Journal of Mathematical Modeling
Vol. 8, No. 4, 2020, pp. 363–376

JMM

Partial correlation screening for varying coefficient models

Mohammad Kazemi[*]
Department of Statistics, Faculty of Mathematical Sciences,
Shahrood University of Technology, Shahrood, Iran
Email(s): m.kazemie64@yahoo.com

Abstract. In this paper, we propose a two-stage approach for feature selection in varying coefficient models with ultra-high-dimensional predictors. Specifically, we first employ partial correlation coefficient for screening, and then penalized rank regression is applied for dimension-reduced varying coefficient models to further select important predictors and estimate the coefficient functions. Simulation studies are carried out to examine the performance of proposed approach. We also illustrate it by a real data example.

Keywords: Big data, feature screening, partial correlation, rank regression.
AMS Subject Classification 2010: 62H12, 62H20, 62J07, 62G08.

1 Introduction

With the remarkable development of modern technology, including computing power and storage, big data of unprecedented size could be collected at a relatively low cost and have appeared in many areas of advanced scientific research ranging from genomic and health science to machine learning and economics. The collected data frequently has an ultra-high dimensionality p that is allowed to diverge at nonpolynomial (NP) rate with the sample size n, namely $\log(p) = O(n^{\rho})$ for some $\rho > 0$. For example, in biomedical research such as genomewide association studies for some mental diseases, millions of SNPs are potential covariates. In such a "large p, small n"

[*] Corresponding author.

Received: 11 February 2020 / Revised: 11 May 2020 / Accepted: 23 May 2020.
DOI: 10.22124/jmm.2020.15692.1379

problem, it is often assumed that only a small number of these covariates contribute to the response, which is called the sparsity principle in the literature. One basic and challenging task is to identify these true important predictors that are associated with the response. When the dimensionality p is ultra-high, the traditional regularized variable selection approaches become ineffective, due to the simultaneous challenges of computational expediency, statistical accuracy and algorithmic stability (see [6]).

To address ultra-high dimensional data with stable computation and accurate selection, Fan and Lv [3] proposed the sure independent screening (SIS) procedure for ultra-high dimensional linear model which utilized the Pearson correlation to rank the importance of each predictor. With the purpose of handling more complex real data, many authors developed the SIS procedure and applied it to various statistical models, such as generalized linear models (see [6,7]), nonparametric additive models (see [1,10,11]) and varying coefficient models (see [5, 14, 16, 17]). These feature screening procedures are based on specialized model and perform well when the underlying model assumptions are correct. Since specifying a correct model for ultra-high dimensional data may be challenging, model-free sure screening procedures are appealing and have been developed. Zhu et al. [21] proposed a sure independent ranking and screening (SIRS) procedure for ultra-high dimensional data in the framework of the general multi-index models. Li et al. [15] proposed a model-free SIS procedure based on the distance correlation. These model-free methods are useful selections when nothing can be known about the underlying model. If we can obtain some of the characteristics of the model based on the information provided by the research background, there may be a better way to take account of those characteristics into the ultra-high dimensional data analysis.

It is well known that nonparametric models are flexible enough to reduce modeling biases. However, they suffer from the so-called "curse of dimensionality". A remarkable simple and powerful nonparametric model for dimensionality reductions is the varying-coefficient model,

$$Y = \boldsymbol{X}^T \boldsymbol{\beta}(U) + \varepsilon, \tag{1}$$

where $\boldsymbol{X} = (X_1, \ldots, X_p)^T$ is the vector of predictors, Y is the response, U is an additional auxiliary variable or index variable, $\boldsymbol{\beta}\,(U)$ consists of p unknown smooth functions of U, and ε is the random noise with conditional mean 0 and finite conditional variance. An intercept term (i.e., $X_0 \equiv 1$) can be introduced if necessary. This model assumes that the variables in the predictor vector $\boldsymbol{X}$ enter the model linearly, meanwhile it allows regression coefficient functions to very smoothly with the index variable.

The model retains general nonparametric characteristics and allows the nonlinear interactions between the index variable U and the predictors. It arises frequently from economics, finance, politics, epidemiology, medical science, ecology, among others. For an overview, see [4].

Some feature selection methods have been developed for varying coefficient models with low-dimensional covariates in literature. Li and Liang [13] proposed a generalized likelihood ratio test to select significant variables with varying effects. Wang et al. [19] developed a regularized estimation procedure based on the basis function approximations and SCAD penalty to simultaneously select significant variables and estimate the nonzero smooth coefficient functions. Wang and Xia [18] proposed a shrinkage method integrating local polynomial regression techniques and LASSO. Nevertheless, these feature selection procedures were developed for the varying coefficient models with fixed-dimensional covariates. However, the above methods may not perform well in ultra-high dimension due to aforementioned challenges. Liu et al. [16] developed a screening method based on conditional correlation (CC-SIS) which relies on kernel estimation.

In this paper, we propose a two-stage method for feature selection and estimation in ultra-high dimensional varying coefficient models. In the first step, we develop a marginal utility for feature screening based on partial correlation to reduce the dimensionality. In the second step, we apply penalized rank regression to perform feature selection and estimation simultaneously. Simulation studies are conducted to evaluate our method under different circumstances.

This paper differs from Liu et al. [16] in several ways. They focused on the two-stage approach consists of (a) reducing the ultra-high dimensionality by using the conditional correlation between the response and predictors and (b) applying KLASSO method (Wang and Xia [18]), for dimension-reduced varying coefficient models, which is very sensitive to outliers and heavy-tailed error distributions. Our work can be regarded as an extension of Liu et al. [16], with differences and our contributions highlighted as follows. In first stage, we develop a new screening procedure using partial correlation instead of conditional correlation. The key benefit of our screening procedure is that the computation is much faster than conditional correlation which relies on kernel estimation. In second stage, we consider the problem of simultaneous variable selection and coefficient estimation in varying coefficient models by using spline approximation and penalization approach based on rank regression, which is robust with respect to heavy tailed errors or outliers in the response.

The remainder of this paper is organized as follows. In Section 2, we introduce the two-step procedure for feature selection and estimation. In Section 3, we consider the computation algorithm of the proposed estimator. Simulation studies are carried out in Section 4 to assess the performance of the proposed method and to compare it with some existing methods. In addition, a data set is used as an illustration of varying coefficient models. Finally, we conclude the paper in Section 5

2 Methodology

2.1 Step 1: Feature screening

The partial correlation $\rho_{u,v.w}$ of U and V after controlling for W is by definition the ordinary correlation between the "residual" variables $U_r = U - \alpha - \beta(W - \mu)$ and $V_r = V - \gamma - \delta(W - \mu)$ where $\mu = E(W)$, $\alpha = E(U)$, $\beta = Cov(U, W)/Var(W)$, $\gamma = E(V)$ and $\delta = Cov(V, W)/Var(W)$. These values of α, β, γ and δ are those which minimize $E_{U,W}\{U - \alpha - \beta(W - \mu)\}^2$ and $E_{V,W}\{V - \gamma - \delta(W - \mu)\}^2$, and so U_r and V_r are the variables U and V after their linear dependence on W has been removed. The above definition of partial correlation implies the definitions of partial variances and partial covariance of U and V as the ordinary variances and covariance of U_r and V_r. Mathematically, the partial correlation coefficient between U and V, when controlling for W, is computed using the following formula

$$\rho_{U,V.W} = \frac{\rho_{UV} - \rho_{UW}\rho_{VW}}{\sqrt{1 - \rho^2_{UW}}\sqrt{1 - \rho^2_{VW}}}, \tag{2}$$

in terms of the pairwise Pearson correlations between U, V and W. Indeed, the sample partial correlation can be calculated easily by replacing the sample pairwise Pearson correlations in (2). It can be reached by the R package `ppcor`.

It is noteworthy that the covariate U is not a conditional variable but an additional variable, i.e., the partial correlation is not in general equal to conditional correlation. The two correlations are equal when the conditional correlation of U and V given W is free of W. By the way, the conditional correlation of U and V given W is not necessarily free of W hence can not in general equal to partial correlation of U and V on W.

In multiple linear regression, the partial correlation measures the strength of the linear relationship between response and one of the predictors after "adjusting" for relationships involving all the other variables. This fact motivated us to develop a feature screening method by ranking the magnitude of partial correlation between response and each predictor X_j controlling for index variable U.

Define the true model index set M and its complement M^c by

$$M = \{1 \leq j \leq p : \beta_j(u) \neq 0 \;\text{ for some } u \in \mathbb{U}\},$$
$$M^c = \{1 \leq j \leq p : \beta_j(u) = 0 \text{ for all } u \in \mathbb{U}\},$$

and the marginal utility for feature screening as

$$\omega_j = \rho^2_{X_j,Y.U}, \qquad j = 1, \ldots, p. \tag{3}$$

Thus the sample estimate of ω_j is

$$\widehat{\omega}_j = \widehat{\rho}^2_{X_j,Y.U}, \qquad j = 1, \ldots, p. \tag{4}$$

Hence, by ranking the $\widehat{\omega}_j$ from largest to smallest, the important predictors are determined by the estimated active set

$$\widehat{M} = \{j : 1 \leq j \leq p \;\text{ s.t. } \widehat{\omega}_j \text{ ranks among the first } d\},$$

where the submodel size d is taken to be smaller than the sample size n. This procedure reduces the dimensionality from p to a possibly much smaller space with model size $d = |\hat{A}|$. Fan and Lv [3] suggested setting $d = [n/\log(n)]$, where $[a]$ refers to the integer part of a.

2.2 Step 2: Post-screening feature selection

In this step, the penalized rank regression is applied to further select important features and estimate the coefficient function $\boldsymbol{\beta}(u)$ in model (1). Suppose that each $\beta_j(u)$, $j = 1, \ldots, p$ can be approximated by B-spline function, that is

$$\beta_j(u) \approx \sum_{k=1}^{K} \gamma_{jk} B_{jk}(u)\;, \qquad j = 1, \ldots, p, \tag{5}$$

where $\{B_{jk}(u)\,,\; k = 1, \ldots, K\}$ denote a B-spline basis from the collection of spline functions of a fixed degree and knot sequence on interval $[0, 1]$ and K is the number of B-spline basis. Suppose that $\boldsymbol{S}_n = \{Y_i,\; \boldsymbol{X}_i,\; U_i\}_{i=1}^n$ is an independent and identically distributed sample from $\{Y, \boldsymbol{X}, U\}$. Following the approximation (5), model (1) becomes

$$Y_i = \sum_{j=1}^{p} \sum_{k=1}^{K} \gamma_{jk} B_{jk}(U_i) X_{ij} + \varepsilon_i, \qquad i = 1, \ldots, n. \tag{6}$$

Let $\boldsymbol{\pi}_j(.) = (B_{j1}(.), \ldots, B_{jK}(.))^T$ be a set of normalized B-spline basis functions. Define $\boldsymbol{\Pi}(U, \boldsymbol{X}) = \left(X_1 \boldsymbol{\pi}_1(U)^T, \ldots, X_p \boldsymbol{\pi}_p(U)^T\right)^T$, $\boldsymbol{\Pi}_i = \boldsymbol{\Pi}(U_i, \boldsymbol{X_i})$,

$\boldsymbol{\pi_{ij}} = \boldsymbol{\pi_j}(U_i), i = 1, \dots, n$ and $\boldsymbol{\pi}(.) = \left(\boldsymbol{\pi}_1(.)^T, \dots, \boldsymbol{\pi}_p(.)^T\right)^T$. So (6) can be written as

$$Y_i \approx \boldsymbol{\mathit{\Pi}}_i^T \boldsymbol{\gamma} + \varepsilon_i, \qquad i = 1, \dots, n. \tag{7}$$

where $\boldsymbol{\gamma}^T = \left(\boldsymbol{\gamma}_1^T, \dots, \boldsymbol{\gamma}_p^T\right)^T$. Based on the above approximation, we obtain the residual at U_i,

$$e_i = Y_i - \boldsymbol{\mathit{\Pi}}_i^T \boldsymbol{\gamma}, \qquad i = 1, \dots, n.$$

By Jaeckel [9], the rank regression is to minimize the dispersion of the residuals. So the parameter $\boldsymbol{\gamma}$ in the basis expansion can be estimated by minimizing

$$Q = \frac{1}{n} \sum_{i<j} |e_i - e_j|. \tag{8}$$

We denote the minimizer of (8) as $\widehat{\boldsymbol{\gamma}}^T = \left(\widehat{\gamma}_1^T \dots \widehat{\gamma}_p^T\right)^T$, where $\boldsymbol{\gamma}_j = (\gamma_{j1}, \dots, \gamma_{jK})^T$.

Suppose that some covariates are irrelevant in the regression, but it is unknown to us which ones are important covariates whose corresponding coefficients are nonzero, and which are unimportant covariates whose corresponding coefficients are zero. However, equation (8) does not have properties of selecting notable predictors, so we turn to the penalized estimation. Let $p_\lambda(.)$ is the SCAD penalty function proposed by Fan and Li [2], and it is defined in terms of its first order derivative as follows

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

where $a > 2$ and λ is a nonnegative penalty parameter and governs variable selection or sparsity of the model. We use $a = 3.7$ as suggested in [2]. Let $\boldsymbol{R}_j$ be a $K \times K$ matrix with entries $r_{kk'} = \int B_{jk}(u) B_{jk'}(u) du$ and $\parallel \gamma_j \parallel_{\boldsymbol{R}_j} = (\boldsymbol{\gamma}_j^T \boldsymbol{R}_j \gamma_j)$. Our main goal is to identify the insignificant components (i.e., $\beta_j(u) \equiv 0$). This can be achieved by shrinking $\parallel \gamma_j \parallel_{\boldsymbol{R}_j}$to zero. We define the proposed estimator $\widehat{\boldsymbol{\gamma}}$ as the minimizer of the following penalized objective function

$$L_n(\gamma) = \frac{1}{n} \sum_{i<j} |e_i - e_j| + \lambda \sum_{j=1}^{p} p_\lambda \left(\parallel \gamma_j \parallel_{\boldsymbol{R}_j} \right). \tag{9}$$

In the proposed procedure, every $\beta_j(u)$ is characterized by a spline coefficient vector $\boldsymbol{\gamma}_j$, $\widehat{\gamma}_j(u) = \boldsymbol{\pi_j}(u)^T \widehat{\boldsymbol{\gamma}}_{\boldsymbol{j}}$, which can be treated as a group. However, different from the SCAD penalization in [2], all components within the same group j receive the same group-specific penalty.

3 Computation algorithm

Leng [12] pointed that the objective function Q in (8) can be seen as Jaeckel's [9] rank dispersion function for Wilcoxon scores. So the first term in (9) can be approximated by

$$\frac{1}{n}\sum_{i<j}|e_i - e_j| \approx \frac{1}{n}\sum_{i<j}\omega_i\,|e_i - \xi|^2,$$

where ξ is the median of $\{e_i\}_{i=1}^{n}$ and

$$\omega_i = \begin{cases} \frac{\frac{R(e_i)}{n+1}-\frac{1}{2}}{e_i-\xi}, & e_i \neq \xi, \\ 0, & \text{otherwise}, \end{cases}$$

where $R(e_i)$ is the rank of e_i among $e_1, \ldots, e_n$. Following [2], we use an iterative local quadratic approximation algorithm to find the minimum of (8). Using a simple Taylor expansion, given an initial estimate $\widetilde{\gamma}_l$ from objective function (8) (equivalently given $\widetilde{\beta}_l$), the weights $\widetilde{\omega}_i$ and the median of residuals, ξ, can be accordingly obtained. We approximate the regularization terms by

$$p_\lambda\left(\|\,\boldsymbol{\gamma}_j\,\|_{\boldsymbol{R}_j}\right) \approx p_\lambda\left(\|\,\widetilde{\boldsymbol{\gamma}}_j\|_{\boldsymbol{R}_j}\right) + \frac{1}{2}\frac{p'_\lambda\left(\|\,\widetilde{\boldsymbol{\gamma}}_j\|_{\boldsymbol{R}_j}\right)}{\|\,\widetilde{\boldsymbol{\gamma}}_j\|_{\boldsymbol{R}_j}}\left(\|\,\boldsymbol{\gamma}_j\,\|^2_{\boldsymbol{R}_j} - \|\,\widetilde{\boldsymbol{\gamma}}_l\|^2_{\boldsymbol{R}_j}\right).$$

Therefore, (9) can be locally approximated (except for constant terms) by

$$L_n(\gamma) \approx \left(\boldsymbol{S} - \boldsymbol{\Pi}'\boldsymbol{\gamma}\right)^T \widetilde{\boldsymbol{W}}\left(\boldsymbol{S} - \boldsymbol{\Pi}'\boldsymbol{\gamma}\right) + \frac{n}{2}\left(\boldsymbol{\gamma}^{\boldsymbol{T}}\boldsymbol{\Omega}\boldsymbol{\gamma}\right), \tag{10}$$

where $\boldsymbol{S}{=}\boldsymbol{Y}{-}\boldsymbol{I}_{n\times 1}$, and

$$\begin{aligned} \boldsymbol{\Omega} &= \operatorname{diag}\left(p'_\lambda\left(\|\,\widetilde{\boldsymbol{\gamma}}_1\|_{\boldsymbol{R}_1}\right)\boldsymbol{R}_1, \ldots, p'_\lambda\left(\|\,\widetilde{\boldsymbol{\gamma}}_p\|_{\boldsymbol{R}_p}\right)\boldsymbol{R}_p\right), \\ \widetilde{\boldsymbol{W}} &= \operatorname{diag}\left(\widetilde{\omega}_1, \ldots, \widetilde{\omega}_p\right). \end{aligned}$$

The algorithm repeatedly solves the minimization criterion (10) and updates $\boldsymbol{\gamma}^{(m)}$ to $\boldsymbol{\gamma}^{m+1}$, $m = 1, 2, \ldots$ until convergence.

4 Numerical studies

4.1 Simulation

For brevity, we refer our screening approach as partial correlation sure independence screening (PC-SIS). In this section, two simulation examples including different varying coefficient models with various scenarios are presented. The first example are allocated to our proposed screening procedure, while in the second one, the capability of the two-stage method is examined. In the former case, the finite sample performance of the PC-SIS is compared with the existing competitors, such as the SIS [3], DC-SIS [15], SIRS [21] and CC-SIS [16]. Throughout our experiments, we set the total number of predictors $p = 1000$, and the covariates u and $\boldsymbol{x} = (x_1, \ldots, x_p)^T$ are generated as follows.

First draw u^* and $\boldsymbol{x}$ from $(u^*, \boldsymbol{x}) \sim N(\mathbf{0}, \Sigma)$, where Σ is a $(p+1) \times (p+1)$ covariance matrix with element $\sigma_{ij} = \rho^{|i-j|},\ i, j = 1, \ldots, p+1$. Then take $u = \Phi(u^*)$, where Φ is the cumulative distribution function of the standard normal distribution. Thus, u follows a uniform distribution $U(0,1)$ and is correlated with $\boldsymbol{x}$, and all the predictors $(x_1, \ldots, x_p)$ are correlated with each other. We consider $\rho = 0, 0.5, 0.8$ for uncorrelated, correlated and high correlated situation among $(u^*, \boldsymbol{x})$. The random error ε is drawn from $N(0,1)$. In our simulation, we consider $d = [n/\log(n)]$.

To implement the two-stage procedure described in this paper, we need to choose some parameters including the spline order, the number of basis K, as well as the regularization parameters λ. We set the spline order to be $q = 4$, which means that cubic splines are used in all numerical examples. For the number of basis functions K, we fixed $K = 6$. This strategy is similar to that commonly used in functional smoothing/functional data analysis literature where the number of knots is chosen to be sufficiently large so that approximation error is small, and the overfitting can be effectively controlled by the penalization terms. To select the regularization parameter λ, we use the BIC criteria.

Example 1. We generate data from the following varying coefficient model.

$$\beta_1(u) = 5\log\left(2 + 3u^2\right), \quad \beta_3(u) = 5\tan(u), \quad \beta_5(u) = 2(u^2+1),$$
$$\beta_7(u) = 2 + u, \quad \beta_9(u) = 3.$$

So the true model index set in this sample is $M = \{1, 3, 5, 7, 9\}$. We set the sample size $n = 200$ and evaluate the performance through the three criteria: "R_j" is the average rank and standard error for an individual true predictor X_j based on 300 repeats; "p_j" is the proportion of submodels

Table 1: The value of R_j of true predictors for Example 1.

ρ	method	R_1	R_3	R_5	R_7	R_9
0	PC-SIS	1.00(0.00)	4.13(11.98)	2.43(0.96)	45.04(32.53)	6.24(27.83)
	SIS	1.00(0.00)	6.19(16.98)	2.43(0.93)	65.26(45.33)	8.40(25.58)
	DC-SIS	1.00(0.00)	4.48(5.00)	2.64(1.12)	52.15(19.65)	5.35(11.44)
	SIRS	1.00(0.00)	4.75(9.50)	2.71(1.21)	41.50(31.45)	4.53(6.90)
	CC-SIS	1.00(0.00)	3.73(5.88)	2.92(1.96)	39.70(98.18)	4.47(8.18)
0.5	PC-SIS	1.02(0.41)	2.98(1.05)	2.55(0.99)	6.42(5.48)	7.57(4.81)
	SIS	1.01(0.17)	2.88(0.86)	3.30(1.03)	9.17(21.94)	9.53(13.10)
	DC-SIS	1.01(0.12)	2.74(0.87)	3.36(1.04)	7.69(8.96)	7.17(2.52)
	SIRS	1.01(0.12)	2.71(0.85)	3.37(1.16)	7.24(6.16)	6.90(2.12)
	CC-SIS	1.03(0.23)	2.71(0.85)	2.86(1.05)	6.49(1.81)	6.51 (1.78)

with size d containing one true X_j in the 300 simulations; "P_a" is the proportion of submodels with size d containing all the true predictors in the 300 simulations.

Table 1 reports the simulation results for average rank and the standard deviation of R_j (the number in parentheses). From Table 1, we can see that SIS is less accurate than DC-SIS, SIRS, CC-SIS and PC-SIS, but our proposed PC-SIS is comparable to the others. It is observed that these four screening methods perform well in most cases.

In addition, Table 2 shows the results of the proportion of submodels in which its performance also indicate the adequacy of PC-SIS. From the simulation results, we can see that DC-SIS, SIRS, CC-SIS and PC-SIS provide nearly the same powerful results. In summary, the performance of CC-SIS is slightly better than others, but the cost of computation is high due to the kernel regression estimate of the conditional correlation. Hence we can conclude that PC-SIS is a competitive with the existing screening methods such as the SIS, DC-SIS, SIRS and CC-SIS.

Example 2. We generate data from two following models.

Model 1. Varying coefficient model with five nonzero varying coefficients and the true model index set $M = \{2, 100, 400, 600, 1000\}$ as follows

$$\beta_2(u) = 2I(u > 0.4), \quad \beta_{100}(u) = 1 + u, \quad \beta_{400}(u) = (2 - 3u)^2,$$
$$\beta_{600}(u) = 2\sin(2\pi u), \quad \beta_{1000}(u) = \mathrm{e}^{u/(u+1)}.$$

Model 2. Partial linear varying coefficient model that the number of con-

Table 2: The proportion value of p_j and p_a for Example 1.

ρ	method	p_1	p_3	p_5	p_7	p_9	p_a
0	PC-SIS	1.000	0.980	1.000	0.780	0.982	0.727
	SIS	1.000	0.983	1.000	0.710	0.953	0.673
	DC-SIS	1.000	0.993	1.000	0.733	0.980	0.716
	SIRS	1.000	0.990	1.000	0.800	0.983	0.783
	CC-SIS	1.000	0.996	1.000	0.806	0.993	0.790
0.5	PC-SIS	1.000	1.000	1.000	0.997	0.996	0.987
	SIS	1.000	1.000	1.000	0.986	0.983	0.970
	DC-SIS	1.000	1.000	1.000	0.986	1.000	0.986
	SIRS	1.000	1.000	1.000	0.990	1.000	0.990
	CC-SIS	1.000	1.000	1.000	1.000	1.000	1.000

stant coefficient is 3 and two coefficients are varying

$$\beta_1\ (u) = 2\sin(2\pi u), \quad \beta_2(u) = 8u(1-u), \quad \beta_3(u) = 2.5,$$
$$\beta_4(u) = 1, \quad \beta_5(u) = 1.5.$$

The true model index set in this model is $M = \{1, 2, 3, 4, 5\}$. We used several criterion to measure the feature selection performance: “NS”: Average number of nonzero estimated components; “AS” is the percentage of occasions on which all the correct variables are included in the selected model; “ES” is the frequency of exactly selecting all true variables and nothing else; “MS” is the percentage of occasions on which some correct variables are missed; “OS” is the frequency of exactly one false variable selected, “RASE” is the square root of average squared error

$$RASE = \left\{ \frac{1}{n_{grid}} \sum_{i=1}^{n_{grid}} \sum_{j=1}^{p} \left(\widehat{\beta}_j\left(u_i\right) - \beta_j\left(u_i\right)\right)^2 \right\}^{\frac{1}{2}}, \tag{11}$$

where $u_1, \ldots, u_{ngrid}$ are the grid points at which the functions $\widehat{\beta}_j\left(u_i\right)$ are evaluated. We report the results based on 200 simulation runs in Table 3.

From Table 3, we observe that for each model, in cases where ρ is large and n is small, the two-step procedure tends to be too greedy, missing some true variables. The small values for MS indicates extremely low false negative rates, and 0% values for OS show zero false positive rates of feature selection. In addition, the RASE of the estimators are very small. Overall, the proposed method has a satisfactory feature selection performance, except in case where the covariates are high correlated ($\rho = 0.8$).

Table 3: Feature selection performance of the two-step procedure.

Model	n	ρ	NS	AS	ES	MS	OS	RASE
Model 1	200	0.5	4.93	93%	93%	7%	0%	0.5661
		0.8	4.54	53%	53%	47%	0%	1.8682
	400	0.5	5	100%	100%	0%	0%	0.2277
		0.8	4.98	97%	97%	3%	0%	0.3175
Model 2	200	0.5	5	100%	100%	0%	0%	0.4087
		0.8	4.47	56%	56%	44%	0%	1.5881
	400	0.5	5	100%	100%	0%	0%	0.1943
		0.8	4.97	97%	97%	3%	0%	0.4621

4.2 Application to real data

We illustrate the performance of the newly proposed method through a real data analysis on Boston Housing Data (see [8]). This dataset contains housing data for 506 census tracts of Boston from the 1970 census. Most empirical results for the housing value equation are based on a common specification,

$$\begin{aligned}\log(MV) = {} & \beta_0 + \beta_1 RM^2 + \beta_2 AGE + \beta_3 \log(DIS) + \beta_4 \log(RAD) + \beta_5 TAX \\ & + \beta_6 PTRATIO + \beta_7 (B - 0.63)^2 + \beta_8 \log(LSTAT) + \beta_9 CRIM \\ & + \beta_{10} ZN + \beta_{11} INDUS + \beta_{12} CHAS + \beta_{13} NOX^2 + \varepsilon,\end{aligned}$$

where the response MV is the median value of owner-occupied homes, the covariates are 13 quantified measurement of its neighborhood whose description can be found in the manual of R package `mlbench`. The common specification uses RM^2 and NOX^2 to get a better fit, and for comparison we take these transformed variables as our input variables.

To exploit the power of varying coefficient model, we take the variable $W = \log(DIS)$, the weighted distances to five employment centers in the Boston region, as the index variable. This allows us to examine how the distance to the business hubs interact with other variables. It is reasonable to assume that the impact of other variables on housing price varies with the distance, which is an important characteristic of the neighborhood, i.e. the geographical accessibility to employment. Interestingly, rank regression selects the following submodel

$$\begin{aligned}\log(MV) = {} & \beta_0(W) + \beta_2(W) AGE + \beta_5(W) TAX + \beta_7(W)(B - 0.63)^2 \\ & + \beta_8(W) \log(LSTAT) + \beta_9(W) CRIM + \beta_{10}(W) ZN \\ & + \beta_{12}(W) CHAS + \beta_{13}(W) NOX^2 + \varepsilon.\end{aligned}$$

We now evaluate the performance of our two-step method in a high dimensional setting. To accomplish this, let $\{Z_1, \ldots, Z_p\}$ be i.i.d. the standard normal random variables and U follow the standard uniform distribution. We then expand the data set by adding the artificial predictors

$$X_j = \frac{Z_j + tU}{1+t}, \qquad j = 14, \ldots, p.$$

Note that $\{W, X_1, \ldots, X_{13}\}$ are the variables in original data set and the variables $\{X_j\}_{j=14}^{p}$ are known to be irrelevant to the housing price, though the maximum spurious correlation of these 987 artificial predictors to the housing price is now small. We take $p = 1000, t = 2$, and randomly select $n = 406$ samples as training set, and compute prediction mean squared error (PE) on the rest 100 samples. We repeat $N = 100$ times and report the average prediction error and model size. Since $\{X_j\}_{j=14}^{p}$ are artificial variables, we also include the number of artificial variables selected by each method as a proxy for false positives. Prediction error, model size and selected noise variables and its standard deviations over 100 repetitions are $0.049(0.010), 9.33(0.866)$ and $0(0)$, respectively. As seen, our two-step method is very effective in filtering noise variables in a high dimensional setting.

5 Conclusions

In this article, we proposed a two-step procedure for feature selection in ultra-high dimensional varying coefficient models by combining partial correlation and rank regression. We examined the finite sample performance of the proposed method via a Monte Carlo simulation study and an illustration through the Boston housing dataset. In conclusion, the proposed method is very useful in high-dimensional scientific discoveries, which can select a parsimonious close-to-truth model and reveal interesting relationship between variables, as illustrated in real data analysis.

Acknowledgements

The author thanks the editor and two anonymous reviewers for their substantial and constructive comments that have resulted in a significant improvement of the manuscript.

References

[1] J. Fan, Y. Feng and R. Song, *Nonparametric independence screening in sparse ultra-high dimensional additive models*, J. Amer. Statist. Assoc. **106** (2011) 544-557.

[2] J. Fan and R. Li, *Variable selection via nonconcave penalized likelihood and it oracle properties*, J. Amer. Statist. Assoc. **96** (2001) 1348-1360.

[3] J. Fan and J. Lv, *Sure independence screening for ultrahigh dimensional feature space (with discussion)*, J. R. Stat. Soc. Ser. B Stat. Methodol. **70** (2008) 849-911.

[4] J. Fan and W. Zhang, *Statistical methods with varying coefficient models*, Stat. Interface. **1** (2008) 179-195.

[5] J. Fan, Y. Ma and W. Dai, *Nonparametric independence screening in sparse ultra-high-dimensional varying coefficient models*, J. Amer. Statist. Assoc. **109** (2014) 1270-1284.

[6] J. Fan, R.J. Samworth and Y. Wu, *Ultrahigh dimensional feature selection: beyond the linear model*, J. Mach. Learn. Res. **10** (2009) 1829–1853.

[7] J. Fan and R. Song, *Sure independence screening in generalized linear models with NP-dimensionality*, Ann. Statist. **38** (2010) 3567-3604.

[8] D. Harrison and D. Rubinfeld, *Hedonic housing prices and the demand for clean air*, J. Environ. Econ. Manag. **5** (1978) 81-102.

[9] L.A. Jaeckel, *Estimating regression coefficients by minimizing the dispersion of residuals*, Ann. Math. Stat. **43** (1972) 1449-1458.

[10] M. Kazemi, D. Shahsavani and M. Arashi, *Variable selection and structure identification for ultrahigh-dimensional partially linear additive models with application to cardiomyopathy microarray data*, Stat. Optim. Inf. Comput. **6**(3) (2018) 373-382

[11] M. Kazemi, D. Shahsavani and M. Arashi, *A sure independence screening procedure for ultra-high dimensional partially linear additive models*, J. Appl. Stat. **46** (2019) 1385-1403.

[12] C. Leng, *Variable selection and coefficient estimation via regularized rank regression*, Statist. Sinica. **20** (2010) 167-181.

[13] R. Li and H. Liang, *Variable selection in semiparametric regression model*, Ann. Statist. **36** (2008) 261–286.

[14] X.J. Li, X.J. Ma and J.X. Zhang, *Robust feature screening for varying coefficient models via quantile partial correlation*, Metrika **80**(1) (2017) 17-49.

[15] R. Li, W. Zhong and L. P. Zhu, *Feature screening via distance correlation learning*, J. Amer. Statist. Assoc. **107** (2012) 1129-1139.

[16] J. Liu, R. Li and R. Wu, *Feature selection for varying coefficient models with ultrahigh-dimensional covariates*, J. Amer. Statist. Assoc. **109** (2014) 266-274.

[17] F. Song, Y. Chen and P. Lai, *Conditional distance correlation screening for sparse ultrahigh-dimensional models*, Appl. Math. Model. **81** (2020) 232-252.

[18] H. Wang and Y. Xia, *Shrinkage estimation of the varying coefficient model*, J. Amer. Statist. Assoc. **104** (2009) 747-757.

[19] L. Wang, H. Li and J. Huang, *Variable selection in nonparametric varying-coefficient models for analysis of repeated measurements*, J. Amer. Statist. Assoc. **103** (2008) 1556-1569.

[20] H. Yang, J. Lv and C. Guo, *Robust variable selection and parametric component identification in varying coefficient models*, Comm. Statist. Theory Methods **45** (2016) 5533-5549.

[21] L. P. Zhu, L. Li, R. Li, L. X. Zhu, *Model-free feature screening for ultrahigh dimensional data*, J. Amer. Statist. Assoc. **106** (2011) 1464-1475.