir (ATV). We converted the response (drug resistance) y_i as $y_i \leftarrow \log(y_i)$ since the responses were heavy tailed. We estimated the variance pa-

Table 4 Computational times [s] in HIV datasets.

Dataset	SI	SI_{inc}
DLV ($n = 732, d = 371$)	54.7	28.5
TDF ($n = 353, d = 348$)	36.4	15.6
ATV ($n = 329, d = 225$)	16.8	8.1

Fig. 10 The confidence intervals using data-splitting approach where data is splitted differently. Each label of horizontal axis is sorted in the order of OMP selection steps: the left side corresponds to step = 1, and the right side corresponds to step = 30. It is quite annoying that both confidence interval and set of selected features are affected by the randomness of splitting.

rameter σ^2 by using $\hat{\sigma}^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$ where $\bar{y}$ is the sample mean of responses. We assumed that the estimated variance $\hat{\sigma}^2$ is greater than the true variance σ^2 , and this leads to wider intervals but type I error is still controlled. We selected $k = 30$ features with the maximum interaction order $r = 5$, and evaluated the confidence intervals for those features with Bonferroni-adjusted significance level $\alpha = 0.05/k$.

Figure 9 showed the confidence intervals of fitted coefficients for selected features. These results indicate that our approaches could successfully identify statistically significant high-order interaction features. Our proposed selective inference approaches

are applicable to genome-wide association studies (GWASs). In GWASs, evaluating combinatorial effects of mutations (or SNPs: single nucleotide polymorphisms) is important because co-occurrence of multiple mutations may have a significant influence on a response even if occurrence of single mutation has no influence [1], [2]. Although several existing pattern mining algorithms [21], [22], [23] can detect interactions of features which are associated with the response, these can NOT control false positive findings, which means that adaptation of the mining results to real world application would be risky. Our approaches can detect them under FCR controlling even if the number of possible combinations of features is extremely large. **Table 4** showed that the computational times of our approaches are feasible and SI_{inc} is strictly faster than SI . Lastly, we applied data-splitting approach to ATV dataset in different splitting. **Figure 10** showed that both confidence interval and set of selected features are affected by the randomness of splitting, that is a drawback of data-splitting approach.

6. Conclusion

In this paper, we extended selective inference framework to a high-order interaction model by introducing a novel computational trick for computing the sampling distribution of test statistic. We demonstrated that our approaches are computationally useful, the statistical power of inference is better than data-splitting approach, and it allows us to address selection bias issue.

References

- [1] Manolio, T.A. and Collins, F.S.: Genes, environment, health, and disease: Facing up to complexity, *Human Heredity*, Vol.63, No.2, pp.63–66 (2006).
- [2] Cordell, H.J.: Detecting gene–gene interactions that underlie human diseases, *Nature Reviews Genetics*, Vol.10, No.6, pp.392–404 (2009).
- [3] Fan, J. and Lv, J.: Sure independence screening for ultrahigh dimensional feature space, *Journal of The Royal Statistical Society B*, Vol.70, pp.849–911 (2008).
- [4] Pati, Y.C., Rezaifar, R. and Krishnaprasad, P.: Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition, *1993 Conference Record of The Twenty-Seventh Asilomar Conference on Signals, Systems and Computers*, pp.40–44, IEEE (1993).
- [5] Tibshirani, R.: Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society, Series B*, Vol.58, pp.267–288 (1996).
- [6] Lee, J.D., Sun, D.L., Sun, Y., Taylor, J.E., et al.: Exact post-selection inference, with application to the lasso, *The Annals of Statistics*, Vol.44, No.3, pp.907–927 (2016).
- [7] Fithian, W., Sun, D. and Taylor, J.: Optimal inference after model selection, arXiv preprint arXiv:1410.2597 (2014).
- [8] Fithian, W., Taylor, J., Tibshirani, R. and Tibshirani, R.: Selective sequential model selection, arXiv preprint arXiv:1512.02565 (2015).
- [9] Tian, X. and Taylor, J.: Asymptotics of selective inference, *Scandinavian Journal of Statistics*, Vol.44, No.2, pp.480–499 (2017).
- [10] Tian, X., Taylor, J., et al.: Selective inference with a randomized response, *The Annals of Statistics*, Vol.46, No.2, pp.679–710 (2018).
- [11] Tibshirani, R.J., Rinaldo, A., Tibshirani, R., Wasserman, L., et al.: Uniform asymptotic inference and the bootstrap after model selection, *The Annals of Statistics*, Vol.46, No.3, pp.1255–1287 (2018).
- [12] Taylor, J. and Tibshirani, R.: Post-selection inference for-penalized likelihood models, *Canadian Journal of Statistics*, Vol.46, No.1, pp.41–61 (2018).
- [13] Yang, F., Barber, R.F., Jain, P. and Lafferty, J.: Selective inference for group-sparse linear models, *Advances in Neural Information Processing Systems*, pp.2469–2477 (2016).
- [14] Barber, R.F., Candès, E.J., et al.: A knockoff filter for high-dimensional selective inference, *The Annals of Statistics*, Vol.47, No.5, pp.2504–2537 (2019).
- [15] Heckman, J.J.: Sample selection bias as a specification error, *Econometrica*, pp.153–161 (1979).

- [16] Suzumura, S., Nakagawa, K., Umez, Y., Tsuda, K. and Takeuchi, I.: Selective inference for sparse high-order interaction models, *Proc. 34th International Conference on Machine Learning-Volume 70*, JMLR.org, pp.3338–3347 (2017).
- [17] Benjamini, Y. and Yekutieli, D.: False discovery rate-adjusted multiple confidence intervals for selected parameters, *Journal of the American Statistical Association*, Vol.100, No.469, pp.71–81 (2005).
- [18] Storey, J.D. et al.: The positive false discovery rate: a Bayesian interpretation and the q-value, *The Annals of Statistics*, Vol.31, No.6, pp.2013–2035 (2003).
- [19] Lee, J.D. and Taylor, J.E.: Exact post model selection inference for marginal screening, *Advances in Neural Information Processing Systems*, pp.136–144 (2014).
- [20] Tibshirani, R.J., Taylor, J., Lockhart, R. and Tibshirani, R.: Exact post-selection inference for sequential regression procedures, *Journal of the American Statistical Association*, Vol.111, No.514, pp.600–620 (2016).
- [21] Nakagawa, K., Suzumura, S., Karasuyama, M., Tsuda, K. and Takeuchi, I.: Safe Pattern Pruning: An Efficient Approach for Predictive Pattern Mining, *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.1785–1794, ACM (2016).
- [22] Saigo, H., Uno, T. and Tsuda, K.: Mining complex genotypic features for predicting HIV-1 drug resistance, *Bioinformatics*, Vol.23, No.18, pp.2455–2462 (2007).
- [23] Tsuda, K.: Entire regularization paths for graph data, *Proc. 24th International Conference on Machine Learning*, pp.919–926, ACM (2007).
- [24] Morishita, S.: Computing optimal hypotheses efficiently for boosting, *Progress in Discovery Science*, pp.471–481, Springer (2002).
- [25] Liu, K., Markovic, J. and Tibshirani, R.: More powerful post-selection inference, with application to the lasso, arXiv preprint arXiv:1801.09037 (2018).
- [26] G’Sell, M.G., Wager, S., Chouldechova, A. and Tibshirani, R.: Sequential selection procedures and false discovery rate control, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Vol.78, No.2, pp.423–444 (2016).
- [27] Rhee, S.-Y., Gonzales, M.J., Kantor, R., Betts, B.J., Ravela, J. and Shafer, R.W.: Human immunodeficiency virus reverse transcriptase and protease sequence database, *Nucleic Acids Research*, Vol.31, No.1, pp.298–303 (2003).

Appendix

A.1 Proofs

A.1.1 Proof of Lemma 1

Proof. To prove the lemma, we first state the polyhedral lemma in Ref. [6] that is slightly general case of $V[\mathbf{y}] = \Sigma$ as follows:

Lemma 4 (Polyhedral Lemma; Ref. [6]). *Suppose $\mathbf{y} \sim N(\boldsymbol{\mu}, \Sigma)$. Let $\mathbf{c} = \Sigma\boldsymbol{\eta}(\boldsymbol{\eta}^\top \Sigma \boldsymbol{\eta})^{-1}$ for any $\boldsymbol{\eta} \in \mathbb{R}^n$, and let $\mathbf{z} = (\mathbf{I}_n - \mathbf{c}\boldsymbol{\eta}^\top)\mathbf{y}$. Then we have*

$$\text{Pol}(S) = \{\mathbf{y} \in \mathbb{R}^n \mid \mathbf{A}\mathbf{y} \leq \mathbf{b}\} \\ = \left\{ \mathbf{y} \in \mathbb{R}^n \mid \begin{array}{l} L(S, \mathbf{z}) \leq \boldsymbol{\eta}^\top \mathbf{y} \leq U(S, \mathbf{z}), \\ N(S, \mathbf{z}) \geq 0 \end{array} \right\},$$

where

$$L(S, \mathbf{z}) = \max_{j: (A\mathbf{c})_j < 0} \frac{b_j - (A\mathbf{z})_j}{(A\mathbf{c})_j}, \quad (\text{A.1a})$$

$$U(S, \mathbf{z}) = \min_{j: (A\mathbf{c})_j > 0} \frac{b_j - (A\mathbf{z})_j}{(A\mathbf{c})_j}, \quad (\text{A.1b})$$

$$N(S, \mathbf{z}) = \max_{j: (A\mathbf{c})_j = 0} b_j - (A\mathbf{z})_j. \quad (\text{A.1c})$$

In addition, $(L(S, \mathbf{z}), U(S, \mathbf{z}), N(S, \mathbf{z}))$ is independent of $\boldsymbol{\eta}^\top \mathbf{y}$.

The polyhedral lemma allows us to construct a pivotal quantity as a truncated normal distribution, that is

$$[F_{\boldsymbol{\eta}^\top \boldsymbol{\mu}, \boldsymbol{\eta}^\top \Sigma}^{L(S, \mathbf{z}), U(S, \mathbf{z})}(\boldsymbol{\eta}^\top \mathbf{y}) \mid \mathbf{y} \in \text{Pol}(S)] \sim \text{Unif}(0, 1),$$

where $\text{Unif}(0, 1)$ denotes the standard (continuous) uniform distribution. For concreteness, $\eta^\top \mathbf{y}$ is the least square estimator $\hat{\beta}_{S,j}$ when $\eta = (X_S^+)^T \mathbf{e}_j$. By using this, the conditional $(1 - \alpha)$ interval $[\ell_{\alpha/2}^{(S,z)}, u_{\alpha/2}^{(S,z)}]$ are defined by

$$F_{\ell_{\alpha/2}^{(S,z)}, \eta^\top \Sigma \eta}^{[L(S,z), U(S,z)]}(\eta^\top \mathbf{y}) = 1 - \alpha/2,$$

$$F_{u_{\alpha/2}^{(S,z)}, \eta^\top \Sigma \eta}^{[L(S,z), U(S,z)]}(\eta^\top \mathbf{y}) = \alpha/2.$$

Since $\mathbf{z}$ is independent of $\eta^\top \mathbf{y}$ (in fact, the direction of two vectors is orthogonal as $\eta^\top \mathbf{z} = 0$), we abbreviate $\mathbf{z}$ that can be integrated out.

The remaining is to show that truncation points in (A.1) are equivalent to

$$L(S) = \eta^\top \mathbf{y} + \theta_L \eta^\top \Sigma \eta \quad (\text{A.2a})$$

where $\theta_L = \min_{\theta \in \mathbb{R}} \theta \quad \text{s.t.} \quad \mathbf{y} + \theta \Sigma \eta \in \text{Pol}(S)$

and

$$U(S) = \eta^\top \mathbf{y} + \theta_U \eta^\top \Sigma \eta \quad (\text{A.2b})$$

where $\theta_U = \max_{\theta \in \mathbb{R}} \theta \quad \text{s.t.} \quad \mathbf{y} + \theta \Sigma \eta \in \text{Pol}(S)$,

respectively. Simple calculation shows that, for any $\theta \in \mathbb{R}$, we have

$$\begin{aligned} \mathbf{y} + \theta \Sigma \eta &\in \text{Pol}(S) \\ \Leftrightarrow A(\mathbf{y} + \theta \Sigma \eta) &\leq \mathbf{b} \\ \Leftrightarrow \theta \cdot A \Sigma \eta &\leq \mathbf{b} - A\mathbf{y}. \\ \Leftrightarrow \begin{cases} \theta \leq (\mathbf{b} - A\mathbf{y})_j / (A \Sigma \eta)_j, & (A \Sigma \eta)_j > 0 \\ \theta \geq (\mathbf{b} - A\mathbf{y})_j / (A \Sigma \eta)_j, & (A \Sigma \eta)_j < 0 \\ 0 \leq (\mathbf{b} - A\mathbf{y})_j, & (A \Sigma \eta)_j = 0 \end{cases} \end{aligned}$$

On the other hand, by the definition of $\mathbf{c}$ and $\mathbf{z}$ in Lemma 4, it is easy to see that

$$L(S) = \eta^\top \mathbf{y} + \eta^\top \Sigma \eta \max_{j: (A \Sigma \eta)_j < 0} \frac{(\mathbf{b} - A\mathbf{y})_j}{(A \Sigma \eta)_j}.$$

Therefore, for each j such that $(A \Sigma \eta)_j < 0$, we have

$$\max_{j: (A \Sigma \eta)_j < 0} \frac{(\mathbf{b} - A\mathbf{y})_j}{(A \Sigma \eta)_j} \leq \theta$$

and thus the minimum possible feasible θ would be

$$\begin{aligned} \theta_L &= \min\{\theta \in \mathbb{R} \mid \mathbf{y} + \theta \Sigma \eta \in \text{Pol}(S)\} \\ &= \max_{j: (A \Sigma \eta)_j < 0} \frac{(\mathbf{b} - A\mathbf{y})_j}{(A \Sigma \eta)_j}. \end{aligned}$$

Similarly, we see that the equivalency of $U(S)$.

To complete the proof, let us consider the covariance matrix $\sigma^2 I_n$ and choose $\eta = (X_S^+)^T \mathbf{e}_j$. In this case, (A.2) can be written as

$$L(S, j) = \hat{\beta}_{S,j} + \theta_L \sigma^2 (X_S^+ X_S^+)^{-1}_{jj}$$

where $\theta_L = \min_{\theta \in \mathbb{R}} \theta \quad \text{s.t.} \quad \mathbf{y} + \theta \sigma^2 (X_S^+)^T \mathbf{e}_j \in \text{Pol}(S)$

and

$$U(S, j) = \hat{\beta}_{S,j} + \theta_U \sigma^2 (X_S^+ X_S^+)^{-1}_{jj}$$

where $\theta_U = \max_{\theta \in \mathbb{R}} \theta \quad \text{s.t.} \quad \mathbf{y} + \theta \sigma^2 (X_S^+)^T \mathbf{e}_j \in \text{Pol}(S)$,

respectively, but we can ignore the scaling factor σ^2 because

$$\begin{aligned} \min\{\theta \in \mathbb{R} \mid \mathbf{y} + \theta (X_S^+)^T \mathbf{e}_j \in \text{Pol}(S)\} \\ = \min\{\theta \sigma^2 \in \mathbb{R} \mid \mathbf{y} + \theta \sigma^2 (X_S^+)^T \mathbf{e}_j \in \text{Pol}(S)\} \end{aligned}$$

and

$$\begin{aligned} \max\{\theta \in \mathbb{R} \mid \mathbf{y} + \theta (X_S^+)^T \mathbf{e}_j \in \text{Pol}(S)\} \\ = \max\{\theta \sigma^2 \in \mathbb{R} \mid \mathbf{y} + \theta \sigma^2 (X_S^+)^T \mathbf{e}_j \in \text{Pol}(S)\}. \end{aligned}$$

□

A.1.2 Proof of Lemma 2

Proof. From Eq. (8), for all possible pairs of $(j', \xi) \in \bar{S}^{(t)} \times \{-1, 1, 0\}$, $t \in [k]$, the constraint $\mathbf{y} + \theta \eta \in \text{Pol}(S^{(k)})$ is written as

$$(-s_{j(t)} \mathbf{x}_{:j(t)} + \xi \mathbf{x}_{:j'})^\top \Gamma^{(t)} (\mathbf{y} + \theta \eta) \leq 0,$$

that is equivalent to

if $(\xi \mathbf{x}_{:j'} - s_{j(t)} \mathbf{x}_{:j(t)})^\top \Gamma^{(t)} \eta < 0$, then

$$\frac{(s_{j(t)} \mathbf{x}_{:j(t)} - \xi \mathbf{x}_{:j'})^\top \Gamma^{(t)} \mathbf{y}}{(\xi \mathbf{x}_{:j'} - s_{j(t)} \mathbf{x}_{:j(t)})^\top \Gamma^{(t)} \eta} \leq \theta, \quad (\text{A.3a})$$

if $(\xi \mathbf{x}_{:j'} - s_{j(t)} \mathbf{x}_{:j(t)})^\top \Gamma^{(t)} \eta > 0$, then

$$\frac{(s_{j(t)} \mathbf{x}_{:j(t)} - \xi \mathbf{x}_{:j'})^\top \Gamma^{(t)} \mathbf{y}}{(\xi \mathbf{x}_{:j'} - s_{j(t)} \mathbf{x}_{:j(t)})^\top \Gamma^{(t)} \eta} \geq \theta. \quad (\text{A.3b})$$

(i) First, the numerators of Eq. (A.3) are non-negative because of Eq. (8). (ii) Second, if the denominators of Eq. (A.3) are negative, then the minimum possible feasible θ would be

$$\max_{t \in [k]} \max_{(j', \xi) \in \bar{S}^{(t)} \times \{-1, 1, 0\}} \frac{(s_{j(t)} \mathbf{x}_{:j(t)} - \xi \mathbf{x}_{:j'})^\top \Gamma^{(t)} \mathbf{y}}{(\xi \mathbf{x}_{:j'} - s_{j(t)} \mathbf{x}_{:j(t)})^\top \Gamma^{(t)} \eta},$$

and the maximum possible feasible θ would be 0. (iii) Similarly, if the denominators of Eq. (A.3) are positive, then the maximum possible feasible θ would be

$$\min_{t \in [k]} \min_{(j', \xi) \in \bar{S}^{(t)} \times \{-1, 1, 0\}} \frac{(s_{j(t)} \mathbf{x}_{:j(t)} - \xi \mathbf{x}_{:j'})^\top \Gamma^{(t)} \mathbf{y}}{(\xi \mathbf{x}_{:j'} - s_{j(t)} \mathbf{x}_{:j(t)})^\top \Gamma^{(t)} \eta},$$

and minimum possible feasible θ would be 0. Since the requirements in (i), (ii) and (iii) must be satisfied for all possible $(j', \xi) \in \bar{S}^{(t)} \times \{-1, 1, 0\}$, $t \in [k]$, θ_L and θ_U are given by Eqs. (16a) and (16b), respectively. □

A.1.3 Proof of Theorem 3

Proof. Nothing that $0 \leq x_{i\tilde{j}} \leq x_{ij} \leq 1$ from Eq. (14), for any descendant node $\tilde{j} \in \text{Des}(j')$. For any vector $\mathbf{v} \in \mathbb{R}^n$, the following inequalities are satisfied

$$\begin{aligned} (\xi \mathbf{x}_{:\tilde{j}} - s_{j(t)} \mathbf{x}_{:j(t)})^\top \mathbf{v} \\ = \sum_{\xi v_i > 0} x_{i\tilde{j}} \xi v_i + \sum_{\xi v_i < 0} x_{i\tilde{j}} \xi v_i - s_{j(t)} \mathbf{x}_{:j(t)}^\top \mathbf{v} \\ \geq \sum_{\xi v_i < 0} x_{i\tilde{j}} \xi v_i - s_{j(t)} \mathbf{x}_{:j(t)}^\top \mathbf{v} \end{aligned}$$

$$\geq \sum_{\xi v_i < 0} x_{ij'} \xi v_i - s_{j^{(t)}} \mathbf{x}_{j^{(t)}}^\top \mathbf{v}, \quad (\text{A.4a})$$

$$\begin{aligned} & (s_{j^{(t)}} \mathbf{x}_{j^{(t)}} - \xi \mathbf{x}_{j'})^\top \mathbf{v} \\ &= s_{j^{(t)}} \mathbf{x}_{j^{(t)}}^\top \mathbf{v} - \sum_{\xi v_i > 0} x_{ij'} \xi v_i - \sum_{\xi v_i < 0} x_{ij'} \xi v_i \\ &\geq s_{j^{(t)}} \mathbf{x}_{j^{(t)}}^\top \mathbf{v} - \sum_{\xi v_i > 0} x_{ij'} \xi v_i \\ &\geq s_{j^{(t)}} \mathbf{x}_{j^{(t)}}^\top \mathbf{v} - \sum_{\xi v_i > 0} x_{ij'} \xi v_i. \end{aligned} \quad (\text{A.4b})$$

We prove the pruning condition (17) in the theorem. (i) First, from Lemma 2, any pairs $(j^{(t)}, j')$ such that $(\xi \mathbf{x}_{j'} - s_{j^{(t)}} \mathbf{x}_{j^{(t)}})^\top \Gamma^{(t)} \boldsymbol{\eta} \geq 0$ are irrelevant to the solution θ_L . Also from Eq. (A.4a),

$$\begin{aligned} & \sum_{i: \xi(\Gamma^{(t)} \boldsymbol{\eta})_i < 0} x_{ij'} \xi(\Gamma^{(t)} \boldsymbol{\eta})_i - s_{j^{(t)}} \mathbf{x}_{j^{(t)}}^\top \Gamma^{(t)} \boldsymbol{\eta} \geq 0 \\ & \Rightarrow (\xi \mathbf{x}_{j'} - s_{j^{(t)}} \mathbf{x}_{j^{(t)}})^\top \Gamma^{(t)} \boldsymbol{\eta} \geq 0. \end{aligned}$$

(ii) Moreover, when $(\xi \mathbf{x}_{j'} - s_{j^{(t)}} \mathbf{x}_{j^{(t)}})^\top \Gamma^{(t)} \boldsymbol{\eta} < 0$, from Eqs. (A.4a) and (A.4b),

$$\begin{aligned} & \frac{(s_{j^{(t)}} \mathbf{x}_{j^{(t)}} - \xi \mathbf{x}_{j'})^\top \Gamma^{(t)} \mathbf{y}}{(\xi \mathbf{x}_{j'} - s_{j^{(t)}} \mathbf{x}_{j^{(t)}})^\top \Gamma^{(t)} \boldsymbol{\eta}} \\ & \leq \frac{\max\{0, s_{j^{(t)}} \mathbf{x}_{j^{(t)}}^\top \Gamma^{(t)} \mathbf{y} - \sum_{\xi(\Gamma^{(t)} \mathbf{y})_i > 0} x_{ij'} \xi(\Gamma^{(t)} \mathbf{y})_i\}}{\min\{0, \sum_{i: \xi(\Gamma^{(t)} \boldsymbol{\eta})_i < 0} x_{ij'} \xi(\Gamma^{(t)} \boldsymbol{\eta})_i - s_{j^{(t)}} \mathbf{x}_{j^{(t)}}^\top \Gamma^{(t)} \boldsymbol{\eta}\}}, \end{aligned}$$

note that the numerators of above inequalities are non-negative because of Eq. (8). By combining (i) and (ii), if Eq. (17) holds, then (j, ℓ') for $\ell' \in \text{Des}(j')$ do not affect the solution of Eq. (16a). The first half of the theorem is proved, and the other half can be shown similarly. $\square$

Shinya Suzumura is currently a Ph.D. student at Department of Computer Science, Nagoya Institute of Technology, Japan. He received B. Eng. and M. Eng. degrees from Chubu University, Japan, in 2011 and 2013, respectively. He joined Yahoo Japan Corporation in 2016. His research interests include interpretable machine learning and optimization for statistical data analysis.

Kazuya Nakagawa received B. Eng. and M. Eng. degrees from Nagoya Institute of Technology, Japan, in 2015 and 2017, respectively. His research interests include machine learning and optimization for statistical data analysis.

Yuta Umezu is an associate professor at Nagasaki University, Japan. He received B. Science, M. Mathematics, and D. Functional Mathematics degrees from Kyushu University, Japan, in 2011, 2013, and 2016, respectively. His research interests include theory and methodology of statistical mathematics and machine

learning and its practical application.

Koji Tsuda received B. Eng., M. Eng., and Ph.D. degrees from Kyoto University, Japan, in 1994, 1995, and 1998, respectively. Subsequently, he joined former Electrotechnical Laboratory (ETL), Tsukuba, Japan, as Research Scientist. When ETL was reorganized as AIST in 2001, he joined newly established Com-

putational Biology Research Center, Tokyo, Japan. In 2000–2001, he worked at GMD FIRST (currently Fraunhofer FIRST) in Berlin, Germany, as Visiting Scientist. In 2003–2004 and 2006–2008, he worked at Max Planck Institute for Biological Cybernetics, Tübingen, Germany, first as Research Scientist and later as Project Leader. Currently, he is Professor at Department of Computational Biology and Medical Sciences, Graduate School of Frontier Sciences, the University of Tokyo. He is also affiliated with National Institute of Material Science (NIMS) and RIKEN Center for Advanced Intelligence Project.

Ichiro Takeuchi is a professor at Nagoya Institute of Technology, Japan. He received B. Eng., M. Eng., and D. Eng. degrees from Nagoya University, Japan, in 1996, 1998, and 2000, respectively. After he worked as a post-doctoral researcher in Montreal, Canada, under Prof. Yoshua Bengio's supervision, he got tenured as-

sistant professor position at Mie University, Japan in 2001, and associate and full professor positions at Nagoya Institute of Technology in 2008 and 2015, respectively. His research interests include theory and algorithm of machine learning and its application to bio-medical and material sciences. Many of his machine learning studies have been published in top conferences of machine learning and data mining such as ICML, NIPS, KDD, while his application studies to bio-medical and material sciences have been published in high-impact journals in each domain.

(Communicated by Toshinori Endo)