RESEARCH ARTICLE

doi: 10.2306/scienceasia1513-1874.2017.43.056

Robust multivariate least angle regression

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

[a] Department of Statistics, College of Administration and Economics, University of Al-Qadisiyah, 50082, Iraq
[b] Institute for Mathematical Research, University Putra Malaysia, 43400 UPM, Serdang, Malaysia
[c] Department of Mathematics, Faculty of Science, University Putra Malaysia, 43400 UPM, Serdang, Malaysia
[d] Department of Applied Statistics, Faculty of Sciences and Engineering, East West University, Aftabnagar, Dhaka-1212, Bangladesh

[*]Corresponding author, e-mail: hssn.sami1@gmail.com, hassan.uraibi@qu.edu.iq

Received 14 Jan 2016
Accepted 1 Dec 2016

ABSTRACT: The least angle regression selection (LARS) algorithms that use the classical sample means, variances, and correlations between the original variables are very sensitive to the presence of outliers and other contamination. To remedy this problem, a simple modification of this algorithm is to replace the non-robust estimates with their robust counterparts. Khan, Van Aelst, and Zamar employed the robust correlation for winsorized data based on adjusted winsorization correlation as a robust bivariate correlation approach for plug-in LARS. However, the robust least angle regression selection has some drawbacks in the presence of multivariate outliers. We propose to incorporate the Olive and Hawkins reweighted and fast consistent high breakdown estimator into the robust plug-in LARS method based on correlations. Our proposed method is tested by using a numerical example and a simulation study.

KEYWORDS: variable selection, least angle regression selection, RFCH, adjusted winsorization

INTRODUCTION

During the last decade the problem of variable selection, which is a result of unreliable data quality and many covariates, has been studied using least angle regression selection (LARS)[1], adaptive lasso[2], robust LARS[3], robust LARS based on S-estimators[4], and sparse partial robust M (SPRM) regression estimators[5]. Most researchers have showed that fitting all possible subsets and sequential methods like stepwise selection are not feasible options, being very time consuming. Furthermore, when the predictors are correlated, these methods not only omit some predictors that exhibit small effects, but may fail to include some covariates that exhibit big effects.

Recently, many selection procedures have been developed to remedy the problem of having many predictors. These use the selection strategy of two-step model building. The first step involves a sequencing that aims to place all candidate covariates in an order such that the more important ones are likely to be placed at the beginning. The second step is a segmentation step whereby a subset of m (determined by the user's experience) candidate variables are carefully examined from the list of variables in the sequencing step in order to select the final model. In this paper, we mainly focus on the sequencing step.

Since the formulation of LARS is based on the classical correlation matrix, it is sensitive to outliers. Khan et al[3] proposed a robust version of LARS based on two approaches (plug-in and data cleaning) of robust bivariate correlation estimates which can be efficiently computed using bivariate winsorization. These types of correlations are robust only to bivariate outliers. However, three- or higher-dimensional outliers may not be detected by univariate and bivariate analyses. Khan, Van Aelst, and Zamar[6] mentioned that the correlation matrix obtained from the pairwise correlation approach may not be positive definite, forcing the use of a correction for positive definiteness in some cases[7]. These problems have motivated us to improve this strategy by using a fast and robust multivariate location and dispersion that is robust to multivariate outliers. Subsequently, a robust correlation matrix will be formulated. Olive and Hawkins[8] suggested using a reweighted fast consistent and high breakdown (RFCH) estimator that uses a standard method for reweighting a fast consistent high breakdown (FCH) estimator and gives an easily computed $\sqrt{n}$-consistent estimator robust to outliers. FCH estimators are based on two attractors, namely, the Devlin-Gnanadesikan-Kettenring (DGK) and the median ball estimators with some kind of location criterion. The RFCH estimator differs from the robust bivariate correlation in that it is robust to multivariate outliers. In this regard, we propose to incorporate the RFCH robust correlation matrix[9] instead of robust bivariate correlation in the establishment of the robust LARS (RLARS) procedure. We will investigate the robustness of the RLARS procedure to different types of outliers, and compare the results with the method proposed by Khan, Van Aelst, and Zamar[3]. For this purpose, we consider the synthetic data that was introduced in Ref. 10.

REWEIGHTED FAST CONSISTENT AND HIGH BREAKDOWN ESTIMATOR

Olive and Hawkins[8] proposed the RFCH estimator as a robust multivariate location and dispersion estimator which is consistent and highly robust to outliers. The algorithm starts by generating a sequence of practical robust estimators from K trial fits, which are called attractors, and are denoted by $(T_1, C_1), \dots, (T_K, C_K)$. Then it uses the concentration technique to obtain the final estimator (T_A, C_A) that minimizes some criterion. The FCH estimator uses the $\sqrt{n}$-consistent DGK[11] estimator and high breakdown median ball (MB) estimator[12]. The classical estimator $(T_{-1,D}, C_{-1,D}) = (\bar{x}, S)$ is used as the initial estimator to obtain the DGK estimator $(T_{K,D}, C_{K,D})$, while the MB estimator $(T_{K,M}, C_{K,M})$ uses $(T_{-1,M}, C_{-1,M}) = (\mathrm{MED}(X), I_p)$ to start with, where $\mathrm{MED}(X)$ is the coordinatewise median. If the DGK location estimator, denoted by $T_{K,D}$, has a greater Euclidean distance than $\mathrm{MED}(X)$, then the FCH uses the MB attractor. The FCH uses the smallest determinant as the location criterion to choose the attractor if

$$\|T_{K,D} - \mathrm{MED}(X)\| \leqslant \mathrm{MED}(D_\mathrm{i}(\mathrm{MED}(X), I_p)). \quad (1)$$

Let (T_A, C_A) be the attractor. Then the location of FCH is $T_\mathrm{F} = T_A$ and the scale is denoted as follows:

$$C_\mathrm{F} = \frac{\mathrm{MED}(D_\mathrm{i}^2(T_A, C_A))}{\chi^2_{(p,0.5)}} C_A, \quad (2)$$

where $D_\mathrm{i}^2(T_A, C_A)$ is the square of the Mahanalobis distance and $\chi^2_{(p,q)}$ is the 100qth percentile of a chi-squared distribution with p degrees of freedom. Olive and Hawkins[8] used two standard reweighting steps for the RFCH estimator. Let $(\hat{\mu}_1, \tilde{\Sigma}_1)$ be the classical estimator applied to the n_1 cases with $D_\mathrm{i}^2(T_\mathrm{FCH}, C_\mathrm{FCH}) \leqslant \chi^2_{(p,0.975)}$ and let

$$\hat{\Sigma}_1 = \frac{\mathrm{MED}(\hat{\mu}_1, \tilde{\Sigma}_1)}{\chi^2_{(p,0.5)}} \tilde{\Sigma}_1. \quad (3)$$

Then let $(T_\mathrm{RFCH}, \tilde{\Sigma}_2)$ be the classical estimator applied to the cases with

$$D_\mathrm{i}^2(\hat{\mu}_1, \tilde{\Sigma}_1) \leqslant \chi^2_{(p,0.975)}, \quad (4)$$

$$C_\mathrm{RFCH} = \frac{D_\mathrm{i}^2(T_\mathrm{RFCH}, \tilde{\Sigma}_2)}{\chi^2_{(p,0.5)}} \tilde{\Sigma}_2. \quad (5)$$

Olive and Hawkins[8] use results from Ref. 13 to prove that the RFCH estimator is a $\sqrt{n}$-consistent estimator of $(\mu, c\Sigma)$ for a large class of elliptically contoured distributions.

PLUG-IN RLARS BASED ON RFCH

Suppose we have $d > 50$ covariates, variables $X_1, \dots, X_d$ represented in matrix X. Consider the response Y as a vector. Let each variable be standardized based on its median and median absolute deviation. Consider the linear model without intercept. The steps for plug-in RLARS are as follows.

(i) It starts with $\hat{\mu} = 0 \in \mathbb{R}^n$. Let μ_A be the current predictor, and $\mathrm{Cor}_\mathrm{RFCH} = (X^\mathrm{t}(Y - \hat{\mu}_\zeta))$ is the vector of current correlations $r_{j,\mathrm{RFCH}} = \mathrm{Cor}_\mathrm{RFCH}(Y - \hat{\mu}_\zeta, X_j)$ where $j = 1, \dots, p$.

(ii) Let ζ denote the active set and initially $\zeta = \phi$. Only the covariates with the largest absolute correlations will be considered to enter ζ. Set $R_\mathrm{RFCH} = \max_j |r_{j,\mathrm{RFCH}}|$,

$$\zeta = \{j : |r_{j,\mathrm{RFCH}}| = R_\mathrm{RFCH}\}, \quad (6)$$

and without loss of generality, $\zeta = 1, \dots, m$.

(iii) Let $s_j = \{\mathrm{sgn}(r_{j,\mathrm{RFCH}}) : j \in \zeta\}$. Then let $X_\zeta \in \mathbb{R}^{(m \times n)}$ be the matrix of active covariates which is constructed by the corresponding signed columns of the design matrix $X, s_j X_j$. Note that the unit vector $u = v_\zeta / \|v_\zeta\|$ makes equal angles with the columns of the X_ζ, where

$$v_\zeta = X_\zeta (X_\zeta^\mathrm{t} X_\zeta)^{-1} 1_\zeta \quad (7)$$

which satisfies

$$X_\zeta^\mathrm{t} u_\zeta = A_\zeta 1_\zeta \quad (8)$$

where $A_\zeta = 1/\|v_\zeta\| \in \mathbb{R}$. LARS modifies the current fit $\hat{\mu}_\zeta$ to

$$\hat{\mu}_\zeta \leftarrow \hat{\mu}_\zeta + \delta \mu_\zeta. \quad (9)$$

Since LARS takes the step in the direction of $s_j X_j$ by a certain distance δ, where δ is a positive number which is chosen as the smallest step to control the speed and greediness of the LARS algorithm, it can be expressed in terms of the correlation between the variables. The step δ should be chosen so that the residual $Y - \hat{Y}$ has equal correlation with $s_j X_j$ and another covariate, say X_k. Consequently, δ bisects the angle between X_j and another inactive covariate X_k with equal correlation. This covariate enters the model and the active set becomes

$$\zeta \leftarrow \zeta \cup \{K\}. \tag{10}$$

Note that for each updated δ,

$$\hat{\mu}(\delta) = \hat{\mu}_\zeta + \delta u_\zeta \tag{11}$$

and so for each $j = 1, \dots, p$ we have

$$\begin{aligned} r_{j,\mathrm{RFCH}}(\delta) &= \mathrm{Cor}_{\mathrm{RFCH}}(Y - \hat{\mu}(\delta)) \\ &= X_j^\mathrm{t}(Y - \hat{\mu}(\delta)) = r_j - \delta a_j, \end{aligned} \tag{12}$$

where $a_j = X_j^\mathrm{t} u_\zeta$. Hence (8) implies

$$|r_{j,\mathrm{RFCH}}(\delta)| = R_{\mathrm{RFCH}} - \delta A_\zeta. \tag{13}$$

The procedure is then repeated for all inactive variables in sequence based on their importance.

SIMULATION

In this section we report on a simulation study similar to Ref. 14. The contaminated observations are simulated in a similar way to Ref. 4, where the correlation between the covariates is weak and no outliers are in the data set. A design matrix coming from a centred multivariate normal distribution with covariance structure $\mathrm{Cov}(X_j, X_K) = \rho^{|j-K|}$ where $\rho = 0$ is considered. The response variable Y is generated using $P = 9$ covariates that have non-zero coefficients and $d - P$ covariates with coefficients equal to zero, where the d are selected to construct the design matrix with dimension 500×50. The non-zero coefficients are selected randomly at each iteration.

We generated 500 simulated data sets using the following.

(i) $Y = X\beta + \varepsilon$, where $\varepsilon \sim N(0,1)$.

(ii) $Y = X\beta + \tilde{\varepsilon}$, where $\tilde{\varepsilon} = 0.90(\varepsilon) + 0.10(\hat{\varepsilon}), \tilde{\varepsilon}$ is contaminated by 10% symmetric normal outliers with the slash distribution, $\hat{\varepsilon} \sim \varepsilon / u(0,1)$.

Table 1 The average of the top potential covariates of 500 simulated data sets for clean data generated by case (i).

Q	$n = 500$		$n = 1000$		$n = 5000$	
	RLARS-Winsor	RLARS-RFCH	RLARS-Winsor	RLARS-RFCH	RLARS-Winsor	RLARS-RFCH
2	2.00	2.00	2.00	2.00	2.00	2.00
3	3.00	3.00	3.00	3.00	3.00	3.00
5	5.00	5.00	5.00	5.00	5.00	5.00
10	8.66	9.00	8.60	9.00	8.00	9.00
20	8.84	9.00	8.88	9.00	8.00	9.00
30	8.90	9.00	8.90	9.00	9.00	9.00
40	8.94	9.00	8.92	9.00	9.00	9.00
50	9.00	9.00	9.00	9.00	9.00	9.00

Table 2 The average of top potential covariates of 500 simulated data set for clean data generated by case (ii).

Q	$n = 500$		$n = 1000$		$n = 5000$	
	RLARS-Winsor	RLARS-RFCH	RLARS-Winsor	RLARS-RFCH	RLARS-Winsor	RLARS-RFCH
2	1.98	2.00	2.00	2.00	2.00	2.00
3	2.98	3.00	3.00	3.00	3.00	3.00
5	4.96	5.00	5.00	5.00	5.00	5.00
10	8.52	9.00	8.80	9.00	8.94	9.00
20	8.76	9.00	8.88	9.00	8.94	9.00
30	8.78	9.00	8.90	9.00	8.96	9.00
40	8.88	9.00	8.94	9.00	8.98	9.00
50	9.00	9.00	9.00	9.00	9.00	9.00

(iii) $Y = X\beta + \dot{\varepsilon}$ where $\dot{\varepsilon} = 0.90(\varepsilon) + 0.10(\ddot{\varepsilon})$ and regression residuals $\dot{\varepsilon}$ are contaminated by 10% asymmetric normal outliers $\ddot{\varepsilon} \sim N(20,1)$, and 10% good observations of all predictors are randomly replaced by another observation which are generated from an $N(50,1)$ distribution to create bad leverage points.

The average number of correctly selected predictors in the top of sequence list which is taken to select the final model. We consider the number of the target potential covariates (predictors) that should appear in the top of the sequence of the selected covariates list, Q, equalling 2, 3, 5, 10, 20, 30, 40, and 50.

A good method is one that is able to select the correct number of potential covariates in the top of the sequence. From the simulation, we chose only 9 potential covariates to construct Y. The results of the simulation are presented in Tables 1, 2, and 3. We observe that the RLARS-RFCH is consistently able to select the correct covariates. It can be seen that the number of selected potential variables for the RLARS-RFCH is consistent with the simulated data. The results are consistent for sample sizes 500, 1000, and 5000.

NUMERICAL EXAMPLE

In this section we use a synthetic data set to illustrate the performance of the RLARS method based

Table 3 The average of top potential covariates of 500 simulated data set for clean data generated by case (iii).

Q	$n = 500$		$n = 1000$		$n = 5000$	
	RLARS-Winsor	RLARS-RFCH	RLARS-Winsor	RLARS-RFCH	RLARS-Winsor	RLARS-RFCH
2	1.98	2.00	2.00	2.00	2.00	2.00
3	2.98	3.00	3.00	3.00	3.00	3.00
5	4.96	5.00	5.00	5.00	5.00	5.00
10	8.52	9.00	8.76	9.00	8.88	9.00
20	8.78	9.00	8.88	9.00	8.94	9.00
30	8.88	9.00	8.90	9.00	8.98	9.00
40	8.96	9.00	8.94	9.00	8.98	9.00
50	9.00	9.00	9.00	9.00	9.00	9.00

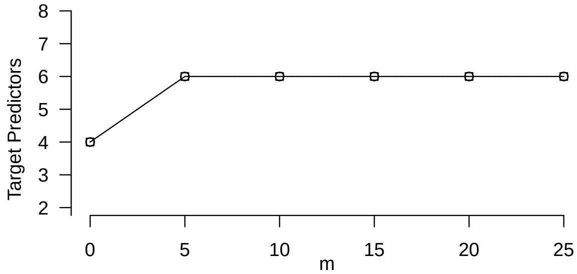

Fig. 1 Average number of correctly selected predictors for various *m* for the original artificial dataset. In this are the remaining figures, circles are data points selected by the RLARS-RFCH method, and squares are points selected by the RLARS-Winsor method.

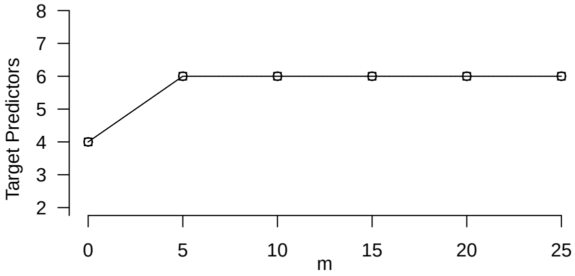

Fig. 2 Average number of correctly selected predictors for various *m* for the artificial data set contaminated with 100 LPs.

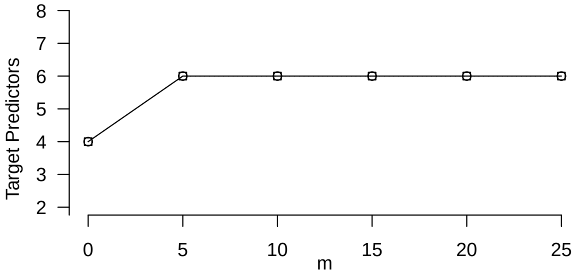

Fig. 3 Average number of correctly selected predictors for various *m* for the artificial data set contaminated with 100 vertical outliers.

on RFCH correlation. The synthetic data is taken from Ref. 10 which presents 1000 observations corresponding to 200 candidate predictors which are labelled from 1–200. Only the predictors 83, 33, 42, 59, 96 and 172 have non-zero coefficients. The response variable is generated from the true potential predictors with non-zero coefficients.

We created 50 data sets from the original one with the same dimensions (1000 × 200). The difference between one data set and another is the contamination of certain covariates. Those covariates are randomly selected to be within 10% of the leverage point (LP) and vertical outliers. The positions of those outliers are selected randomly. For each data set, we consider *m* potential covariates (i.e., the model size is *m*) that appear in the top of the sequencing step of LARS, such that $m = 5, 10, 15, 20, 25$ are recorded.

The best method is the one that includes all or most target variables, as a function of model size that is considered according to *m* values. The performance of the LARS method based on the RFCH estimator and adjusted winsorization is shown in Figs. 1, 2, 3, and 4. It can be observed that in Fig. 1 for clean data, both the adjusted winsorization correlation and the RFCH correlation can correctly identify the same number of predictors. We noted that when the LPs and vertical outliers are present together in the data set, our proposed method performs better than the Khan, Van Aelst, and Zamar algorithm[3]. As can be seen from Fig. 4, the RLARS method based on Winsor fails to select all or most target predictors. On the other hand, RLARS method based on RFCH selects all target predictors regardless of the outlier's position.

DISCUSSION

It is evident from Figs. 1–3 that the results from our method match those from the Khan, Van Aelst, and Zamar algorithm. When $m = 5$, both methods select 5 target variables, and select all target variables when $m = 10$, 15, 20, and 25. Note that when the LPs and vertical outliers are present together in the data set, our proposed method is better than that of Khan, Van Aelst, and Zamar. As can be seen from 4, the RLARS procedure based on Winsor fails to select all or most of the target predictors. On the other hand, the RLARS procedure based on the RFCH estimator selects all target predictors regardless of the outlier's position.

The main focus of this article was to propose a

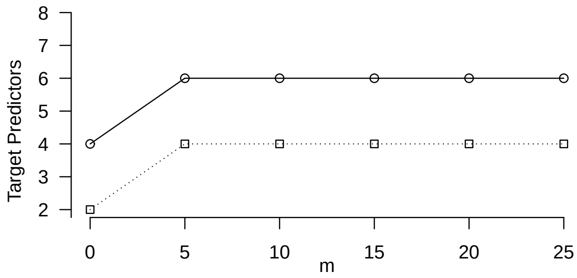

Fig. 4 Average number of correctly selected predictors for various *m* for the artificial data set contaminated with 100 LPs and 100 vertical outliers.

new method for linear regression model selection based on a RLARS-RFCH method. The empirical studies show that the performance of the RLARS-RFCH and RLARS-Winsor methods are equally good for clean data. However, in the presence of contaminated data or outliers, the RLARS-Winsor is less efficient whereas the RLARS-RFCH is very successful at selecting the correct covariates. Hence we can consider the RLARS-RFCH as a better variable selection technique and recommend using this technique particularly when outliers are present in the data.

See http://www.researchgate.net/publication/304056513 for the *R* code used in this article.

Acknowledgements: The authors thank the Ministry of Science, Technology & Innovation, Malaysia for supporting this work under eScienceFund Research Grant No. 06-01-04-SF1764.

REFERENCES

1. Efron B, Hastie T, Iain J, Tibshirani R (2004) Least angle regression. *Ann Stat* **32**, 407–99.
2. Zou H (2006) The adaptive lasso and its oracle properties. *J Am Stat Assoc* **101**, 1418–29.
3. Khan J, Van Aelst S, Zamar R (2007) Robust linear model selection based on least angle regression. *J Am Stat Assoc* **102**, 1289–99.
4. Agostinelli C, Barrera M (2010) Robust model selection with LARS based on S-estimators. In: *Proceedings of COMPSTAT2010*, pp 69–78.
5. Hoffman I, Serneels S, Filzmoser P, Croux C (2015) Sparse partial robust M regression. *Chemometr Intell Lab Syst* **149A**, 50–9.
6. Khan J, Van Aelst S, Zamar R (2007) Building a robust linear model with forward selection and stepwise procedures. *Comput Stat Data Anal* **52**, 239–48.
7. Alqallaf F, Konis K, Martin R, Zamar R (2002) Scalable robust covariance and correlation estimates for data mining. In: *Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp 14–23.
8. Olive D, Hawkins D (2010) Robust multivariate location and dispersion. Preprint, see www.math.siu.edu/olive/preprints.htm.
9. Uraibi H, Midi H, Rana S (2016) Selective overview of forward selection in terms of robust correlations. *Comm Stat Simul Comput* (in press).
10. Lin D, Foster D, Ungar L (2011) VIF regression: A fast regression algorithm for large data. *J Am Stat Assoc* **106**, 232–47.
11. Devlin S, Gnanadesikan R, Kettenring J (1981) Robust estimation of dispersion matrices and principal components. *J Am Stat Assoc* **76**, 354–62.
12. Olive D (2004) A resistant estimator of multivariate location and dispersion. *Comput Stat Data Anal* **46**, 99–102.
13. Lopuhaa HP (1999) Asymptotics of reweighted estimators of multivariate location and scatter. *Ann Stat* **27**, 1638–65.
14. Meinshausen N, Meier L, Buhlmann P (2009) *P*-values for high-dimensional regression. *J Am Stat Assoc* **104**, 1671–81.