The Annals of Statistics
2017, Vol. 45, No. 2, 897–922
DOI: 10.1214/16-AOS1474

INTERACTION PURSUIT IN HIGH-DIMENSIONAL MULTI-RESPONSE REGRESSION VIA DISTANCE CORRELATION[1]

BY YINFEI KONG*, DAOJI LI†, YINGYING FAN‡ AND JINCHI LV‡

California State University, Fullerton, University of Central Florida† and University of Southern California‡*

Feature interactions can contribute to a large proportion of variation in many prediction models. In the era of big data, the coexistence of high dimensionality in both responses and covariates poses unprecedented challenges in identifying important interactions. In this paper, we suggest a two-stage interaction identification method, called the interaction pursuit via distance correlation (IPDC), in the setting of high-dimensional multi-response interaction models that exploits feature screening applied to transformed variables with distance correlation followed by feature selection. Such a procedure is computationally efficient, generally applicable beyond the heredity assumption, and effective even when the number of responses diverges with the sample size. Under mild regularity conditions, we show that this method enjoys nice theoretical properties including the sure screening property, support union recovery and oracle inequalities in prediction and estimation for both interactions and main effects. The advantages of our method are supported by several simulation studies and real data analysis.

1. Introduction. Recent years have seen a surge of interests on interaction identification in the high-dimensional setting by many researchers. For instance, Hall and Xue [8] proposed a recursive approach to identify important interactions among covariates, where all p covariates are first ranked by the generalized correlation and then only the top $p^{1/2}$ ones are retained to construct pairwise interactions of order $O(p)$ for further screening and selection of both interactions and main effects. A forward selection based screening procedure was introduced in [9] for identifying interactions in a greedy fashion under the heredity assumption. Such an assumption in the strong sense requires that an interaction between two covariates should be included in the model only if both main effects are important, while the weak version relaxes such a constraint to the presence of at least one main effect being important. Regularization methods have also been used for interaction selection under the heredity assumption. See, for example, [3, 19] and [1]. Under the inverse modeling framework, [11] proposed a new method, called the

Received December 2015.

[1] Supported by NSF CAREER Awards DMS-09-55316 and DMS-11-50318 and a grant from the Simons Foundation.

MSC2010 subject classifications. Primary 62H12, 62J02; secondary 62F07, 62F12.

Key words and phrases. Interaction pursuit, distance correlation, square transformation, multi-response regression, high dimensionality, sparsity.

sliced inverse regression for interaction detection (SIRI), which can detect pairwise interactions among covariates without the heredity assumption. The theoretical development in [11] relies primarily on the joint normality assumption on the covariates. The innovated interaction screening procedure was introduced in [7] for high-dimensional nonlinear classification with no heredity assumption.

Although the aforementioned methods can perform well in many scenarios, they may have two potential limitations. First, those approaches assume mainly interaction models with a single response, while the coexistence of multiple responses becomes increasingly common in the big data era. Second, those developments are usually built upon the strong or weak heredity assumption, or the normality assumption, which may not be satisfied in certain real applications.

To enable broader applications in practice, in this paper we consider the following high-dimensional multi-response interaction model

$$\mathbf{y} = \boldsymbol{\alpha} + \mathbf{B}_{\mathbf{x}}^T \mathbf{x} + \mathbf{B}_{\mathbf{z}}^T \mathbf{z} + \mathbf{w}, \tag{1}$$

where $\mathbf{y} = (Y_1, \dots, Y_q)^T$ is a q-dimensional vector of responses, $\mathbf{x} = (X_1, \dots, X_p)^T$ is a p-dimensional vector of covariates, $\mathbf{z}$ is a $p(p-1)/2$-dimensional vector of all pairwise interactions between covariates X_j's, $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_q)^T$ is a q-dimensional vector of intercepts, $\mathbf{B}_{\mathbf{x}} \in \mathbb{R}^{p \times q}$ and $\mathbf{B}_{\mathbf{z}} \in \mathbb{R}^{[p(p-1)/2] \times q}$ are regression coefficient matrices for the main effects and interactions, respectively, and $\mathbf{w} = (W_1, \dots, W_q)^T$ is a q-dimensional vector of random errors with mean zero and being independent of $\mathbf{x}$. Each response in this model is allowed to have its own regression coefficients, and to simplify the presentation, the covariate vector $\mathbf{x}$ is assumed to be centered with mean zero. Commonly encountered is the setting of high dimensionality in both responses and covariates, where the numbers of responses and covariates, q and p, can diverge with the sample size. It is of practical importance to consider sparse models in which the rows of the coefficient matrices $\mathbf{B}_{\mathbf{x}}$ and $\mathbf{B}_{\mathbf{z}}$ are sparse with only a fraction of nonzeros. We aim at identifying the important interactions and main effects, which have nonzero regression coefficients, that contribute to the responses.

Interaction identification in the multi-response interaction model (1) with large p and q is intrinsically challenging. The difficulties include the high dimensionality in responses, the high computational cost caused by the existence of a large number of interactions among covariates, and the technical challenges associated with the complex model structure. The idea of variable screening can speed up the computation. Yet, under model setting (1) most existing variable screening methods based on the marginal correlation may no longer work. To appreciate this, let us consider a specific case of model (1) with only one response

$$Y = \alpha + \sum_{j=1}^{p} \beta_j X_j + \sum_{k=1}^{p-1} \sum_{\ell=k+1}^{p} \gamma_{k\ell} X_k X_\ell + W, \tag{2}$$

where all the notation is the same as therein with $\mathbf{B}_{\mathbf{x}} = (\beta_j)_{1 \le j \le p}$ and $\mathbf{B}_{\mathbf{z}} = (\gamma_{k\ell})_{1 \le k \le p-1, k+1 \le \ell \le p}$. For simplicity, assume that the covariates $X_1, \dots, X_p$ are independent of each other. Then under the above model setting (2), it is easy to see that

$$E(Y|X_j) = \alpha + \beta_j X_j. \tag{3}$$

This representation shows that if some covariate X_j is an unimportant main effect with $\beta_j = 0$, then the conditional mean of Y given X_j is free of X_j, regardless of whether X_j contributes to interactions or not. When such a covariate X_j indeed appears in an important interaction, variable screening methods based on the marginal correlations of Y and X_k's are not capable of detecting X_j if the heredity assumption fails to hold. As a consequence, there is an important need for new proposals of interaction screening. When the covariates are correlated, the conditional mean (3) may depend on X_j indirectly through correlations with other covariates when $\beta_j = 0$. Such a relationship can, however, be still weak if the correlations between X_j and other covariates are weak.

To address the aforementioned challenges, we suggest a new two-stage approach to interaction identification, named the interaction pursuit via distance correlation (IPDC), exploiting the idea of interaction screening and selection. In the screening step, we first transform the responses and covariates and then perform variable screening based on the transformed responses and covariates. Such a transformation enhances the dependence of responses on covariates that contribute to important interactions or main effects. The novelty of our interaction screening method is that it aims at recovering variables that contribute to important interactions instead of finding these interactions directly, which reduces the computational cost substantially from a factor of $O(p^2)$ to $O(p)$. To take advantage of the correlation structure among multiple responses, we build our marginal utility function using the distance correlation proposed in [18]. After the screening step, we conduct interaction selection by constructing pairwise interactions with the retained variables from the first step, and applying the group regularization method to further select important interactions and main effects for the multi-response model in the reduced feature space.

The main contributions of this paper are twofold. First, the suggested IPDC method provides a computationally efficient approach to interaction screening and selection in ultra-high dimensional interaction models. Such a procedure accommodates the model setting with a diverging number of responses, and is generally applicable without the requirement of the heredity assumption. Second, our procedure is theoretically justified to be capable of retaining all covariates that contribute to important interactions or main effects with asymptotic probability one, the so-called sure screening property [5, 15], in the screening step. In the selection step, it is also shown to enjoy nice sampling properties for both interactions and main effects such as the support union recovery and oracle inequalities in prediction and estimation. In particular, there are two key messages that are delivered in this paper: a separate screening step for interactions can significantly enhance the screening performance if one aims at finding important interactions, and screening interaction variables can be more effective and efficient than screening interactions directly due to the noise accumulation. The former message is elaborated more with a numerical example presented in Section C.1 of the Supplementary Material [12].

The rest of the paper is organized as follows. Section 2 introduces the new interaction screening approach and studies its theoretical properties. We illustrate the advantages of the proposed procedure using several simulation studies in Section 3 and a real data example in Section 4. Section 5 discusses some possible extensions of our method. The proofs of main results are relegated to the Appendix. The details about the post-screening interaction selection, additional numerical studies and additional proofs of main results as well as additional technical details are provided in the Supplementary Material.

2. A new interaction screening approach.

2.1. *Motivation of the new method.* To facilitate the presentation, we call $X_k X_\ell$ an important interaction if the corresponding row of $\mathbf{B}_\mathbf{z}$ is nonzero, and X_k an active interaction variable if there exists some $1 \le \ell \ne k \le p$ such that $X_k X_\ell$ is an important interaction. Denote by $\mathcal{I}$ the set of all important interactions. Similarly, X_j is referred to as an important main effect if its associated row of $\mathbf{B}_\mathbf{x}$ is nonzero. It is of crucial importance to identify both the set $\mathcal{A}$ of all active interaction variables and the set $\mathcal{M}$ of all important main effects.

Before presenting our main ideas, let us revisit the specific example (2) discussed in the Introduction. A phenomenon mentioned there is that variable screening methods using the marginal correlations between the response and covariates can fail to detect active interaction variables that have no main effects. We now consider the square transformation for the response. Some standard calculations (see Section E.1 of the Supplementary Material) yield

$$\begin{aligned} E(Y^2|X_j) = {} & \left[\beta_j^2 + \sum_{k=1}^{j-1} \gamma_{kj}^2 E(X_k^2) + \sum_{\ell=j+1}^{p} \gamma_{j\ell}^2 E(X_\ell^2)\right] X_j^2 \\ & + 2\left[\beta_j \alpha + \sum_{k=1}^{j-1} \beta_k \gamma_{kj} E(X_k^2) + \sum_{\ell=j+1}^{p} \beta_\ell \gamma_{j\ell} E(X_\ell^2)\right] X_j + C_j, \end{aligned}$$

where C_j is some constant that is free of X_j. This shows that the conditional mean $E(Y^2|X_j)$ is linear in X_j^2 as long as X_j is an active interaction variable, that is, γ_{kj} or $\gamma_{j\ell} \ne 0$ for some k or ℓ, regardless of whether it is also an important main effect or not. In fact, we can see from the above representation that the coefficient of X_j^2 reflects the cumulative contribution of covariate X_j to response Y as both an interaction variable and a main effect.

Motivated by the above example, we consider the approach of screening interaction variables via some marginal utility function for the transformed variables Y^2

and X_j^2, with the square transformation applied to both the response and covariates. Such an idea has been exploited in [6] for interaction screening in the setting of single-response interaction models. To rank the relative importance of features, they calculated the Pearson correlations between Y^2 and X_j^2. This idea is, however, no longer applicable when there are multiple responses, since the Pearson correlation is not well defined for the pair of q-vector $\mathbf{y}$ of responses with $q > 1$ and covariate X_j. A naive strategy is to screen the interaction variables for each response Y_k with $1 \le k \le q$ using the approach of [6]. Such a naive procedure can suffer from several potential drawbacks. First, it may be inefficient and can result in undesirable results since the correlation structure among the responses $Y_1, \dots, Y_q$ is completely ignored. Second, when q is large it may retain too many interaction variables in total, which can in turn cause difficulty in model interpretation and high computational cost when further selecting active interaction variables.

To address the afore-discussed issues, we propose to construct the marginal utility function exploiting the distance correlation introduced in [18]. More specifically, we identify the set of all active interaction variables $\mathcal{A}$ by ranking the distance correlations between the squared covariates X_j^2 and the squared response vector $\mathbf{y} \circ \mathbf{y}$, where $\circ$ denotes the Hadamard (componentwise) product of two vectors. The distance correlation

$$\operatorname{dcorr}(\mathbf{u}, \mathbf{v}) = \frac{\operatorname{dcov}(\mathbf{u}, \mathbf{v})}{\sqrt{\operatorname{dcov}(\mathbf{u}, \mathbf{u}) \operatorname{dcov}(\mathbf{v}, \mathbf{v})}}$$

is well defined for any two random vectors $\mathbf{u} \in \mathbb{R}^{d_u}$ and $\mathbf{v} \in \mathbb{R}^{d_v}$ of arbitrary mixed dimensions, where the distance covariance between $\mathbf{u}$ and $\mathbf{v}$ is given by

$$\operatorname{dcov}^2(\mathbf{u}, \mathbf{v}) = \frac{1}{c_{d_u} c_{d_v}} \int_{\mathbb{R}^{d_u + d_v}} \frac{|\varphi_{\mathbf{u},\mathbf{v}}(\mathbf{s}, \mathbf{t}) - \varphi_{\mathbf{u}}(\mathbf{s}) \varphi_{\mathbf{v}}(\mathbf{t})|^2}{\|\mathbf{s}\|^{d_u+1} \|\mathbf{t}\|^{d_v+1}} \, d\mathbf{s} \, d\mathbf{t}.$$

Here, $c_m = \pi^{(m+1)/2} / \Gamma\{(m+1)/2\}$ is the half area of the unit sphere $S^m \subset \mathbb{R}^{m+1}$, $\varphi_{\mathbf{u},\mathbf{v}}(\mathbf{s}, \mathbf{t})$, $\varphi_{\mathbf{u}}(\mathbf{s})$, and $\varphi_{\mathbf{v}}(\mathbf{t})$ are the characteristic functions of $(\mathbf{u}, \mathbf{v})$, $\mathbf{u}$, and $\mathbf{v}$, respectively, and $\|\cdot\|$ denotes the Euclidean norm. Compared to the Pearson correlation, it also has the advantage that the distance correlation of two random vectors is zero if and only if they are independent. Moreover, the distance correlation of two univariate Gaussian random variables is a strictly increasing function of the absolute value of the Pearson correlation between them. See [18] for more properties and discussions of the distance correlation, and [10] for a fast algorithm for computing the distance correlation.

It is worth mentioning that [14] introduced a model-free feature screening procedure based on the distance correlations of the original response and covariates. Their method is applicable to the cases of multiple responses and grouped covariates. Yet we found that the use of distance correlations for the transformed response vector and covariates, $\mathbf{y} \circ \mathbf{y}$ and X_j^2, can result in improved performance in interaction variable screening. The specific example considered in Section 2.1

provides some intuitive explanation of this phenomenon. To further illustrate this point, we generated 200 data sets from the following simple interaction model:

$$Y = X_1 X_2 + W, \tag{4}$$

where the covariate vector $\mathbf{x} = (X_1, \ldots, X_p)^T \sim N(\mathbf{0}, \boldsymbol{\Sigma})$ with $p = 1000$ and $\boldsymbol{\Sigma} = (\rho^{|j-k|})_{1 \le j,k \le p}$, ρ ranging in $(-1, 1)$ measures the correlation level among covariates, and the random error $W \sim N(0, 1)$ is independent of $\mathbf{x}$. As shown in Figure 1, the sample distance correlation between X_1^2 and Y^2 is much larger than that between X_1 and Y. For covariates having weak correlation with active interaction variables X_1 and X_2, such as X_{10} and X_{1000}, the square transformation does not in-

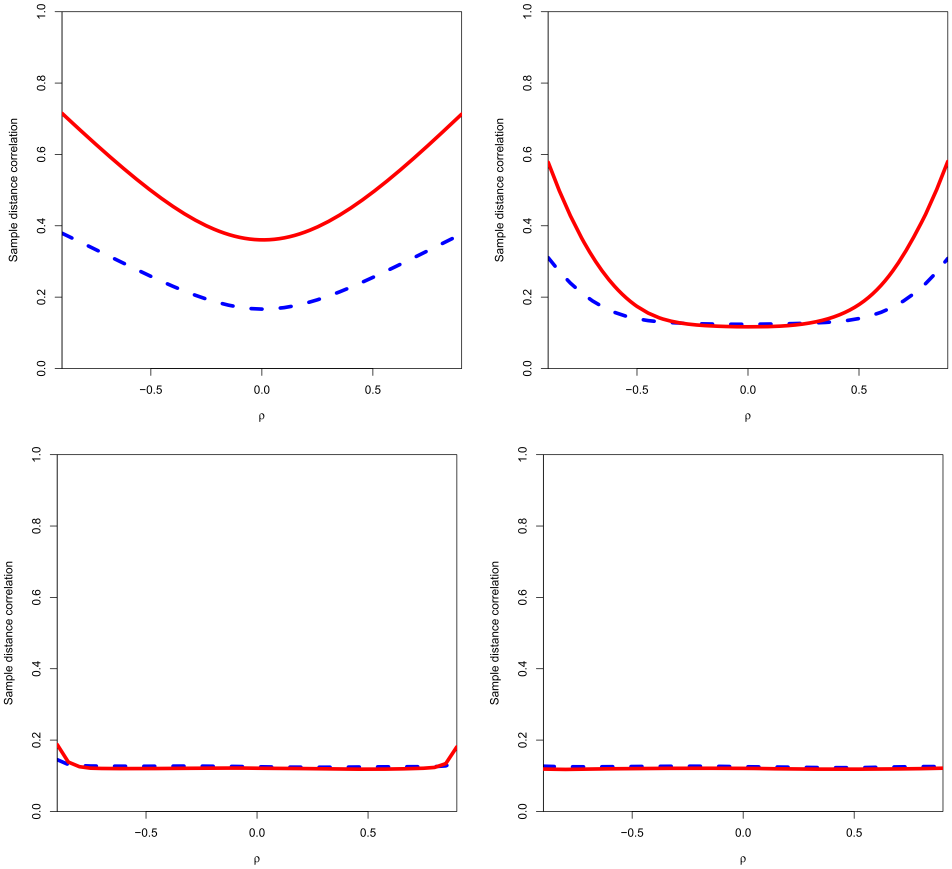

FIG. 1. *Plots of sample distance correlations as a function of correlation level* ρ *based on model* (4). *Top left*: $\widehat{\mathrm{dcorr}}(X_1^2, Y^2)$ (*solid*) *and* $\widehat{\mathrm{dcorr}}(X_1, Y)$ (*dashed*); *top right*: $\widehat{\mathrm{dcorr}}(X_3^2, Y^2)$ (*solid*) *and* $\widehat{\mathrm{dcorr}}(X_3, Y)$ (*dashed*); *bottom left*: $\widehat{\mathrm{dcorr}}(X_{10}^2, Y^2)$ (*solid*) *and* $\widehat{\mathrm{dcorr}}(X_{10}, Y)$ (*dashed*); *bottom right*: $\widehat{\mathrm{dcorr}}(X_{1000}^2, Y^2)$ (*solid*) *and* $\widehat{\mathrm{dcorr}}(X_{1000}, Y)$ (*dashed*).

crease their distance correlations with the response. The numerical studies in Sections 3 and 4 also confirm the advantages of our method over the procedure in [14].

2.2. *Interaction screening.* Suppose we have a sample $(\mathbf{y}_i, \mathbf{x}_i)_{i=1}^n$ of n independent and identically distributed (i.i.d.) observations from $(\mathbf{y}, \mathbf{x})$ in the multi-response interaction model (1). For each $1 \le k \le p$, denote by $\operatorname{dcorr}(X_k^2, \mathbf{y} \circ \mathbf{y})$ the distance correlation between the squared covariate X_k^2 and squared response vector $\mathbf{y} \circ \mathbf{y}$. The idea of the screening step of our IPDC procedure is to rank the importance of the interaction variables X_k using the sample version of distance correlations $\operatorname{dcorr}(X_k^2, \mathbf{y} \circ \mathbf{y})$. Similarly, we conduct screening of main effects based on the sample version of distance correlations $\operatorname{dcorr}(X_j, \mathbf{y})$ between covariates X_j and response vector $\mathbf{y}$.

For notational simplicity, we write $X_k^* = X_k^2$, $\widetilde{\mathbf{y}} = \mathbf{y}/\sqrt{q}$, and $\mathbf{y}^* = \widetilde{\mathbf{y}} \circ \widetilde{\mathbf{y}} = \mathbf{y} \circ \mathbf{y}/q$. Define two population quantities

$$\omega_k^* = \frac{\operatorname{dcov}^2(X_k^*, \mathbf{y}^*)}{\sqrt{\operatorname{dcov}^2(X_k^*, X_k^*)}} \quad \text{and} \quad \omega_j = \frac{\operatorname{dcov}^2(X_j, \widetilde{\mathbf{y}})}{\sqrt{\operatorname{dcov}^2(X_j, X_j)}} \tag{5}$$

with $1 \le k, j \le p$ for interaction variables and main effects, respectively. 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 distance covariances based on the sample $(\mathbf{y}_i, \mathbf{x}_i)_{i=1}^n$. According to [18], the sample distance covariance between any two random vectors $\mathbf{u}$ and $\mathbf{v}$ based on a sample $(\mathbf{u}_i, \mathbf{v}_i)_{i=1}^n$ is given by

$$\widehat{\operatorname{dcov}}^2(\mathbf{u}, \mathbf{v}) = \widehat{S}_1 + \widehat{S}_2 - 2\widehat{S}_3,$$

where the three quantities are defined as $\widehat{S}_1 = n^{-2} \sum_{i,j=1}^n \|\mathbf{u}_i - \mathbf{u}_j\| \|\mathbf{v}_i - \mathbf{v}_j\|$, $\widehat{S}_2 = [n^{-2} \sum_{i,j=1}^n \|\mathbf{u}_i - \mathbf{u}_j\|][n^{-2} \sum_{i,j=1}^n \|\mathbf{v}_i - \mathbf{v}_j\|]$, and $\widehat{S}_3 = n^{-3} \sum_{i,j,k=1}^n \|\mathbf{u}_i - \mathbf{u}_k\| \|\mathbf{v}_j - \mathbf{v}_k\|$. In view of

$$\operatorname{dcorr}^2(X_k^2, \mathbf{y} \circ \mathbf{y}) = \operatorname{dcorr}^2(X_k^*, \mathbf{y}^*) = \omega_k^* / \{\operatorname{dcov}^2(\mathbf{y}^*, \mathbf{y}^*)\}^{1/2}$$

and

$$\operatorname{dcorr}^2(X_j, \mathbf{y}) = \operatorname{dcorr}^2(X_j, \widetilde{\mathbf{y}}) = \omega_j / \{\operatorname{dcov}^2(\widetilde{\mathbf{y}}, \widetilde{\mathbf{y}})\}^{1/2},$$

the procedure of screening the interaction variables and main effects via distance correlations $\operatorname{dcorr}(X_k^2, \mathbf{y} \circ \mathbf{y})$ and $\operatorname{dcorr}(X_j, \mathbf{y})$ suggested above is equivalent to that of thresholding the quantities ω_k^*'s and ω_j's, respectively.

More specifically, in the screening step of IPDC we estimate the sets of important main effects $\mathcal{M}$ and active interaction variables $\mathcal{A}$ as

$$\widehat{\mathcal{M}} = \{1 \le j \le p : \widehat{\omega}_j \ge \tau_1\} \quad \text{and} \quad \widehat{\mathcal{A}} = \{1 \le k \le p : \widehat{\omega}_k^* \ge \tau_2\}, \tag{6}$$

where τ_1 and τ_2 are some positive thresholds. With the set $\widehat{\mathcal{A}}$ of retained interaction variables, we construct a set of pairwise interactions

$$\widehat{\mathcal{I}} = \{(k, l) : 1 \le k < l \le p \text{ and } k, l \in \widehat{\mathcal{A}}\}. \tag{7}$$

This gives a new interaction screening procedure. It is worth mentioning that the set of constructed interactions $\widehat{\mathcal{I}}$ tends to overestimate the set of all important interactions $\mathcal{I}$ since the goal of the first step of IPDC is screening interaction variables. Such an issue can be addressed in the selection step of IPDC investigated in Section B of the Supplementary Material.

2.3. *Sure screening property.* We now study the sampling properties of the newly proposed interaction screening procedure. Some mild regularity conditions are needed for our analysis.

CONDITION 1. *Both* $\operatorname{dcov}(X_k, X_k)$ *and* $\operatorname{dcov}(X_k^2, X_k^2)$ *are bounded away from zero uniformly in* k.

CONDITION 2. *There exists some constant* $c_0 > 0$ *such that* $E\{\exp(c_0 X_k^2)\}$ *and* $E\{\exp(c_0\|\mathbf{y}\|/\sqrt{q})\}$ *are uniformly bounded.*

CONDITION 3. *There exist some constants* $c_1, c_2 > 0$ *and* $0 \le \kappa_1, \kappa_2 < 1/2$ *such that* $\min_{j\in\mathcal{M}} \omega_j \ge 3c_1 n^{-\kappa_1}$ *and* $\min_{k\in\mathcal{A}} \omega_k^* \ge 3c_2 n^{-\kappa_2}$.

Condition 1 is a basic assumption requiring that the distance variances of covariates X_k and squared covariates X_k^2 are at least of a constant order. Conditions 2 and 3 are analogous to the regularity conditions in [14]. In particular, Condition 2 controls the tail behavior of the covariates and responses, which facilitates the derivation of deviation probability bounds. Condition 3 also shares the same spirit as Condition 3 in [5], 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 signal strength for active interaction variables and important main effects, respectively. With these regularity conditions, we establish the sure screening property of IPDC in the following theorem.

THEOREM 1. *Under Conditions 1–2 with* $\log p = o(n^{\eta_0})$ *for* $\eta_0 = \min\{(1 - 2\kappa_1)/3, (1 - 2\kappa_2)/5\}$, *there exists some positive constant* C *such that*

$$P\Big(\max_{1\le j\le p} |\widehat{\omega}_j - \omega_j| \ge c_1 n^{-\kappa_1}\Big) \le O\big(\exp\{-Cn^{(1-2\kappa_1)/6}\}\big), \tag{8}$$

$$P\Big(\max_{1\le k\le p} |\widehat{\omega}_k^* - \omega_k^*| \ge c_2 n^{-\kappa_2}\Big) \le O\big(\exp\{-Cn^{(1-2\kappa_2)/10}\}\big). \tag{9}$$

Assume in addition that Condition 3 holds and set $\tau_1 = 2c_1 n^{-\kappa_1}$ *and* $\tau_2 = 2c_2 n^{-\kappa_2}$. *Then we have*

$$P(\mathcal{M} \subset \widehat{\mathcal{M}} \text{ and } \mathcal{I} \subset \widehat{\mathcal{I}}) = 1 - O\{\exp(-Cn^{\eta_0/2})\}. \tag{10}$$

Theorem 1 reveals that the IPDC enjoys the sure screening property that all active interaction variables and all important main effects can be retained in the reduced model with high probability. In particular, we see that it can handle ultra-high dimensionality with $\log p = o(n^{\eta_0})$. A comparison of the deviation probability bounds in (8) and (9) shows that interaction screening is generally more challenging, and thus needs more restrictive constraint on dimensionality p than main effect screening; see the probability bound (15) and its main effect counterpart for more details. It is also seen that when the marginal signal strength for interactions and main effects becomes stronger, the sure screening property of IPDC holds for higher dimensionality p.

For any feature screening procedure, it is of practical importance to control the dimensionality of the reduced feature space, since feature selection usually follows the screening for further selection of important features in such a space. We next investigate such an aspect for IPDC. Let s_1 and s_2 be the cardinalities of sets of all important main effects $\mathcal{M}$ and all active interaction variables $\mathcal{A}$, respectively. With the thresholds $\tau_1 = 2c_1 n^{-\kappa_1}$ and $\tau_2 = 2c_2 n^{-\kappa_2}$ specified in Theorem 1, we introduce two sets of unimportant main effects and inactive interaction variables

$$\mathcal{M}_1 = \{j \in \mathcal{M}^c : \omega_j \geq c_1 n^{-\kappa_1}\} \quad \text{and} \quad \mathcal{A}_1 = \{k \in \mathcal{A}^c : \omega_k^* \geq c_2 n^{-\kappa_2}\} \tag{11}$$

that are of significant marginal effects. Denote by s_3 and s_4 the cardinalities of these two sets $\mathcal{M}_1$ and $\mathcal{A}_1$, respectively. Larger values of s_3 and s_4 indicate more difficulty in the problem of interaction and main effect screening in the high-dimensional multi-response interaction model (1).

THEOREM 2. *Assume that all the conditions of Theorem 1 hold and set $\tau_1 = 2c_1 n^{-\kappa_1}$ and $\tau_2 = 2c_2 n^{-\kappa_2}$. Then we have*

$$\begin{aligned} &P\{|\widehat{\mathcal{M}}| \leq s_1 + s_3 \text{ and } |\widehat{\mathcal{I}}| \leq (s_2 + s_4)(s_2 + s_4 - 1)/2\} \\ &\quad = 1 - O\{\exp(-Cn^{\eta_0/2})\} \end{aligned} \tag{12}$$

for some positive constant C.

Theorem 2 quantifies how the size of the reduced model for interactions and main effects is related to the thresholding parameters τ_1 and τ_2, and the cardinalities of the two sets $\mathcal{M}_1$ and $\mathcal{A}_1$. In particular, we see that when $s_i = O(n^{\delta_i})$ with some constants $\delta_i \geq 0$ for $1 \leq i \leq 4$, the total number of retained interactions and main effects in the reduced feature space can be controlled as $O(n^{\delta})$ with $\delta = \max\{\delta_1 \vee \delta_3, 2(\delta_2 \vee \delta_4)\}$, where $\vee$ denotes the maximum of two values. In contrast, the dimensionality p is allowed to grow nonpolynomially with sample size n in the rate of $\log p = o(n^{\eta_0})$ with $\eta_0 = \min\{(1 - 2\kappa_1)/3, (1 - 2\kappa_2)/5\}$. The reduced model size can fall below the sample size and be a smaller order of n when both $\max\{\delta_1, \delta_3\} < 1$ and $\max\{\delta_2, \delta_4\} < 1/2$ are satisfied. The post-screening interaction selection and its sampling properties are further investigated in Section B of the Supplementary Material.

3. Simulation studies. We illustrate the finite-sample performance of our method using several simulation examples. Two sets of models are considered for the single-response case and the multi-response case, respectively. This section evaluates the screening performance, while the post-screening selection performance is investigated in Section C.2 of the Supplementary Material.

3.1. *Screening in single-response models.* We begin with the following four high-dimensional single-response interaction models:

$$\text{Model 1: } Y = 2X_1 + 2X_2 + X_1X_2 + W,$$

$$\text{Model 2: } Y = 2X_1 + 3X_1X_2 + 3X_1X_3 + W,$$

$$\text{Model 3: } Y = 3X_1X_2 + 3X_1X_3 + W,$$

$$\text{Model 4: } Y = 3\mathbb{I}(X_{12} \geq 0) + 2X_{22} + 3X_1X_2 + W,$$

where all the notation is the same as in (1) and $\mathbb{I}(\cdot)$ denotes the indicator function. The covariate vector $\mathbf{x} = (X_1, \dots, X_p)^T$ is sampled from the distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$ with covariance matrix $\boldsymbol{\Sigma} = (\rho^{|j-k|})_{1 \leq j,k \leq p}$ for $\rho \in (-1, 1)$, and the error term $W \sim N(0, 1)$ is generated independently of $\mathbf{x}$ to form an i.i.d. sample of size $n = 200$. For each of the four models, we further consider three different settings with $(p, \rho) = (2000, 0.5)$, $(5000, 0.5)$ and $(2000, 0.1)$, respectively. In particular, Models 2 and 3 are adapted from simulation scenarios 2.2 and 2.3 in Jiang and Liu [11], whereas Model 4 is adapted from simulation example 2.b of Li, Zhong and Zhu [14] and accounts for model misspecification since without any prior information, our working model treats X_{12} as a linear predictor instead of $\mathbb{I}(X_{12} \geq 0)$. We see that Model 1 satisfies the strong heredity assumption and Model 2 obeys the weak heredity assumption, while Models 3 and 4 violate the heredity assumption since none of the active interaction variables are important main effects.

We compare the interaction and main effect screening performance of the IPDC with the SIS [5], DCSIS [14], SIRI [11], IP [6] and iFORT and iFORM [9]. Like IPDC, SIRI and IP were developed for screening interaction variables and main effects separately. In particular, SIRI is an iterative procedure, while all others are noniterative ones. For a fair comparison, we adopt the initial screening step described in Section 2.3 of Jiang and Liu [11] to implement SIRI in a noniterative fashion, and keep the top ranked covariates. Since the SIS is originally designed only for main effect screening and the original DCSIS screens variables without the distinction between main effects and interaction variables, we construct pairwise interactions based on the covariates recruited by SIS and DCSIS, and refer to the resulting procedures as SIS2 and DCSIS2, respectively, to distinguish them from the original ones. It is worth mentioning that the SIS2 shares a similar spirit to the TS-SIS procedure proposed in [13], where the difference is that the latter constructs pairwise interactions between the main effects retained by SIS and all p covariates. Following the suggestions of Fan and Lv [5] and Li, Zhong and Zhu

[14], we keep the top $[n/(\log n)]$ variables after ranking for each screening procedure. We examine the screening performance by the proportions of important main effects, important interactions and all of them being retained by each screening procedure over 100 replications.

Table 1 reports the screening results of different methods. In Model 1, all screening methods are able to retain almost all important main effects and interactions across all three settings. The IPDC outperforms SIS2, DCSIS2, SIRI, IP, iFORT and iFORM in Models 2–4 over all three settings. It is seen that SIS2 can barely identify important interactions for those three models. The advantage of IPDC over SIS2, DCSIS2 and SIRI is most pronounced when the heredity assumption is violated as in Models 3 and 4. We also observe significant improvement of IPDC over IP in many of those model settings. When the dimensionality increases from 2000 to 5000 (settings 1 and 2), the problem of interaction and main effect screening becomes more challenging as indicated by the drop of the screening probabilities. Compared to others, IPDC consistently performs well.

It is interesting to observe that in view of settings 1 and 3, the interaction screening performance can be improved in the presence of a higher level of correlation among covariates. One possible explanation is that high correlation among covariates may increase the dependence of the response on the interaction variables, and thus benefit interaction screening. For instance, in Model 2 due to the correlation between the interaction variable X_2 (or X_3) and main effect X_1, the response Y depends on X_2 (or X_3) not only directly through the interaction X_1X_2 (or X_1X_3) but also indirectly through the main effect X_1. Therefore, in this case high correlation among covariates can boost the performance of interaction screening. Similar phenomenon has been documented for DCSIS in the literature; see, for example, Models 1.b and 1.c in Table 2 of Li, Zhong and Zhu [14].

3.2. *Screening in multi-response model.* We next consider the setting of interaction model with multiple responses and specifically Model 5 with $q = 10$ responses:

$$\begin{aligned}
Y_1 &= \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + W_1,\\
Y_2 &= \beta_4 X_1 + \beta_5 X_2 + \beta_6 X_1 X_3 + W_2,\\
Y_3 &= \beta_7 X_1 + \beta_8 X_2 + \beta_9 X_6 X_7 + W_3,\\
Y_4 &= \beta_{10} X_1 + \beta_{11} X_2 + \beta_{12} X_8 X_9 + W_4,\\
Y_5 &= \beta_{13} X_6 X_7 + \beta_{14} X_8 X_9 + W_5,\\
Y_6 &= \beta_{15} X_1 + \beta_{16} X_2 + \beta_{17} X_1 X_2 + W_6,\\
Y_7 &= \beta_{18} X_1 + \beta_{19} X_2 + \beta_{20} X_1 X_3 + W_7,\\
Y_8 &= \beta_{21} X_1 + \beta_{22} X_2 + \beta_{23} X_6 X_7 + W_8,\\
Y_9 &= \beta_{24} X_1 + \beta_{25} X_2 + \beta_{26} X_8 X_9 + W_9,\\
Y_{10} &= \beta_{27} X_6 X_7 + \beta_{28} X_8 X_9 + W_{10},
\end{aligned}$$

TABLE 1

Proportions of important main effects, important interactions and all of them retained by different screening methods. For SIS2, DCSIS2 and SIRI, interactions are constructed using the top $[n/(\log n)]$ covariates ranked by their marginal utilities with the response.

	Model 1				**Model 2**				**Model 3**			**Model 4**			
Method	X_1	X_2	X_1X_2	**All**	X_1	X_1X_2	X_1X_3	**All**	X_1X_2	X_1X_3	**All**	X_{12}	X_{22}	X_1X_2	**All**
							Setting 1: $(p, \rho) = (2000, 0.5)$								
SIS2	1.00	1.00	1.00	1.00	0.95	0.48	0.23	0.20	0.08	0.04	0.04	0.93	1.00	0.05	0.05
iFORT	1.00	1.00	1.00	1.00	0.67	0.00	0.00	0.00	0.00	0.00	0.00	0.56	1.00	0.00	0.00
iFORM	1.00	1.00	1.00	1.00	0.85	0.07	0.02	0.00	0.00	0.00	0.00	0.53	1.00	0.00	0.00
DCSIS2	1.00	1.00	1.00	1.00	1.00	1.00	0.91	0.91	1.00	0.68	0.68	0.99	1.00	0.92	0.91
SIRI	1.00	1.00	1.00	1.00	1.00	0.99	0.88	0.87	1.00	0.73	0.73	0.89	1.00	0.86	0.78
IP	1.00	1.00	1.00	1.00	0.95	1.00	0.87	0.83	1.00	0.90	0.90	0.93	1.00	1.00	0.93
IPDC	1.00	1.00	1.00	1.00	1.00	1.00	0.99	0.99	1.00	0.99	0.99	0.99	1.00	1.00	0.99
							Setting 2: $(p, \rho) = (5000, 0.5)$								
SIS2	1.00	1.00	1.00	1.00	0.94	0.39	0.17	0.15	0.03	0.00	0.00	0.86	1.00	0.03	0.02
iFORT	1.00	1.00	1.00	1.00	0.62	0.00	0.00	0.00	0.00	0.00	0.00	0.40	1.00	0.00	0.00
iFORM	1.00	1.00	1.00	1.00	0.85	0.04	0.01	0.00	0.00	0.00	0.00	0.41	1.00	0.00	0.00
DCSIS2	1.00	1.00	1.00	1.00	1.00	0.99	0.80	0.79	1.00	0.46	0.46	0.99	1.00	0.86	0.85
SIRI	1.00	1.00	1.00	1.00	1.00	0.99	0.81	0.80	1.00	0.63	0.63	0.83	1.00	0.84	0.71
IP	1.00	1.00	1.00	1.00	0.94	1.00	0.73	0.69	1.00	0.85	0.85	0.86	1.00	1.00	0.86
IPDC	1.00	1.00	1.00	1.00	1.00	1.00	0.96	0.96	1.00	0.98	0.98	0.99	1.00	1.00	0.99
							Setting 3: $(p, \rho) = (2000, 0.1)$								
SIS2	1.00	1.00	1.00	1.00	1.00	0.08	0.04	0.00	0.02	0.00	0.00	0.97	1.00	0.00	0.00
iFORT	1.00	1.00	1.00	1.00	0.89	0.00	0.00	0.00	0.00	0.00	0.00	0.64	1.00	0.00	0.00
iFORM	1.00	1.00	1.00	1.00	0.92	0.06	0.01	0.01	0.01	0.00	0.00	0.62	1.00	0.00	0.00
DCSIS2	1.00	1.00	1.00	1.00	1.00	0.71	0.72	0.55	0.19	0.11	0.00	1.00	1.00	0.06	0.06
SIRI	1.00	1.00	1.00	1.00	1.00	0.58	0.58	0.34	0.35	0.37	0.16	0.86	1.00	0.19	0.15
IP	1.00	1.00	0.99	0.99	1.00	0.64	0.64	0.38	0.79	0.75	0.58	0.97	1.00	0.98	0.95
IPDC	1.00	1.00	1.00	1.00	1.00	0.80	0.79	0.62	0.93	0.90	0.84	1.00	1.00	1.00	1.00

where all the notation and setup are the same as in Section 3.1, the covariate vector $\mathbf{x} = (X_1, \dots, X_p)^T$ is sampled from distribution $N(\mathbf{0}, \mathbf{\Sigma})$ with covariance matrix $\mathbf{\Sigma} = (0.5^{|j-k|})_{1 \le j,k \le p}$, and the error vector $\mathbf{w} = (W_1, \dots, W_q)^T \sim N(\mathbf{0}, I_q)$ is independent of $\mathbf{x}$. The nonzero regression coefficients β_k with $1 \le k \le 28$ for all important main effects and interactions are generated independently as $\beta_k = (-1)^U \text{Uniform}(1, 2)$, where U is a Bernoulli random variable with success probability 0.5 and Uniform$(1, 2)$ is the uniform distribution on $[1, 2]$. For simplicity, we consider only the setting of $(n, p, \rho) = (100, 1000, 0.5)$. In Model 5, covariates X_1 and X_2 are both active interaction variables and important main effects, whereas covariates X_3 and X_j with $6 \le j \le 9$ are active interaction variables only.

To simplify the presentation, we examine only the proportions of active interaction variables and important main effects retained by different screening procedures. A direct application of SIS to each response Y_k with $1 \le k \le q$ results in q marginal correlations for each covariate X_j with $1 \le j \le p$. We thus consider two modifications of SIS to deal with multi-response data. Specifically, we exploit two new marginal measures, $\max_{1 \le k \le q} |\widehat{\text{corr}}(Y_k, X_j)|$ and $\sum_{k=1}^{q} |\widehat{\text{corr}}(Y_k, X_j)|$, to quantify the importance of covariates X_j, where $\widehat{\text{corr}}$ denotes the sample correlation. We refer to these two methods as SIS.max and SIS.sum, respectively. The SIRI and IP are not included for comparison in this model since both methods were not designed for multi-response models, while the DCSIS is still applicable since the distance correlation is well defined in such a scenario.

Since feature screening is more challenging in multi-response models, we implement IPDC in a slightly different fashion than in single-response models. Recall that in Section 3.1, IPDC screens interaction variables and main effects separately, and keeps the top $[n/(\log n)]$ of each type of variables. For Model 5, we take a union of these two sets of variables and regard an active interaction variable or important main effect as being retained if such a variable belongs to the union, which can contain up to $2[n/(\log n)]$ variables. Consequently we construct pairwise interactions of all variables in the union. To ensure fair comparison, we keep the top $2[n/(\log n)]$ variables for the other screening methods SIS.max, SIS.sum and DCSIS.

Table 2 summarizes the screening results under Model 5. We see that all methods perform well in recovering variables X_1, X_2 and X_3. Yet only IPDC is able to retain active interaction variables $X_6, \dots, X_9$ with large probability.

3.3. *Screening in multi-response model with discrete covariates.*

We now turn to the scenario of multi-response interaction model with mixed covariate types and specifically Model 6 with $q = 50$ responses and $(n, p) = (100, 1000)$:

$$Y_1 = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_1 X_2 + \beta_6 X_3 X_4 + W_1,$$

$$Y_2 = \beta_7 X_1 + \beta_8 X_2 + \beta_9 X_3 + \beta_{10} X_4 + \beta_{11} X_1 X_3 + \beta_{12} X_4 X_5 + W_2,$$

$$Y_3 = \beta_{13} X_1 + \beta_{14} X_2 + \beta_{15} X_3 + \beta_{16} X_4 + \beta_{17} X_4 X_5 + \beta_{18} X_9 X_{13} + W_3,$$

TABLE 2
Proportions of important main effects and active interaction variables retained by different screening methods

Method	X_1	X_2	X_3	X_6	X_7	X_8	X_9
SIS.max	1.00 (0.00)	1.00 (0.00)	0.98 (0.01)	0.12 (0.03)	0.18 (0.04)	0.13 (0.03)	0.08 (0.03)
SIS.sum	1.00 (0.00)	1.00 (0.00)	0.99 (0.01)	0.17 (0.04)	0.17 (0.04)	0.17 (0.04)	0.17 (0.04)
DCSIS	1.00 (0.00)	1.00 (0.00)	1.00 (0.00)	0.61 (0.05)	0.57 (0.05)	0.72 (0.05)	0.68 (0.05)
IPDC	1.00 (0.00)	1.00 (0.00)	0.99 (0.01)	0.91 (0.03)	0.90 (0.03)	0.95 (0.02)	0.90 (0.03)

$$Y_4 = \beta_{19}X_1 + \beta_{20}X_2 + \beta_{21}X_3 + \beta_{22}X_4 + \beta_{23}X_9X_{12} + \beta_{24}X_{12}X_{13} + W_4,$$
$$Y_5 = \beta_{25}X_9X_{12} + \beta_{26}X_9X_{13} + \beta_{27}X_{12}X_{13} + W_5,$$

and the remaining nine groups of five responses are defined in a similar way to how $Y_6, \dots, Y_{10}$ were defined in Model 5 in Section 3.2, that is, repeating the support of each response but with regression coefficients β_k generated independently from the same distribution as in Model 5. There are several key differences with Model 5. We consider higher response dimensionality $q = 50$, higher population collinearity level $\rho = 0.8$, and larger true model sizes for the responses. The covariates $X_1, \dots, X_p$ are sampled similarly as in Model 5, but the even numbered covariates are further discretized. More specifically, each even numbered covariate is assigned values 0, 1 or 2 if the original continuous covariate takes values below 0, between 0 and 1.5, or above 1.5, respectively, and then centered with mean zero. These discrete covariates are included in the model because in real applications some covariates can also be discrete. For instance, the covariates in the single nucleotide polymorphism (SNP) data are typically coded to take values 0, 1 and 2. In addition, the random errors $W_1, \dots, W_q$ are sampled independently from the t-distribution with 5 degrees of freedom. Thus, Model 6 involves both a non-Gaussian design matrix with mixed covariate types and a non-Gaussian error vector.

We list in Table 3 the screening performance of all the methods as in Section 3.2. Note that the standard errors are omitted in this table to save space. Comparing

TABLE 3
Proportions of important main effects and active interaction variables retained by different screening methods

Method	X_1	X_2	X_3	X_4	X_5	X_9	X_{12}	X_{13}
SIS.max	1.00	1.00	1.00	1.00	1.00	0.65	0.44	0.24
SIS.sum	1.00	1.00	1.00	1.00	1.00	0.76	0.45	0.25
DCSIS	1.00	1.00	1.00	1.00	1.00	0.93	0.28	0.80
IPDC	1.00	1.00	1.00	1.00	1.00	0.92	0.67	0.85

these results to those in Table 2, we see that the problem of interaction screening becomes more difficult in this model. This result is reasonable since the scenario of Model 6 is more challenging than that of Model 5. Nevertheless, IPDC still improves over other methods in retaining active interaction variables X_9, X_{12} and X_{13}.

4. Real data analysis. We further evaluate the performance of our proposed procedure on a multivariate yeast cell-cycle data set from [17], which can be accessed in the R package "spls". This data set has been studied in [4] and [2]. Our goal is to predict how much mRNA is produced by 542 genes related to the yeast cell's replication process. For each gene, the binding levels of 106 transcription factors (TFs) are recorded. The binding levels of the TFs play a role in determining which genes are expressed and help detail the process behind eukaryotic cell-cycles. Messenger RNA is collected for two cell-cycles for a total of eighteen time points. Thus, this data set has sample size $n = 542$, number of covariates $p = 106$ and number of response $q = 18$, with all variables being continuous.

Considering the relatively large sample size, we use 30% of the data as training and the rest as testing, and repeat such random splitting for 100 times. We follow the same screening and selection procedures as in the simulation study for the setting of multiple responses in Section 3. Similarly, we take a union of the set of retained interaction variables and the set of retained main effects. For fair comparison, we keep $2[n/(\log n)] = 62$ variables in the screening procedures of SIS.max, SIS.sum and DCSIS, and use those variables to construct pairwise interactions for the selection step.

Table 4 presents the results on the prediction error and selected model size. Paired t-tests of prediction errors on the 100 splits of IPDC-GLasso against SIS.max-GLasso, SIS.sum-GLasso and DCSIS-GLasso result in p-values $2.86 \times$

TABLE 4
Means and standard errors (in parentheses) of prediction error as well as numbers of selected main effects and interactions for each method in yeast cell-cycle data

		Model size	
Method	**PE ($\times 10^{-3}$)**	**Main**	**Interaction**
SIS.max-GLasso	224.05 (1.20)	73.73 (7.96)	755.35 (61.38)
SIS.sum-GLasso	223.42 (1.17)	50.76 (3.97)	764.68 (63.52)
DCSIS-GLasso	223.93 (1.16)	63.67 (7.47)	705.11 (61.46)
IPDC-GLasso	220.44 (1.14)	113.78 (9.74)	801.70 (54.86)
SIS.max-GLasso-Lasso	226.66 (1.45)	47.56 (3.66)	327.32 (19.38)
SIS.sum-GLasso-Lasso	225.07 (1.48)	50.76 (3.97)	319.25 (19.95)
DCSIS-GLasso-Lasso	226.40 (1.43)	47.12 (3.92)	306.18 (20.75)
IPDC-GLasso-Lasso	222.43 (1.39)	56.08 (3.33)	300.93 (15.17)

10^{-11}, 1.70×10^{-13} and 1.15×10^{-14}, respectively. Moreover, paired t-tests of prediction errors on the 100 splits of IPDC-GLasso-Lasso against SIS.max-GLasso-Lasso, SIS.sum-GLasso-Lasso and DCSIS-GLasso-Lasso give p-values 2.92×10^{-4}, 2.30×10^{-2} and 9.73×10^{-5}, respectively. These results show significant improvement of our method over existing ones.

5. Discussions.

We have investigated the problem of interaction identification in the setting where the numbers of responses and covariates can both be large. Our suggested two-stage procedure IPDC provides a scalable approach with the idea of interaction screening and selection. It exploits the joint information among all the responses by using the distance correlation in the screening step and the regularized multi-response regression in the selection step. One key ingredient is the use of the square transformation to responses and covariates for effective interaction screening. The established sure screening and model selection properties enable its broad applicability beyond the heredity assumption.

Although we have focused our attention on the square transformation of the responses and covariates due to its simplicity and the motivation discussed in Section 2.1, it is possible that other functions can also work for the idea of IPDC. It would be interesting to investigate and characterize what class of functions is optimal for the purpose of interaction screening.

Like all independence screening methods using the marginal utilities including the SIS and DCSIS, our feature screening approach may fail to identify some important interactions or main effects that are marginally weakly related to the responses. One possible remedy is to exploit the idea of the iterative SIS proposed in [5] which has been shown to be capable of ameliorating the SIS. Recently, [20] also introduced an iterative DCSIS procedure and demonstrated that it can improve the finite-sample performance of the DCSIS. The theoretical properties of these iterative feature screening approaches are, however, less well understood. It would be interesting to develop an effective iterative IPDC procedure for further improving on the IPDC and investigate its sampling properties. For more flexible modeling, it is also of practical importance to extend the idea of IPDC to high-dimensional multi-response interaction models in the more general settings of the generalized linear models, nonparametric models and survival models, as well as other single-index models and multi-index models. These possible extensions are beyond the scope of the current paper and will be interesting topics for future research.

APPENDIX: PROOFS OF MAIN RESULTS

We provide the main steps of the proof of Theorem 1 and the proof of Theorem 2 in this Appendix. Some intermediate steps of the proof of Theorem 1 and additional technical details are included in the Supplementary Material. Hereafter, we denote by $\widetilde{C}_i$ with $i \geq 0$ some generic positive constants whose values may vary from line to line.

A.1. Proof of Theorem 1. The proof of Theorem 1 consists of two parts. The first part establishes the exponential probability bounds for $\widehat{\omega}_j - \omega_j$ and $\widehat{\omega}_k^* - \omega_k^*$, and the second part proves the sure screening property.

Part 1. We first prove inequalities (8) and (9), which give the exponential probability bounds for $\widehat{\omega}_j - \omega_j$ and $\widehat{\omega}_k^* - \omega_k^*$, respectively. Since the proofs of (8) and (9) are similar, here we focus on (9) to save space. Recall that

$$\omega_k^* = \frac{\mathrm{dcov}^2(X_k^*, \mathbf{y}^*)}{\sqrt{\mathrm{dcov}^2(X_k^*, X_k^*)}} \quad \text{and} \quad \widehat{\omega}_k^* = \frac{\widehat{\mathrm{dcov}}^2(X_k^*, \mathbf{y}^*)}{\sqrt{\widehat{\mathrm{dcov}}^2(X_k^*, X_k^*)}}.$$

The key idea of the proof is to show that for any positive constant $\widetilde{C}$, there exist some positive constants $\widetilde{C}_1, \dots, \widetilde{C}_4$ such that

$$\begin{aligned} &P\Bigl(\max_{1\le k\le p} \bigl|\widehat{\mathrm{dcov}}^2(X_k^*, \mathbf{y}^*) - \mathrm{dcov}^2(X_k^*, \mathbf{y}^*)\bigr| \ge \widetilde{C} n^{-\kappa_2}\Bigr) \\ &\quad \le p\widetilde{C}_1 \exp\bigl\{-\widetilde{C}_2 n^{(1-2\kappa_2)/5}\bigr\} + \widetilde{C}_3 \exp\bigl\{-\widetilde{C}_4 n^{(1-2\kappa_2)/10}\bigr\}, \end{aligned} \tag{13}$$

$$\begin{aligned} &P\Bigl(\max_{1\le k\le p} \bigl|\widehat{\mathrm{dcov}}^2(X_k^*, X_k^*) - \mathrm{dcov}^2(X_k^*, X_k^*)\bigr| \ge \widetilde{C} n^{-\kappa_2}\Bigr) \\ &\quad \le p\widetilde{C}_1 \exp\bigl\{-\widetilde{C}_2 n^{(1-2\kappa_2)/5}\bigr\} \end{aligned} \tag{14}$$

for all n sufficiently large. Once these two probability bounds are obtained, it follows from Conditions 1–2 and Lemmas 2–3 and 6 that

$$\begin{aligned} &P\Bigl(\max_{1\le k\le p} |\widehat{\omega}_k^* - \omega_k^*| \ge c_2 n^{-\kappa_2}\Bigr) \\ &\quad \le O\bigl(p \exp\bigl\{-C_1 n^{(1-2\kappa_2)/5}\bigr\} + \exp\bigl\{-C_2 n^{(1-2\kappa_2)/10}\bigr\}\bigr) \\ &\quad \le O\bigl(\exp\bigl\{-C n^{(1-2\kappa_2)/10}\bigr\}\bigr), \end{aligned} \tag{15}$$

where C_1, C_2 and C are some positive constants, and the last inequality follows from the condition that $\log p = o(n^{\eta_0})$ with $\eta_0 = \min\{(1-2\kappa_1)/3, (1-2\kappa_2)/5\}$.

It thus remains to prove (13) and (14). Again we concentrate on (13) since (14) can be shown using similar arguments. Define $\phi(X_{1k}^*, X_{2k}^*) = |X_{1k}^* - X_{2k}^*|$ and $\psi(\mathbf{y}_1^*, \mathbf{y}_2^*) = \|\mathbf{y}_1^* - \mathbf{y}_2^*\|$. According to [18], we have

$$\mathrm{dcov}^2(X_k^*, \mathbf{y}^*) = T_{k1} + T_{k2} - 2T_{k3} \quad \text{and} \quad \widehat{\mathrm{dcov}}^2(X_k^*, \mathbf{y}^*) = \widehat{T}_{k1} + \widehat{T}_{k2} - 2\widehat{T}_{k3},$$

where $T_{k1} = E[\phi(X_{1k}^*, X_{2k}^*)\psi(\mathbf{y}_1^*, \mathbf{y}_2^*)]$, $T_{k2} = E[\phi(X_{1k}^*, X_{2k}^*)]E[\psi(\mathbf{y}_1^*, \mathbf{y}_2^*)]$, $T_{k3} = E[\phi(X_{1k}^*, X_{2k}^*)\psi(\mathbf{y}_1^*, \mathbf{y}_3^*)]$, and

$$\widehat{T}_{k1} = n^{-2} \sum_{i,j=1}^n \phi(X_{ik}^*, X_{jk}^*)\psi(\mathbf{y}_i^*, \mathbf{y}_j^*),$$

$$\widehat{T}_{k2} = \Biggl[n^{-2} \sum_{i,j=1}^n \phi(X_{ik}^*, X_{jk}^*)\Biggr]\Biggl[n^{-2} \sum_{i,j=1}^n \psi(\mathbf{y}_i^*, \mathbf{y}_j^*)\Biggr],$$

$$\widehat{T}_{k3} = n^{-3} \sum_{i=1}^{n} \sum_{j,l=1}^{n} \phi(X_{ik}^*, X_{jk}^*) \psi(\mathbf{y}_i^*, \mathbf{y}_l^*).$$

It follows from the triangle inequality that

$$\begin{aligned} &\max_{1 \le k \le p} \left|\widehat{\mathrm{dcov}}^2(X_k^*, \mathbf{y}^*) - \mathrm{dcov}^2(X_k^*, \mathbf{y}^*)\right| \\ &\quad \le \max_{1 \le k \le p} |\widehat{T}_{k1} - T_{k1}| + \max_{1 \le k \le p} |\widehat{T}_{k2} - T_{k2}| + 2 \max_{1 \le k \le p} |\widehat{T}_{k3} - T_{k3}|. \end{aligned} \tag{16}$$

To establish the probability bound for the term $\max_{1 \le k \le p} |\widehat{\mathrm{dcov}}^2(X_k^*, \mathbf{y}^*) - \mathrm{dcov}^2(X_k^*, \mathbf{y}^*)|$, it is sufficient to bound each term on the right-hand side above. To enhance the readability, we proceed with three main steps.

Step 1. We start with the first term $\max_{1 \le k \le p} |\widehat{T}_{k1} - T_{k1}|$. An application of the Cauchy–Schwarz inequality gives

$$T_{k1} \le \{E[\phi^2(X_{1k}^*, X_{2k}^*)] E[\psi^2(\mathbf{y}_1^*, \mathbf{y}_2^*)]\}^{1/2}.$$

It follows from the triangle inequality that

$$\phi(X_{1k}^*, X_{2k}^*) = |X_{1k}^* - X_{2k}^*| \le |X_{1k}^*| + |X_{2k}^*| = X_{1k}^2 + X_{2k}^2 \tag{17}$$

and

$$\psi(\mathbf{y}_1^*, \mathbf{y}_2^*) = \|\mathbf{y}_1^* - \mathbf{y}_2^*\| \le \|\mathbf{y}_1^*\| + \|\mathbf{y}_2^*\| \le \|\widetilde{\mathbf{y}}_1\|^2 + \|\widetilde{\mathbf{y}}_2\|^2, \tag{18}$$

in view of $\mathbf{y}_1^* = \widetilde{\mathbf{y}}_1 \circ \widetilde{\mathbf{y}}_1$ and the fact that $\|\mathbf{a} \circ \mathbf{a}\| \le \|\mathbf{a}\|^2$ for any $\mathbf{a} \in \mathbb{R}^q$. By (17), we have $E[\phi^2(X_{1k}^*, X_{2k}^*)] \le E\{2(X_{1k}^4 + X_{2k}^4)\} = 4E(X_{1k}^4)$. Similarly, it holds that $E[\psi^2(\mathbf{y}_1^*, \mathbf{y}_2^*)] \le 4E(\|\widetilde{\mathbf{y}}_1\|^4)$. Combining these results leads to $0 \le T_{k1} \le 4\{E(X_{1k}^4)E(\|\widetilde{\mathbf{y}}_1\|^4)\}^{1/2}$. By Condition 2, $E(X_{1k}^4)$ and $E(\|\widetilde{\mathbf{y}}_1\|^4)$ are uniformly bounded by some positive constant for all $1 \le k \le p$. Thus, for any positive constant $\widetilde{C}$, $|T_{k1}/n| < \widetilde{C}n^{-\kappa_2}/8$ holds uniformly for all $1 \le k \le p$ when n is sufficiently large.

Let $\widehat{T}_{k1}^* = n(n-1)^{-1}\widehat{T}_{k1} = \{n(n-1)\}^{-1} \sum_{i \ne j} \phi(X_{ik}^*, X_{jk}^*) \psi(\mathbf{y}_i^*, \mathbf{y}_j^*)$. Then we have

$$|\widehat{T}_{k1} - T_{k1}| \le n^{-1}(n-1)\left|\widehat{T}_{k1}^* - T_{k1}\right| + |T_{k1}/n| \le \left|\widehat{T}_{k1}^* - T_{k1}\right| + \widetilde{C}n^{-\kappa_2}/8$$

for all $1 \le k \le p$, which entails

$$P\Big(\max_{1 \le k \le p} |\widehat{T}_{k1} - T_{k1}| \ge \widetilde{C}n^{-\kappa_2}/4\Big) \le P\Big(\max_{1 \le k \le p} \left|\widehat{T}_{k1}^* - T_{k1}\right| \ge \widetilde{C}n^{-\kappa_2}/8\Big) \tag{19}$$

for sufficiently large n. Thus, it is sufficient to bound $\widehat{T}_{k1}^* - T_{k1}$.

Since X^*_{ik} and $\mathbf{y}^*_i$ are generally unbounded, we apply the technique of truncation in the technical analysis. Define

$$\begin{aligned}
\widehat{T}^*_{k1,1} &= \{n(n-1)\}^{-1}\sum_{i\neq j}\phi(X^*_{ik}, X^*_{jk})\psi(\mathbf{y}^*_i, \mathbf{y}^*_j)\mathbb{I}\{\phi(X^*_{ik}, X^*_{jk}) \le M_1\}\\
&\qquad \times \mathbb{I}\{\psi(\mathbf{y}^*_i, \mathbf{y}^*_j) \le M_2\},\\
\widehat{T}^*_{k1,2} &= \{n(n-1)\}^{-1}\sum_{i\neq j}\phi(X^*_{ik}, X^*_{jk})\psi(\mathbf{y}^*_i, \mathbf{y}^*_j)\mathbb{I}\{\phi(X^*_{ik}, X^*_{jk}) \le M_1\}\\
&\qquad \times \mathbb{I}\{\psi(\mathbf{y}^*_i, \mathbf{y}^*_j) > M_2\},\\
\widehat{T}^*_{k1,3} &= \{n(n-1)\}^{-1}\sum_{i\neq j}\phi(X^*_{ik}, X^*_{jk})\psi(\mathbf{y}^*_i, \mathbf{y}^*_j)\mathbb{I}\{\phi(X^*_{ik}, X^*_{jk}) > M_1\},
\end{aligned}$$

where $\mathbb{I}\{\cdot\}$ denotes the indicator function and the thresholds $M_1, M_2 > 0$ will be specified later. Then we have $\widehat{T}^*_{k1} = \widehat{T}^*_{k1,1} + \widehat{T}^*_{k1,2} + \widehat{T}^*_{k1,3}$. Consequently, we can rewrite T_{k1} as $T_{k1} = T_{k1,1} + T_{k1,2} + T_{k1,3}$ with

$$\begin{aligned}
T_{k1,1} &= E\big[\phi(X^*_{1k}, X^*_{2k})\psi(\mathbf{y}^*_1, \mathbf{y}^*_2)\mathbb{I}\{\phi(X^*_{1k}, X^*_{2k}) \le M_1\}\mathbb{I}\{\psi(\mathbf{y}^*_1, \mathbf{y}^*_2) \le M_2\}\big],\\
T_{k1,2} &= E\big[\phi(X^*_{1k}, X^*_{2k})\psi(\mathbf{y}^*_1, \mathbf{y}^*_2)\mathbb{I}\{\phi(X^*_{1k}, X^*_{2k}) \le M_1\}\mathbb{I}\{\psi(\mathbf{y}^*_1, \mathbf{y}^*_2) > M_2\}\big],\\
T_{k1,3} &= E\big[\phi(X^*_{1k}, X^*_{2k})\psi(\mathbf{y}^*_1, \mathbf{y}^*_2)\mathbb{I}\{\phi(X^*_{1k}, X^*_{2k}) > M_1\}\big].
\end{aligned}$$

Clearly, $\widehat{T}^*_{k1,1}$, $\widehat{T}^*_{k1,2}$, and $\widehat{T}^*_{k1,3}$ are unbiased estimators of $T_{k1,1}$, $T_{k1,2}$ and $T_{k1,3}$, respectively. Therefore, it follows from Bonferroni's inequality that

$$\begin{aligned}
&P\Big(\max_{1\le k\le p}|\widehat{T}^*_{k1} - T_{k1}| \ge \widetilde{C}n^{-\kappa_2}/8\Big)\\
&\qquad \le \sum_{j=1}^{3} P\Big(\max_{1\le k\le p}|\widehat{T}^*_{k1,j} - T_{k1,j}| \ge \widetilde{C}n^{-\kappa_2}/24\Big).
\end{aligned}\tag{20}$$

In what follows, we will provide details on deriving an exponential tail probability bound for each term on the right-hand side above.

Step 1.1. We first consider $\widehat{T}^*_{k1,1} - T_{k1,1}$. For any $\delta > 0$, by Markov's inequality we have

$$P\big(\widehat{T}^*_{k1,1} - T_{k1,1} \ge \delta\big) \le \exp(-t\delta)\exp(-tT_{k1,1})E\big[\exp\big(t\widehat{T}^*_{k1,1}\big)\big]\tag{21}$$

for $t > 0$. Let $h(X^*_{1k}, \mathbf{y}^*_1; X^*_{1k}, \mathbf{y}^*_2) = \phi(X^*_{1k}, X^*_{2k})\psi(\mathbf{y}^*_1, \mathbf{y}^*_2)\mathbb{I}\{\phi(X^*_{1k}, X^*_{2k}) \le M_1\}\mathbb{I}\{\psi(\mathbf{y}^*_1, \mathbf{y}^*_2) \le M_2\}$ be the kernel of the U-statistic $\widehat{T}^*_{k1,1}$ and define

$$\begin{aligned}
&W(X^*_{1k}, \mathbf{y}^*_1; \ldots; X^*_{nk}, \mathbf{y}^*_n)\\
&\qquad = m^{-1}\big\{h(X^*_{1k}, \mathbf{y}^*_1; X^*_{1k}, \mathbf{y}^*_2) + h(X^*_{3k}, \mathbf{y}^*_3; X^*_{4k}, \mathbf{y}^*_4) + \cdots\\
&\qquad\quad + h(X^*_{2m-1,k}, \mathbf{y}^*_{2m-1}; X^*_{2m,k}, \mathbf{y}^*_{2m})\big\},
\end{aligned}\tag{22}$$

where $m = \lfloor n/2 \rfloor$ is the integer part of $n/2$. According to the theory of U-statistics [16], Section 5.1.6, any U-statistic can be expressed as an average of averages of i.i.d. random variables. This representation gives

$$\widehat{T}^*_{k1,1} = (n!)^{-1} \sum_{n!} W\bigl(X^*_{i_1 k}, \mathbf{y}^*_{i_1}; \ldots; X^*_{i_n k}, \mathbf{y}^*_{i_n}\bigr),$$

where $\sum_{n!}$ represents the summation over all possible permutations $(i_1, \ldots, i_n)$ of $(1, \ldots, n)$. An application of Jensen's inequality yields that for any $t > 0$,

$$\begin{aligned} E\bigl[\exp(t\widehat{T}^*_{k1,1})\bigr] &= E\biggl\{\exp\biggl[(n!)^{-1} \sum_{n!} tW\bigl(X^*_{i_1 k}, \mathbf{y}^*_{i_1}; \ldots; X^*_{i_n k}, \mathbf{y}^*_{i_n}\bigr)\biggr]\biggr\} \\ &\le E\biggl\{(n!)^{-1} \sum_{n!} \exp\bigl[tW\bigl(X^*_{i_1 k}, \mathbf{y}^*_{i_1}; \ldots; X^*_{i_n k}, \mathbf{y}^*_{i_n}\bigr)\bigr]\biggr\} \\ &= E\bigl\{\exp\bigl[tW\bigl(X^*_{1k}, \mathbf{y}^*_1; \ldots; X^*_{nk}, \mathbf{y}^*_n\bigr)\bigr]\bigr\} \\ &= E^m\bigl\{\exp\bigl[tm^{-1}h\bigl(X^*_{1k}, \mathbf{y}^*_1; X^*_{2k}, \mathbf{y}^*_2\bigr)\bigr]\bigr\}, \end{aligned}$$

where the last equality follows from (22). The above inequality together with (21) leads to

$$P\bigl(\widehat{T}^*_{k1,1} - T_{k1,1} \ge \delta\bigr) \le \exp(-t\delta) E^m\bigl\{e^{tm^{-1}[h(X^*_{1k}, \mathbf{y}^*_1; X^*_{2k}, \mathbf{y}^*_2) - T_{k1,1}]}\bigr\}.$$

Note that $E[h(X^*_{1k}, \mathbf{y}^*_1; X^*_{2k}, \mathbf{y}^*_2) - T_{k1,1}] = 0$ and

$$-T_{k1,1} \le h\bigl(X^*_{1k}, \mathbf{y}^*_1; X^*_{2k}, \mathbf{y}^*_2\bigr) - T_{k1,1} \le M_1 M_2 - T_{k1,1}.$$

Hence, it follows from Lemma 9 that

$$P\bigl(\widehat{T}^*_{k1,1} - T_{k1,1} \ge \delta\bigr) \le \exp\bigl[-t\delta + t^2 M_1^2 M_2^2/(8m)\bigr]$$

for any $t > 0$. Minimizing the right-hand side above with respect to t gives $P(\widehat{T}^*_{k1,1} - T_{k1,1} \ge \delta) \le \exp(-2m\delta^2/M_1^2 M_2^2)$ for any $\delta > 0$. Similarly, we can show that $P(\widehat{T}^*_{k1,1} - T_{k1,1} \le -\delta) \le \exp(-2m\delta^2/M_1^2 M_2^2)$. Therefore, it holds that

$$P\bigl(\bigl|\widehat{T}^*_{k1,1} - T_{k1,1}\bigr| \ge \delta\bigr) \le 2\exp\bigl(-2m\delta^2/M_1^2 M_2^2\bigr).$$

Recall that $m = \lfloor n/2 \rfloor$. If we set $M_1 = n^{\xi_1}$ and $M_2 = n^{\xi_2}$ with some positive constants ξ_1 and ξ_2, then for $\delta = \widetilde{C} n^{-\kappa_2}/24$ with any positive constant $\widetilde{C}$, there exists some positive constant $\widetilde{C}_1$ such that when n is sufficiently large,

$$P\bigl(\bigl|\widehat{T}^*_{k1,1} - T_{k1,1}\bigr| \ge \widetilde{C} n^{-\kappa_2}/24\bigr) \le 2\exp\bigl(-\widetilde{C}^2 \widetilde{C}_1 n^{1-2\kappa_2-2\xi_1-2\xi_2}\bigr)$$

for all $1 \le k \le p$. This along with Bonferroni's inequality entails

$$\begin{aligned} &P\Bigl(\max_{1\le k\le p} \bigl|\widehat{T}^*_{k1,1} - T_{k1,1}\bigr| \ge \widetilde{C} n^{-\kappa_2}/24\Bigr) \\ &\qquad \le 2p\exp\bigl(-\widetilde{C}^2 \widetilde{C}_1 n^{1-2\kappa_2-2\xi_1-2\xi_2}\bigr). \end{aligned} \tag{23}$$

Step 1.2. We next deal with $\widehat{T}^*_{k1,2} - T_{k1,2}$. Note that

$$0 \le T_{k1,2} \le M_1 E[\psi(\mathbf{y}_1^*, \mathbf{y}_2^*)\mathbb{I}\{\psi(\mathbf{y}_1^*, \mathbf{y}_2^*) > M_2\}]$$

for all $1 \le k \le p$. It follows from the Cauchy–Schwarz inequality that

$$\begin{aligned} &E[\psi(\mathbf{y}_1^*, \mathbf{y}_2^*)\mathbb{I}\{\psi(\mathbf{y}_1^*, \mathbf{y}_2^*) > M_2\}] \\ &\quad \le [E[\psi^2(\mathbf{y}_1^*, \mathbf{y}_2^*)]P\{\psi(\mathbf{y}_1^*, \mathbf{y}_2^*) > M_2\}]^{1/2}. \end{aligned} \tag{24}$$

In view of (18), we see that

$$\begin{aligned} E[\psi^2(\mathbf{y}_1^*, \mathbf{y}_2^*)] &\le E[(\|\widetilde{\mathbf{y}}_1\|^2 + \|\widetilde{\mathbf{y}}_2\|^2)^2] \\ &\le E[2(\|\widetilde{\mathbf{y}}_1\|^4 + \|\widetilde{\mathbf{y}}_2\|^4)] = 4E(\|\widetilde{\mathbf{y}}_1\|^4) \end{aligned} \tag{25}$$

and the probability term in (24) is bounded from above by

$$\begin{aligned} P(\|\widetilde{\mathbf{y}}_1\|^2 + \|\widetilde{\mathbf{y}}_2\|^2 > M_2) &\le P(\|\widetilde{\mathbf{y}}_1\|^2 > M_2/2) + P(\|\widetilde{\mathbf{y}}_2\|^2 > M_2/2) \\ &= 2P(\|\widetilde{\mathbf{y}}_1\| > \sqrt{M_2/2}) \\ &\le 2\exp(-c_0\sqrt{M_2/2})E[\exp(c_0\|\widetilde{\mathbf{y}}_1\|)], \end{aligned} \tag{26}$$

where c_0 is a positive constant given in Condition 2 and the last inequality follows from Markov's inequality. Combining inequalities (24)–(26) and by Condition 2, we obtain

$$E[\psi(\mathbf{y}_1^*, \mathbf{y}_2^*)\mathbb{I}\{\psi(\mathbf{y}_1^*, \mathbf{y}_2^*) > M_2\}] \le \widetilde{C}_2 \exp(-2^{-1}c_0\sqrt{M_2/2}) \tag{27}$$

and thus $0 \le T_{k1,2} \le \widetilde{C}_2 M_1 \exp(-2^{-1}c_0\sqrt{M_2/2})$, where $\widetilde{C}_2$ is some positive constant. Recall that $M_1 = n^{\xi_1}$ and $M_2 = n^{\xi_2}$. Then for any positive constant $\widetilde{C}$, it holds that

$$0 \le T_{k1,2} \le \widetilde{C}_2 n^{\xi_1} \exp(-2^{-3/2}c_0 n^{\xi_2/2}) \le \widetilde{C}n^{-\kappa_2}/48$$

for all $1 \le k \le p$ when n is sufficiently large. This inequality gives

$$P\Big(\max_{1\le k\le p}|\widehat{T}^*_{k1,2} - T_{k1,2}| \ge \widetilde{C}n^{-\kappa_2}/24\Big) \le P\Big(\max_{1\le k\le p}|\widehat{T}^*_{k1,2}| \ge \widetilde{C}n^{-\kappa_2}/48\Big) \tag{28}$$

for all n sufficiently large.

Note that for all $1 \le k \le p$, $|\widehat{T}^*_{k1,2}|$ is uniformly bounded from above by $M_1[n(n-1)]^{-1}\sum_{i\ne j}\psi(\mathbf{y}_i^*, \mathbf{y}_j^*)\mathbb{I}\{\psi(\mathbf{y}_i^*, \mathbf{y}_j^*) > M_2\}$. Thus, in view of (27), applying Markov's inequality yields that for any $\delta > 0$,

$$\begin{aligned} &P\Big(\max_{1\le k\le p}|\widehat{T}^*_{k1,2}| \ge \delta/2\Big) \\ &\quad \le P\Big\{M_1[n(n-1)]^{-1}\sum_{i\ne j}\psi(\mathbf{y}_i^*, \mathbf{y}_j^*)\mathbb{I}\{\psi(\mathbf{y}_i^*, \mathbf{y}_j^*) > M_2\} \ge \delta/2\Big\} \end{aligned}$$

$$
\begin{aligned}
&\le (\delta/2)^{-1} E\Big\{ M_1 [n(n-1)]^{-1} \sum_{i \ne j} \psi(\mathbf{y}_i^*, \mathbf{y}_j^*) \mathbb{I}\{\psi(\mathbf{y}_i^*, \mathbf{y}_j^*) > M_2\} \Big\} \\
&= (\delta/2)^{-1} M_1 E\big[\psi(\mathbf{y}_1^*, \mathbf{y}_2^*) \mathbb{I}\{\psi(\mathbf{y}_1^*, \mathbf{y}_2^*) > M_2\}\big] \\
&\le (\delta/2)^{-1} M_1 \widetilde{C}_2 \exp\big(-2^{-1} c_0 \sqrt{M_2/2}\big).
\end{aligned}
$$

Since $M_1 = n^{\xi_1}$ and $M_2 = n^{\xi_2}$, setting $\delta = \widetilde{C} n^{-\kappa_2}/24$ in the above inequality entails

$$
P\Big(\max_{1 \le k \le p} |\widehat{T}_{k1,2}^*| \ge \widetilde{C} n^{-\kappa_2}/48\Big) \le 48 \widetilde{C}^{-1} \widetilde{C}_2 n^{\kappa_2 + \xi_1} \exp\big(-2^{-3/2} c_0 n^{\xi_2/2}\big).
$$

Combining this inequality with (28) gives

$$
\begin{aligned}
&P\Big(\max_{1 \le k \le p} |\widehat{T}_{k1,2}^* - T_{k1,2}| \ge \widetilde{C} n^{-\kappa_2}/24\Big) \\
&\qquad \le 48 \widetilde{C}^{-1} \widetilde{C}_2 n^{\kappa_2 + \xi_1} \exp\big(-2^{-3/2} c_0 n^{\xi_2/2}\big).
\end{aligned} \tag{29}
$$

Step 1.3. We finally handle the term $\widehat{T}_{k1,3}^* - T_{k1,3}$ and show that it satisfies

$$
\begin{aligned}
&P\Big(\max_{1 \le k \le p} |\widehat{T}_{k1,3}^* - T_{k1,3}| \ge \widetilde{C} n^{-\kappa_2}/24\Big) \\
&\qquad \le 48 p \widetilde{C}^{-1} \widetilde{C}_3 n^{\kappa_2} \exp\big(-8^{-1} c_0 n^{\xi_1}\big)
\end{aligned} \tag{30}
$$

with $\widetilde{C}_3$ some positive constant in Section D.1 of the Supplementary Material.

Combining the results in (20), (23) and (29)–(30) leads to

$$
\begin{aligned}
&P\Big(\max_{1 \le k \le p} |\widehat{T}_{k1}^* - T_{k1}| \ge \widetilde{C} n^{-\kappa_2}/8\Big) \\
&\quad \le 2p \exp\big(-\widetilde{C}^2 \widetilde{C}_1 n^{1 - 2\kappa_2 - 2\xi_1 - 2\xi_2}\big) + 48 p \widetilde{C}^{-1} \widetilde{C}_3 n^{\kappa_2} \exp\big(-8^{-1} c_0 n^{\xi_1}\big) \\
&\qquad + 48 \widetilde{C}^{-1} \widetilde{C}_2 n^{\kappa_2 + \xi_1} \exp\big(-2^{-3/2} c_0 n^{\xi_2/2}\big).
\end{aligned}
$$

Let $\xi_1 = (1 - 2\kappa_2)/3 - 2\eta$ and $\xi_2 = 3\eta$ with some $0 < \eta < (1 - 2\kappa_2)/6$. Then we have

$$
\begin{aligned}
P\Big(\max_{1 \le k \le p} |\widehat{T}_{k1}^* - T_{k1}| \ge \widetilde{C} n^{-\kappa_2}/8\Big) &\le p \widetilde{C}_1 \exp\big\{-\widetilde{C}_2 n^{(1-2\kappa_2)/3 - 2\eta}\big\} \\
&\quad + \widetilde{C}_3 \exp\big\{-\widetilde{C}_4 n^{3\eta/2}\big\},
\end{aligned} \tag{31}
$$

where $\widetilde{C}_1, \dots, \widetilde{C}_4$ are some positive constants. This inequality along with (19) yields

$$
\begin{aligned}
P\Big(\max_{1 \le k \le p} |\widehat{T}_{k1} - T_{k1}| \ge \widetilde{C} n^{-\kappa_2}/4\Big) &\le p \widetilde{C}_1 \exp\big\{-\widetilde{C}_2 n^{(1-2\kappa_2)/3 - 2\eta}\big\} \\
&\quad + \widetilde{C}_3 \exp\big\{-\widetilde{C}_4 n^{3\eta/2}\big\}.
\end{aligned} \tag{32}
$$

Step 2. For the second term $\max_{1\le k\le p}|\widehat{T}_{k2}-T_{k2}|$, we show in Section D.2 of the Supplementary Material that

$$P\Big(\max_{1\le k\le p}|\widehat{T}_{k2}-T_{k2}|\ge \widetilde{C}n^{-\kappa_2}/4\Big)\le \sum_{k=1}^{p}P\big(|\widehat{T}_{k2}-T_{k2}|\ge \widetilde{C}n^{-\kappa_2}/4\big) \tag{33}$$
$$\le p\widetilde{C}_5\exp\{-\widetilde{C}_6 n^{(1-2\kappa_2)/5}\}$$

holds, where $\widetilde{C}_5$ and $\widetilde{C}_6$ are some positive constants.

Step 3. We further prove that the third term $\widehat{T}_{k3}-T_{k3}$ satisfies

$$P\Big(\max_{1\le k\le p}|\widehat{T}_{k3}-T_{k3}|\ge \widetilde{C}n^{-\kappa_2}/4\Big)\le p\widetilde{C}_1\exp\{-\widetilde{C}_2 n^{(1-2\kappa_2)/3-2\eta}\} \tag{34}$$
$$+\widetilde{C}_3\exp\{-\widetilde{C}_4 n^{3\eta/2}\}$$

with $\widetilde{C}_1,\dots,\widetilde{C}_4$ some positive constants in Section D.3 of the Supplementary Material.

Combining inequalities (16) and (32)–(34) and setting $\eta=(1-2\kappa_2)/15$ entail

$$P\Big\{\max_{1\le k\le p}|\widehat{\mathrm{dcov}}^2(X_k^*,\mathbf{y}^*)-\mathrm{dcov}^2(X_k^*,\mathbf{y}^*)|\ge \widetilde{C}n^{-\kappa_2}\Big\} \tag{35}$$
$$\le p\widetilde{C}_1\exp\{-\widetilde{C}_2 n^{(1-2\kappa_2)/5}\}+\widetilde{C}_3\exp\{-\widetilde{C}_4 n^{(1-2\kappa_2)/10}\}$$

with $\widetilde{C}_1,\dots,\widetilde{C}_4$ some positive constants, which completes the proof for the first part of Theorem 1.

Part 2. We now proceed to prove the second part of Theorem 1. The main idea is to build the probability bounds for two events $\{\mathcal{M}\subset\widehat{\mathcal{M}}\}$ and $\{\mathcal{I}\subset\widehat{\mathcal{I}}\}$. We first bound $P(\mathcal{M}\subset\widehat{\mathcal{M}})$. Define an event $\Omega_1=\{\max_{j\in\mathcal{M}}|\widehat{\omega}_j-\omega_j|<c_1n^{-\kappa_1}\}$. Then by Condition 3, conditional on the event Ω_1 we have $\widehat{\omega}_j\ge 2c_1n^{-\kappa_1}$ for all $j\in\mathcal{M}$, which gives

$$P(\mathcal{M}\subset\widehat{\mathcal{M}})\ge P(\Omega_1)=1-P\big(\Omega_1^c\big) \tag{36}$$
$$=1-P\Big(\max_{j\in\mathcal{M}}|\widehat{\omega}_j-\omega_j|\ge c_1n^{-\kappa_1}\Big).$$

Following similar arguments as for proving (15), it can be shown that there exist some positive constants $\widetilde{C}_5$ and $\widetilde{C}_6$ such that

$$P\Big(\max_{j\in\mathcal{M}}|\widehat{\omega}_j-\omega_j|\ge c_1n^{-\kappa_1}\Big)=O\big(s_1\exp\{-\widetilde{C}_5n^{(1-2\kappa_1)/3}\}$$
$$+\exp\{-\widetilde{C}_6n^{(1-2\kappa_1)/6}\}\big),$$

where s_1 is the cardinality of $\mathcal{M}$. This inequality together with (36) yields

$$P(\mathcal{M}\subset\widehat{\mathcal{M}})\ge 1-O\big(s_1\exp\{-\widetilde{C}_5n^{(1-2\kappa_1)/3}\}+\exp\{-\widetilde{C}_6n^{(1-2\kappa_1)/6}\}\big). \tag{37}$$

We next bound $P(\mathcal{I} \subset \widehat{\mathcal{I}})$. Note that $P(\mathcal{I} \subset \widehat{\mathcal{I}}) \geq P(\mathcal{A} \subset \widehat{\mathcal{A}})$ since conditional on the event $\{\mathcal{A} \subset \widehat{\mathcal{A}}\}$ it holds that $\{\mathcal{I} \subset \widehat{\mathcal{I}}\}$. Define an event $\Omega_2 = \{\max_{k\in\mathcal{A}} |\widehat{\omega}_k^* - \omega_k^*| < c_2 n^{-\kappa_2}\}$. Then by Condition 3, we have $\widehat{\omega}_k \geq 2c_2 n^{-\kappa_2}$ for all $k \in \mathcal{A}$ conditional on the event Ω_2, which leads to $P(\mathcal{A} \subset \widehat{\mathcal{A}}) \geq P(\Omega_2)$. Combining these results yields

$$P(\mathcal{I} \subset \widehat{\mathcal{I}}) \geq P(\Omega_2) = 1 - P(\Omega_2^c) = 1 - P\Bigl(\max_{k\in\mathcal{A}} |\widehat{\omega}_k^* - \omega_k^*| \geq c_2 n^{-\kappa_2}\Bigr). \tag{38}$$

Using similar arguments as for proving (15) shows that there exist some positive constants $\widetilde{C}_7$ and $\widetilde{C}_8$ such that

$$\begin{aligned} &P\Bigl(\max_{k\in\mathcal{A}} |\widehat{\omega}_k^* - \omega_k^*| \geq c_2 n^{-\kappa_2}\Bigr) \\ &\quad = O\bigl(s_2 \exp\{-\widetilde{C}_7 n^{(1-2\kappa_2)/5}\} + \exp\{-\widetilde{C}_8 n^{(1-2\kappa_2)/10}\}\bigr), \end{aligned}$$

where s_2 is the cardinality of $\mathcal{A}$. This together with (38) entails

$$P(\mathcal{I} \subset \widehat{\mathcal{I}}) \geq 1 - O\bigl(s_2 \exp\{-\widetilde{C}_7 n^{(1-2\kappa_2)/5}\} + \exp\{-\widetilde{C}_8 n^{(1-2\kappa_2)/10}\}\bigr). \tag{39}$$

Finally, combining (37) and (39), we obtain

$$\begin{aligned} P(\mathcal{M} \subset \widehat{\mathcal{M}} \text{ and } \mathcal{I} \subset \widehat{\mathcal{I}}) &\geq P(\mathcal{M} \subset \widehat{\mathcal{M}}) + P(\mathcal{I} \subset \widehat{\mathcal{I}}) - 1 \\ &\geq 1 - O\bigl(s_1 \exp\{-\widetilde{C}_5 n^{(1-2\kappa_1)/3}\} + \exp\{-\widetilde{C}_6 n^{(1-2\kappa_1)/6}\}\bigr) \\ &\quad - O\bigl(s_2 \exp\{-\widetilde{C}_7 n^{(1-2\kappa_2)/5}\} + \exp\{-\widetilde{C}_8 n^{(1-2\kappa_2)/10}\}\bigr) \\ &\geq 1 - O\bigl(\exp\{-C n^{\eta_0/2}\}\bigr), \end{aligned}$$

where C is some positive constant, and the last inequality follows from the facts $s_1, s_2 \leq p$ and the condition that $\log p = o(n^{\eta_0})$ with $\eta_0 = \min\{(1-2\kappa_1)/3, (1-2\kappa_2)/5\}$. This concludes the proof for the second part of Theorem 1.

A.2. Proof of Theorem 2. Define an event $\Omega_3 = \{\max_{1\leq j\leq p} |\widehat{\omega}_j - \omega_j| < c_1 n^{-\kappa_1}\}$. For any $j \in \mathcal{M}^c$, if $\omega_j < c_1 n^{-\kappa_1}$ and $|\widehat{\omega}_j - \omega_j| < c_1 n^{-\kappa_1}$, then we have $\widehat{\omega}_j < 2c_1 n^{-\kappa_1}$. Thus, conditional on the event Ω_3, the cardinality of $\{j : \widehat{\omega}_j \geq 2c_1 n^{-\kappa_1} \text{ and } j \in \mathcal{M}^c\}$ cannot exceed that of $\{j : \omega_j \geq c_1 n^{-\kappa_1} \text{ and } j \in \mathcal{M}^c\}$. This entails that the cardinality of $\{j : \widehat{\omega}_j \geq 2c_1 n^{-\kappa_1}\}$ is no larger than that of $\{j : \omega_j \geq c_1 n^{-\kappa_1} \text{ and } j \in \mathcal{M}^c\} \cup \mathcal{M}$, which is in turn bounded from above by $|\mathcal{M}| + s_3$. Thus, it follows from (8) in Theorem 1 that

$$\begin{aligned} P\bigl(|\widehat{\mathcal{M}}| \leq |\mathcal{M}| + s_3\bigr) \geq P(\Omega_3) &= 1 - P\Bigl(\max_{1\leq j\leq p} |\widehat{\omega}_j - \omega_j| \geq c_1 n^{-\kappa_1}\Bigr) \\ &\geq 1 - O\bigl(\exp\{-C n^{(1-2\kappa_1)/6}\}\bigr). \end{aligned}$$

Similarly, we can show that

$$P\bigl\{|\widehat{\mathcal{I}}| \leq (|\mathcal{A}| + s_4)(|\mathcal{A}| + s_4 - 1)/2\bigr\} \geq 1 - O\bigl(\exp\{-C n^{(1-2\kappa_2)/10}\}\bigr).$$

Combining these two probability bounds yields

$$
\begin{aligned}
P\{|\widehat{\mathcal{M}}| &\le |\mathcal{M}| + s_3 \text{ and } |\widehat{\mathcal{I}}| \le (|\mathcal{A}| + s_4)(|\mathcal{A}| + s_4 - 1)/2\} \\
&\ge 1 - O(\exp\{-Cn^{(1-2\kappa_1)/6}\}) - O(\exp\{-Cn^{(1-2\kappa_2)/10}\}) \\
&\ge 1 - O(\exp\{-Cn^{\eta_0/2}\}),
\end{aligned}
$$

where C is some positive constant, and the last inequality follows from the condition that $\log p = o(n^{\eta_0})$ with $\eta_0 = \min\{(1-2\kappa_1)/3, (1-2\kappa_2)/5\}$. This completes the proof of Theorem 2.

Acknowledgments. Yinfei Kong and Daoji Li contributed equally to this work. The authors would like to thank the Co-Editor, Associate Editor and referees for their valuable comments that have helped improve the paper significantly. Part of this work was completed while the last two authors visited the Departments of Statistics at University of California, Berkeley and Stanford University. They sincerely thank both departments for their hospitality.

SUPPLEMENTARY MATERIAL

Supplementary material to "Interaction pursuit in high-dimensional multi-response regression via distance correlation" (DOI: 10.1214/16-AOS1474SUPP; .pdf). Due to space constraints, the details about the post-screening interaction selection, additional numerical studies, some intermediate steps of the proof of Theorem 1 and additional technical details are provided in the Supplementary Material [12].

REFERENCES

[1] BIEN, J., TAYLOR, J. and TIBSHIRANI, R. (2013). A LASSO for hierarchical interactions. *Ann. Statist.* **41** 1111–1141. MR3113805

[2] CHEN, L. and HUANG, J. Z. (2012). Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. *J. Amer. Statist. Assoc.* **107** 1533–1545. MR3036414

[3] CHOI, N. H., LI, W. and ZHU, J. (2010). Variable selection with the strong heredity constraint and its oracle property. *J. Amer. Statist. Assoc.* **105** 354–364. MR2656056

[4] CHUN, H. and KELEŞ, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 3–25. MR2751241

[5] FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **70** 849–911. MR2530322

[6] FAN, Y., KONG, Y., LI, D. and LV, J. (2016). Interaction pursuit with feature screening and selection. Preprint. Available at arXiv:1605.08933.

[7] FAN, Y., KONG, Y., LI, D. and ZHENG, Z. (2015). Innovated interaction screening for high-dimensional nonlinear classification. *Ann. Statist.* **43** 1243–1272. MR3346702

[8] HALL, P. and XUE, J.-H. (2014). On selecting interacting features from high-dimensional data. *Comput. Statist. Data Anal.* **71** 694–708. MR3132000

[9] HAO, N. and ZHANG, H. H. (2014). Interaction screening for ultrahigh-dimensional data. *J. Amer. Statist. Assoc.* **109** 1285–1301. MR3265697

[10] HUO, X. and SZÉKELY, G. J. (2016). Fast computing for distance covariance. *Technometrics*. To appear.

[11] JIANG, B. and LIU, J. S. (2014). Variable selection for general index models via sliced inverse regression. *Ann. Statist.* **42** 1751–1786. MR3262467

[12] KONG, Y., LI, D., FAN, Y. and LV, J. (2016). Supplement to "Interaction pursuit in high-dimensional multi-response regression via distance correlation." DOI:10.1214/16-AOS1474SUPP.

[13] LI, J., ZHONG, W., LI, R. and WU, R. (2014). A fast algorithm for detecting gene-gene interactions in genome-wide association studies. *Ann. Appl. Stat.* **8** 2292–2318. MR3292498

[14] LI, R., ZHONG, W. and ZHU, L. (2012). Feature screening via distance correlation learning. *J. Amer. Statist. Assoc.* **107** 1129–1139. MR3010900

[15] LV, J. (2013). Impacts of high dimensionality in finite samples. *Ann. Statist.* **41** 2236–2262. MR3127865

[16] SERFLING, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York. MR0595165

[17] SPELLMAN, P. T., SHERLOCK, G., ZHANG, M. Q., IYER, V. R., ANDERS, K., EISEN, M. B., BROWN, P. O., BOTSTEIN, D. and FUTCHER, B. (1998). Combined expression trait correlations and expression quantitative trait locus mapping. *Mol. Biol. Cell* **9** 3273–3297.

[18] SZÉKELY, G. J., RIZZO, M. L. and BAKIROV, N. K. (2007). Measuring and testing dependence by correlation of distances. *Ann. Statist.* **35** 2769–2794. MR2382665

[19] YUAN, M., JOSEPH, V. R. and ZOU, H. (2009). Structured variable selection and estimation. *Ann. Appl. Stat.* **3** 1738–1757. MR2752156

[20] ZHONG, W. and ZHU, L. (2015). An iterative approach to distance correlation-based sure independence screening. *J. Stat. Comput. Simul.* **85** 2331–2345. MR3339301

Y. KONG
DEPARTMENT OF INFORMATION SYSTEMS
AND DECISION SCIENCES
MIHAYLO COLLEGE OF BUSINESS AND ECONOMICS
CALIFORNIA STATE UNIVERSITY, FULLERTON
FULLERTON, CALIFORNIA 92831
USA
E-MAIL: yikong@fullerton.edu

D. LI
DEPARTMENT OF STATISTICS
UNIVERSITY OF CENTRAL FLORIDA
ORLANDO, FLORIDA 32816-2370
USA
E-MAIL: daoji.li@ucf.edu

Y. FAN
J. LV
DATA SCIENCES AND OPERATIONS DEPARTMENT
MARSHALL SCHOOL OF BUSINESS
UNIVERSITY OF SOUTHERN CALIFORNIA
LOS ANGELES, CALIFORNIA 90089
USA
E-MAIL: fanyingy@marshall.usc.edu
jinchilv@marshall.usc.edu