Communications of the Korean Statistical Society
2011, Vol. 18, No. 6, 733–739

DOI: http://dx.doi.org/10.5351/CKSS.2011.18.6.733

Weighted Least Absolute Deviation Lasso Estimator

Kang-Mo Jung[1,a]

[a]Department of Informatics & Statistics, Kunsan National University

Abstract

The linear absolute shrinkage and selection operator(Lasso) method improves the low prediction accuracy and poor interpretation of the ordinary least squares(OLS) estimate through the use of L_1 regularization on the regression coefficients. However, the Lasso is not robust to outliers, because the Lasso method minimizes the sum of squared residual errors. Even though the least absolute deviation(LAD) estimator is an alternative to the OLS estimate, it is sensitive to leverage points. We propose a robust Lasso estimator that is not sensitive to outliers, heavy-tailed errors or leverage points.

Keywords: Heavy-tailed errors, Lasso, leverage points, outliers, robust estimator, weight least absolute deviation.

1. Introduction

In a linear regression model, the ordinary least squares(OLS) estimate is usually used to estimate the regression coefficients through the minimization of the sum of squared errors, because it is simple and unbiased. However, the OLS estimate has low prediction accuracy and poor interpretation. Prediction accuracy can be improved by shrinking some regression coefficients even though we sacrifice a little bit of bias. Poor interpretation can be resolved through the selection of a sparse representation with a smaller subset of coefficients (Tibshirani, 1996).

The linear absolute shrinkage and selection operator(Lasso) method is proposed by Tibshirani (1996) to estimate the parameters by shrinking some coefficients and setting others to zero in linear regression model. The Lasso estimator retains the good features of both subset selection and ridge regression (Hoerl and Kennard, 1970) that stabilizes estimates by placing a restriction on coefficients. The difference between Lasso and ridge regression is the penalty on the regression coefficients. Lasso employs the L_1 penalty while the ridge regression uses the L_2 penalty. The L_1 regularization tends to produce extremely sparse solutions; subsequently, the the Lasso method attracts more interests in model selection (Zhao and Yu, 2006).

The OLS estimate can be distorted when the error has a heavy-tailed distribution or outliers. It is well known that the OLS estimate is not robust to even a single outlier. Many robust estimators have been proposed to address the problem. One of them is the least absolute deviation(LAD) estimator that has $\sqrt{n}$-consistency and asymptotic normality without assuming the distribution of errors (Pollard, 1991). The Lasso estimate is obtained by minimizing the sum of squared residuals. Then it will be significantly degraded in a noise situation. Wang *et al.* (2007) proposed a robust regression shrinkage and selection method that can do regression shrinkage and selection like Lasso and is also resistant to outliers or heavy-tailed errors like LAD.

[1] Professor, Department of Informatics & Statistics, Kunsan National University, Kunsan 573-701, Korea.
E-mail: kmjung@kunsan.ac.kr

The LAD estimator is robust to points with large residuals called regression outliers; however, it is known that the LAD estimator is sensitive to leverage points (Croux *et al.*, 2003). Giloni *et al.* (2006) proposed a version of the LAD estimator that is not sensitive to leverage points by redescending the leverage points. A robust Lasso is proposed in this study and it adapts to the weighted LAD estimator instead of the LAD estimator so that the outliers or leverage points can be effectively suppressed.

The rest of the article is organized as follows. In Section 2 we review Lasso, LAD-Lasso and propose the weighted LAD-Lasso. The statistical properties of the proposed estimator are described. Section 3 presents a simulation results under several situations. Finally Section 4 concludes the article.

2. Weighted Absolute Shrinkage and Selection

2.1. Weighted LAD-Lasso

Consider the linear regression model

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i, \quad i = 1, \ldots, n, \tag{2.1}$$

where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ is the p-dimensional regression predictor, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is the regression coefficient vector, and ϵ_i is the independently identically distributed random errors with median 0.

The OLS estimate $\hat{\boldsymbol{\beta}}^{ols}$ minimizes

$$\text{RSS} = \frac{1}{2} \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^T \boldsymbol{\beta} \right)^2 .$$

Despite its simplicity and unbiasedness, the OLS estimator is not optimal in a predictive point of view. To get the better prediction with sacrificing the unbiasedness of the estimator Hoerl and Kennard (1970) introduced ridge regression obtained by penalizing the L_2 norm of the regression coefficients

$$\hat{\boldsymbol{\beta}}^{ridge}(\lambda) = \text{argmin}_\beta \left(\text{RSS} + n\lambda \sum_{j=1}^{p} |\boldsymbol{\beta}|^2 \right),$$

where $\sum |\boldsymbol{\beta}|^2$ is the L_2 penalty on $\boldsymbol{\beta}$ and $\lambda \geq 0$ is the tuning parameter that balances goodness-of-fit and model complexity. However, the ridge regression does not shrink unnecessary coefficients to zero. Tibshirani (1996) proposed the Lasso estimator obtained by minimizing

$$\hat{\boldsymbol{\beta}}^{lasso}(\lambda) = \text{argmin}_\beta \left(\text{RSS} + n\lambda \sum_{j=1}^{p} \left| \beta_j \right| \right). \tag{2.2}$$

Because it can shrink some regression coefficients to zero, the Lasso estimator can get a sparse regression model. Since Lasso uses the same tuning parameter for all regression coefficients, the Lasso estimate produces biases for especially large coefficients (Fan and Li, 2001). The adaptive Lasso was introduced by Zou (2006) for linear regression

$$\hat{\boldsymbol{\beta}}^{alasso}(\lambda) = \text{argmin}_\beta \left(\text{RSS} + n \sum_{j=1}^{p} \lambda_j \left| \beta_j \right| \right) \tag{2.3}$$

in which adaptive weights are used for penalizing different coefficients.

It is well known that the OLS estimate is not robust to outliers or heavy-tailed errors in a response variable. Like the OLS estimate the Lasso suffers from unusual data points. To get a robust version of the Lasso estimator Wang *et al.* (2007) proposed a LAD-lasso that uses the sum of the absolute values of the residuals instead of RSS

$$\hat{\boldsymbol{\beta}}^{lad\text{-}lasso}(\lambda) = \text{argmin}_{\beta}\left(\sum_{i=1}^{n}\left|y_i - \mathbf{x}_i^T\boldsymbol{\beta}\right| + n\sum_{j=1}^{p}\lambda_j\left|\beta_j\right|\right). \tag{2.4}$$

The LAD estimator in linear regression and the LAD-Lasso estimator are both resistant to outliers or heavy-tailed errors; however, the LAD estimator is not robust to observations with unusual predictor values (Giloni *et al.*, 2006). Then he proposed a weighted LAD estimator in linear regression.

To alleviate the sensitivity to leverage points we consider a robust version of the Lasso estimator

$$\hat{\boldsymbol{\beta}}^{wlad\text{-}lasso}(\lambda) = \text{argmin}_{\beta}\left(\sum_{i=1}^{n} w_i(\mathbf{x}_i)\left|y_i - \mathbf{x}_i^T\boldsymbol{\beta}\right| + n\sum_{j=1}^{p}\lambda_j\left|\beta_j\right|\right), \tag{2.5}$$

where the weight w_i depends on the space of predictors. Giloni *et al.* (2006) suggested that the weight w_i was taken to be inversely proportional to the distance from the clean subset. We adopt the weight

$$w_i = \min\left(1, \frac{\chi^2_{q,0.05}}{\text{RD}_i^2}\right),$$

where $\chi^2_{q,0.05}$ is the upper 5% critical value of a chi-squared distribution with q degrees of freedom and RD_i^2 is a robust version of the Mahalanobis distance that can be written by $(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})$ (Croux *et al.*, 2003). Here $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ are a robust estimator of the mean vector and the covariance matrix for the predictors $\mathbf{x}_1, \ldots, \mathbf{x}_n$, respectively (Rousseeuw and Zomeren, 1990). The weighted LAD estimator will not be seriously influenced by leverage points, because the weights for them are alleviated as the value of the robust distance RD_i increases. The resulting estimator in (2.5) provides a sparse representation of a regression model and is also reliable to outliers or leverage points.

We can easily find the weighted LAD-Lasso in (2.5). We reformulate the data set $\{(y_i^*, \mathbf{x}_i^*)\}$ as

$$(y_i^*, \mathbf{x}_i^*) = \begin{cases} (w_i y_i, w_i\mathbf{x}_i), & \text{for } i = 1, \ldots, n, \\ (0, n\lambda_{i-n}\mathbf{e}_{i-n}), & \text{for } i = n+1, \ldots, n+p, \end{cases}$$

where $\mathbf{e}_j$ is the unit vector having 0 except the j^{th} element one (Wang *et al.*, 2007). Then the weighted LAD-Lasso estimator can be written by

$$\hat{\boldsymbol{\beta}}^{wlad\text{-}lasso}(\lambda) = \text{argmin}_{\beta}\sum_{i=1}^{n+p}\left|y_i^* - \mathbf{x}_i^*\boldsymbol{\beta}\right|. \tag{2.6}$$

Consequently, we can use a standard LAD program (the function *rq* in R program) without computation effort.

2.2. Statistical properties

Under mild conditions on the errors and the predictor variables in (2.1) (See assumptions A and B in Wang *et al.* (2007)), the weighted LAD-Lasso estimator in (2.6) yields the statistical properties as

the $\sqrt{n}$-consistency and the sparsity. The properties of the LAD-Lasso estimator can be preserved, because the weighted LAD-Lasso estimator can be modeled by $\tilde{y}_i = \tilde{\mathbf{x}}_i^T\boldsymbol{\beta} + w_i\epsilon_i$, where $\tilde{\mathbf{x}}_i = w_i\mathbf{x}_i$ and $\tilde{y}_i = w_i y_i$. Thus the statistical properties of the weighted LAD-Lasso estimation follow them of the LAD-Lasso estimation.

Let $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$ where p_0 is the length of $\boldsymbol{\beta}_1$. Without loss of generality, assume that $\boldsymbol{\beta}_2 = \mathbf{0}$. Its corresponding weighted LAD-Lasso estimator is denoted by $\hat{\boldsymbol{\beta}}^{wlad\text{-}lasso} = (\hat{\boldsymbol{\beta}}_1^{wlad\text{-}lasso\,T}, \hat{\boldsymbol{\beta}}_2^{wlad\text{-}lasso\,T})^T$. Furthermore let $a_n = \max\{\lambda_j|\beta_j \neq 0\}$ and $b_n = \min\{\lambda_j|\beta_j = 0\}$. The theorem in Wang *et al.* (2007) implies that the LAD-Lasso estimator has the same asymptotic distribution as the LAD estimator obtained under the true model. Therefore, Theorem 1 in Giloni *et al.* (2006) yields the following theorem.

Theorem 1. *Suppose that* $(\mathbf{x}_i, y_i), i = 1, \ldots, n$ *are independently identically distributed. Under the assumptions A and B in Wang et al. (2007) if* $\sqrt{n}a_n \to 0$, $\sqrt{n}b_n \to \infty$ *and* $\max_i w_i = O(1), \min_i w_i = O(1)$*, then the weighted LAD-Lasso estimator* $\hat{\boldsymbol{\beta}}^{wlad\text{-}lasso} = (\hat{\boldsymbol{\beta}}_1^{wlad\text{-}lasso\,T}, \hat{\boldsymbol{\beta}}_2^{wlad\text{-}lasso\,T})^T$ *satisfies that*

(i) $P(\hat{\boldsymbol{\beta}}_2^{wlad\text{-}lasso} = \mathbf{0}) \to 1$ *and*

(ii) $\sqrt{n}(\hat{\boldsymbol{\beta}}_1^{wlad\text{-}lasso} - \boldsymbol{\beta}_1)$ *is asymptotically* p_0*-variate normal with mean* $\mathbf{0}$ *and covariance matrix* $\mathbf{Q}^{-1}(\mathbf{X}^T\mathbf{W}\mathbf{X})\mathbf{Q}^{-1}/(4f(0)^2)$*, where* $\mathbf{Q} = \lim_{n\to\infty}\mathbf{X}^T\mathbf{W}\mathbf{X}/n$, $\mathbf{W} = \mathrm{diag}(w_i)$ *and* $f(t)$ *is the density function of* ϵ_i*.*

2.3. Tuning parameter

To find a good tuning parameters is an important issue in penalized estimation methods. The values of tuning parameters can be chosen by optimizing the performance via cross-validation and generalized cross validation (Fan and Li, 2001). Zou (2006) in the model (2.3) used the tuning parameters by the reciprocal of the absolute value of the OLS estimate. Wang *et al.* (2007) used the tuning parameter by minimizing a BIC-type objective function. The tuning parameter can be obtained by

$$\hat{\lambda}_j = \frac{\log n}{n|\tilde{\beta}_j|}, \tag{2.7}$$

where $\tilde{\beta}_j$ is the unpenalized LAD estimate for β_j in the regression model (2.1). We use the tuning parameter (2.7) where $\tilde{\beta}_j$ is the unpenalized weighted LAD estimate.

3. Simulation

In this section we conducted simulation in various situations to show the effectiveness of the weighted LAD-Lasso estimate which is resistant to heavy-tailed errors, outliers, or leverage points. We numerically compare the proposed method with the LAD-Lasso estimate, the adaptive Lasso estimate, and the best subset selection. All simulations are carried out using R codes.

We consider the model

$$y = \mathbf{x}^T\boldsymbol{\beta} + \sigma\epsilon,$$

where $\boldsymbol{\beta} = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$ and ϵ is generated from a heavy-tailed distribution. The component of $\mathbf{x}$ is a multivariate normal with mean $\mathbf{0}$ and the correlation ρ_{ij} between x_i and x_j, where $\rho_{ij} = 0.5^{|i-j|}$. We considered three types of error distributions: the standard normal, the standard double exponential

Table 1: Simulation results for t_3 errors and no leverage points

σ	n	Method	Correct	Incorrect	AMAD
1	50	wlad-lasso	3.48(1.17)	0.00(0.00)	0.310(0.118)
		lad-lasso	3.50(1.14)	0.00(0.00)	0.300(0.123)
		alasso	4.70(0.63)	0.02(0.14)	0.385(0.217)
		oracle	5.00(0.00)	0.00(0.00)	0.282(0.131)
	100	wlad-lasso	3.41(1.06)	0.00(0.00)	0.182(0.085)
		lad-lasso	3.34(1.06)	0.00(0.00)	0.180(0.082)
		alasso	4.84(0.40)	0.01(0.10)	0.233(0.147)
		oracle	5.00(0.00)	0.00(0.00)	0.176(0.085)
	200	wlad-lasso	3.31(1.11)	0.00(0.00)	0.141(0.055)
		lad-lasso	3.32(1.21)	0.00(0.00)	0.140(0.054)
		alasso	4.85(0.44)	0.00(0.00)	0.189(0.110)
		oracle	5.00(0.00)	0.00(0.00)	0.135(0.058)
$\sqrt{3}$	50	wlad-lasso	2.82(1.24)	0.01(0.10)	0.644(0.265)
		lad-lasso	2.97(1.04)	0.00(0.00)	0.629(0.265)
		alasso	4.63(0.68)	0.22(0.46)	0.791(0.469)
		oracle	5.00(0.00)	0.01(0.10)	0.537(0.248)
	100	wlad-lasso	3.11(1.24)	0.00(0.00)	0.370(0.145)
		lad-lasso	3.01(1.21)	0.00(0.00)	0.373(0.146)
		alasso	4.76(0.57)	0.05(0.22)	0.511(0.300)
		oracle	5.00(0.00)	0.00(0.00)	0.345(0.159)
	200	wlad-lasso	2.97(1.14)	0.00(0.00)	0.259(0.096)
		lad-lasso	2.91(0.98)	0.00(0.00)	0.256(0.094)
		alasso	4.75(0.54)	0.01(0.10)	0.355(0.159)
		oracle	5.00(0.00)	0.00(0.00)	0.235(0.093)

and t distribution with 3 degrees of freedom (t_3). Two different values for σ are tested for 1 and $\sqrt{3}$. The sample sizes are considered by $n = 50, 100$ and 200. The considered model is used in Tibshirani (1996) and Fan and Li (2001). We also consider the contaminated data with leverage points about 20% to show the robustness of the proposed estimator to leverage points.

Our simulation data consist of a training set and an independent test set. The regression coefficients in (2.1) are estimated for training data only, and the test error on the test data set is computed (where the sample size of a test data set is 1000). For each case, 100 simulation replications are carried out to evaluate the performance of the weighted LAD-Lasso estimate. The simulation results are summarized in Tables 1–4 that include the column labeled "Correct" presents the average number of correctly estimated zeros, and the column labeled "Incorrect" means the average number of coefficients erroneously set to zero in the same manner as done by Tibshirani (1996) and Wang *et al.* (2007). In addition, Tables 1–4 include the average of the mean absolute deviations(AMAD) evaluated on the test data set. The number in the parenthesis is the sample standard deviation.

The simulation results for t_3 errors with no leverage points is summarized in Table 1. It can be seen that the performance of the weighted LAD-Lasso (*wlad-lasso*) is similar to that of the LAD-Lasso (*lad-lasso*), because the weights w_i for the predictors may become 1 when the data set does not have leverage points. Even though the adaptive Lasso (*alasso*) is very efficient in respect to model complexity, the AMAD values of the adaptive Lasso are larger than them of the weighted LAD-Lasso and the LAD-lasso. In addition, the AMAD values of the weighted LAD-Lasso draw closer to the optimal AMAD values (*oracle*) as the sample size becomes larger regardless of the spread of the errors. The results on the correct number of zeros implies that the estimation methods based on the least absolute deviation errors are inclined to demonstrate overfitting effects and that the adaptive lasso is the best estimation on variable selection procedures with a little bit of underfitting effects. As expected the standard deviation of the AMAD values become larger when σ becomes larger.

Table 2: Simulation results for the standard normal errors and 20% leverage points

σ	n	Method	Correct	Incorrect	AMAD
$\sqrt{3}$	50	wlad-lasso	2.93(1.27)	0.16(0.42)	1.321(1.006)
		lad-lasso	1.49(0.95)	0.45(0.59)	3.164(0.195)
		alasso	3.08(1.13)	1.44(0.76)	3.163(0.225)
		oracle	5.00(0.00)	0.03(0.17)	0.681(0.313)
	100	wlad-lasso	3.11(1.21)	0.02(0.14)	0.715(0.672)
		lad-lasso	1.25(0.96)	0.30(0.56)	3.017(0.130)
		alasso	2.47(1.05)	0.94(0.75)	3.002(0.134)
		oracle	5.00(0.00)	0.00(0.00)	0.463(0.203)
	200	wlad-lasso	2.88(1.27)	0.01(0.10)	0.461(0.289)
		lad-lasso	1.11(0.82)	0.11(0.35)	2.927(0.090)
		alasso	1.79(1.16)	0.52(0.69)	2.905(0.090)
		oracle	5.00(0.00)	0.00(0.00)	0.367(0.138)

Table 3: Simulation results for the standard double exponential errors and 20% leverage points

σ	n	Method	Correct	Incorrect	AMAD
$\sqrt{3}$	50	wlad-lasso	2.76(1.14)	0.18(0.46)	1.226(1.042)
		lad-lasso	1.58(0.98)	0.52(0.63)	3.187(0.220)
		alasso	3.37(0.97)	1.57(0.79)	3.182(0.223)
		oracle	5.00(0.00)	0.02(0.14)	0.615(0.370)
	100	wlad-lasso	3.34(1.14)	0.01(0.10)	0.627(0.614)
		lad-lasso	1.28(0.94)	0.31(0.53)	3.054(0.142)
		alasso	2.61(1.05)	0.97(0.78)	3.044(0.140)
		oracle	5.00(0.00)	0.00(0.00)	0.416(0.186)
	200	wlad-lasso	3.20(1.17)	0.00(0.00)	0.512(0.531)
		lad-lasso	1.01(0.82)	0.14(0.40)	2.923(0.098)
		alasso	2.14(1.14)	0.78(0.77)	2.925(0.102)
		oracle	5.00(0.00)	0.00(0.00)	0.329(0.140)

We next conducted the simulation for three types of errors with 20% leverage points. The results for $\sigma = 1$ and $\sqrt{3}$ are similar. We summarized the results in Tables 2 and 4 for only $\sigma = \sqrt{3}$, because the data for $\sqrt{3}$ are much contaminated by regression outliers or leverage points.

Table 2 summarizes the simulation results for the standard normal errors and 20% leverage points. Seeing "Correct" term implies that in model complexity the weighted LAD-lasso and the adaptive Lasso are comparable when the sample size is small. However, the weighted LAD-lasso method is better than the adaptive Lasso method as the sample size increases. If we only consider the model complexity, the LAD-lasso is the worst sparse estimation among three estimations. The adaptive lasso demonstrates underfitting effects in the sample size 50, because the average number of incorrect zeros is 1.44 in Table 2. Even though the weighted LAD-Lasso has underfitting effects in small sample size which is the smallest effects among three methods, the effect disappears when the sample size becomes larger. Table 2 shows that in model error the weighted LAD-Lasso is the most efficient estimation regardless of the spread of errors and the sample size, since we considered the estimation method reducing the influence of leverage points.

Table 3 presents the simulation results for the standard double exponential errors and 20% leverage points are summarized. In addition, Table 4 considers the t_3 errors and 20% leverage points. The results are similar to Table 2. Thus, the weighted LAD-Lasso performs amazingly well for thin- or thick-tailed errors even when the predictor variables have leverage points.

Tables 1 to 4 shows that the weighted LAD-Lasso is very robust to heavy-tailed errors and leverage points. Especially the weighted LAD-Lasso presents high prediction accuracy among compared estimators. When the data with large sample size have leverage points, the weighted LAD-Lasso provides

Table 4: Simulation results for t_3 errors and 20% leverage points

σ	n	Method	Correct	Incorrect	AMAD
$\sqrt{3}$	50	wlad-lasso	2.56(1.21)	0.22(0.46)	1.475(1.096)
		lad-lasso	1.71(0.98)	0.55(0.66)	3.207(0.226)
		alasso	3.54(0.95)	1.72(0.79)	3.286(0.274)
		oracle	5.00(0.00)	0.06(0.28)	0.714(0.377)
	100	wlad-lasso	3.05(1.13)	0.03(0.17)	0.587(0.597)
		lad-lasso	1.35(0.86)	0.29(0.50)	2.988(0.167)
		alasso	2.83(1.04)	1.15(0.85)	3.009(0.186)
		oracle	5.00(0.00)	0.00(0.00)	0.518(0.227)
	200	wlad-lasso	2.74(1.15)	0.01(0.10)	0.570(0.445)
		lad-lasso	1.00(0.71)	0.08(0.27)	2.935(0.097)
		alasso	2.26(1.09)	0.79(0.74)	2.951(0.124)
		oracle	5.00(0.00)	0.00(0.00)	0.404(0.157)

a sparse model rather than the adaptive Lasso. Thus, the weighted LAD-Lasso is a best estimator in the sense of prediction accuracy and model complexity.

4. Concluding Remarks

In this paper we proposed a robust estimator based on a weighted least absolute deviation criterion with penalizing the l_1 norm of regression coefficients. We show good performance from simulation under various situations by combining the tail shape of errors, the strength of spread of errors and leverage points. Especially the proposed estimator is very robust to the contaminated data by heavy tailed errors, outliers, or leverage points.

References

Croux, C., Filzmoser, P., Pison, G. and Rousseeuw, P. J. (2003). Fitting multiplicative models by robust alternating regressions, *Statistics and Computing*, **13**, 23–36.

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

Giloni, A., Simonoff, J. S. and Sengupta, B. (2006). Robust weighted LAD regression, *Computational Statistics and Data Analysis*, **50**, 3124–3140.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems, *Technometrics*, **12**, 55–67.

Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators, *Econometric Theory*, **7**, 186–199.

Rousseeuw, P. J. and van Zomeren, B. C. (1990). Unmasking multivariate outliers and leverage points, *Journal of the American Statistical Association*, **85**, 633–639.

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

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

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso, *Journal of Machine Learning Research*, **7**, 2541–2563.

Zou, H. (2006) The adaptive Lasso and its oracle properties, *Journal of the American Statistical Association*, **101**, 1418–1429.

Received September 2011; Accepted October 2011