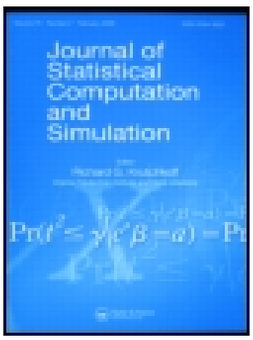

Journal of Statistical Computation and Simulation

ISSN: 0094-9655 (Print) 1563-5163 (Online) Journal homepage: https://www.tandfonline.com/loi/gscs20

Ultrahigh dimensional feature screening for additive model with multivariate response

Shishi Liu, Xiangjie Li & Jingxiao Zhang

To cite this article: Shishi Liu, Xiangjie Li & Jingxiao Zhang (2020): Ultrahigh dimensional feature screening for additive model with multivariate response, Journal of Statistical Computation and Simulation, DOI: 10.1080/00949655.2019.1703371

To link to this article: https://doi.org/10.1080/00949655.2019.1703371

JOURNAL OF STATISTICAL COMPUTATION AND SIMULATION
https://doi.org/10.1080/00949655.2019.1703371

Ultrahigh dimensional feature screening for additive model with multivariate response

Shishi Liu, Xiangjie Li and Jingxiao Zhang

Center for Applied Statistics, School of Statistics, Renmin University of China, Beijing, People's Republic of China

ABSTRACT

We consider feature screening for ultrahigh dimensional additive model with multivariate response in this paper. A new method named generalized correlation based projection screening is proposed by using generalized correlation between each predictor and multivariate response. The sure screening and ranking consistency properties are established under some regularized conditions for the proposed procedure. In addition, we construct an iterative version of the proposed screening procedure to enhance the finite sample screening performance. Both simulation studies and the real data analysis demonstrate that the proposed method works effectively.

KEYWORDS

Feature screening; additive model; multivariate response; generalized correlation; sure screening; ranking consistency

AMS SUBJECT CLASSIFICATIONS

62G07; 65D30; 62F07

1. Introduction

Advances in modern technology enable people to collect abundant data with ultrahigh dimensional predictors. Examples can be found in text data, gene expression data, imaging data, combinatorial chemistry data and so on. In the high dimension situation, there exist many difficulties in analysis such as rank deficient normal equation in linear regression prediction, overfitting and high computation complexity, etc. Hence, extracting useful variables from high dimensional data has been widely discussed for decades. Many penalization methods have been developed to deal with these problems, including Lasso [1], SCAD [2], elastic net [3], Dantzig selector [4] and MCP [5]. However, when the number of predictors is far beyond the sample size, these penalization methods incur challenges in speed, stability and accuracy, as pointed out in Fan and Lv [6] and Fan et al. [7]. Thus sure independence screening (SIS) method proposed by Fan and Lv [6] has emerged as a natural way to simplify the ultrahigh dimensional problem by effectively eliminating unimportant predictors at first.

CONTACT Jingxiao Zhang zhjxiaoruc@163.com Center for Applied Statistics, School of Statistics, Renmin University of China, Beijing, 100872, People's Republic of China.

Supplemental data for this article can be accessed here. https://doi.org/10.1080/00949655.2019.1703371

The good numerical performances and novel theoretical properties have made SIS popular in ultrahigh dimensional analysis. Consequently, SIS method and its extensions have been applied to various settings. Fan and Song [8] extended the SIS method to ultrahigh dimensional generalized linear models and developed a maximum marginal likelihood screening procedure. Hall and Miller [9] extended the marginal Pearson correlation learning by considering polynomial transformations of predictors, which was further extended by Fan et al. [10] who considered nonparametric independence screening (NIS) in sparse ultrahigh dimensional additive models. Huang et al. [11] approximated additive components by B-spline bases and applied the adaptive group Lasso to select nonzero components. Zhu et al. [12] proposed a sure independence ranking and screening (SIRS) method in multi-index semiparametric models. Li et al. [13] proposed a robust screening procedure based on Kendall τ rank correlation (RRCS) for transformation models. Li et al. [14] developed a model-free feature screening method based on distance correlation (DC-SIS), which could be used for multivariate response and grouped predictors. Lin et al. [15] proposed a nonparametric function-correlative feature screening procedure without assumption on structural relationship between response and predictor. Chang et al. [16] constructed empirical likelihood locally conjuncted with marginal nonparametric regressions to identify predictors that locally contribute to the response. There are still many other screening methods based on other frameworks, including varying coefficient models [17,18], ultrahigh dimensional discriminant analysis [19], quantile regression models [20,21], and so forth.

No methods above except DC-SIS can handle multivariate response, but practitioners sometimes encounter the cases of several response variables in application. It will incur some information lost if we simply conduct screening procedure on each response without considering the correlation among the responses. Therefore, Li et al. [22] proposed a projection screening procedure for ultrahigh dimensional linear regression model with multivariate response by projecting each predictor on the linear space spanned by responses. However, they only concerned with linear relationship between each predictor and multivariate response.

In this paper, we consider the ultrahigh dimensional additive model with multivariate response to relax the linear restriction and aim to propose a new feature screening procedure measuring the relationship between the predictor and all responses simultaneously. We also develop an iterative algorithm to enhance the performance of our screening methods in practice. The proposed procedure is simple and fast to implement computationally. Moreover, we show that the proposed method, GCPS possesses sure screening property and ranking consistency property under certain conditions.

The rest of the paper is organized as follows. In Section 2, we introduce the generalized correlation-based projection screening (GCPS) procedure and other two screening index in comparison for additive model with multivariate response and further study the theoretical properties of GCPS. In Section 3, some numerical studies are conducted to evaluate the performance of our proposed methods. A real data analysis is shown in Section 4. We conclude the paper with a brief discussion in Section 5. All proofs and more simulation results are given in supplementary material.

2. Generalized correlation based projection screening (GCPS)

Suppose that we have a random sample $\{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i=1}^n$ from an additive model with multivariate response

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_q \end{pmatrix} = \begin{pmatrix} m_{11}(x_1) & m_{12}(x_2) & \cdots & m_{1p}(x_p) \\ m_{21}(x_1) & m_{22}(x_2) & \cdots & m_{2p}(x_p) \\ \vdots & \vdots & & \vdots \\ m_{q1}(x_1) & m_{q2}(x_2) & \cdots & m_{qp}(x_p) \end{pmatrix} \mathbb{1} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_q \end{pmatrix}, \tag{1}$$

where $\mathbb{1} = (1, 1, \ldots, 1)^\top$ is a p-dimensional vector.

Let $\boldsymbol{y} = (y_1, \ldots, y_q)^\top$ denote a q-dimensional response vector and $\boldsymbol{x} = (x_1, \ldots, x_p)^\top$ is a p-dimensional predictor vector. $\mathbf{X}$ is $n \times p$ sample matrix of $\boldsymbol{x}$; $\mathbf{Y}$ is $n \times q$ sample matrix of $\boldsymbol{y}$. $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_q)^\top$ represents random error with mean zero. For identifiability, we assume $\{m_{kj}(x_j)\}_{j,k}$ have mean zero, and response y_k's have mean zero consequently. Note that response y_k's do not have to be centered as additive models (1) may have an intercept term for each y_k. For general y_k's, we can deal with $Y_{ik} - \bar{Y}_k$ instead of Y_{ik} where $\bar{Y}_k$ is the sample mean of y_k. Hence, it has no influence on screening procedure whether y_k's are centered or not.

Consider the optimization problem

$$\min_{\boldsymbol{m}_k} E\big(y_k - \boldsymbol{m}_k(\boldsymbol{x})\big) \quad \text{such that} \quad \boldsymbol{m}_k(\boldsymbol{x}) = \sum_{j=1}^{p} m_{kj}(x_j).$$

By the theory of projection, optimizer m_{kj}'s are obtained as follows (see [23]),

$$m_{kj}(x) = E\left(\left(y_k - \sum_{l \neq j} m_{kl}(x_l)\right) \mid x_j = x\right), \quad j = 1, \ldots, p, \quad k = 1, \ldots, q.$$

Since we aim to identify the important predictors in additive model (1) rather than estimate each m_{kj}, we follow Fan et al. [10] and consider the following marginal nonparametric regression problems,

$$\min_{f_{kj} \in L^2(P)} E\big(y_k - f_{kj}(x_j)\big)^2,$$

where P denotes the joint distribution of $(\boldsymbol{x}, y_k)$. Obviously, the minimizer of the optimization problem above is the conditional expectation

$$f_{kj}(x_j) = E(y_k \mid x_j).$$

Using f_{kj} instead of m_{kj} helps to avoid the 'curse of dimensionality' since $f_{kj}(x_j)$ is just the solution to a marginal problem.

Note that $f_{kj} \in L^2(P)$ and we employ spline approximation to obtain sample estimates of f_{kj}'s for its convenience in both utilization and theoretical derivation, see [10,11,17,24], etc. Let $\mathcal{S}_n$ be the finite dimensional space of polynomial splines of degree $\ell \geq 1$ and $\{\psi_{jm} : m = 1, \ldots, d_n\}$ are normalized spline basis functions of $\mathcal{S}_n$ with $\|\psi_{jm}\|_\infty \leq 1$, where $\|\cdot$

$\|_\infty$ represents the sup norm and d_n, which is related to the sample size n, denotes the dimension of finite spline space $\mathcal{S}_n$. The subscript of $\mathcal{S}_n$ indicates that $\mathcal{S}_n$ depends on the sample size n for its dimension d_n associated with n. For any $f_{nkj} \in \mathcal{S}_n$, we have

$$f_{nkj}(x) = \sum_{m=1}^{d_n} \beta_{kjm}\psi_{jm}(x), \quad j = 1, \ldots, p, \quad k = 1, \ldots, q,$$

for some coefficients $\{\beta_{kjm}\}_{m=1}^{d_n}$. Under some smoothness restrictions, f_{kj} can be well approximated by some f_{nkj} in $\mathcal{S}_n$.

2.1. Generalized correlation

In order to construct a reasonable screening index which can well describe the relationship between predictor and multivariate response in additive model, we first review the generalized measure of correlation between predictor x_j and response y_k proposed by Hall and Miller [9],

$$\sup_{h\in\mathcal{H}} \frac{\text{cov}(h(x_j), y_k)}{\sqrt{\text{var}(h(x_j))\text{var}(y_k)}}, \tag{2}$$

where $\mathcal{H}$ is a vector space of functions. When restricting $\mathcal{H}$ to linear function space, the generalized correlation between x_j and y_k is the absolute value of conventional Pearson correlation coefficient. More generally, one could take $\mathcal{H}$ to be the vector space generated by any given set of functions h so that we could describe nonlinear correlation between x_j and y_k. Note that generalized correlation between y_k and x_j is actually the Pearson correlation between y_k and its best linear predictor in the function space of x_j. Because $var(y_k)$ in (2) does not depend on j, it would not affect the rank of each predictor x_j with respect to their associations to y_k. Therefore, we shall work instead with

$$\phi_{kj} = \sup_{h\in\mathcal{H}} \frac{\text{cov}(h(x_j), y_k)}{\sqrt{\text{var}(h(x_j))}}. \tag{3}$$

It might be quite onerous to compute the sample version of ϕ_{kj} since $\mathcal{H}$ can be infinite dimensional function space. However, under the assumption that $\mathcal{H}$ is a finite dimensional function space and assume the existence of $h \in \mathcal{H}$ that achieves ϕ_{kj}, the maximizer of sample estimate $\hat{\phi}_{kj}$ is the solution to the marginal least square regression problem [9],

$$\min_{h\in\mathcal{H}} \frac{1}{n}\sum_{i=1}^{n}(Y_{ik} - h(X_{ij}))^2. \tag{4}$$

This simplifies the computation problem in a wide range of cases.

Inspired by the generalized correlation (3) above, we consider the generalized correlation between x_j and y_k in additive model with multivariate response similarly and deal

with

$$\phi_{kj} = \sup_{f_{nkj} \in \mathcal{S}_n} \frac{\mathrm{cov}(f_{nkj}(x_j), y_k)}{\sqrt{\mathrm{var}(f_{nkj}(x_j))}}.$$

According to (4), the sample version of f_{nkj} can be expressed as

$$\min_{f_{nkj} \in \mathcal{S}_n} \mathbb{P}_n(y_k - f_{nkj}(x_j))^2 = \min_{\boldsymbol{\beta}_{kj} \in \mathbb{R}^{d_n}} \mathbb{P}_n(y_k - \Psi_j^\top \boldsymbol{\beta}_{kj})^2, \tag{5}$$

where $\Psi_j \equiv \Psi_j(x_j) = (\psi_1(x_j), \ldots, \psi_{d_n}(x_j))^\top$ is the d_n-dimensional basis functions and $\boldsymbol{\beta}_{kj} = (\beta_{kj1}, \ldots, \beta_{kjd_n})$ denotes the d_n-dimensional coefficients of basis functions. We denote $\psi_{jm} \equiv \psi_m(x_j)$ for short. $\mathbb{P}_n(\cdot)$ is utilized to represent sample mean of $(\cdot)$. For example, $\mathbb{P}_n(y_k - f_{nkj}(x_j))^2 = 1/n \sum_{i=1}^n (Y_{ik} - f_{nkj}(X_{ij}))^2$; $\mathbb{P}_n \Psi_j \Psi_j^\top = 1/n \Psi_{nj} \Psi_{nj}^\top$, where Ψ_{nj} is a $d_n \times n$ sample matrix of Ψ_j whose column vectors are $(\psi_1(X_{ij}), \ldots, \psi_{d_n}(X_{ij}))^\top$, $i = 1, \ldots, n$; $\mathbb{P}_n \Psi_j y_k = \frac{1}{n} \Psi_{nj} Y_k$, where $Y_k = (Y_{1k}, \ldots, Y_{nk})^\top$ is a n-dimensional sample vector of response y_k. By solving the marginal regression (5), we have

$$\hat{\boldsymbol{\beta}}_{kj} = (\mathbb{P}_n \Psi_j \Psi_j^\top)^{-1} \mathbb{P}_n \Psi_j y_k.$$

Consequently, the estimation of $f_{nkj}(x_j)$ is

$$\hat{f}_{nkj}(x_j) = \Psi_j^\top (\mathbb{P}_n \Psi_j \Psi_j^\top)^{-1} \mathbb{P}_n \Psi_j y_k, \quad j = 1, \ldots, p, \quad k = 1, \ldots, q,$$

and we define the population version of f_{nkj} correspondingly as

$$f_{nkj}(x_j) = \Psi_j^\top (E \Psi_j \Psi_j^\top)^{-1} E \Psi_j y_k,$$

where $E(\cdot)$ is the expectation of $(\cdot)$ under true model. Since each y_k has mean zero, it is easy to show that f_{nkj}'s have mean zero.

So the generalized correlation between x_j and y_k can be obtained by

$$\phi_{kj} = \frac{E f_{nkj}(x_j) y_k}{\sqrt{E f_{nkj}^2(x_j)}}.$$

Note that $E f_{nkj} y_k = E f_{nkj}^2$ and $\mathbb{P}_n \hat{f}_{nkj} y_k = \mathbb{P}_n \hat{f}_{nkj}^2$ (see [10]). It is obvious that $\phi_{kj} \geq 0$. Since square transformation is monotone increasing over $[0, \infty)$, it has no big impacts on our ranking-based method introduced below if we utilize $E f_{nkj}(x_j) y_k$ rather than

$$\frac{E f_{nkj}(x_j) y_k}{\sqrt{E f_{nkj}^2(x_j)}} = (E f_{nkj}^2(x_j))^{1/2}.$$

Thus, we take

$$\phi_{kj} = E f_{nkj}(x_j) y_k \quad \text{and} \quad \hat{\phi}_{kj} = \mathbb{P}_n \hat{f}_{nkj}(x_j) y_k$$

as generalized correlation in the following context.

2.2. Screening method

We aim to investigate correlation between predictor x_j and response vector $\boldsymbol{y}$ instead of considering correlation between predictor x_j and responses y_k individually. Li et al. [22] proposed a projection screening (PS) procedure for linear model with multivariate response, which projected each predictor x_j onto the linear space spanned by multivariate response $\boldsymbol{y}$ and took the norm-squared of this projection as a marginal utility screening index as follows,

$$E(x_j\boldsymbol{y}^\top)(E\boldsymbol{y}\boldsymbol{y}^\top)^{-1}E(\boldsymbol{y}x_j). \tag{6}$$

Analogously, in the additive model with multivariate response context, we consider generalized correlation ϕ_{kj} between x_j and y_k and replace the conventional correlation part $Ex_j\boldsymbol{y}^\top$ with ϕ_{kj}'s in the screening index (6), which leads us to construct a marginal utility screening index for model (1), i.e.

$$\omega_j = \boldsymbol{\phi}_j^\top(E\boldsymbol{y}\boldsymbol{y}^\top)^{-1}\boldsymbol{\phi}_j, \quad j = 1, \ldots, p, \tag{7}$$

where $\boldsymbol{\phi}_j = (\phi_{1j}, \ldots, \phi_{qj})^\top$. Intuitively, if x_j and all y_k's are independent, then generalized correlation ϕ_{kj} would be 0, $k = 1, \ldots, q$, and $\omega_j = 0$. If x_j and some y_k's are generalized correlated, there exist some non-zero components of vector $\boldsymbol{\phi}_j$, hence ω_j would be positive. The property above allows us to use ω_j conducting a feature screening procedure for additive model with multivariate response (1).

Assume the sample version of $(E\boldsymbol{y}\boldsymbol{y}^\top)^{-1}$ exists when $q \leq n$. The sample estimate of ω_j in (7) is defined as

$$\hat{\omega}_j = \hat{\boldsymbol{\phi}}_j^\top(\mathbb{P}_n\boldsymbol{y}\boldsymbol{y}^\top)^{-1}\hat{\boldsymbol{\phi}}_j, \quad j = 1, \ldots, p, \tag{8}$$

where $\hat{\boldsymbol{\phi}}_j = (\hat{\phi}_{1j}, \ldots, \hat{\phi}_{qj})^\top$.

Let $\mathcal{M}_k = \{j : Em_{kj}^2(x_j) \neq 0,\ 1 \leq j \leq p\}$ be the true predictors subset of the kth additive model, $k = 1, \ldots, q$ and define the whole true predictors subset as $\mathcal{M} = \{j : Em_{kj}^2(x_j) \neq 0,\ 1 \leq j \leq p, 1 \leq k \leq q\}$. Denote by $\mathcal{M}_k^c = \{1, \ldots, p\} \setminus \mathcal{M}_k$ the unimportant predictors subset of the kth additive model and $\mathcal{M}^c = \{1, \ldots, p\} \setminus \mathcal{M}$ be the whole unimportant predictors subset.

To illustrate the intuition of screening index (7), we introduce another two methods named Naive-GCS1 and Naive-GCS2 below for comparison.

Consider the kth additive model $y_k = \sum_{j=1}^p m_{kj}(x_j) + \varepsilon_k$ firstly,

$$\begin{aligned} Em_{kj}(x_j)y_k &= E\left(m_{kj}(x_j)\left(\sum_{l=1}^p m_{kl}(x_j) + \varepsilon_k\right)\right) \\ &= Em_{kj}^2(x_j) + \sum_{l\neq j, l\in\mathcal{M}_k} Em_{kj}(x_j)m_{kl}(x_l) + Em_{kj}(x_j)\varepsilon_k. \end{aligned}$$

It can be easily seen that $Em_{kj}(x_j)y_k$ is a good measurement of $Em_{kj}^2(x_j)$ when the second and third terms equal to 0. Consider some ideal conditions as follows,

(1) $Em_{kj}m_{kl} = 0$, for $j = 1, \ldots, p, l \in \mathcal{M}, k = 1, \ldots, q$,

(2) $Em_{kj}\varepsilon_k = 0$, for $j = 1, \ldots, p$ and $k = 1, \ldots, q$,
(3) $E\varepsilon_k\varepsilon_l = 0$, for $k \neq l, k, l = 1, \ldots, q$, i.e. $E\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^\top = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_q^2)$,

and define a screening index called naive generalized correlation based screening method 1 (Naive-GCS1),

$$\text{Naive-GCS1 index} : \omega_j^{(1)} = \boldsymbol{\phi}_j^\top \boldsymbol{\phi}_j = \sum_{k=1}^{q} \phi_{kj}^2.$$

Naive-GCS1 is equivalent to the sum of squares of marginal generalized correlation ϕ_{kj} between x_j and all y_k's, which shall well approximate $\sum_{k=1}^{q}(Em_{kj}y_k)^2$ under those ideal conditions. Besides, Naive-GCS1 can be seen as the sum of squares of Ef_{nkj}^2, as we mentioned before that $\phi_{kj} = Ef_{nkj}y_k = Ef_{nkj}^2$, and so it is also equivalent to the sum of squares of NIS index in Fan et al. [10].

However, those ideal conditions can hardly hold in realistic problems. If the ideal conditions are violated, Naive-GCS1 may suffer since it does not take the scales of y_k's into consideration and may sometimes overlap the effect of important predictors associated with those y_k's with small magnitude. Concerning with the magnitudes of response variables, we turn to define another screening index, naive generalized correlation-based screening method 2 (Naive-GCS2),

$$\text{Naive-GCS2 index} : \omega_j^{(2)} = \boldsymbol{\phi}_j^\top \mathbf{D}_{\boldsymbol{y}}^{-1} \boldsymbol{\phi}_j = \sum_{k=1}^{q} \frac{\phi_{kj}^2}{Ey_k^2},$$

where $\mathbf{D}_{\boldsymbol{y}} = diag(Ey_1^2, \ldots, Ey_q^2)$. Naive-GCS2 is equivalent to the weighted sum of squares of marginal generalized correlation ϕ_{kj} between x_j and all y_k's with weights Ey_k^2, $k = 1, \ldots, q$. $\mathbf{D}_{\boldsymbol{y}}^{-1}$ in Naive-GCS2 index may relieve the problem mentioned above.

It is obvious that Naive-GCS1 and Naive-GCS2 consider correlation between predictor x_j and multi-responses $\boldsymbol{y}$ separately and neglect the correlation among responses. In some actual situations, multi-responses are likely to be associated with some of others. To make full use of these correlations, we consider treating responses $\boldsymbol{y}$ as a whole. We decompose $Ey_ky_{k'}$ as an illustration,

$$\begin{aligned} Ey_ky_{k'} &= E\left(\left(\sum_{j=1}^{p} m_{kj}(x_j) + \varepsilon_k\right)\left(\sum_{l=1}^{p} m_{k'l}(x_l) + \varepsilon_{k'}\right)\right) \\ &= \sum_{j \in \mathcal{M}_k \cap \mathcal{M}_{k'}} Em_{kj}(x_j)m_{k'j}(x_j) + \sum_{j \neq l, j \in \mathcal{M}_k, l \in \mathcal{M}_{k'}} Em_{kj}(x_j)m_{k'l}(x_l) \\ &\quad + \sum_{j \in \mathcal{M}_k} Em_{kj}(x_j)\varepsilon_{k'} + \sum_{l \in \mathcal{M}_{k'}} Em_{k'l}(x_l)\varepsilon_k + E\varepsilon_k\varepsilon_{k'}. \end{aligned}$$

When $m_{kj}(x_j)m_{k'j}(x_j), j \in \mathcal{M}_k \cap \mathcal{M}_{k'}$ contributes to a nonzero $Ey_ky_{k'}$, responses y_k and $y_{k'}$ have correlated tendency caused by some same components of m_{kj} and $m_{k'j}$ with respect to predictor $x_j, j \in \mathcal{M}_k \cap \mathcal{M}_{k'}$. This correlated effect may result in that the contribution of $x_j, j \in \mathcal{M}_k \cap \mathcal{M}_{k'}$ is calculated repeatedly in Naive-GCS1 and Naive-GCS2 screening

index, which makes it unfair for other important predictors $j \notin \mathcal{M}_k \cap \mathcal{M}_{k'}$. To eliminate the effect of correlation among responses and conduct a more fair screening index, generalized correlation based projection screening (GCPS) index is proposed,

$$\text{GCPS} : \omega_j = \boldsymbol{\phi}_j^\top (E\boldsymbol{y}\boldsymbol{y}^\top)^{-1}\boldsymbol{\phi}_j.$$

$(E\boldsymbol{y}\boldsymbol{y}^\top)^{-1}$ in GCPS index is able to reduce the influence of the problem mentioned above.

For a random sample $\{(\mathbf{X}_i, \mathbf{Y}_i)\}_{i=1}^n$, we calculate $\hat{\omega}_j$ using (8) and select a subset of variables

$$\widehat{\mathcal{M}} = \{j : \hat{\omega}_j \geq cn^\tau\},$$

where c and τ are pre-determined thresholds defined in Condition (C9) below. In practice, we can find a pre-determined size ν_n to select the same subset

$$\widehat{\mathcal{M}} = \{j : \hat{\omega}_j \text{ is among the first } \nu_n \text{ largest of all}\},$$

where the size ν_n is taken to be smaller than the sample size n. The default value of ν_n is usually taken as $[n/\log(n)]$, where $[a]$ is the integer part of a, see Fan and Lv [6].

2.3. Iterative generalized correlation-based projection screening (IGCPS) procedure

It has been pointed out in [6] and [12] that marginal feature screening for ultrahigh dimensional linear model may encounter three main difficulties in practice. These problems also occur in additive model scenario. First, unimportant predictors which are highly correlated with important predictors could be selected with higher priority than those important predictors with slightly weaker correlation to responses. Second, important predictors that are marginally uncorrelated but jointly correlated with responses could not be selected by GCPS procedure. Third, correlation among predictors could make it harder to screen out all the important predictors. To relieve effect of these problems, we adopt an iterative GCPS procedure comprising the following steps to enhance the performance of the GCPS method.

Step 1. Start with an important subset $\mathcal{A} = \emptyset$. First apply GCPS and select ν_1 predictors with the first ν_1 largest $\hat{\omega}_j$ values, where ν_1 is pre-determined. Add the indices of these ν_1 predictors to $\mathcal{A}$. Let $|\mathcal{A}| = \nu_{\mathcal{A}}$ and denote the associated $n \times \nu_{\mathcal{A}}$ sample matrix by $\mathbf{X}_{\mathcal{A}}$.

Step 2. Let $\mathcal{A}^c = \{1, \ldots, p\} \setminus \mathcal{A}$ and denote the remaining $n \times (p - \nu_{\mathcal{A}})$ sample matrix by $\mathbf{X}_{\mathcal{A}^c}$. Then define $\mathbf{X}_{new} = \{\mathbf{I}_n - \mathbf{X}_{\mathcal{A}}(\mathbf{X}_{\mathcal{A}}^\top \mathbf{X}_{\mathcal{A}})^{-1}\mathbf{X}_{\mathcal{A}}\}\mathbf{X}_{\mathcal{A}^c}$, and apply again GCPS procedure to $\mathbf{X}_{new}$ and $\mathbf{Y}$. Select ν_2 predictors similarly and update the $\mathcal{A}$ by adding the indices of ν_2 selected predictors.

Step 3. Repeat Step 2 until the total number of selected predictors reaches a pre-determined number $\nu_n = [n/\log(n)]$, for example. The final selected predictor subset is $\mathcal{A}$.

Instead of working with $(\mathbf{X}_{\mathcal{A}^c}, \mathbf{Y}_{new})$ used in ISIS procedure, we fit the $(\mathbf{X}_{new}, \mathbf{Y})$ in the next step. This can weaken the priority of those unimportant predictors which are highly correlated with responses because of their association with selected predictors. The important predictor that marginally uncorrelated but jointly correlated with responses is possible to survive in the following steps. Furthermore, dealing with $\mathbf{X}_{new}$ helps eliminating the

correlation between selected predictors and unselected predictors. As a result, the iterative screening procedure can address those three difficulties to some extent.

Similarly, Naive-GCS1 and Naive-GCS2 can also conduct an iterative procedure, and we call them Naive-IGCS1 and Naive-IGCS2 respectively.

2.4. Theoretical properties

Theoretical properties of GCPS method are shown in this section. We allow the dimension of responses q and the dimension of predictors p to diverge with the sample size n and denote them as q_n and p_n respectively. Let s_n denote the non-sparsity size of true model $\mathcal{M}$ and s_n can also grow with the sample size n. Without loss of generality, let $[a, b]$ be the support of each x_j.

To establish the sure screening property, that is, all important predictors are included in the reduced model with probability approaching 1 as sample size n goes to infinity, we employ the normalized B-spline basis functions and assume the following additional conditions:

(C1) The nonparametric marginal projections $\{f_{kj}\}$ belong to a class of functions $\mathcal{F}$ whose rth derivative $f^{(r)}$ exists and is Lipschitz of order ι,

$$\mathcal{F} = \left\{ f : \left| f^{(r)}(s) - f^{(r)}(t) \right| \leq K|s-t|^{\iota}, s, t \in [a, b] \right\},$$

for some positive constant K, where r is a non-negative integer and $\iota \in (0, 1]$ such that $d = r + \iota > 0.5$.

(C2) The marginal density function g_j of x_j satisfies $0 < K_1 \leq g_j(x_j) \leq K_2 < \infty$ on $[a, b]$ for $1 \leq j \leq p$ for some constants K_1 and K_2.

Under Conditions (C1)–(C2), the following facts hold when ℓ, the degree of basis function, is no less than d in Condition (C1), i.e. $\ell \geq d$.

Fact 1. There exists a positive constant C_1 such that $\|f_{kj} - f_{nkj}\|^2 \leq C_1 d_n^{-2d}, j = 1, \ldots, p$, $k = 1, \ldots, q$, where $\|f\|^2 = Ef^2$ and d_n is the number of basis functions. (Stone [25])

Fact 2. There exists a positive constant D_5 such that $E\psi_{jm}^2(X_{ij}) \leq D_5 d_n^{-1}, j = 1, \ldots, p$, $m = 1, \ldots, d_n$, where d_n is the number of basis functions. (Stone [25]; Huang et al. [11])

Fact 3. There exist some positive constants D_3 and D_4 such that (Zhou et al. [26])

$$D_3 d_n^{-1} \leq \lambda_{\min}(E\Psi_j \Psi_j^{\top}) \leq \lambda_{\max}(E\Psi_j \Psi_j^{\top}) \leq D_4 d_n^{-1},$$

where $\lambda_{\min}$ and $\lambda_{\max}$ denote the minimum and maximum eigenvalue of the matrix.

Conditions (C1) and (C2) are standard in the context of additive models, they are commonly required for spline approximation to f_{kj}.

(C3) $\min_{j \in \mathcal{M}} Ef_{kj}^2 \geq c_1 d_n n^{-2\kappa}$, for some $0 < \kappa < d/(2d+1)$ and $c_1 > 0$.

(C4) There exists a positive constant c_1 and $\xi \in (0, 1)$ such that $d_n^{-2d-1} \leq c_1(1 - \xi) n^{-2\kappa}/C_1$.

Condition (C3) means the true components should have significant marginal effects, which is quite reasonable for the purpose of screening. Condition (C4) is required to show the good approximation of f_{nkj} to the marginal projection f_{kj}. Besides, when conditions above are satisfied, we have the following fact

Fact 4. $\min_{j \in \mathcal{M}} \|f_{nkj}\|^2 \geq c_1 \xi d_n n^{-2\kappa}$. (Fan et al. [10])

Moreover, we also need the following conditions to establish the sure screening property.

(C5) $\|\boldsymbol{m}(\boldsymbol{x})\|_\infty < B_1$ for some positive constant B_1, where $\|\cdot\|_\infty$ is the sup norm.

(C6) The random error $\{\varepsilon_{ik}\}_{i=1}^n$ are i.i.d. with conditional mean 0 and for any $B_2 > 0$, there exists a positive constant B_3 such that $E[\exp(B_2|\varepsilon_{ik}|)|\mathbf{X}_i] < B_3$.

(C7) There exist some positive constants D_1 and D_2, such that

$$D_1 q_n^{-1} \le \lambda_{\min}(E\boldsymbol{y}\boldsymbol{y}^\top) \le \lambda_{\max}(E\boldsymbol{y}\boldsymbol{y}^\top) \le D_2 q_n^{-1}.$$

(C8) Each y_k satisfies the sub-exponential tail probability uniformly. That is, there exists $s_0 > 0$, such that for $0 \le s < s_0$,

$$\max_{1\le m,k\le q_n} E\exp(sy_m y_k) < \infty.$$

(C9) There exist some positive constants $c > 0$, $0 < \alpha < 1/2$, and $0 \le \tau < 1/2 - \alpha$, such that $\min_{j\in\mathcal{M}} \omega_j \ge 2cn^{-\tau}$.

(C10) $q_n = O(n^\beta)$, $d_n = O(n^\gamma)$, where β and γ satisfy $\min\{\beta, \gamma, 1-3\gamma, 1-2\alpha-2\tau-8\beta-12\gamma\} > 0$. Furthermore, $\gamma \ge 2\kappa/(1+2d)$ with consideration of Condition (C4). (see α in Lemma 5.3 in supplementary material.)

Conditions (C5)–(C7) are required to ensure the uniform control in probability inequality. Condition (C5) might be strict, but it is often used in theoretical establishment in additive models (see Fan et al. [10], etc). Condition (C6) is mild and Condition (C7) is commonly used (see Li et al. [22], etc). Condition (C8) is used to control the tail distribution of responses to ensure the sure screening property. Condition (C9) means that the signal of important predictors cannot be too weak. Condition (C10) states that the number of the responses q_n and the number of spline bases d_n can grow with sample size n under certain constraints. The following theorem provides the sure screening property of the proposed GCPS method.

Theorem 1: *Under Conditions* (C1)–(C8)*, and* (C10)*, there exist some positive constants* b_5 *and* b_6 *such that*

$$\begin{aligned} &P\left(\max_{1\le j\le p_n} |\hat{\omega}_j - \omega_j| \ge cn^{-\tau}\right) \\ &\quad\le O(6p_n n^{2\beta}\exp(-c_2 n^{1-2\alpha-4\beta}) + 4p_n n^{1+2\beta} m_1 \exp(-s_1 n^\alpha) \\ &\qquad + 12p_n n^{\beta+2\gamma}\exp(-c_7 n^{1-3\gamma}) + 2p_n n^{2\beta}\exp(-b_6 n^{1-2\alpha-2\tau-8\beta-12\gamma}) \\ &\qquad + p_n(16n^{\beta+\gamma} + 4n^{\beta+2\gamma})\exp(-b_5 n^{1-2\tau-8\beta-11\gamma})). \end{aligned}$$

In addition, if Condition (C9) *holds, we have*

$$\begin{aligned} P\left(\mathcal{M} \subseteq \widehat{\mathcal{M}}\right) \ge 1 &- O(6s_n n^{2\beta}\exp(-c_2 n^{1-2\alpha-4\beta}) + 4s_n n^{1+2\beta} m_1 \exp(-s_1 n^\alpha) \\ &+ 12s_n n^{\beta+2\gamma}\exp(-c_7 n^{1-3\gamma}) + 2s_n n^{2\beta}\exp(-b_6 n^{1-2\alpha-2\tau-8\beta-12\gamma}) \\ &+ s_n(16n^{\beta+\gamma} + 4n^{\beta+2\gamma})\exp(-b_5 n^{1-2\tau-8\beta-11\gamma})), \end{aligned}$$

where s_n *is the non-sparsity size of* $\mathcal{M}$.

The proof of Theorem 1 is shown in supplementary material. Theorem 1 ensures that the probability of true model being selected into the screened submodel by GCPS procedure tends to 1 with an exponential rate. It follows from Theorem 1 that we can handle the NP-dimensionality $\log p_n = O(n^a)$, where $a < \min\{\alpha, 1-3\gamma, 1-2\alpha-2\tau-8\beta-12\gamma\}$. It shows that the number of responses q_n and spline bases d_n also affect the order of dimension of p_n: the smaller q_n and d_n are, the higher dimensionality we can handle, while d_n cannot be too small since the approximation error of f_{nkj} should not be too large.

Zhu et al. [12] proposed the ranking consistency property for independence screening. We also investigate the ranking consistency property of GCPS method here by imposing the following condition:

(C11) $\liminf_{p\to\infty}\{\min_{j\in\mathcal{M}}\omega_j - \max_{j\in\mathcal{M}^c}\omega_j\} \geq \delta > 0$.

Condition (C11) is strict that it requires the screening index is able to separate important and unimportant predictors well in the population level (see [19]), but it can provide a stronger theoretical result compared to sure screening property.

Theorem 2: *Suppose Conditions* (C1)–(C11) *hold, then we have*

$$\lim_{n\to\infty}\inf\{\min_{j\in\mathcal{M}}\hat{\omega}_j - \max_{j\in\mathcal{M}^c}\hat{\omega}_j\} \geq \frac{\delta}{2}, \quad a.s.$$

The proof of Theorem 2 is also shown in supplementary material. Theorem 2 shows that the proposed screening index can separate important and unimportant predictors well in the sample level under some conditions.

3. Simulation

In this section, we illustrate the finite sample performance of our screening method through several simulation examples. Here we compare the performance of GCPS with DC-SIS [14] and Naive-SIS1, Naive-SIS2, PS (three methods in [22]) to demonstrate the effectiveness of our method in detecting nonlinearity. In addition, performances of Naive-GCS1, Naive-GCS2 (defined in Section 2) are represented to show the necessity to consider multi-responses as a whole. Besides, Naive-ISIS1, Naive-ISIS2 and IPS (iterative versions of Naive-SIS1, Naive-SIS2 and PS respectively, see [22]) are listed to verify that the proposed method, GCPS, performs better than those iterative screening methods. Moreover, comparison with Naive-IGCS1, Naive-IGCS2 and IGCPS shows that iterative procedure can modify screening performance under some circumstance. We run 200 replications in each example and consider the following criteria to evaluate the screening performance.

- R_j: The average rank of x_j in terms of the sorted list by the screening procedure based on 200 replications.
- MMS: The minimum model size needed to include all important predictors.
- P_a: The proportion of all important predictors being selected into the submodel with size ν_n.
- P_j: The proportion of important predictor x_j being selected into the submodel with size ν_n.

Example 1: Following Li et al. [22], we consider a multi-responses linear model firstly,

$$\begin{aligned} y_1 &= \beta_{11}x_1 + \beta_{12}x_2 + \cdots + \beta_{15}x_5 + \varepsilon_1, \\ &\cdots\cdots \quad \cdots\cdots \\ y_{q_n} &= \beta_{q_n,1}x_1 + \beta_{q_n,2}x_2 + \cdots + \beta_{q_n,5}x_5 + \varepsilon_{q_n}. \end{aligned}$$

The dimension of multivariate response is defined by $q_n = [n^\alpha]$, $\alpha \sim U(0.2, 0.9)$, and $(n, p) = (400, 4000)$. The true predictor set is $\mathcal{M} = \{x_1, \ldots, x_5\}$. $\mathbf{X}_i$ is drawn independently from distribution $N(\mathbf{0}, \Sigma)$, where $\Sigma = (\sigma_{jl})_{p\times p}$ is the covariance matrix with $\sigma_{jl} = 0.8^{|j-l|}$. The nonzero coefficient follows that $\beta_{kj} = (-1)^{W_{kj}}(\log(\sqrt{n}) + |Z_{kj}|), k = 1, \ldots, q$, $j = 1, \ldots, 5$, where W_{kj} is a binary random variable with $P(W_{kj} = 1) = 0.6$ and $Z_{kj} \sim N(0, 1)$. The error term ε_k is generated from $N(0, 1)$ or $t(3)$.

Example 2: Following Fan et al. [10], we consider two linear models as follows

$$\begin{aligned} y_1 &= 3x_1 + 3x_2 + \varepsilon_1, \\ y_2 &= 2x_3 + 2x_4 + \varepsilon_2, \end{aligned}$$

with $(n, p) = (400, 2000)$. The error term ε_k follows $N(0, 1)$ or $t(3)$. The true predictor set is $\mathcal{M} = \{x_1, \ldots, x_4\}$. $\mathbf{X}_{i,\{2,4\}^c}$ is generated from $N(\mathbf{0}, \Sigma)$, where Σ is a $(p-2)\times(p-2)$ covariance matrix with diagonal elements $\sigma_{jj} = 1$ and off-diagonal elements $\sigma_{jl} = 0.6$, $l \neq j$. $x_2 = -\frac{1}{3}x_1^3 + \epsilon$, $x_4 = \frac{1}{2}\exp(-x_3) + \epsilon$, where $\epsilon \sim N(0, 1)$. Example 2 is designed to show that projection screening (PS) method proposed by Li et al. [22] based on Pearson correlation in linear models may suffer when linear models contain non-negligible nonlinear effect.

Example 3: Example 3 is to illustrate the performance of all screening procedures when responses are strongly correlated. For simplicity of notations, let $g_1(x) = e^{\sin(\pi x)}$, $g_2(x) = e^x(2x-1)^2$, $g_3(x) = \sin(2\pi x)/[2 - \sin(2\pi x)]$, $g_4(x) = 0.1\sin(2\pi x) + 0.2\cos(2\pi x) + 0.3\sin^2(2\pi x) + 0.4\cos^3(2\pi x) + 0.5\sin^3(2\pi x)$. We generate data from the following additive model:

$$\begin{aligned} y_1 &= 5g_1(x_1) + 3g_2(x_2) + 4g_3(x_3) + 6g_4(x_4) + \varepsilon_1, \\ y_2 &= 6g_1(x_1) + 5g_2(x_2) + 4g_3(x_3) + 3g_4(x_4) + \sqrt{0.86}\varepsilon_2, \\ y_3 &= 4g_1(x_1) + 7g_2(x_2) + 5g_3(x_3) + 3g_4(x_4) + \sqrt{0.5184}\varepsilon_3. \end{aligned}$$

And $\mathbf{X}_i$ is simulated according to

$$X_{ij} = \frac{W_{ij} + tU_i}{1+t}, \quad j = 1, \ldots, p,$$

where $W_{i1}, \ldots, W_{ip}$ and U_i are i.i.d. Unif(0,1) and $cor(X_{ij}, X_{il}) = t^2/(1+t^2)$, $j \neq l$. $(n, p) = (200, 2000)$. The true predictor set is $\mathcal{M} = \{x_1, \ldots, x_4\}$. ε_k follows $N(0, 1)$ or $t(3)$. Here we take $t = 1$ and the pairwise correlation is 0.5.

Example 4: We generate data from the additive model below:

$$\begin{aligned} y_1 &= 2(x_1^2 - 1) + x_2 + \varepsilon_1, \\ y_2 &= x_1 + 3\sin\left(\frac{\pi}{2}x_3\right) + \varepsilon_2, \\ y_3 &= \frac{1}{3}x_2^2 + \frac{1}{2}\exp(x_3) + \varepsilon_3, \\ y_4 &= x_4^2 + 2x_5 + \varepsilon_4, \end{aligned}$$

with $(n, p) = (200, 3000)$. The true predictor set is $\mathcal{M} = \{x_1, \ldots, x_5\}$. And $\mathbf{X}_i \sim N(\mathbf{0}, \Sigma)$, where $\Sigma = (\sigma_{jl})_{p\times p}$, $\sigma_{jj} = 1$; $\sigma_{jl} = 0.8$, $1 \le j, l \le 3$; $\sigma_{4j} = \sigma_{l4} = \sigma_{5j} = \sigma_{l5} = 0$; $\sigma_{jl} = 0.3$, $j, l \ge 6$. Consequently, x_4 and x_5 are uncorrelated with other predictors. The error terms ε_k's follow $N(0, 1)$ or $t(3)$. In this example, it might be challenging for methods based on Pearson correlation in screening x_4 out.

Example 5: Following the simulation model of Fan and Lv [6] and Li et al. [22], we consider

$$\begin{aligned} y_1 &= 5x_1 + 5x_2 + 5x_3 - 15x_4\sqrt{\rho} + \varepsilon_1, \\ y_2 &= 4x_1 + 6x_2 + 8x_3 - 18x_4\sqrt{\rho} + \varepsilon_2, \\ y_3 &= 5x_1 + 4x_2 + 6x_3 - 15x_4\sqrt{\rho} + \varepsilon_3, \end{aligned}$$

with $(n, p) = (200, 2000)$. The true predictor set is $\mathcal{M} = \{x_1, \ldots, x_4\}$. $\mathbf{X}_i \sim N(\mathbf{0}, \Sigma)$, where $\Sigma = (\sigma_{jl})_{p\times p}$, $\sigma_{jj} = 1$; $\sigma_{4j} = \sigma_{l4} = \sqrt{\rho}$; $\sigma_{jl} = 0.5$, $j, l \ge 4$ and $j \ne l$. Here we take $\rho = 0.5$ such that x_4 is marginally uncorrelated with y_k at population level. ε_k follows $N(0, 1)$ or $t(3)$. Example 5 is to show that iterative procedures would enhance screening performance under this situation.

Example 6: In this example, we consider simultaneous equation model widely used in economics,

$$\boldsymbol{y}^\top \Gamma = \boldsymbol{x}^\top B + \boldsymbol{\varepsilon}^\top,$$

where $\boldsymbol{y}$ is a q-dimensional response vector, $\boldsymbol{x}$ is a p-dimensional predictor vector, Γ is a $q \times q$ matrix, B is a $p \times q$ coefficient matrix and $\boldsymbol{\varepsilon}$ is a q-dimensional error term. The simultaneous equation model states that a response variable is not only determined by predictors but also affected by other responses. When Γ has full rank, we can simulate data by $\boldsymbol{y}^\top = \boldsymbol{x}^\top B\Gamma^{-1} + \boldsymbol{\varepsilon}^\top\Gamma^{-1}$. We consider generalizing the simultaneous equation model to simultaneous equation additive model and constructing one as follows:

$$\begin{aligned} y_1 - 3y_3 &= 5g_1(x_1) + 3g_2(x_2) + 4g_3(x_3) + 6g_4(x_4) + \varepsilon_1, \\ y_2 &= 6g_1(x_1) + 5g_2(x_2) + 4g_3(x_3) + 3g_4(x_4) + \sqrt{0.86}\varepsilon_2, \\ -y_1 + y_3 &= 4g_1(x_1) + 7g_2(x_2) + 5g_3(x_3) + 3g_4(x_4) + \sqrt{0.5184}\varepsilon_3. \end{aligned}$$

with $(n, p) = (200, 2000)$. The true predictor set is $\mathcal{M} = \{x_1, \ldots, x_4\}$. The rest settings are same to Example 3. Example 6 will show corresponding results of each method under this kind of model misspecification.

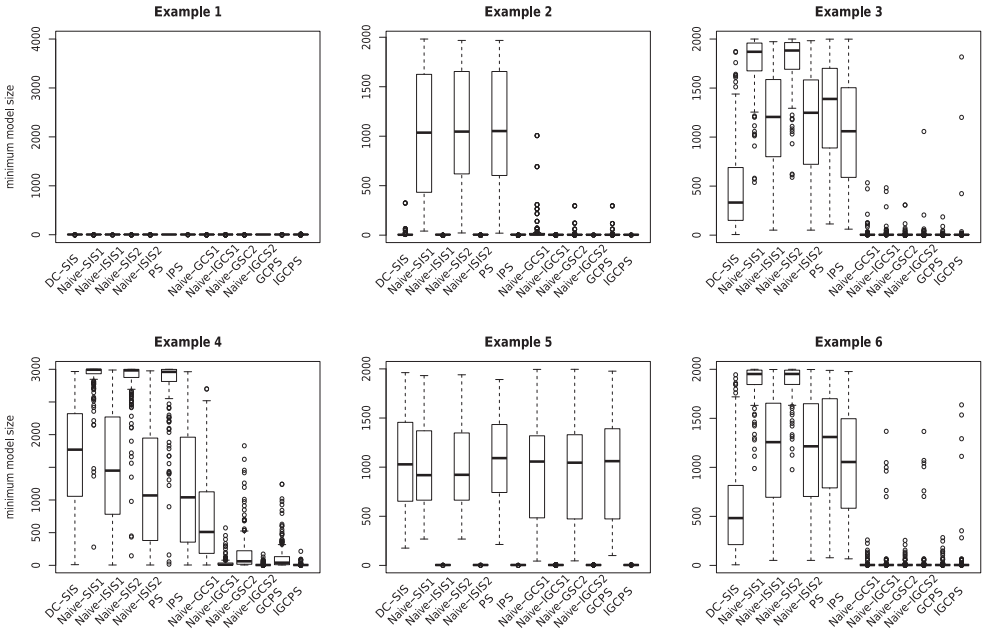

Figure 1. The boxplots of minimum model size (MMS) for Examples 1–6 where $\varepsilon \sim$ N(0,1).

Here following the suggestion of Fan and Lv [6], we keep the top $\nu_n = [n/\log(n)]$ variables after ranking in each screening procedure. The results of minimum model size (MMS) using different screening procedures for Examples 1–6 where error term ε_k's are generated from standard normal distribution N(0,1) are shown in Figure 1. Table 1 gives a summary of average ranks of true predictors for Examples 1–6 under the situation where ε_k's follow standard normal distribution. Table 2 reports the proportions of true predictors retained in the screened submodel using different screening methods with ε_k's follow N(0,1).

The model in Example 1 is a linear model with multivariate response, so it is not surprised that all methods perform well under this setting as demonstrated in Figure 1 and Tables 1–2. In addition, the diverging dimension of responses q_n under the restriction of Condition (C10) will not bring about many difficulties for all methods.

As for Example 2, it is a linear model with nonlinear effects. It is obviously shown in Figure 1 and Tables 1–2 that Naive-SIS1, Naive-SIS2 and PS methods which based on Pearson correlation can hardly screen out x_1 and x_3, while DC-SIS, Naive-GCS1, Naive-GSC2 and GCPS perform better in this situation. Furthermore, Naive-GCS2 and GCPS perform a little better than DC-SIS with respect to x_1. The two methods, Naive-GSC2 and GCPS, have almost the same performances in this setting for there is little correlation between response variables. Example 2 illustrates that methods based on Pearson correlation may get in trouble with the overlapping effects caused by non-linear factors in linear models. Moreover, all iterative methods outperform non-iterative methods. Naive-ISIS1, Naive-ISIS2 and IPS methods work well in this situation for it is a linear model here and iterative procedure can help detect non-linear associated variables efficiently.

Table 1. The average rank R_j of true predictor x_j for Examples 1–6. The ε's are generated from $N(0,1)$ distribution. '–' denotes there is no corresponding content because each example has different $\mathcal{M}$.

Example	Method	R_1	R_2	R_3	R_4	R_5
Example 1	DC-SIS	4.240	2.340	1.640	2.540	4.380
	Naive-SIS1	4.260	2.400	1.560	2.540	4.400
	Naive-ISIS1	3.940	2.440	1.580	2.720	4.340
	Naive-SIS2	3.740	2.460	1.980	2.740	4.180
	Naive-ISIS2	3.480	2.540	2.080	3.020	4
	PS	3.340	2.860	2.640	2.920	3.240
	IPS	3.240	2.960	2.700	2.960	3.220
	Naive-GCS1	4.420	2.460	1.400	2.540	4.340
	Naive-IGCS1	4.520	2.520	1.400	2.640	4.680
	Naive-GCS2	4.280	2.420	1.580	2.520	4.340
	Naive-IGCS2	4	2.500	1.600	2.680	4.220
	GCPS	4.340	2.460	1.680	2.440	4.240
	IGCPS	4.420	2.840	1.960	2.640	4.520
Example 2	DC-SIS	3.180	1	14.82	2	–
	Naive-SIS1	821.1	1	422.9	2	–
	Naive-ISIS1	3.980	1	4.900	2	–
	Naive-SIS2	898.6	1.040	341.5	1.960	–
	Naive-ISIS2	3.980	1.040	4.820	1.960	–
	PS	903.0	1.040	342.0	1.960	–
	IPS	3.980	1.040	4.780	1.960	–
	Naive-GCS1	2.080	1	62.84	14.56	–
	Naive-IGCS1	2.080	1	4.260	3.020	–
	Naive-GCS2	2.500	1	14.56	3.200	–
	Naive-IGCS2	2.500	1	4	2.560	–
	GCPS	2.500	1	14.44	3.180	–
	IGCPS	2.500	1	4	2.560	–
Example 3	DC-SIS	137.7	472.8	1.425	1.595	–
	Naive-SIS1	982.8	1696	1.565	12.36	–
	Naive-ISIS1	957.8	339.4	3.515	33.53	–
	Naive-SIS2	983.6	1715	1.505	14.67	–
	Naive-ISIS2	944.0	328.4	3.105	39.96	–
	PS	938.1	736.3	8.900	1.265	–
	IPS	938.2	182.6	4.565	1.265	–
	Naive-GCS1	3.225	14.92	2.635	1.090	–
	Naive-IGCS1	24.79	39.63	2.635	1.090	–
	Naive-GCS2	2.990	9.050	2.640	1.180	–
	Naive-IGCS2	13.24	29.19	2.640	1.180	–
	GCPS	2.505	3.045	4.395	1.010	–
	IGCPS	2.505	8.310	6.500	1.010	–
Example 4	DC-SIS	1.940	3	1.060	1686	398.6
	Naive-SIS1	2.020	2.820	1.160	2895	682.6
	Naive-ISIS1	2.020	2.820	1.160	1496	6.645
	Naive-SIS2	2.140	2.830	1.030	2828	63.50
	Naive-ISIS2	2.140	2.830	1.030	1246	4
	PS	1.980	2.980	1.040	2749	7.050
	IPS	1.980	2.980	1.040	1190	4
	Naive-GCS1	1.080	2.985	1.935	115.4	703.0
	Naive-IGCS1	1.080	2.985	1.935	14.14	36.11
	Naive-GCS2	1.085	3	1.915	19.85	183.7
	Naive-IGCS2	1.085	3	1.915	5.270	10.11
	GCPS	1.025	2.990	1.985	12.74	126.8
	IGCPS	1.025	2.990	1.985	5.250	9.135
Example 5	DC-SIS	2.860	11.74	1.040	1044	–
	Naive-SIS1	2.860	5.160	1.060	1030	–
	Naive-ISIS1	2.860	2.180	1.060	4.020	–
	Naive-SIS2	2.680	5.640	1.080	1027	–
	Naive-ISIS2	2.660	2.340	1.080	4.040	–

(*continued*).

Table 1. Continued.

Example	Method	R_1	R_2	R_3	R_4	R_5
	PS	1.720	2.900	1.420	1107	–
	IPS	1.720	2.900	1.420	4	–
	Naive-GCS1	4.640	10.94	1.060	979.5	–
	Naive-IGCS1	3.920	2.400	1.060	4	–
	Naive-GCS2	3.760	11.62	1.060	979.2	–
	Naive-IGCS2	3.740	2.460	1.060	4	–
	GCPS	2.860	2.840	1.020	1020	–
	IGCPS	2.860	2.380	1.020	4	–
Example 6	DC-SIS	166.3	518.2	1.125	2.610	–
	Naive-SIS1	957.8	1830	2.070	39.91	–
	Naive-ISIS1	1002	448.9	2.225	189.1	–
	Naive-SIS2	952.5	1812	2.535	35.51	–
	Naive-ISIS2	1019	468.9	3.090	176.4	–
	PS	928.5	799.5	9.635	1.305	–
	IPS	966.5	213.5	5.760	1.280	–
	Naive-GCS1	12.59	6.975	1.875	2.175	–
	Naive-IGCS1	62.10	10.12	1.835	1.990	–
	Naive-GCS2	4.985	9.645	2.005	2.020	–
	Naive-IGCS2	17.20	19.72	1.945	1.855	–
	GCPS	5.500	2.975	1.790	3.735	–
	IGCPS	26.30	4.915	1.820	9.850	–

In Example 3, responses share the same nonlinear components but with different coefficients. The responses are strongly correlated in this setting. We can find from Figure 1 that Naive-SIS1, Naive-SIS2 and PS methods are not applicable in this context any more, and DC-SIS also performs badly in this situation. These four procedures can screen out x_3 and x_4, but always miss x_1 and x_2 as shown in Tables 1 and 2. Naive-GCS1, Naive-GCS2 and GCPS improve a lot in screening out x_1 and x_2 in that generalized correlation does well in measuring non-monotone correlation between predictor and response, as pointed out in [9]. It is noteworthy that iterative procedures based on Pearson correlation, Naive-ISIS1, Naive-ISIS2 and IPS, also get in trouble in screening out x_1 and x_2. Hence, the proposed method is quite necessary in this context. Generally, GCPS and IGCPS perform best among these methods when multi-responses have correlation to some extent.

Example 4 illustrates the context that each response has their own additive terms. Recall that x_1, x_2 and x_3 are highly connected, while x_4 and x_5 are uncorrelated with other predictors. x_4 represents a non-linear effect, but x_5 brings a linear one. All methods perform well on x_1-x_3 and x_5, but it occurs challenges on x_4 which is uncorrelated with other predictors and corresponds to a non-linear term in the meantime, so that Naive-SIS1, Naive-SIS2, PS and DC-SIS methods can hardly screen it out. Variable x_5, though independent from other predictors, shares a linear marginal signal which is easy for PS to capture. Hence, we can see that PS outperforms DC-SIS, Naive-GCS1, Naive-GCS2 and GCPS on x_5. GCPS performs relatively better with respect to the tough variable x_4 among non-iterative methods in this scenario. Note that IGCPS procedure modifies the screening performance a lot on x_5.

In Example 5, we can tell from Figure 1 and Tables 1–2 that iterative screening procedures will significantly relieve the problem when some predictor is marginally uncorrelated but jointly correlated with responses.

Table 2. The selecting proportion P_j's of true predictors and P_a for Examples 1–6. The ε's follow $N(0,1)$ distribution. '–' denotes there is no corresponding content because each example has different $\mathcal{M}$.

Example	Method	P_1	P_2	P_3	P_4	P_5	P_a
Example 1	DC-SIS	1	1	1	1	1	1
	Naive-SIS1	1	1	1	1	1	1
	Naive-ISIS1	1	1	1	1	1	1
	Naive-SIS2	1	1	1	1	1	1
	Naive-ISIS2	1	1	1	1	1	1
	PS	1	1	1	1	1	1
	IPS	1	1	1	1	1	1
	Naive-GCS1	1	1	1	1	1	1
	Naive-IGCS1	1	1	1	1	1	1
	Naive-GCS2	1	1	1	1	1	1
	Naive-IGCS2	1	1	1	1	1	1
	GCPS	1	1	1	1	1	1
	IGCPS	1	1	1	1	1	1
Example 2	DC-SIS	1	1	0.960	1	–	0.960
	Naive-SIS1	0.080	1	0.340	1	–	0.020
	Naive-ISIS1	1	1	1	1	–	1
	Naive-SIS2	0.100	1	0.480	1	–	0.040
	Naive-ISIS2	1	1	1	1	–	1
	PS	0.100	1	0.500	1	–	0.040
	IPS	1	1	1	1	–	1
	Naive-GCS1	1	1	0.860	0.980	–	0.860
	Naive-IGCS1	1	1	1	1	–	1
	Naive-GCS2	1	1	0.960	1	–	0.960
	Naive-IGCS2	1	1	1	1	–	1
	GCPS	1	1	0.960	1	–	0.960
	IGCPS	1	1	1	1	–	1
Example 3	DC-SIS	0.535	0.125	1	1	–	0.075
	Naive-SIS1	0.020	0	1	0.950	–	0
	Naive-ISIS1	0.035	0.260	0.990	0.840	–	0.010
	Naive-SIS2	0.015	0	1	0.930	–	0
	Naive-ISIS2	0.025	0.280	0.990	0.825	–	0
	PS	0.030	0.090	0.960	1	–	0
	IPS	0.030	0.485	0.985	1	–	0.020
	Naive-GCS1	1	0.915	1	1	–	0.915
	Naive-IGCS1	0.960	0.860	1	1	–	0.825
	Naive-GCS2	1	0.940	1	1	–	0.940
	Naive-IGCS2	0.975	0.895	1	1	–	0.880
	GCPS	1	1	1	1	–	1
	IGCPS	1	0.990	0.980	1	–	0.970
Example 4	DC-SIS	1	1	1	0.005	0.205	0.005
	Naive-SIS1	1	1	1	0	0.085	0
	Naive-ISIS1	1	1	1	0.025	0.985	0.025
	Naive-SIS2	1	1	1	0	0.725	0
	Naive-ISIS2	1	1	1	0.060	1	0.060
	PS	1	1	1	0.005	0.970	0.005
	IPS	1	1	1	0.045	1	0.045
	Naive-GCS1	1	1	1	0.635	0.065	0.065
	Naive-IGCS1	1	1	1	0.950	0.765	0.760
	Naive-GCS2	1	1	1	0.925	0.365	0.365
	Naive-IGCS2	1	1	1	0.990	0.945	0.935
	GCPS	1	1	1	0.945	0.485	0.480
	IGCPS	1	1	1	0.995	0.955	0.950
Example 5	DC-SIS	1	0.980	1	0	–	0
	Naive-SIS1	1	0.980	1	0	–	0
	Naive-ISIS1	1	1	1	1	–	1
	Naive-SIS2	1	0.980	1	0	–	0
	Naive-ISIS2	1	1	1	1	–	1
	PS	1	1	1	0	–	0

(*continued*).

Table 2. Continued.

Example	Method	P_1	P_2	P_3	P_4	P_5	P_a
	IPS	1	1	1	1	–	1
	Naive-GCS1	0.980	0.980	1	0	–	0
	Naive-IGCS1	1	1	1	1	–	1
	Naive-GCS2	1	0.980	1	0	–	0
	Naive-IGCS2	1	1	1	1	–	1
	GCPS	1	1	1	0	–	0
	IGCPS	1	1	1	1	–	1
Example 6	DC-SIS	0.400	0.145	1	0.995	–	0.050
	Naive-SIS1	0.020	0	0.995	0.770	–	0
	Naive-ISIS1	0.025	0.385	0.990	0.595	–	0.005
	Naive-SIS2	0.020	0	0.995	0.800	–	0
	Naive-ISIS2	0.020	0.325	0.990	0.630	–	0.005
	PS	0.020	0.080	0.960	1	–	0
	IPS	0.015	0.415	0.975	1	–	0.015
	Naive-GCS1	0.940	0.980	1	1	–	0.925
	Naive-IGCS1	0.825	0.985	1	1	–	0.820
	Naive-GCS2	0.990	0.960	1	1	–	0.950
	Naive-IGCS2	0.945	0.965	1	1	–	0.925
	GCPS	0.980	0.995	1	0.995	–	0.970
	IGCPS	0.935	0.995	1	0.995	–	0.925

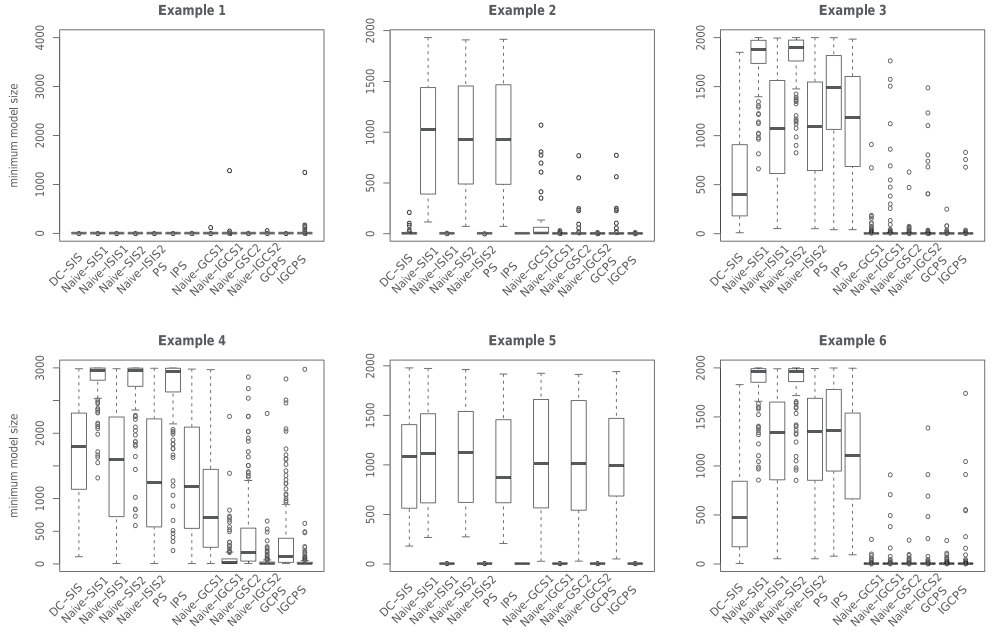

Figure 2. The boxplots of minimum model size (MMS) for Examples 1–6 where $\varepsilon \sim$ t(3).

Note that when Γ in Example 6 has full rank, the simultaneous equation additive model can be seen as a new additive model with multivariate response and new error terms are not independent but associated with others. The transformation matrix Γ shall change marginal signal magnitudes of each important predictor. Methods based on Pearson correlation still suffer on x_1 and x_2. Other screening methods perform poorly compared to the results in Example 3. In this scenario, GCPS keeps its advantage and outperforms other

Table 3. The average rank R_j of the true predictors for Examples 1–6. The ε_k's follow $t(3)$ distribution. '–' denotes there is no corresponding content because each example has different $\mathcal{M}$.

Example	Method	R_1	R_2	R_3	R_4	R_5
Example 1	DC-SIS	4.040	2.260	1.620	2.700	4.420
	Naive-SIS1	4.080	2.240	1.580	2.700	4.440
	Naive-ISIS1	3.920	2.260	1.600	2.920	4.320
	Naive-SIS2	3.720	2.420	2.040	2.780	4.180
	Naive-ISIS2	3.500	2.540	2.080	3.060	3.900
	PS	2.980	2.820	2.960	3.020	3.260
	IPS	2.880	3	3.040	3.140	3.060
	Naive-GCS1	6.440	4.260	3.140	4.480	6.900
	Naive-IGCS1	6.800	28.94	2.140	3.580	6.120
	Naive-GCS2	4.220	2.400	1.540	2.560	4.560
	Naive-IGCS2	4.460	2.960	1.540	2.760	5.480
	GCPS	4.200	2.540	1.580	2.520	4.280
	IGCPS	12.14	30.54	1.580	6.180	5.880
Example 2	DC-SIS	3.360	1	17.80	2	–
	Naive-SIS1	832.3	1	490.1	2	–
	Naive-ISIS1	4	1	5.140	2	–
	Naive-SIS2	835.9	1.060	399.7	1.940	–
	Naive-ISIS2	4	1.060	4.960	1.940	–
	PS	828.3	1.060	403.1	1.940	–
	IPS	4	1.060	5	1.940	–
	Naive-GCS1	2.280	1	113.0	26.68	–
	Naive-IGCS1	2.280	1	6.760	2.920	–
	Naive-GCS2	2.500	1.020	49.26	5.340	–
	Naive-IGCS2	2.500	1.020	4.600	2.580	–
	GCPS	2.500	1.020	49.82	5.360	–
	IGCPS	2.500	1.020	4.600	2.580	–
Example 3	DC-SIS	145.6	453.3	1.465	1.635	–
	Naive-SIS1	1045	1703	2.045	18.60	–
	Naive-ISIS1	1000	350.5	4.705	47.37	–
	Naive-SIS2	1041	1725	1.975	21.86	–
	Naive-ISIS2	1001	344.0	3.93	56.30	–
	PS	1061	911.0	7.825	4.325	–
	IPS	972.4	239.8	7	39.41	–
	Naive-GCS1	5.175	11.34	2.685	1.090	–
	Naive-IGCS1	29.04	19.45	4.01	1.09	–
	Naive-GCS2	4.095	7.565	2.710	1.130	–
	Naive-IGCS2	13.65	18.48	3.86	1.13	–
	GCPS	5.905	4.155	6.690	1.015	–
	IGCPS	4.950	5.560	10.49	1.015	–
Example 4	DC-SIS	1.885	3	1.115	1670	570.5
	Naive-SIS1	1.935	2.805	1.260	2842	791.6
	Naive-ISIS1	1.935	2.805	1.260	1534	28.50
	Naive-SIS2	2.055	2.830	1.120	2770	222.5
	Naive-ISIS2	2.055	15.28	1.120	1355	6.205
	PS	1.985	2.930	1.090	2684	95.25
	IPS	1.985	12.27	1.090	1314	6.345
	Naive-GCS1	1.170	3.060	1.865	157.9	902.1
	Naive-IGCS1	1.170	18.81	2.280	15.22	84.43
	Naive-GCS2	1.140	3.025	1.875	38.64	415.9
	Naive-IGCS2	1.140	17.23	1.875	7.750	36.82
	GCPS	1.135	3.040	1.890	28.54	328.4
	IGCPS	1.560	20.39	1.890	7.325	31.99
Example 5	DC-SIS	3.500	2.480	1.080	1030	–
	Naive-SIS1	3.100	2.460	1.020	1088	–
	Naive-ISIS1	2.640	2.460	1.020	4	–
	Naive-SIS2	2.800	2.520	1.040	1083	–
	Naive-ISIS2	2.560	2.520	1.040	4	–
	PS	1.760	2.880	1.360	1012	–

(*continued*).

Table 3. Continued.

Example	Method	R_1	R_2	R_3	R_4	R_5
	IPS	1.760	2.880	1.360	4	–
	Naive-GCS1	4.320	2.400	1.020	1050	–
	Naive-IGCS1	2.880	2.600	1.020	4	–
	Naive-GCS2	3.720	2.540	1.020	1054	–
	Naive-IGCS2	2.760	2.720	1.020	4	–
	GCPS	2.940	2.460	1.060	1036	–
	IGCPS	2.940	2.600	1.060	4	–
Example 6	DC-SIS	236.1	516.6	1.165	2.155	–
	Naive-SIS1	1012	1790	1.860	32.75	–
	Naive-ISIS1	1018	509.5	2.610	188.2	–
	Naive-SIS2	1003	1775	1.930	29.99	–
	Naive-ISIS2	1012	573.4	4.210	157.8	–
	PS	1031	988.4	9.920	3.395	–
	IPS	975.3	298.6	9.310	32.91	–
	Naive-GCS1	25.44	8.225	2.020	2.130	–
	Naive-IGCS1	88.09	20.76	1.860	1.945	–
	Naive-GCS2	8.420	11.78	2.075	1.900	–
	Naive-IGCS2	29.82	24.48	2	1.795	–
	GCPS	13.20	2.960	2.245	4.020	–
	IGCPS	57.53	3.995	1.760	18.00	–

procedures. Furthermore, iterative methods do not improve the screening performances in Example 6 for changes of marginal signal magnitudes and less stability in computation process. Naive-IGCS2 and IGCPS still offer acceptable results anyway.

Through Examples 1–6, we conclude that the proposed method works well in identifying non-linear effect in multi-responses additive model, especially when responses are associated with each other. And iterative procedure may improve screening performance to some degree.

Moreover, we consider heavy tail error terms for Examples 1–6 and the corresponding screening results are summarized in Figure 2 and Tables 3–4, which report the minimum model size (MMS), the average rank of true predictors and the selecting proportions of true predictors using different screening methods in Examples 1–6, respectively.

Similar to the conclusion made before, Naive-GCS1, Naive-GCS2 and GCPS methods perform better than the other procedures in additive model with nonlinear terms. Compared to the situation of normal distributed error term in Figure 1, it can be seen that heavy tail error terms affect all the screening results more or less. Methods based on additive models are more sensitive to heavy tail error because of the nonparametric nature of models. The performances of methods based on Pearson correlation are also affected by heavy tail error to some extent. Besides, DC-SIS procedure suffers a little from heavy tail error. Similar conclusions can be obtained by investigating Tables 3–4 and we will not overtalk about it here.

To sum up, through investigating Figures 1–2 and Tables 1–4, it is shown that GCPS method has its edge in additive models with multivariate response, especially when responses are strongly correlated and the association between predictor and response is non-monotone. Furthermore, we also add an extra Example 7 to illustrate whether the proposed methods still work well if additive structure is violated. For details, please see Example 7 in supplementary material.

Table 4. The selecting proportion P_j's and P_a for Examples 1–6. ε_k's follow $t(3)$ distribution. '–' denotes there is no corresponding content because each example has different $\mathcal{M}$.

Example	Method	P_1	P_2	P_3	P_4	P_5	P_a
Example 1	DC-SIS	1	1	1	1	1	1
	Naive-SIS1	1	1	1	1	1	1
	Naive-ISIS1	1	1	1	1	1	1
	Naive-SIS2	1	1	1	1	1	1
	Naive-ISIS2	1	1	1	1	1	1
	PS	1	1	1	1	1	1
	IPS	1	1	1	1	1	1
	Naive-GCS1	0.980	0.980	0.980	0.980	0.980	0.980
	Naive-IGCS1	1	0.980	1	1	1	0.980
	Naive-GCS2	1	1	1	1	1	1
	Naive-IGCS2	1	1	1	1	1	1
	GCPS	1	1	1	1	1	1
	IGCPS	0.960	0.960	1	0.980	0.980	0.900
Example 2	DC-SIS	1	1	0.920	1	–	0.920
	Naive-SIS1	0.120	1	0.200	1	–	0
	Naive-ISIS1	1	1	1	1	–	1
	Naive-SIS2	0.120	1	0.300	1	–	0
	Naive-ISIS2	1	1	1	1	–	1
	PS	0.120	1	0.300	1	–	0
	IPS	1	1	1	1	–	1
	Naive-GCS1	1	1	0.760	0.900	–	0.760
	Naive-IGCS1	1	1	1	1	–	1
	Naive-GCS2	1	1	0.880	0.980	–	0.880
	Naive-IGCS2	1	1	1	1	–	1
	GCPS	1	1	0.880	0.980	–	0.880
	IGCPS	1	1	1	1	–	1
Example 3	DC-SIS	0.470	0.110	1	1	–	0.055
	Naive-SIS1	0.010	0	0.990	0.900	–	0
	Naive-ISIS1	0.040	0.260	0.980	0.785	–	0
	Naive-SIS2	0.010	0	0.990	0.890	–	0
	Naive-ISIS2	0.035	0.260	0.985	0.765	–	0
	PS	0.010	0.045	0.960	0.975	–	0
	IPS	0.040	0.500	0.970	0.925	–	0.025
	Naive-GCS1	0.990	0.970	1	1	–	0.960
	Naive-IGCS1	0.950	0.925	0.995	1	–	0.875
	Naive-GCS2	0.995	0.975	1	1	–	0.970
	Naive-IGCS2	0.970	0.960	0.995	1	–	0.925
	GCPS	0.990	0.990	0.980	1	–	0.960
	IGCPS	0.980	0.975	0.965	1	–	0.925
Example 4	DC-SIS	1	1	1	0	0.170	0
	Naive-SIS1	1	1	1	0	0.125	0
	Naive-ISIS1	1	1	1	0.010	0.915	0.010
	Naive-SIS2	1	1	1	0	0.505	0
	Naive-ISIS2	1	0.995	1	0.040	0.990	0.040
	PS	1	1	1	0	0.780	0
	IPS	1	0.995	1	0.045	0.985	0.045
	Naive-GCS1	1	1	1	0.605	0.070	0.070
	Naive-IGCS1	1	0.985	0.995	0.930	0.590	0.585
	Naive-GCS2	1	1	1	0.860	0.245	0.240
	Naive-IGCS2	1	0.990	1	0.975	0.790	0.790
	GCPS	1	1	1	0.900	0.305	0.300
	IGCPS	0.995	0.990	1	0.980	0.800	0.800
Example 5	DC-SIS	0.980	1	1	0	–	0
	Naive-SIS1	1	1	1	0	–	0
	Naive-ISIS1	1	1	1	1	–	1
	Naive-SIS2	1	1	1	0	–	0
	Naive-ISIS2	1	1	1	1	–	1
	PS	1	1	1	0	–	0

(*continued*).

Table 4. Continued.

Example	Method	P_1	P_2	P_3	P_4	P_5	P_a
	IPS	1	1	1	1	–	1
	Naive-GCS1	0.980	1	1	0.020	–	0.020
	Naive-IGCS1	1	1	1	1	–	1
	Naive-GCS2	0.980	1	1	0.020	–	0.020
	Naive-IGCS2	1	1	1	1	–	1
	GCPS	1	1	1	0	–	0
	IGCPS	1	1	1	1	–	1
Example 6	DC-SIS	0.390	0.145	1	1	–	0.055
	Naive-SIS1	0.025	0	0.990	0.790	–	0
	Naive-ISIS1	0.040	0.295	0.985	0.660	–	0
	Naive-SIS2	0.030	0	0.995	0.800	–	0
	Naive-ISIS2	0.035	0.270	0.985	0.680	–	0
	PS	0.020	0.035	0.965	0.990	–	0
	IPS	0.030	0.320	0.965	0.955	–	0.005
	Naive-GCS1	0.870	0.970	1	0.995	–	0.850
	Naive-IGCS1	0.800	0.955	1	1	–	0.760
	Naive-GCS2	0.955	0.955	1	1	–	0.915
	Naive-IGCS2	0.925	0.955	1	1	–	0.885
	GCPS	0.935	0.990	0.995	0.995	–	0.930
	IGCPS	0.875	0.985	1	0.990	–	0.850

4. Real data analysis

In this section, we consider the yeast cell cycle gene expression data and chromatin immunoprecipitation (ChIP-chip) data [27,28] to illustrate the application of our proposed method. Identifying TFs related to cell cycle regulation is of great interest generally. This data set measures the amount of transcription (mRNA) and physical binding of transcription factors (TFs). Cell cycle gene expression data consists of data sets from three different experiments of about 800 genes. Here we use the part based on experiment of α factor which measures mRNA levels every 7 minutes covering 2 cell cycle periods with 18 records. ChIP-chip data contains the binding information of 106 TFs implying genome's regulation by binding to specific sequences. Excluding missing values, we obtain $n = 542$ cell cycle genes with 18 records and ChIP-chip data of 106 TFs. We take those 18 measurements as q response variables and 106 TFs data as p predictors, respectively. The data set is available in the R package *spls* with a detailed description.

We randomly split the data into a training set of size $n_{train} = 271$ and a test set of size $n_{test} = 271$. All screening procedures were carried out on the training set and selected a submodel of size $[n/\log(n)]$. Then, a penalized method for additive model proposed by Ravikumar et al. [24] was employed. The R package *SAM* was used in this step. Prior to analysis, we normalized predictors to be of mean 0 and variance 1. Here we compared the performances of Naive-SIS1, Naive-SIS2, PS, DC-SIS, Naive-GCS1, Naive-GCS2, GCPS and Naive-ISIS1, Naive-ISIS2, IPS, Naive-IGCS1, Naive-IGCS2, IGCPS. This process was repeated 100 times.

There are 21 experimentally verified TFs cell cycle related [29] and we compared the number of confirmed TFs being selected in the screening stage of all screening methods firstly, as depicted in Figure 3. The more confirmed TFs were screened out, the better the method performed. We could tell from Figure 3 that Naive-SIS1, Naive-SIS2 and Naive-ISIS1, Naive-ISIS2 methods performed the worst for that these methods screened out relatively less verified TFs among 100 replications. DC-SIS performed rather stable with

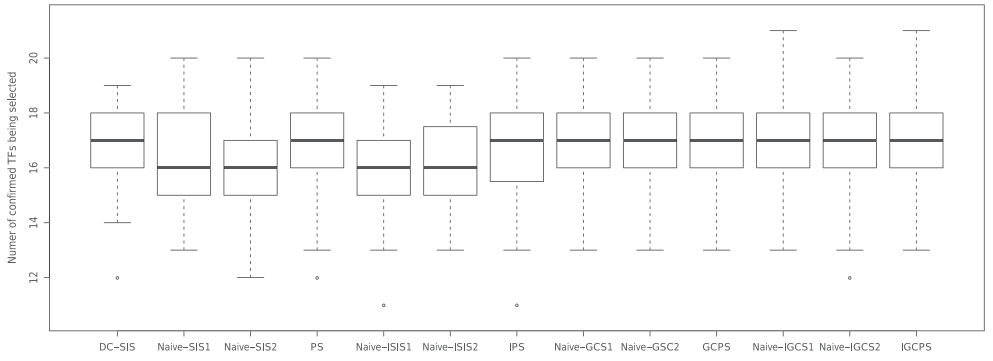

Figure 3. The number of confirmed TFs being selected in the screening stage of each method.

Table 5. Average Prediction Error over 100 replications and their standard deviations for the screening methods listed below.

Method	Prediction Error
DC-SIS	0.177876(0.010432)
Naive-SIS1	0.176984(0.008578)
Naive-SIS2	0.177102(0.008693)
PS	0.177021(0.008600)
Naive-ISIS1	0.177583(0.008571)
Naive-ISIS2	0.177815(0.008440)
IPS	0.177676(0.008554)
Naive-GCS1	0.175933(0.009380)
Naive-GCS2	0.175983(0.009414)
GCPS	0.175967(0.009381)
Naive-IGCS1	0.174814(0.009145)
Naive-IGCS2	**0.174725**(0.009047)
IGCPS	0.174777(0.009129)

a tight boxplot, but it hardly screened out more than 20 confirmed TFs compared to other methods. Difference among PS and Naive-GCS1, Naive-GCS2, GCPS was little, but Naive-IGCS1 and IGCPS modified the performance of Naive-GCS1 and GCPS for their longer upper tail.

Moreover, in order to further evaluate the performances of these screening methods, we used the test set to elucidate the prediction performances through prediction mean square error (PE) of each screening method, where $PE = \|\mathbf{Y}_{test} - \hat{\boldsymbol{m}}(\mathbf{X}_{test})\|_F^2/(qn_{test})$. $\mathbf{Y}_{test}$ and $\mathbf{X}_{test}$ denote the test part of sample data $\mathbf{Y}$ and $\mathbf{X}$ respectively. $\|\cdot\|_F$ represents the Frobenius norm and n_{test} is the sample size of test set.

Table 5 shows the average values and their associated standard deviations over 100 replications. It is illustrated that the average prediction error level of methods based on linear model and DC-SIS was relatively higher than those based on additive model. Besides, iterative procedures, Naive-IGCS1, Naive-IGCS2 and IGCPS, did improve the performances a little of Naive-GCS1, Naive-GCS2 and GCPS.

Combined with the screening results listed above, Naive-SIS1, Naive-SIS2 and their corresponding iterative procedures screened out less verified TFs and predicted slightly

poorly; PS and IPS performed well in the screening stage but obtained a slightly larger average prediction error; DC-SIS could hardly select as much confirmed TFs as other methods and had a higher prediction error level; Naive-GCS1, Naive-GCS2 and GCPS and their iterative versions outperformed other methods with respect to prediction error. Naive-IGCS1 did well in the screening stage but reached a higher prediction error compared to Naive-IGCS2 and IGCPS. Though obtained the best prediction, Naive-IGCS2 screened out less confirmed TFs in the screening stage. IGCPS performed best overall because of its good results in both screening and prediction stages.

5. Conclusion and discussion

In this paper, we developed the generalized correlation based projection screening (GCPS) method for ultrahigh dimensional additive model with multivariate response. The screening criterion ω_j was constructed by taking generalized correlations between x_j and y_k's and the correlations among y_k's into consideration simultaneously. We further established sure screening property and ranking consistency property of GCPS under some regularized conditions. An iterative GCPS procedure was proposed to relieve some difficulties in marginally feature screening for additive model. Simulation studies and a real data analysis illustrated the finite sample performance of the proposed method.

Here we have considered basic additive models that allow functions of individual variables, and it is natural to consider interactions and extend to select interactive terms. Moreover, inspired by Chip-seq gene data sets, it is interesting to investigate variable screening in functional data frameworks. These possible extensions are beyond this paper and will be interesting topics for future research.

Acknowledgments

The authors thank the editors and two referees for their valuable comments and suggestions.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

The research was supported by Fundamental Research Funds for Central Universities, and the Research Funds of Renmin University of China (No. 18XNI010).

References

[1] Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B (Stat Methodol). 1996;58:267–288.
[2] Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc. 2001;96(456):1348–1360.
[3] Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B (Stat Methodol). 2005;67(2):301–320.
[4] Candes E, Tao T. The Dantzig selector: statistical estimation when p is much larger than n. Ann Stat. 2007;35:2313–2351.
[5] Zhang CH. Nearly unbiased variable selection under minimax concave penalty. Ann Stat. 2010;38(2):894–942.

[6] Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. J R Stat Soc Ser B (Stat Methodol). 2008;70(5):849–911.
[7] Fan J, Samworth R, Wu Y. Ultrahigh dimensional feature selection: beyond the linear model. J Mach Learn Res. 2009;10:2013–2038.
[8] Fan J, Song R. Sure independence screening in generalized linear models with NP-dimensionality. Ann Stat. 2010;38(6):3567–3604.
[9] Hall P, Miller H. Using generalized correlation to effect variable selection in very high dimensional problems. J Comput Graph Stat. 2009;18(3):533–550.
[10] Fan J, Feng Y, Song R. Nonparametric independence screening in sparse ultra-high-dimensional additive models. J Am Stat Assoc. 2011;106(494):544–557.
[11] Huang J, Horowitz JL, Wei F. Variable selection in nonparametric additive models. Ann Stat. 2010;38(4):2282–2313.
[12] Zhu LP, Li L, Li R, et al. Model-free feature screening for ultrahigh-dimensional data. J Am Stat Assoc. 2011;106(496):1464–1475.
[13] Li G, Peng H, Zhang J, et al. Robust rank correlation based screening. Ann Stat. 2012;40:1846–1877.
[14] Li R, Zhong W, Zhu L. Feature screening via distance correlation learning. J Am Stat Assoc. 2012;107(499):1129–1139.
[15] Lin L, Sun J, Zhu L. Nonparametric feature screening. Comput Stat Data Anal. 2013;67:162–174.
[16] Chang J, Tang CY, Wu Y. Local independence feature screening for nonparametric and semiparametric models by marginal empirical likelihood. Ann Stat. 2016;44(2):515–539.
[17] Fan J, Ma Y, Dai W. Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models. J Am Stat Assoc. 2014;109(507):1270–1284.
[18] Liu J, Li R, Wu R. Feature selection for varying coefficient models with ultrahigh-dimensional covariates. J Am Stat Assoc. 2014;109(505):266–274.
[19] Cui H, Li R, Zhong W. Model-free feature screening for ultrahigh dimensional discriminant analysis. J Am Stat Assoc. 2015;110(510):630–641.
[20] He X, Wang L, Hong HG. Quantile-adaptive model-free variable screening for high-dimensional heterogeneous data. Ann Stat. 2013;41(1):342–369.
[21] Ma S, Li R, Tsai CL. Variable screening via quantile partial correlation. J Am Stat Assoc. 2017;112:650–663.
[22] Li X, Cheng G, Wang L, et al. Ultrahigh dimensional feature screening via projection. Comput Stat Data Anal. 2017;114:88–104.
[23] Härdle WK, Müller M, Sperlich S, et al. Nonparametric and semiparametric models. Berlin: Springer Series in Statistics; 2004.
[24] Ravikumar P, Lafferty J, Liu H, et al. Sparse additive models. J R Stat Soc Ser B (Stat Methodol). 2009;71(5):1009–1030.
[25] Stone CJ. Additive regression and other nonparametric models. Ann Stat. 1985;13:689–705.
[26] Zhou S, Shen X, Wolfe DA. Local asymptotics for regression splines and confidence regions. Ann Stat. 1998;26(5):1760–1782.
[27] Lee TI, Rinaldi NJ, Robert F, et al. Transcriptional regulatory networks in saccharomyces cerevisiae. Science. 2002;298(5594):799–804.
[28] Spellman PT, Sherlock G, Zhang MQ, et al. Comprehensive identification of cell cycle cregulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell. 1998;9(12):3273–3297.
[29] Wang L, Chen G, Li H. Group SCAD regression analysis for microarray time course gene expression data. Bioinformatics. 2007;23(12):1486–1494.