Penalised robust estimators for sparse and high-dimensional linear models

Umberto Amato
Istituto per la Microelettronica e Microsistemi, National Research Council
Via Pietro Castellino 111, 80131 Napoli (Italy)

Anestis Antoniadis
Laboratoire Jean Kuntzmann, Université Joseph Fourier
BP 53, 38041 Grenoble Cedex 09 (France), and
Department of Statistical Sciences, University of Cape Town
Rondebosch 7701, Cape Town, South Africa

Italia De Feis*
Istituto per le Applicazioni del Calcolo 'M. Picone', National Research Council
Via Pietro Castellino 111, 80131 Napoli (Italy)

Irene Gijbels
Department of Mathematics and Leuven Statistics Research Center (LStat), KU Leuven
Celestijnenlaan 200B, Box 2400, B-3001 Leuven (Belgium)

Abstract

We introduce a new class of robust mean M-estimators for performing simultaneous parameter estimation and variable selection in high-dimensional regression models. We first explain the motivations for the key ingredient of our procedures which are inspired by regularization methods used in wavelet thresholding procedures in noisy signal processing. The derived penalized estimation procedures are shown to enjoy theoretically the oracle property both in the classical finite dimensional case as well as the high-dimensional case when the number of variables p is not fixed but can grow with the sample size n, and to achieve optimal asymptotic rates of convergence. A fast accelerated proximal gradient (APG) algorithm, of coordinate descent type, is proposed and implemented for computing the estimates and appears to be surprisingly efficient in solving the corresponding regularization problems including the case for ultra high-dimensional data where $p \gg n$. Finally, a very extensive simulation study and some real data analysis, compare several recent existing M-estimation procedures with the ones proposed in the paper, and demonstrate their

*e-mail i.defeis@iac.cnr.it; telephone +39 081 6132390; fax +39 081 6132597

utility and their advantages. Supplementary materials for this article are available from the authors.

Keywords: Contamination; Outliers; High-dimensional regression; Variable selection; Wavelet thresholding; Nonconvex penalties; Regularization.

AMS 2000 subject classification: Primary 62H12 – 62G08; secondary 62G10.

Noname manuscript No.
(will be inserted by the editor)

Penalised robust estimators for sparse and high-dimensional linear models

Received: date / Accepted: date

Abstract We introduce a new class of robust mean M-estimators for performing simultaneous parameter estimation and variable selection in high-dimensional regression models. We first explain the motivations for the key ingredient of our procedures which are inspired by regularization methods used in wavelet thresholding procedures in noisy signal processing. The derived penalized estimation procedures are shown to enjoy theoretically the oracle property both in the classical finite dimensional case as well as the high-dimensional case when the number of variables p is not fixed but can grow with the sample size n, and to achieve optimal asymptotic rates of convergence. A fast accelerated proximal gradient (APG) algorithm, of coordinate descent type, is proposed and implemented for computing the estimates and appears to be surprisingly efficient in solving the corresponding regularization problems including the case for ultra high-dimensional data where $p \gg n$. Finally, a very extensive simulation study and some real data analysis, compare several recent existing M-estimation procedures with the ones proposed in the paper, and demonstrate their utility and their advantages. Supplementary materials for this article are available from the authors.

Keywords Contamination; Outliers; High-dimensional regression; Variable selection; Wavelet thresholding; Nonconvex penalties; Regularization.
AMS 2000 subject classification: Primary 62H12 – 62G08; secondary 62G10.

1 Introduction

Penalised regression estimators are a popular tool for the analysis of sparse and high-dimensional data sets. However, penalised regression estimators defined using an unbounded loss function can be very sensitive to the presence of outlying observations and high leverage outliers. Moreover, it is challenging to detect outliers in high-dimensional data sets. Thus, robust estimators for sparse and high-dimensional linear regression models are in need, especially in cases 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 k/n, is low. Our paper is concerned with such type of regression scenarios, including cases of diverging number of parameters.

An initial motivation for our approach are the wavelet procedures developed in the literature (see Antoniadis (2007); see also Chang and Qu (2004), Fadili and Bullmore (2005), Gannaz (2007)) and briefly reviewed below, to estimate both the linear and the non-linear component in a semi-parametric partially linear regression model with unknown regression coefficients, an unknown nonparametric function, sparsely represented in the wavelet domain, for the non-linear component, and unobservable Gaussian distributed random errors.

More precisely, assume that responses $Z_1, \ldots, Z_n$ are observed at deterministic equidistant points $t_i = \frac{i}{n}$ of an univariate variable such as time and for fixed values $\mathbf{A}_i$, $i = 1, \ldots, n$, of some

Address(es) of author(s) should be given

p-dimensional explanatory variable and that the relation between the response and predictor values is modelled by a Partially Linear Model (PLM):

$$Z_i = \mathbf{A}_i^T \boldsymbol{\beta} + f(t_i) + u_i, \qquad i = 1 \dots n, \tag{1.1}$$

where $\boldsymbol{\beta}$ is an unknown p-dimensional real parameter vector and $f(\cdot)$ is an unknown real-valued function; the u_i's are i.i.d. normal errors with mean 0 and variance σ^2 and superscript "T" denotes the transpose of a vector or matrix. We suppose at first that $n > p$ and that $A = [\mathbf{A}_1, \dots, \mathbf{A}_n]^{\mathsf{T}}$ has full rank p. Later we will consider the case when $\boldsymbol{\beta}$ is sparse, too, with p possibly greater than n. To begin (but this will be relaxed later) assume that the sample size is $n = 2^J$ for some positive integer J. Given the observed data $(Z_i, \mathbf{A}_i)_{i=1\dots n}$, the aim is to estimate from the data the vector $\boldsymbol{\beta}$ and the function f. Most of methods developed in the literature to estimate the components of such semi-parametric partially linear model assume that the unknown nonparametric component $f(t)$ is smooth. In reality, such a strong assumption may not be satisfied. To deal with cases of a less-smooth nonparametric component, to key characteristics of variations in f and to exploit its sparse wavelet coefficients representation, the wavelet thresholding procedures assume that f belongs to a Besov space $\mathcal{B}^s_{\pi,r}([0;1])$ with $s + 1/\pi - 1/2 > 0$, the last condition ensuring in particular that evaluation of f at a given point makes sense (see e.g. Antoniadis (2007)).

In matrix notation, the PLM model specified by (1.1) can be written as

$$\mathbf{Z} = A\boldsymbol{\beta} + \mathbf{F} + \mathbf{U}, \tag{1.2}$$

where $\mathbf{Z} = \big(Z_1, \dots, Z_n\big)^T$, $A^T = \big(\mathbf{A}_1, \dots, \mathbf{A}_n\big)$ is the $p \times n$ design matrix, and $\mathbf{F} = \big(f(t_1), \dots f(t_n)\big)^T$. The noise vector $\mathbf{U} = \big(u_1, \dots, u_n\big)^T$ is a Gaussian vector with mean 0 and variance matrix $\sigma^2 I_n$. Now let $\mathbf{Y} = W\mathbf{Z}$, $X = WA$, $\boldsymbol{\gamma} = W\mathbf{F}$ and $\boldsymbol{\epsilon} = W\mathbf{U}$, where W is the discrete wavelet transform operator. Pre-multiplying (1.1) by W, we obtain the transformed model

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\gamma} + \boldsymbol{\epsilon}. \tag{1.3}$$

The orthogonality of the DWT matrix W ensures that the transformed noise vector $\boldsymbol{\epsilon}$ is still distributed as a Gaussian white noise with variance $\sigma^2 I_n$. Hence, the representation of the model in the wavelet domain not only allows to retain the partly linear structure of the model but also to exploit in an efficient way the sparsity of the wavelet coefficients in the representation of the nonparametric component.

With the basic understanding of wavelet based shrinkage rules and their links to the general concept of regularisation with appropriately chosen penalty functions, covered in depth in Section 3 of Antoniadis (2007) and due to the sparsity of γ_i, the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ in model (1.3) are estimated by penalised least squares with appropriate penalties. For some further discussion on possible penalties and their properties of convexity, and smoothness at zero, see for example, Antoniadis et al. (2011). To be specific, the wavelet based estimators are defined as follows:

$$(\hat{\boldsymbol{\beta}}_n, \hat{\boldsymbol{\gamma}}_n) = \underset{(\boldsymbol{\beta},\boldsymbol{\gamma})}{\operatorname{argmin}} \left\{ J_n(\boldsymbol{\beta},\boldsymbol{\gamma}) = \sum_{i=1}^{n} \frac{1}{2}(Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta} - \boldsymbol{\gamma}_i)^2 + \sum_{i=1}^{n} p_M(|\gamma_i|\,) \right\}, \tag{1.4}$$

for a given regularisation parameter M and the associated penalty $p_M(\cdot)$. The penalty term in the above expression penalises only the wavelet coefficients of the nonparametric part of the model. Remember that the ℓ^1-penalty is associated with the soft thresholding rule.

Let us now have a closer look at the minimisation of the criterion J_n stated in (1.4) with an ℓ^1 penalty. The function J_n in (1.4) is jointly convex in $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. Its simple form suggests that an alternating optimisation can be applied: given $\boldsymbol{\gamma}$, the optimal $\boldsymbol{\beta}$ is the ordinary least squares (OLS) estimate from regression of $\mathbf{Y} - \boldsymbol{\gamma}$ on $\boldsymbol{X}$; given $\boldsymbol{\beta}$, this ℓ^1-penalised problem is orthogonal and separable in $\boldsymbol{\gamma}$, and the optimal $\boldsymbol{\gamma}$ can be obtained by soft-thresholding. More precisely, as pointed in Antoniadis (2007), for a fixed value of $\boldsymbol{\beta}$, the criterion $J_n(\boldsymbol{\beta}, \cdot)$ is minimum at

$$\tilde{\gamma}_i(\boldsymbol{\beta}) = \operatorname{sign}(Y_i - \mathbf{X}_i^T\boldsymbol{\beta}) \left(|Y_i - \mathbf{X}_i^T\boldsymbol{\beta}| - M\right)_+ \tag{1.5}$$

Therefore, finding $\hat{\boldsymbol{\beta}}_n$, a solution to problem (1.4), amounts in finding $\hat{\boldsymbol{\beta}}_n$ minimizing the criterion $J_n(\boldsymbol{\beta}, \tilde{\boldsymbol{\gamma}}(\boldsymbol{\beta}))$. However, note that

$$J_n(\boldsymbol{\beta}, \tilde{\boldsymbol{\gamma}}(\boldsymbol{\beta})) = \sum_{i=1}^{n} \rho_M(Y_i - \mathbf{X}_i^T \boldsymbol{\beta}) \tag{1.6}$$

where ρ_M is Huber's cost functional with threshold M, defined by:

$$\rho_M(u) = \begin{cases} u^2/2 & \text{if } |u| \leq M, \\ M\ |u| - M^2/2 & \text{if } |u| > M. \end{cases} \tag{1.7}$$

For the seek of completeness, we sketch below how to derive the above statement. Minimizing expression (1.4) with respect to γ_i is equivalent in minimizing $j(\gamma_i) := \frac{1}{2}(Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta} - \gamma_i)^2 + M|\gamma_i|$. The first order conditions for this are : $j'(\gamma_i) = \gamma_i - (Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta}) + \text{sign}(\gamma_i)\ M = 0$ where j' denotes the derivative of j. Now,

- if $\gamma_i \geq 0$, then $j'(\gamma_i) = 0$ if and only if $\gamma_i = Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta} - M$. Hence, if $Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta} \leq M$, $\gamma_i = 0$ and otherwise $\gamma_i = Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta} - M$.
- if $\gamma_i \leq 0$, $j'(\gamma_i)$ is zero if and only if $\gamma_i = Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta} + M$; therefore, if $Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta} \geq -M$, $\gamma_i = 0$ and otherwise $\gamma_i = Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta} + M$.

This proves that for a fixed value of $\boldsymbol{\beta}$, the criterion (1.4) is minimal for $\tilde{\boldsymbol{\gamma}}(\boldsymbol{\beta})$ given by expression (1.5). If we now replace $\boldsymbol{\gamma}$ in the objective function J_n we obtain $J_n(\boldsymbol{\beta}, \tilde{\boldsymbol{\gamma}}(\boldsymbol{\beta})) = \sum_{i=1}^n \frac{1}{2}(Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta} - \tilde{\gamma}_i)^2 + M\sum_{i=1}^n |\tilde{\gamma}_i|$. Now denoting by I the set $I := \{j = 1, \dots n; \quad |Y_j - \boldsymbol{X}_j\boldsymbol{\beta}| < M\}$, and by I^C its complement, we find that $J_n(\boldsymbol{\beta}, \tilde{\boldsymbol{\gamma}}(\boldsymbol{\beta})) = \frac{1}{2}\sum_I (Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta})^2 + \frac{1}{2}\sum_{I^C} M^2 + M\sum_{I^C}\Big(|Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta}| - M\Big)$ by replacing $\tilde{\gamma}_i$ with (1.5), which is exactly Huber's functional.

This result allows the computation of the estimators $\hat{\boldsymbol{\beta}}_n$ and $\hat{\boldsymbol{\gamma}}_n$ in the alternating optimisation described above by iteratively updating the above two steps. Regarding the estimation of $\boldsymbol{\beta}$, it can be obtained in a non-iterative fashion and, with an appropriate value of M, it is in fact the M-estimate associated to Huber's ψ function. This was extended by She and Owen (2011) to Huber's method with concomitant scale estimation, allowing also an estimation of σ, the scale of the noise, and yielding their Θ-IPOD procedure for outlier detection and a robust estimate of $\boldsymbol{\beta}$.

To reduce the adverse effects of outliers on parameter estimators and to deal with observations deviating from the model assumptions in linear regression problems, many robust methods have been proposed in the literature. A widely recognized strategy, typified by Huber's method, is to use general M-estimation procedures (see e.g. Huber (1981)) that are flexible and generalise straightforwardly to multi-parameter problems. A robust estimator should not be much affected by a small fraction of outliers and should be relatively efficient when compared to the ordinary least squares (OLS) estimate when the error distribution is exactly normal and there are no outliers. A popular quantitative measure of an estimator's robustness, introduced by Donoho and Huber (1983), is the finite-sample replacement breakdown point. Very loosely speaking, the finite-sample replacement breakdown point of an estimator is the maximum fraction of outliers that the estimator may tolerate without being completely ruined. However the monotone M-estimate corresponding to Huber's loss is not resistant to high leverage outliers (outliers in the predictors) because it never rejects gross outliers that have moderate or high leverage. Its breakdown point is $1/n$.

Another line of work aims at trimming outliers by minimizing the residuals over a selected subset. A key motivation for trimmed approaches is that convex loss functions with linear tail growth (such as the Huber loss) are not robust enough and have a breakdown point close to 0. Remarkably, the median of least squares residual avoids this problem, reaching breakdown point of nearly 50%; the approach is equivalent to 'trimming' a portion of the largest residuals. However, for such hard trimming procedures it is not possible, in a single step, to achieve a small asymptotic variance (high efficiency) and 50 % breakdown point simultaneously. The interested reader may find in Cerioli et al. (2018) a nice and extensive discussion on the pivotal role played by the trimming proportion on such estimates. The above remarks lead to the consideration

of sparse Least Squares (sparse LTS) for robust high-dimensional linear models (Alfons et al. (2013)) where the high breakdown point property for sparse LTS is established. On the other hand, asymptotic properties of such high-dimensional trimmed approaches are difficult to analyse and we will restrict our attention hereafter to M-estimation.

We will derive consistency and asymptotic normality results for a class of redescending M-estimators that is large enough to include both S and MM-estimators (see, e.g. Maronna and Yohai (1981)) and that can be tuned to have large breakdown point. Recall that a redescending M-estimator is an estimator associated to a robust loss whose influence function ψ tends to zero after some finite point (redescending) or at infinity (weakly redescending). To reject gross outliers, redescending or weakly redescending ψ functions are advocated, corresponding to a class of thresholdings offering little shrinkage for large components or using nonconvex penalties for solving the sparsity problem (1.4). Now, generally, for a given thresholding function δ_M (see Antoniadis (2007)) one can always find the corresponding penalty function (see Proposition 3.2 in Antoniadis (2007)). For example, the soft, hard, MCP, SCAD (smoothly clipped absolute deviation; see e.g. Antoniadis and Fan (2001) and Fan and Li (2001)), nonnegative garotte, Tukey's bisquare thresholding functions correspond to ℓ_1, Hard, MCP, SCAD, NNG and Tukey's penalty functions, respectively. Minimising the criterion (1.4) yields a sparse $\hat{\boldsymbol{\gamma}}$ for outlier detection and a robust estimate of $\boldsymbol{\beta}$. In fact, adapting Proposition 4.1 of She and Owen (2011) allows us to show that the estimates of (1.4) with such penalties are robust M-estimates with redescending or weakly redescending functions ψ_M, with the corresponding thresholding rule satisfying $\delta_M(t) + \psi_M(t) = t$, for all t. For the sake of completeness we state here the proposition and give its proof in the Appendix.

Proposition 1 *For any thresholding rule $\delta_M(\cdot)$, minimising the criterion (1.4) yields a sparse estimate $\hat{\boldsymbol{\gamma}}$ for outlier detection and an estimate of $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$, which is an M-estimate associated with an influence (redescending or weakly redescending) function ψ_M, that satisfies*

$$\delta_M(t) + \psi_M(t) = t, \quad \forall t. \tag{1.8}$$

All the above discussion explains therefore why, when we focus on regression functions that are expressible as M-estimators for estimation and variable selection for data drawn from sparse high-dimensional linear models and contaminated by outliers in both the response and the covariates, we will use specific influence functions ψ_M derived from such penalties with parameters M that control robustness, tuned sometimes in a data-dependent way.

Despite the large amount of existent work on robust M-estimators, research on high-dimensional regression estimators has mostly been limited to penalised likelihood-based approaches. To our knowledge, fewer papers have addressed high-dimensional M-estimation, and the M-estimators studied in those papers appear to be finite-sample versions of globally convex functions. Inspired by M-estimators, such as those arising in classical robust regression, we study penalised versions of robust regression estimators with losses that only possess convex curvature over local regions, leading to regularised M-estimators with highly nonconvex loss functions. A related paper which is relevant to our work is the one of Fan et al. (2017) which focuses exclusively on penalised robust regression with the Huber loss function.

The remainder of our paper is organised as follows. In Section 2, we provide the basic background concerning M–estimators and introduce various robust loss functions and regularisers/penalties to be discussed in the sequel. In Section 3, we present results concerning statistical consistency and oracle properties on regularised versions of some low-dimensional robust regression estimators, while Section 4 is devoted to establish statistical consistency and oracle properties of robust high-dimensional M-estimators for specific combinations of robust estimators and suitable nonconvex regularisers. Section 5 provides a two-step optimisation algorithm for computing the estimators. Simulations and numerical studies are given in Section 6 and the results are compared with several existing alternatives in the recent literature. In addition, Section 7 presents an application to real data. The results indicate that these data contain outliers such that robust methods are necessary for analysis. Some facts on influence functions and technical proofs are presented in the Appendices.

2 M-estimators and penalised M-estimators

To focus on penalised M-estimators, we consider a linear regression model:

$$\mathbf{Y} = X\boldsymbol{\beta}^* + \boldsymbol{\epsilon} \tag{2.1}$$

with $\mathbf{Y} = (Y_1, \ldots, Y_n)^T \in \mathbb{R}^n$ a vector of responses, $X \in \mathbb{R}^{n\times p}$ a design matrix, $\boldsymbol{\beta}^* \in \mathbb{R}^p$ a vector of parameters, the noise vector $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)^T \in \mathbb{R}^n$ having i.i.d, zero-mean components each with distribution $F = F_\epsilon$ and a density function f_ϵ. The design may be either deterministic or random. In the random design case, letting $X = [\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n]^\mathsf{T}$, we will assume that $\boldsymbol{X}_i$, $i = 1, \ldots, n$ are mean-zero i.i.d. p-dimensional covariate random vectors, independent of the noise vector. Therefore $X\boldsymbol{\beta}^*$ in eq. (2.1) is the mean of $\mathbf{Y}$ in the deterministic case or the conditional mean of $\mathbf{Y}$ on $\boldsymbol{X}$ in the random design case.

For $p < n$ this is a standard linear regression model. However, when p is large or diverges with n , in particular when $\log(p) \simeq n^\alpha$ with $0 < \alpha < 1$, a form of sparsity is imposed on the model parameters $\boldsymbol{\beta}^*$, i.e., it is imposed that $\text{supp}(\boldsymbol{\beta}^*) = \{1 \le j \le p : {\boldsymbol{\beta}^*}_j \neq 0\}$ is such that $|\text{supp}(\boldsymbol{\beta}^*)| = k^*$ is relatively small. When $\boldsymbol{\beta}^*$ is sparse so that most of its elements are zero or negligible, finding the non-negligible elements of $\boldsymbol{\beta}^*$, the so-called variable selection problem, is of particular importance.

We already discussed in the Introduction some early work on sparsity inducing estimators in signal processing and their extensions for high dimensional linear regression, including penalised least squares (LS) estimators with various penalties including l_1-penalty, Lasso (Tibshirani (1996)), concave penalty, SCAD (Fan and Li (2001)), MCP (Zhang (2010)), adaptive l_1 penalty (Zou (2006)), nonnegative garotte (Gijbels and Vrinssen (2015)), elastic net penalty (Zou and Hastie (2005)), and many more. When the error distribution F_ϵ deviates from the normal distribution, the l_2 loss function is typically changed to the log-likelihood $-\log f_\epsilon$, leading to penalised log-likelihood estimators. Unfortunately, in real life situations the error distribution F_ϵ is unknown and methods that adapt to many different distributions are needed. Following classical literature on M-estimators, penalised robust methods such as penalised quantile regression (Belloni and Chernozhukov (2011)), penalised Least Absolute Deviation estimator (Wang (2013)), RA-Lasso estimator (Fan et al. (2017)), robust adaptive Lasso (Avella Medina and Ronchetti (2014)) and many more, have been proposed. These methods penalise an empirical loss function $\mathcal{L}_n$ in the following manner

$$\hat{\boldsymbol{\beta}}(\lambda) = \underset{\boldsymbol{\beta}\in\mathbb{R}^\mathbf{p}, \|\boldsymbol{\beta}\|_1 < R}{\text{argmin}} \; \mathcal{L}_n(\boldsymbol{\beta}) + p_\lambda(\boldsymbol{\beta}), \tag{2.2}$$

where $\mathcal{L}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n \ell(Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta})$ for a chosen loss ℓ and for a suitable penalty function p_λ. The reason assuming the side condition $\|\boldsymbol{\beta}\|_1 \le R$ for some positive number R is that since we are interested in cases where the loss and the penalty may be nonconvex, such a condition will guarantee the existence of local/global optima. We will also require $\|\boldsymbol{\beta}^*\|_1 \le R$, so that the true regression vector $\boldsymbol{\beta}^*$ is feasible for the optimisation.

In this paper, we will consider loss functions $\mathcal{L}_n$ that satisfy

$$\mathbb{E}\left[\nabla\mathcal{L}_n(\boldsymbol{\beta}^*)\right] = 0, \tag{2.3}$$

where ∇h denotes the gradient or subgradient of a function h. When the population-level loss $\mathcal{L}(\boldsymbol{\beta}) := \mathbb{E}[\mathcal{L}_n(\boldsymbol{\beta})]$ is a convex function, equation (2.3) implies that $\boldsymbol{\beta}^*$ is a global optimum of $\mathcal{L}(\boldsymbol{\beta})$. When $\mathcal{L}$ is nonconvex, the condition (2.3) ensures that $\boldsymbol{\beta}^*$ is at least a stationary point.

In high-dimensional robust linear regression, typical properties discussed are model selection consistency, oracle properties and tight upper bounds on the statistical estimation error (e.g., Bradic et al. (2011), Lambert-Lacroix and Zwald (2011), Wang et al. (2014), Chen et al. (2014), Fan et al. (2017), Loh (2017)). Recall that an estimator is said to have the *oracle property* if the estimated coefficients corresponding to zero coefficients of the true regression parameter are set to zero with probability tending to one, while at the same time the coefficient corresponding to

non-zero coefficients of the true regression parameter are estimated with the same asymptotic efficiency as if we knew the correct model in advance. Our goal hereafter is to develop conditions under which certain stationary points of the optimisation problem (2.2) are statistically consistent estimators for $\boldsymbol{\beta}^*$ as n goes to ∞. Most existing work has been primarily reduced to the tools that are intrinsic to Huber's M-estimators, with the exception of Loh and Wainwright (2012) and Loh (2017). Therefore, to address cases where both loss function and penalty are nonconvex, we will rely on Loh's results when necessary.

2.1 Robust M-estimators

Recall that in classical robust regression theory an M-estimator is defined to be a solution to the score equation

$$\sum_{i=1}^{n} \psi_d \left(\frac{Y_i - \boldsymbol{X}_i^T \boldsymbol{\beta}}{\sigma} \right) \boldsymbol{X}_i = \mathbf{0}, \tag{2.4}$$

where d is a general tuning parameter of the ψ function and σ is a scale associated to the residuals $y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}$. A common practice is to fix σ at an initial robust M-estimate of scale and then optimise over $\boldsymbol{\beta}$ (see Hampel et al. (1986)). Unless otherwise specified, we consider equation (2.4) as constraining $\boldsymbol{\beta}$ with σ fixed. If σ is replaced by a robust $\sqrt{n}$-consistent estimate of σ, the results remain unchanged as we will see in the proofs. Using a good σ is only important for the iteration steps when computing the estimates to make the iterative algorithms convergent at a linear rate. For convenience of notation, whenever is not necessary, we will not explicitly use σ in the notation of the generic loss function ℓ below.

Let ρ_d be the robust loss on a scaled individual observation pair $(Y_i, \boldsymbol{X}_i)$. The scaled M-estimator is then given as the minimiser of

$$\mathcal{L}_n(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \rho_d((Y_i - \boldsymbol{X}_i^T \boldsymbol{\beta})/\sigma) = \frac{1}{n} \sum_{n=1}^{n} \ell(Y_i - \boldsymbol{X}_i^T \boldsymbol{\beta}). \tag{2.5}$$

Note that, when the design is deterministic, we have

$$\mathbb{E}\left[\nabla \mathcal{L}_n(\boldsymbol{\beta}^*)\right] = \mathbb{E}\left[\frac{1}{n\sigma} \sum_{i=1}^{n} \ell'(\epsilon_i) \boldsymbol{X}_i \right] = \frac{1}{n\sigma} \sum_{i=1}^{n} \mathbb{E}\left[\psi_d(\epsilon_i/\sigma)\right] \cdot \boldsymbol{X}_i = 0,$$

since the errors are mean zero i.i.d random variables and the influence function ψ_d is odd. When the design is random then

$$\begin{aligned} \mathbb{E}\left[\nabla \mathcal{L}_n(\boldsymbol{\beta}^*)\right] &= \mathbb{E}\left[\frac{1}{n\sigma} \sum_{i=1}^{n} \ell'(Y_i - \boldsymbol{X}_i^T \boldsymbol{\beta}^*) \boldsymbol{X}_i \right] = \frac{1}{n\sigma} \sum_{i=1}^{n} \mathbb{E}\left[\ell'(\epsilon_i) \boldsymbol{X}_i\right] \\ &= \frac{1}{n\sigma} \sum_{i=1}^{n} \mathbb{E}\left[\psi_d(\epsilon_i/\sigma)\right] \cdot \mathbb{E}\left[\boldsymbol{X}_i\right] = 0, \end{aligned}$$

since ϵ_i and $\boldsymbol{X}_i$ are zero mean i.i.d. and independent. Therefore the condition (2.3) is always satisfied. In particular, the maximum likelihood estimator corresponds to the choice $\ell(u) = -\log f_\epsilon(u)$, where f_ϵ is the probability density function of the additive errors ϵ_i. Note that when $\epsilon_i \sim N(0, \sigma^2)$, the MLE corresponds to the choice $\ell(u) = \frac{u^2}{2\sigma^2}$, and the resulting scaled loss function is convex.

The loss functions that we will adopt in this paper include the Huber loss, Tukey's biweight, MCP, LAD, nonnegative garrote loss, Welsh loss and Cauchy loss (see Appendix 1). Although second and third derivatives do not always exist for the above loss functions, a unifying property is that the derivative ℓ' is *bounded* and odd in each case. This turns out to be an important property for robustness of the resulting estimator. Indeed, intuitively, we may view a solution $\hat{\boldsymbol{\beta}}_\lambda$

of the optimisation problem (2.2) as an approximate sparse solution to the estimating equation $\nabla\mathcal{L}_n(\boldsymbol{\beta}) = 0$, or equivalently,

$$\frac{1}{n}\sum_{i=1}^{n}\ell'(Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta})\boldsymbol{X}_i = 0. \tag{2.6}$$

When $\boldsymbol{\beta} = \boldsymbol{\beta}^*$, equation (2.6) becomes

$$\frac{1}{n}\sum_{i=1}^{n}\ell'(\epsilon_i)\boldsymbol{X}_i = 0. \tag{2.7}$$

In particular, if a pair $(\boldsymbol{X}_i, Y_i)$ satisfies the linear model (2.1) but ϵ_i is an outlier, its contribution to the sum in equation (2.7) is bounded when ℓ' is bounded, lessening the contamination effect of gross outliers.

Since a *redescending M-estimator* has the additional property that there exists $x_r > 0$ such that $|\ell'(u)| = 0$, for all $|u| \geq x_r$, where x_r is a *finite rejection point*, outliers $(\boldsymbol{X}_i, Y_i)$ with $|\epsilon_i| \geq x_r$ will be completely eliminated from the summand in equation (2.7). When *weakly redescending M-estimators* are used, terms in summand in equation (2.7) corresponding to errors in the tails from a subgaussian distribution or a distribution with light-tails will be also highly downweighted. However, note that redescending and weakly redescending M-estimators are always nonconvex.

2.2 Nonconvex penalties

We already have provided in our introduction some details on appropriate penalties that enhance sparsity. Indeed, the performance of the regularised M-estimator in the high-dimensional case not only depends on the robust loss used but also on the penalty and the corresponding regularisation parameter. To select a good penalty function, Antoniadis and Fan (2001) and Fan and Li (2001) proposed three principles that a good penalty function should satisfy: unbiasedness, in which there is no over-penalisation of large coefficients to avoid unnecessary modelling biases; sparsity, as the resulting penalised least-squares estimators should follow a thresholding rule such that insignificant coefficients should be set to zero to reduce model complexity; and continuity to avoid instability and large variability in model prediction. The interested reader is referred to Theorem 1 of Antoniadis and Fan (2001) which gives the necessary and sufficient conditions on a penalty function for the solution of the penalised least-squares problem to be thresholding, continuous and approximately unbiased for large values of their argument. Following these developments and also some extra conditions on robust regularisers introduced in Loh and Wainwright (2015) we may summarise our requirements on the penalties p_λ to be as follows:

Assumption 1 (Amenable penalties) *The penalty is coordinate-separable:*

$$p_\lambda(\boldsymbol{\beta}) = \sum_{j=1}^{p} p_\lambda(\beta_j),$$

for some scalar function $p_\lambda : \mathbb{R} \mapsto \mathbb{R}$. In addition:

(i) The function $t \mapsto p_\lambda(t)$ is symmetric around zero and $p_\lambda(0) = 0$.
(ii) The function $t \mapsto p_\lambda(t)$ is nondecreasing on $\mathbb{R}^+$.
(iii) The function $t \mapsto \frac{p_\lambda(t)}{t}$ is nonincreasing on $\mathbb{R}^+$.
(iv) The function $t \mapsto p_\lambda(t)$ is differentiable for $t \neq 0$.
(v) $\lim_{t\to 0^+} p'_\lambda(t) = \lambda$.
(vi) There exists $\mu > 0$ such that the function $t \mapsto p_\lambda(t) + \frac{\mu}{2}t^2$ is convex.
(vii) There exists $\xi \in (0, \infty)$ such that $p'_\lambda(t) = 0$ for all $t \geq \xi\lambda$.

If p_λ satisfies conditions (i)–(vi) of Assumption 1, we say that p_λ is μ-*amenable*. If p_λ also satisfies condition (vii), we say that p_λ is (μ, ξ)-*amenable* (see Loh and Wainwright (2015)). In particular, if p_λ is μ-amenable, then $q_\lambda(t) := \lambda|t| - p_\lambda(t)$ is everywhere differentiable. Defining the vector version $q_\lambda : \mathbb{R}^p \to \mathbb{R}$ accordingly, it is easy to see that $\frac{\mu}{2}\|\beta\|_2^2 - q_\lambda(\beta)$ is convex.

Some examples of amenable regularizers are the smoothly clipped absolute deviation (SCAD) penalty (see Antoniadis and Fan (2001) and Fan and Li (2001)), the minimax concave penalty (MCP) (see Zhang (2010)) and the standard ℓ_1-penalty. The SCAD penalty with fixed parameter $a > 2$ is (μ, ξ)-amenable, with $\mu = \frac{1}{a-1}$ and $\xi = a$. The MCP regularizer is (μ, ξ)-amenable, with $\mu = \frac{1}{\gamma}$ and $\xi = \gamma$. The ℓ_1-penalty $p_\lambda(t) = \lambda|t|$ is an example of a regulariser that is 0-amenable, but not $(0, \xi)$-amenable, for any $\xi < \infty$.

3 Penalised linear models and robust penalised estimators in low-dimensional cases

We start from the classical linear regression model to describe the regularised regression methods. We consider the linear regression model (2.1) introduced in the previous section and assume for the moment that the distribution of the errors $F = F_\epsilon$ is symmetric around 0 with mean 0 and variance σ^2. If the response Y is not centered, the intercept may be efficiently estimated by the empirical median or mean of the observations and therefore without any loss of generality we may assume that the regression model has zero intercept.

An active line of research focuses on variable selection and recovery of a sparse vector β^* using penalised least squares with penalties that are singular at the origin, resulting in thresholding and shrinking unnecessary coefficients to 0. Common choices to perform variable selection in regression analysis in such situations are the ℓ_1 penalised least square Lasso Tibshirani (1996), the Least Angle Regression (LARS) Efron et al. (2004) closely related to the Lasso, and concave penalised regularisation methods such as SCAD Fan and Li (2001) or MCP Zhang (2010), the latter with penalties functions that are spiked at zero but flatten for large values of the coefficients (as opposed to the constant increase of the l_1 penalty of Lasso or LARS) yielding sparse solutions where large non-zero values are estimated with little bias. LARS provides an ordering in which the covariates enter the regression model. This sequence is usually the same as in Lasso but obtained in a computationally efficient way.
These methods are useful in many situations, particularly when most regression coefficients are assumed to be null. The Lasso has been studied under at least three common criteria: (i) model selection criteria, meaning the correct recovery of the support set $S = \{j \in \{1, 2, ..., p\} : \beta_j^* \neq 0\}$ of the model vector $\boldsymbol{\beta}^*$; (ii) l_2 estimation errors $||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*||_2^2$, where $\hat{\boldsymbol{\beta}}$ is the estimate of $\boldsymbol{\beta}^*$; and (iii) prediction error $||X\hat{\boldsymbol{\beta}} - X\boldsymbol{\beta}^*||_2^2$.

The Lasso estimator is defined by

$$\hat{\boldsymbol{\beta}}(\lambda_n) = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ ||Y - X\boldsymbol{\beta}||_2^2 + \lambda_n ||\boldsymbol{\beta}||_1 \right\}, \tag{3.1}$$

where $\lambda_n \geq 0$ is the tuning parameter which controls the amount of regularization applied to the estimate. Setting $\lambda_n = 0$ reverses the Lasso problem to Ordinary Least Squares (OLS) which minimizes the unregularised empirical loss.

The Lasso estimator has two nice properties, namely, (i) it generates sparse models by means of ℓ_1 regularization and (ii) it is also computationally feasible (see the LARS algorithm). The asymptotic behavior of Lasso-type estimators in low-dimensional cases has been studied by Knight and Fu (2000) for fixed p and $\boldsymbol{\beta}^*$ as $n \to \infty$. In particular, they have shown that under some regularity conditions on the design, $\lambda_n = o(n)$ is sufficient for consistency in the sense that $\hat{\boldsymbol{\beta}}(\lambda_n) \to_P \boldsymbol{\beta}^*$, and λ_n should grow more slowly (i.e. $\lambda_n = O(\sqrt{n})$) for asymptotic normality of the Lasso estimator. On the l_2 estimation error front, the Lasso has been shown to achieve l_2 convergence rates of $(k^\star \log p)/n$ (see e.g. van de Geer (2008), Bickel et al. (2009), and Meinshausen and Yu (2009), to cite only a few) which is the minimax optimal rate. Other work focuses on the convergence rates of $||X\hat{\boldsymbol{\beta}}(\lambda_n) - X\boldsymbol{\beta}^*||_2^2$ (see for example Bunea et al. (2007) and Bunea (2008)). However, even if p is fixed, there does not exist a sequence of the tuning parameter λ_n which can lead to both variable selection consistency and asymptotic normality (see Fan and Li (2001) and Zou (2006)). This mean that the Lasso or LARS estimates do not have the oracle property even when the minimum of the nonzero coefficients of the regression parameter is bounded below.

Note also that for Lasso the optimal λ is known to depend on σ_ϵ even if the loss function ℓ does not require calibration to the scale of the errors. A poor estimation of the scale of the observations may lead to a sub-optimal grid for the regularisation parameter λ and more stringent conditions on the design matrix X. One of the first papers to consider the unknown error variance case for the Lasso was Städler et al. (2010) who suggested the following penalised loss function for introducing unknown variance into the Lasso framework:

$$\frac{||Y - X\boldsymbol{\beta}||_2^2}{\sigma^2} + \frac{\lambda_n}{\sigma}||\boldsymbol{\beta}||_1 + n\log\sigma \tag{3.2}$$

Interestingly, Antoniadis (2010) and Sun and Zhang (2010) found that the resulting joint estimate of this formulation may give biased estimates for the noise level and proposed the scaled Lasso which is equivalent to the *square-root Lasso* of Belloni et al. (2011) which doesn't require a preliminary estimation of σ. The corresponding estimation procedure is called SSLasso. Interestingly enough the loss function associated to SSLasso is a penalised version of Huber's concomitant loss function, and so may be viewed as performing robust low-dimensional regression.

A way to improve variable selection accuracy and gain oracle properties is by reducing the bias of Lasso via the adaptive Lasso procedure Zou (2006) which solves the following weighted l_1 regularization problem for some $\alpha \in (0, 1)$:

$$\hat{\boldsymbol{\beta}}(\lambda_n) = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ ||Y - X\boldsymbol{\beta}||_2^2 + \lambda_n \sum_{j=1}^{n} |\hat{w}_j|^{-\alpha}|\beta_j| \right\}, \tag{3.3}$$

where $\hat{\mathbf{w}}$ is an estimator of $\boldsymbol{\beta}$ (for example, the solution of the standard unweighted Lasso with regularisation parameter λ or a ridge estimate of $\boldsymbol{\beta}$). A low-dimensional analysis in Zou (2006) shows that the adaptive Lasso solution can achieve the oracle property asymptotically. A high dimensional analysis of this procedure was given in Huang et al. (2008). For variable selection consistency and oracle properties to hold, the adaptive lasso requires strong conditions in terms of the minimal strength of nonzero components of $\boldsymbol{\beta}^*$ and some extra conditions on the design matrix X.

Although least-squares penalised regression estimates in low-dimensional situations present excellent properties, they may not be reliable if the errors have a more or less heavy-tailed distribution or if outliers exist in responses or/and predictors. It is indeed formally shown in Alfons et al. (2013) that the breakdown point of the Lasso is $1/n$, that is, only one single outlier in the response can make the Lasso estimate completely unreliable. Therefore, robust alternatives are needed. For these cases, especially when outliers are found only in the responses, least absolute deviation (LAD) regression has been proven to be a useful and robust alternative to the OLS regression. Since the least-squares loss used in the Lasso or adaptive Lasso is very sensitive to outliers, to obtain a robust lasso-type estimator, one may further modify the lasso objective function into the following LAD-Lasso criterion, which is also appropriate in a high-dimensional setup:

$$\hat{\boldsymbol{\beta}}(\lambda_n) = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \left\{ ||Y - X\boldsymbol{\beta}||_1 + \lambda_n||\boldsymbol{\beta}||_1 \right\}. \tag{3.4}$$

The estimation consistency of the penalised LAD estimator is discussed for example in Wang et al. (2007) and Lambert-Lacroix and Zwald (2011), where the number of variables p is assumed to be fixed. The LAD-Lasso estimator produces consistent variable selection and extensive simulation studies in Wang et al. (2007) demonstrate the satisfactory sample performance of the LAD-Lasso. It is worth noting that the LAD-Lasso is particularly well-suited to heavy-tailed error distributions. However, if the outliers occur in the explanatory variables (leverage points) the performance of the LAD-Lasso regression estimator is not better than the OLS estimators. As Wang and Leng (2007) point out, combining the LAD and the LASSO methods can only produce estimators that are resistant to the outliers in the response variable. To deal with variable selection with outliers occurring also in the explanatory variables the weighted LAD-Lasso regression estimation has been proposed by Arslan (2012) where some theoretical properties are derived for

the finite case p under appropriate conditions on the design and the noise distribution. Alfons et al. (2013) proposed another approach that is robust with respect to high leverage points, by adding an l_1 penalty on the coefficient estimates to the well-known least trimmed squares (LTS) estimator. They prove that sparse LTS has a high breakdown point and that somehow it can also be interpreted as a trimmed version of the lasso. In a simulation study, they show that the sparse-LTS can be resistant to multiple regression outliers, including leverage points. The Sparse-LTS estimator can be also calculated for $p > n$. However, Alfons et al. (2013) do not provide any asymptotic theory for their estimator.

In Gijbels and Vrinssen (2015) and Gijbels et al. (2017) the nonnegative garrote method is robustified for outliers in the response and in the covariates by using robust alternatives to the least squares regression estimator, such as the S-estimator and the least trimmed squares estimator. These authors also develop a re-weighting step that is related with the MM-estimator of Yohai (see, e.g. Maronna and Yohai (1981)), to increase the efficiency of the proposed robust nonnegative garrote method under the normal error model. Their robust nonnegative garrote method (NNG) performs quite well and often outperforms other available methods for linear regression models.

Other robust versions of the lasso have been considered in the literature. Rosset and Zhu (2004) proposed a Huber-type loss function, which requires knowledge of the residual scale. To adapt for different magnitude of the errors in the regression model (2.1) and robustify the estimation, Fan et al. (2017), inspired by the proposal of Rosset and Zhu (2004), propose to use in the estimating equation (2.2) the Huber loss $\ell = \rho_{1/d}$, $d > 0$:

$$\rho_d(x) = \begin{cases} \frac{1}{2}x^2 & \text{if } |x| \leq \frac{1}{d} \\ \frac{1}{d}(|x| - \frac{1}{2d}) & \text{if } |x| > \frac{1}{d} \end{cases}.$$

The Huber loss is quadratic for small values of x and linear for large values of x. The parameter d controls the blending of quadratic and linear penalisation. The least squares and the LAD can be regarded as two extremes of the Huber loss for $d = 0$ and $d = \infty$, respectively. A first difference with respect to Rosset and Zhu's traditional Huber's loss, is the parameter d that Fan *et al.* regard as a tuning parameter. If it converges to zero it reduces the biases of estimating the mean. On the other hand, d can not shrink too fast in order to maintain the robustness. In practice, d needs to be tuned by some data-driven method. By letting d vary, Fan et al. (2017) call $\rho_d(x)$ the robust approximate quadratic (RA-quadratic) loss. Using this loss, the resulting estimation problem (2.2) becomes the following convex optimisation problem:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \ \frac{1}{n}\sum_{i=1}^{n} \rho_d(Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta}) + \lambda_n \sum_{j=1}^{p} |\beta_j|. \tag{3.5}$$

The second difference is that their method can handle also the high-dimensional case and will be discussed in the sections that follow.

Again, l_1 penalised Huber type estimators are not robust with respect to leverage points, that is, outliers in the predictor space, and can handle outliers only in the response variable. The main competitor of the sparse LTS is robust least angle regression, called RLARS, and proposed in Khan et al. (2007). They develop a robust version of the LARS algorithm, essentially replacing correlations by a robust type of correlation, to sequence and select the most important predictor variables. Then a non sparse robust regression estimator is applied to the selected predictor variables. RLARS seems to be robust with respect to response outliers and to leverage points, but RLARS has not been asymptotically studied in the literature, the main reason being that the procedure is not based on the minimisation of a clearly defined objective function.

Note that since our goal is to develop conditions under which certain stationary points of the optimisation problem (2.2) are statistically consistent estimators for $\boldsymbol{\beta}^*$ for losses that may be non-convex and penalties that are not necessary l_1, we have reviewed only the most relevant methods. Finally note that in all the above methods that rely upon the l_1 penalty, one may use

adaptive versions of the objective function as it is done for the adaptive lasso, since this usually just amounts in modifying in an appropriate way the design matrix X and the regression vector.

3.1 Robust estimation with nonconvex losses and adaptive l_1 penalties when $p < n$.

In this subsection we prove consistency and asymptotic normality results when $p < n$ for a class of adaptive l_1 penalized redescending M-estimation equations including an estimate of scale that is wide enough to include both S and MM-estimators.

We consider the linear model (2.1), with centered $\mathbf{Y}$, with p finite and fixed, a deterministic design $\boldsymbol{X}_i \in \mathbb{R}^p$, $i = 1, \ldots, n$ and a true parameter vector $\boldsymbol{\beta}^*$ that is sparse, in the sense that only few k^* components (with $k^* < p$) are nonzero. For simplicity, we will assume $\boldsymbol{\beta}_0^* = (\boldsymbol{\beta}_I^*, \boldsymbol{\beta}_{II}^*)$, where $\boldsymbol{\beta}_I^* \in \mathbb{R}^{k^*}$, $\boldsymbol{\beta}_{II}^* \in \mathbb{R}^{p-k^*}$, all the coordinates of $\boldsymbol{\beta}_I^* \in \mathbb{R}^{k^*}$ are non-zero and all the coordinates of $\boldsymbol{\beta}_{II}^* \in \mathbb{R}^{p-k^*}$ are zero. We already mentioned in the previous section many existing methods that have been used and studied in the literature for handling such models and we are going to illustrate their behaviour in a later section devoted to simulations. Before handling the case of high-dimensional linear models with such losses and other amenable penalties in the general case 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 $k^\star/n$, is low, we would like to study theoretically, in an asymptotic sense, the family of estimators derived by solving the following class $\mathcal{R}$ of optimisation problems:

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \ \frac{1}{n}\sum_{i=1}^{n} \rho_\lambda \left(\frac{(Y_i - \boldsymbol{X}_i^T \boldsymbol{\beta})}{\hat{s}_n} \right) + \mu_n \sum_{j=1}^{p} \hat{w}_j |\beta_j|, \tag{3.6}$$

where ρ_λ is among one of the losses mentioned in the previous sections (MCP, NNG, Tukey, Welsh or Cauchy), $\hat{s}_n$ is a robust estimate of scale and the adaptive weights $\hat{w}_j = 1/|\hat{\beta}_j^{\mathrm{S}}|$ are based on some recently proposed penalized S-type estimates of the coefficients (see Cohen Freue et al. (2018)). The scale σ can be estimated for example by the median absolute deviation of the residuals from the full model fitted by S-estimation (see Rousseeuw and Yohai (1984)) or any other S-scale estimator. We do not consider here the absolute deviation loss or losses related to l_1 or the Huber loss since, as we already said, these have been extensively studied in the recent literature. In order to obtain the asymptotic rates of estimates solving (3.6) and to show that they achieve the oracle property, we will assume that

- A1. $\frac{1}{n} X^T X \to V$ where V is a symmetric non-negative defined matrix with its main $k^* \times k^*$ diagonal block V_1 positive-definite.
- A2. $\max_{1 \le i \le n} \|\mathbf{X}_i\|_2 / \sqrt{n} \to 0$ as $n \to \infty$
- A3. The errors ϵ_i are i.i.d with a common symmetric density f_ϵ admitting a zero mean and a finite variance σ^2.
- A4. Given the variance of the noise, the robust scale estimate $\hat{s}_n$ is $\sqrt{n}$ probability consistent.

Assumptions (A1) and (A2) are classical. Assumption (A1) is a standard assumption for consistency of the least squares estimator (see Zou (2006) and Lambert-Lacroix and Zwald (2011)). Assumption (A2) can be seen as a "compacity assumption": it is satisfied if the covariates are supposed to be bounded (see Lambert-Lacroix and Zwald (2011)). Condition (A3) is natural without prior knowledge on the distribution of the errors and ensures that the noise is not degenerated and has finite variance.

Let $\hat{\boldsymbol{\beta}}_I$ stand for the first k^* coordinates of $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_{II}$ for the remaining $p - k^*$. The following theorem shows that, as long as $n\mu_n \to \infty$ but $\sqrt{n}\,\mu_n \to 0$ as $n \to \infty$, the adaptive l_1 penalised robust estimators defined by (3.6) are asymptotically normal and variable selection consistent. More precisely we can state the following theorem whose proof is given in the Appendix.

Theorem 1 *Assume (A1)–(A4) hold. Then*

(i) $\sqrt{n}(\hat{\boldsymbol{\beta}}_I - \boldsymbol{\beta}_I) \xrightarrow{d} N_{k^*}\left(\mathbf{0}, \frac{a(\psi_\lambda, F_\epsilon)}{b(\psi, F_\epsilon)^2} \sigma^{*2} V_1^{-1}\right)$ *where*

$$a(\psi_\lambda, F_\epsilon) = \mathbb{E}_{F_\epsilon} \psi_\lambda^2(\epsilon)$$

and

$$b(\psi_\lambda, F_\epsilon) = \mathbb{E}_{F_\epsilon} \psi'_\lambda(\epsilon).$$

(ii)

$$\mathbb{P}\left(\hat{\boldsymbol{\beta}}_I = \boldsymbol{\beta}^*_I\right) \to 1,$$

and

$$\mathbb{P}\left(\hat{\boldsymbol{\beta}}_{II} = \mathbf{0}_{p-k^\star}\right) \to 1.$$

Under similar settings we already mentioned available results in the recent literature for a large variety of available methods such as robust Lasso, robust adaptive Lasso, robust LAD-Lasso, WLAD-Lasso, robust NNG-Lasso, Sparse LTS, RLARS and penalised Huber. With Theorem 1 we get asymptotic normality and variable selection consistency in the case of p fixed for criteria within the class $\mathcal{R}$ including Tukey's, Welsh's, Cauchy's, nonnegative garrote and MCP and using an adaptive l_1 penalty and an estimation of the scale by some robust S-estimator as for example Fan et al. (2017) suggest. One could also use the square-root Lasso estimate of σ suggested by Antoniadis (2010) and studied by Sun and Zhang (2012). These authors proved that the resulting estimate of σ is consistent for the "oracle" estimator. Interestingly, the scaled Lasso estimators for the regression coefficients and error variance are scale equivariant. We need now to proceed to the general case using Loh's results.

4 Robust estimation with nonconvex losses and nonconvex penalties in high-dimensional sparse linear models

In the previous section, the ambient dimension p stayed fixed while the number of observations n tends to infinity. In contrast, the analysis in this section is all within a high-dimensional framework, in which the couple (n, p), as well as other problem parameters, such as the sparsity k^* may all be allowed to tend to infinity. We will therefore assume in this section that n, p_n and k^*_n are such that $n \geq c_0 k^*_n \log p_n$, for a sufficiently large constant c_0. By known information-theoretic results of Wainwright (2009), this type of lower bound is required for any method to recover the support of a k^*-sparse signal, hence is not a limiting restriction.

Inspired by the theory on high-dimensional robust estimators developed recently by Loh and Wainwright (2015) we give some sufficient conditions under which optima of regularised robust M-estimators with separable penalties are statistically consistent, even in the presence of heavy-tailed errors and outlier contamination. The conditions involve a bound on the derivatives of the robust loss functions, as well as their restricted strong convexity in a neighbourhood of constant radius around the true parameter vector $\boldsymbol{\beta}^*$, and the conclusions are given in terms of the tails of the error distribution.

The restricted strong convexity (RSC) requirement of the loss functions is an important requirement. Denote $\widehat{\boldsymbol{\Delta}} = \widehat{\boldsymbol{\beta}}_{\lambda_n} - \boldsymbol{\beta}^*$ the difference between an optimal solution $\widehat{\boldsymbol{\beta}}_{\lambda_n}$ and the true parameter, and consider the loss difference $\mathcal{L}_n(\widehat{\boldsymbol{\beta}}_{\lambda_n}) - \mathcal{L}_n(\boldsymbol{\beta}^*)$. In the classical setting, under fairly mild conditions, one expects that the loss difference should converge to zero as the sample size n increases. It is important to note, however, that such convergence on its own is *not sufficient* to guarantee that $\widehat{\boldsymbol{\beta}}_{\lambda_n}$ and $\boldsymbol{\beta}^*$ are close or, equivalently, that $\boldsymbol{\Delta}$ is small. Rather, the closeness depends on the curvature of the loss function, as illustrated in Figure 1. The standard way to ensure that a function is "not too flat" is via the notion of strong convexity. However restricted strong convexity traditionally involves a global condition on the behaviour of the loss function. Due to the highly nonconvex behaviour of the robust regression functions of interest, we will assume only a *local* condition of restricted strong convexity in the sequel (see Negahban et al. (2012)).

We may now exploit the general statistical results of Loh (2017) concerning stationary points of the high-dimensional robust M-estimators when the loss function satisfies restricted strong

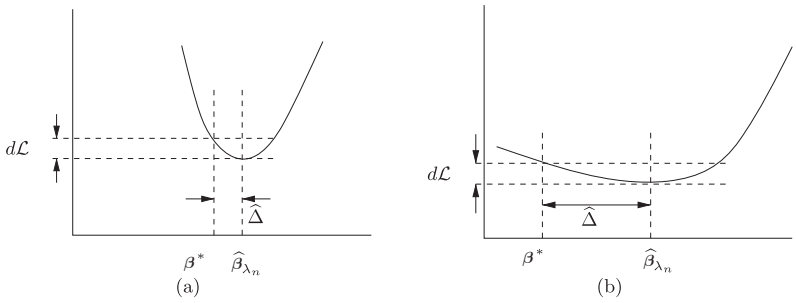

Fig. 1 Role of curvature in distinguishing parameters. (a) Loss function has high curvature around $\widehat{\boldsymbol{\Delta}}$. A small excess loss $d\mathcal{L}_n = |\mathcal{L}_n(\widehat{\boldsymbol{\beta}}_{\lambda_n}) - \mathcal{L}_n(\boldsymbol{\beta}^*)|$ guarantees that the parameter error $\widehat{\boldsymbol{\Delta}}$ is also small. (b) A less desirable setting, in which the loss function has relatively low curvature around the optimum.

convexity and the penalty used is μ-amenable, which for the sake of completeness we will summarise in a theorem restated below. The assertions of this theorem then hold with high probability provided that some conditions on the distributions of the covariates and error terms are satisfied and that the appropriate conditions are checked for the specific M-estimators and the penalties we have adopted. Recall that $\widetilde{\boldsymbol{\beta}}$ is a *stationary point* of the objective function in (2.2) if

$$\langle \nabla \mathcal{L}_n(\widetilde{\boldsymbol{\beta}}) + \nabla p_\lambda(\widetilde{\boldsymbol{\beta}}),\, \boldsymbol{\beta} - \widetilde{\boldsymbol{\beta}} \rangle \geq 0,$$

for all feasible $\boldsymbol{\beta}$, where with a slight abuse of notation, we write $\nabla p_\lambda(\widetilde{\boldsymbol{\beta}}) = \lambda \operatorname{sign}(\widetilde{\boldsymbol{\beta}}) - \nabla q_\lambda(\widetilde{\boldsymbol{\beta}})$ (recall that q_λ is differentiable by our assumptions on μ-amenability of the penalties we are using). Assume that there exist $\alpha, \tau > 0$ and a radius $r > 0$ such that the following restricted strong convexity assumption of Loh (2017) is satisfied:

$$\langle \nabla \mathcal{L}_n(\boldsymbol{\beta}_1) - \nabla \mathcal{L}_n(\boldsymbol{\beta}_2),\, \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2 \rangle \geq \alpha \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_2^2 - \tau \frac{\log p}{n} \|\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2\|_1^2, \tag{4.1}$$

for all $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbb{R}^p$ such that $\|\boldsymbol{\beta}_1 - \boldsymbol{\beta}^*\|_2, \|\boldsymbol{\beta}_2 - \boldsymbol{\beta}^*\|_2 \leq r$.

We can now state the following Theorem of Loh (2017) which guarantees that stationary points within the local region where the loss function satisfies restricted strong convexity are statistically consistent.

Theorem 2 *Suppose $\mathcal{L}_n$ satisfies the RSC condition (4.1) with $\boldsymbol{\beta}_2 = \boldsymbol{\beta}^*$ and the penalty p_λ is μ-amenable, with $\frac{3}{4}\mu < \alpha$. Suppose $n \geq Cr^2 \cdot k^* \log p$ for some constant $C > 0$, that $\|\boldsymbol{\beta}^*\|_1 \leq R$ and*

$$\lambda \geq \max\left\{ 4\|\nabla \mathcal{L}_n(\boldsymbol{\beta}^*)\|_\infty,\; 8\tau R \frac{\log p}{n} \right\}. \tag{4.2}$$

A stationary point $\widetilde{\boldsymbol{\beta}}$ of the objective function in (2.2) such that $\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq r$ exists and satisfies the bounds

$$\|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \leq \frac{24\lambda\sqrt{k}}{4\alpha - 3\mu}, \qquad \text{and} \qquad \|\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_1 \leq \frac{96\lambda k}{4\alpha - 3\mu}.$$

Note that the statement of Theorem 2 is entirely deterministic and the distributional properties of the covariates and error terms in the linear model come into play when verifying that the inequality (4.2) and the RSC condition (4.1) hold with high probability under a prescribed sample size scaling. When r is chosen to be a constant and $\frac{n}{k^* \log p} = o(1)$, as is the case in the robust regression settings that we are interested in, all stationary points within the constant-radius region are actually guaranteed to fall within a shrinking ball of radius $\mathcal{O}\left(\sqrt{\frac{k^* \log p}{n}}\right)$ centered around $\boldsymbol{\beta}^*$.

When the design is deterministic we will assume that $\|\frac{1}{n}\sum \boldsymbol{X}_i\|_\infty \leq D\sqrt{\frac{\log p}{n}}$ for some constant D which is much more general than assumption (A2) of the previous section. When the design is random we will assume that $\boldsymbol{X}$ is sub-Gaussian with parameter σ_X^2. It is then easy

to see that one may find constants $c_1 > 0$, $c_2 > 0$, κ_1 and $c > 0$ such that with probability at least $1 - c_1 \exp(-c_2 \log p)$, the loss function $\mathcal{L}_n$ satisfies the bound

$$\|\nabla \mathcal{L}_n(\boldsymbol{\beta}^*)\|_\infty \leq c\kappa_1 \sigma_X \sqrt{\frac{\log p}{n}}.$$

Indeed, we have

$$\|\nabla \mathcal{L}_n(\boldsymbol{\beta}^*)\|_\infty = \left\| \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{X}_i \cdot \ell'(\epsilon_i)) \right\|_\infty .$$

We have already seen that under our assumptions on the design (deterministic or random) for the linear model (2.1) and for the robust losses within the class $\mathcal{R}$, condition (2.3) is always satisfied. Note also that all influence functions ℓ' that we consider are bounded by some constant κ_1. In the deterministic design case the result is trivial. In the random design case, the variables $\boldsymbol{X}_i$ are i.i.d sub-Gaussian. Hence, the desired bound holds by using standard concentration results for i.i.d. sums of sub-Gaussian variables. Therefore the gradient bound (4.2) is established with high probability under fairly mild assumptions.

When the errors are drawn from a sub-Gaussian distribution with an appropriate bound on its variance, or when the errors have a contaminated distribution with an appropriate bound on the fraction of outliers one may quantify the parameters of the RSC condition by an additional parameter $T > 0$ which is treated as a fixed constant. Since for the class $\mathcal{R}$ of robust loss functions that we consider we also have $\ell''(u) > -K$ for all $u \in \mathbb{R}$ and some constant $K > 0$, all conditions of Loh's Proposition 2 are true and therefore the local RSC condition for the loss functions holds with probability at least $1 - c\exp(-c_2 \log p)$ for some constant $c > 0$ as soon as $n > C\, k^* \log p$. This allows to get oracle properties and asymptotic normality of our penalised robust M-estimators by relying upon Theorem 2 and Corollary 1 of Loh (2017). This ends our discussion on consistency properties of the penalised M-estimators we consider in this paper.

5 Algorithms for numerical optimisation

We first present some theory for the composite gradient descent algorithm, including rates of convergence for regularised problems. We then describe a two-step algorithm, which is guaranteed to converge to a stationary point within the local region where the RSC condition holds, even when the M-estimator is non-convex.

5.1 Composite gradient descent

In order to obtain stationary points of the objective function (2.2), we will use the composite gradient descent algorithm (see Nesterov (2007)). Denoting $\bar{\mathcal{L}}_n(\boldsymbol{\beta}) := \mathcal{L}_n(\boldsymbol{\beta}) - q_\lambda(\boldsymbol{\beta})$, we may rewrite the minimisation problem as

$$\hat{\boldsymbol{\beta}}_\lambda \in \underset{\|\boldsymbol{\beta}\|_1 \leq R}{\operatorname{argmin}} \left\{ \bar{\mathcal{L}}_n(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1 \right\}.$$

Then the composite gradient iterates are given by

$$\boldsymbol{\beta}^{t+1} \in \underset{\|\boldsymbol{\beta}\|_1 \leq R}{\operatorname{argmin}} \left\{ \frac{1}{2} \left\| \boldsymbol{\beta} - \left(\boldsymbol{\beta}^t - \frac{\nabla \bar{\mathcal{L}}_n(\boldsymbol{\beta}^t)}{\eta} \right) \right\|_2^2 + \frac{\lambda}{\eta} \|\boldsymbol{\beta}\|_1 \right\}, \tag{5.1}$$

where η is the step-size parameter. Defining the soft-thresholding operator $S_{\lambda/\eta}(\boldsymbol{\beta})$ componentwise according to

$$S^j_{\lambda/\eta} := \operatorname{sign}(\beta_j) \left(|\beta_j| - \frac{\lambda}{\eta} \right)_+ ,$$

a simple calculation shows that the iterates (5.1) take the form

$$\beta^{t+1} = S_{\lambda/\eta}\left(\boldsymbol{\beta}^t - \frac{\nabla\bar{\mathcal{L}}_n(\boldsymbol{\beta}^t)}{\eta}\right). \tag{5.2}$$

Note here, that when the scale σ is unknown, its estimation is only useful to setup the appropriate step size η. Theorem 3 of Loh (2017) guarantees that the composite gradient descent algorithm will converge at a linear rate to point near $\boldsymbol{\beta}^*$ as long as the initial point $\boldsymbol{\beta}^*$ is chosen close enough to $\boldsymbol{\beta}^*$ under mild conditions, that are satisfied when the loss function ℓ appearing in the definition of the M-estimator has a bounded second derivative and when q_λ is convex, as is the case for the SCAD and MCP penalties. The theorem may be applied to situations where q_λ is nonconvex, given an appropriate quadratic bound on the Taylor remainder of q_λ but we will not pursue this here.

The above guarantee that the composite gradient descent algorithm will converge quickly to a desirable stationary point if the initial point is chosen within a constant radius of the true regression vector. As in Fan et al. (2017) we can propose a two-step algorithm that may be applied to optimise high-dimensional robust M-estimators. Even when the regression function is nonconvex, the two-step procedure defined below will always converge to a stationary point that is statistically consistent for $\boldsymbol{\beta}^*$ under the scaling $n \gtrsim k^* \log p$.

(1) Run composite gradient descent using a convex regression function ℓ with convex ℓ_1-penalty, such that ℓ' is bounded (for example Huber's loss with l_1 penalty as in Fan et al. (2017)).
(2) Use the output of step (1) to initialise composite gradient descent on the desired high-dimensional M-estimator.

Remark 1 Note here that when using the estimates studied in Section 3 (case $p < n$) with robust losses but l_1 penalties, the optimisation algorithm is identical to the one above but with the function $q_\lambda(\boldsymbol{\beta}) = 0$ (since the ℓ_1 penalty is 0-amenable) so in this case $\bar{\mathcal{L}}_n(\boldsymbol{\beta}) = \mathcal{L}_n(\boldsymbol{\beta})$ and everything is separable since the ℓ_1 penalty is coordinate-separable. More precisely, we have

$$\mathcal{L}_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\rho_d((Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta})/\sigma) \text{ and } \nabla\mathcal{L}_n(\boldsymbol{\beta}^*) = \frac{1}{n\sigma}\sum_{i=1}^{n}[\psi_d(\epsilon_i/\sigma)]\cdot[\boldsymbol{X}_i],$$

and the iterates (5.1) take again the form

$$\beta^{t+1} = S_{\lambda/\eta}\left(\boldsymbol{\beta}^t - \frac{\nabla\bar{\mathcal{L}}_n(\boldsymbol{\beta}^t)}{\eta}\right).$$

The scale $\hat{\sigma}$ is only relevant in the iterative step size η of the algorithm and is estimated by the median absolute deviation of the residuals from the full model fitted by S-estimation. In the adaptive version of the algorithm, with a weighted version of the penalty coordinates, the design matrix X is transformed to a new design matrix $\tilde{X} = XD^{-1}$, with D the diagonal matrix with diagonal elements the weights $\hat{w}_j$ used in the adaptive penalty, the resulting optimisation problem for $\tilde{\boldsymbol{\beta}} := D\boldsymbol{\beta}$ is solved using the composite gradient descent algorithm, and the obtained estimate $\hat{\tilde{\boldsymbol{\beta}}}$ is back-transformed to $D^{-1}\hat{\tilde{\boldsymbol{\beta}}}$.

6 Simulation Results

Throughout this section we will assume without any loss of generality that our model does not contain an intercept. To meet the model assumptions, prior to any calculations, all the columns of $\mathbf{X}$ are centered and scaled using the median and the normalised median absolute deviation respectively. The response vector $\mathbf{y}$ is centered using the median. At the end, the final estimates are expressed in the original coordinates.

For all adaptive estimators that follow, the adaptive weights used were the reciprocal of an initial penalized elastic net S-Lasso estimator of Cohen Freue et al. (2018). The penalisation parameter for S-Lasso estimator, γ_n, is chosen via robust 5-fold cross-validation, as described in Cohen Freue et al. (2018). The robust scale estimate associated to the optimal S-Lasso estimator

is obtained by the median absolute deviation of the residuals from the full model fitted by S-estimation (see Rousseeuw and Yohai (1984)) and is denoted by s_n. Note that according to Theorem 3 and the remarks above Theorem 1 of Smucler and Yohai (2017), the S-Lasso has a high breakdown point and both the S-Lasso and s_n are $\sqrt{n}$-consistent estimates of $\boldsymbol{\beta}^*$ and σ, as long as the penalisation parameter γ_n of the S-Lasso estimator satisfies $\gamma_n = O(\sqrt{n})$.

To compare the performance with regards to prediction accuracy and variable selection properties of our class $\mathcal{R}$ of methods with existing procedures in the literature, we mainly used an ℓ_1 penalty. Our comparison analyses the following methods:

- The standard LASSO estimator. The penalisation parameter for this estimator was chosen using a k-fold CV criterion with $k = 10$. The estimator was calculated using the `cv.ncvreg()` function from the `ncvreg R` package Breheny (2018).
- The adaptive SSLASSO estimator of Javanmard and Montanari (2014). The associated Lasso regularization parameter is estimated by sqrt-lasso (see e.g. Antoniadis (2010) or Sun and Zhang (2010)). The penalisation parameter for de-biasing the lasso for this estimator is chosen iteratively on a given grid.
- The adaptive Lasso (ADLASSO) estimator using the `adalasso()` function from the `parcor R` package Kraemer and Schaefer (2014). The penalisation parameter for this estimator is chosen using a k-fold CV criterion with $k = 10$.
- The Sparse-LTS (SLTS) of Alfons et al. (2013). The penalisation parameter for this estimator is chosen using a BIC-type criterion as advocated by the authors. The estimator was calculated using the `sparseLTS()` function from the `robustHD R` package Alfons (2016).
- The adaptive HUBER Lasso estimator. The estimator was calculated using a function inspired by the `CVXR R` package Fu et al. (2019).
- The adaptive TUKEY estimator calculated with Tukey's bisquare function (our own R-code).
- The adaptive MCP estimator (our own R-code).
- The LADLASSO estimator of Wang et al. (2007) using the function `ladlasso()` of the `perryExamples R` package Alfons (2014). Note that for each simulation the estimator depends on a random seed which has to be set at different values for each run. (There is a confusion in the package in using global and local environment variables).
- The robust least angle regression estimator, called RLARS proposed in Khan et al. (2007) using their `Rlars()` function.
- The RA-LASSO estimator of Fan et al. (2017) with tuning parameters for the Huber loss function chosen by cross-validation. We have translated in R their Matlab code and used an equivalent `RALasso()` function.
- The adaptive WELSH estimator (our own R-code).
- The adaptive CAUCHY estimator (our own R-code).

R-codes implementing the above methods as well as the simulations and the scripts that produce the plots and the tables in the paper are available upon request from the authors.

6.1 Simulations for the case p fixed ($p < n$)

We set $p = 15$, the $\mathbf{X}_i$ are generated from a p-dimensional Gaussian distribution with mean 0 and covariance Σ with $\Sigma_{i,j} = 0.5^{|i-j|}$, creating correlated predictor variables. Prior to any calculations, all columns of X are centered and scaled. The number of nonzero coefficients k^* is set equal to 5, with $\boldsymbol{\beta}^\star = (1.5, 0.5, 0, 1, 0, 0, 1.5, 0, 0, 0, 1, 0, 0, 0, 0)^T$.

To evaluate and compare the estimators we generate two independent samples of size $n = 100$ of the model

$$Y_i = \boldsymbol{X}_i^T \boldsymbol{\beta}^* + \epsilon_i, \quad 1 \leq i \leq n$$

with i.i.d. random errors ϵ_i. 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. The error terms follow a mean zero normal distribution with standard deviation $\sigma = 0.5$.

We use the mean squared prediction error (MSPE) to evaluate the prediction accuracy of the estimates with optimal tuning values of the regularisation parameters. We have also evaluated 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), that is, the fraction of coefficient erroneously not set to zero (the fraction is evaluated with respect to p). Both FPR and FNR should be as small as possible for sparse estimators. False negatives in general have a stronger effect on the MSPE than false positives. Indeed a false negative suggests that important information is not used for prediction, whereas a false positive merely adds a bit of variance.

We apply contamination schemes taken from Alfons et al. (2013). Several simulation scenarios have been used to get a good idea of the various methods, ranging from "good" data (normal or symmetric non-normal errors and no outliers) to data with outliers in or both the covariates and the response. To be more precise, we consider the following 5 scenarios:

Scenario 0 Data are generated without any outliers with a design and a response generated as described above.

Scenario 1 Data are generated without any outliers. Again the design matrix is constructed as above but the i.i.d errors in the responses in the linear model are drawn from a mixture of two centered Gaussians

$$(1-\pi)\mathcal{N}(0,\sigma)+\pi\mathcal{N}(0,9\sigma)$$

where π denotes the mixture proportion that we set to 10%.

Scenario 2 Data are generated with outliers only in the response. Again the design matrix is constructed as above but the responses in the linear model are drawn from a Gaussian mixture,

$$(1-\pi)\mathcal{N}(0,\sigma)+\pi\mathcal{N}(20,\sigma)$$

where π denotes the mixture proportion that we set to 10%.

Scenario 3 To evaluate the robustness of the estimators for the case of high-leverage outliers, we introduce contaminations in the covariates. To generate the data we use the same scenario as in Scenario 2 for generating the data with errors from a mixture distribution, but we further contaminate the training sample as follows. A 10% proportion of the observations have their design contaminated with high-leverage values, by shifting the corresponding predictor variables by an amount drawn independently from a $\mathcal{N}(50,1)$ distribution. Note that we always only contaminate the training sample and not the testing sample.

Scenario 4 To evaluate the robustness of the estimators for the case of high-leverage outliers and outliers in the response, we introduce contaminations in both the response and the covariates. To generate the data we use a similar scenario to scenario 2 with errors from the mixture distribution $(1-\pi)\mathcal{N}(0,\sigma)+\pi\mathcal{N}(10,\sigma)$, where π denotes the mixture proportion that we set to $0\%, 5\%, 10\%, 15\%, 20\%, 25\%$, but we further contaminate the training sample by adding to the corresponding predictor variables a proportion π of high leverage values from an independent $\mathcal{N}(50,1)$ distribution.

Simulations are repeated $N = 25$ times to keep computation times reasonably low.

6.2 Results

We now present the results of our simulation study for the case p fixed ($p < n$). All numerical summaries are rounded to two decimal places. Table 1 displays the results for Scenarios 0 through 3. We compute the average, the median and the standard error of the MSPEs and the FNR and FPR statistics.

In Scenario 0, which is the scenario without contamination with normal errors, from the perspective of variable selection, all methods present a false negative ratio equal to 0, and seem to identify the significant variables correctly. LASSO, SSLASSO, LADLASSO, SLTS and RALASSO, while selecting significant variables correctly, also select more noise variables (present a larger FPR), compared to the other remaining methods that generally select fewer noise variables (with a lower FPR). From the prediction point of view, as can be seen in Table 1 or in the left-upper panel of Figure 2, we see that tall procedures perform well in terms of MSPE in the uncontaminated case. SSLASSO, LADLASSO, SLTS, and to a lesser extent LASSO and RALASSO, have higher average MSPEs than the other methods, probably due to their larger FPRs. Note that LASSO, compared to ADLASSO, shows, as expected, a higher average MSPE due to the lack of adaptive weights.

In Scenario 1, which is the other scenario without contamination but with symmetric non-normal errors, from the perspective of variable selection, all methods present a false negative ratio equal to 0 except ADLASSO, and seem to identify the significant variables correctly. LADLASSO, SSLASSO, SLTS LASSO and RALASSO, while selecting significant variables correctly, also select more noise variables (present a larger FPR), compared to the other remaining methods that generally select fewer noise variables (with a lower FPR). From the prediction point of view, as can be seen in Table 1 or in the right-upper panel of Figure 2, we see that all procedures perform well in terms of MSPE in the uncontaminated case. SSLASSO, LASSO, ADLASSO, and to a lesser extent LADLASSO, SLTS and RALASSO, have higher average MSPEs than the other methods, probably due to their larger FPRs, or in the case of ADLASSO to the presence of a no zero FNR.

For Scenario 2 (simulation design without leverage – presence of y-outliers only) in Table 1, it is clear that LASSO, SSLASSO, ADLASSO and to a lesser extent LADLASSO are affected by the presence of outliers and exhibit poor performance in terms of both variable selection and prediction accuracy. In contrast, the remainder of the compared methods are confirmed to be robust to y-outliers and, among them, BIWEIGHT, NNG, CAUCHY, WELSH, MCP and HUBER achieve lower MSPEs. This is also shown in the left-lower panel of Figure 2.

Finally, the right-hand side of Table 1 reports simulation results for Scenario 3 (the Scenario with both x- and y-outliers). When leverage points are introduced in addition to y-outliers the performance of LASSO, SSLASSO, ADLASSO, HUBER, LADLASSO and RALASSO suffers greatly from the outliers both in terms of variable selection and in terms of prediction accuracy. RLARS, followed by BIWEIGHT, SLTS and MCP seem to perform better than all the other methods, maintaining desirable MSPEs and are shown to be resistant to both $x-$outliers and y-outliers, as expected. As evidenced in the right-lower panel of Figure 2, these four methods consistently dominate the others by achieving much lower MSPEs.

Table 1 Results for all the simulation scenarios, with normal $\mathcal{N}(0, 0.5)$, $(0.9, 0.1)$ mixture of $\mathcal{N}(0, 0.5)$ and $\mathcal{N}(0, 4.5)$ and $(0.9, 0.1)$ mixture of $\mathcal{N}(0, 0.5)$ and $\mathcal{N}(20, 0.5)$ distributed errors. MSPE, FNR and FPR, averaged over 25 replications are reported for each estimator.

	MSPE.0	FNR.0	FPR.0	MSPE.1	FNR.1	FPR.1	MSPE.2	FNR.2	FPR.2	MSPE.3	FNR.3	FPR.3
LASSO	0.280	0	0.253	0.464	0	0.299	7.600	0.088	0.187	12.154	0.264	0.107
SSLASSO	0.296	0	0.667	0.604	0	0.667	5.473	0	0.667	8.258	0	0.667
ADLASSO	0.271	0	0.021	0.422	0.003	0.069	4.507	0.173	0.093	8.065	0.280	0.059
HUBER	0.270	0	0.011	0.275	0	0.037	0.270	0	0.024	6.679	0.048	0.187
BIWEIGHT	0.271	0	0.008	0.270	0	0.019	0.268	0	0.005	0.523	0.011	0.008
MCP	0.273	0	0.008	0.269	0	0.008	0.270	0	0.005	0.523	0.011	0.011
LADLASSO	0.293	0	0.667	0.315	0	0.667	0.443	0	0.667	5.638	0	0.667
RLARS	0.279	0	0.072	0.274	0	0.051	0.276	0	0.091	0.296	0	0.504
SLTS	0.311	0	0.347	0.299	0	0.379	0.310	0	0.395	0.480	0.003	0.264
RALASSO	0.285	0	0.275	0.296	0	0.176	0.289	0	0.221	5.181	0.099	0.507
WELSH	0.270	0	0	0.267	0	0	0.270	0	0	0.595	0.013	0
NNG	0.273	0	0.021	0.275	0	0.045	0.268	0	0.008	0.916	0.027	0.013
CAUCHY	0.270	0	0	0.268	0	0	0.269	0	0	0.596	0.013	0

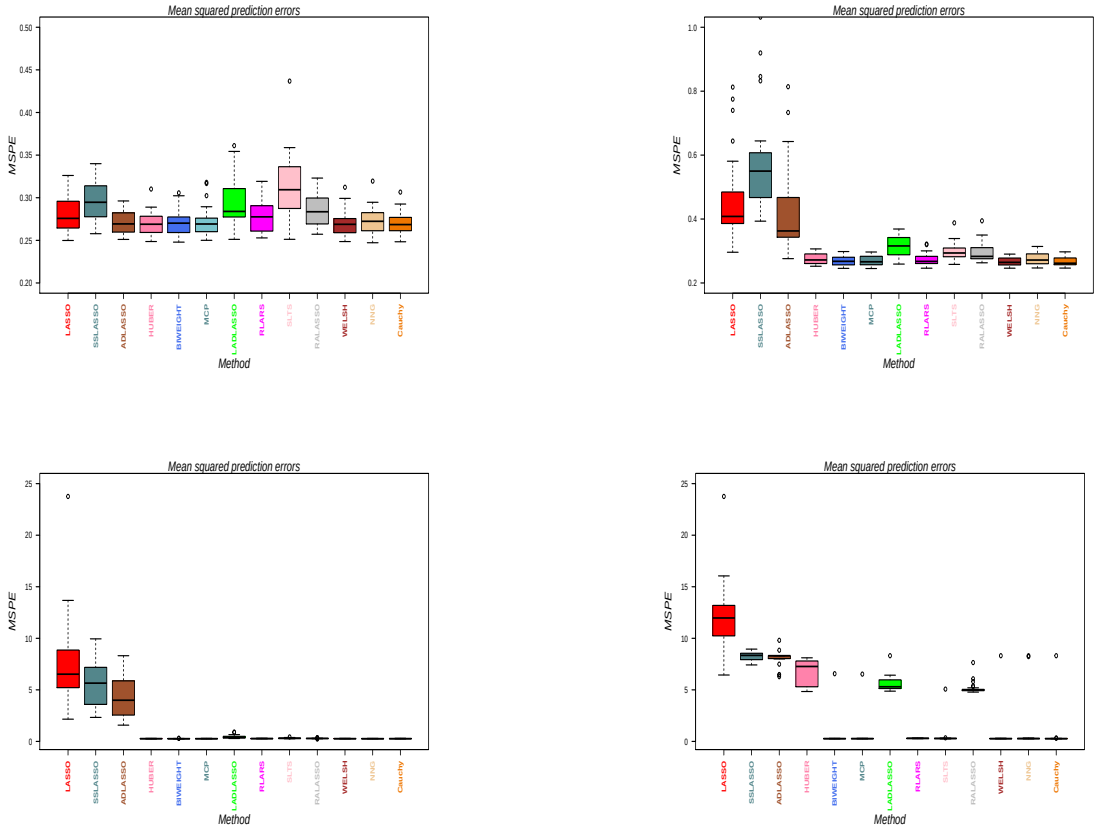

Fig. 2 Mean squared prediction errors for $p = 15$ and $n = 100$. (Left-upper panel) Scenario 0, (right-upper panel) Scenario 1, (left-lower panel) Scenario 2, (right-lower panel) Scenario 3.

The above analysis is confirmed by the simulation runs from Scenario 4 which aim to examine the behavior of the estimation procedures at other contamination levels. Figure 3 displays the averaged MSPEs for all methods at various contamination levels for both the covariates and the response, when $p = 15$ and $n = 100$.

One observes that LASSO is not resistant to any percentage of outliers; its average MSPE, and to a lesser extent the average MSPE of LADLASSO, ADLASSO, SSLASSO, RALASSO and HUBER increase gradually immediately after 5% contamination, while for the other methods the increase is noted after 15% contamination. From the left panel in Figure 3, RLARS followed by SLTS, and then CAUCHY, WELSH and MCP perform well over all methods maintaining a satisfactory MSPE value even when the contamination level for both the covariates and the response reaches 25%.

6.3 Simulations for the high-dimensional with $k \log p = o(n)$

To save space and for comparing our results with those of the low dimensional case we report here the results when using again an ℓ_1 penalty. Note however that according to our Remark 1, the changes to make when using any of the other penalties considered in this paper are minimal. To satisfy the requirements of Theorem 2 we adopt a Monte-Carlo simulation design with $p = 128$, sparsity level $k = 5$ and a sample size n such that $n/(k \log(p)) \simeq 8$, i.e. $n = 200$.

All simulations were generated according to the same four scenarios used for the low dimensional case. Simulations are repeated $N = 25$ times to keep the computational time reasonably low. While n is of the same order as p, but still larger than p we were able to use the same methods

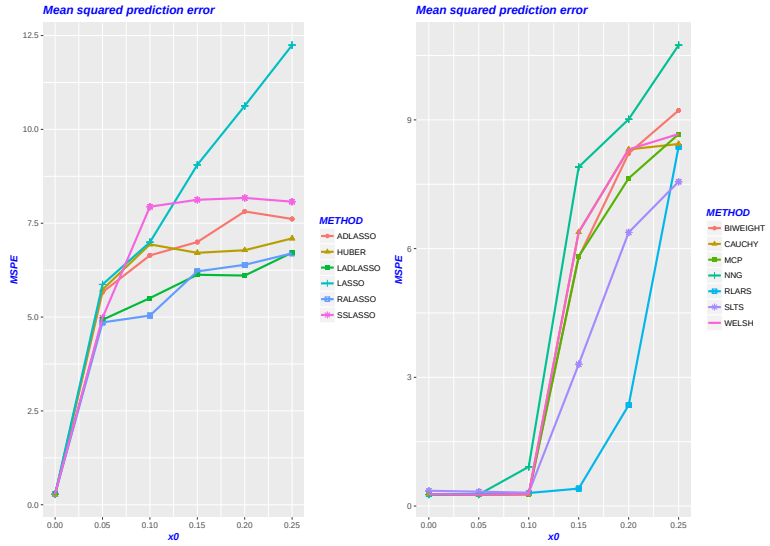

Fig. 3 MSPEs as a function of contamination sizes for each of the estimators for the fourth Scenario, with $p = 15$, $n = 100$. MSPEs are averaged over 25 replications.

in class $\mathcal{R}$ with an ℓ_1 penalty.

Table 2 contains the simulation results for the data generated under Scenarios 0, 1, 2, and 3, with $p = 128$, $n = 200$, while the corresponding MPSEs are displayed in Figure 4.

In Scenario 0, SSLASSO, SLTS and, to a lesser extent, LADLASSO display the highest average MSPEs because they suffer from efficiency problems when the data contain no outliers. It is also clear that ADLASSO, MCP, CAUCHY, WELSH, HUBER, BIWEIGHT and NNG achieve a good performance in terms of variable selection and prediction accuracy, while the other methods present a slight overfitting problem (high FPR).

In Scenario 1, SSLASSO, and, to a lesser extent, LASSO, SLTS, ADLASSO and LADLASSO display the highest average MSPEs because they suffer from efficiency problems when the data contain no outliers. Moreover, except ADLASSO, they show a slight overfitting problem (high FPR). ADLASSO is the only method with a FNR different from 0. It is also clear that MCP, CAUCHY, WELSH, HUBER, BIWEIGHT and NNG achieve a good performance in terms of variable selection and prediction accuracy.

With responses generated via a mixture of two Gaussian distributions in Scenario 2, the prediction power of SSLASSO, LASSO and ADLASSO break down. Similar to the case of $p = 15$, in this high dimensional case BIWEIGHT, CAUCHY, WELSH, MCP, NNG, HUBER and RALASSO still maintain substantially lower MSPEs. Note however that both LADLASSO and SLTS display much higher FPR's.

The far right-hand side of Table 2 reports simulation results for Scenario 3 in this high-dimensional case. Similarly to the low dimensional case, when leverage points are introduced in addition to y-outliers, LASSO, SSLASSO, ADLASSO, HUBER, LADLASSO and RALASSO suffer from the presence of x-outliers in terms of prediction accuracy and to a lesser extent in terms of variable selection. BIWEIGHT, MCP, WELSH and CAUCHY perform better than other methods, maintaining desirable MSPEs and reasonable variable selection power. This is also evidenced in the right-lower panel of Figure 4.

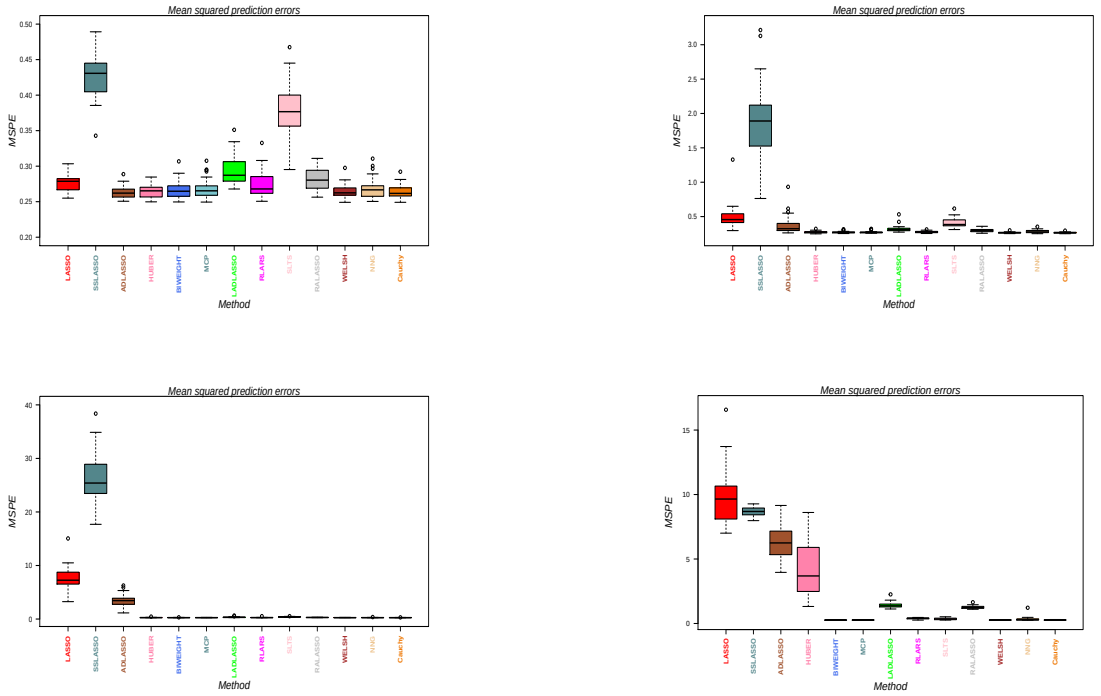

Fig. 4 Mean squared prediction errors for $p = 128$ and $n = 200$. (Left-upper panel) Scenario 0, (right-upper panel) Scenario 1, (left-lower panel) Scenario 2, (right-lower panel) Scenario 3.

Table 2 Results for all the simulation scenarios, with normal $\mathcal{N}(0, 0.5)$, $(0.9, 0.1)$ mixture of $\mathcal{N}(0, 0.5)$ and $\mathcal{N}(0, 4.5)$ and $(0.9, 0.1)$ mixture of $\mathcal{N}(0, 0.5)$ and $\mathcal{N}(20, 0.5)$ distributed errors. MSPE, FNR and FPR, averaged over 25 replications are reported for each estimator.

	MSPE.0	FNR.0	FPR.0	MSPE.1	FNR.1	FPR.1	MSPE.2	FNR.2	FPR.2	MSPE.3	FNR.3	FPR.3
LASSO	0.276	0	0.092	0.500	0	0.094	7.825	0.009	0.102	9.850	0.018	0.063
SSLASSO	0.426	0	0.961	1.921	0	0.961	26.295	0	0.961	8.664	0	0.961
ADLASSO	0.263	0	0.002	0.384	0.001	0.014	3.452	0.017	0.012	6.436	0.024	0.036
HUBER	0.266	0	0.008	0.274	0	0.009	0.280	0	0.027	4.453	0.004	0.063
BIWEIGHT	0.267	0	0.009	0.273	0	0.008	0.271	0	0.007	0.267	0	0.001
MCP	0.270	0	0.010	0.274	0	0.007	0.270	0	0.003	0.268	0	0.002
LADLASSO	0.294	0	0.961	0.324	0	0.961	0.348	0	0.961	1.415	0	0.961
RLARS	0.275	0	0.022	0.276	0	0.010	0.283	0	0.014	0.387	0.006	0.093
SLTS	0.378	0	0.452	0.412	0	0.522	0.400	0	0.500	0.355	0	0.300
RALASSO	0.281	0	0.023	0.299	0	0.045	0.294	0	0.035	1.262	0	0.380
WELSH	0.265	0	0.0003	0.267	0	0	0.270	0	0	0.268	0	0
NNG	0.270	0	0.017	0.282	0	0.018	0.272	0	0.015	0.348	0	0.034
CAUCHY	0.264	0	0	0.266	0	0	0.269	0	0	0.270	0	0

In Figure 5 we show the resulting MSPEs for all methods at various contamination levels for both the covariates and the response on data generated according to the high dimensional version of Scenario 4. Similar conclusions to the low-dimensional case hold, with the exception of SLTS which exhibits a high increase in MSPE values when the contamination level increases.

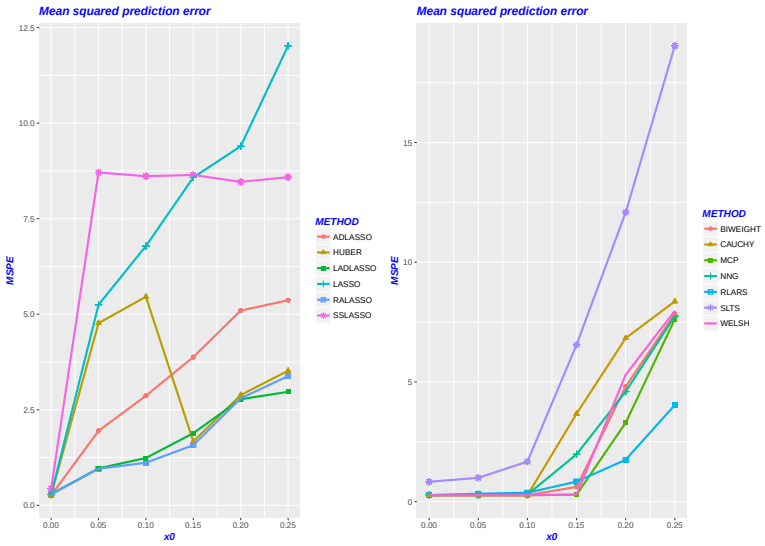

Fig. 5 MSPEs as a function of contamination sizes for each of the estimators for the fourth Scenario, with $p = 128$, $n = 200$. MSPEs are averaged over 25 replications.

6.4 Simulations for the high-dimensional case with $p \gg n$

To be complete, we are adding this set of simulations to examine the behavior of the various robust estimators in a high dimensional case that we did not analyse theoretically in the previous sections. To run the simulations we have used the same scenarios as before, but this time with $n = 128$ and $p = 200$. Again, simulation runs are repeated $N = 25$ times to keep the computational time reasonably low.

Table 3 and Figure 6 display the simulation results for the data generated under Scenarios 0, 1, 2, and 3, using $n = 128$ and $p = 200$. As can be seen from these, the conclusions in terms of prediction accuracy of all the used procedures are comparable with those obtained in the high-dimensional case when $n > p$. Sometimes, some of the methods give few extreme MSPE values and this is due to a bad, hopefully rare, choice of their regularization parameters. However, the selection results differ significantly.

Table 3 Results for all the simulation scenarios, with normal $\mathcal{N}(0, 0.5)$, $(0.9, 0.1)$ mixture of $\mathcal{N}(0, 0.5)$ and $\mathcal{N}(0, 4.5)$ and $(0.9, 0.1)$ mixture of $\mathcal{N}(0, 0.5)$ and $\mathcal{N}(20, 0.5)$ distributed errors. MSPE, FNR and FPR, averaged over 25 replications are reported for each estimator.

	MSPE.0	FNR.0	FPR.0	MSPE.1	FNR.1	FPR.1	MSPE.2	FNR.2	FPR.2	MSPE.3	FNR.3	FPR.3
LASSO	0.331	0	0.037	0.662	0.0002	0.080	9.369	0.010	0.051	11.609	0.015	0.030
SSLASSO	0.770	0	0.975	4.847	0	0.975	82.119	0	0.975	9.753	0	0.975
ADLASSO	0.274	0	0	0.433	0.001	0.003	5.456	0.016	0.005	7.081	0.017	0.017
HUBER	0.293	0	0.043	0.359	0	0.085	0.310	0	0.057	3.446	0.0002	0.051
BIWEIGHT	0.290	0.0002	0.012	0.351	0	0.021	0.307	0	0.022	0.273	0	0.002
MCP	0.323	0	0.008	0.346	0	0.014	0.347	0	0.006	0.274	0	0.002
LADLASSO	0.604	0	0.975	0.629	0	0.975	1.154	0	0.975	1.692	0	0.975
RLARS	0.293	0	0.025	0.268	0	0.005	0.270	0	0.006	0.731	0.005	0.075
SLTS	0.513	0	0.334	0.531	0	0.352	0.501	0	0.352	0.493	0	0.373
RALASSO	0.374	0	0.054	0.374	0	0.049	0.368	0	0.051	1.142	0	0.237
WELSH	0.622	0.001	0.0002	1.315	0.003	0.0002	0.673	0.001	0.0002	0.274	0	0
NNG	0.295	0	0.053	0.334	0	0.093	0.305	0	0.065	0.552	0	0.032
CAUCHY	0.285	0	0.0002	0.280	0	0.0002	0.273	0	0.001	0.290	0	0

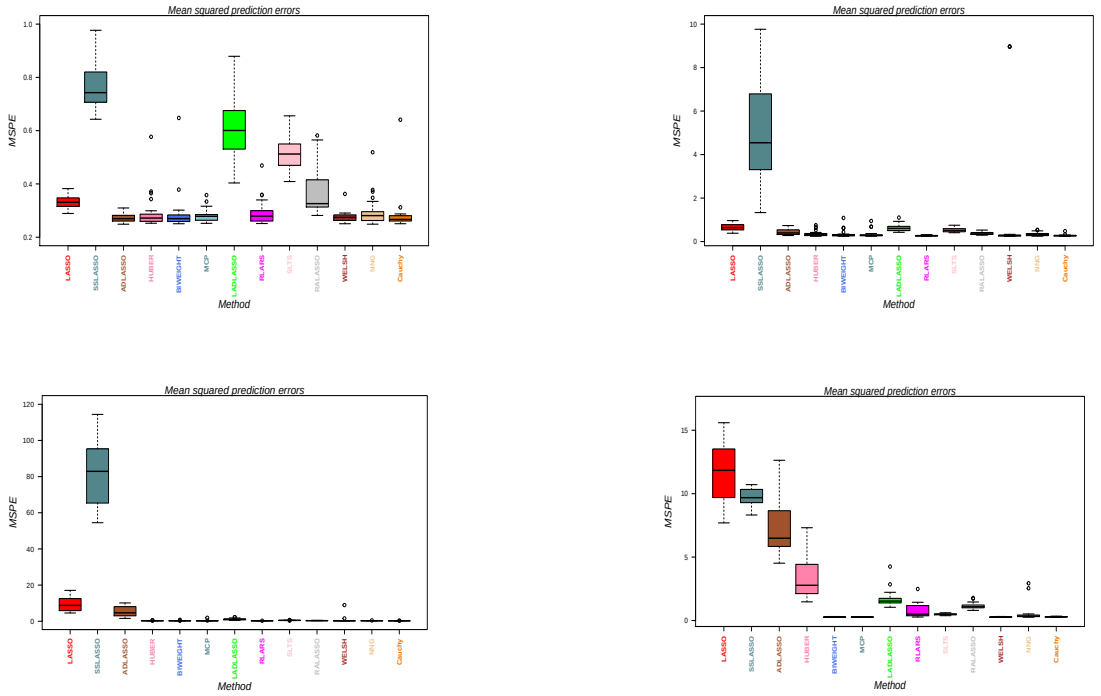

Fig. 6 Mean squared prediction errors for $p = 200$ and $n = 128$. (Left-upper panel) Scenario 0, (right-upper panel) Scenario 1, (left-lower panel) Scenario 2, (right-lower panel) Scenario 3.

Figure 7 displays the resulting MSPEs for all methods at various contamination levels for both the covariates and the response on data generated according to the high dimensional version of Scenario 4. Similar conclusions to the high dimensional case with $n < p$ can be reached, again with the exception of SLTS which continues to exhibit a high increase in MSPE values when the contamination level increases.

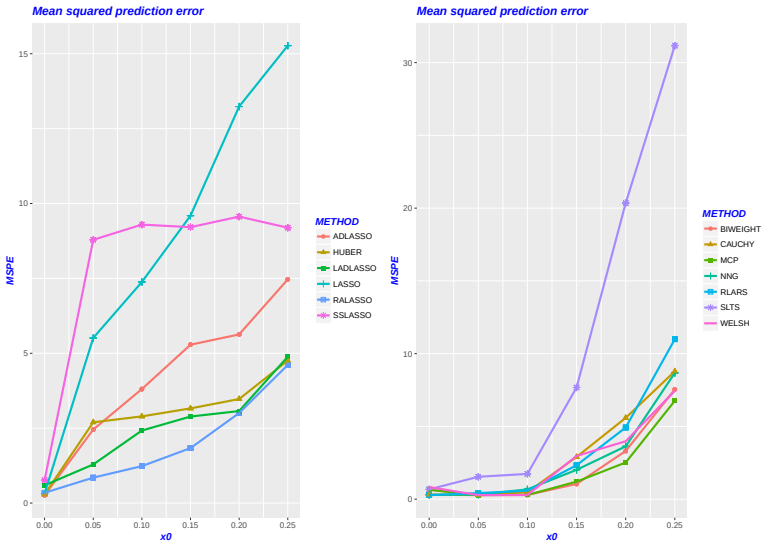

Fig. 7 MSPEs as a function of contamination sizes for each of the estimators for the fourth Scenario, with $p = 200$, $n = 128$. MSPEs are averaged over 25 replications.

6.5 Simulations summary

Estimators from the procedures belonging to the class $\mathcal{R}$ such as BIWEIGHT, MCP, WELSH, NNG and CAUCHY show the best overall performance in this simulation study. It is also confirmed that the LASSO is not robust to outliers and that while the LADLASSO still sustains vertical outliers, it is not robust against bad leverage points.

In the low dimensional case $p < n$, whenever a calibration of the noise level σ is necessary, it can be estimated by the median absolute deviation of the residuals from the full model fitted by S-estimation (see Rousseeuw and Yohai (1984)) or any other S-scale estimator. When p is comparable or larger than n and the signal is sufficiently sparse one can estimate the noise level in the spirit of the squared-root LASSO, but simulations have shown that, while the resulting estimates of σ are theoretically consistent for the "oracle" estimator, they are not efficient with data contaminated with outliers. We thus prefer using an estimate of σ by a robust S-estimator as in RALASSO (see Fan et al. (2017)). As illustrated by the simulations, a good estimation of σ is essential for choosing the regularization parameter properly in the penalized versions of the various procedures, especially in the high-dimensional cases. However, we feel that the problem of defining an optimal estimator of σ in the ultra-high dimensional case still requires a separate investigation. To conclude, let us say that the methods we used for estimating the scale in our simulations are still useful for practical data analysis.

7 Real data examples

In this section we use two real data sets to further illustrate the performance of the robust regression procedures analyzed in the paper. Remember that in practice, we often do not know the number of covariates that are needed in the model.

7.1 Boston housing data

We first apply the robust procedures to the Boston housing price dataset available at `https://archive.ics.uci.edu/ml/datasets/Housing`, which is commonly used as an example for regressions. The original Boston housing data was gathered by Harrison and Rubinfeld (1978) and comprises the housing price and other attributes of 506 suburb areas of Boston from the 1970 census. This dataset is particularly of interest for robust regression analysis as it contains outliers and skewed variables. In this section we have used a corrected version of the data by Pace and Gilley (1997) with additional spatial information. The data, named BostonHousing2, are available at the `mlbench` repository at `https://cran.r-project.org/web/packages/mlbench/`. For each census track, there is an observation of the corrected median value of owner-occupied homes per 1000 USD and 15 independent covariates which are briefly explained in Table 4.
We have used the logarithm of $CMEDV$ as the response variable and used the following regression model, which contains 18 candidate predictors, using the factors described in Table 4:

$$\begin{aligned}\log(CMEDV) = {} & \beta_0 + \beta_1 LON + \beta_2 LAT + \beta_3 CRIM + \beta_4 ZN + \beta_5 INDUS \\ & + \beta_6 CHAS + \beta_7 NOX^2 + \beta_8 RM^2 + \beta_9 AGE + \beta_{10} \log(DIS) \\ & + \beta_{11} \log(RAD) + \beta_{12} TAX + \beta_{13} PTRATIO + \beta_{14} B \\ & + \beta_{15} \log(LSTAT) + \beta_{16} LAT \cdot LON + \beta_{17} LAT^2 + \beta_{18} LON^2.\end{aligned}$$

We standardize all variables in Table 4, except DIS, RAD, $LSTAT$. For these three variables, we first take log and then standardize them.

We applied all the methods described in Section 6 to the data. The predictive behaviour of the estimators was assessed through five-fold CV. Prediction and variable selection results are summarized in Table 5. As we can observe, ADLASSO, LADLASSO, NNG and CAUCHY select similar subsets of variables with an MSPE estimated error of the same order while LASSO, SSLASSO, HUBER and SLTS select all or too many predictors and therefore do not perform an efficient or even any variable selection. A further examination of the QQ-plots of the residuals

Table 4 Boston housing data variables and descriptions.

Name	Description
$CMEDV$	Corrected median value of owner-occupied homes per 1000 US $
$CRIM$	per capita crime rate by town
ZN	proportion of residential land zoned for lots over 25000 sq.ft
$INDUS$	proportion of non-retail business acres per town
$CHAS$	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
NOX	nitric oxides concentration (parts per 10 million)
RM	average number of rooms per dwelling
AGE	proportion of owner-occupied units built prior to 1940
DIS	weighted distances to five Boston employment centres
RAD	index of accessibility to radial highways
TAX	full-value property-tax rate per 10000 US $
$PTRATIO$	pupil-teacher ratio by town
B	$1000(B-0.63)^2$ where B is the proportion of blacks by town
$LSTAT$	percentage of lower status of the population
LON	longitude of census tract
LAT	latitude of census tract

for each of the methods (not displayed here to save some space), confirm the presence of many outliers in the dataset and supports the use of ADLASSO, LADLASSO, NNG and CAUCHY. This findings are largely consistent with variables commonly used in the literature.

Table 5 Results on Boston Housing Data.

Method	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	MSPE
LASSO	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	0.049
SSLASSO	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	0.049
ADLASSO	x		x			x	x	x		x	x	x	x	x	x				0.050
HUBER	x	x	x		x	x	x	x	x	x	x	x	x	x	x	x	x	x	0.056
BIWEIGHT	x		x			x	x	x	x	x	x	x	x	x	x		x		0.077
MCP			x			x	x	x	x	x	x	x	x	x	x		x		0.081
LADLASSO	x		x			x	x	x	x		x	x	x	x	x		x	x	0.055
RLARS	x	x					x	x		x	x	x	x	x	x		x		0.066
SLTS	x	x	x		x	x	x	x	x	x	x	x	x	x	x	x	x	x	0.082
RALASSO	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x		x	x	0.053
WELSH			x			x	x	x	x	x	x	x	x	x	x		x		0.080
NNG	x	x	x			x	x	x	x	x	x	x	x	x	x		x		0.054
CAUCHY			x			x	x	x	x	x	x	x	x	x	x		x		0.059

7.2 Glass vessels data

We also analyzed a data set collected from electron-probe X-ray microanalysis (EPXMA) of 180 archaeological glass vessels of the 16th-17th century, see Janssens et al. (1998). This data set has been analyzed in several other papers on high-dimensional robust linear regression with leverage points (see e.g. Maronna (2011) and Smucler and Yohai (2017)). In this data set each of the $n = 180$ glass vessels is represented by a X-ray spectrum on 1920 frequencies (KeV) and for each vessel the contents of thirteen chemical compounds are registered. We fit a linear model where the response variable is the content of the 13th chemical compound (PbO) and the predictors are the $p = 486$ X-ray detected intensities on each glass vessel obtained by retaining only the intensities x_{ij}, $i = 1, \ldots, 180$, with j between 15 and 500, since all other intensities among the 1920 are almost null.

As discussed in Maronna (2011), the dataset contains clear outliers. We applied all the methods described in Section 6 to the data. Following Smucler and Yohai (2017)) for tuning parameter selection, we chose the parameter λ in our algorithm via five-fold cross-validation using an estimate of the scale of the residuals. Note that our theorems are stated with λ equal to $\sqrt{\frac{\log p}{n}}$ times universal constants, but in practice, choosing λ in a data-driven manner leads to better predictive performance. The predictive accuracy of the estimators was assessed using five-fold cross-validation. Results on the number of selected variables and on the MSPE are summarized in Table 6.

Table 6 MSPE on Glass vessels data.

Method	Number of selected variables	MSPE
LASSO	17	0.020
SSLASSO	41	0.066
ADLASSO	1	0.778
HUBER	2	0.832
BIWEIGHT	2	0.906
MCP	2	0.909
LADLASSO	79	1.377
RLARS	11	0.536
SLTS	79	0.725
RALASSO	25	2.042
WELSH	2	0.909
NNG	2	0.891
CAUCHY	2	0.905

The LASSO selects 17 variables, the SSLASSO selects 41 and LADLASSO and SLTS select 79, RLARS selects 11 and RALASSO selects 25. All the other penalized robust estimators produce models that are much sparser. Concerning the prediction accuracy LASSO, SSLASSO, RLARS, SLTS and ADLASSO show the best behavior (ADLASSO yields sparse vector with only one nonzero component corresponding to frequency number 230), HUBER, NNG, CAUCHY, WELSH and MCP follow in that order. However, the latter, producing more sparse models with reasonable prediction accuracy, are much easier to interpret. Both our theoretical findings and the extensive numerical applications on synthetic and real datasets confirm the good behavior in terms of predictive accuracy and variable selection of our procedures

8 Conclusion

This paper introduces some novel regularized M-estimation approaches for robust estimation in high dimensional linear regression model obtained by studying penalised versions of robust regression M-estimators with losses that only possess convex curvature over local regions, thus leading to regularised M-estimators with highly nonconvex loss functions. Both the theoretical properties of the resulting robust estimation procedures, under fixed and high-dimensional regression, as well as their numerical behavior, in terms of prediction accuracy and variable selection, are confirmed by the extensive simulation study provided in the paper on synthetic data drawn from sparse high-dimensional linear models and contaminated by outliers in both the response and the covariates.

Computationally, all our estimation procedures are shown to rely on a proximal iterative algorithm (APG), of a coordinate descent (CD) type, that appears to be surprisingly fast and efficient in solving the corresponding regularization problems and which is directly applicable to both the fixed and high-dimensional regression settings with little modifications.

Finally, an interesting direction, that was only addressed numerically but not theoretically in the paper is how to obtain good asymptotic results of our estimates when n, the sample size, is much greater that p, the number of predictors, involved in the sparse high-dimensional linear regression case. Another interesting question is what procedures one might use to identify the

outliers in such contaminated data. When the task is outlier detection in the mean shift linear regression problem with a sparse mean shift parameter, penalisation methods such as those developed in She and Owen (2011) or Kong et al. (2018) may be extended to high-dimensional settings. However when the proportion of the outliers is large or there are high leverage outliers, these procedures break down and cannot identify the outliers correctly. Addressing such problems is an interesting topic that the authors wish to address in future research.

Acknowledgments

The authors thank the Editor and a referee for their constructive comments and helpful suggestions, which improved the paper. They also would like to thank E. Smucler for sharing the archaeological dataset used in the examples. Part of this work was completed while A. Antoniadis and I. Gijbels were visiting the Istituto per le Applicazioni del Calcolo "M. Picone", National Research Council, Naples, Italy. I. Gijbels gratefully acknowledges financial support from the GOA/12/014 project of the Research Fund KU Leuven, Belgium.

Appendix 1: Definitions of ρ, ψ and thresholding δ functions (used in this paper)

Preamble

Unless otherwise stated, most of the following definitions of functions are standard, bur few definitions differ sometimes slightly, by the different way of *standardising* them. To avoid confusion, we first define ψ- and ρ-functions.

Definition 1 A *ψ-function* is a piecewise continuous function $\psi : \mathbb{R} \to \mathbb{R}$ such that

(1) ψ is odd, i.e., $\psi(-x) = -\psi(x)\ \forall x$,
(2) $\psi(x) \geq 0$ for $x \geq 0$, and $\psi(x) > 0$ for $0 < x < x_r := \sup\{\tilde{x} : \psi(\tilde{x}) > 0\}$ $(x_r > 0$, possibly $x_r = \infty)$.
(3) Its slope is 1 at 0, i.e., $\psi'(0) = 1$.

Note that (3) is not strictly required mathematically, but we use it for standardisation in those cases where ψ is continuous at 0. Then, it also follows (from (1)) that $\psi(0) = 0$, and we require $\psi(0) = 0$ also for the case where ψ is discontinuous in 0, as it is, e.g., for the M-estimator defining the median.

Definition 2 A *ρ-function* can be represented by the following integral of a ψ-function,

$$\rho(x) = \int_0^x \psi(u)du \tag{8.1}$$

which entails that $\rho(0) = 0$ and ρ is an even function.

A ψ-function is called *redescending* if $\psi(x) = 0$ for all $x \geq x_r$ for $x_r < \infty$, and x_r is often called *rejection point*. Corresponding to a redescending ψ-function, one may associate a loss function $\tilde{\rho}$, a version of ρ standardised such as to attain maximum value one. Formally,

$$\tilde{\rho}(x) = \rho(x)/\rho(\infty). \tag{8.2}$$

Note that $\rho(\infty) = \rho(x_r) \equiv \rho(x),\ \forall\, |x| \geq x_r$. $\tilde{\rho}$ is a ρ-function as defined in Maronna (2011) and has been called χ function in other contexts. For example, in package `robustbase` (see Maechler and al (2017)), `Mchi(x, *)` computes $\tilde{\rho}(x)$, whereas `Mpsi(x, *, deriv=-1)` ("(-1)-st derivative" is the primitive or antiderivative) computes $\rho(x)$, both according to the above definitions.

Weakly redescending ψ functions. Note that the above definition does require a finite rejection point x_r. But there exist $\psi(\cdot)$ functions having $x_r = \infty$, e.g. $\psi_C(x) := s(x)/2$ with $s(x) = 2x/(1+x^2)$ score function for the Cauchy $(= t_1)$ distribution and hence $\psi_C(\cdot)$ is *not* a redescending ψ-function in the above sense. For this reason we call ψ-functions fulfilling $\lim_{x\to\infty} \psi(x) = 0$ *weakly redescending*. Note that they'd naturally fall into two sub categories, namely the one with a *finite* ρ-limit, i.e. $\rho(\infty) := \lim_{x\to\infty} \rho(x)$, and those for which $\rho(x)$ is unbounded even though $\rho' = \psi$ tends to zero.

Note: An alternative slightly more general definition of *redescending* would only require $\rho(\infty) := \lim_{x\to\infty} \rho(x)$ to be finite.

Monotone ψ-Functions

Monotone ψ-functions lead to convex ρ-functions such that the corresponding M-estimators are defined uniquely. Historically, the "Huber function" has been the first ψ-function, proposed by Huber (1964).

Huber

The family of Huber functions is defined as

$$\rho_M(x) = \begin{cases} \frac{1}{2}x^2 & \text{if } |x| \le M \\ M(|x| - \frac{M}{2}) & \text{if } |x| > M \end{cases},$$
$$\psi_M(x) = \begin{cases} x & \text{if } |x| \le M \\ M\ \text{sign}(x) & \text{if } |x| > M \end{cases}.$$

The constant M for 95% efficiency of the regression estimator is 1.345.

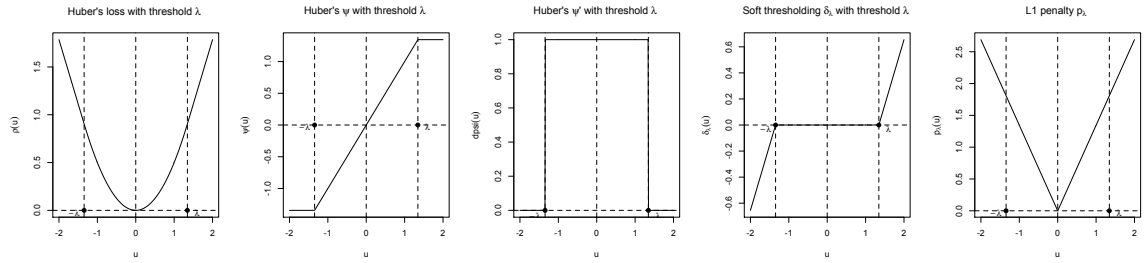

Fig. 8 Huber family of functions using tuning parameter $M = \lambda = 1.345$.

Redescenders

All the ψ-functions below, unless stated differently, are redescending, i.e. with finite "rejection point" $x_r = \sup\{t; \psi(t) > 0\} < \infty$. We recall here their definition and we visualize them in the following subsections.

Tukey's bisquare

Tukey's bisquare (aka "biweight") family of functions (see Tukey (1960)) is defined as

$$\tilde{\rho}_M(x) = \begin{cases} 1 - \left(1 - (x/M)^2\right)^3 & \text{if } |x| \le M \\ 1 & \text{if } |x| > M \end{cases},$$

with derivative $\tilde{\rho}'_M(x) = 6\psi_M(x)/M^2$ where,

$$\psi_M(x) = x\left(1 - \left(\frac{x}{M}\right)^2\right)^2 \cdot I_{\{|x| \le M\}}.$$

The constant M for 95% efficiency of the regression estimator is 4.685 and the constant for a breakdown point of 0.5 of the S-estimator is 1.548.

Hampel

The Hampel family of functions (see Hampel et al. (1986)) is defined as

$$\tilde{\rho}_{a,b,r}(x) = \begin{cases} \frac{1}{2}x^2/C & |x| \le a \\ \left(\frac{1}{2}a^2 + a(|x| - a)\right)/C & a < |x| \le b \\ \frac{a}{2}\left(2b - a + (|x| - b)\left(1 + \frac{r-|x|}{r-b}\right)\right)/C & b < |x| \le r \\ 1 & r < |x| \end{cases},$$

$$\psi_{a,b,r}(x) = \begin{cases} x & |x| \le a \\ a\ \text{sign}(x) & a < |x| \le b \\ a\ \text{sign}(x)\frac{r-|x|}{r-b} & b < |x| \le r \\ 0 & r < |x| \end{cases},$$

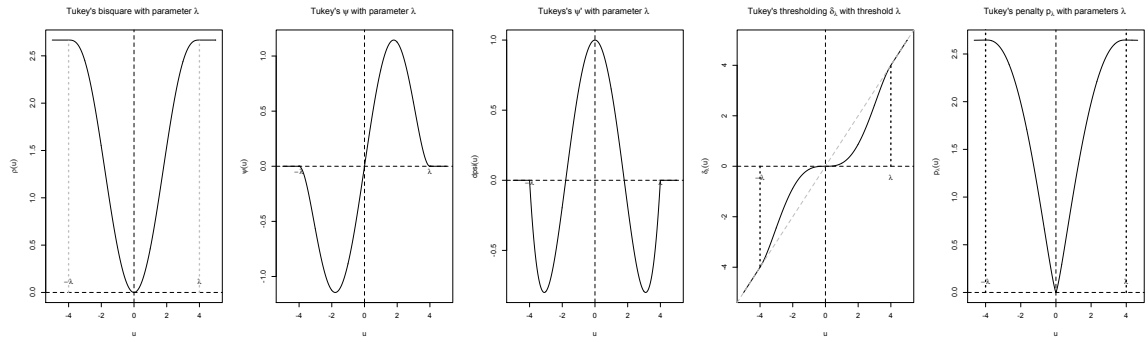

Fig. 9 Bisquare family functions using tuning parameter $M = \lambda$.

where $C := \rho(\infty) = \rho(r) = \frac{a}{2}(2b - a + (r - b)) = \frac{a}{2}(b - a + r)$.
By this standardization, ψ has slope 1 in the center. The slope of the redescending part ($x \in [b, r]$) is $-a/(r - b)$. If it is set to $-\frac{1}{2}$, as recommended sometimes, one has

$$r = 2a + b.$$

When restricting ourselves to a two-parameter family of Hampel functions with $a = b = M$ and $r = \gamma M$, where $\gamma > 1$ hence a redescending slope of $-\frac{1}{3}$, and varying M to get the desired efficiency or breakdown point, the resulting functions are those associated to the MCP penalties of Zhang (2010).

The constant M for 95% efficiency of the regression estimator is 0.9016085 and the one for a breakdown point of 0.5 of the S-estimator is 0.2119163.

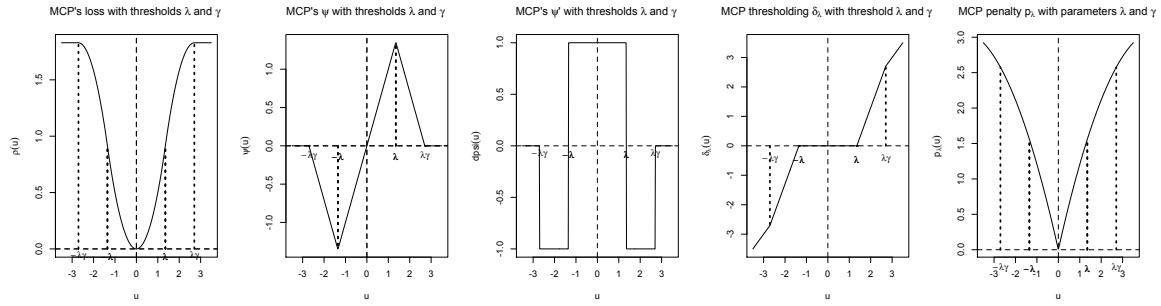

Fig. 10 MCP family of functions using tuning parameter $M = \lambda$ and γ.

Weak Redescenders

Cauchy loss

The Cauchy loss has also been propagated as "Lorentzian merit function" in regression for outlier detection. We have

$$\rho_M(u) = \frac{M^2}{2} \log\left(1 + \frac{u^2}{M^2}\right).$$

Note that ρ_M is *nonconvex*. When $M = 1$, the function $\rho_M(u)$ is proportional to the MLE for the t-distribution with one degree of freedom (a heavy-tailed distribution). This suggests that for heavy-tailed distributions, nonconvex loss functions may be more desirable from the point of

view of statistical efficiency, although optimisation becomes more difficult. For the Cauchy loss, we have

$$\psi_M(u) = \frac{u}{1+u^2/M^2}, \qquad \text{and} \qquad \psi'_M(u) = \frac{1-u^2/M^2}{(1+u^2/M^2)^2}.$$

In particular, $|\psi_M(u)|$ is maximized when $u^2 = M^2$, so $\|\psi_M\|_\infty \leq \frac{M}{2}$. We may also check that $\|\psi'_M\|_\infty \leq 1$ and $\|\psi''_M\|_\infty \leq \frac{3}{2M}$. The Cauchy ψ-functions fulfill $\lim_{x\to\infty} \psi(x) = 0$ so they are *weakly redescending*. The constant M for 95% efficiency of the regression estimator is 2.3849 (see Rey (1983)).

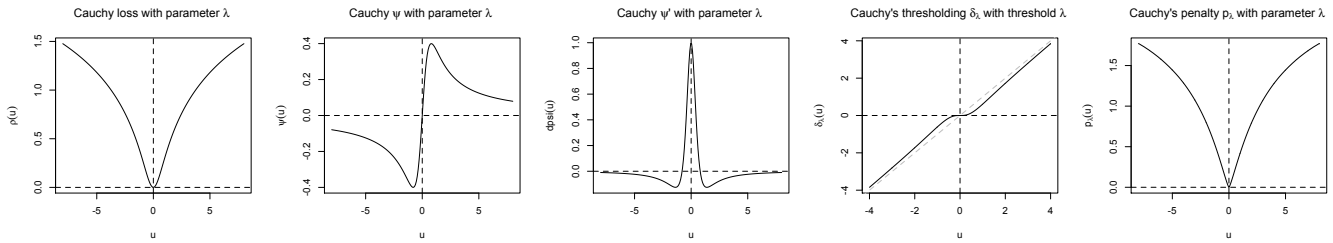

Fig. 11 Cauchy family of functions using tuning parameter $M = \lambda$.

Nonnegative garrote

The NNG functions are defined as (see Rodriguez (2017)),

$$\rho_M(x) = \begin{cases} 0.5x^2 & \text{if } |x| \leq M \\ 0.5M^2(1 + 2\log\left(\frac{|x|}{M}\right) & \text{if } |x| > M \end{cases}.$$

$$\psi_M(x) = \begin{cases} x & \text{if } |x| \leq M \\ M^2/|x| & \text{if } |x| > M \end{cases}.$$

The constant M for 95% efficiency of the regression estimator is 2.0 and the constant for a breakdown point of 0.5 of the S-estimator is 0.199.

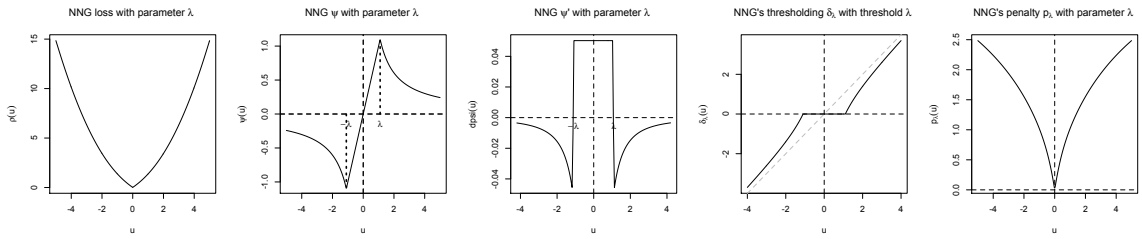

Fig. 12 Nonnegative garrote family of functions using tuning parameter $M = \lambda$.

In particular, ψ_M is maximised when $u = M$ so $\|\psi_M\|_\infty \leq M$. We may also check that $\|\psi'_M\|_\infty \leq 1$ and $\|\psi''_M\|_\infty$ is finite. The NNG ψ-functions also fulfill $\lim_{x\to\infty} \psi(x) = 0$ so they are *weakly redescending*.

Welsh

The Welsh functions (see Dennis and Welsch (1978)) are defined as,

$$\begin{aligned}\rho_M(x) &= 1 - \exp\left(-\left(x/M\right)^2/2\right)\\ \psi_M(x) &= M^2\rho'_M(x) = x\exp\left(-\left(x/M\right)^2/2\right)\\ \psi'_M(x) &= \left(1-\left(x/M\right)^2\right)\exp\left(-\left(x/M\right)^2/2\right)\end{aligned}$$

The constant M for 95% efficiency of the regression estimator is 2.9846 (see Rey (1983)) and the constant for a breakdown point of 0.5 of the S-estimator is 0.577.

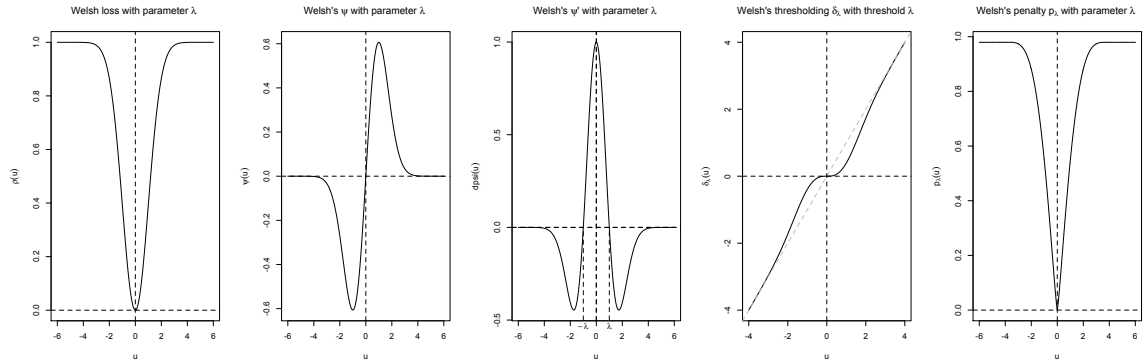

Fig. 13 Welsh family of functions using tuning parameter $M = \lambda$.

`"Welsh"` does *not* have a finite rejection point, but *does* have bounded ρ, and hence well defined $\rho(\infty)$.

Appendix 2: Technical proofs

Proof of Proposition 1

By definition, for any estimate $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\gamma}})$ minimising the criterion (1.4) with a penalty associated to a re-descending or weakly re-descending function, $\hat{\boldsymbol{\gamma}}$ is a fixed point of $\boldsymbol{\gamma} = \delta_M(\mathbf{H}\boldsymbol{\gamma} + (\boldsymbol{I} - \mathbf{H})\boldsymbol{y})$, and $\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T(\boldsymbol{y} - \hat{\boldsymbol{\gamma}})$, where $\mathbf{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ is the hat matrix associated to $\boldsymbol{X}$. It follows that

$$\begin{aligned}\boldsymbol{X}^T\psi_M(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) &= \boldsymbol{X}^T\psi_M(\boldsymbol{y} - \mathbf{H}(\boldsymbol{y} - \hat{\boldsymbol{\gamma}})\\ &= \boldsymbol{X}^T(((\boldsymbol{I} - \mathbf{H})\boldsymbol{y} + \mathbf{H}\hat{\boldsymbol{\gamma}}) - \delta_M((\boldsymbol{I} - \mathbf{H})\boldsymbol{y} + \mathbf{H}\hat{\boldsymbol{\gamma}}))\\ &= \boldsymbol{X}^T((\boldsymbol{I} - \mathbf{H})\boldsymbol{y} + \mathbf{H}\hat{\boldsymbol{\gamma}} - \hat{\boldsymbol{\gamma}})\\ &= \boldsymbol{X}^T(\boldsymbol{I} - \mathbf{H}) \cdot (\boldsymbol{y} - \hat{\hat{\boldsymbol{\gamma}}}) = \mathbf{0},\end{aligned}$$

and so $\hat{\boldsymbol{\beta}}$ is an M-estimate associated with ψ_M. □

Proof of Theorem 1

For any $\boldsymbol{\beta} \in \mathbb{R}^p$ and any $\sigma > 0$, we write $\boldsymbol{\beta} = \boldsymbol{\beta}^* + \mathbf{u}_n/\sqrt{n}$ and $\sigma = \sigma^* + \delta_n/\sqrt{n}$ where $\boldsymbol{\beta}^*$ and σ^* are the true location and scale parameters of the linear regression model (2.1). To avoid complicate notation we will suppress hereafter the index n. Note that by our assumptions the sequences $\mathbf{u}$ and δ are bounded. The loss function involved in Theorem 1 is

$$J_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n}\rho_\lambda\left(\frac{(Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta})}{\sigma}\right) + \mu_n\sum_{j=1}^{p}\hat{w}_j|\beta_j|,$$

and may be replaced for the optimisation by a function of $\mathbf{u}$ and δ defined by

$$\Psi_n(\mathbf{u},\delta) = \sum_{i=1}^{n} \rho_\lambda \left(\frac{(Y_i - \boldsymbol{X}_i^T(\boldsymbol{\beta}^* + \mathbf{u}/\sqrt{n}))}{\sigma^* + \delta/\sqrt{n}} \right) + \sqrt{n}\mu_n \sum_{j=1}^{p} \hat{w}_j \sqrt{n}|\beta_j^* + u_j/\sqrt{n}|. \tag{8.3}$$

Let

$$G_n(\mathbf{u},\delta) := \Psi_n(\mathbf{u},\delta) - \Psi(\mathbf{0},\delta) = A_n(\mathbf{u},\delta) + B_n(\mathbf{u},\delta)$$

where the first term $A_n(\mathbf{u},\delta)$ involves the summation terms in G_n related to ρ_λ while the second term B_n involves the summation terms related to the weighted differences of absolute values. Since both $\mathbf{u}$ and δ are bounded, and by the properties of the involved robust losses we can use a Taylor expansion with a remainder up to order 2 (denoted R_2 and R below) to get

$$\begin{aligned} A_n(\mathbf{u},\delta) = &-\frac{2\sqrt{n}}{\sigma^*}\left(\frac{1}{n}\sum_{i=1}^{n}\psi_\lambda(\epsilon_i/\sigma^*)\boldsymbol{X}_i^T\right)\mathbf{u} + \frac{1}{\sigma^{*2}}\mathbf{u}^T\left(\frac{1}{n}\sum_{i=1}^{n}\psi'_\lambda(\epsilon_i/\sigma^*)\boldsymbol{X}_i\boldsymbol{X}_i^T\right)\mathbf{u} \\ &+ \frac{2}{\sigma^{*2}}\left(\frac{1}{n}\sum_{i=1}^{n}\psi_\lambda(\epsilon_i/\sigma^*)\boldsymbol{X}_i^T + \frac{1}{n}\sum_{i=1}^{n}\frac{\epsilon_i}{\sigma^*}\psi'_\lambda(\epsilon_i/\sigma^*)\boldsymbol{X}_i\boldsymbol{X}_i^T\right)\mathbf{u}\,\delta \\ &+ 2\sum_{i=1}^{n} R_2\left((\frac{\boldsymbol{X}_i\mathbf{u}}{\sqrt{n}}, \frac{\delta}{\sqrt{n}})\right) - 2\sum_{i=1}^{n} R\left(\frac{\delta}{\sqrt{n}}\right). \end{aligned}$$

Since $\mathbb{E}_{F_\epsilon}(\psi_\lambda(\epsilon)) = 0$ and $\mathrm{var}_{F_\epsilon}(\psi_\lambda(\epsilon)) < \infty$ the central limit theorem yields

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n}\psi_\lambda(\epsilon_i\sigma^*)\boldsymbol{X}_i^T\right) \xrightarrow{d} \mathcal{N}(\mathbf{0}, a(\psi_\lambda, F_\epsilon)V).$$

For all robust losses considered in the theorem, it is easy to see that $\mathrm{var}_{F_\epsilon}\psi'_\lambda(\epsilon))$ is also finite, and therefore by assumption (A2), $\mathrm{var}_{F_\epsilon}\left(\frac{1}{n}\sum_{i=1}^{n}\psi'_\lambda(\epsilon_i/\sigma^*)\boldsymbol{X}_i\boldsymbol{X}_i^T\right) \to \mathbf{0}$. It follows by the law of large numbers that

$$\frac{1}{n}\sum_{i=1}^{n}\psi'_\lambda(\epsilon_i/\sigma^*)\boldsymbol{X}_i\boldsymbol{X}_i^T \xrightarrow{p} V\,\mathbb{E}_{F_\epsilon}(\psi'_\lambda(\epsilon)) \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n}\psi_\lambda(\epsilon_i/\sigma^*)\boldsymbol{X}_i^T \xrightarrow{p} \mathbf{0}.$$

For all robust losses considered in the theorem the derivatives ψ'_λ of the corresponding influence functions are even and by symmetry around 0 of the distribution F_ϵ we have $\mathbb{E}_{F_\epsilon}(\epsilon\,\psi'_\lambda(\epsilon_i/\sigma^*)) = 0$ and $\mathrm{var}_{F_\epsilon}(\epsilon\psi'_\lambda(\epsilon)) = \mathbb{E}_{F_\epsilon}(\epsilon^2\,\psi'^2_\lambda(\epsilon_i/\sigma^*)) \le T$, where T is a finite number since ψ'_λ are bounded. It follows that

$$\frac{1}{n}\sum_{i=1}^{n}\epsilon_i\psi'_\lambda(\epsilon_i/\sigma^*))\boldsymbol{X}_i^T \xrightarrow{p} \mathbf{0}$$

and therefore the term involving the multiplication $\mathbf{u}\delta$ in the expression of $A_n(\mathbf{u},\delta)$ converges to 0 in probability as n goes to ∞. By the remainder theorem on Taylor approximations, by Assumption (A2) and by the fact that both $\mathbf{u}$ and δ are bounded it follows easily that both $\sum_{i=1}^{n} R_2\left(\frac{\boldsymbol{X}_i\mathbf{u}}{\sqrt{n}}, \frac{\delta}{\sqrt{n}}\right)$ and $\sum_{i=1}^{n} R\left(\frac{\delta}{\sqrt{n}}\right)$ tend to 0 as n goes to ∞. The above imply that

$$A_n(\mathbf{u},\delta) \xrightarrow{d} \mathcal{N}\left(\frac{1}{\sigma^{*2}}b(\psi_\lambda, F_\epsilon)\mathbf{u}^T V\mathbf{u}, \frac{4}{\sigma^{*2}}a(\psi_\lambda, F_\epsilon)\mathbf{u}^T V\mathbf{u}\right).$$

Let's consider now the asymptotic behaviour of $B_n(\mathbf{u},\delta)$. The analysis of this term follows closely the one by Zou (2006) on the adaptive lasso. The only difference to be noted here is that we are using weights based on MM-estimators of $\boldsymbol{\beta}$ and a $\sqrt{n}$-consistent S-estimator $\hat{s_n}$ of σ^*. It follows that $\sqrt{n}(\hat{s}_n - \sigma^*) = \delta_n \xrightarrow{p} 0$ as n goes to ∞ and also that $\hat{w}_j = 1/|\hat{\beta}_j^{\mathrm{MM}}| \xrightarrow{p} 1/|\beta_j^*|$ for any $\beta_j^* \neq 0$, by the consistency of MM-estimates and preliminary S-estimates as discussed in Smucler and Yohai (2017). When $\beta_j^* = 0$, then the corresponding term involved in $B_n(\mathbf{u},\delta)$ is $\sqrt{n}(|\beta_j^* + \frac{u_j}{\sqrt{n}}| - |\beta_j^*|) = |u_j|$ and, since $\sqrt{n}\hat{\beta}_j^{\mathrm{MM}} = \mathcal{O}_p(1)$, $n\mu_n(\sqrt{n}\hat{\beta}_j^{\mathrm{MM}})^{-1}|u_j| \xrightarrow{d} \infty$. We may

then mimic completely the analysis of Zou (2006) (see also Lambert-Lacroix and Zwald (2011)) for the term $B_n(\mathbf{u}, \delta)$ to obtain the asymptotic normality result (i) of the theorem.

Assertion (ii) follows similarly from KKT conditions satisfied by $\hat{\boldsymbol{\beta}}_{II}$, the above asymptotic normality result of $\hat{\boldsymbol{\beta}}$, the $\sqrt{n}$-consistency rates of the MM-estimators of $\boldsymbol{\beta}^*$ and σ^* and the behaviour of the influence functions of the robust losses used in the theorem. □

References

Alfons A (2014) perryExamples: Examples for integrating prediction error estimation into regression models. R package version 0.1.0

Alfons A (2016) robustHD: Robust Methods for High-Dimensional Data. R package version 0.5.1

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

Antoniadis A (2007) Wavelet methods in statistics: some recent developments and their applications. Stat Surv 1:16–55

Antoniadis A (2010) Comments on: ℓ_1-penalization for mixture regression models [mr2677722]. TEST 19(2):257–258

Antoniadis A, Fan J (2001) Regularization of wavelet approximations. J Amer Statist Assoc 96(455):939–967, with discussion and a rejoinder by the authors

Antoniadis A, Gijbels I, Nikolova M (2011) Penalized likelihood regression for generalized linear models with nonquadratic penalties. The Annals of the Institute of Statistical Mathematics 63(3):585–615

Arslan O (2012) Weighted lad-lasso method for robust parameter estimation and variable selection in regression. Computational Statistics & Data Analysis 56(6):1952–1965

Avella Medina MA, Ronchetti E (2014) Robust and consistent variable selection for generalized linear and additive models. Technical report 310, University of Geneva

Belloni A, Chernozhukov V (2011) ℓ_1-penalized quantile regression in high-dimensional sparse models. Ann Statist 39(1):82–130

Belloni A, Chernozhukov V, Wang L (2011) Square-root lasso: pivotal recovery of sparse signals via conic programming. Biometrika 98(4):791–806

Bickel PJ, Ritov Y, Tsybakov AB (2009) Simultaneous analysis of lasso and Dantzig selector. Ann Statist 37(4):1705–1732

Bradic J, Fan J, Wang W (2011) Penalized composite quasi-likelihood for ultrahigh dimensional variable selection. J R Stat Soc Ser B Stat Methodol 73(3):325–349

Breheny P (2018) ncvreg: Regularization Paths for SCAD and MCP Penalized Regression Models. R package version 3.11-0

Bunea F (2008) Consistent selection via the lasso for high dimensional approximating regression models. In: Clarke B, Ghosal S (eds) Pushing the Limits of Contemporary Statistics: Contributions in Honor of Jayanta K. Ghosh, Collections, vol 3, Institute of Mathematical Statistics, Beachwood, Ohio, USA, pp 122–137, DOI 10.1214/074921708000000101

Bunea F, Tsybakov A, Wegkamp M (2007) Sparsity oracle inequalities for the Lasso. Electron J Stat 1:169–194

Cerioli A, Riani M, Atkinson AC, Corbellini A (2018) The power of monitoring: how to make the most of a contaminated multivariate sample. Stat Methods Appl 27:589–594

Chang X, Qu L (2004) Wavelet estimation of partially linea rmodels. Computational statistics and data analysis 47(1):31–48

Chen Z, Tang ML, Gao W, Shi NZ (2014) New robust variable selection methods for linear regression models. Scand J Stat 41(3):725–741

Cohen Freue GV, Kepplinger D, Salibian-Barrera M, Smucler E (2018) Proteomic biomarker study using novel robust penalized elastic net estimators, submitted to the Annals of Applied Statistics

Dennis JEJ, Welsch RE (1978) Techniques for nonlinear least squares and robust regression. Communications in Statistics - Simulation and Computation 7(4):345–359

Donoho D, Huber PJ (1983) The notion of breakdown point. In: A Festschrift for Erich L. Lehmann, Wadsworth Statist./Probab. Ser., Wadsworth, Belmont, CA, pp 157–184

Efron B, Hastie T, Johnstone I, Tibshirani R (2004) Least angle regression. Ann Statist 32(2):407–499

Fadili J, Bullmore E (2005) Penalized partially linear models using sparse representation with an application to fmri time series. IEEE Transactions on signal processing 53(9):3436–3448

Fan J, Li R (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. J Amer Statist Assoc 96(456):1348–1360

Fan J, Li Q, Wang Y (2017) Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. J R Stat Soc Ser B Stat Methodol 79(1):247–265

Fu A, Narasimhan B, Diamond S, Miller J (2019) Cvxr: Disciplined convex optimization. Journal of Statistical Software to appear

Gannaz I (2007) Robust estimation and wavelet thresholding in partially linear models. Stat Comput 17(4):293–310

van de Geer SA (2008) High-dimensional generalized linear models and the lasso. Ann Statist 36(2):614–645

Gijbels I, Vrinssen I (2015) Robust nonnegative garrote variable selection in linear regression. Comput Statist Data Anal 85:1–22

Gijbels I, Verhasselt A, Vrissen I (2017) Consistency and robustness properties of the s-nonnegative garrote estimator. Statistics 51(4):921–947

Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA (1986) Robust statistics. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics, John Wiley & Sons, Inc., New York, the approach based on influence functions

Harrison D, Rubinfeld D (1978) Hedonic prices and the demand for clean air. Journal of Environmental Economics and Management 5:81–102

Huang J, Ma S, Zhang CH (2008) Adaptive Lasso for sparse high-dimensional regression models. Statist Sinica 18(4):1603–1618

Huber PJ (1964) Robust estimation of a location parameter. Ann Math Statist 35:73–101

Huber PJ (1981) Robust statistics. John Wiley & Sons, Inc., New York, wiley Series in Probability and Mathematical Statistics

Janssens KH, Deraedt I, Schalml O, Veeckman J (1998) Composition of 15-17th century archaeological glass vessels excavated in antwerp, belgium. Mikrochim Acta [SupplJ] 15:253–267

Javanmard A, Montanari A (2014) Confidence intervals and hypothesis testing for high-dimensional regression. J Mach Learn Res 15(1):2869–2909

Khan JA, Van Aelst S, Zamar RH (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

Kong D, Bondell H, Wu Y (2018) Fully efficient robust estimation, outlier detection, and variable selection via penalized regression. Statistica Sinica 28:1031–1062

Kraemer N, Schaefer J (2014) parcor: Regularized estimation of partial correlation matrices. R package version 0.2-6

Lambert-Lacroix S, Zwald L (2011) Robust regression through the Huber's criterion and adaptive lasso penalty. Electron J Stat 5:1015–1053

Loh PL (2017) Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. Ann Statist 45(2):866–896

Loh PL, Wainwright MJ (2012) High-dimensional regression with noisy and missing data: provable guarantees with nonconvexity. Ann Statist 40(3):1637–1664

Loh PL, Wainwright MJ (2015) Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. Journal of Machine Learning Research 16:559–616

Maechler M, al (2017) robustbase: Basic Robust Statistics. R package version 0.92-8

Maronna RA (2011) Robust ridge regression for high-dimensional data. Technometrics 53(1):44–53, supplementary materials available online

Maronna RA, Yohai VJ (1981) Asymptotic behavior of general M-estimates for regression and scale with random carriers. Z Wahrsch Verw Gebiete 58(1):7–20

Meinshausen N, Yu B (2009) Lasso-type recovery of sparse representations for high-dimensional data. Ann Statist 37(1):246–270

Negahban SN, Ravikumar P, Wainwright MJ, Yu B (2012) A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. Statist Sci 27(4):538–557

Nesterov Y (2007) Gradient methods for minimizing composite objective function. Discussion Paper 2007076, Center for Operations Research and Econometrics (CORE). Université Catholique de Louvain

Pace RK, Gilley OW (1997) Using the spatial configuration of the data to improve estimation. Journal of the Real Estate Finance and Economics 14:333–340

Rey W (1983) Introduction to Robust and Quasi-Robust Statistical Methods. Springer, Berlin Heidelberg

Rodriguez P (2017) A two-term penalty function for inverse problems with sparsity constrains. In: EUSIPCO 17, pp 2185–2189

Rosset S, Zhu J (2004) Discussion on least angle regression. Ann Statist 32(2):459–475

Rousseeuw P, Yohai V (1984) Robust regression by means of S-estimators. In: Robust and nonlinear time series analysis (Heidelberg, 1983), Lect. Notes Stat., vol 26, Springer, New York, pp 256–272

She Y, Owen AB (2011) Outlier detection using nonconvex penalized regression. J Amer Statist Assoc 106(494):626–639

Smucler E, Yohai VJ (2017) Robust and sparse estimators for linear regression models. Comput Stat Data Anal 111(C):116–130

Städler D, Bühlmann PJ, van De Geer N (2010) ℓ_1-penalization for mixture regression models. TEST 19(2):209–256

Sun T, Zhang CH (2010) Comments on: ℓ_1-penalization for mixture regression models [mr2677722]. TEST 19(2):270–275

Sun T, Zhang CH (2012) Scaled sparse linear regression. Biometrika 99(4):879–898

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J Roy Statist Soc Ser B 58(1):267–288

Tukey JW (1960) A survey of sampling from contaminated distributions. In: Contributions to probability and statistics, Stanford Univ. Press, Stanford, Calif., pp 448–485

Wainwright MJ (2009) Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. IEEE Trans Inf Theor 55(12):5728–5741

Wang H, Leng C (2007) Unified lasso estimation by least squares approximation. Journal of the American Statistical Association 102:1039–1048

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

Wang L (2013) The L_1 penalized LAD estimator for high dimensional linear regression. J Multivariate Anal 120:135–151

Wang Z, Liu H, Zhang T (2014) Optimal computational and statistical rates of convergence for sparse nonconvex learning problems. Ann Statist 42(6):2164–2201

Zhang CH (2010) Nearly unbiased variable selection under minimax concave penalty. Ann Statist 38(2):894–942

Zou H (2006) The adaptive lasso and its oracle properties. J Amer Statist Assoc 101(476):1418–1429

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J R Stat Soc Ser B Stat Methodol 67(2):301–320