Robust Estimation and Variable Selection for High-Dimensional Linear Regression

S. Li[1], **Y. Qin**[1*], Y. Li[2] and Y. Yu[1]

[1] *University of Cincinnati, Cincinnati, Ohio, USA 45221; qinyn@ucmail.uc.edu, lis6@mail.uc.edu, yuyu@ucmail.uc.edu.*
[2] *Renmin University of China, Beijing, China 100872; yang.li@ruc.edu.cn.*
[*] *Presenting author*

Keywords. *Penalized estimation; Adaptive lasso; High-dimensional data.*

In this article, we introduce a class of robust linear regression estimators for variable selection in presence of outliers. Consider a linear model $y_i = \mathbf{x}_i^T\boldsymbol{\beta} + \epsilon_i$, $i = 1, \ldots, n$ where $(y_i, \mathbf{x}_i)$ represents the ith observation and $y_i \in \mathbb{R}, \mathbf{x}_i \in \mathbb{R}^d$. $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d) \in \mathbb{R}^d$ is an unknown regression coefficient vector. ϵ_i is an iid random error that is independent from $\mathbf{x}_i$ and follows a symmetric parametric distribution $f(\cdot)$ with mean 0 and constant variance σ^2. $f(\cdot)$ is assumed to be a Gaussian probability density function. Usually some of the elements in $\boldsymbol{\beta}$ are zeros. To select only important variables and estimate their coefficients robustly, we propose the following penalized likelihood estimator,

$$\hat{\boldsymbol{\beta}}_t = \arg\max_{\beta\in\mathbb{R}^d} \left\{ \underbrace{\sum_{i=1}^{n} \ln_t(f(y_i - \mathbf{x}_i^T\boldsymbol{\beta}))}_{\text{robustified likelihood}} - \underbrace{n\sum_{j=1}^{d} p_{\lambda_{nj}}(\beta_j)}_{\text{penalty}} \right\},$$

where $\ln_t(\cdot)$ is defined as: $\ln_t(u) = \ln(t) + \sum_{k=1}^{K} \frac{\ln^{(k)}(t)}{k!}(u-t)^k$ if $u < t$, and $\ln_t(u) = \ln(u)$ if $u \geq t$ or $t = 0$. Here, $t \geq 0$ is a tuning parameter and $\ln_t(u)$ is essentially a K-th order Taylor expansion of $\ln(u)$ for $u < t$. By introducing this tuning parameter t, we robustify the log-likelihood function so that it becomes insensitive to perturbation to the data (**?**). Note that when $t \to 0$, $\ln_t(u) \to \ln(u)$, therefore, the proposed estimator includes the penalized least square estimator as a special case with $t = 0$.

When solving the optimization program, we essentially solve a weighted likelihood equation where observations that disagree with the assumed model receive low weights. For example, when $K = 1$, the first order condition on the robustified likelihood becomes $0 = \sum_{i=1}^{n}[\frac{\partial}{\partial\boldsymbol{\beta}} \ln(f(y_i - \mathbf{x}_i^T\boldsymbol{\beta}))] \min(1, f(y_i - \mathbf{x}_i^T\boldsymbol{\beta})/t)$. Therefore, observations whose likelihoods are below t (which more likely turn out to be outliers) receive only partial weights whereas other observations receive full weights.

Table 1: Monte Carlo Simulation

n	Method	CSR	Under Fitted	Over Fitted	Miss Fitted	Model Error	
						Median	MAD
100	Proposed	0.995	0	0.005	0	0.053	0.027
	LAD	0.989	0	0.011	0	0.097	0.050
200	Proposed	1.000	0	0.000	0	0.024	0.012
	LAD	0.998	0	0.002	0	0.043	0.020
400	Proposed	1.000	0	0.000	0	0.012	0.006
	LAD	1.000	0	0.000	0	0.021	0.011

In the linear regression setting, the proposed penalized estimator obtains remarkable robustness when data is contaminated and still performs well when the model is correctly specified. One can control the estimator's robustness by adjusting t. When $t \to 0$, the proposed estimator becomes the traditional penalized least square estimator. When t is sufficiently large, the proposed estimator becomes the penalized minimum L_2 distance estimator. With a moderate t, the proposed estimator can be considered as a mixture of penalized Kullback-Leibler distance estimation and penalized L_2 distance estimation, where the former is known for its desirable asymptotic properties and the latter is known for its remarkable robustness.

We further show that the proposed estimator is consistent and enjoys oracle property. We also establish the bound of L_2 norm of the estimation error. Furthermore, the proposed estimator achieves the highest asymptotic breakdown point of 1/2 and is equipped with a bounded influence function. In addition, we have proposed a method for adaptively selecting the tuning parameter t to guarantee the robustness as well as the its asymptotic properties.

We conduct simulation studies to demonstrate the advantage of the proposed method over the traditional method in Table **??**. We generate $\mathbf{x}_i \overset{i.i.d.}{\sim} 0.8N(\mathbf{0}, \Omega_1)+0.2N(\mathbf{2}, \Omega_2)$ where $\Omega_1 = \mathbf{I}_d, \Omega_2 = \Sigma_{d\times d}$ with $\{\Sigma\}_{ij} = 0.5^{|i-j|}$ and $\epsilon_i \overset{i.i.d.}{\sim} 0.8N(0,1) + 0.2N(10, 6^2)$, and obtain y_i by the linear regression. We apply both the proposed estimator and the least absolute deviations (LAD) estimator with adaptive lasso penalty on the simulated data and compare their variable selection performance based on the proportions of correctly selecting (i.e. CSR), under fitting, over fitting, and miss fitting the true model. We also compare the estimation performance based on model error, $(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T E[\mathbf{x}\mathbf{x}^T](\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})$. As the table shows, the proposed method outperforms LAD in different scenarios in terms of both selection accuracy and estimation error.

References

Wang, X., Jiang, Y., Huang, M., Zhang, H. (2013). Robust Variable Selection with Exponential Squared Loss, *J. Amer. Statist. Assoc.*, 108(502), 632-643