The Outlier-Corrected-Data-Adaptive Lasso: A New Robust Estimator for the Independent Contamination Model

Jasin Machkour[1], Bastian Alt[2], Michael Muma[1], Abdelhak M. Zoubir[1]

[1] Signal Processing Group
Technische Universität Darmstadt
Merckstr. 25, 64283 Darmstadt, Germany
Email: {machkour,muma,zoubir}@spg.tu-darmstadt.de

[2] Bioinspired Communication Systems
Technische Universität Darmstadt
Rundeturmstr. 12, 64283 Darmstadt, Germany
bastian.alt@bcs.tu-darmstadt.de

***Abstract*—Many of today's signal processing tasks consider sparse models where the number of explanatory variables exceeds the sample size. When dealing with real-world data, the presence of impulsive noise and outliers must also be accounted for. Accurate and robust parameter estimation and consistent variable selection are needed simultaneously. Recently, some popular robust methods have been adapted to such complex settings. Especially, in high dimensional settings, however, it is possible to have a single contaminated predictor being responsible for many outliers. The amount of outliers introduced by this predictor easily exceeds the breakdown point of any existing robust estimator. Therefore, we propose a new robust and sparse estimator, the Outlier-Corrected-Data-(Adaptive) Lasso (OCD-(A) Lasso). It simultaneously handles highly contaminated predictors in the dataset and performs well under the classical contamination model. In a numerical study, it outperforms competing Lasso estimators, at a largely reduced computational complexity compared to its robust counterparts.**

I. Introduction

Many problems in data analysis and signal processing require sparse model estimation to handle large datasets in terms of model interpretation, including the case, where the number of explanatory variables p is larger than the sample size n. In these settings accurate parameter estimation and consistent variable selection are needed simultaneously. A classical method is the least absolute shrinkage and selection operator (Lasso) [1]

$$\hat{\boldsymbol{\beta}}_{\text{lasso}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1, \tag{1}$$

where $\mathbf{y} \in \mathbb{R}^n$ is the vector of responses, $\mathbf{X} \in \mathbb{R}^{n\times p}$ the predictor matrix with n denoting the number of observations and p the number of predictors, $\boldsymbol{\beta} \in \mathbb{R}^p$ the vector of parameters to be estimated and λ a non-negative real number. Based on the choice of the tuning parameter λ, this approach shrinks the coefficients towards zero, which allows for a bias-variance trade-off. Zou showed that the Lasso variable selection can be inconsistent, so that the oracle properties do not hold and proposed the adaptive Lasso [2]

$$\hat{\boldsymbol{\beta}}_{\text{lasso}}^{\text{ad}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\sum_{j=1}^{p}\hat{w}_j|\beta_j|, \tag{2}$$

where $\hat{w}_j = 1/|\hat{\beta}_j|^\gamma$ ($\gamma > 0$) are non-negative weights depending on $\hat{\boldsymbol{\beta}}$, which is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\beta}$.

When dealing with real-world data, the presence of impulsive noise and outliers [3], [4], [5], [6], [7] must also be accounted for. Recently, the MM-Lasso and adaptive MM-Lasso were introduced to robustify against outliers [8]. The objective function of the MM estimator is

$$\hat{\boldsymbol{\beta}}_{\text{MM}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \sum_{i=1}^{n}\rho\left(\frac{r_i(\boldsymbol{\beta})}{s_n(\mathbf{r}(\hat{\boldsymbol{\beta}}_1))}\right). \tag{3}$$

Here, $\rho(\cdot)$ is a robustifying function, see, e.g. [3]; $\mathbf{r}(\hat{\boldsymbol{\beta}}_1) = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}_1$ is the residual of an S-estimator [3] whose estimates $\hat{\boldsymbol{\beta}}_1$ have the property of minimizing a robust M-scale $s_n(\mathbf{r}(\boldsymbol{\beta}))$ that satisfies

$$\frac{1}{n}\sum_{i=1}^{n}\rho\left(\frac{r_i(\boldsymbol{\beta})}{s_n}\right) = b,$$

where b is usually chosen such that consistency under the Gaussian distribution is obtained. For MM-(adaptive) Lasso (3) is extended by the penalty terms of (1) and (2), respectively [8].

High breakdown point estimators like the MM or MM-Lasso estimator perform well in the Tukey-Huber Contamination Model (THCM) [3], which assumes, that a certain fraction of data-points, corresponding to rows in $\mathbf{X}$, are contaminated. Unfortunately, many datasets, e.g. in atmospheric inverse problems [9], [10], contain highly and independently *contaminated predictors*. This type of contamination can be represented by the Independent Contamination Model (ICM) [11]. In contrast to the THCM, it assumes, that a certain fraction of entries in each predictor is contaminated, where the contamination contained in predictor i is independent of that in predictor j, if $i \neq j$. Thus, in the ICM, the majority of the data points might be contaminated, causing classical robust estimators to break down.

Contributions In this paper, we propose a new robust Lasso type estimator, the Outlier-Corrected-Data-(Adaptive) Lasso (OCD-(A) Lasso). This estimator is able to deal with outliers that follow the classical THCM, as well as with ICM

ISBN 978-0-9928626-7-1 © EURASIP 2017

outliers, which contaminate individual cells of the regression matrix $\mathbf{X}$. The contaminated cells are detected and replaced by interpolated values. For this, we introduce an outlyingness measure that combines Stahel Donoho Outlyingness (SDO) and Predictor Outlyingness (PO). We propose an algorithm to compute the associated weights that is based on sampling from a p-dimensional unit hypersphere. We provide numerical experiments, which show that the proposed OCD-(A) Lasso outperforms the classical OLS Lasso [2], MM Lasso [8], and adaptive MM Lasso [8]. An advantage of the proposed OCD-(A) Lasso is that its computational complexity is much lower compared to its recently proposed robust counterparts.

The remainder of the paper is organized as follows. As a motivation, we show how ICM outliers affect existing robust estimators in Section II. Section III introduces the OCD-(A) Lasso and Section IV describes the algorithm to compute the weights that define the outlyingness of each cell in the regression matrix. In Section V, we provide our simulation results, while Section VI concludes the paper.

II. Towards a New Paradigm in Modeling Outliers in High-Dimensional Datasets

A highly valuable property of the THCM is that the percentage of contaminated rows in the data-matrix stays unchanged under affine transformations, that is, if the random vector $\mathbf{X}$ follows the THCM, then the affine transformed vector $\tilde{\mathbf{X}} = \mathbf{A}\mathbf{X} + \mathbf{b}$ also follows the THCM. Unfortunately, the ICM is not affine equivariant [11]. The lack of affine equivariance has a far reaching consequence for the ICM, which is referred to as ”outlier propagation”. Outlier propagation means that an outlying cell in a predictor may spread over other components of the corresponding data point, e.g. by linearly combining the predictors in a regression model. From these considerations, we can calculate the probability of a data point in a p-dimensional dataset being contaminated by the formula

$$P_{\text{cont,row}} = 1 - (1 - \epsilon)^p, \tag{4}$$

where ϵ is the assumed fraction of contamination in each predictor. For any high breakdown point estimator, tuned to have the highest possible breakdown point (50%), Fig. 1 illustrates the dependence of tolerable predictor contamination on the dimension p. Already the dimension $p = 13$ requires the contamination of each predictor to be lower than 5%, in order not to exceed the highest possible breakdown point.

III. Outlier-Corrected-Data-(Adaptive) Lasso (OCD-(A) Lasso)

We propose to remove outlier contaminated cells in $\mathbf{X}$ and calculate interpolated values for the removed cells for each predictor. After outlier removal a classical Lasso is applied. Detecting the outliers requires the computation of an outlyingness matrix that is based on the following weights.

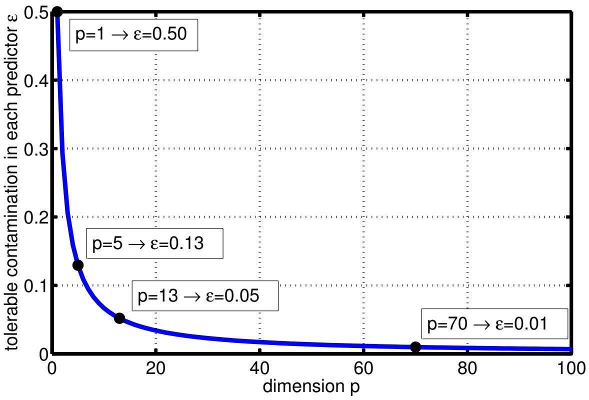


Fig. 1. Tolerable fraction of outlier contamination of every predictor as a function of the dimension p.

A. Weights to Incorporate Stahel Donoho Outlyingness (SDO) and Predictor Outlyingness (PO)

We apply the concept of Stahel Donoho Outlyingness (SDO) [12] and adjust it in a similar vein to the Adjusted Stahel Donoho Outlyingness of [13]. Let $\mathbf{B} = \{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n\} \subset \mathbb{R}^p$ be a set of n observations. Then, the robust Stahel Donoho outlyingness is given by

$$r(\mathbf{b}_i, \mathbf{X}) = \sup_{a \in S_p} \frac{|\mathbf{a}^T \mathbf{b}_i - \operatorname{med}(\mathbf{a}^T \mathbf{X})|}{\operatorname{mad}(\mathbf{a}^T \mathbf{X})}, \quad i = 1, \ldots, n, \tag{5}$$

where $S_p = \{\mathbf{a} \in \mathbb{R}^p : \|\mathbf{a}\|_2 = 1\}$ and $\operatorname{med}(\cdot)$ and $\operatorname{mad}(\cdot)$ denote the median and the median absolute deviation. Since we assume that most data points flagged by the SDO as outliers are not outlying in all of their components, the SDO can be extended by also taking into account the outlyingness of the predictors. The idea that has been introduced in [13] in a similar vein, is to adjust the SDO of every observation by taking into account the outlyingness of every single predictor. This gives us the Predictor Outlyingness (PO)

$$c_j = \frac{1}{n} \sum_{i=1}^{n} \frac{|x_{ij} - \operatorname{med}(\mathbf{x}_j)|}{\operatorname{mad}(\mathbf{x}_j)}, \quad j = 1, \ldots, p. \tag{6}$$

Combining both, the SDO and the PO, we introduce an outlyingness-matrix, whose (i, j)-th element is

$$r_{ij} = \alpha r_i + (1 - \alpha) c_j, \quad i = 1, \ldots, n, \quad j = 1, \ldots, p. \tag{7}$$

We chose the tuning parameter $\alpha \in [0, 1]$ to be 0.5 throughout this paper, to equally weight the SDO and the PO, in order to perform well in both contamination models, i.e. the THCM and the ICM. Finally, we apply a weight function $w(\cdot)$, e.g. the Huber weight function

$$w(r) = \mathbb{1}_{(r \le c_{\text{huber}})} + (c_{\text{huber}}/r)^2 \mathbb{1}_{(r > c_{\text{huber}})}, \tag{8}$$

with c_{huber} being a tuning constant, to each cell of the outlyingness-matrix defined in (7). Each entry of the outlier-detection-matrix is used to determine if the corresponding cell in the predictor matrix $\mathbf{X}$ is outlying or not.

B. OCD-(A) Lasso

Let Ω_j denote the set of indices belonging to the j-th predictor and having the cardinal number $|\Omega_j| = n$, $C_j = \{i : |w(r_{ij})| < t_j\} \subset \Omega_j$ the subset of indices belonging to the cells of the j-th predictor, which are flagged as outliers, and $S_j = \Omega_j \backslash C_j$ the set of indices belonging to the unflagged cells of the j-th predictor[1]. With this notation and t_j being a threshold value, the following algorithm describes the proposed OCD Lasso. A further example is given in the simulations.

Algorithm 1: (OCD-(A) Lasso)

1) If $i \in C_j$ and $i \in \{1, n\}$, then replace x_{ij} by the median of the j-th predictor.
2) If $i \in C_j \setminus \{1, n\}$, then replace x_{ij} by a value obtained by linear interpolation:

$$x_{ij} := x_{kj} + i \cdot \frac{x_{kj} - x_{lj}}{k - l}, \tag{9}$$

where $k, l \in S_j$ are chosen, so that $l < i < k$ and $d := k - l$ is minimized.
3) Repeat Steps 1 and 2 for every predictor $j \in \{1, \ldots p\}$ to obtain the outlier corrected matrix $\mathbf{X}_{\text{OCD}}$.
4) Carry out the Lasso and/or the adaptive Lasso based on $\mathbf{X}_{\text{OCD}}$.

IV. Computation of the Weights

The main challenge of calculating the weights, is to compute the supremum in the SDO, because the cardinal number of S_p is infinite and the objective function is non-convex. Therefore, we need to apply a random search algorithm to obtain an approximation of the supremum. We chose to take a subsample from S_p by sampling from a $(p-1)$-dimensional unit-hypersphere [14]. We use Algorithm 2 to obtain S_p, which we require in Eq. (5).

Algorithm 2: (Sampling From a $(p-1)$-Dimensional Unit Hypersphere)

1) Generate p vectors with k entries

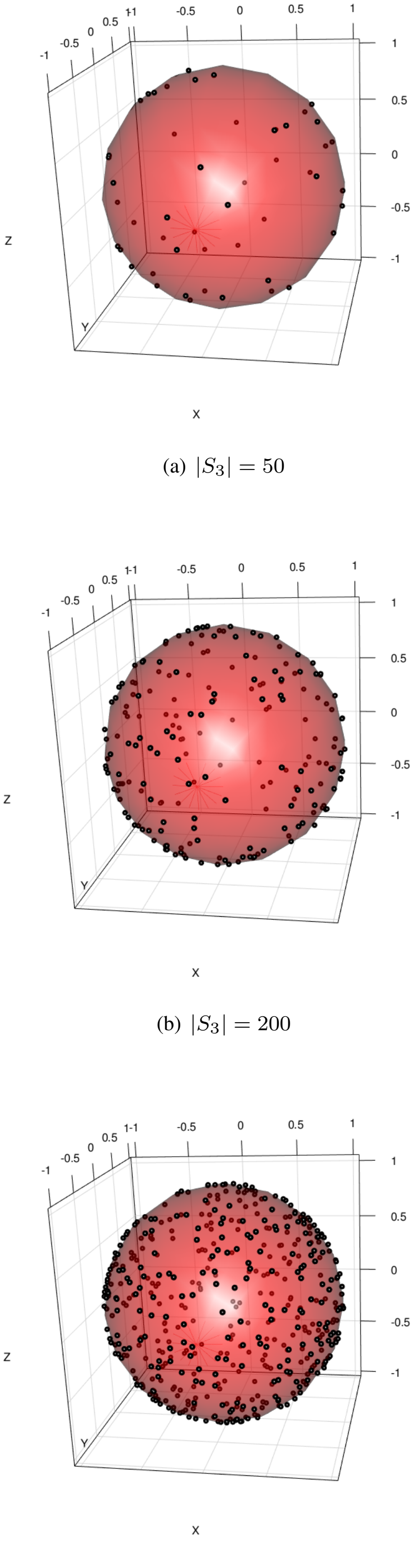
$$\mathbf{x}_j = (x_{1j}, x_{2j}, \ldots, x_{kj})^T, \quad j = 1, \ldots, p, \tag{10}$$

where $x_{ij} \sim \mathcal{N}(0, 1)$.
2) Calculate k p-dimensional vectors

$$\mathbf{a}_i = \sum_{j=1}^{p} \frac{x_{ij}}{\sqrt{x_{i1}^2 + x_{i2}^2 + \ldots + x_{ip}^2}} \cdot \mathbf{e}_j, \quad i = 1, \ldots, k, \tag{11}$$

where $\mathbf{e}_j$ is the j-th unit vector.
3) Set $S_p := \{\mathbf{a}_i \in \mathbb{R}^p : i \in \{1, \ldots, k\}\}$.

For increasing p, a larger subsample is needed. Our simulations indicate that a subsample size of approximately 1,500 for $p \leq 15$ and 10,000 for $p \leq 50$ suffices in the case of our proposed estimator. Figure 2 illustrates the results of our



(a) $|S_3| = 50$

(b) $|S_3| = 200$

(c) $|S_3| = 500$

Fig. 2. Illustration of the Results of Algorithm 2 for S_3.

[1] It is possible to define the set C_j in different ways. E.g., if we assume that there are no more than 20% of outliers in a single predictor, C_j can be defined as: $C_j = \{i : |w(r_{ij})| \leq (|w(r_{ij})|)_{\lceil 0.2 \cdot n \rceil : n}\}$, where $(|w(r_{ij})|)_{1:n} \leq (|w(r_{ij})|)_{2:n} \leq \ldots \leq (|w(r_{ij})|)_{n:n}$ is the order statistic.

ISBN 978-0-9928626-7-1 © EURASIP 2017

algorithm for S_3 and (a) $|S_3| = 50$, (b) $|S_3| = 200$ and (c) $|S_3| = 500$.

Theorem 1
The random vector $\mathbf{a}_i$, $i = 1, \ldots, n$, generated by Algorithm 2 has the uniform distribution on the $(p-1)$-dimensional unit hypersphere

$$K^{p-1} := \{\mathbf{x} \in \mathbb{R}^p : \|\mathbf{x}\|_2 = 1\}. \tag{12}$$

Proof of Theorem 1
We need to show that $\mathbf{a}_i$ lies on the $(p-1)$-dimensional unit hypersphere and that $\mathbf{a}_i$ is uniformly distributed there.

1)

$$\mathbf{a}_i = \sum_{j=1}^{p} \frac{x_{ij}}{\sqrt{x_{i1}^2 + x_{i2}^2 + \ldots + x_{ip}^2}} \cdot \mathbf{e}_j = \frac{\mathbf{x}_i}{\|\mathbf{x}_i\|}$$

$$\Rightarrow \qquad \|\mathbf{a}_i\| = \left\| \frac{\mathbf{x}_i}{\|\mathbf{x}_i\|} \right\| = \frac{\|\mathbf{x}_i\|}{\|\mathbf{x}_i\|} = 1$$

This shows that $\mathbf{a}_i$ lies on the $(p-1)$-dimensional unit hypersphere.

2) The only distribution on the $(p-1)$-dimensional hypersphere, which is invariant under rotations, is the uniform distribution [15]. So, we only need to show that the distribution of $\mathbf{a}_i$ is invariant under rotations and the proposition follows.

Let $\mathbf{R}$ be an orthogonal matrix with determinant 1. Then

$$\mathbf{a}_{i,\mathbf{R}} = \frac{\mathbf{R}\mathbf{x}_i}{\|\mathbf{R}\mathbf{x}_i\|} = \frac{\mathbf{R}\mathbf{x}_i}{\|\mathbf{x}_i\|}.$$

The second equation follows from the fact that the scalar product norm is preserved under orthogonal mappings. With $\mathbf{x}_i \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I}_p)$ and therefrom $\frac{\mathbf{x}_i}{\|\mathbf{x}_i\|} \sim \mathcal{N}_p(\mathbf{0}, \mathbf{I}_p)$ and the invariance of the standard normal distribution under rotations, the proposition follows.

■

V. Simulations

To evaluate the OCD Lasso estimator, we conduct Monte-Carlo experiments with a setup similar to [16], which is one of the few papers that consider the ICM. As benchmark comparison we consider the classical OLS Lasso [2], MM Lasso [8], and adaptive MM Lasso [8]. For the MM Lasso and adaptive MM Lasso we use the S-ridge estimator as an initialization and use a bisquare ρ function with a clipping constant $c = 3.44$, which are the standard settings as proposed in [8]. For the OCD Lasso we use a threshold value of $t_j = \left(\frac{c_{\text{huber}}}{10 \cdot \text{mad}(\mathbf{x}_j)}\right)^2$, with $c_{\text{huber}} = \min(\sqrt{\chi_p^2(.5)}, 4)$. We use $k = 10^5$ samples to calculate the supremum in the SDO. For all methods, the penalty parameter λ is chosen as the value that provided the lowest mean squared error compared to the ground truth. This is done to evaluate the performance independent of the choice of λ. Obviously, in practical situations, methods like cross validation or information criteria are applied instead.

A. Experiment 1: $n > p$, cell-wise outliers, uncorrelated predictors

We consider $n = 100$ observations with $p = 15$ regression coefficients taking values $\beta_j = j/p$ for $j = 1, ..., 5, 10, ..., 15$, and to obtain a sparse setup, $\beta_j = 0$ for $j = 6, ..., 9$. The predictors $\mathbf{x}_i$ are independent and identically normally distributed, i.e., $\boldsymbol{\mu} = \mathbf{0}, \boldsymbol{\Sigma} = \boldsymbol{I}$. The responses y_i are created according to

$$y_i = \mathbf{x}_i \boldsymbol{\beta} + e_i, \ \forall i \in \{1, ..., n\}, \tag{13}$$

where the error e_i is zero-mean independent and identically normal distributed with variance $\sigma^2 = 0.5^2$. To simulate cell-wise outliers in the predictor matrix, we contaminate 0%, 1%, 2%, 5% and 10% of the entries of the regression matrix randomly with samples drawn from a contaminating distribution $\mathcal{N}(0, 100^2)$.

To measure the accuracy of the parameter estimation, we calculate the average mean squared error (MSE) over $R = 100$ Monte Carlo experiments:

$$\text{MSE} = \frac{1}{R} \sum_{r=1}^{R} \frac{1}{p} \sum_{j=1}^{p} (\hat{\beta}_j^{(r)} - \beta_j)^2,$$

where $\hat{\beta}_j^{(r)}$ is the jth element of the parameter estimate in the rth Monte Carlo experiment.

Table I shows that the OLS Lasso breaks down for only 1% cell-wise outliers in the regression matrix. Further, even the high-breakdown MM Lasso and adaptive MM Lasso do not perform much better for cell-wise outliers.

B. Experiment 2: $p > n$, column-wise outliers, correlated predictors

We consider a challenging setup of $n = 25$ observations with $p = 50$ regression coefficients taking values $\beta_1 = 1, \beta_2 = 1.5, \beta_3 = 2, \beta_{4,...,10} = 0.3, \beta_{11,...,50} = 0$. The model is thus highly sparse. $\mathbf{X}$ is generated from a p-dimensional normal distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, where the elements of $\boldsymbol{\Sigma}$ are defined by $0.5^{|i-j|}, 0 \leq i, j \leq p$, i.e., the predictors are correlated. The responses are created according to (13). In this experiment, we consider column-wise outliers that are defined as follows. A set of predictors is randomly selected with probability ϵ. For the selected predictors, 30 % of the entries are contaminated by additive outliers from a contaminating distribution $\mathcal{N}(0, 25^2)$.

To evaluate the prediction accuracy we calculate the root mean squared prediction error (RMSPE)

$$\text{RMSPE} = \sqrt{\frac{1}{n} \|\mathbf{y}_{\text{test}} - \mathbf{X}_{\text{test}} \hat{\boldsymbol{\beta}}\|_2^2},$$

where the test data $\mathbf{y}_{\text{test}}$, $\mathbf{X}_{\text{test}}$ is generated as described above, but leaving out the contamination. Table II, which provides the averages over 100 Monte Carlo experiments shows that the proposed OCD Lasso outperforms its competitors for all contamination probabilities.

ISBN 978-0-9928626-7-1 © EURASIP 2017

	$\epsilon = 0.00$	$\epsilon = 0.01$	$\epsilon = 0.02$	$\epsilon = 0.05$	$\epsilon = 0.1$
OLS Lasso	0.31	27.02	28.88	29.82	29.90
MM Lasso	0.27	16.33	24.24	29.39	29.76
MM Ad. Lasso	0.27	16.32	24.24	29.37	29.76
OCD Lasso	0.31	2.13	4.01	16.95	26.58

TABLE I

$n \cdot \mathrm{MSE}$ OF DIFFERENT REGRESSION ESTIMATORS FOR UNCORRELATED PREDICTORS WITH CELL-WISE OUTLIERS OF PROPORTION ϵ

	$\epsilon = 0.00$	$\epsilon = 0.1$	$\epsilon = 0.2$
OLS Lasso	1.29	1.91	2.18
MM Lasso	1.63	1.92	2.48
MM Ad. Lasso	1.55	1.71	2.41
OCD Lasso	1.22	1.43	1.79

TABLE II

RMSPE OF DIFFERENT REGRESSION ESTIMATORS FOR CORRELATED PREDICTORS AND $p = 50, n = 25$ WITH ICM OUTLIERS, WHERE PREDICTORS ARE RANDOMLY CONTAMINATED WITH PROBABILITY ϵ. ONLY FOR THESE PREDICTORS, 30 % OF THE ENTRIES ARE CONTAMINATED BY ADDITIVE OUTLIERS.

C. Computation Time

A major benefit of the OCD Lasso compared to the MM Lasso is the much lower computation time. The reason is the iterative nature of the MM Lasso and adaptive MM Lasso compared to the possibility of calculating the whole regularization path at once for the OCD Lasso. Eq. (1) is efficiently solved by the LARS algorithm [17]. Table III provides the average computation times, the number of Lasso problems that are solved during the estimation process, and if the whole regularization path can be calculated at once using a LARS type algorithm.

	comp. time	# of Lasso	LARS
OLS Lasso	0.05s	1	☑
MM Lasso	127.41s	3192	□
MM Ad. Lasso	212.69s	6662	□
OCD Lasso	1.81s	1	☑

TABLE III

AVERAGE COMPUTATION TIMES ON AN INTEL(R) CORE(TM) I7-4510U WITH 8 GB RAM FOR ONE MONTE CARLO RUN, NUMBER OF REQUIRED LASSO COMPUTATIONS, AND POSSIBILITY FOR FAST COMPUTATION USING LARS [17].

VI. Conclusion

In [18], P. J. Rousseeuw *et al.* state that "*recently researchers have come to realize that the outlying rows paradigm is no longer sufficient for modern high-dimensional datasets*". We take this statement of leading researchers in the field of robust statistics as a warning, especially when applying robust methods in high dimensional settings. The presented OCD Lasso is our first step in this direction, we will next consider a real data application, as well as the extension of the proposed outlyingness measure to other Lasso estimators.

Acknowledgements

The work of M. Muma was supported by the project HANDiCAMS which acknowledges the financial support of the Future and Emerging Technologies (FET) programme within the Seventh Framework Programme for Research of the European Commission (HANDiCAMS), under FET-Open grant number: 323944.

References

[1] R. Tibshirani, "Regression Shrinkage and Selection via the Lasso," *J. Roy. Stat. Soc. B. Met.*, pp. 267–288, 1996.

[2] H. Zou, "The Adaptive Lasso and its Oracle Properties," *J. Am. Stat. Assoc.*, vol. 101, no. 476, pp. 1418–1429, 2006.

[3] A.M. Zoubir, V. Koivunen, Y. Chakhchoukh, and M. Muma, "Robust Estimation in Signal Processing: A Tutorial-Style Treatment of Fundamental Concepts," *IEEE Signal Process. Mag.*, vol. 29, no. 4, pp. 61–80, July 2012.

[4] R. Couillet, F. Pascal, and J.W. Silverstein, "Robust estimates of covariance matrices in the large dimensional regime," *IEEE Trans. Inf. Theory*, vol. 60, no. 11, pp. 7269–7278, Nov 2014.

[5] F. Pascal, Y. Chitour, and Yihui Q., "Generalized robust shrinkage estimator and its application to stap detection problem," *IEEE Trans. Signal Process.*, vol. 62, no. 21, pp. 5640–5651, Nov 2014.

[6] W.-J. Zeng, H.-C. So, and Jiang X., "Outlier-robust greedy pursuit algorithms in ℓ_p-space for sparse approximation," *IEEE Trans. Signal Process.*, vol. 64, no. 1, pp. 60–75, Jan 2016.

[7] E. Ollila, "Adaptive lasso based on joint m-estimation of regression and scale," in *Proc. Eur. Signal Process. Conf*, Aug 29 - Sep 2, 2016, pp. 2191–2195.

[8] E. Smucler and V.J. Yohai, "Robust and sparse estimators for linear regression models," *Computation. Stat.*, vol. 111, pp. 116–130, 2017.

[9] M. Martinez-Camara, B. Béjar Haro, A. Stohl, and M. Vetterli, "A Robust Method for Inverse Transport Modeling of Atmospheric Emissions Using Blind Outlier Detection," *Geosci. Model Dev.*, vol. 7, no. 5, pp. 2303–2311, 2014.

[10] M. Martinez-Camara, M. Muma, A.M. Zoubir, and M. Vetterli, "A new robust and efficient estimator for ill-conditioned linear inverse problems with outliers," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, April 2015, pp. 3422–3426.

[11] F. Alqallaf, S. Van Aelst, V.J. Yohai, and R.H. Zamar, "Propagation of outliers in multivariate data," *Ann. Statist.*, pp. 311–331, 2009.

[12] D. Donoho, "Breakdown Properties of Multivariate Location Estimators," Tech. Rep., Ph.D. qualifying paper. Harvard University, 1982.

[13] S. Van Aelst, E. Vandervieren, and G. Willems, "Stahel-Donoho Estimators with Cellwise Weights," *J. Stat. Comput. Sim.*, vol. 81, no. 1, pp. 1–27, 2011.

[14] W. P. Petersen and A. Bernasconi, "Uniform sampling from an n-sphere," Tech. Rep., Swiss Center for Scientific Computing, ETH, Zürich.

[15] J. Dongarra, K. Madsen, and J. Wasniewski, *Applied Parallel Computing: State of The Art in Scientific Computing*, vol. 3732, Springer Science & Business Media, 2006.

[16] V. Öllerer, A. Alfons, and C. Croux, "The shooting S-estimator for robust regression," *Computational Statistics*, vol. 31, no. 3, pp. 829–844, 2016.

[17] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, "Least Angle Regression," *Ann. Statist.*, vol. 32, no. 2, pp. 407–499, 2004.

[18] P. J. Rousseeuw and W. Van den Bossche, "Detecting deviating data cells," *arXiv preprint arXiv:1601.07251*, 2016.