ISSN 1022 - 2502, E-ISSN 2408- 8528
Dhaka Univ. J. Sci. 71(2): 134-141, 2023 (July) DOI: https://doi.org/10.3329/dujs.v71i2.69122

Robust Multiple Linear Backward Elimination Regression

Md Siddiqur Rahman[1]*and Sabina Sharmin[2]

[1]*Department of Statistics, Jagannath University, Dhaka-1100, Bangladesh*
[2]*Department of Statistics, University of Dhaka, Dhaka-1000, Bangladesh*

(Received : 23 January 2023; Accepted : 20 July 2023)

Abstract

For building a linear prediction model, robust Backward Elimination (RBE) algorithm, which is computationally useful and scalable to high-dimensional large datasets, is introduced in this investigation. Backward Elimination (BE) can be stated in terms of sample correlations and simple RBE can be obtained by swapping out these correlations with their corresponding robust counterparts. The robust correlation for winsorized data was employed based on the adjusted winsorized correlation as a robust bivariate correlation. In another study, the Spearman rank correlation was employed as a robust bivariate correlation. However, the RBE has some drawbacks in the presence of multivariate outliers. In this article, the usage of FastMCD (Fast minimum covariance determinant)-based correlation is proposed in BE to reduce the influence of outlying data points. We call this proposed method BEmcd. A comprehensive simulation study was conducted to evaluate the performance of BEmcd with that of RBE based on winsorized correlation and Spearman rank correlation. Simulations and an application of actual data demonstrate the outstanding performance of BEmcd.

Keywords: Computational complexity, Multivariate outliers, Robust model selection, Slash contamination, Winsorization

I. Introduction

For a limited set of candidate predictors, we can select a robust linear prediction model by calculating robust versions of selection criteria[1,2,3], best subset regression[4,5], or first CV with MCD[6]. Morgenthaler et al.[7] developed a selection method that identifies both the proper model structure and unusual observations. One major disadvantage of majority of robust model selection techniques is that the computational burden (also known as all possible subsets regression) grows rapidly as the number of subsets increases. One exceptional model selection procedure based on the Wald test[8]requires only estimates from the full model to be computed. However, the goal is frequently to select a subset of a large number of possible predictors, and fitting the entire model may be impossible.

When there are many candidate predictors, A parsimonious set of candidate predictors must be chosen in order to effectively predict a response variable. Backward elimination (BE), one of the traditional step-by-step model-building techniques, is often used for this reason[9,10]. Our strategy is to use the BE model to sequence the candidate predictors to form a list, with the best predictors at the top. BE has been described using sample correlations and proposed a robust version of BE (RBE) that is based on two approaches to robust bivariate correlation estimates: adjusted winsorized correlation and Spearman rank correlation[11,12,13]. These two types of correlations are robust only to bivariate outliers. However, multidimensional outliers may be missed by univariate as well as bivariate studies. Furthermore, the matrix of correlation generated by the adjusted winsorized correlation technique might not be positively definite, necessitating its use in some situations. These issues prompted us to improve this selection criterion by employing a fast and robust multivariate location and dispersion method that is resistant to multivariate outliers. The fast minimum covariance determinant (FastMCD) approach is a computationally efficient and highly multivariate robust estimator of location and scatter[13]. We propose to use robust correlation obtained from the FastMCD scatter matrix for sequencing candidate predictors with BE, referred to as BEmcd.

A short list of first-ranked predictors (which is equal to or somewhat higher than the total number of predictors in the ultimate model) can be derived from the sequence, from which a final model can be derived using a robust regression estimator. The sequence is the primary focus of this paper.

The remaining sections of the article are organized as follows: The BE and RBE algorithms were discussed in section 2. Fast MCD-based correlation was presented in section 3. Section 4 presents a Monte Carlo simulation study comparing the performances of RBE based on adjusted bivariate winsorized correlation, Spearman's rank correlation, and Fast MCD-based correlation. There is a real-data application in Section 5. Section 6 brings us to a close.

II. BE Algorithm Expressed in Correlations

The benefit of BE is the ability to generate the sequence of covariates from the matrix of correlations in the data. Let Y be the n-dimensional standardized response variable and $X_1, X_2, \cdots, X_p$ be n-dimensional standardized predictor

*Author for correspondence. e-mail: sabina.sharmin@du.ac.bd

variables with zero mean and unit variance. The BE procedure begins with the entire model and pulls out one covariate at each step and then replaces the covariate at the end of the sequence. Let r_{jY} denote the correlation between X_j and Y. R_X is the predictors' correlation matrix. Without losing generality, we assume that the absolute partial correlation of X_1 with Y is the lowest after removing the linear influence of $X_2, X_3, \cdots, X_p$ on X_1. The first predictor X_1, known as the inactive predictor, is then removed from the model and placed at the end of the sequence. To identify the inactive predictor (let us say, X_1), we require the partial correlation of X_1 and Y after removing the linear influence of $X_2, X_3, \cdots, X_p$ on X_1, denoted by $r_{1y.23\cdots p}$.

BE steps in correlations

We summarize the BE algorithm based on the correlations among the initial variables as follows[11]:

1. Let A represent the subset of all predictors and Srepresent the subset that excludes the jth predictor. To eliminate the first predictor, say X_{m_1}, compute the partial correlation $r_{jY.S}$ of X_j and Y after removing the linear influence of the predictor that belongs to S on X_j. Determine $m_1 = \text{argmin}|r_{jY.S}|$.

2. Let C represent a subset carrying $(l-1)$ predictors took away from A after $(l-1)$ steps $(l = 2,3,\cdots)$, and S represent the subset excluding the jth predictor and C. To remove the kth predictor, say X_{m_l}, calculate the partial correlation $r_{jY.S}$ between X_j and Y after removing the linear effect of $X_{m_1}, X_{m_2}, \cdots, X_{m_{(l-1)}}$ on X_j, and then determine $m_l = \text{argmin}|r_{jY.S}|$.

The 'weakest' predictor (among the rest of the predictors) is identified at every BE step and placed to the left of the predictors in S. We have a sequence of all predictors that are now in S at the $(p-1)$ steps.

Robustification of BE algorithm

BE algorithm has been described with regard to sample means, standard deviations, and correlations[11]. It is well known that the sample mean and sample standard deviation are affected by outliers or other contaminations. In this situation, Pearson's correlation r becomes non-robust when outliers may occur in either x_i or y_i or in both (x_i, y_i). Consequently, the presence of outliers and other contaminations in data destroys the estimate r of good data and may change its sign[14,15]. On the solution considered this issue, the robustness literature shows a variety of approaches to robust correlations obtained from the robust covariance matrix. Thus, the classical building blocks (mean, standard deviation and Pearson's correlation) of BE algorithm are replaced by their corresponding robust counterparts. Initial standardization has two simple options for quickly computed robust center and scale measures: median (med) and median absolute deviation (mad).

III. FastMCD-based Correlation

The minimum covariance determinant (MCD) estimator is a very much robust and affine equivariant estimator of multivariate location and scatter[16,17]. In multivariate location and scatter setting, let Y be the n-dimensional response variable in multivariate location and scatter settings, and $\boldsymbol{X} = (X_1, X_2, \cdots, X_p)$be p predictor variables each of size n. Suppose that the observations in X are drawn from a sample of a unimodal distribution that is elliptically symmetric with an unknown mean vector $\boldsymbol{\mu}$ and a positive definite covariance matrix $\boldsymbol{\Sigma}$. The traditional tolerance ellipse is then described as the collection of d-dimensional points $\boldsymbol{x}$ with Mahalanobis distance

$$MD(x) = d(\boldsymbol{x}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}) = \sqrt{(\boldsymbol{x} - \hat{\boldsymbol{\mu}})'(\hat{\boldsymbol{\Sigma}})^{-1}(\boldsymbol{x} - \hat{\boldsymbol{\mu}})} \quad (1)$$

equals $\sqrt{\chi^2_{(p,0.975)}}$. $\hat{\boldsymbol{\mu}}$ is the sample mean vector, and $\hat{\boldsymbol{\Sigma}}$ is the sample covariance matrix. The robust tolerance ellipse according to robust distances is

$$RD(x) = d(\boldsymbol{x}, \hat{\boldsymbol{\mu}}_{MCD}, \hat{\boldsymbol{\Sigma}}_{MCD}), \quad (2)$$

where $\hat{\boldsymbol{\mu}}_{MCD}$ is the MCD estimator of location and $\hat{\boldsymbol{\Sigma}}_{MCD}$ is the MCD covariance estimator.

The raw MCD estimator with a tuning constant of$n/2 \leq h \leq n$, $(h > p$ and $n > 2p)$ is $(\hat{\boldsymbol{\mu}}_0, \hat{\boldsymbol{\Sigma}}_0)$, where $\hat{\boldsymbol{\mu}}_0$ is the mean of h observations that have the smallest sample covariance matrix, and $\hat{\boldsymbol{\Sigma}}_0$ is the matching covariance matrix multiplied by a constant factor c_0. To get consistency in the normal distribution, c_0 equals $\alpha/F_{\chi^2_{p+2}}(q_\alpha)$ with $\alpha = \lim_{n\to\infty} h(n)/n$, and q_α the α quantile of the χ^2_p distribution[18]. The MCD estimator is the most robust when $h = \lfloor (n+p+1)/2 \rfloor$, where $\lfloor a \rfloor$ is the greatest integer less than or equal to a.

To boost efficiency while retaining high robustness, a weighting step can be applied[19,20]. For the MCD, this yields

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{MCD}} = \frac{\sum_{i=1}^{n} \boldsymbol{w}(\boldsymbol{d_i^2})\boldsymbol{x_i}}{\sum_{i=1}^{n} \boldsymbol{w}(\boldsymbol{d_i^2})},$$

$$\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{MCD}} = \boldsymbol{c_1}\frac{\mathbf{1}}{\boldsymbol{n}}\sum_{\boldsymbol{i=1}}^{\boldsymbol{n}} \boldsymbol{w}\left(\boldsymbol{d_i^2}\right)(\boldsymbol{x_i} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{MCD}})(\boldsymbol{x_i} - \widehat{\boldsymbol{\mu}}_{\boldsymbol{MCD}})' \quad (3)$$

with $\boldsymbol{d_i} = \boldsymbol{d}(\boldsymbol{x}, \widehat{\boldsymbol{\mu}}_{\mathbf{0}}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{0}})$, an appropriate weight function $\boldsymbol{w}$ and again a consistency factor $\boldsymbol{c_1}$. Effective choice of $\boldsymbol{w}$ is such that $\boldsymbol{w}(\boldsymbol{d^2}) = \boldsymbol{I}\left(\boldsymbol{d^2} \leq \boldsymbol{\chi^2_{(p,0.975)}}\right)$. Huber[21]shows that if $\boldsymbol{\alpha} = \mathbf{0.5}$, the weighted step enhances the efficiency almost 45.5% for $\boldsymbol{p} = \mathbf{2}$ as well as 82% for $\boldsymbol{p} = \mathbf{10}$. The efficiency can be raised by taking a greater$\boldsymbol{\alpha}$ as $\boldsymbol{\alpha} = \mathbf{0.75}$.

The robust correlation between $\boldsymbol{X_i}$, and$\boldsymbol{X_j}$ as from the MCD scatter matrix as

$$\boldsymbol{r_{ij}} = \frac{\boldsymbol{s_{ij}}}{\sqrt{\boldsymbol{s_{ii}}\boldsymbol{s_{jj}}}}.$$

with $\boldsymbol{s_{ij}}$ is the $(\boldsymbol{i},\boldsymbol{j})^{\text{th}}$ element of the MCD scatter matrix.

The accurate MCD estimator is difficult to compute due to the need to evaluate every$\binom{\boldsymbol{n}}{\boldsymbol{h}}$ subset of size $\boldsymbol{h}$. Rousseeuw and Driessen[17]developed a quite efficient FastMCD algorithm. The algorithm's key element is the C(concentration)-step:

Take $\boldsymbol{X} = \{\boldsymbol{x_1}, \boldsymbol{x_2}, \cdots, \boldsymbol{x_n}\}$ *be a sample of size* $\boldsymbol{n}$ *and* $\boldsymbol{H_1} \subset \{\mathbf{1}, \mathbf{2}, \cdots, \boldsymbol{n}\}$ *be a subset of size* $\boldsymbol{h}$. *For data in* $\boldsymbol{H_1}$, *calculate empirical mean and covariance matrix* $\widehat{\boldsymbol{\mu}}_{\mathbf{1}}$ *and* $\widehat{\boldsymbol{\Sigma}}_{\mathbf{1}}$. *If* $\left|\widehat{\boldsymbol{\Sigma}}_{\mathbf{1}}\right| > \mathbf{0}$, *define the relative distances* $\boldsymbol{d_1}(\boldsymbol{i}) \coloneqq \boldsymbol{d}\left(\boldsymbol{x_i}, \widehat{\boldsymbol{\mu}}_{\mathbf{1}}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{1}}\right)$ *for* $\boldsymbol{i} = \mathbf{1}, \mathbf{2}, \cdots, \boldsymbol{n}$. *Now consider* $\boldsymbol{H_2}$ *such that* $\{\boldsymbol{d_1}(\boldsymbol{i}); \boldsymbol{i} \in \boldsymbol{H_2}\} \coloneqq \{\boldsymbol{d_1}(\boldsymbol{i})_{\mathbf{1:}\boldsymbol{n}}, \boldsymbol{d_1}(\boldsymbol{i})_{\mathbf{2:}\boldsymbol{n}}, \cdots, \boldsymbol{d_1}(\boldsymbol{i})_{\boldsymbol{h:n}}\}$ *where* $\boldsymbol{d_1}(\boldsymbol{i})_{\mathbf{1:}\boldsymbol{n}} \leq \boldsymbol{d_1}(\boldsymbol{i})_{\mathbf{2:}\boldsymbol{n}} \leq \cdots \leq \boldsymbol{d_1}(\boldsymbol{i})_{\boldsymbol{h:n}}$. *Compute* $\widehat{\boldsymbol{\mu}}_{\mathbf{2}}$ *and* $\widehat{\boldsymbol{\Sigma}}_{\mathbf{2}}$ *from data set* $\boldsymbol{H_2}$. *Then* $\left|\widehat{\boldsymbol{\Sigma}}_{\mathbf{2}}\right| \leq \left|\widehat{\boldsymbol{\Sigma}}_{\mathbf{1}}\right|$ *and equality holds iff* $\widehat{\boldsymbol{\mu}}_{\mathbf{2}} = \widehat{\boldsymbol{\mu}}_{\mathbf{1}}$ *and* $\widehat{\boldsymbol{\Sigma}}_{\mathbf{2}} = \widehat{\boldsymbol{\Sigma}}_{\mathbf{1}}$.

If $\left|\widehat{\boldsymbol{\Sigma}}_{\mathbf{1}}\right| > \mathbf{0}$, the C-steps give a new subset of size $\boldsymbol{h}$with covariance matrix $\widehat{\boldsymbol{\Sigma}}_{\mathbf{2}}$ such that $\left|\widehat{\boldsymbol{\Sigma}}_{\mathbf{2}}\right| \leq \left|\widehat{\boldsymbol{\Sigma}}_{\mathbf{1}}\right|$.

Iterate C-steps until $\left|\widehat{\boldsymbol{\Sigma}}_{\mathbf{new}}\right| = \left|\widehat{\boldsymbol{\Sigma}}_{\mathbf{old}}\right|$. The sequence of determinants derived in this manner has to converge in a limited number of steps, and it does so quickly in practice.The global minimum of the MCD objective function is not guaranteed to be the final value $\left|\widehat{\boldsymbol{\Sigma}}_{\mathbf{new}}\right|$ of the iteration process. Therefore, an approximation of the MCD solution can be derived by considering a large number of initial options for $\boldsymbol{H_1}$ and applying C-steps to each, retaining the solution with the smallest determinant.

In order to build an initial subset $\boldsymbol{H_1}$, we first generate a random subset $\boldsymbol{S}$ of size $(\boldsymbol{p}+\mathbf{1})$ and compute $\widehat{\boldsymbol{\mu}}_{\mathbf{0}}$and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{0}}$, where $\widehat{\boldsymbol{\mu}}_{\mathbf{0}}$is the empirical mean and $\widehat{\boldsymbol{\Sigma}}_{\mathbf{0}}$ is the covariance matrix. (If $\left|\widehat{\boldsymbol{\Sigma}}_{\mathbf{0}}\right| = \mathbf{0}$, $\boldsymbol{S}$ grows by adding observations until $\left|\widehat{\boldsymbol{\Sigma}}_{\mathbf{0}}\right| > \mathbf{0}$.) Compute the distances $\boldsymbol{d_0^2}(\boldsymbol{i}) \coloneqq \boldsymbol{d^2}\left(\boldsymbol{x_i}, \widehat{\boldsymbol{\mu}}_{\mathbf{0}}, \widehat{\boldsymbol{\Sigma}}_{\mathbf{0}}\right)$ for $\boldsymbol{i} = \mathbf{1}, \mathbf{2}\cdots, \boldsymbol{n}$. The $\boldsymbol{h}$ observations that have the shortest distance $\boldsymbol{d_0}$ make up the initial subset $\boldsymbol{H_1}$.This strategy produces better initial subsets than simply taking random subsets of size $\boldsymbol{h}$.

Each C-step includes the computation of a covariance matrix, its relevant determinant, and the accompanying distances. Using fewer C-steps significantly improves the algorithm's speed. For a small n, this process is very quick but as n grows, the amount of computing time grows because each C-step requires calculating n more distances. FastMCD splits the data set for large n, avoiding any calculations on the whole set[21]. It should be noted that the FastMCD method it self is affine equivariant. For $\alpha = 0.75$,the breakdown point of MCD-based correlation is 25%. Bernholdand Fischer[22] show that the computational complexity of MCD correlation is $\mathcal{O}\left(n^{1+p(p+3)/2}\right)$.

IV. A Simulation Study

To evaluate the effectiveness of BEmcd, a simulation study equivalent to the one conducted by Frank and Friedman is carried out[23]. Out of a total of $\boldsymbol{a} = \mathbf{9}$ target predictors (predictors with non-zero coefficients), $\boldsymbol{p} = \mathbf{50}$ candidate predictors are considered. Three correlation structures exist among the target predictors: nocorrelation, moderate correlation, and high correlation.

For the no-correlation case, independent predictors $X_j \sim N(0,1)$ are considered, and the a target predictors having coefficients (7, 6, 5) cycled three times are used to generate the response variable Y. The standard deviation of the error term is set to have a signal to noise ratio of 2.

For the correlation scenario, three independent latent variables $L_i \sim N(0,1), (i = 1,2,3)$ are introduced,which are accountable for the systematic variation in both of the response and the target predictors. The model is

$$Y = 7L_1 + 6L_2 + 5L_3 + \varepsilon = signal + \sigma\varepsilon,$$

where $\varepsilon \sim N(0,1)$, and $\sigma = \sqrt{110}/2$. For $a = 9$, a set of 50 predictors is created as follows:

$X_{(k-1)3+i} = L_i + \delta e_{(k-1)3+i}$; $\quad i = 1,2,3, \quad k = 1,2,3$ and $X_j = u_j, j = 10, \cdots, 50$.

where all $\boldsymbol{e_{(k-1)3+i}}$ and $\boldsymbol{u_j}$ are independent standard normal variates. For $\boldsymbol{\delta = 1}$, the real correlation between the predictors produced by the same latent variable is 0.5 (moderate correlation case). For $\boldsymbol{\delta = 0.5}$, the real correlation between the predictors produced by the same latent variable is 0.8 (high correlation case).

To permit for a fraction Δ of outliers, we generate using

$\boldsymbol{\varepsilon \sim (1-\Delta)N(0,1) + \Delta G}$, where $\boldsymbol{G}$ is another distribution rather than $\boldsymbol{N(0,1)}$.

We consider Δ= 0.05,0.10, 0.15, and 0.20.

We consider the following scenarios:

1. For the moderate correlation case, all noise predictors are contaminated with $\boldsymbol{N}(50,1)$ and corresponding response values with $\boldsymbol{N}(500,1)$.
2. For high correlation cases, outliers are given by symmetric slash contamination: $\boldsymbol{G \sim N(5,1)/uniform(0,1)}$ and outliers are given to all noise predictors, and corresponding response values as $\boldsymbol{G \sim N(50,1)/uniform(0,1)}$.
3. For a centered multivariate normal distribution with covariance structure $\boldsymbol{COV(X_j, X_k) = \rho^{|j-k|}}$where $\boldsymbol{\rho = 0.5}$ is considered. Outliers are given like (2).

We generated 1000 independent data sets, each having a size of 500. For each of the simulated data sets, the covariates were sequenced using BEmcd, BEr, and RBE.

To provide a summary of the simulated findings for each of the sequences, the number t_m of targeted predictors contained in the first m sequenced variables was determined, with m ranging from 1 to 25.

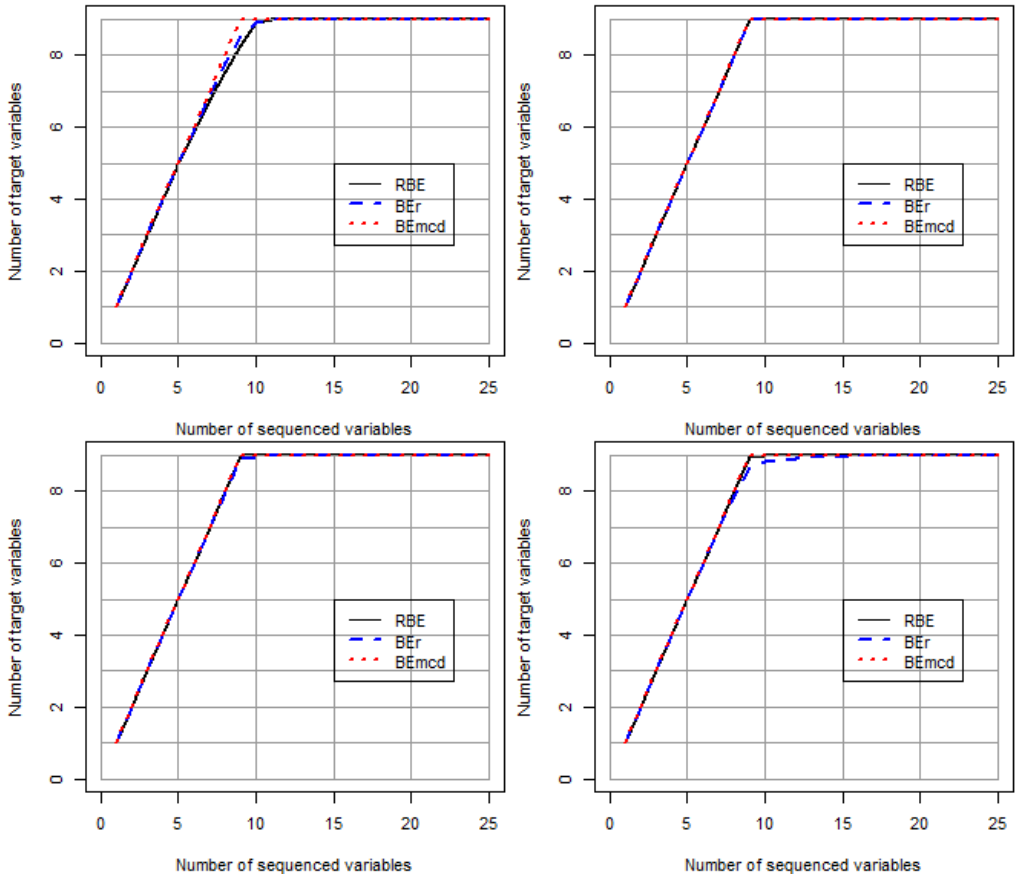

Fig. 1. Recall curves for no correlation case: 5% outliers (upper left), 10% outliers (upper right), 15% outliers (lower left), and 20% outliers (lower right)

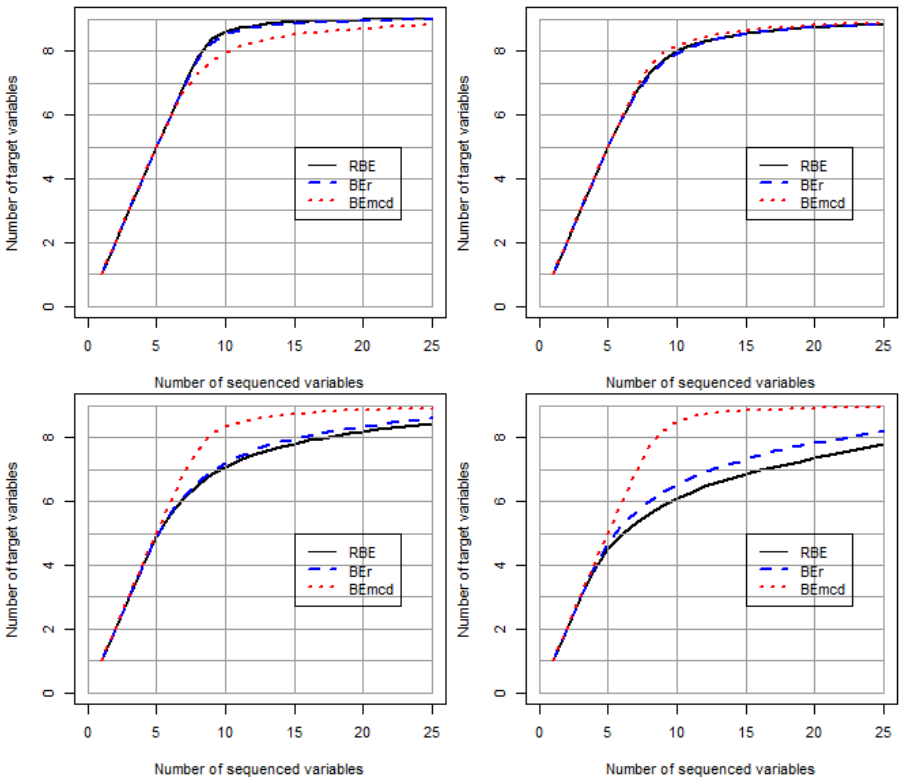

Fig. 2. Recall curves for moderate correlation case: 5% outliers (upper left),10% outliers (upper right), 15% outliers (lower left), and 20% outliers (lower right)

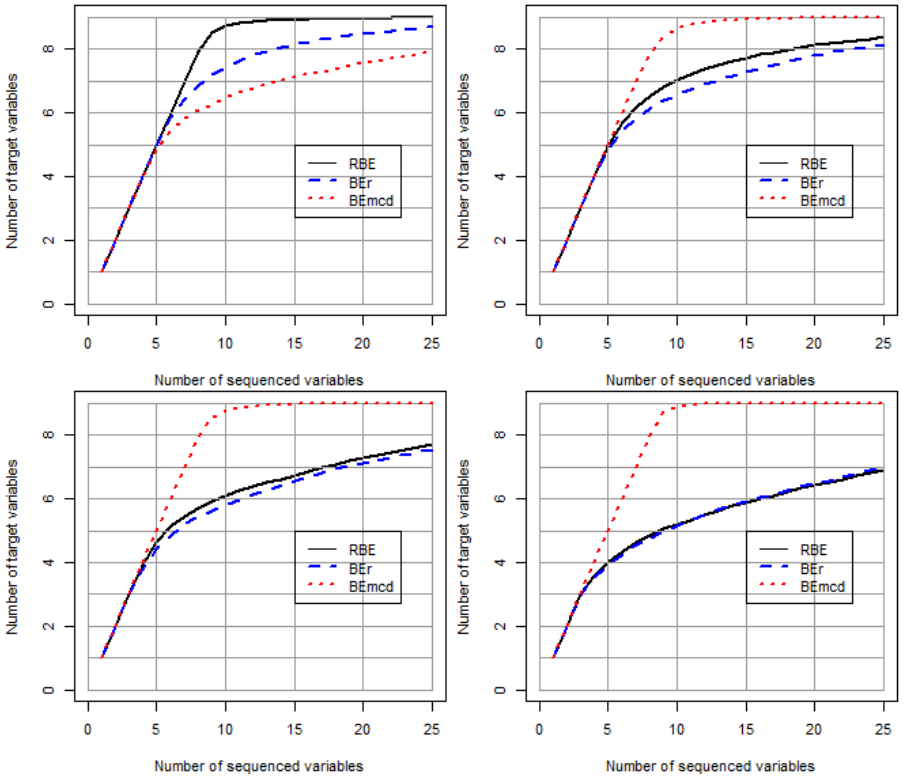

Fig. 3. Recall curves for a high correlation case with symmetric slash contamination: 5% outliers (upper left), 10% outliers (upper right), 15% outliers (lower left), and 20% outliers (lower right).

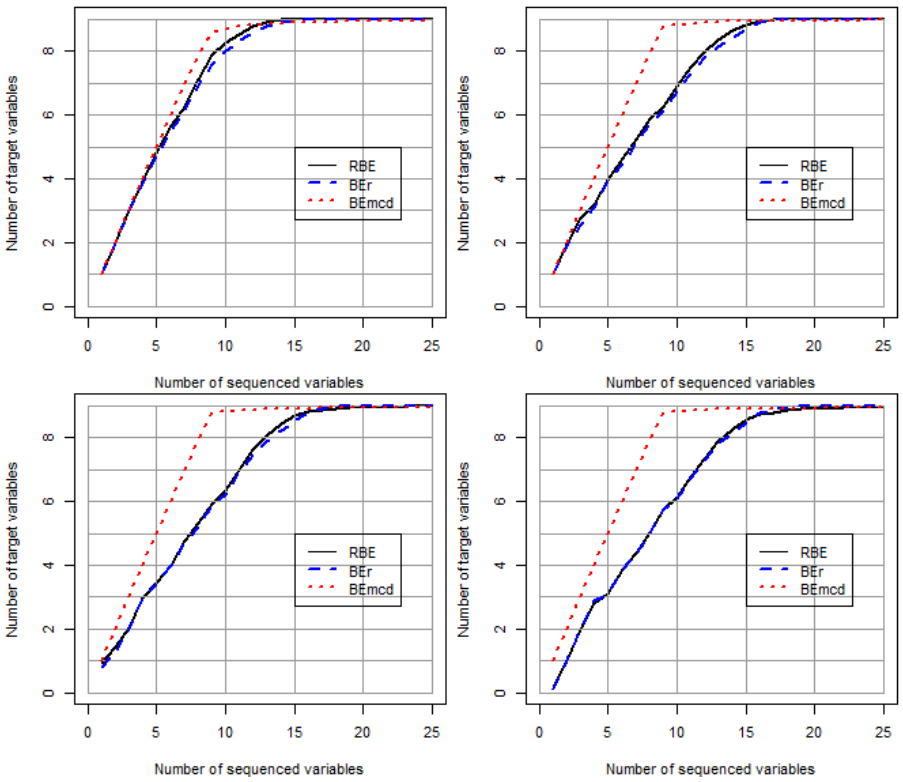

Fig. 4. Recall curves for multivariate outliers with symmetric slash contamination: 5% outliers (upper left), 10% outliers (upper right), 15% outliers (lower left), and 20% outliers (lower right).

All pieces of Fig. 1 show that the three procedures perform equally well for sequencing the important predictor variables for various contamination cases.

All pieces of Fig. 2 show the recall curves for predictors with moderate correlation and contamination cases. The recall curve obtained in the upper left piece of Fig. 2 shows that BEmcd performs slightly worse than RBE and BEr with 5% contamination, whereas BEmcd performs better than RBE and BEr with larger contamination proportions (upper right, lower left, and lower right pieces of Fig. 2). Even with 20% contamination, BEmcd with $m = 10, t_m$=8.5$(a = 9)$ already yields a recall proportion of around 94.4%.

All pieces of Fig. 3 show the recall curves for predictors with high correlation and symmetric slash contamination cases. The recall curve obtained in the upper left piece of Fig. 3 shows that BEmcd performs slightly worse than RBE and BEr with 5% contamination, whereas BEmcd performs much better than RBE and BEr with higher contamination proportions. Even with 20% contamination, BEmcd with $m = 9, t_m$=8.74$(a = 9)$ yields a recall proportion of around 97%. BEmcd with $m = 10$ already produces a recall proportion of around 99%.

All pieces of Fig. 4 show the recall curves for predictors with multivariate outliers and a symmetric slash distribution. All recall curves show that the BEmcd is consistently able to select the correct predictors. BEmcd with $m = 9, t_m$=8.8 $(a = 9)$ already produces a recall percentage of around 98%.

V. Determination of Reduced Ret and Final Prediction Model

To get a stable sequence, the BEmcd algorithm is repeated 100 times, and then the predictors are sequenced by their ranking orders. After sequencing all the predictor variables using the RBE, BEmcd, and BEr algorithms, the first m top ranking predictors form a reduced set for each sequence. When a reduced set is obtained, we can go to the segmentation step to obtain the final prediction model. The condensed set should be large enough to contain the majority of the important predictors while remaining manageable enough to prevent the segmentation step from being rendered ineffective. In practice, the number of predictors required in the model is frequently unknown. Thus, we employ a graphical tool called the learning curve to get the length of the condensed set. We start with the first predictor in the sequence and grow the number of predictors throughout the sequence, fitting a robust regression model each time to compute a robust R^2 measure like $R^2 = 1 - med(e^2)/mad^2(y)$, where e is the vector of

residuals from the robust fit[24]. We obtain a learning curve by plotting these R^2 values against the number of predictors[25]. The length of the condensed set, m, is selected at the point at which the learning curve no longer has a significant slope.

One reasonable approach is to execute all conceivable subsets regression of this "condensed set" using proper selection criteria (e.g, RAIC, RC_p, RFPE, RCV, all possible subset regression and robust bootstrap). We employall possible subset regression to fitall possible models and display some of the best candidates based on adjusted R-squared or the robust version of C_p, Mallows' ${RC_p}^2$. RC_p is defined as $RC_p = \frac{W_p}{\hat{\sigma}^2} - (U_p - V_p)$, where $W_p = \sum \hat{w}_i^2 r_i^2$ is the weighted sum of squares of residuals, $\hat{\sigma}^2$ is a robust consistent estimator of σ^2 from the entire model, and U_p and V_p are constants that depend on p and the weight function, w.

VI. Application to Real Data

A true data set was used in this section to compare the performance of BEmcd to that of RBE and BEr. As an example of real data, we use the information on wave energy converters (WECs) in two real wave scenarios from Australia's southern coast (Adelaide and Tasmania) stored in the UCI Machine Learning Repository (https://archive.ics.uci.edu/ml/datasets.php). This data set contains 49 variables. We excluded the last non-dimensional variable from the data set. We consider the first variable (V1) as the response. The rest of the 47 variables are the predictors which are numbered from 1 to 47. The data set obtained from Adelaide was used as training data and that from Tasmania as test data. Each data set contains missing observations. After eliminating the missing rows, training and test data sets contain 24251 observations and 18454 observations, respectively. The reduced sets obtained from the sequences of RBE, BEmcd and BEr are (32 16 30 46 47 31 21 45 33 13 415 38 24), (32 16 47 31 45 14 38 33 41) and (32 16 30 46 21 47 31 45 41 33 5 13 38 24), respectively. The final models obtained using all possible subset regression over the above shortlists are (32 16 30 46 47 31 21 45 33 13 38), (32 16 47 31 45 14 38 41 33) and (32 16 30 46 21 47 31 45 33 13 38), respectively. The corresponding Mallows' C_ps are 12.00, 10.00, and 10.00, respectively. The final model obtained by RBE and BEr are the same and each contains 11 predictors, whereas BEmcd contains only 9 predictors.

Using robust five-fold cross-validation (CV) with 100 replications, the 5% trimmed MTMSPEs are obtained as 128.93, 127.99, and 128.93, respectively for the final models obtained from RBE, BEmcd, and BEr. The predicted outcomes for several methods are shown in Fig.5.

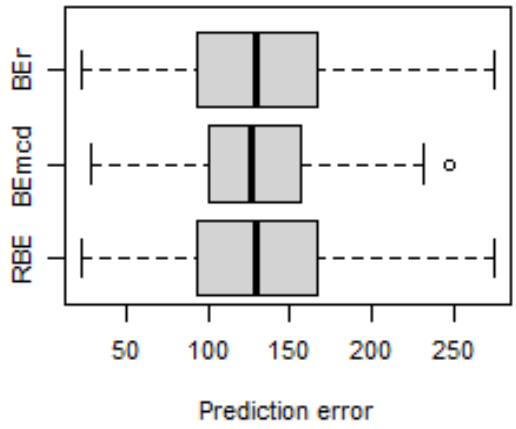

Fig. 5. Prediction error of root-trimmed mean squared with 5% trimming estimated through repeated five-fold robust CV with 100 repetitions.

Table 1. WECs data set: root trimmed mean squared prediction error (RTMSPE) with 1%, 5% and 10% trimming estimates, root median squared prediction error (RMSPE), standard deviation (SD), interquartile range (IQR) and normalized median absolute deviation about median (MAD).

Algorithm	RTMSPE			RMSPE	SD	IQR	MAD
	1%	5%	10%				
RBE	303.80	282.85	265.21	215.45	301.36	416.94	311.01
BEmcd	288.23	269.00	252.74	206.97	296.91	411.21	306.80
BEr	303.80	282.85	265.21	215.45	301.36	416.94	311.01

In Table 1, we assess the predictive power of different final models fitted by MM regression method.

For the given data set, we observe that the final models obtained via the algorithms RBE and BEr perform equally, whereas the final model obtained via BEmcd results in better prediction performance compared to the others. From Fig.5, it is also clear that BEmcd outperforms the other methods.

VII. Conclusions

This paper considers the issue that occurs when choosing a linear prediction model for sizable, high-dimensional data sets

that could be clean or could have a fraction of contamination. RBE and BEr are two popular robust algorithms. In this study, we propose a robust BE algorithm based on Fast MCD based correlations (BEmcd). The outstanding performance of the proposed algorithm is demonstrated by a simulation study and a real-data application comparing the performances of RBE and BEr. For data sets with multivariate outliers, the BEmcd algorithm outperforms at different contamination levels, while RBE and BEr decline gradually.

References

1. Ronchetti, E., 1985. Robust Model Selection in Regression. *Statistics and Probability Letters*, **3,** 21-23.

2. Ronchetti, E., and R. G. Staudte, 1994. A Robust Version of Mallow's Cp. *Journal of the American Statistical Association*, **89,** 550-559.

3. Yohai, V. J., 1997. A New Robust Model Selection Criterion for Linear Models: RFPE, *unpublished manuscript*.

4. Furnival, G. M., and R. W. Wilson, 1974. Regressions by leaps and bounds. *Technometrics*, **16,** 499-511.

5. Ronchetti, E., C. Field, and W. Blanchard, 1997. Robust Linear Model Selection by Cross-Validation. *Journal of the American Statistical Association*, **92,** 1017-1023.

6. Hubert, M., and S. Engelen, 2004. *Robust PCA and Classification in Biosciences. Bioinformatics.* 1728-1736.

7. Morgenthaler, S., R. E. Welsch, and A. Zenide, 2003. Algorithms for Robust Model Selection in Linear Regression in *Theory and Applications of Recent Robust Methods*, *eds*. Conference paper. pp. 196-206

8. Sommer, S., and R. M. Huggins, 1996. Variable Selection Using the Wald Test and Robust Cp, *Journal of the Royal Statistical Society*, Ser. B, **45,** 15-29.

9. Gatu, C., and E. J. Kontoghiorghes, 2006. Branch-and-bound algorithms for computing the best subset regression models. *Journal of Computational and Graphical Statistics*, **15,** 139-156.

10. Weisberg, S., 2014. Applied Linear Regression (3rd ed.). *New York: Wiley-Interscience*. 221-232.

11. Rahman, M. S., and J. A. Khan, 2014. Building a Robust Linear Model with Backward Elimination Procedure. *The Dhaka University Journal of Science*, **62,** 87-93.

12. Rahman, M. S., 2015. Backward Elimination Procedure for Linear Model Building Using Spearman's Rank Correlation. *Jahangirnagar University Journal of Science*, **38(2),** 11-22.

13. Spearman, C., 1904. General intelligence objectively determined and measured. *Am J Psychol*, **15,** 201–293.

14. Devlin, S. J., R. Gnanadesikan, and J. R. Kettenring, 1981. Robust Estimation of Dispersion Matrices and Principal Components. *Journal of the American Statistical Association*, **76,** 354-362.

15. Gnanadesikan, R., and J. R. Kettenring, 1972. Robust estimates, residuals and outlier detection with multiresponse data. *Biometrics*, **28,** 81-124

16. Rousseeuw, P. J., 1985. *Multivariate Estimation with High Breakdown Point. Vol. B*, eds. W. Grossmann, G. Pflug, I Vincze, and W. Wertz Dordrechet: Reidel, pp 283-297.

17. Rousseeuw, P. J., and K. V. Driessen, 1999. A First Algorithm for the Minimum Covariance Determinant Estimator. *Technometrics*, **41,** 212-223.

18. Corux, C., and G. Haesbroeck, 1999. Influence Function and Efficiency of the Minimum Covariance Determinant Scatter Matrix Estimator. *Journal of Multivariate Analysis*, **71,** 161-190.

19. Lopuhaä, H., 1999. Asymptotics of Reweighted Estimators of Multivariate Location and Scatter. *The Annals of Statistics*, **27,** 1638-1665.

20. Lopuhaä, H., and P. J. Rousseeuw, 1991. Breakdown Points of Affine Equivariant Estimators of Multivariate Location and Covariance Matrices. *The Annals of Statistics*, **19,** 229-248.

21. Hubert, M., M. Debruyne, and P. J. Rousseeuw, 2017. Minimum Covariance Determinant and Extensions. *Journal of the American Statistical Association*, **88,** 1273–1283.

22. Bernhold, T, and P. Fischer, 2004. The Complexity of Computing the MCD-Estimator. *Theoretical Computer Science*, **326,** 383-398.

23. Frank, I., and J. H. Friedman, 1993. A Statistical View of Some Chemometrics Regression Tools. *Technometrics*, **35,** 109-148.

24. Rousseeuw, P. J., 1984. Least Median of Squares Regression. *Journal of the American Statistical Association*, **79,** 871-880.

25. Corux, C., P. Filzmoser, and P. J. Rousseeuw 2003. Fitting Multiplicative Models by Robust Alternating Regressions. *Statistics and Computing*, **13,** 23-36.