STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **6**, September 2018, pp 373–382.
Published online in International Academic Press (www.IAPress.org)

Variable selection and structure identification for ultrahigh-dimensional partially linear additive models with application to cardiomyopathy microarray data

Mohammad Kazemi, Davood Shahsavani, Mohammad Arashi*

Department of Statistics, Faculty of Mathematical Sciences, Shahrood University of Technology, Shahrood, Iran.

Abstract In this paper, we introduce a two-step procedure, in the context of ultrahigh-dimensional additive models, to identify nonzero and linear components. We first develop a sure independence screening procedure based on the distance correlation between predictors and marginal distribution function of the response variable to reduce the dimensionality of the feature space to a moderate scale. Then a double penalization based procedure is applied to identify nonzero and linear components, simultaneously. We conduct extensive simulation experiments to evaluate the numerical performance of the proposed method and analyze a cardiomyopathy microarray data for an illustration. Numerical studies confirm the fine performance of the proposed method for various semiparametric models.

Keywords Dimensionality Reduction, Partially Linear Additive Model, Structure Identification, Sure Screening, Variable Selection

AMS 2010 subject classifications 62J07, 62G05, 62G08

DOI: 10.19139/soic.v6i3.577

1. Introduction

Suppose we have a random sample $(y_i, x_{i1}, \ldots, x_{ip}), 1 \leq i \leq n$, where y_i is the response variable and $(x_{i1}, \ldots, x_{ip})$ is a p-dimensional covariate vector. Consider the *partially linear additive model* (PLAM)

$$y_i = \sum_{j \epsilon S_1} \beta_j x_{ij} + \sum_{j \epsilon S_2} f_j(x_{ij}) + \varepsilon_i, 1 \leq i \leq n, \tag{1}$$

where S_1 and S_2 are mutually exclusive and complementary subsets of $\{1, ..., p\}$, $\{\beta_j : j \epsilon S_1\}$ are the regression coefficients of covariates, $\{f_j : j \epsilon S_2\}$ are unknown smooth functions and the model error ε has conditional mean zero and finite variance σ^2. To ensure identifiability of the nonparametric functions, we assume that $E[f_j(X_j)] = 0$ for $j \epsilon S_2$. Estimation and variable selection for PLAM have been well studied in literature, for example, [16, 12, 7, 1, 17].

The use of model (1) is based on the assumption that the linear and nonlinear parts are known in advance. However, such prior information is usually unavailable, especially when the number of covariates is large. Thus, in addition to distinguish nonzero components, it is of great interest to develop some efficient methods to identify linear components from nonlinear ones.

Zhang et al. [21] studied the model selection using two penalties, simultaneously, to identify the zero and linear components in PLAM. They did not prove any selection consistency results for general partially linear models.

*Correspondence to: Mohammad Arashi (Email: m_arashi_stat@yahoo.com). Department of Statistics, Faculty of Mathematical Sciences, Shahrood University of Technology, IRAN.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

Motivated by this, Huang et al. [9] proposed a semiparametric regression pursuit method for distinguishing linear from nonlinear components using a group MCP penalty. Lian [13] provided a way to determine linear components by using SCAD penalty. This was a new usage of SCAD in which no variable selection is performed. Lian [14] successfully identified nonzero and linear components of model (1) by applying a two-fold SCAD penalty in the additive regression. Lian et al. [15] proposed another two penalty procedure in high dimensional setting using adaptive group LASSO in which insignificant predictors and parametric components were simultaneously identified.

However, with the rapid development of data collecting technologies, the PLAM often face the challenge of ultrahigh-dimensionality. The aforementioned regularization methods may not perform well for ultrahigh-dimensional data due to the simultaneous challenges of computational expediency, statistical accuracy, and algorithmic stability ([4]). Thus a natural question is how to identify the truly relevant predictors and parametric components in such ultrahigh-dimensional PLAM.

In this paper, this question is addressed on the basis of a two-stage procedure: (1) reducing the dimension from ultrahigh to moderate using a fast and efficient independence screening procedure; (2) identify nonzero and linear components from the screened submodel by a double penalization based procedure. Regrading the first stage, Fan and LV [3] first advocated the *sure independence screening* (SIS) method for the ultrahigh-dimensional linear models based on the Pearson correlation learning. Many authors further developed the SIS method and applied it to various statistical models, such as generalized linear models ([4, 6]), nonparametric additive models (NIS, [5]). Furthermore, in order to avoid the specification of a particular model structure, Zhu et al. [23] proposed a *sure independent ranking and screening* (SIRS) procedure for ultrahigh-dimensional data in the framework of the general multi-index models. Thereafter, a model-free SIS based on the distance correlation (DC-SIS) was developed by Li et al. [10]. Li et al.[11] proposed a *robust rank correlation screening* (RRCS) method based on the Kendall-correlation coefficient between response and predictors.

We propose a more robust approach, called *robust distance correlation sure independence screening* (RDC-SIS), to reduce dimensionality which ranks each covariate through its distance correlation with the marginal distribution function of the response variable. This method is model-free and we can expect that the procedure works well for skew or heavy tailed response variable. This procedure is a modification of DC-SIS proposed by Li et al. [10] in which Y is replaced by $F(Y)$. In the second stage, a double penalization based method is applied to refine the screened submodel and identify nonzero and linear components, simultaneously.

The rest of this paper is organized as follows. In Section 2, a modification of DC-SIS procedure is introduced for dimension reduction. Afterwards, a doubly penalized estimation method is explained in details in Section 3. We just focus on the SCAD penalty ([2]) but other penalties such as the LASSO ([19]) and MCP ([20]) could also be applied. In section 4, simulation studies are carried out to assess the performance of the proposed method and to compare it with some existing methods. A real data example is used for illustration in Section 5.

2. Screening Procedure (RDC-SIS)

At the start of the analysis, we do not know which component functions are linear or actually zero and thus the following general additive model is used initially:

$$Y = \sum_{j=1}^{p} f_j(X_j) + \epsilon. \tag{2}$$

We consider the problem of nonlinear variable screening in ultrahigh-dimensional feature space. The goal is to rapidly reduce the dimension of the covariate space p to a moderate scale via a computationally convenient procedure. We propose a robust feature screening procedure for model (2) using distance correlation between predictors and marginal distribution function of response variable. A review of distance correlation is as follows.

Sz'ekely, Rizzo, and Bakirov [18] introduced distance correlation as a measurement of dependence between two random vectors. The distance correlation between random vectors U and V with finite first moments is a

nonnegative value which is defined by

$$dcorr(U,V) = \frac{dcov(U,V)}{\sqrt{dcov(U,U)dcov(V,V)}}, \tag{3}$$

if $dcov(U,U)dcov(V,V) > 0$, and equals 0 otherwise. They stated that $dcov^2(U,V) = S_1 + S_2 - 2S_3$, where $S_j, j = 1,2,$ and 3, are defined as: $S_1 = E\{|U-\tilde{U}||V-\tilde{V}|\}, S_2 = E\{|U-\tilde{U}|\}E\{|V-\tilde{V}|\}, S_3 = E\{E(|U-\tilde{U}||U)E(|V-\tilde{V}||V)\}$, and $(\tilde{U},\tilde{V})$ is an independent copy of (U,V). In the category of model-free feature screening procedures for ultrahigh-dimensional setting, Li et.al [10] developed a sure independence screening method (DC-SIS) based on the distance correlation (DC-SIS) and showed that the DC-SIS has the sure screening property. Two remarkable properties of the distance correlation motivate them to use it in a feature screening procedure. The first one is the relationship between the distance correlation and the Pearson correlation coefficient. For two univariate normal random variables U and V, with the Pearson correlation coefficient ρ, Szekely, Rizzo, and Bakirov [18] showed that $dcorr(U,V)$ is strictly increasing in $|\rho|$. This property implies that the distance correlation based feature screening procedure is equivalent to the marginal Pearson correlation learning for linear regression with normally distributed predictors and random error. The second remarkable property of the distance correlation is that $dcorr(U,V) = 0$ if and only if U and V are independent. Distance correlation has properties of a true dependence measure, analogous to Pearson correlation ρ. It has the advantage that it can detect nonlinear relationships which are ignored by marginal correlation. For more details about distance correlation, see Szekely, Rizzo, and Bakirov [18].

Our new measure of correlation between Y and X_k is proposed by substituting $F(Y)$ instead of Y, i.e., marginal utility measure for predictor ranking is as:

$$\omega_k = dcorr(X_k, F(Y)), \qquad k = 1,\ldots,p, \tag{4}$$

where $F(Y)$ is the marginal distribution function of Y. We note that two univariate random variables U and V are independent if and only if U and $h(V)$, a strictly monotone transformation of V, are independent. This implies that $dcorr(X,F(Y)) = 0$ if and only if X and Y are independent. Furthermore, when the response is the skew or heavy-tailed, it can be expected that this procedure has a good performance. This screening procedure is a model-free in which one does not need to specify a model structure between the predictors and the response.

It is desirable to derive an estimator of ω_k based on the n independent and identical observations. Suppose that we have a random sample, $(X_i,Y_i)_{i=1}^n$, from the nonparametric additive model (2). We estimate $S_1, S_2,$ and S_3 through the usual moment estimation as follows:

$$\begin{aligned}
\hat{S}_{k,1} &= \tfrac{1}{n^2}\textstyle\sum_{i=1}^n\sum_{j=1}^n |X_{ik}-X_{jk}|\,|F_n(Y_i)-F_n(Y_j)|,\\
\hat{S}_{k,2} &= \tfrac{1}{n^2}\textstyle\sum_{i=1}^n\sum_{j=1}^n |X_{ik}-X_{jk}|\,\tfrac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n |F_n(Y_i)-F_n(Y_j)|,\\
\hat{S}_{k,3} &= \tfrac{1}{n^3}\textstyle\sum_{i=1}^n\sum_{j=1}^n\sum_{l=1}^n |X_{ik}-X_{lk}|\,|F_n(Y_j)-F_n(Y_l)|,
\end{aligned}$$

where $F_n(y) = \frac{1}{n}\sum_{i=1}^n I(Y_i \le y)$ is the empirical distribution function of Y. Thus, a natural estimator of $dcov^2(X_k,F(Y))$ is given by

$$\widehat{dcov^2}(X_k,F(Y)) = \hat{S}_{k,1} + \hat{S}_{k,2} - 2\hat{S}_{k,3}.$$

Similarly, we can define the sample distance covariances $dcov(X_k,X_k)$ and $dcov(F(Y),F(Y))$. Accordingly, the sample distance correlation between X_k and $F(Y)$ can be defined by

$$\widehat{dcorr}(X_k,F(Y)) = \frac{\widehat{dcov}(X_k,F(Y))}{\sqrt{\widehat{dcov}(X_k,X_k)}\sqrt{\widehat{dcov}(F(Y),F(Y))}}.$$

Thus we estimate ω_k with $\hat{\omega}_k = \widehat{dcorr}(X_k,F(Y))$. We select a set of important predictors with large $\hat{\omega}_k$. That is, the screened sub-model be determined by,

$$\hat{M}_{\nu_n} = \{1 \le k \le p : \hat{\omega}_k \ge \nu_n\}, \tag{5}$$

where ν_n is a pre-determined positive number. This procedure reduces the dimensionality from p to a possibly much smaller space with model size $d = |\hat{M}_{\nu_n}|$.

3. Variable Selection and Structure Identification

In section 2, a screening procedure is proposed to reduce the model size from a very large value p to a moderate scale d by specifying sensible threshold parameters ν_n, although it is difficult to choose in practice. A practical way is to select the top d variables by ranking marginal utilities. The choice of d plays a very important role in the screening stage. Fan and Lv [3] recommended $d = [?n/log(n)]$ as a sensible choice. Zhao and Li (2012) proposed an approach to select d for Cox models by controlling false positive rate. A larger value of the specified d would give a greater chance to include inactive variables. This can be solved by a penalty-based variable selection procedure given below.

Now, suppose that d variables are selected in the screening stage. Consider a joint nonparametric additive model $Y = \sum_{j=1}^{d} f_j(X_j) + \epsilon$. B-spline basis is used to approximate each of unknown smooth functions, i.e., $f_j(x) \approx \sum_{k=1}^{K} b_{jk}B_{jk}(x)$ for $j = 1, ..., d$. We use the two-fold penalization procedure to automatically identify different types of components, i.e., the coefficient $b = (b_1^T, \dots b_d^T)^T, b_j = (b_{j1}, \dots b_{jK})^T, j = 1, \dots, d$, are estimated in the following optimization problem

$$\hat{b} = \arg\min_b \frac{1}{2}\sum_{i=1}^{n}\left(Y_i - \mu - \sum_{j=1}^{d}\sum_{k=1}^{K} b_{jk}B_{jk}(X_{ij})\right)^2 + n\sum_{j=1}^{d} p_{\lambda_1}(\|b_j\|_{A_j}) + n\sum_{j=1}^{d} p_{\lambda_2}(\|b_j\|_{D_j}), \tag{6}$$

where two penalties $p_{\lambda_1}(.)$ and $p_{\lambda_2}(.)$ are used to identify the zero and the linear coefficients, respectively, with two regularization parameters λ_1 and λ_2, and A_j and D_j are two $K \times K$ matrices, $\|b_j\|_{A_j} = \left(b_j^T A_j b_j\right)^{\frac{1}{2}}, \|b_j\|_{D_j} = \left(b_j^T D_j b_j\right)^{\frac{1}{2}}$. There is some flexibility in choosing A_j and D_j but one requirement is that $\|b_j\|_{A_j} = 0$ if only if $\sum_k b_{jk}B_{jk}(x) \equiv 0$ and $\|b_j\|_{D_j} = 0$ if only if $\sum_k b_{jk}B_{jk}(x)$ is a linear function, so that the two penalties can be used to identify zero and linear components, respectively. One natural choice is $A_j = \{\int_0^1 B_{jk}(x)B_{jk'}(x)dx\}_{k,k'=1}^{K}$ and $D_j = \{\int_0^1 B''_{jk}(x)B''_{jk'}(x)dx\}_{k,k'=1}^{K}$ so that $\|b_j\|_{A_j} = \|\sum_k b_{jk}B_{jk}(x)\|$ and $\|b_j\|_{D_j} = \|\sum_k b_{jk}B''_{jk}(x)\|$. Let

$$Z_j = \begin{pmatrix} B_{j1}(X_{1j}) & B_{j2}(X_{1j}) & \cdots & B_{jK}(X_{1j}) \\ \vdots & \vdots & & \vdots \\ B_{j1}(X_{nj}) & B_{j2}(X_{nj}) & \cdots & B_{jK}(X_{nj}) \end{pmatrix}_{n\times K},$$

$Z = (Z_1, ..., Z_d)$ and $Y = (Y_1, ..., Y_n)$. Then (6) can be written in matrix form as

$$\hat{b} = \arg\min_b \frac{1}{2}\|Y - Zb\|^2 + n\sum_{j=1}^{d} p_{\lambda_1}\|b_j\|_{A_j}) + n\sum_{j=1}^{d} p_{\lambda_2}(\|b_j\|_{D_j}). \tag{7}$$

For later use we denote the objective function on the right hand side (7) as $Q(b)$. There are different way to specify the penalty functions, but here we only focus on the SCAD penalty function, defined by its first derivative

$$P'_{a,\lambda}(x) = \lambda\left\{I(|x| \le \lambda) + \frac{(a\lambda - |x|)_+}{(a-1)\lambda} I(|x| > \lambda)\right\}, \qquad x \ge 0, \tag{8}$$

with $a > 2$ and $P'_{a,\lambda}(0) = 0$. We use $a = 3.7$ as suggested in [2].

To find the minimum of (7) for fixed tuning parameters, we use the iterative local quadratic approximation (LQA) proposed in [2]. Using a simple Taylor expansion, given an initial estimate b_j^0, if $\|b_j\|_{A_j} > 0$ and $\|b_j\|_{D_j} > 0$, we approximate the penalty terms by

$$p_{\lambda_1}(\|b_j\|_{A_j}) \approx p_{\lambda_1}\left(\|b_j^{(0)}\|_{A_j}\right) + \frac{1}{2}\frac{p'_{\lambda_1}\left(\|b_j^{(0)}\|_{A_j}\right)}{\|b_j^{(0)}\|_{A_j}}\left\{\|b_j\|_{A_j}^2 - \|b_j^{(0)}\|_{A_j}^2\right\},$$

and

$$p_{\lambda_2}\left(\|b_j\|_{D_j}\right) \approx p_{\lambda_2}\left(\|b_j^{(0)}\|_{D_j}\right) + \frac{1}{2}\frac{p'_{\lambda_2}\left(\|b_j^{(0)}\|_{D_j}\right)}{\|b_j^{(0)}\|_{D_j}}\left\{\|b_j\|_{D_j}^2 - \|b_j^{(0)}\|_{D_j}^2\right\}.$$

After removing some irrelevant terms, the criterion becomes

$$Q(b) = \frac{1}{n}\|Y - Zb\|^2 + \frac{1}{2}b^T(\Omega_1 + \Omega_2)b \tag{9}$$

for two $dK \times dK$ matrices Ω_1 and Ω_2 which are defined by

$$\Omega_1 = \text{diag}\left(\frac{p'_{\lambda_1}\left(\|b_1^{(0)}\|_{A_1}\right)}{\|b_1^{(0)}\|_{A_1}}A_1, \ldots, \frac{p'_{\lambda_1}\left(\|b_d^{(0)}\|_{A_d}\right)}{\|b_d^{(0)}\|_{A_d}}A_d\right)$$

and

$$\Omega_2 = \text{diag}\left(\frac{p'_{\lambda_2}\left(\|b_1^{(0)}\|_{A_1}\right)}{\|b_1^{(0)}\|_{D_1}}D_1, \ldots, \frac{p'_{\lambda_2}\left(\|b_d^{(0)}\|_{A_d}\right)}{\|b_d^{(0)}\|_{D_d}}D_d\right)$$

Note that (9) is a quadratic function and thus there exists a closed-form solution. Then the updating equation given the current estimate $b^{(0)}$ is

$$b = \left(Z^TZ + n(\Omega_1 + \Omega_2)\right)^{-1}Z^TY. \tag{10}$$

The algorithm repeatedly solves the minimization criterion (9) and updates $b^{(m)}$ to $b^{(m+1)}$, $m = 0, 1, \ldots$ until the convergence is satisfied. That is, in the m-th iteration, we solve (9), where Ω_1 and Ω_2 are as defined above but with b_j^0 replaced by the current estimate $b_j^{(m)}$. The solution obtained from (9) is the new estimate $b^{(m+1)}$. During the iterations, as soon as some $\|b_j\|_{A_j}$ (respectively, $\|b_j\|_{D_j}$) drops below a certain threshold (10^{-6} in our implementation), the component is identified as a zero function (respectively, linear function).

4. Simulation Studies

In this section, some simulation studies have been conducted to asses the finite sample performance of our methods. We first consider three models to illustrate our proposed screening procedure (RDC-SIS). The performance of the RDC-SIS is then compared with the existing competitors, such as DC-SIS ([10]), SIRS ([23]), SIS ([3]) and NIS ([5]).

To evaluate the performance of proposed mehod, three criteria are considered. The first criterion is the minimum model size (denoted by *M*), that is the smallest number of covariates needed to ensure that all the active variables are selected. To get better inference, the 5%, 25%, 50%, 75% and 95% quantiles of *M* out of 200 replications were also presented. The second criterion (denoted by P_j) is the empirical probability that the active covariate X_j is selected, when the threshold $d = 2\lfloor n/log(n)\rfloor$ is adopted. The last criterion is the proportion (denoted by *S*) of truly active predictors that are identified by the screening procedure. Note that the first criterion does not need to specify a threshold. The more reliable screening procedure, the closer *M* value to the number of active predictor and also the closer *S* and P_j value to 1.

We also carry out some Monte Carlo studies to assess the effectiveness of our two stage proposed method to separation of the linear and nonlinear components and to identify insignificant covariates simultaneously in partial linear additive models of non-polynomial (NP) dimensionality based on double penalization.

It is also needed to find a data-driven procedure to choose the regularization parameters λ_1 and λ_2. We use the BIC-type criterion which is defined as

$$\log(\frac{1}{n}\|Y - Z\hat{b}_\lambda\|^2) + d_1\frac{\log(n/K)}{n/K} + d_2\frac{\log n}{n}, \tag{11}$$

Table 1. Five quantiles of minimum model size M, the empirical probability P_j and the proportion of S in Model 1.

ε	c	method	M					P					S
			5%	25%	50%	75%	95%	1	2	3	4	5	
$N(0,1)$	1	RDC-SIS	5	5	5	5	5	1.00	1.00	1.00	1.00	1.00	1.00
		DC-SIS	5	5	5	5	5	1.00	1.00	1.00	1.00	1.00	1.00
		SIS	5	5	5	5	5	1.00	1.00	1.00	1.00	1.00	1.00
		SIRS	5	5	5	5	5	1.00	1.00	1.00	1.00	1.00	1.00
	2	RDC-SIS	5	5	5	5	5	1.00	1.00	1.00	1.00	1.00	1.00
		DC-SIS	5	5	5	5	5	1.00	1.00	1.00	1.00	1.00	1.00
		SIS	5	5	5	5	5	1.00	1.00	1.00	1.00	1.00	1.00
		SIRS	5	5	5	5	5	1.00	1.00	1.00	1.00	1.00	1.00
$t(1)$	1	RDC-SIS	5	5	5	5	19	1.00	1.00	1.00	1.00	0.99	0.99
		DC-SIS	5	7	19	109	584	0.82	0.84	0.86	0.77	0.71	0.67
		SIS	35	467	806	916	966	0.20	0.21	0.21	0.20	0.16	0.10
		SIRS	5	5	5	6	25	1.00	1.00	0.99	0.99	0.98	0.98
	2	RDC-SIS	5	5	5	5	5	1.00	1.00	1.00	1.00	1.00	1.00
		DC-SIS	5	5	5	6	156	0.95	0.94	0.94	0.94	0.90	0.87
		SIS	5	90	494	865	981	0.44	0.43	0.44	0.42	0.30	0.23
		SIRS	5	5	5	5	5	1.00	1.00	1.00	1.00	1.00	1.00

where $\hat{b}_\lambda$ is the minimizer of (7) for given $\lambda = (\lambda_1, \lambda_2)$, d_1 is the number of nonparametric components and d_2 is the number of parametric components, both for the given λ.

Example 1. Consider three models:

Model 1: $Y = c\beta^T X + \sigma\varepsilon$,
Model 2: $Y = X_1 + 2X_2^2 + 3X_3^3 + 4X_4^4 + \varepsilon$,
Model 3: $Y = X_1^2 + (2 + sin(X_2))^2 + (1 + X_3)^{-3} + (X_4^2 + X_4 - 1)^{-1} + X_5 + \varepsilon$,

where $\beta = (1, 0.8, 0.6, 0.4, 0.2, 0, \ldots, 0)^T$ takes grid values and $\sigma^2 = 6.83$. Model 1 is adapted from [23]. We varied the constant c to control the signal-to-noise ratio. We choose $c = 1$ and 2, with the corresponding $R^2 = 50\%$ and 80%. The vector of covariates $X = (X_1, \ldots, X_p)$ was generated from the multivariate normal distribution with mean 0 and the covariance matrix $\Sigma = (\sigma_{ij})_{p\times p}$ with $\sigma_{ii} = 1$ and $\sigma_{ij} = 0.8^{|i-j|}$ for $i \neq j$. We considered three error ε distributions, $N(0,1)$, a t-student ($t(1)$) and a skew normal distribution (SN(0,1,2)). In this example, we set the sample size $n = 200$ and the total number of predictors $p = 1000$. We repeat each scenario 200 times and the results are given in Tables 1-3.

From Table 1, when the random error has a normal distribution, for both cases $c = 1$, and $c = 2$, all four screening methods perform quit well. However, it is consistently superior for the heavy-tailed error distribution. When the error has a t-distribution with one degree of freedom, however, the performance of DC-SIS nad SIS quickly deteriorates, while our method continues to perform well. DC-SIS do not work well in identifying active covariates. Compared with the SIRS method, the performances of the RDC-SIS and SIRS are equally good in all the considered scenarios; both of them deliver more satisfactory results than the DC-SIS and SIS procedures. The SIS has little chance to identify the important predictors.

Table 2 indicates that, for all types of distribution error, the performance of the proposed RDC-SIS is very well and outperform other methods. All P_j and S of the RDC-SIS is equal 1. Thus, all active predictors can perfectly be selected into the resulting model across all three different error distributions. We can see that the DC-SIS procedure is comparable to the RDC-SIS method for all types of distribution error.

In model 3, where the model has a more complex structure, RDC-SIS is effective in identifying the number of active variables for all types of errors; while DC-SIS and NIS fail to identify some important predictors. The NIS has no chance for to identify X_3 and little chance to identify other active covariates. RDC-SIS and SIRS have similar performances and are equally well. Both of them outperform the NIS and DC-SIS procedures.

Table 2. Five quantiles of minimum model size M, the empirical probability P_j and the proportion of S in Model 2.

ε	method	M					P				S
		5%	25%	50%	75%	95%	1	2	3	4	
$N(0,1)$	RDC-SIS	4	4	4	4	4	1.00	1.00	1.00	1.00	1.00
	DC-SIS	4	5	5	6	9	0.99	1.00	1.00	1.00	0.99
	NIS	5	5	6	11	102	0.93	1.00	1.00	1.00	0.93
	SIRS	4	4	5	19	299	1.00	1.00	1.00	0.85	0.85
$t(1)$	RDC-SIS	4	4	4	4	5	1.00	1.00	1.00	1.00	1.00
	DC-SIS	4	5	5	7	64	0.94	0.96	0.98	0.98	0.94
	NIS	5	6	11	82	864	0.65	0.81	0.85	0.89	0.55
	SIRS	4	4	7	34	468	0.99	1.00	1.00	0.78	0.69
SN	RDC-SIS	4	4	4	4	4	1.00	1.00	1.00	1.00	1.00
	DC-SIS	4	5	5	6	8	0.99	1.00	1.00	1.00	0.99
	NIS	5	5	6	11	104	0.92	1.00	1.00	1.00	0.92
	SIRS	4	4	5	23	345	1.00	1.00	1.00	0.87	0.87

Table 3. Five quantiles of minimum model size M, the empirical probability P_j and the proportion of S in Model 3.

ε	method	M					P					S
		5%	25%	50%	75%	95%	1	2	3	4	5	
$N(0,1)$	RDC-SIS	5	5	5	5	7	1.00	1.00	1.00	1.00	1.00	1.00
	DC-SIS	221	499	680	821	994	0.13	0.16	0.33	0.14	0.15	0.02
	NIS	576	776	877	954	994	0.07	0.09	0.00	0.06	0.09	0.00
	SIRS	5	5	5	5	9	1.00	1.00	1.00	1.00	1.00	1.00
$t(1)$	RDC-SIS	5	5	5	6	11	1.00	1.00	1.00	1.00	1.00	1.00
	DC-SIS	225	500	679	821	961	0.14	0.15	0.33	0.14	0.15	0.01
	NIS	576	776	877	950	994	0.07	0.09	0.00	0.06	0.09	0.00
	SIRS	5	5	5	6	15	1.00	1.00	1.00	1.00	1.00	1.00
SN	RDC-SIS	5	5	5	5	6	1.00	1.00	1.00	1.00	1.00	1.00
	DC-SIS	214	500	679	821	961	0.13	0.16	0.33	0.14	0.15	0.02
	NIS	576	776	877	954	994	0.07	0.09	0.00	0.06	0.09	0.00
	SIRS	5	5	5	5	8	1.00	1.00	1.00	1.00	1.00	1.00

Example 2. In this example, we first apply the RDC-SIS method to reduce dimensionality, and then fit a partial linear model (PLAM) where two penalty is used to simultaneously identify nonzero and linear components. We generated data from the model

$$Y = \sum_{j=1}^{p} f_j(X_j) + \varepsilon, \tag{12}$$

where $f_1(x) = 3sin(2\pi x)/(2 - sin(2\pi x))$, $f_2(x) = 6x(1-x)$, $f_3(x) = 2x$, $f_4(x) = x$, $f_5(x) = -x, f_j(x) = 0$ for $j > 5$. Thus the true number of nonparametric components is 2 and the true number of linear components is 3. To generate covariates, we let X_j be marginally standard normal with correlations given by $cov(X_i, X_j) = 0.5^{|i-j|}$. To illustrate the efficiency of our proposed method, the following two different error distributions are considered: standard normal distribution N(0, 1) and t-distribution with three degree of freedom, t_3, which is used to produce heavy-tailed distribution. We performed simulations with $n = 200, 400, p = 1000, 2000$ and the results are summarized in Table 4.

We used several criterion to measure the model identification performance: "NN": average number of nonlinear components selected; "NNT":average number of nonlinear components selected that are truly nonlinear; NL: average number of linear components selected; "NLT":average number of linear components selected that are truly linear. The numbers in parenthesis are the corresponding standard errors. The simulation results indicate that the proposed two-stage method is effective in estimating and identifying nonzero components as well as linear components from nonlinear ones simultaneously.

Table 4. Model identification results for Example 2.

p	n	Error	NN	NNT	NL	NLT
1000	200	N(0,1)	2.31(0.96)	2(0)	2.93(0.88)	2.74(0.32)
		$t(3)$	2.52(0.88)	1.97(0.12)	3.37(1.03)	2.71(0.71)
	400	N(0,1)	2.20(0.64)	2(0)	3.12(0.59)	2.96(0.26)
		$t(3)$	2.49(0.54)	1.99(0.08)	3.28(0.66)	2.74(0.75)
2000	200	N(0,1)	2.46(0.84)	2(0)	3.12(0.94)	2.84(0.64)
		$t(3)$	2.77(0.79)	1.98(0.17)	3.24(1.01)	2.76(0.83)
	400	N(0,1)	2.33(0.81)	2(0)	3.04(0.94)	2.92(0.64)
		$t(3)$	2.54(0.74)	1.99(0.06)	2.97(0.85)	2.81(0.63)

5. Data Analysis

To illustrate the usefulness of the suggested strategies for ultrahigh-dimensional data in the semiparametric regression model, we consider cardiomyopathy microarray data. This data set has attracted considerable attention and been systematically investigated by many researchers. They aim to identify the influential genes that affect the overexpression of a G protein-coupled receptor, called Ro1, in mice. The Ro1 expression level was measured for 30 specimens, and the predictors, the genetic expression levels, were obtained for $p = 6319$ genes.

This data set has been studied by many researchers. Hall and Miller [8] showed that both genes Msa.2877.0 and Msa.1166.0 are particularly important using the generalized correlation. Li et al. [10] used the DC-SIS procedure that ranks two genes, Msa.2134.0 and Msa.2877.0, at the top. Li et al. [11] showed that Msa.1166.0 and Msa.7019.0 are particularly important using the RRCS procedure. The NIS procedure in [5] ranks two genes, labeled as Msa.2877.0 and Msa.1166.0, at the top. The RDC-SIS procedure ranks two genes, Msa.2134.0 and Msa.2877.0, in the top, which is similar to DC-SIS and SIRS of Zhu et al. [23]. The scatter plots of Y versus these two gene expression levels with cubic spline fit curves in Figure 1 indicate clearly the existence of nonlinear patterns. It can be noted that the distance correlation has the advantage that it can detect nonlinear relationships which are ignored by marginal correlation. Our proposed RDC-SIS procedure shows the advantage that it detected two important genes having nonlinear relationships with Ro1, which might be ignored by some other methods.

A natural question arises: which screening procedure does perform better in terms of ranking? Following Li et al.[11], to compare the performance of these procedures, we fit three different additive models as follows:

$$Y = g_{k1}(X_{k1}) + g_{k2}(X_{k2}) + \varepsilon_k, \qquad k = 1, 2, 3,$$

where X_{k1} and X_{k2} are the top two genes, g_{k1} and g_{k2} are two unknown link functions, ε is an error term. We fit g_{k1} and g_{k2} by using the "gam" function in the R "mgcv" package, where "gam" can be used to fit a generalized additive model to data. We also measure the performance of goodness of fit by the adjusted R^2 values and the explained deviance, where deviance implies the proportion of the null deviance explained by the proposed model, with a larger value indicating better performance.

The RDC-SIS, corresponding to $k = 1$, regards Msa.2134.0 and Msa.2877.0 as the two predictors, while the generalized correlation ranking proposed by Hall and Miller [8], corresponding to $k = 2$, regards Msa.2877.0 and Msa.1166.0 as predictors in the above model. Also the RRCS proposed by Li et al.[11], corresponding to $k = 3$, regards Msa.1166.0 and Msa.5758.0 as the two predictors. The RDC-SIS method clearly achieves better performance with the adjusted R^2 of 96.8% and the deviance explained of 98.3%, in contrast to the adjusted R^2 of 84.5% and the deviance explained of 86.6% for the generalized correlation ranking method, and the adjusted R^2 of 77.9% and the deviance explained of 81.5% for the RRCS method.

From the above, we can conclude that the RDC-SIS is an efficient method for dimension reduction in ultrahigh dimensional data. For variable selection and structure identification in cardiomyopathy dataset, we first applied RDC-SIS to reduce the covariate dimension to the size of $3[n/\log(n)] = 24$. After cleaning, two-fold SCAD penalty was used to identify parametric and non parametric parts. We have identified 4 genes of linear effects and 13 genes of nonlinear effects. The genes of linear effects are Msa.1920.0, Msa.2877.0, Msa.5595.0 and Msa.741.0.

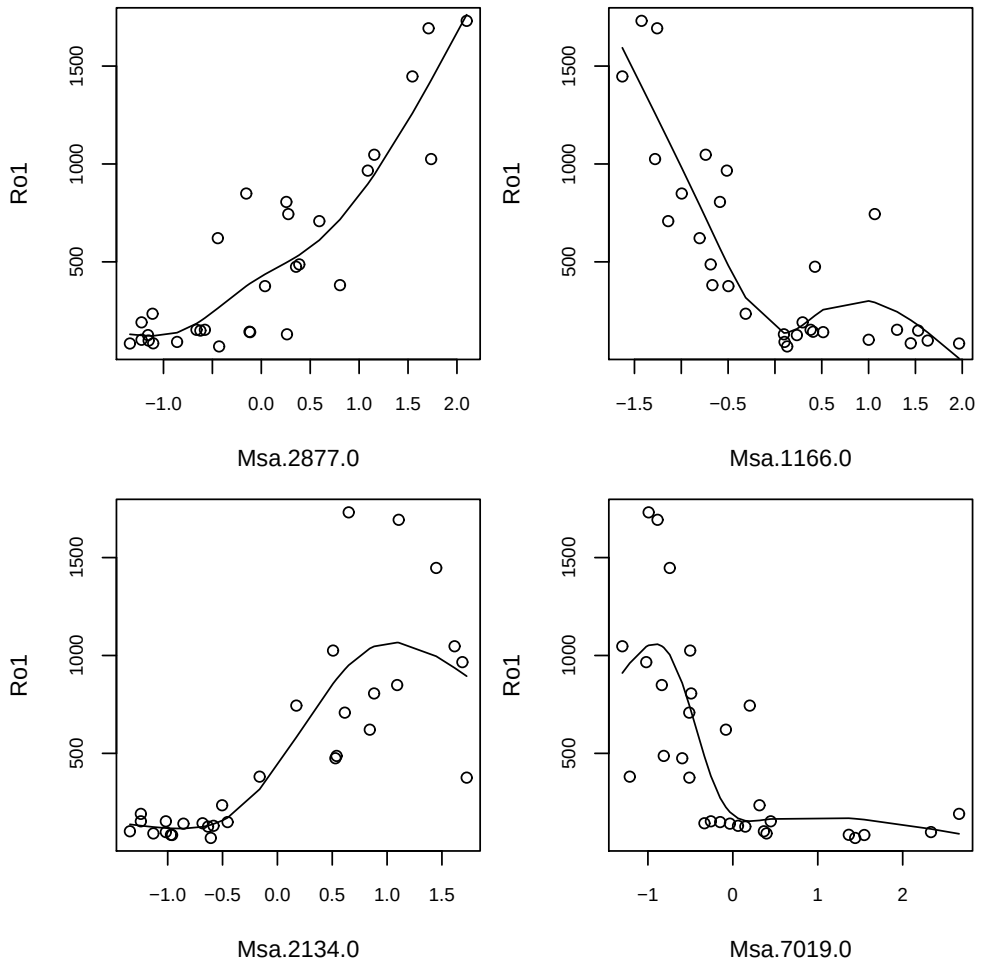

Figure 1. The scatter plots and corresponding cubic-spline fit curves of the relationship between the important genes (Msa.2877.0, Msa.1166.0, Msa.2134.0, Msa.7019.0) and the outcome (Ro1) based on $n = 30$ specimens.

The genes of nonlinear effects are Msa.10180.0, Msa.963.0, Msa.5727.0, Msa.1166.0, Msa.15442.0, Msa.1590.0, Msa.2134.0, Msa.2400.0, Msa.7019.0, Msa.28021.0, Msa.5583.0, Msa.26025.0 and Msa.15405.0.

6. Conclusion

In this paper, we developed a two stage procedure for variable selection and structure identification in ultrahigh-dimensional partially linear additive models. In the first stage, a sure independence procedure was used to dimension reduction from ultrahigh to moderate scale. This procedure ranks covariates trough their distance correlation with marginal distribution function of response variable. The proposed methodology has been supported by numeric examples and a real data analysis. This method is model-free and is robust for skew or heavy tailed response variable. In second stage, in order to distinguish linear and nonlinear parts and to identify insignificant covariates simultaneously, we used a double penalization based procedure.

Acknowledgments

The authors would like to thank two anonymous reviewers for their constructive comments which improved the presentation of paper. Third author M. Arashi's work is based on the research supported in part by the National Research Foundation of South Africa (Grant NO. 109214).

REFERENCES

1. J. Du, G. Li, and H. Peng, *Variable selection for semiparametric partially linear Covariate-Adjusted Regression Models*, Comm. Statist. Theo. Meth., vol. 44, no. 3, pp. 2809–2826, 2015.
2. J. Fan, and R. Li, *Variable selection via nonconcave penalized likelihood and its oracle properties,* J. Amer. Statist. Assoc., vol. 96, pp. 1348–1360, 2001.
3. J. Fan, and J. Lv, *Sure independence screening for ultrahigh dimensional feature space,* J. R. Statist. Soc. Ser. B Stat. Meth., vol. 70, no. 5, pp. 849-911, 2008.
4. J. Fan, R.J. Samworth, and Y. Wu, *Ultrahigh dimensional feature selection: beyond the linear model,* J. Mach. Learn. Res., vol. 10, pp. 1829–1853, 2009.
5. J. Fan, Y. Feng, and R. Song, *Nonparametric independence screening in sparse ultrahigh-dimensional additive models,* J. Amer. Statist. Assoc., vol. 106, pp. 544–557, 2011.
6. J. Fan, and R. Song, *Sure independence screening in generalized linear models with NP-dimensionality,* Ann. Statist., vol.6, pp. 3567-3604, 2010.
7. J. Guo, M. Tang, M. Tian, and K. Zhu, *Variable selection in high-dimensional partially linear additive models for composite quantile regression,* Comp. Statist. Data Anal., vol. 65, pp. 56–67, 2013.
8. P. Hall, H. Miller, *Using generalized correlation to effect variable selection in very high dimensional problems,* J. Computnl Graph. Statist., vol. 18, pp. 533C-550, 2009.
9. J. Huang, F. Wei, and S. Ma, *Semiparametric regression pursuit.* Statist. Sinica , vol. 22, pp. 1403–1426, 2012.
10. R.Z. Li, W. Zhong, and L.P. Zhu, *Feature screening via distance correlation learning,* J. Amer. Statist. Assoc., vol. 107, pp. 1129–1139, 2012.
11. G. Li, H. Peng, J. Zhang, and J. Zhu, *Robust rank correlation based screening,* Ann. Statist. vol. 40, pp. 1846C-1877, 2012.
12. H. Lian, *Variable selection in high-dimensional partly linear additive models,* J. Nonparametric Statist., vol. 24, no. 4, pp. 825–839, 2012.
13. H. Lian, *Shrinkage estimation for identification of linear components in additive models,* Statist. Prob. Lett., vol. 82, pp. 225–231, 2012.
14. H. Lian, X. Chen, and JY. Yang, *Identification of partially linear structure in additive models with an application to gene expression prediction from sequences,* Biometrics, vol. 68, pp. 437-C445, 2012.
15. H. Lian, H. Liang, and D. Ruppert, *Separation of covariates into nonparametric and parametric parts in high-dimensional partially linear additive models,* Statistica Sinica, vol. 25, pp. 591–607, 2015.
16. X. Liu, L. Wang, and H. Liang, *Estimation and variable selection for semiparametric additive partial linear models,* Statist. Sinica., vol. 21, pp. 1225–1248, 2011.
17. J. Lv, H. Yang, and C. Guo, *Variable selection in partially linear additive models for modal regression,* Comm. Statist. Sim. Comp., DOI: 10.1080/03610918.2016.1171346, 2016.
18. G. J. Szekely, M. L. Rizzo, and N. K. Bakirov, *Measuring and testing dependence by correlation of distances,* Annals of Statistics, vol. 35, pp. 2769–2794, 2007.
19. R. Tibshirani, *Regression shrinkage and selection via the lasso,* J. Royal. Statist. Soc. Ser. B, vol. 58, pp. 267–288, 1996.
20. C. H. Zhang, *Nearly unbiased variable selection under the minimax concave penalty,* Ann. Statist., vol. 83, pp. 894–942, 2010.
21. H. H. Zhang, G. Cheng, and Y. Liu, *Linear or Nonlinear? Automatic structure discovery for partially linear models,* J. Amer. Statist. Assoc., vol. 106, pp. 1099-1112, 2011.
22. S. D. Zhao, and Y. Li, *Principled sure independence screening for Cox models with ultrahigh-dimensional covariates,* J. Mult. Anal., vol. 105, pp. 397–411, 2012.
23. L. P. Zhu, L. Li, R. Li, and L. X. Zhu, *Model-free feature screening for ultrahigh dimensional data,* J. Amer. Statist. Assoc., vol. 106, pp. 1464–1475, 2011.