Statistica Sinica **29** (2019), 1065-1081
doi:https://doi.org/10.5705/ss.202016.0549

ASYMPTOTICS FOR REDESCENDING M-ESTIMATORS IN LINEAR MODELS WITH INCREASING DIMENSION

Ezequiel Smucler

Universidad de Buenos Aires - CONICET

Abstract: This paper deals with the asymptotic statistical properties of a class of redescending M-estimators in linear models with increasing dimension. This class is large enough to include popular high breakdown point estimators such as S-estimators and MM-estimators, which were not covered by existing results in the literature. We prove consistency assuming only that $p/n \to 0$ and asymptotic normality essentially if $p^3/n \to 0$, where p is the number of covariates and n is the sample size.

Key words and phrases: Dimension asymptotics, M-estimators, MM-estimators, robust regression, S-estimators.

1. Introduction

The growing number of statistical problems with a large number of parameters has motivated the study of the asymptotic properties of estimators for statistical models with a number of parameters that diverges with the sample size. For the case of linear regression, consider a sequence of regression models

$$y_{i,n} = \mathbf{x}_{i,n}^T \boldsymbol{\beta}_{0,n} + u_{i,n}, \quad 1 \leq i \leq n$$

where $y_{i,n} \in \mathbb{R}$, $\mathbf{x}_{i,n} \in \mathbb{R}^{p_n}$ is a vector of fixed predictor variables, $\boldsymbol{\beta}_{0,n} \in \mathbb{R}^{p_n}$ is to be estimated and $u_{i,n}$ are i.i.d. random variables defined in a common probability space with distribution function F_0. u will denote a random variable with distribution F_0. We consider the case in which p_n may tend to infinity with n at a certain rate. To lighten the notation, we drop the n subscript from $y_{i,n}, \mathbf{x}_{i,n}, \boldsymbol{\beta}_{0,n}, p_n$ and $u_{i,n}$.

It is well known that the Least Squares estimator of β_0 is not robust, that is, it can be ruined by a small number of extreme outliers in the data, and it is very inefficient when the errors are heavy-tailed. This fact has led to the development of robust estimators. A general framework for estimation in the linear model is provided by M-estimators. The notion of an M-estimator was first introduced in the landmark paper Huber (1964) for the case of the estimation of a location

parameter and extended to the linear model in Huber (1973). Given a suitably chosen loss function ρ, the corresponding regression M-estimator is defined, see for example Section 4.4 of Maronna, Martin and Yohai (2006), as

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} \rho\left(\frac{r_i(\boldsymbol{\beta})}{\hat{\sigma}_n}\right), \tag{1.1}$$

where $r_i(\boldsymbol{\beta}) = y_i - \mathbf{x}_i^T \boldsymbol{\beta}$ and $\hat{\sigma}_n$ is an estimate of scale of the errors u_i, that may be estimated a priori or simultaneously. The scale estimate in (1.1) is needed to make the resulting regression estimator scale equivariant. For example, $\hat{\sigma}_n$ could be the median of the absolute values of the residuals of some initial regression estimator. For the case of a convex and differentiable loss function, (1.1) is essentially equivalent to

$$\sum_{i=1}^{n} \psi\left(\frac{r_i(\hat{\boldsymbol{\beta}})}{\hat{\sigma}_n}\right) \mathbf{x}_i = \mathbf{0}, \tag{1.2}$$

where $\psi = \rho'$; see Section 4.4 of Maronna, Martin and Yohai (2006). In this case, the resulting M-estimator is called a monotone regression M-estimator. When ψ tends to zero at infinity the resulting estimator is called a redescending regression M-estimator and, in this case, some solutions of (1.2) may not correspond to solutions of (1.1).

The robustness of an estimator can be measured by its stability when a small fraction of the observations is arbitrarily replaced by outliers that may not follow the assumed model. A robust estimator should not be much affected by a small fraction of 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. It can be shown that any regression equivariant estimator has a breakdown point of at most $1/2$. See, for example, Section 5.4.1 of Maronna, Martin and Yohai (2006). On the other hand, the breakdown point of monotone regression M-estimators is zero; see Section 5.16.1 of Maronna, Martin and Yohai (2006). Moreover, monotone regression M-estimators may be highly inefficient when the errors are heavy tailed. These facts have motivated the study of M-estimators defined using bounded, and hence non-convex, loss functions, since they can be tuned to have the maximal breakdown point of $1/2$, and be simultaneously highly efficient when the errors are normal and when they are heavy-tailed.

A brief history of the study of the asymptotic properties of M-estimators for

linear regression models with a diverging number of parameters goes as follows. To the best of our knowledge, the first analysis of this problem appears in Huber (1973). In Huber (1973), Huber studied the asymptotic properties of monotone regression M-estimators defined without using an estimate of scale. Motivated by problems in X-ray crystallography, Huber proposed to study the properties of these estimators when $p = p_n \to \infty$ and proved asymptotic normality when $p^3/n \to 0$. This result was improved by Yohai and Maronna (1979), who obtained similar results assuming only $p^{5/2}/n \to 0$. Carroll (1982) extended this result to heteroscedastic linear models. Portnoy (1984) and Portnoy (1985) studied the asymptotic properties of the solutions of M-estimating equations, (1.2), without including an estimate of scale and proved consistency and asymptotic normality assuming $(p \log p)/n \to 0$ and $(p \log n)^{3/2}/n \to 0$ respectively. Mammen (1989) obtained similar results, but assuming only $(p^{3/2} \log n)/n \to 0$. Welsh (1989), Bai and Wu (1994a) and Bai and Wu (1994b) further improved these results by relaxing the regularity conditions imposed on ρ or the rate of growth of p. He and Shao (2000) studied M-estimators of general parametric models with increasing dimension. More recently, El Karoui et al. (2013), Bean et al. (2013), El Karoui (2013), Donoho and Montanari (2016), and Nevo and Ritov (2016) have studied the asymptotic properties of monotone M-estimators when $p/n \to m \in (0, 1)$.

A related line of work is that of penalized M-estimators in the context of sparse high-dimensional linear models. Li, Peng and Zhu (2011) studied the asymptotic properties of penalized M-estimators defined using a convex loss function and a general penalty term, assuming that $p/n \to 0$. Bradic (2016) and Loh (2017) also studied the asymptotic properties of penalized M-estimators, but allowing for p to be possibly much greater than n.

None of these results are directly applicable to M-estimators defined using a bounded loss function, or to high-breakdown point estimators such as S-estimators (Rousseeuw and Yohai (1984)) or MM-estimators (Yohai (1987)). The only available result is that of Davies (1990), who proved the consistency of S-estimators assuming $(p \log n)/n \to 0$. In this paper, we prove consistency and asymptotic normality results for a class of redescending M-estimators that is large enough to include both S and MM-estimators. More precisely, we prove the consistency of the estimators under very general assumptions and requiring only that $p/n \to 0$ and we prove their asymptotic normality essentially when $p^3/n \to 0$.

The rest of this paper is organized as follows. In Section 2 we present the class of estimators we study and we show that S and MM-estimators belong

to this class. Moreover, we state and discuss the assumptions needed to prove our results, and compare our assumptions with those previously considered in the literature. In Section 3 we state our main results. Section 4 includes a simulation study of the finite-sample performance of two estimators that are covered by our theoretical results. Finally, in Section 5 we provide some conclusions. The Supplementary Material to this article contains the proofs of all our results.

2. Definitions and Assumptions

We begin by defining the class of estimators for which we prove results. We consider

$$L_n(\boldsymbol{\beta}, \hat{\sigma}_n) = \sum_{i=1}^{n} \rho_1 \left\{ \frac{r_i(\boldsymbol{\beta})}{\hat{\sigma}_n} \right\}, \tag{2.1}$$

and study the class of estimators defined by

$$\hat{\boldsymbol{\beta}} = \arg\min_{\beta \in \mathbb{R}^p} L_n(\boldsymbol{\beta}, \hat{\sigma}_n) \tag{2.2}$$

where $\hat{\sigma}_n$ is an estimate of the scale of the errors, that we assume satifies

$$\hat{\sigma}_n \xrightarrow{p} s_0 \tag{2.3}$$

for some deterministic positive value s_0, and ρ_1 is a bounded ρ-function in the sense of Maronna, Martin and Yohai (2006). In detail, ρ is said to be a ρ-function if (i) ρ is bounded, with $\lim_{t\to\infty} \rho(t) = 1$ and $\rho(0) = 0$; (ii) ρ is even and continuous; (iii) $\rho(t)$ is a non-decreasing function of $|t|$; (iv) $\rho(t_2) < \lim_{t\to\infty} \rho(t)$ and $0 \leq t_1 < t_2$ imply $\rho(t_1) < \rho(t_2)$.

Hence, $\hat{\boldsymbol{\beta}}$ as defined by (2.2) is a regression M-estimator defined using a particular type of bounded loss function. To make the notation lighter, we drop the $\hat{\sigma}_n$ argument form the definition of L_n, keeping it implicit.

Two commonly used bounded ρ-functions are Tukey's Bisquare loss function, given by

$$\rho_c^B(t) = 1 - \left\{ 1 - \left(\frac{t}{c} \right)^2 \right\}^3 I\{|t| \leq c\}, \tag{2.4}$$

and Welsh's loss function, given by

$$\rho_c^W(t) = 1 - \exp\left(- \left(\frac{t}{c} \right)^2 \right), \tag{2.5}$$

where $c > 0$ is some tuning constant, that can be chosen to give the resulting M-estimator of regression a given asymptotic efficiency at the normal distribution. For example, the tuning constants needed for 85% efficiency at the normal

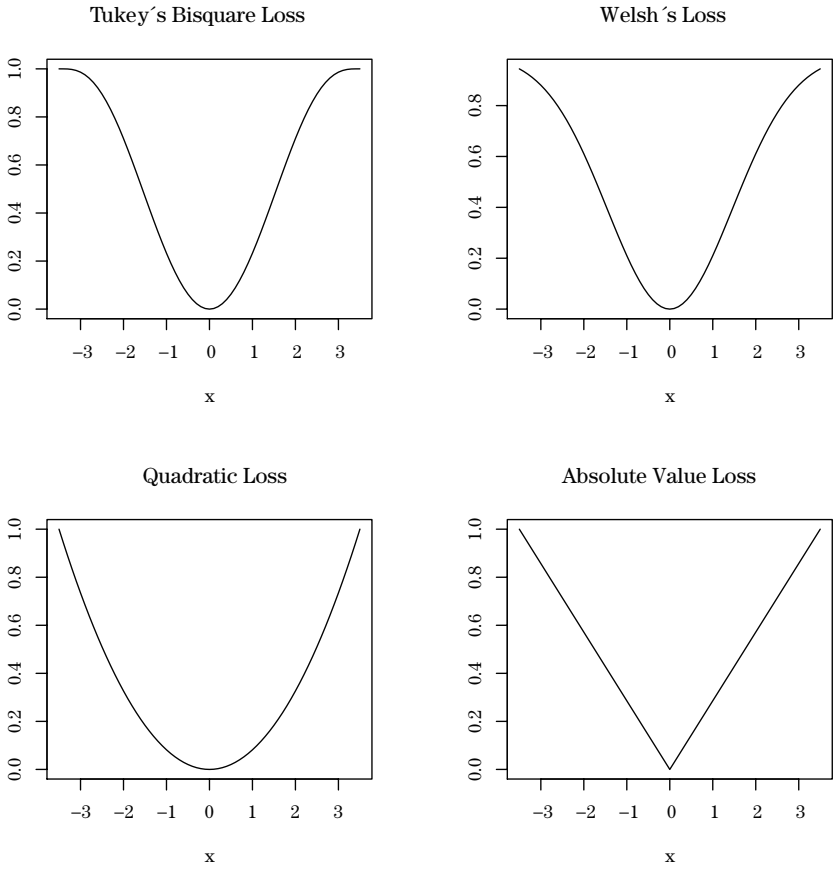

Figure 1. Plots of several loss functions.

distributions are 3.44 for Tukey's Bisquare and 1.46 for Welsh's loss. In Figure 1 we show plots of $\rho^B_{3.44}$, $\rho^W_{1.46}$, the absolute value loss that defines the Least Absolute Deviations estimator and the quadratic loss that defines the Least Squares estimator. The absolute value and the quadratic loss were standardized so as to have a maximum value of 1 over the interval $[-3.5, 3.5]$.

Even though to prove our theoretical results we only need (2.3) to hold, using a robust estimate of scale in (2.1) is crucial for obtaining robust regression estimators. The intuition behind using a bounded loss function is to give small weights to outliers, that is, observations with 'large' standardized residuals. The robust scale estimate used to standardize the residuals gives an indication of the typical size of the residuals when no outliers are present in the data. Moreover, a scale estimate is needed to make $\hat{\boldsymbol{\beta}}$ scale equivariant, and the breakdown point of $\hat{\sigma}_n$ affects the breakdown point of $\hat{\boldsymbol{\beta}}$. Hence, robust scale estimates play an

important role in robust regression. See for example Section 4.4.2 of Maronna, Martin and Yohai (2006). In Lemma 3 we give an example of a robust estimate of scale that satisfies (2.3).

A large class of robust estimates of scale is given by M-estimates of scale. Let ρ_0 be a bounded ρ-function. Given a vector $\mathbf{v} = (v_1, \ldots, v_n)$ and $0 < b < 1$, the corresponding M-estimate of scale $\hat{\sigma}_n^M(\mathbf{v})$ is defined, see Yohai (1987) for example, as the value $s > 0$ that is the solution of

$$\frac{1}{n}\sum_{i=1}^{n}\rho_0\left(\frac{v_i}{s}\right) = b, \tag{2.6}$$

if $\#\{i : v_i = 0\} < (1-b)n$, and as zero otherwise. We use the notation $\hat{\sigma}_n^M()$ to refer to the function whose value when evaluated at a vector $\mathbf{v}$ is $\hat{\sigma}_n^M(\mathbf{v})$. In Section 3.2.2 of Maronna, Martin and Yohai (2006) it is shown that the breakdown point of the M-estimate of scale is $\min(b, 1-b)$. In practice, b is usually taken to be $1/2$, so that the M-estimate of scale has maximal breakdown point. Then, one can choose ρ_0 such that $E\rho_0(v) = 1/2$ for v standard normal, to achieve consistency for the standard deviation in the case of normal observations. For example, one can take $\rho_0 = \rho^B_{1.54}$, where ρ^B is Tukey's Bisquare loss, (2.4).

2.1. S and MM-estimators

We show that S and MM-estimators are included in the class of estimators defined by (2.2).

S-estimators, introduced in Rousseeuw and Yohai (1984), are regression estimators that can be tuned to have a high breakdown point. They are defined by

$$\hat{\boldsymbol{\beta}}_S = \arg\min_{\beta\in\mathbb{R}^p}\hat{\sigma}_n^M(\mathbf{r}(\boldsymbol{\beta})), \tag{2.7}$$

where $\mathbf{r}(\boldsymbol{\beta}) = (r_1(\boldsymbol{\beta}), \ldots, r_n(\boldsymbol{\beta}))$ and $\hat{\sigma}_n^M()$ is an M-estimator of scale. If $\hat{\sigma}_n^M()$ is defined using $b = 1/2$, then $\hat{\boldsymbol{\beta}}^S$ has breakdown point equal to $1/2$, but S-estimators cannot combine a maximal breakdown point with arbitrarily high-efficiency at the normal distribution. Let ρ_0 be the ρ-function used to define $\hat{\sigma}_n^M()$. Then, S-estimators satisfy

$$\hat{\boldsymbol{\beta}}_S = \arg\min_{\beta\in\mathbb{R}^p}\sum_{i=1}^{n}\rho_0\left(\frac{r_i(\boldsymbol{\beta})}{\hat{\sigma}_n^M(\mathbf{r}(\hat{\boldsymbol{\beta}}_S))}\right),$$

see Section 5.6.1 of Maronna, Martin and Yohai (2006). In Lemma 3 we show that $\hat{\sigma}_n^M(\mathbf{r}(\hat{\boldsymbol{\beta}}_S))$ converges in probability to a positive value and hence S-estimators are included in the class of estimators defined by (2.2), since they are M-estimators

defined using a ρ-function and an estimate of scale that converges in probability to a positive constant.

MM-estimators, introduced in Yohai (1987), are regression estimators that can be tuned to attain both a high breakdown point and an arbitrarily high asymptotic efficiency at the normal distribution. Suppose $\hat{\boldsymbol{\beta}}_1$ is a highly robust, but not necessarily highly efficient, initial estimator. In practice, $\hat{\boldsymbol{\beta}}_1$ is usually an S-estimator, tuned to have maximal breakdown point. Let $\hat{\sigma}_n^M()$ be an M-estimator of scale defined using a bounded ρ-function ρ_0 and b. Let ρ_1 be another ρ-function that satisfies $\rho_1(t) \leq \rho_0(t)$ for all t. Then the MM-estimator is defined by

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

It can be shown, see Maronna, Martin and Yohai (2006), that $\hat{\boldsymbol{\beta}}_{MM}$ has a breakdown point that is at least as high as that of $\hat{\boldsymbol{\beta}}_1$. The original definition of MM-estimators is actually more general, but for technical convenience we work with this definition. Note that if we take $\hat{\boldsymbol{\beta}}_1$ as an S-estimator, then $\hat{\boldsymbol{\beta}}_{MM}$ is included in the class of estimators defined by (2.2). As we will see in the simulation study included in Section 4, MM-estimators can be tuned to be simultaneously efficient at the normal distribution and at heavy tailed distributions such as Student's t, while at the same time being resistant to the presence of outliers in the data.

2.2. Assumptions

We now state and discuss the assumptions needed to prove our results regarding the theoretical properties of (2.2).

R0. ρ_0 is a ρ-function and, for some $m > 0$, $\rho_0(t) = 1$ if $|t| \geq m$.

R1. ρ_1 is a continuously differentiable ρ-function. Let ψ_1 be the derivative of ρ_1. Then $\psi_1(t)$ and $t\psi_1(t)$ are bounded.

R2. ρ_1 is a three-times continuously differentiable ρ-function. Let ψ_1 be the derivative of ρ_1. Then $\psi_1(t)$, $\psi_1'(t)$, $\psi_1''(t)$, $t\psi_1(t)$, $t\psi_1'(t)$ and $t\psi_1''(t)$ are bounded.

These are additional conditions on the loss functions. We used ρ_0 in (2.6) to define M-estimates of scale and ρ_1 was used in (2.2) to define the estimators we are studying. Conditions R0 and R1 are satisfied by, for example, Tukey's

Bisquare loss function. Condition R2 is a strengthening of condition R1, and it is needed to obtain the asymptotic distribution of the estimators. It is satisfied by, for example, Welsh's loss (2.5) and by

$$\rho(t) = 1 - \left(1 - t^2\right)^4 I\left\{|t| \leq 1\right\},$$

which is similar to Tukey's Bisquare loss.

F0. F_0 has a density, f_0, that is an absolutely continuous, even, monotone decreasing function of $|t|$, and a strictly decreasing function of $|t|$ in a neighbourhood of 0.

Here F0 does not require finite moments from F_0, the distribution of the errors. The condition is clearly satisfied by the normal distribution, but also by heavy tailed distributions, such as Student's t-distribution. As shown in Lemma 6 in the Supplement, F0 and R2 together imply that $E\psi_1'(u/s_0) > 0$, a fact that is needed to obtain the rate of convergence of the estimators.

X0. $p < [n(1-b)]$ for all n, where b is the constant used in (2.6).

X1. a) There exists a constant $M > 0$ such that $(1/n)\sum_{i=1}^n \|\mathbf{x}_i\|^2 \leq pM$ for all n.

b) There exists a constant $B > 0$ such that $\max_{i \leq n} \|\mathbf{x}_i\| \leq Bn$ for all n.

Condition X0 is needed in the proof of the consistency of the scale estimate provided by the S-estimator. To prove the consistency of the regression estimators we need $p/n \to 0$. To obtain the rate of consistency of the estimators we need $(p \log n)/n \to 0$. Note that $(p \log n)/n \to 0$ is no stronger than $(p \log p)/n \to 0$, paraphrasing Portnoy (1984): if $p \leq \sqrt{n}$, $(p \log n)/n \leq (\log n)/\sqrt{n} \to 0$; while if $p \geq \sqrt{n}$, $(p \log n)/n \leq (2p \log p)/n$. X1 is needed to obtain the rate of convergence of the estimators. X1 a) holds when the covariates are standardized. X1 b) appears in Portnoy (1984) and holds, for example, if all the covariates are bounded and $p < n$, which we assume throughout this paper. To illustrate whether a condition on the design is reasonable, it is usual to show that if the predictors were sampled from some distribution, say the multivariate normal for example, then the condition holds with high probability; see for example Yohai and Maronna (1979) and Portnoy (1985). If $\mathbf{X}_i$, $i = 1, \ldots, n$, are independent and identically distributed random vectors in $\mathbb{R}^p$ such that for some C, $EX_{i,j}^2 \leq C$ for all i, j and n, then X1 holds in probability for $\mathbf{X}_i$ if $p/n \to 0$; see Section 4 of Portnoy (1984).

Let $\gamma_{1,n}$ and $\gamma_{2,n}$ be the smallest and largest eigenvalues of $\mathbf{\Sigma}_n = (1/n)\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T$.

X2. $\mathbf{\Sigma}_n$ is non-singular for all n and $\tau = \sup_n \gamma_{2,n} < \infty$.

Condition X2 is common in the literature and appears in, for example, Portnoy (1985) and Welsh (1989). See also Bai and Wu (1994a). It is needed to obtain the rate of consistency of the estimators.

For $0 < \alpha < 1$, let

$$\lambda_n(\alpha) = \min_{\mathcal{A}\subset\{1,\ldots,n\},\#\mathcal{A}=[n\alpha]} \left\{ \min_{\boldsymbol{\theta},\|\boldsymbol{\theta}\|=1} \left(\max_{i\in\mathcal{A}} |\mathbf{x}_i^T\boldsymbol{\theta}| \right) \right\}.$$

X3. For some $0 < \alpha < 1$, $\liminf_n \lambda_n(\alpha) > 0$.

The function $\lambda_n(\alpha)$ that appears in X3 was introduced in Davies (1990). For $\mathcal{A} \subset \{1,\ldots,n\}$ with $\#\mathcal{A} = [n\alpha]$ let $\Sigma(\mathcal{A}) = (1/[n\alpha])\sum_{i\in\mathcal{A}} \mathbf{x}_i\mathbf{x}_i^T$. Let $\gamma_{1,n}(\mathcal{A})$ be the smallest eigenvalue of $\Sigma(\mathcal{A})$. Take $\boldsymbol{\theta}$ with $\|\boldsymbol{\theta}\| = 1$. Then

$$\boldsymbol{\theta}^T\Sigma(\mathcal{A})\boldsymbol{\theta} \le \max_{i\in\mathcal{A}} |\mathbf{x}_i^T\boldsymbol{\theta}|^2.$$

Hence $\gamma_{1,n}(\mathcal{A}) \le \min_{\|\boldsymbol{\theta}\|=1} \max_{i\in\mathcal{A}} |\mathbf{x}_i^T\boldsymbol{\theta}|^2$ and

$$\min_{\mathcal{A}\subset\{1,\ldots,n\},\#\mathcal{A}=[n\alpha]} \gamma_{1,n}(\mathcal{A}) \le \lambda_n(\alpha)^2.$$

It follows that $\liminf_n \lambda_n(\alpha) > 0$ holds if the smallest eigenvalues of the covariance matrices formed from any subsample of size $[n\alpha]$ are uniformly bounded away from zero. An extended discussion of X3 can be found in the Supplement. The following lemma gives necessary conditions for $\liminf_n \lambda_n(\alpha) > 0$ to hold.

Lemma 1. *Assume* X1 *a) holds. Then, if* $\liminf_n \lambda_n(\alpha) > 0$ *for some* $0 < \alpha < 1$, *there exist positive numbers* η_1, η_2 *and* n_0 *such that*

$$\frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^T I\{\|\mathbf{x}_i\| < \eta_1\sqrt{p}\} - \eta_2\mathbf{I}_p$$

is positive definite for all $n \ge n_0$.

If X1 and X3 hold, by Lemma 1 we have that $\inf_n \gamma_{1,n} > 0$.

For $\mathbf{z} \in \mathbb{R}^p$ and $c > 0$, let $I(\mathbf{z}, c) = \{i = 1,\ldots,n : |\mathbf{x}_i^T\mathbf{z}| \le c\}$, let $\mathcal{B}(\delta)$ be the ball in $\mathbb{R}^p$ centered at zero with radius δ, and let $\mathcal{S}^*$ be the sphere centered at zero with radius 1.

X4. For any $c > 0$ there are constants $a > 0$, $\delta > 0$ and $C > 0$ such that for all $\beta \in \mathcal{B}(\delta)$, $\mathbf{z} \in \mathcal{S}^*$, and n, $\sum_{i\in J} \left(\mathbf{x}_i^T\mathbf{z}\right)^2 \ge an$, where $J = I(\beta, c) \cap I(\mathbf{z}, C)$.

X5. For any $c > 0$ and $\varepsilon > 0$ there are constants $\delta^{'} > 0$ and $C > 0$ such that for all $\beta \in \mathcal{B}(\delta^{'})$, $\mathbf{z} \in \mathcal{S}^*$, and n, $\sum_{i \notin J} \left(\mathbf{x}_i^T \mathbf{z}\right)^2 \leq \varepsilon n$, where $J = I(\beta, c) \cap I(\mathbf{z}, C)$.

X6. $\max_{i \leq n} \|\mathbf{x}_i\|^2 = o(n/p^2)$.

X4 and X5 were introduced in Portnoy (1984) where they appear as X1 and X2. Portnoy (1984) showed that these conditions hold in probability if the covariates are sampled from an appropriate distribution in $\mathbb{R}^p$, such as a scale mixture of standard multivariate normals, and $(p \log n)/n \to 0$. X4 and X5 are used in Lemma 7, a result that is needed in the proof of the rate of convergence of the estimators. This lemma shows that, very loosely speaking, $L_n(\beta)$ is convex in a neighbourhood of the true regression parameter with probability tending to one.

X6 is needed in the proof of the asymptotic normality of the estimators. It holds, for example, if the covariates are bounded and $p^3/n \to 0$. This is the rate of growth of p allowed by the asymptotic normality result of Huber (1973).

3. Main Results

In this section, we state and prove our main results.

We make extensive use of the tools from empirical processes theory that appear in Pollard (1989) and van der vaart and Wellner (1996). The results in Pollard (1989), in particular the maximal inequalities of Theorem 4.2, are stated and proved for i.i.d. random variables. The maximal inequality of Theorem 2.14.1 of van der vaart and Wellner (1996) is stated and proved for i.i.d. random variables. In Theorem 1 we extend Theorem 4.2 of Pollard (1989) to make it directly applicable to our scenario of interest, where the observations are of the form $(\mathbf{v}_i, \mathbf{z}_i)$, $i = 1, \ldots, n$, for $\mathbf{v}_1, \ldots, \mathbf{v}_n$ i.i.d. random vectors in $\mathbb{R}^m$ and $\mathbf{z}_1, \ldots, \mathbf{z}_n$ fixed vectors in $\mathbb{R}^d$.

We first introduce some notation. Let $\varepsilon > 0$. Let $\mathcal{H}$ be a class of functions defined on $\mathbb{R}^d$ and let $\|.\|$ be a pseudo-norm on $\mathcal{H}$. The capacity number of $\mathcal{H}$, $D(\varepsilon, \mathcal{H}, \|.\|)$, is the largest N such that there exists $h_1, \ldots, h_N$ in $\mathcal{H}$ with $\|h_i - h_j\| > \varepsilon$ for all $i \neq j$. The capacity number is also called the packing number in the literature. Given Q, a probability measure on $\mathbb{R}^d$ with finite support, let $\|.\|_{2,Q}$ be the $L^2(Q)$ pseudo-norm.

Theorem 1. *Let $\mathbf{z}_1, \ldots, \mathbf{z}_n$ be fixed vectors in $\mathbb{R}^d$. Let $\mathbf{v}_1, \ldots, \mathbf{v}_n$ be i.i.d. random vectors in $\mathbb{R}^m$. Let $\mathcal{H}$ be a class of functions defined in $\mathbb{R}^{m+d}$ and taking values*

in $\mathbb{R}$. *Assume* $\mathcal{H}$ *has envelope* H *that satisfies*

$$\left(\frac{1}{n}\right)\sum_{i=1}^{n} EH^2(\mathbf{v}_i, \mathbf{z}_i) < \infty$$

and that $\mathcal{H}$ *contains the zero function. Furthermore, assume that there exists a decreasing function* $D(\varepsilon)$ *that satisfies* $\int_0^1 \{\log D(\varepsilon)\}^{1/2}\, d\varepsilon < \infty$, *such that for all* $0 < \varepsilon < 1$ *and any probability measure on* $\mathbb{R}^{m+d}$ *with finite support* Q *with* $\|H\|_{2,Q} > 0$, $D(\varepsilon\|H\|_{2,Q}, \mathcal{H}, \|.\|_{2,Q}) \le D(\varepsilon)$. *Then*

(i)

$$E\sup_{h\in\mathcal{H}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{h(\mathbf{v}_i,\mathbf{z}_i) - Eh(\mathbf{v}_i,\mathbf{z}_i)\}\right| \le M\left\{\frac{1}{n}\sum_{i=1}^{n} EH^2(\mathbf{v}_i,\mathbf{z}_i)\right\}^{1/2}\left[\int_0^1 \{\log D(\varepsilon)\}^{1/2}\, d\varepsilon\right],$$

(ii)

$$E\sup_{h\in\mathcal{H}}\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{h(\mathbf{v}_i,\mathbf{z}_i) - Eh(\mathbf{v}_i,\mathbf{z}_i)\}\right|^2 \le M\frac{1}{n}\sum_{i=1}^{n} EH^2(\mathbf{v}_i,\mathbf{z}_i)\ \left[\int_0^1 \{\log D(\varepsilon)\}^{1/2}\, d\varepsilon\right]^2,$$

where $M > 0$ *is a fixed universal constant.*

The following lemma is a key result in the proof of the consistency of the estimators. Lemma 2 of Davies (1990) is a similar result, but requires $(p\log n)/n \to 0$. The improved rate provided by Lemma 2 explains the difference in the rate of growth of p our consistency result requires, $p/n \to 0$, and the rate required by Davies' result, $(p\log n)/n \to 0$.

Lemma 2. *Assume* ρ *is a* ρ*-function. Consider the class of functions*

$$\mathcal{H} = \left\{h_{s,\boldsymbol{\theta}}(t,\mathbf{x}) = \rho\left(\frac{t - \mathbf{x}^T\boldsymbol{\theta}}{s}\right) :\ \boldsymbol{\theta}\in\mathbb{R}^p\,, s > 0\right\}.$$

Then

$$E\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^{n}\{h(u_i,\mathbf{x}_i) - Eh(u,\mathbf{x}_i)\}\right| \le M\sqrt{\frac{p}{n}},$$

where $M > 0$ *is a constant depending only on* ρ. *In particular, if* $p/n \to 0$,

$$\sup_{h\in\mathcal{H}}\left|\frac{1}{n}\sum_{i=1}^{n}\{h(u_i,\mathbf{x}_i) - Eh(u,\mathbf{x}_i)\}\right| \xrightarrow{p} 0.$$

The following lemma proves the consistency of $\hat{\sigma}_n^M(\mathbf{r}(\hat{\boldsymbol{\beta}}_S))$, where $\hat{\boldsymbol{\beta}}_S$ is an S-estimator as defined in (2.7).

Lemma 3. *Assume* R0, F0 *and* X0 *hold and that* $p/n \to 0$. *If* f_0 *is strictly decreasing on the non-negative real numbers, then* $\hat{\sigma}_n^M\{\mathbf{r}(\hat{\boldsymbol{\beta}}_S)\} \xrightarrow{P} s(F_0)$, *where* $s(F_0)$ *is the positive solution of* $E\rho_0(u/s) = b$.

In particular, if ρ_0 is chosen to satisfy $E\rho_0(u) = b$ for u with standard normal distribution, $\hat{\sigma}_n^M(\mathbf{r}(\hat{\boldsymbol{\beta}}_S))$ is a consistent estimator of the standard deviation in the case of normal errors.

We are now ready to state the consistency of the estimators defined by (2.2).

Theorem 2 (Consistency)**.** *If* R1 *and* F0 *hold and* $p/n \to 0$, *then, for any* $0 < \alpha < 1$, $\lambda_n(\alpha)\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| \xrightarrow{P} 0$.

Note that Theorem 2 together with X3 entails that $\hat{\beta}$ is consistent. In the following theorem, we derive its rate of convergence.

Theorem 3 (Rate of convergence)**.** *If* R2, F0 *and* X1-X5 *hold and* $(p \log n)/n \to 0$, *then* $\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = O_P(\sqrt{p/n})$

If, under the assumptions of Theorem 3, we further assume that $\max_{i \le n} \|\mathbf{x}_i\|^2 = o(n/p)$ it follows that $\max_{i \le n} |\mathbf{x}_i^T(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)| \xrightarrow{p} 0$. Hence, Theorem 2 of Mammen (1989) can be applied to obtain asymptotic expansions for S-estimators.

Next, we derive the asymptotic distribution of $\hat{\beta}$.

Theorem 4 (Asymptotic distribution)**.** *Let* $\mathbf{a}_n$ *be a vector in* $\mathbb{R}^p$ *satisfying* $\|\mathbf{a}_n\| = 1$. *Let* $r_n^2 = \mathbf{a}_n^T \boldsymbol{\Sigma}_n^{-1} \mathbf{a}_n$. *If* R2, F0 *and* X1-X6 *hold and* $(p \log n)/n \to 0$, *then*

$$\sqrt{n} r_n^{-1} \mathbf{a}_n^T \left(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \right) \xrightarrow{d} N\left(0, s_0^2 \frac{a(\psi_1)}{b(\psi_1)^2}\right),$$

where $a(\psi_1) = E\psi_1^2(u/s_0)$ *and* $b(\psi_1) = E\psi_1'(u/s_0)$.

4. Simulation Study

The simulation study in this section aims at showing the usefulness of the class of estimators considered in the previous section, in particular of MM-estimators, in dealing with outliers in the data and heavy tailed errors. We compare the finite-sample performance with regards to robustness and efficiency of the following estimators.

LS The Least Squares estimator. This is the maximum likelihood estimator for the case of normal errors.

LAD The Least Absolute Deviations estimator. This is a monotone M-estimator and also the maximum likelihood estimator for the case of double exponential errors.

Tukey An MM-estimator, defined using Tukey's loss function, (2.4), and tuned to have an 85% asymptotic efficiency when the errors are normally distributed, with tuning constant $c = 3.44$. The initial estimator is an S-estimator defined using $b = 1/2$ and Tukey's loss function with tuning constant equal to 1.54. Hence, the initial S-estimator has maximal breakdown point and the associated scale estimate is consistent in the case of normal errors.

Welsh An MM-estimator, defined using Welsh's loss function, (2.5), and tuned to have an 85% asymptotic efficiency when the errors are normally distributed, with tuning constant $c = 1.46$. The initial estimator is the same as that of the last estimator.

Both MM-estimators are included in the class of estimators defined in Section 2 for which we proved asymptotic results. Tuning for an 85% efficiency at the normal distribution is chosen because that value of efficiency provides a good trade-off between robustness and efficiency, see Section 5.9 of Maronna, Martin and Yohai (2006). All computations were performed in `R`. To compute the MM-estimators we used the `robust` and `robustbase` packages.

We generated 500 Monte Carlo replications of a linear model with p predictors and n observations. We considered three possible combinations of (p, n): $(5, 40), (50, 500)$ and $(100, 1{,}500)$. We considered normally distributed predictors. Since all the estimator considered were regression, affine and scale equivariant, there is no loss in generality in taking the predictors to be i.i.d standard normal, and the true regression parameter $\boldsymbol{\beta}_0$ to be zero. We first considered two possible error distributions, normal (light tailed) and Student's t-distribution with three degrees of freedom (heavy tailed). Let $\hat{\boldsymbol{\beta}}^i$ be the result of one of the estimators being compared in the i-th Monte Carlo replication and let $\hat{\boldsymbol{\beta}}^i_{ML}$ be the Maximum Likelihood estimator computed using the i-th replication of the data. We measure the finite sample efficiency as $(\sum_{i=1}^{500} \|\hat{\boldsymbol{\beta}}^i_{ML}\|^2)/(\sum_{i=1}^{500} \|\hat{\boldsymbol{\beta}}^i\|^2)$.

Results are shown in Table 1. The efficiency of LS in the case of normal errors is not included in the table, since it is always 1. For normal errors, it is seen that the finite sample efficiency of the estimators is close to the asymptotic one, 85% for the MM-estimators and 64% for the LAD estimator. For the MM-estimator defined using Tukey's loss, the efficiency is seen to increase as the ratio

Table 1. Efficiencies of the estimators for errors with standard normal distribution and Student's t-distribution with 3 degrees of freedom.

	Normal			Student			
(p, n)	Tukey	Welsh	LAD	Tukey	Welsh	LAD	LS
$(5, 40)$	0.78	0.83	0.63	0.88	0.92	0.80	0.57
$(50, 500)$	0.81	0.82	0.65	0.93	0.96	0.81	0.55
$(100, 1{,}500)$	0.82	0.83	0.64	0.94	0.97	0.80	0.53

Table 2. Maximum MSE of the estimators under contamination.

(p, n)	Tukey	Welsh	LAD	LS
$(5, 40)$	0.57	0.67	3.15	6.05
$(50, 500)$	0.69	0.80	1.92	5.92
$(100, 1{,}500)$	0.60	0.74	1.49	5.54

p/n decreases. For the case of errors with Student's t-distribution with three degrees of freedom, the LS estimator is very inefficient, the LAD estimator does well and the MM-estimators are highly efficient. For both error distributions, the MM-estimators are more efficient than the LAD estimator. The MM-estimator defined using Welsh's loss is more efficient than the one defined using Tukey's loss. As we will see, the price to pay for this increase in efficiency is a loss in robustness.

To measure the robustness of the estimators to outliers, we introduced contamination in the data. We only considered normal errors. In each Monte Carlo replication, we contaminated 10% of the data by replacing, for $i = 1, \ldots, [0.1n]$, $\mathbf{x}_i$ with $(5, 0, \ldots, 0)$ and y_i with $5k$, where k, the outlier size, was moved in a grid between 0 and 3 with step 0.1. We then computed for each estimator the maximum mean squared error over all outlier sizes. Results are shown in Table 2. It is seen that the LS estimator and the LAD estimator are heavily affected by the outliers, more so in the case of a relatively high p/n ratio. On the other hand, both MM-estimators are seen to be resistant to the contamination. Note however that the MM-estimator defined using Welsh's loss has maximum MSEs that are around 10% higher than those of the MM-estimator defined using Tukey's loss. Finally, in Figure 2 a plot of the MSEs as a function of the outlier size for $(p, n) = (5, 40)$ is shown. The resulting curves for both MM-estimators are similar, and flatten out quickly. On the other hand, the performance of the LS and LAD estimators continues to deteriorate as the outlier size increases.

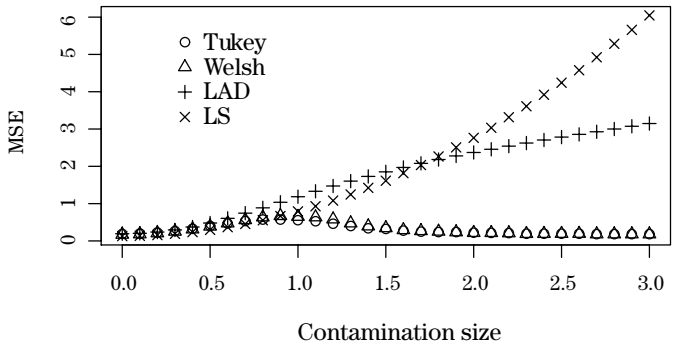

Figure 2. MSE as function of outlier sizes for $(p, n) = (5, 40)$ and normal errors.

5. Discussion

Recent papers, such as Bean et al. (2013) and Donoho and Montanari (2016), have studied the asymptotic properties of monotone M-estimators when $p/n \to m \in (0, 1)$. Extending their results to redescending M-estimators is part of our current research. We should note however, that the computation of redescending M-estimators is in general challenging, since it involves minimizing non-convex functions.

Recently Smucler and Yohai (2017) proposed regularized versions of MM-estimators, in an effort to obtain robust estimators that perform variable selection and work even when $p > n$. They studied their asymptotic properties only in the fixed p regime. Extending their results to linear models with increasing dimension is an interesting problem, requiring further research.

Supplementary Materials

The Supplementary Material contains the proofs of all the results stated in the paper, and a discussion of one of the main assumptions needed to prove the results.

Acknowledgment

Supported in part by Grant PIP 112-201101-00339 from CONICET and by a CONICET Doctoral Fellowship. This paper is based on the author's Ph.D. dissertation at the University of Buenos Aires. The author would like to express his gratitude to his advisor Victor J. Yohai, and to Graciela Boente, Daniela Rodriguez and Mariela Sued for their support and helpful suggestions. The

author would also like to thank two reviewers whose insightful comments and suggestions led to important improvements in this paper, both in presentation and content.

References

Bai, Z. D. and Wu, Y. (1994a). Limiting behavior of M-estimators of regression coefficients in high dimensional linear models I. Scale dependent case. *J. Multivariate Anal.* **51**, 211–239.

Bai, Z. D. and Wu, Y. (1994b). Limiting behavior of M-estimators of regression-coefficients in high dimensional linear models II. Scale-invariant case. *J. Multivariate Anal.* **51**, 240–251.

Bean, D., Bickel, P., El Karoui, N. and Yu, B. (2013). Optimal M-estimation in high-dimensional regression. *Proc. Natl. Acad. Sci. USA* **110**, 14563–14568.

Bradic, J. (2016). Robustness in sparse high-dimensional linear models: Relative efficiency and robust approximate message passing. *Electron. J. Stat.* **10**, 3894–3944.

Carroll, R. (1982). Robust estimation in certain heteroscedastic linear models when there are many parameters. *J. Statist. Plann. Inference* **7**, 1–12.

Davies, L. (1990). The asymptotics of S-estimators in the linear regression model. *Ann. Statist.* **18**, 1651–1675.

Donoho, D. L. and Huber, P. J. (1983). The notion of breakdown point. In *A Festschrift for Erich L. Lehmann* (Edited by P. J. Bickel, K. A. Doksum and J. L. Hodges), 157–185. Wadsworth, Belmont.

Donoho, D. L. and Montanari, A. (2016). High dimensional robust M-estimation: Asymptotic variance via approximate message passing. *Probab. Theory Related Fields* **166**, 935–969.

El Karoui, N. (2013). Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators : rigorous results. *ArXiv e-prints*. Available at `https://arxiv.org/abs/1311.2445`.

El Karoui, N., Bean, D., Bickel, P. J., Lim, C. and Yu, B. (2013). On robust regression with high-dimensional predictors. *Proc. Natl. Acad. Sci. USA* **110**, 14557–14562.

He, X. and Shao, Q. (2000). On parameters of increasing dimensions. *J. Multivariate Anal.* **73**, 120 – 135.

Huber, P. J. (1964). Robust estimation of a location parameter. *Ann. Math. Statist.* **35**, 73–101.

Huber, P. J. (1973). Robust regression: Asymptotics, conjectures and Monte Carlo. *Ann. Statist.* **1**, 799–821.

Li, G., Peng, H. and Zhu, L. (2011). Nonconcave penalized M-estimation with a diverging number of parameters. *Statist. Sinica* **21**, 391–419.

Loh, P. (2017). Statistical consistency and asymptotic normality for high-dimensional robust M-estimators. *Ann. Statist.* **45**, 866–896.

Mammen, E. (1989). Asymptotics with increasing dimension for robust regression with applications to the bootstrap. *Ann. Statist.* **17**, 382–400.

Maronna, R. A., Martin, D. R. and Yohai, V. J. (2006). *Robust Statistics: Theory and Methods*. Wiley, Chichester.

Nevo, D. and Ritov, Y. (2016). On Bayesian robust regression with diverging number of predictors. *ArXiv e-prints. Electron. J. Stat.* **10**, 3045–3062.

Pollard, D. (1989). Asymptotics via empirical processes. *Statist. Sci.* **4**, 341–354.

Portnoy, S. (1984). Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. I. Consistency. *Ann. Statist.* **12**, 1298–1309.

Portnoy, S. (1985). Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large. II. Normal approximation. *Ann. Statist.* **13**, 1403–1417.

Rousseeuw, P. J. and Yohai, V. J. (1984). Robust regression by means of S-estimators. In *Robust and Nonlinear Time Series Analysis* (Edited by J. Franke, W. Härdle and D. Martin), 256–272. Springer US, New York.

Smucler, E. and Yohai, V. J. (2017). Robust and sparse estimators for linear regression models. *Comput. Statist. Data Anal.* **111**, 116–130.

van der vaart, A. W. and Wellner, J. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer-Verlag New York, New York.

Welsh, A. H. (1989). On M-processes and M-estimation. *Ann. Statist.* **17**, 337–361.

Yohai, V. J. (1987). High breakdown-point and high efficiency robust estimates for regression. *Ann. Statist.* **15**, 642–656.

Yohai, V. J. and Maronna, R. A. (1979). Asymptotic behavior of M-estimators for the linear model. *Ann. Statist.* **7**, 258–268.

Instituto de Cálculo, Universidad de Buenos Aires - CONICET, Ciudad Universitaria, Pabellón 2, Buenos Aires 1426, Argentina.

E-mail: esmucler@ic.fcen.uba.ar

(Received December 2016; accepted November 2017)