5-22-2019

Forward Selection via Distance Correlation

Ty Adams
adamstj1@rose-hulman.edu

FORWARD SELECTION VIA DISTANCE CORRELATION FOR VARIABLE SELECTION

A DISSERTATION
SUBMITTED TO THE DEPARTMENT OF MATHEMATICS
OF ROSE-HULMAN INSTITUTE OF TECHNOLOGY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
BACHELOR OF SCIENCE

Ty Adams
May 2019

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Bachelor of Science.

() Principal Adviser

Approved by the Department of Mathematics.

Contents

1 Introduction **1**

2 Theoretical Background **3**
2.1 Pearson's Correlation 3
2.2 Classical Forward Selection 3
2.3 Distance Correlation 4
2.4 False Selection Rate 5

3 Forward Selection via Distance Correlation **7**

4 The Selection of α_{min} **10**

5 Numerical Simulations **12**
5.1 Linear Models 12
5.2 Non-Linear Models 13

6 Results **14**

7 Conclusions **18**

List of Tables

1.1 Diabetes study: A subset of data from[4]. It is of interest to determine which variables are related to the response. 2

3.1 Theoretical set of Partial Distance Correlation for an example 7
3.2 Theoretical set of Partial Distance Correlation for an example 7
3.3 The full model after forward selection via distance correlation ran . . 8
3.4 Monotone ordering of Correlations for the full model 8
3.5 Unique correlation thresholds to consider and their respective number of variables allowed in the model 8
3.6 The false selection rates for the correlation thresholds still in consideration 8

6.1 variables selected by each of the methods tested for the diabetes data 17

List of Figures

2.1 Plot of $Y = X^2$ showing that Pearson's correlation of zero does not imply distance correlation of 0 5

4.1 Average and Median Density plots of Correlations 10
4.2 Minimum and Maximum Density plots of Correlations 11

6.1 Summarization of classification measures and false selection rates for linear case with independent predictors; we compare the adaptive lasso (ALASSO), classical forward selection (FS), and forward selection tuned with distance correlation (DCFS). 15
6.2 Summarization of classification measures and false selection rates for linear case with correlated predictors; we compare the adaptive lasso (ALASSO), classical forward selection (FS), and forward selection tuned with distance correlation (DCFS). 15
6.3 Summarization of classification measures and false selection rates for non-linear sine case; we compare the adaptive lasso (ALASSO), classical forward selection (FS), and forward selection tuned with distance correlation (DCFS). 16
6.4 Summarization of classification measures and false selection rates for non-linear spread case; we compare the adaptive lasso (ALASSO), classical forward selection (FS), and forward selection tuned with distance correlation (DCFS). 16

Chapter 1

Introduction

Variable selection is the process of selecting a set predictor variables that are related to the response from a larger set of possible predictor variables. This method helps with trying to predict a response, when given other characteristics that may be able to explain the data observed. There are several different automated ways to determine which set of predictors are best, but two well known methods are adaptive LASSO and forward selection. While these methods are good, they have one underlying assumption, they assume linearity between the response and the predictors considered. Since this is the case, they struggle to fit nonlinear data well.

There are few methods for performing variable selection with nonlinearity. We propose a variable selection procedure based on distance correlation, a nonlinear correlation structure. This is because, forward selection uses Pearson's correlation coefficient, which is a linear form of correlation, and the most well known. With this being the case, we propose using distance correlation in a forward selection procedure.

As a motivating example, consider the diabetes study described in [4]. Table 1 shows a subset of the data. The age, sex, body mass index, blood pressure, and six blood serum measurements, were obtained for $n = 442$ diabetes patients, as well as a response y, which is a quantitative measure of disease progression a year after the baseline. The goal for this research was to construct a predictive model for the response y, based on the other variables, that would suggest which covariates were important factors in disease progression.

Table 1.1: Diabetes study: A subset of data from[4]. It is of interest to determine which variables are related to the response.

	AGE	SEX	BMI	BP	Serum measurements						Response
Patient	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	y
1	59	2	32.1	101	157	93.2	38	4	4.9	87	151
2	48	1	21.6	87	183	103.2	70	3	3.9	69	75
3	72	2	30.5	93	156	93.6	41	4	4.7	85	141
4	24	1	25.3	84	198	131.4	40	5	4.9	89	206
5	50	1	23.0	101	192	125.4	52	4	4.3	80	135
6	23	1	22.6	89	139	64.8	61	2	4.2	68	97
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
441	36	1	30.0	95	201	125.2	42	5	5.1	85	220
442	36	1	19.6	71	250	133.2	97	3	4.6	92	57

The rest of this paper is organized as follows. Chapter 2 will discuss the theoretical background of Pearson's Correlation, distance correlation, forward selection, and false selection rate. Forward selection via distance correlation will be established in Chapter 3. Chapter 4 will present the numerical simulations that were ran to test the model. Chapter 5 will discuss the results of the numerical simulations. Lastly, Chapter 6 will be our conclusions.

Chapter 2

Theoretical Background

2.1 Pearson's Correlation

The most well known measure of correlation, is Pearson's correlation coefficient. Pearson's correlation coefficient measures the strength of a linear relationship between two variables. Pearson's correlation coefficient, r, is calculated the following way:

$$Cov(X, Y) = \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y]$$

$$Var(x) = Cov(X, X)$$

$$r = \frac{Cov(X, Y}{\sqrt{Var(x)Var(Y)}}$$

Pearson's correlation is bounded between -1 and 1, which can be shown by using the Cauchy-Schwartz Inequality. A negative correlation relates to an inverse linear relationship between the two variables, while a positive correlation relates to a direct linear relationship.

2.2 Classical Forward Selection

Classical forward selection is the most well known model and variable selection algorithms. The framework for forward selection uses Pearson's correlation coefficient and partial correlation to build a model from no predictors included, and adds one in at a time. Partial correlation, is the same as Pearson's correlation, but it takes into account other variables, holding them constant for both the response and predictor considered [7].

Consider an example where we have a response variable y and 10 predictors variables we want to consider. The first step in the algorithm will select the predictor

that has the highest correlation with the response, lets call this x_1. The next step is to consider the partial correlations of the rest of the predictors with the response y, which means we hold x_1 constant for both the predictor considered and y. This step is then repeated until some threshold is met, correlation falls below a certain value. While, this is not the only way to do forward selection, but it is the one that will most resemble our new method.

2.3 Distance Correlation

Distance correlation is a measure of dependence between two variables that measures the distance between their two characteristic functions. This was first introduced in 2007 by Szekely, Rizzo, and Bakirov [2]. To get a correlation though, you first need to define a covariance/variance. The way that the distance covariance is defined is by the following: $||f_{X,Y}(t,s) - f_X(t)f_Y(s)||$ where f is the respective characteristic function of the given variables[2]. Since characteristic functions always exist, the covariance is always defined, and therefore correlation always exist. Sometimes there is not a closed form characteristic function for a variable. However, a nice empirical way to estimate the distance covariance is the following:

$$\begin{aligned} a_{kl} &= ||X_k - X_l||_p & \bar{a}_{k.} &= \frac{1}{n}\sum_{l=1}^{n} a_{kl} \\ \bar{a}_{.l} &= \frac{1}{n}\sum_{k=1}^{n} a_{kl} & \bar{a}_{..} &= \frac{1/n^2}{\sum_{k,l=1}^{n}} a_{kl} \\ A_{kl} &= a_{kl} - \bar{a}_{k.} - \bar{a}_{.l} - \bar{a}_{..} \end{aligned} \tag{2.1}$$

The distance covariance is then estimated by the following:

$$\mathscr{V}(X,Y) = \sqrt{\frac{1}{n^2}\sum_{k,l=1}^{n} A_{kl}B_{kl}} \tag{2.2}$$

where B_{kl} is similarly defined as A_{kl} but for Y instead of X [2]. Then the distance correlation is calculated the following way:

$$\mathscr{D}(X,Y) = \frac{\mathscr{V}(X,Y)}{\sqrt{\mathscr{V}(X,X)\mathscr{V}(Y,Y)}} \tag{2.3}$$

This yields some nice properties.

1. $0 \leq \mathscr{D}(X,Y) \leq 1$.
2. $\mathscr{D}(X,Y) = 0 \; iff \; X$ and Y Independent.

3. If $\mathscr{D}(X, Y) = 1$, then there exist a vector a, a nonzero real number b, and an orthogonal matrix C, such that $Y = a + bXC$.

A proof of all of these properties can be found in the paper by Szekely, Rizzo, and Bakirov [2].

To further establish that distance correlation and Pearson's correlation are different, lets examine Figure 2.1. This is a plot of $X = Unif(-1, 1)$ and $Y = X^2$. This is a classic example of where Pearson's correlation is zero, but distance correlation is non-zero.

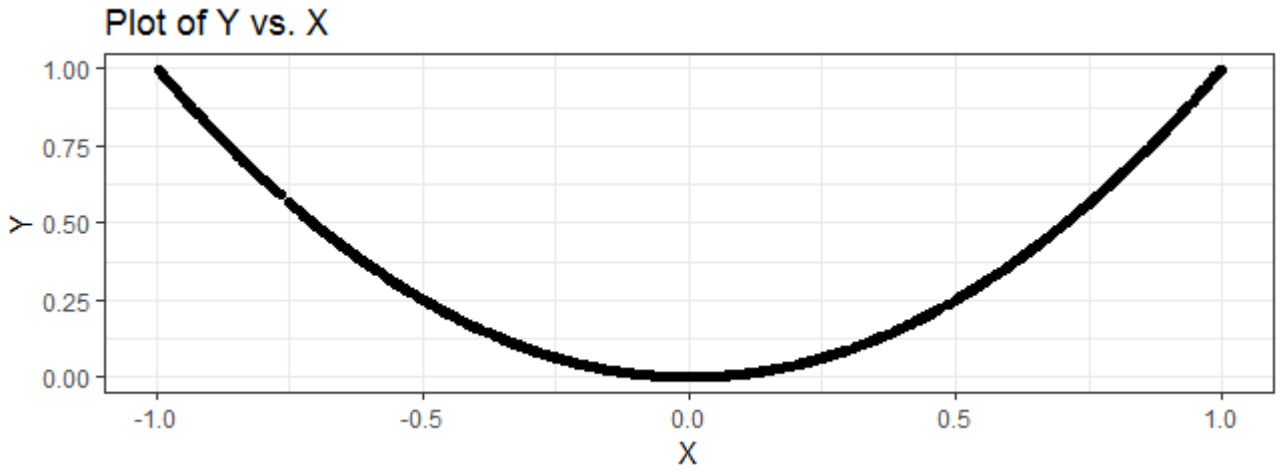

Figure 2.1: Plot of $Y = X^2$ showing that Pearson's correlation of zero does not imply distance correlation of 0

While distance correlation is sufficient for identifying a single variable related to the response, we need a form that can account for variables that have already been identified. In 2013, a conditional form of distance correlation was introduced by Wang, Wen, Pan, and Huang [3]. This can be calculated similarly to distance correlation, except using conditional characteristic functions. An empirical estimate is available in R in the CDCSIS package, which was developed by Wang, Wen, Pan, and Huang[3].

2.4 False Selection Rate

False Selection Rate, first introduced by Wu, Boos and Stefanski [1], is a method that creates pseudovariables and runs the model many times with the real variables and psuedovariables and tracks the number of psuedovariables that enter the model.

The main goal is to use the resulting information to minimize the rate at which uninformative variables enter the model, hence the name False Selection Rate(FSR).

There are two competing ways to calculate the FSR; expectation ratio (γ_{ER}), or ratio of expectations (γ_{RE}). The expression below shows these two ideas.

$$\gamma_{ER} = \mathbf{E}[\frac{U(\alpha)}{1 + S(\alpha)}] \tag{2.4}$$

$$\gamma_{RE} = \frac{\mathbf{E}[U(\alpha)]}{\mathbf{E}[1 + S(\alpha)]} \tag{2.5}$$

where $U(\alpha)$ is the number of uninformative variables that entered the model at a given correlation of α and $S(\alpha)$ is the total size of the model at the same given correlation. For this paper, we are going to continue with using the expectation ratio method. We can estimate γ_{ER} by the following equation:

$$\hat{\gamma}_{ER} = \frac{\hat{k}_u(\alpha)\bar{U}_p^*(\alpha)/k_p}{1 + S(\alpha)} \tag{2.6}$$

Where $\hat{k}_u(\alpha) = k_t - S(\alpha)$ with k_t defined to be the total number of variables considered, and $\bar{U}_p^*(\alpha)/k_p$ is defined as the average number of psuedovariables that entered the model after multiple replications of this process. The correlation α needs to be specified, since this allows us to control the size of models we would like to consider. While we are using the same estimate for $\hat{\gamma}_{ER}$ as Wu, Boos and Stefanski [1], we need to use the opposite formulation for the estimate of the optimal α. This is done the following way:

$$\hat{\alpha}_{ER} = \inf_{\alpha \geq \alpha_{min}} \{\alpha : \hat{\gamma}_{ER}(\alpha) \leq \gamma_0\} \tag{2.7}$$

Where γ_0 is a fixed value for the FSR that we are trying to control. This optimization constraint makes sure that we are not exceeding a set false selection rate in our models and helps determines the correlations that give us these model sizes.

Chapter 3

Forward Selection via Distance Correlation

Forward selection via distance correlation is very similar to classical forward selection, but with distance correlation and conditional distance correlation, rather than Pearson's correlation and partial correlation. To help with understanding the procedure, consider the following example; We have a response (y) and 10 possible predictors ($X_1, X_2, ...X_{10}$)).

The procedure starts by running a forward selection algorithm that adds in variables based on their distance correlation with the response, using conditional distance correlation for selecting the order of the rest of the predictors after selecting the first variable is selected. The order is determined by having the largest correlation with the response at that step. For the theoretical example proposed above, lets assume we get the distance correlation between the response and each of the predictors in Table 3.1. Then the variable that would be allowed to enter the model is X_3. Then for the next step consider the conditional correlations in table 3.2. The next variable to enter would be X_1. This is repeated until all of the variables have entered the model, creating the full model.

Table 3.1: Theoretical set of Partial Distance Correlation for an example

Variable	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
Distance Correlation	0.1	0.2	0.9	0.7	0.3	0.8	0.6	0.4	0.5	0.0

Table 3.2: Theoretical set of Partial Distance Correlation for an example

Variable	X_1	X_2	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
Distance Correlation	0.85	0.10	0.45	0.15	0.80	0.70	0.25	0.40	0.35

Table 3.3: The full model after forward selection via distance correlation ran

Variable	X_3	X_1	X_9	X_{10}	X_7	X_4	X_5	X_8	X_6	X_2
Distance Correlation	0.80	0.85	0.25	0.20	0.35	0.40	0.30	0.05	0.01	0.00

Table 3.4: Monotone ordering of Correlations for the full model

Variable	X_3	X_1	X_9	X_{10}	X_7	X_4	X_5	X_8	X_6	X_2
Correlation	0.80	0.85	0.25	0.20	0.35	0.40	0.30	0.05	0.01	0.00
Monotone Correlation	0.80	0.80	0.25	0.20	0.20	0.20	0.20	0.05	0.01	0.00

After choosing the order of entry for a full model, which can be seen in table 3.3 along with their associated correlations, The correlations are to be ordered monotonically. Table 3.4 shows the ordering of the alphas monotonically. Then a cut of the size of models to consider is made by choosing a correlation threshold (α_{min}). α_{min} is to be chosen such that not all variables are included nor excluded. The selection of α_{min} is further discussed in the next section. For our example lets consider $\alpha_{min} = 0.06$, then we would keep a model of the first seven variables to enter as the maximum size to consider. The result of these can be seen in table 3.5, which contains the unique correlation thresholds to consider and the model size associated with them.

Table 3.5: Unique correlation thresholds to consider and their respective number of variables allowed in the model

Correlation	0.80	0.25	0.20
# of Variables	2	3	7

Next in the procedure is to perform FSR as described earlier for each of the correlation thresholds still in consideration, on the forward selection via distance correlation algorithm. This will produce false selection rates for each of the considered correlation thresholds. continuing with our example, table 3.6 gives some theoretical results of this procedure on the made up data.

Table 3.6: The false selection rates for the correlation thresholds still in consideration

Correlation	0.80	0.25	0.20
False selection Rate	0.1	0.5	0.8

Lastly, we select the minimum correlation threshold that provides a false selection rate that is below some predetermined false selection rate initial value γ_0. Continuing our example, lets assume we chose $\gamma_0 = 0.15$, then we would chose the model with

the correlation threshold of 0.8, which includes the first two variables to enter the model. Therefore, forward selection via distance correlation procedure would return that the variables selected were x_3 and x_1 in that order.

Chapter 4

The Selection of α_{min}

From the previous section, we need to determine what α_{min} should be set at to neither include or exclude all of the possible predictors from the model. To determine the appropriate value for α, simulations were ran using R, which took a response variable and 10 noise variables, variables that are uninformative, while allowing all the variables to enter the model. This was done using a multivariate normal distribution with $n = 250$, $p = 11$, and $m = 250$, where n is the number of observations, p is the number of variables generated, and $m = 250$ the number of replications ran. In one case, the correlation structure was set to make all of the variables independent from each other and the first variable was used as the response. In the other case the correlation structure was set to make the first variable independent from the rest, while the others had a correlation of 0.6 between them. Figures 4.1 and 4.2 show summaries of the max, min, median, and mean densities of the simulation data.

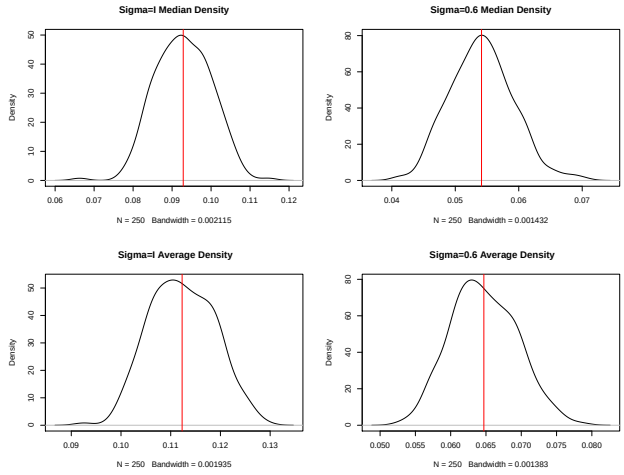

Figure 4.1: Average and Median Density plots of Correlations

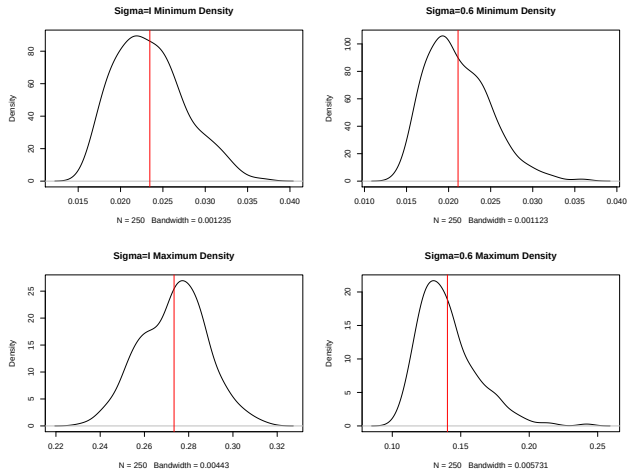

Figure 4.2: Minimum and Maximum Density plots of Correlations

From figures 4.1 and 4.2, we can see that the median and mean densities for the slightly related predictors is between 0.055 and 0.065, meaning that for case, we see about half of the variables entering when the correlation threshold is set within that range. looking at the max and min density plots in figure 4.2, we can see that the density of the minimum correlation never goes above 0.04, so we would like a minimum correlation threshold to be above that. We can also see that the maximum density never goes below approximately 0.09, so we need a minimum correlation threshold below that. Using this information , we chose $\alpha_{min} = 0.06$ as our minimum correlation threshold, since it satisfies all of the above statements, which means we will always considers model sizes of 1-9.

Chapter 5

Numerical Simulations

These simulations are used to compare the new method of using forward selection via distance correlation against well known methods of adaptive lasso and standard forward selection with a significance level of $\alpha = 0.05$ methods. We considered two different scenarios, linear and non-linear that each contain two cases. Each of the simulations were run with $p = 10$ predictors, $n = 250$ observations, and $m = 250$ replications. These were chosen due to computational limitations that we faced.

5.1 Linear Models

The first type of models that were considered, were linear models. These models were chosen to provide direct comparison to the well known linear methods, since distance correlation should be able to perform well in both cases. The predictors in each case were derived from a multivariate normal distribution. The first case considered had a correlation structure of independence and the second case had a compound symmetric correlation structure with the off diagonal values being 0.6. Using these predictors, the following form for the response was used:

$$y = -4x_1 + 2x_2 + x_3 + \varepsilon \tag{5.1}$$

where $\varepsilon \sim N(0, \sigma^2)$ where σ^2 was determined by setting the theoretical r-squared of the model to 0.8 and using the following formula:

$$\frac{(1 - R^2)(b^t \Sigma b)}{R^2} \tag{5.2}$$

where b is the known linear regression coefficients. The response and the predictors were used in the three different variable selection methods, and output of the variables that entered the model was recorded.

5.2 Non-Linear Models

The second scenario we considered was adding non-linear terms to the model, since distance correlation does not need to have the assumption of linearity imposed. These cases also are used to show the robustness of our method. The first non-linear case we considered was introducing spread and an indicator variable into the model. We generated the predictors using a multivariate normal distribution with an independent correlation structure. The response variable was formulated the following way.

$$y = -4x_1 + 2x_2 + \mathbf{1}(x_3 < 0) + exp(x_4)\varepsilon \tag{5.3}$$

this formulation of the response was obtained from the article, Feature Screening via Distance Correlation by Li, Zhong, and Zhu [6]. The non-linearity exists in the indicator function as well as in the error. It should be noted that x_4 is not even in the mean model.

The second non-linear case that we considered was introducing the sine function into the model. For this, the data was generated the following way:

$$X_1 = Unif(0, \frac{\pi}{2})$$
$$X_2 = 2 + 5\sin(X_1)$$

The rest of the predictors were generated using a multivariate normal distribution with a compound symmetric correlation structure with the off diagonal entries being 0.6. The response for this was formulated the following way:

$$y = -4\sin(x_1) - 2x_2 + \varepsilon \tag{5.4}$$

Chapter 6

Results

From the numerical simulations, the variables that entered the model were kept track of. One way to help determine how well a variable selection algorithm performed is the F1 score. F1 scores are commonly used in classification literatutre, and we are treating variable selection in this case as a classification problem. The F1 score determines how well the algorithm is at selecting the right sized model (Precision),and how accurate it is at picking up the correct number of informative variables (Recall)[5]. An F1 score is calculated the following way:

$$F1 = 2\frac{Precision * Recall}{Precision + Recall}$$
$$Precision = \frac{TruePositives}{TotalpredictedPositives}$$
$$Recall = \frac{TruePositives}{TruePositives + FalseNegative}$$

Another way we can asses the performance of our method is to calculate the false selection rate. This can be calculated by finding the total number of falsely predicted variables, over the number of variables that were considered false variables. This will allow us to asses if we are picking up any incorrect variables that do not have any effect on the response.

Figures 6.1 and 6.2 show the false selection rate and the F1 score for each of the linear cases mentioned above, for each of the 250 replications. The F1 score and false selection rate were calculated for each of the three procedures.Figures 6.3 and 6.4 show the false selection rate for each of the three procedures as well as the F1 score for each of the 250 replications, for the two nonlinear cases mentioned above.

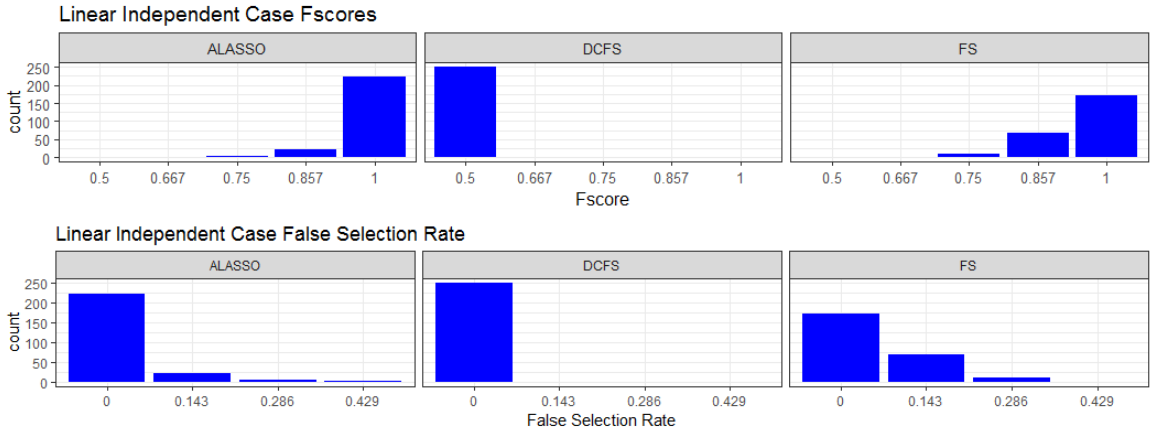

Figure 6.1: Summarization of classification measures and false selection rates for linear case with independent predictors; we compare the adaptive lasso (ALASSO), classical forward selection (FS), and forward selection tuned with distance correlation (DCFS).

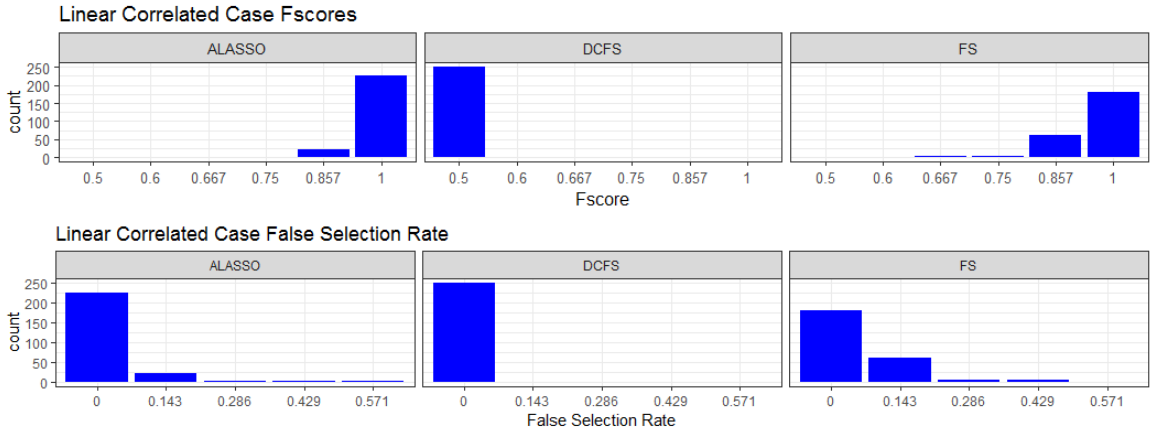

Figure 6.2: Summarization of classification measures and false selection rates for linear case with correlated predictors; we compare the adaptive lasso (ALASSO), classical forward selection (FS), and forward selection tuned with distance correlation (DCFS).

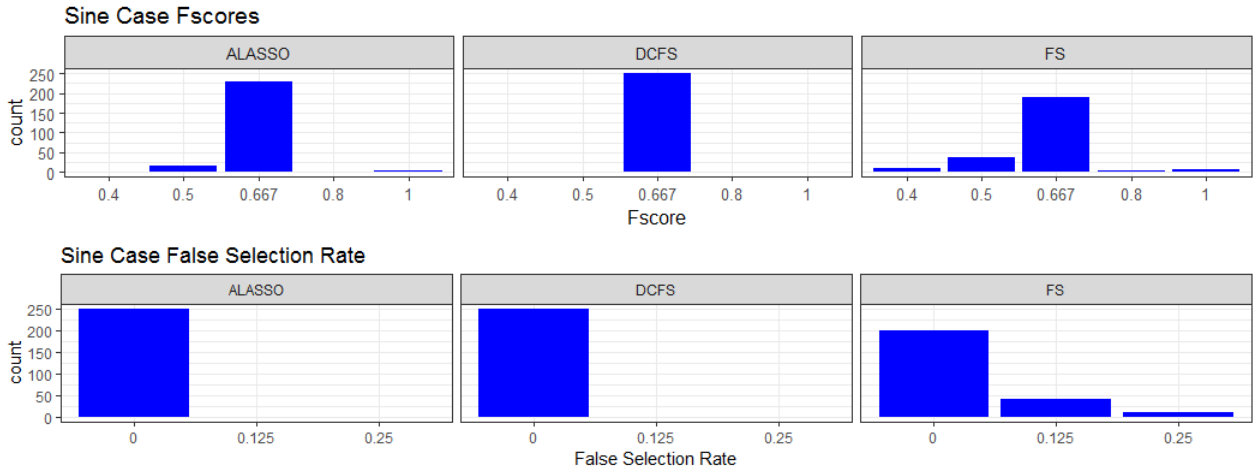

Figure 6.3: Summarization of classification measures and false selection rates for non-linear sine case; we compare the adaptive lasso (ALASSO), classical forward selection (FS), and forward selection tuned with distance correlation (DCFS).

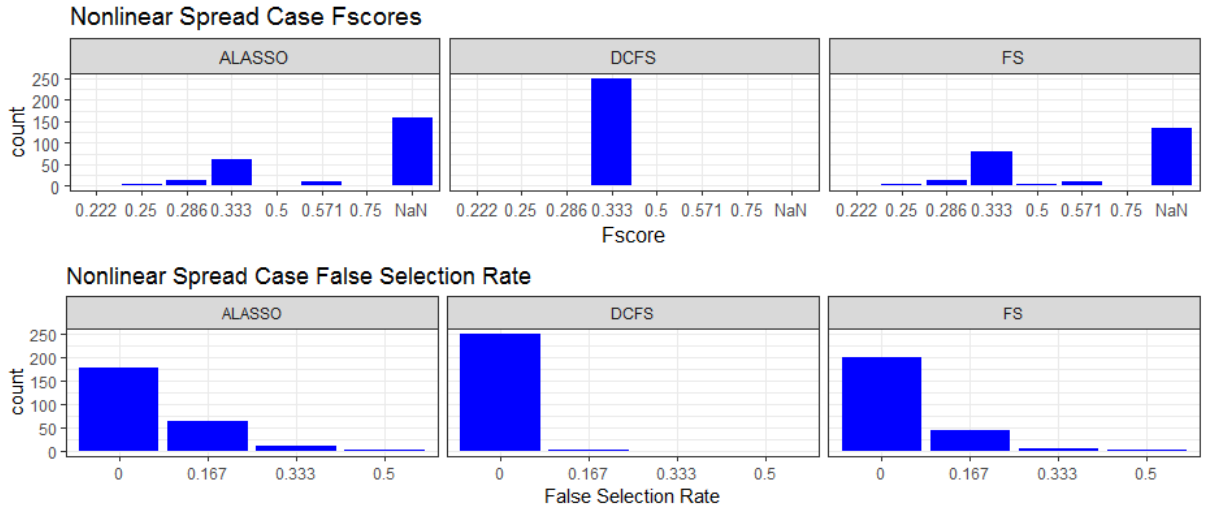

Figure 6.4: Summarization of classification measures and false selection rates for non-linear spread case; we compare the adaptive lasso (ALASSO), classical forward selection (FS), and forward selection tuned with distance correlation (DCFS).

Table 6.1 shows the results from running forward selection via distance correlation, classical forward selection, and adaptive LASSO on the diabetes data set introduced earlier. We can see that the model sizes vary drastically between the procedures.

Table 6.1: variables selected by each of the methods tested for the diabetes data

Method	Variables Selected in Order
ALASSO	$x_2, x_3, x_4, x_5, x_6, x_8, x_9$
FS	$x_3, x_9, x_4, x_5, x_2, x_6$
DCFS	x_9

Chapter 7

Conclusions

As we can see from the graphics in the results section, the method of using distance correlation in forward selection does really well at keeping the false selection rate at 0, which means it never lets in a false variable into the model. When it comes to classification scores, such as F1, we can see that it does well at picking up a single important variable. Both of these occurrences can possibly be due to the fact that we only consider small pool of predictors, due to computational constraints.

We can also see this for the Diabetes data. Both the adaptive lasso and classical forward selection methods pick up 7 and 6 variables respectively, while distance correlation in forward selection only picks up 1 variable. This leads us to determine that we may have been to conservative when choosing our cutoff for the false selection rate.

Some limitations of our study is that we were computationally constrained by computer power on the servers we were using. The simulations took over a month to completely finish, even though we were considering relatively small data sets with only 2500 entries. This being the case, we may need to consider making a more efficient algorithm for performing the forward selection via distance correlation. Another reason this may have occurred is that we are using false selection rate that causes us to run the algorithm many time over again on larger data. This caused the computation time to largely increase.

For future work, we would like to take a look into developing a different way of determining when to stop the model, rather than using false selection rate. This could lead to a much more efficient procedure in the future, allowing us to try it on larger data sets. There do exist ways to get a test statistic for this procedure, but it seems like it will just be as computationally intensive. After all, we would like to rerun our studies using a less conservative cutoff for false selection rate and to determine a better way to determine the cut off in general. Lastly, being able to run this procedure on larger data sets would be beneficial to test whether or not this

method can accurately predict what variables to include in the model.

Bibliography

[1] Yujun Wu, Dennis D. Boos, and Leonard A. Stefanski, *Controlling Variable Selection by the Addition of Pseudovariables*, American Statistical Association **102** (2007).

[2] Gabor J. Szekely, Maria L. Rizzo, and Nail K. Bakirov, *Measuring and Testing Dependence by Correlation of Distances*, The Annals of Statistics **35** (2007).

[3] Xueqin Wang, Canhong Wen, Wenliang Pan, and Mian Huang, *Sure Independence Screening Adjusted for Confounding Covariates with Ultrahigh-dimensional Data*, Statistica Sinica (2013).

[4] Bradley Efron, Trevor Hastie, Iain Johnstone, and Robert Tibshirani, *Least Angle Regression*, The Annals of Statistics **32** (2004).

[5] Koo Ping Shung, *Accuracy, Precision, Recall or F1?* (Mar 15, 2018), `https://towardsdatascience.com/accuracy-precision-recall-or-f1-331fb37c5cb9`.

[6] Runze Li, Wei Zhong, and Liping Zhu, *Feature Screening via Distance Correlation Learning*, American Statistical Association **107** (2012).

[7] Runze Li, Jingyuan Liu, and Lejia Lou, *Variable Selection via Partial Correlation*, Statistica Sinica **27** (2017).