Journal of Biometrics & Biostatistics

ISSN: 2155-6180

Research Article

Open Access

Classical and Robust Forward Selection: A Simulation Study and Real Data Application

Moushumi Pervin[1]* **and Md. Siddiqur Rahman**[2]

[1]*Department of Business Administration, Bangladesh Army International University of Science and Technology, Cumilla, Bangladesh*
[2]*Department of Statistics, Jagannath University, Dhaka, Bangladesh*

Abstract

In order to use any method as a model selection algorithm, it is needed to check the adequacy and stability of a model selected by the algorithm. Adequacy of a robust model was not checked by giving outliers in various ways. In this paper, several contamination cases have been introduced to check the adequacy of the robust model selected by robust forward selection (RFS). In each of the contamination case, the performance of RFS has been compared to standard forward selection (FS) through a simulation study. The adequacy and stability of the robust model has also been checked through a real data application. Based on simulation study and real data application, RFS has much better performance compared to standard FS.

Keywords: Adjusted winsorization; Computational burden; Forward selection; Pairwise correlation; Robust forward selection

Introduction

When the number d of candidate predictor variables is small, a linear prediction model can be chosen by computing a reasonable criterion (e.g., AIC, Mallow's C_p, BIC, CV) for all possible subsets of the predictor variables. However, the computational burden of this approach increases very quickly as d increases. This is one of the main reasons why step-by-step algorithm like forward selection (FS) is popular. But when the data sets contain outliers and other contaminations, classical FS procedure yields poor results, and often fails to select important covariates that would have been chosen if there were no outliers and other contaminations in the data sets. On the other hand, it is not logical to predict future outliers without having knowledge about the radical mechanism that produces these outliers.

Classical FS algorithm has been expressed in terms of sample means, variances and correlations [1]. Robust FS is obtained by replacing these classical sample quantities by their robust counterparts [1]. As a stopping rule, partial *F*-test procedure is used. The focus of this work is not to fit a final model but to check the adequacy and stability of the robust model. Khan et al. [1] checked the adequacy of the robust model by giving outliers to the noise variables and their corresponding response values. In this study, the adequacy of the robust model has been checked by giving outliers in several ways through a simulation study.

The rest of the paper is organized as follows. In §2, the classical FS is reviewed. In §3, the robust version of the algorithm is reviewed. In §4, a simulation study is presented for the comparison of robust FS and standard FS. In §5, a real data application is presented. And finally, §6 is the conclusion.

FS Algorithm Expressed in terms of Correlations

Let $X_1, X_2, \ldots, X_d$ be n dimensional predictor variables and Y be the n dimensional response variable. Each variable is standardized with mean 0 and variance 1. The FS procedure begins with the assumption that there are no predictor variables in the model other than the intercept. The first predictor (X_1, say) selected for entry into the equation is the one which has the largest absolute correlation $|r_{1Y}|$ with Y, and the residual vector $Y - r_{1Y}X_1$ is obtained. For entering all the other predictor variables into the competition, they are 'adjusted for X_1'. That is, each X_j is regressed on X_1, and the corresponding residual vector $Z_{j.1}$ (which is orthogonal to X_1) is obtained. The correlations of these $Z_{j.1}$ with the residual vector $Y - r_{1Y}X_1$ called the partial correlations between X_j and Y 'adjusted for X_1' decide the next variable (X_2, say) to enter the regression model. All the other predictor variables are then 'adjusted for the first two selected variables X_1 and X_2' for entering into further competition, and so on. This procedure of adding one predictor variable at each step is continued, until a stopping criterion is met. Let the correlation between X_j and Y be r_{jY} and R_X be the correlation matrix of the predictors $X_1, X_2, \ldots, X_d$ Let us assume, without loss of generality, X_1has the maximum absolute correlation with Y Then, X_1 is the first variable that enters the regression model. The predictors in the current regression model are *active* predictors *a*. The remaining candidate predictors (*d-a*) are inactive predictors. The second predictor X_2 (say) that enters the regression model is the one that has the maximum absolute partial correlation $|r_{jY.1}|$ with Y.

FS steps in correlations

FS algorithm is summarized in terms of correlations among the original variables as follows [1]:

1. To select the first covariate X_{m1}, determine m_1=argmax $|r_j|$

2. To select the kth covariate (k=2,3,…), calculate $\tilde{r}_{jY.m_1 \cdots m_{(k-1)}}$, which is proportional to the partial correlation between X_j and Y adjusted for $X_{m1}\ldots,X_{m(k-1)}$ and then determine $m_k = \arg\max |\tilde{r}_{jY.m_1 \cdots m_{(k-1)}}|$.

Stopping rule for FS

At each FS step, once the "best" covariate (among the remaining covariates) is identified, a partial *F*-test can be performed to decide

*Corresponding author: Moushumi Pervin, Department of Business Administration, Bangladesh Army International University of Science and Technology, Cumilla, Bangladesh, Tel: +8801674234869; E-mail: mst.moushumi55@gmail.com

Received May 23, 2018; **Accepted** June 14, 2018; **Published** June 20, 2018

Citation: Pervin M, Rahman MS (2018) Classical and Robust Forward Selection: A Simulation Study and Real Data Application. J Biom Biostat 9: 404. doi: 10.4172/2155-6180.1000404

whether to include this covariate in the model (and continue the process) or to stop. The new "best" covariate enters the model only if the partial F-value, denoted by F_{partial}, is greater than $F(0.90,1,n\text{-}k\text{-}1)$ (say), where k is the current size of the model including the new covariate. Here again, the required quantities can be expressed in terms of correlations among the original variables, which is shown below.

When $(k\text{-}1)$ covariates $X_1,X_2,\ldots,X_{k-1}$ are already in the model, and without loss of generality X_k has the largest absolute partial correlation with Y after adjusting for $X_1,X_2,\ldots,X_{k-1}$ the partial F-statistic for X_k can be expressed as:

$$F_{\text{partial}}=\frac{(n-k-1)\,\tilde{r}^2_{kY.12\cdots(k-1)}}{1-r^2_{Y1}-\tilde{r}^2_{2Y.1}-\cdots-\tilde{r}^2_{kY.12\cdots(k-1)}}\,.$$

Robustification of FS Algorithm

Simple robustification of FS algorithm is achieved by replacing the non-robust ingredients of FS algorithm by their robust counterparts [1,2]. For the initial standardization, the choices of first countable robust center and scale measures are straight forward: median (med) and median absolute deviation (mad). Most available robust correlation estimators are computed from the d-dimensional data and therefore are very time consuming [3]. On the other hand, robust univariate approaches [4] are very sensitive to correlation outliers (outliers that are not detected by univariate analyses but affect the classical correlation).

One solution is to derive correlations among pairs of variables from an affine-equivariant bivariate covariance estimator. A computationally efficient choice is the bivariate M-estimator defined by Maronna [5]. Alternatively, the robust correlation estimator of Gnanadesikan and Kettnring [6] or the related orthogonalized Gnanadesikan Kettnring estimator [7] can be used. For very large, high-dimensional data sets, we need an even faster robust correlation estimator. Huber [4] introduced the idea of univariate winsorization of the data and suggested that classical correlation coefficients be calculated from the winsorized data. Alqallaf, et al. [8] re-examined this approach for the estimation of individual elements of a high-dimensional correlation matrix. For n univariate observations $X_1,X_2,\ldots,X_n$, the transformation is given by

$$u_i=\psi_c\,((x_i\text{-}\mathrm{med}(x_i))/\mathrm{mad}(x_i)),\ i=1,2,\ldots,n,$$

where the Huber score function $_c(x)$ is defined as $\psi_c(x)=min\{max\{-c,x\},c\}$, with c a tuning constant chosen by the user (e.g., c=2 or c=2.5). Note that in our case, $\mathrm{med}(x_i)=0$ and $\mathrm{mad}(x_i)=1$, because med and mad are used to robustly standardize the data. This univariate winsorization approach can be computed very rapidly, but unfortunately it does not take into account the orientation of the bivariate data.

To remedy this problem, Khan et al. [2] proposed a *bivariate winsorization* of the data based on an initial robust bivariate correlation matrix R_0 and a corresponding tolerance ellipse. Outliers are shrunken to the border of this ellipse by using the bivariate transformation $\mathbf{u}=\min\left(\sqrt{c/D(\mathbf{x})},\,1\right)\mathbf{x}$ with $x=(x_1,x_2)^t$ Here $D(x)$ is the Mahalanobis distance based on R_0. Notice that c is a tuning constant that was chosen to be c=5.99 the 95% quantile of the χ^2_2 distribution. The choice of R_0 is discussed later.

The initial correlation estimate

Choosing an appropriate initial correlation matrix R_0 is essential for bivariate winsorization. In principle, we could use any robust bivariate scatter estimate, but for computational convenience, Khan et al. [2] proposed a new method called adjusted winsorization. This method considers quadrants relative to the coordinate-wise medians (which is considered as 0 due to the robust standardization of the data) and uses two tuning constants to perform univariate winsorization of the data. A larger tuning constant, c_1, is used to winsorize the points lying in the two diagonally opposed quadrants that contain majority of the standardized data (called the "major quadrants"). A smaller tuning constant c_2 is used to winsorize the remaining data in the other two quadrants. In this article, we used $c_{1=2}$ and $c_2=\sqrt{h}c_1$, where $h=n_2/n_1,n_1$ is the number of observations in the major quadrants and $n_2=n\text{-}n_1$. The initial correlation matrix R_0 is obtained by computing the classical correlation matrix of the adjusted winsorized data. The adjusted winsorization handles correlation outliers much better than univariate winsorization. By using bivariate winsorization, the outliers are shrunken to the boundary of the larger ellipsoid and thus appropriately down-weighted so that a robust correlation estimate is obtained. Although the initial adjusted winsorization and the resulting bivariate winsorization are not affine-equivariant, they can be computed very rapidly and can appropriately handle correlation outliers

Stopping rule for RFS

The classical correlations in the partial F statistic are replaced by their robust counterparts to form a robust partial F statistic. For stopping rule, standard F distribution is used as in §2.

A Simulation Study

A simulation study is accomplished analogous to Khan et al. [1] so that the performance of robust FS and classical FS can be compared. To perform the simulation study, d=50 candidate predictor variables are considered out of which a=9 or a=15 are non-zero target predictor variables. No-correlation case and two different correlation cases i.e., moderate-correlation and high-correlation cases which exist among the target predictor variables are considered. These cases are described below:

For the no-correlation case, independent predictor variable $X_j\sim N(0,1)$ is considered and the response variable Y is generated using the a non-zero predictor variables with coefficients [1,2,9] which is repeated three times for a=9 and five times for a=15. The rest of the candidate predictors are considered as noise variables whose coefficients are zero. The variance of the error term is chosen in such a way that the signal to noise ratio equals to 2.

For the moderate-correlation case, three latent variables are introduced which are responsible for the systematic variation of both the response and the covariates, but are not active covariates. The linear model is created as follows:

$$Y=7L_1+6L_2+5L_3+\varepsilon=\text{Singnal}+\varepsilon,$$

where $L_j\sim N(0,1)$, i=1,2,3 and ε is a normal variable with mean 0 and standard deviation $\sigma=\frac{\sqrt{110}}{2}$. When a=9, a set of candidate predictor variables d=50 is created as follows. Let,

$$X_{ji}=L_i+e_{ij},\ i=1,2,3;\ j=1,2,3 \text{ and } X_k=u_k,\ k=1,2,3,\ldots,41,$$

where all $e_{ij}\sim N(0,1)$ and $u_k\sim N(0,1)$. Thus, the true correlation between these covariates is 0.5.

Similarly, for the high-correlation case, a similar linear model is created as in moderate-correlation case and a set of candidate predictor variables d=50 for a=9 is created as follows. Let,

$X_{ij}=L_i + \delta e_{ij}$, i=1,2,3; j=1,2,3 and $X_k=u_{k,}$ k=1,2,3,…,41.

Here, δ is a fixed constant which is chosen to be 0.5 so that the correlation between these covariates is 0.80.

For the no-correlation, moderate-correlation and high-correlation cases, 1000 data sets each of size 200 is generated. Each data set is randomly divided into a training sample of size 100 and a test sample of size 100. We considered data without outliers, and also with 10% and 15% outliers or bad leverage points. 10% and 15% of bad leverage points are obtained by generating the errors with mean 50 and standard deviation 1. While at the same time all or parts of the corresponding predictor variables are contaminated. The contaminated predictor variables are generated with mean 500 and standard deviation 1.

Process of contamination of the training data

For contamination of the training data, at first a number of rows is randomly chosen and the covariates of these rows were replaced by large positive numbers. The corresponding response values were also replaced by large positive numbers. The corresponding response values were also replaced by large numbers. To contaminate the training sample with 10% of bad leverage points, the probability that any specific row of the training sample will be contaminated is 1-(1-p)^z. That is, $p = 1 - \exp\left(\frac{\ln(0.90)}{z}\right)$, where z is the number of predictor variables whose values and their corresponding response values are contaminated. Similarly, for 15% of bad leverage points, the probability will be $p = 1 - \exp\left(\frac{\ln(0.85)}{z}\right)$.

For each of the no-correlation, moderate-correlation and high-correlation cases, the training data sets are contaminated in different ways for measuring the adequacy of the robust model. Different cases of contaminations are given below:

- Case 1: All candidate predictor variables are contaminated.
- Case 2: All active predictor variables plus 5 first noise variables are contaminated.
- Case 3: All active predictor variables are contaminated.
- Case 4: All active predictor variables related to the two most important latent variables L_1,L_2 are contaminated.
- Case 5: Two active predictor variables related to each of the three latent variables L_1,L_2 and L_3 are contaminated.
- Case 6: Most important active predictor variables plus first 10 noise variables are contaminated.
- Case 7: First three active predictor variables related to the most important latent variable L_1 are contaminated.

At first the training data is used for fitting the obtained models by applying each of the classical and robust FS methods. Then the test data is used for testing the significance of the fitted models. Both the classical and robust models are fitted by using a regression *MM*-estimator because of its high breakdown point which is 0.5, and high efficiency at the normal distribution [10].

For each simulated data set, 10% trimmed mean of squared prediction error on the test sample is recorded. The average, standard deviation (SD) and median absolute deviation (mad) of the three quantities i.e., mean squared prediction error (MSPE), noise variables and target variables are shown in parentheses.

At first the performance of the classical and robust methods in clean data for the no-correlation, moderate-correlation and high-correlation cases is presented.

Table 1 depicts that the performance of classical FS and robust FS is comparable for the no-correlation, moderate-correlation and high correlation cases in clean data.

All candidate predictor variables

In this case, the values of the d=50 candidate predictor variables and their corresponding response values are contaminated. Table 2 represents the results for the no-correlation case in contaminated data. It shows that the test error produced by robust FS is much smaller than for the classical FS for both 10% and 15% outliers cases. The median absolute deviations and standard deviations are much smaller for the robust method than for the classical one. Also, the model selected by robust FS contains less noise variables than the classical FS. At the same time, more we increase the percentage of outliers in the training data, more the robust method performs very well while the performance of classical method is quite poor. Because classical FS selects more noise variables in the final model as the percentage of bad leverage points is increased. On the other hand, robust method selects less noise variables for the cases of 10% and 15% outliers. For example, for a=5 when we increase outliers from 10% to 15%, the average of the noise variables decreases from 0.9 to 0.6. Thus, we say that the robust method fits the final model with a small number of predictor variables by producing less test error compared to the classical method.

Tables 3 and 4 present the results for the moderate-correlation and high-correlation cases respectively. Here, the conclusions of the results for both the correlation cases are same as the no-correlation case [11-14].

All active predictor variables plus 5 first noise variables

In this case, the first 14 and 20 predictor variables are contaminated for a=9 and a=15 respectively, and the corresponding values of the

Cases	Method	*a*=9			*a*=15		
		MSPE	Noise	Target	MSPE	Noise	Target
No-correlation	FS	80.8	4.6	9	145.4	4.0	15
		(17.9)	(2.4)	(0.05)	(35.1)	(2.3)	(0.5)
		(16.3)	(1.5)	(0)	(30.7)	(3.0)	(0)
	RFS	89.9	10.0	9	164.3	9.3	14.8
		(23.1)	(7.2)	(0.1)	(46.1)	(6.6)	(0.7)
		(21.2)	(5.9)	(0)	(21.2)	(5.9)	(0)
Moderate-correlation	FS	58.1	4.8	6.5	52.1	4.1	8.3
		(12.4)	(2.5)	(1.1)	(11.1)	(2.3)	(1.3)
		(11.9)	(3.0)	(1.5)	(9.7)	(3.0)	(1.5)
	RFS	60.4	6.7	6.5	54.2	6.0	8.4
		(13.5)	(3.9)	(1.1)	(11.8)	(3.7)	(1.6)
		(12.6)	(3.0)	(1.5)	(11.1)	(3.0)	(1.5)
High-correlation	FS	37.4	4.7	5.8	34.8	4.0	6.7
		(7.3)	(2.5)	(1.0)	(7.0)	(2.3)	(1.1)
		(7.3)	(3.0)	(1.5)	(6.4)	(3.0)	(1.5)
	RFS	39.8	7.8	5.9	37.7	7.5	7.2
		(8.8)	(5.6)	(1.1)	(27.8)	(6.1)	(2.2)
		(8.0)	(4.5)	(1.5)	(8.6)	(4.5)	(1.5)

Table 1: Performance of the classical FS and robust FS in clean data for the no-correlation, moderate-correlation and high-correlation cases.

Outliers	Method	*a*=9			*a*=15		
		MSPE	Noise	Target	MSPE	Noise	Target
10%	FS	222.9	11.4	6.0	421.1	9.9	9.5
		(144.5)	(3.0)	(1.6)	(191.0)	(2.7)	(2.3)
		(103.8)	(3.0)	(1.5)	(161.6)	(3.0)	(3.0)
	RFS	95.4	0.7	8.0	289.7	0.9	8.8
		(46.6)	(1.2)	(1.5)	(100.8)	(1.2)	(2.9)
		(33.1)	(1.5)	(1.5)	(105.0)	(1.5)	(3.0)
15%	FS	340.5	14.1	5.4	545.5	12.1	8.8
		(212.0)	(3.1)	(1.6)	(232.9)	(3.0)	(2.1)
		(192.3)	(3.0)	(1.5)	(216.4)	(3.0)	(1.5)
	RFS	141.7	0.5	6.2	370.3	0.6	5.8
		(62.6)	(0.9)	(2.0)	(101.2)	(1.0)	(2.7)
		(64.8)	(0)	(3.0)	(96.1)	(1.5)	(3.0)

Table 2: Case 1: Performance of the classical FS and robust FS in contaminated data for no-correlation case.

Outliers	Method	*a*=9			*a*=15		
		MSPE	Noise	Target	MSPE	Noise	Target
10%	FS	80.3	11.22	3.3	67.9	9.6	5.1
		(42.3)	(3.0)	(1.4)	(35.0)	(2.7)	(2.0)
		(20.2)	(3.0)	(1.5)	(16.0)	(3.0)	(1.5)
	RFS	59.7	1.3	4.2	55.5	1.0	4.9
		(12.1)	(1.5)	(1.0)	(11.2)	(1.3)	(1.2)
		(11.6)	(1.5)	(1.5)	(10.5)	(1.5)	(1.5)
15%	FS	105.2	14.1	3.6	86.2	12.0	6.0
		(136.6)	(3.1)	(1.5)	(132.4)	(2.8)	(2.2)
		(32.8)	(3.0)	(1.5)	(24.6)	(3.0)	(3.0)
	RFS	61.1	0.7	3.6	58.5	0.6	4.1
		(12.5)	(1.0)	(0.9)	(11.6)	(1.0)	(1.1)
		(11.2)	(1.5)	(1.5)	(10.9)	(0)	(1.5)

Table 3: Case 1: Performance of the classical FS and robust FS in contaminated data for moderate-correlation case.

Outliers	Method	*a*=9			*a*=15		
		MSPE	Noise	Target	MSPE	Noise	Target
10%	FS	54.4	11.3	3.5	46.3	9.6	5.2
		(34.3)	(3.0)	(1.4)	(23.0)	(2.7)	(2.0)
		(19.4)	(3.0)	(1.5)	(15.0)	(3.0)	(1.5)
	RFS	37.3	1.1	3.5	36.3	0.9	3.8
		(7.3)	(1.4)	(0.8)	(6.8)	(1.4)	(1.0)
		(6.6)	(1.5)	(0)	(6.6)	(1.5)	(1.5)
15%	FS	69.0	14.1	3.8	53.8	11.9	6.3
		(84.4)	(2.9)	(1.5)	(70.1)	(2.9)	(2.2)
		(29.1)	(3.0)	(1.5)	(19.5)	(3.0)	(3.0)
	RFS	37.5	0.7	3.2	36.8	0.6	3.3
		(7.6)	(1.0)	(0.5)	(7.3)	(1.0)	(0.7)
		(6.6)	(1.5)	(0)	(6.8)	(0)	(0)

Table 4: Case 1: Performance of the classical FS and robust FS in contaminated data for high-correlation case.

response variable are also contaminated. Table 5 represents the results for the no-correlation case. It shows that the robust FS has less MSPE than the classical FS when *a*=9 active covariates are considered for both 10% and 15% outliers cases. Also the robust method fits the model with less noise variables than the classical method. But when we increased the number of active predictor variables from *a*=9 to *a*=15 in the model, the robust FS produces more test error than the classical FS. Despite of producing more test error, robust method includes less noise variables in the final model than the classical method. The conclusions of the results for the moderate correlation and high-correlation cases are also similar as in the no-correlation case.

In all other contamination cases, the conclusions of the simulation results are similar as in Case 1 and Case 2. So, the results of other contamination cases are not included.

Real Data Application

In this section, a real-data set is used to evaluate the robustness and scalability of robust FS method.

Breast cancer data

This data set was used for the KDD-cup 2008. We considered the

Outliers	Method	*a*=9			*a*=15		
		MSPE	Noise	Target	MSPE	Noise	Target
10%	FS	128.9	6.9	8.0	258.3	6.0	12.5
		(70.2)	(2.5)	(1.1)	(103.1)	(2.3)	(1.9)
		(47.7)	(3.0)	(1.5)	(92.4)	(1.5)	(1.5)
	RFS	90.1	0.8	8.5	286.3	0.9	9.4
		(36.4)	(1.3)	(1.1)	(94.1)	(1.2)	(2.9)
		(26.6)	(1.5)	(0)	(98.2)	(1.5)	(3.0)
15%	FS	155.0	7.4	7.7	311.6	6.4	11.9
		(73.5)	(2.5)	(1.2)	(120.9)	(2.3)	(1.8)
		(63.7)	(3.0)	(1.5)	(105.4)	(3.0)	(1.5)
	RFS	124.7	0.5	7.2	349.2	0.5	6.6
		(57.7)	(0.9)	(1.8)	(97.0)	(1.0)	(2.7)
		(55.2)	(0)	(1.5)	(98.5)	(0)	(3.0)

Table 5: Case 2: Performance of the classical FS and robust FS in contaminated data for no-correlation case.

training data set which consists of a total of *n*=102,294 candidates, each described by 117 feature variables. We used the first 101 feature variables with a total of *n*=50,000 observations in our analysis. The first variable is considered as the response variable, and the remaining 100 variables are considered as the candidate predictor variables. *n*=50,000 observations are divided into a training sample of size *n*=25,000 and a test sample of size *n*=25,000. When the classical FS and robust FS are applied to this training data set, the classical FS selects a huge model with the following 63 covariates:

(32, 59, 19, 42, 33, 23, 11, 89, 14, 81, 30, 51, 8, 9, 66, 17, 34, 5, 39, 79, 25, 99, 43, 58, 57, 1, 31, 65, 15, 24, 36, 78, 45, 82, 92, 6, 10, 62, 53, 7, 88, 68, 71, 22, 13, 46, 16, 60, 44, 91, 90, 4, 64, 76, 63, 67, 72, 38, 29, 84, 83, 28, 12),

While the robust FS selects a model with only 13 covariates as follows:

(32, 59, 42, 19, 14, 90, 31, 71, 68, 66, 39, 96, 75).

The 10% trimmed mean of squared prediction error for both the methods are 0.005. That is, the robust FS fits a good model by using only 13 predictor variables, while the classical FS does the same thing by using 63 predictor variables.

We really don't know whether this data set is contaminated or not. To check the scalability and robustness of robust FS, this data set is contaminated in three different ways. These contamination cases are described below:

Case 1: The response variable is contaminated.

This data set is contaminated by replacing one small value of the response variable (say 24532*th* value 0.1070976) by a large value 100. In the contaminated data set, classical FS selects a different model with the following 12 covariates:

(32, 59, 42, 19, 23, 11, 33, 91, 14, 8, 43, 71),

while robust FS selects the same model as before containing same number of predictor variables.

Case 2: The predictor variables are contaminated.

This data set is contaminated by replacing one small value of the predictor variable 32 (say 15023*th* value 0.5222431) by a large value 522. When both the classical FS and robust FS methods are applied to this contaminated data set, robust method selects the same model as before but classical FS selects a different model containing 62 covariates.

Case 3: Both the response and predictor variables are contaminated.

This data set is contaminated by replacing one small value of the predictor variables 20 (say 9001*th* value 0.692335) by a large value 160 and the corresponding value of the response variable (0.1446717) by a large value 180. Again in this case, robust FS selects the same model as before, but classical FS selects a model containing 52 covariates which is different from the previous model.

Conclusions

In this study, we considered the problem which arises when we select a linear prediction model for large high-dimensional data sets that may be clean or possibly contain a fraction of contaminations. At the same time, our goal was to achieve robustness and scalability. The performance of the classical FS and robust FS is compared through a simulation study and real data application. In simulated data sets, the performance of robust FS is comparable to standard FS for the no-correlation, moderate-correlation and high-correlation cases in clean data. We also compared the performance of robust FS and classical FS by contaminating the simulated data sets in different ways. Robust FS has performed much better than standard FS. As we increased the percentage of bad leverage points in the simulated data sets, the robust FS has performed much better than the standard FS for the no-correlation, moderate-correlation and high-correlation cases. In almost all the contamination cases, the classical FS produced more test error, and also included more noise variables than the robust FS. In some contamination cases, the robust FS produced almost same test error but included less noise variables than the classical FS. Overall the performance of robust FS is better than the classical FS. In real data set, when we replaced some observations by bad leverage points, the model selected by classical FS changes frequently and produces more test error than robust FS. From the simulation study and real data example, it is proved that the robust FS outperforms the classical FS in contaminated data.

References

1. Khan JA, Van Aelst S, Zamar RH (2007) Building a Robust Linear Model with Forward Selection and Stepwise Procedures. Computational Statistics and Data Analysis 52: 239-248.
2. Khan JA, Van Aelst S, Zamar RH (2007) Robust Linear Model Selection Based on Least Angle Regression. Journal of American Statistical Association 102: 1289-1299.
3. Rousseeuw PJ, Leroy AM (1987) Robust Regression and Outlier Detection. Wiley-Interscience, New York.
4. Huber PJ (1981) Robust Statistics. Wiley, New York.

5. Maronna RA (1976) Robust M-estimators of Multivariate Location and Scatter. The Annals of Statistics 4: 51-67.

6. Gnanadesikan R, Kettenring JR (1992) Robust Estimates, Residuals, and Outlier Detection with Multiresponse Data. Biometrica 28: 81-124.

7. Maronna RA, Zamar RH (2002) Robust Estimates of Location and Dispersion for High-Dimensional Datasets. Technometrics 44: 307-317.

8. Alqallaf FA, Konis KP, Martin RD, Zamar RH (2002) Scalable Robust Covariance and Correlation Estimates for Data Mining. Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, pp: 14-23.

9. Mallows CL (1973) Some Comments on Cp. Technometrics 15: 661-675.

10. Yohai VJ (1987) High Breakdown Point and High Efficiency Robust Estimates for Regression. The Annals of Statistics 15: 642-656.

11. Akaike H (1970) Statistical Predictor Identification. Annals of the Institute of Statistical Mathematics, 22: 203-217.

12. Schwartz G (1978) Estimating the Dimensions of a Model. The Annals of Statistics 6: 461-464.

13. Shao J (1993) Linear Model Selection by Cross-Validation. Journal of the American Statistical Association 88: 486-494.

14. Weisberg S (1985) Applied Linear Regression (2ndedn), Wiley-Interscience, New York.