Academic Journal of Nawroz University (AJNU), Vol.9, No.2, Jan 2020

https://doi.org/10.25007/ajnu.v9n2a818

Comparison Some Robust Regularization Methods in Linear Regression via Simulation Study

[1] Sherzad M. Ajeel, [2] Hussein A. Hashem
[1,2] Department of Mathematics, College of Sciences, University of Duhok, Kurdistan Region, Iraq

ABSTRACT

In this paper, we reviewed some variable selection methods in linear regression model. Conventional methodologies such as the Ordinary Least Squares (OLS) technique is one of the most commonly used method in estimating the parameters in linear regression. But the OLS estimates performs poorly when the dataset suffer from outliers or when the assumption of normality is violated such as in the case of heavy-tailed errors. To address this problem, robust regularized regression methods like Huber Lasso (Rosset and Zhu, 2007) and quantile regression (Koenker and Bassett ,1978] were proposed. This paper focuses on comparing the performance of the seven methods, the quantile regression estimates, the Huber Lasso estimates, the adaptive Huber Lasso estimates, the adaptive LAD Lasso, the Gamma-divergence estimates, the Maximum Tangent Likelihood Lasso (MTE) estimates and Semismooth Newton Coordinate Descent Algorithm (SNCD) Huber loss estimates.

Keywords: High-dimensional regression; Lasso; Regularization; Robust regression; Variable selection.

1. Introduction

Variable selection is important for high-dimensional data analysis in many research areas such as biology, signal processing and collaborative filtering. For example, microarray experiments allow one to measure thousands of variables (genes, proteins) simultaneously. The data sets generated by these experiments are generally very large in terms of the number of predictors (p) and often small in terms of the number of biological samples (n). In regression analysis, this problem is often termed as the "large p and small n problem" ($p \gg n$) and presents a major barrier to traditional statistical methods.

With the development of computer and data collection technologies, the database sizes continue to grow and various statistical methodologies have been developed over the past several decades to cope with the challenges presented by these data. In particular, there are major challenges in parameter estimation, model and variable selection. Several robust regression methods have been proposed for fitting multiple regression models, especially for the case when $p \geq n$ where the least squares method could not be used.

Tibshirani (1996) proposed Lasso (Least Absolute Shrinkage and Selection Operator), that minimizes the residual sum of squares subject to an L_1-norm constraint. The Lasso penalty results into some coefficients being estimated to completely zero, thus performing estimation and variable selection simultaneously. Following from the seminal paper of Tibshirani (1996), various extensions of Lasso were developed, for example adaptive Lasso (Zou, 2006), Smoothly Clipped Absolute Deviation (SCAD) (Fan and Li, 2001), etc.

Quantile regression, introduced by Koenker and Bassett (1978), could be used when an estimate of the various quantiles (such as the median) of a conditional distribution is of interest. This allows one to see and

compare how some quantiles of the response variable may be more affected by some predictor variables than other quantiles.

Some methods have combined regularized and robust regression methods in order to perform variable selection in high-dimensional data with outliers. For example, Rosset and Zhu (2007) proposed the Huber Lasso method which combines the Huber's criterion loss with a Lasso penalty. The LAD-adaptive Lasso method is proposed by Wang et al. (2007), combining the idea of Least Absolute Deviance (LAD) and adaptive Lasso. Fujisawa and Eguchi (2008) introduce the Gamma divergence for regression. It measures the difference between two conditional probability density functions. Lambert-Lacroix and Zwald (2011) developed the Huber's Criterion with adaptive Lasso which combines the Huber's loss function and adaptive Lasso penalty. Yi and Huang (2016) developed an algorithm, called Semismooth Newton Coordinate Descent (SNCD), to obtain a better efficiency and scalability for computing the solution paths of penalized quantile regression. Qin et .al (2017) proposed a method called Maximum Tangent Likelihood Estimation (MTE). In the next section we will give an overview some regularized and robust regression methods.

2. Methods

We start from the standard model for multiple linear regression to describe the regression regularization methods. Let the data $(x_1, y_1), \ldots, (x_n, y_n)$, and the design matrix denoted by $\boldsymbol{X} = (x_1^T, \ldots, x_n^T)^T$, the general linear model is usually written as

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

Here $\beta = (\beta_1, \ldots, \beta_p)^T$ are the regression coefficients $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^T \sim N(0, \sigma^2 I_n)$ are the random errors, x_i are the regressors for observation i, $i = 1, \ldots, n$ and $y = (y_1, \ldots, y_n)^T$. The ordinary least squares (OLS) method estimates β by minimizing the residual squared error, i.e. $\hat{\beta}_{OLS} = \min_{\beta}\{(y - \boldsymbol{X}\beta)^T(y - \boldsymbol{X}\beta)\}$.

In general, OLS tends to give estimators with low biases but high variances, and better prediction accuracy can usually be obtained by lowering the variance with a little increased bias.

2.1 Lasso Regression

In order to reduce the estimator's variance and to carry out variable selection, Tibshirani (1996) introduced Least Absolute Shrinkage and Selection Operator, also known as Lasso, was a new method for linear model estimation. The inventor Tibshirani (1996) described it as follows: "The Lasso minimizes the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant". In other words, Lasso is a regression shrinkage method typically used in models with large number of variables but relatively few observations. The main purpose of Lasso is to perform variable selection while fitting the regression line to the data. This is done by shrinking certain coefficients but in addition setting some of the coefficients also to zero. Lasso performs a L_1 regularization by adding a penalty to the objective under optimization. This penalty is the sum of absolute value of coefficients and determines which coefficients to shrink and how much. The Lasso estimate $\hat{\beta}$ is defined by:

$$\hat{\beta}_{lasso} = \min_{\beta}\left\{\sum_{i=1}^{n}\left(y_i - \sum_j \beta_j x_{ij}\right)^2\right\}, \text{ s.t. } \sum_{j=1}^{p}|\beta_j| \leq t, \; t \geq 0. \qquad (2)$$

An equivalent form of the Lasso is,

$$\hat{\beta}_{lasso} = \min_{\beta}\left\{\sum_{i=1}^{n}\left(y_i - \sum_j \beta_j x_{ij}\right)^2 + \lambda\sum_j|\beta_j|\right\}, \qquad (3)$$

or

$$\hat{\beta}_{lasso} = \min_{\beta}\|y - x\beta\|_2^2 + \lambda\|\beta\|_1 . \qquad (4)$$

lambda is the parameter deciding the weight on minimizing the RSS compared to the penalty term that is the sum of absolute value of coefficients.

2.2 Adaptive Lasso

To remedy the problem of the lack of the oracle property, the adaptive Lasso estimator was proposed (Zou, 2006)

$$\hat{\beta}_{\text{adaptive Lasso}} = \min_{\beta}\left\{\sum_{i=1}^{n}\left(y_i - \sum_j \beta_j x_{ij}\right)^2 + \lambda \sum_{j=1}^{p} \widehat{w}|\beta_j|\right\} \quad (5)$$

Where $\widehat{w}_j (j = 1, \dots, p)$ are the adaptive data-driven weights, which can be estimated by $\widehat{w}_j = \frac{1}{|\hat{\beta}_j|^{\gamma}}$,where γ is a positive constant and $\hat{\beta}_j$ is an initial consistent of β obtained through least squares or ridge regression if multicolinearity is important (Ogutu,2012) . The optimal value $\gamma > 0$ and λ can be simultaneously selected from a gride of values, with values of γ selected from {0.5, 1, 2}, using two-dimensional cross-validation. The weights allow the adaptive Lasso to apply different amounts of shrinkage to different coefficients and hence to more severely penalize coefficients with small values.

2.3 Huber Lasso

When the regression response suffers from outliers, the performance of Lasso can be poor. A first attempt to solve this problem has been done by Rosset and Zhu (2007) and Wang et al. (2007). Rosset and Zhu (2007) combine the idea of Huber's criterion as loss function and Lasso penalty. They fix the penalty to be the L_1-penalty and use Huber's loss function with fixed M. That is

$$\hat{\beta}_{Huber\ lasso} = \min_{\beta} \sum_{i=1}^{n} \rho(y_i - x_i^T \beta) + \lambda \sum_{j=1}^{p} |\beta_j| , \quad (6)$$

where $\rho(t) = \begin{cases} t^2 & if \quad |t| \le M \\ 2M|t| - M^2 & if \quad |t| > M \end{cases}$.

2.4 Adaptive Huber Lasso

Lambert-Lacroix and Zwald (2011) proposed the Huber's Criterion with adaptive Lasso which combines the idea of Huber's criterion as loss function and adaptive Lasso penalty, defined by

$$\hat{\beta}_{Hadl} = \min_{\beta} \mathcal{L}_{\rho}(\beta, s) + \lambda \sum_{j=1}^{p} \widehat{w}_j^{Hadl} |\beta_j|$$

where $\widehat{w}_j^{ladl} = \left(\widehat{w}_1^{Hladl}, \dots, \widehat{w}_p^{Hladl}\right)$ is a known weights vector and the Huber's criterion is defined by

$$\mathcal{L}_{\rho}(\beta, s) = \begin{cases} ns + \sum_{i=1}^{n} \rho\left(\frac{y_i - \sum_{j=1}^{p} \beta_j x_{ij}}{s}\right) s & if \quad s > 0, \\ 2M \sum_{i=1}^{n} \left|y_i - \sum_{j=1}^{p} \beta_j x_{ij}\right| & if \quad s = 0, \\ +\infty & if \quad s < 0, \end{cases}$$

where $\rho(t)$ is defined as (6), $s > 0$ is a scale parameter for the distribution. The $\rho(t)$ definition shows how the loss is quadratic for small residuals but it becomes linear for large residuals, thus penalizing outliers. Also this method has been used for regression problems in a number of applications and has shown robustness against outliers. The constant M depends on the level of noise and outliers in the data and is often set to the value $M = 1.345$.

2.5 LAD-Lasso

To obtain a robust Lasso-type estimator, The LAD Lasso method is developed. The LAD -Lasso can be written as (Wang et al., 2007).

$$\hat{\beta}_{Lad\ lasso} = \min_{\beta} \sum_{i=1}^{n} \left|y_i - \sum_{j=1}^{p} \beta_j x_{ij}\right| + \lambda \sum_{j=1}^{p} |\beta_j| \quad (7)$$

As can be seen, the LAD-Lasso criterion combines the LAD criterion and the Lasso penalty, and hence the resulting estimator is expected to be robust against outliers and also to enjoy a sparse representation.

2.6 Adaptive LAD-LASSO

We consider the following LAD-Lasso criterion, which combines Zou's adaptive LASSO, to perform consistent variable selection, with LAD regression, to perform robust estimation in the presence of heavy-tailed errors (Lambert-Lacroix and Zwald, 2011)

$$\hat{\beta}_{ladl} = \min_{\beta} \sum_{i=1}^{n} \left|y_i - \sum_{j=1}^{p} \beta_j x_{ij}\right| + \lambda \sum_{j=1}^{p} \widehat{w}_j^{ladl} |\beta_j| \quad (8)$$

where $\widehat{w}_j^{ladl} = \left(\widehat{w}_1^{ladl}, \dots, \widehat{w}_p^{ladl}\right)$ is a known weights vector. In this model the estimator is robust to outliers because the squared loss has been replaced by the l_1-loss.

2.7 Variable Selection in Quantile Regression

Ordinary least squares regression estimates the mean response as a function of the predictors. As an alternative, least absolute deviation (LAD) regression estimates the conditional median function, which has

been shown to be resistant to response outliers and more efficient when the errors have heavy tails. In the seminal paper of Koenker and Bassett (1978), they generalized the idea of LAD regression and introduced quantile regression (QR) to estimate the conditional quantile function of the response. QR not only inherits the good properties of LAD regression but also provides much more information about the conditional distribution of the response variable. A brief review of quantile regression models is as follows.

Given the data$(x_1, y_1),. \ . \ ., (x_n, y_n)$, unlike the mean regression model (1) which models the conditional mean $E(y|X) = X\beta$.

Koenker and Bassett (1978) proposed the linear quantile regression model for the θth quantile $(0 < \theta < 1)$ as

$$y_i = x_i^T\beta + u_i, \ i = 1,. \ . \ ., n \qquad (9)$$

Where $\beta = (\beta_1,. \ . \ ., \beta_p)^T \in R^p$ and $u_i's$ are independent with their θth quantiles equal to zero.

Based on different choices of θ, quantile regression gives a more flexible and comprehensive modelling of the relationship between response variables $y_i's$ and regressors $x_i's$. Note that when $\theta = 0.5$, this reduces to the least absolute deviation regression or median regression, which is known for its robustness to outliers. In general, quantile regression with a given $\theta \in (0, 1)$ is also recognized as being robust to outliers. Moreover, one important advantage of quantile regression is that it makes no distributional assumption to the error terms $u_i's$ other than their quantile. It can be shown that the coefficients β can be estimated consistently by the solution to the following minimization problem

$$\min_{\beta} \sum_{i=1}^{n} \rho_\theta(y_i - x_i^T\beta) \qquad (10)$$

where $\rho(.)$ is an outlier resistant loss function called the objective function

$$\rho_\theta(t) = \begin{cases} \theta t & if \ \ t \geq 0 \\ -(1-\theta)t & if \ \ t < 0 \end{cases}, \text{where } 0 < \theta < 1. \qquad (11)$$

The first use of regularization in quantile regression is made by (Koenker, 2004), which put the Lasso penalty on the random effects in a mixed-effect quantile regression model to shrink the random effects towards zero.

2.8 Regression via Gamma-Divergence

The Gamma divergence for regression was first proposed by Fujisawa and Eguchi (2008). It measures the difference between two conditional probability density functions. The other type of the Gamma divergence for regression was proposed by Kawashima and Fujisawa (2017), in which the treatment of the base measure on the explanatory variable was changed. In this section, we briefly review Gamma divergences for regression and present the corresponding parameter estimation (Fujisawa and Eguchi, 2008).

Suppose that $g(x, y), g(y|x)$ and $g(x)$ are the underlying probability density functions of (x, y),y given x and x, respectively. Let $f(y|x)$ be another parametric conditional probability density function of y given x. Let us define the *Gamma*-cross-entropy for regression by:

$$d_\gamma\big(g(y|x), f(y|x); g(x)\big) = -\frac{1}{\gamma} log \int \left(\int g(y|x) f(y|x)^\gamma \, dy \right) g(x) dx + \frac{1}{1+\gamma} log \int \left(\int (y|x)^{1+\gamma} \, dy \right) g(x) dx = -\frac{1}{\gamma} log \int \int f(y|x)^\gamma g(x, y) dx dy + \frac{1}{1+\gamma} log \int (\int (y|x)^{1+\gamma} \, dy) g(x) dx \quad for \ \gamma > 0 \qquad (12)$$

The *Gamma*-divergence for regression is defined by:

$$D_\gamma\big(g(y|x), f(y|x); g(x)\big) = -d_\gamma\big(g(y|x), g(y|x); g(x)\big) + d_\gamma\big(g(y|x), f(y|x); g(x)\big) \qquad (13)$$

Let $f(y|\, x;\, \theta)$ be the conditional probability density function of y given x with parameter θ. The target parameter can be considered by:

$$\theta_\gamma^* = \min_\theta D_\gamma(g(y|x), f(y|x; \theta); g(x) = \min_\theta d_\gamma\big(g(y|x), f(y|x; \theta); g(x)\big) \qquad (14)$$

when $g(y|x) = f(y|x; \theta^*)$, we have $\theta_\gamma^* = \theta^*$.

Let $(x_1, y_1), \ldots \ , (x_n, y_n)$ be the observations randomly drawn from the underlying distribution $g(x, y)$. Using

the formula (12), the γ −cross-entropy for regression, $d_\gamma\big(g(y|x), f(y|x;\theta); g(x)\big)$, can be empirically estimated by:

$$\bar{d}_\gamma\big(f(y|x;\theta)\big) = -\frac{1}{\gamma} log\left\{\frac{1}{n}\sum_{i=1}^{n} f(y_i|x_i;\theta)^\gamma\right\} + \frac{1}{1+\gamma} log\left\{\frac{1}{n}\sum_{i=1}^{n} f(y_i|x_i;\theta)^{\gamma+1} dy\right\}.$$

By virtue of (13), we define the γ-estimator by :

$$\hat{\theta}_\gamma = \underset{\theta}{argmin}\, \bar{d}_\gamma\big(f(y|x;\theta)\big)$$

Fujisawa and Eguchi (2008) proposed three procedures to estimate the parameters, MM algorithm for sparse *Gamma* regression, sparse *Gamma* linear regression and robust cross-validation.

2.9 Maximum Tangent Likelihood Estimation (MTE)

Qin et .al (2017) proposed the method of Maximum Tangent Likelihood Estimation (MTE) as

$\check{\beta} = \max_\beta \sum_{i=1}^n ln_t\left(f(\boldsymbol{z}_i:\beta)\right)$, where $\{\boldsymbol{z}_i\}_{i=1}^n = \{y_i, \boldsymbol{X}_i^T\}_{i=1}^n$ represents the response variable and covariates, and f represents the normal distribution with zero mean, and $f(\boldsymbol{z}_i:\beta) = f(y - \boldsymbol{X}_i^T\beta)$. However, the performance of such an estimator usually degrades drastically even if a small proportion of data is contaminated.

The robust statistical procedure should perform nearly optimally when model assumptions are valid and still maintain good performance when the assumptions are violated. The penalized maximum tangent likelihood estimation (penalized MTE) for variable selection as

$$\hat{\beta} = \max_\beta\left\{\sum_{i=1}^n ln_t\left(f(\boldsymbol{z}_i:\beta)\right) - n\sum_{j=1}^n p_{\lambda_n}(|\beta_j|)\right\} \quad (15)$$

where the function $lnt(.)$ is defined as

$$ln_t(u) = \begin{cases} \ln(u), & if\ u > t \\ \ln(t) + \sum_{k=1}^p \frac{\partial^k \ln(v)}{\partial v^k}\,|_{v=t} \frac{(u-t)^k}{k!} & if\ 0 \le u \le t \end{cases}. \quad (16)$$

Here $t \ge 0$ is a tuning parameter. $ln_t(u)$ is essentially a $p-th$ order Taylor expansion of $\ln(u)$ for $0 \le u < t$.

2.10 Semismooth Newton Coordinate Descent Algorithm (SNCD)

Yi and Huang (2016) developed an algorithm, called Semismooth Newton Coordinate Descent (SNCD), to obtain a better efficiency and scalability for computing the solution paths of penalized quantile regression. They also provide an R package called *hqreg*. Moreover, this package also obtains Lasso of (Tibshirani (1996)), Ridge of (Hoerl and Kennard (1970)) and Elastic Net of (Zou and Hastie (2005)) estimators in the quantile regression models. The *hqreg* functions give the solution path while the *quantreg* package of Koenker (2013) computes a single solution.

3. Simulation Study

In this section, we compare some robust regularized regression methods in low-dimensional with sparse and non-sparse coefficients($p = 15, n = 100$) and high-dimensional with sparse coefficients ($p = 100$, $n = 50$)settings. For the sparse settings we use a classical simulation setting, e.g. Bradic and Fan (2011), where $y = \beta_0 + x\beta + u$, with $\beta_0 = 0$ and $\beta = (3, 1.5, 0, 0, 2, 0, \ldots, 0)$ (the sparse case) and for non-sparse setting we use $\beta_j = 0.1$ (the dense case) for all j .We draw the independent variables x from a multivariate normal distribution, $N(0, \Sigma_x)$. The pairwise covariance between x_i and x_j is set to be $(\Sigma_x)_{ij} = r^{|i-j|}$. For the error u, we choose a range of distributions in order to test the robustness of the methods to departures from normality. In particular, we consider the following cases: $u \sim N(0,1)$, Double Exponential (DE), Beta distribution $\beta(2,3)$, t-distribution (t_3) with 3 degrees of freedom, Chi square $(\chi^2_{(3)})$ with 3 degrees of freedom and mixture normal distributions. We design a mixture normal distribution with large outliers, similar to Lambert-Lacroix and Zwald (2011), by drawing 90% of the data from $a\ N(0,1)$ distribution and 10% from a $N(0, 1000)$ distribution. Under all these cases, we

compare the regularized regression methods described in the previous section, namely adaptive LAD Lasso, Huber Lasso with their adaptive version (Xu and Ying, 2010; Lambert-Lacroix and Zwald, 2011), quantile regression Koenker and Bassett (1978), Gamma divergence (Fujisawa and Eguchi, 2008),Maximum Tangent Likelihood Estimation (MTE) (Qin et .al, 2017) and Semismooth Newton Coordinate Descent (SNCD) (Yi and Huang, 2016). For Huber Lasso we use the R implementations provided by Rosset and Zhu (2007), for the adaptive LAD Lasso and adaptive Huber Lasso methods we adapt some of the functions in the *parcor* R package. For the adaptive versions of the methods, we define the weights using the corresponding non-adaptive Lasso versions with a penalty parameter chosen to optimize a BIC criterion. For quantile regression and SNCD methods we use the R package *hqreg*, for Gamma divergence we use the R package *gamreg* and for MTE method we use the R package *MTE*

3.1 Example 1: low-dimensional with sparse coefficients

The best subset selection and the lasso estimators have a special, useful property. Their solutions are sparse, i.e., at a solution $\hat{\beta}$ we will have $\beta_j = 0$ for many components $j \epsilon\{1,. \ . \ . \ ,p\}$. We consider a low-dimensional data with sparse coefficients set with $p = 15$ and $n = 100$. Figure 1, table 1 and table 2 report the results of the simulation. We consider both the case of low correlation ($r = 0.5$) and that of high correlation ($r = 0.95$) of the predictors. The top panels report the median model error over 100 iterations (similar results for the mean error), with the model error computed by$(\hat{\beta} - \beta)^T S_x(\hat{\beta} - \beta)$, where $\hat{\beta}$ are the estimated parameters and S_x the sample covariance. The bottom panels report the true positives that are the number of correctly found non-zero coefficients. Here three corresponds to the case of all non-zero coefficients being correctly detected.

Our results show that: quantile Lasso and the Semismooth Newton Coordinate Descent SNCD methods do not perform well when the predictors are highly correlated; the adaptive LAD and the Gamma divergence methods outperform all others methods for all error distributions.

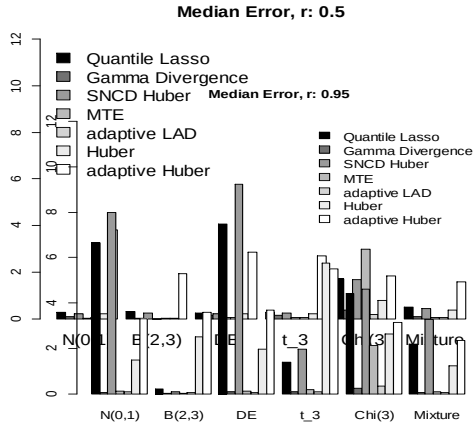
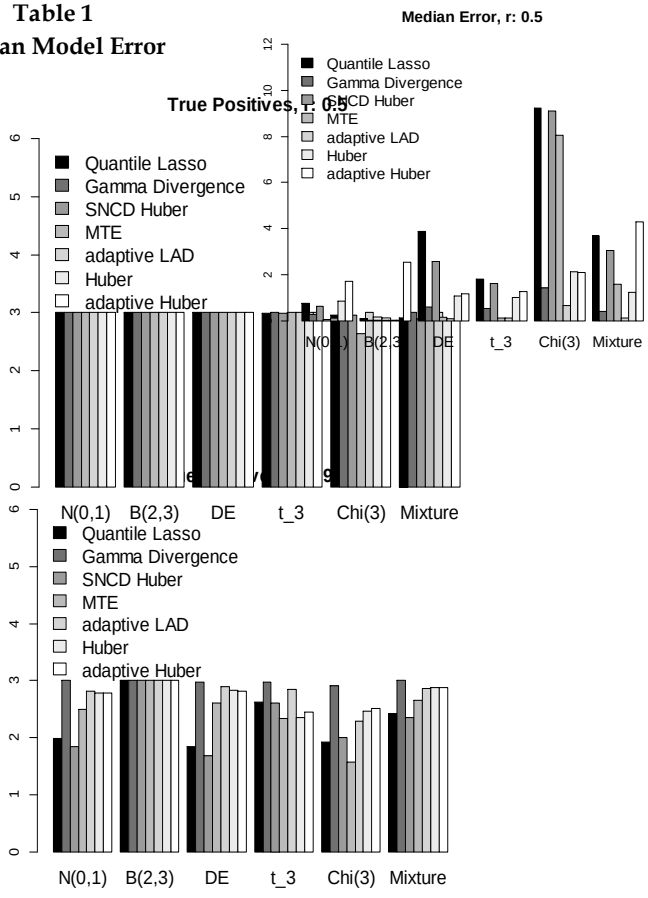


Figure 1: Comparison of robust regression Lasso methods under different error distributions, for low (left) and high (right) correlated predictors. The top panels plot the median model error over 100 replications for example 1 and the bottom panels the average true positives when $p = 15$ and $n = 100$

Table 1
Median Model Error



averaged over 100 replications for the case: $\mathbf{p = 15, n = 100, r = 0.5}$ and β values as in example (1), Best method indicated in bold.

	Quantile Lasso	MTE	Huber	*Gamma Divergence*	*Adaptive LAD*	*adaptive Huber*	SNCD Huber
$N(0,1)$	0.271	**0.027**	0.216	0.070	0.043	3.810	0.205
$B(2,3)$	0.306	**0.002**	0.008	0.003	0.004	1.940	0.254
DE	0.252	0.050	0.199	0.136	**0.042**	2.865	0.200
t_3	0.272	**0.049**	0.206	0.133	0.057	2.702	0.250
$Chi(3)$	1.715	2.994	0.784	0.355	**0.181**	1.841	1.674
Mixtur	0.495	0.038	0.377	0.091	**0.062**	1.594	0.438

Table 2

Median Model Error averaged over 100 replications for the case: $\boldsymbol{p = 15, n = 100, r = 0.95}$ and $\boldsymbol{\beta}$ values as in example (1). Best method indicated in bold.

	Quantile Lasso	MTE	Huber	*Gamma Divergence*	*Adaptive LAD*	*adaptive Huber*	SNCD Huber
$N(0,1)$	6.653	0.124	1.499	**0.048**	0.076	3.315	7.975
$B(2,3)$	0.204	0.007	2.510	**0.003**	0.036	3.590	0.091
DE	7.483	0.128	1.948	0.075	**0.051**	3.697	9.245
t_3	1.401	0.175	5.749	0.079	**0.076**	5.509	1.952
$Chi(3)$	4.414	**2.119**	2.630	0.252	0.342	3.138	4.600
Mixture	2.199	0.069	1.225	0.065	**0.056**	2.345	3.277

3.2 Example 2: high-dimensional with sparse coefficients

We consider a similar setting to simulation 3.1 but with different sample size and number of predictors. In particular, we consider a high- dimensional example with sparse coefficients with $p = 100$ and $n = 50$. Given the setup of the simulation, this a very sparse problem in which most of the coefficients are zero. Figure 2, table 3 and table 4 present the results of the simulation. The top panels report the median model error over 100 replications, with the model error computed in the same way as in Figure 1. The bottom panels report the true positive that is the number of correctly classified non-zero coefficients.

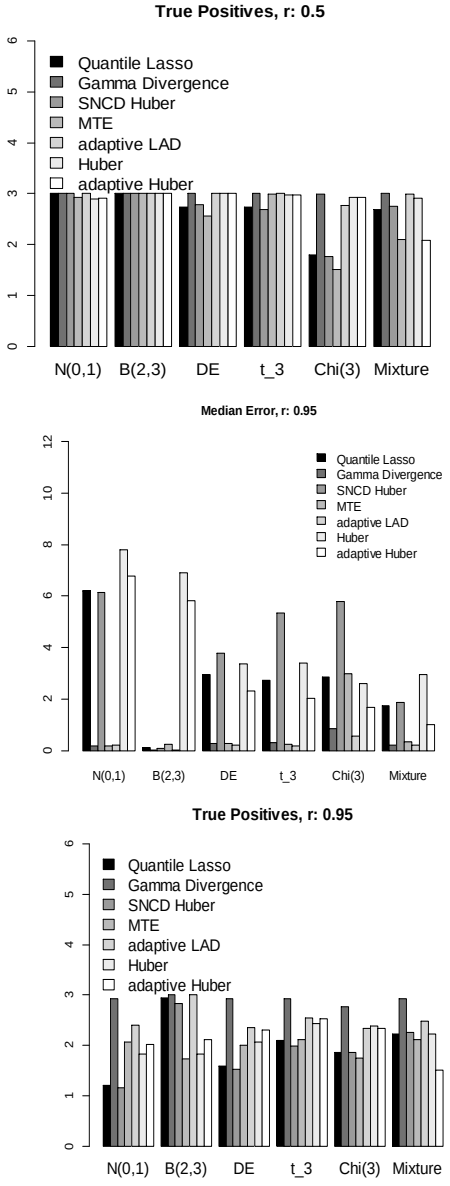


Figure 2: Comparison of robust regression Lasso methods under different error distributions, for low (left) and high (right) correlated predictors. The top panels plot the median model error over 100 replications for example 2 and the bottom panels the average true positives when $\boldsymbol{p = 100}$ and $\boldsymbol{n = 50}$.

Table 3
Median Model Error averaged over 100 replications for the case: $p = 100, n = 50, r = 0.5$ and β values as in example (2). Best method indicated in bold.

	Quantile Lasso	MTE	Huber	*Gamma Divergence*	*Adaptive LAD*	*adaptive Huber*	SNCD Huber
$N(0,1)$	0.765	**0.056**	0.840	0.275	0.074	1.710	0.629
$B(2,3)$	0.068	0.010	0.017	0.011	**0.004**	2.545	0.030
DE	3.867	0.140	1.074	0.601	**0.078**	1.165	2.579
t_3	1.804	**0.116**	0.994	0.520	0.125	1.274	1.628
$Chi(3)$	9.218	8.040	2.111	1.433	**0.662**	2.090	9.116
Mixture	3.691	1.587	1.217	0.414	**0.128**	4.288	3.049

Table 4
Median Model Error averaged over 100 for the case: $p = 100, n = 50, r = 0.95$ and β values as in example (2). Best method indicated in bold.

	Quantile Lasso	MTE	Huber	*Gamma Divergence*	*Adaptive LAD*	*adaptive Huber*	SNCD Huber
$N(0,1)$	6.198	0.186	7.794	**0.181**	0.199	6.766	6.155
$B(2,3)$	0.113	0.235	6.901	0.006	**0.004**	5.822	0.084
DE	2.948	0.288	3.372	0.275	**0.206**	2.329	3.778
t_3	2.727	0.242	3.394	0.292	**0.189**	2.032	5.328
$Chi(3)$	2.871	2.987	2.610	0.849	**0.549**	1.681	5.776
Mixture	1.754	0.337	2.960	0.220	**0.203**	0.995	1.875

The results support the performance of the methods: quantile Lasso and Semismooth Newton Coordinate Descent (SNCD) do not perform well when the predictors are highly correlated, the adaptive LAD and the Gamma divergence methods outperform all others method ones as departures from normality increase. This is particularly evident for the cases of the $\beta(2,3)$ and $\chi^2_{(3)}$ simulation, which have a severe departure from normality.

3.3 Example 3: low- dimensional with non-sparse coefficients

In this paper, "non-sparsity" is in the sense that only a few regression coefficients are large and the rest are small but not necessary to be zero. In order to investigate the performance of robust regularized regression methods in example 2, we set up a new simulation where we have $\beta_j = 0.1$ (the dense case) for all j, that is a non-sparse situation. Figure 3 reports the median model error over 100 replications for the case $p = 15$ and $n = 100$.

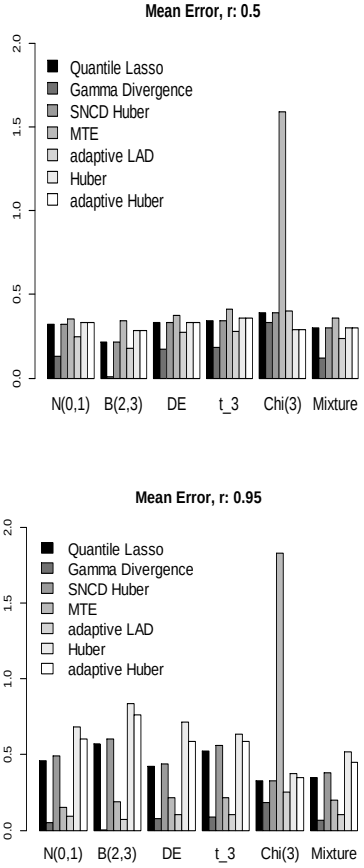


Figure 3: Comparison of robust regression Lasso methods under different error distributions, for low (left) and high (right) correlated predictors. The plot show the median model error over 100 replications for example 3 when $p = 15$ and $n = 100$.

Table 5
Median Model Error averaged over 100 replications for the case: $p = 15, n = 100, r = 0.5$ and β values as in example (3). Best method indicated in bold.

	Quantile Lasso	MTE	Huber	*Gamma Divergence*	*Adaptive LAD*	*adaptive Huber*	SNCD Huber
$N(0,1)$	0.320	0.356	0.332	**0.101**	0.237	0.332	0.320
$B(2,3)$	0.216	0.363	0.284	**0.006**	0.171	0.284	0.213
DE	0.334	0.373	0.336	**0.155**	0.274	0.336	0.338
t_3	0.344	0.416	0.360	**0.161**	0.273	0.360	0.344
$Chi(3)$	0.360	1.105	**0.289**	0.338	0.363	**0.289**	0.363
Mixture	0.301	0.340	0.303	**0.114**	0.223	0.303	0.301

Table 6
Median Model Error averaged over 100 replications for the case: $p = 15, n = 100, r = 0.95$ and β values as in example (3). Best method indicated in bold.

	Quantile Lasso	MTE	Huber	*Gamma Divergence*	*Adaptive LAD*	*adaptive Huber*	SNCD Huber
$N(0,1)$	0.416	0.127	0.673	**0.046**	0.088	0.554	0.447
$B(2,3)$	0.564	0.198	0.815	**0.005**	0.067	0.759	0.595
DE	0.324	0.203	0.655	**0.066**	0.092	0.532	0.392
t_3	0.438	0.199	0.613	**0.070**	0.092	0.609	0.472
$Chi(3)$	0.294	**1.140**	0.292	0.156	0.213	0.240	0.296
Mixture	0.329	0.207	0.471	**0.054**	0.095	0.401	0.330

From results in Figure 3, table 5 and table 6 our simulation study confirms that the performances of the adaptive LAD and the Gamma divergence methods are closer. Furthermore, the results show how MTE is the worst performing method in case of departure from normality especially when the predictors highly correlated.

4. Concluding remarks

Many approaches are developed in statistics that rely on the assumption of normality. These approaches are not suited to data that show clear departures from normality. This is often the case when data are contaminated, resulting in the presence of outliers. In this paper, we have considered recently developed robust regularized regression methods and, such as the Huber or LAD methods. In a high dimensional setting, when $p \geq n$. In a simulation study, we show how the adaptive LAD and the Gamma divergence methods are superior to other robust methods, particularly for cases where there is a large departure from normality.

5. References:

1. Bradic, J., Fan, J. and Wang, W. (2011). Penalized composite quasi-likelihood for ultrahigh-dimensional variable selection. *J. R. Stat. Soc. Ser. B, 73, 325-349.*
2. Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc., 96 1348-1360.*
3. Fujisawa, H. and Eguchi, S. (2008). Robust parameter estimation with a small bias against heavy contamination, *J. Multivar. Anal., 99(9), 2053-2081.*
4. Hoerl, A.E. and Kennard, R.W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics , 12 (1), 55–67.*
5. Kawashima, K., Fujisawa, H. (2017). Robust and Sparse Regression via γ Divergence. *J. of entropy and infor. studies 19(11).*
6. Koenker, R. (2004). Quantile regression for longitudinal data. *J. Multivar. Anal. 91,74–89.*
7. Koenker, R. (2013). quantreg: Quantile Regression. R package version 5.05. R Foundation for Statistical Computing: Vienna) Available at: http://CRAN. R-project. org/package= quantreg.
8. Koenker, R. and G. W. Bassett (1978). Regression quantiles. *Econometrica 46, 33–50*
9. Lambert-Lacroix, S. and Zwald, L. (2011). Robust regression through the Huber's criterion and adaptive Lasso penalty. *Electronic J. of Statist. 5,*
10. Ogutu, J.O., T. Schulz-Streeck, and H.P. Piepho. (2012). Genomic selection using regularized linear regression models: Ridge regression, Lasso, elastic net and their extensions. *BMC Proceedings 6*:S10.
11. Qin, Y., Li, S. and Yu, Y. (2017). Penalized Maximum Tangent Likelihood Estimation and Robust Variable Selection.https://arxiv.org/pdf/1708.05439.pdf.
12. Rosset, S. and Zhu, J. (2007). Piecewise linear regularized solution paths. *The Annals of Statist. 35 (3)*, 1012–1030.
13. Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B, 58, 267–288.*
14. Wang, H., Li, G., and Jiang, G. (2007). Robust regression shrinkage and consistent variable selection through the LAD-Lasso. *J. of Business & Economic Statistics 25*, 347 - 355.
15. Yi, C. Huang, J. (2016). Semismooth Newton Coordinate Descent Algorithmfor Elastic-Net Penalized Huber Loss Regression and Quantile Regression. *J. of Com. and Graph. Statist. 3.* 547–557
16. Xu, J. and Ying, Z. (2010). Simultaneous estimation and variable selection in median regression using Lasso-type penalty. *Ann. of the Inst. of Statist. Math. 62*, 487–514.
17. Zou, H. (2006). The adaptive Lasso and its oracle properties. *J. Amer. Statist. Assoc., 101, 1418-1429.*
18. Zou, H. and Hastie, t. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B 67 301-320.*