

VARIABLE SELECTION IN SUFFICIENT DIMENSION REDUCTION

Lexin Li*, R. Dennis Cook

School of Statistics, University of Minnesota

Christopher J. Nachtsheim

Carlson School of Management, University of Minnesota

Minneapolis, MN, 55455, USA

Abstract

In high-dimensional regression and data mining problems, variable selection can both significantly improve the prediction accuracy and facilitate the model interpretation. We propose a novel variable selection approach based on sufficient dimension reduction theory. The proposed method does not require parametric model specification, thus applies naturally to any prediction model. It also enjoys computational simplicity, therefore lends itself to large-scale data and real-time applications. This report summarizes the work to date.

1 Introduction

For the supervised learning problem with a univariate response y on a number of quantitative predictors $x \in \mathbb{R}^p$, the goal is to understand how the conditional distribution of $y | x$ changes as a function of the value of x . The theory of *sufficient dimension reduction* has been developed (Cook, 1998, Li, 2000) to replace x with a lower-dimensional *linearly transformed features* $\eta^T x$, where η is a $p \times q$ matrix with $q \leq p$. Specifically, we intend to find η such that

$$y \perp\!\!\!\perp x | \eta^T x \tag{1}$$

With η known, the data analysis can be carried out in the transformed lower-dimensional space $\eta^T x$ without any loss of information on the targeted characteristics of the conditional distribution of $y | x$. A set of estimation approaches to find η without requiring any pre-specified parametric model are available. The theory has been proved successful in a large number of real data applications; the reduction in dimension is often substantial: reduced predictors with dimension equal to 1 or 2 are common in practice.

*Corresponding author. School of Statistics, University of Minnesota. 313 Ford Hall, 224 Church St. S.E. Minneapolis, MN 55455. Tel: +1-612-624-8559. Email: lexinl@stat.umn.edu

Given a large number of predictors, it is reasonable to assume that some variables are irrelevant while some are highly correlated with others. Therefore it is desired to screen out both irrelevant and redundant features. There are several advantages of doing so. First, the model in the reduced lower-dimensional space is more interpretable; this is especially useful when a large portion of variables are noninformative and can be excluded from the modeling. Second, the prediction accuracy of the model can often be significantly improved in the reduced space, since the high-dimensional modeling inevitably suffers from the curse of dimensionality.

We propose a novel variable selection approach in this paper within the theoretical framework of sufficient dimension reduction. Specifically, we assume the existence of a subset of variables in x , denoted as x_1 , with x_2 representing the rest of predictors in x , such that,

$$y \perp\!\!\!\perp x_1 \mid x_2 \tag{2}$$

With x_1 known, the second-stage analysis can be conducted in the reduced feature space x_2 without loss of information on the conditional distribution of $y \mid x$. The proposed method identifies elements in x_1 without any model pre-specification, therefore applies naturally to any prediction model. It also demands relatively simple computational efforts thus is eligible for large-scale data mining applications.

Cook and Weisberg (1982), and Cook (1987) first studied the added-variable plots in the linear regression context, where the plots were used to examine the effect of inclusion of x_1 on the linear model of $y \mid x$ after x_2 being present. Cook (1994, 1996) extended the ideas of added-variable plots to general regression problems. The variable selection method proposed in this paper is based on the work of added-variable plots and is an extension of this graphical approach to an automated numerical procedure. The rest of the article is organized as follows. Section 2 develops the proposed selection approach for the cases when the response y is quantitative. Section 3 shows a number of simulation examples. Section 4 addresses the situations when y is categorical, followed by the simulations in Section 5. Discussion is presented in Section 6.

2 Sufficient variable selection for continuous response

2.1 Development

Let $S_{y|x}(\eta)$ denote the central dimension reduction subspace (Cook, 1998) for the regression of y on x . Partition $x^T = (x_1^T, x_2^T)$, where x_i is p_i dimensional, $i = 1, 2$, and $p_1 + p_2 = p$. Partition $\eta^T = (\eta_1^T, \eta_2^T)$ accordingly so that the row dimension of η_i corresponds to the dimension of x_i , $i = 1, 2$. We intend to study $S_{y|x}$ through the central subspace of marginal regression $S_{y|x_1}$. The relationship between $S_{y|x}$ and $S_{y|x_1}$ depends on the relation between the central subspace $S_{y|x_1}$ and the space spanned by the columns of η_1 , denoted by $S(\eta_1)$. If these two spaces are the same, i.e., $S_{y|x_1} = S(\eta_1)$, we then have

$$S_{y|x} \subseteq S(\eta_1) \oplus S(\eta_2) \subseteq S_{y|x_1} \oplus \mathbb{R}^{p_2} \tag{3}$$

An application of (3) is that, if $y \perp\!\!\!\perp x_1$ then $S_{y|x_1} = S\{0\}$, so that x_1 can be removed from the analysis of y on x . This is the basis for our proposed variable selection approach.

Next we explore the conditions that make $S_{y|x_1} = S(\eta_1)$. First we have the following proposition (Cook, 1998, Proposition 7.3).

Proposition 2.1 If $x_1 \perp\!\!\!\perp \eta_2^T x_2 \mid \eta_1^T x_1$, then $S_{y|x_1} \subseteq S(\eta_1)$.

The condition in proposition 2.1 alone is not sufficient to conclude that $S_{y|x_1} = S(\eta_1)$. Cook (1994) gave an example to show that it is possible for $S_{y|x_1}$ to be a proper subset of $S(\eta_1)$. However, he also suggested that the regression problems in which $S_{y|x_1}$ is a proper subset of $S(\eta_1)$ might be the exception rather than the rule in practice. Henceforth we will assume $S_{y|x_1} = S(\eta_1)$ whenever $S_{y|x_1} \subseteq S(\eta_1)$ in the rest of the paper.

A special case for the condition in proposition 2.1 to be true is that $x_1 \perp\!\!\!\perp x_2$. However both the conditions $x_1 \perp\!\!\!\perp \eta_2^T x_2 \mid \eta_1^T x_1$ and $x_1 \perp\!\!\!\perp x_2$ seem rather restrictive from a practical point of view. To increase the degree of applicability of the proposition, we modify the predictors to satisfy the condition while preserving $S(\eta_1)$. Define the population predictor residual

$$r_{1|2} = x_1 - \mathbb{E}(x_1|x_2) \quad (4)$$

If the regression of x_1 on x_2 follows an additive-location regression such that $x_1 - \mathbb{E}(x_1|x_2) \perp\!\!\!\perp x_2$, and we modify the predictors from x_1 to $r_{1|2}$, the condition for proposition 2.1 holds. More generally, we have the next proposition.

Proposition 2.2 If $r_{1|2} \perp\!\!\!\perp x_2 \mid \eta_1^T r_{1|2}$, then $y \perp\!\!\!\perp r_{1|2} \mid \eta_1^T r_{1|2}$, or equivalently, $S_{y|r_{1|2}} \subseteq S(\eta_1)$.

Proof Firstly, by Proposition 4.5, Cook (1998), we have

$$\begin{aligned} y \perp\!\!\!\perp x \mid \eta^T x &\Rightarrow y \perp\!\!\!\perp (r_{1|2}, x_2) \mid \eta_1^T x_1 + \eta_2^T x_2 \\ &\Rightarrow y \perp\!\!\!\perp (r_{1|2}, x_2) \mid (\eta_1^T x_1 + \eta_2^T x_2, \eta_1^T r_{1|2}, x_2) \\ &\Rightarrow y \perp\!\!\!\perp r_{1|2} \mid (\eta_1^T x_1 + \eta_2^T x_2, \eta_1^T r_{1|2}, x_2) \end{aligned} \quad (5)$$

Secondly, we observe that

$$\begin{aligned} \eta_1^T x_1 + \eta_2^T x_2 &= \eta_1^T (r_{1|2} + \mathbb{E}(x_1) + \mathbb{E}(x_1|x_2)) + \eta_2^T x_2 \\ &= \eta_1^T r_{1|2} + (\eta_1^T \mathbb{E}(x_1|x_2) + \eta_2^T x_2) + \eta_1^T \mathbb{E}(x_1) \end{aligned}$$

which is a function of $(\eta_1^T r_{1|2}, x_2)$. Therefore we have

$$y \perp\!\!\!\perp \eta_1^T x_1 + \eta_2^T x_2 \mid (\eta_1^T r_{1|2}, x_2) \quad (6)$$

By (5), (6), and Conditional Independence Proposition 4.6, Cook (1998), we obtain that

$$y \perp\!\!\!\perp r_{1|2} \mid (\eta_1^T r_{1|2}, x_2) \quad (7)$$

Finally, by (7), the condition $r_{1|2} \perp\!\!\!\perp x_2 \mid \eta_1^T r_{1|2}$, and Proposition 4.6 again, we conclude that

$$y \perp\!\!\!\perp r_{1|2} \mid \eta_1^T r_{1|2} \quad (8)$$

Therefore $S(\eta_1)$ is a dimension reduction subspace of regression of y on $r_{1|2}$ by (8). Since $S_{y|r_{1|2}}$ is the corresponding central dimension reduction subspace, we have $S_{y|r_{1|2}} \subseteq S(\eta_1)$.

Proposition 2.2 states that if $r_{1|2} \perp\!\!\!\perp x_2 \mid \eta_1^T r_{1|2}$, then $(y, x_2) \perp\!\!\!\perp r_{1|2} \mid \eta_1^T r_{1|2}$. This implies that any function of (y, x_2) is independent of $r_{1|2}$ given $\eta_1^T r_{1|2}$. In particular, define the population residual from modeling y as a function of x_2 alone

$$r_{y|2} = y - E(y|x_2) \quad (9)$$

Following the condition in proposition 2.2, we have

Corollary 2.3 If $r_{1|2} \perp\!\!\!\perp x_2 \mid \eta_1^T r_{1|2}$, then $r_{y|2} \perp\!\!\!\perp r_{1|2} \mid \eta_1^T r_{1|2}$, or equivalently, $S_{r_{y|2}|r_{1|2}} \subseteq S(\eta_1)$.

Again, we will assume $S_{r_{y|2}|r_{1|2}} = S(\eta_1)$ whenever $S_{r_{y|2}|r_{1|2}} \subseteq S(\eta_1)$. Moving from y to $r_{y|2}$ has the effect of reducing the gross variation in y in the graphical representation. A direct application of the corollary 2.3 is that, if we know $r_{y|2} \perp\!\!\!\perp r_{1|2}$, together with the condition $r_{1|2} \perp\!\!\!\perp x_2 \mid \eta_1^T r_{1|2}$, we have $S(\eta_1) = S_{r_{y|2}|r_{1|2}} = S\{0\}$, which in turn implies that x_1 can be removed from the analysis of y on x .

To further relax the requirement of $r_{1|2} \perp\!\!\!\perp x_2 \mid \eta_1^T r_{1|2}$, we let $p_1 = 1$. Under this situation, we have the next proposition.

Proposition 2.4 Given (a) $p_1 = 1$, (b) $r_{y|2} \perp\!\!\!\perp r_{1|2}$, (c) whenever $S_{r_{y|2}|r_{1|2}} \subseteq S(\eta_1)$, we assume $S_{r_{y|2}|r_{1|2}} = S(\eta_1)$, then $\eta_1 = 0$, or say, x_1 can be removed from analysis of y on x .

Proof The proof is completed by inducing the contradiction when assuming $\eta_1 \neq 0$. By condition (a) and assumption $\eta_1 \neq 0$, we get $r_{1|2} \perp\!\!\!\perp x_2 \mid \eta_1^T r_{1|2}$. Following Corollary 2.3, we have $r_{y|2} \perp\!\!\!\perp r_{1|2} \mid \eta_1^T r_{1|2}$, or equivalently, $S_{r_{y|2}|r_{1|2}} \subseteq S(\eta_1)$. Then by condition (c), we have $S_{r_{y|2}|r_{1|2}} = S(\eta_1)$. Meanwhile, condition (b) suggests that $S_{r_{y|2}|r_{1|2}} = S\{0\}$, which in turn implies that $\eta_1 = 0$. However this contradicts the assumption that $\eta_1 \neq 0$.

Proposition 2.4 is the theoretical basis for our proposed variable selection approach. Operationally, we employ a backward one-variable-at-a-time elimination procedure. For a given predictor x_1 , and the rest of predictors x_2 , we compute the residuals $r_{y|2}$ in (9) and $r_{1|2}$ in (4). We then check the independence between these two one-dimensional variables $r_{y|2}$ and $r_{1|2}$. If they are independent, we delete x_1 , and repeat the whole process for the remaining predictors. We stop when no more predictors can be removed. The next algorithm summaries this selection procedure, and is to be referred to as the *added-variable selection approach* in the rest of the paper.

Algorithm I

1. Let x_{cand} denote the current set of candidate predictors. At the start of the algorithm x_{cand} contains all predictors in the input space. Let p_{cand} denote the number of variables in x_{cand} .
2. Let x_i denote the i th predictor in x_{cand} , $i = 1, \dots, p_{cand}$.

- (a) Partition x_{cand} as $\tilde{x}_1 = x_i, \tilde{x}_2 = x_{-i}$, where x_{-i} denotes variables in x_{cand} excluding x_i .
 - (b) Compute the residuals $r_{y|2}$ and $r_{1|2}$ by applying the selected smoothing approaches to the fittings of y on \tilde{x}_2 and \tilde{x}_1 on \tilde{x}_2 respectively.
 - (c) Test the hypothesis $H_0 : r_{y|2} \perp\!\!\!\perp r_{1|2}$. Store the p-value of this test.
3. Identify the predictor with the largest p-value in step 2. If this p-value exceeds a per-specified threshold, drop the associated predictor from x_{cand} , decrement p_{cand} and repeat step 2; otherwise stop.

It is interesting to notice that, the condition we check in the added-variable selection approach is

$$r_{y|2} \perp\!\!\!\perp r_{1|2} \tag{10}$$

If this condition holds, the conclusion we draw is

$$\begin{aligned} y \perp\!\!\!\perp x_1 \mid x_2 &\Leftrightarrow y - \mathbb{E}(y \mid x_2) \perp\!\!\!\perp x_1 - \mathbb{E}(x_1 \mid x_2) \mid x_2 \\ &\Leftrightarrow r_{y|2} \perp\!\!\!\perp r_{1|2} \mid x_2 \end{aligned} \tag{11}$$

We conclude the conditional independence in (11) from the independence in (10). However, given arbitrarily three random variables U, V , and W , the independence $U \perp\!\!\!\perp V$ does *not* necessarily imply the conditional independence $U \perp\!\!\!\perp V \mid W$.

2.2 Another view

Next we examine a semi-parametric model and apply the Taylor expansion to gain some insight to the proposed added-variable selection approach. Specifically, consider the model of the following form

$$y = f(x_1, x_2) + g(x_1, x_2) \varepsilon \tag{12}$$

where $\varepsilon \perp\!\!\!\perp (x_1, x_2)$, $\mathbb{E}(\varepsilon) = 0$, $\text{Var}(\varepsilon) = 1$, $x_1 \in \mathbb{R}$, and $x_2 \in \mathbb{R}^{p-1}$. Some direct calculation yields

$$\begin{aligned} r_{y|2} &= f(x_1, x_2) + g(x_1, x_2) \varepsilon - \mathbb{E}(f(x_1, x_2) \mid x_2) - \mathbb{E}(g(x_1, x_2) \varepsilon \mid x_2) \\ &= f(x_1, x_2) + g(x_1, x_2) \varepsilon - \mathbb{E}(f(x_1, x_2) \mid x_2) - \mathbb{E}(g(x_1, x_2) \mid x_2) \mathbb{E}(\varepsilon) \\ &= f(x_1, x_2) - \mathbb{E}(f(x_1, x_2) \mid x_2) + g(x_1, x_2) \varepsilon \end{aligned} \tag{13}$$

Next we apply Taylor expansion of the mean function $f(x_1, x_2)$ with respect to x_1 around the point $x_1 = 0$, we get

$$f(x_1, x_2) = f(0, x_2) + \dot{f}(0, x_2) x_1 + \frac{1}{2!} \ddot{f}(0, x_2) x_1^2 + R \tag{14}$$

$$\mathbb{E}(f(x_1, x_2) \mid x_2) = f(0, x_2) + \dot{f}(0, x_2) \mathbb{E}(x_1 \mid x_2) + \frac{1}{2!} \ddot{f}(0, x_2) \mathbb{E}(x_1^2 \mid x_2) + \mathbb{E}(R \mid x_2) \tag{15}$$

Substitute (14) and (15) into (13), we get

$$r_{y|2} = \dot{f}(0, x_2) r_{1|2} + \frac{1}{2!} \ddot{f}(0, x_2) [x_1^2 - \mathbb{E}(x_1^2 | x_2)] + [R - \mathbb{E}(R | x_2)] + g(x_1, x_2) \varepsilon \quad (16)$$

We then assume $g(x_1, x_2) = \sigma^2$, a constant, and examine several cases based on (16).

Case I: If x_1 is not present in $f(x_1, x_2)$, then the first partial derivative \dot{f} and all the subsequent partial derivatives should equal to 0. Therefore there should exhibit no dependence between $r_{y|2}$ and $r_{1|2}$.

Case II: If x_1 is linear in $f(x_1, x_2)$, then the second partial derivative \ddot{f} and all the subsequent partial derivatives should equal to 0. In this case, the relation between $r_{y|2}$ and $r_{1|2}$ is essentially linear, and many tests are available for picking out the linear dependence.

Case III: If x_1 is nonlinear in $f(x_1, x_2)$, then higher (than first) order partial derivatives are nonzero and the dependence between $r_{y|2}$ and $r_{1|2}$ involves higher (than linear) order correlation. In this case we need an independence test which is able to detect the high order correlations of two random variables.

2.3 Smoothing

The implementation of the added-variable approach involves the smoothing procedures to obtain the residuals $r_{y|2}$ and $r_{1|2}$. In principle, we can apply any smoothing function to estimate $\mathbb{E}(y | x_2)$ and $\mathbb{E}(x_1 | x_2)$ and to obtain estimates of $r_{y|2}$ and $r_{1|2}$. In practice, however, we choose to use the most simple smoothing function, the linear smoothing. Define the population OLS residuals

$$e_{y|2} = y - \mathbb{E}(y) - \beta_{y|2}^T (x_2 - \mathbb{E}(x_2)) \quad (17)$$

$$e_{1|2} = x_1 - \mathbb{E}(x_1) - \beta_{1|2}^T (x_2 - \mathbb{E}(x_2)) \quad (18)$$

where $\beta_{y|2}$ and $\beta_{1|2}$ are the population OLS vectors of regression of y on x_2 and x_1 on x_2 respectively. We apply the independence test on $e_{y|2}$ and $e_{1|2}$, in place of $r_{y|2}$ and $r_{1|2}$. There are several advantages of doing so. First the computation of OLS estimates is inexpensive, therefore the proposed approach is applicable to the large-scale data problems. Second, more complicated smoothing function, e.g., a full quadratic smoothing, requires more free parameters, thus demands a larger number of observations. This may serve to restrict application of the variable selection approach to the problems when there are a relatively small number of observations compared with the number of predictors.

The next proposition justifies the use of $e_{y|2}$ and $e_{1|2}$ in place of $r_{y|2}$ and $r_{1|2}$. Its proof follows immediately that of Proposition 2.2.

Proposition 2.5 If $e_{1|2} \perp\!\!\!\perp x_2 | \eta_1^T e_{1|2}$, then $y \perp\!\!\!\perp e_{1|2} | \eta_1^T e_{1|2}$, or equivalently, $S_{y|e_{1|2}} \subseteq S(\eta_1)$.

Following the same development as for $r_{y|2}$ and $r_{1|2}$, we can check the independence between $e_{y|2}$ and $e_{1|2}$ to infer the exclusion of x_1 .

2.4 Independence test

Another component in the proposed added-variable selection approach is the independence test of two one-dimensional random variables. Two tests are examined in our simulation study below. One is based on the data-driven rank tests by Kallenberg and Ledwina (1999), and the other is a modified version of the χ^2 test.

Kallenberg and Ledwina (1999) proposed testing the independence of random variables X and Y by detecting the correlation between the marginal distributions $F(X)$ and $G(Y)$. Their test statistic is of the form

$$\sum_{(r,s) \in \Lambda} V(r,s) = \sum_{(r,s) \in \Lambda} \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^n b_r \left(\frac{R_i - 1/2}{n} \right) \cdot b_s \left(\frac{S_i - 1/2}{n} \right) \right\}^2 \quad (19)$$

where $\{b_r\}$ are orthonormal Legendre polynomials on $[0, 1]$, Λ is the set of all pairs of orders (r, s) of Legendre polynomials under consideration, and R, S are ranks of X, Y respectively. Notice that the first term $V(1, 1)$ corresponds to the linear correlation between $F(X)$ and $G(Y)$, and if $\Lambda = \{(1, 1)\}$, then the test statistic in (19) becomes the Spearman's rank test statistic (differing by a constant). Generally, the term $V(r, s)$ reflects the correlation of r th order polynomial of $F(X)$ and s th order polynomial of $G(Y)$. To choose a proper set of orders of polynomials for Λ , a modified BIC selection rule was recommended. The limiting distribution of the proposed test statistic was obtained under some mild conditions, therefore we can compute the p-value of the independence test.

We also propose a heuristic approach, called gridded χ^2 test, for testing the independence of two one-dimensional variables X and Y . Specifically, a χ^2 test is applied to the observations divided into k^2 classes determined by $k - 1$ equally spaced values of X and Y , and this test is repeated for a sequence of k ranging from 2 to a pre-specified maximum number of classes in each variable, called K . Each χ^2 test produces one p-value, and the 25% quartile of all these $K - 1$ p-values is chosen as the heuristic measure of the independence between X and Y . This 25%-quartile-rule has been shown to perform well in our simulations; meanwhile we set $K = 10$ in all simulations.

3 Simulation examples I

Example 1.1 The first example is taken from Li (1991). Let $p = 10$ and $x_1, \dots, x_{10}, \varepsilon$ be independent standard normal random variables. Consider two regression models

$$\begin{aligned} y_1 &= x_1(x_1 + x_2 + 1) + 0.5\varepsilon \\ y_2 &= \frac{x_1}{0.5 + (x_2 + 1.5)^2} + 0.5\varepsilon \end{aligned}$$

The sample size is set as $n = 200$. Two versions of added-variable selection approaches are applied, one with rank test and the other with gridded χ^2 test. These two versions will be labeled as AVC1 and AVC2 selection rules in the rest of simulation examples for brevity. The threshold for p-value in step 3 of Algorithm I is set as 0.1. AIC and BIC selection rules are implemented for the purpose of comparisons. All results are based on 10 replications.

The criteria are the percentages of times a selection rule picks out x_1 , x_2 , (x_1, x_2) , and only (x_1, x_2) respectively among all replications. Figure 1 shows the barplot of those criteria when four selection rules AVC1, AVC2, AIC, and BIC are applied to response y_1 . Figure 2 is the bar plot for response y_2 .

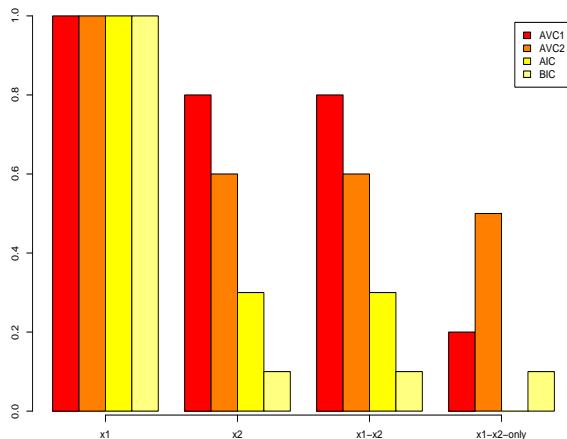


Figure 1. y_1

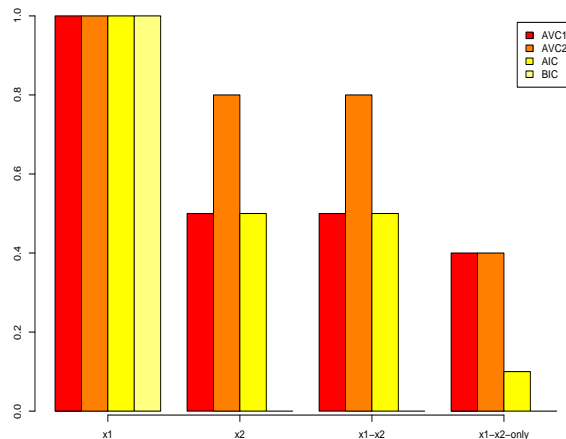


Figure 2. y_2

Example 1.2 The next example explicitly introduces a nonlinear relationship among predictors. Let $p = 10$, and $x_4, \dots, x_{10}, \varepsilon$ are independent standard normal variables. Three relevant features x_1, x_2, x_3 and response are generated as

$$\begin{aligned}
 x_1 &\sim U(0, 1) \\
 x_2 &= \log(x_1) + e_1, \quad e_1 \sim U(-0.5, 0.5) \\
 x_3 &= \exp(x_1) + x_2^2 + e_2, \quad e_2 \sim N(0, 0.1^2) \\
 y &= (x_1 - x_2 + x_3)^2 + e_3, \quad e_3 \sim N(0, 0.1^2)
 \end{aligned}$$

The sample size is set as $n = 100$. The criteria are the percentages of times a selection rule selects x_1 , x_2 , x_3 , (x_1, x_2, x_3) , and only (x_1, x_2, x_3) among 10 replications. Figure 3 shows the barplot of these criteria when four selection rules are applied.

Example 1.3 The next example examines the performance of selection rules when there is little linear trend in $E(y|x)$. p is chosen to be 15, with x_1, \dots, x_{15} as independent standard normal variables. The responses are generated as

$$\begin{aligned}
 y_1 &= (x_1 + x_2 + x_3)^2 + \varepsilon_1 \\
 y_2 &= (10x_1 + 5x_2 + x_3)^2 + \varepsilon_2
 \end{aligned}$$

where $\varepsilon_1, \varepsilon_2$ are $\text{Normal}(0, 0.5^2)$, and are independent of x 's. $n = 100$. Figure 4 and 5 are barplots for y_1 and y_2 respectively.

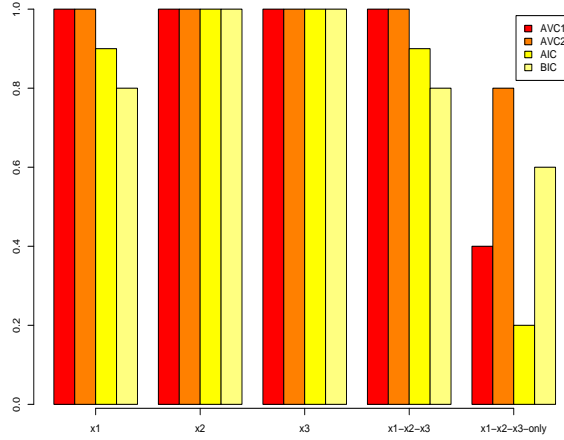


Figure 3.

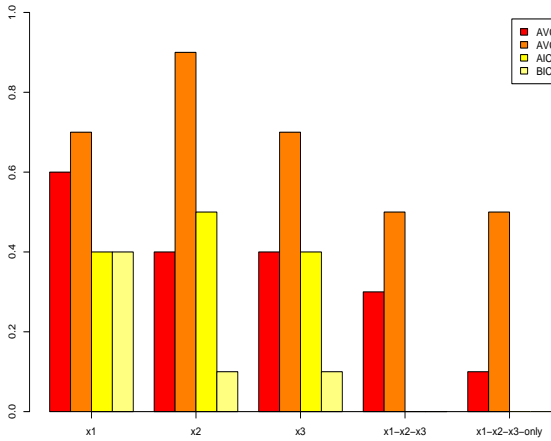


Figure 4. y_1

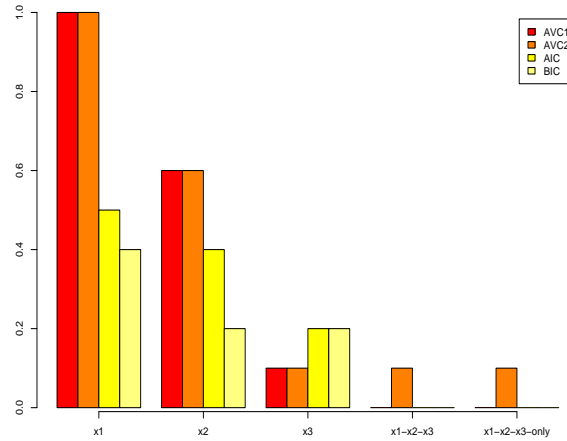


Figure 5. y_2

Example 1.4 Relevant features can appear in both $E(y|x)$ and $\text{Var}(y|x)$. This example examines whether the selection rules can pick out features in the variance structure. $p = 10$, and $n = 100$. $x_1, \dots, x_{10}, \varepsilon$ are independent standard normal variables. The response is generated as

$$y = \beta_1^T x + 0.2 \cdot (0.8 \beta_2^T x + 3)^2 \cdot \varepsilon$$

where $x = (x_1, \dots, x_{10})$, $\beta_1 = (1, 1, 0, 0, 0, 0, 0, 0, 0, 0)^T$, and $\beta_2 = (1, 0, 1, 0, 0, 0, 0, 0, 0, 0)^T$. Figure 6 is corresponding barplot.

Summary of simulation examples: Based on the barplots shown in Figures 1 to 6, it is clear that the proposed added-variable selection approach successfully identifies the relevant predictors, while AIC and BIC selection rules fail to do so.

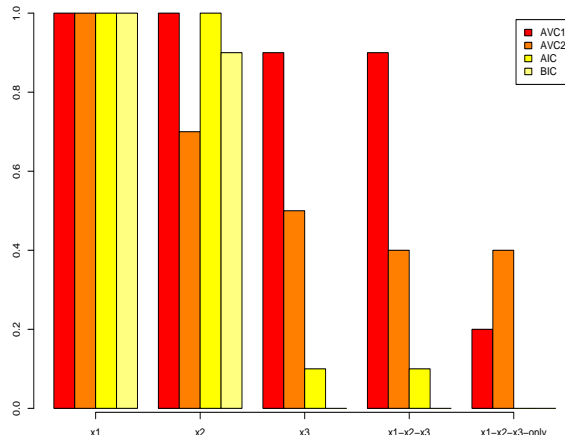


Figure 6. y

4 Sufficient variable selection for discrete response

4.1 Development

When the response y is categorical, we follow similar development of added-variable selection approach as for continuous response, but employ a slightly different treatment. In this paper we mainly focus on the binary response situations, which correspond to the most commonly seen two-class classification problems in practice.

For the binary response $y = (0, 1)$, it can be trivially shown that

$$y \perp\!\!\!\perp x \quad \text{if and only if} \quad x | (y = 0) \stackrel{d}{=} x | (y = 1) \quad (20)$$

where $\stackrel{d}{=}$ denotes the equivalence of distributions. Following Proposition 2.2 we have the next proposition

Proposition 4.1 Given (a) $p_1 = 1$, (b) $r_{1|2} | (y = 0) \stackrel{d}{=} r_{1|2} | (y = 1)$, (c) whenever $S_{y|r_{1|2}} \subseteq S(\eta_1)$, we assume $S_{y|r_{1|2}} = S(\eta_1)$, then $\eta_1 = 0$, or say, x_1 can be removed from analysis of y on x .

Proof The proof is completed by inducing the contradiction when assuming $\eta_1 \neq 0$. By condition (a) and assumption $\eta_1 \neq 0$, we get $r_{1|2} \perp\!\!\!\perp x_2 | \eta_1^T r_{1|2}$. Following Proposition 2.2, we have $y \perp\!\!\!\perp r_{1|2} | \eta_1^T r_{1|2}$, or equivalently, $S_{y|r_{1|2}} \subseteq S(\eta_1)$. Then by condition (c), we have $S_{y|r_{1|2}} = S(\eta_1)$. Meanwhile, condition (b) suggests that $S_{y|r_{1|2}} = S\{0\}$, which in turn implies that $\eta_1 = 0$. However this contradicts the assumption that $\eta_1 \neq 0$.

Proposition 4.1 is the theoretical basis of the added-variable selection approach for binary response problems. A backward one-variable-at-a-time elimination procedure is used. The predictor residual $r_{1|2}$ is estimated by OLS residual $e_{1|2}$. Kolmogorov-Smirnov test is employed to test distribution equivalence condition in Proposition 4.1 (b). An algorithm

similar to Algorithm I can be developed analogously for the binary response problems.

Algorithm II

1. Let x_{cand} denote the current set of candidate predictors. At the start of the algorithm x_{cand} contains all predictors in the input space. Let p_{cand} denote the number of variables in x_{cand} .
2. Let x_i denote the i th predictor in x_{cand} , $i = 1, \dots, p_{cand}$.
 - (a) Partition x_{cand} as $\tilde{x}_1 = x_i$, $\tilde{x}_2 = x_{-i}$, where x_{-i} denotes variables in x_{cand} excluding x_i .
 - (b) Compute the residual $r_{1|2}$ by applying the selected smoothing approach to the fitting of \tilde{x}_1 on \tilde{x}_2 .
 - (c) Test the hypothesis $H_0 : r_{1|2} | (y = 0) \stackrel{d}{=} r_{1|2} | (y = 1)$ using Kolmogorov-Smirnov test. Store the p-value for this test.
3. Identify the predictor with the largest p-value in step 2. If this p-value exceeds a perspecified threshold, drop the associated predictor from x_{cand} , decrement p_{cand} and repeat step 2; otherwise stop.

4.2 Another view

Next we assume the design matrix x to be orthonormal, and the elements of x are independent with each other. We then adopt a semi-parametric model to gain some insight to the proposed variable selection approach. Specifically, we consider the model

$$\begin{aligned}
 g(x) &= g(x_1, x_2) = g(x_2) \\
 p(x) &= \frac{1}{1 + \exp(-g(x))} \\
 y &= \begin{cases} 1 & \text{with probability } p(x) \\ 0 & \text{with probability } 1 - p(x) \end{cases}
 \end{aligned}$$

where $x = (x_1^T, x_2^T)^T \in \mathbb{R}^p$, and $g(x)$ is a function determining the decision boundary and depends only on x_2 . Under this setting, we have $y \perp\!\!\!\perp x_1 \mid x_2$, and $x_1 \perp\!\!\!\perp x_2$, and hence, $y \perp\!\!\!\perp x_1$. Therefore,

$$\frac{f(r_{1|2} \mid y = 1)}{f(r_{1|2} \mid y = 0)} = \frac{f(x_1 \mid y = 1)}{f(x_1 \mid y = 0)} = \frac{f(x_1)}{f(x_1)} = 1 \tag{21}$$

where $f(\cdot)$ stands for the density function.

On the other hand, we know

$$\begin{aligned}
 \log \frac{f(y = 1 \mid x)}{f(y = 0 \mid x)} &= \log \frac{p(x)}{1 - p(x)} = g(x_2) \\
 \log \frac{f(y = 1 \mid x)}{f(y = 0 \mid x)} &= \log \frac{f(x \mid y = 1)}{f(x \mid y = 0)} + \log \frac{f(y = 1)}{f(y = 0)}
 \end{aligned}$$

Therefore,

$$\log \frac{f(x | y = 1)}{f(x | y = 0)} = g(x_2) - \log \frac{f(y = 1)}{f(y = 0)} \quad (22)$$

In addition, we have

$$f(x | y) = f(x_1, x_2 | y) = f(x_1 | x_2, y) f(x_2 | y) = f(x_1 | x_2) f(x_2 | y)$$

where the last equality comes from the fact that $y \perp\!\!\!\perp x_1 | x_2$. Henceforth,

$$\log \frac{f(x | y = 1)}{f(x | y = 0)} = \log \frac{f(x_1 | x_2) f(x_2 | y = 1)}{f(x_1 | x_2) f(x_2 | y = 0)} = \log \frac{f(x_2 | y = 1)}{f(x_2 | y = 0)} \quad (23)$$

Substitute (23) back to (22), we obtain

$$\log \frac{f(x_2 | y = 1)}{f(x_2 | y = 0)} = g(x_2) - \log \frac{f(y = 1)}{f(y = 0)}$$

Since $x_1 \perp\!\!\!\perp x_2$, we finally have

$$\frac{f(r_{2|1} | y = 1)}{f(r_{2|1} | y = 0)} = \frac{f(x_2 | y = 1)}{f(x_2 | y = 0)} = \exp \left\{ g(x_2) - \log \frac{f(y = 1)}{f(y = 0)} \right\} \quad (24)$$

Equation (21) states that if x_1 is not in the model, then the ratio of two density functions $f(r_{1|2} | y = 1)$ and $f(r_{1|2} | y = 0)$ should be 1; while equation (24) implies that if x_2 is indeed in the model, then the ratio of two density functions $f(r_{2|1} | y = 1)$ and $f(r_{2|1} | y = 0)$ is a function of x_2 .

5 Simulation examples II

Example 2.1 The first example follows the setting in the previous section. Specifically, let $p = 10$, and x_1, \dots, x_{10} be independent standard normal random variables. Consider the binary response model

$$\begin{aligned} g(x) &= (x_1 + x_2 + x_3)^2 - 2 \\ p(x) &= \frac{1}{1 + \exp(-g(x))} \\ y &\sim \text{Bernoulli}(p(x)) \end{aligned}$$

Under this setting, the decision boundary is determined by $(x_1 + x_2 + x_3)^2 - 2 = 0$. There are about half observations of y taking value 1 and half 0. A large sample size is chosen as $n = 3000$. The added-variable selection rule for discrete response will be labeled as AVD in the rest of the section for brevity. The threshold for p-value in step 3 of Algorithm II is set as 0.1, the same as for continuous response cases. AIC and BIC selection rules are again implemented for comparisons. The criteria are the percentages of times a selection rule picks out $x_1, x_2, x_3, (x_1, x_2, x_3)$, and only (x_1, x_2, x_3) among 10 replications. Figure 7 shows the

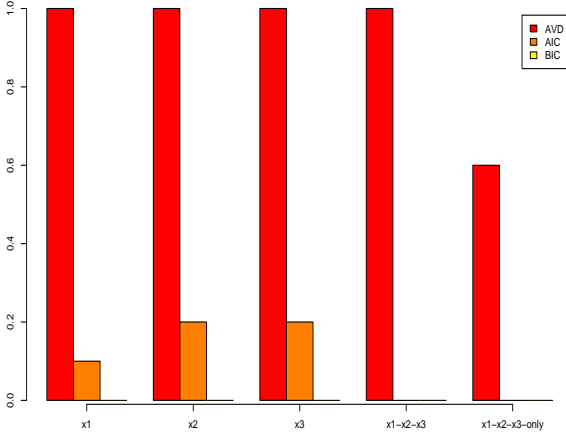


Figure 7.

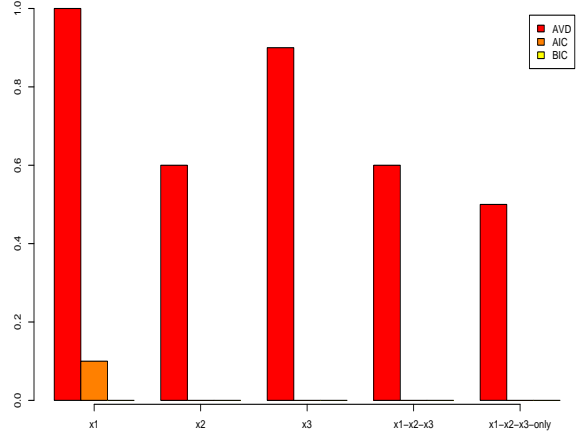


Figure 8.

barplot of the criteria when AVD, AIC, and BIC selection rules are applied.

Example 2.2 The next example employs the following nonlinear decision boundary $x_1/(x_2 + x_3) - 5 = 0$, which yields that most observations of y are 0, and very a few are 1. In addition, $p = 10, n = 3000$.

$$\begin{aligned}
 g(x) &= \frac{x_1}{x_2 + x_3} - 5 \\
 p(x) &= \frac{1}{1 + \exp(-g(x))} \\
 y &\sim \text{Bernoulli}(p(x))
 \end{aligned}$$

Three selection rules AVD, AIC, and BIC are applied with 10 replications. Figure 8 is the barplot constructed in the same manner as Figure 7.

Example 2.3 In this example, we still set $n = 3000, p = 10$ with x_4, \dots, x_{10} following the standard normal distribution, but introduce nonlinear relations among x_1, x_2 and x_3 . Figure 9 shows the corresponding barplot

$$\begin{aligned}
 x_1 &\sim \text{U}(0, 1) \\
 x_2 &= \log(x_1) + e_1, \quad e_1 \sim \text{U}(-0.3, 0.3) \\
 x_3 &= \frac{10x_1}{5 + x_2} + e_2, \quad e_2 \sim \text{N}(0, 0.1^2) \\
 g(x) &= \exp(x_1) + x_2 + x_3^2 - 3 \\
 p(x) &= \frac{1}{1 + \exp(-g(x))} \\
 y &\sim \text{Bernoulli}(p(x))
 \end{aligned}$$

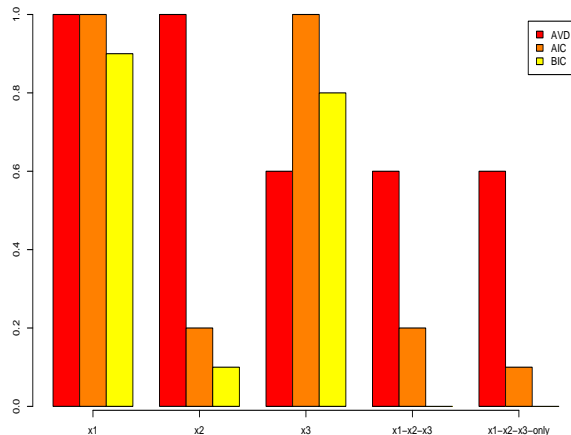


Figure 9.

Summary of simulation examples: Based on the barplots shown in Figures 7 to 9, it is clear that, when the response is binary, the proposed added-variable selection approach successfully identifies the relevant predictors, while AIC and BIC selection rules fail again.

6 Discussion

We have found that the proposed added-variable selection approach useful during the exploratory data analysis stage. The approach does not require parametric model specification, thus applies naturally to any prediction model. It also enjoys computational simplicity, therefore lends itself to large-scale data and real-time applications. In the high-dimensional regression and data mining problems, the effective variable selection can both significantly improve the prediction accuracy and facilitate the model interpretation.

There are many selection rules in addition to AIC, BIC for variable subset selection. However most of those rules work only for the linear regression problems. There is also another class of variable subset selection approaches which are mainly data-driven. Two most commonly seen examples include cross validation and bootstrap. It is noteworthy that those approaches actually do variable selection and model fitting simultaneously, while our proposed approach focus on the variable selection without fitting any model. This difference corresponds to that of a wrapper-based selection approach and a filter-based selection approach in machine learning research. Details of comparison can be found in Koller and Sahami (1996).

Variable subset selection is regarded as a discrete selection procedure, because individual predictor only has two states in the model, present or absent. Alternatively, a class of continuous variable selection approaches through coefficient shrinkage have been proposed. Some representative examples include the ridge regression, the least absolute shrinkage and selection operator (Lasso, Tibshirani, 1996), and the nonconcave penalized likelihood selection (Fan and Li, 2001).

7 Reference

- Cook, R.D. (1987). Influence assessment, *Journal of Applied Statistics*, **14**, 117-131.
- Cook, R.D. (1994). On the interpretation of regression plots. *Journal of the American Statistical Association*, **89**, 177-190.
- Cook, R.D. (1996). Graphics for regressions with a binary response. *Journal of the American Statistical Association*, **91**, 983-992.
- Cook, R.D. (1998). *Regression Graphics: Ideas for Studying Regressions Through Graphics*, New York: Wiley.
- Cook, R.D., and Weisberg, S. (1982). *Residuals and Influence in Regression*. London: Chapman and Hall.
- Fan, J., Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**, 1348-1360.
- Li, K.C. (1991). Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, **86**, 316-327.
- Li, K.C. (2000). High dimensional data analysis via the SIR / PHD approach. Unpublished manuscript dated April 6, 2000 obtained at the Internet site www.stat.ucla.edu/~kcli/sir-PHD.pdf.
- Chen, C.H., and Li, K.C. (1998). Can SIR be as popular as multiple linear regression? *Statistica Sinica*, **8**, 289-316.
- Kallenberg, W.C.M., and Ledwina T. (1999). Data-driven rank tests for independence. *Journal of the American Statistical Association*, **94**, 285-301.
- Koller, D., and Sahami, M. (1996). Toward optimal feature selection. *Machine Learning: Proceedings of the Thirteenth International Conference*. Morgan Kaufmann.
- Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society, Series B.* **58**, 267-288.
- Tibshirani, R., and Knight, K. (1999). The covariance inflation criterion for adaptive model selection. *Journal of the Royal Statistical Society, Series B.* **61**, 529-546.
- Xing, E.P., Jordan, M.I., and Karp, R.M. (2001). Feature selection for high-dimensional genomic microarray data. *Machine Learning: Proceedings of the Eighteenth International Conference*, San Mateo, CA: Morgan Kaufmann.