Selective Inference for Sparse High-Order Interaction Models

Shinya Suzumura [1] **Kazuya Nakagawa** [1] **Yuta Umezu** [1] **Koji Tsuda** [2 3] **Ichiro Takeuchi** [1 3]

Abstract

Finding statistically significant high-order interactions in predictive modeling is important but challenging task because the possible number of high-order interactions is extremely large (e.g., $> 10^{17}$). In this paper we study feature selection and statistical inference for sparse high-order interaction models. Our main contribution is to extend recently developed selective inference framework for linear models to high-order interaction models by developing a novel algorithm for efficiently characterizing the selection event for the selective inference of high-order interactions. We demonstrate the effectiveness of the proposed algorithm by applying it to an HIV drug response prediction problem.

1. Introduction

Finding statistically reliable high-order interaction features in predictive modeling has been important challenging task. 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 (Manolio & Collins, 2006; Cordell, 2009). A major challenge in prediction modeling with high-order interaction features is the exponentially expanded feature space. If one has a dataset with d original variables and takes into account interactions up to order r, the model has $D := \sum_{\rho=1}^{r} \binom{d}{\rho}$ features (e.g., for $d = 10,000$, $r = 5$, $D > 10^{17}$). Unless both d and r are fairly small, D is extremely large. 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* (Fan & Lv, 2008), *orthogonal matching pursuit* (Pati et al., 1993), LASSO (Tibshirani, 1996), 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 very recently. The main challenge in statistical inference of sparse models is that, if the data is used for selecting a subset of features, this selection event must be taken into account in the following inference stage. Otherwise, the inference results are distorted by so-called *selection bias*, and false positive errors cannot be controlled at desired levels. This problem is refereed to as *selective inference* or *post selection inference* (Benjamini & Yekutieli, 2005; Benjamini et al., 2009; Berk et al., 2013). After the seminal work by Lee et al. (2016), significant progress has been recently made on selective inference for sparse linear models (Fithian et al., 2014b; Lee & Taylor, 2014; Fithian et al., 2015; Tian & Taylor, 2015; Taylor & Tibshirani, 2016; Yang et al., 2016; Barber & Candès, 2016).

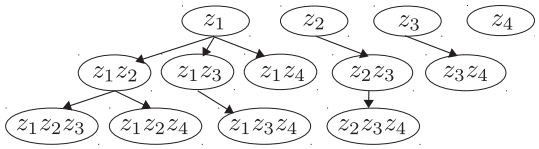


Figure 1. Example of the tree structure among high-order interaction features when $d = 4$ and $r = 3$.

In this paper, we study feature selection and statistical inference for sparse high-order interaction models. Unfortunately, neither existing feature selection methods nor existing selective inference methods can be applied to sparse high-order interaction models because the computational costs of these existing methods at least linearly depend on the number of features D. The main contribution in this paper is to develop computationally efficient algorithms for these two tasks when the original variables are represented in $[0, 1]^d$. Our main idea is to exploit the underlying tree structure of high-order interaction features as depicted in

[1]Nagoya Institute of Technology, Nagoya, Japan [2]University of Tokyo, Tokyo, Japan [3]RIKEN, Tokyo, Japan. Correspondence to: Ichiro Takeuchi <takeuchi.ichiro@nitech.ac.jp>.

Proceedings of the 34^{th} International Conference on Machine Learning, Sydney, Australia, PMLR 70, 2017.

Figure 1. In feature selection tasks, it allows us to efficiently identify interaction features that have no chance to be selected. In statistical inference tasks, it allows us to efficiently identify interaction features that do not affect the results of the selective inference.

We demonstrate the effectiveness of the proposed methods through numerical experiments both on synthetic and real datasets. In the latter, we apply the proposed method to HIV dataset in (Rhee et al., 2003), where the goal is to identify statistically significant high-order interactions of multiple gene mutations that are significantly associated with HIV drug responses.

Related works and our contributions Methods for efficiently finding high-order interaction features and properly evaluating their statistical significances have long been desired in many scientific studies.

In the past decade, feature selection for interaction models has been studied in the context of sparse learning (Choi et al., 2010; Hao & Zhang, 2014; Bien et al., 2013). None of these works have a special computational trick for handling exponentially large number of interaction features, which makes their empirical evaluations restricted up-to second order interactions. One commonly used heuristic in the context of interaction modeling is to introduce a prior knowledge such as *strong heredity assumption* where, e.g., an interaction term $z_1 z_2$ would be selected only when both of z_1 and z_2 are selected. Such a heuristic restriction is helpful for reducing the number of interaction terms to be considered. However, in many scientific studies, researchers are primarily interested in finding interactions even when their main effects alone do not have any association with the response. The idea of considering a tree structure among interaction features has been commonly used n data mining literature (Kudo et al., 2005; Saigo et al., 2006; Nakagawa et al., 2016). However, it is difficult to properly assess the statistical significances of the selected features by these mining techniques.

One traditional approach to assessing the statistical significances of selected features is *multiple testing correction (MTC)*. In the context of DNA microarray studies, many MTC procedures for high-dimensional data have been proposed (Tusher et al., 2001; Dudoit et al., 2003). An MTC approach for statistical evaluation of high-order interaction features was recently studied in (Terada et al., 2013; Llinares-López et al., 2015). A main drawback of MTC is that they are highly conservative when the number of candidate features increases. Another common approach is *data-splitting (DS)*. (Fithian et al., 2014a). In DS approach, we split the data into two subsets, and use one for feature selection and another for statistical inference, which enables us to remove the selection bias. However, performances of DS approach is clearly weak both in selection and inference stages because only a part of the available data is used in each stage. In addition, it is quite annoying that different set of features would be selected if data is splitted differently. Recently, much attention has been paid to selective inference for sparse linear models. The basic idea of selective inference is to make inferences conditional on a feature selection event. Lee et al. (2016) recently proposed a practical selective inference framework for a class of feature selection algorithms.

The main contribution in this paper is to extend the selective inference framework into sparse high-order interaction models by introducing novel computational algorithms. To the best of our knowledge, there are no other existing works for sparse high-order interaction models in which the statistical significances of the fitted coefficients are properly evaluated in non-asymptotic sense.

Notations We use the following notations in the remainder. For any natural number n, we define $[n] := \{1, \dots, n\}$. A vector and a matrix is denoted such as $\boldsymbol{v} \in \mathbb{R}^n$ and $M \in \mathbb{R}^{n \times m}$, respectively. The index function is written as $\mathbf{1}\{z\}$ which returns 1 if z is true, and 0 otherwise. The sign function is written as $\mathrm{sgn}(z)$ which returns 1 if $z \geq 0$, and -1 otherwise. An $n \times n$ identity matrix is denoted as I_n.

2. Preliminaries

2.1. Problem setup

Consider a regression problem with a response $Y \in \mathbb{R}$ and d-dimensional original covariates $\boldsymbol{z} = [z_1, \dots, z_d]^\top$ by the following high-order interaction model up to r-th order

$$Y = \sum_{j_1 \in [d]} \alpha_{j_1} z_{j_1} + \sum_{\substack{(j_1, j_2) \in [d] \times [d] \\ j_1 \neq j_2}} \alpha_{j_1, j_2} z_{j_1} z_{j_2} + \cdots + \sum_{\substack{(j_1, \dots, j_r) \in [d]^r \\ j_1 \neq \dots \neq j_r}} \alpha_{j_1, \dots, j_r} z_{j_1} \cdots z_{j_r} + \varepsilon, \quad (1)$$

where αs are the coefficients and ε is a random noise. We assume that each original covariate $z_j, j \in [d]$ is defined in a domain $[0, 1]$. Here, values 1 and 0 respectively might be interpreted as the existence and the non-existence of a certain property, and values between them indicate the "degree" of existence. High-order interaction features thus represent co-existence of multiple properties. For example, if we are interested in interactions among age, body mass index (BMI), and a mutation in a certain gene, we may code

some covariates as

$$z_{j_1} := \begin{cases} 1 & \text{if BMI } > 30, \\ (\text{BMI} - 15)/(30 - 15) & \text{if BMI } \in [15, 30], \\ 0 & \text{if BMI } < 15, \end{cases}$$

$z_{j_2} := \mathbf{1}\{\text{mutation in the gene}\}$.

Then, e.g., an interaction term $z_{j_1} z_{j_2}$ represents the co-existence of high BMI and a mutation in the gene.

The high-order interaction model Eq.(1) has in total $D := \sum_{\rho \in [r]} \binom{d}{\rho}$ features. Let us write the mapping from the original covariates $\boldsymbol{z} := [z_1, \ldots, z_d]^\top \in \mathbb{R}^d$ to the high-order interaction features $\boldsymbol{x} := [x_1, \ldots, x_D]^\top \in \mathbb{R}^D$ as $\boldsymbol{\phi} : [0,1]^d \to [0,1]^D, \boldsymbol{z} \mapsto \boldsymbol{x}$,, i.e.,

$$\boldsymbol{x} := \boldsymbol{\phi}(\boldsymbol{z}) = [z_1, \ldots, z_d, z_1 z_2, \ldots, z_{d-1} z_d, \ldots, z_1 \cdots z_k, \ldots, z_{d-r+1} \cdots z_d]^\top$$

Then, the high-order interaction model Eq.(1) is simply written as a D-dimensional linear model

$$y = \boldsymbol{\beta}^\top \boldsymbol{x} = \beta_1 x_1 + \cdots + \beta_D x_D,$$

where $\beta_1, \ldots, \beta_D$ are D coefficients corresponding to $\alpha_{j_1}, \ldots, \alpha_{j_1, \ldots, j_r}$ in Eq.(1). Since a high-order interaction feature is a product of original covariates defined in $[0,1]$, the range of each feature $x_j, j \in [D]$ is also $[0,1]$.

The original training set is denoted as $\{(\boldsymbol{z}_i, y_i) \in [0,1]^d \times \mathbb{R}\}_{i \in [n]}$, while the expanded training set is written as $\{(\boldsymbol{x}_i, y_i) \in [0,1]^D \times \mathbb{R}\}_{i \in [n]}$. The latter is also denoted as $(X, \boldsymbol{y}) \in [0,1]^{n \times D} \times \mathbb{R}^n$ where each row of X is $\boldsymbol{x}_i \in \mathbb{R}^d$ and each element of $\boldsymbol{y}$ is y_i. Furthermore, the j-th column of X is written as $\boldsymbol{x}_{\cdot j}, j \in [D]$. We denote the pseudo inverse of X as $X^+ := (X^\top X)^{-1} X^\top$.

Our goal is to identify statistically significant high-order interaction terms that have large impacts on the response Y by identifying regression coefficients αs which are significantly deviated from zero. Unfortunately, since the number of coefficients αs to be fitted would be far greater than the sample size n, traditional least-square estimation theory cannot be used for making statistical inferences on the fitted model. We thus consider first to perform feature selection and then to make statistical inference only for the selected features based on selective inference approach.

2.2. Selective inference for sparse linear models

In this section, we briefly review the selective inference framework for sparse linear models developed by Lee et al. (2016). Selective inference is developed for two stage methods, where a subset of features is *selected* in the first stage, and *inferences* are made only on the selected features in the second stage. A key finding by Lee et al. (2016) is that, if the first selection stage is described as a *linear selection event*, then exact statistical inference of the fitted coefficients conditional on the selection event can be done.

Consider a linear regression model $\boldsymbol{y} = X\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}$, where $\boldsymbol{\beta}^* \in \mathbb{R}^D$ is the true coefficients and $\boldsymbol{\varepsilon}$ is distributed according to $\mathrm{N}(\boldsymbol{0}, \sigma^2 I)$ with known variance σ^2.

Feature selection stage Suppose that, in the first feature selection stage, a subset of features $S \subseteq [D]$ are selected. The selective inference framework in Lee et al. (2016) can be applied to feature selection algorithms whose selection process can be characterized by a set of linear inequalities in the form of $A\boldsymbol{y} \leq \boldsymbol{b}$ with a certain matrix A and a certain vector $\boldsymbol{b}$ that do not depend on $\boldsymbol{y}$. 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 $\boldsymbol{y}$ is observed in a polytope $\mathrm{Pol}(S) := \{\boldsymbol{y} \in \mathbb{R}^n \mid A\boldsymbol{y} \leq \boldsymbol{b}\}$. In Lee & Taylor (2014) and Lee et al. (2016), marginal screening, OMP and LASSO are shown to be linear selection events, indicating that the selective inference framework can be applied to statistical testing of the selected features by these algorithms.

Statistical inference stage Consider a hypothesis testing for the j-th selected feature in S

$$\mathrm{H}_{0,j} : \beta^*_{S,j} = 0 \quad \text{vs.} \quad \mathrm{H}_{1,j} : \beta^*_{S,j} \neq 0. \tag{2}$$

The least-square estimator of the linear model only with the selected features S is written as $\hat{\boldsymbol{\beta}}_S = (X_S^\top X_S)^{-1} X_S^\top \boldsymbol{y}$.

If we consider the case where S is NOT selected from the data, i.e., independent of $\boldsymbol{y}$, then, under the null hypothesis H_0, the sampling distribution of each fitted coefficient is

$$\hat{\beta}_{S,j} \sim \mathrm{N}(0, \sigma^2_{S,j}), \text{ where } \sigma^2_{S,j} := \sigma^2 (X_S^\top X_S)^{-1}_{jj}. \tag{3}$$

For two-sided test at level α, if the critical values $\ell_{\alpha/2}$ and $u_{\alpha/2}$ are chosen to be the lower and the upper $\alpha/2$ points of the sampling distribution in Eq.(3), then the *type I error* at level α is controlled as

$$\Pr(\hat{\beta}_{S,j} \notin [\ell_{\alpha/2}, u_{\alpha/2}]) \leq \alpha \tag{4}$$

On the other hand, when S is selected from the data as we consider here, we would like to control the following *selective type I error*

$$\begin{aligned} &\Pr(\hat{\beta}_{S,j} \notin [\ell^{(S,j)}_{\alpha/2}, u^{(S,j)}_{\alpha/2}] \mid \{S \text{ is selected}\}) \\ =&\Pr(\hat{\beta}_{S,j} \notin [\ell^{(S,j)}_{\alpha/2}, u^{(S,j)}_{\alpha/2}] \mid \boldsymbol{y} \in \mathrm{Pol}(S)) \leq \alpha \end{aligned} \tag{5}$$

by appropriately selecting the *adjusted* critical values $\ell^{(S,j)}_{\alpha/2}$ and $u^{(S,j)}_{\alpha/2}$, where the selection event $\{S \text{ is selected}\}$ is

written as $\boldsymbol{y} \in \mathrm{Pol}(S)$ in the case of a linear selection event. Lee et al. (2016) derived how to compute these adjusted critical values as formally stated in the following lemma.

Lemma 1. *If the critical values are computed as*

$$\ell^{(S,j)}_{\alpha/2} := (F^{[L(S,j),U(S,j)]}_{0,\sigma^2_{S,j}})^{-1}(\alpha/2), \tag{6a}$$

$$u^{(S,j)}_{\alpha/2} := (F^{[L(S,j),U(S,j)]}_{0,\sigma^2_{S,j}})^{-1}(1-\alpha/2), \tag{6b}$$

then the selective type I error is controlled as in Eq. (5), where $F^{[L,U]}_{\mu,\sigma^2}$ is the cumulative distribution function of a truncated Normal distribution $\mathrm{TN}(\mu,\sigma^2,L,U)$, i.e.,

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

and the truncation points are obtained, by using the observed $\hat{\beta}_{S,j}$ and $\boldsymbol{y}$, as

$$L(S,j) = \hat{\beta}_{S,j} + \theta_L (X_S^\top X_S)^{-1}_{jj}, \tag{7a}$$
$$\textit{where } \theta_L := \min_{\theta\in\mathbb{R}} \theta \textit{ s.t. } \boldsymbol{y} + \theta (X_S^+)^\top \boldsymbol{e}_j \in \mathrm{Pol}(S),$$

$$U(S,j) = \hat{\beta}_{S,j} + \theta_U (X_S^\top X_S)^{-1}_{jj}, \tag{7b}$$
$$\textit{where } \theta_U := \max_{\theta\in\mathbb{R}} \theta \textit{ s.t. } \boldsymbol{y} + \theta (X_S^+)^\top \boldsymbol{e}_j \in \mathrm{Pol}(S).$$

The proof of Lemma 1 is is presented in Appendix A although it is easily proved by using the results in Lee et al. (2016). See Lee et al. (2016) for more general statement about the selective inference framework.

Eq.(7) indicates that the truncation points are obtained by considering the interval where the test statistic $\hat{\beta}_{S,j}$ can move within the polyhedron $\mathrm{Pol}(S)$. Figure 2 schematically illustrates that, when we make inferences conditional on a linear selection event S, the sampling distribution is defined within the polytope $\mathrm{Pol}(S)$, and it follows a truncated normal distribution when $\boldsymbol{y}$ is normally distributed.

Unfortunately, we cannot directly apply this selective inference framework to high-order interaction models because the polytope $\mathrm{Pol}(S)$ is characterized by extremely large number of linear inequalities, and the optimization problems in Eq.(7) are hard to solve.

3. Feature selection for interaction models

In this section, we present two feature selection algorithms for high-order interaction models. Since the number of features D is extremely large, existing feature selection algorithms for linear models cannot be directly applied to interaction models. In this paper, we study marginal screening (MS) and orthogonal matching pursuit (OMP) as examples of feature selection algorithms.

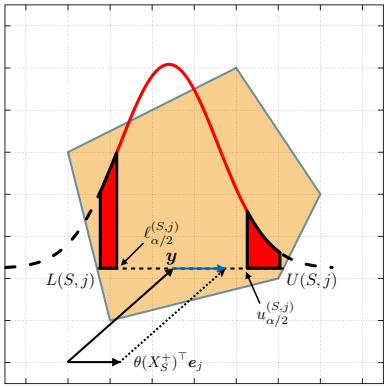


Figure 2. An illustration of polyhedral lemma. The polyhedron represents the selection event and truncation points can be computed by optimizing θ along with the direction $(X_S^+)^\top \boldsymbol{e}_j$. In addition, critical values can be obtained by computing Eq.(6), and the red region shows the rejection region of the test in Eq.(2).

3.1. MS for interaction models

Consider selecting the top k interaction features from all the D interaction features that have marginal strong correlations with the response. Noting that each feature is defined in $[0,1]$ and the value indicates (the degree of) the existence of a certain property, we consider a score $\boldsymbol{x}_{\cdot j}^\top \boldsymbol{y}, j \in [D]$ for each of the D features, and select the top k features according to their absolute scores $|\boldsymbol{x}_{\cdot j}^\top \boldsymbol{y}|$. We denote the index set of the selected k features by S, and that of the unselected $\bar{k} := D - k$ features by $\bar{S} := [D] \setminus S$.

Since D is extremely large, we cannot compute the score for each interaction feature. We exploit the tree structure among interaction patterns as depicted in Figure 1.

Definition 2. *(Descendant features) For each $j \in [D]$, let $Des(j) \subseteq [D]$ be the set of features corresponding to the descendant nodes in the tree including j itself.*

Lemma 3. *Consider an interaction feature $\boldsymbol{x}_{\cdot j}, j \in [D]$, whose indices are represented in a tree structure as depicted in Figure1. Then, for any node $j \in [D]$ in the tree,*

$$|\boldsymbol{x}_{\cdot \tilde{j}}^\top \boldsymbol{y}| \le \max \left\{ \sum_{i:y_i>0} x_{ij} y_i, -\sum_{i:y_i<0} x_{ij} y_i \right\} \tag{8}$$

for all $\tilde{j} \in Des(j)$.

The proof of Lemma 3 is presented in Appendix A. Lemma 3 tells that, for a descendant feature $\boldsymbol{x}_{\cdot \tilde{j}}, (j,\tilde{j}) \in S \times Des(j)$, an upper bound of the absolute score $|\boldsymbol{x}_{\cdot \tilde{j}}^\top \boldsymbol{y}|$ can be computed based on its parent feature $\boldsymbol{x}_{\cdot j}$.

We note that this simple upper bound has been used in some data mining studies such as Saigo et al. (2006); Kudo et al.

(2004); Nakagawa et al. (2016). When we search over the tree, if the upper bound in Eq.(8) is smaller than the current k-th largest score at a certain node j, then we can quit searching over its descendant nodes $\tilde{j} \in Des(j)$.

As pointed out in Lee & Taylor (2014), feature selection processes of marginal screening is a linear selection event, i.e., characterized by a set of linear constraints. The event that k features in S are selected, and $\bar{k}$ features in $\bar{S}$ are not selected is rephrased as $|\boldsymbol{x}_{\cdot j}^\top \boldsymbol{y}| \geq |\boldsymbol{x}_{\cdot j'}^\top \boldsymbol{y}|$, $\forall\ (j, j') \in S \times \bar{S}$. Let $s_j := \mathrm{sgn}(\boldsymbol{x}_{\cdot j}^\top \boldsymbol{y}), j \in S$. Then, the above feature selection event is rewritten with the sign constraints of the selected features by the following $2k\bar{k}+k$ constraints

$$(-s_j \boldsymbol{x}_{\cdot j} - \boldsymbol{x}_{\cdot j'})^\top \boldsymbol{y} \leq 0, \quad \forall\ (j, j') \in S \times \bar{S}, \tag{9a}$$
$$(-s_j \boldsymbol{x}_{\cdot j} + \boldsymbol{x}_{\cdot j'})^\top \boldsymbol{y} \leq 0, \quad \forall\ (j, j') \in S \times \bar{S}, \tag{9b}$$
$$-s_j \boldsymbol{x}_{\cdot j}^\top \boldsymbol{y} \leq 0, \quad \forall\ j \in S. \tag{9c}$$

These constraints are written as $A\boldsymbol{y} \leq \boldsymbol{0}$ with a matrix $A \in \mathbb{R}^{(2k\bar{k}+k)\times n}$. Unfortunately, finding $\theta_{\min}$ and $\theta_{\max}$ by naively solving the optimization problems in Eq.(7) is computationally difficult because the polyhedron $\mathrm{Pol}(S)$ is characterized by the extremely large number of constraints. For example, when $d = 10,000$, $r = 5$, $k = 10$, the number of linear inequalities that defines the polyhedron $\mathrm{Pol}(S)$ is $2k\bar{k} + k > 10^{19}$.

3.2. OMP for interaction models

Orthogonal matching pursuit (OMP) is a well-known iterative feature selection method (Pati et al., 1993). At each iteration, the most correlated feature with the residual of the current model which is fitted via least-squares method by using the features selected in earlier steps.

Consider again selecting k interaction features by OMP. Let $[(1), \ldots, (h)]$ be the sequence of the indices of the selected features from step 1 to step h for $h \in [k]$, and define $S_h := \{(1), \ldots, (h)\}$. Before step $h+1$, we have already selected h features $\boldsymbol{x}_{\cdot j}, j \in S_h$. Using these h features, the current n-dimensional model output is written as $\sum_{j\in[h]} \hat{\beta}_{S_h,(j)} \boldsymbol{x}_{\cdot(j)}$, where the coefficients $\hat{\beta}_{S_h,(j)}, j \in [h]$ are estimated by least-squares method. Denoting by Γ_{S_h} the $n \times h$ matrix whose j-th column is $\boldsymbol{x}_{\cdot(j)}$, the least square estimates are written as $\hat{\boldsymbol{\beta}}_{S_h} := [\hat{\beta}_{S_h,(1)}, \ldots, \hat{\beta}_{S_h,(h)}]^\top = (\Gamma_{S_h})^+ \boldsymbol{y}$. Then, at the $h+1$ step, we consider the correlation between the residual vector $\boldsymbol{r}_h := \boldsymbol{y} - \Gamma_{S_h} \hat{\boldsymbol{\beta}}_{S_h}$ and a feature $\boldsymbol{x}_{\cdot j'}$ for $j' \in \bar{S}_h$, and find the one that maximizes the absolute correlation $|\boldsymbol{x}_{\cdot j'}^\top \boldsymbol{r}_h|$ among them. Here, since the number of remaining features $|\bar{S}_h| = D - h$ is still extremely large, it is hard to compute all these $D-h$ correlations. To overcome this difficulty, we can simply use Lemma 3 just by replacing $\boldsymbol{y}$ with the current residual $\boldsymbol{r}_h$. Specifically, for a descendant feature $\boldsymbol{x}_{\cdot\tilde{j}}, \tilde{j} \in Des(j)$, an upper bound of $|\boldsymbol{x}_{\cdot\tilde{j}}^\top \boldsymbol{r}_h|$ is given as

$$|\boldsymbol{x}_{\cdot\tilde{j}}^\top \boldsymbol{r}_h| \leq \max \left\{ \sum_{i:r_{h,i}>0} x_{ij} r_{h,i}, - \sum_{i:r_{h,i}<0} x_{ij} r_{h,i} \right\}.$$

At each iteration, when we search over the tree, if the upper bound is smaller than the current largest correlation, then, in the same way as the case of MS, we can quit searching over its descendant nodes $j' \in Des(j)$.

It is also pointed out in Lee & Taylor (2014) that a feature selection process of OMP is linear selection event. At step h, the event that the (h)-th feature is selected is formulated as $|\boldsymbol{x}_{\cdot(h)}^\top \boldsymbol{r}_h| \geq |\boldsymbol{x}_{\cdot j'}^\top \boldsymbol{r}_h|$, for all $j' \in \bar{S}_h$. Let $P_{S_h} := I_n - \Gamma_{S_h} \Gamma_{S_h}^+$. Then, the above selection event is rewritten as a set of linear inequalities with respect to $\boldsymbol{y}$

$$(-s_{(h)} \boldsymbol{x}_{\cdot(h)} - \boldsymbol{x}_{\cdot j'})^\top P_{S_h} \boldsymbol{y} \leq 0, \forall\ j' \in \bar{S}_h, \tag{10a}$$
$$(-s_{(h)} \boldsymbol{x}_{\cdot(h)} + \boldsymbol{x}_{\cdot j'})^\top P_{S_h} \boldsymbol{y} \leq 0, \forall\ j' \in \bar{S}_h, \tag{10b}$$
$$-s_{(h)} \boldsymbol{x}_{\cdot(h)}^\top P_{S_h} \boldsymbol{y} \leq 0, \tag{10c}$$

where $s_{(h)} = \mathrm{sgn}(\boldsymbol{x}_{\cdot(h)}^\top \boldsymbol{r}_h)$. By combining all the linear selection events in k steps, the entire selection event of the OMP is characterized by $\sum_{h\in[k]}(2(D-h)+1)$ linear inequalities in $\mathbb{R}^n$. In practice, it is computationally intractable to handle these extremely large number of linear inequalities.

4. Selective inference for interaction models

In this section, we present an efficient selective inference algorithm for high-order interaction models, which is our main contribution.

The discussion in §3 suggests that it would be hard to compute critical values for selective inference in Eq.(6) because the selection event $\boldsymbol{y} \in \mathrm{Pol}(S)$ is characterized by extremely large number of inequalities. Our basic idea for addressing this 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 among high-order interaction features.

4.1. Marginal screening

We consider k trees for each of the k selected features. Each tree consists of a set of nodes corresponding to each of the non-selected features $j' \in \bar{S}$. For a pair $(j, j') \in S \times \bar{S}$, the j'-th node in the j-th tree corresponds to the linear inequalities Eqs.(9a) and (9b). When we search over these k trees, we introduce a novel pruning strategy by deriving a condition such that, if the j'-th node in the j-th tree satisfies certain conditions, then all the $(j, \tilde{j}')$-th inequalities for all

$\tilde{j}' \in Des_j(j')$ are guaranteed to be irrelevant to the selective inference results because they do not affect the optimal solutions in Eq.(7), where we define $Des_j(j')$ be all the features corresponding to the descendant node of j' in the j-th tree.

Lemma 4. *Let $\boldsymbol{\eta} := (X_S^+)^\top \boldsymbol{e}_j$. The solutions of the optimization problems in (7) are respectively written as*

$$\theta_L = -\min\{\theta_L^{(a)}, \theta_L^{(b)}, \theta_L^{(c)}\},$$
$$\theta_U = -\max\{\theta_U^{(a)}, \theta_U^{(b)}, \theta_U^{(c)}\},$$

where

$$\theta_L^{(a)} := \min_{\substack{(j,j')\in S\times\bar{S},\\ (s_j\boldsymbol{x}_{\cdot j}+\boldsymbol{x}_{\cdot j'})^\top\boldsymbol{\eta}>0}} \frac{(s_j\boldsymbol{x}_{\cdot j}+\boldsymbol{x}_{\cdot j'})^\top\boldsymbol{y}}{(s_j\boldsymbol{x}_{\cdot j}+\boldsymbol{x}_{\cdot j'})^\top\boldsymbol{\eta}}, \tag{11a}$$

$$\theta_U^{(a)} := \max_{\substack{(j,j')\in S\times\bar{S},\\ (s_j\boldsymbol{x}_{\cdot j}+\boldsymbol{x}_{\cdot j'})^\top\boldsymbol{\eta}<0}} \frac{(s_j\boldsymbol{x}_{\cdot j}+\boldsymbol{x}_{\cdot j'})^\top\boldsymbol{y}}{(s_j\boldsymbol{x}_{\cdot j}+\boldsymbol{x}_{\cdot j'})^\top\boldsymbol{\eta}}, \tag{11b}$$

$$\theta_L^{(b)} := \min_{\substack{(j,j')\in S\times\bar{S},\\ (s_j\boldsymbol{x}_{\cdot j}-\boldsymbol{x}_{\cdot j'})^\top\boldsymbol{\eta}>0}} \frac{(s_j\boldsymbol{x}_{\cdot j}-\boldsymbol{x}_{\cdot j'})^\top\boldsymbol{y}}{(s_j\boldsymbol{x}_{\cdot j}-\boldsymbol{x}_{\cdot j'})^\top\boldsymbol{\eta}}, \tag{11c}$$

$$\theta_U^{(b)} := \max_{\substack{(j,j')\in S\times\bar{S},\\ (s_j\boldsymbol{x}_{\cdot j}-\boldsymbol{x}_{\cdot j'})^\top\boldsymbol{\eta}<0}} \frac{(s_j\boldsymbol{x}_{\cdot j}-\boldsymbol{x}_{\cdot j'})^\top\boldsymbol{y}}{(s_j\boldsymbol{x}_{\cdot j}-\boldsymbol{x}_{\cdot j'})^\top\boldsymbol{\eta}}, \tag{11d}$$

$$\theta_L^{(c)} := \min_{\substack{j\in S,\\ s_j\boldsymbol{x}_{\cdot j}^\top\boldsymbol{\eta}>0}} \frac{s_j\boldsymbol{x}_{\cdot j}^\top\boldsymbol{y}}{s_j\boldsymbol{x}_{\cdot j}^\top\boldsymbol{\eta}}, \quad \theta_U^{(c)} := \max_{\substack{j\in S,\\ s_j\boldsymbol{x}_{\cdot j}^\top\boldsymbol{\eta}<0}} \frac{s_j\boldsymbol{x}_{\cdot j}^\top\boldsymbol{y}}{s_j\boldsymbol{x}_{\cdot j}^\top\boldsymbol{\eta}}.$$

The proof of Lemma 4 is presented in Appendix A.

Lemma 5. *For any triplet $(j, j', \tilde{j}') \in S \times \bar{S} \times Des_j(j')$,*

$$L_E^{(a)} := s_j\boldsymbol{x}_{\cdot j}^\top\boldsymbol{y} + \sum_{i:y_i<0} x_{ij'}y_i \le (s_j\boldsymbol{x}_{\cdot j}+\boldsymbol{x}_{\cdot\tilde{j}'})^\top\boldsymbol{y}, \tag{12a}$$

$$U_E^{(a)} := s_j\boldsymbol{x}_{\cdot j}^\top\boldsymbol{y} + \sum_{i:y_i>0} x_{ij'}y_i \ge (s_j\boldsymbol{x}_{\cdot j}+\boldsymbol{x}_{\cdot\tilde{j}'})^\top\boldsymbol{y}, \tag{12b}$$

$$L_D^{(a)} := s_j\boldsymbol{x}_{\cdot j}^\top\boldsymbol{\eta} + \sum_{i:\eta_i<0} x_{ij'}\eta_i \le (s_j\boldsymbol{x}_{\cdot j}+\boldsymbol{x}_{\cdot\tilde{j}'})^\top\boldsymbol{\eta}, \tag{12c}$$

$$U_D^{(a)} := s_j\boldsymbol{x}_{\cdot j}^\top\boldsymbol{\eta} + \sum_{i:\eta_i>0} x_{ij'}\eta_i \ge (s_j\boldsymbol{x}_{\cdot j}+\boldsymbol{x}_{\cdot\tilde{j}'})^\top\boldsymbol{\eta}, \tag{12d}$$

$$L_E^{(b)} := s_j\boldsymbol{x}_{\cdot j}^\top\boldsymbol{y} - \sum_{i:y_i>0} x_{ij'}y_i \le (s_j\boldsymbol{x}_{\cdot j}-\boldsymbol{x}_{\cdot\tilde{j}'})^\top\boldsymbol{y}, \tag{12e}$$

$$U_E^{(b)} := s_j\boldsymbol{x}_{\cdot j}^\top\boldsymbol{y} - \sum_{i:y_i<0} x_{ij'}y_i \ge (s_j\boldsymbol{x}_{\cdot j}-\boldsymbol{x}_{\cdot\tilde{j}'})^\top\boldsymbol{y}, \tag{12f}$$

$$L_D^{(b)} := s_j\boldsymbol{x}_{\cdot j}^\top\boldsymbol{\eta} - \sum_{i:\eta_i>0} x_{ij'}\eta_i \le (s_j\boldsymbol{x}_{\cdot j}-\boldsymbol{x}_{\cdot\tilde{j}'})^\top\boldsymbol{\eta}, \tag{12g}$$

$$U_D^{(b)} := s_j\boldsymbol{x}_{\cdot j}^\top\boldsymbol{\eta} - \sum_{i:\eta_i<0} x_{ij'}\eta_i \ge (s_j\boldsymbol{x}_{\cdot j}-\boldsymbol{x}_{\cdot\tilde{j}'})^\top\boldsymbol{\eta}. \tag{12h}$$

The proof of Lemma 5 is presented in Appendix A.

Theorem 6. *(i) Consider solving the optimization problem in Eq.(11a), and let $\hat{\theta}_L^{(a)}$ be the current optimal solution, i.e., we know that the optimal $\theta_L^{(a)}$ is at least no greater than $\hat{\theta}_L^{(a)}$. If*

$$\{U_D^{(a)} < 0\} \cup \{L_D^{(a)} > 0, L_E^{(a)} < 0, L_E^{(a)}/L_D^{(a)} > \hat{\theta}_L^{(a)}\} \cup \{L_D^{(a)} > 0, L_E^{(a)} > 0, L_E^{(a)}/U_D^{(a)} > \hat{\theta}_L^{(a)}\}$$

is true, then the $(j, \tilde{j}')$-th constraint in Eq. (9a) for any $(j, j', \tilde{j}') \in S \times \bar{S} \times Des_j(j')$ does not affect the optimal solution in Eq.(11a).

(ii) Next, consider solving the optimization problem in Eq.(11c), and let $\hat{\theta}_L^{(b)}$ be the current optimal solution. If

$$\{U_D^{(b)} < 0\} \cup \{L_D^{(b)} > 0, L_E^{(b)} < 0, L_E^{(b)}/L_D^{(b)} > \hat{\theta}_L^{(b)}\} \cup \{L_D^{(b)} > 0, L_E^{(b)} > 0, L_E^{(b)}/U_D^{(b)} > \hat{\theta}_L^{(b)}\}$$

is true, then the $(j, \tilde{j}')$-th constraint in Eq. (9b) for any $(j, j', \tilde{j}') \in S \times \bar{S} \times Des_j(j')$ does not affect the optimal solution in Eq.(11c).

(iii) Furthermore, consider solving the optimization problem in Eq.(11b), and let $\hat{\theta}_U^{(a)}$ be the current optimal solution. If

$$\{L_D^{(a)} > 0\} \cup \{U_D^{(a)} < 0, L_E^{(a)} < 0, L_E^{(a)}/U_D^{(a)} < \hat{\theta}_U^{(a)}\} \cup \{U_D^{(a)} < 0, L_E^{(a)} > 0, L_E^{(a)}/L_D^{(a)} < \hat{\theta}_U^{(a)}\}$$

is true, then the $(j, \tilde{j}')$-th constraint in Eq. (9a) for any $(j, j', \tilde{j}') \in S \times \bar{S} \times Des_j(j')$ does not affect the optimal solution in Eq.(11b).

(iv) Finally, consider solving the optimization problem in Eq.(11d), and let $\hat{\theta}_U^{(b)}$ be the current optimal solution. If

$$\{L_D^{(b)} > 0\} \cup \{U_D^{(b)} > 0, L_E^{(b)} < 0, L_E^{(b)}/U_D^{(b)} < \hat{\theta}_U^{(b)}\} \cup \{U_D^{(b)} > 0, L_E^{(b)} < 0, L_E^{(b)}/L_D^{(b)} < \hat{\theta}_U^{(b)}\}$$

is true, then the $(j, \tilde{j}')$-th constraint in Eq. (9b) for any $(j, j', \tilde{j}') \in S \times \bar{S} \times Des_j(j')$ does not affect the optimal solution in Eq.(11d).

The proof of Theorem 6 is presented in Appendix. Note that all the conditions in Theorem 6 can be checked at the j'-th node in each tree. 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 for high-order interaction models even the number of constraints that defines the selection event is extremely large. As we demonstrate in the experiment section, these pruning conditions are quite effective in practice. For example, we can perform selective inference for an interaction models with $d = 10,000$, $r = 5$, $k = 10$ in a few seconds.

4.2. Orthogonal matching pursuit (OMP)

As we discuss in the previous section, the selection event at each iteration of OMP has same form as MS. Therefore, we can derive similar pruning conditions as in Theorem 6 for OMP. Due to the space limitation, we deffer the corresponding lemma and the theorem for OMP in Appendix B.

5. Experiments

We demonstrate the performance of the selective inference for high-order sparse interaction models by numerical experiments on synthetic datasets and a real dataset.

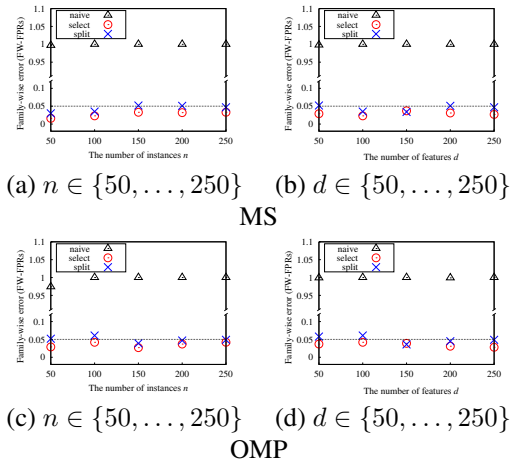
5.1. Experiments on synthetic datasets

First, we compared selective inference (`select`) with naive (`naive`) and data-splitting (`split`) on synthetic datasets. In `naive`, the critical values of the selected k features were naively computed without any selection bias correction mechanisms as in Eq. (4). In `split`, the dataset was first divided into two equally sized sets, and one of them was used for selection stage, and the other was used for inference stage. Note that the errors controlled by these methods are individual false positive rate for each of the selected features (although `naive` actually cannot control it), we applied Bonferroni correction within the k selected features, i.e., we reject the hypothesis in Eq. (2) with the significance level α/k where $\alpha = 0.05$, and we refer this error as family-wise false positive rates (FW-FPRs).

The synthetic dataset was generated as follows. In the experiments for comparing FW-FPRs, we generated the training instances $(\boldsymbol{z}_i, y_i) \in [0,1]^d \times \mathbb{R}$ independently at random for each $i \in [n]$. The original covariates $\boldsymbol{z}_i$ were randomly generated so that it contains $d(1-\zeta)$ 1s on average, where $\zeta \in [0,1]$ is an experimental parameter for representing the sparsity of the dataset, while the response y_i was randomly generated from a Normal distribution $N(0, \sigma^2)$. In the experiments for comparing true positive rates (TPRs) the response y_i was randomly generated from a Normal distribution $N(\boldsymbol{\mu}(X), \sigma^2 I)$, where, for each row of $\boldsymbol{\mu}(X)$ is defined as $\mu(\boldsymbol{z}_i) = 2z_1z_2z_3$ in the experiments for MS, $\mu(\boldsymbol{z}_i) = 0.5z_1 - 2z_2z_3 + 3z_4z_5z_6$ in the experiments for OMP. We investigated the performances by changing various experimental parameters. We set the baseline parameters as $n = 100$, $d = 100$, $k = 5$, $r = 5$, $\alpha = 0.05$, $\sigma = 0.5$, and $\zeta = 0.6$.

5.1.1. False positive rates

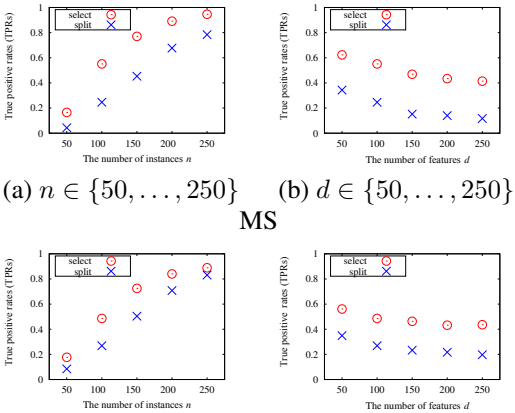
Figure 3 shows the FW-FPRs when varying the number of transactions $n \in \{50, 100, \ldots, 250\}$, the number of original covariates $d \in \{50, 100, \ldots, 250\}$. In all cases, the FW-FPRs of `naive` were far greater than the desired significance level $\alpha = 0.05$, indicating that the selection bias is harmful. The FW-FPRs of the other two approaches `select` and `split` were successfully controlled.



(a) $n \in \{50, \ldots, 250\}$ (b) $d \in \{50, \ldots, 250\}$ MS

(c) $n \in \{50, \ldots, 250\}$ (d) $d \in \{50, \ldots, 250\}$ OMP

Figure 3. False positive rates (FPRs).



(a) $n \in \{50, \ldots, 250\}$ (b) $d \in \{50, \ldots, 250\}$ MS

(c) $n \in \{50, \ldots, 250\}$ (d) $d \in \{50, \ldots, 250\}$ OMP

Figure 4. True positive rates (TPRs).

5.1.2. True positive rates

Figure 4 shows the TPRs of `select` and `split` (we omit `naive` because it cannot control FPRs). Here, TPRs are defined as the probability of finding truly correlated interaction features. In all the setups, the TPRs of `select` were much greater than `split`. Note that the performances of `split` would be worse than `select` both in the selection and the inference stages. The risk of failing to select truly correlated features in `split` would be higher than `select` because only half of the data would be used in the selection stage. Similarly, the statistical power in the inference stage in `split` would be smaller than `select` because the sample size is smaller.

Table 1. Computation times [sec]

	MS				OMP			
	with computational trick		without computational trick		with computational trick		without computational trick	
n	$\zeta = 0.8$	$\zeta = 0.9$	$\zeta = 0.8$	$\zeta = 0.9$	$\zeta = 0.8$	$\zeta = 0.9$	$\zeta = 0.8$	$\zeta = 0.9$
100	4.68×10^{-2}	1.80×10^{-2}	1.37×10^{2}	1.31×10^{2}	2.33×10^{-1}	5.85×10^{-2}	8.83×10^{2}	8.28×10^{2}
500	1.74×10^{-1}	9.07×10^{-2}	1.80×10^{2}	1.36×10^{2}	1.01×10^{0}	3.74×10^{-1}	1.33×10^{3}	8.60×10^{2}
1000	3.38×10^{-1}	1.54×10^{-1}	2.65×10^{2}	1.41×10^{2}	3.18×10^{0}	7.27×10^{-1}	2.15×10^{3}	9.07×10^{2}
5000	2.33×10^{0}	6.61×10^{-1}	1.05×10^{3}	2.57×10^{2}	6.20×10^{1}	3.48×10^{0}	1.00×10^{4}	2.05×10^{3}
10000	5.04×10^{0}	1.55×10^{0}	2.06×10^{3}	5.12×10^{2}	1.24×10^{2}	9.00×10^{0}	1.98×10^{4}	4.63×10^{3}
d	$\zeta = 0.8$	$\zeta = 0.9$	$\zeta = 0.8$	$\zeta = 0.9$	$\zeta = 0.8$	$\zeta = 0.9$	$\zeta = 0.8$	$\zeta = 0.9$
100	4.40×10^{-2}	1.77×10^{-2}	1.47×10^{2}	1.31×10^{2}	2.41×10^{-1}	6.02×10^{-2}	8.86×10^{2}	8.20×10^{2}
500	5.06×10^{-1}	1.64×10^{-1}	≥ 1 day	≥ 1 day	3.52×10^{1}	9.83×10^{0}	≥ 1 day	≥ 1 day
1000	1.23×10^{0}	3.74×10^{-1}	≥ 1 day	≥ 1 day	3.01×10^{2}	1.66×10^{2}	≥ 1 day	≥ 1 day
5000	1.53×10^{1}	2.88×10^{0}	≥ 1 day	≥ 1 day	≥ 1 day	1.92×10^{3}	≥ 1 day	≥ 1 day
10000	3.70×10^{1}	6.16×10^{0}	≥ 1 day	≥ 1 day	≥ 1 day	5.98×10^{4}	≥ 1 day	≥ 1 day

Table 2. The numbers of significant high-order interactions of multiple mutations in HIV datasets.

	MS					OMP				
Data	1^{st}	2^{nd}	3^{rd}	4^{th}	Time[s]	1^{st}	2^{nd}	3^{rd}	4^{th}	Time[s]
NNRTI ($d = 371$)										
dlv($n = 732$)	1				.495	2				18.0
efv($n = 734$)					.732	5				13.7
nvp($n = 746$)	4	1			.774	8				17.4
NRTI ($d = 348$)										
3tc($n = 633$)	1	2			.257	4				15.1
abc($n = 628$)	5	13	7	2	.238	9				11.7
azt($n = 630$)	2	5	3	1	.231	5				17.5
d4t($n = 630$)	4	11	6	1	.215	7	1	3		13.7
ddi($n = 632$)	2	1			.234	6				12.1
tdf($n = 353$)					.230	3	1			26.4
PI ($d = 225$)										
apv($n = 768$)	3	6	1		.188	9				6.5
atv($n = 329$)	1	3	2		.150	3	1			5.0
idv($n = 827$)	1	6	3		.437	9				6.2
lpv($n = 517$)	4	4	1		.275	11				6.1
nfv($n = 844$)	5	7	1		.455	15				5.8
rtv($n = 795$)	5	7	2		.183	10	1			5.6
sqv($n = 826$)	1	3	2		.623	7	1			7.8

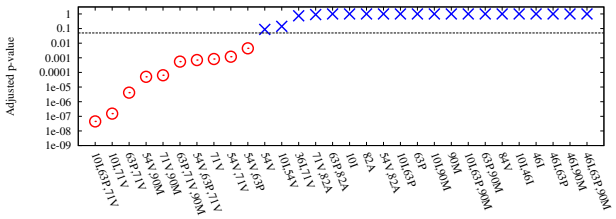
5.1.3. Computational efficiency

Table 1 shows the computation times in seconds for the selective inference approach with and without the computational tricks described in §4 for various values of the number of transactions $n \in \{100, \ldots, 10,000\}$, the number of original covariates $d \in \{100, \ldots, 10,000\}$, and the sparsity rates $\zeta \in \{0.8, 0.9\}$ (we terminated the search if the time exceeds 1 day). It can be observed from the table that, if we use the computational trick, the selective inferences can be conducted with reasonable computational costs except for $d \geq 5,000$ and $\zeta = 0.8$ cases with OMP. When the computational trick was not used, the cost was extremely large. Especially when the number of original covariates d is larger than 100, we could not complete the search within 1 day. From the results, we conclude that computational trick described in §4 is indispensable for selective inferences for sparse high-order interaction models.

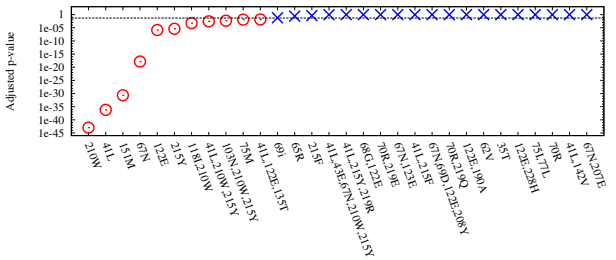
5.2. Application to HIV drug resistance data

We applied the selective inference approach to HIV-1 sequence data obtained from Stanford HIV Drug Resistance Database (Rhee et al., 2003). The goal here is to find statistically significant high-order interactions of multiple mutations (up to $r = 5$ order interactions) that are highly associated with the drug resistances. We selected $k = 30$ features, and evaluated the statistical significances of these features by the selective inference framework. Table 2 shows the numbers of 1st, 2nd, 3rd and 4th order interactions that were statistically significant after Bonferroni correction, i.e., significance level is set to be α/k with $\alpha = 0.05$. (there were no statistically significant 5th order interactions).

Figure 5 shows the degree of significances in the form of *adjusted p-values* after Bonferroni correction in increasing order on `idv` and `d4t` datasets by MS and OMP scenario, respectively. These results indicate that the selective inference approach could successfully identify statistically significant high-order interactions of multiple mutations.



(a) `idv dataset (MS)`



(b) `d4t dataset (OMP)`

Figure 5. The list of Bonferroni-adjusted selective p-values of $k = 30$ selected high-order interactions of multiple mutations on two HIV datasets.

Acknowledgements

This work was partially supported by MEXT KAKENHI (17H00758, 16H06538), JST CREST (JPMJCR1302, JPMJCR1502), RIKEN Center for Advanced Intelligence Project, and JST support program for starting up innovation-hub on materials research by information integration initiative.

References

Barber, Rina Foygel and Candès, Emmanuel J. A knockoff filter for high-dimensional selective inference. arXiv preprint arXiv:1602.03574, 2016.

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

Benjamini, Yoav, Heller, Ruth, and Yekutieli, Daniel. Selective inference in complex research. *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 367(1906): 4255–4271, 2009.

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

Bien, J., Taylor, J. E., and Tibshirani, R. A LASSO for hierarchical interactions. *Journal of The Royal Statistical Society B*, 41:1111–1141, 2013.

Choi, N.H., Li, W., and Zhu, J. Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association*, 105:354–364, 2010.

Cordell, Heather J. Detecting gene–gene interactions that underlie human diseases. *Nature Reviews Genetics*, 10 (6):392–404, 2009.

Dudoit, Sandrine, Shaffer, Juliet Popper, and Boldrick, Jennifer C. Multiple hypothesis testing in microarray experiments. *Statistical Science*, pp. 71–103, 2003.

Fan, J. and Lv, J. Sure independence screening for ultrahigh dimensional feature space. *Journal of The Royal Statistical Society B*, 70:849–911, 2008.

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

Fithian, William, Sun, Dennis, and Taylor, Jonathan. Optimal inference after model selection. arXiv preprint arXiv:1410.2597, 2014b.

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

Hao, Ning and Zhang, Hao Helen. Interaction screening for ultrahigh-dimensional data. *Journal of the American Statistical Association*, 109(507):1285–1301, 2014.

Kudo, T., Maeda, E., and Matsumoto, Y. An application of boosting to graph classification. In *Advances in Neural Information Processing Systems*, 2005.

Kudo, Taku, Maeda, Eisaku, and Matsumoto, Yuji. An application of boosting to graph classification. In *Advances in neural information processing systems*, pp. 729–736, 2004.

Lee, Jason D and Taylor, Jonathan E. Exact post model selection inference for marginal screening. In *Advances in Neural Information Processing Systems*, pp. 136–144, 2014.

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

Llinares-López, Felipe, Sugiyama, Mahito, Papaxanthos, Laetitia, and Borgwardt, Karsten. Fast and memory-efficient significant pattern mining via permutation testing. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 725–734. ACM, 2015.

Manolio, Teri A and Collins, Francis S. Genes, environment, health, and disease: facing up to complexity. *Human heredity*, 63(2):63–66, 2006.

Nakagawa, Kazuya, Suzumura, Shinya, Karasuyama, Masayuki, Tsuda, Koji, and Takeuchi, Ichiro. Safe pattern pruning: An efficient approach for predictive pattern mining. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1785–1794. ACM, 2016.

Pati, Yagyensh Chandra, Rezaiifar, Ramin, and Krishnaprasad, PS. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Signals, Systems and Computers, 1993. 1993 Conference Record of The Twenty-Seventh Asilomar Conference on*, pp. 40–44. IEEE, 1993.

Rhee, Soo-Yon, Gonzales, Matthew J, Kantor, Rami, Betts, Bradley J, Ravela, Jaideep, and Shafer, Robert W. Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic acids research*, 31 (1):298–303, 2003.

Saigo, H., Uno, T., and Tsuda, K. Mining complex genotypic features for predicting hiv-1 drug resistance. *Bioinformatics*, 24:2455—2462, 2006.

Taylor, Jonathan and Tibshirani, Robert. Post-selection inference for l1-penalized likelihood models. arXiv preprint arXiv:1602.07358, 2016.

Terada, Aika, Okada-Hatakeyama, Mariko, Tsuda, Koji, and Sese, Jun. Statistical significance of combinatorial regulations. *Proceedings of the National Academy of Sciences*, 110(32):12996–13001, 2013.

Tian, Xiaoying and Taylor, Jonathan. Asymptotics of selective inference. arXiv preprint arXiv:1501.03588, 2015.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996.

Tusher, Virginia Goss, Tibshirani, Robert, and Chu, Gilbert. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, 98(9):5116–5121, 2001.

Yang, Fan, Barber, Rina Foygel, Jain, Prateek, and Lafferty, John. Selective inference for group-sparse linear models. *arXiv preprint arXiv:1607.08211*, 2016.