Journal of the Korean Statistical Society 46 (2017) 390–403

Contents lists available at ScienceDirect

Journal of the Korean Statistical Society

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

Variable selection and parameter estimation via WLAD–SCAD with a diverging number of parameters

Yanxin Wang [a,*], Li Zhu [b]

[a] School of Science, Ningbo University of Technology, 315211, Ningbo, China

[b] School of Applied Mathematics, Xiamen University of Technology, 361024, Xiamen, China

ARTICLE INFO

Article history:
Received 7 June 2015
Accepted 5 December 2016
Available online 14 January 2017

AMS 2000 subject classifications:
primary 62J07
secondary 62F35

Keywords:
WLAD–SCAD
Robust regularization
Oracle property
Variable selection

ABSTRACT

In this paper, we focus on the variable selection based on the weighted least absolute deviation (WLAD) regression with the diverging number of parameters. The WLAD estimator and the smoothly clipped absolute deviation (SCAD) are combined to achieve robust parameter estimation and variable selection in regression simultaneously. Compared with the LAD–SCAD method, the WLAD–SCAD method will resist the heavy-tailed errors and outliers in explanatory variables. Furthermore, we obtain consistency and asymptotic normality of the estimators under certain appropriate conditions. Simulation studies and a real example are provided to demonstrate the superiority of the WLAD–SCAD method over the other regularization methods in the presence of outliers in the explanatory variables and the heavy-tailed error distribution.

1. Introduction

It is well known that variable selection and feature extraction are basic problems in high-dimensional and massive data analysis. In a traditional linear regression setting, many selection criteria, e.g., Akaike information criterion (AIC) and Bayesian information criterion (BIC) have been extensively used in practice. Recently, various shrinkage methods have been developed, which include but are not limited to the LASSO (Tibshirani, 1996; Zou, 2006), SCAD (Fan & Li, 2001) and Dantzig selector (Candès & Tao, 2007). Yet, most existing methods such as penalized least-squares or penalized likelihood (Fan & Lv, 2011) are designated for light-tailed distributions. The robustness of the aforementioned methods has not yet been thoroughly studied and well understood.

Robust regularization methods such as the least absolute deviation (LAD) regression and quantile regression have been used for variable selection in the case of fixed dimensionality (Li & Zhu, 2008; Wu & Liu, 2009; Zou & Yuan, 2008). Belloni and Chernozhukov (2011) studied the L_1-penalized quantile regression in high-dimensional sparse models where the dimensionality could be larger than the sample size. The LAD regression and the LASSO methods have been combined (LAD-LASSO) to carry out robust parameter estimation and variable selection simultaneously (Wang, Li, & Jiang, 2007; Xu & Ying, 2010). Li, Peng, and Zhu (2011) investigate the asymptotic properties of a nonconcave penalized M-estimator in sparse high-dimensional linear regression models. Wang (2013) investigate the L_1 penalized least absolute deviation method in the high-dimensional sparse linear regression model. The quantile regression for analyzing heterogeneity in ultrahigh dimension is introduced by Wang, Wu, and Li (2012). Wang and Li (2009) propose efficient shrinkage estimators, using the idea of rank

* Corresponding author.
E-mail addresses: wyxinbj@163.com (Y. Wang), zhulwhu@163.com (L. Zhu).

http://dx.doi.org/10.1016/j.jkss.2016.12.003
1226-3192/

regression. Fan, Fan, and Barut (2014) introduced the penalized quantile regression with the weighted L_1-penalty (WR-LASSO) for robust regularization. Feng, Zou, Wang, Wei, and Chen (2015) propose a robust variable selection method in varying coefficient model.

The LAD regression method is particularly well-suited to the heavy-tailed error distributions. However, it is well-known that the LAD regression estimation method is only resistant to the outlier in the response variable, but not resistant to the leverage points. As Wang and Leng (2007) point out, combining the LAD and the LASSO method can only produce estimators that are only resistant to the outliers in the response variable. If the outliers occur in the explanatory variables (leverage points) the performance of the LAD regression estimators is not better than the ordinary least squares (OLS) regression estimators so that the performance of the LAD-LASSO estimators will not be better than the ordinary LASSO estimators. To deal with the outliers in the explanatory variables the weighted LAD (WLAD) regression estimation has been proposed (Ellis & Morgenthaler, 1992; Giloni, Simonoff, & Sengupta, 2006). Correspondingly, Arslan (2012) proposed weighted LAD-LASSO method for robust parameter estimation and variable selection in regression.

In this paper we will propose a weighted version of the LAD–SCAD method to find the robust regression estimators and select the appropriate predictors. In our proposal we will combine the WLAD regression criterion and the SCAD penalty. The WLAD criterion will downweight the leverage points and be resistant to the outliers in the response variable so that the resulting regression estimators will be less sensitive to the leverage points and the outliers. Different from Arslan (2012), for the choice of the weights, in this paper we present the weights selection method based on the concept of "the decontamination subset" (Giloni, Simonoff et al., 2006). In theory, we shown that the LAD–SCAD estimator has the so-called "oracle property"; it is able to select variables consistently, and the estimators of nonzero coefficients have the same asymptotic distribution as they would if the zero coefficients were known in advance. Furthermore, we investigated the properties of the WLAD–SCAD estimator.

The rest of the paper is organized as follows. In Section 2 we introduce the WLAD–SCAD regression method to achieve robust parameter estimation and variable selection simultaneously in a regression analysis. We discuss some of its theoretical properties in Section 3. In Section 4, we describe the algorithm used to compute the WLAD–SCAD estimator and the criterion used to choose the regularization parameter. In Section 5 we provide simulation studies and a real example to demonstrate the performance of the proposed method. Concluding remarks are given in Section 6. The proofs of the main results are relegated to Appendix.

2. Weighted LAD–SCAD

2.1. LAD estimator

Consider the linear regression model

$$y_i = \mathbf{x}_i^T \beta_n + \varepsilon_i, \quad i = 1, 2, \ldots, n, \tag{1}$$

where $\beta_n = (\beta_{n1}, \ldots, \beta_{np_n})^T$ is the regression parameter, $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip_n})^T$ is the p_n-dimensional covariate vector, where $y_i \in \mathbb{R}$ is the response variable, and ε_i are the i.i.d. random errors. Here, the subscript is used to make it explicit that both the covariates and the parameters may change with n.

The most popular way of estimating β_n is to minimize the OLS criterion

$$\sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \beta_n)^2, \tag{2}$$

which yields the estimator $\hat{\beta}_n = (X^T X)^{-1} X^T \mathbf{y}$, where X is the $n \times p_n$ matrix whose ith row is $\mathbf{x}_i^T$ with rank p_n, and $\mathbf{y} = (y_1, y_2, \ldots, y_n)^T$ is the response vector. The usual assumption for using the OLS method is that the random errors ε_i are normally distributed with mean zero and variance σ^2. However, the OLS is not a robust method, because it is sensitive to outliers and is much less efficient if the error distribution has heavier tails than the normal distribution. A robust method provides a useful and stable alternative that is not sensitive to outliers.

Huber (1973) introduced M-estimation of β_n, which is defined as any value of $\hat{\beta}_n$ that minimizes

$$\sum_{i=1}^{n} \rho(y_i - \mathbf{x}_i^T \beta_n) \tag{3}$$

with a suitable choice of function ρ. Important examples include L_q regression estimate with $\rho(x) = |x|^q$, $1 \le q \le 2$. If $q = 1$, then the minimizer of (3) is called the least absolute deviation

$$\sum_{i=1}^{n} |y_i - \mathbf{x}_i^T \beta_n|. \tag{4}$$

However, it is also well known that the LAD regression method is resistant to the outliers in response variable, but it is very sensitive to the outlying observations in the explanatory variables. To correct this problem of the LAD regression method