

# Building a Robust Linear Model with Forward Selection and Stepwise Procedures

Jafar A. Khan<sup>a</sup>, Stefan Van Aelst<sup>b</sup>, Ruben H. Zamar<sup>c,\*</sup>,

<sup>a</sup>*Dept. of Statistics, University of British Columbia, Canada*

<sup>b</sup>*Dept. of Applied Mathematics and Computer Science, Ghent University, Belgium*

<sup>c</sup>*Dept. of Statistics, University of British Columbia, Canada.*

---

## Abstract

Classical step-by-step algorithms, such as forward selection (FS) and stepwise (SW) methods, are computationally suitable, but yield poor results when the 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 and computationally efficient versions of FS and SW are proposed. Since FS and SW 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 proposed robust methods have much better performance compared to standard FS and SW. Also, they are computationally very suitable and scalable to large high-dimensional datasets.

*Key words:* Stepwise algorithm; computational complexity; robust model selection; pairwise robust correlation.

---

---

\* Corresponding author. Dept. of Statistics, University of British Columbia, 333 - 6356 Agricultural Road, Vancouver, BC, Canada V6T 1Z2. E-mail: ruben@stat.ubc.ca, phone: 1-604-822-3167, fax: 1-604-822-6960.

## 1 Introduction

When the number  $d$  of candidate covariates is small, one can choose a linear prediction model by computing a reasonable criterion (e.g.,  $C_p$ , AIC, FPE or cross-validation error) for all possible subsets of the predictors. However, as  $d$  increases, the computational burden of this approach (sometimes referred to as *all possible subsets regression*) increases very quickly. This is one of the main reasons why step-by-step algorithms like forward selection (FS) or stepwise (SW) are popular. See for example Furnival and Wilson (1974); Gatu and Kontoghiorghes (2006) and Weisberg (1985, Chapter8).

Unfortunately, classical FS or SW procedures yield poor results when the data are contaminated. These algorithms attempt to select the covariates that will fit well all the cases (including the outliers), and often fail 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.

We argue that it is not reasonable to attempt to predict the future outliers without knowledge of the underlying mechanism that produces them. Therefore, our goal is to develop robust step-by-step algorithms that will select important variables in the presence of outliers, and predict well the future non-outlying cases.

We show that the list of variables selected by classical FS and SW procedures are functions of sample means, variances and correlations. We express the two classical algorithms in terms of these sample quantities, and replace them by robust counterparts to obtain simple robust versions of the algorithms. Once the covariates are selected (by using these simple robust selection algorithms), we can use a robust regression estimator on the final model.

Robust correlation matrix estimators for  $d$ -dimensional datasets 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  $n$ . On the other hand, only a few of the  $d$  covariates are typically included in the final model, and the computation of the whole  $d$ -dimensional correlation matrix at once will unnecessarily increase the numerical complexity of the otherwise computationally suitable step-by-step algorithms.

To avoid this complexity, we use an affine-equivariant bivariate M-estimator of scatter to obtain robust correlation estimates for all pairs of variables, and combine these to construct a robust correlation matrix. We call this the

pairwise correlation approach. Interestingly, this pairwise approach for robust correlation matrix estimation is not only computationally suitable, but it is also more convenient (compared to the full  $d$ -dimensional approach) for robust step-by-step algorithms. The reason is as follows. The sample correlation matrix ( $R$ , say) has the property that the correlation matrix of a subset of variables can be obtained by simply taking the appropriate submatrix of  $R$ . This property allows us to compute only the required correlations at each step of the algorithm. With the robust pairwise correlation approach we keep this property.

Affine equivariance and regression equivariance are considered to be important properties for robust regression estimators (see, e.g., Rousseeuw and Leroy, 1987). However, these properties are not required in the context of variable selection, because we do not consider general linear transformations of the given covariates. The only transformations that should not affect the selection result are linear transformations of individual variables, i.e., shifts and scale changes. Variable selection methods are often based on correlations among the 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.

It should be emphasized that with our approach we consider the problem of “selecting” a list of important predictors, but we do not yet “fit” the selected model. 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 (Yohai, 1987) that is resistant to all types of outliers. Note that we always use models with intercept.

Robust selection criteria to compare a set of models have been proposed in the robustness literature. Important examples are Ronchetti (1985); Ronchetti and Staudte (1994); Maronna, Martin, and Yohai (2006), and Ronchetti, Field, and Blanchard (1997) which introduced robust versions of AIC,  $C_p$ , FPE and cross-validation, respectively. Sommer and Huggins (1996) proposed robust model selection based on Wald tests. Morgenthaler, Welsch, and Zenide (2003) constructed a selection technique to simultaneously identify the correct model structure as well as unusual observations. However, most of these papers do not propose any strategy to select the set of models that are to be compared, and often suggest using the time-consuming all-subsets approach. Even when these criteria are used in a step-by-step approach such as FS or SW, the selection procedure remains time-consuming, because a time demanding robust fit needs to be estimated for each model. No literature is available yet on efficient robust counterparts of the step-by-step algorithms.

The rest of the paper is organized as follows. In Section 2 we decompose the FS and SW procedures in terms of the correlation matrix of the data. In Section 3, we present robust versions of these algorithms, along with their numerical complexities. Section 4 presents a Monte Carlo study that compares our robust methods with the classical ones by their predicting powers. Section 5 contains a real-data application. Section 6 is the conclusion.

## 2 FS and SW Algorithms Expressed in Correlations

In this section we review the classical FS and SW selection procedures. For clarity of exposition, we show how both procedures can be expressed only in terms of classical correlations between pairs of variables.

### 2.1 FS expressed in correlations

Let  $X_1, \dots, X_d$  be  $n$ -dimensional vectors representing the covariates, and  $Y$  be the  $n$ -dimensional vector representing the response. Let the variables be standardized using their mean and standard deviation. The FS procedure selects the covariate ( $X_1$ , say) that has the largest absolute correlation  $|r_{1y}|$  with  $Y$ , and calculates the residual vector  $Y - r_{1y}X_1$ . All the other covariates are then ‘adjusted for  $X_1$ ’ and entered into competition. That is, each  $X_j$  is regressed on  $X_1$ , and the corresponding residual vector  $Z_{j,1}$  (which is orthogonal to  $X_1$ ) is obtained. The correlations of these  $Z_{j,1}$  with the residual vector  $Y - r_{1y}X_1$ , which are also called “the partial correlations between  $X_j$  and  $Y$  adjusted for  $X_1$ ,” decide the next variable to enter the regression model, and so on. We need  $(d - 1)$  steps to get the ordering of all  $d$  predictors.

The reason behind the ‘orthogonalization,’ that is, the construction of  $Z_{j,1}$  from  $X_j$ , is that the algorithm measures what ‘additional’ contribution  $X_j$  makes in explaining the variability of  $Y$ , when  $X_j$  joins  $X_1$  in the regression model. The  $R^2$  produced by  $(X_1, Z_2)$  is the same as the  $R^2$  produced by  $(X_1, X_2)$ , and the orthogonalization ensures maximum  $R^2$  at each FS step.

Let  $r_{jy}$  denote the correlation between  $X_j$  and  $Y$ , and  $R_X$  be the correlation matrix of the covariates  $X_1, \dots, X_d$ . Suppose w.l.o.g. that  $X_1$  has the maximum absolute correlation with  $Y$ . Then,  $X_1$  is the first variable that enters the regression model. The predictors that are in the current regression model are called *active* predictors. The remaining candidate predictors are called the *inactive* predictors. We now need the partial correlations between  $X_j$  ( $j \neq 1$ ) and  $Y$  adjusted for  $X_1$ , denoted by  $r_{jy,1}$ . The second covariate  $X_2$  (say) that enters the regression model is then the covariate that has maximal partial

correlation  $r_{jy,1}$  with  $Y$ .

**The partial correlations  $r_{jy,1}$  expressed in original correlations.** Each inactive covariate  $X_j$  should be regressed on  $X_1$  to obtain the residual vector  $Z_{j,1}$  as follows

$$Z_{j,1} = X_j - \beta_{j1}X_1, \quad (1)$$

where

$$\beta_{j1} = \frac{1}{n}X_1^t X_j = r_{j1}. \quad (2)$$

Moreover, we have

$$\frac{1}{n}Z_{j,1}^t Y = \frac{1}{n}(X_j - \beta_{j1}X_1)^t Y = r_{jy} - r_{j1}r_{1y}, \quad (3)$$

and

$$\frac{1}{n}Z_{j,1}^t Z_{j,1} = \frac{1}{n}(X_j - \beta_{j1}X_1)^t (X_j - \beta_{j1}X_1) = 1 - r_{j1}^2. \quad (4)$$

The partial correlation  $r_{jy,1}$  is given by

$$r_{jy,1} = \frac{Z_{j,1}^t (Y - \beta_{y1}X_1)/n}{\sqrt{Z_{j,1}^t Z_{j,1}/n} \text{SD}(Y - \beta_{y1}X_1)}. \quad (5)$$

Note that the factor  $\text{SD}(Y - \beta_{y1}X_1)$  in the denominator of (5) is independent of the covariates  $X_j$ ; ( $j = 2, \dots, d$ ) being considered. Hence, when selecting the covariate  $X_j$  that maximizes the partial correlation  $r_{jy,1}$ , this constant factor can be ignored. This reduces computations and therefore is more time efficient. It thus suffices to calculate

$$\tilde{r}_{jy,1} = \frac{Z_{j,1}^t (Y - \beta_{y1}X_1)/n}{\sqrt{Z_{j,1}^t Z_{j,1}/n}}, \quad (6)$$

where  $\tilde{r}_{jy,1}$  is proportional to the actual partial correlation. Since  $Z_{j,1}$  and  $X_1$  are orthogonal and by using (3) and (4),  $\tilde{r}_{jy,1}$  can be rewritten as follows

$$\tilde{r}_{jy,1} = \frac{Z_{j,1}^t Y/n}{\sqrt{Z_{j,1}^t Z_{j,1}/n}} = \frac{r_{jy} - r_{j1}r_{1y}}{\sqrt{1 - r_{j1}^2}}. \quad (7)$$

Now, suppose w.l.o.g. that  $X_2$  (or, equivalently,  $Z_{2,1}$ ) is the new active covariate, because it minimizes  $\tilde{r}_{jy,1}$  (and thus also the partial correlation  $r_{jy,1}$ ). All the inactive covariates should now be orthogonalized with respect to  $Z_{2,1}$ .

**Orthogonalization of  $Z_{j,1}$  wrt  $Z_{2,1}$ .** Each inactive vector  $Z_{j,1}$  should be regressed on  $Z_{2,1}$  to obtain the residual vector  $Z_{j,12}$  as follows

$$Z_{j,12} = Z_{j,1} - \beta_{j2,1}Z_{2,1}.$$

Here,

$$\begin{aligned}
\beta_{j2.1} &= \frac{Z_{2.1}^t Z_{j.1}/n}{Z_{2.1}^t Z_{2.1}/n} \\
&= \frac{X_2^t Z_{j.1}/n}{Z_{2.1}^t Z_{2.1}/n} \quad [\text{because of orthogonality}] \\
&= \frac{X_2^t (X_j - r_{j1} X_1)/n}{Z_{2.1}^t Z_{2.1}/n} \quad [\text{Using (1) and (2)}] \\
&= \frac{r_{2j} - r_{21} r_{j1}}{1 - r_{21}^2} \quad [\text{using (squared) denominator of (7) for } j = 2].
\end{aligned} \tag{8}$$

Thus,  $\tilde{r}_{jy.1}$  and  $\beta_{j2.1}$  are expressed in terms of original correlations.

**Lemma 1.** Given that

$$\tilde{r}_{jy.1\dots(k-1)} = \frac{Z_{j.1\dots(k-1)}^t Y/n}{\sqrt{Z_{j.1\dots(k-1)}^t Z_{j.1\dots(k-1)}/n}}, \quad \text{for } k = 2, \dots, (d-1); j \text{ inactive,} \tag{9}$$

and

$$\beta_{jh.1\dots(h-1)} = \frac{Z_{h.1\dots(h-1)}^t Z_{j.1\dots(h-1)}/n}{Z_{h.1\dots(h-1)}^t Z_{h.1\dots(h-1)}/n}, \quad \text{for } h = 2, \dots, k; j \text{ inactive,} \tag{10}$$

are functions of original correlations, the following quantities can be expressed as functions of original correlations: (a)  $\tilde{r}_{jy.1\dots k}$  and (b)  $\beta_{j(k+1).1\dots k}$ .

**Proof.** Here,  $\tilde{r}_{jy.1\dots(k-1)}$  determines the next active covariate  $X_k$  (or, equivalently,  $Z_{k.1\dots(k-1)}$ ). Orthogonalization of the remaining inactive predictors is obtained by

$$Z_{j.1\dots k} = Z_{j.1\dots(k-1)} - \beta_{jk.1\dots(k-1)} Z_{k.1\dots(k-1)}. \tag{11}$$

Now,

$$\tilde{r}_{jy.1\dots k} = \frac{Z_{j.1\dots k}^t Y/n}{\sqrt{Z_{j.1\dots k}^t Z_{j.1\dots k}/n}}. \tag{12}$$

Hence, it follows from (9)-(11) that  $\tilde{r}_{jy.1\dots k}$  can be expressed as a function of the original correlations which proves Part (a) of the lemma. The proof of part (b) is similar.

### 2.1.1 FS steps in correlations

We can now summarize the FS algorithm in terms of correlations among the original variables as follows:

- (1) To select the first covariate  $X_{m_1}$ , determine  $m_1 = \operatorname{argmax} |r_j|$ .
- (2) To select the  $k$ th covariate  $X_{m_k}$  ( $k = 2, 3, \dots$ ), calculate  $\tilde{r}_{jy.m_1\dots m_{(k-1)}}$ , which is proportional to the partial correlation between  $X_j$  and  $Y$  adjusted for  $X_{m_1}, \dots, X_{m_{(k-1)}}$ , and then determine  $m_k = \operatorname{argmax} |\tilde{r}_{jy.m_1\dots m_{(k-1)}}|$ .

### 2.1.2 Stopping rule

At each FS step, once the “active” covariate is identified, we can perform a partial F-test to decide whether to include this covariate in the model (and continue the process) or to stop. The “active” covariate enters the model only if the partial F-value, denoted by  $F_{\text{partial}}$ , is greater than  $F(0.95, 1, n - k - 1)$  (say), where  $k$  is the current size of the model including the “active” covariate. These partial F-tests have been shown to be useful stopping rules for the classical step-by-step procedures. Note, however, that they are no formal hypothesis tests anymore, because a sequence of these partial F-tests is performed which introduces pre-test complications. Therefore, we call this the partial F-rule for now on. The quantities required for the partial F-rules can be expressed in terms of correlations among the original variables, as shown below.

Suppose that  $X_1$  is already included in the model, and  $X_2$  has the largest absolute partial correlation with  $Y$  after adjusting for  $X_1$ . To decide whether  $X_2$  should be included in the model we perform a partial F-rule as follows:

$$\begin{aligned}
 F_{\text{partial}} &= \frac{(Y - \beta_{y1}X_1)^t(Y - \beta_{y1}X_1) - (Y - \beta_{y1}X_1 - \beta_{y2.1}Z_{2.1})^t(Y - \beta_{y1}X_1 - \beta_{y2.1}Z_{2.1})}{(Y - \beta_{y1}X_1 - \beta_{y2.1}Z_{2.1})^t(Y - \beta_{y1}X_1 - \beta_{y2.1}Z_{2.1})/(n - 3)} \\
 &= \frac{(n - 3) (2 \beta_{y2.1}Z_{2.1}^t Y/n - \beta_{y2.1}^2 Z_{2.1}^t Z_{2.1}/n)}{1 - r_{1y}^2 - (2 \beta_{y2.1}Z_{2.1}^t Y/n - \beta_{y2.1}^2 Z_{2.1}^t Z_{2.1}/n)} \\
 &= \frac{(n - 3) (\beta_{y2.1}Z_{2.1}^t Y/n)}{1 - r_{1y}^2 - \beta_{y2.1}Z_{2.1}^t Y/n} \\
 &= \frac{(n - 3) \tilde{r}_{2y.1}^2}{1 - r_{1y}^2 - \tilde{r}_{2y.1}^2},
 \end{aligned}$$

where  $\tilde{r}_{2y.1}$  is expressed in correlations in (7).

Similarly, when  $(k - 1)$  covariates  $X_1, \dots, X_{k-1}$  are already in the model, and w.l.o.g.  $X_k$  has the largest absolute partial correlation with  $Y$  after adjusting for  $X_1, \dots, X_{k-1}$ , the partial F-rule for  $X_k$  can be expressed as:

$$F_{\text{partial}} = \frac{(n - k - 1) \tilde{r}_{ky.1 \dots (k-1)}^2}{1 - r_{1y}^2 - \tilde{r}_{2y.1}^2 - \dots - \tilde{r}_{ky.1 \dots (k-1)}^2}.$$

## 2.2 SW expressed in correlations

The SW algorithm (Weisberg, 1985, Chapter 8) is the same as the FS procedure up to the second step. When there are at least two covariates in the model, at each subsequent SW step we either (a) add a covariate, or (b) drop a covariate, or (c) exchange two covariates, or (d) stop.

To decide whether to add a covariate, the partial correlations of each inactive covariate  $X_j$  with  $Y$  can be computed as in the case of FS (see Equation 12) to perform a partial F-rule (see Section 2.1.2). To decide whether to drop an “active” covariate, we can pretend that the active covariate under consideration entered the model last, and calculate its partial correlations with  $Y$  (see Equation 12, subscripts modified) to perform a partial F-rule (Section 2.1.2, subscripts modified).

Once an “active” covariate is dropped, the “orthogonalizations” of the other covariates (active or inactive) with this covariate that were used before to derive the partial correlations become irrelevant, and the order of the other active covariates in the model cannot be determined. Fortunately, this does not create a problem to decide the next covariate, because, for example,  $r_{jy.346} = r_{jy.643}$ . Therefore, we can update all relevant calculations considering the currently active covariates in any order.

**Stopping criteria for SW.** Unlike the FS algorithm where a stopping criterion is “optional” (we may choose to sequence all the covariates), SW has to have a built-in stopping rule, because at each step we have to decide whether to add one covariate and/or delete another. We may choose two different theoretical F percentiles as the inclusion and deletion criteria, e.g.,  $F(0.95, 1, n - k_1 - 1)$  and  $F(0.90, 1, n - k_2 - 1)$ , respectively, where  $k_1$  and  $k_2$  are the model sizes after inclusion and before deletion.

## 3 Robustification of FS and SW algorithms

In the last section we expressed the FS and SW algorithms in terms of sample means, variances and correlations. Because of these non-robust building blocks, these algorithms are sensitive to contamination in the data. A simple robustification of these algorithms can be achieved by replacing the non-robust ingredients of the algorithms by their robust counterparts. For the initial standardization of the variables, 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 con-

suming (see, e.g., Rousseeuw and Leroy, 1987). On the other hand, robust univariate approaches (see, e.g., Huber, 1981) are very sensitive to correlation outliers (outliers that are not detected by univariate analyses but affect the classical correlation).

One solution is to derive correlations among pairs of variables from an affine-equivariant bivariate covariance estimator. A computationally efficient choice is a bivariate M-estimator proposed by Maronna (1976). Maronna's bivariate M-estimator of the location vector  $\mathbf{t}$  and scatter matrix  $\mathbf{V}$  is defined as the solution of the system of equations:

$$\frac{1}{n} \sum_i u_1(d_i)(\mathbf{x}_i - \mathbf{t}) = \mathbf{0}$$

and

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

where  $d_i^2 = (\mathbf{x}_i - \mathbf{t})' \mathbf{V}^{-1} (\mathbf{x}_i - \mathbf{t})$ , and  $u_1$  and  $u_2$  are functions satisfying a set of general assumptions. The estimator is affine equivariant and has breakdown point  $1/3$  in two dimensions (Maronna, 1976). To further simplify computations, we use the coordinatewise median as the bivariate location estimate 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  $\chi_2^2$  distribution. Finally, FS and SW algorithms are 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 2. Since our 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. We also verified this empirically in a small simulation study.

### 3.1 Numerical complexity of the algorithms

If we sequence all  $d$  covariates, the standard FS procedure requires  $\mathcal{O}(nd^2)$  time. However, when applied with a stopping criterion, the complexity of FS depends on the number of covariates selected in the model. Assuming that the model size will not exceed a certain number  $m < d$ , the complexity of FS is less than or equal to  $\mathcal{O}(ndm)$ . Similarly, the maximum complexity of SW is  $\mathcal{O}(n(dm + m^2)) = \mathcal{O}(ndm)$ .

Since we used the coordinatewise median as the bivariate location estimate, the

correlation based on Maronna’s M-estimate can be computed in  $\mathcal{O}(n \log n + bn)$  time, where  $b$  is the number of iterations required. Assuming that  $b$  does not exceed  $\mathcal{O}(\log n)$  (convergence was achieved after 3 to 5 iterations in our simulations), the complexity of this estimate is  $\mathcal{O}(n \log n)$ . As a result, the maximum complexity of robust FS is  $\mathcal{O}((n \log n)dm)$ , and the maximum complexity of robust SW is  $\mathcal{O}((n \log n)(dm + m^2)) = \mathcal{O}((n \log n)dm)$ .

Though *all possible subsets regression* is expected to select a better model (with respect to predictive power) than any step-by-step algorithm, its computational burden is extremely high for large values of  $d$ , since it requires the fitting of all  $2^d - 1$  submodels. The complexity of the classical algorithms of this type is  $\mathcal{O}(2^d nd^2)$ . Since robust model selection methods proposed so far uses *all possible subsets regression*, the complexity of the existing robust algorithms is  $\mathcal{O}(2^d nd^2)$  multiplied by the number of iterations required for the robust fits.

We can consider an alternative approach for the robustification of FS and SW that obtains MM-estimates for each model under consideration and uses robust FPE (Maronna, Martin, and Yohai, 2006) as the stopping criterion (see also Section 4). The numerical complexity of this MM-RFPE method is greater than  $\mathcal{O}(nd^2 + 2^m nm^2)$ , because we cannot recycle the calculations of any particular step for the next step, or avoid the fitting of the full model.

### 3.2 Limitation of the proposed algorithms

The robust FS and SW procedures based on robust pairwise correlations proposed are resistant to bivariate (correlation) outliers. However, they 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 (see, e.g., Alqallaf et al., 2002).

It should be emphasized here that these are very small prices to pay to make the selection of covariates possible for large values of  $d$ . For example, in our simulations (presented later) we used  $d = 50$ . It is impossible to apply *all possible subsets regression* on a dataset of this dimension. If one robust fit takes 0.001 cpu second, we would need  $2^{50} * 0.001 / (3600 * 24 * 365)$  years to select the final model.

## 4 A simulation study

To compare our robust methods with the classical ones, we carried out a simulation study similar to Frank and Friedman (1993). The total number of variables is  $d = 50$ . A small number  $a = 9$  or  $a = 15$  of them are nonzero covariates. We considered 2 correlation structures of these nonzero covariates: “no correlation” case and “moderate correlation” case, which are described below.

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  $a = 9$ , and five times for  $a = 15$ . 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

$$Y = 7L_1 + 6L_2 + 5L_3 + \epsilon = \text{Signal} + \epsilon, \quad (13)$$

where  $L_i \sim N(0, 1)$ , and  $\epsilon$  is a normal error not related to the latent variables. The variance of  $\epsilon$  is chosen such that the signal-to-noise ratio equals 2, that is  $\text{Var}(\epsilon) = 110/4$ . The nonzero 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 1000 datasets 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 a 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 10% for  $a = 15$ , and 11.6% for  $a = 9$ .

For each of the 4 selection procedures (2 classical and 2 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 (Yohai,

1987) to fit the models obtained by either of the robust methods. Though the robust correlations used for the selection of covariates allow us to obtain the robust regression coefficients, we used an MM-estimator on the final model to obtain a fully robust fit that is also resistant to high-dimensional outliers.

Table 1

Performance of the classical and robust methods in clean and contaminated data for no-correlation case. The average mean squared prediction error (MSPE) on the test set and the average number of noise variables (Noise) selected are shown.

Data	Method	$a = 9$		$a = 15$	
		MSPE	Noise	MSPE	Noise
Clean	FS	55.6 (11.6)	5.0 (2.4)	107.0 (21.7)	4.6 (2.3)
	SW	55.8 (11.8)	4.8 (2.3)	108.1 (22.1)	4.3 (2.1)
	Rob FS	56.5 (12.4)	5.1 (2.6)	109.9 (21.6)	4.8 (2.4)
	Rob SW	56.7 (12.8)	4.9 (2.5)	108.4 (22.4)	4.6 (2.3)
Contam	FS	161.8 (38.1)	13.6 (3.0)	296.7 (75.3)	11.9 (2.8)
	SW	162.5 (37.5)	13.4 (2.8)	297.9 (75.9)	11.7 (2.7)
	Rob FS	72.5 (13.9)	2.1 (2.4)	124.1 (19.9)	1.2 (1.8)
	Rob SW	72.6 (13.8)	2.1 (2.3)	124.2 (20.8)	1.2 (1.7)

Table 2

Performance of the classical and robust methods in clean and contaminated data for moderate-correlation case. The average mean squared prediction error (MSPE) on the test set and the average number of noise variables (Noise) selected are shown.

Data	Method	$a = 9$		$a = 15$	
		MSPE	Noise	MSPE	Noise
Clean	FS	59.7 (12.0)	4.9 (2.4)	50.2 (9.3)	4.3 (2.2)
	SW	60.3 (12.3)	4.8 (2.3)	51.2 (9.7)	4.2 (2.1)
	Rob FS	60.4 (12.2)	5.1 (2.6)	51.5 (10.3)	4.7 (2.5)
	Rob SW	61.1 (12.8)	5.0 (2.5)	52.8 (10.5)	4.6 (2.4)
Contam	FS	157.6 (40.8)	13.6 (3.1)	134.5 (32.9)	11.7 (2.9)
	SW	158.4 (41.3)	13.4 (3.0)	136.3 (33.3)	11.6 (2.8)
	Rob FS	94.9 (27.9)	2.5 (2.9)	78.9 (23.7)	1.6 (2.9)
	Rob SW	95.1 (27.8)	2.4 (2.8)	79.3 (23.4)	1.5 (2.6)

For each simulated dataset, we recorded the size of the selected model (including the noise variables selected), the number of noise variables in the model, and the mean squared prediction error (MSPE) on the test sample.

Table 1 shows the average (sd) of the MSPE, and the number of noise variables selected in the model over all generated datasets for the no-correlation case. In general, FS performs as good as SW, and robust FS performs as good as robust SW. For the clean data, the performance of robust FS (SW) is comparable to standard FS (SW). For the contaminated data, the MSPE produced by robust methods is much smaller than for the classical methods. Also, the models obtained by robust methods contain less noise variables than the classical methods.

Table 2 presents the results for the moderate-correlation case. We obtain the same conclusions as in the no-correlation case.

We applied the MM-RFPE method on 50 simulated datasets, and obtained similar results compared to the robust FS and SW. However, the computational burden of the MM-RFPE method is at least 100 times larger than that of our proposed methods in the above setup. For larger values of  $d$ , the computational burden of MM-RFPE will dramatically increase, making this method infeasible.

## 5 Example

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

**Particle data.** This quantum physics dataset was used for the KDD-Cup 2004. Each of  $n = 50000$  data-points (rows) describes one “example” (particle generated in a high energy collider experiment). There are 80 variables in the data: Example ID, class of the example (positive examples are denoted by 1, negative examples by 0), and 78 feature measurements. We considered only the feature variables in our analysis. We deleted 13 of the features (either because they have a large number of missing values, or they are degenerate with all observations equal to 0), and used the first feature as the response. Thus, we have 64 covariates and one response. Though this analysis may not be of particular scientific interest, it will demonstrate the scalability and robustness of our algorithms.

We first applied the four algorithms (FS, SW, Rob FS and Rob SW) to a randomly selected training sample of size  $n = 5000$ . The remaining 45000 cases constitute the test sample. The classical FS and SW (with  $F_{0.9}$  criterion) both select a huge model with the following 25 covariates:

(2, 60, 58, 18, 8, 4, 51, 53, 1, 59, 5, 20, 10, 6, 62, 19, 38, 46, 39, 47, 21, 36, 50, 48, 37).

With the  $F_{0.95}$  criterion, the model has 23 covariates. On the other hand,

robust FS and SW (with either  $F_{0.9}$  or  $F_{0.95}$  criterion) select a model with only one covariate,  $X_1$ . Since the intercept and the residuals scale of this model, as well as the median and mad of  $X_1$  are all exactly equal to zero, we conclude that  $X_1 = Y = 0$  for more than 50% (in fact, 85.6%) of the cases. This clearly suggests the following “two-stage” robust prediction strategy.

Because of the unusual pattern detected by the robust method, we considered the part of the training data (528 cases) for which  $X_1 \neq 0$  for further investigation. We applied robust FS on this part, and selected the following set of covariates: (62, 5, 8, 58, 24). The final robust prediction rule is as follows. If  $X_1 = 0$ , predict  $Y = 0$ . If  $X_1 \neq 0$ , predict  $Y$  using the robust fit based on the 5 covariates above.

We used the selected classical and robust models to predict the test data outcomes. The 1% and 5% trimmed means of squared prediction errors for the classical and (robust) models are: 0.110 (0.032) and 0.012 (0.001), respectively. That is, the robust model with fewer covariates predicts 99% of the data better than the huge classical model.

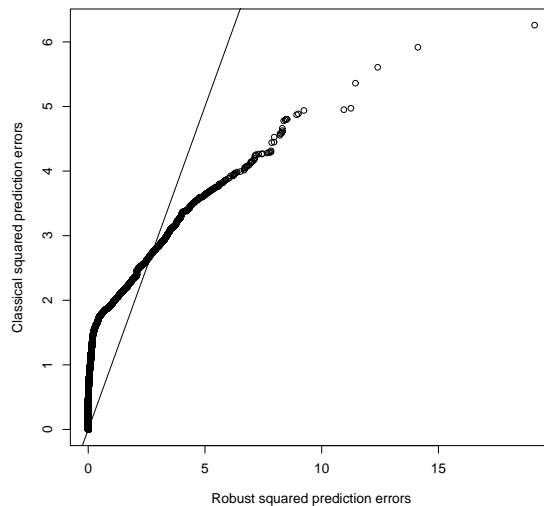


Fig. 1. QQplot of the classical squared prediction errors against the robust squared prediction errors for the test data.

Figure 1 shows the QQplot of the classical squared prediction errors (vertical axis) against the robust squared prediction errors (horizontal axis) for the test data. For the 439 cases with the largest squared errors (<1% of the test cases), the robust errors are larger than the classical ones.

Note that given the unusual patterns in this dataset, the MM-RFPE method cannot be used. In fact, the computation of the MM-estimate for the full model crashes (returns an error message).

To illustrate the scalability and stability of our algorithm we also used a training sample of size  $n = 25000$ . This time, classical FS and SW select a model of 30 covariates, and robust FS and SW both select one covariate, in this case  $X_2$  instead of  $X_1$ . We notice that  $X_1$  and  $X_2$  have robust correlations 0.82 and  $-0.85$  with  $Y$ , respectively.

## 6 Conclusions

FS and SW are popular and computationally suitable algorithms for building linear prediction models, but they are sensitive to outliers. We expressed these algorithms in terms of sample means, variances and correlations, and obtained simple robust versions of FS and SW 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 both 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 methods may be sensitive to three- or higher-dimensional outliers, this is a very small price to pay to make the selection of covariates possible for large values of  $d$ .

Our robust methods have much better performance compared to the standard FS and SW algorithms. Also, they are computationally very suitable, and scalable to large dimensions.

## 7 Acknowledgement

We thank the referees and editors for useful references and suggestions. The research of Stefan Van Aelst was supported by a grant of the Fund for Scientific Research-Flanders (FWO-Vlaanderen). The research of Ruben H. Zamar and Jafar A. Khan was supported by NSERC.

## References

Alqallaf, F.A., Konis, K.P., Martin, R.D., and Zamar, R.H., 2002. Scalable robust covariance and correlation estimates for data mining. In: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Edmonton, Alberta, 14-23.

- Frank, I., and Friedman, J.H., 1993. A statistical view of some chemometrics regression tools. *Technometrics*, 35, 109-148.
- Gatu, C., and Kontoghiorghes, E.J., 2006. Branch-and-bound algorithms for computing the best subset regression models. *J. Comput. Graph. Statist.*, 15, 139-156.
- Furnival, G., and Wilson, R. 1974. Regression by leaps and bounds. *Technometrics*, 16, 499-511.
- Huber, P.J., 1981. *Robust statistics*. Wiley, New York.
- Maronna, R.A., 1976. Robust M-estimators of multivariate location and scatter. *Ann. Statist.*, 4, 51-67.
- Maronna, R.A., Martin, R.D., and Yohai, V.J., 2006. *Robust statistics: Theory and methods*. John Wiley and Sons.
- Morgenthaler, S., Welsch, R. E. and Zenide, A., 2003. Algorithms for robust model selection in linear regression. In: M. Hubert, G. Pison, A. Struyf, and S. Van Aelst (Eds.), *Theory and Applications of Recent Robust Methods*, Birkhäuser-Verlag, Basel (Switzerland), 195-206.
- Ronchetti, E., 1985. Robust model selection in regression. *Statist. Prob. Letters*, 3, 21-23.
- Ronchetti, E., Field, C. and Blanchard, W., 1997. Robust linear model selection by cross-validation. *J. Amer. Statist. Assoc.*, 92, 1017-1023.
- Ronchetti, E., and Staudte, R.G., 1994. A robust version of Mallows's  $C_p$ . *J. Amer. Statist. Assoc.*, 89, 550-559.
- Rousseeuw, P.J., and Leroy, A.M., 1987. *Robust regression and outlier detection*. Wiley, New York.
- Sommer, S., and Huggins, R.M., 1996. Variable selection using the Wald test and robust  $C_p$ . *J. R. Statist. Soc. B*, 45, 15-29.
- Weisberg, S., 1985. *Applied linear regression*. (2nd ed.). Wiley, New York.
- Yohai, V.J., 1987. High breakdown point and high efficiency robust estimates for regression. *Ann. Statist.*, 15, 642-656.