High-Dimensional Interaction Detection with False Sign Rate Control *

Daoji Li[1], Yinfei Kong[1], Yingying Fan[2], and Jinchi Lv[2]

[1]Department of Information Systems and Decision Sciences, College of Business and Economics, California State University, Fullerton
[2]Data Sciences and Operations Department, Marshall School of Business, University of Southern California

Abstract

Identifying interaction effects is fundamentally important in many scientific discoveries and contemporary applications, but it is challenging since the number of pairwise interactions increases quadratically with the number of covariates and that of higher-order interactions grows even faster. Although there is a growing literature on interaction detection, little work has been done on the prediction and false sign rate on interaction detection in ultrahigh-dimensional regression models. This paper fills such a gap. More specifically, in this paper we establish some theoretical results on interaction selection for ultrahigh-dimensional quadratic regression models under random designs. We prove that the examined method enjoys the same oracle inequalities as the lasso estimator and further admits an explicit bound on the false sign rate. Moreover, the false sign rate can be asymptotically vanishing. These new theoretical characterizations are confirmed by simulation studies. The performance of our proposed approach is further illustrated through a real data application.

*This work was supported by NSF CAREER Award DMS-1150318, NSF Grant DMS-1953356, a grant from the Simons Foundation, Adobe Data Science Research Award, and 2020 individual Award (0358220) from the Innovative Research and Creative Activities Grant at California State University, Fullerton. The authors would like to thank the Joint Editor, Associate Editor, and referees for their constructive comments that have helped improve the paper significantly.

Running title: Interaction Detection

Key words: Interaction screening; Interaction selection; High dimensionality; Sure independence screening; Nonconvex learning; False sign rate

1 Introduction

Understanding how features interact with each other is fundamentally important in many scientific discoveries and contemporary applications, especially in areas such as medicine, genetics, and cancer studies (Xu et al., 2004; Musani et al., 2007; Cordell, 2009; Gosik et al., 2018). Identifying important interactions can also help improve model interpretability and prediction. With rapid development of information technologies, high-dimensional data are increasingly encountered in many scientific fields. Recent years have also seen a surge of interests on high-dimensional data in business and economics (Fan et al., 2016; Belloni et al., 2018; Cattaneo et al., 2019; Uematsu and Tanaka, 2019; Ke et al., 2020; Zheng et al., 2021). However, interaction identification with high-dimensional data poses great challenges since the number of pairwise interactions increases quadratically with the number of covariates.

Generally speaking, most of existing approaches for interaction models with continuous response can be categorized into two types: one-step approaches and stagewise approaches. A typical idea of one-step approaches is to select the main and interaction terms simultaneously by using regularization methods with specifically designed penalty functions or imposing inequality or convex constraints; see, for example, Yuan et al. (2009), Choi et al. (2010), Bien et al. (2013), Yan and Bien (2017) and references therein. Recently, Zhao and Leng (2016) presented a unified analysis on the convergence rate for a class of penalized estimators for quadratic regression models with random design. These methods can suffer from prohibitively high computational cost because they need to deal with complex penalty structures or multiple inequality constraints and thus are not feasible for ultrahigh-dimensional data.

The stagewise selection approaches first reduce the number of interactions and main

effects to a moderate scale and then select important interactions and main effects in the reduced feature space. There are two different ways for existing stagewise selection approaches. The first way is to impose the heredity assumption. The strong heredity assumption requires that an interaction between two covariates be included in the model only if both main effects are important, while the weak one relaxes such as a constraint to the presence of at least one main effect being important. Examples include the two-step recursive approach (Hall and Xue, 2014), the forward selection based procedure (Hao and Zhang, 2014), and the two-stage regularization method based on the lasso (Hao et al., 2018). However, the heredity assumptions can be violated or difficult to verify in real applications (Ritchie et al., 2001; Cordell, 2009; Gosik et al., 2018). This motivates the second way of stagewise approaches for interaction detection, which do not require the heredity assumptions. For example, Jiang and Liu (2014) considered interaction detection for the sliced inverse index models by screening interaction variables, which are variables contributing to interactions, instead of main effects and established the sure screening property under the normality assumption of covariates on each slice. Kong et al. (2017) introduced the method of interaction pursuit with distance correlation to select important interactions in high-dimensional multi-response regression models, which exploits feature screening applied to transformed variables with distance correlation (Székely et al., 2007; Li et al., 2012) followed by feature selection. See Section 2.1 for detailed description on the method of interaction pursuit with distance correlation. More recently, Tian and Feng (2021) suggested a new framework for variable screening via random subspace ensembles, which evaluates the contribution of variables through the joint contributions in different subspaces and can be used for interaction screening without the heredity assumptions.

Although there is a growing literature on interaction detection as discussed above, little work has been done on the prediction and false sign rate on interaction detection. In this paper, we establish some global theoretical results on the oracle inequalities and false sign rate on interaction selection for a class of stagewise selection approaches with random design in high-dimensional settings. Unlike selective inference or conditional

inference in which theoretical results are conditional on the selected model obtained from the first stage, our goal is to establish global bounds for prediction error, estimation error, and false sign rate. The major challenge in establishing these global theoretical results is that the set of selected predictors after the first stage is random and we need to control all possible random sets.

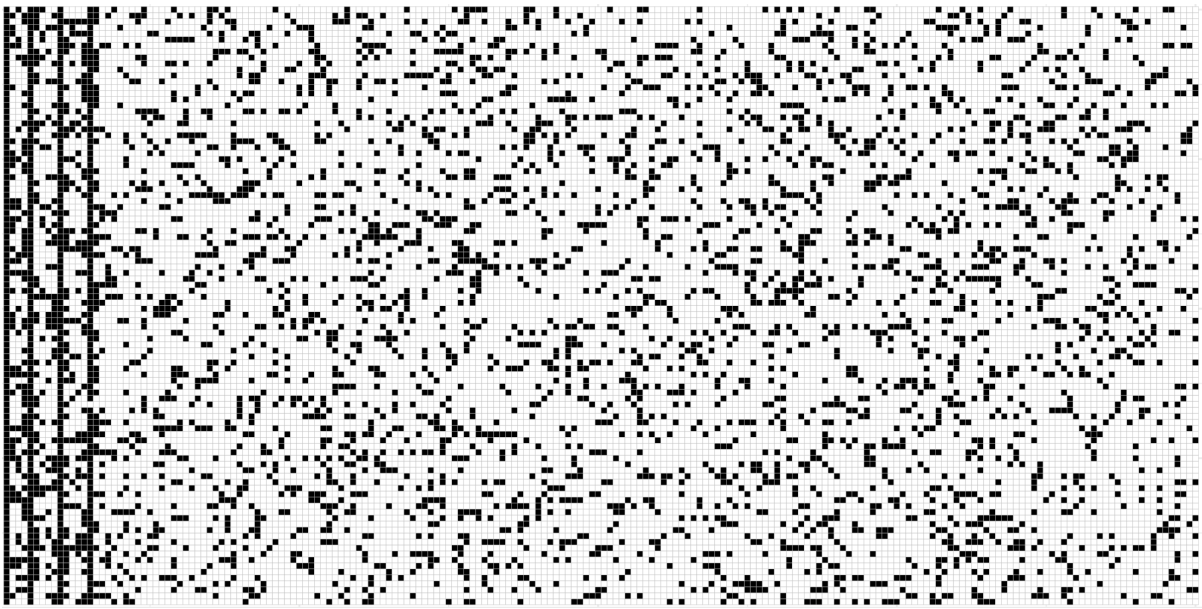

Figure 1: Heat map of retained variables after the feature screening step of the interaction pursuit approach in Kong et al. (2017). Each row represents one repetition while each column represents one variable. Black dots denote retained variables.

To further illustrate this point, consider the interaction model $Y = 3X_1X_5 + 3X_{10}X_{15} + \varepsilon$. There are four important variables X_1, X_5, X_{10}, and X_{15}. We simulated data in the same way as that for model 4 in Section 3 except that $p = 200$ and $\rho = 0.5$. Figure 1 shows the heat map of the union of the retained main effects and interaction variables after the feature screening step of the interaction pursuit approach in Kong et al. (2017). As shown in Figure 1, the retained set after the feature screening step is random. Therefore, existing results on the prediction and false sign rate for fixed design in the literature are no longer applicable here. To address this issue, we establish new theory which accounts for the uncertainty related to the random set of variables resulted from the first stage (e.g., the screening step). Our theoretical results are obtained by first showing

that the sparse eigenvalue condition and restricted eigenvalue condition hold for the full augmented random design matrix (which includes both main effects and interactions) with high probability, where the covariates in main effects are not required to follow the Gaussian or sub-Gaussian distributions. This result plays a key role in our technical analysis for the main effect and interaction selection. It can also be of independent interest for other two-step procedures.

Our goal in this paper is to establish more comprehensive theory on feature selection for a class of stagewise selection approaches for quadratic regression models with random design in high-dimensional settings. We prove that the resulting estimator enjoys the oracle inequalities under the prediction loss and L_d loss, with $1 \leq d \leq 2$, as well as an asymptotically vanishing bound on the false sign rate. The key novelty of our selection results is that we allow for random support for global inference instead of local inference conditional on the selected models. Thus we can avoid the use of sample splitting. In contrast, other existing methods usually need to split the data into two parts to obtain the same results, which can lead to loss of efficiency.

Among the existing literature, a most closely related paper is Zhao and Leng (2016). Our work differs significantly from theirs in the following aspects: First, Zhao and Leng (2016) provided only the ℓ_1 bound for estimation error, while we establish the ℓ_d bound with $d \in [1, 2]$ for estimation error and bounds for prediction error and false sign rate. To the best of our knowledge, our result on false sign rate for interaction models is new to the literature. Second, Zhao and Leng (2016) considered the interaction model (1) without the first stage (for example, the screening step), and thus their results cannot be applied to stagewise approaches; while our results are applicable to both one-step and stagewise approaches. Third, we have weaker assumptions on the dependence structure among covariates and allow heavier-tailed distributions in covariates and error. As a result, our theory covers broader scenarios. We will provide more detailed comparisons in Section 2.4 after we present the conditions and results.

The rest of the paper is organized as follows. Section 2 provides one condition to

characterize the key idea of a class of stagewise selection approaches, exploits the regularization methods to further select important interactions and main effects in reduced feature space, and studies the theoretical properties on variable selection. Sections 3 and 4 demonstrate the advantage of our proposed approach through simulation studies and a real data application, respectively. We discuss some extensions of our work in Section 5. The proofs of some main results are relegated to the Appendix. Additional simulation studies and technical details are provided in the supplementary material.

2 Interaction Detection

We will first introduce the model setting and then discuss one condition which holds for most of stagewise approaches. We also exploit the regularization methods to further select important interactions and main effects in reduced feature space, and study the theoretical properties on variable selection in this Section.

2.1 Model Setting

Consider the interaction model

$$Y = \alpha_0 + \sum_{j=1}^{p} \beta_j X_j + \sum_{k=1}^{p-1} \sum_{\ell=k+1}^{p} \gamma_{k\ell} X_k X_\ell + \varepsilon, \tag{1}$$

where Y is the response variable, $x = (X_1, \ldots, X_p)^T$ is a p-vector of random covariates X_j's, α_0 is the intercept, β_j's and $\gamma_{k\ell}$'s are regression coefficients for main effects and interactions, respectively, and ε is the mean zero random error independent of X_j's. Throughout the paper, we assume that $E(X_j) = 0$ for each random covariate X_j. Otherwise, consider the interaction model (1) with each X_j replaced by $X_j - E(X_j)$. Without loss of generality, we also assume that $\mathrm{var}(X_j) = 1$.

Denote by $(\beta_{0,j})_{1\le j\le p}$ and $(\gamma_{0,k\ell})_{1\le k<\ell\le p}$ the true regression coefficient vectors for main effects and interactions, respectively. To ease the presentation, throughout the paper $X_k X_\ell$ is referred to as an *important interaction* if its regression coefficient $\gamma_{0,k\ell}$ is

nonzero, and X_k is called an *active interaction variable* if there exists some $1 \leq \ell \neq k \leq p$ such that $X_k X_\ell$ is an important interaction. Define three sets of indices

$$\begin{aligned} \mathcal{I} &= \{(k,\ell) : 1 \leq k < \ell \leq p \text{ with } \gamma_{0,k\ell} \neq 0\}, \\ \mathcal{A} &= \{1 \leq k \leq p : (k,\ell) \text{ or } (\ell,k) \in \mathcal{I} \text{ for some } \ell\}, \\ \mathcal{B} &= \{1 \leq j \leq p : \beta_{0,j} \neq 0\}. \end{aligned} \tag{2}$$

The set $\mathcal{I}$ contains all important interactions and set $\mathcal{A}$ consists of all active interaction variables, while set $\mathcal{B}$ is comprised of all important main effects. We combine sets $\mathcal{A}$ and $\mathcal{B}$, and define the set of important features as $\mathcal{M} = \mathcal{A} \cup \mathcal{B}$. It is straightforward to see that sets $\mathcal{A}$, $\mathcal{I}$, and $\mathcal{M}$ are invariant under affine transformations $X_j^{new} = b_j(X_j - a_j)$ with $a_j \in \mathbb{R}$ and $b_j \in \mathbb{R} \setminus \{0\}$ for $1 \leq j \leq p$. Thus there is no issue of identifiability when recovering interactions in $\mathcal{I}$ and variables in $\mathcal{M}$.

For high-dimensional interaction models with multivariate response, which includes our model (1) as a special case, Kong et al. (2017) proposed a new interaction screening approach, where variables in $\mathcal{A}$ and $\mathcal{B}$ are identified by distance correlations $\text{dcorr}(X_j^2, Y^2)$ and $\text{dcorr}(X_j, Y)$, respectively. Here dcorr stands for distance correlation (Székely et al., 2007) between two random vectors. They showed that this screening method enjoys the sure screening property (Fan and Lv, 2008), meaning that all important interactions and all covariates that contribute to important interactions or main effects can be retained with asymptotic probability one. However, the global theoretical properties of this stage-wise method of screening followed by selection is not well-understood, even in the single response setting. In this paper, we intend to provide such theoretical guarantee under the interaction model (1). We will focus on the prediction and false sign rate on feature selection for random design when p grows exponentially with n.

Assume that we are given a sample $\{(x_i^T, Y_i), i = 1, \ldots, n\}$ of n independent and identically distributed observations from (x^T, Y) in interaction model (1). We rewrite

interaction model (1) in the matrix form

$$y = \alpha_0 1_n + Z\theta + \varepsilon, \tag{3}$$

where $y = (Y_1, \ldots, Y_n)^T$ is the response vector, 1_n is an n-dimensional column vector with all elements being 1, $\theta = (\theta_1, \ldots, \theta_q)^T$ is a parameter vector consisting of $q = p(p+1)/2$ regression coefficients β_j and $\gamma_{k\ell}$, Z is the corresponding $n \times q$ augmented design matrix incorporating the covariate vectors for X_j's and their interactions in columns, and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T$ is the error vector. Hereafter, for the simplicity of presentation and theoretical derivations, we slightly abuse the notation and still use y and Z to denote the de-meaned response vector and column de-meaned design matrix, respectively, which leads to $\alpha_0 = 0$.

As discussed before, there are many developments on stagewise approaches for interaction detection. A common idea for most of stagewise approaches is to reduce the number of interactions and main effects to a moderate scale and then select important interactions and main effects in the reduced feature space. The key to guaranteeing the success of the selection stage in these approaches is to ensure that all important interactions and main effects are retained in the first stage with probability tending to one. This property is called sure screening property in the feature screening literature (Fan and Lv, 2008). Thus, it is reasonable to directly assume that the sure screening property holds if one has no preference on which method should be used to reduce the number of interactions and main effects in the first stage.

2.2 Condition 1 and Verification

Let $\widehat{\mathcal{I}}$ and $\widehat{\mathcal{M}}$ be the estimators of $\mathcal{I}$ and $\mathcal{M}$ in the first stage (for example, screening step) of a user-specified stagewise approach for interaction detection. In general, both $\widehat{\mathcal{I}}$ and $\widehat{\mathcal{M}}$ can still contain many noise variables. Throughout this paper, we make the following assumption.

Condition 1. *There exist some constant* $C > 0$ *and* $0 < \eta < 1$ *such that*

$$P\Big(\mathcal{I} \subset \widehat{\mathcal{I}} \quad \textit{and} \quad \mathcal{M} \subset \widehat{\mathcal{M}}\Big) = 1 - o\Big(n^{-C}\Big)$$

for $\log p = o(n^{\eta})$.

As discussed before, Condition 1 is a sure screening property, which holds for many stagewise interaction detection approaches, such as those in Hall and Xue (2014); Jiang and Liu (2014); Hao and Zhang (2014); Kong et al. (2017); Zhou et al. (2019). We show in Lemma 1 in Appendix B that Condition 1 holds under some sufficient conditions.

2.3 Interaction Models in Reduced Feature Space

Denote by $\mathcal{H}$ a subset of $\{1, \ldots, q\}$ given by the features in $\widehat{\mathcal{M}}$ and interactions in $\widehat{\mathcal{I}}$. To estimate the true value $\theta_0 = (\theta_{0,1}, \ldots, \theta_{0,q})^T$ of the parameter vector θ, we can consider the reduced feature space spanned by the $q_1 = |\widehat{\mathcal{I}}| + |\widehat{\mathcal{M}}|$ columns of the augmented design matrix Z in $\mathcal{H}$, thanks to the sure screening property in Condition 1. Here $|\mathcal{G}|$ stands for the cardinality of a set $\mathcal{G}$. When the model dimensionality is reduced to a moderate scale q_1, one can apply any favorite variable selection procedure for effective selection of important interactions and main effects and efficient estimation of the corresponding coefficients. There is a large literature on the developments of various variable selection methods. Among all approaches, two classes of regularization methods, the convex ones and the concave ones, have been extensively investigated; see, for example, Tibshirani (1996), Fan and Li (2001), Candes and Tao (2007) and references therein. To combine the strengths of both classes, Fan and Lv (2014) introduced the combined L_1 and concave regularization method and established the oracle risk inequalities and bound on the false sign rate for the main-effects-only linear models with fixed design. More specifically, they considered the regularization problem

$$\min_{\beta \in \mathbb{R}^p} \left\{ (2n)^{-1} \|y - X\beta\|_2^2 + \lambda_0 \|\beta\|_1 + \|p_\lambda(\beta)\|_1 \right\},$$

where X is $n \times p$ fixed design matrix with main effects only, $\beta = (\beta_1, \ldots, \beta_p)^T$ with β_j's being regression coefficients for these main effects, $\lambda_0 \geq 0$ is the regularization parameter for the L_1-penalty, $p_\lambda(\beta) = p_\lambda(|\beta|) = (p_\lambda(|\beta_1|), \ldots, p_\lambda(|\beta_p|))^T$, and $p_\lambda(t)$ is an increasing concave penalty function on $[0, \infty)$ indexed by regularization parameter $\lambda \geq 0$.

Following Fan and Lv (2014), we consider the following combined L_1 and concave regularization problem

$$\min_{\theta \in \mathbb{R}^q, \theta_{\mathcal{H}^c} = 0} \left\{ (2n)^{-1} \|y - Z\theta\|_2^2 + \lambda_0 \|\theta_*\|_1 + \|p_\lambda(\theta_*)\|_1 \right\}, \tag{4}$$

where $\theta_{\mathcal{H}^c}$ denotes a subvector of θ given by components in the complement $\mathcal{H}^c$ of the reduced set $\mathcal{H}$, $\theta_* = (\theta_1^*, \ldots, \theta_q^*)^T = D\theta = n^{-1/2}(\|\tilde{z}_1\|_2\theta_1, \ldots, \|\tilde{z}_q\|_2\theta_q)^T$ is the coefficient vector corresponding to the design matrix with each column rescaled to have L_2-norm $n^{1/2}$, and $\tilde{z}_j$ is the jth column of the augmented random design matrix Z and D is a $q \times q$ diagonal matrix with diagonal entries $D_{jj} = n^{-1/2}\|\tilde{z}_j\|_2$ for $j = 1, \ldots, q$. Intuitively, the L_1 component $\lambda_0\|\theta_*\|_1$ reflects the minimum amount of regularization for suppressing the noise in prediction, while the concave component $\|p_\lambda(\theta_*)\|_1$ serves to adapt the model sparsity for variable selection.

Solving problem (4) without the constraint $\theta_{\mathcal{H}^c} = 0$ is equivalent to feature selection using regularization method in $q = p(p+1)/2$ dimensions without the first stage. One advantage of having the first stage is that the computational cost of solving problem (4) in q_1 dimensions is generally substantially reduced compared to that of solving the same problem in q dimensions. However, substantial theoretical challenges arise in investigating the asymptotic properties of the resulting regularized estimator in (4). As discussed in the Introduction section, due to the first stage, the set of selected predictors after the first stage is random and we have to control all possible random sets when establishing the global theoretical results of the resulting regularized estimator in (4) with high-dimensional settings. In addition, Fan and Lv (2014) considered linear models with deterministic design matrix and no interactions, whereas we now need to study the interaction model with random design matrix. The presence of both interactions and

additional randomness requires more delicate technical analyses.

2.4 Asymptotic Properties of Interaction and Main Effect Selection

Before presenting the theoretical results, we state some mild regularity conditions that are needed in our analysis. For a vector a, the usual vector ℓ_d norm is denoted by $\|a\|_d$ ($d = 0$ or $d \in [1, 2]$). Without loss of generality, assume that the first $s = \|\theta_0\|_0$ components of the true regression coefficient vector θ_0 in (3) are nonzero. Throughout the paper, the regularization parameter for the L_1 component is fixed to be $\lambda_0 = c_0\{(\log q_1)/n^{\alpha_1\alpha_2/(\alpha_1+2\alpha_2)}\}^{1/2}$ with c_0 some positive constant, where α_1 and α_2 are given in Condition 2. Some insights into this choice of λ_0 will be provided at the end of this subsection. Denote by $p_{\mathrm{H},\lambda}(t) = 2^{-1}\{\lambda^2 - (\lambda - t)_+^2\}$, $t \geq 0$, the hard-thresholding penalty, where $(\cdot)_+$ denotes the positive part of a number.

Condition 2. *There exist constants $\alpha_1, \alpha_2, c_1 > 0$ such that for any $t > 0$, $P(|X_j| > t) \leq c_1 \exp(-c_1^{-1}t^{\alpha_1})$ for each $1 \leq j \leq p$ and $P(|\varepsilon| > t) \leq c_1 \exp(-c_1^{-1}t^{\alpha_2})$, and $\mathrm{var}(X_j^2)$ are uniformly bounded away from zero.*

Condition 3. *There exist some constants $\kappa_0, \kappa, L_1, L_2 > 0$ such that with probability $1 - a_n$ satisfying $a_n = o(1)$, it holds that $\min_{\|\delta\|_2=1,\, \|\delta\|_0<2s} n^{-1/2}\|Z\delta\|_2 \geq \kappa_0$,*

$$\min_{\delta\neq 0,\, \|\delta_2\|_1\leq 7\|\delta_1\|_1} \left\{n^{-1/2}\|Z\delta\|_2/(\|\delta_1\|_2 \vee \|\delta_3\|_2)\right\} \geq \kappa$$

for $\delta = (\delta_1^T, \delta_2^T)^T \in \mathbb{R}^q$ with $\delta_1 \in \mathbb{R}^s$ and δ_3 a subvector of δ_2 consisting of the s largest components in magnitude, and D_{jj}'s are bounded between $L_1 \leq L_2$, where $a \vee b = \max\{a, b\}$.

Condition 4. *The concave penalty satisfies that $p_\lambda(t) \geq p_{\mathrm{H},\lambda}(t)$ on $[0, \lambda]$, $p'_\lambda\{(1-c_3)\lambda\} \leq \min\{\lambda_0/4, c_3\lambda\}$ for some constant $c_3 \in [0, 1)$, and $-p''_\lambda(t)$ is decreasing on $[0, (1 - c_3)\lambda]$. Moreover, $\min_{1\leq j\leq s} |\theta_{0,j}| > L_1^{-1} \max\{(1 - c_3)\lambda, 2L_2\kappa_0^{-1}p_\lambda^{1/2}(\infty)\}$ with $p_\lambda(\infty) = \lim_{t\to\infty} p_\lambda(t)$.*

The first part of Condition 2 is a usual assumption to control the tail behavior of the covariates and error, which is important for ensuring the sure screening property of our

procedure. Similar assumptions have been made in such work as Fan and Song (2010) and Barut et al. (2016). The scenario of $\alpha_1 = \alpha_2 = 2$ corresponds to the case of sub-Gaussian covariates and error, including distributions with bounded support and light tails.

Condition 3 is similar to Condition 1 in Fan and Lv (2014) for the case of deterministic design matrix, except that the design matrix is now random in our setting and also augmented with interactions. We provide in Section 2.5 some sufficient conditions ensuring that Condition 3 holds. Parallel to our Condition 3, Zhao and Leng (2016) required the restricted eigenvalue condition:

$$\min_{|J|\le s,\, J\subseteq\{1,\ldots,q\}} \min_{\|\delta_{J^c}\|_1 \le k_0'\|\delta_J\|_1} \left\{ n^{-1/2}\|Z\delta\|_2 / \|\delta_J\|_2 \right\} = M(k_0', s) > 0,$$

where k_0' is some positive constant. They also required that the eigenvalues of $\mathrm{cov}(Z)$ are bounded. It is seen that our condition on Z is generally weaker than theirs.

Similar to Condition 2 in Fan and Lv (2014), our Condition 4 ensures that the concave penalty $p_\lambda(t)$ satisfies the hard-thresholding property, requires that its tail grows relatively slowly, and puts a constraint on the minimum signal strength. The hard-thresholding property means that the resulting estimator has the same feature as the hard-thresholding estimator: each component is either zero or of magnitude larger than some value. See §3.1 of Fan and Lv (2014) for more discussions on hard-thresholding property.

The following theorem presents the selection properties of the resulting regularized estimator $\widehat{\theta} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_q)^T$. In particular, we present an explicit bound on the number of falsely discovered signs $\mathrm{FS}(\widehat{\theta}) = |\{1 \le m \le q : \mathrm{sgn}(\widehat{\theta}_m) \ne \mathrm{sgn}(\theta_{0,m})\}|$, which is a stronger measure on variable selection outcome than the total number of false positives and false negatives. To simplify the presentation, hereafter, q_1 is implicitly understood as $\max(n, q_1)$.

Theorem 1. *Assume that* $\log p = o\{n^{\alpha_1\alpha_2/(\alpha_1+2\alpha_2)}\}$ *with* $\alpha_1\alpha_2/(\alpha_1+2\alpha_2) \le 1$, *Conditions 1–4 hold, and* $p_\lambda(t)$ *is continuously differentiable. Then the global minimizer* $\widehat{\theta}$ *of (4) has the hard-thresholding property that each component is either zero or of magnitude larger than* $(1 - c_3)\lambda$, *and with probability at least* $1 - a_n - o(n^{-C} + q_1^{-c_4})$ *for some positive*

constant c_4, it satisfies simultaneously that

$$
\begin{aligned}
& n^{-1/2}\left\|Z(\widehat{\theta}-\theta_0)\right\|_2 = O(\kappa^{-1}\lambda_0 s^{1/2}), \\
& \left\|\widehat{\theta}-\theta_0\right\|_d = O(\kappa^{-2}\lambda_0 s^{1/d}), \quad d\in[1,2], \\
& \mathrm{FS}(\widehat{\theta}) = O\left\{\kappa^{-4}(\lambda_0/\lambda)^2 s\right\},
\end{aligned}
$$

and furthermore $\mathrm{sgn}(\widehat{\theta}) = \mathrm{sgn}(\theta_0)$ *if* $\lambda \geq 56(1-c_3)^{-1}\kappa^{-2}\lambda_0 s^{1/2}$. *Moreover, the same results hold with probability at least* $1-a_n-o(\max\{n,q\}^{-c_4})$ *for the regularized estimator* $\widehat{\theta}$ *without the first stage, that is, without the constraint* $\theta_{\mathcal{H}^c}=0$ *in (4). In that case, Condition 1 is not needed.*

Theorem 1 shows that if the tuning parameter λ satisfies $\lambda_0/\lambda \to 0$, then the number of falsely discovered signs $\mathrm{FS}(\widehat{\theta})$ is of order $o(s)$ and thus the false sign rate $\mathrm{FS}(\widehat{\theta})/s$ is asymptotically vanishing with probability tending to one by noticing that q_1 is implicitly understood as $\max(n, q_1)$. Recall that $\lambda_0 = c_0\{(\log q_1)/n^{\alpha_1\alpha_2/(\alpha_1+2\alpha_2)}\}^{1/2}$. All bounds for prediction and estimation losses and false sign rate in Theorem 1 become larger as q_1 increases. Thus, effective screening and dimension reduction in the first stage can improve accuracy in prediction and estimation and reduce false sign rate. We also observe that the bounds for prediction and estimation losses are independent of the tuning parameter λ for the concave penalty.

As shown in Theorem 1, λ_0 plays a crucial role in characterizing the rates of convergence for the regularized estimator $\widehat{\theta}$. Such a parameter basically measures the maximum noise level in interaction models. In particular, the exponent $\alpha_1\alpha_2/(\alpha_1+2\alpha_2)$ is a key parameter that reflects the level of difficulty in the problem of interaction selection. This quantity is determined by three sources of heavy-tailedness: covariates themselves, their interactions, and the error. In this paper we have focused on the more challenging case of $\alpha_1\alpha_2/(\alpha_1+2\alpha_2) \leq 1$. Such a scenario includes two specific cases: 1) sub-Gaussian covariates and sub-Gaussian error, that is, $\alpha_1=\alpha_2=2$ and 2) sub-Gaussian covariates and sub-exponential error, that is, $\alpha_1=2, \alpha_2=1$. We remark that in the lighter-tailed

case of $\alpha_1\alpha_2/(\alpha_1+2\alpha_2) > 1$, one can simply set $\lambda_0 = c_0\{(\log q_1)/n\}^{1/2}$ and inequalities in Theorem 1 remain to hold by resorting to Lemma 6 in Section E.5 of the supplementary material using similar arguments in the proof of Theorem 1 in Appendix A.

The results in Theorem 1 are also applicable to the regularized estimator without the first stage. In that case, Condition 1 is not needed. Such a one-step procedure was investigated in Zhao and Leng (2016) under the same interaction model (1) with a different class of penalty functions. They showed that the ℓ_1 estimation error bound for the resulting penalized estimator is in the order of $O\left(s\{(\log q)/n\}^{1/2}\right)$ under the assumption that both covariates and model error are sub-Gaussian (that is, $\alpha_1 = \alpha_2 = 2$ in our paper). Under the same distribution assumption, our ℓ_1 estimation error is $O\left(s\{(\log q_1)/n^{2/3}\}^{1/2}\right)$. It is seen that our result is tighter than that of Zhao and Leng (2016) when $\log\{p(p+1)/2\}$ is higher order of $n^{1/3}\log\{\max(n, q_1)\}$, has the same order as theirs when $\log\{p(p+1)/2\}$ is of order $n^{1/3}\log\{\max(n, q_1)\}$, and is weaker than theirs when $\log\{p(p+1)/2\}$ is smaller order of $n^{1/3}\log\{\max(n, q_1)\}$. In the last scenario, our weaker result can be seen as the price we need to pay in exchange for weaker conditions but more general results. In addition, Zhao and Leng (2016) required that $\log(q) = o(n^{1/3})$ with $q = p(p+1)/2$, while we can allow for $\log(p) = o(n^{2/3})$ when $\alpha_1 = \alpha_2 = 2$ even without the first stage. Thus we allow for higher dimensionality. Moreover, the bound for estimation error in Zhao and Leng (2016) holds only when the tuning parameter in their penalty is no less than $c\{(\log q)/n\}^{1/2}$ with some positive constant c, while our bounds for prediction and estimation errors are independent of the tuning parameter for the concave penalty when our tuning parameter λ_0 for the L_1 component is fixed.

As discussed in the Introduction, there are many other one-step procedures in the literature. However, most of those methods impose various constraints on coefficients to enforce the heredity assumption and focus on the low- or moderate-dimensional settings; see, for example, Yuan et al. (2009), Choi et al. (2010), Bien et al. (2013), Yan and Bien (2017), and references therein. These methods are not directly suitable for high-dimensional quadratic regression models for at least two reasons. First, the compu-

tational cost of these methods increases substantially when the number of covariates p grows, mainly because of the $O(p^2)$ added interactions. The computational cost can be extremely high for ultrahigh-dimensional data where p grows at an exponential rate of the sample size n. Second, because of the large number of interactions, selecting important main effects and interactions is more challenging due to spurious correlation and noise accumulation. It is worth mentioning that Tang et al. (2020) proposed a new ADMM based method, which can deal with high-dimensional data and requires no heredity assumptions. However, their theoretical results focus on the estimation and selection of interaction effects, and there are no theoretical results on prediction error, false sign rate, and main effect estimation and selection.

Our work is also related to Fan and Lv (2014) in the sense that the same type of penalty function is used. However, we consider linear interaction model with random design while Fan and Lv (2014) considered linear model with fixed design. Thus the results in Fan and Lv (2014) are not directly applicable to stagewise procedure such as the ones considered here. Also, the presence of random design and interactions poses additional significant theoretical difficulties. Moreover, the bounds in Fan and Lv (2014) depend on $\{(\log p)/n\}^{1/2}$, while our bounds depend on $\{(\log q_1)/n^{\alpha_1\alpha_2/(\alpha_1+2\alpha_2)}\}^{1/2}$, which involves the rate parameters controlling tail probability of each covariate and the model error. Thus, our results can characterize how each bound depends on the distribution of covariates and model error.

2.5 Verification of Condition 3

Since Condition 3 is a key assumption for proving Theorem 1, we provide some sufficient conditions that ensure this assumption on the augmented random design matrix $Z = (z_1, \ldots, z_q)$. Denote by Γ the population covariance matrix of the augmented covariate vector consisting of p main effects X_j's and $p(p-1)/2$ interactions $X_k X_\ell$'s.

Condition 5. *There exists some constant $K > 0$ such that for $\delta = (\delta_1^T, \delta_2^T)^T \in \mathbb{R}^q$,*

$$\min_{\|\delta\|_2=1,\, \|\delta\|_0<2s} \delta^T \Gamma \delta \geq K \quad \textit{and} \quad \min_{\delta\neq 0,\, \|\delta_2\|_1 \leq 7\|\delta_1\|_1} \delta^T \Gamma \delta / \left(\|\delta_1\|_2^2 \vee \|\delta_3\|_2^2\right) \geq K,$$

where $\delta_1 \in \mathbb{R}^s$ and δ_3 is a subvector of δ_2 consisting of the s largest components in magnitude.

Condition 5 is satisfied if the smallest eigenvalue of Γ is assumed to be bounded away from zero. Such a condition is in fact much weaker than the minimum eigenvalue assumption, since it is the population version of a mild sparse eigenvalue assumption and the restricted eigenvalue assumption. The following theorem shows that under some mild assumptions, Condition 3 holds for the full augmented design matrix Z and thus holds naturally for any $n \times q_2$ sub-design matrix with $q_2 \leq q$ given by the first stage.

Theorem 2. *Assume that Condition 5 holds, there exist some constants $\alpha_1, c_1 > 0$ such that for any $t > 0$, $P(|X_j| > t) \leq c_1 \exp(-c_1^{-1} t^{\alpha_1})$ for each $j \in \{1, \ldots, p\}$, $s = O(n^{\xi_0})$, and $\log p = o(n^{\min\{\alpha_1/4,\, 1\}-2\xi_0})$ with constant $0 \leq \xi_0 < \min\{\alpha_1/8, 1/2\}$. Then Condition 3 holds with probability $1 - a_n$ satisfying $n^{\min\{\alpha_1/4,\, 1\}-2\xi_0} = O(-\log a_n)$.*

3 Simulation Studies

We design two simulation studies to verify the theoretical results in this paper. In study 1, we consider the following five interaction models:

$$\begin{aligned}
&\text{Model 1}: Y = 2X_1 + 2X_5 + 3X_1X_5 + \varepsilon_1,\\
&\text{Model 2}: Y = 2X_1 + 2X_{10} + 3X_1X_5 + \varepsilon_2,\\
&\text{Model 3}: Y = 2X_{10} + 2X_{15} + 3X_1X_5 + \varepsilon_3,\\
&\text{Model 4}: Y = 3X_1X_5 + 3X_{10}X_{15} + \varepsilon_4,\\
&\text{Model 5}: Y = 2X_1 + 2X_{10} + 2X_{20} + 2X_{30} + 2X_{40} +\\
&\qquad\qquad 3X_1X_5 + 3X_1X_{10} + 3.5X_5X_{15} + 3.5X_{10}X_{15} + \varepsilon_5,
\end{aligned}$$

where each covariate $X_j = W_j + U_j$ for $1 \leq j \leq p$, the covariate vector $(W_1, \ldots, W_p)^T \sim N(0, \Sigma)$ with $\Sigma = (\rho^{|j-k|})_{1 \leq j,k \leq p}$, and U_j's are independent and identically distributed and follow the uniform distribution on $[-0.5, 0.5]$. The errors $\varepsilon_1 \sim t_{(3)}$, $\varepsilon_2 \sim t_{(4)}$, $\varepsilon_3 \sim t_{(4)}$, $\varepsilon_4 \sim t_{(8)}$, and $\varepsilon_5 \sim t_{(12)}$ are independent of $x = (X_1, \ldots, X_p)^T$. The first two models (models 1 and 2) satisfy the heredity assumption (either strong or weak), the next two models (models 3 and 4) do not obey such an assumption, and the last model has larger number of important variables. A sample of size n was randomly generated from each of the five models. We fixed the sample size and dimensionality at $(n, p) = (300, 5000)$ and considered two different correlation levels $\rho = 0$ and 0.5 for these five models. We repeated each experiment 100 times.

We compare performance of different stagewise selection approaches, each of which is a two-step method, where in the first step a variable screening is employed to reduce the number of interactions and main effects, and in the second step a variable selection method is used to further select important interactions and main effects. For the screening step, we employed several recent feature screening procedures: the sure independence screening (Fan and Lv, 2008), feature screening via distance correlation (Li et al., 2012), variable selection via sliced inverse regression (Jiang and Liu, 2014), and interaction pursuit via distance correlation (Kong et al., 2017), respectively. Since this paper focuses on interaction models with one single response, we can also consider the method of interaction pursuit via the Pearson correlation for screening, which is exactly the same as interaction pursuit via distance correlation, except for the replacement of distance correlation with the Pearson correlation when identifying the variables in $\mathcal{A}$ and $\mathcal{B}$.

Both methods in Fan and Lv (2008) and Li et al. (2012) are not particularly designed for interaction models and each returns a set of variables without distinguishing between important main effects and active interaction variables. Thus for each of those methods, we construct interactions using all possible pairwise interactions of the recruited variables, and refer to the resulting procedures as SIS2 and DCSIS2, respectively, to distinguish them from the original ones. By doing so, the strong heredity assumption is enforced.

For the method of interaction pursuit with the Pearson correlation, we retain the top $[n/(\log n)]$ variables in each of sets $\widehat{\mathcal{A}}$ and $\widehat{\mathcal{B}}$ defined in (B.2), respectively. The features in the union set $\widehat{\mathcal{M}} = \widehat{\mathcal{A}} \cup \widehat{\mathcal{B}}$ are used as main effects, while variables in set $\widehat{\mathcal{A}}$ are used to build interactions in the selection stage. We use the same procedure for the method of interaction pursuit via distance correlation in Kong et al. (2017). To ensure a fair comparison, the numbers of variables kept in other methods are all equal to the cardinality of $\widehat{\mathcal{M}}$, which is up to $2[n/(\log n)]$. Since the screening step is not the main focus of our current paper, we present all screening results in the supplementary material.

The method of variable selection via sliced inverse regression or SIRI for short (Jiang and Liu, 2014) is an iterative procedure that alternates between a large-scale variable screening step and a moderate-scale variable selection step when p is large. Here we use the full iterative procedure as described in Jiang and Liu (2014) and thus no additional variable selection step is needed for this method. For other screening methods, we can employ regularization methods such as the Lasso (Tibshirani, 1996) and the combined L_1 and concave method (Fan and Lv, 2014) to select important interactions and main effects after the screening stage. As shown in Fan and Lv (2014), different choices of the concave penalty gave rise to similar performance. We thus implemented the combined L_1 and smooth integration of counting and absolute deviation method (Lv and Fan, 2009), for simplicity. We also tried Lasso penalty but it generally yields inferior results. Thus we only report the approach of interaction pursuit with the Pearson correlation followed by Lasso penalty as a representative. For the method of SIRI (Jiang and Liu, 2014), we added an additional refitting step using the selected variables to calculate model performance measures. The oracle procedure based on the true underlying interaction model, or Oracle for short, was used as a benchmark for comparisons.

To ease the presentation, denote by IP the screening method of interaction pursuit with the Pearson correlation, IPDC the the screening method of interaction pursuit with the distance correlation, and L_1+SICA the combined L_1 and smooth integration of counting and absolute deviation method for selection. The approach of SIS2 fol-

lowed by the L_1+SICA method is referred to as SIS2-L_1+SICA for short. All other combinations of screening and selection methods are defined similarly. We include the following seven methods to assess the variable selection performance: SIS2-L_1+SICA, DC-SIS2-L_1+SICA, SIRI, IPDC-L_1+SICA, IP-Lasso, IP-L_1+SICA, and Oracle. The cross-validation was used to select tuning parameters for all the methods, except that BIC was applied to the procedures with the combined L_1 and smooth integration of counting and absolute deviation method for computational efficiency since two regularization parameters are involved.

Table 1: Variable selection results for study 1 with $(n, p, \rho) = (300, 5000, 0.5)$. Reported values are medians and robust standard deviations (in parentheses) of three performance measures: PE, prediction error; FS, falsely discovered signs; and Time, running time in seconds. 0^* means that the corresponding value is small than 0.001.

	SIS2-L_1+SICA	DC-SIS2-L_1+SICA	SIRI	IPDC-L_1+SICA	IP-Lasso	IP-L_1+SICA	Oracle
Model 1							
PE	3.1 (0.8)	3.2 (1.0)	3.0 (0.0)	3.1 (1.8)	3.6 (0.3)	3.1 (0.8)	2.9 (0.2)
FS	0 (4.5)	1 (7.8)	0 (0.0)	0 (14.1)	109 (20.5)	0 (4.5)	0 (0)
Time	728.0 (60.2)	716.5 (47.6)	806.1 (54.5)	114.3 (15.4)	7.4 (1.0)	124.2 (12.8)	0* (0*)
Model 2							
PE	19.1 (3.2)	2.1 (0.3)	2.2 (0.0)	2.1 (0.4)	2.5 (0.2)	2.1 (0.3)	2.0 (0.1)
FS	27 (9.0)	0 (3.4)	3 (3.0)	0 (5.4)	100 (20.5)	0 (3.0)	0 (0)
Time	741.8 (40.6)	732.7 (41.3)	801.9 (51.8)	109.3 (10.2)	7.0 (0.7)	117.8 (8.5)	0* (0*)
Model 3							
PE	20.8 (3.0)	20.0 (3.7)	13.2 (0.6)	2.1 (0.3)	2.4 (0.2)	2.1 (0.2)	2.0 (0.2)
FS	29.5 (10.4)	27 (13.4)	7 (5.2)	0 (2.7)	97 (18.3)	0 (2.2)	0 (0)
Time	766.4 (25.0)	758.8 (29.2)	788.3 (57.1)	111.2 (9.1)	6.8 (0.6)	118.4 (7.1)	0* (0*)
Model 4							
PE	36.4 (3.3)	21.6 (13.2)	12.4 (7.7)	1.4 (0.1)	1.6 (9.1)	1.4 (9.3)	1.3 (0.0)
FS	33.5 (14.6)	28 (13.1)	6 (5.2)	0 (0.0)	71 (39.9)	0 (4.9)	0 (0)
Time	764.9 (28.3)	769.0 (27.2)	231.4 (131.3)	97.2 (6.4)	6.4 (1.0)	106.6 (11.8)	0* (0*)
Model 5							
PE	67.0 (13.5)	54.7 (24.1)	6.4 (34.3)	2.0 (3.5)	37.4 (28.9)	35.6 (30.7)	2.0 (0.0)
FS	21 (9.0)	16 (13.8)	22 (8.6)	6 (1.9)	119.5 (27.6)	11 (13.8)	0 (0.0)
Time	776.4 (48.1)	773.1 (41.3)	472.9 (81.1)	111.1 (19.3)	9.5 (1.9)	145.0 (14.8)	0* (0*)

To evaluate the variable selection performance of each method, we employed three measures: the prediction error, falsely discovered signs, and running time in seconds. The prediction error was calculated using an independent test sample of size 10,000. The medians and robust standard deviations of these measures were calculated based on 100 simulations for different models in the study 1. The robust standard deviation is defined as the interquartile range divided by 1.34. The experiments were conducted on two identical servers with 3.07 GHz Intel Core with 24 processors and 64 GB memory.

Parallel implementation was applied on the 100 simulations so the individual running time may seem long but the entire experiment can finish in much shorter time than sequential implementation. Since the selection results for study 1 under two settings $(n, p, \rho) = (300, 5000, 0)$ and $(n, p, \rho) = (300, 5000, 0.5)$ are similar, we only present the selection results for study 1 with $(n, p, \rho) = (300, 5000, 0.5)$ here to save space. For the selection results for study 1 with $(n, p, \rho) = (300, 5000, 0)$, see Table 6 in the supplementary material.

In view of Table 1, we see that when the strong heredity assumption holds (model 1), most methods performed well with the SIRI method following closely the oracle procedure. In model 2 with the weak heredity assumption, all methods, except the SIS2-L_1+SICA method, performed fairly well. In the cases when the heredity assumption does not hold (models 3 and 4), those variable selection methods based on the interaction pursuit via distance correlation or the Pearson correlation (the IPDC-L_1+SICA, IP-Lasso, and IP-L_1+SICA methods) still mimicked the oracle procedure and uniformly outperformed the other methods over all settings. The inflated robust standard deviations, relative to medians, in model 4 were due to the relatively low sure screening probabilities (see Tables 3 and 4 in the supplementary material. When the sure screening probability is low, a non-negligible number of replications can have nonzero false negatives, which inflated the corresponding prediction errors. Model 5 has five main effects and four interactions, some of which satisfy the heredity assumption while others do not. The method based on interaction pursuit via distance correlation (the IPDC-L_1+SICA method) performs the best and closest to the oracle procedure. The SIRI method performs better than the methods of interaction pursuit via Pearson correlation (the IP-Lasso and IP-L_1+SICA methods) but not the method of interaction pursuit via distance correlation (the IPDC-L_1+SICA method) in model 5.

To assess the performance of each method with increasing sample size n and dimensionality p, we consider study 2. More specifically, we consider model 4 with the same settings except for $(n, p) = (300, 10000)$ and $(400, 10000)$, respectively, in study 2. See

Tables 5 and 7 in the supplemental material for corresponding screening and selection results. We can observe similar conclusions as those for model 4 in study 1.

4 Real Data Application

We further evaluate the performance of our method on a supermarket data set, which was also analyzed in Wang (2009), Hao and Zhang (2014), and Hao et al. (2018). The data set contains a total of $n = 464$ daily sale records from a major supermarket in northern China. Each record includes the number of customers on a particular day, denoted as Y, and the sale amounts of $p = 6,398$ products on the same day, denoted as $X_1, \cdots, X_p$. The response Y and covariates $X_1, \cdots, X_p$ have already been standardized to have zero mean and unit variance. The goal is to identify the products that significantly contribute to the prediction of the daily number of customers, which can be useful for the supermarket manager to make promotion strategies.

Following Hao et al. (2018), we randomly split the data into a training set of size 400 and a test set of size 64 to evaluate the prediction performance of different methods. Using the training set, we retained the top $[n/(\log n)]$ variables in each of sets $\widehat{\mathcal{A}}$ and $\widehat{\mathcal{B}}$ in the screening stage, and used the features in the union set $\widehat{\mathcal{M}} = \widehat{\mathcal{A}} \cup \widehat{\mathcal{B}}$ as main effects and variables in set $\widehat{\mathcal{A}}$ to build interactions in the selection stage. We then calculated the number of selected main effects (size.main), the number of selected interactions (size.inter), and the out-of-sample R^2 on the test data. We repeated the random splits 100 times, with the average performance presented in Table 2. The results of the RAMP with different tuning methods were extracted from Hao et al. (2018), where RAMP-GIC is RAMP with the tuning parameter selected by GIC (Fan and Tang, 2013). Other procedures for the RAMP are defined similarly.

Table 2 shows that both IP-Lasso and IP-L_1+SICA yield slightly higher out-of-sample R^2 than the RAMP based methods. In addition, the out-of-sample R^2 values for iFORT and iFORM (reported in Table 8 of Hao and Zhang (2014)) are 88.91(0.17) and 88.66(0.18), respectively, with standard errors included in parentheses. This indi-

Table 2: Mean selection and prediction results on the supermarket data set over 100 random splits. The standard errors are in parentheses.

	size.main	size.inter	R^2
RAMP-AIC	229.12(1.68)	94.53(1.06)	0.9048(0.0023)
RAMP-BIC	101.17(3.25)	34.36(1.65)	0.9118(0.0020)
RAMP-EBIC	29.27(1.01)	3.07(0.29)	0.8967(0.0031)
RAMP-GIC	30.71(0.92)	3.20(0.30)	0.9008(0.0028)
IP-Lasso	53.00(0.30)	147.30(2.02)	0.9191(0.0020)
IP-L_1+SICA	52.20(0.28)	121.85(0.93)	0.9206(0.0020)

cates that both IP-Lasso and IP-L_1+SICA also outperform iFORT and iFORM in terms of the out-of-sample R^2.

5 Discussion

Our theoretical analysis has shown that the regularized estimator given by the interaction pursuit approach enjoys the same asymptotic properties as the lasso estimator, but with improved sparsity and false sign rate, in ultrahigh-dimensional quadratic regression models under random design. To simplify the technical presentation, our analysis has focused on the linear pairwise interaction models. It would be interesting to extend these selection results to other general model frameworks such as the generalized linear models, nonparametric models, and survival models with interactions.

A Proofs of Theorems 1 and 2

We provide the detailed proofs of Theorems 1 and 2 in Appendix A and give some sufficient conditions in Appendix B to ensure that Condition 1 holds. Additional simulation studies and technical details are provided in the supplementary material. In particular, q_1 in the proofs of Theorems 1 and 2 is implicitly understood as $\max(n, q_1)$ to simplify the notation.

A.1 Proof of Theorem 1

Recall that $Z = (\tilde{z}_1, \ldots, \tilde{z}_q)$ is the corresponding $n \times q$ augmented design matrix incorporating the covariate vectors for X_j's and their interactions in columns, where $\tilde{z}_j = \tilde{x}_j = (X_{1j}, \ldots, X_{nj})^T$ for $1 \le j \le p$ is the jth covariate vector and $\tilde{z}_j$ for $p+1 \le j \le q = p(p+1)/2$ is $\tilde{x}_k \circ \tilde{x}_\ell$ with some $1 \le k < \ell \le p$ and $\circ$ denoting the Hadamard (componentwise) product. We rescale the augmented design matrix Z such that each column has L_2-norm $n^{1/2}$, and denote by $\widetilde{Z} = ZD^{-1}$ the resulting matrix, where $D = \mathrm{diag}\{D_{11}, \ldots, D_{qq}\}$ with $D_{jj} = n^{-1/2}\|\tilde{z}_j\|_2$ is a diagonal scale matrix.

Define the event $\mathcal{E}_4 = \{L_1 \le \min_{1\le j\le q} |D_{jj}| \le \max_{1\le j\le q} |D_{jj}| \le L_2\}$, where L_1 and L_2 are two positive constants defined in Condition 3. Then by the assumption in Condition 3, event $\mathcal{E}_4$ holds with probability at least $1 - a_n$. In what follows, we will condition on the event $\mathcal{E}_4$.

Note that conditional on $\mathcal{E}_4$, we have

$$\|Z\delta\|_2 \sim \|\widetilde{Z}\delta\|_2, \tag{A.1}$$

where the notation $f_n \sim g_n$ means that the ratio f_n/g_n is bounded between two positive constants. Thus, conditional on $\mathcal{E}_4$, Condition 3 holds with matrix Z replaced with $\widetilde{Z}$. More specifically, with probability at least $1 - a_n$, it holds that

$$\min_{\|\delta\|_2=1,\, \|\delta\|_0<2s} n^{-1/2}\|\widetilde{Z}\delta\|_2 \ge \widetilde{\kappa}_0, \quad \min_{\delta\neq 0,\, \|\delta_2\|_1\le 7\|\delta_1\|_1} \left\{ n^{-1/2}\|\widetilde{Z}\delta\|_2 / (\|\delta_1\|_2 \vee \|\delta_3\|_2) \right\} \ge \widetilde{\kappa}, \tag{A.2}$$

where $\widetilde{\kappa}_0$ and $\widetilde{\kappa}$ are two positive constants depending only on κ, κ_0, L_1, and L_2. In addition, conditional on $\mathcal{E}_4$, the desired results in Theorem 1 are equivalent to those with Z and θ replaced by $\widetilde{Z}$ and $\theta^* = D\theta$, respectively. Thus, we only need to work with the design matrix $\widetilde{Z}$ and reparameterized parameter vector θ^*.

We will use sets to index vectors and matrices. For example, $\delta_{\mathcal{H}}$ denotes the subvector of $\delta \in \mathbb{R}^q$ formed by its components with indices in $\mathcal{H}$, and $\widetilde{Z}_{\mathcal{H}}$ represents the submatrix of $\widetilde{Z} \in \mathbb{R}^{n\times q}$ formed by its columns in $\mathcal{H}$. By the definitions of $\widetilde{Z}_{\mathcal{H}}$ and $\delta_{\mathcal{H}}$, we see that

the two inequalities in (A.2) with $\widetilde{Z}$ and δ replaced by $\widetilde{Z}_{\mathcal{H}}$ and $\delta_{\mathcal{H}}$, respectively, still hold with probability at least $1 - a_n$. By examining the proof of Theorem 1 in Fan and Lv (2014), in order to prove Theorem 1 in our paper, it suffices to show that the following inequality

$$\|n^{-1}\widetilde{Z}_{\mathcal{H}}^T\boldsymbol{\varepsilon}_{\mathcal{H}}\|_\infty > \lambda_0/2 \tag{A.3}$$

holds with probability at most $a_n + o(p^{-c_4})$, where $\lambda_0 = c_0\{(\log q_1)/n^{\alpha_1\alpha_2/(\alpha_1+2\alpha_2)}\}^{1/2}$ for some constant $c_0 > 0$ and c_4 is some arbitrarily large positive constant depending on c_0. Then with (A.3), following the proof of Theorem 1 in Fan and Lv (2014), we can obtain that all results in Theorem 1 hold with probability at least $1 - a_n - o(p^{-c_4})$.

It remains to prove (A.3). We first show that $\|n^{-1}\widetilde{Z}_{\mathcal{H}}^T\boldsymbol{\varepsilon}_{\mathcal{H}}\|_\infty > \lambda_0/2$ holds with an overwhelming probability. To this end, note that an application of the Bonferroni inequality gives

$$P(\|n^{-1}\widetilde{Z}_{\mathcal{H}}^T\boldsymbol{\varepsilon}_{\mathcal{H}}\|_\infty > \lambda_0/2|\mathcal{E}_4) \le \sum_{j=1}^{q_1} P(|n^{-1}\tilde{z}_j^T\boldsymbol{\varepsilon}| > L_1\lambda_0/2|\mathcal{E}_4) \tag{A.4}$$

for any $\lambda_0 > 0$. The key idea is to construct an upper bound for $P(|n^{-1}\tilde{z}_j^T\boldsymbol{\varepsilon}| > L_1\lambda_0/2|\mathcal{E}_4)$. We claim that such an upper bound is $\widetilde{C}_1\exp\{-\widetilde{C}_2 n^{\alpha_1\alpha_2/(\alpha_1+2\alpha_2)}\lambda_0^2\}$ for any $0 < L_1\lambda_0 < 2$, where $\widetilde{C}_1$ and $\widetilde{C}_2$ are some positive constants. To prove this, we consider the following two cases.

Case 1: $j \in \{1, 2\ldots, p\}$. In this case, $\tilde{z}_j = (X_{1j}, \ldots, X_{nj})^T$. Thus $n^{-1}\tilde{z}_j^T\boldsymbol{\varepsilon} = n^{-1}\sum_{i=1}^n X_{ij}\varepsilon_i$. By Lemma 2, for any $t > 0$, we have

$$P(|X_{ij}\varepsilon_i| > t) \le 2c_1\exp\{-c_1^{-1}t^{\alpha_1\alpha_2/(\alpha_1+\alpha_2)}\}$$

for all $1 \le i \le n$ and $1 \le j \le p$. Note that $E(X_{ij}\varepsilon_i) = 0$. Thus it follows from Lemma 6 that there exist some positive constants $\widetilde{C}_3$ and $\widetilde{C}_4$ such that

$$P(|n^{-1}\tilde{z}_j^T\boldsymbol{\varepsilon}| > L_1\lambda_0/2|\mathcal{E}_4) \le \widetilde{C}_3\exp\{-\widetilde{C}_4 n^{\min\{\alpha_1\alpha_2/(\alpha_1+\alpha_2),1\}}\lambda_0^2\}$$

for all $0 < L_1\lambda_0 < 2$.

Case 2: $j \in \{p+1, \ldots, q\}$. In this case, $\tilde{z}_j = (X_{1k}X_{1\ell}, \ldots, X_{nk}X_{n\ell})^T$. Thus $n^{-1}\tilde{z}_j^T\boldsymbol{\varepsilon} = n^{-1}\sum_{i=1}^n X_{ik}X_{i\ell}\varepsilon_i$ with some $1 \le k < \ell \le p$. By Lemma 2, for any $t > 0$, we have $P(|X_{ik}X_{i\ell}\varepsilon_i| > t) \le 4c_1 \exp\{-c_1^{-1}t^{\alpha_1\alpha_2/(\alpha_1+2\alpha_2)}\}$ for all $1 \le i \le n$ and $1 \le k < j \le p$. Note that $E(X_{ik}X_{i\ell}\varepsilon_i) = 0$. Thus it follows from Lemma 6 and $\alpha_1\alpha_2/(\alpha_1+2\alpha_2) \le 1$ that there exist some positive constants $\widetilde{C}_5$ and $\widetilde{C}_6$ such that

$$P(|n^{-1}\tilde{z}_j^T\boldsymbol{\varepsilon}| > L_1\lambda_0/2|\mathcal{E}_4) \le \widetilde{C}_5 \exp\{-\widetilde{C}_6 n^{\alpha_1\alpha_2/(\alpha_1+2\alpha_2)}\lambda_0^2\}$$

for all $0 < L_1\lambda_0 < 2$.

Under the assumption that $\alpha_1\alpha_2/(\alpha_1+2\alpha_2) \le 1$, we have $\alpha_1\alpha_2/(\alpha_1+2\alpha_2) \le \min\{\alpha_1\alpha_2/(\alpha_1+\alpha_2), 1\}$. Thus combining Cases 1 and 2 above along with (A.4) leads to

$$\begin{aligned} P(\|n^{-1}\widetilde{Z}_{\mathcal{H}}^T\boldsymbol{\varepsilon}_{\mathcal{H}}\|_\infty > \lambda_0/2|\mathcal{E}_4) &\le \sum_{j=1}^{q_1} P(|n^{-1}\tilde{z}_j^T\boldsymbol{\varepsilon}| > L_1\lambda_0/2|\mathcal{E}_4) \\ &\le \widetilde{C}_1 q_1 \exp\{-\widetilde{C}_2 n^{\alpha_1\alpha_2/(\alpha_1+2\alpha_2)}\lambda_0^2\} \end{aligned}$$

for all $0 < L_1\lambda_0 < 2$, where $\widetilde{C}_1 = \max\{\widetilde{C}_3, \widetilde{C}_5\}$ and $\widetilde{C}_2 = \min\{\widetilde{C}_4, \widetilde{C}_6\}$.

Set $\lambda_0 = c_0\{(\log q_1)/n^{\alpha_1\alpha_2/(\alpha_1+2\alpha_2)}\}^{1/2}$ with $c_0 > (\widetilde{C}_2)^{-1/2}$ some positive constant. Then $0 < L_1\lambda_0 < 2$ for all n sufficiently large. Thus, with the above choice of λ_0, it holds that

$$P(\|n^{-1}\widetilde{Z}_{\mathcal{H}}^T\boldsymbol{\varepsilon}_{\mathcal{H}}\|_\infty > \lambda_0/2|\mathcal{E}_4) \le o(q_1^{-c_4}),$$

where c_4 is some positive constant. Note that $P(A) \le P(A|B) + P(B^c)$ for any events A and B with $P(B) > 0$. Thus,

$$P(\|n^{-1}\widetilde{Z}_{\mathcal{H}}^T\boldsymbol{\varepsilon}_{\mathcal{H}}\|_\infty > \lambda_0/2) \le P(\|n^{-1}\widetilde{Z}_{\mathcal{H}}^T\boldsymbol{\varepsilon}_{\mathcal{H}}\|_\infty > \lambda_0/2|\mathcal{E}_4) + P(\mathcal{E}_4^c) \le o(q_1^{-c_4}) + a_n,$$

which completes the proof of Theorem 1.

A.2 Proof of Theorem 2

We first prove that the diagonal entries D_{jj}'s of the scale matrix D are bounded between two positive constants $L_1 \leq L_2$ with significant probability. Since $P(|X_{ij}| > t) \leq c_1 \exp(-c_1^{-1} t^{\alpha_1})$ for any $t > 0$ and all $1 \leq i \leq n$ and $1 \leq j \leq p$, by Lemma 7 and noting that $E(X_{ij}^2) = 1$, there exist some positive constants $\widetilde{C}_1$ and $\widetilde{C}_2$ such that

$$\begin{aligned} P(1/2 \leq n^{-1/2}\|\tilde{x}_j\|_2 \leq \sqrt{7}/2) &= P\{-3/4 \leq n^{-1}\sum_{i=1}^{n}[X_{ij}^2 - E(X_{ij}^2)] \leq 3/4\} \\ &= 1 - P\{|n^{-1}\sum_{i=1}^{n}[X_{ij}^2 - E(X_{ij}^2)]| > 3/4\} \geq 1 - \widetilde{C}_1 \exp(-\widetilde{C}_2 n^{\min\{\alpha_1/2,1\}}) \end{aligned} \tag{A.5}$$

for all $1 \leq j \leq p$ where $\tilde{x}_j = (X_{1j}, \ldots, X_{nj})^T$.

Since $\mathrm{var}(X_{ik}X_{i\ell})$ is a diagonal entry of the population covariance matrix Γ, it follows from Condition 5 that $\mathrm{var}(X_{ik}X_{i\ell}) \geq K > 0$ for all $1 \leq k < \ell \leq p$. Thus, there exists a constant $0 < K_0 \leq 1$ such that $E(X_{ik}^2 X_{i\ell}^2) \geq \mathrm{var}(X_{ik}X_{i\ell}) \geq K > K_0$ for all $1 \leq k < \ell \leq p$. Meanwhile, it follows from $X_{ik}^2 X_{i\ell}^2 \leq (X_{ik}^4 + X_{i\ell}^4)/2$ and Lemma 3 that $E(X_{ik}^2 X_{i\ell}^2) \leq \widetilde{C}_3$, where $\widetilde{C}_3 \geq K_0$ is some positive constant. Note that $P(|X_{ij}| > t) \leq c_1 \exp(-c_1^{-1} t^{\alpha_1})$ for any $t > 0$ and all $1 \leq i \leq n$ and $1 \leq j \leq p$. Thus it follows from Lemma 7 that there exist some positive constants $\widetilde{C}_4$ and $\widetilde{C}_5$ such that for all $1 \leq k < \ell \leq p$,

$$\begin{aligned} & P\left\{(K_0/4)^{-1/2} \leq n^{-1/2}\|\tilde{x}_k \circ \tilde{x}_\ell\|_2 \leq (7\widetilde{C}_3/4)^{-1/2}\right\} \\ \geq & P\left\{(K_0/4)^{-1/2} \leq n^{-1/2}\|\tilde{x}_k \circ \tilde{x}_\ell\|_2 \leq (3K_0/4 + \widetilde{C}_3)^{-1/2}\right\} \\ \geq & P\Big\{\big|n^{-1}\sum_{i=1}^{n}[X_{ik}^2 X_{i\ell}^2 - E(X_{ik}^2 X_{i\ell}^2)]\big| \leq 3K_0/4\Big\} \qquad \text{(A.6)} \\ = & 1 - P\Big\{\big|n^{-1}\sum_{i=1}^{n}[X_{ik}^2 X_{i\ell}^2 - E(X_{ik}^2 X_{i\ell}^2)]\big| > 3K_0/4\Big\} \\ \geq & 1 - \widetilde{C}_4 \exp(-\widetilde{C}_5 n^{\min\{\alpha_1/4,1\}}). \qquad \text{(A.7)} \end{aligned}$$

Let $L_1 = 2^{-1}\min\{1, K_0^{1/2}\} = 2^{-1}K_0^{1/2}$ and $L_2 = \sqrt{7}/2\max\{1, \widetilde{C}_3^{1/2}\}$. Then combining

(A.5) with (A.6) yields that the following inequality

$$L_1 \leq \min_{1\leq j\leq q} |D_{jj}| \leq \max_{1\leq j\leq q} |D_{jj}| \leq L_2, \tag{A.8}$$

holds with probability at least $1 - \widetilde{C}_1 p \exp(-\widetilde{C}_2 n^{\min\{\alpha_1/2,1\}}) - \widetilde{C}_4 p^2 \exp(-\widetilde{C}_5 n^{\min\{\alpha_1/4,1\}})$, which shows that D_{jj}'s are bounded away from zero and infinity with large probability.

We proceed to show that the first two parts of Theorem 2 hold with significant probability. For any $0 < \epsilon < 1$, define an event $\mathcal{E}_5 = \{\|n^{-1}Z^TZ - \Gamma\|_\infty \leq \epsilon\}$, where $\|\cdot\|_\infty$ stands for the entrywise matrix infinity norm, and Z and Γ are defined in Section 2.5. Recall that $q = p(p+1)/2$. Since $P(|X_{ij}| > t) \leq c_1 \exp(-c_1^{-1}t^{\alpha_1})$ for any $t > 0$ and all $1 \leq i \leq n$ and $1 \leq j \leq p$, it follows from Lemma 7 that there exist some positive constants $\widetilde{C}_6$ and $\widetilde{C}_7$ such that

$$\begin{aligned} P(\mathcal{E}_5) &= 1 - P(|(n^{-1}Z^TZ - \Gamma)_{jk}| > \epsilon \text{ for some } (j,k) \text{ with } 1 \leq j,k \leq q) \\ &\geq 1 - \sum_{j=1}^{q}\sum_{k=1}^{q} P(|(n^{-1}Z^TZ - \Gamma)_{jk}| > \epsilon) \geq 1 - \widetilde{C}_6 q^2 \exp(-\widetilde{C}_7 n^{\min\{\alpha_1/4,1\}}\epsilon^2) \end{aligned} \tag{A.9}$$

for any $0 < \epsilon < 1$, where A_{jk} denotes the (j,k)-entry of a matrix A.

Next, we show that conditional on the event $\mathcal{E}_5$, the desired inequalities in Theorem 2 hold. From now on, we condition on the event $\mathcal{E}_5$. Note that $(n^{-1/2}\|Z\delta\|_2)^2 = \delta^T(n^{-1}Z^TZ - \Gamma)\delta + \delta^T\Gamma\delta$. Let δ_J be the subvector of δ formed by putting all nonzero components of δ together. For any δ satisfying $\|\delta\|_2 = 1$ and $\|\delta\|_0 < 2s$, by the Cauchy-Schwarz inequality we have

$$|\delta^T(n^{-1}Z^TZ - \Gamma)\delta| \leq \epsilon\|\delta\|_1^2 = \epsilon\|\delta_J\|_1^2 \leq \epsilon\|\delta_J\|_0\|\delta_J\|_2^2 = \epsilon\|\delta\|_0\|\delta\|_2^2 < 2s\epsilon. \tag{A.10}$$

It follows that $(n^{-1/2}\|Z\delta\|_2)^2 > \delta^T\Gamma\delta - 2s\epsilon$ for any δ satisfying $\|\delta\|_2 = 1$ and $\|\delta\|_0 < 2s$. Thus we derive

$$\min_{\|\delta\|_2=1,\|\delta\|_0<2s} (n^{-1/2}\|Z\delta\|_2)^2 \geq \min_{\|\delta\|_2=1,\|\delta\|_0<2s} (\delta^T\Gamma\delta) - 2s\epsilon \geq K - 2s\epsilon, \tag{A.11}$$

where the last inequality follows from Condition 5.

Meanwhile, for any $\delta \neq 0$ we have

$$\left(\frac{n^{-1/2}\|Z\delta\|_2}{\|\delta_1\|_2 \vee \|\delta_3\|_2}\right)^2 = \frac{\delta^T(n^{-1}Z^TZ-\Gamma)\delta}{\|\delta_1\|_2^2 \vee \|\delta_3\|_2^2} + \frac{\delta^T\Gamma\delta}{\|\delta_1\|_2^2 \vee \|\delta_3\|_2^2} + \frac{\delta^T\Gamma\delta}{\|\delta\|_2^2}$$
$$\geq \frac{\delta^T(n^{-1}Z^TZ-\Gamma)\delta}{\|\delta_1\|_2^2 \vee \|\delta_3\|_2^2} + \frac{\delta^T\Gamma\delta}{\|\delta\|_2^2}.$$

Under the additional condition $\|\delta_2\|_1 \leq 7\|\delta_1\|_1$, by the first inequality of (A.10) it holds that

$$\left|\frac{\delta^T(n^{-1}Z^TZ-\Gamma)\delta}{\|\delta_1\|_2^2 \vee \|\delta_3\|_2^2}\right| \leq \frac{\epsilon\|\delta\|_1^2}{\|\delta_1\|_2^2} = \frac{\epsilon(\|\delta_1\|_1+\|\delta_2\|_1)^2}{\|\delta_1\|_2^2} \leq \frac{64\epsilon\|\delta_1\|_1^2}{\|\delta_1\|_2^2} \leq 64s\epsilon,$$

where the last inequality follows from the Cauchy-Schwarz inequality. This entails that

$$\left(\frac{n^{-1/2}\|Z\delta\|_2}{\|\delta_1\|_2 \vee \|\delta_3\|_2}\right)^2 \geq \frac{\delta^T\Gamma\delta}{\|\delta\|_2^2} - 64s\epsilon$$

for any $\delta \neq 0$ with $\|\delta_2\|_1 \leq 7\|\delta_1\|_1$. Thus, by Condition 5 we have

$$\min_{\delta\neq 0,\, \|\delta_2\|_1 \leq 7\|\delta_1\|_1} \left(\frac{n^{-1/2}\|Z\delta\|_2}{\|\delta_1\|_2 \vee \|\delta_3\|_2}\right)^2 \geq \min_{\delta\neq 0,\, \|\delta_2\|_1 \leq 7\|\delta_1\|_1} \frac{\delta^T\Gamma\delta}{\|\delta\|_2^2} - 64s\epsilon \geq K - 64s\epsilon. \tag{A.12}$$

Recall that $s = O(n^{\xi_0})$ with $0 \leq \xi_0 < \min\{\alpha_1/8, 1/2\}$ by assumption and thus $s \leq \widetilde{C}_8 n^{\xi_0}$ for some positive constant $\widetilde{C}_8$. Take $\epsilon = Kn^{-\xi_0}/\widetilde{C}_9$ with $\widetilde{C}_9$ some sufficiently large positive constant such that $\epsilon \in (0,1)$ and $K - 64s\epsilon > 0$. In view of (A.8), (A.9), (A.11), and (A.12), since $\log p = o(n^{\min\{\alpha_1/4,1\}-2\xi_0})$ by assumption, we obtain that

$$a_n = \widetilde{C}_1 p\exp(-\widetilde{C}_2 n^{\min\{\alpha_1/2,1\}}) + \widetilde{C}_4 p^2\exp(-\widetilde{C}_5 n^{\min\{\alpha_1/4,1\}})$$
$$+ \widetilde{C}_6 q^2\exp(-\widetilde{C}_7 K^2\widetilde{C}_9^{-2} n^{\min\{\alpha_1/4,1\}-2\xi_0}) = o(1)$$

with the above choice of ϵ, and that with probability at least $1-a_n$, the desired results in the theorem hold with $\kappa_0 = K(1-2\widetilde{C}_8/\widetilde{C}_9)$ and $\kappa = K(1-64\widetilde{C}_8/\widetilde{C}_9)$. This concludes the proof of Theorem 2.

B Verification of Condition 1

Kong et al. (2017) considered high-dimensional interaction models with multi-response $\widetilde{Y} = (Y_1, \ldots, Y_m)^T$ and proposed a two-stage interaction identification method, called the interaction pursuit via distance correlation, which exploits feature screening applied to transformed variables with distance correlation (Székely et al., 2007) followed by feature selection. They showed the screening step in the method of interaction pursuit with distance correlation enjoys the sure screening property for any positive integer m; see Theorem 1 in Kong et al. (2017) for details. Thus, Condition 1 in our paper holds under those conditions for Theorem 1 in Kong et al. (2017). Next, we would like to show that for the interaction model (1) the method of interaction pursuit with the Pearson correlation can also enjoy the sure screening property.

Following the idea of the interaction pursuit via distance correlation in Kong et al. (2017), we can identify the set of active interaction variables $\mathcal{A}$ by ranking the marginal correlations $\mathrm{corr}(X_k^2, Y^2)$ in magnitude, and then retaining the top ones with absolute correlations bounded from below by some positive threshold. Similarly, we can identify the set of important main effects $\mathcal{B}$ through the marginal correlations $\mathrm{corr}(X_j, Y)$. Define two population quantities

$$\omega_k = \frac{\mathrm{cov}(X_k^2, Y^2)}{\sqrt{\mathrm{var}(X_k^2)}} \quad \text{and} \quad \omega_j^* = \frac{\mathrm{cov}(X_j, Y)}{\sqrt{\mathrm{var}(X_j)}} \tag{B.1}$$

with $1 \le k, j \le p$ for interaction variables and main effects, respectively. Observe that $\mathrm{corr}(X_k^2, Y^2) = \omega_k / \{\mathrm{var}(Y^2)\}^{1/2}$ and $\mathrm{corr}(X_j, Y) = \omega_j^* / \{\mathrm{var}(Y)\}^{1/2}$. Denote by $\widehat{\omega}_k$ and $\widehat{\omega}_j^*$ the empirical versions of ω_k and ω_j^*, respectively, constructed by plugging in the corresponding sample covariances based on the sample $\{(X_{i1}, \ldots, X_{ip}, Y_i), i = 1, \ldots, n\}$. Then in the screening step of the interaction pursuit, we estimate the sets of active interaction variables $\mathcal{A}$ and important main effects $\mathcal{B}$ as

$$\widehat{\mathcal{A}} = \{1 \le j \le p : |\widehat{\omega}_k| \ge \tau\} \quad \text{and} \quad \widehat{\mathcal{B}} = \left\{1 \le j \le p : |\widehat{\omega}_j^*| \ge \widetilde{\tau}\right\}, \tag{B.2}$$

where τ and $\widetilde{\tau}$ are some positive thresholds. Based on the retained interaction variables in $\widehat{\mathcal{A}}$, we can construct all pairwise interactions as

$$\widehat{\mathcal{I}} = \left\{(k,\ell) : k,\ell \in \widehat{\mathcal{A}} \text{ and } k < \ell\right\}. \tag{B.3}$$

Finally the set of important features $\mathcal{M}$ can then be estimated as $\widehat{\mathcal{M}} = \widehat{\mathcal{A}} \cup \widehat{\mathcal{B}}$. Although our approach for estimating the set $\mathcal{B}$ is the same as SIS, the theoretical developments on the screening property for main effects are distinct from those in Fan and Lv (2008) due to the presence of interactions in our model.

It is worth mentioning that $\widehat{\mathcal{I}}$ generally provides an overestimate of the set of important interactions $\mathcal{I}$, in the sense that some interactions in the constructed set $\widehat{\mathcal{I}}$ may be unimportant ones. This is, however, not an issue for the purpose of interaction screening and will be addressed later in the selection step of our method. We would like to remark that this screening step is similar to the screening step of Kong et al. (2017) except that the distance correlation is used in Kong et al. (2017). Next, we provide some sufficient conditions that ensures Condition 1 also holds for the method of interaction pursuit via Person correlation. To this end, we provide the following sufficient conditions.

Condition 6. *There exist constants* $0 \le \xi_1, \xi_2 < 1$ *such that* $s_1 = |\mathcal{I}| = O(n^{\xi_1})$ *and* $s_2 = |\mathcal{B}| = O(n^{\xi_2})$, *and* $|\beta_0|, \|\boldsymbol{\beta}_0\|_\infty, \|\boldsymbol{\gamma}_0\|_\infty = O(1)$ *with* $\|\cdot\|_\infty$ *denoting the vector* L_∞*-norm.*

Condition 7. *There exist some constants* $0 \le \kappa_1, \kappa_2 < 1/2$ *and* $c_2 > 0$ *such that* $\min_{k\in\mathcal{A}} |\omega_k| \ge 2c_2 n^{-\kappa_1}$ *and* $\min_{j\in\mathcal{B}} |\omega_j^*| \ge 2c_2 n^{-\kappa_2}$.

Condition 6 allows the numbers of important interactions and important main effects to grow with the sample size n, and imposes an upper bound on the magnitude of true regression coefficients. Clearly, Condition 6 entails that the number of active interaction variables is at most $2s_1$, that is, $|\mathcal{A}| \le 2s_1$.

Condition 7 puts constraints on the minimum marginal correlations, through different forms, for active interaction variables and important main effects, respectively. It is

analogous to Condition 3 in Fan and Lv (2008), and can be understood as an assumption on the minimum signal strength in the feature screening setting. Smaller constants κ_1 and κ_2 correspond to stronger marginal signals. This condition is crucial for ensuring that the marginal utilities carry enough information about the active interaction variables and important main effects.

To gain more insights into Condition 7, consider the specific case of $(X_1, \ldots, X_p)^T \sim N(0, I_p)$. Note that $\text{var}(X_k^2)$ are uniformly bounded by Condition 2. Then it can be shown that the constraint of $\min_{k\in\mathcal{A}} |\omega_k| \geq 2c_2 n^{-\kappa_1}$ in Condition 7 is equivalent to that of

$$\min_{k\in\mathcal{A}} \left(\beta_{0,k}^2 + \sum_{j=1}^{k-1} \gamma_{0,jk}^2 + \sum_{\ell=k+1}^{p} \gamma_{0,k\ell}^2\right) \geq cn^{-\kappa_1},$$

where c is some positive constant which may be different from c_2. Thus Condition 7 can be understood as constraints imposed indirectly on the true nonzero regression coefficients.

Under these conditions, Lemma 1 in the supplementary material shows that the sample estimates of the marginal utilities are sufficiently close to the population ones with significant probability, and establishes the sure screening property for both interaction and main effect screening. Thus Condition 1 holds under these conditions.

References

Barut, E., J. Fan, and A. Verhasselt (2016). Conditional sure independence screening. *Journal of the American Statistical Association 111*, 1266–1277.

Belloni, A., V. Chernozhukov, D. Chetverikov, C. Hansen, and K. Kato (2018). High-dimensional econometrics and regularized GMM. *arXiv preprint*, arXiv:1806.01888.

Bien, J., J. Taylor, and R. Tibshirani (2013). A lasso for hierarchical interactions. *The Annals of Statistics 41*, 1111–1141.

Candes, E. and T. Tao (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics 35*, 2313–2351.

Cattaneo, M. D., M. Jansson, and X. Ma (2019). Two-step estimation and inference with possibly many included covariates. *The Review of Economic Studies 86*, 1095–1122.

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

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

Fan, J., A. Furger, and D. Xiu (2016). Incorporating global industrial classification standard into portfolio allocation: A simple factor-based large covariance matrix estimator with high-frequency data. *Journal of Business & Economic Statistics 34*, 489–503.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*, 1348–1360.

Fan, J. and J. Lv (2008). Sure independence screening for ultrahigh dimensional feature space (with discussion). *Journal of the Royal Statistical Society, Series B 70*, 849–911.

Fan, J. and R. Song (2010). Sure independence screening in generalized linear models with NP-dimensionality. *The Annals of Statistics 38*, 3567–3604.

Fan, Y. and J. Lv (2014). Asymptotic properties for combined L_1 and concave regularization. *Biometrika 101*, 57–70.

Fan, Y. and C. Y. Tang (2013). Tuning parameter selection in high dimensional penalized likelihood. *Journal of the Royal Statistical Society, Series B*, 531–552.

Gosik, K., L. Sun, V. M. Chinchilli, and R. Wu (2018). An ultrahigh-dimensional mapping model of high-order epistatic networks for complex traits. *Current Genomics 19*, 384–394.

Hall, P. and J.-H. Xue (2014). On selecting interacting features from high-dimensional data. *Computational Statistics & Data Analysis 71*, 694–708.

Hao, N., Y. Feng, and H. H. Zhang (2018). Model selection for high-dimensional quadratic regression via regularization. *Journal of the American Statistical Association 113*, 615–625.

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

Jiang, B. and J. S. Liu (2014). Variable selection for general index models via sliced inverse regression. *The Annals of Statistics 42*, 1751–1786.

Ke, Y., H. Lian, and W. Zhang (2020). High-dimensional dynamic covariance matrices with homogeneous structure. *Journal of Business & Economic Statistics*, to appear.

Kong, Y., D. Li, Y. Fan, and J. Lv (2017). Interaction pursuit in high-dimensional multi-response regression via distance correlation. *The Annals of Statistics 45*, 897–922.

Li, R., W. Zhong, and L. Zhu (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association 107*, 1129–1139.

Lv, J. and Y. Fan (2009). A unified approach to model selection and sparse recovery using regularized least squares. *The Annals of Statistics 37*, 3498–3528.

Musani, S. K., D. Shriner, N. Liu, R. Feng, C. S. Coffey, N. Yi, H. K. Tiwari, and D. B. Allison (2007). Detection of gene× gene interactions in genome-wide association studies of human population data. *Human Heredity 63*, 67–84.

Ritchie, M. D., L. W. Hahn, N. Roodi, L. R. Bailey, W. D. Dupont, F. F. Parl, and J. H. Moore (2001). Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *The American Journal of Human Genetics 69*, 138–147.

Székely, G. J., M. L. Rizzo, and N. K. Bakirov (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics 35*, 2769–2794.

Tang, C. Y., E. X. Fang, and Y. Dong (2020). High-dimensional interactions detection with sparse principal hessian matrix. *Journal of Machine Learning Research 21*, 1–25.

Tian, Y. and Y. Feng (2021). RaSE: A variable screening framework via random subspace ensembles. *arXiv preprint arXiv:2102.03892*.

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

Uematsu, Y. and S. Tanaka (2019). High-dimensional macroeconomic forecasting and variable selection via penalized regression. *The Econometrics Journal 22*, 34–56.

Wang, H. (2009). Forward regression for ultra-high dimensional variable screening. *Journal of the American Statistical Association 104*, 1512–1524.

Xu, J., C. D. Langefeld, S. L. Zheng, E. M. Gillanders, B.-l. Chang, S. D. Isaacs, A. H. Williams, K. E. Wiley, L. Dimitrov, D. A. Meyers, et al. (2004). Interaction effect of PTEN and CDKN1B chromosomal regions on prostate cancer linkage. *Human Genetics 115*, 255–262.

Yan, X. and J. Bien (2017). Hierarchical sparse modeling: A choice of two group lasso formulations. *Statistical Science 32*, 531–560.

Yuan, M., V. R. Joseph, and H. Zou (2009). Structured variable selection and estimation. *The Annals of Applied Statistics 3*, 1738–1757.

Zhao, J. and C. Leng (2016). An analysis of penalized interaction models. *Bernoulli 22*, 1937–1961.

Zheng, Z., J. Lv, and W. Lin (2021). Nonsparse learning with latent variables. *Operations Research 69*, 346–359.

Zhou, M., M. Dai, Y. Yao, J. Liu, C. Yang, and H. Peng (2019). BOLT-SSI: A statistical approach to screening interaction effects for ultra-high dimensional data. *arXiv preprint*, arXiv:1902.03525.