Dhaka Univ. J. Sci. 62(2): 87-93, 2014 ((July)

Building a Robust Linear Model with Backward Elimination Procedure

Md Siddiqur Rahman[1] and Jafar A. Khan[2]

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

[2]*Department of Statistics, Biostatistics and Informatics, Dhaka University, Dhaka-1000, Bangladesh*

(Received: 30 January 2013; Accepted: 20 March 2014)

Abstract

For building a linear prediction model, Backward Elimination (BE) is a computationally suitable stepwise procedure for sequencing the candidate predictors. This method yields poor results when data contain outliers and other contaminations. Robust model selection procedures, on the other hand, are not computationally efficient or scalable to large dimensions, because they require the fitting of a large number of submodels. Robust version of BE is proposed in this study, which is computationally very suitable and scalable to large high-dimensional data sets. Since BE can be expressed in terms of sample correlations, simple robustifications are obtained by replacing these correlations by their robust counterparts. A pairwise approach is used to construct the robust correlation matrix – not only because of its computational advantages over the d-dimensional approach, but also because the pairwise approach is more consistent with the idea of step-by-step algorithms. The performance of the proposed robust method is much better than standard BE.

Key words: Computational complexity, Pairwise robust correlation, Robust model selection, Stepwise procedure, Winsorization.

I. Introduction

When the number d of candidate covariates is small, one can choose a linear prediction model by computing a reasonable criterion (e.g., Mallows C_P, AIC, FPE or cross-validation error) for all possible subsets of the predictors. However, as d increases, the computational burden of this approach increases very quickly. This is one of the main reasons why step-by-step model-building algorithms like Backward Elimination (BE) or Stepwise (SW) are popular[1,2,3].

Classical BE procedure yields poor results when the data are contaminated. This algorithm attempts to select the covariates that will fit well all the cases (including the outliers), and often fails to select the model that would have been chosen if those outliers were not present in the data. Moreover, aggressive deletion of outliers is not desirable, because we may end up deleting a lot of observations which are outliers only with respect to predictors that will not be in the model. Our goal is to develop a robust-step-by-step algorithm that will select important variables in the presence of outliers, and predict well the future non-outlying cases.

Available literature on robust model selection focuses mainly on robustification of selection criteria to compare all possible subsets of covariates. Important examples are Ronchetti[4], Ronchetti *et al*.[5], Maronna *et al*.[6], and Ronchetti *et al.*[7] which introduced robust versions of AIC, C_P, FPE and cross-validation, respectively. Sommer *et al*.[8] proposed robust model selection based on Wald tests. Morgenthaler *et al*.[9] constructed a selection technique to simultaneously identify the correct model structure as well as any unusual observations. A major drawback of most robust model selection methods is that they are very time-consuming, because they require the robust fitting of a large number of submodels.

We show that the list of variables selected by classical BE procedure is a function of sample means, variances and correlations. We express the classical algorithm in terms of these quantities, and replace them by robust counterparts to obtain simple robust version of the algorithm. Once the covariates are selected by using this simple robust selection algorithm, we can use a robust regression estimator on the final model.

Robust correlation matrix estimators for d-dimensional data sets are usually derived from affine-equivariant, robust estimators of scatter. Hence, this is very time-consuming, particularly for large values of d. Moreover, the computation of such robust correlation matrices becomes unstable when the dimension d is large compared to the sample size n. To avoid this complexity, we use an affine-equivariant bivariate *M*-estimator of scatter proposed by Khan *et al*.[10] to obtain robust correlation estimates for all pairs of variables, and combine these to construct a robust correlation matrix which is called the pairwise correlation approach. Interestingly, this pairwise approach is computationally suitable as well as more convenient for robust step by step algorithms.

Variable selection methods are often based on correlations among variables. Therefore, robust variable selection procedures need to be robust against correlation outliers, that is, outliers that affect the classical correlation estimates but can not be detected by looking at the individual variables separately. Our approach based on pairwise correlations is robust against correlation outliers and thus suitable for robust variable selection. We consider the problem of "selecting" a list of important predictors. The final model resulting from the selection procedure usually contains only a small number of predictors compared to the initial dimension d, when d is large. Therefore, to robustly fit the final model we propose to use a highly robust regression estimator such as an MM-estimator proposed by Yohai[11] that is resistant to all types of outliers. Note that we always use models with intercept.

The rest of the article is organized as follows. In section II, we decompose the classical BE procedure in terms of the correlation matrix of the data. In section III, we present

*Author for Correspondence. *e-mail*: rsiddiq11@yahoo.com

robust version of this algorithm, along with its numerical complexities. In section IV, we present a Monte Carlo study that compares our robust method with the classical one by their predicting powers. Section V contains a real-data application. We conclude in section VI.

II. BE Algorithm Expressed in Correlations

In this section we review the classical BE procedure. For clarity of exposition, we show how this procedure can be expressed only in terms of correlations between pair of variables.

BE expressed in correlations

Let the d covariates $X_1, X_2, \cdots, X_d$ and the response Y be standardized using their mean and standard deviation. Let r_{jY} denote the correlation between X_j and Y, and R_X be the correlation matrix of the covariates. We call the predictors that are in the current regression model "active" predictors. Suppose without loss of generality X_1 has the minimum absolute partial correlation with Y after eliminating the linear effect of $X_2, X_3, \cdots, X_d$ on X_1. Then, X_1 is the first variable that is dropped from the regression model. This candidate predictor is called "inactive" predictor. Thus to find out the inactive predictor (say, X_1), we need to compute the partial correlation between X_1 and Y after eliminating the linear effect of $X_2, X_3, \cdots, X_d$ on X_1 and we denote this partial correlation by $r_{1Y.23\cdots d}$.

The partial correlations expressed in terms of original correlations

$$\text{Let } X_1 = \gamma_2 X_2 + \gamma_3 X_3 + \cdots + \gamma_d X_d + \xi \quad (1)$$

Thus, the residue vector $Z_{1.23\cdots d}$ is as follows

$$Z_{1.23\cdots d} = X_1 - \gamma_2 X_2 - \gamma_3 X_3 - \cdots - \gamma_d X_d \quad (2)$$

$$\text{Thus, } r_{1Y.23\cdots d} = \frac{Z^t_{1.23\cdots d} Y / n}{\sqrt{Z^t_{1.23\cdots d} Z_{1.23\cdots d} / n}\sqrt{V(Y)}} = \frac{Z^t_{1.23\cdots d} Y / n}{\sqrt{Z^t_{1.23\cdots d} Z_{1.23\cdots d} / n}} \quad (3)$$

Using (2), the numerator of (3) can be written as

$$\begin{aligned} & Z^t_{1.23\cdots d} Y / n \\ &= \frac{1}{n}\left(X_1 - \gamma_2 X_2 - \gamma_3 X_3 - \cdots - \gamma_d X_d\right)^t Y \\ &= r_{1Y} - \sum_{i=2}^{d} \gamma_i r_{iY} \end{aligned} \quad (4)$$

Using (2), the squared denominator of (3) can be written as

$$\begin{aligned} & \frac{1}{n} Z^t_{1.23\cdots d} Z_{1.23\cdots d} \\ &= \frac{1}{n}\left(X_1 - \gamma_2 X_2 - \gamma_3 X_3 - \cdots - \gamma_d X_d\right)^t \left(X_1 - \gamma_2 X_2 - \gamma_3 X_3 - \cdots - \gamma_d X_d\right) \\ &= 1 + \sum_{i=2}^{d} \gamma_i^2 - 2\sum_{i=2}^{d} \gamma_i r_{1i} + \sum_{\substack{i,j \\ i,j\neq 2}}^{d} \gamma_i \gamma_j r_{ij} \end{aligned} \quad (5)$$

Hence, (3) becomes in the form

$$r_{1Y.23\cdots d} = \frac{r_{1Y} - \sum_{i=2}^{d} \gamma_i r_{iY}}{\sqrt{1 - 2\sum_{i=2}^{d} \gamma_i r_{1i} + \sum_{\substack{i,j \\ i,j\neq 1}}^{d} \gamma_i \gamma_j r_{ij}}} \quad (6)$$

Now the normal equations for the model (1) are

$$\begin{aligned} \sum\left(\gamma_2 X_{2i} + \gamma_3 X_{3i} + \cdots + \gamma_d X_{di}\right) X_{2i} &= \sum X_{1i} X_{2i} \\ \sum\left(\gamma_2 X_{2i} + \gamma_3 X_{3i} + \cdots + \gamma_d X_{di}\right) X_{3i} &= \sum X_{1i} X_{3i} \\ &\vdots \\ \sum\left(\gamma_2 X_{2i} + \gamma_3 X_{3i} + \cdots + \gamma_d X_{di}\right) X_{di} &= \sum X_{1i} X_{di} \end{aligned} \quad (7)$$

From (7) we have,

$$\begin{aligned} \gamma_2 r_{22} + \gamma_3 r_{32} + \cdots + \gamma_d r_{d2} &= r_{12} \\ \gamma_2 r_{23} + \gamma_3 r_{33} + \cdots + \gamma_d r_{d3} &= r_{13} \\ &\vdots \\ \gamma_2 r_{2d} + \gamma_3 r_{3d} + \cdots + \gamma_d r_{dd} &= r_{1d} \end{aligned} \quad (8)$$

From equations (8) we have the following form:

$$\begin{pmatrix} \gamma_2 \\ \gamma_3 \\ \vdots \\ \gamma_d \end{pmatrix} = \begin{pmatrix} r_{22} & r_{23} & \cdots & r_{2d} \\ r_{32} & r_{33} & \cdots & r_{3d} \\ \vdots & \vdots & \ddots & \vdots \\ r_{d2} & r_{d3} & \cdots & r_{dd} \end{pmatrix}^{-1} \begin{pmatrix} r_{12} \\ r_{13} \\ \vdots \\ r_{1d} \end{pmatrix} \quad (9)$$

Thus, $\gamma_2, \gamma_3, \cdots, \gamma_d$ and hence $r_{1Y.23\cdots d}$ are expressed in terms of original correlations.

In general, the partial correlation between X_l and Y after eliminating the linear effect of $X_1, X_2, \cdots X_{l-1}, X_{l+1}, \cdots, X_d$ on X_l can be written as

$$r_{lY.12\cdots(l-1),(l+1)\cdots d} = \frac{r_{lY} - \sum_{\substack{i=1 \\ i\neq l}}^{d} \gamma_i r_{lY}}{\sqrt{1 - 2\sum_{\substack{i=1 \\ i\neq l}}^{d} \gamma_i r_{li} + \sum_{\substack{i,j \\ i,j\neq l}}^{d} \gamma_i \gamma_j r_{ij}}} \quad (10)$$

BE steps in correlations

BE algorithm is summarized in terms of correlations among the original variables as follows:

1. Let D be the set of all covariates and P be the subset not containing jth covariate. To remove the first covariate X_{m1}, let us calculate partial correlation $r_{jY.P}$ between X_j and Y after eliminating the linear effect of covariate belonging to P on X_j. Determine $m_1 = \arg\min\left|r_{jY.P}\right|$.

2. Let C be a subset containing $(k-1)$ variables that has been removed from D after $(k-1)$ steps $(k = 2, 3, \cdots)$ and P be the subset not containing jth covariate and C. To remove the kth covariate X_{mk}, $r_{jY.P}$ between X_j and Y may be calculated after eliminating the linear effect of $X_{m1}, X_{m2}, \cdots, X_{m(k-1)}$ on X_j, and then determine $m_k = \arg\min\left|r_{jY.P}\right|$.

Stopping rule

At each BE step, once the "weakest" covariate (among the remaining covariates) is identified, we can perform a partial F-test to decide whether to drop this covariate from the model (and continue the process) or to stop. The new "weakest" covariate is dropped from the model only if the partial F-value, denoted by F_{partial}, is smaller than $F(0.95, 1, n-k-1)$ (say), where k is the current size of the model excluding the new covariate. Here, the required quantities can be expressed in terms of correlations among the original variables, as shown below.

Suppose that 2 covariates X_1, X_2 are in the model, and X_2 has the smallest absolute partial correlation with *Y* after adjusting X_2 for X_1. To decide whether X_2 should be deleted from the model we perform a partial F-test using the statistic F_{Partial} given by

$$F_{\text{Partial}} = \frac{A^t A - B^t B}{B^t B/(n-3)}$$

where, $A = Y - \beta_{Y1} X_1$ and $B = Y - \beta_{Y1} X_1 - \beta_{Y2.1} Z_{2.1}$.

$$= \frac{(n-3)\left(2\beta_{Y2.1} Z_{2.1}^t Y/n - \beta_{Y2.1}^2 Z_{2.1}^t Z_{2.1}/n\right)}{1 - r_{1Y}^2 - 2\beta_{Y2.1} Z_{2.1}^t Y/n + \beta_{Y2.1}^2 Z_{2.1}^t Z_{2.1}/n}$$

$$= \frac{(n-3)\left(\beta_{Y2.1} Z_{2.1}^t Y/n\right)}{1 - r_{1Y}^2 - \beta_{Y2.1} Z_{2.1}^t Y/n} = \frac{(n-3) r_{2Y.1}^2}{1 - r_{1Y}^2 - r_{2Y.1}^2}$$

where $r_{2Y.1}$ is expressed in correlations in (10)

Similarly, when k covariates $X_1, X_2, \cdots, X_k$ are in the model, and w.l.o.g. X_k has the smallest absolute partial correlation with Y after adjusting X_k for $X_1, X_2, \cdots, X_{k-1}$, the partial F-statistic for X_k can be expressed as

$$F_{\text{Partial}} = \frac{(n-k) r_{kY.123\cdots(k-1)}^2}{1 - r_{1Y}^2 - r_{2Y.1}^2 - r_{3Y.12}^2 - \cdots - r_{kY.123\cdots(k-1)}^2} \quad (11)$$

III. Robustification of BE Algorithm

In the last section, the BE algorithm has been expressed in terms of sample means, variances and correlations. Because of these non-robust building blocks, this algorithm is sensitive to contamination in the data. A simple robustification of this algorithm can be achieved by replacing the non-robust ingredients of the algorithm by its robust counterparts. For the initial standardization, the choices of fast computable robust center and scale measures are straightforward: median (med) and median absolute deviation (mad). As mentioned earlier, most available robust correlation estimators are computed from the *d*-dimensional data and therefore are very time consuming[12]. Robust univariate approaches[13] are very sensitive to correlation outliers.

One solution is to derive correlations among pairs of variables from an affine equivariant covariance estimator. A computationally efficient choice is a bivariate M-estimator proposed by Maronna[14]. Maronna's bivariate M-estimator of the location vector $\boldsymbol{t}$ and scatter matrix $\boldsymbol{V}$ is defined as the solution of the system of equations:

$$\frac{1}{n}\sum_i u_1(d_i)(\boldsymbol{x_i} - \boldsymbol{t}) = \boldsymbol{0},$$

and

$$\frac{1}{n}\sum_i u_2(d_i^2)(\boldsymbol{x}_i - \boldsymbol{t})(\boldsymbol{x}_i - \boldsymbol{t})' = \boldsymbol{V},$$

where $d_i^2 = (\boldsymbol{x_i} - \boldsymbol{t})' \boldsymbol{V}^{-1} (\boldsymbol{x_i} - \boldsymbol{t})$, and u_1 and u_2 are functions satisfing a set of general assumptions. The estimator is affine equivarient and has breakdown point 1/3 in two dimensions[14]. To further simplify computations, we use the coordinatewise median as the bivariate location estimate[10] and only use the second equation to estimate the scatter matrix and hence the correlation. In this equation we used the function $u_2(t) = \min(c/t, 1)$ with $c = 9.21$, the 99% quantile of a χ_2^2 distribution. Finally, BE algorithm is implemented using these robust pairwise correlations.

Robust stopping rule

We replace the classical correlations in the partial F statistic by their robust counterparts to form a robust partial F statistic. For the stopping rule, we use the standard F distribution as in section II. Since robust pairwise correlation estimator (due to the choice of the constant c) behaves very similar to the classical correlation estimator in the absence of outliers, the standard F distribution seems appropriate.

Time-complexity of the algorithms

Since classical BE procedure sequences all the d covariates, it requires $O(nd^2)$ time. So the complexity of BE is $O(nd^2)$. Since we used the coordinatewise median as the bivariate location estimate, the correlation based on Maronna's M-estimate can be computed in $O(n\log n + bn)$ time, (b is the number of iterations required). Assuming that b does not exceed $O(n\log n)$ (convergence was achieved after 3 to 5 iterations in our simulations), the complexity of this estimate is $O(n\log n)$. So, the complexity of robust BE is $O((n\log n)d^2)$. Classical BE takes approximately 20 seconds with Dual CPU T3400, while robust BE takes approximately 25 seconds for implementation. This is a very small price to pay in order to achieve robustness. It should be mentioned that existing robust algorithms would take several days.

Limitation of the proposed algorithm

The robust BE procedure based on robust pairwise correlations is resistant to bivariate (correlation) outliers. However, it may be sensitive to three or higher-dimensional outliers, that is, outliers that are not detected by univariate and bivariate analyses. Also, the correlation matrix obtained from the pairwise correlation approach may not be positive definite, forcing the use of correction for positive definiteness in some cases[15].

IV. A Simulation Study

To compare the robust method with the classical one, we carried out a simulation study similar to Frank and Friedman[16]. The total number of variables is $d = 50$. A small number $a = 9$ of them are non-zero covariates. We considered 2 correlation structures of these non-zero covariates: "no correlation" case and "moderate correlation" case.

For the no-correlation case (a true correlation of 0 between the covariates), independent predictors $X_j \sim N(0,1)$ are considered, and Y is generated using the a non-zero covariates, with coefficients (7, 6, 5) repeated three times for $a = 9$. The variance of the error term is chosen such that the signal-to-noise ratio equals 2.

For the moderate-correlation case, we considered 3 independent 'unknown' processes, represented by latent variables $L_i, i = 1, 2, 3,$ which are responsible for the systematic variation of both the response and the covariates. The model is

$$Y = 7L_1 + 6L_2 + 5L_3 + \sigma \in = \text{Signal} + \sigma \in, \qquad (12)$$

where L_i and $\in$ are independent standard normal variables. The value of σ is chosen such that the signal-to-noise ratio equals 2, that is $Var(\sigma \in) = 110/4$. The non-zero covariates are divided in 3 equal groups, with each group related to exactly one of the latent variables by the following relation

$$X_j = L_i + \delta_j,$$

where $\delta_j \sim N(0,1)$. Thus, we have a true correlation of 0.5 between the covariates generated with the same latent variable.

For each case we generated 5000 data sets each of which was randomly divided into a training sample of size 100 and a test sample of size 100.

Contamination of the training data

Each of the $d - a$ noise variables are contaminated independently. Each observation of a noise variable is assigned probability 0.003 of being replaced by a large number. If this observation is contaminated, then the corresponding observation of Y is also replaced by a large number to generate bad leverage point. Thus, the probability that any particular row of the training sample data matrix will be contaminated is $1-(1-0.003)^{d-a}$, which is approximately 11.6% for $a = 9$.

For each of the 2 selection procedures (1 classical and 1 robust), we fitted the selected model (including the intercept) on the training data, and then used it to predict the test data outcomes. We used a regression MM estimator[11] to fit the model obtained by the robust method, because of its high breakdown point and high efficiency at the normal model. For each simulated data set, we recorded (1) the average squared prediction error on the test sample considering $m\,(m = 1, 2, \cdots, 30)$ first sequenced variables in the model, and (2) the total number of target variables selected in the model.

To summarize the simulation results, the average (SD) of mean squared prediction error (MSPE) on the test set, and the number t_m of target variables included in the first m sequenced variables was determined for each sequence, with m ranging from 1 to 30. The left panels of Figs. 1 through 4 (Panels 1(a), 2(a), 3(a) and 4(a)) show the MSPE (over 5000 data sets) and the right panels of Figs. 1 through 4 (Panels 1(b), 2(b), 3(b) and 4(b)) show the average of t_m (over 5000 data sets).

For the clean data, the MSPE produced by robust and classical methods shown in left panels of Figs. 1 and 3 (Panels 1(a) and 3(a)) are almost the same. Also, the robust and classical methods contain almost the same average of t_m shown in right panels of Figs. 1 and 3 (Panels 1(b) and 3(b)). For the contaminated data, the MSPE produced by robust method is much smaller than for the classical method shown in panels of Figs. 2 and 4 ((Panels 2(a) and 4(a)). Also, the model obtained by robust met hod contains more target variables than the classical method shown in panels of Figs. 2 and 4 (Panels 2(b) and 4(b)).

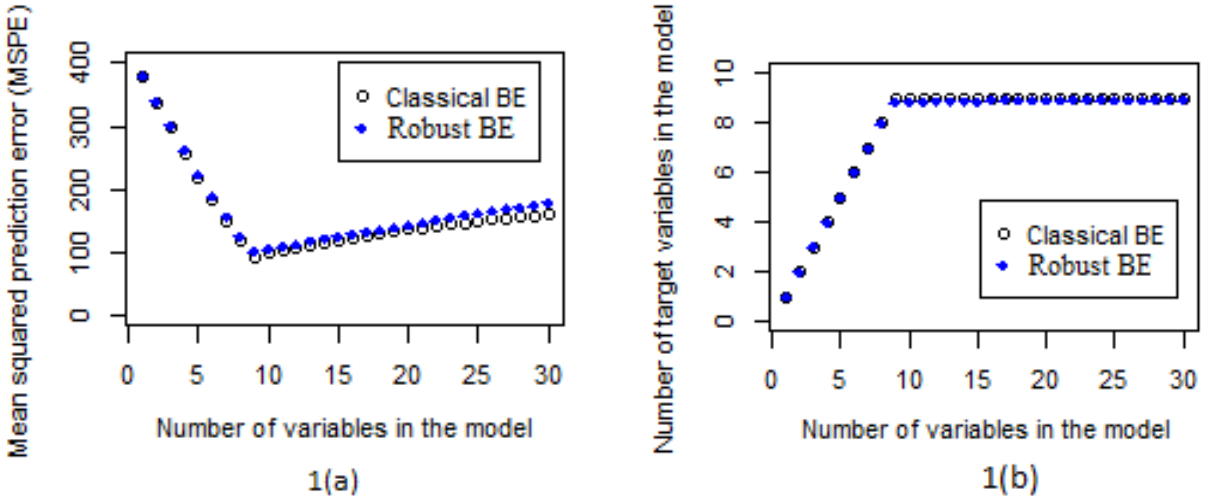

Fig. 1. (a). MSPE for Classical and Robust BE Methods for $a = 9$ non-zero covariates (no correlation case) in clean data.
(b). Average of t_m for Classical and Robust BE methods for $a = 9$ non-zero covariates (no correlation case) in clean data.

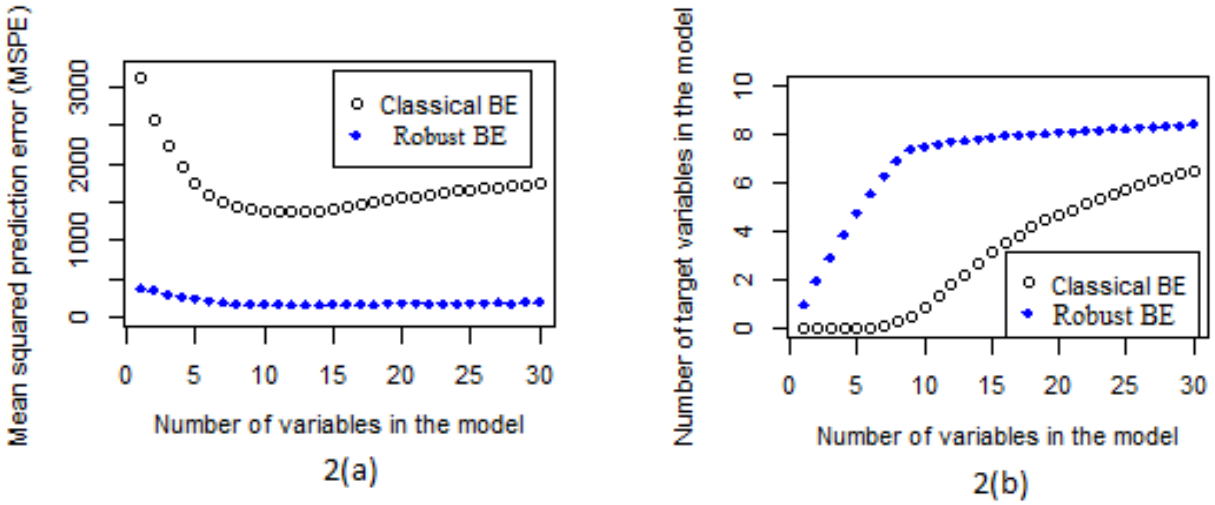

Fig. 2. (a). MSPE for Classical and Robust BE Methods for $a = 9$ non-zero covariates (no correlation case) in contaminated data.
(b). Average of t_m for Classical and Robust BE methods for $a = 9$ non-zero covariates (no correlation case) in contaminated data.

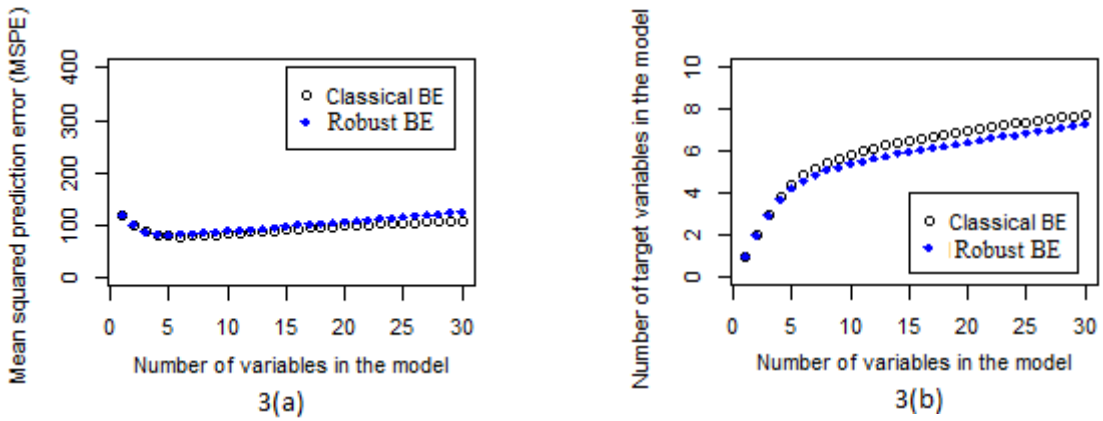

Fig. 3. (a). MSPE for Classical and Robust BE Methods for $a = 9$ non-zero covariates (correlation case) in clean data.
(b). Average of t_m for Classical and Robust BE methods for $a = 9$ non-zero covariates (correlation case) in clean data.

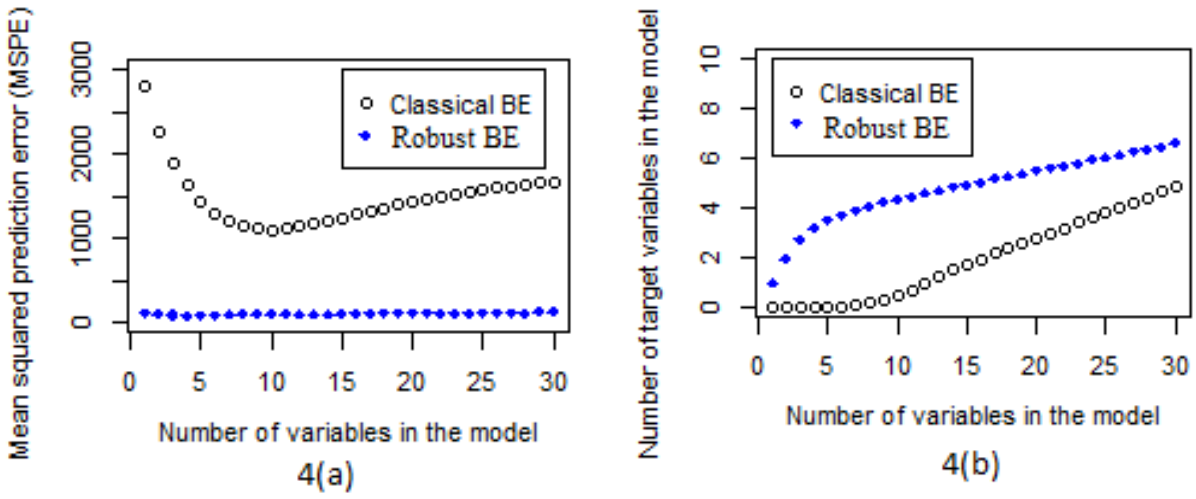

Fig. 4. (a). MSPE for Classical and Robust BE Methods for $a=9$ non-zero covariates (correlation case) in contaminated data.
(b). Average of t_m for Classical and Robust BE methods for $a=9$ non-zero covariates (correlation case) in contaminated data.

V. Real Data Application

In this section, we used a real-data example to show the robustness and scalability of our algorithm.

Executive data

This data set is obtained from Mendenhall *et al.*[17]. The annual salary of 100 executives is recorded as well as 10 potential predictors (7 quantitative and 3 qualitative) such as education, experience etc. We label the candidate predictors from 1 to 10. Classical BE (with $F_{0.95}$ as the deletion criterion) selects the covariates: (1, 3, 4, 2, 5, 9). Robust BE (also with $F_{0.95}$ as deletion criterion) selects almost the same model (1, 3, 4, 2, 5) except the last covarite of model selected by the classical BE.

We then contaminated the data by replacing one small value of predictor 1 (less than 5) by a large value 100. When BE is applied to the contaminated data, it now selects a larger set of variables: (7, 3, 4, 2, 1, 10, 6). Thus, changing a single number in the data set drastically changes the selected model. On the other hand, robust BE selects almost the same model, (1, 3, 4, 2), when applied to the contaminated data set.

VI. Conclusions

BE is a popular and computationally suitable algorithm for building linear prediction models, but they are sensitive to outliers. We express this algorithm in terms of sample means, variances and correlations, and obtained a simple robust version of BE by replacing these sample quantities by their robust counterparts.

For the construction of the robust correlation matrix of the required covariates we used robust correlation estimates between pairs of variables, because it is computationally suitable, and more convenient for (robust) step-by-step algorithms. We used robust correlations derived from Maronna's bivariate M-estimator of the scatter matrix. Though our method may be sensitive to three– or higher–dimensional outliers, this is a very little small price to pay to make the selection of covariates for large values of d.

Our robust method has much better performance compared to the classical BE algorithm. Also it is computationally very suitable, and scalable to large dimensions.

References

1. Furnival, G. and R. Wilson, 1974. Regression by Leaps and Bounds. *Technometrics*. **16**, 499-511.
2. 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.
3. Weisberg, S., 1985. Applied Linear Regression. (2nd ed.), Wiley, New York.
4. Ronchetti, E., 1985. Robust Model Selection in Regression. *Statistics and Probability Letters*, **3**, 21-23.
5. Ronhetti, E., and R. G. Staudte, 1994. A Robust Version of Mallows's C_p. *Journal of the American Statistical Association*, **89**, 550-559.
6. Maronna, R. A., R. D. Martin, and V. J. Yohai, 2006. Robust Statistics: Theory and Methods, John Wiley and Sons.
7. Ronhetti, E., C. Field, and W. Blanchard, 1997. Robust Linear Model Selection by Cross-Validation. *Journal of the American Statistical Association*. **92**, 1017-1023.
8. Sommer, S., and R. M. Huggins, 1996. Variable Selection Using the Wald Test and Robust C_p. *Journal of the Royal Statistical Society*, Ser. B, **45**, 15-29.
9. 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. M Hubert, G. Pison, A. Struyf, and S. Van Aelst, Basel, Switzerland: Birkhauser-Verlag, 195-206.

10. Khan, J. A., S. Van Aelst, R. H. Zamar, 2007. Robust Linear Model Selection Based on Least Angle Regression. *Journal of the American Statistical Association*, **102**, 1289-1299.
11. Yohai, V. J., 1987. High Breakdown Point and High Efficiency Robust Estimates for Regression. *The Annals of Statistics*. **15**, 642-656.
12. Rousseeuw, P. J., and A. M. Leroy, 1987. *Robust Regression and Outlier Detection*, New York: Wiley-Interscience.
13. Huber, P. J., 1981. *Robust Statistics.* Wiley, New York
14. Maronna, R. A., 1976. Robust M-estimators of Location and Scatter. *The Annals of Statistics*, **4**, 51-67.
15. Alqallaf, F. A., K. P. Konis, R. D. Martin, and R. H. Zamar, 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, 14-23.
16. Frank, I, and J. H. Friedman, 1993. A Statistical View of Some Chemometrics Regression Tools. *Technometrics*, **35**, 109-148, New York: Springer-Verlag.
17. Mendenhall, W., and T. Sincich, 2003. A second Course in Statistics: Regression *Analysis. (6th ed.) Pearson Education, Inc., New Jersey.*