Statistics and Its Interface Volume 11 (2018) 317–325

# Interaction screening by partial correlation

Yue Selena Niu, Ning Hao, and Hao Helen Zhang*

Interaction effects between predictors can play an important role in improving prediction and model interpretation for regression models. However, it is both statistically and computationally challenging to discover informative interactions for high dimensional data. Variable screening based on marginal information is popular for identifying important predictors, but it is mainly used for main-effect-only models. In this paper, we study interaction screening for high dimensional quadratic regression models. First, we show that the direct generalization of main-effect screening to interaction screening can be incorrect or inefficient, as it overlooks the intrinsic relationship between main effects and interactions. Next, we propose a main-effect-adjusted interaction screening procedure to select interactions while taking into account main effects. This new unified framework can be employed with multiple types of correlation measures, such as Pearson correlation coefficients, nonparametric rank-based measures including Spearman's and Kendall's correlation coefficients. Efficient algorithms are developed for each correlation measure to make the screening procedure scalable to high dimensional data. Finally, we illustrate performance of the new screening procedure by simulation studies and an application to a retinopathy study.

AMS 2000 subject classifications: Primary 62H20, 62F07; secondary 62J05.
Keywords and phrases: High dimensional data, Interaction effects, Marginal statistic, Quadratic regression, Rank correlation, Variable screening.

## 1. INTRODUCTION

Interaction terms naturally appear in classical models for experimental design and polynomial regression. In practice, models containing interaction effects are more flexible and powerful than main-effects-only models in capturing complex data structures, as they can improve both prediction accuracy and model interpretability. Recently, detecting interaction effects for high dimensional data has received much attention in the literature, partially due to its important applications in genetics; see [5, 17] for overviews. Interaction selection is challenging for high dimensional data. To facilitate implementation, computationally less intensive procedures are generally preferable in practice. For example, two-stage approaches are popular choices ([20, 19]), mainly due to their fast computation and effective dimension reduction. However, these procedures rely on hierarchical model assumptions [8], which might be violated or are sometimes hard to justify in real applications. Moreover, [1] pointed out that it is often difficult to determine the thresholding rule or the model size at the first stage, which is crucial to the success of two-stage procedures. Another popular strategy is to fit a joint model containing both main and interaction effects subject to penalty constraints; see [15, 21, 4, 1], among others. However, these methods are typically computational expensive or infeasible to analyze high dimensional data.

In this paper, we consider the problem of interaction screening via marginal statistics. When the number of features is large, one common strategy to screen out noise features is to rank features based on their marginal statistics, such as the marginal Pearson correlation coefficient between each feature and the response variable. Marginal approaches to main effects screening have gained much attention since the seminal paper [7]. Other relevant works include but are not limited to [22, 12, 13]. However, the problem of interaction screening for high dimensional data has been much less studied. This work aims to fill the gap. One simple and straightforward idea for interaction screening is to treat main effects and interactions equally as separate features and rank all the features based on their marginal statistics. However, this naive method can be problematic in practice, since it ignores the intrinsic relationship between main effects and interactions. In fact, we find out that it is usually helpful to take into account parental main effects when evaluating the importance of interaction terms to the response. This in turn suggests a new "marginal statistic" for interaction effects. Motivated by this, we propose a main-effect-adjusted screening approach, called Interaction Screening by Partial Correlation (ISPC), for ranking and screening interaction effects.

The proposed ISPC provides a general framework to enhance any standard correlation coefficient and make it suitable for assessing interaction effects. In the paper, we develop the ISPC for three commonly used correlation measures, including Pearson correlation coefficient, Spearman's, and Kendall's rank correlation coefficients. The advantage of the proposed interaction screening method is twofold. First, it is computationally scalable for big data sets with many features. Although we need go over all the pairs, the marginal statistic is easy and fast to calculate by using the proposed algorithms. Its implementation never requires stor-

*Corresponding Author.

ing the whole design matrix of interaction effects. Therefore, the procedure contributes a convenient and effective tool for high dimensional interaction screening. Second, by directly screening interactions, the ISPC procedure does not require parental main effects to be strong in order to detect important interactions. Compared to two-stage methods reviewed in [8], the ISPC approach does not rely on the hierarchical model assumption and is more flexible. In particular, this feature makes it superior to two-stage methods when the signal carried by main effects is weak.

The rest of this paper is organized as follows. In Section 2, we first consider the naive approach to interaction screening and discuss its drawbacks. Then we propose a new main-effect-adjusted interaction screening framework based on a variety of correlation measures. In Sections 3 and 4, we investigate the proposed screening procedures using numerical studies. Section 5 contains final remarks. Technical details are presented in the Appendix.

## 2. METHODS

### 2.1 Notations

Given data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$, which are independent and identically distributed (IID) copies of the pair $(X, Y)$, where $X = (X_1, ..., X_p)^\top$ is a $p$-dimensional predictor vector and $Y$ is the response, we consider a linear model with two-way interaction terms, or quadratic model, by assuming

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \gamma_{11} X_1^2 + \\ &\quad \gamma_{12} X_1 X_2 + \cdots + \gamma_{pp} X_p^2 + \varepsilon. \end{aligned} \tag{1}$$

In model (1), $\beta_0$, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^\top$, $\boldsymbol{\gamma} = (\gamma_{11}, \gamma_{12}, ..., \gamma_{pp})^\top$ are unknown parameters. The predictors $\{X_j\}_{j=1}^p$, $\{X_j^2\}_{j=1}^p$, and $\{X_j X_k\}_{1 \le j < k \le p}$ are main effects, quadratic effects, and two-way interaction effects, respectively. For convenience, we call $X_j$ and $X_k$ the parents of $X_j X_k$. Let $\mathbf{X} = (\mathbf{x}_1, \cdots, \mathbf{x}_n)^\top$ and $\mathbf{y} = (y_1, \cdots, y_n)^\top$ be the $n \times p$ design matrix of main effects and the response vector, respectively. Here we assume $\mathbf{y}$ is centered and $\mathbf{X}$ is standardized to mean zero and variance one column-wisely. For any subset $\mathcal{A} \subset \{1, \cdots, p\}$, $\mathbf{X}_\mathcal{A}$ is the submatrix of $\mathbf{X}$ with columns indexed by $\mathcal{A}$. In particular, $\mathbf{X}_j$ is the $j$th column vector of $\mathbf{X}$. Moreover, define $\mathbf{X}^{\circ 2} = \mathbf{X} \circ \mathbf{X}$ as $n \times \frac{p(p+1)}{2}$ matrix consisting of all pairwise products of column vectors of $\mathbf{X}$. That is, $\mathbf{X}^{\circ 2} = (\mathbf{X}_1 \circ \mathbf{X}_1, \mathbf{X}_1 \circ \mathbf{X}_2, ..., \mathbf{X}_p \circ \mathbf{X}_p)$, where, for column vectors, $\circ$ means entry-wise product. Denote by $\mathbf{Z}$ the matrix obtained by standardizing $\mathbf{X}^{\circ 2}$ column-wisely. We use $\lfloor a \rfloor$ to denote the largest integer no greater than $a$.

### 2.2 Naive approach to interaction screening

In literature, a variety of screening techniques have been recently developed, and the following is a brief review. To start with, we first consider the Pearson correlation used in sure independence screening (SIS) of [7]. Recall that $\mathbf{y}$ is centered and $\mathbf{X}_j$'s are standardized by our convention. So the marginal sample Pearson correlation $\widehat{\mathrm{Corr}}(Y, X_j)$ is proportional to $\omega_j = \mathbf{X}_j^\top \mathbf{y}$. Denote $\boldsymbol{\omega} = \mathbf{X}^\top \mathbf{y}$. The SIS procedure screens variables by ranking and thresholding $\boldsymbol{\omega}$. That is, a submodel

$$\widehat{\mathcal{M}}_\lambda = \{j : |\omega_j| > \lambda\}$$

is selected by SIS. The parameter $\lambda$ can be chosen by the order statistic $|\omega|_{(K)}$ for a fixed model size $K$, (e.g., $K = \lfloor n/\log n \rfloor$) or by other data-adaptive tuning criteria.

Similar to screening main effects, the goal of interaction screening is to screen out unimportant interaction terms in (1) while keeping important ones. A naive extension of the SIS to interaction screening would be to screen interactions based on $\boldsymbol{\Omega} = (\mathbf{Z})^\top \mathbf{y}$. Note that $\boldsymbol{\Omega}$ is a $p(p+1)/2$ dimensional vector with entries $\Omega_{jk} = \mathbf{Z}_{jk}^\top \mathbf{y}$, $1 \le j \le k \le p$, where $\mathbf{Z}_{jk}$ is a standardized vector from $\mathbf{X}_j \circ \mathbf{X}_k$. A direct interaction screening (DIS) procedure selects a model

$$\widehat{\mathcal{I}}_\lambda = \{(j, k) : |\Omega_{jk}| > \lambda\}.$$

Although the naive approach seems natural and intuitive, it has some drawbacks. In particular, this DIS approach totally ignores the intrinsic relationships between main effects and interaction effects. In other words, when the effect of $X_j X_k$ is evaluated, the effects of its parents $X_j$ and $X_k$ are not taken into account. As a result, the DIS tends to give suboptimal screening results. For example, when the data are skewed and $\mathrm{Corr}(X_j, X_j X_k) \neq 0$, the DIS is barely effective for interaction screening. To elaborate, consider the following toy example.

**A Motivating Example.** Consider the model $Y = X_1 + X_2 + a X_1 X_2 + \epsilon$, where $\epsilon$ is an independent noise. Furthermore, assume $X_j = W_j^2 - 1$, $j = 1, 2$, where $(W_1, W_2)^\top$ are jointly normal, and marginally standard normal with correlation $\rho \neq 0$. A simple calculation shows that

$$\begin{aligned} \mathrm{Corr}(Y, X_1 X_2) &= c_1 \mathrm{Cov}(Y, X_1 X_2) \\ &= c_1 \{16\rho^2 + a(20\rho^4 + 32\rho^2 + 4)\}, \end{aligned} \tag{2}$$

where $c_1 = [\mathrm{Var}(X_1 X_2)\mathrm{Var}(Y)]^{-\frac{1}{2}} > 0$. Then there are two facts: (i) $\mathrm{Corr}(Y, X_1 X_2) = c_1 16\rho^2 \neq 0$ when $a = 0$; (ii) $\mathrm{Corr}(Y, X_1 X_2) = 0$ when

$$a = -\frac{16\rho^2}{20\rho^4 + 32\rho^2 + 4}.$$

Fact (i) suggests that $X_1 X_2$ may be labeled as "important" by the naive approach, when it is actually not predictive to the response. Fact (ii) suggests that $X_1 X_2$ may be labeled as "unimportant" when it is truly important. In either case, the naive correlation ranking for interactions does not work even for this simple example.

In short, the naive screening procedure DIS fails to account for intrinsic correlations between interaction terms

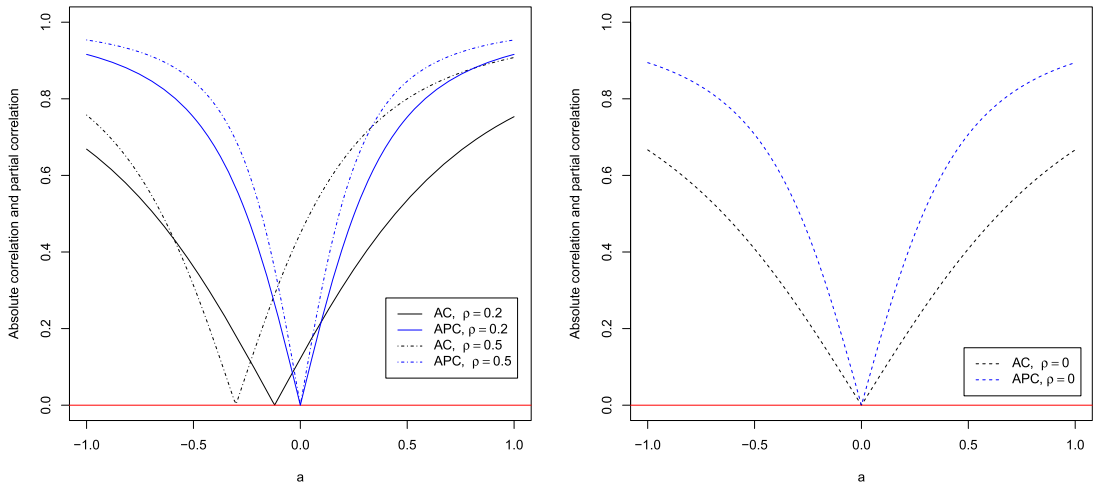


*Figure 1. Plots of absolute correlation (AC) and absolute partial correlation (APC) with respect to the coefficient $a$ for the toy example. Left, $\rho = 0.2$ and 0.5; Right, $\rho = 0$.*

and their parents. This motivates us to develop an alternative method which takes into account main effects when evaluating interaction effects and can improve accuracy for interaction screening.

## 2.3 Main-effect-adjusted interaction screening

To improve the naive correlation ranking method, we consider the partial correlation between $Y$ and $X_jX_k$ conditional on $X_j$ and $X_k$, denoted by $\text{pCorr}(Y, X_jX_k|X_j, X_k)$, or $\text{pCorr}(Y, X_j^2|X_j)$ for a quadratic term. Formally speaking, the partial correlation between two random variables $X$ and $Y$ given a set of $q$ controlling variables $\mathbf{Z} = (Z_1, Z_2, ..., Z_q)$, denoted by $\text{pCorr}(X, Y|\mathbf{Z})$, is the correlation between the residuals $R_X$ and $R_Y$ resulting from the linear regression of $X$ with $\mathbf{Z}$ and of $Y$ with $\mathbf{Z}$, respectively. When $q = 1$, it is called the first-order partial correlation. When $q = 2$, it is called the second-order partial correlation.

To see advantages of the partial correlation approach, let us revisit the example in Section 2.2. It is easy to calculate the partial correlation as

$$\text{pCorr}(Y, X_1X_2|X_1, X_2) = \frac{a}{\sqrt{a^2 + c_2}}, \tag{3}$$

where $c_2$ is a positive constant. (We refer to the appendix for the calculation of equations (2) and (3).) In particular, $\text{pCorr}(Y, X_1X_2|X_1, X_2) = 0$ when $a = 0$ and $\big|\text{pCorr}(Y, X_1X_2|X_1, X_2)\big| \to 1$ as $|a| \to \infty$. This suggests that we can eliminate the influence of parental main effects using partial correlation when conducting interaction screening.

To make a better illustration, we compare in Figure 1 the absolute correlation (AC) and the absolute partial correlation (APC) with respect to the coefficient $a$ in the toy example for $\rho = 0.2$, 0.5 (left) and $\rho = 0$ (right), respectively. We observe that the APC score is not zero as long as $a \neq 0$, but this does not hold for the AC score if $\rho \neq 0$. Moreover, the APC score is typically larger than the AC score when $a$ is away from zero, as shown in the right plot when $\rho = 0$. This means that partial correlation is more powerful than correlation for detecting signals in interaction screening. Similar patterns also hold for quadratic effects $\{X_j^2\}_{j=1}^n$.

In practice, the sample partial correlation can be calculated easily. In the following, we propose a new procedure called interaction screening by partial correlation (ISPC). Here we conduct screening for both interaction and quadratic effects together, but one can certainly screen them separately.

**Interaction Screening by Partial Correlation (ISPC):**

1. Calculate the standardized interaction effects $\mathbf{Z}$. In other words, standardize the columns of $\mathbf{X}$, calculate interaction effects $\mathbf{X} \circ \mathbf{X}$, and standardize $\mathbf{X} \circ \mathbf{X}$ to obtain $\mathbf{Z}$.
2. Calculate the sample partial correlation $\mathbf{P}$ as

$$P_{jk} = \begin{cases} \widehat{\text{pCorr}}(Y, X_jX_k|X_j, X_k), & 1 \le j \le k \le p; \\ \widehat{\text{pCorr}}(Y, X_j^2|X_j), & 1 \le j \le p. \end{cases}.$$

3. Determine a threshold $\lambda$ and obtain a model

$$\hat{\mathcal{I}}_\lambda = \{(j, k) : |P_{jk}| > \lambda\}.$$

Theoretically speaking, one main advantage of the ISPC procedure is that it conducts interaction screening by taking into account main effects, which overcomes drawbacks of the naive approach. Furthermore, compared to two-stage methods, the ISPC does not require the underlying model to obey the hierarchical structure, therefore it is more flexible and can be applied even when the model hierarchy is

violated. Computationally, the ISPC is easy to implement and the algorithm is scalable to very high dimensional data. As shown in Section 2.4, the ISPC does not require storage of the matrix $\mathbf{Z}$, which makes the computation fast and feasible.

**Invariance Property of ISPC.** For the DIS procedure, it is crucial to center to main effects first before calculating the marginal correlation of interactions. The reason is that $\mathrm{Corr}(Y, X_iX_j)$ is not invariant with respect to translations. That is, in general, $\mathrm{Corr}(Y, X_jX_k) \neq \mathrm{Corr}(Y, (X_j+a_j)(X_k+a_k))$ when $\mathrm{Corr}(Y, X_j)\cdot\mathrm{Corr}(Y, X_k) \neq 0$. On the other hand, the partial correlation employed by ISPC is invariant of arbitrary coding transformation $X_j \to b_jX_j + a_j$, $b_j > 0$. It is another reason why ISPC is preferable.

## 2.4 Extension of ISPC to nonparametric rank correlation

In the above, we proposed the ISPC based on the standard Pearson correlation coefficient, which measures the strength of linear relationship between variables. In this section, we will extend the ISPC idea to nonparametric correlation coefficients.

Besides Pearson product-moment correlation, there are two classical measures of association between variables, Spearman's and Kendall's rank correlation coefficients. These two nonparametric versions of correlation can achieve about 91% efficiency of their parametric counterpart to test whether the correlation coefficient $\rho = 0$ when a normal assumption is satisfied [9], and they are more robust against heavy tailed distributions. Moreover, they are invariant of monotonic transformation and therefore useful to reveal a complex relationship between the response and covariates. For example, [12] studied Kendall's rank correlation for screening main effects, based on the model $Y = f(\sum_{j=1}^{p} \beta_p X_p + \varepsilon)$ with an arbitrary monotonic function $f$. Therefore, it is desirable to generalize the ISPC procedure to these nonparametric correlation coefficients.

For Spearman's correlation, there is no direct nonparametric interpretation for partial correlation. Since Spearman's rank correlation is equivalent to Pearson's correlation computed with ranks of the data points [18], it is a convention to compute the sample Spearman's partial correlation by calculating the sample Pearson's partial correlation of ranks. Following this convention, we can conduct ISPC easily with Spearman's partial correlation.

For Kendall's correlation, [11] defined the first-order partial rank correlation in the nonparametric context and showed a surprising result that the well-known formula for Pearson's partial correlation still holds. That is, for three random variables $U_1$, $U_2$, $U_3$, the following holds

$$\tau_{12\cdot 3} = \frac{\tau_{12} - \tau_{13}\tau_{23}}{\sqrt{1-\tau_{13}^2}\sqrt{1-\tau_{23}^2}}, \tag{4}$$

where $\tau_{ij}$ is the Kendall's rank correlation between $U_i$ and $U_j$, and $\tau_{12\cdot 3}$ is the Kendall's partial correlation between $U_1$ and $U_2$ conditioning on $U_3$. Therefore, this formula can be iteratively used to calculate higher-order partial rank correlation coefficients. For example, for four random variables $U_1$, ..., $U_4$, a second-order partial correlation can be calculated by

$$\tau_{12\cdot 34} = \frac{\tau_{12\cdot 3} - \tau_{14\cdot 3}\tau_{24\cdot 3}}{\sqrt{1-\tau_{14\cdot 3}^2}\sqrt{1-\tau_{24\cdot 3}^2}}, \tag{5}$$

where $\tau_{12\cdot 34}$ is the Kendall's partial correlation between $U_1$ and $U_2$ conditioning on $U_3$ and $U_4$. If $\Gamma$ is the inverse of correlation matrix of $\{U_j\}_{j=1}^4$, an equivalent formula is

$$\tau_{12\cdot 34} = -\frac{\Gamma_{12}}{\sqrt{\Gamma_{11}\Gamma_{22}}}. \tag{6}$$

To summarize, all three versions of correlation coefficients considered here satisfy formulas (4–6), which can be used to calculate sample partial correlation. We employ the ISPC by using Spearman's and Kendall's partial correlation coefficients, and call the procedures as ISPC-S and ISPC-K respectively. Also, it is straightforward to implement DIS with these rank correlations, which are denoted as DIS-S and DIS-K.

# 3. COMPUTATION ALGORITHMS FOR HIGH DIMENSIONAL DATA

Though it is straightforward to implement the DIS and ISPC, it is necessary to accelerate the computation by some techniques when the number of covariates is overwhelmingly large. When $p$ is really large, it may not be possible to store the entire matrix $\mathbf{Z}$ due to limited computer memory, which is a bottleneck of many interaction selection algorithms. For the DIS or ISPC, we do not need to store $\mathbf{Z}$ or even the $p(p+1)/2$ vector of all marginal statistics, say $\mathbf{P}$, as it targets only on the top elements. Therefore, in the screening process, we only need to identify and update the top $K$ elements or those elements above a pre-specified threshold for the marginal statistic. For example, when $p$ is $10^5$ or larger, it might not be possible to store all $p(p+1)/2$ marginal statistics for a desktop. But our algorithms still work. Given the data $\{\mathbf{X}, \mathbf{y}\}$, the following is the computational algorithm to implement DIS, when a model size $K$ is specified.

**Computational Algorithm for DIS.**

1. Let $j = 1$ and `MS` be a NULL vector to store the absolute marginal correlation, and the threshold `t`$= 0$.
2. Calculate the absolute sample correlation between $\mathbf{X}_j \circ \mathbf{X}_{\{j:p\}}$ and $\mathbf{y}$, where $\{j : p\}$ denotes the set $\{j, j+1, ..., p\}$. If none of the absolute marginal correlation is above the threshold `t`, go to step 4.
3. Combine `MS` with the absolute marginal correlation coefficients larger than `t`. Rank `MS` to identify the top $K$ elements, which are stored as the new `MS`. Set `t` as the minimal element of `MS`.
4. Let $j = j + 1$. Go to step 2 if $j \leq p$; otherwise, stop.

It is slightly more time-consuming to find partial correlation than correlation. To further accelerate the computation of ISPC, we can avoid calculating all second order partial correlations. To elaborate, we first introduce a lemma, which shows how to control the magnitude of higher-order partial correlations by correlation. For a set of random variables $U_1$, $U_2$, ..., $U_N$, let $\tau_{jk}$ be the correlation (of possibly all the three types considered in this paper) between $U_j$ and $U_k$, and $\tau_{jk\cdot\mathcal{K}}$ be the partial correlation between $U_j$ and $U_k$, conditional on $\{U_\ell|\ell \in \mathcal{K}\}$.

**Lemma 1.** *For a set of random variables $U_1$, $U_2$, ..., $U_N$, if $|\tau_{jk} = \mathrm{Corr}(U_j, U_k)| < \delta$ for all $1 \le j \le k \le N$, then all the absolute $m$th order partial correlation can be controlled by $\frac{\delta}{1-m\delta}$ when $(m+1)\delta < 1$. That is, for $j$, $k$ and index set $\mathcal{K}$, where $|\mathcal{K}| = m$ and $j, k \notin \mathcal{K}$, $|\tau_{jk\cdot\mathcal{K}}| < \frac{\delta}{1-m\delta}$.*

We are particularly interested in the second-order partial correlation $\mathrm{pCorr}(Y, X_jX_k|X_j, X_k)$ whose absolute value is bounded by $\frac{\delta}{1-2\delta}$ when the absolute correlation between every pair is bounded by $\delta$. To implement ISPC, we can first re-rank the features by their absolute marginal correlation coefficients with the response. The following is the computational algorithm for implementing the ISPC procedure.

**Computational Algorithm for ISPC.**

0. Index the features based on their absolute marginal correlation coefficients with the response so that $|\widehat{\mathrm{Corr}}(Y, X_j) \ge \widehat{\mathrm{Corr}}(Y, X_k)|$ when $j > k$.
1. Let $j = 1$ and `MS` be a NULL vector to store the marginal statistics, and the threshold $\mathtt{t} = 0$.
2. If the absolute marginal correlation between $\mathbf{X}_j$ and $\mathbf{y}$ is smaller than $\frac{\mathtt{t}}{1+2\mathtt{t}}$, then go to 2a; otherwise, go to 2b.

2a. Calculate the absolute sample correlation between $\mathbf{X}_j \circ \mathbf{X}_{\{j:p\}}$ and $\mathbf{y}$ and find the index set of interactions whose marginal correlation is above $\frac{\mathtt{t}}{1+2\mathtt{t}}$. Calculate the partial correlation for interactions in the index set. Go to step 4 directly if the index set is empty.

2b. Calculate the partial correlation for all interactions $\mathbf{X}_j \circ \mathbf{X}_{\{j:p\}}$.

3. Combine `MS` with the absolute marginal correlation coefficients larger than `t`. Rank `MS` to obtain the top $K$ elements, which are stored as the new `MS`. Set `t` as the minimal element of `MS`.
4. Let $j = j + 1$. Go to step 2 if $j \le p$; otherwise, stop.

Along the computation, the threshold `t` gets larger and the marginal correlation between $\mathbf{X}_j$ and $\mathbf{y}$ gets smaller. Once the condition in 2 holds, we can avoid calculating all the partial correlations and save a lot of time. In many scenarios, this trick makes the computation time of ISPC comparable to that of DIS. By using these techniques, we can implement DIS and ISPC with `R` program and handle quite large data sets with a desktop. If a target threshold instead of the model size $K$ is given, we can set `t` to the threshold directly in these algorithms.

As a final remark, the whole procedure replies only on the marginal statistics, so parallel computing can be further used to accelerate the computation for extremely large data sets, which many other screening methods may not be able to handle.

## 4. SIMULATION STUDIES

We demonstrate the finite sample performance of the proposed ISPC procedures under a variety of settings. Furthermore, they are compared with the naive DIS methods, which does not take into account main effects during interaction screening. In all the tables, we denote Pearson correlation screening methods by DIS and ISPC, Kendall's correlation screening methods by DIS-K and ISPC-K, and Spearman's correlation screening methods by DIS-S and ISPC-S, respectively. Three examples are designed. In each example, we implement all the methods with 100 replicates and report the average performance on identifying important interactions.

**Example 1** (Gaussian design)**.** Generate $n$ IID pairs $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ based on the model

$$Y = X_1 - 2X_2 + 2X_4 + X_1X_2 - X_3X_4 + \varepsilon, \tag{7}$$

where $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$, $\mathbf{\Sigma} = (\sigma_{jk})$ with $\sigma_{jk} = \rho^{|j-k|}$, and $\varepsilon \sim \mathcal{N}(0, 1)$ is independent of all covariates. Let $n = 300, \rho = 0.5$, and two dimension settings with $p = 600, p = 2000$. The index of important interaction effects is $\mathcal{I}^* = \{(1, 2), (3, 4)\}$.

**Example 2** (Non-Gaussian design)**.** Consider the same model (7). Let $\mathbf{X} = (X_1, ..., X_p)^\top$ be a random vector with $X_j = (W_j^2 - 1)/\sqrt{2}$ when $1 \le j \le 10$ and $X_j$'s IID from $\mathcal{N}(0, 1)$ when $11 \le j \le p$, where $W = (W_1, ..., W_{10})^\top \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ and $\mathbf{\Sigma} = (\sigma_{jk})$ with $\sigma_{jk} = \rho + (1 - \rho)\delta_{(j=k)}$. Let $n = 300, \rho = 0.5$, and two dimension settings with $p = 600, p = 2000$.

For both examples, we implement all the methods and select $K$ interactions with largest marginal statistics. In particular, we have tried three different values of $K = C\lfloor n/\log n \rfloor$ for $C$ =1, 2, and 3. The sure screening probabilities for the interaction terms $(X_1X_2)$ and $(X_3X_4)$ by all the procedures are summarized in Table 1. It is observed that, as the value of $K$ increases, the sure screening probabilities for important interactions for all the methods increase a little bit, but at the cost of an increasing false positive rate. This pattern is expected since more terms are identified in the screening process when a larger $K$ is used. Since the gain in sure screening probabilities is not that substantial when $C$ increases from 1 to 3, we recommend using a small value, say, $C = 1$, in order to control the false positive rate in these examples.

In Table 1, the last column "Average" is the average of the first four columns, which is a summary of the overall screening accuracy. Let us focus on $C = 1$ from now on. In

Table 1. *Sure screening probabilities for important interactions in Examples 1 and 2.*

| | | $p = 600$ | | $p = 2000$ | | |
|---|---|---|---|---|---|---|
| | | $X_1X_2$ | $X_3X_4$ | $X_1X_2$ | $X_3X_4$ | Average |
| Example 1 $K = \lfloor n/\log n \rfloor$ | DIS | 96% | 97% | 96% | 91% | 95.0% |
| | ISPC | 99% | 100% | 100% | 100% | 99.8% |
| | DIS-K | 85% | 76% | 75% | 53% | 72.3% |
| | ISPC-K | 89% | 87% | 83% | 70% | 82.3% |
| | DIS-S | 85% | 74% | 72% | 48% | 69.8% |
| | ISPC-S | 94% | 98% | 92% | 80% | 91.0% |
| Example 1 $K = 2\lfloor n/\log n \rfloor$ | DIS | 98% | 97% | 98% | 92% | 96.3% |
| | ISPC | 99% | 100% | 100% | 100% | 99.8% |
| | DIS-K | 89% | 78% | 80% | 60% | 76.8% |
| | ISPC-K | 92% | 92% | 88% | 74% | 86.5% |
| | DIS-S | 88% | 78% | 77% | 53% | 74% |
| | ISPC-S | 95% | 98% | 94% | 86% | 93.3% |
| Example 1 $K = 3\lfloor n/\log n \rfloor$ | DIS | 99% | 98% | 99% | 93% | 97.3% |
| | ISPC | 100% | 100% | 100% | 100% | 100% |
| | DIS-K | 92% | 84% | 83% | 64% | 80.8% |
| | ISPC-K | 94% | 94% | 91% | 77% | 89% |
| | DIS-S | 90% | 80% | 81% | 59% | 77.5% |
| | ISPC-S | 97% | 99% | 94% | 88% | 94.5% |
| Example 2 $K = \lfloor n/\log n \rfloor$ | DIS | 71% | 28% | 57% | 33% | 47.3% |
| | ISPC | 88% | 78% | 81% | 74% | 80.3% |
| | DIS-K | 63% | 95% | 33% | 87% | 69.5% |
| | ISPC-K | 58% | 56% | 32% | 44% | 47.5% |
| | DIS-S | 46% | 90% | 26% | 83% | 61.3% |
| | ISPC-S | 58% | 57% | 33% | 42% | 47.5% |
| Example 2 $K = 2\lfloor n/\log n \rfloor$ | DIS | 76% | 37% | 61% | 36% | 52.5% |
| | ISPC | 90% | 80% | 84% | 80% | 83.5% |
| | DIS-K | 67% | 96% | 40% | 90% | 73.3% |
| | ISPC-K | 63% | 67% | 40% | 46% | 54% |
| | DIS-S | 55% | 94% | 28% | 86% | 65.8% |
| | ISPC-S | 68% | 68% | 45% | 48% | 57.3% |
| Example 2 $K = 3\lfloor n/\log n \rfloor$ | DIS | 78% | 38% | 66% | 39% | 55.3% |
| | ISPC | 93% | 82% | 85% | 84% | 86% |
| | DIS-K | 71% | 97% | 42% | 95% | 76.3% |
| | ISPC-K | 68% | 73% | 43% | 48% | 58% |
| | DIS-S | 61% | 95% | 33% | 87% | 69% |
| | ISPC-S | 73% | 74% | 49% | 48% | 61% |

Example 1, the ISPC-type procedure shows consistent improvement over the corresponding DIS-type procedures in terms of the average sure screening probability, and the improvement is quite substantial for Kendall's and Spearman's correlation coefficients. The ISPC works the best by achieving as high as 99.8% sure screening probability in average. In Example 2, Pearson's correlation works better than the nonparametric rank correlation methods. Again, the ISPC is the best among all by achieving in average 80.3% sure screening probability. In Example 2, DIS is better than ISPC for rank correlation methods in identifying $X_3X_4$. This is the only case that DIS is better than ISPC in our entire numerical studies, which might be due to the underlying data generating process. Overall speaking, that partial correlation based screening methods are effective in identifying interactions.

Table 2. *Average true positive rate (TPR) and standard errors for Example 3*

| | $p = 600$ | | $p = 2000$ | |
|---|---|---|---|---|
| | TPR | SE | TPR | SE |
| DIS | 0.51 | 0.21 | 0.36 | 0.17 |
| ISPC | 0.59 | 0.22 | 0.44 | 0.21 |
| DIS-K | 0.78 | 0.11 | 0.73 | 0.11 |
| ISPC-K | 0.84 | 0.10 | 0.77 | 0.11 |
| DIS-S | 0.77 | 0.10 | 0.71 | 0.11 |
| ISPC-S | 0.84 | 0.10 | 0.76 | 0.11 |

**Example 3** (A Challenging Example)**.** Consider a complex data generation process where the number of important interaction terms is not fixed, but instead, it varies from one data set to another. The purpose of this example is to evaluate the performance of the proposed procedures throughout different scenarios. The design matrix is the same as in Example 2. Let $\mathcal{S} = \{1, 2, 3, 11, 12, 13\}$. Consider the model

$$Y = \sum_{j\in\mathcal{S}} \beta_j X_j + \sum_{j\le k\in\mathcal{S}} \gamma_{jk} X_j X_k + \varepsilon, \tag{8}$$

where all the coefficients $\{\beta_j | j \in \mathcal{S}\}$ and $\{\gamma_{jk} | j \le k \in \mathcal{S}\}$ are independently chosen from $\{-1, 0, 1\}$ with equal probability. That is, there are up to 21 nontrivial interaction effects in the data generating process. Let $n = 1000$, $p = 600$ or 2000. We fix the model size $K = \lfloor n/\log n \rfloor$ for all methods.

Given any data set, for each screening procedure, define its true positive rate (TPR) as the ratio of the number of selected important interactions over the total number of important interactions. Table 2 presents the average TPR over 100 data sets and the corresponding standard error, for three DIS-type and three ISPC-type procedures. It is observed that each ISPC procedure performs better than its DIS counterpart by achieving a higher TPR. In this case, nonparametric rank partial correlation works significantly better than Pearson partial correlations. And the ISPC-S and ISPC-K appear to be equally best among all the procedures.

## 5. REAL DATA EXAMPLE

It is very challenging to identify predictive interaction effects for modern high dimensional and complex data. To illustrate our proposed methods, we analyze a rat microarray expression data set [16], which has been analyzed by [10, 6]. For this data set, 120 12-week-old male rat offsprings were selected for tissue harvesting and microarray analysis. The microarrays used to analyze the RNA from the eyes of these animals contain more than 31,000 different probes. For each probe set, the intensity values were normalized to obtain summary expression values. Following [16, 6], we focused on only the 18,976 probes that are expressed in the eye tissue. In [10, 6], they were interested in identifying the genes that

Table 3. Sure screening probabilities of interaction effects

| | | $p = 400$ | | $p = 2000$ | | |
|---|---|---|---|---|---|---|
| | | $X_1X_2$ | $X_3X_4$ | $X_1X_2$ | $X_3X_4$ | Average |
| Analysis 1 ($K = 25$) | DIS | 33% | 25% | 13% | 16% | 21.8% |
| | ISPC | 44% | 47% | 38% | 28% | 39.3% |
| | DIS-K | 39% | 40 % | 13% | 24% | 29.0% |
| | ISPC-K | 41% | 49% | 27% | 23% | 35.0% |
| | DIS-S | 32% | 39% | 13% | 20% | 26.0% |
| | ISPC-S | 51% | 57% | 28% | 31% | 41.8% |
| Analysis 2 ($K = 120$) | DIS | 40% | 42% | 21% | 23% | 31.5% |
| | ISPC | 52% | 55% | 49% | 40% | 49.0% |
| | DIS-K | 54% | 55% | 26% | 32% | 41.8% |
| | ISPC-K | 64% | 69% | 38% | 39% | 52.5% |
| | DIS-S | 48% | 55% | 21% | 30% | 38.5% |
| | ISPC-S | 68% | 76% | 49% | 46% | 59.8% |

are relevant to the gene TRIM32, which has been found to cause Bardet-Biedl syndrome [3].

In general, the underlying important interaction effects are unknown, so we analyze this data set in the following two ways. First, we simulate a response $Y$ using a known model. Since the true model and important interaction terms are known, we can compare the ISPC and the DIS performance in terms of their interaction screening accuracy. Generate the response $Y$ using the real data and by a quadratic model, so that we can test the sure screening property. To be more precise, we first standardize the data set and obtain a $120 \times 18{,}976$ matrix. Then we randomly choose $p = 400$, or 2,000 probe sets from 18,976 ones for each replicate, and generate response $Y$ by the same quadratic model considered in Example 1. We repeat 100 times and report the sure screening probabilities for important interaction terms in Table 3. We use $K = \lfloor n/\log n \rfloor = 25$ and $K = n = 120$ for all the methods.

Overall speaking, the ISPC-type procedures consistently give better performance than the DIS-type procedures, for both Pearson correlation and nonparametric correlations. Though the sure screening probabilities are not so high as in simulations, the performance is still reasonable given that the sample size $n = 120$ is very small. In order to improve the coverage probability further, we may lower the threshold or use iterative screening method etc. One limitation is that there are so many spurious interactions on the top, which is not surprising given the huge number of total interactions versus the small sample size.

Second, we analyze the raw data where the truth is unknown. We report the interaction terms selected by the screening procedures, which provide a short list for scientists to conduct further validations. We treat gene TRIM32 as the response variable and try to identify top interactions associated with it by all six screening procedures. The analysis is based on the entire data set, which contains $p = 18{,}976$ genes of $n = 120$ samples. The total number of gene interaction pairs is $p(p+1)/2 = 180{,}053{,}776 \approx 1.8 \times 10^8$, therefore the total dimension is ultra-high.

In Table 4 we list top 5 pairs of gene interactions by six screening procedures. We observe that there are some variations in the top lists, which is not surprising considering an extremely large number of interactions and high correlations among genes. Among all the identified interactions, some pairs are selected frequently by multiple procedures, so they are deemed more "interesting". For example, two pairs of interactions, 1373599_at*1374388_at and 1370952_at*1373599_at, are both identified by four screening procedures out of six. Furthermore, we notice that gene 1373599_at is very active in working with other genes, as it is involved with many interactions in the top lists. We point out that these findings are just based on statistical analysis, and they need to be further validated by scientists in labs. On the other hand, the screening procedure is helpful to narrow down the number of research targets to a few top ranked pairs from $1.8 \times 10^8$ candidates.

## 6. DISCUSSION

Marginal screening is a powerful and computationally efficient technique for variable screening in high dimensional data analysis. Its effectiveness depends on many factors including distribution tails of the covariates and the noise, the

Table 4. Top interactions associated with gene TRIM32

| | DIS | DIS-K | DIS-S |
|---|---|---|---|
| Top Selected interactions | 1371995_at*1387793_at | 1372260_at*1373599_at | 1373599_at*1374388_at |
| | 1371995_at*1384708_at | 1373599_at*1374388_at | 1372260_at*1373599_at |
| | 1372369_at*1386344_at | 1370952_at*1373599_at | 1370952_at*1373599_at |
| | 1377455_at*1383417_at | 1371578_at*1377887_at | 1369583_at*1373599_at |
| | 1371995_at*1398873_at | 1369583_at*1373599_at | 1371578_at*1377887_at |
| | ISPC | ISPC-K | ISPC-S |
| Top Selected interactions | 1367746_at*1370303_at | 1377455_at*1391190_at | 1377455_at*1391190_at |
| | 1371995_at*1391643_at | 1370952_at*1373599_at | 1370952_at*1373599_at |
| | 1372318_at*1391628_at | 1387393_at*1391932_at | 1375233_at*1381886_at |
| | 1370266_at*1372318_at | 1372260_at*1373599_at | 1373599_at*1374388_at |
| | 1398859_at*1384620_at | 1373599_at*1374388_at | 1373599_at*1388145_at |

correlation structure among covariates, and the true model sparsity. In this paper, we discuss how to use the model structure to enhance effectiveness of interaction screening. We find that it is helpful to utilize the hierarchical structure when conducting interaction screening, and the screening procedure based on partial correlation outperforms simple correlation ranking. The proposed strategy is widely applicable to complex models. As a conclusion, we suggest that one take parental effects into account when calculating a marginal statistic of an interaction effect during screening.

## APPENDIX SECTION

### A.1 Calculation of (2) and (3)

Simple calculation shows that

$$E(X_m) = 0, \quad E(X_m^2) = 2, \quad m = 1, 2; \quad E(X_1X_2) = 2\rho^2,$$
$$E(X_1^2X_2) = E(X_1X_2^2) = 8\rho^2, \quad E(X_1^2X_2^2) = 4+32\rho^2+24\rho^4.$$

It follows that

$$\begin{aligned} & \mathrm{Cov}(Y, X_1X_2) \\ = \; & \mathrm{Cov}(X_1, X_1X_2) + \mathrm{Cov}(X_2, X_1X_2) + a\mathrm{Var}(X_1X_2) \\ = \; & 8\rho^2 + 8\rho^2 + a(4 + 32\rho^2 + 24\rho^4 - (2\rho^2)^2) \\ = \; & 16\rho^2 + a(20\rho^4 + 32\rho^2 + 4), \end{aligned}$$

which leads to (2).

To calculate (3), we first write

$$X_1X_2 = b_1X_1 + b_2X_2 + T,$$

where $\mathrm{Cov}(X_m, T) = 0$, $m = 1, 2$. So $Y = (1 + ab_1)X_1 + (1 + ab_2)X_2 + aT + \epsilon$, and the partial correlation

$$\begin{aligned} & \mathrm{pCorr}(Y, X_1X_2|X_1, X_2) \\ = \; & \frac{\mathrm{Cov}(T, aT + \epsilon)}{\sqrt{\mathrm{Var}(T)\mathrm{Var}(aT + \epsilon)}} \\ = \; & \frac{a\mathrm{Var}(T)}{\sqrt{\mathrm{Var}(T)(a^2\mathrm{Var}(T) + \mathrm{Var}(\epsilon))}} \\ = \; & \frac{a}{\sqrt{a^2 + \frac{\mathrm{Var}(\epsilon)}{\mathrm{Var}(T)}}}, \end{aligned}$$

which leads to (3). In particular,

$$\mathrm{Var}(T) = 20\rho^4 + 32\rho^2 + 4 - \frac{64\rho^4}{1 + \rho^2}.$$

### A.2 Proof of Lemma 1

For the first-order partial correlation,

$$|\tau_{jk\cdot\ell}| \;=\; \left| \frac{\tau_{jk} - \tau_{j\ell}\tau_{k\ell}}{\sqrt{1 - \tau_{j\ell}^2}\sqrt{1 - \tau_{k\ell}^2}} \right|$$

$$\begin{aligned} \leq \; & \frac{\delta + \delta^2}{1 - \delta^2} \\ = \; & \frac{\delta}{1 - \delta} \\ = \; & \delta_1. \end{aligned}$$

By the same technique, all second-order partial correlation is no more than $\frac{\delta_1}{1-\delta_1} = \frac{\delta}{1-2\delta}$. And Lemma 1 holds by induction.

## ACKNOWLEDGEMENTS

This research is supported in part by National Science Foundations DMS-1309507, DMS-1418172, DMS-1722691, and NSFC-11571009. The authors thank the Editor, AE, and reviewers for their helpful comments and suggestions.

*Received 9 February 2017*

## REFERENCES

[1] Bien, J., Simon, N., and Tibshirani, R. (2015). Convex hierarchical testing of interactions. *The Annals of Applied Statistics* **9**, 27–42. MR3341106

[2] Bien, J., Taylor, J., and Tibshirani, R. (2013). A lasso for hierarchical interactions. *The Annals of Statistics* **41**, 1111–1141. MR3113805

[3] Chiang, A. P., Beck, J. S., Yen, H.-J., Tayeh, M. K., Scheetz, T. E., Swiderski, R. E., Nishimura, D. Y., Braun, T. A., Kim, K.-Y. A. and Huang, J. (2006). Homozygosity mapping with SNP arrays identifies trim32, an e3 ubiquitin ligase, as a bardet-biedl syndrome gene (bbs11). *Proceedings of the National Academy of Sciences* **103**, 6287–6292.

[4] Choi, N. H., Li, W., and Zhu, J. (2010). Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association* **105**, 354–364. MR2656056

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

[6] Fan, J., Feng, Y., and Song, R. (2011). Nonparametric independence screening in sparse ultra-high-dimensional additive models. *Journal of the American Statistical Association* **106**, 544–557. MR2847969

[7] Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**, 849–911. MR2530322

[8] Hao, N. and Zhang, H. H. (2017). A note on high dimensional linear regression with interactions. *The American Statistician*, **to appear**. doi:10.1080/00031305.2016.1264311.

[9] Hotelling, H. and Pabst, M. R. (1936). Rank correlation and tests of significance involving no assumption of normality. *The Annals of Mathematical Statistics* **7**, 29–43.

[10] Huang, J., Horowitz, J. L., and Wei, F. (2010). Variable selection in nonparametric additive models. *Annals of Statistics* **38**, 2282–2313. MR2676890

[11] Kendall, M. G. (1942). Partial rank correlation. *Biometrika* **32**, 277–283.

[12] Li, G., Peng, H., Zhang, J. and Zhu, L. (2012). Robust rank correlation based screening. *The Annals of Statistics* **40**, 1846–1877. MR3015046

[13] Li, R., Zhong, W., and Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association* **107**, 1129–1139. MR3010900

[14] Moore, J., Asselberg, F., and William, S. (2010). Bioinformatics challenges for genome-wide association studies. *Bioinfomatics* **26**, 445–455.

[15] Park, M. Y. and Hastie, T. (2008). Penalized logistic regression for detecting gene interactions. *Biostatistics* **9**, 30–50.
[16] Scheetz, T. E., Kim, K.-Y. A., Swiderski, R. E., Philp, A. R., Braun, T. A., Knudtson, K. L., Dorrance, A. M., DiBona, G. F., Huang, J. and Casavant, T. L. (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences* **103**, 14429–14434.
[17] Van Steen, K. (2012). Travelling the world of gene-gene interactions. *Briefings in bioinformatics* **13**, 1–19.
[18] Wackerly, D., Mendenhall, W., and Scheaffer, R. (2007). *Mathematical statistics with applications*. Cengage Learning.
[19] Wu, J., Devlin, B., Ringquist, S., Trucco, M., and Roeder, K. (2010). Screen and clean: A tool for identifying interactions in genome-wide association studies. *Genetic Epidemiology* **34**, 275–285.
[20] Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E., and Lange, K. (2009). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* **25**, 714–721.
[21] Zhao, P., Rocha, G., and Yu, B. (2009). The composite absolute penalties family for grouped and hierarchical variable selection. *Annals of Statistics* **37**, 3468–3497. MR2549566
[22] Zhu, L.-P., Li, L., Li, R., and Zhu, L.-X. (2011). Model-free feature screening for ultrahigh-dimensional data. *Journal of the American Statistical Association* **106**, 1464–1475. MR2896849

Yue Selena Niu
Department of Mathematics
University of Arizona
Tucson, AZ 85721
USA
E-mail address: yueniu@math.arizona.edu

Ning Hao
Department of Mathematics
University of Arizona
Tucson, AZ 85721
USA
E-mail address: nhao@math.arizona.edu

Hao Helen Zhang
Department of Mathematics
University of Arizona
Tucson, AZ 85721
USA
E-mail address: hzhang@math.arizona.edu