Selective overview of forward selection in terms of robust correlations

Hassan S. Uraibi,[a,c] Habshah Midi,[b] and Sohel Rana[b]

[a] Laboratory of Computational Statistics and Operations Research,

Institute for Mathematical Research, UPM. 43400 Serdang, Selangor, Malaysia

[b] Department of Mathematics, Faculty of Science, UPM,43400 Serdang, Selangor, Malaysia

[c] Dept. of Statistics, College of Administration & Economics,University of Al-Qadisiyah, Iraq

hssn.sami1@gmail.com

Key Words: forward selection, RFCH, adjusted winsorization, robust correlation, robust variable selection.

ABSTRACT

Forward selection is a very effective variable selection procedure for selecting a parsimonious subset of covariates from a large number of candidate covariates. Detecting the type of outlying observations, such as vertical outliers or leverage points, and the forward selection procedure are inseparable problems. For robust variable selection, a crucial issue is whether the outliers are univariate, bivariate, or multivariate. This paper uses a $\sqrt{n}$ consistent robust multivariate dispersion estimator to obtain robust correlation estimators used to establish robust forward selection procedures that outperform methods that use robust bivariate correlations. The usefulness of our proposed procedure is studied with a numerical example and a simulation study. The result shows the proposed method has scalability and the ability to deal with univariate, bivariate and multivariate outlying observations including leverage points or vertical outliers, and the new method outperforms previously published methods of robust forward selection.

1. INTRODUCTION

As a consequence of the rapid development in computer and networking technologies in

recent years, the process of collecting large scale information has become easy. When there are a large number of variables, the curse of dimensionality is a major challenge for researchers. The challenge can be classified into two directions, the cost of the calculation and the time consumed. This issue that preoccupied the researchers led them to benefit from the proposed bivariate location and scatter estimators which have been utilized in the variable selection procedure to reduce the time of computation by clear mathematical calculations.

It is well known that the algorithms of classical variable selection were greedy and unstable, and that little changes in the data might result in at least one covariate to be chosen instead of another (Heterberge et al. 2008). In the last few years, considerable attention has been paid to improving and extending the general framework of Forward selection (FS) which is a very effective step-by-step procedure for choosing a useful subset from a lot of candidate covariates.

Hastie et al. (2001) proposed forward stagewise which a stable variable selection procedure that picks the same first covariate as FS and it changes the identical coefficient by a small amount and then it pays another small step for the variable that has the highest correlation with the current residuals. Unlike FS, forward stagewise takes many tiny steps to move toward the final model, and tends to obtain order between variables (Khan et al., 2007b). Efron et al. (2004) introduced Least Angle Regression (LARS) which is an algorithmic framework using the forward stagewise with lasso (Tibshirani, 1996) and boosting (Freund and Schapire, 1997) algorithms.

It is clear that the correlation coefficient is an essential issue in the previous variable selection methods, because the correlation coefficient between the response Y and the covariate X can be expressed in terms of orthogonal design as geometric angle $\theta_{X,Y}$ which is defined as follows:

$$Cos(\theta_{X,Y}) = \frac{< X, Y >}{\|X\|\|Y\|} = \frac{COV(X,Y)}{\sqrt{Var(X)}\sqrt{Var(Y)}} = Cor(X,Y) \quad (1)$$

where the normalization $\|X\|$ and $\|Y\|$ are 1 (unit length) and $\sum_i Y_i = 0,\ \sum_i X_i = 0\,,\ i = 1, \ldots, m,$ and m is a positive integer number greater than zero.

Unfortunately, the classical correlation coefficients are sensitive to the presence of outliers or other contamination. Consequently, where the correlation coefficient is non-robust the favourite variable selection procedure is non-robust too. One solution considered to overcome this issue was to propose a robust correlation method. The robustness literature shows a variety of approaches on robust correlations and a few of publications that addressed or discussed the issue of robust univariate and bivariate correlations e.g.,(Alqallaf et al., 2002; Maronna, 1976; Khan et al. 2007b).

Khan et al. (2007a) and Khan et al. (2007b) incorporated the robust bivariate correlation with FS and LARS respectively. The authors suggested a pairwise correlation approach, based on Maronna's M-estimates of the multivariate location and scatter matrix (Maronna, 1976) for FS and adjusted Winsorization in terms of correlations used with LARS. The major drawback of bivariate correlation is resistance only to bivariate outliers. However, three or higher-dimensional outliers may not be detected by univariate and bivariate analyses. Alfons et al. (2011) suggested a context-sensitive algorithm after LARS to overcome the problem of multivariate outliers. The authors seek, to do firstly dimensional reduction, and then weighting the observations, the last step the all possible subsets are used. This procedure is impossible with the classical forward selection.

Most of the authors in the robustness literature use a multivariate location and dispersion estimator, such as Fast-MCD (Rousseeuw and Van Driessen, 1999) to overcome the problem of outliers in high dimensional data since MCD is impractical to calculate. Olive and Hawkins (2010) proposed the Reweighted Fast Consistent High breakdown estimator (RFCH) which is backed by theory, and faster than Fast-MCD (Zhang et al. (2012)). Moreover it is feasible

option for many robust applications, e.g. Alkenani and Yu (2013) and Ozdemir and Wilcox (2012).

However, several practical questions arise when dealing with variable selection in terms of correlations, for instance, is it important to find robust multivariate correlation approach? Is it crucial to establish robust multivariate forward selection when the multivariate outliers are present in a data. Since the adjusted Winsorization correlations is only robust to bivariate outliers, it is very imperative to develop robust correlations which resistant to multivariate outliers. As such, we proposed robust multivariate correlations based on RFCH. Subsequently, this paper incorporates the correlations from the RFCH estimator instead of the adjusted Winsorization correlations in the development of Robust multivariate Forward Selection (RFS) procedure.

We will investigate the resistance of the RFS procedure to various types of outlier scenarios, and compare the results with the classical FS in terms of correlations and robust forward selection based on the adjusted Winsorization correlations (Khan et al. (2007b). The remainder of the paper is organized into four sections, the RFCH and the competing estimator are presented in section 2. The section 3 describes the robust forward selection in terms of robust correlation. To evaluate the performance of RFS procedure with the competing methods we consider an example and simulation study that will be discussed in section 4. The conclusion is reported in section 5.

2. BIVARIATE AND MULTIVARIATE ROBUST CORRELATION

The choice of an appropriate initial correlation matrix is an important issue and an essential part for robust FS procedure. In this section we review two approaches, one is the adjusted Winsorization correlation estimate and the multivariate location and dispersion RFCH estimator.

2.1 Adjusted Winsorization Correlation

Khan et al.(2007b) proposed an adjusted Winsorization correlation that is more resistant to bivariate outliers. They developed the univariate Winsorization correlation that was introduced by Alqallaf et al.(2002), by resolving the effect of bivariate outliers. Two tuning constants are put forward to overcome the problem of having more outliers: a tuning constant C_1 for the two quadrants that are bounded by the $2C_1 \times 2C_2$ square that contains the majority of the standardized data, and a smaller tuning constant C_2 for the other two quadrants. The initial correlation is obtained by computing the classical correlation of the adjusted Winsorized data.

Let $(X_i, Y_i), i = 1,2, \ldots,n$, be a random sample from a bivariate distribution with location parameters μ_X and μ_Y, and scale parameter σ_X and σ_Y, respectively.

(1) Standardized $X_i$and Y_i by their location parameters and scales.

$$\hat{X}_i = \frac{X_i - Med_X}{MAD_X}$$

$$\hat{Y}_i = \frac{Y_i - Med}{MAD_Y}$$

(2) The Huber psi function is given by

$$\psi(\hat{X}_i) = \begin{cases} \hat{X}_i & if\ |\hat{X}_i| \le C_1 \\ C_1\ sign(\hat{X}_i) & if\ |\hat{X}_i| > C_1 \end{cases}$$

$$\psi(\hat{Y}_i) = \begin{cases} \hat{Y}_i & if\ |\hat{Y}_i| \le C_1 \\ C_1\ sign(\hat{Y}_i) & if\ |\hat{Y}_i| > C_1 \end{cases}$$

where $C_1 = 2$

This univariate winsorization approach does not take into account the orientation of bivariate data. As a result, univariate correlation coefficient may be affected by some outliers which appear on the square boundaries. To overcome this problem (Khan et al., 2007b) adjusted the winsorized data as follows:

Suppose that $a_i = \psi(\hat{X}_i).\psi(\hat{Y}_i)$ and a_i^- is equivalent to $(a_i < 0)$. The remaining values of a_i be positive which is denoted by a_k^+. Let the n_1 and n_2 be the subset size of a_j^- and a_k^+ respectively. The smaller constant C_2 for the points in the two minor quadrants of $2C_1 \times 2C_2$ square.

$$C_2 = \begin{cases} C_1.\sqrt{\dfrac{n_1}{n_2}} & if\ n_1 \le n_2 \\ C_1.\sqrt{\dfrac{n_2}{n_1}} & if\ n_1 > n_2 \end{cases}$$

The Huber Psi function is used again to get rid of the effect of remaining outliers which do not exceed the C_1 point.

$$X_m^* = \begin{cases} \psi\left(\psi(\hat{X}_j)\right) = \begin{cases} \psi(\hat{X}_j) & \text{if } |a_i^-| \le C_2 \\ C_2\ \text{sign}(\psi(\hat{X}_j)) & \text{if } |a_i^-| > C_2 \end{cases} \\ \psi\left(\psi(\hat{X}_k)\right) = \begin{cases} \psi(\hat{X}_k) & \text{if } |a_i^-| \le C_2 \\ C_2\ \text{sign}(\psi(\hat{X}_k)) & \text{if } |a_i^-| > C_2 \end{cases} \end{cases},$$

$$Y_m^* = \begin{cases} \psi\left(\psi(\hat{Y}_j)\right) = \begin{cases} \psi(\hat{Y}_j) & \text{if } |a_i^-| \le C_2 \\ C_2\ \text{sign}(\psi(\hat{Y}_j)) & \text{if } |a_i^-| > C_2 \end{cases} \\ \psi\left(\psi(\hat{Y}_k)\right) = \begin{cases} \psi(\hat{Y}_k) & \text{if } |a_i^-| \le C_2 \\ C_2\ \text{sign}(\psi(\hat{Y}_k)) & \text{if } |a_i^-| > C_2 \end{cases} \end{cases},$$

where $j = 1, \dots, n_1$; and $k = 1, \dots, n_2$ and $m = 1, \dots, n = (n_1 + n_2)$

Then finding the simple Huber winsorized correlation is given by

$$r_w = \frac{\sum_{m=1}^{n} X_m^* . Y_m^* - \sum_{m=1}^{n} X_m^* . \sum_{m=1}^{n} Y_m^* .}{\sqrt{\sum_{m=1}^{n} X_m^{*2} - \frac{\left(\sum_{m=1}^{n} X_m^*\right)^2}{n}} . \sqrt{\sum_{m=1}^{n} Y_m^{*2} - \frac{\left(\sum_{m=1}^{n} Y_m^*\right)^2}{n}}} \quad (2)$$

(3) Suppose the initial bivariate variance-covariance matrix is the diagonal identity matrix. The adjusted Winsorized correlations are plugged into the initial bivariate variance-covariance matrix by their counterparts, with respect to the unit variance and the $cov(X_m^*, Y_m^*) = cor(X_m^*, Y_m^*)$. The combination between X_m^* and $, Y_m^*$ is included in one matrix, denoted $XY = (X_m^*, Y_m^*)^t$. To reduce the effect of outliers, the XY matrix is transformed using the bivariate transformation $u = min\left(\sqrt{\left(\frac{c}{D(XY)}\right)}, 1\right).XY,$ where $D(XY)$ is the Mahalanobis distance based on an initial bivariate correlation matrix. If the data follows a multivariate normal distribution, the squared Mahalanobis distance follows a χ_2^2 distribution. Here the tuning constant c = 5:99, the 95% quantile of the χ_2^2 distribution. Finally, find a simple correlation between each two variables based on clean data.

2.2 Reweighted Fast Consistent and High breakdown (RFCH)

Olive and Hawkins (2010) proposed a robust $\sqrt{n}$ consistent estimator that is called, the Fast Consistent, and High breakdown (FCH) estimator. The FCH employs the $\sqrt{n}$ consistent DGK estimator and the high breakdown Median Ball (MB) estimator) as attractors. The algorithm starts by generating a sequence of practical robust estimators from K trial fits. These are called attractors and are denoted by $(T_1, C_1), \ldots, (T_K, C_K)$. The concentration technique is then used to obtain the final estimator (T_A, C_A).

The classical estimator $(T_{-1,D}, C_{-1,D}) = (\bar{x}, S)$ is used as an initial estimator to get the DGK estimator $(T_{K,D}, C_{K,D})$, while the MB estimator $(T_{K,M}, C_{K,M})$ uses $(T_{-1,M}, C_{-1,M}) = (MED(X), I_p)$ as a start, where $MED(X)$ is the coordinate-wise median. If the DGK location estimator $T_{K,D}$ has a greater Euclidean distance from $MED(X)$ than half of the data, FCH uses the MB attractor. The FCH uses the smallest determinant as the dispersion criterion to choose the attractor, otherwise $\|T_{K,D} - MED(X)\|$.

Let (T_A, C_A) be the attractor used, then the location of FCH is $T_F = T_A$ and the scale is as follows:

$$C_F = \frac{MED(D_i^2((T_A, C_A)))}{\chi^2_{(p,0.5)}} \times C_A \quad (3)$$

where $\chi^2_{(p,0.5)}$ is the 50th percentile of a chi-square distribution with p degrees of freedom.

Olive and Hawkins (2010) used two standard reweighting steps for the RFCH estimator. Let $(\hat{\mu}_1, \widetilde{\Sigma}_1)$ be the classical estimator applied to the n_1 cases with $D_i^2(T_{FCH}, C_{FCH}) \leq \chi^2_{(p,0.975)}$ and let

$$\hat{\Sigma}_1 = \frac{MED(D_i^2(\hat{\mu}_1, \tilde{\Sigma}_1))}{\chi^2_{(p,0.5)}} \times \tilde{\Sigma}_1 \quad (4)$$

Then let $(T_{RFCH}, \widetilde{\Sigma}_2)$ be the classical estimator applied to the cases, with $D_i^2(\hat{\mu}_1, \widetilde{\Sigma}_1) \leq \chi^2_{(p,0.975)}$, and let

$$C_{RFCH} = \frac{MED(D_i^2((T_{RFCH}, \widetilde{\Sigma}_2)))}{\chi^2_{(p,0.5)}} \times \widetilde{\Sigma}_2 \quad (5)$$

Olive and Hawkins use results from Lopuhaa (1999) to prove that RFCH is a $\sqrt{n}$ consistent estimator of $(\mu, c\Sigma)$ for a large class of elliptically contoured distribution. However, the RFCH dispersion estimator is only conjectured to be high breakdown. The conjecture has not yet been proven.

3. FORWARD SELECTION BASED ON ROBUST CORRELATION

Two robust forward selection procedures were performed in order to verify the validity of the robustness against various types of outliers. The first one follows Khan et al. (2007a) but rather than using Maronna's M-estimates of the multivariate location and scatter matrix, we apply the adjusted Winsorization correlation which was also proposed by (Khan et al., 2007b). The adjusted Winsorization correlation is replaced by the correlation matrix based on RFCH estimator in the second approach.

Suppose we have d covariate variables $X_1, \ldots, X_d$, where $d \leq 50$, represented in a matrix X, and the response Y as a vector. Let each variable be robustly standardized based on its median and MAD. The FS.RFCH consists of the following steps:

Step 1. Split the original dataset into two subsets, a training subset and a test subset, and then scale the variables of both sets:

$$\widehat{X}_{itr} = \frac{X_{itr} - Med_{X_{itr}}}{MAD_{X_{itr}}} \quad , \quad \widehat{Y}_{tr} = \frac{Y_{itr} - Med_{Y_{itr}}}{MAD_{Y_{itr}}}$$

$$\widehat{X}_{its} = \frac{X_{its} - Med_{X_{its}}}{MAD_{X_{its}}} \quad , \quad \widehat{Y}_{ts} = \frac{Y_{its} - Med_{Y_{its}}}{MAD_{Y_{its}}}$$

Step 2. Define the active and inactive sets as follows: $\text{active} = \phi$ and $\text{inactive} = \{1,2,3, \ldots, d\}$

Step 3. Calculate the robust correlation favour estimator (adjusted Winsorzation or RFCH).

Step 4. Determine $m_{1=}|r_{j\ \hat{Y}_{tr}}|$, which is the highest absolute correlation between each covariate and the response variable $\hat{Y}_{tr}$,where $j = 1,2 \ldots d$, in order to choose the first covariate to be in the active vector. The remaining variables will be in the inactive vector. Say the first covariate is the active variable, which is written as $\text{active} \leftarrow \{1\}$, $\text{inactive} \leftarrow \{2,3, \ldots, d\}$, and then find the F test based on the correlations can be computed by

$$F_c = \frac{(n-2)r^2_{1\ \hat{Y}_{tr}}}{1-r^2_{1\ \hat{Y}_{tr}}} \tag{6}$$

If $F_c > F_{t(\alpha=0.05)}(1, n-2)$ then continue to the next step, otherwise stop. In fact, in this step we can also calculate the p-value for the model that relies on the value of F_c.

Step 5. Regress the $\hat{X}^*_{1\,ts}$ in the active vector on $\hat{Y}_{ts}$ using an FMM-estimation, and then weight the observations using the bisquare weighted function and compute the probability as follows:

$$Pr_i = \frac{w_i}{\sum w_i} \tag{7}$$

This procedure tries to assign zero weight to the outliers.

Multiplying $\hat{X}^*_{1\,ts}$ and $\hat{Y}_{ts}$ by Pr_i yields $\hat{X}^{*R}_{1ts}$ and $\hat{Y}^{*R}_{ts}$, where R stands for the remaining data after deleting the outlying observations. The fitted value based on the first covariate in the active set should be written such that $\hat{y}^R_{ts} = r_{1\ \hat{Y}_{tr}} \times \hat{X}^{*R}_{1ts}$ and the prediction error is the mean square of $\hat{Y}^{*R}_{ts} - \hat{y}^R_{ts}$.

Step 6. To select a new active covariate, find the partial correlation between the inactive covariates with the $\hat{Y}_{tr}$ vector adjusted for the active covariate. Say $j = 2, \ldots, d$, then the partial correlation is defined as $r_{j\hat{Y}_{tr}.1}$. Then determine $m_{2=}|r_{j\hat{Y}_{tr}.1}|$ to select the second candidate covariate for the active covariates vector (say $\hat{X}^*_2$).

Step 7. Select an appropriate robust criterion to test the significance of adding a new covariate $\widehat{X}_2^*$ to $\widehat{Y}_{tr}$. Here the Partial F-test (P. F) is used to decide whether the new covariate should be added to $\widehat{Y}_{tr}$ or the algorithm should be stopped:

$$P.F_c = \frac{(n-3)r^2_{\widehat{Y}_{tr.1}}}{1-r^2_{1\,\widehat{Y}_{tr}}-r^2_{2\,\widehat{Y}_{tr.1}}} \tag{8}$$

If $P.F_c > F_{t(\alpha=0.05)}(2, n-3)$ then continue to the next step, otherwise stop. The p-value for the model is computed. Calculate the robust prediction error again in the same way as in step (5), with the intention of fitting all covariates in the active set.

Step 8. Repeat steps 6 and 7 and use equation (7) to find $P.F_c$, then continue until the null hypothesis is accepted.

At each step of the forward selection, once one covariate among the remaining covariates is considered for inclusion in the model, several selection criteria are employed to decide whether to include this covariate in the model and continue the process or to stop. Three robust selection criteria, namely the Partial F-test that was introduced by Khan et al. (2007a), the robust prediction error and the p-value of the total model, are applied to confirm whether the algorithm has selected the correct model.

4. EXAMPLE AND SIMULATION

Three methods are considered to evaluate the accuracy of the selection of the best model, namely classical FS, robust FS.Winso, and robust FS.RFCH. The Partial F-statistic criterion used by Khan et al. (2007a) for stopping the algorithm is used. The Partial F-statistic is robust if a robust correlation matrix is used, but is not robust if the classical correlation is used. The best covariate that enters the model is the one that has P:F greater than $F(0:95;\ k;\ n-k-1)$. The performance of the classical and the two proposed methods is evaluated according to three criteria. First, the method should select the correct variables (for real data this is similar to the standard model selected in the previous study (1,3,4,2,5) and should have the optimal

values for the Robust Prediction Error (RPE) and the signi cance p-value. The method that produces this model is better than the others. The optimal value of RPE is not necessarily the smallest one, because the RP E that is used with real data is the square root of the mean square error multiplied by the length of the nal model. This procedure is put forward to avoid the effect of the trade-o between bias and variance. In this case the method that selects an under-fitted model might produce the smallest RPE, and it would definitely be wrong to take a decision based on this result. By contrast, in the case when the final model is over-fitted, the RP E will be higher than the RP E of the correct model. In the simulation study, because we know the correct covariates in the empirical test, we propose a new criterion that takes into account the effect of the final model length and the number of the correct covariates in it. We call this criterion the Optimal Prediction Error (OP E), and it is given by

$$\mathrm{OPE} = \sqrt{\mathrm{MSE_e}} \times \frac{\mathrm{L}}{\mathrm{P_v}} \tag{9}$$

where $\mathrm{MSE_e}$ is the mean squares error of the final model, L is the final model size and $\mathrm{P_v}$ is the number of correct covariates in the final model.

4.1 Example

Data on executives are taken from Mendenhall et al. (1996), who present the annual salary of 100 executives corresponding to 10 potential predictors (7 quantitative and 3 qualitative) such as education and experience, and these are used to evaluate the proposed method. The candidate predictors are labelled from 1 to 10. This data is also used by Khan et al. (2007a); they are clean data (no outliers) and there is no multicollinearity. The original dataset is modified to have Leverage Points (LP) and Vertical Outliers (VO).

Figure 1 shows the modifying effect of vertical outliers on the data. Table I shows the results for the data on executive pay without any contamination. All three methods select the covariates 1,3,4,2,5 as the best model. It can be observed that the three methods show a minimum value for RP E when the full model includes the covariates 1,3,4,2,5, and that the

Partial F-test (P:F) values are greater than the F Table (F:T) values. We show that the results for RPE for the robust FS.Winso, the robust FS.RFCH and the improved methods are very close, and all three methods select the standard model.

To investigate the effect of a single leverage point on the variable selection, we follow Khan et al. (2007a) by replacing one small value of predictor 1 (which was less than 5) by the large value 100. The results in Table II show that our proposed method and the robust FS.Winso method select the same model and meet the three criteria. However, the FS method fails to select the standard model. The FS.Winso procedure is robust against a single outlier.

Next, we contaminated the first covariate variable 1 by 10% leverage points. We replaced 10 good observations by 100 randomly selected values for predictor 1. As can be seen from Table III, the classical method and the robust FS.Winso method fail to select the standard model. The classical method chooses covariate 3, while the robust FS.Winso method selects the covariates 3,10 as the best model. It is interesting to see that our proposed method (FS.RFCH) selects the standard model 1,3,4,2,5 with the least values of RPE.

To investigate the effect of vertical outliers, we contaminated with a single V.outlier by replacing the observation number 4 in Y by 100. The results in Table IV show that the classical method failed to select any covariate and to stop, while the robust FS.Winso and robust FS.RFCH selected the standard model with very close results against the criteria.

In the same way, we created 10% vertical outliers by substituting observations arbitrary selected in Y with 100. The results in Table V show the reliability of FS.RFCH against the vertical outliers.

In the previous examples, we focused only on the covariate of the first predictor. Next we investigated the effect of 10% LP and 10% vertical outliers on covariates 1, 2, 7, 8, and 10. We distributed the LP on the first five predictors by replacing one randomly selected good

observation by a value of 100 in each predictor, and then we replaced 10 randomly selected observations in the y direction by 100, with the condition that there was a different level of LP.

Figure 2 above explains the effect of vertical outliers and leverage points in this scenario of contamination. Table VI shows the results for the three methods that are used for comparison. The results are similar to the previous results in that only the FS.RFCH method selects the standard model, while the other methods select incorrect models. The last simulation was the same as the previous simulation, but the 10% vertical outliers and leverage points lay on the same level. Table VII shows our proposed procedure still doing well, and FS.Winso in this simulation produces an under-fitting model by losing only one variable.

4.2 Simulation Study

A simulation study similar to that in Meinshausen and Buhlmann (2010) for the case when there is no correlation was carried out to investigate the behaviour of our proposed method compared with the classical FS and FS.Winso. The design matrix came from a centred multivariate normal distribution with covariance structure $Cov(X_j, X_k) = \rho^{|j-k|}$, where in this study we consider $\rho = 0$ when $k \neq j$ (no multicollinearity). The total of independent standard normal covariates is $d = 30$. We select the rst ve ($a = 5$) or nine ($p = 9$) as non-zero covariates (or active predictors) to create two linear models as follows:

$$\mathrm{Y} = 5\mathrm{X}_1 + 4\mathrm{X}_2 + 3\mathrm{X}_3 + 2\mathrm{X}_4 + \mathrm{X}_5 + \sigma_{\mathrm{e}} \tag{10}$$

$$\mathrm{Y} = 9\mathrm{X}_1 + 8\mathrm{X}_2 + 7\mathrm{X}_3 + \cdots + \mathrm{X}_9 + \sigma_{\mathrm{e}} \tag{11}$$

We follow Alfons et al. (2011) in choosing $_{\mathrm{e}}$ so that the signal-to-ratio ($Signois$) is defined as follows:

$$Signois = \sqrt{\frac{Var(X_1+X_2+\cdots+X_P)}{Var(\sigma e\ )}} \tag{12}$$

The remaining standard normal covariates $(d - p)$ are considered as noise. 500 datasets were generated, each of which was randomly divided into two samples, a training sample of size 250

and a test sample of size 250. Each process was repeated for 500 simulation runs. The following scenarios were considered:

(1) No outliers

(2) Vertical outliers (VO): the contamination is given by replacing 10% of randomly selected σ_e by a large number.

(3) Univariate Leverage Points (ULP) and VO : as in 1, but one of the active predictors was contaminated with 10% of leverage points.

(4) Bivariate Leverage Points (BLP) and VO: this was similar to 3, but the contamination included two active predictors with 10% leverage points.

(5) Multivariate Leverage Points (MLP) and VO: the contamination in y is given by replacing a certain percentage of randomly selected observations by a large number, and for each such observation some or all active predictors are also replaced by a large number.

The selection criterion consists of two stages. At the first stage, the P.F statistics (should be robust where the correlations in use are robust too) values are used to select the best model of the training sample in each iteration, and then the robust RPE for the best model selected is calculated based on the test sample. For each simulated dataset we recorded the model size, the number of non-zero coefficients and the number of noise variables in the selected model.

To evaluate the performance of the three methods, the average of the noise variables, denoted as AV.Noise, the average of the model size, denoted as AV.Model Size, the average of non-zero coefficients, denoted as AV.Nonzero, and the average of OPE, denoted as AV.OPE are recorded over all training sets. The least value for AV.RPE (the average of RPE) is found where the particular method selects the standard model many times more than other methods. All methods fitted the selected model without including the intercept on the training dataset, and then they used this to predict the test dataset outcomes. The FMM-estimator introduced by

(Yohai, 1987) is used to fit the model obtained by the robust methods, and then the weights of the bisquare function are used. Only the non-zero weighted observations are used to nd the RPE with the test set. The best method is the one that has the least values for AV.RPE and Av.Noise. Moreover, we have to include the numbers of non-zero and zero coefficients in our consideration. Table VIII presents the RPE, the AV.Nonzero, the AV.Noise, the AV.Model Size and the AV.OPE of 500 simulation runs.

For the clean data (0% outliers) the performance of FS.RFCH is as good as FS and better than FS.Winso, whether $p = 5$ or 9. It selects the lowest noise variables, has a reasonable average model size, and has the lowest AV.RPE and the lowest AV.OPE. Although FS.Winso selected a higher average of non-zero covariates, the difference does not exceed the appearance or disappearance of a single variable. In the cases with 10% of vertical outliers, and 10% of vertical outliers and univariate or bivariate LP, present in the dataset, Table VIII shows that both FS.RFCH and FS.Winso performed better than FS. The difference in the performance of three methods is evident in Table VIII when the dataset was contaminated by 10% vertical outliers and multivariate leverage points. The proposed method was more consistent in the selection of non-zero covariates, in the number of noise variables, and in the optimal prediction error. Note that the FS.Winso performs much better than FS and that its performance is very close to FS.RFCH when univariate and bivariate outliers appear in the dataset.

5. CONCLUSION

Based on the results, it can be concluded that the improvement of the performance of the forward selection procedure has been very successful, considering that it is quite convincing, and thus the following conclusions can be drawn. The main target of this study is achieved where as the robust forward selection based on RFCH is more reliable than one based on bivariate correlation where multivariate outliers are existing in the data set.

The comparison has clearly shown that the FS and FS.Winso methods tend to over- t even with clean data, in contrast to FS.RFCH which selects the reasonable model. We have noticed that this problem has been inherent to FS.Winso method even though the univariate and bivariate LP points were present in the data. On the other hand, we noticed that FS.RFCH was more stable and consistent than the other methods. Summing from the results of real data and simulation, it can be noted that the controlling the error selection procedure in the FS.RFC algorithm is an important reason behind stability feature.

Consequently, it can be concluded that the adjusted Winsorization correlation matrix is sensitive outliers. The correlation matrix from the robust variance-covariance matrix in RFCH is more resistant to univariate, bivariate and multivariate outliers, for the outlier configurations considered in the simulation. The proposed robust FS.RFCH procedure can be readily used in practice as a remedy for the problem of having more outliers in the original dataset.

This article was concerned the sample size $n \geq 10 \times d$ (Real data and Simulated data) and $n > 5 \times d$ (training and test sets) where d is the number of predictors. However, the results should be applicable also to one case high dimensional data, that when the sample size $n \geq c \times d$ where $c > 5$ .Therefore on the basis of the promising ndings presented in this paper, variable selection for high dimensional data based on RFCH and will be presented in future papers.

BIBLIOGRAPHY

Alfons, A., Baaske, W. E, Filzmoser, P., Mader, W.,& Wieser, R.. (2011). Robust variable selection with application to quality of life research. Statistical Methods & Applications, 20(1), 65-82.

Alkenani, A., & Yu, K. (2013). A comparative study for robust canonical correlation methods. Journal of Statistical Computation and Simulation, 83(4), 692-720.

Alqallaf, F. A., Konis, K. P, Martin, R D., & Zamar, R. H. (2002). Scalable robust covariance and correlation estimates for data mining. Paper presented at the Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R.. (2004). Least angle regression. The Annals of statistics, 32(2), 407-499.

Ozdemir, A, & Wilcox, R. (2012). New results on the small-sample properties of some robust univariate estimators of location. Communications in Statistics-Simulation and Computation, 41(9), 1544-1556.

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of computer and system sciences, 55(1), 119-139.

Hastie, T., Tibshirani, R. & Friedman, J. (2001), The Elements of Statistical Learning; Data mining, Inference and Prediction, Springer Verlag, New York.

Hesterberg, T., Choi, Nam H., Meier, L., & Fraley, C.. (2008). Least angle and L1 penalized regression: A review. Statistics Surveys, 2, 61-93.

Khan, J. A, Van A. S., & Zamar, R. H. (2007a). Building a robust linear model with forward selection and stepwise procedures. Computational Statistics & Data Analysis, 52(1), 239-248.

Khan, J. A, Van A., S., & Zamar, R. H. (2007b). Robust linear model selection based on least angle regression. Journal of the American Statistical Association, 102(480), 1289-1299.

Lopuhaa, H. P. (1999). Asymptotics of reweighted estimators of multivariate location and scatter. Annals of Statistics, 1638-1665.

Maronna, R. A. (1976). Robust M-estimators of multivariate location and scat-ter. The Annals of Statistics, 4: 5167.

Meinshausen, N.,& Buhlmann, P. (2010). Stability selection. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 72(4), 417-473.

Mendenhall, W., Sincich, T., & Boudreau, N. S. (1996). A second course in statistics: regression analysis (Vol. 5): Prentice Hall Upper Saddle River New Jersey New Jersey.

Olive, D. J, & Hawkins, D. M. (2010). Robust multivariate location and dispersion. Preprint, see (www. math. siu. edu/olive/preprints. htm).

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 267-288.

Yohai, V. J. (1987).High breakdown-point and high e ciency robust estimates for regression. The Annals of Statistics, 642-656.

Zhang, J., Olive, D. J, & Ye, P.. (2012). Robust covariance matrix estimation with canonical correlation analysis. International Journal of Statistics and Probability, 1(2), p119.