Robust sparse regression in high dimensions

Andreas Alfons[1,2,*], Christophe Croux[2], Viktoria Öllerer[2], and Sarah Gelper[1]
[1]Erasmus University Rotterdam, Rotterdam, The Netherlands
[2]KU Leuven, Leuven, Belgium
[*]Corresponding author: Andreas Alfons, e-mail: alfons@ese.eur.nl

Extended abstract

Due to the increasing availability of data sets with a large number of variables, sparse model estimation is a topic of high importance in modern data analysis. Sparse regression allows to improve prediction performance by variance reduction and increase interpretability of the resulting models due to the smaller number of explanatory variables. If the number of explanatory variables is larger than the number of observations, it is not even possible anymore to apply traditional methods such as least squares due to the rank deficiency of the design matrix. Regularization allows to overcome such computational difficulties. Appropriate regularization techniques thereby yield sparse coefficient estimates.

Let $y_1, \ldots, y_n$ be the observations on the response, and let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be the p-dimensional observations on the predictor variables. We consider the linear regression model

$$y_i = \mathbf{x}_i' \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{1}$$

with regression parameter $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ and error terms $\varepsilon_i \sim N(0, \sigma)$. Tibshirani (1996) proposed the least absolute shrinkage and selection operator (lasso), which adds an L_1 penalty on the coefficients to the least squares objective function. Thus the lasso estimate of $\boldsymbol{\beta}$ with regularization parameter λ is defined as

$$\hat{\boldsymbol{\beta}}_{\text{lasso}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n} (y_i - \mathbf{x}_i' \boldsymbol{\beta})^2 + n\lambda \sum_{j=1}^{p} |\beta_j|. \tag{2}$$

The L_1 penalty has the desirable property that some coefficients are shrunken to exactly zero, resulting in sparse model estimates. Other penalties share this property, e.g., the smoothly clipped absolute deviation (SCAD) penalty by Fan and Li (2001).

However, another common problem in applied statistics is the presence of outliers in the data. Since the lasso uses the least squares loss function, it is highly influenced by such outliers. More robust alternatives have therefore been developed in the literature. Most of those are penalized M-estimators (e.g., Rosset and Zhu, 2004; Wang et al., 2007; van de Geer, 2008; Li et al., 2011), which are robust against outliers in the response variable but not against outliers in the predictor space. Robustness against the latter can be achieved by regularizing suitable robust regression methods such as least trimmed squares (LTS; Rousseeuw and Van Driessen, 2006).

By combining the LTS objective function with the L_1 penalty, Alfons et al. (2013) introduced the sparse least trimmed squares estimator, or sparse LTS for short. It is given by

$$\hat{\boldsymbol{\beta}}_{\text{sparseLTS}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{h} \left(\boldsymbol{r}^2(\boldsymbol{\beta})\right)_{i:n} + h\lambda \sum_{j=1}^{p} |\beta_j|, \tag{3}$$

where $(\boldsymbol{r}^2(\boldsymbol{\beta}))_{1:n} \leq \ldots \leq (\boldsymbol{r}^2(\boldsymbol{\beta}))_{n:n}$ are the order statistics of the squared residuals and $h \leq n$. Since the limit case $h = n$ yields the lasso solution, sparse LTS can be interpreted as a trimmed version of the lasso.

Sparse LTS has been shown to perform well with respect to model selection and prediction when the data are contaminated. We perform further numerical experiments to better understand the behavior of sparse LTS and other sparse regression methods under increasing levels of contamination. Concerning theoretical robustness properties, Alfons et al. (2013) assessed a family of L_1 penalized regression methods by computing their breakdown point (i.e., the maximum percentage of outliers that an estimator can withstand). Another important measure of robustness is the influence function (Hampel et al., 1986). It measures the influence that an observation has on a statistical functional at a given model distribution. We derive influence functions for a general class of regularized regression estimators.

Keywords: Influence function, outliers, regularization, sparse least trimmed squares

References

Alfons, A., C. Croux, and S. Gelper (2013). Sparse least trimmed squares regression for analyzing high-dimensional large data sets. *The Annals of Applied Statistics 7*(1), 226–248.

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association 96*(456), 1348–1360.

Hampel, F., E. Ronchetti, P. Rousseeuw, and W. Stahel (1986). *Robust Statistics: The Approach Based on Influence Functions*. New York: John Wiley & Sons.

Li, G., H. Peng, and L. Zhu (2011). Nonconcave penalized M-estimation with a diverging number of parameters. *Statistica Sinica 21*(1), 391–419.

Rosset, S. and J. Zhu (2004). Discussion of "Least angle regression" by Efron, B., Hastie, T., Johnstone, I., Tibshirani, R. *The Annals of Statistics 32*(2), 469–475.

Rousseeuw, P. and K. Van Driessen (2006). Computing LTS regression for large data sets. *Data Mining and Knowledge Discovery 12*(1), 29–45.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B 58*(1), 267–288.

van de Geer, S. (2008). High-dimensional generalized linear models and the lasso. *The Annals of Statistics 36*(2), 614–645.

Wang, H., G. Li, and G. Jiang (2007). Robust regression shrinkage and consistent variable selection through the LAD-lasso. *Journal of Business & Economic Statistics 25*(3), 347–355.