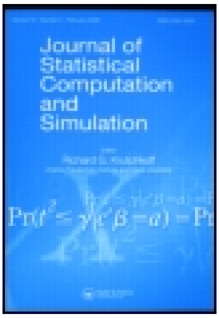

Journal of Statistical Computation and Simulation

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/gscs20

Variable selection in regression using maximal correlation and distance correlation

C. Deniz Yenigün[a] & Maria L. Rizzo[b]

[a] Faculty of Business Administration, Bilkent University, Ankara 06800, Turkey

[b] Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, OH, USA

Published online: 18 Mar 2014.

To cite this article: C. Deniz Yenigün & Maria L. Rizzo (2014): Variable selection in regression using maximal correlation and distance correlation, Journal of Statistical Computation and Simulation, DOI: 10.1080/00949655.2014.895354

To link to this article: http://dx.doi.org/10.1080/00949655.2014.895354

PLEASE SCROLL DOWN FOR ARTICLE

Journal of Statistical Computation and Simulation, 2014
http://dx.doi.org/10.1080/00949655.2014.895354

Variable selection in regression using maximal correlation and distance correlation

C. Deniz Yenigün[a]* and Maria L. Rizzo[b]

[a] *Faculty of Business Administration, Bilkent University, Ankara 06800, Turkey;* [b] *Department of Mathematics and Statistics, Bowling Green State University, Bowling Green, OH, USA*

(*Received 26 April 2013; accepted 13 February 2014*)

In most of the regression problems the first task is to select the most influential predictors explaining the response, and removing the others from the model. These problems are usually referred to as the *variable selection problems* in the statistical literature. Numerous methods have been proposed in this field, most of which address linear models. In this study we propose two variable selection criteria for regression based on two powerful dependence measures, maximal correlation and distance correlation. We focus on these two measures since they fully or partially satisfy the Rényi postulates for dependence measures, and thus they are able to detect nonlinear dependence structures. Therefore, our methods are considered to be appropriate in linear as well as nonlinear regression models. Both methods are easy to implement and they perform well. We illustrate the performances of the proposed methods via simulations, and compare them with two benchmark methods, stepwise Akaike information criterion and lasso. In several cases with linear dependence all four methods turned out to be comparable. In the presence of nonlinear or uncorrelated dependencies, we observed that our proposed methods may be favourable. An application of the proposed methods to a real financial data set is also provided.

Keywords: variable selection; nonlinear regression; maximal correlation; distance correlation

1. Introduction

Recent improvements in data collection technologies give rise to complex regression problems where the number of candidate predictor variables explaining the response variable may be very large. In most of these regression problems the main task is to select the most important predictors explaining the response, and removing the others from the model. There is an extensive statistical literature in this type of screening problem, usually referred to as the *variable selection problem*. In some situations, identifying the most important variables may be the only concern. More often, the model is used for predictions where one wishes to avoid bias in estimating coefficients, and also wishes to get stable models where small changes in data does not result in entirely different models and predictions. In the existing variable selection literature, most of the work is devoted to (generalized) linear models.

In this study we propose two novel variable selection methods for linear and nonlinear regression models. The tools we employ in our methods are two powerful dependence measures, maximal correlation (MC) and distance correlation (DC). Our rationale for employing these measures

*Corresponding author. Email: yenigun@bilkent.edu.tr is that MC fully satisfies, and DC partially satisfies the Rényi [1] postulates for dependence measures. This means in addition to having several other desirable properties, both measures are able to detect nonlinear dependence structures. Therefore variable selection criteria based on these two measures are considered to be appropriate in linear and nonlinear regression models. For simplicity, we employ both measures in a stepwise regression setting as alternative comparison criteria for predictor variables to enter the model. The proposed stepwise procedures can easily be implemented. However, our proposed model selection criteria could alternately be applied via other procedures, such as stagewise regression. We carried out extensive simulations in order to compare the performances of the proposed methods with common methods such as stepwise Akaike information criterion (AIC) and lasso. The performances of the proposed methods and the benchmark methods turned out to be comparable for cases with linear dependence. When we introduced nonlinear or uncorrelated dependencies, we observed that our methods perform better.

The rest of the paper is organized as follows. In Section 2, we give a general review of the variable selection problem. In Section 3, we review the Rényi postulates for dependence measures, and give the definitions of MC and DC. We describe the proposed methods in Section 4, followed by an illustration and an application to a real data set in Section 5. The simulation results are given in Section 6, and the paper concludes in Section 7.

2. Variable selection

Consider the linear regression model

$$Y = X\beta + \epsilon, \tag{1}$$

where Y is a vector of length n representing the response variable, X is an n by p design matrix, β is a vector of length p containing regression coefficients, and ϵ is a vector of length n containing independent normal noise terms. The essential goal in variable selection is to divide X into the set of active terms X_A and the set of inactive terms X_I. Older variable selection methods such as *stepwise regression* and *all-subsets regression* can be classified as *subset selection methods*. These methods simply pick predictors and estimate the model coefficients using standard techniques such as least squares or maximum likelihood. In general, there are two important issues in subset selection methods. The first issue is to find a reasonable comparison criterion for two candidates for X_A. The most commonly used criteria are the AIC given by Sakamoto et al.,[2] the *Bayesian information criterion* (BIC) given by Schwarz,[3] and Mallows' C_p given by Mallows.[4] All three criteria are similar in that they all seek a balance between lack of fit and complexity. As an alternative to these criteria, it is common to use a computationally intensive criteria such as cross validation or predictive residual sum of squares. The second issue is computational. Note that if there are k predictor variables, there are $2^k - 1$ possible subsets which are candidates for X_A. Then one has to come up with a reasonable computational method to deal with this potentially large number of comparisons. Perhaps the most commonly used method for reducing this search space is *forward stepwise regression*. In short, a forward stepwise regression procedure considers all predictor variables individually in the first step, and finds the one that minimizes a given comparison criterion such as AIC. For the remaining steps, new terms are added such that the comparison criterion is minimum. When all the terms have entered the model, or when addition of a new term increases the selection criterion, the procedure stops. Other commonly used methods are modifications of forward stepwise regression such as *backward stepwise, stagewise, or leaps-and-bounds regression*. For a general treatment of all these classical subset selection methods, see, for example, Miller.[5]

More recent methods such as *ridge regression*, *lasso* [6] and *least angle regression* [7] provide an alternative to the subset selection methods summarized above, and they may be classified as *shrinkage methods*. The advantage of shrinkage methods is that they are not only concerned with variable selection since they employ all the candidate predictor variables, but they also modify the estimation procedures for the coefficients. Ridge regression does not perform a variable selection, but it produces smaller coefficients for the unimportant predictors than they would have under ordinary least squares. The lasso is similar to ridge regression, but it reduces some of the coefficients to zero and yields a natural variable selection. Least angle regression (LAR) is similar to stagewise regression, but much faster.

As various fields require variable selection techniques for analysing complex data structures, variable selection is still a very active research area with a stronger emphasis on the analysis of ultra high-dimensional data. Recent methods include the *smoothly clipped absolute deviation*,[8] the *elastic net*,[9] *adaptive lasso*,[10] and *sure independent screening*.[11] Similar to our study,[12] consider employing DC in variable selection, and propose the DC-based sure independence screening method. As a very good review paper on variable selection, see Fan and Lv.[13]

In this study we propose two forward stepwise procedures. The first procedure uses *partial MC* and the second procedure uses *partial DC* as the comparison criterion. Both criteria will be explained in detail in Section 4.

3. MC and DC

In virtually any field of statistics, there is a need for measuring the dependence between random variables. There are several measures of dependence in the statistical literature, which can be classified into three groups. The first group is the bivariate correlation based measures such as the *product moment correlation*, *Kendall's* τ, and *rank correlation*. All these correlations intend to measure the strength of the relationships between two variables, which usually refers to the strength of the tendency to move in the same direction. In the second group are the dependence measures based on distribution or density functions. For example, the problem of measuring the dependence between two variables can be considered as the problem of measuring the distance between their joint distribution function and the product of their marginal distribution functions. The third group consists of the dependence measures for cross classifications such as the *mean square contingency* (chi-square statistic) and *Goodman and Kruskal's* λ *and* τ. For a comprehensive survey on dependence measures, see Liebetrau.[14]

In this section we review the two dependence measures that we will employ in variable selection in regression, MC and DC. MC can be considered in the first group of dependence measures as described above, and DC can be considered in the second group. We are particularly interested in these measures since MC fully satisfies, and DC partially satisfies the Rényi postulates for dependence measures. Introduced by Rényi in 1959, these postulates are generally accepted as a complete list of postulates that a good dependence measure must satisfy. We first list the postulates, then we give definitions of MC and DC, indicating which postulates they satisfy. Since both measures are sensitive to nonlinear dependence structures, variable selection criteria based on them are considered to be appropriate in nonlinear regression models.

Consider two random variables X and Y defined on a given probability space. According to Rényi,[1] a measure of dependence $\delta(X, Y)$ of these variables should satisfy the following postulates.

(A) $\delta(X, Y)$ is defined for any X,Y neither of which is constant with probability 1.
(B) $\delta(X, Y) = \delta(Y, X)$.
(C) $0 \leq \delta(X, Y) \leq 1$.

(D) $\delta(X,Y)=0$ if and only if X and Y are independent.
(E) $\delta(X,Y)=1$ if either $X=g(Y)$ or $Y=f(X)$, where $g(\cdot)$ and $f(\cdot)$ are Borel-measurable functions.
(F) If the Borel-measurable functions $g(\cdot)$ and $f(\cdot)$ map the real axis in a one-to-one way to itself, then $\delta(f(X),g(Y))=\delta(X,Y)$.
(G) If the joint distribution of X and Y is normal, then $\delta(X,Y)=|\rho(X,Y)|$, where $\rho(X,Y)$ is the product moment correlation coefficient of X and Y.

After listing these seven postulates, Rényi [1] considers five classical bivariate measures of dependence, and notes that of these five, only MC satisfies all of the seven postulates.

3.1. *Maximal correlation*

The MC S between X and Y is defined as

$$S(X,Y)=\sup_{f,g}\rho(f(X),g(Y)), \tag{2}$$

where the supremum is taken over all Borel-measurable functions of X and Y with finite and positive variance. Here $\rho(U,V)$ denotes the product moment correlation coefficient between the random variables U and V. As mentioned above, MC satisfies all seven postulates given by Rényi.[1] Here, perhaps we must give more emphasis on postulate D, indicating that the MC vanishes if and only if the variables are independent. Note that the commonly used product moment correlation satisfies B, C, and G only. The important postulate D is not satisfied, in other words, two variables may be uncorrelated but dependent. This is one of the well-known drawbacks of product moment correlation.

The MC is introduced by Gebelein,[15] and received considerable attention in the statistical literature. Rényi [1] gave the conditions such that the MC can be attained. Csáki and Fishcher [16] computed MC for a number of examples. Koyak [17] considered a multivariate analog of MC. For random variables that take only a finite number of values, Sethuraman [18] gave a procedure to estimate the MC from the sample, and gave the asymptotic distribution of this estimate under the null hypothesis of independence. Dembo et al. [19] and Novak [20] studied the MC between partial sums of independent and identically distributed random variables. More recently, Yenigün et al. [21] considered the computation of MC in contingency tables and proposed an independence test.

MC is an attractive measure of dependence, however, since there does not always exist functions $f_0(x)$ and $g_0(x)$ such that $S(X,Y)=\rho(f_0(X),g_0(Y))$, MC cannot be evaluated explicitly except for special cases. If this equality holds for some f_0 and g_0, we say that *the MC of X and Y can be attained.*

Let $\mathcal{L}_X^2$ denote the Hilbert space of all random variables of the form $f(X)$ for which $E(f(X))=0$ and Var$(f(X))$ is finite. Similarly, let $\mathcal{L}_Y^2$ denote the Hilbert space of all random variables of the form $g(Y)$ for which $E(g(Y))=0$ and Var$(g(Y))$ is finite. For any $f=f(X)\in\mathcal{L}_X^2$, consider the transformation

$$Af=E[E(f(X)|Y)|X]. \tag{3}$$

Rényni [1] shows that if the transformation A defined in Equation (3) is completely continuous, then the MC between X and Y is attained for $f_0(X)$ and $g_0(Y)$, where f_0 is an eigenfunction belonging to the greatest eigenvalue $S^2=S^2(X,Y)$ of A and $g_0(Y)=S^{-1}E(f_0(X)|Y)$. Rényni [1] also notes that if the dependence between X and Y is *regular* and the mean square contingency is finite, then the transformation A is completely continuous. Here, *regular* dependence of the variables means that the joint distribution of the variables is absolutely continuous with respect to the direct product of their distributions.

As noted above, the problem of computing MC can be mathematically intractable, thus it cannot be evaluated analytically except in special cases. However, a practical approach to this problem is provided by Breiman and Friedman,[22] which is easily applied for estimating the MC from data. We provide some insight about this algorithm in Section 4.1.

3.2. *Distance correlation*

DC is a recent and powerful dependence measure introduced by Székely et al.[23] For all distributions with finite first moments, DC generalizes the idea of correlation in two fundamental ways. Firstly, DC is defined for variables in arbitrary dimensions, it is not limited to the bivariate case. Secondly, DC vanishes if and only if the variables are independent. DC satisfies the Rényi postulates *A, B, C, D*. The remaining postulates are partly satisfied. Postulate *E* is satisfied for linear functions and *F* is satisfied for orthogonal transformations. As for *G*, if X and Y are bivariate normal, R is a function of ρ. The formal definitions of distance covariance and DC is given in [23].

Consider a random sample $(\mathbf{X}, \mathbf{Y}) = \{(X_k, Y_k) : k = 1, \dots, n\}$ from the joint distribution of random vectors X in $\mathbb{R}^p$ and Y in $\mathbb{R}^q$. The empirical distance covariance $\mathcal{V}_n(\mathbf{X}, \mathbf{Y})$ is the nonnegative number defined by

$$\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y}) = \frac{1}{n^2} \sum_{k,l=1}^{n} A_{kl} B_{kl}, \tag{4}$$

where

$$A_{kl} = a_{kl} - \bar{a}_{k\cdot} - \bar{a}_{\cdot l} + \bar{a}_{\cdot\cdot},$$
$$B_{kl} = b_{kl} - \bar{b}_{k\cdot} - \bar{b}_{\cdot l} + \bar{b}_{\cdot\cdot}.$$

Here

$$a_{kl} = \|X_k - X_l\|_p, \quad b_{kl} = \|Y_k - Y_l\|_q, \quad k, l = 1, \dots, n,$$

$\|a\|_d$ is the Euclidean norm of a in $\mathbb{R}^d$, and the subscript $\cdot$ denotes that the mean is computed for the index that it replaces. Similarly, the empirical distance variance $\mathcal{V}_n(\mathbf{X})$ is the nonnegative number defined by

$$\mathcal{V}_n^2(\mathbf{X}) = \mathcal{V}_n^2(\mathbf{X}, \mathbf{Y}) = \frac{1}{n^2} \sum_{k,l=1}^{n} A_{kl}^2. \tag{5}$$

The empirical DC $\mathcal{R}_n(\mathbf{X}, \mathbf{Y})$ is the square root of

$$\mathcal{R}_n^2(\mathbf{X}, \mathbf{Y}) = \frac{\mathcal{V}_n^2(\mathbf{X}, \mathbf{Y})}{\sqrt{\mathcal{V}_n^2(\mathbf{X}) \mathcal{V}_n^2(\mathbf{Y})}}. \tag{6}$$

The DC statistic is implemented in the energy [24] package for R Project for Statistical Computing,[25] available under general public licence. More recent work on DC include.[12, 26–29]

4. Proposed methods

In this section we propose two stepwise variable selection methods, one based on MC, and one based on DC. We will describe the MC-based method in full detail, the DC-based method will be a straightforward modification of the former. We begin with a few remarks on computational details.

4.1. *Computational details*

As noted in Section 3, it is typically not easy to compute MC explicitly except some special cases. Therefore, we will use an algorithm by Breiman and Friedman [22] for estimating MC from data, which is quite practical to apply as the algorithm is implemented in the acepack package [30] for R. Breiman and Friedman [22] consider the problem of replacing the response variable Y and the predictor variables $X_1, \dots, X_p$ by functions $\theta(Y)$ and $\phi_1(X_1), \dots, \phi_p(X_p)$. Given the sample only, they discuss a procedure for estimating the functions θ^* and $\phi_1^*, \dots, \phi_p^*$ that minimize $e^2 = E\{[\theta(Y) - \sum_{j=1}^{p} \phi_j(X_j)]^2\}/\mathrm{var}[\theta(Y)]$ while making minimal assumptions on data distribution and the form of the functions. Their algorithm is referred to as the alternating conditional expectations (ACE) algorithm. For the bivariate case, θ^* and ϕ^* satisfy $\rho(\theta^*, \phi^*) = \max_{\theta,\phi} \rho(\theta(Y), \phi(X))$, and thus their algorithm provides an estimate of MC between two variables. In this paper all MCs are computed using the ACE algorithm, as implemented in acepack. The second dependence measure we consider in this paper, DC, can easily be implemented using dcor function in the energy [24] package for R. In both stepwise regression procedures described below, we use the cross-validation error of the response variable to compare the steps, where the cross-validation is obtained using the leave-one-out approach.

4.2. *Stepwise regression using MC*

We first define *partial MC*. Consider random variables X, Y, and a possibly vector-valued random variable Z. Given Z, the partial MC between X and Y is computed as follows:

(1) Regress X on Z, denote the error terms by R_X.
(2) Regress Y on Z, denote the error terms by R_Y.
(3) The MC between R_X and R_Y is the partial MC between X and Y, given Z.

Then we can define a stepwise regression procedure, using MC as follows:

(1) Consider all candidate predictor variables individually and find the one which has the largest MC with the dependent variable.
(2) For the remaining steps, add one more term such that the partial MC with the dependent variable, given the previously entered variable(s), is largest.
(3) Stop when all terms have entered the model. The step with the smallest cross-validation error is the selected model.

4.3. *Stepwise regression using DC*

Simply replace the MCs with DCs in the above procedure in order to define a *partial DC* and the stepwise regression using DC.

Remark 1 More precisely, we have defined a ‘linear partial MC’ and a ‘linear partial DC’ as a natural definition for the problem of variable selection in regression.

5. Illustration and application to real data

In this section we first provide an illustration of the proposed methods on a classical data set. We then consider an application of our methods, as well as two commonly used variable selection methods, on a more timely data set we compiled for this study.

5.1. *Illustration: Swiss fertility data*

As an illustration of the proposed methods, we consider the Swiss fertility data which consists of standardized fertility measures and socio-economic indicators for each of 47 French-speaking provinces of Switzerland in about 1888.[31] Here, the response variable is a common fertility measure (*Fertility*), and the candidate predictors are percentage of males involved in agriculture as occupation (*Agriculture*), percentage of draftees receiving highest mark on army examination (*Examination*), percentage of education beyond primary school for draftees (*Education*), percentage of Catholic population (*Catholic*), and live births who live less than 1 year (*Infant Mortality*). The result of stepwise selection for MC and DC criteria are illustrated in Figure 1. For the MC-based method, the order of entering the model is *Education, Catholic, Infant Mortality, Agriculture*, and *Examination*. According to the cross-validation errors, this method excludes *Examination* from the model. As for the DC-based method, the order of entering the model is *Examination, Infant Mortality, Education, Catholic*, and *Agriculture*. This method includes all candidate predictors in the model. Note that in this example the model returned by the MC-based method has a lower cross-validation error.

5.2. *Application: S&P 500 returns*

Here, we consider an application of the proposed methods on a real data set, and compare them with the commonly used variable selection methods stepwise AIC and lasso.[6] The data set consists of the monthly returns of S&P 500 index and the values of 11 candidate predictors observed between January 1989 and December 2007 ($n = 216$). The data set is available upon contact with

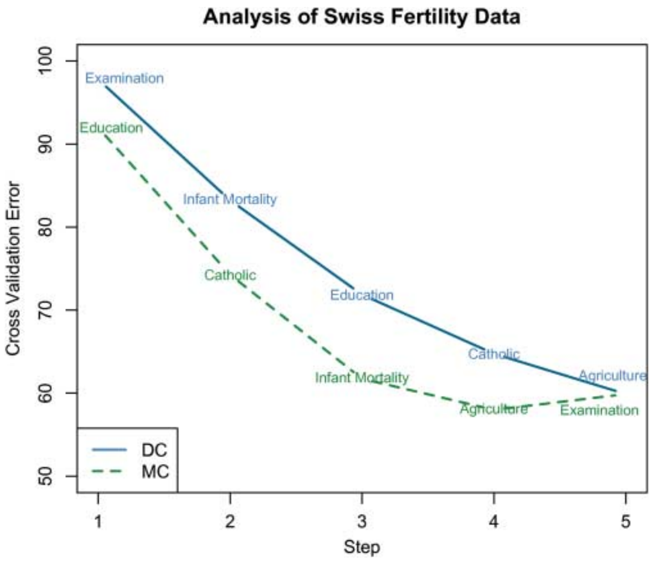

Figure 1. Analysis of Swiss fertility data. Cross-validation errors for each step are displayed for DC and MC.

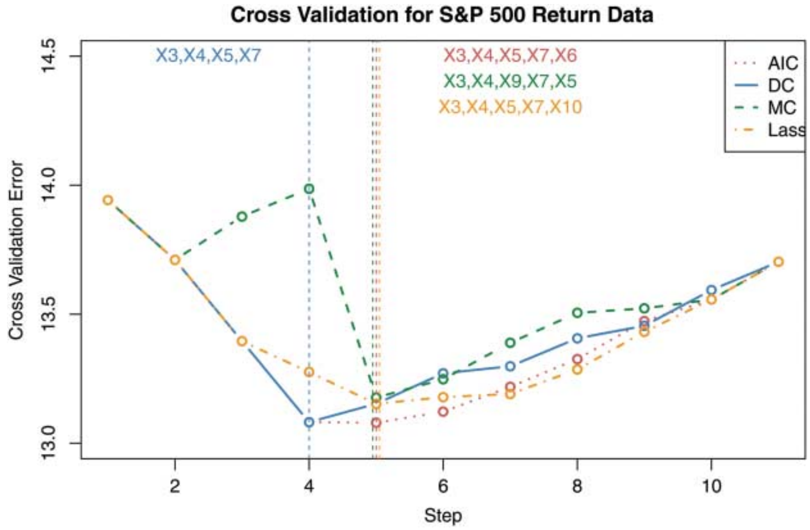

Figure 2. Analysis of S&P 500 monthly returns data. Cross-validation errors for each step are displayed for stepwise AIC, DC, MC, and lasso methods. Vertical lines indicate the selected models.

the authors. The response variable here is the monthly returns, and the candidate predictors are dividend yield (X_1), earnings yield (X_2), volatility index (X_3), unexpected volatility (X_4), inflation rate (X_5), change in inflation rate (X_6), 90-day treasury bill (X_7), industrial production index growth (X_8), credit spread (X_9), term spread (X_{10}), and yield spread (X_{11}). The steps and cross-validation errors of all four methods considered in this study are summarized in Figure 2. For each method, the returned model is indicated by the related colour and style. The DC method returns the model $Y \sim X_3, X_4, X_5, X_7$, which has the smallest cross-validation error among all other models returned. All these predictor variables are common in the remaining three models. The MC method adds X_9, AIC adds X_6, and lasso adds X_{10} to the model.

6. Simulation results

We illustrate the performance of the proposed variable selection methods with an extensive simulation study, where we compare our methods with the commonly used stepwise AIC and lasso. We consider six cases with different dependence structures between the predictors and the response variable. The cases are set up such that there are p candidate predictor variables, but only q of them ($q < p$) have direct influence on the response variable. We first present the results in detail for $p = 8$, then we summarize our findings for $p = 20$. For each case we generate $N = 100$ samples of size $n = 100$, perform variable selection using all four methods under consideration, and report the frequencies of the selected models for each method. In what follows, $N(a, b)$ denotes the normal distribution with mean a and variance b^2, $U(a, b)$ denotes the continuous uniform distribution on (a, b).

6.1. *Results for p = 8*

Case 1 Linear relations. In this case we consider a total of $p = 8$ candidate predictors having independent standard normal distributions, $q = 3$ of which are related with the dependent variable via the linear model (1), where $\beta = [1, 1, 1, 0, 0, 0, 0, 0]$ and $\epsilon \sim N(0, \sigma = 2)$.

Case 2 Nonlinear relations. We consider a total of $p = 8$ candidate predictors from the following distributions: $X_1 \sim N(0,1)$, $X_2 \sim N(0,2)$, $X_3 \sim U(-1.5,1.5)$, $X_4,\dots,X_8 \sim U(-1,1)$. The first $q = 4$ are related with the dependent variable via:

$$Y = \log[4 + \sin(3X_1) + \sin(X_2) + X_3^2 + X_4 + 0.1\epsilon],$$

where $\epsilon \sim N(0,\sigma = 1)$.

Case 3 Dependent but uncorrelated variables. We consider a total of $p = 8$ candidate predictors from the following distributions: $X_1 \sim N(0,1.4)$, $X_2 \sim U(-1.7,1.7)$, $X_3 \sim N(0,0.8)$, $X_4,\dots,X_8 \sim N(0,1)$. Let us define $Y_1,\dots,Y_3$ as follows:

$$Y_1 = |X_1|, \quad Y_2 = X_2^2, \quad Y_3 = X_3^2.$$

It can be shown that the pairs (X_i, Y_i), $i = 1,2,3$, are uncorrelated. We define the dependent variable as

$$Y = |X_1| + X_2^2 + X_3^2.$$

Case 4 Constant collinearity among predictors. We consider a total of $p = 8$ candidate predictors from a multivariate normal distribution, $X \sim N_P(0, \Sigma)$, where

$$\Sigma = \begin{bmatrix} 1 & \theta & \cdots & \theta \\ \theta & 1 & \cdots & \theta \\ \vdots & \vdots & \ddots & \vdots \\ \theta & \theta & \cdots & 1 \end{bmatrix}.$$

We set $\theta = 0.6$. The first $q = 3$ of these variables are related with the dependent variable via:

$$Y = X\beta + \epsilon,$$

where $\beta = [1,1,1,0,0,0,0,0]$ and $\epsilon \sim N(0,\sigma = 2)$.

Case 5 Toeplitz-type collinearity among predictors. This is the same as Case 4, but

$$\Sigma = \begin{bmatrix} 1 & \theta & \theta^2 & \cdots & \theta^{p-1} \\ \theta & 1 & \theta & \cdots & \theta^{p-2} \\ \theta^2 & \theta & 1 & \cdots & \theta^{p-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \theta^{p-1} & \theta^{p-2} & \theta^{p-3} & \cdots & 1 \end{bmatrix}.$$

Case 6 A generalized linear model (Gamma Regression) In this case we generate the response and the predictors from a generalized linear model, namely, gamma regression. A total of $p = 8$ candidate predictor variables follow standard normal distribution, $q = 3$ of them are related with the response via the linear predictor $L = X\beta$, where $\beta = (0.25,0.25,0.25,0,0,0,0,0)$, and X is the $n \times p$ matrix representing the p predictor variables. The link function is the log function, thus the mean vector of the responses are $\hat{\mu} = e^L$. Then the responses are generated from a gamma distribution with mean $\hat{\mu}$ and unit variance. In this case we used the response residuals for computing partial MC and partial DC.

For each case we present two graphs. The first graph gives proportions of the three most frequent models returned by each method, plus a ratio we call the *hit rate*, the rate of models containing all q true predictors. The second graph contains the individual proportions of each candidate predictor variable to be included in the models returned.

Figures 3–8 summarize our simulation results. In Case 1, linear relations, all methods seem to perform well as the hit rates and individual proportions for detecting the true predictors are

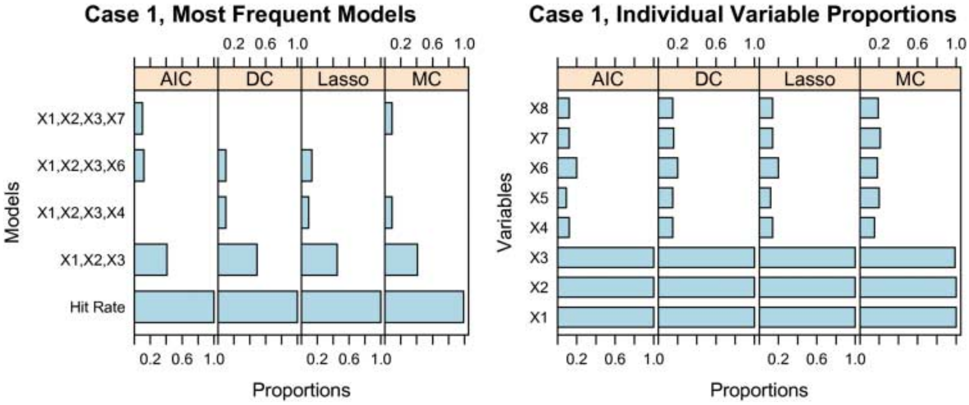

Figure 3. Simulation results for Case 1, linear relations. True model is X1, X2, and X3.

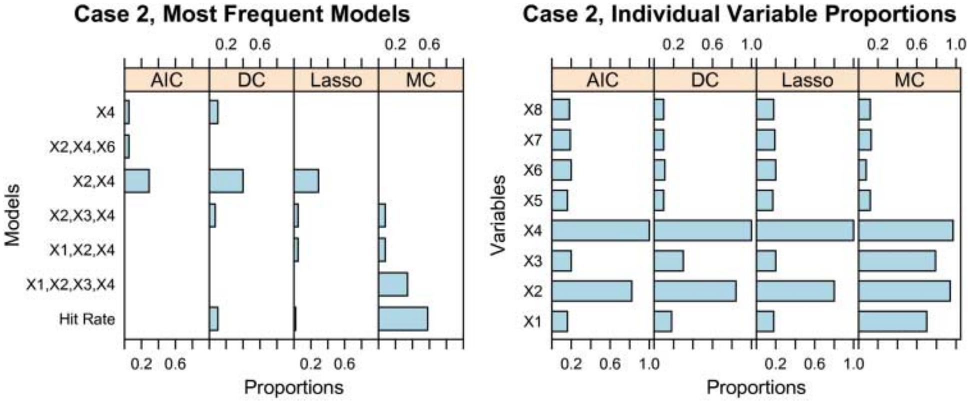

Figure 4. Simulation results for Case 2, nonlinear relations. True model is X1, X2, X3, and X4.

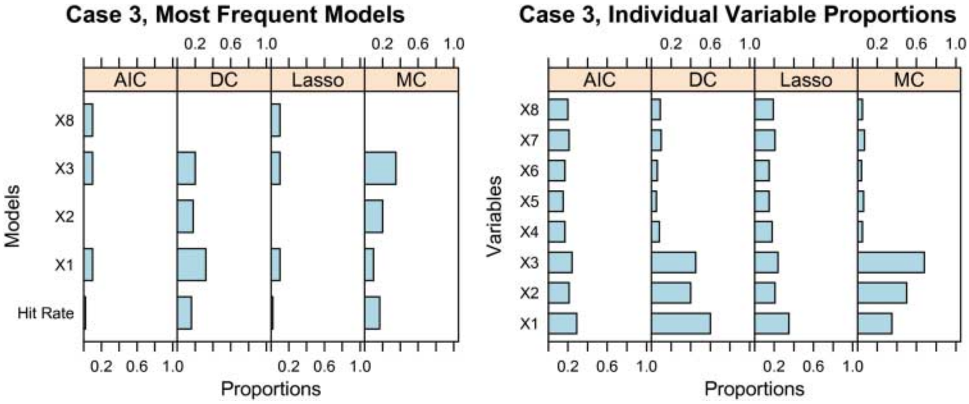

Figure 5. Simulation results for Case 3, dependent but uncorrelated variable. True model is X1, X2, and X3.

very high. The proportion of returning the exact true model is largest for DC method. In Case 2, nonlinear relations, the performances go down for all methods. Here, MC method stands out with the largest hit rate and proportion of returning the exact model. The remaining three methods seem to be unable to detect the true predictors X_1 and X_3. In Case 3, dependent but correlated

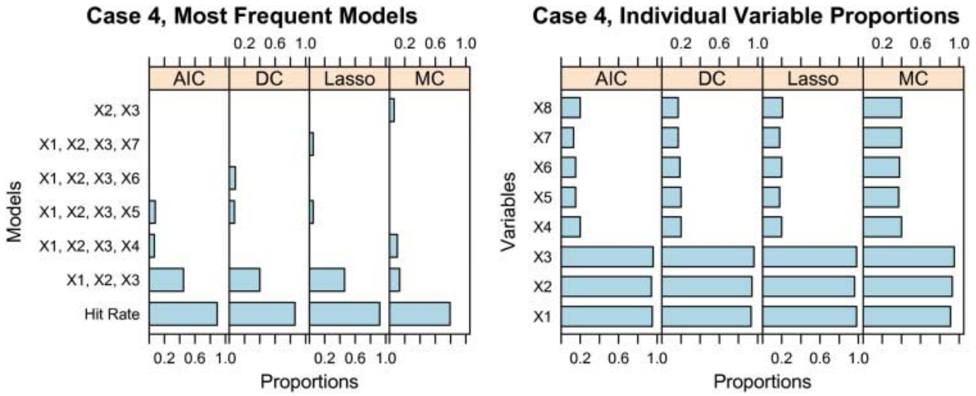

Figure 6. Simulation results for Case 4, constant collinearity among predictors. True model is X1, X2, and X3.

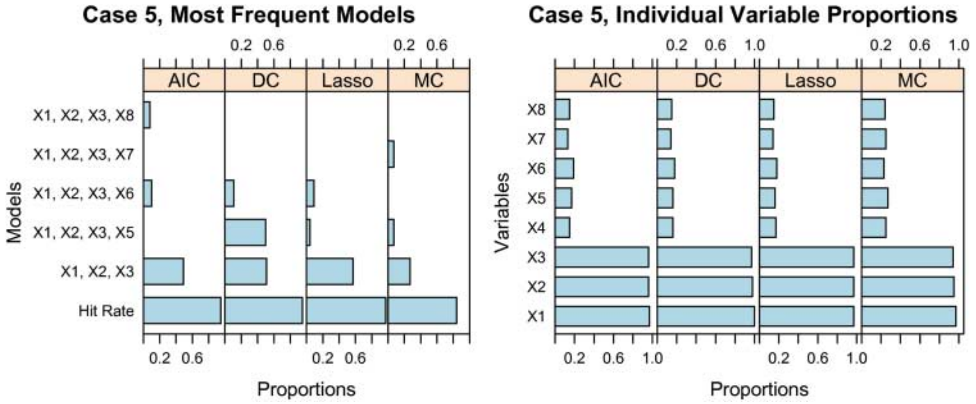

Figure 7. Simulation results for Case 5, Toeplitz type collinearity among predictors. True model is X1, X2, and X3.

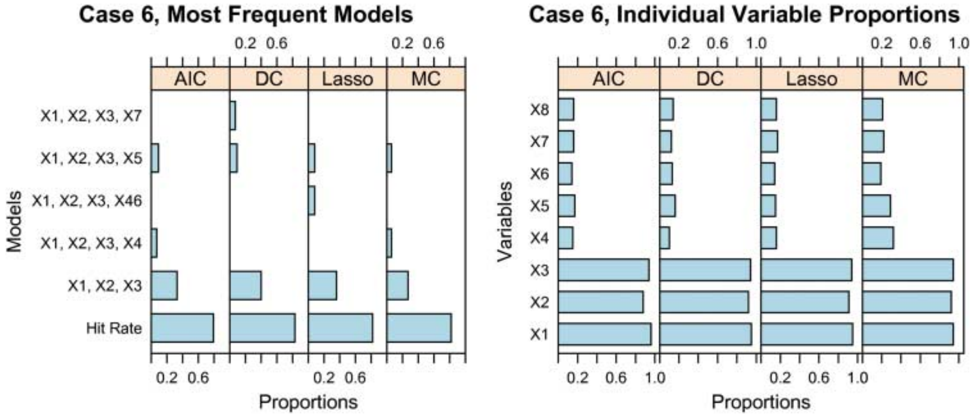

Figure 8. Simulation results for Case 6, gamma regression. True model is X1, X2, and X3.

variables, hit rates are larger for MC and DC methods, but the ability to detect the exact model for all methods is very low for all methods. In terms of individual detection proportions, MC and DC outperform the benchmark methods (AIC and lasso). In Cases 4 and 5, constant collinearity and Toeplitz-type collinearity among the predictors, the performances of all four methods are good

Table 1. Simulation results for models in dimension $p = 20$.

Case	Method	Model	Percentage	X_1	X_2	X_3	X_4
1	AIC	X_1, X_2, X_3	3	100	100	100	–
	DC	X_1, X_2, X_3	10	100	100	100	–
	Lasso	X_1, X_2, X_3	5	100	100	100	–
	MC	X_1, X_2, X_3	24	99	96	97	–
2	AIC	X_2, X_4	3	20	75	18	99
	DC	X_2, X_4	11	45	86	63	99
	Lasso	X_2, X_4	3	20	76	18	99
	MC	X_1, X_2, X_3, X_4	16	75	91	78	96
3	AIC	X_1	2	24	2	19	–
	DC	X_1	24	73	51	48	–
	Lasso	X_1	2	24	18	21	–
	MC	X_3	34	44	46	62	–
4	AIC	X_1, X_2, X_3	4	97	94	94	–
	DC	X_1, X_2, X_3	10	96	93	93	–
	Lasso	X_1, X_2, X_3	17	98	96	94	–
	MC	X_1, X_2, X_{12}	2	85	76	86	–
5	AIC	X_1, X_2, X_3, X_{20}	6	100	98	99	–
	DC	X_1, X_2, X_3	11	99	98	99	–
	Lasso	X_1, X_2, X_3	10	100	100	100	–
	MC	X_1, X_2, X_3	12	86	92	84	–
6	AIC	X_1, X_2, X_3	6	88	87	86	–
	DC	X_1, X_2, X_3	15	90	01	87	–
	Lasso	X_1, X_2, X_3	8	88	89	86	–
	MC	X_1, X_2, X_3	7	79	78	78	–

Note: The percentage of the most frequent model returned by each method is given, along with the individual detection percentage for each true predictor variable. The true models are $X_1, X_2,$ andX_3 for all cases except Case 2, where the true model is $X_1, X_2, X_3,$ and X_4.

and similar, except the MC method is a little weaker. In Case 6, gamma regression, we observe that the DC method has the highest frequency of detecting the exact model.

The simulation results indicate that for cases with linear dependence with no or some collinearity (Cases 1, 4, 5), our methods are comparable with the benchmark methods. When we introduce nonlinear dependence structures between the response and predictors (Case 2), we observed that the MC-based method outperforms the others. When we defined the response variable as a linear combination of variables that are dependent but uncorrelated with the responses (Case 3), both proposed methods outperformed the benchmark methods. This is because both MC and DC vanish if and only if the two variables are independent. When the underlying model is the gamma regression model which also has a nonlinear nature (Case 6), we observe that the DC outperforms all other methods.

6.2. *Results for p* $= 20$

In this section we increase the number of candidate predictor variables to 20 and study the effect of the increased dimension on the performances of the variable selection methods discussed above. We consider the same cases, except we now look for the same number of true predictors among 20 predictors generated in the same fashion. Due to larger number of variables, rather than giving the full simulation details as we did for $p = 8$, here we only give a summary of our findings. Table 1 presents the percentage of the most frequent model returned by each method, along with the individual detection percentage for each true predictor variable. For Cases 2, 3 and 4, increasing the number of candidate predictors does not seem to effect the empirical comparison results to a great extent. However, AIC severely suffers from the increased dimension in Cases 1, 5 and 6; lasso severely suffers in Cases 1 and 6; and the MC method severely suffers in Case 6. In Cases 1 and 2, the MC method performs significantly better than the others. In all of the cases studied for $p = 8$ and $p = 20$, we can say that the DC method either has the best or an acceptable performance.

7. Conclusions

In this study we considered two new variable selection methods for linear and nonlinear regression models. The first method is a stepwise procedure which uses the partial MC as the criterion for adding a variable to the model at each step. The second method is similar, except it uses partial DC as the criterion. Both methods perform well and they can be easily implemented. We carried out an extensive simulation study to compare the performances of the proposed methods with two benchmark methods, stepwise AIC and lasso. In many cases, the performances of the proposed methods were comparable with the benchmark methods. In the presence of nonlinear or uncorrelated dependencies, our methods turned out to perform better. We also observed that the benchmark methods may suffer from increased number of candidate predictor variables, while the proposed methods still have an acceptable performance.

Acknowledgements

The authors thank to Mustafa Fırat for compiling the data set studied in Section 5.

Funding

Deniz Yenigün's research was supported by The Scientific and Technological Research Council of Turkey, [grant 110T668].

References

[1] Rényi A. On measures of dependence. Acta Math Acad Sci Hungar. 1959;10:441–451.
[2] Sakamoto Y, Ishiguro M, Kitagawa G. Akaike information criterion statistics. Tokyo: D. Reidel; 1986.
[3] Schwarz G. Estimating the dimension of a model. Ann Stat. 1978;6:461–464.
[4] Mallows CL. Some Comments on C_p. Technometrics. 1973;15(4):661-675.
[5] Miller, A. Subset selection in regression. London: Chapman and Hall/CRC; 2002.
[6] Tibshirani R. Regression shrinkage and selection via LASSO. J R Stat Soc B. 1996;58:267–288.
[7] Efron B, Hastie T, Johnstone I, Tibshirani R. Least angle regression (with discussion). Ann Stat. 2004;32:409–499.
[8] Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc. 2001;96:1348–1360.
[9] Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc B. 2005;67:301–320.
[10] Zou H. The adaptive lasso and its oracle properties. J Am Stat Assoc. 2006;101:1418–1429.
[11] Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space (with discussion). J R Stat Soc B. 2008;70:849–911.
[12] Li R, Zhong W, Zhu L. Feature screening via distance correlation learning. J Am Statist Assoc. 2012;107(499):1129–1139.
[13] Fan J, Lv J. A selective overview of variable selection in high dimensional feature space. Statist Sin. 2010;20(1): 101–148.
[14] Liebetrau AM. Measures of association. London: Sage Publications; 2005.
[15] Gebelein, H. Das statistische problem der korrelation als variations - und eigenwerthproblem und sein zusammenhang mit der ausgleichsrechnung. Z Angew Math Mech. 1941;21:364–379.
[16] Csáki P, Fischer J. On the general notion of maximum correlation. Magyar Tudományos Akad Mat Kutató Intézetenk Közleményei (publ. Math. Inst. Hungar Acad Sci). 1963;8:27–51.
[17] Koyak R. On measuring internal dependence in a set of random variables. Ann Stat. 1987;15:1215–1228.

[18] Sethuraman J. The asymptotic distribution of Rényi maximal correlation. Comm Stat Theory Methods. 1990;19:4291–4298.

[19] Dembo A, Kagan A, Shepp L. Remarks on the maximum correlation coefficient. Bernoulli. 2001;7:343–350.

[20] Novak, S. On Gebelein's correlation coefficient. Stat Probab Lett. 2004;69:299–303.

[21] Yenigün CD, Székely GJ, Rizzo ML. A test of independence in two way contingency tables based on maximal correlation. Comm Stat: Theory Methods. 2011;40:2225–2242.

[22] Breiman L, Friedman J. Estimating optimal transformations for multiple regression and correlation (with discussion). J Amer Statist Assoc. 1985;80:580–619.

[23] Székely GJ, Rizzo ML, Bakirov NK. Measuring and testing dependence by correlation of distances. Ann Stat. 2007;35:2769–2794.

[24] Rizzo ML, Székely GJ. energy: E-statistics (energy statistics). R package version 1.5.0; 2013.

[25] R Development Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2013, ISBN 3-900051-07-0. Available from: http://www.R-project.org.

[26] Székely G, Rizzo ML. Brownian distance correlation. Ann Appl Stat. 2009;3(4):1236–1265.

[27] Székely GJ, Rizzo ML. On the uniqueness of distance covariance. Stat Probab Lett. 2012;82:2278–2282.

[28] Dueck J, Edelmann D. Gneiting T, Richards D. The affinely invariant distance correlation. arXiv:1210.2482v1 [math.ST] 9 Oct 2012.

[29] Lyons R. Distance covariance in metric spaces. Ann Prob. 2013;41(5):3051–3696.

[30] Spector P, Friedman J, Tibshirani R, Lumley T. acepack: ACE and AVAS methods for choosing regression transformations. R package version 1.3-3.2; 2012.

[31] Mosteller F, Tukey JW. Data analysis and regression: a second course in statistics. Reading, MA: Addison-Wesley; 1977.