Computational Statistics and Data Analysis 111 (2017) 116–130

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Robust and sparse estimators for linear regression models

Ezequiel Smucler *, Victor J. Yohai

Universidad de Buenos Aires - CONICET, Ciudad Universitaria, Pabellón 2, Buenos Aires 1428, Argentina

ARTICLE INFO

Article history:
Received 29 March 2016
Received in revised form 25 January 2017
Accepted 2 February 2017
Available online 16 February 2017

Keywords:
Robust regression
MM-estimators
Lasso
Oracle property
Sparse linear models

ABSTRACT

Penalized regression estimators are popular tools for the analysis of sparse and high-dimensional models. However, penalized regression estimators defined using an unbounded loss function can be very sensitive to the presence of outlying observations, especially to high leverage outliers. The robust and asymptotic properties of ℓ_1-penalized MM-estimators and MM-estimators with an adaptive ℓ_1 penalty are studied. For the case of a fixed number of covariates, the asymptotic distribution of the estimators is derived and it is proven that for the case of an adaptive ℓ_1 penalty, the resulting estimator can have the *oracle property*. The advantages of the proposed estimators are demonstrated through an extensive simulation study and the analysis of real data sets. The proofs of the theoretical results are available in the Supplementary material to this article (see Appendix A).

1. Introduction

In this paper, we consider the problem of robust and sparse estimation for linear regression models. In modern regression analysis, sparse and high-dimensional estimation scenarios where the ratio of the number of predictor variables to the number of observations, say p/n, is high, but the number of actually relevant predictor variables to the number of observations, say s/n, is low, have become increasingly common in areas such as bioinformatics and chemometrics. Outlier identification and robustness issues are difficult even when p is of moderate size. Traditional robust regression estimators do not produce sparse models and can have a bad behaviour with regard to robustness and efficiency when p/n is high, see Maronna and Yohai (2015) and Smucler and Yohai (2015). Moreover, they cannot be calculated for $p > n$. Thus, robust regression methods for high-dimensional data are in need.

Modern approaches to estimation in sparse and high-dimensional linear regression models include penalized least squares (LS) estimators, e.g. the LS-Bridge estimator of Frank and Friedman (1993) and the LS-SCAD estimator of Fan and Li (2001). LS-Bridge estimators are penalized least squares estimators in which the penalization function is proportional to the qth power of the ℓ_q norm with $q > 0$. They include as special cases the LS-Lasso of Tibshirani (1996) ($q = 1$) and the LS-Ridge of Hoerl and Kennard (1970) ($q = 2$). The LS-SCAD estimator is a penalized least squares estimator in which the penalization function, the smoothly clipped absolute deviation (SCAD), is a function with several interesting theoretical properties.

The theoretical properties of penalized least squares estimators have been extensively studied in the past years. Of special note is the so called *oracle property* defined in Fan and Li (2001): An estimator is said to have the oracle property if the estimated coefficients corresponding to zero coefficients of the true regression parameters are set to zero with probability tending to one, while at the same time the coefficients corresponding to non-zero coefficients of the true regression parameter are estimated with the same asymptotic efficiency we would have if we knew the correct model in advance.

* Corresponding author.
E-mail addresses: ezequiels.90@gmail.com (E. Smucler), vyohai@dm.uba.ar (V.J. Yohai).

http://dx.doi.org/10.1016/j.csda.2017.02.002
0167-9473/

Knight and Fu (2000) derive the asymptotic distribution of LS-Bridge estimators in the classical regression scenario of fixed p. The LS-Lasso estimator is not variable selection consistent unless rather stringent conditions are imposed on the design matrix, and thus in general does not possess the oracle property; see Zou (2006) and Bühlmann and van de Geer (2011) for details. Moreover, the LS-Lasso estimator has a bias problem: it can excessively shrink large coefficients. To remedy this issue, Zou (2006) introduced the adaptive LS-Lasso, in which adaptive weights are used for penalizing different coefficients and showed that the adaptive Lasso can have the oracle property. As Zou (2006) points out, adaptive LS-Lasso estimators can be computed using any of the algorithms available to compute LS-Lasso estimators.

Penalized least squares estimators are not robust and may be highly inefficient under heavy tailed errors. In an attempt to remedy this issue, penalized M-estimators defined using a convex loss function have been proposed. For example, in Wang et al. (2007) the authors propose to take the absolute value loss and a Lasso type penalty, they call the resulting estimator LAD-Lasso. See also Li et al. (2011) and Lambert-Lacroix and Zwald (2011). Estimators based on ranks have also been proposed, see for example Johnson and Peng (2008) and Leng (2010). Zou and Yuan (2008) proposed the adaptive Lasso Penalized Composite Quantile Regression estimator. All of the aforementioned estimators aim at robustness towards outliers in the response variable and/or when heavy-tailed errors are present. Unfortunately, they are not robust with respect to contamination in the predictor variables.

Khan et al. (2007) proposed a robust version of the LARS procedure, see Efron et al. (2004), and called it RLARS. However, since the RLARS procedure is not based on the minimization of a clearly defined objective function, a theoretical analysis of its properties is difficult. In Wang and Li (2009) the authors proposed a weighted Wilcoxon-type smoothly clipped absolute deviation (WW-SCAD) estimator. Maronna (2011) introduced S-Ridge and MM-Ridge estimators: ℓ_2-penalized S- and MM-estimators. However, ℓ_2-penalized regression estimators do not produce sparse models. Alfons et al. (2013) proposed the Sparse-LTS estimator, a least trimmed squares estimator with an ℓ_1 penalization. See also Öllerer et al. (2016), Alfons et al. (2016) and Öllerer et al. (2015). Wang et al. (2013) proposed a penalized regression estimator based on an exponential squared loss function (ESL-Lasso). Gijbels and Vrinssen (2015) proposed nonnegative garrote versions of several robust regression estimators, including MM and S-estimators. In Loh (2015), the author studied the theoretical properties of penalized regression M-estimators in the $p \gg n$ regime. Unfortunately, these results are not directly applicable to the estimators we study in this paper.

In this paper, we study the robust and asymptotic properties of MM-Lasso and adaptive MM-Lasso estimators: ℓ_1-penalized MM-estimators and MM-estimators with an adaptive ℓ_1 penalty. We obtain lower bounds on their breakdown points. We derive the asymptotic distribution of the estimators and prove that adaptive MM-Lasso estimators can have the oracle property. Even though we derive our asymptotic results for fixed p, MM-Lasso and adaptive MM-Lasso estimators can be computed for $p > n$. In extensive simulations, we compare the performance of the MM-Lasso and adaptive MM-Lasso estimators with that of several competitors. In all the scenarios considered our proposed estimators compare favourably to the competitors. Finally, we apply our proposed estimators to two real data sets.

The rest of this paper is organized as follows. In Section 2 we review the definition and some of the most important properties of MM and S-estimators. In Section 3 we define MM-Lasso and adaptive MM-Lasso estimators, we study their robust and asymptotic theoretical properties and we describe an algorithm to compute them. In Section 4 we conduct an extensive simulation. In Section 5 we apply the aforementioned estimators to two real data sets. Conclusions are provided in Section 6. Finally, the proofs of all our results are given in the Supplementary material to this article (see Appendix A).

2. MM and S-estimators

We consider a linear regression model with random carriers: we observe $(\mathbf{x}_i^{\mathrm{T}}, y_i)$ $i = 1, \ldots, n$, i.i.d. $(p+1)$-dimensional vectors, where y_i is the response variable and $\mathbf{x}_i \in \mathbb{R}^p$ is a vector of random carriers, satisfying

$$y_i = \mathbf{x}_i^{\mathrm{T}} \boldsymbol{\beta}_0 + u_i \quad \text{for } i = 1, \ldots, n, \tag{1}$$

where $\boldsymbol{\beta}_0 \in \mathbb{R}^p$ is to be estimated and u_i is independent of $\mathbf{x}_i$. For $\boldsymbol{\beta} \in \mathbb{R}^p$ let $\mathbf{r}(\boldsymbol{\beta}) = (r_1(\boldsymbol{\beta}), \ldots, r_n(\boldsymbol{\beta}))$, where $r_i(\boldsymbol{\beta}) = y_i - \mathbf{x}_i^T \boldsymbol{\beta}$. Some of the coefficients of $\boldsymbol{\beta}_0$ may be zero, and thus the corresponding carriers do not provide relevant information to predict y. We do not know in advance the set of indices corresponding to coefficients that are zero, and it may be of interest to estimate it. For simplicity, we will assume $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{0,I}, \boldsymbol{\beta}_{0,II})$, where $\boldsymbol{\beta}_{0,I} \in \mathbb{R}^s$, $\boldsymbol{\beta}_{0,II} \in \mathbb{R}^{p-s}$, all the coordinates of $\boldsymbol{\beta}_{0,I} \in \mathbb{R}^s$ are non-zero and all the coordinates of $\boldsymbol{\beta}_{0,II} \in \mathbb{R}^{p-s}$ are zero.

Let F_0 be the distribution of the errors u_i, G_0 the distribution of the carriers $\mathbf{x}_i$ and H_0 the distribution of $(\mathbf{x}_i^{\mathrm{T}}, y_i)$. Then H_0 satisfies

$$H_0(\mathbf{x}, y) = G_0(\mathbf{x}) F_0(y - \mathbf{x}^{\mathrm{T}} \boldsymbol{\beta}_0). \tag{2}$$

Let $\mathbf{x}_I$ stand for the first s coordinates of $\mathbf{x}$ and let $G_{0,I}$ be its distribution. For $\mathbf{b} \in \mathbb{R}^p$ and $q > 0$ we note

$$\|\mathbf{b}\|_q = \left(\sum_{j=1}^{p} |b_j|^q \right)^{1/q}$$

and $\|\mathbf{b}\| = \|\mathbf{b}\|_2$. Throughout this paper, a ρ-function will refer to a bounded ρ-function, in the sense of Maronna et al. (2006). That is, we will say that ρ is a ρ-function if: (i) ρ is even, continuous and bounded, (ii) $\rho(x)$ is a nondecreasing function of $|x|$, (iii) $\rho(0) = 0$, (iv) $\lim_{x\to\infty} \rho(x) = 1$ and (v) If $\rho(v) < 1$ and $0 \le u < v$ then $\rho(u) < \rho(v)$. A popular choice of ρ-functions is Tukey's Bisquare family of functions given by

$$\rho_c^B(u) = 1 - \left(1 - \left(\frac{u}{c}\right)^2\right)^3 I(|u| \le c), \tag{3}$$

where $c > 0$ is a tuning constant. If ρ is a differentiable ρ-function we will let $\psi = \rho'$.

Given a sample $\mathbf{u} = (u_1, \dots, u_n)$ from some distribution F and $0 < b < 1$ the corresponding M-estimate of scale $s_n(\mathbf{u})$ is defined by

$$s_n(\mathbf{u}) = \inf\left\{s > 0 : \frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{u_i}{s}\right) \le b\right\}.$$

It is easy to prove that $s_n(\mathbf{u}) > 0$ if and only if $\#\{i : u_i = 0\} < (1-b)n$, and in this case

$$\frac{1}{n}\sum_{i=1}^{n} \rho\left(\frac{u_i}{s_n(\mathbf{u})}\right) = b.$$

A quantitative measure of an estimator's robustness, introduced by Donoho and Huber (1983), is the finite-sample replacement breakdown point. Loosely speaking, the finite-sample replacement breakdown point of an estimator is the maximum fraction of outliers that the estimator may tolerate without losing all meaning. For a regression estimator, this measure is defined as follows. Given a sample $\mathbf{z}_i = (\mathbf{x}_i^{\mathrm{T}}, y_i)$, $i = 1, \dots, n$, let $\mathbf{Z} = \{\mathbf{z}_1, \dots, \mathbf{z}_n\}$ and let $\hat{\boldsymbol{\beta}}(\mathbf{Z})$ be a regression estimator. The finite-sample replacement breakdown point of $\hat{\boldsymbol{\beta}}$ is then defined as $FBP(\hat{\boldsymbol{\beta}}) = m^*/n$, where

$$m^* = \max\left\{m \ge 0 : \hat{\boldsymbol{\beta}}(\mathbf{Z}_m) \text{ is bounded for all } \mathbf{Z}_m \in \mathcal{Z}_m\right\},$$

and $\mathcal{Z}_m$ is the set of all data sets with at least $n - m$ elements in common with $\mathbf{Z}$.

Given a sample $(\mathbf{x}_i^{\mathrm{T}}, y_i)$, $i = 1, \dots, n$ from the model given in (1), Rousseeuw and Yohai (1984) define the S-estimator as

$$\hat{\boldsymbol{\beta}}_S = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} s_n(\mathbf{r}(\boldsymbol{\beta})),$$

where s_n is a M-estimate of scale. Fasano et al. (2012) derive the asymptotic distribution of S-estimators under very general conditions. S-estimators can always be tuned so as to attain the maximum possible finite-sample replacement breakdown point for regression equivariant estimators. However, S-estimators cannot combine high breakdown point with high efficiency at the normal distribution, see Hossjer (1992).

Let $(\mathbf{x}_i^{\mathrm{T}}, y_i)$, $i = 1, \dots, n$, be a sample satisfying (1), and ρ_0 and ρ_1 be two ρ-functions satisfying $\rho_1 \le \rho_0$. Then Yohai (1987) defines the MM-estimator as

$$\hat{\boldsymbol{\beta}}_{MM} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{n} \rho_1\left(\frac{r_i(\boldsymbol{\beta})}{s_n(\mathbf{r}(\hat{\boldsymbol{\beta}}_1))}\right),$$

where $\hat{\boldsymbol{\beta}}_1$ is a consistent and high breakdown point estimator of $\boldsymbol{\beta}_0$ and $s_n(\mathbf{r}(\hat{\boldsymbol{\beta}}_1))$ is the M-estimate of scale of the residuals of $\hat{\boldsymbol{\beta}}_1$, defined using ρ_0 and b.

Yohai (1987) proves that under general conditions MM-estimators are strongly consistent for $\boldsymbol{\beta}_0$ and furthermore

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{MM} - \boldsymbol{\beta}_0) \xrightarrow{d} N_p\left(\mathbf{0}, s(\boldsymbol{\beta}_0)^2 \frac{a(\psi_1, F_0)}{b(\psi_1, F_0)^2} \mathbf{V}_{\mathbf{x}}^{-1}\right), \tag{4}$$

where $\mathbf{V}_{\mathbf{x}} = E_{G_0}(\mathbf{x}\mathbf{x}^{\mathrm{T}})$, $\psi_1 = \rho_1'$, $s(\boldsymbol{\beta}_0)$ is defined by

$$E_{H_0}\rho_0\left(\frac{y - \mathbf{x}^T\boldsymbol{\beta}_0}{s(\boldsymbol{\beta}_0)}\right) = b, \qquad a(\psi, F_0) = E_{F_0}\psi^2\left(\frac{u}{s(\boldsymbol{\beta}_0)}\right) \quad \text{and}$$

$$b(\psi, F_0) = E_{F_0}\psi'\left(\frac{u}{s(\boldsymbol{\beta}_0)}\right).$$

Besides, he shows that ρ_1 can be chosen so that the resulting MM-estimator simultaneously has a normal asymptotic efficiency as close to one as desired and a breakdown point greater than or equal to that of the initial estimator.

3. MM-Lasso and adaptive MM-Lasso estimators

Given a sample $(\mathbf{x}_i^{\mathrm{T}}, y_i)$, $i = 1, \ldots, n$, $\gamma_n \geq 0$, a ρ-function ρ_0 and $0 < b < 1$, the S-Ridge estimator is defined as (see Maronna, 2011)

$$\hat{\boldsymbol{\beta}}_{PS} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} s_n^2(\mathbf{r}(\boldsymbol{\beta})) + \frac{\gamma_n}{n}\|\boldsymbol{\beta}\|,$$

where $s_n(\mathbf{r}(\boldsymbol{\beta}))$ is the residual scale estimate defined using ρ_0 and b. If the model contains an intercept, then it is not penalized.

It is easy to see that

$$\|\hat{\boldsymbol{\beta}}_{PS}\| \leq \|\hat{\boldsymbol{\beta}}_S\|, \tag{5}$$

where $\hat{\boldsymbol{\beta}}_S$ is the S-estimator defined using ρ_0 and b.

Given another ρ-function ρ_1 that satisfies $\rho_1 \leq \rho_0$ and $\lambda_n \geq 0$ we define the MM-Lasso estimator as

$$\hat{\boldsymbol{\beta}}_B = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{n} \rho_1\left(\frac{r_i(\boldsymbol{\beta})}{s_n(\mathbf{r}(\hat{\boldsymbol{\beta}}_1))}\right) + \lambda_n\|\boldsymbol{\beta}\|_1 \tag{6}$$

where $\hat{\boldsymbol{\beta}}_1$ is a strongly consistent initial estimator of $\boldsymbol{\beta}_0$. Clearly, the robustness of the MM-Lasso estimator will depend heavily on the robustness of the initial estimator. If the model contains an intercept, then it is not penalized.

Note that, for any fixed $\lambda_n > 0$, by definition of $\hat{\boldsymbol{\beta}}_B$ we have that

$$\begin{aligned}
\lambda_n\|\hat{\boldsymbol{\beta}}_B\|_1 &\leq \sum_{i=1}^{n} \rho_1\left(\frac{r_i(\hat{\boldsymbol{\beta}}_B)}{s_n(\mathbf{r}(\hat{\boldsymbol{\beta}}_1))}\right) + \lambda_n\|\hat{\boldsymbol{\beta}}_B\|_1 \\
&\leq \sum_{i=1}^{n} \rho_1\left(\frac{r_i(\mathbf{0}_p)}{s_n(\mathbf{r}(\hat{\boldsymbol{\beta}}_1))}\right) + \lambda_n\|\mathbf{0}_p\|_1 \\
&\leq n,
\end{aligned} \tag{7}$$

since $\rho_1 \leq 1$. Hence, $\|\hat{\boldsymbol{\beta}}_B\|_1 \leq n/\lambda_n$. This immediately implies that for any fixed $\lambda_n > 0$, the breakdown point of $\hat{\boldsymbol{\beta}}_B$ is equal to 1.

Given $\varsigma > 0$ and $\iota_n \geq 0$ we define the adaptive MM-Lasso estimator as

$$\hat{\boldsymbol{\beta}}_A = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \sum_{i=1}^{n} \rho_1\left(\frac{r_i(\boldsymbol{\beta})}{s_n(\mathbf{r}(\hat{\boldsymbol{\beta}}_1))}\right) + \iota_n \sum_{j=1}^{p} \frac{|\beta_j|}{|\hat{\beta}_{2,j}|^{\varsigma}}, \tag{8}$$

where $\hat{\boldsymbol{\beta}}_2$ is a strongly consistent initial estimator of $\boldsymbol{\beta}_0$. Clearly if $\hat{\beta}_{2,j} = 0$ for some j, then $\hat{\boldsymbol{\beta}}_{A,j} = 0$. If the model contains an intercept, then it is not penalized. Note that for coefficients corresponding to large coefficients of $\hat{\boldsymbol{\beta}}_2$, the adaptive MM-Lasso employs a small penalty; this ameliorates the bias issues associated with the ℓ_1 penalty.

Wang et al. (2013) prove that their estimator can have the highest possible breakdown point among regression equivariant estimators, but it must be noted that their estimator is not regression equivariant. Alfons et al. (2013) show that the breakdown point of the Sparse-LTS estimator is $(n-h)/n$, where $n-h$ is the number of trimmed observations, and prove that the breakdown point of any ℓ_1-penalized M-estimator defined using a convex loss function is 0. In particular, the breakdown points of the LS-Lasso and the LAD-Lasso are 0. Note that it follows immediately from (5) that for any γ_n, the finite-sample breakdown point of $\hat{\boldsymbol{\beta}}_{PS}$ is at least as high as that of $\hat{\boldsymbol{\beta}}_S$. It follows from (7) that the breakdown point of $\hat{\boldsymbol{\beta}}_B$ is 1. We believe that this result hints at the possibility that the breakdown point may not be an adequate measure of robustness for penalized estimators. More generally, one could argue that the breakdown point is not an adequate measure of robustness for estimators that are not regression equivariant. See Davies and Gather (2005) for a discussion of this point. Nonetheless, in Theorem 1 we prove that for any fixed $\iota_n > 0$, the breakdown point of $\hat{\boldsymbol{\beta}}_A$ is greater than or equal to the breakdown point of $\hat{\boldsymbol{\beta}}_2$.

Theorem 1. *If $\iota_n > 0$ is fixed, then $FBP(\hat{\boldsymbol{\beta}}_A) \geq FBP(\hat{\boldsymbol{\beta}}_2)$.*

Note that if $\hat{\boldsymbol{\beta}}_2 = \hat{\boldsymbol{\beta}}_B$, then $FBP(\hat{\boldsymbol{\beta}}_A) = 1$ whenever $\lambda_n, \iota_n > 0$. In practice, γ_n, λ_n and ι_n may be chosen via some data-driven procedure such as cross-validation. In this case, the breakdown points of the resulting MM-Lasso and adaptive MM-Lasso estimators may be lower. The robustness of the resulting estimators will depend sorely on the robustness of the cross-validation scheme, and hence the use of robust residual scales as objective functions, instead of the classical root mean squared error, is crucial.

3.1. Asymptotics

We now describe the set-up to study the asymptotic properties of S-Ridge, MM-Lasso and adaptive MM-Lasso estimators. We will assume that

B1. (a) ρ_0 and ρ_1 are twice continuously differentiable and eventually constant.
(b) Let $\psi_1 = \rho_1'$. Then $E_{F_0}\psi_1'\left(u/s(\boldsymbol{\beta}_0)\right) > 0$.
B2. $\mathbb{P}\left(\mathbf{x}^T\boldsymbol{\beta} = 0\right) < 1 - b$ for all non-zero $\boldsymbol{\beta} \in \mathbb{R}^p$.
B3. F_0 has an even density, f_0, that is a monotone decreasing function of $|u|$ and a strictly decreasing function of $|u|$ in a neighbourhood of 0.
B4. G_0 has finite second moments and $\mathbf{V}_{\mathbf{x}} = E_{G_0}\mathbf{x}\mathbf{x}^{\mathrm{T}}$ is non-singular.

A family of ρ-functions that satisfies [B1] (a) is Tukey's Bisquare family of functions, given in (3). Condition [B2] is needed in the proof of the consistency of the estimators. Note that condition [B3] does not require finite moments from F_0. Thus, extremely heavy tailed error distributions, such as Cauchy's distribution, can be easily seen to satisfy [B3]. However, [B3] does impose a rather stringent symmetry assumption on the error distribution. This requirement greatly simplifies the asymptotic treatment of the estimators and is usual in robust statistics. [B4] is a moment condition that is usual in the literature.

Propositions 1 and 2, which can be found in the Supplementary material to this article (see Appendix A), prove the strong consistency of S-Ridge, MM-Lasso and adaptive MM-Lasso estimators and the $\sqrt{n}$-consistency of MM-Lasso and adaptive MM-Lasso estimators.

In practice, we will use the S-Ridge estimator of Maronna (2011) as the initial estimate $\hat{\boldsymbol{\beta}}_1$ in (6) and (8). Note that according to Proposition 1 and the remarks above Theorem 1, the S-Ridge is a high breakdown point and consistent estimate of $\boldsymbol{\beta}_0$, as long as the penalization parameter satisfies $\gamma_n = o(n)$.

Remark 1. From now on, we will assume that the initial estimator used to define the penalty weights for the adaptive MM-Lasso estimator, $\hat{\boldsymbol{\beta}}_2$, is $\sqrt{n}$-consistent. For example, according to Proposition 2, we could take $\hat{\boldsymbol{\beta}}_2$ to be some MM-Lasso estimator calculated with $\lambda_n = O(\sqrt{n})$.

Let $\hat{\boldsymbol{\beta}}_{A,I}$ stand for the first s coordinates of $\hat{\boldsymbol{\beta}}_A$ and $\hat{\boldsymbol{\beta}}_{A,II}$ for the remaining $p-s$. The following theorem shows that, as long as $\varsigma > 0$ adaptive MM-Lasso estimators can be variable selection consistent.

Theorem 2. *Let* $(\mathbf{x}_i^{\mathrm{T}}, y_i)$, $i = 1, \ldots, n$, *be i.i.d. observations with distribution* H_0, *which satisfies* (2). *Assume* [B1]–[B4] *hold. Then if* $\iota_n = O(\sqrt{n})$ *and* $\iota_n n^{(\varsigma-1)/2} \to \infty$,

$$\mathbb{P}\left(\hat{\boldsymbol{\beta}}_{A,II} = \mathbf{0}_{p-s}\right) \to 1.$$

Next we derive the asymptotic distribution of $\hat{\boldsymbol{\beta}}_{A,I}$.

Theorem 3. *Let* $(\mathbf{x}_i^{\mathrm{T}}, y_i)$, $i = 1, \ldots, n$, *be i.i.d. observations with distribution* H_0, *which satisfies* (2). *Assume* [B1]–[B4] *hold. Then if* $\iota_n/\sqrt{n} \to 0$ *and* $\iota_n n^{(\varsigma-1)/2} \to \infty$

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_{A,I} - \boldsymbol{\beta}_{0,I}) \xrightarrow{d} N_s\left(\mathbf{0}, s(\boldsymbol{\beta}_0)^2 \frac{a(\psi_1, F_0)}{b(\psi_1, F_0)^2}\mathbf{V}_{\mathbf{x}_I}^{-1}\right).$$

Here $s(\boldsymbol{\beta}_0)$, $a(\psi, F_0)$ *and* $b(\psi, F_0)$ *are as in* (4) *and* $\mathbf{V}_{\mathbf{x}_I} = E_{G_0}\mathbf{x}_I\mathbf{x}_I^T$.

Theorem 2 together with Theorem 3 proves that $\hat{\boldsymbol{\beta}}_A$ can have the oracle property as long as $\varsigma > 0$. That is: the estimated coefficients corresponding to null coordinates of the true regression parameters are set to zero with probability tending to 1, while at the same time the coefficients corresponding to non-null coordinates of the true regression parameter are estimated with the same asymptotic efficiency we would have if we had applied a non penalized MM-estimator to the relevant carriers only.

In Theorem 4 we derive the asymptotic distribution of $\hat{\boldsymbol{\beta}}_B$. Similar results can be obtained for ℓ_q-penalized MM-estimators.

Theorem 4. *Let* $(\mathbf{x}_i^{\mathrm{T}}, y_i)$, $i = 1, \ldots, n$, *be i.i.d. observations with distribution* H_0, *which satisfies* (2). *Assume* [B1]–[B4] *hold and* $\lambda_n/\sqrt{n} \to \lambda_0$. *Then*

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_B - \boldsymbol{\beta}_0) \xrightarrow{d} \arg\min(R),$$

where

$$R(\mathbf{z}) = -\mathbf{z}^T\mathbf{W} + \frac{1}{2s(\boldsymbol{\beta}_0)^2}b(\psi_1, F_0)\mathbf{z}^T\mathbf{V}_\mathbf{x}\mathbf{z} + \lambda_0\sum_{j=1}^{p}(z_j sgn(\beta_{0,j})I(\beta_{0,j} \neq 0) + |z_j|I(\beta_{0,j} = 0)),$$

and $\mathbf{W} \sim N_p\left(\mathbf{0}, (a(\psi_1, F_0)/s(\boldsymbol{\beta}_0)^2)\mathbf{V}_\mathbf{x}\right)$. *Here* $s(\boldsymbol{\beta}_0)$, $a(\psi, F_0)$ *and* $b(\psi, F_0)$ *are as in* (4).

Note that if $\lambda_0 = 0$, $\hat{\boldsymbol{\beta}}_B$ has the same asymptotic distribution as the corresponding non penalized MM-estimator. If $\lambda_0 > 0$, the asymptotic distributions of the coordinates of $\hat{\boldsymbol{\beta}}_B$ corresponding to null coefficients of $\boldsymbol{\beta}_0$ put positive probability at zero, the proof is essentially the same as the one that appears in pages 1361–1362 of Knight and Fu (2000). However, one can show that

$$\limsup \mathbb{P}\left(\hat{\boldsymbol{\beta}}_{B,II} = \mathbf{0}_{p-s}\right) \leq c < 1,$$

for some c. The proof is essentially the same as the proof of Proposition 1 of Zou (2006).

3.2. Computation

In this section, we describe an algorithm to obtain approximate solutions of (6), i.e. MM-Lasso estimators. Throughout this section we will assume that our model, (1), contains an intercept, and that the first coordinate of each $\mathbf{x}_i$ equals 1. Let $\mathbf{X}$ be the matrix with $\mathbf{x}_i$ as rows.

Prior to any calculations, all the columns of $\mathbf{X}$, except the first one, are centred and scaled using the median and the normalized median absolute deviation respectively. The response vector $\mathbf{y}$ is centred using the median. At the end, the final estimates are expressed in the original coordinates.

We take the S-Ridge estimator of Maronna (2011), which we note $\hat{\boldsymbol{\beta}}_{PS}$, as the initial estimate in (6). The penalization parameter for the S-Ridge estimator, γ_n, is chosen via robust 5-fold cross-validation, as described in Maronna (2011). Let $s_n = s_n(\mathbf{r}(\hat{\boldsymbol{\beta}}_{PS}))$.

Let $w(u) = \psi_1(u)/u$, where ψ_1 is the derivative of ρ_1. For a given $\boldsymbol{\beta}$, let $\omega_i = w(r_i(\boldsymbol{\beta})/s_n)$. Suppose λ_n is given and let $\mathbf{W}$ be the diagonal matrix formed by $\sqrt{\omega_1}, \ldots, \sqrt{\omega_n}$. Let $\mathbf{y}^* = \mathbf{W}\mathbf{y}$ and $\mathbf{X}^* = \mathbf{W}\mathbf{X}$. Let $\hat{\boldsymbol{\beta}}_B$ be the MM-Lasso estimator. It is easy to show that $\hat{\boldsymbol{\beta}}_B$ satisfies

$$\mathbf{X}^{*T}(\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta}) - \lambda_n s_n^2 \begin{pmatrix} 0 \\ sign(\beta_2) \\ \vdots \\ sign(\beta_{p+1}) \end{pmatrix} \stackrel{s}{=} \mathbf{0}_{p+1},$$

where $\stackrel{s}{=}\mathbf{0}_{p+1}$ stands for a change of sign. Note that the first column of $\mathbf{X}^*$ equals $\mathbf{k}^* = (\sqrt{\omega_1}, \ldots, \sqrt{\omega_n})$. For each $j = 2, \ldots, p+1$ let $\mathbf{x}^{*(j)}$ be the jth column of $\mathbf{X}^*$ and let

$$\eta_j = \frac{\mathbf{k}^{*T}\mathbf{x}^{*(j)}}{\|\mathbf{k}^*\|^2}.$$

Then $\mathbf{x}^{*(j)}$ can be decomposed as the sum of two vectors: $\eta_j\mathbf{k}^*$, in the direction of $\mathbf{k}^*$, and $\mathbf{x}^{*\perp(j)} = \mathbf{x}^{*(j)} - \eta_j\mathbf{k}^*$, orthogonal to $\mathbf{k}^*$. Let $\mathbf{X}^{*\perp}$ be the matrix with columns $\mathbf{x}^{*\perp(2)}, \ldots, \mathbf{x}^{*\perp(p+1)}$. It is easy to show that $\hat{\boldsymbol{\beta}}_B$ satisfies

$$\mathbf{k}^*\mathbf{y}^* - \|\mathbf{k}^*\|^2(\beta_1 + \eta_2\beta_2 + \cdots + \eta_{p+1}\beta_{p+1}) = 0 \tag{9}$$

$$\mathbf{X}^{*\perp T}(\mathbf{y}^* - \mathbf{X}^{*\perp}\boldsymbol{\beta}) - \lambda_n s_n^2 \begin{pmatrix} sign(\beta_2) \\ \vdots \\ sign(\beta_{p+1}) \end{pmatrix} \stackrel{s}{=} \mathbf{0}_p. \tag{10}$$

We note that if $\mathbf{k}^*$, $\mathbf{y}^*$ and $\mathbf{X}^{*\perp T}$ were known, $\hat{\boldsymbol{\beta}}_{B,2},\ldots, \hat{\boldsymbol{\beta}}_{B,p+1}$ could be estimated by solving Eq. (10) using some algorithm to solve Lasso-type problems, e.g. the LARS procedure or Coordinate Descent Optimization, without including an intercept. Then $\hat{\boldsymbol{\beta}}_{B,1}$ could be solved easily from (9).

The fact that $\mathbf{k}^*$, $\mathbf{y}^*$ and $\mathbf{X}^{*\perp T}$ depend on $\hat{\boldsymbol{\beta}}_B$ suggests an iterative procedure, as is usual in robust statistics. Starting from $\hat{\boldsymbol{\beta}}_{PS}$ we iteratively solve Eq. (10) using the LARS algorithm without including an intercept and then solve for the intercept in (9). Call $\boldsymbol{\beta}^{(i)}$ the estimate at the ith iteration. Convergence is declared when

$$\frac{\|\boldsymbol{\beta}^{(i+1)} - \boldsymbol{\beta}^{(i)}\|}{\|\boldsymbol{\beta}^{(i)}\|} \leq \delta,$$

where δ is some fixed tolerance parameter. In our simulations we took $\delta = 10^{-4}$.

Regarding the computation of adaptive MM-Lasso estimators, we note that solving (8) is equivalent to solving

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{p+1}} \sum_{i=1}^{n} \rho_1\left(\frac{y_i - \hat{\mathbf{x}}_i^T\boldsymbol{\beta}}{s_n}\right) + \iota_n \sum_{j=2}^{p+1} |\beta_j|,$$

where $\hat{\mathbf{x}}_{i,1} = 1$, $\hat{\mathbf{x}}_{i,j} = \mathbf{x}_{i,j}|\hat{\beta}_{2,j}|^{\varsigma}$ for $j = 2, \ldots, p+1$ and taking $\hat{\beta}_{A,1} = \hat{\beta}_1$ and $\hat{\beta}_{A,j} = \hat{\beta}_j|\hat{\beta}_{2,j}|^{\varsigma}$ for $j = 2, \ldots, p+1$. Hence, our procedure to compute MM-Lasso estimators can be used to compute adaptive MM-Lasso estimators, simply applying the routine to the data with weighed carriers. To calculate our proposed adaptive MM-Lasso estimator, we take $\hat{\boldsymbol{\beta}}_1 = \hat{\boldsymbol{\beta}}_{PS}$, $\hat{\boldsymbol{\beta}}_2 = \hat{\boldsymbol{\beta}}_B$ and $\varsigma = 1$.

In practice, we chose the ρ-functions used to calculate the initial S-Ridge estimator, the MM-Lasso estimator and the adaptive MM-Lasso estimator of the form $\rho_0 = \rho^B_{c_0}$ and $\rho_1 = \rho^B_{c_1}$ where $c_1 \geq c_0$ and ρ^B_c is as in (3). The tuning constants c_0 and c_1 are chosen as in Maronna (2011).

The penalization parameter for $\hat{\boldsymbol{\beta}}_B$, λ_n, is chosen over a set of candidates via robust 5-fold cross validation, using a τ-scale of the residuals as the objective function. The τ-scale was introduced by Yohai and Zamar (1988) to measure in a robust and efficient way the largeness of the residuals in a regression model. The set of candidate lambdas is taken as 30 equally spaced points between 0 and $\lambda_{\max}$, where $\lambda_{\max}$ is approximately the minimum λ such that all the coefficients of $\hat{\boldsymbol{\beta}}_B$ except the intercept are zero. To estimate $\lambda_{\max}$ we first robustly estimate the maximal correlation between $\mathbf{y}$ and the columns of $\mathbf{X}$ using bivariate winsorization as advocated by Khan et al. (2007). We use this estimate as an initial guess for $\lambda_{\max}$ and then improve it using a binary search. If $p > n$, then 0 is excluded from the candidate set. The penalization parameter for $\hat{\boldsymbol{\beta}}_A$, ι_n, is chosen using the same scheme used to choose λ_n.

The initial S-Ridge estimate is calculated using our own adaption of Maronna's MATLAB code to C++. To solve Eq. (10) we use the FastLasso() function from the `robustHD` R package. We use the `foreach` R package for parallel computations when it comes to finding optimal penalization parameters via cross-validation. This provided a significant reduction in computing times in computers with several cores. Extensive parts of our computer code are written in C++ and interfaced with R using the `RcppArmadillo` package. An R package that includes the functions to calculate the estimators we propose is available at http://esmucler.github.io/mmlasso/.

4. Simulations

In this section, we compare the performance with regard to prediction accuracy and variable selection properties of

- The MM-Lasso estimator described in the previous section.
- The adaptive MM-Lasso (adaMM-Lasso) estimator described in the previous section.
- The MM Nonnegative Garrote (MM-NNG) estimator of Gijbels and Vrinssen (2015). The estimator was calculated using R code provided by the authors.
- The ESL-Lasso estimator. The estimator was calculated using MATLAB code provided by the authors.
- The Sparse-LTS. The penalization parameter for this estimator was chosen using a BIC-type criterion as advocated by the authors. The estimator was calculated using the sparseLTS() function from the `robustHD` R package.
- The Wilcoxon-SCAD estimator (WW-SCAD). The estimator was calculated using R code provided by the authors.
- The LAD-Lasso estimator. The penalization parameter was chosen using 5-fold cross validation using the median of the absolute value of the residuals as the objective function. We implemented this estimator in R.
- The LS-Lasso estimator (Lasso). The penalization parameter for this estimator was chosen using 5-fold cross validation using the sum of the squared residuals as the objective function. The estimator was calculated using the lars() function from the `lars` R package.
- The adaptive LS-Lasso estimator (adaLasso). We used as an initial estimator an LS-Lasso estimator, calculated as above. Both the initial and the final penalization parameters were chosen using 5-fold cross validation using the sum of the squared residuals as the objective function. The estimator was calculated using the lars() function from the `lars` R package.
- The Maximum Likelihood Oracle estimator (Oracle), that is, the Maximum Likelihood estimator applied to the relevant carriers only. When the errors follow a normal distribution, this is the Least Squares estimators applied to the relevant carriers only. Note that in any case, this is not a feasible estimator, and is included for benchmarking purposes only.
- For the contaminated scenarios, we will also include the Oracle MM estimator: an MM-estimator, calculated with Tukey's Bisquare function and tuned to have 85% normal efficiency, applied to the relevant carriers only. The estimator was calculated using the lmRob() function from the `robust` R package. Once again, note that this is not a feasible estimator, and is included for benchmarking purposes only.

4.1. Scenarios

To evaluate the estimators we generate two independent samples of size n of the model $y = \mathbf{x}^T\boldsymbol{\beta}_0 + u$, plus an intercept that is equal to zero. The first sample, called the training sample, is used to fit the estimates and the second sample, called the testing sample, is used to evaluate the prediction accuracy of the estimates. We considered three possible distributions for the errors: a zero mean normal distribution, Student's t-distribution with three degrees of freedom ($t(3)$) and Student's t-distribution with one degree of freedom ($t(1)$). The first case corresponds to the classical scenario of normal errors, the second case has heavy-tailed errors and the third case has extremely heavy-tailed errors. For the first two cases we use the prediction root mean squared error (RMSE) to evaluate the prediction accuracy of the estimates. For the third case, since Student's t-distribution with one degree of freedom does not have a finite first moment, we use the median of the absolute value (MAD) of the prediction residuals as a measure of the estimators prediction accuracy. We also evaluate the variable selection performance of the estimators by calculating the false negative ratio (FNR), that is, the fraction of coefficients erroneously set to zero, and the false positive ratio (FPR), the fraction of coefficients erroneously not set to zero.

We consider the following seven scenarios for the sample size, the number of covariates, $\boldsymbol{\beta}_0$ and the distribution of the carriers.

1. We take $p = 8$, $n = 40$ and $\boldsymbol{\beta}_0$ given by: component 1 is 3, component 2 is 1.5, component 6 is 2 and the rest of the coordinates are set to zero. We take $\mathbf{x} \backsim N_p(\mathbf{0}, \boldsymbol{\Sigma})$ with $\Sigma_{i,j} = \rho^{|i-j|}$ with $\rho = 0.5$. For the case of normally distributed errors, we take the standard deviation of the errors to be $\sigma = 3$.
2. We take $p = 30$, $n = 100$ and $\boldsymbol{\beta}_0$ given by: components 1–5 are 2.5, components 6–10 are 1.5, components 11–15 are 0.5 and the rest are zero. We take $\mathbf{x} \backsim N_p(\mathbf{0}, \boldsymbol{\Sigma})$ with $\Sigma_{i,j} = \rho^{|i-j|}$ with $\rho = 0.95$. For the case of normally distributed errors, we take the standard deviation of the errors to be $\sigma = 1.5$.
3. We take $p = 200$, $n = 100$ and $\boldsymbol{\beta}_0$ given by: components 1–5 are 2.5, components 6–10 are 1.5, components 11–15 are 0.5 and the rest are zero. The first 15 covariates $(x_1, \dots, x_{15})$ and the remaining 185 covariates $(x_{16}, \dots, x_{200})$ are independent. The first 15 covariates have a zero mean multivariate normal distribution. The pairwise correlation between the ith and jth components of $(x_1, \dots, x_{15})$ is $\rho^{|i-j|}$ with $\rho = 0.95$ for $i, j = 1, \dots, 15$. The final 185 covariates have a zero mean multivariate normal distribution. The pairwise correlation between the ith and jth components of $(x_{16}, \dots, x_{200})$ is $\rho^{|i-j|}$ with $\rho = 0.95$ for $i, j = 16, \dots, 200$. For the case of normally distributed errors, we take the standard deviation of the errors to be $\sigma = 1.5$.
4. We take $p = 250$, $n = 50$ and $\boldsymbol{\beta}_0$ given by: component 1 is 3, component 2 is 1.5, component 6 is 2 and the rest of the coordinates are set to zero. We take $\mathbf{x} \backsim N_p(\mathbf{0}, \boldsymbol{\Sigma})$ with $\Sigma_{i,j} = \rho^{|i-j|}$ and $\rho = 0.5$. For the case of normally distributed errors, we take the standard deviation of the errors to be $\sigma = 3$.
5. This scenario is just like Scenario 1, but with the first coordinate of $\boldsymbol{\beta}_0$ equal to 3, and the rest equal to 0.
6. This scenario is just like Scenario 3, but with coordinates 1–5 of $\boldsymbol{\beta}_0$ equal to 2.5, coordinates 7–8 equal to 1.5, and the rest equal to 0.
7. We take $p = 50$, $n = 100$ and $\boldsymbol{\beta}_0$ with all its components equal to 0.1. We take $\mathbf{x} \backsim N_p(\mathbf{0}, \boldsymbol{\Sigma})$ with $\Sigma_{i,j} = \rho^{|i-j|}$ and $\rho = 0.8$. For the case of normally distributed errors, we take the standard deviation of the errors to be $\sigma = 1$.

In Scenario 1 we have a moderately high p/n ratio. In Scenario 2 we have $p < n$ and high p/n ratio and in Scenarios 3 and 4 we have $p > n$. Scenario 1 was analysed by Tibshirani (1996) and Fan and Li (2001). Scenarios 2 and 3 were analysed in Huang et al. (2008). Scenarios 5 and 6 are the same as Scenarios 1 and 3, but with half the number of non-zeros coordinates. In Scenario 7, we have a very high p/n ratio and a non-sparse real regression parameter; in this case we do not include 'Oracle' estimators or compute FPRs.

Since the MM-NNG, the ESL-Lasso and the WW-SCAD can only be computed for $p < n$, we only compute them for Scenarios 1, 2, 5 and 7. The MM-NNG estimator could not be computed for Scenarios 1, 5 and 7 because the program crashed. The ESL-Lasso could not be computed for Scenario 7, because the program crashed.

To evaluate the robustness of the estimators for the case of high-leverage outliers, we introduce contamination in all seven scenarios, for the case of normal errors. Note that we only contaminate the training sample and not the testing sample. We take $m = [0.1n]$ and for $i = 1, .., m$ we set $y_i = 5y_0$ and $\mathbf{x}_i = (5, 0, \dots, 0)$. We moved y_0 in a uniformly spaced grid between 0 and 3 with step 0.1 and then between 3 and 10 with step 1. To summarize the results for the contaminated scenarios we report for each estimator the maximum RMSE, FNR and FPR over all outlier sizes y_0.

The number of Monte Carlo replications for the uncontaminated scenarios was $M = 500$. The number of Monte Carlo replications for contaminated scenarios was reduced to $M = 100$, to keep computation times reasonably low.

4.2. Results

We now present the results of our simulation study. All results are rounded to two decimal places. Tables 1–4 show the results for Scenarios 1 through 7, without contamination. In Tables 5–8 we show the results for Scenarios 1 through 7 under high-leverage contamination.

The results may be summarized as follows.

- The MM-Lasso and the adaptive MM-Lasso estimators show a good behaviour for all three error distributions, with prediction errors generally close to those of the Oracle. Moreover, they show the best overall behaviour in the presence of high leverage outliers. As expected, the adaptive MM-Lasso has a lower FPR than the MM-Lasso, but a higher FNR.
- The ESL-Lasso estimator has an acceptable performance in Scenario 1 and a good performance in Scenario 5, but the worst performance of all the estimators for Scenario 2.

Table 1
Results for normal, $t(3)$ and $t(1)$ distributed errors for Scenarios 1 and 2. RMSE, MAD, FNR and FPR, averaged over 500 replications are reported for each estimator.

Scenario	Normal			$t(3)$			$t(1)$		
	RMSE	FNR	FPR	RMSE	FNR	FPR	MAD	FNR	FPR
1									
MM-Lasso	3.42	0.04	0.52	1.77	0	0.52	1.36	0.01	0.50
adaMM-Lasso	3.43	0.09	0.27	1.75	0	0.20	1.32	0.02	0.21
ESL-Lasso	4.09	0.41	0.06	1.94	0.04	0.04	1.58	0.10	0.02
Sparse-LTS	3.92	0.03	0.82	1.91	0	0.85	1.44	0	0.69
WW-SCAD	3.46	0.25	0.10	1.75	0	0.11	1.97	0.37	0.04
LAD-Lasso	3.51	0.05	0.54	1.82	0	0.57	1.40	0	0.56
Lasso	3.33	0.02	0.43	1.84	0	0.46	9.9	0.38	0.28
adaLasso	3.28	0.06	0.26	1.82	0.01	0.29	10	0.46	0.19
Oracle	3.15	0	0	1.69	0	0	1.16	0	0
2									
MM-Lasso	1.69	0.13	0.21	1.75	0.10	0.27	1.28	0.15	0.17
adaMM-Lasso	1.77	0.26	0.09	1.80	0.21	0.09	1.39	0.29	0.06
MM-NNG	1.97	0.40	0.09	1.95	0.35	0.09	1.58	0.46	0.08
ESL-Lasso	9.73	0.71	0.14	9.97	0.68	0.15	6.92	0.71	0.12
Sparse-LTS	2.25	0	1	2.14	0	1	1.78	0.01	0.97
WW-SCAD	1.83	0.41	0.08	1.89	0.37	0.08	1.95	0.61	0.08
LAD-Lasso	1.73	0.13	0.25	1.78	0.10	0.23	1.32	0.15	0.23
Lasso	1.74	0.11	0.22	1.90	0.12	0.27	10.7	0.55	0.21
adaLasso	1.74	0.21	0.13	1.94	0.22	0.14	10.7	0.72	0.09
Oracle	1.63	0	0	1.73	0	0	1.27	0	0

Table 2
Results for normal, $t(3)$ and $t(1)$ distributed errors for Scenarios 3 and 4. RMSE, MAD, FNR and FPR, averaged over 500 replications are reported for each estimator.

Scenario	Normal			$t(3)$			$t(1)$		
	RMSE	FNR	FPR	RMSE	FNR	FPR	MAD	FNR	FPR
3									
MM-Lasso	1.92	0.16	0.08	1.89	0.11	0.06	1.42	0.16	0.05
adaMM-Lasso	1.94	0.29	0.03	1.89	0.22	0.02	1.47	0.31	0.01
Sparse-LTS	1.89	0.13	0	1.98	0.11	0	1.47	0.15	0
LAD-Lasso	1.81	0.14	0.07	1.86	0.11	0.06	1.41	0.16	0.05
Lasso	1.88	0.11	0.21	2.12	0.12	0.22	6.33	0.57	0.10
adaLasso	2.06	0.18	0.13	2.29	0.19	0.14	6.75	0.70	0.10
Oracle	1.64	0	0	1.77	0	0	1.29	0	0
4									
MM-Lasso	4.05	0.12	0.07	1.99	0	0.06	2.05	0.08	0.05
adaMM-Lasso	3.99	0.18	0.03	1.80	0.01	0.01	1.79	0.12	0.02
Sparse-LTS	4.72	0.26	0.12	2.60	0.04	0.09	2.22	0.07	0.11
LAD-Lasso	4.02	0.09	0.10	2.03	0	0.10	1.96	0.06	0.08
Lasso	3.67	0.05	0.07	2.04	0.01	0.07	30.5	0.62	0.03
adaLasso	3.97	0.06	0.06	2.26	0.01	0.06	31.3	0.64	0.02
Oracle	3.13	0	0	1.67	0	0	1.12	0	0

- The MM-NNG estimator works well in Scenario 2, with or without outliers.
- The Sparse-LTS estimator shows the best overall behaviour for Scenarios 3 and 6. However, it is very much affected by high leverage outliers in Scenarios 2 and 7. Note also that its FPR is rather high in Scenarios 1 and 2.
- The WW-SCAD estimator works well in Scenarios 1, 2, and 5 with normal or $t(3)$ errors. However, its prediction error is somewhat higher than those of the other robust estimators for $t(1)$ errors. Moreover, its performance can be significantly worse than those of the MM-Lasso and adaptive MM-Lasso when there are high leverage outliers.
- The LAD-Lasso estimator shows a good behaviour in all scenarios when there are no added outliers, but its performance can be significantly worse than those of the MM-Lasso and adaptive MM-Lasso when there are high leverage outliers in the data. See the results in Table 5 for Scenario 2, for example.
- The Lasso and the adaptive Lasso show the best overall behaviour for the case of normal errors and a good behaviour for the case of $t(3)$ errors, in both cases with prediction errors close to the Oracle. However they are heavily affected by outliers or extremely heavy tailed errors such as $t(1)$ distributed errors.

Table 3
Results for normal, $t(3)$ and $t(1)$ distributed errors for Scenarios 5 and 6. RMSE, MAD, FNR and FPR, averaged over 500 replications are reported for each estimator.

Scenario	Normal			$t(3)$			$t(1)$		
	RMSE	FNR	FPR	RMSE	FNR	FPR	MAD	FNR	FPR
5									
MM-Lasso	3.29	0.01	0.35	1.73	0.00	0.33	1.24	0.00	0.32
adaMM-Lasso	3.29	0.01	0.21	1.72	0.00	0.16	1.20	0.00	0.12
ESL-Lasso	3.43	0.15	0.05	1.70	0.00	0.04	1.16	0.00	0.02
Sparse-LTS	3.83	0.00	0.67	1.87	0.00	0.64	1.36	0.00	0.53
WW-SCAD	3.16	0.01	0.09	1.69	0.00	0.11	1.32	0.12	0.05
LAD-Lasso	3.39	0.01	0.45	1.75	0.00	0.43	1.26	0.00	0.41
Lasso	3.22	0.00	0.24	1.79	0.00	0.28	5.30	0.34	0.21
adaLasso	3.18	0.00	0.17	1.75	0.00	0.18	5.32	0.36	0.14
Oracle	3.09	0.00	0.00	1.67	0.00	0.00	1.09	0.00	0.00
6									
MM-Lasso	1.78	0.04	0.07	1.76	0.01	0.05	1.28	0.04	0.04
adaMM-Lasso	1.77	0.11	0.02	1.75	0.05	0.01	1.28	0.13	0.01
Sparse-LTS	1.70	0.01	0.01	1.78	0.00	0.01	1.27	0.02	0.01
LAD-Lasso	1.73	0.03	0.06	1.77	0.01	0.06	1.29	0.03	0.05
Lasso	1.77	0.01	0.19	1.97	0.02	0.20	7.24	0.49	0.08
adaLasso	1.94	0.03	0.12	2.15	0.05	0.12	7.58	0.63	0.04
Oracle	1.56	0.00	0.00	1.68	0.00	0.00	1.12	0.00	0.00

Table 4
Results for normal, $t(3)$ and $t(1)$ distributed errors for Scenario 7. RMSE, MAD and FNR, averaged over 500 replications are reported for each estimator.

Scenario	Normal		$t(3)$		$t(1)$	
	RMSE	FNR	RMSE	FNR	MAD	FNR
7						
MMLasso	1.19	0.46	1.84	0.52	1.33	0.61
adaMM-Lasso	1.25	0.65	1.89	0.70	1.41	0.76
Sparse-LTS	1.94	0.00	2.68	0.00	2.64	0.03
WW-SCAD	1.33	0.74	2.01	0.79	1.73	0.88
LAD-Lasso	1.22	0.45	1.86	0.51	1.36	0.58
Lasso	1.22	0.50	1.93	0.62	5.98	0.92
adaLasso	1.27	0.69	1.96	0.76	6.21	0.95

Table 5
Results for Scenarios 1 and 2 with normal errors and 10% contaminated observations. Maximum RMSEs, FNRs and FPRs over all outlier sizes are averaged over 100 replications.

Scenario	Max. RMSE	Max. FNR	Max. FPR
1			
MM-Lasso	4.38	0.11	0.55
adaMM-Lasso	4.49	0.23	0.32
ESL-Lasso	4.87	0.60	0.19
Sparse-LTS	4.92	0.07	0.95
WW-SCAD	5.69	0.54	0.18
LAD-Lasso	4.61	0.41	0.52
Lasso	5.78	0.27	0.49
adaLasso	6.14	0.36	0.33
Oracle MM	3.71	0	0
2			
MM-Lasso	2.02	0.20	0.35
adaMM-Lasso	2.11	0.36	0.21
MM-NNG	2.33	0.48	0.24
ESL-Lasso	13.69	0.83	0.21
Sparse-LTS	3.18	0	1
WW-SCAD	3.36	0.55	0.21
LAD-Lasso	3.31	0.29	0.34
Lasso	3.05	0.25	0.26
adaLasso	3.24	0.41	0.15
Oracle MM	2.09	0	0

Table 6
Results for Scenarios 3 and 4 with normal errors and 10% contaminated observations. Maximum RMSEs, FNRs and FPRs over all outlier sizes are averaged over 100 replications.

Scenario	Max. RMSE	Max. FNR	Max. FPR
3			
MM-Lasso	2.48	0.21	0.15
adaMM-Lasso	2.72	0.37	0.05
Sparse-LTS	2.14	0.22	0
LAD-Lasso	3.63	0.39	0.08
Lasso	20.25	0.64	0.15
adaLasso	13.03	0.79	0.06
Oracle MM	2.09	0	0
4			
MM-Lasso	4.97	0.36	0.08
adaMM-Lasso	5.08	0.45	0.04
Sparse-LTS	5.40	0.47	0.11
LAD-Lasso	5.13	0.45	0.10
Lasso	6.04	0.42	0.07
adaLasso	7.89	0.45	0.06
Oracle MM	3.68	0	0

Table 7
Results for Scenarios 5 and 6 with normal errors and 10% contaminated observations. Maximum RMSEs, FNRs and FPRs over all outlier sizes are averaged over 100 replications.

Scenario	Max. RMSE	Max. FNR	Max. FPR
5			
MM-Lasso	4.14	0.20	0.48
adaMM-Lasso	4.17	0.30	0.31
ESL-Lasso	4.31	0.87	0.18
Sparse-LTS	4.85	0.16	0.90
WW-SCAD	4.92	0.67	0.17
LAD-Lasso	4.25	0.27	0.51
Lasso	5.69	0.09	0.44
adaLasso	6.03	0.20	0.34
Oracle MM	3.50	0.00	0.00
6			
MM-Lasso	2.29	0.11	0.13
adaMM-Lasso	2.34	0.25	0.05
Sparse-LTS	1.97	0.17	0.01
LAD-Lasso	3.59	0.36	0.09
Lasso	14.13	0.94	0.09
adaLasso	14.09	0.97	0.03
Oracle-MM	1.99	0.00	0.00

Table 8
Results for Scenario 7 with normal errors and 10% contaminated observations. Maximum RMSEs, FNRs and FPRs over all outlier sizes are averaged over 100 replications.

Scenario	Max. RMSE	Max. FNR
7		
MM-Lasso	1.33	0.51
adaMM-Lasso	1.41	0.70
Sparse-LTS	4.06	0.01
WW-SCAD	8.72	0.31
LAD-Lasso	1.89	0.69
Lasso	5.34	0.85
adaLasso	5.85	0.87

In Fig. 1 we show the RMSEs of the estimators as a function of the outlier size for Scenario 1. The MM-Lasso has the overall best behaviour, followed closely by the adaptive MM-Lasso. Note that even though the maximum RMSE of the LAD-Lasso is in this case similar to that of the MM-Lasso, Fig. 1 shows that the overall behaviour of the MM-Lasso is substantially better.

Finally, we calculated the computing times of the estimators for Scenarios 1–4, for the case of normal errors and no contamination. Since the computing times for the adaptive MM-Lasso and the MM-Lasso, and for Lasso and the adaptive Lasso were very similar we only report the results for the adaptive MM-Lasso and the adaptive Lasso. Computing times were

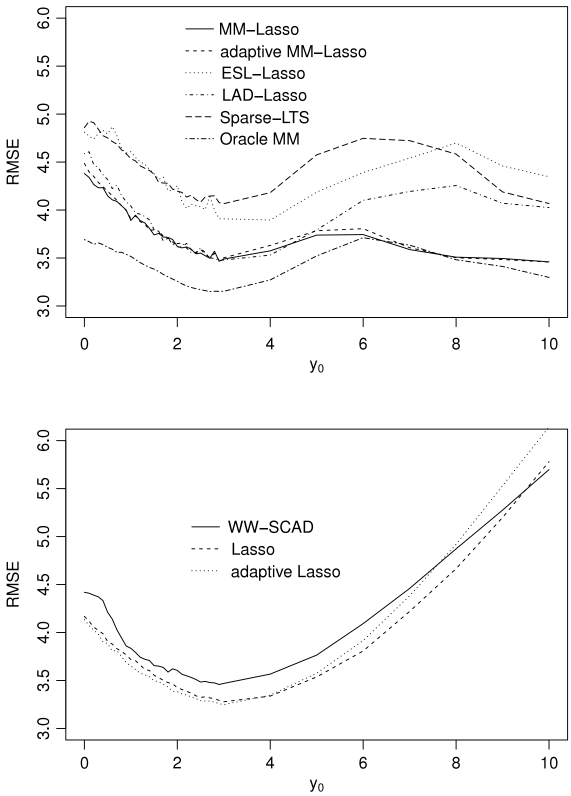

Fig. 1. RMSEs as a function of outlier sizes for each of the estimators for the first scenario, with $p = 8$, $n = 40$, normal errors and 10% contamination. RMSEs are averaged over 100 replications.

Table 9
Computing times in seconds for the estimators, averaged over 100 replications.

	Scenario 1	Scenario 2	Scenario 3	Scenario 4
adaMM-Lasso	3.46	7.58	44.59	9.05
MM-NNG	–	22.04	–	–
ESL-Lasso	0.30	0.93	–	–
Sparse-LTS	0.78	1.77	31.12	27.35
WW-SCAD	0.27	9.73	–	–
LAD-Lasso	2.22	2.37	10.31	10.25
adaLasso	0.05	0.14	0.99	0.30

averaged over 100 replications and calculations were performed on a 3.07×4 GHz Intel Core i7 PC. Results are shown in Table 9. It is clear that the adaptive Lasso is orders of magnitude faster than the other estimators. The Sparse-LTS is generally faster than the adaptive MM-Lasso, except for Scenario 4, where the adaptive MM-Lasso is almost three times faster.

5. Two real data sets

5.1. X-ray microanalysis of archaeological glass vessels data

First we analyse data from an electron-probe X-ray microanalysis of archaeological glass vessels, where each of $n = 180$ glass vessels is represented by a spectrum on 1920 frequencies. For each vessel the contents of thirteen chemical compounds are registered. This data set appears in Janssens et al. (1998), and was previously analysed in Maronna (2011). We fit a linear model where the response variable is the content of the 13*th* chemical compound (PbO) and the carriers are the 1920 frequencies measures on each glass vessel. Since for frequencies below 15 and above 500 the values of x_{ij} are almost null and show very little variability, we keep frequencies 15–500, so that we have $p = 486$. We apply the MM-Lasso, the adaptive MM-Lasso, the Sparse-LTS, the LAD-Lasso, the Lasso and the adaptive Lasso estimators to the data.

The MM-Lasso selects seven variables. The adaptive MM-Lasso drops three of the variables selected by the MM-Lasso. The Sparse-LTS selects three variables. The LAD-Lasso selects five variables. The Lasso selects seventy one variables, the adaptive Lasso selects forty nine.

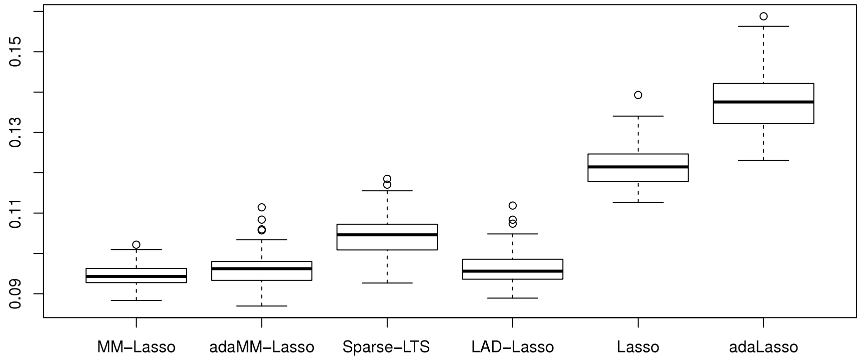

Fig. 2. Boxplot of the cross-validated τ-scale of the prediction residuals for the X-ray microanalysis data.

Table 10
Predictors selected for the air pollution data.

Predictor	MM-Lasso	adaMM-Lasso	MM-NNG	Sparse-LTS	LAD-Lasso	Lasso	adaLasso
JanTemp	1	0	1	1	1	1	1
JulyTemp	0	0	0	1	0	0	0
RelHum	0	0	0	1	1	0	0
Rain	1	1	1	1	1	1	1
Edu	1	1	1	1	1	1	1
PopDens	1	1	1	1	1	1	0
%NonWhite	1	1	1	1	1	1	1
%WhiteCollar	1	0	0	1	1	1	0
Pop	1	0	0	1	1	0	0
Pop/House	1	1	0	1	1	0	0
Income	0	0	0	1	0	0	0
LogHCPot	0	0	0	1	0	0	0
LogNOxPot	0	0	1	1	1	1	1
LogS02Pot	1	1	0	1	1	1	1

To assess the prediction accuracy of the estimators, we used 5-fold cross-validation. The criterion used was a τ-scale of the residuals, calculated as in Maronna and Zamar (2002). We repeated the procedure for 100 random splits. A boxplot of the τ-scales is shown in Fig. 2. The MM-Lasso, the adaptive MM-Lasso, the Sparse-LTS and the LAD-Lasso have similar behaviours, whereas the Lasso and the adaptive Lasso have higher prediction errors.

5.2. Air pollution data

Now, we analyse a data set on air pollution and mortality in 60 metropolitan areas in the United States. The data is available at http://lib.stat.cmu.edu/DASL/Datafiles/SMSA.html and was also analysed in Gijbels and Vrinssen (2015). The response variable is age-adjusted mortality. The 14 predictor variables are: mean January temperature (in degrees Fahrenheit), mean July temperature (in degrees Fahrenheit), relative humidity, annual rainfall (in inches), median education, population density, percentage of non-whites, percentage of white collar workers, population, population per household, median income, hydrocarbon pollution potential, nitrous oxide pollution potential and sulphur dioxide pollution potential. Observation number 21 contains two missing values and hence was removed. Since the pollution variables were skewed, we transformed them to logarithmic scale.

We applied all the estimators considered in the previous section to the data. WW-SCAD and ESL-Lasso could not be included because the programs crashed.

Table 10 shows the predictor variables selected by each estimator. Annual rainfall, median education and percentage of non-whites are selected by all the estimators. Note that Sparse-LTS selects all the variables.

To assess the prediction accuracy of the estimators, we used 5-fold cross-validation. We repeated the procedure for 100 random splits. The criterion used was again a τ-scale of the residuals. A boxplot of the τ-scales is shown in Fig. 3. We see that the results for all the estimators, except the Sparse-LTS, are similar, with the MM-Lasso and adaptive MM-Lasso having a slightly better performance than the competitors.

6. Conclusions

We have studied the robust and asymptotic properties of MM-Lasso and adaptive MM-Lasso regression estimators. We derived the asymptotic distribution of MM-Lasso estimators and proved that adaptive MM-Lasso estimators can have the oracle property.

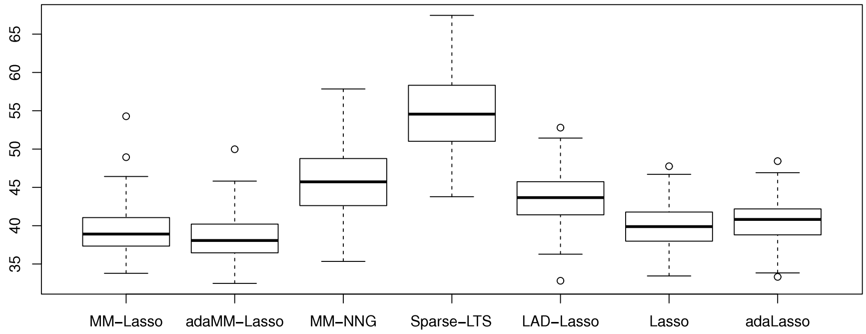

Fig. 3. Boxplot of the cross-validated τ-scale of the prediction residuals for the air pollution data.

Our simulation study shows that, in the scenarios considered in this paper, the MM-Lasso and adaptive MM-Lasso estimators provide the best balance between prediction accuracy and sparse modelling for uncontaminated samples, and stability in the presence of high leverage outliers. The adaptive MM-Lasso reduces the false positive ratio of the MM-Lasso, with the unpleasant, and foreseeable, side effect of an increase in the false negative ratio.

Acknowledgements

Research was partially supported by Grants W276 from Universidad de Buenos Aires, PIPs 112-2008-01-00216 and 112-2011-01-00339 from CONICET and PICT 2011-0397 from ANPCYT, Argentina. This paper is part of Ezequiel Smucler's Ph.D. thesis at Universidad de Buenos Aires; he was supported by a Doctoral Fellowship from CONICET. The author's would like to thank an Associate Editor and two reviewers whose thoughtful comments led to an improved version of this paper.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.csda.2017.02.002.

References

Alfons, A., Croux, C., Gelper, S., 2013. Sparse least trimmed squares regression for analyzing high-dimensional large data sets. Ann. Appl. Stat. 7 (1), 226–248.
Alfons, A., Croux, C., Gelper, S., 2016. Robust groupwise least angle regression. Comput. Statist. Data Anal. 93 (C), 421–435.
Bühlmann, P., van de Geer, S., 2011. Statistics for High-Dimensional Data: Methods, Theory and Applications. In: Springer Series in Statistics, Springer, Berlin, Heidelberg.
Davies, P.L., Gather, U., 2005. Breakdown and groups. Ann. Statist. 33 (3), 977–1035.
Donoho, D.L., Huber, P.J., 1983. The notion of breakdown point. In: Bickel, P.J., Doksum, K.A., Hodges, Jr., J.L. (Eds.), A Festschrift for Erich L. Lehmann. Wadsworth, pp. 157–185.
Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., 2004. Least angle regression. Ann. Statist. 32 (2), 407–499.
Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Amer. Statist. Assoc. 96 (456), 1348–1360.
Fasano, M.V., Maronna, R., Sued, M., Yohai, V.J., 2012. Continuity and differentiability of regression M functionals. Bernoulli 18 (4), 1284–1309.
Frank, I.E., Friedman, J.H., 1993. A statistical view of some chemometrics regression tools. Technometrics 35 (2), 109–135.
Gijbels, I., Vrinssen, I., 2015. Robust nonnegative garrote variable selection in linear regression. Comput. Statist. Data Anal. 85, 1–22.
Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12 (1), 55–67.
Hossjer, O., 1992. On the optimality of S-estimators. Statist. Probab. Lett. 14 (5), 413–419.
Huang, J., Horowitz, J.L., Ma, S., 2008. Asymptotic properties of bridge estimators in sparse high-dimensional regression models. Ann. Statist. 36 (2), 587–613.
Janssens, K.H., Deraedt, I., Schalm, O., Veeckman, J., 1998. Composition of 15–17th Century Archaeological Glass Vessels Excavated in Antwerp, Belgium. Springer, Vienna, Vienna, pp. 253–267.
Johnson, B.A., Peng, L., 2008. Rank-based variable selection. J. Nonparametr. Stat. 20 (3), 241–252.
Khan, J.A., Aelst, S.V., Zamar, R.H., 2007. Robust linear model selection based on least angle regression. J. Amer. Statist. Assoc. 102 (480), 1289–1299.
Knight, K., Fu, W., 2000. Asymptotics for lasso-type estimators. Ann. Statist. 28 (5), 1356–1378.
Lambert-Lacroix, S., Zwald, L., 2011. Robust regression through the hubers criterion and adaptive lasso penalty. Electron. J. Stat. 5, 1015–1053.
Leng, C., 2010. Variable selection and coefficient estimation via regularized rank regression. Statist. Sinica 20 (1), 167–181.
Li, G., Peng, H., Zhu, L., 2011. Nonconcave penalized m-estimation with a diverging number of parameters. Statist. Sinica 21 (1), 391–419.
Loh, P.-L., 2015. Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. ArXiv e-prints.
Maronna, R.A., 2011. Robust ridge regression for high-dimensional data. Technometrics 53 (1), 44–53.
Maronna, R.A., Martin, D.R., Yohai, V.J., 2006. Robust Statistics: Theory and Methods. Wiley.
Maronna, R.A., Yohai, V.J., 2015. High finite-sample efficiency and robustness based on distance-constrained maximum likelihood. Comput. Statist. Data Anal. 83, 262–274.
Maronna, R.A., Zamar, R.H., 2002. Robust estimates of location and dispersion for high-dimensional datasets. Technometrics 44 (4), 307–317.
Öllerer, V., Alfons, A., Croux, C., 2016. The shooting s-estimator for robust regression. Comput. Statist. 31 (3), 829–844.
Öllerer, V., Croux, C., Alfons, A., 2015. The influence function of penalized regression estimators. Statistics 49 (4), 741–765.
Rousseeuw, P.J., Yohai, V.J., 1984. Robust regression by means of S-estimators. In: Franke, J., Härdle, W., Martin, D. (Eds.), Robust and Nonlinear Time Series Analysis. Springer, US, pp. 256–272.
Smucler, E., Yohai, V.J., 2015. Highly robust and highly finite sample efficient estimators for the linear model. In: Nordhausen, K., Taskinen, S. (Eds.), Modern Nonparametric, Robust and Multivariate Methods: Festschrift in Honour of Hannu Oja. Springer International Publishing, Cham, pp. 91–108.
Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. 58 (1), 267–288.

Wang, X., Jiang, Y., Huang, M., Zhang, H., 2013. Robust variable selection with exponential squared loss. J. Amer. Statist. Assoc. 108 (502), 632–643.
Wang, L., Li, R., 2009. Weighted wilcoxon-type smoothly clipped absolute deviation method. Biometrics 65 (2), 564–571.
Wang, H., Li, G., Jiang, G., 2007. Robust regression shrinkage and consistent variable selection through the lad-lasso. J. Bus. Econom. Statist. 25 (3), 347–355.
Yohai, V.J., 1987. High breakdown-point and high efficiency robust estimates for regression. Ann. Statist. 15 (2), 642–656.
Yohai, V.J., Zamar, R.H., 1988. High breakdown-point estimates of regression by means of the minimization of an efficient scale. J. Amer. Statist. Assoc. 83 (402), 406–413.
Zou, H., 2006. The adaptive lasso and its oracle properties. J. Amer. Statist. Assoc. 101 (476), 1418–1429.
Zou, H., Yuan, M., 2008. Composite quantile regression and the oracle model selection theory. Ann. Statist. 36 (3), 1108–1126.