Robust Variable Selection Using Least Angle Regression and Elemental Set Sampling

Lauren McCann and Roy Welsch*

Sloan School of Management, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Room E53-383, Cambridge, MA 02139

Abstract

In this paper we address the problem of selecting variables or features in a regression model in the presence of both additive (vertical) and leverage outliers. Since variable selection and the detection of anomalous data are not separable problems, we focus on methods that select variables and outliers simultaneously. For selection, we use the fast forward selection algorithm, LARS, which is not robust. To achieve robustness to additive outliers, we append a dummy variable identity matrix to the design matrix and allow both real variables and additive outliers to be in the selection set. For leverage outliers, we use these selection methods on samples of elemental sets in a manner similar to that used in high breakdown robust estimation. Bagging is then used to stabilize the selection results. We conclude by comparing our results to several other selection methods of varying computational complexity and robustness and discussing the extension of our methods to situations where the number of variables exceeds the number of observations.

Keywords: Robust regression; Variable selection; LARS; Outliers; Elemental sets

1. Introduction

In many areas where regression is used, the emphasis is on model and variable (feature) selection. Model selection generally focuses on the ability to predict well (preferably on

* Corresponding author. Tel.: +1-617-253-6601
E-mail addresses: futed@mit.edu, rwelsch@mit.edu

out-of-sample data) with less emphasis on getting the variables in order of importance or exactly right. Variable or feature selection places more emphasis on finding the correct variables. Of course, variable selection is one way to accomplish model selection.

We are concerned with selection methods that are robust to outliers. There are a number of approaches in the literature such as Ronchetti et al. (1997), Morgenthaler et al. (2004), and Müller and Welsh (2005). These papers contain references to a number of others. A Bayesian approach is discussed in Raftery et al. (1997). The available methods are, generally, computationally expensive, lack high breakdown properties, and run into problems if there are more variables than observations.

Another difficulty arises because robustness and variable selection are not necessarily exchangeable, i.e., selection may affect what is considered to be an outlier and vice versa. Since there are a large number of fairly fast selection algorithms available, it is tempting to select first and ask robustness questions later. There are also a variety of fairly expensive high breakdown robust estimators available. These could be used on the full model, weights between zero and one placed on each observation, and then standard variable selection would be run with these weights fixed.

High breakdown robust regression generally requires some form of sampling (looking for "good" subsets of the data). For a discussion see Rousseeuw and Van Driessen (2000). If we are going to sample, and then select variables on those samples, a fast selection algorithm is required. Least angle regression (LARS) developed by Efron et al. (2004) provides a very fast way to do forward selection. LARS addresses the problem that standard least-squares forward selection can be overly greedy and can also easily provide the LASSO (Tibshirani, 1996) sequence of variables to enter the model. It is, however, not

robust to outliers since it depends on the computation of the pairwise correlation matrix of the explanatory variables. Khan et al. (2005), hereafter sometimes called KVZ, provide several approaches to addressing the selection problem with LARS in the presence of anomalous data. They propose replacing the non-robust correlation matrix by a (fast) robust version. This, however, requires treating outliers before selection, something we would like to avoid as much as possible.

In this paper we use LARS to both select variables and outliers simultaneously by adopting the dummy variable approach contained in Morgenthaler et al. (2004) combined with the elemental set sampling that is essential for high breakdown robust estimation (Rousseeuw and Van Driessen, 2000).

2. Vertical or Additive Outliers

If the regression data contained only additive outliers (called vertical outliers by Rousseeuw and Van Driessen (2000)), then we could start with the usual regression model

$$\mathbf{y} = \mathbf{X}\beta + \boldsymbol{\varepsilon} \qquad (1)$$

with $\mathbf{y}$ $n \times 1$, $\mathbf{X}$ $n \times p$, β $p \times 1$ and $\boldsymbol{\varepsilon}$ $n \times 1$ and append an $n \times n$ identity matrix to $\mathbf{X}$ to form a new model

$$\mathbf{y} = \mathbf{Z}\delta + \boldsymbol{\varepsilon} \qquad (2)$$

where $\mathbf{Z} = \mathbf{X} \mid \mathbf{I}$ is $n \times (p + n)$ and δ is $(p + n) \times 1$. Of course, this problem is under determined and traditional least-squares estimation will not work. However, if we have a variable selection procedure (such as forward selection) that would accommodate more

variables than cases, then we could proceed to simultaneously select variables and outliers. LARS is a reasonable (and fast) candidate, the only limitation being that no more than n out of the $n + p$ possible explanatory and dummy variables can be selected.

If we assume that our data has no more than 50% contamination (outliers), then we will need to select a maximum of $n/2$ additive outliers. Therefore $n/2 + p$ should be less than n, which limits the number of "real" explanatory variables we can consider to $p < n / 2$. The default level of contamination for many robust regression programs (SAS, S+, etc.) is 25% and we shall use that as our default in what follows. This implies $p < 3n/4$. There are ways to relax these restrictions on p, and we comment on this at the end of this paper.

3. Variable Selection

Given the problem (2) with dummy variables appended, LARS will provide a sequence of models of size 1, 2, . . . , q where q is chosen to be less than or equal to n. With 25% contamination q is $p + n/4$ if p is less than $3n/4$. For each of the q models, we obtain LARS estimated regression coefficients. We can also compute least-squares (LS) regression coefficients for each of those models and models with dummy variables included will automatically exclude observations corresponding to the dummy variables selected. One natural approach to selecting variables for the LS fit is to use t-statistics based on $n - (p + n/4) = 3n/4 - p$ degrees of freedom. When the number of degrees of freedom is too small, other approaches are needed and we return to this later on.

We call this algorithm LARSD-T and the parameters to be chosen are the level of contamination (25% in our examples) and the *t*-statistic threshold for deciding if a (real) variable should be retained out of the *q* real and dummy variables selected. We also considered removing any non-significant dummy variables (as determined by the *t*-statistic) and refitting. This is called LARSD-T2.

4. High Breakdown Methods

When the data contained leverage outliers (Rousseeuw and Van Driessen, 2000) appending the dummy variable matrix to **X** will, in general, not provide a high breakdown robust method. If variable selection is not an issue, high breakdown methods such as S or MM estimators begin with a least trimmed squares (LTS) algorithm such as the one discussed in Rousseeuw and Van Driessen (2000). These algorithms depend on the sampling of elemental sets of size *p* (or perhaps a little larger) in order to find a "good" subset of the data that minimizes the LTS fitting criterion.

Since these methods are designed to address both vertical and leverage outliers, there would not appear to be a need for the dummy variables. We would take a sample of size *p* (or larger), use LARS to find *p* models, compute a robust prediction error (median absolute deviations from the median or MAD) on the remaining $n - p$ observations, repeat this process, say 2,000 times, choose the best model based on prediction error, and select the variables contained in that model.

We also considered both sampling and appending dummy variables. This was designed to help in cases where our sample was, say, clean of leverage outliers but still had some additive outliers.

5. Selecting the Variables When Sampling

When sampling cases with a sample size of p, we will have the p models to consider from each sample multiplied by the number of samples we use. Picking the best model based on a robust prediction error is unlikely to lead to the best selection of variables due to sampling variability. We always caution users of C_p, etc. to look at a number of models nearest the lowest C_p or nearest to p. In data-mining applications, models within one standard deviation (based on cross-validated sampling error) and of lower complexity are often considered (Hastie et al., 2001).

We face a similar problem and keep track of the best one percent of models based on the MAD prediction error of the models obtained from LARS and over all samples. The next problem is to select the "best" variables from these models. We adopt the bagging idea of majority rules (Breiman, 1996). If a variable appears in 50% or more of the top one percent of models, we declare the variable to be in the final model.

6. Simulation Results

Since Khan et al. (2005) and other authors have used the basic simulation design of Ronchetti et al. (1997), we also use it here. Four error distributions are considered: N(0,1) (e1), 93% N(0,1) and 7% N(0,5) (e2), slash or N(0,1)/U(0,1) (e3), and 90% N(0,1) and 10% N(30, 1) (e4).

There are two matrices of explanatory variables each containing six variables with sample size of 60. In the first, the variables are all independently drawn from a uniform (0,1) distribution. In the second, two rows are replaced by a leverage points (0,3,3,3,3,3) and (0,5,5,5,5,5). There are five non-zero regression coefficients each with a *t*-statistic of six under model (e1). All columns of data are centered and scaled robustly using the median for location and MAD for scaling. We consider a selected model correct if we got five and only the five non-zero variables in the model. There are other figures of merit we could use and we will consider some in a later section of this paper.

Our results are summarized in Table A. LS-CV is a cross-validated LS selection method due to Shao (1992). BIF-CV is a bounded influence cross-validated selection method due to Ronchetti et al. (1997). LARSD-T and LARSD-T2, described earlier, use a Bonferroni *t*-value for selection with $\alpha = 0.05/6$. LARS-S6-CV and LARS-S8-CV draw 2000 samples of size 6 (or 8) and let LARS select six models (with one up to six variables) on each sample. These six models are also fit using LS. The best (lowest) MAD prediction score from either the LARS or LS coefficients is retained. Checking both the LARS and LS coefficients always gave better results. See Meinshausen (2005) for a theoretical justification of this result. From these 12,000 models, variables were selected as described in the section on variable selection. We also created a variation on LARS-S6-CV and LARS-S8-CV which we call LARSD-S6-CV and LARSD-S8-CV. The letter "D" denotes that the variables available for selection by LARS include dummies as well as the real explanatory variables.

Since the LARS algorithm can also produce LASSO (Tibshirani, 1996) coefficients and models, we also tried this in some cases. Neither the LASSO nor LARS dominated across the error models and leverage situations.

BIF-CV is based on bounded-influence robust estimation which has breakdown bound of order $1 / p$. Since there are just two leverage points and at most 10% contamination, BIF-CV is probably not breaking down except possibly for the slash distribution.

LARSD-T and LARSD-T2 provide results similar to BIF-CV (without all possible subset selection and cross-validation). Ironically, they do better on slash and not as well on e4, which is the additive outlier error model. We are fitting for 25% contamination (15 dummies) and, thus, being overly protective in this case. For slash, we are making better use of our 25% contamination protection, while BIF-CV is probably breaking down in more cases. This advantage seems to go away when we remove insignificant dummy variables as in LARSD-T2 case.

The estimators using samples of size 6 and 8 cannot be expected to do as well in "nice" situations since they are attempting to provide more protection. Since those estimators look for "good" subsets of the data and then perform selection (in our case via LARS), we could improve efficiency on (e1) by the appropriate use of a fast algorithm for S-regression like that proposed by Salibian-Barrera and Yohai (2005). This work is in progress.

We note, however, that going from the minimal sample size of 6 to 8 provides a dramatic boost in our results, especially in the non-leverage cases. The downside is that we would, technically, need more samples to get a "clean" subset of size 8 rather than size 6. For leverage cases, sampling with 6 or 8 is very effective, as it should be. There appears to be little to gain from including dummy variables once sampling is used.

Table A

	Uniform				Leverage			
Method	e1	e2	e3	e4	e5	e6	e7	e8
LS-CV	188	47	0	3	186	44	0	1
BIF-CV	157	160	10	167	169	166	7	172
LARSD-T	178	167	28	152	178	186	41	154
LARSD-T2	185	181	12	150	187	182	9	154
LARS-S6-CV	132	128	8	124	167	143	12	157
LARS-S8-CV	169	156	10	156	182	174	24	175
LARSD-S6-CV	130	122	3	109	167	152	14	155
LARSD-S8-CV	161	149	17	140	172	168	36	168

7. Comparison with KVZ

Khan et al. (2005) use a modification of the Ronchetti et al. (1997) simulation design. KVZ keep the error distributions the same as well as the non-leverage design matrix. The leverage design matrix has just one leverage row (5, 5, 3, 3, 3, 3). There are only three non-zero regression coefficients with values 7, 5, and 3. Following Khan et al. (2005), two performance measures were considered, exact (E) and global (G). The exact measure gives the percentage of times a procedure chooses the non-zero variables first and in their true order (7, 5, 3). The global measure gives the percentage of times a procedure chooses the non-zero variables first, but any order is allowed.

We compare our results with those of Khan et al. (2005) in Table B. The first six rows are taken from their paper. The W stands for Winsorized, the P for plug-in, and the C for cleaning. They also considered an M-estimator approach, but the results are similar and the method is more computationally complex.

Table B

	Uniform				Leverage			
Method	e1	e2	e3	e4	e1	e2	e3	e4
LARSE	97	86	11	8	0	1	1	2
LARSG	100	89	26	24	0	2	5	7
WPE	96	97	58	78	92	85	46	59
WPG	99	99	77	89	94	86	61	68
WCE	96	98	54	82	96	94	52	83
WCG	99	99	76	92	98	96	71	92
LARSD-TE	95	95	61	83	96	95	65	85
LARSD-TG	100	100	80	88	100	99	85	91

We see from Table B that using LARS with the added dummy variables and *t*-statistics for final selection compares quite favorably to the Khan et al. (2005) procedures on problems of this size and level of contamination. Without sacrificing efficiency, we make some gains in the slash and additive outlier cases. Appending dummy variables and using LARS is a simple way to achieve robust selection where leverage outliers are not a major concern. The sparse nature of the dummy variable matrix means that there is little additional computational complexity beyond LARS itself.

8. Conclusion

Although it is wise to be cautious about generalizations from a small set of simulation results, it appears that LARS when coupled with either dummy variables or row sampling, can provide computationally efficient robust selection procedures that are reasonably efficient at the Gaussian model.

When *p* approaches *n* and even exceeds it, as is often the case in bioinformatics feature selection, adding dummy variables or sampling with LARS will not work directly.

Morgenthaler et al. (2004) showed that ridge regression with dummy variables can be an effective robust selection tool but ridge is more costly than LARS. However, ridge regression can be used for $p \geq n$, but leaves many non-zero coefficients. The elastic net ideas of Zou and Hastie (2005), which adds a $p \times p$ diagonal matrix of rows to the $n \times p$ data matrix and then uses LARS (LASSO) to select variables, can be used with appended columns of dummy variables. Both the sparsity of the $p \times p$ appended row matrix and the appended dummy variable matrix is important to computational feasibility. The elastic net also makes sampling feasible with or without dummy variables.

References

Breiman, L., 1996, Bagging predictors, Machine Learning, 24 (2), 123-140.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004, Least Angle Regression, Annals of Statistics, 32, 407-499.

Hastie, T., Tibshirani, R., and Friedman, J., 2001, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York.

Khan, J., Van Aelst, S. and Zamar, R., 2005, Robust linear model selection based on Least Angle Regression, Technical Report, Department of Statistics, University of British Columbia.

Meinshausen, N., 2005, LASSO with Relaxation, Technical Report 129, Seminar füer Statistik, ETH Zürich.

Morgenthaler, S., Welsch, R.E. and Zenide, A., 2004, 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*, Series: Statistics for Industry and Technology, Birkhauser Verlag Basel, Switzerland, 195-206.

Müller, S. and Welsh, A. H., 2005, Outlier Robust Model Selection in Linear Regression, Journal of the American Statistical Association, 100, 1297-1310.

Raftery, A. E., Madigan, D. and Hoeting, J. A., 1997, Bayesian model averaging for linear regression models, Journal of the American Statistical Association, 92, 179-191.

Ronchetti, E., Field, C. and Blanchard, W., 1997, Robust Linear Model Selection by Cross-Validation, Journal of the American Statistical Association, 92, 1017–1032.

Rousseeuw, P. J. and Leroy, A. M., 1987, *Robust Regression and Outlier Detection*, Wiley, New York.

Rousseeuw, P.J. and Van Driessen, K., 2000, An Algorithm for Positive-Breakdown Regression Based on Concentration Steps. In W. Gaul, O. Opitz, and M. Schader (Eds.), *Data Analysis: Scientific Modeling and Practical Application*, Springer-Verlag, New York, 335-346.

Salibian-Barrera, M. and Yohai, V., 2006, A fast algorithm for S-regression estimates, to appear in the Journal of Computational and Graphical Statistics.

Shao, J., 1993, Linear Model Selection by Cross-Validation, Journal of the American Statistical Association 88, 486–494.

Tibshirani, R., 1996, Regression Shrinkage and Selection via the Lasso, Journal of the Royal Statistical Society, Series B, 5, 267–288.

Zou, H. and Hastie, T., 2005, Regularization and variable selection via the elastic net, J.R. Statist. Soc. B, 67, Part 2, 301-320.

CSDA 1/31/06b