COMSTA: 6418 Model 3Gsc pp. 1–15 (col. fig: NIL)

ARTICLE IN PRESS

Computational Statistics and Data Analysis xx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

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

Robust and sparse estimators for linear regression models

Ezequiel Smucler *, Victor J. Yohai

Instituto de Calculo, Universidad de Buenos Aires-CONICET, Ciudad Universitaria, Pabellon 2, Buenos Aires 1426, Argentina

ARTICLE INFO

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

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

ABSTRACT

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

1. Introduction

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

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

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

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

http://dx.doi.org/10.1016/j.csda.2017.02.002
0167-9473/© 2017 Elsevier B.V. All rights reserved.

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

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

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

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

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

2. MM and S-estimators

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

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

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

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

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

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

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

and $\|\mathbf{b}\| = \|\mathbf{b}\|_2$. Throughout this paper, a ρ-function will refer to a bounded ρ-function, in the sense of Maronna et al. (2006). That is, we will say that ρ is a ρ-function if: (i) ρ is even, continuous and bounded, (ii) $\rho(x)$ is a nondecreasing