Fast and More Powerful Selective Inference for Sparse High-Order Interaction Model

Diptesh Das,[1*] Vo Nguyen Le Duy,[1,2] Hiroyuki Hanada,[2] Koji Tsuda,[2,3,4] Ichiro Takeuchi,[1,2*]

[1]Nagoya Institute of Technology, Japan,
[2]RIKEN, Japan,
[3]The University of Tokyo, Japan,
[4]NIMS, Japan.

Abstract

Automated high-stake decision-making, such as medical diagnosis, requires models with high interpretability and reliability. We consider the sparse high-order interaction model as an interpretable and reliable model with a good prediction ability. However, finding statistically significant high-order interactions is challenging because of the intrinsically high dimensionality of the combinatorial effects. Another problem in data-driven modeling is the effect of "cherry-picking" (i.e., selection bias). Our main contribution is extending the recently developed parametric programming approach for selective inference to high-order interaction models. An exhaustive search over the cherry tree (all possible interactions) can be daunting and impractical, even for small-sized problems. We introduced an efficient pruning strategy and demonstrated the computational efficiency and statistical power of the proposed method using both synthetic and real data.

Introduction

Although blackbox models such as deep neural network models generally have a high predictive performance, they are difficult to interpret; hence, they are often considered unreliable. Therefore, for tasks that require high-stake decision-making, such as medical diagnosis, models with higher interpretability and reliability are required. We consider the sparse high-order interaction model (SHIM) as an interpretable and reliable model with a good prediction ability. Considering a regression problem with a response y and m original covariates $z_1, \ldots, z_m$, an example SHIM up to 4^{th} order interactions can be written as

$$y = \beta_1 z_3 + \beta_2 z_5 + \beta_3 z_2 z_6 + \beta_4 z_1 z_2 z_5 z_9, \qquad (1)$$

where $\beta_1, \beta_2, \beta_3, \beta_4$ are the model parameters (or coefficients). Such a SHIM (Das et al. 2019; Rendle 2010; Hall 1999) has practical importance, including identifying complex genotypic features for HIV-1 drug resistance (Saigo, Uno, and Tsuda 2007). HIV-1 evolves in the human body, and exposure to certain drugs causes mutations that lead to drug resistance. Structural biological studies show that the association of multiple mutations along with some crucial single mutations can best describe the complex biological phenomenon of drug resistance (Vivet-Boudou et al. 2006; Iversen et al. 1996; Rhee et al. 2006).

The goal of this study is to fit a SHIM, such as that in (1), to the given data and subsequently conduct a statistical significance test to judge the reliability of the model parameters. However, unless the original dimension and the order of interactions are small, fitting a high-order interaction model can be challenging, requiring some computational tricks to avoid combinatorial effects. Another challenge of data-driven modeling is understanding the reliability of the findings because the model might have cherry-picked the strong associations given a particular realization of the data. This is called the "cherry-picking" effect, or selection bias (Taylor and Tibshirani 2015). A traditional statistical inference, which assumes that the statistical model and the target for which inferences are conducted must be fixed *a priori*, cannot be used for this problem. Any inference conducted after model selection suffers from selection bias unless it is corrected.

Related work: Several approaches have been suggested in the literature to address the problem of selection bias (Fithian, Sun, and Taylor 2014; Fithian et al. 2015; Choi, Taylor, and Tibshirani 2017; Tian and Taylor 2018; Chen and Bien 2020; Hyun et al. 2018; Loftus and Taylor 2014, 2015; Panigrahi, Taylor, and Weinstein 2016; Tibshirani et al. 2016; Yang et al. 2016; Liu, Markovic, and Tibshirani 2018). A particularly notable approach is the *conditional* selective inference (SI) introduced in the seminal study conducted by Lee et al. (2016). The basic idea of conditional SI is to make inferences on a data-driven hypothesis conditional on the selection event that the hypothesis is selected. Lee et al. (2016) first proposed conditional SI methods for the selected features using Lasso. Their basic idea is to characterize the selection event by a polytope, i.e., a set of linear inequalities, in the sample space. When a selection event can be characterized by a polytope, the practical computational methods developed by these authors can be used to make inferences of the selected hypotheses conditional on the selection events.

However, the conditional SI framework based on a polytope has a serious drawback, called the *over-conditioning*

[*] Correspondence: das.diptesh@nitech.ac.jp, takeuchi.ichiro@nitech.ac.jp

issue. In that case, extra events must be introduced to characterize the selection event by a single polytope, which is known to be *statistically sub-optimal* or leads to a loss of statistical power (Fithian, Sun, and Taylor 2014). For example, to characterize the LASSO selection event by a polytope in Lee et al. (2016), the author explained that one needs to consider conditioning on the sign of the LASSO model parameters in addition to the selected model. The study by Suzumura et al. (2017), who first applied polytope-based SI into a high-order interaction model for which a high-order interaction feature is sequentially added, also suffers from this problem. As a solution to the case of LASSO, Lee et al. (2016) proposed taking the union of all possible signs of the selected features. However, unless the number of the selected features is small, it is computationally expensive, and in the case of a SHIM-type problem, it will be impractical because of the combinatorial effects.

Recently, Le Duy and Takeuchi (2021) introduced a homotopy method to resolve the over-conditioning issue, which essentially leads to a minimally conditioned SI for Lasso. The homotopy method exploits the piecewise linearity of the model coefficients $\beta(\tau)$ as a function of scalar τ. This enables one to simply use a linear interpolation between the change points of each piece of piecewise linear paths to compute the exact solution $\beta(\tau)$, thus avoiding the computational burden of solving the exact optimization problem for each and every τ between the change points of every linear piece. A well-known homotopy algorithm in statistical machine learning is the LARS-LASSO algorithm to construct the exact regularization paths of LASSO solutions (Efron et al. 2004). Recently there have been other studies which also exploited the homotopy method in the context of conditional SI to improve the statistical efficiency (Duy and Takeuchi 2021b; Sugiyama, Le Duy, and Takeuchi 2021; Sugiyama et al. 2021; Duy and Takeuchi 2021a) .

Our basic idea for identifying statistically reliable high-order interaction features in a sparse modeling framework is to employ an *exact homotopy*-based SI method for the SHIM. Unfortunately, the computational cost of applying the exact homotopy method to the SHIM increases exponentially and becomes intractable unless the size of the selected features and the maximum order of interactions are small (Mairal and Yu 2012). Several methods have already been proposed for fitting the SHIM (Saigo et al. 2009; Tsuda 2007; Nakagawa et al. 2016). These existing approaches mainly consider a tree of high-order interactions (or patterns) and exploit the tree anti-monotonicity property to derive efficient branch and bound pruning strategies in order to avoid the combinatorial computation burdens of the SHIM.

Contribution: Our main contribution in this paper is to introduce a "homotopy mining" method by exploiting the best of both homotopy and (pattern) mining methods for the conditional SI of the SHIM. This approach is motivated by the exact regularization path computation algorithm for graph data (Tsuda 2007), which is considered as a homotopy method with respect to the regularization parameter. In the algorithm of our proposed method, we use two types of homotopy mining methods, one for fitting a SHIM to the observed dataset (which is essentially the same as the approach in (Tsuda 2007)), and the other for computing the sampling distribution of the test statistic conditional on the selection event. Interestingly, these two types of homotopy mining methods share many common properties such as branch and bound techniques for pruning a high-order interaction tree (see Fig.1(a)). We applied our proposed method to synthetic and real-world HIV1 drug resistance data and demonstrated in the results section that we can quantify the statistical significance of high-order interaction features in the forms of p-values and confidence intervals without any computational or statistical approximations. In an experimental study of the inference stage, we showed that a single traversal of a search space of more than 10^9 high-order interaction terms (sample size = 625) took less than 317 s (worst case) and 110 s (best case) on average using an Intel Xeon Gold 6130 CPU @ 2.10 GHz. We extended this framework to solve the Elastic Net problem (Zou and Hastie 2005). It was not trivial as we could not follow the commonly used practice of data augmentation. The combinatorial effects of the SHIM prohibit the stacking of extra rows as the method of data augmentation. Our implementation is available at https://github.com/DipteshDas/SI-SHIM.

Problem Statement

Consider a regression problem with a response vector $y \in \mathbb{R}^n$ and m original covariate vectors $z_1, \dots, z_m$, where $z_j \in \mathbb{R}^n$ and $j \in [m] = \{1, ..., m\}$. A high-order interaction model up to the d^{th} order is then written as follows:

$$\begin{aligned} y = \sum_{j_1 \in [m]} \xi_{j_1} z_{j_1} + \sum_{\substack{(j_1, j_2) \in [m] \times [m] \\ j_1 \neq j_2}} \xi_{j_1, j_2} z_{j_1} z_{j_2} + \cdots \\ + \sum_{\substack{(j_1, \dots, j_d) \in [m]^d \\ j_1 \neq \dots \neq j_d}} \xi_{j_1, \dots, j_d} z_{j_1} \cdots z_{j_d}, \end{aligned} \quad (2)$$

where $z_{j_1} \cdots z_{j_d}$ is the element-wise product and scalar ξ represents the coefficient. In this study, we consider each element of the original covariate vector $z_j \in \mathbb{R}^n$ is defined in the domain $[0, 1]^n$. To simplify the notation, it is convenient to write the high-order interaction model in (2) using the following matrix of concatenated vectors of all high-order interactions:

$$X = [\underbrace{z_1, \dots, z_m}_{1^{\text{st}} \text{ order}}, \cdots, \underbrace{z_1 \dots z_d, \dots, z_{m-d+1} \cdots z_m}_{d^{\text{th}} \text{ order}}] \in \mathbb{R}^{n \times p},$$

where $p := \sum_{\kappa=1}^{d} \binom{m}{\kappa}$. Similarly, the coefficient vector associated with all possible high-order interaction terms can be written as follows:

$$\beta := [\underbrace{\xi_1, \dots, \xi_m}_{1^{\text{st}} \text{ order}}, \cdots, \underbrace{\xi_{1,\dots,d}, \dots, \xi_{m-d+1,\dots,m}}_{d^{\text{th}} \text{ order}}]^\top \in \mathbb{R}^p.$$

The high-order interaction model (2) is then simply written as a linear model $y = X\beta$. Unfortunately, p can be prohibitively large unless both m and d are fairly small. In SHIM, we consider a sparse estimation of a high-order interaction model. An example of SHIM is as follows:

$$y = \xi_3 z_3 + \xi_5 z_5 + \xi_{2,6} z_2 z_6 + \xi_{1,2,5,9} z_1 z_2 z_5 z_9. \quad (3)$$

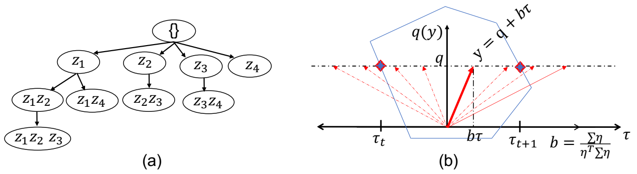


Figure 1: (a) A tree of patterns has been constructed by exploiting the hierarchical structure of high-order interaction features. (b) The conditional data space y of LASSO is restricted to a line parameterized by a scalar τ.

To quantify the reliability, the goal of this study is to fit a SHIM such as in (3) and test the statistical significance of the coefficients of the selected model (in the above example, $\xi_3, \xi_5, \xi_{2,6}, \xi_{1,2,5,9}$). Unfortunately, both the fitting and testing of a SHIM are non-trivial because, unless both m and d are very small, a high-order interaction model will have a significantly large number of parameters to be considered. Several algorithms for fitting a sparse high-order interaction model have been proposed in the literature (see Introduction). A common approach adopted in these existing works is to exploit the hierarchical structure of high-order interaction features. In other words, a tree structure as in Fig. 1(a) is considered and a branch-and-bound strategy is employed in order to avoid handling all the exponentially increasing number of high-order interaction features.

Here, we introduce an algorithm for conditional SI to quantify the statistical significance of the fitted coefficients of a SHIM such as $\xi_3, \xi_5, \xi_{2,6}, \xi_{1,2,5,9}$ in the forms of p-values or confidence intervals by applying a homotopy-based SI. However, owing to the extremely large number of features in (2), it is difficult to characterize the selection event for homotopy-based SI. To overcome this challenge, we develop a *homotopy mining* method that effectively combines the homotopy method and branch-and-bound strategy in the cherry tree. Before delving into our proposed method, we briefly overview the conditional SI.

Selective Inference and Homotopy Method

We present conditional selective inference (SI), which was introduced in Lee et al. (2016), and then demonstrate that an *optimal (i.e., minimally conditioned)* conditional SI can be conducted with a homotopy method. In the conditional SI framework, we assume that the design matrix X is fixed; the response vector y is a realization of the random response vector $Y \sim N(\mu, \Sigma)$, where $\mu \in \mathbb{R}^n$ is unknown mean vector; and $\Sigma \in \mathbb{R}^{n \times n}$ is a covariance matrix that is known or estimated from external data. In this framework, we do not assume a "true" relationship between X and μ, but consider a case in which the data analyst adopts the SHIM as a reasonable approximation model to describe the relationship.

Let $\mathcal{A}$ be the set of selected features by solving the SHIM fitting problem. With a slight abuse of notation, we also write this set of features as $\mathcal{A}(y)$ in order to emphasize that the set of features $\mathcal{A}$ is obtained when y is observed. This notation enables us to consider $\mathcal{A}(y')$ as the set of features that would be selected when a different response vector y' is observed. Furthermore, $\mathcal{A}(Y)$ represents the "random" set of features selected from the "random" response vector Y.

Given $\mathcal{A}$, we consider the best linear approximation of μ with the selected features. For $j \in \mathcal{A}$, let $\beta^*_j := e_j^\top (X_\mathcal{A}^\top X_\mathcal{A})^{-1} X_\mathcal{A}^\top \mu$ be the j^{th} population coefficient of the best linear approximation model fitted only with the selected features. Here, $e_j \in \mathbb{R}^{|\mathcal{A}|}$ is a vector with 1 at the j^{th} component and 0 otherwise. In the conditional SI framework, we consider the following hypothesis test:

$$\mathrm{H}_0 : \beta^*_j = 0 \quad \text{v.s.} \quad \mathrm{H}_1 : \beta^*_j \neq 0, \; j \in \mathcal{A}. \tag{4}$$

By defining $\eta := X_{\mathcal{A}(Y)}(X_{\mathcal{A}(Y)}^\top X_{\mathcal{A}(Y)})^{-1} e_j$, we can write $\beta^*_j = \eta^\top \mu$ with $Y = y$. It is therefore reasonable to use $\eta^\top Y$ as the test statistic for the test (4). The (unconditional) sampling distribution of $\eta^\top Y$ is highly complicated and intractable because η also depends on the random response vector Y through the selected features $\mathcal{A}(Y)$. The basic idea of conditional SI is to consider the sampling distribution of the test statistic conditional on the selection event, that is, $\eta^\top Y \mid \{\mathcal{A}(Y) = \mathcal{A}\}$. By further conditioning on the nuisance component $q(Y) = (I_n - b\eta^\top)Y$ with $b := \Sigma\eta(\eta^\top \Sigma \eta)^{-1}$ which is independent of the test statistic $\eta^\top Y$, Lee et al. (2016) showed that the conditional sampling distribution of $\eta^\top Y \mid \{\mathcal{A}(Y) = \mathcal{A}, q(Y) = q\}$ follows a truncated normal distribution

$$\eta^\top Y \mid \{\mathcal{A}(Y) = \mathcal{A}, q(Y) = q\} \sim F^{\mathcal{T}}_{\eta^T\mu, \eta^T \sum \eta}, \tag{5}$$

where $F^{\mathcal{T}}_{\tilde{\mu}, \tilde{\sigma}^2}$ is the c.d.f. of the truncated normal distribution with mean $\tilde{\mu}$, variance $\tilde{\sigma}^2$, and truncation region $\mathcal{T}$; and q is the observed nuisance component defined as $q = (I_n - b\eta^\top)y$. The reason why we condition on the nuisance component is that it is considered to be fixed because it is not of immediate interest. One more important point is that, under the assumption of Gaussian noise, the nuisance component is independent of the test statistic. The $q(Y)$ in our paper corresponds to the vector z in the seminal SI (SI) paper of Lee et al. (2016) (Sec 5, Eq 5.2 and Theorem 5.2 in (Lee et al. 2016)). Unfortunately, identifying the conditional data space $\{\mathcal{A}(Y) = \mathcal{A}, q(Y) = q\}$ is a challenging problem.

In Lee et al. (2016), the authors developed a practical algorithm to compute the truncated normal distribution by further conditioning on the signs of the selected features in $\mathcal{A}$. Although the validity of the inference can be maintained with this additional conditioning on the signs, it turns out that the power of the inference is *suboptimal* with this over-conditioning (Fithian, Sun, and Taylor 2014). Le Duy and Takeuchi (2021) recently developed an algorithm to resolve this issue by using the homotopy method. They showed that the homotopy method can be efficiently used to fully characterize the conditional data space and does not suffer from the computational burden of taking the unions of all possible signs. Their method is statistically more powerful as it does not over-condition on the signs. In particular, they considered a parametrized response vector (see Fig. 1 (b))

$$y(\tau) := q + b\tau \tag{6}$$

for a scalar parameter $\tau \in \mathbb{R}$, and solved the continuum of optimal solutions when the response vector y is replaced with $y(\tau)$ by using the homotopy method. Basically, they showed that the LASSO solutions $(\beta(\tau))$ with a variable response $(y(\tau))$ is piecewise linear in τ, and exploited this property to develop a homotopy algorithm to construct the exact regularization path $\tau \mapsto \beta(\tau)$. The details regarding this homotopy method are provided in the proposed method section. Therefore, we can redefine the conditional data space in (5) as

$$\mathcal{T} = \{\tau \in \mathbb{R} \mid \mathcal{A}(y(\tau)) = \mathcal{A}(y)\}. \tag{7}$$

This enables us to completely identify the truncation region of the truncated normal sampling distribution (by exactly identifying the pieces of the piecewise linear path $\tau \mapsto \beta(\tau)$, where $\mathcal{A}(y(\tau)) = \mathcal{A}(y)$) and compute the selective p-value

$$P_j^{\text{selective}} = 2\,\min\{\pi_j, 1-\pi_j\}, \pi_j = 1 - F^{\mathcal{T}}_{0,\eta^\top\Sigma\eta}(\eta^\top y). \tag{8}$$

Similarly, one can obtain $1-\alpha$ confidence interval $\mathcal{C}_\alpha$ for any $\alpha \in [0,1]$ such that

$$\mathbb{P}(\beta_j^* \in \mathcal{C}_\alpha \big| \{\mathcal{A}(Y) = \mathcal{A}, q(Y) = q\}) = 1-\alpha.$$

Unfortunately, in the case of a SHIM, because the number of high-order interaction features is exponentially large, we cannot use the same homotopy method. In the following section, we present the *homotopy mining algorithm* by exploiting the best of both *"homotopy"* and *"pattern mining"* methods. This enables us to compute the conditional sampling distribution (5) of the fitted SHIM coefficients by effectively combining the homotopy method and the branch-and-bound method used in pattern mining.

Proposed Method

In this study, we propose a similar *"homotopy-mining"* approach for model selection and inference. The homotopy method refers to an optimization framework for solving a sequence of parameterized optimization problems. The basic idea of our homotopy mining approach is to consider the following optimization problem with a parameterized response vector $y(\tau)$ in (6)

$$\beta(\lambda,\tau) = \underset{\beta\in\mathbb{R}^p}{\arg\min}\;\; \mathcal{F}_{\lambda,\tau}(\beta) := \frac{1}{2}\left\|y(\tau) - X\beta\right\|^2 + \lambda\left\|\beta\right\|_1, \tag{9}$$

where $\tau \in \mathbb{R}$ is a scalar parameter, λ is the regularization parameter for L_1-regularization, and the objective function $\mathcal{F}_{\lambda,\tau}(\beta)$ is parameterized by both τ and λ. Homotopy mining enables us to solve a sequence of parameterized optimization problems in the form of (9) by effectively combining the homotopy and mining methods.

To extend the homotopy selective inference framework for a SHIM, we first need to solve (9) for a fixed τ and target λ using the observed data and obtain an active set $\mathcal{A}$. Now, $\forall j \in \mathcal{A}$, we need to construct the exact solution path characterized by τ and then identify the conditional data space in (7) by identifying the intervals of τ on the solution path. This exact solution path can be constructed in a manner similar to the LARS-LASSO algorithm (Efron et al. 2004; Mairal and Yu 2012) using an efficient step size calculation. Here, we define the exact regularization paths $\lambda \mapsto \beta(\lambda)$ for a fixed τ as the "*λ-path*" and $\tau \mapsto \beta(\tau)$ for a fixed λ as the "*τ-path*", respectively. Then, these two paths of the SHIM can be constructed in a similar fashion as stated below:

• A model selection of the SHIM can be achieved using the exact regularization path algorithm

$$\lambda_0 > \lambda_1 > \cdots > \lambda_{\min} \Rightarrow \{\beta(\lambda_0), \beta(\lambda_1), \cdots, \beta(\lambda_{\min})\}. \tag{10}$$

• For inference, we can obtain a similar path algorithm

$$\tau_0 > \tau_1 > \cdots > \tau_{\min} \Rightarrow \{\beta(\tau_0), \beta(\tau_1), \cdots, \beta(\tau_{\min})\}, \tag{11}$$

where the sequences of λ and τ represent the breakpoints of the homotopy method (Efron et al. 2004; Mairal and Yu 2012). Equations (10) and (11) have similar problem structures, with the only difference being that in (10), we find the solution path characterized by the regularization parameter λ, whereas in (11), we find the solution path characterized by τ. Basically, what we need to characterize the selection event is finding those breakpoints (e.g., τ_0, τ_3, τ_8) along the τ-line where the active set remains the same as the observed one, that is, $\mathcal{A}(y) = \mathcal{A}(y(\tau_0)) = \mathcal{A}(y(\tau_3)) = \mathcal{A}(y(\tau_8))$. However, computing the exact solution paths for such a SHIM is a challenging task because of the exponentially expanded feature space (Le Morvan and Vert 2018; Suzumura et al. 2017). Efficient computational methods are required at both the selection and inference stages. Hence, we considered a tree structure (see Fig.1 (a)) of the interaction terms (or patterns) and proposed a tree pruning strategy to make it computationally tractable. In the next section, we present the main technical details of characterizing the conditional data space in (7) by using the homotopy-mining method.

Characterization of Truncation Region in SHIM

The optimal condition of (9) can be written as

$$X^\top\big(X\beta(\lambda,\tau) - y(\tau)\big) + \lambda s(\lambda,\tau) = 0,$$
$$\text{where}\;\; s_j(\lambda,\tau) \in \begin{cases}\{-1,+1\} & \text{if}\;\; \beta_j(\lambda,\tau) \neq 0,\\ [-1,+1] & \text{if}\;\; \beta_j(\lambda,\tau) = 0,\end{cases} \tag{12}$$

and $j \in [p]$. Let us define the active set of features as $\mathcal{A}(y(\tau)) = \{j \in [p] : \;\beta_j(\lambda,\tau) \neq 0\}$.

The τ-path (fixed λ). Because λ is fixed, we drop it from the notation. Now, consider two real values τ_t and τ_{t+1} $(\tau_{t+1} > \tau_t)$ at which the active set does not change and their signs also remain the same. For notational simplicity, we denote $\mathcal{A}_{\tau_t} = \mathcal{A}(y(\tau_t))$. Then, one can write from (12)

$$\beta_{\mathcal{A}_{\tau_t}}(\tau_{t+1}) - \beta_{\mathcal{A}_{\tau_t}}(\tau_t) = \nu_{\mathcal{A}_{\tau_t}}(\tau_t) \times (\tau_{t+1} - \tau_t), \tag{13}$$

$$\lambda s_{\mathcal{A}^c_{\tau_t}}(\tau_{t+1}) - \lambda s_{\mathcal{A}^c_{\tau_t}}(\tau_t) = \gamma_{\mathcal{A}^c_{\tau_t}}(\tau_t) \times (\tau_{t+1} - \tau_t), \tag{14}$$

where $\nu_{\mathcal{A}_{\tau_t}}(\tau) = (X^\top_{\mathcal{A}_{\tau_t}} X_{\mathcal{A}_{\tau_t}})^{-1} X^\top_{\mathcal{A}_{\tau_t}} b$ and $\gamma_{\mathcal{A}^c_{\tau_t}}(\tau) = X^\top_{\mathcal{A}^c_{\tau_t}} b - X^\top_{\mathcal{A}^c_{\tau_t}} X_{\mathcal{A}_{\tau_t}} \nu_{\mathcal{A}_{\tau_t}}(\tau)$ remain constant for all real values of $\tau \in [\tau_t, \tau_{t+1})$. Therefore, $\beta(\tau)$ and $\lambda s(\tau)$ are piecewise linear in τ for a fixed λ. If $\tau_{t+1} > \tau_t$ is the next zero crossing point, then either of the following two events occurs • A zero variable becomes non-zero, that is, $\exists j \in$

$\mathcal{A}^c_{\tau_t}$ s.t. $|x_j^\top(y(\tau_{t+1}) - X_{\mathcal{A}_{\tau_t}}\beta_{\mathcal{A}_{\tau_t}}(\tau_{t+1}))| = \lambda$ or, • A nonzero variable becomes zero, that is, $\exists j \in \mathcal{A}_{\tau_t}$ s.t. $\beta_j(\tau_t) \neq 0$ and $\beta_j(\tau_{t+1}) = 0$. Overall, the next change in the active set occurs at $\tau_{t+1} = \tau_t + \Delta_j$, where

$$\Delta_j = \min(\Delta_j^1, \Delta_j^2) = \min\left(\min_{j \in \mathcal{A}_{\tau_t}}\Big(-\frac{\beta_j(\tau_t)}{\nu_j(\tau_t)}\Big)_{++}, \min_{j \in \mathcal{A}^c_{\tau_t}}\Big(\lambda\frac{\text{sign}(\gamma_j(\tau_t)) - s_j(\tau_t)}{\gamma_j(\tau_t)}\Big)_{++}\right) . \quad (15)$$

Here, we use the convention that for any $a \in \mathbb{R}$, $(a)_{++} = a$ if $a > 0$ and ∞ otherwise. However, determining the step size (Δ_j) of the τ-path can be challenging for SHIM-type problems. Hence, we need efficient computational methods to make it practically feasible. In the following section, we present an efficient tree pruning strategy that considers the tree structure of the interaction terms. Here, each node of the tree represents an interaction term. A similar pruning strategy already exists in the literature to solve the λ-path of the LASSO in the context of graph mining (Tsuda 2007).

Tree Pruning

We first define the following tree anti-monotonicity property which we will use in the subsequent part of this article.

Property 1. A tree is constructed in such a way that for any pair of nodes (ℓ, ℓ'), where ℓ is the ancestor of ℓ', i.e., $\ell \subset \ell'$, the following conditions are satisfied

$$x_{i\ell'} = 1 \implies x_{i\ell} = 1 \text{ and, } x_{i\ell} = 0 \implies x_{i\ell'} = 0, \forall i \in [n].$$

Now considering the τ-path of the LASSO, the equicorrelation condition for any active feature $k \in \mathcal{A}_{\tau_{t+1}}$ at a fixed λ can be written as

$$\left|x_k^\top(y(\tau_{t+1}) - X\beta(\tau_{t+1}))\right| = \lambda.$$

Therefore at a fixed λ, any non-active feature $\ell \in \mathcal{A}^c_{\tau_t}$ becomes active at τ_{t+1} if

$$\left|x_\ell^\top\left(y(\tau_{t+1}) - X_{\mathcal{A}_{\tau_t}}(\beta_{\mathcal{A}_{\tau_t}}(\tau_t) + \Delta_\ell\nu_{\mathcal{A}_{\tau_t}}(\tau_t))\right)\right| = \left|x_k^\top\left(y(\tau_{t+1}) - X_{\mathcal{A}_{\tau_t}}(\beta_{\mathcal{A}_{\tau_t}}(\tau_t) + \Delta_\ell\nu_{\mathcal{A}_{\tau_t}}(\tau_t))\right)\right|,$$

or $$|\rho_\ell(\tau_t, \tau_{t+1}) - \Delta_\ell\eta_\ell(\tau_t)| = |\rho_k(\tau_t, \tau_{t+1}) - \Delta_\ell\eta_k(\tau_t)|, \quad (16)$$

where the left hand side (l.h.s.) corresponds to $\ell \in \mathcal{A}^c_{\tau_t}$ and the right hand side (r.h.s.) corresponds to $k \in \mathcal{A}_{\tau_t}$. Here, we define $\rho_\ell(\tau_t, \tau_{t+1}) = x_\ell^\top\Big(y(\tau_{t+1}) - X_{\mathcal{A}_{\tau_t}}\beta_{\mathcal{A}_{\tau_t}}(\tau_t)\Big)$ and $\eta_\ell(\tau_t) = x_\ell^\top X_{\mathcal{A}_{\tau_t}}\nu_{\mathcal{A}_{\tau_t}}(\tau_t)$. The r.h.s. of (16) has a lower bound, that is,

$$|\rho_k(\tau_t, \tau_{t+1}) - \Delta_\ell\eta_k(\tau_t)| \geq |\rho_k(\tau_t, \tau_{t+1})| - \Delta_\ell|\eta_k(\tau_t)|,$$

and the l.h.s. of (16) has an upper bound, i.e.,

$$|\rho_\ell(\tau_t, \tau_{t+1}) - \Delta_\ell\eta_\ell(\tau_t)| \leq |\rho_\ell(\tau_t, \tau_{t+1})| + \Delta_\ell|\eta_\ell(\tau_t)|.$$

Therefore, for equation (16) to obtain a solution, the following condition needs to be satisfied:

$$|\rho_\ell(\tau_t, \tau_{t+1})| + \Delta_\ell|\eta_\ell(\tau_t)| \geq |\rho_k(\tau_t, \tau_{t+1})| - \Delta_\ell|\eta_k(\tau_t)|. \quad (17)$$

Algorithm 1: τ-path

```
Input: $Z, \lambda, b, q, [\tau_{\min}, \tau_{\max}]$
Initialization: $t = 0, \tau_t = \tau_{\min}, \mathcal{T} = \{\tau_t\}, \beta(\tau_t) = 0$
$y(\tau_t) = q + b\tau_t, \quad \mathcal{A}_{\tau_t}, \beta_{\mathcal{A}_{\tau_t}}(\tau_t) \leftarrow \lambda$-path$(Z, y(\tau_t), \lambda)$
$\nu_{\mathcal{A}_{\tau_t}}(\tau_t) = (X^\top_{\mathcal{A}_{\tau_t}} X_{\mathcal{A}_{\tau_t}})^{-1} X^\top_{\mathcal{A}_{\tau_t}} b, \quad \nu_{\mathcal{A}^c_{\tau_t}}(\tau_t) = 0$
while $(\tau_t < \tau_{max})$ do
    Compute step-length $\Delta_j \leftarrow$ Equation (15)
    If $\Delta_j = \Delta_j^1$, remove $j$ from $\mathcal{A}_{\tau_t}$
    If $\Delta_j = \Delta_j^2$, add $j$ into $\mathcal{A}_{\tau_t}$
    Update: $\tau_{t+1} \leftarrow \tau_t + \Delta_j, \mathcal{T} = \mathcal{T} \cup \{\tau_{t+1}\}$,
        $\beta_{\mathcal{A}_{\tau_{t+1}}}(\tau_t) \leftarrow \beta_{\mathcal{A}_{\tau_t}}(\tau_t) + \Delta_j\nu_{\mathcal{A}_{\tau_t}}(\tau_t)$,
        $y(\tau_{t+1}) = q + b\tau_{t+1}$,
        $\nu_{\mathcal{A}_{\tau_{t+1}}}(\tau_{t+1}) = (X^\top_{\mathcal{A}_{\tau_{t+1}}} X_{\mathcal{A}_{\tau_{t+1}}})^{-1} X^\top_{\mathcal{A}_{\tau_{t+1}}} b$,
        $\nu_{\mathcal{A}^c_{\tau_{t+1}}}(\tau_{t+1}) = 0$
end while
Output: $\mathcal{T}, \{\mathcal{A}_{\tau_t}\}_{\tau_t \in \mathcal{T}}$
```

If the above condition (17) is not satisfied, then equation (16) does not have any solution, which can be used as a pruning condition. Therefore, the pruning condition is written as

$$|\rho_\ell(\tau_t, \tau_{t+1})| + \Delta_\ell|\eta_\ell(\tau_t)| < |\rho_k(\tau_t, \tau_{t+1})| - \Delta_\ell|\eta_k(\tau_t)|. \quad (18)$$

If $\Delta_\ell^* = \min_{t \in \{1,2,\ldots,\ell\}}\{\Delta_t\}$, is the current minimum step size, then we can further simplify (18) as stated in Lemma 1.

Lemma 1 *$\forall \ell' \supset \ell, \Delta_{\ell'} > \Delta_\ell^*$ if*

$$b_{\ell,w(\tau_t)} + \Delta_\ell^* b_{\ell,\theta} + \Delta_\ell^* b_{\ell,v(\tau_t)} < |\rho_k(\tau_t)| - \Delta_\ell^*|\theta_k| - \Delta_\ell^*|\eta_k(\tau_t)|, \quad (19)$$

where, $b_{\ell,w(\tau_t)} := \max\big\{\sum_{w_i(\tau_t)<0}|w_i(\tau_t)|x_{i\ell}, \sum_{w_i(\tau_t)>0}|w_i(\tau_t)|x_{i\ell}\big\}$; $b_{\ell,\theta}$ and $b_{\ell,v(\tau_t)}$ are similarly defined. Here, $w(\tau_t) = y(\tau_t) - X_{\mathcal{A}_{\tau_t}}\beta_{\mathcal{A}_{\tau_t}}(\tau_t), v(\tau_t) = X_{\mathcal{A}_{\tau_t}}\nu_{\mathcal{A}_{\tau_t}}(\tau_t)$ and $\theta_\ell = x_\ell^\top b$.

Therefore, Lemma 1 states that if $\exists \ell$ s.t. $\ell \subset \ell'$ and the condition in (19) is satisfied then one can safely ignore the subtree with ℓ as the root node. The complete algorithm for the τ-path is given in Algorithm 1.

Complexity analysis: If $X \in \mathbb{R}^{n \times p}$ is full rank and for a given λ, the τ-path of $\beta(\tau), \forall \tau \in [-\infty, +\infty]$ is well defined, then the worst-case complexity of the τ-path is 3^p. Note that the sign $s(\tau) \in \{-1, 0, +1\}^p$ of the coefficients $\beta(\tau)$ does not change between any two consecutive kinks of the piece-wise linear path $\tau \mapsto \beta(\tau)$. Hence, for p patterns the number of linear segments in τ-path is bounded by 3^p. For details see Mairal and Yu (2012). However, fortunately, it has been well-recognized that this worst-case rarely happens in practice (Le Duy and Takeuchi 2021) and, this is also evident from our experimental results (Table 2).

Extension for Elastic Net

We extended our proposed method to elastic net and solved the following optimization problem:

$$\beta(\lambda, \tau) \in \arg\min_{\beta \in \mathbb{R}^p} \frac{1}{2}\|y(\tau) - X\beta\|_2^2 + \frac{1}{2}\alpha\|\beta\|_2^2 + \lambda\|\beta\|_1 .$$

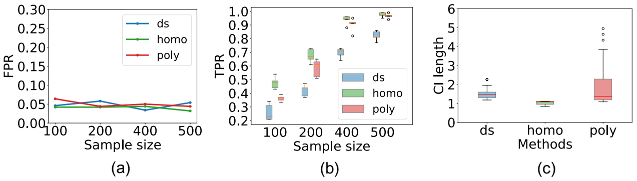


Figure 2: Demonstration of the statistical power of three selection bias correction methods (ds, data splitting; homo, homotopy; poly, polytope) using synthetic data experiments. (a) and (b) False positive rates (FPR) and the true positive rates (TPR) for different sample sizes, and (c) shows the distribution of the confidence interval lengths (CI length).

Note that here the only change compared to the LASSO is the addition of an $\alpha I_{|\mathcal{A}_{\tau_t}|}$ term to the expression of the direction vector, i.e. $\nu_{\mathcal{A}_{\tau_t}}(\tau_t) = (X_{\mathcal{A}_{\tau_t}}^\top X_{\mathcal{A}_{\tau_t}} + \alpha I_{|\mathcal{A}_{\tau_t}|})^{-1} X_{\mathcal{A}_{\tau_t}}^\top b$. Similar to the LASSO (19), the pruning condition for the τ-path of ElNet can be written as follows: $\exists \ell,$ such that $\forall \ell' \supset \ell, \Delta_{\ell'} > \Delta_\ell^*$, if

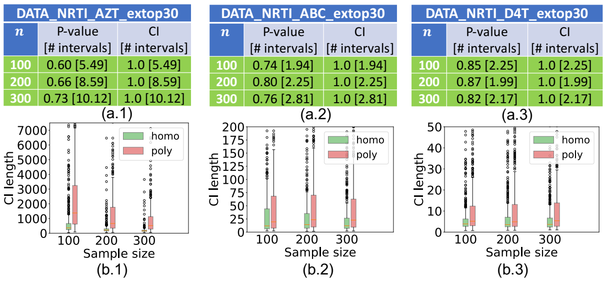
$$b_{\ell,w(\tau_t)} + \Delta_\ell^* b_{\ell,\theta} + \Delta_\ell^* b_{\ell,v(\tau_t)} < |\bar{\rho}_k(\tau_t)| - \Delta_\ell^* |\theta_k| - \Delta_\ell^* |\bar{\eta}_k(\tau_t)|, \quad (20)$$

where $\bar{\rho}_k(\tau_t) = \sum_{i=1}^n w_i(\tau_t) x_{ik} - \alpha\beta_k$, $\bar{\eta}_k(\tau_t) = \sum_{i=1}^n v_i(\tau_t) x_{ik} + \alpha\nu_k$.

Experiments

Comparison of Statistical Powers

Synthetic data: We generated random samples $(z_i, y) \in [0,1]^m \times \mathbb{R}$ in such a way that $100m(1-\zeta)\%$ of $z_i \in \mathbb{R}^m$ contain a value of 1 on average. Here, $\zeta \in [0,1]$ is the sparsity controlling parameter. The response $y_i \in \mathbb{R}$ is randomly generated from a normal distribution $N(0, \sigma^2)$. For a comparison of the false positive rates (FPRs), true positive rates (TPRs), and confidence intervals (CIs) across different methods, we generated a design matrix for a fixed sparsity parameter $\zeta = 0.95$. During all experiments, the significance level was set at 0.05. To compare the TPRs, we considered a true model of up to third-order interactions, which is defined as $\mu(x_i) = 0.5z_1 - 2z_2z_3 + 3z_4z_5z_6$. The response y_i is accordingly generated from $N(\mu(X), \sigma^2 I)$. For the comparison of FPRs, we set $\beta_j = 0, \forall j \in \mathbb{R}^p$. We compared both FPRs and TPRs across three different methods, i.e. (ds, data splitting; homo, homotopy; poly, polytope) for four different sample sizes $n \in [100, 200, 400, 500]$. We generated TPRs and FPRs over 100 trials for all three methods and repeated the experiments for 5 times. The results are shown in Fig. 2(a) and Fig. 2(b), respectively. It can be seen that all SI methods can properly control the FPRs under 0.05. Regarding the TPR comparison, homotopy has the highest power, which is obvious because it is minimally conditioned compared to polytope which suffers from the over conditioning issue. Comparing the TPRs of the data splitting (ds) and homotopy (homo), it can be seen that the TPRs of homo are always greater than those of ds. Note that in ds, only half of the data are used for selection and the remaining half are used for the inference. Therefore, compared to homo, ds has a higher risk of failing to identify truly correlated features in the selection stage and similarly suffers from low statistical power in the inference stage. The results of the CIs are shown in Fig. 2(c). Here, we used the same true model of the TPR experiments and reported the average CIs over 100 trials across different methods. The results of the CIs are consistent with the findings of the TPRs.



DATA_NRTI_AZT_extop30

n	P-value [# intervals]	CI [# intervals]
100	0.60 [5.49]	1.0 [5.49]
200	0.66 [8.59]	1.0 [8.59]
300	0.73 [10.12]	1.0 [10.12]

(a.1)

DATA_NRTI_ABC_extop30

n	P-value [# intervals]	CI [# intervals]
100	0.74 [1.94]	1.0 [1.94]
200	0.80 [2.25]	1.0 [2.25]
300	0.76 [2.81]	1.0 [2.81]

(a.2)

DATA_NRTI_D4T_extop30

n	P-value [# intervals]	CI [# intervals]
100	0.85 [2.25]	1.0 [2.25]
200	0.87 [1.99]	1.0 [1.99]
300	0.82 [2.17]	1.0 [2.17]

(a.3)

Figure 3: Comparison of statistical powers (homotopy vs polytope). (a.1-a.3) The percentage of cases where selection bias corrected p-values and confidence interval lengths of the proposed method (homotopy) were smaller than those of the existing method (polytope) in the random sub-sampling experiments. (b.1-b.3) The distributions of the confidence interval lengths of the same experiments. The numbers inside the brackets represent the average number of intervals along the τ-line considered for the homotopy method. Note that in the case of polytope only one such interval is considered.

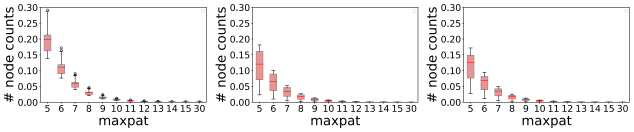
Real data: We obtained HIV-1 sequence data from the Stanford HIV Drug Resistance Database (Rhee et al. 2003). In our experiment, we used six NRTIs, one NNRTIs and three PIs drugs. However, because of the space limitations, we reported only the results of three NRTIs drugs. To demonstrate the statistical efficacy of the proposed homotopy method over the existing polytope method, we generated random sub-samples of the 10 drug data as follows. First, we created a dataset consisting of the top 30 mutations from each of the 10 drug data. Because most of the columns contain zeros, we sorted them based on the number of 1s present in each column and selected the top-30 columns as our starting set. Then, from this starting set, we considered random sub-samples of five features for three different sample sizes ($n \in \{100, 200, 300\}$). Here, we considered randomization without replacement for both sample and features selection. We generated 100 samples and repeated the experiments for five times; hence, we generated 500 samples in total. Figure 3 demonstrates the percentage of times homotopy produced smaller p-values and CI lengths than polytope. This also depicts the distributional difference of the CI lengths between homotopy and polytope. These results clearly demonstrate that homotopy is statistically more powerful than the existing polytope method.

Comparison of Computational Efficiencies

To demonstrate the computational efficiency of the proposed pruning strategy for the τ-path, we applied our homotopy mining method with and without pruning on the HIV NRTI

d	Search space (# nodes)	With pruning			Without pruning		
		1^{st}	2^{nd}	3^{rd}	1^{st}	2^{nd}	3^{rd}
5	174436	19.53 ± 7.91	8.69 ± 2.71	9.45 ± 3.91	25.29 ± 2.50	34.80 ± 1.19	23.34 ± 1.88
6	768211	49.47 ± 23.41	20.39 ± 7.85	22.74 ± 14.74	126.96 ± 8.61	125.29 ± 2.14	127.97 ± 4.80
7	2804011	92.15 ± 44.27	39.28 ± 17.08	43.43 ± 32.73	450.24 ± 28.50	447.59 ± 22.15	447.19 ± 37.69
8	8656936	151.71 ± 76.53	57.69 ± 26.67	64.59 ± 50.52	> 1 day	> 1 day	> 1 day
9	22964086	209.23 ± 101.84	74.58 ± 36.24	84.83 ± 62.48	> 1 day	> 1 day	> 1 day
10	53009101	239.28 ± 126.02	75.40 ± 35.09	90.38 ± 69.36	> 1 day	> 1 day	> 1 day
11	107636401	289.99 ± 140.12	98.19 ± 49.19	116.44 ± 90.12	> 1 day	> 1 day	> 1 day
12	194129626	303.73 ± 149.75	105.11 ± 52.70	118.91 ± 92.44	> 1 day	> 1 day	> 1 day
13	313889476	311.92 ± 154.93	108.31 ± 54.27	122.97 ± 95.32	> 1 day	> 1 day	> 1 day
14	459312151	317.30 ± 154.47	98.02 ± 51.63	114.18 ± 82.37	> 1 day	> 1 day	> 1 day
15	614429671	318.91 ± 155.89	100.52 ± 53.37	116.06 ± 84.58	> 1 day	> 1 day	> 1 day
None	1073741823	317.09 ± 155.92	110.0 ± 55.31	126.35 ± 97.08	> 1 day	> 1 day	> 1 day

Table 1: Computation time (in s) with and without pruning for first, second and third order interactions. Here, the computation time is measured against different maximum pattern size (d) constraints. The last row corresponds to the case in which "d" is not specified, and the entire search space is used for exploration. All computation times were measured on an Intel Xeon Gold 6230 CPU @ 2.10GHz.



(a.1) 1^{st} order interaction (a.2) 2^{nd} order interaction (a.3) 3^{rd} order interaction

Figure 4: Distribution of the fraction of total nodes traversed against different maximum pattern size (d) constraints while applying the proposed pruning method during the construction of the τ-path. (a.1) - (a.3) The results for the first, second and third order interaction terms.

D4T drug resistance data with the same starting set of top-30 mutations (i.e. $m = 30$) as used to demonstrate the statistical power. Although we varied the d from 5 to m, high-order interaction terms up to the third-order appeared in $\mathcal{A}$. We compared both the number of nodes traversed (Fig.4) and the time taken (Table 1) against a different maximum interaction order d during the construction of the τ-path of each test statistic direction. Empirically, pruning was found to be more effective for the τ-path of high-order interaction terms compared to that of the singleton terms, and the power of pruning increases as the order of the interaction increases.

Therefore, we reported the average number of nodes and average time taken separately for first, second and third order interaction terms. It can be observed that the pruning is more effective at the deeper nodes of the tree and saturates after a certain depth of the tree. This is evident as the sparsity of the data increases at the deeper nodes, and the pruning exploits the anti-monotonicity of high-order interaction terms constructed as tree of patterns. In the case of the homotopy method without pruning, we stopped the execution of the program if the τ-path was not finished in one day. From Table 1, it can be observed that without the tree pruning, the construction of the τ-path is not practical as we progress to the deeper nodes of the tree because of the generation of an exponential number of high-order interaction terms. The τ-path without pruning took more than a day beyond $d = 7$, whereas the maximum time taken by the τ-path with pruning was approximately 317 s on average, even when no d constraint was imposed. In Table. 2 we demonstrated the computational advantage of the proposed homotopy mining method over exiting method on conditioning on model (Lee et al. 2016). Here, we considered an active set ($\mathcal{A}$) of size 20. The Lee et al. (2016) method needs to consider the union of all possible signs in the observed active set ($\mathcal{A}$) in order to condition on the model. However, our homotopy mining needs to consider only ~ 120 polytopes (worst case).

High-order interactions	Homotopy (# kinks)	Polytope (# polytopes)
1^{st}	104.15 ± 10.73	2^{20}
2^{nd}	101.0 ± 4.64	2^{20}
3^{rd}	78.33 ± 24.69	2^{20}

Table 2: Comparison of computational efficiencies of the proposed homotopy method against existing polytope method. The "# kinks" represents the average number of kinks encountered in the τ-path for each test statistic direction, whereas the "# polytopes" represents the number of all possible signs needed to condition on the model.

Conclusions

In this paper, we presented an algorithm for testing a sparse high-order interaction model (SHIM) using the framework of conditional selective inference (SI). The algorithm was developed by effectively combining the homotopy and branch-and-bound tree mining methods to deal with the combinatorial computational burden of the SHIM and also to improve the statistical power.

Acknowledgements

This work was partially supported by MEXT KAKENHI (20H00601, 16H06538), JST CREST (JPMJCR21D3), JST

Moonshot R&D (JPMJMS2033-05), NEDO (JPNP18002, JPNP20006) and RIKEN Center for Advanced Intelligence Project.

References

Chen, S.; and Bien, J. 2020. Valid inference corrected for outlier removal. *Journal of Computational and Graphical Statistics*, 29(2): 323–334.

Choi, Y.; Taylor, J.; and Tibshirani, R. 2017. Selecting the number of principal components: Estimation of the true rank of a noisy matrix. *The Annals of Statistics*, 2590–2617.

Das, D.; Ito, J.; Kadowaki, T.; and Tsuda, K. 2019. An interpretable machine learning model for diagnosis of Alzheimer's disease. *PeerJ*, 7: e6543.

Duy, V. N. L.; and Takeuchi, I. 2021a. Exact Statistical Inference for the Wasserstein Distance by Selective Inference. arXiv:2109.14206.

Duy, V. N. L.; and Takeuchi, I. 2021b. More Powerful Conditional Selective Inference for Generalized Lasso by Parametric Programming. arXiv:2105.04920.

Efron, B.; Hastie, T.; Johnstone, I.; Tibshirani, R.; et al. 2004. Least angle regression. *The Annals of statistics*, 32(2): 407–499.

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

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

Hall, M. A. 1999. *Correlation-based feature selection for machine learning*. Ph.D. diss., Department of Computer Science, Waikato University, New Zealand.

Hyun, S.; Lin, K. Z.; G'Sell, M.; and Tibshirani, R. J. 2018. Post-selection inference for changepoint detection algorithms with application to copy number variation data. *Biometrics*.

Iversen, A.; Shafer, R. W.; Wehrly, K.; Winters, M. A.; Mullins, J. I.; Chesebro, B.; and Merigan, T. C. 1996. Multidrug-resistant human immunodeficiency virus type 1 strains resulting from combination antiretroviral therapy. *Journal of Virology*, 70(2): 1086–1090.

Le Duy, V. N.; and Takeuchi, I. 2021. Parametric programming approach for more powerful and general lasso selective inference. In *International Conference on Artificial Intelligence and Statistics*, 901–909. PMLR.

Le Morvan, M.; and Vert, J.-P. 2018. WHInter: A Working set algorithm for High-dimensional sparse second order Interaction models. In *International Conference on Machine Learning*, 3635–3644. PMLR.

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

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

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

Loftus, J. R.; and Taylor, J. E. 2015. Selective inference in regression models with groups of variables. *arXiv preprint arXiv:1511.01478*.

Mairal, J.; and Yu, B. 2012. Complexity analysis of the lasso regularization path. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, 1835–1842.

Nakagawa, K.; Suzumura, S.; Karasuyama, M.; Tsuda, K.; and Takeuchi, I. 2016. 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*, 1785–1794.

Panigrahi, S.; Taylor, J.; and Weinstein, A. 2016. Bayesian post-selection inference in the linear model. *arXiv preprint arXiv:1605.08824*, 28.

Rendle, S. 2010. Factorization machines. In *2010 IEEE International conference on data mining*, 995–1000. IEEE.

Rhee, S.-Y.; Gonzales, M. J.; Kantor, R.; Betts, B. J.; Ravela, J.; and Shafer, R. W. 2003. Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic acids research*, 31(1): 298–303.

Rhee, S.-Y.; Taylor, J.; Wadhera, G.; Ben-Hur, A.; Brutlag, D. L.; and Shafer, R. W. 2006. Genotypic predictors of human immunodeficiency virus type 1 drug resistance. *Proceedings of the National Academy of Sciences*, 103(46): 17355–17360.

Saigo, H.; Nowozin, S.; Kadowaki, T.; Kudo, T.; and Tsuda, K. 2009. gBoost: a mathematical programming approach to graph classification and regression. *Machine Learning*, 75(1): 69–89.

Saigo, H.; Uno, T.; and Tsuda, K. 2007. Mining complex genotypic features for predicting HIV-1 drug resistance. *Bioinformatics*, 23(18): 2455–2462.

Sugiyama, K.; Le Duy, V. N.; and Takeuchi, I. 2021. More powerful and general selective inference for stepwise feature selection using homotopy method. In *International Conference on Machine Learning*, 9891–9901. PMLR.

Sugiyama, R.; Toda, H.; Duy, V. N. L.; Inatsu, Y.; and Takeuchi, I. 2021. Valid and Exact Statistical Inference for Multi-dimensional Multiple Change-Points by Selective Inference. arXiv:2110.08989.

Suzumura, S.; Nakagawa, K.; Umezu, Y.; Tsuda, K.; and Takeuchi, I. 2017. Selective inference for sparse high-order interaction models. In *International Conference on Machine Learning*, 3338–3347. PMLR.

Taylor, J.; and Tibshirani, R. J. 2015. Statistical learning and selective inference. *Proceedings of the National Academy of Sciences*, 112(25): 7629–7634.

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

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

Tsuda, K. 2007. Entire Regularization Paths for Graph Data. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, 919–926. New York, NY, USA: Association for Computing Machinery. ISBN 9781595937933.

Vivet-Boudou, V.; Didierjean, J.; Isel, C.; and Marquet, R. 2006. Nucleoside and nucleotide inhibitors of HIV-1 replication. *Cellular and Molecular Life Sciences CMLS*, 63(2): 163–186.

Yang, F.; Foygel Barber, R.; Jain, P.; and Lafferty, J. 2016. Selective inference for group-sparse linear models. *Advances in neural information processing systems*, 29.

Zou, H.; and Hastie, T. 2005. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2): 301–320.