NEW ROBUST LASSO METHOD BASED ON RANKS

Hyon-Jung Kim[⋆,†]

[⋆]University of Tampere
Department of Information Science
FI-33014 University of Tampere

Esa Ollila[†] *and Visa Koivunen*[†]

[†]Aalto University
Dept. of Signal Processing and Acoustics
FI-00076 Aalto, Finland

ABSTRACT

The LASSO (Least Absolute Shrinkage and Selection Operator) has been a popular technique for simultaneous linear regression estimation and variable selection. Robust approaches for LASSO are needed in the case of heavy-tailed errors or severe outliers. We propose a novel robust LASSO method that has a non-parametric flavor: it solves a criterion function based on ranks of the residuals with LASSO penalty. The criterion is based on pairwise differences of residuals in the least absolute deviation (LAD) loss leading to a bounded influence function. With the ℓ_1-criterion we can easily incorporate other penalties such as fused LASSO for group sparsity and smoothness. For both methods, we propose efficient algorithms for computing the solutions. Our simulation study and application examples (image denoising, prostate cancer data analysis) show that our method outperform the usual LS/LASSO methods for either heavy-tailed errors or outliers, offering better variable selection than another robust competitor, LAD-LASSO method.

Index Terms— LASSO, penalized regression, sparse regression, group sparsity, robust

1. INTRODUCTION

We consider the classic linear model $\mathbf{y} = \alpha\mathbf{1}_n + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbf{X} = \begin{pmatrix}\mathbf{x}_1 & \cdots & \mathbf{x}_n\end{pmatrix}^\top$ is a known full rank $n \times p$ design matrix (matrix of predictors), and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ is the unknown vector of regression coefficients with the unknown intercept term $\alpha \in \mathbb{R}$. The primary interest is to estimate the unknown parameters, $\boldsymbol{\beta}$ and α where $\mathbf{y} \in \mathbb{R}^n$ is the observed response and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ denotes the additive noise. However, in many practical applications, the linear system is *underdetermined* ($p > n$) or $p \approx n$ and the least squares estimate (LSE) can not be computed (non-unique solutions) or is subject to a very high-variance. Another problem with the LSE arises when there are outliers or the noise exhibits a heavy-tailed non-Gaussian feature. The usual solution to ill-posed systems is to regularize the regression coefficients (i.e. control how large they can grow).

After centering the response and the predictors, the popular LASSO [1] solves ℓ_1-penalized LS regression problem,

$$\hat{\boldsymbol{\beta}}_{\ell_2}(\lambda) = \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1 \qquad (1)$$

where λ is the *shrinkage (penalty) parameter*, $\|\cdot\|$ denotes the usual Euclidean (ℓ_2-)norm on vectors and $\|\cdot\|_1$ denotes the ℓ_1-norm, i.e., $\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^p |\beta_j|$. Note that Lasso is known as *basis pursuit* [2] in signal processing which is commonly expressed via constrained counterpart of equation (1) in the form $\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_1$ s.t. $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 \leq \gamma$, where γ depends on λ. As $\lambda \in [0, \infty)$ varies, the solution $\hat{\boldsymbol{\beta}}_{\ell_2}(\lambda)$ traces out a path in $\mathbb{R}^p$, with $\hat{\boldsymbol{\beta}}_{\ell_2}(0)$ corresponding to the conventional LSE. Note that (1) does not involve the intercept α, which is computed after computing $\hat{\boldsymbol{\beta}}(\lambda)$. The larger the value of λ the greater is the amount of shrinkage for the coefficients which are shrunk all the way to zero.

Several robust LASSO regression approaches have been proposed. For example, [3] advocates penalizing the LAD objective function (LAD-LASSO) and [4] uses the least trimmed squares criterion (LTS-LASSO), whereas [5,6] consider a penalized M-estimator based on Huber's loss function. In this paper, we propose a penalized rank regression estimator (Rank-LASSO) based on the rank dispersion function with Wilcoxon scores [7]. This new objective function corresponds to LAD objective function of pairwise differences of the residuals. It offers improved efficiency under normality and better model selection performance compared to the LAD-LASSO. Recently penalties that enforce smoothness and group/block sparsity [8,9] have received a lot of attention and we extend our approach to the case of using fused Lasso penalty [8].

The paper is organized as follows. In Section 2, we illustrate our new Rank-LASSO method in detail. In particularly, we address the important problem of how the path of solutions evolve for λ ranging over a grid from $[0, \lambda_M]$. λ_M denotes the smallest value of λ that shrinks all the coefficients to zero, and is computed efficiently with the optimal selection of the penalty parameter using the *Bayes Information Criterion* (BIC) [10]. Section 3 introduces an extension to Fused LASSO penalty and data applications and simulation study follow in Section 4. Section 5 concludes.

978-0-9928626-3-3/15/$31.00 ©2015 IEEE

2. RANK-LASSO

The LAD-LASSO [3] is a simple popular choice for the robust LASSO method which uses the penalized LAD cost function of residuals: $\hat{\boldsymbol{\beta}}_{\ell_1}(\lambda) = \arg\min_{\boldsymbol{\beta}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_1 + \lambda\|\boldsymbol{\beta}\|_1$, where the responses and the predictors are assumed to be centered: the predictor variables $\mathbf{x}_{i1}$'s are centered to have zero mean, whereas y_i's are centered to have median equal to zero. Note that $\hat{\boldsymbol{\beta}}_{\ell_1}(0)$ is the conventional LAD regression estimator. However, the LAD-loss is well-known to be inefficient under normality having 63.6 % efficiency.

In this paper, we propose a new robust extension of LASSO based on the pairwise differences of the residuals. It has a bounded influence function, yet attaining high efficiency under normality:

$$\hat{\boldsymbol{\beta}}_R(\lambda) = \arg\min_{\boldsymbol{\beta}} \sum_{i<j} |e_i - e_j| + \lambda\|\boldsymbol{\beta}\|_1, \qquad (2)$$

where $e_i = e_i(\boldsymbol{\beta}) = y_i - \mathbf{x}_i^\top\boldsymbol{\beta}$ denotes the ith residual for a candidate $\boldsymbol{\beta}$ and the summation ranges over all $N = n(n-1)/2$ pairwise differences ($1 \le i < j \le n$). With the pairwise differences of the residuals in the objective function, our method does not require any preprocessing steps such as centering of the responses and predictors. Note that the utilized objective function is well-known in nonparametric statistics. It co-incides with Jaeckel's (1972) [7] rank regression dispersion function for Wilcoxon scores: $\sum_{i=1}^n R(e_i)e_i = (1/2)\sum_{i<j}|e_i - e_j|$, where $R(e_i)$ is the centered rank (Wilcoxon score) of e_i among the residuals $e_1, \ldots, e_n$.

Note that the objective function (2) can be written as

$$\sum_{i<j} |y_i - y_j - (\mathbf{x}_i - \mathbf{x}_j)^\top\boldsymbol{\beta}| = \sum_{i=1}^N |\tilde{y}_i - \tilde{\mathbf{x}}_i^\top\boldsymbol{\beta}|$$

where

$$\tilde{\mathbf{y}} = \begin{pmatrix} y_1 - y_2 \\ \vdots \\ y_{n-1} - y_n \end{pmatrix}, \quad \tilde{\mathbf{X}} = \begin{pmatrix} (\mathbf{x}_1 - \mathbf{x}_2)^\top \\ \vdots \\ (\mathbf{x}_{n-1} - \mathbf{x}_n)^\top \end{pmatrix}, \qquad (3)$$

are based on pairwise differences of y_i's/$\mathbf{x}_i$'s. If we define the *augmented measurements* as

$$\tilde{\mathbf{y}}_a = \begin{pmatrix} \tilde{\mathbf{y}} \\ \mathbf{0} \end{pmatrix}, \quad \tilde{\mathbf{X}}_a(\lambda) = \begin{pmatrix} \tilde{\mathbf{X}} \\ \lambda\mathbf{I}_p \end{pmatrix},$$

where $\mathbf{0}$ is a p-vector of zeros and $\mathbf{I}$ denotes an $p \times p$ identity matrix, then it is easy to verify that penalized objective function (2) can be recast into a LAD criterion: $\hat{\boldsymbol{\beta}}_R(\lambda) = \arg\min_{\boldsymbol{\beta}} \|\tilde{\mathbf{y}}_a - \tilde{\mathbf{X}}_a(\lambda)\boldsymbol{\beta}\|_1$. Since this is a convex optimization problem, its global minimizer can be found efficiently; see for [11–14] different approaches. Furthermore, uniqueness properties of LAD estimator are inherited by Rank-Lasso.

Let us now discuss estimation of the intercept α. In the penalized regression, the first preprocessing step is to center the response and the predictor variables, so that the model has no intercept term. For example $\hat{\boldsymbol{\beta}}_{\ell_2}(\lambda)$ in (1) is solved for centered data ($y_i \leftarrow y_i - \bar{y}$ and $\mathbf{x}_i \leftarrow \mathbf{x}_i - \bar{\mathbf{x}}$ for $i = 1, \ldots, n$), and the intercept is estimated at the last stage by minimizing the non-penalized LS criterion function, $\hat{\alpha}_{\ell_2}(\lambda) = \arg\min_\alpha \sum_i (\hat{e}_i(\lambda) - \alpha)^2$. The *non-centered* residuals are denoted by $\hat{e}_i(\lambda) = y_i - \mathbf{x}_i^\top\hat{\boldsymbol{\beta}}_{\ell_2}(\lambda)$, where y_i and $\mathbf{x}_i$ are not centered when computing the intercept.

In Rank-LASSO, centering the data is not needed, because the objective function is based on the pairwise differences of residuals with which the intercept term cancels out naturally. Here, the above approach of finding the intercept as a minimizer of the non-penalized ℓ_1 objective function based $\hat{e}_i - \hat{e}_j$ is not possible. Therefore, instead of using pairwise differences, we use the pairwise averages, and compute the intercept as

$$\hat{\alpha}_R(\lambda) = \arg\min_\alpha \sum_{i<j} \left| \frac{\hat{e}_i(\lambda) + \hat{e}_j(\lambda)}{2} - \alpha \right|$$

where $\hat{e}_i(\lambda) = y_i - \mathbf{x}_i^\top\hat{\boldsymbol{\beta}}_R(\lambda)$. Hence the solution $\hat{\alpha}_R$ is the Hodges-Lehmann median of the estimated residuals $\hat{e}_i$.

Next, we focus on finding an optimal λ and the path of solutions $\hat{\boldsymbol{\beta}}(\lambda)$. Since $\hat{\boldsymbol{\beta}}(\lambda)$ does not have a closed-form expression, it is common to compute $\hat{\boldsymbol{\beta}}(\lambda)$ in a grid of λ values, $[\lambda] = (\lambda_0, \lambda_1, \ldots, \lambda_M)$, where λ_0 is a small value close to zero (no penalty) and λ_M is commonly chosen to be the smallest value of λ that shrinks all the coefficients to zero. Then, the coefficient paths $\hat{\beta}_i(\lambda)$ is displayed over the grid of λ values, for example, in Figure 1 for a data set with $p = 8$ regression coefficients. For the Rank-LASSO, the value λ_M can be computed as

$$\lambda_M = \|\tilde{\mathbf{X}}^\top \mathrm{sign}(\tilde{\mathbf{y}})\|_\infty. \qquad (4)$$

where $\|\cdot\|_\infty$ denotes the ℓ_∞-norm on vectors (e.g. $\|\mathbf{a}\|_\infty = \max_i |a_i|$), and $\mathrm{sign}(\tilde{\mathbf{y}})$ denotes a vector of marginal signs of $\tilde{\mathbf{y}}$, i.e., its ith component is $\mathrm{sign}(\tilde{y}_i)$. The sign function is defined as $\mathrm{sign}(x) = -1, 1, 0$, if $x < 0, > 0, = 0$, respectively. We compute $\hat{\boldsymbol{\beta}}_R(\lambda)$ using the pathwise coordinatewise descent algorithm [14, p. 306] or the iteratively re-weighted LS algorithm [12] (IRWLS) as both algorithms permit using an initial guess of iteration. A natural initial guess for computing $\hat{\boldsymbol{\beta}}_R(\lambda_i)$ is the previously computed value $\hat{\boldsymbol{\beta}}_R(\lambda_{i-1})$. If the columns of $\mathbf{x}_i$ are incoherent (i.e., having modest correlations), then coordinatewise optimization approach often provides the fastest computation. Similar results were also reported in [14].

After computing a path of solutions over a grid, we select the optimal value in the grid $[\lambda]$ as the one minimizing BIC,

$$\lambda^\star = \arg\min_{\lambda \in [\lambda]} 2n \ln\hat{\sigma}(\lambda) + \mathrm{df}(\lambda) \cdot \ln n,$$

where the number of the degrees of freedom of the model, $\mathrm{df}(\lambda)$ is defined as the number of nonzero elements in $\hat{\boldsymbol{\beta}}(\lambda)$, and $\hat{\sigma}(\lambda)$ denotes the scale estimate of the error terms. For Rank-LASSO estimate, the natural scale statistics is Gini's mean difference of residuals, $\hat{\sigma}(\lambda) = \frac{1}{N}\sum_{i<j}|\hat{e}_i(\lambda) - \hat{e}_j(\lambda)|$. The Rank-LASSO estimate based on BIC is then $\hat{\boldsymbol{\beta}}^{\star}_R = \hat{\boldsymbol{\beta}}_R(\lambda^{\star})$ and $\hat{\alpha}^{\star}_R = \hat{\alpha}_R(\lambda^{\star})$.

3. FUSED RANK-LASSO

In order to enforce block-sparsity and smoothness, we introduce a rank-based extension of the fused LASSO (FL) criterion of [8]. This leads to a new penalized criterion:

$$\min_{\boldsymbol{\beta}} \sum_{i<j}|e_i - e_j| + \lambda\|\boldsymbol{\beta}\|_1 + \lambda_2\sum_{j=2}^{p}|\beta_j - \beta_{j-1}|,$$

where $\lambda_1, \lambda_2 \geq 0$ form a pair of fixed regularization parameters. Note that again the optimization problem is convex and hence a global solution $\hat{\boldsymbol{\beta}}_{\mathrm{R}}(\lambda_1, \lambda_2)$, referred to as Rank-FLASSO, can be computed efficiently. Note that if $\lambda_2 = 0$, then we obtain the Rank-LASSO solution presented in Section 2. FL-penalty encourages flatness of the magnitudes as a function of j and the local constancy of the coefficient profile. Important practical applications of FL-penalization can found in areas such as protein mass spectroscopy, microarray gene expression [8] or in image denoising. We provide an example in Section 4, illustrating the effectiveness of our method in image denoising.

The benefit of our ℓ_1 criterion (in contrast to ℓ_2-criterion used in [8]) is the flexibility of FL-penalty that can be incorporated into our Rank(LAD) framework with minimal programming effort. Namely, let us define a $(p-1)\times p$ matrix $\mathbf{X}_f$ as $\mathbf{X}_f = (\mathbf{u}_2 - \mathbf{u}_1 \ \cdots \ \mathbf{u}_p - \mathbf{u}_{p-1})^{\top}$, where $\mathbf{u}_i$'s denote basis vectors of $\mathbb{R}^p$ (having a 1 at its ith element and 0's elsewhere). Let the *augmented fused measurements* be

$$\tilde{\mathbf{y}}_{fa} = \begin{pmatrix}\tilde{\mathbf{y}}_a \\ \mathbf{0}\end{pmatrix}, \quad \tilde{\mathbf{X}}_{fa}(\lambda_1,\lambda_2) = \begin{pmatrix}\tilde{\mathbf{X}}_a(\lambda_1) \\ \lambda_2\mathbf{X}_f\end{pmatrix},$$

where $\mathbf{0}$ is a $(p-1)$-vector of zeros. It can be easily verified that $\hat{\boldsymbol{\beta}}_R(\lambda_1,\lambda_2) = \arg\min_{\boldsymbol{\beta}}\|\tilde{\mathbf{y}}_{fa} - \tilde{\mathbf{X}}_{fa}(\lambda_1,\lambda_2)\boldsymbol{\beta}\|_1$, i.e., Rank-FLASSO solves a standard LAD criterion based on the augmented fused measurements.

4. NUMERICAL EXAMPLES

4.1. Prostate cancer data

We consider the benchmark prostate cancer data set ($n = 97, p = 8$) used in many text-books; see [15]. The interest is in exploring a relationship between measurements on the level of prostate-specific antigen with a number of clinical measures in men who were about to receive a radical prostatectomy. The predictor variables, denoted `lcavol`, `lweight`, `age`, `lbph`, `svi`, `lcp`, `gleason`, `pgg45`, are explained in [15].

The first column of Figure 1 shows the coefficient paths of $\hat{\boldsymbol{\beta}}(\lambda)$ as λ ranges from $(0, \lambda_M)$ for the LASSO, LAD-LASSO and Rank-LASSO. We tested the effect of an outlier by changing y_1 to $y_1^* = 10\max(|y_i|_{i=1}^n)$ and recomputing the estimates. Thus, we have only a single (vertical) outlier of moderate size in the data set. The second column of Figure 1 shows the coefficient paths when an outlier is present. It is instructive to compare LAD-LASSO and Rank-LASSO since both are based on ℓ_1 criterion. As can be seen LAD approach yields coefficient paths that are non-monotone and highly non-smooth with a visible zigzag feature. One can observe that in the outlier-free case (left plot), Rank coefficient paths can be described as smoother and monotone versions of LAD coefficient paths. However, when an outlier is present, the LAD coefficient paths have changed more than Rank coefficient paths. Although both methods are robust, Rank is *more stable locally*, i.e., a small change in λ does not imply large effect on the solution. This example illustrates the increased stability and robustness of Rank-LASSO compared to LAD-LASSO.

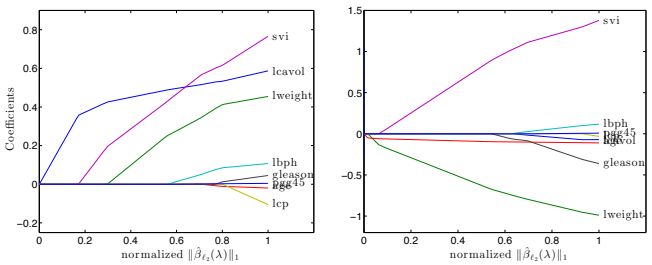
The LSE and the LASSO solutions chosen by BIC method are shown below:

	LSE	LASSO	Rank	LAD
intercept	.669	.355	1.191	.273
`lcavol`	.587	.516	.568	.511
`lweight`	.454	.345	.399	.492
`age`	-.020		-.019	-.022
`lbph`	.107	.050	.126	.123
`svi`	.766	.566	.605	.760
`lcp`	-.105		-.078	-.096
`gleason`	.045			.119
`pgg45`	.005	.001	.006	.004

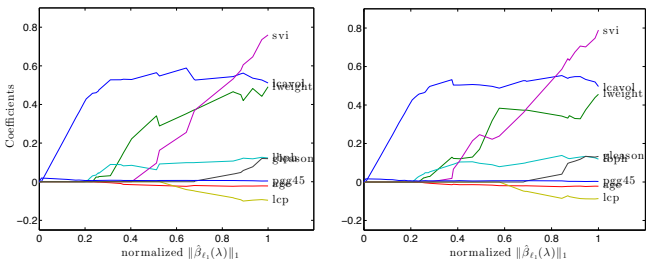
Note that both LASSO and Rank-LASSO select `gleason` as a non-significant predictor, but the BIC solution for LAD-LASSO is very conservative: the smallest value λ_0 on the grid (no shrinkage) was chosen as the optimal value. The BIC may not work well with LAD-LASSO due to its non-smooth coefficient paths: a very dense grid of penalty values should be chosen in order to capture the rapidly changing differences in solutions.

4.2. Image denoising example

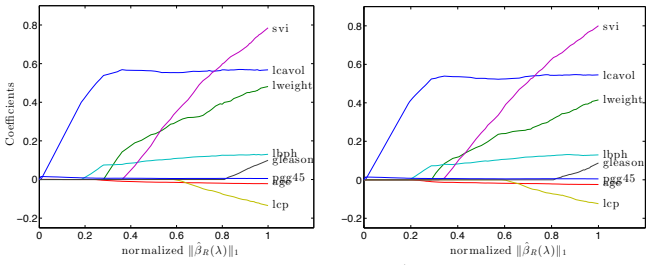
Figure 2(a) shows a signal $\mathbf{s}$ of $n = 400$ measurements obtaining values $\{0, 1, 2, 3\}$ in blocks of varying length and the noisy signal $\mathbf{y} = \mathbf{s} + \mathbf{n}$ on top right panel. Note that the noise-free signal $\mathbf{s}$ is sparse, 43.75% of the measurements are equal to zero. The data sets are reshaped into 20×20 gray-value images shown at the bottom plots. Then, given the knowledge of the noisy signal/image alone, the objective is to find a good approximation of the original noise-free signal/image. Due to block-sparse nature of the signal, the FL-penalty can offer ef-



(a) LASSO $\hat{\boldsymbol{\beta}}_{\ell_2}(\lambda)$



(b) LAD-LASSO $\hat{\boldsymbol{\beta}}_{\ell_1}(\lambda)$



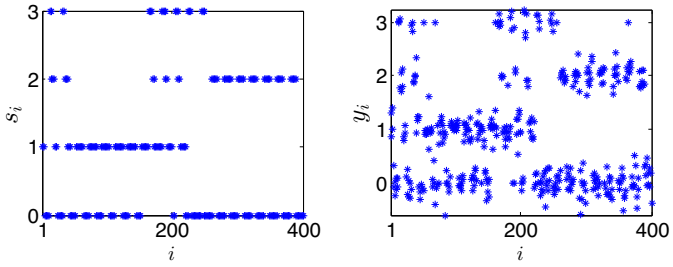
(c) Rank-LASSO $\hat{\boldsymbol{\beta}}_R(\lambda)$

Fig. 1. Coefficient paths for prostate cancer data without outlier (1st column) and with an outlier (2nd column).

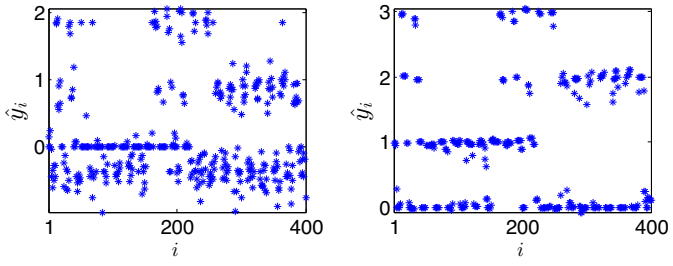
ficient signal/image denoising. We computed the LASSO and fused Rank-LASSO signal approximations given the knowledge of noisy signal $\mathbf{y}$ and the prediction matrix $\mathbf{X} = \mathbf{I}_n$. The best signal approximation obtained by the LASSO $\hat{\mathbf{s}}_{\ell_2}(\lambda)$ and the fused Rank-LASSO $\hat{\mathbf{s}}_R(\lambda_1, \lambda_2)$ are shown at the bottom panels of Figure 2. For both of the methods, we did an extensive grid search of penalty parameters λ and (λ_1, λ_2) and the denoised images in Figure 2(c), (d) are the solutions that had the smallest mean squared error (MSE) between the solution $\hat{\mathbf{s}}(\lambda)$ and the true noise-free signal $\mathbf{s}$. The minimum MSE was 0.0179 for Rank-FLASSO and 0.7013 for LASSO. This drastic difference in denoising is due to the fact that fused Rank-LASSO has successfully exploited the spatial smoothness (block sparsity). As the result it provided significantly better signal approximation than LASSO, as is well illustrated in Figure 2.

4.3. Variable selection

In the simulation, the covariates x_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, p$ are generated as i.i.d. normal random variates. We



(a) noise free data $\mathbf{s}$ (b) observed noisy data $\mathbf{y}$



(c) LASSO $\hat{\mathbf{s}}(\lambda)$ (d) Rank-FLASSO $\hat{\mathbf{s}}(\lambda_1, \lambda_2)$

Fig. 2. The signal $\mathbf{s}$ in (a) represents a vectorized gray-scale image of squares; (b) signal $\mathbf{s}$ with added noise $\mathbf{n}$. The denoised images shown in (c) and (d) are optimal solutions having smallest MSE with the original signal over extensive grid search of penalty parameters. Rank-FLASSO obtained much better approximation ($10\times$ smaller MSE than LASSO).

set $p = 15$ and $n = 75$, or 200. The coefficient vector $\boldsymbol{\beta} = (\beta_i)$ is given by $\beta_1 = 1.5$, $\beta_2 = 2.0$, $\beta_3 = 2.5$ and $\beta_i = 0$ for $4 \leq \beta_i \leq p$. The measurement vector $\mathbf{y}$ is generated according to the linear model (with intercept $\alpha = 0$) where the errors are from either the Gaussian distribution $\mathcal{N}(0, \sigma^2)$ or the Cauchy distribution $\mathrm{Cau}(0, \sigma)$. In the former case σ is the variance and in the latter case (as the variance does not exist for Cauchy) the median absolute deviation (MAD) $\sigma_{\text{MAD}} = \mathrm{Med}(|\varepsilon_i|)$. In both cases, we fix $\sigma = 0.1$,

	Method	RMSPE	CMS	FPR	FNR
$n = 75$	Oracle	.100	1	0	0
	LASSO	.107	.32	.13	0
	LAD-LASSO	.110	.09	.26	0
	Rank-LASSO	.108	.29	.15	0
$n = 200$	Oracle	.1004			
	LASSO	.1030	.45	.08	0
	LAD-LASSO	.1038	.20	.16	0
	Rank-LASSO	.1031	.42	.09	0
$n = 75$	Oracle	0.099			
	LASSO	0.918	.34	0.10	0.13
	LAD-LASSO	0.208	.56	0.05	0.01
	Rank-LASSO	0.209	.64	0.04	0.01
$n = 200$	Oracle	0.100			
	LASSO	0.947	0.46	0.05	0.15
	LAD-LASSO	0.109	0.80	0.02	0
	Rank-LASSO	0.114	0.90	0.01	0

Table 1. Root mean squared prediction error (MSPE), the false positive rate (FPR), the false negative rate (FNR) and percentage of correct model selection, averaged over 250 runs, for Gaussian (top) and Cauchy (bottom) errors.

moderate signal to noise ratio (SNR) $= 6$ dB. We use BIC also for LASSO and LAD-LASSO for which the natural scale statistic $\hat{\sigma}(\lambda)$ is the sample standard deviation of the residuals, $\hat{\sigma}(\lambda) = (\frac{1}{n}\sum_i \hat{e}_i(\lambda)^2)^{1/2}$, and the sample mean absolute deviation $\hat{\sigma}(\lambda) = \frac{1}{n}\sum_i |\hat{e}_i(\lambda)|$, respectively.

We evaluate the estimators by their *root mean squared prediction error*, $\text{RMSPE}(\hat{\boldsymbol{\beta}}) = (\frac{1}{n}\sum_{i=1}^{n}(y_i^* - (\mathbf{x}_i^*)^\top\hat{\boldsymbol{\beta}})^2)^{1/2}$ as in [4], where an additional test data set $(\mathbf{y}^*, \mathbf{X}^*)$ is generated from the respective sampling schemes (without outliers) for each MC trial. The RMSPE of the oracle estimator, which uses the true coefficient vector $\boldsymbol{\beta}$, is computed as a point of reference for the evaluated methods. To assess the performance of model selections, we examine the false positive rate (FPR), the false negative rate (FNR) and the percentage of the correct model selection (CMS). (A false positive indicates that the coefficient whose value is zero in the true model is estimated to be non-zero.) In the Gaussian case, in terms of RMSPE and CMS the LASSO gives only slightly better performance than Rank-LASSO, but Rank-LASSO has much better performance than LAD-LASSO. With the Cauchy errors, the LASSO clearly gives the worst performance, whereas the robust methods show similar results in terms of root MSPE. Concerning the correct model selection, for $n = 200$, the Rank-LASSO had correctly identified the true model 90 % time of the trials, which significantly outperforms the LAD-LASSO (80 %).

5. CONCLUSIONS

We developed a novel robust LASSO (and fused LASSO) estimator based on Jaeckel's rank dispersion function with efficient Wilcoxon scores. This method is robust and efficient even under Gaussian errors. Simulation studies illustrate that our method offers excellent model selection performance both in Gaussian and Cauchy errors. In regard to selecting the correct model, it always outperformed the LAD-LASSO. Another advantage over LAD-LASSO is that it provides smooth paths of th regression coefficients. This also supports the fact why BIC model selection works well with Rank-LASSO but not with LAD-LASSO.

REFERENCES

[1] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Royal Stat. Soc., Ser. B*, vol. 58, pp. 267–288, 1996.

[2] S.S. Chen, D. L. Donoho, and M. A. Saunders, "Atomic decomposition by basis pursuit," *SIAM J. Sci. Comput.*, vol. 20, no. 1, pp. 33–61, 1998.

[3] H. Wang, G. Li, and G. Jiang, "Robust regression shrinkage and consistent variable selection through the LAD-Lasso," *J. Bus. Econ. Stat.*, vol. 25, pp. 347–355, 2007.

[4] A. Alfons, C. Croux, and S. Gelper, "Sparse least trimmed squares regression for analyzing high-dimensional large data sets," *Ann. Appl. Stat.*, vol. 7, no. 1, pp. 226–248, 2013.

[5] A. B. Owen, "A robust hybrid of lasso and ridge regression," *Contemporary Mathematics*, vol. 443, pp. 59–72, 2007.

[6] X. Chen, J. Wang, and M.J. McKeown, "Asymptotic analysis of robust LASSOs in the presence of noise with large variance," *IEEE Trans. Signal Process.*, vol. 56, no. 10, pp. 5131–5149, 2010.

[7] L. A. Jaeckel, "Estimating regression coefficients by minimizing the dispersion of the residuals," *The Annals of Mathematical Statistics*, vol. 43, no. 5, pp. 1449–1458, 1972.

[8] M. Tibshirani, R. Saunders, S. Rosset, J. Zhu, and K. Knight, "Sparsity and smoothness via the fused lasso," *J. Royal Stat. Soc., Ser. B*, vol. 67, no. 1, pp. 91–108, 2005.

[9] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *J. Royal Stat. Soc., Ser. B*, vol. 68, no. 1, pp. 49–67, 2006.

[10] G. Schwarz, "Estimating the dimension of a model," *Ann. Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[11] I Barrodale and F. D. K. Roberts, "An improved algorithm for discrete ℓ_1 linear approximation," *SIAM Journal on Numerical Analysis*, vol. 10, no. 5, pp. 839–848, 1973.

[12] E. J. Schlossmacher, "An iterative technique for absolute deviations curve fitting," *J. Amer. Stat. Assoc.*, vol. 68, no. 344, pp. 857–859, 1973.

[13] Y. Li and G.R. Arce, "A maximum likelihood approach to least absolute deviation regression," *EURASIP J. Adv. Signal Process.*, vol. 2004, pp. 1762–1769, 2004.

[14] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani, "Pathwise coordinate optimization," *Ann. Appl. Stat.*, vol. 1, no. 2, pp. 302–332, 2007.

[15] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning*, Springer, New York, 2001.