Robust Linear Model Selection Based on Least Angle Regression

Stefan Van Aelst, Jafar A. Khan, and Ruben H. Zamar

Abstract

In this paper we consider the problem of building a linear prediction model when the number d of candidate covariates is large and the data possibly contains anomalies that are difficult to visualize and clean. We aim at predicting the non-outlying cases. Therefore, we need a method that is robust and scalable at the same time. We consider the stepwise algorithm LARS which is computationally very efficient, but is sensitive to outliers. We introduce two different approaches to robustify LARS. The *plug-in* approach replaces the classical correlations in LARS by robust correlation estimates. The *cleaning* approach first transforms the dataset by shrinking the outliers toward the bulk of the data (which we call *multivariate Windsorization*), and then applies LARS on the transformed data. In particular, the plug-in approach is a time-efficient and scalable procedure for robust linear model selection, which we call *robust LARS*. We propose to use bootstrap to stabilize the results obtained by robust LARS. We recommend the use of robust bootstrapped LARS to sequence a number of predictors to form a *reduced set* from which a more refined model can be selected.

KEY WORDS: Stepwise algorithm; Robust prediction; Computational complexity; Windsorization; Bootstrapping.

Jafar A. Khan is PhD candidate and Ruben H. Zamar is Professor, Department of Statistics, University of British Columbia, Vancouver, British Columbia, Canada V6T-1Z2. Stefan Van Aelst is Assistant Professor, Department of Applied Mathematics and Computer Science, Ghent University, Krijgslaan 281 S9, B-9000 Gent, Belgium.

1 INTRODUCTION

Robust model selection has not received much attention in the robustness literature. Seminal papers that address this issue include Ronchetti (1985) and Ronchetti and Staudte (1994) which introduced robust versions of the selection criteria AIC and C_p, respectively. Yohai (1997) proposed a robust Final Prediction Error (FPE) criterion (for Splus documentation). Morgenthaler, Welsh, and Zenide (2003) constructed a selection technique to simultaneously identify the correct model structure as well as unusual observations. Ronchetti, Field, and Blanchard (1997) proposed robust model selection by cross-validation. A major drawback of most robust model selection methods is that they are very time consuming, as they require robust fitting of a large number of submodels. One exception is a model selection procedure based on the Wald test (Sommer and Huggins 1996) which requires the computation of estimates only from the full model. However, often the purpose is to select a subset of a large number d of possible predictors, and the fitting of the 'full model' may not be feasible.

Our goal is to develop a selection procedure that can find the important predictors in a large list of candidate predictors. Our model selection strategy proceeds in two steps. The first step - which we call *sequencing* - quickly screens out unimportant variables to form a "reduced set" for further consideration. The second step - which we call *segmentation* - carefully examines the predictors in the reduced set for possible inclusion in the prediction model. For the segmentation in the second step the aforementioned robust selection techniques can be used because the set of candidate predictors has been reduced to a feasible size. Thus, the goal of the first step is a drastic reduction of the number of candidate predictors. The input variables are sequenced to form a list such that the good predictors appear in the beginning. The first m variables of the list then form the reduced set from which the prediction model will be obtained. This paper focuses on the construction of the sequence and the determination of appropriate values of m for the reduced set. The probability that the reduced set contains all the important variables increases with m. Unfortunately, also the computational cost of the second step increases with m. Therefore, we aim to determine a "reasonable " value of m which is large enough to include most important variables but not so large as to make the second

step impractical or unfeasible.

One strategy for sequencing the candidate predictors is to use one of the several available stepwise or stagewise procedures such as forward selection (FS) (see, e.g., Weisberg 1985, chap. 8) or stagewise forward selection (SFS) (see Hastie, Tibshirani, and Friedman 2001, chap. 10). We focus on a powerful technique recently proposed by Efron, Hastie, Johnstone, and Tibshirani (2004) called least angle regression (LARS) which is computationally very efficient.

Since LARS is based on sample means, variances and correlations (as will be shown later), it yields poor results when the data is contaminated. This is a potentially serious deficiency. Therefore, we propose several approaches to strengthen the robustness properties of LARS without affecting its computational efficiency too much and compare their behavior.

The rest of the paper is organized as follows. In Section 2 we express the LARS procedure in terms of the correlation matrix of the data. In Section 3, we illustrate LARS' sensitivity to outliers and introduce two different approaches to robustify LARS. A small simulation study is also presented here to compare the performance and the computing time of LARS to those of the two robust approaches. In Section 4, we investigate the selection of the size of the reduced set of candidate predictors. Section 5 proposes to use bootstrap to stabilize the results obtained by robust LARS. Section 6 introduces "learning curves" as a graphical tool to choose the size of the reduced set. Section 7 contains some real-data applications. Section 8 concludes and the Appendix contains some technical derivations.

2 LEAST ANGLE REGRESSION

Efron et al. (2004) proposed Least Angle Regression which is closely related to the stagewise forward selection (SFS) and LASSO (Tibshirani 1996) procedures. LARS provides an ordering in which the variables enter the model. This sequence is usually the same as in LASSO or SFS but obtained in a computationally efficient way.

The SFS procedure enters variables in small steps in the regression model to prevent

correlated predictors from being excluded from the top of the sequence. However, this method is often time consuming due to the fact that often a large number of small steps are taken in the direction of the same variable. LARS solves this problem by analytically determining the optimal step size for each variable.

Another convenient feature of LARS is that the resulting sequence of the covariates can be derived from the correlation matrix of the data (without the observations themselves). Let $Y, X_1, \ldots, X_d$ be the variables, standardized using their mean and standard deviation. Let r_j denote the correlation between X_j and Y, and R_X be the correlation matrix of the covariates $X_1, \ldots, X_d$. Suppose that X_m has the maximum absolute correlation r with Y and denote $s_m = \text{sign}(r_m)$. Then, X_m becomes the first *active variable* and the current prediction $\hat{\boldsymbol{\mu}} \leftarrow \mathbf{0}$ should be modified by moving along the direction of $s_m X_m$ upto a certain distance γ that can be expressed in terms of correlations between the variables (see Appendix A for details). By determining γ, LARS simultaneously identifies the new covariate that will enter the model, that is the second active variable.

As soon as we have more than one active variable, LARS modifies the current prediction along the *equiangular direction*, that is the direction that has equal angle (correlation) with all active covariates. Moving along this direction ensure that the current correlation of each active covariate with the residual decreases equally. Let A be the set of subscripts corresponding to the active variables. In Appendix B the standardized equiangular vector B_A is derived. Note that we do not need the direction B_A itself to decide which covariate enters the model next. We only need the correlation of all variables (active and inactive) with B_A. These correlations can be expressed in terms of the correlation matrix of the variables as shown in Appendix B. LARS modify the current prediction by moving along B_A upto a certain distance γ_A which, again, can be determined from the correlations of the variables (see Appendix C).

We now summarize the LARS algorithm in terms of correlations r_j between X_j and Y, and the correlation matrix R_X of the covariates:

1. Set the active set, $A = \emptyset$, and the sign vector $\boldsymbol{s}_A = \emptyset$.

2. Determine $m = \text{argmax}|r_j|$, and $s_m = \text{sign}\{r_m\}$. Let $r = s_m r_m$.

3. Put $A \leftarrow A \cup \{m\}$, and $\boldsymbol{s}_A \leftarrow \boldsymbol{s}_A \cup \{s_m\}$.

4. Calculate $a = [\mathbf{1}'_A (D_A R_A D_A)^{-1} \mathbf{1}_A]^{-1/2}$, where $\mathbf{1}_A$ is a vector of 1's, $D_A = \text{diag}(\boldsymbol{s}_A)$, and R_A is the submatrix of R_X corresponding to the active variables. Calculate $\boldsymbol{w}_A = a\ (D_A R_A D_A)^{-1} \mathbf{1}_A$, and $a_j = (D_A \boldsymbol{r}_{jA})' \boldsymbol{w}_A$, for $j \in A^c$, where $\boldsymbol{r}_{jA}$ is the vector of correlations between X_j and the active variables. (Note that, when there is only one active covariate X_m, the above quantities simplify to $a = 1$, $w = 1$, and $a_j = r_{jm}$.)

5. For $j \in A^c$, calculate $\gamma_j^+ = (r - r_j)/(a - a_j)$, and $\gamma_j^- = (r + r_j)/(a + a_j)$, and let $\gamma_j = \min(\gamma_j^+, \gamma_j^-)$. Determine $\gamma = \min\{\gamma_j,\ j \in A^c\}$, and m, the index corresponding to the minimum $\gamma = \gamma_m$. If $\gamma_m = \gamma_m^+$, set $s_m = +1$. Otherwise, set $s_m = -1$. Modify $r \leftarrow r - \gamma a$, and $r_j \leftarrow r_j - \gamma a_j$, for $j \in A^c$.

6. Repeat steps 3, 4 and 5.

3 ROBUST LARS

From the results in Section 2, it is not surprising to see that LARS is sensitive to contamination in the data. To illustrate this, we use a dataset on executives obtained from Mendenhall and Sincich (2003). 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. LARS sequences the covariates in the following order: $(1, 3, 4, 2, 5, 6, 9, 8, 10, 7)$. We contaminate the data by replacing one small value of predictor 1 (less than 5) by the large value 100. When LARS is applied to the contaminated data, we obtain the following completely different sequence of predictors: $(\mathbf{7}, 3, \mathbf{2}, \mathbf{4}, 5, \mathbf{1}, \mathbf{10}, \mathbf{6}, \mathbf{8}, \mathbf{9})$. Predictor 7, which was selected last (10th) in the clean data, now enters the model first. The position of predictor 1 changes from first to sixth. Predictors 2 and 4 interchange their places. Thus, changing a single number in the data set completely changes the predictor sequence, which illustrates the sensitivity of LARS to contamination.

We now introduce two approaches to robustify the LARS procedure which we call the *plug-in* and *cleaning* approaches respectively.

3.1 Robust Plug-in

The plug-in approach consists of replacing the non-robust building blocks of LARS (mean, variance and correlation) by robust counterparts. The choices of fast computable robust center and scale measures are straightforward: median (med) and median absolute deviation (mad). Unfortunately, good available robust correlation estimators are computed from the d-dimensional data and therefore are very time consuming (see Rousseeuw and Leroy 1987). Robust pairwise approaches (see Huber 1981) are not affine equivariant and therefore are sensitive to two-dimensional outliers. One solution is to use robust correlations derived from a pairwise affine equivariant covariance estimator. A computationally efficient choice is a bivariate M-estimator as defined by Maronna (1976). Alternatively, a bivariate correlation estimator can be computed from bivariate Windsorized data. Both methods will be explained in detail below.

3.1.1 M Plug-in

Maronna's bivariate M-estimator of the location vector $\boldsymbol{t}$ and scatter matrix $\boldsymbol{V}$ is a highly robust and computationally efficient estimator. It is defined as the solution of the system of equations:

$$\begin{aligned} \frac{1}{n}\sum_i u_1(d_i)(\boldsymbol{x}_i - \boldsymbol{t}) &= \mathbf{0}, \\ \frac{1}{n}\sum_i u_2(d_i^2)(\boldsymbol{x}_i - \boldsymbol{t})(\boldsymbol{x}_i - \boldsymbol{t})' &= \boldsymbol{V}, \end{aligned}$$

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 satisfying a set of general assumptions. The estimator is affine equivariant and has breakdown point 1/2 in two dimensions. 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 χ^2_2 distribution. The bivariate correlations are then

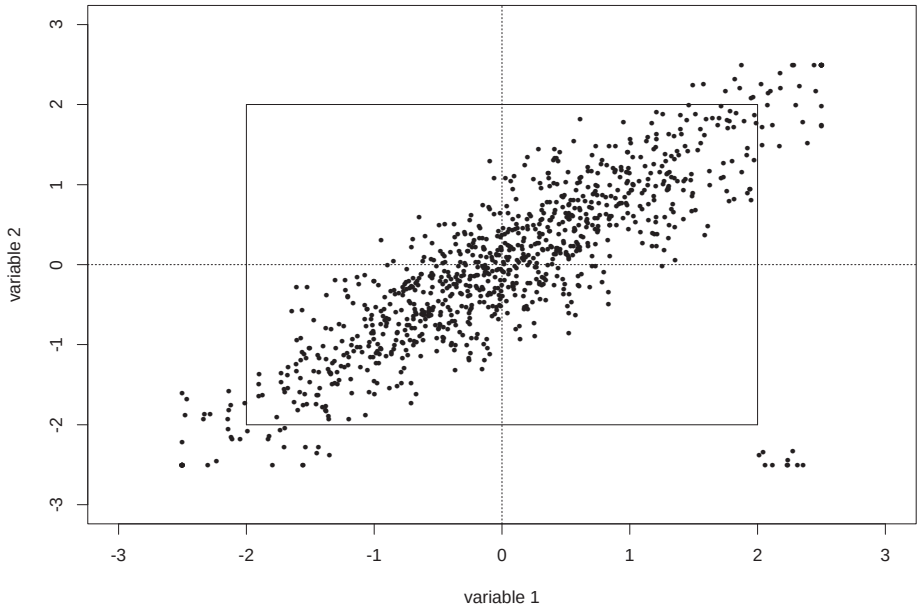


Figure 1: Limitation of separate univariate Windsorizations ($c = 2$).

ensembled to form a $d \times d$ correlation matrix R. Finally, LARS is applied to this robust correlation matrix. We call this the **M plug-in** method.

3.1.2 W Plug-in

For very large, high-dimensional data we need an even faster robust correlation estimator. Huber (1981) introduced the idea of one-dimensional Windsorization of the data, and suggested that classical correlation coefficients be calculated from the transformed data. Alqallaf, Konis, Martin, and Zamar (2002) re-examined this approach for the estimation of individual elements of a large-dimension correlation matrix. For n univariate observations x_1, x_2 $\ldots$, x_n, the transformation is given by $u_i = \psi_c((x_i - \text{med}(x_i))/\text{mad}(x_i)), \;\; i = 1, 2, \ldots, n$, where the Huber score function $\psi_c(x)$ is defined as $\psi_c(x) = \min\{\max\{-c, x\}, c\}$, with c a tuning constant chosen by the user, e.g., $c = 2$ or $c = 2.5$. This one-dimensional Windsorization approach is very fast to compute but unfortunately it does not take into account the orientation of the bivariate data. It

merely brings the outlying observations to the boundary of a $2c \times 2c$ square, as shown in Figure 1. This plot clearly shows that the univariate approach does not resolve the effect of the obvious outliers at the bottom right which are shrunken to the corner $(2, -2)$, and thus are left almost unchanged.

To remedy this problem, we propose a *bivariate Windsorization* of the data based on an initial tolerance ellipse for the majority of the data. Outliers are shrunken to the border of this ellipse by using the bivariate transformation $\boldsymbol{u} = \min(\sqrt{c/D(\boldsymbol{x})},\ 1)\ \boldsymbol{x}$ with $\boldsymbol{x} = (x_1, x_2)^t$. Here $D(\boldsymbol{x})$ is the Mahalanobis distance based on an initial bivariate correlation matrix R_0. For the tuning constant c we used $c = 5.99$, the 95% quantile of the χ^2_2 distribution. We call this the **W plug-in** method. The choice of R_0 will be discussed below.

Figure 2 shows bivariate Windsorizations for both the complete data set of Figure 1 and the data set excluding the outliers. The ellipse for the contaminated data is only slightly larger than that for the clean data. By using bivariate Windsorization the outliers are shrunken to the boundary of the larger ellipsoid.

The initial correlation estimate. Choosing an appropriate initial correlation matrix R_0 is an essential part of bivariate Windsorization. For computational simplicity we can choose the estimate based on univariate Windsorization explained above. However, we propose an adjusted Windsorization method that is more resistant to bivariate outliers. This method uses two tuning constants. A tuning constant c_1 for the two quadrants that contain the majority of the standardized data and a smaller tuning constant c_2 for the other two quadrants. For example, c_1 is taken equal to 2 or 2.5 as before and $c_2 = hc_1$ where $h = n_2/n_1$ with n_1 the number of observations in the major quadrants and $n_2 = n - n_1$. We use $c_1 = 2$ in this paper.

Figure 3 shows how the adjusted Windsorization deals with bivariate outliers, which are now shrunken to the boundary of the smaller square. Thus, adjusted-Windsorization handles bivariate outliers much better than univariate Windsorization. The initial correlation matrix R_0 is obtained by computing the classical correlation matrix of the adjusted Windsorized data.

Note that the correlations based on both univariate- and adjusted-Windsorized data

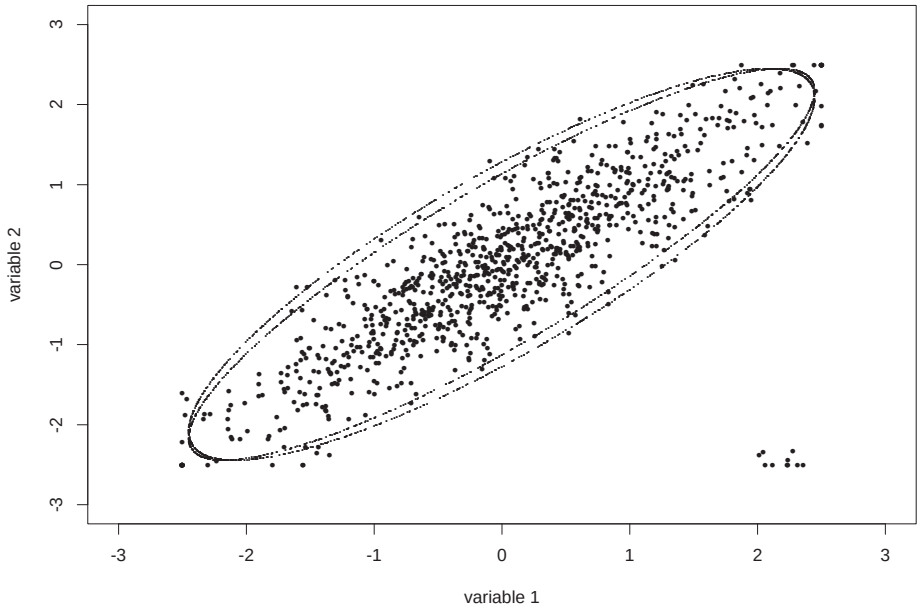


Figure 2: Bivariate Windsorizations for clean and contaminated data.

can be computed in $\mathcal{O}(n \log n)$ time. The adjusted-Windsorized estimate takes slightly more time for a particular n, but is much more accurate in the presence of bivariate outliers as shown above. Bivariate-Windsorized estimate and Maronna's M-estimate also require $\mathcal{O}(n \log n)$ time, but Maronna's M-estimate has a larger multiplication factor depending on the number of iterations required. Thus for large n, the bivariate-Windsorized estimate is much faster to compute than Maronna's M-estimate. Figure 4 shows for each of the four correlation estimates the mean cpu times in seconds (based on 100 replicates) for 5 different sample sizes: 10000, 20000, 30000, 40000 and 50000. These results confirm that the bivariate-Windsorized estimate is faster to compute than Maronna's M-estimate and the difference increases with sample size. Numerical results (not presented here) showed that the bivariate-Windsorized estimate is almost as accurate as Maronna's M-estimate also in the presence of contamination. Note that both the univariate Windsorized and adjusted Windsorized correlations are very fast to compute.

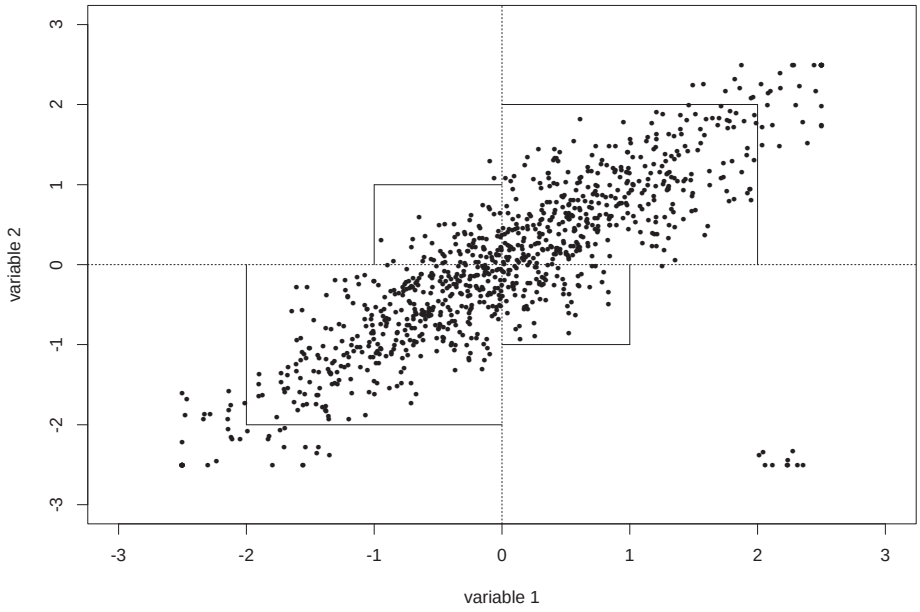


Figure 3: Adjusted Windsorization (for initial estimate r_0) with $c = 2$.

3.2 Robust Data Cleaning

If the dimension d is not extremely large, an alternative approach to robustifying LARS is to apply it on cleaned data. For example, each standardized d-dimensional data point $\boldsymbol{x} = (x_1, \ldots, x_d)^t$ can be replaced by its Windsorized counterpart $\boldsymbol{u} = \min(\sqrt{c/D(\boldsymbol{x})},\ 1)\ \boldsymbol{x}$ in the d-dimensional space. Here $D(\boldsymbol{x}) = \boldsymbol{x}^t V^{-1} \boldsymbol{x}$, is the Mahalanobis distance of $\boldsymbol{x}$ based on V, a fast computable, robust initial correlation matrix. A reasonable choice for the tuning distance c is $c = \chi^2_d(0.95)$, the 95% quantile of the χ^2_d distribution.

The initial correlation matrix V. The choice of the initial correlation matrix V is an essential part of the Windsorization procedure. Most available high-breakdown, affine-equivariant methods are inappropriate for our purposes because they are too computationally intensive. Therefore, we resort to pairwise approaches, that is methods in which each entry of the correlation matrix is estimated separately (see Alqallaf et al. 2002). As before we will use a bivariate M-estimator or the bivariate windsorized estimator to calculate the correlations in V. The resulting methods are called **M-cleaning**

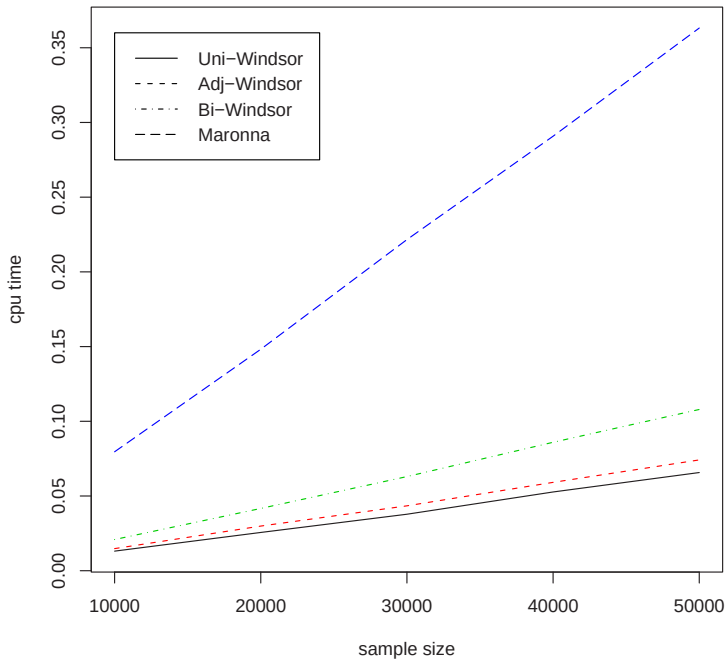


Figure 4: Numerical complexity of different correlation estimates.

and **W-cleaning** respectively.

3.3 Simulations

To investigate the performance and stability of the four proposed methods we consider a simulation study involving a small number of variables. We used the following design (see Ronchetti et al. 1997). The error distributions considered are $(e1)$ standard normal, $(e2)$ 93% from standard normal and 7% from $N(0, 5^2)$, $(e3)$ slash, that is standard normal divided by a uniform on $(0, 1)$, and $(e4)$ 90% from standard normal and 10% from $N(30, 1)$.

Two design matrices are considered: the uniform design for which the columns are generated from a uniform distribution on $(0, 1)$, and the leverage design which is the same as the uniform design except that it contains a leverage point. Six variables are used from which the first three are nonzero and in order of importance. The true regression

Table 1: Percentage of correct sequences.

	Uniform				Leverage			
Method	$e1$	$e2$	$e3$	$e4$	$e1$	$e2$	$e3$	$e4$
LARS E	97	86	11	8	0	1	1	2
LARS G	100	89	26	24	0	2	5	7
M plug-in E	95	97	53	87	96	96	49	87
M plug-in G	99	99	74	95	99	99	68	95
W plug-in E	96	97	58	78	92	85	46	59
W plug-in G	99	99	77	89	94	86	61	68
M cleaning E	96	98	55	89	96	97	50	87
M cleaning G	99	99	77	97	100	98	73	97
W cleaning E	96	98	54	82	96	94	52	83
W cleaning G	99	99	76	92	98	96	71	92

coefficients for the nonzero variables are 7, 5 and 3, respectively. The sample size equals $n = 60$ and we generated 200 data sets for each setting. We used two performance measures which we call exact (E) and global (G). The exact measure gives the percentage of times a procedure sequences the important variables in front and in their true order. The global measure gives the percentage of times a procedure sequences the important variables in front in any order.

Table 1 shows the simulation results. For error distribution $e1$ (standard normal), the performance of the robust methods is almost as good as that of LARS. For the heavy tailed distributions the robust methods drastically outperform LARS. Overall we see from Table 1 that the plug-in approaches are almost as stable as the computationally more expensive data cleaning approaches. Comparing the M and W approaches for both the plug-in and data cleaning procedures, it is reassuring to see that the computationally faster W approach (see Figure 5 below) is almost as stable as the M approach.

Finally, we compare the computational complexity of the different methods. The standard LARS procedure sequences all d covariates in only $\mathcal{O}(nd^2)$ time. The plug-in

and cleaning procedures based on M-estimators both require $\mathcal{O}((n \log n)d^2)$ time. Based on Windsorization these procedures also require $\mathcal{O}((n \log n)d^2)$ time, but with a much smaller multiplication factor. Moreover, if we are only interested in sequencing the top fraction of a large number of covariates, then the plug-in approach will be much faster than the cleaning approach, because the plug-in approach only calculates the required correlations along the way instead of the 'full' correlation matrix. In this case, the complexity for plug-in methods reduces to $\mathcal{O}((n \log n)dm)$, where m is the number of variables being sequenced.

Figure 5 shows the mean cpu times based on 10 replicates for LARS, W plug-in and M plug-in for different dimensions d with a fixed sample size $n = 2000$. The times required by the cleaning methods are not shown because they were similar to the plug-in times since we sequenced all the covariates. As in Figure 4, we see that the approaches based on M-estimators are more time consuming than the Windsorization approaches. The difference increases fast with dimension.

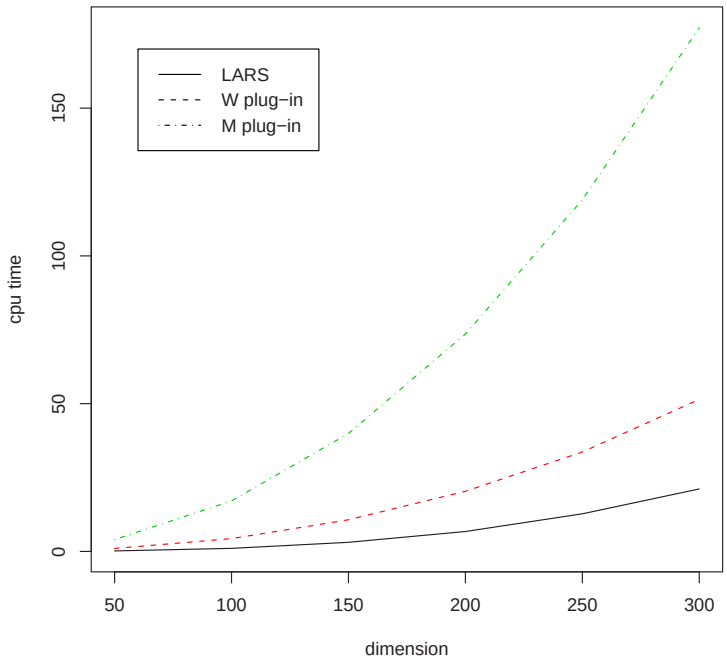


Figure 5: Complexity of the different techniques.

The cleaning approaches perform slightly better than the plug-in approaches when the number of variables is relatively small, and much smaller than the number of cases (see Table 1). However, plug-in approaches are less time-consuming when only a part of the predictors are sequenced. Since W plug-in has a reasonable performance compared to the other methods and has favorable computing times, this method is to be preferred for large, high-dimensional datasets. The performance of W plug-in will be studied further in the next sections and we will call this method *robust LARS* from now on.

4 SIZE OF THE REDUCED SET

To obtain a good final model, it is important to choose an appropriate value of m, the size of the reduced set that is kept from the sequencing step and will be used for segmentation. The reduced set should be large enough to include most of the important variables, but not so large as to make the segmentation step impractical. Several factors can be important when determining the size m such as d, the total number of variables, the sample size n, the unknown number of non-zero variables in the optimal model, the correlation structure of the covariates, and of course also time and feasibility of the segmentation step. For example, for high-dimensional datasets, including only 1% of the variables in the reduced set may make the segmentation step already infeasible.

To investigate what values of m are appropriate, we carry out a simulation study similar to Frank and Friedman (1993). We consider 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

$$y = 5l_1 + 4l_2 + 3l_3 + \epsilon = \text{Signal} + \epsilon, \tag{1}$$

where $l_i \sim N(0, 1)$, and ϵ is a normal error not related to the latent variables. The variance of ϵ is chosen such that the signal-to-noise ratio equals 2, that is $\text{Var}(\epsilon) = 50/4$. The total number of variables equals $d = 100$. A small number $a = 9$ or $a = 15$ of these covariates are linked with the latent variables in (1). These covariates are divided in 3 equal groups, with each group related to exactly one of the latent variables by the

following relation

$$x_i = l_j + \delta_i,$$

where $\delta_i \sim N(0, \sigma_i^2)$. The value of σ_i^2 determines the correlation structure of the nonzero covariates. We consider 3 correlation structures: "high correlation" case (a true correlation of 0.9 between the covariates generated with the same latent variable), "moderate correlation" case (a true correlation of 0.5), and "no correlation" case. For each simulation we generated 100 samples of size $n = 150$. Outliers were added by giving the noise term a large positive mean (asymmetric error). We considered four different levels of contamination: $0, 5, 10$ and 20%.

For the high-correlation and moderate-correlation cases, though "a" of the covariates are linked to the response y through the latent variables, it is not clear which of these covariates should be considered important for explaining y. Even when the true pairwise correlations of the covariates are zero (no-correlation case), the "best" model not necessarily includes all of the a non-zero coefficients because of the bias-variance trade-off. Therefore, for each simulated dataset we first find the "best" model among all possible subsets of the non-zero covariates that has the minimum prediction error estimated by 5-fold cross-validation.

For each simulated dataset, we determine the "recall proportion", i.e., the proportion of important variables (in the sense that they are in the "best" model by cross-validation) that are captured (recalled) by LARS/robust LARS for a fixed size of the reduced sequence.

For $a = 9$, Figure 6 plots the average recall proportion against the size of the reduced set for the three correlation structures. In each plot, the 4 curves with the same line type correspond to the 4 levels of contamination, higher curves correspond to lower levels of contamination. These plots show that, for each correlation structure considered, we can capture the important variables if the percentage of variables in the reduced set is 9 or 10. Robust LARS performs as good as LARS for clean data, and much better than LARS for contaminated data.

Figure 7 plots the average recall proportion against the size of the reduced set for the moderate-correlation case with $a = 15$. This plot can be compared with Figure 6b

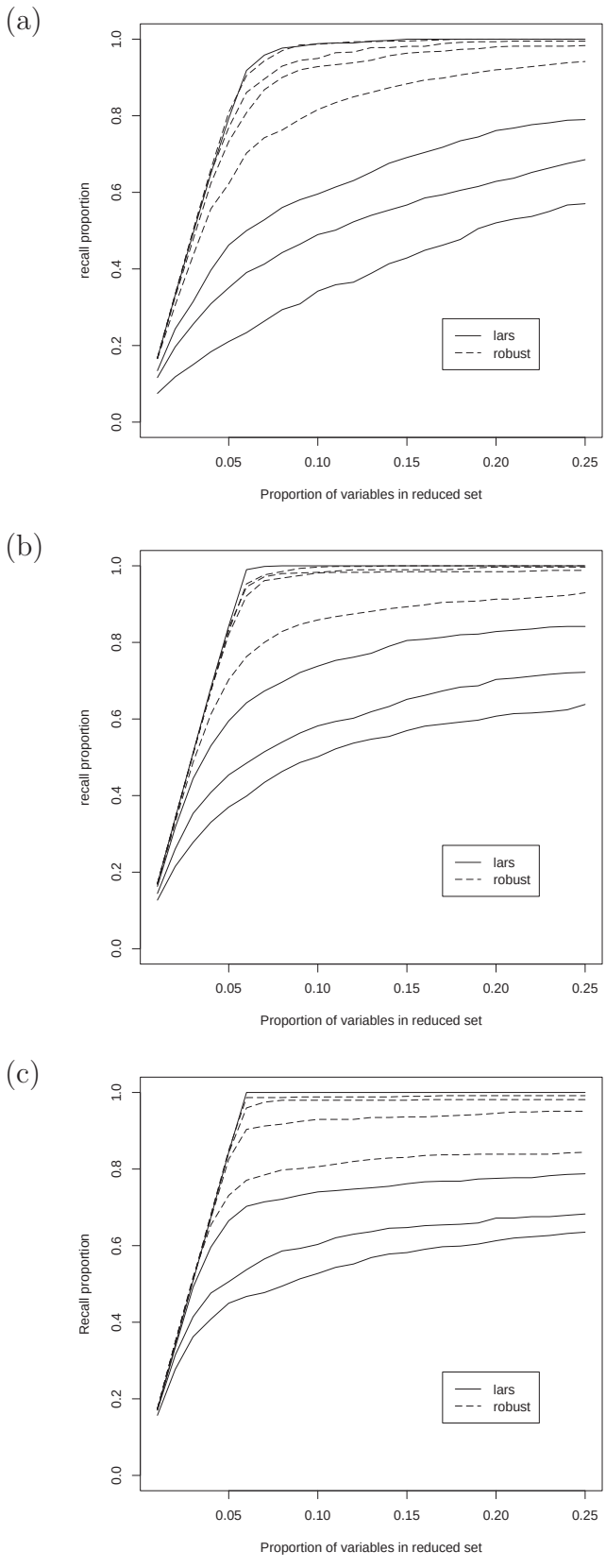


Figure 6: Recall curves for $a = 9$ and (a) no correlation; (b) low correlation; (c) high correlation.

to see how the increase in the number of nonzero variables affects the recall proportions. In both cases, we observe that the average recall proportions stop increasing even before the size m of the reduced set exceeds the number a of non-zero variables.

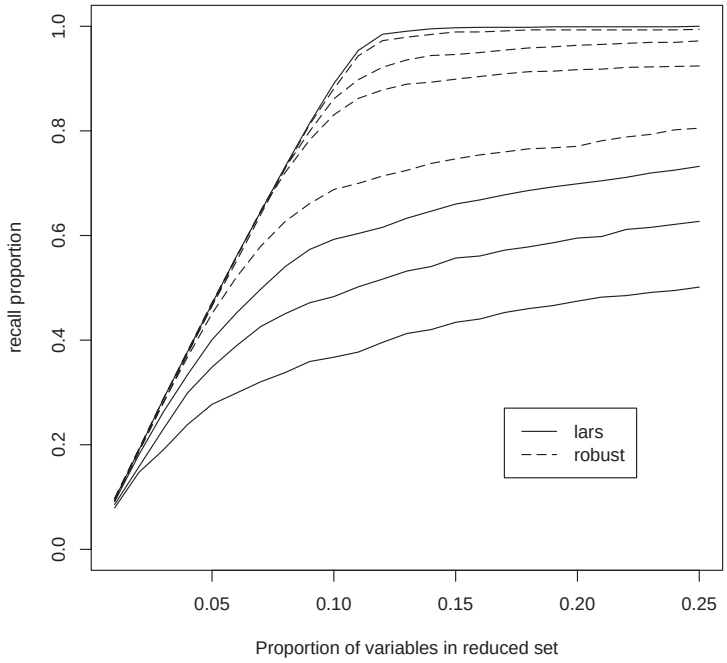


Figure 7: Recall curves for $a = 15$ and moderate correlation.

5 BOOTSTRAPPED SEQUENCING

To obtain more stable and reliable results we can combine robust LARS with bootstrap. Therefore, we generate a number B of bootstrap samples from the dataset, and use (robust) LARS to obtain the corresponding sequence of covariates for each of these bootstrap samples. Each sequence ranks the covariates from 1 to d. For each covariate we can take the average over these B ranks, and the m covariates with the smallest average ranks then form the reduced set.

When resampling from a high-dimensional dataset (compared to the sample size, e.g.

$n = 150, d = 100$) the probability of obtaining singular samples becomes very high. Note that even the original sample may already be singular or the dimension d of the data may exceed the sample size. In these cases it will be impossible to sequence all covariates. We can easily overcome this problem by sequencing only the first $m_0 < d$ of the covariates for each bootstrap sample, where preferably $m_0 \geq m$. We then rank the covariates according to the number of times (out of B) they are actually sequenced. When ties occur, the order of the covariates is determined according to the average rank in the sequences. In our simulations, we generated $B = 100$ bootstrap samples from each of the 100 simulated datasets. We sequenced the first 25 covariates in each bootstrap sample.

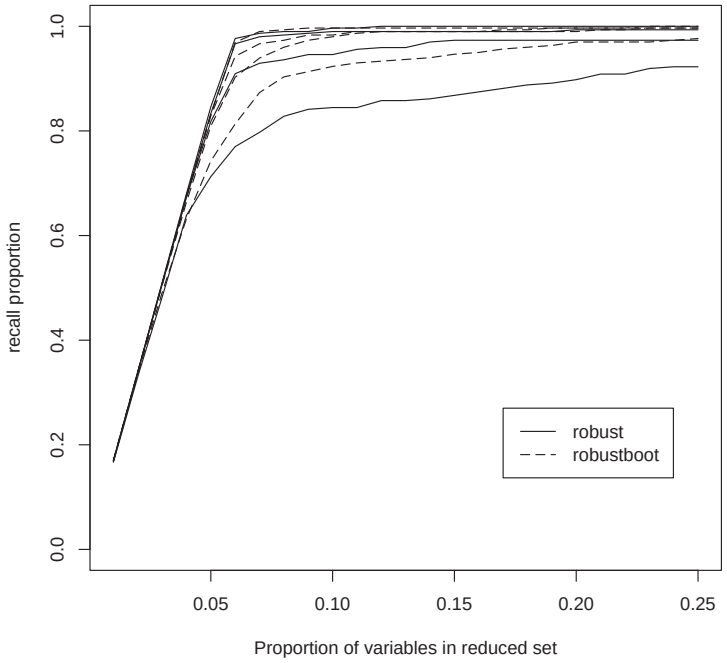
Figure 8 shows the recall curves obtained by robust LARS (solid lines) and robust bootstrapped LARS (dotted lines) for covariates with moderate correlation. The recall curves obtained by robust bootstrapped LARS perform better than the initial robust LARS curves for all levels of contamination, the difference being larger with larger contamination proportions. This confirms that by applying the bootstrap we obtain more stable and reliable results. Even with 20% of contamination, robust bootstrapped LARS with $m = 10$ ($a = 9$) or $m = 15$ ($a = 15$) already yields a recall proportion around 90%.

To investigate what minimum number of bootstrap samples is required to obtain significant improvement over robust LARS, we also tried $B = 10$, 20 and 50 in the above setups. In each case, $B = 10$ and $B = 20$ do not yield much improvement, while with $B = 50$ the results obtained are almost as stable as with $B = 100$.

6 LEARNING CURVES

Although the simulation results in the previous sections suggested that it suffices to select the size of the reduced set equal to or slightly larger than the number of predictors in the final model, we usually have no information about the number of predictors that is needed. Hence, a graphical tool to select the size of the reduced set would be useful. The following plot can be constructed to determine a reasonable size for the reduced set. Starting from a model with only 1 variable (the first one in the sequence), we increase the number of variables according to the sequence obtained and each time fit a robust regression model

(a)



(b)

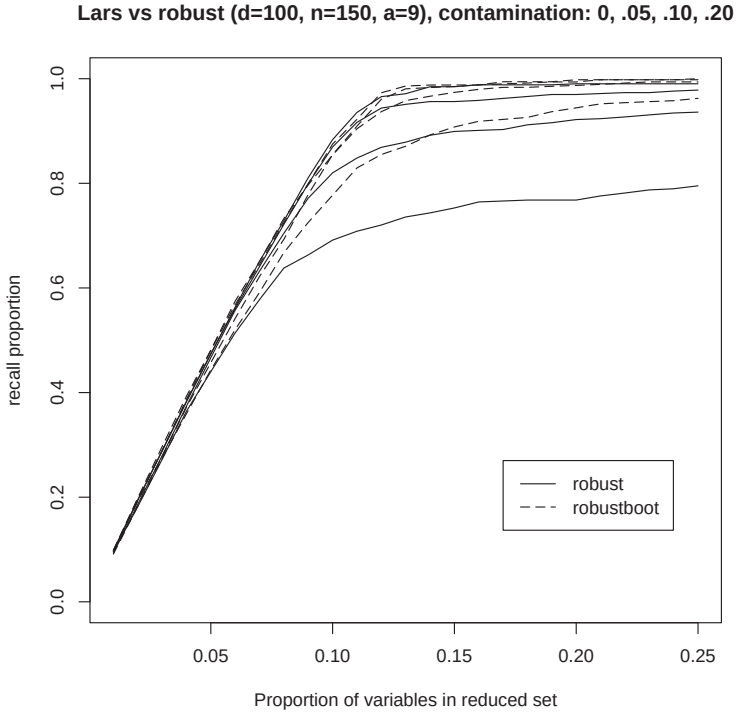


Figure 8: Recall curves for robust LARS and bootstrapped robust LARS for covariates with moderate correlation and (a) $a = 9$; (b) $a = 15$.

to compute a robust R^2 measure such as $R^2 = 1 - \text{Median}(\boldsymbol{e}^2)/\text{MAD}^2(y)$, where $\boldsymbol{e}$ is the vector of residuals from the robust fit. We then plot these robust R^2 values against the number of variables in the model to obtain a *learning curve*. The size of the reduced set can be selected as the point where the learning curve does not have a considerable slope anymore.

A problem that can occur with a robust R^2 measure is that, unlike its classical counterpart, it is not always a nondecreasing function of the number of covariates. This can be resolved as follows. If the robust R^2 at any step is smaller than that of the preceding step, then fit a robust simple linear regression of the residuals from the preceding step on the newly selected covariate. The residuals obtained from this fit can be used to compute another robust R^2 value. We then use the larger of the two values.

To investigate the performance of learning curves, we consider a dataset on air pollution and mortality in 60 Metropolitan areas in the United States. The response variable is the age-adjusted mortality. There are 14 potential predictors, numbered from 1 to 14. Since row 21 contains 2 missing values, we drop this observation from the data. Based on robust data exploration we identified 4 clear outliers that correspond to the four metropolitan areas in California. We applied 5-fold cross-validation (CV) to this dataset without the four outliers, and obtained the "best model" that has the following 7 covariates: $(2, 3, 4, 6, 7, 10, 13)$.(The order of the covariates in not relevant here.)

Robust bootstrapped LARS applied to this dataset (including the outliers) produced the sequence $(7, 5, 13, 4, 6, 3, 2, 10, 9, 1, 14, 11, 8, 12)$. We used this sequence and fitted Least Median of Squares (Rousseeuw 1984) regressions to obtain the robust R^2 values. Figure 9 shows the corresponding learning curve. This plot suggests a reduced set of size 8. It is encouraging to notice that the reduced set (first 8 covariates in the sequence above) contains all 7 predictors selected in the "best model" obtained by CV.

7 EXAMPLES

In this section we use two real datasets to evaluate the performance of robust (bootstrapped) LARS. The demographic data example further explores the idea of "learning

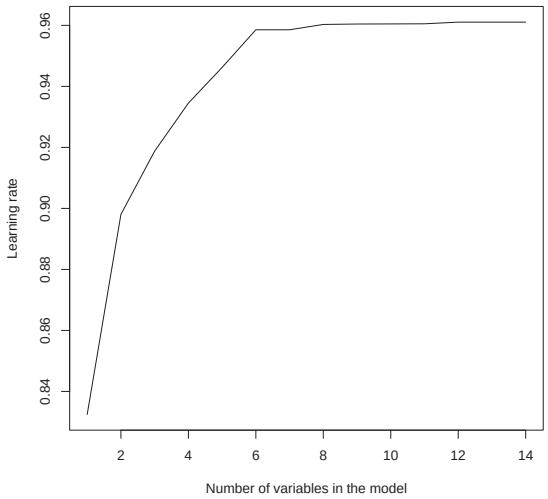


Figure 9: Learning curve for Pollution data.

curves" to choose the size of the reduced set. We then use a large dataset (protein data) to demonstrate the scalability as well as stability of robust LARS.

Demographic data. This dataset contains demographical information on the 50 states of the United States for 1980. The response variable of interest is the murder rate per 100,000 residents. There are 25 predictors which we number from 1 to 25. Exploration of the data using robust estimation and graphical tools revealed one clear outlier. We applied 5-fold CV to this dataset without the outlier, and obtained the "best of 25" model that has the following 15 covariates $(1, 2, 3, 5, 6, 8, 9, 10, 16, 17, 18, 19, 21, 24, 25)$.

Figure 10 shows the learning curve for the Demographic data based on robust bootstrapped LARS. This plot suggests a reduced set of size 12 which include the covariates: $(22, 20, 4, 15, \mathbf{10}, \mathbf{2}, \mathbf{19}, \mathbf{25}, \mathbf{8}, \mathbf{18}, \mathbf{6}, \mathbf{24})$. The boldface numbers correspond to covariates in the sequence that are also in the model obtained by CV. The number of "hits" is 8 out of 12.

We applied 5-fold CV to the clean data using the reduced set of size 12 obtained by robust bootstrapped LARS. The model selected in this case has the following 9 covariates: $(22, 20, 4, 15, 2, 10, 25, 18, 24)$. To compare this "best of 12" model with the "best of 25" model above, we estimated the prediction errors of these two models 1000 times using

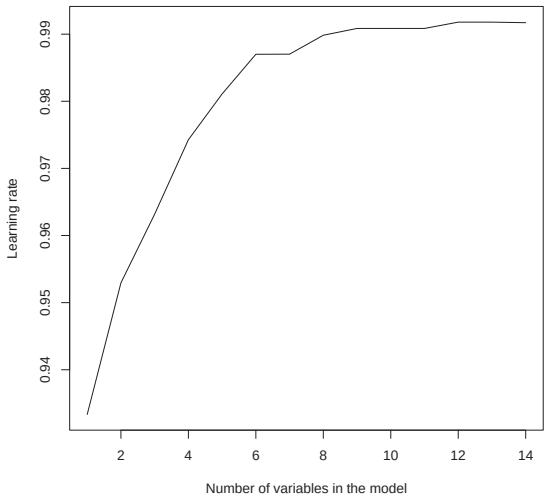


Figure 10: Learning curve for Demographic data.

5-fold CV. The two density curves are shown in Figure 11. The "best of 12" model has a mean error of 204.8 (median error 201.5) while the "best of 25" model has a mean error of 215.9 (median error 202.0). Also, the standard deviations (mads) of the errors are 25.6 (22.7) and 74.6 (31.4), respectively. (Some of the "best of 25" errors are very large and not included in the plot.) Thus, robust bootstrapped LARS gives more stable results in this high-variability dataset. It should be mentioned here that we needed almost 10 days to find the "best of 25" model, while "best of 12" model requires less than 5 minutes including the time needed to sequence the covariates by robust bootstrapped LARS. (CV on m covariates is $2^{(d-m)}$ times faster than CV on d covariates.)

Protein data. This dataset of $n = 145751$ protein sequences was used for the KDD-Cup 2004. Each of the 153 blocks corresponds to a native protein, and each data-point of a particular block is a candidate homologous protein. There are 75 variables in the dataset: the block number (categorical) and 74 measurements of protein features. We replace the categorical variable by block indicator variables, and use the first feature as the response. Though this analysis may not be of particular scientific interest, it will demonstrate the scalability and stability of the robust LARS algorithm.

We used the package R to apply robust LARS to this dataset, and obtained a reduced

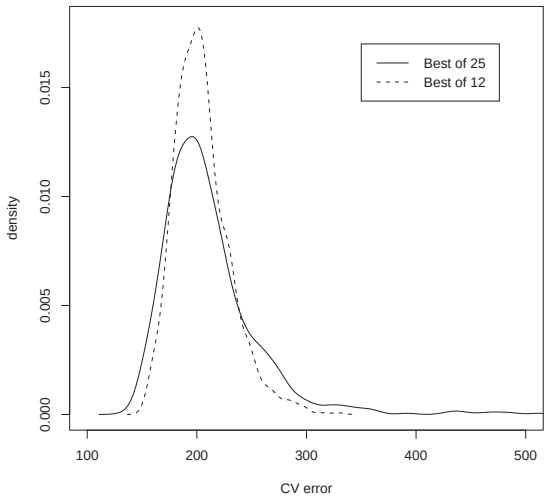


Figure 11: Error densities for two best models for Demographic data.

set of size 25 from $d = 225$ covariates (152 block indicators + 73 features) in only 30 minutes. Given the huge computational burden of other robust variable selection procedures, our algorithm maybe considered extremely suitable for computations of this magnitude.

For a thorough investigation of the performance of robust LARS with this dataset, we select 5 blocks with a total of $n = 4141$ protein sequences. These blocks were chosen because they contain the highest proportions of homologous proteins (and hence the highest proportions of potential outliers). We split the data of each block into two almost equal parts to get a training sample of size $n = 2072$ and a test sample of size $n = 2069$. The number of covariates is $d = 77$, with 4 block indicators (variables $1 - 4$) and 73 features. We apply robust bootstrapped LARS with $B = 100$ bootstrap samples and we sequence the first 25 variables of each bootstrap sample. The resulting learning curve is shown in Figure 12.

This plot suggests that a drastic reduction to a small number of predictors can be performed, e.g. m=5 or m=10. The first 10 predictors found by robust bootstrapped LARS are $(14, 13, 5, 76, 73, 8, 7, 40, 46, 51)$. The covariates in this sequence are almost the same as those obtained with the whole dataset (not shown). The standard LARS

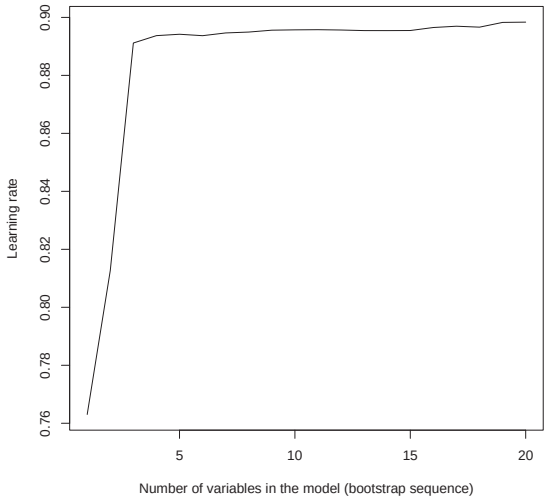


Figure 12: Learning curve for Protein data.

produced the sequence $(14, 13, 5, 8, 7, 76, 18, 65, 2, 46)$. Note that the two sequences are quite different. For example, if we select a model from the first five predictors, then only 3 predictors are contained in both sequences. Using MM-estimators and robust AIC, the best model selected from the first five variables of the robust sequence contains variables $(14, 13, 5, 76)$ while the best model out of the first 10 predictors contains variables $(14, 13, 5, 76, 40)$. Hence only 1 variable is added.

Using classical AIC, the best model selected from the first 5 variables of the LARS sequence contains variables $(14, 13, 5, 8)$. Variable 76 of the corresponding robust model is replaced by Variable 8. The best model from the first 10 predictors contains variables $(14, 13, 5, 8, 76, 2)$. Note that 2 variables are added to the list compared to 1 variable in the robust case.

We fitted the 4 best models using the training data, and then used them to predict the test data outcomes. The 1%, 5% and 10% trimmed means of prediction errors for the smaller robust (classical) model are : 114.92 (117.49), 92.77 (95.66) and 74.82 (78.19), respectively. The corresponding quantities for the larger robust (classical) model are: 114.37 (115.46), 92.43 (94.84) and 74.34 (76.50), respectively. Notice that the robust models always outperform the classical models.

8 CONCLUSIONS

LARS is a very effective, time-efficient model building tool, but is not resistant to outliers. We introduced two different approaches to construct robust versions of the LARS technique. The plug-in approach replaces the classical Pearson correlations in LARS by easily computable robust correlation estimates. The cleaning approach first transforms of dataset by shrinking the outliers towards the bulk of the data, and then applies LARS on the transformed data. Both approaches use robust pairwise correlations estimates which can be computed efficiently using bivariate-Windsorization or bivariate M-estimates.

The data cleaning approach is limited in use because the sample size needs to be (much) larger than the number of candidate predictors to ensure that the resulting correlation matrix will be positive definite. Moreover, the data cleaning approach is more time consuming than the plug-in approach, certainly when only part of the predictors is being sequenced. Since the plug-in approach has good performance, is faster to compute and more widely applicable, we prefer this method. Comparing bivariate M-estimates with bivariate Windsorization we showed that the latter is faster to compute with important time differences when the number of candidate predictors becomes high.

We propose to use the robust LARS technique to sequence the candidate predictors and as such identify a reduced set of most promising predictors from which a more refined model can be selected in a second segmentation step. We recommend combining W plug-in with bootstrap to obtain more stable and reliable results. The reduced sets obtained by robust bootstrapped LARS contain more of the important covariates than the reduced sets obtained by initial robust LARS.

It is important to select the number of predictors to use for the second step. This number is a trade-off between success-rate, that is the number of important predictors captured in the reduced set, and feasibility of the segmentation step. Our simulation study indicated that the reduced set can have size comparable to the actual number of relevant candidate predictors. However, this number is usually unknown. To still get an idea about an appropriate size for the reduced set we introduced a learning curve that plots robust R^2 values versus dimension. An appropriate size can be selected as the dimension corresponding to the point where the curve starts to level off.

APPENDIX: TECHNICAL DERIVATIONS

A. Determination of γ for One Active Covariate

Assume that the first selected covariate is $+X_m$. The current prediction $\hat{\boldsymbol{\mu}} \leftarrow \mathbf{0}$ should be modified as

$$\hat{\boldsymbol{\mu}} \leftarrow \gamma\, X_m.$$

The distance γ should be such that the modified residual $(y - \hat{\boldsymbol{\mu}})$ will have equal correlation with $+X_m$ and another signed covariate X_j. We have

$$\text{cor}(y - \hat{\boldsymbol{\mu}}, X_m) = \frac{X_m'(y - \gamma X_m)/n}{\text{SD}(y - \gamma X_m)} = \frac{r - \gamma}{\text{SD}(y - \gamma X_m)}, \tag{A.1}$$

and

$$\text{cor}(y - \hat{\boldsymbol{\mu}}, +X_j) = \frac{X_j'(y - \gamma X_m)/n}{\text{SD}(y - \gamma X_m)} = \frac{r_j - \gamma r_{jm}}{\text{SD}(y - \gamma X_m)}. \tag{A.2}$$

Equating (A.1) to (A.2), we have

$$\gamma(+X_j) = \frac{r - r_j}{1 - r_{jm}}. \tag{A.3}$$

Similarly, equating (A.1) with the correlation of modified residual and $-X_j$ we have

$$\gamma(-X_j) = \frac{r + r_j}{1 + r_{jm}}. \tag{A.4}$$

We should take the minimum of (A.3) and (A.4) and minimum over all inactive (not yet selected) j. The signed covariate that will enter the model at this point is determined alongwith.

B. Quantities Related to Equiangular Vector B_A

Here, A is the set of 'active' subscripts. Let $X_A = (\cdots\ s_l X_l\ \cdots),\ \ l \in A$, where s_l is the sign of X_l as it enters the model. The standardized equiangular vector B_A is obtained using the following three conditions. B_A is a linear combination of the active signed predictors.

$$B_A = X_A\, \boldsymbol{w}_A\ , \ \text{ where } \boldsymbol{w}_A \text{ is a vector of weights.} \tag{A.5}$$

B_A has unit variance:

$$\frac{1}{n} B_A' B_A = 1. \tag{A.6}$$

B_A has equal correlation (a, say) with each of the active predictors. Since the covariates and B_A are standardized,

$$\frac{1}{n} X_A' B_A = a\ \mathbf{1}_A\ ,\ \ \mathbf{1}_A \text{ is a vector of 1's.} \tag{A.7}$$

Using equation (A.5) in equation (A.6), we have

$$\frac{1}{n} \boldsymbol{w}_A' X_A' X_A \boldsymbol{w}_A = 1,$$

so that

$$\boldsymbol{w}_A' R_A^{(s)} \boldsymbol{w}_A = 1, \tag{A.8}$$

where $R_A^{(s)}$ is the correlation matrix of the active signed variables. Using (A.5) in (A.7), we have

$$R_A^{(s)} w_A = a\mathbf{1}_A,$$

so that the weight vector w_A can be expressed as

$$w_A = a\ (R_A^{(s)})^{-1} \mathbf{1}_A.$$

Let R_A be the correlation matrix the unsigned active covariates, i.e., R_A is a submatrix of R_X. Let $\boldsymbol{s}_A$ be the vector of signs of the active covariates (we get the sign of each covariate as it enters the model). We have

$$w_A = a\ (D_A R_A D_A)^{-1} \mathbf{1}_A, \tag{A.9}$$

where D_A is the diagonal matrix whose diagonal elements are the elements of s_A. Finally, using equation (A.9) in equation (A.8), we get

$$a = [\mathbf{1}_A' (D_A R_A D_A)^{-1} \mathbf{1}_A]^{-1/2}. \tag{A.10}$$

The correlation of an inactive covariate X_j with B_A, denoted by a_j, can be expressed as follows

$$a_j = \frac{1}{n} X_j' B_A = \frac{1}{n} X_j' X_A \boldsymbol{w}_A = (D_A \boldsymbol{r}_{jA})' \boldsymbol{w}_A, \tag{A.11}$$

where $\boldsymbol{r}_{jA}$ is the vector of correlation coefficients between the inactive covariate X_j and the (unsigned) selected covariates. Thus, we need only (a part of) the correlation matrix of the data (not the observations themselves) to determine the above quantities.

C. Determination of γ for Two or More Active Covariates

Let us update $r \leftarrow (r - \gamma)$, see (A.1), and $r_j \leftarrow (r_j - \gamma r_{jm})$, see (A.2).

The correlation of an active covariate with the 'current' residual $y - \hat{\boldsymbol{\mu}}$ is $r/\mathrm{SD}(y - \hat{\boldsymbol{\mu}})$, and the correlation of the active covariate with the current equiangular vector B_A is 'a'. Therefore, the correlation between an active covariate and the 'modified' residual $(y - \hat{\boldsymbol{\mu}} - \gamma_A B_A)$ is

$$\frac{r - \gamma_A\, a}{\mathrm{SD}(y - \hat{\boldsymbol{\mu}} - \gamma_A B_A)}.$$

An inactive covariate $+X_j$, $j \in A^c$, has correlation $r_j/\mathrm{SD}(y - \hat{\boldsymbol{\mu}})$ with the 'current' residual, and it has correlation a_j with B_A. Therefore, the correlation between $+X_j$, $j \in A^c$, and the 'modified' residual is

$$\frac{r_j - \gamma_A\, a_j}{\mathrm{SD}(y - \hat{\boldsymbol{\mu}} - \gamma_A B_A)}.$$

Equating the above two quantities, we get

$$\gamma_A(+X_j) = (r - r_j)/(a - a_j). \tag{A.12}$$

Similarly,

$$\gamma_A(-X_j) = (r + r_j)/(a + a_j). \tag{A.13}$$

We have to choose the minimum possible γ_A over all inactive covariates. Note that when A has only one covariate, (A.12) and (A.13) reduce to (A.3) and (A.4), respectively.

REFERENCES

Alqallaf, F. A., Konis, K. P., Martin, R. D., and Zamar, R. H. (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.

Efron, B. E., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," *The Annals of Statistics*, 32, 407-451.

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

Hastie, T., Tibshirani, R., and Friedman, J. (2001), *The Elements of Statistical Learning*, New York: Springer-Verlag.

Huber, P. J. (1981), *Robust Statistics*, New York: Wiley.

Maronna, R. A. (1976), "Robust M-estimators of Multivariate Location and Scatter," *The Annals of Statistics*, 4, 51-67.

Mendenhall, W., and Sincich, T. (2003), *A Second Course in Statistics: Regression Analysis* (6th ed.), New Jersey: Pearson Education, Inc.

Morgenthaler, S., Welsch, R. E., and Zenide, A. (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): Birkhäuser-Verlag.

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

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

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

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

Rousseeuw, P. J., and Leroy, A.M. (1987), *Robust Regression and Outlier Detection*, New York: Wiley-Interscience.

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

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Ser. B, 58, 267-288.

Weisberg, S. (1985), *Applied Linear Regression* (2nd ed.), New York: Wiley-Interscience.

Yohai, V. J. (1997), "A New Robust Model Selection Criterion for Linear Models: RFPE," unpublished manuscript.