Journal of Inequalities and Applications
a SpringerOpen Journal

RESEARCH Open Access

M-estimation in high-dimensional linear model

Kai Wang[1] and Yanling Zhu[1*]

*Correspondence:
zhuyanling99@126.com
[1]School of Statistics and Applied Mathematics, Anhui University of Finance and Economics, Bengbu, P.R. China

Abstract

We mainly study the M-estimation method for the high-dimensional linear regression model and discuss the properties of the M-estimator when the penalty term is a local linear approximation. In fact, the M-estimation method is a framework which covers the methods of the least absolute deviation, the quantile regression, the least squares regression and the Huber regression. We show that the proposed estimator possesses the good properties by applying certain assumptions. In the part of the numerical simulation, we select the appropriate algorithm to show the good robustness of this method.

MSC: 62F12; 62E15; 62J05

Keywords: M-estimation; High-dimensionality; Variable selection; Oracle property; Penalized method

1 Introduction

For the classical linear regression model $Y = X\beta + \varepsilon$, we are interested in the problem of variable selection and estimation, where $Y = (y_1, y_2, \ldots, y_n)^T$ is the response vector, $X = (X_1, X_2, \ldots, X_{p_n}) = (x_1, x_2, \ldots, x_n)^T = (x_{ij})_{n\times p_n}$ is an $n \times p_n$ design matrix, and $\varepsilon = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T$ is a random vector. The main topic is how to estimate the coefficients vector $\beta \in \mathrm{R}_n^p$ when p_n increases with sample size n and many elements of β equal zero. We can transfer this problem into a minimization of a penalized least squares objective function

$$\hat{\beta}_n = \arg\min_{\beta} Q_n(\beta_n), \quad Q_n(\beta_n) = \|Y - X\beta_n\|^2 + \sum_{j=1}^{p_n} p_{\lambda_n}(|\beta_{nj}|),$$

where $\|\cdot\|$ is the l_2 norm of the vector, λ_n is a tuning parameter, and $p_{\lambda_n}(|t|)$ a penalty term. It is well known that the least squares estimation is not robust, especially when in the data there exist abnormal values or the error term has a heavy-tailed distribution.

In this paper we consider the loss function to be the least absolute deviation, i.e., we minimize the following objective function:

$$\hat{\beta}_n = \arg\min_{\beta} Q_n(\beta_n), \quad Q_n(\beta_n) = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - x_i^T\beta_n\right| + \sum_{j=1}^{p_n} p_{\lambda_n}(|\beta_{nj}|),$$

where the loss function is the least absolute deviation (LAD for short) that does not need the noise to obey a gaussian distribution and be more robust than a least squares estimation. In fact, the LAD estimation is a special case of the M-estimation, which was mentioned by Huber (1964, 1973, 1981) [1–3] firstly and which can be obtained by minimizing the objective function

$$Q_n(\beta_n) = \frac{1}{n}\sum_{i=1}^{n}\rho\left(y_i - x_i^T\beta_n\right),$$

where the function ρ can be selected. For example, if we choose $\rho(x) = \frac{1}{2}x^2 1_{|x|\le c} + (c|x| - c^2/2)1_{|x|>c}$, where $c > 0$, the Huber estimator can be obtained; if we choose $\rho(x) = |x|^q$, where $1 \le q \le 2$, L_q estimator will be obtained, with two special cases: LAD estimator for $q = 1$ and OLS estimator for $q = 2$. If we choose $\rho(x) = \alpha x^+ + (1-\alpha)(-x)^+$, where $0 < \alpha < 1$, $x^+ = \max(x, 0)$, we call it a quantile regression, and we can also get the LAD estimator for $\alpha = 1/2$ especially.

When p_n approaches infinity as n tends to infinity, we assume that the function ρ is convex and not monotone, and the monotone function φ is the derivative of ρ. By imposing the appropriate regularity conditions, Huber (1973), Portnoy (1984) [4], Welsh (1989) [5] and Mammen (1989) [6] have proved that the M-estimator enjoyed the properties of consistency and asymptotic normality, where Welsh (1989) gave the weaker condition imposed on φ and the stronger condition on p_n/n. Bai and Wu [7] further pointed out that the condition on p_n could be a part of the integrable condition imposed on the design matrix. Moreover, He and Shao (2000) [8] studied the asymptotic properties of the M-estimator in the case of a generalized model setting and the dimension p_n getting bigger and bigger. Li (2011) [9] obtained the Oracle property of the non-concave penalized M-estimator in high-dimensional model with the condition of $p_n \log n/n \to 0$, $p_n^2/n \to 0$, and proposed RSIS to make a variable selection by applying a rank sure independence screening method in the ultra high-dimensional model. Zou and Li (2008) [10] combined a penalized function and a local linear approximation method (LLA) to prove that the obtained estimator enjoyed good asymptotic properties, and they demonstrated that this method improved the computational efficiency of a local quadratic approximation (LQA) in a simulation.

Inspired by this, in this paper we consider the following problem:

$$\hat{\beta}_n = \arg\min_{\beta_n} Q_n(\beta_n), \quad Q_n(\beta_n) = \frac{1}{n}\sum_{i=1}^{n}\rho\left(y_i - x_i^T\beta_n\right) + \sum_{j=1}^{p_n} p'_{\lambda_n}\left(|\tilde{\beta}_{nj}|\right)|\beta_{nj}|, \tag{1.1}$$

where $p'_{\lambda_n}(\cdot)$ is the derivative of the penalized function, and $\tilde{\beta}_n = (\tilde{\beta}_{n1}, \tilde{\beta}_{n2}, \ldots, \tilde{\beta}_{np_n})^T$ is the non-penalized estimator.

In this paper, we assume that the function ρ is convex, hence the objective function is still convex and the obtained local minimizer is a global minimizer.

2 Main results

For convenience, we first give some notations. Let $\beta_0 = (\beta_{01}, \beta_{02}, \ldots, \beta_{0p})^T$ be the true parameter. Without loss of generality, we assume the first k_n coefficients of the covariates are nonzero, then there are $p_n - k_n$ covariates with zero coefficients. $\beta_0 = (\beta_{0(1)}^T, \beta_{0(2)}^T)^T$, $\hat{\beta}_n = (\hat{\beta}_{n(1)}^T, \hat{\beta}_{n(2)}^T)^T$ correspondingly. For the given symmetric matrix Z, denote by $\lambda_{\min}(Z)$

and $\lambda_{\max}(Z)$ the minimum and maximum eigenvalue of Z, respectively. Denote $\frac{X^TX}{n} := D$ and $D = \begin{pmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{pmatrix}$, where $D_{11} = \frac{1}{n}X_{(1)}^T X_{(1)}$. Finally, we denote $c_n = \max\{|p'_{\lambda_n}(|\tilde{\beta}_{nj}|)| : \tilde{\beta}_{nj} \neq 0, 1 \leq j \leq p_n\}$.

Next, we state some assumptions which will be needed in the following results.

- (A_1) The function ρ is convex on R, and its left derivative and right derivative $\varphi_+(\cdot)$, $\varphi_-(\cdot)$ satisfies $\varphi_-(t) \leq \varphi(t) \leq \varphi_+(t)$, $\forall t \in R$.
- (A_2) The error term ε is i.i.d, and the distribution function F of ε_i satisfies $F(S) = 0$, where S is the set of discontinuous points of φ.

  Moreover, $E[\varphi(\varepsilon_i)] = 0$, $0 < E[\varphi^2(\varepsilon_i)] = \sigma^2 < \infty$, and $G(t) \equiv E[\varphi(\varepsilon_i + t)] = \gamma t + o(|t|)$, where $\gamma > 0$. Besides these, we assume that $\lim_{t\to 0} E[\varphi(\varepsilon_i + t) - \varphi(\varepsilon_i)]^2 = 0$.
- (A_3) There exist constants τ_1, τ_2, τ_3, τ_4 such that $0 < \tau_1 \leq \lambda_{\min}(D) \leq \lambda_{\max}(D) \leq \tau_2$ and $0 < \tau_3 \leq \lambda_{\min}(D_{11}) \leq \lambda_{\max}(D_{11}) \leq \tau_4$.
- (A_4) $\lambda_n \to 0$ $(n \to \infty)$, $p_n = O(n^{1/2})$, $c_n = O(n^{-1/2})$.
- (A_5) Let z_i be the transpose of the ith row vector of $X_{(1)}$, such that $\lim_{n\to\infty} n^{-\frac{1}{2}} \times \max_{1\leq i\leq n} z_i^T z_i = 0$.

It is worth mentioning that conditions (A_1) and (A_2) are classical assumptions for an M-estimation in a linear model, which can be found in many references, for example Bai, Rao and Wu (1992) [11] and Wu (2007) [12]. The condition (A_3) is frequently used for a sparse model in the linear model regression theory, which requires that the eigenvalues of the matrices D and D_{11} are bounded. The condition (A_4) is weaker than that in previous references. Under the condition (A_4) we broaden the order of p_n to $n^{1/2}$, but Huber (1973) and Li, Peng and Zhu (2011) [9] required that $p_n^2/n \to 0$, Portnoy (1984) required $p_n \log p_n/n \to 0$, and Mammen (1989) required $p_n^{3/2} \log p_n/n \to 0$. Compared with these results, it is obvious that our sparse condition is much weaker. The condition (A_5) is the same as that in Huang, Horowitz and Ma (2008) [13], which is used to prove the asymptotic properties of the nonzero part of M-estimation.

Theorem 2.1 (Consistency of estimator) *If the conditions* (A_1)–(A_4) *hold, there exists a non-concave penalized M-estimation* $\hat{\beta}_n$*, such that*

$$\|\hat{\beta}_n - \beta_0\| = O_P\big((p_n/n)^{1/2}\big).$$

Remark 2.1 From Theorem 2.1, we can see that there exists a global M-estimation $\hat{\beta}_n$ if we choose the appropriate tuning parameter λ_n; moreover, this M-estimation is $(n/p_n)^{1/2}$-consistent. This convergence rate is the same as that in the work of Huber (1973) and Li, Peng and Zhu (2011).

Theorem 2.2 (The sparse model) *If the conditions* (A_1)–(A_4) *hold and* $\lambda_{\min}(D) > \lambda_{\max}(\frac{1}{n} \times \sum_{i=1}^n J_i J_i^T)$*, for the non-concave penalized M-estimation* $\hat{\beta}_n$ *we have*

$$P(\hat{\beta}_{n(2)} = 0) \to 1.$$

Remark 2.2 By Theorem 2.2, we see that under the suitable conditions the global M-estimation of the zero-coefficient variables goes to zero with a high probability when n is large enough. This also shows that the model is sparse.

Theorem 2.3 (Oracle property) *If the conditions* (A_1)–(A_5) *hold and* $\lambda_{\min}(D) > \lambda_{\max}(\frac{1}{n} \times \sum_{i=1}^{n} J_i J_i^T)$, *with probability converging to one, the non-concave penalized M-estimation* $\hat{\beta}_n = (\hat{\beta}_{n(1)}^T, \hat{\beta}_{n(2)}^T)^T$ *has the following properties*:

(1) (*The consistency of the model selection*) $\hat{\beta}_{n(2)} = 0$;

(2) (*Asymptotic normality*)

$$\sqrt{n} s_n^{-1} u^T (\hat{\beta}_{n(1)} - \beta_{0(1)}) = \sum_{i=1}^{n} n^{-1/2} s_n^{-1} \gamma^{-1} u^T D_{11} z_i^T \varphi(\varepsilon_i) + o_P(1) \xrightarrow{\mathrm{d}} N(0,1),$$

where $s_n^2 = \sigma^2 \gamma^{-1} u^T D_{11}^{-1} u$, *and* u *is any* k_n *dimensional vector such that* $\|u\| \le 1$. *Meanwhile,* z_i *is the transpose of the ith row vector of a* $k_n \times k_n$ *matrix* $X_{(1)}$.

Remark 2.3 From Theorem 2.3, the M-estimation enjoys the Oracle property, that is, the M-estimator can correctly select covariates with nonzero coefficients with probability converging to one and the estimators of the nonzero coefficients has the same asymptotic distribution that they would have if the zero coefficients were known in advance.

Remark 2.4 In Fan and Peng (2004) [14], the authors showed that the non-concave penalized M-estimation has the property of consistency with the condition $p_n^4/n \to 0$, and enjoyed the property of asymptotic normality with the condition $p_n^5/n \to 0$. By Theorems 2.1–2.3, we can see that the corresponding conditions we impose are quite weak.

3 Proofs of main results

The proof of Theorem 2.1 Let $\alpha_n = (p_n/n)^{1/2} + p_n^{1/2} c_n$. For any p_n-dimensional vector u with $\|u\| = C$, we only need to prove that there exists a great enough positive constant C such that

$$\liminf_{n\to\infty} P\Big\{ \inf_{\|u\|=C} Q_n(\beta_0 + \alpha_n u) > Q_n(\beta_0) \Big\} \ge 1 - \varepsilon,$$

for any $\varepsilon > 0$, that is, there at least exists a local minimizer $\hat{\beta}_n$ such that $\|\hat{\beta}_n - \beta_0\| = O_P(\alpha_n)$ in the closed ball $\{\beta_0 + \alpha_n u : \|u\| \le C\}$.

Firstly, by the triangle inequality we get

$$\begin{aligned}
&Q_n(\beta_0 + \theta u) - Q_n(\beta_0) \\
&\quad = \frac{1}{n} \sum_{i=1}^{n} \big[\rho\big(y_i - x_i^T(\beta_0 + \alpha_n u)\big) - \rho\big(y_i - x_i^T \beta_0\big) \big] + \sum_{j=1}^{p_n} p'_{\lambda_n}\big(|\tilde{\beta}_{nj}|\big)\big(|\beta_{0j} + \alpha_n u_j| - |\beta_{0j}|\big) \\
&\quad \ge \frac{1}{n} \sum_{i=1}^{n} \big[\rho\big(y_i - x_i^T(\beta_0 + \alpha_n u)\big) - \rho\big(y_i - x_i^T \beta_0\big) \big] - \alpha_n \sum_{j=1}^{p_n} p'_{\lambda_n}\big(|\tilde{\beta}_{nj}|\big)|u_j| \\
&\quad := T_1 + T_2,
\end{aligned}$$

where $T_1 = \frac{1}{n}\sum_{i=1}^{n}[\rho(y_i - x_i^T(\beta_0 + \alpha_n u)) - \rho(y_i - x_i^T\beta_0)]$, $T_2 = -\alpha_n \sum_{j=1}^{p_n} p'_{\lambda_n}(|\tilde{\beta}_{nj}|)|u_j|$. Noticing that

$$\begin{aligned} T_1 &= \frac{1}{n}\sum_{i=1}^{n}\big[\rho\big(y_i - x_i^T(\beta_0 + \alpha_n u)\big) - \rho\big(y_i - x_i^T\beta_0\big)\big] \\ &= \frac{1}{n}\sum_{i=1}^{n}\big[\rho\big(\varepsilon_i - \alpha_n x_i^T u\big) - \rho(\varepsilon_i)\big] \\ &= \frac{1}{n}\sum_{i=1}^{n}\int_0^{-\alpha_n x_i^T u}\big[\varphi(\varepsilon_i + t) - \varphi(\varepsilon_i)\big]\,dt - \frac{1}{n}\alpha_n\sum_{i=1}^{n}\varphi(\varepsilon_i)x_i^T u \\ &:= T_{11} + T_{12}, \end{aligned} \tag{3.1}$$

where $T_{11} = \frac{1}{n}\sum_{i=1}^{n}\int_0^{-\alpha_n x_i^T u}[\varphi(\varepsilon_i + t) - \varphi(\varepsilon_i)]\,dt$, $T_{12} = -\frac{1}{n}\alpha_n\sum_{i=1}^{n}\varphi(\varepsilon_i)x_i^T u$. Combining with the Von-Bahr–Esseen inequality and the fact that $|T_{12}| \le \frac{1}{n}\alpha_n\|u\|\|\sum_{i=1}^{n}\varphi(\varepsilon_i)x_i\|$, we instantly have

$$E\left[\left\|\sum_{i=1}^{n}\varphi(\varepsilon_i)x_i\right\|^2\right] \le n\sum_{i=1}^{n}E\big[\|\varphi(\varepsilon_i)x_i\|^2\big] = n\sum_{i=1}^{n}E[\varphi^2(\varepsilon_i)x_i^T x_i \le n^2 p_n\sigma^2,$$

hence

$$|T_{12}| = O_P\big(\alpha_n p_n^{1/2}\big)\|u\| = O_P\big(\big(p_n^2/n\big)^{1/2}\big). \tag{3.2}$$

Secondly for T_{11}, let $T_{11} = \sum_{i=1}^{n}A_{in}$, where $A_{in} = \frac{1}{n}\int_0^{-\alpha_n x_i^T u}[\varphi(\varepsilon_i + t) - \varphi(\varepsilon_i)]\,dt$, so

$$T_{11} = \sum_{i=1}^{n}\big[A_{in} - E(A_{in})\big] + \sum_{i=1}^{n}E(A_{in}) := T_{111} + T_{112}.$$

We can easily obtain $E(T_{111}) = 0$. From the Von-Bahr–Esseen inequality, the Schwarz inequality and the condition (B_3), it follows that

$$\begin{aligned} \operatorname{var}(T_{111}) = \operatorname{var}\left(\sum_{i=1}^{n}A_{in}\right) &\le \frac{1}{n}\sum_{i=1}^{n}E\left(\int_0^{-\alpha_n x_i^T u}\big[\varphi(\varepsilon_i + t) - \varphi(\varepsilon_i)\big]\,dt\right)^2 \\ &\le \frac{1}{n}\sum_{i=1}^{n}\big|\alpha_n x_i^T u\big|\left|\int_0^{-\alpha_n x_i^T u}E\big[\varphi(\varepsilon_i + t) - \varphi(\varepsilon_i)\big]^2\,dt\right| \\ &= \frac{1}{n}\sum_{i=1}^{n}o_P(1)\big(\alpha_n x_i^T u\big)^2 = \frac{1}{n}o_P(1)\alpha_n^2\sum_{i=1}^{n}u^T x_i x_i^T u \\ &= o_P(1)\alpha_n^2 u^T D u \le \lambda_{\max}(D)o_P(1)\alpha_n^2\|u\|^2 = o_P\big(\alpha_n^2\big)\|u\|^2, \end{aligned}$$

so together with the Markov inequality this yields

$$P\big(|T_{111}| > C_1\alpha_n\|u\|\big) \le \frac{\operatorname{var}(T_{111})}{C_1^2\alpha_n^2\|u\|^2} \le \frac{o_P(\alpha_n^2)\|u\|^2}{C_1^2\alpha_n^2\|u\|^2} \to 0 \quad (n \to \infty),$$

hence

$$T_{111} = o_P(\alpha_n)\|u\|. \tag{3.3}$$

As for T_{112},

$$\begin{aligned} T_{112} &= \sum_{i=1}^{n} E(A_{in}) = \frac{1}{n}\sum_{i=1}^{n}\int_0^{-\alpha_n x_i^T u}\left[\gamma t + o\left(|t|\right)\right]dt \\ &= \frac{1}{n}\sum_{i=1}^{n}\left(\frac{1}{2}\gamma\alpha_n^2 u^T x_i x_i^T u + o_P(1)\alpha_n^2 u^T x_i x_i^T u\right) \\ &= \frac{1}{2}\gamma\alpha_n^2 u^T D u + o_p(1)\alpha_n^2 u^T D u \\ &\geq \left[\frac{1}{2}\gamma\lambda_{\min}(D) + o_P(1)\right]\alpha_n^2\|u\|^2. \end{aligned} \tag{3.4}$$

Finally, considering T_2, we can easily obtain

$$T_2 \leq (p_n)^{1/2}\alpha_n \max\left\{\left|p'_{\lambda_n}\left(|\tilde{\beta}_{nj}|\right)\right|, 1 \leq j \leq k_n\right\}\|u\| = (p_n)^{1/2}\alpha_n c_n\|u\| \leq \alpha_n^2\|u\|. \tag{3.5}$$

This together with (3.1)–(3.5) shows that we can choose a great enough constant C such that T_{111} and T_2 are controlled by T_{112}, from which it follows that there at least exists a local minimizer $\hat{\beta}_n$ such that $\|\hat{\beta}_n - \beta_0\| = O_P(\alpha_n)$ in the closed ball $\{\beta_0 + \alpha_n u : \|u\| \leq C\}$. □

The proof of Theorem 2.2 From Theorem 2.1, as long as we choose a great enough constant C and appropriate α_n, then $\hat{\beta}_n$ will be in the ball $\{\beta_0 + \alpha_n u : \|u\| \leq C\}$ with probability converging to one, where $\alpha_n = (p_n/n)^{1/2} + p_n^{1/2}c_n$. For any p_n-dimensional vector β_n, now we denote $\beta_n = (\beta_{n(1)}^T, \beta_{n(2)}^T)^T$, $\beta_{n(1)} = \beta_{0(1)} + \alpha_n u_{(1)}$, $\beta_{n(2)} = \beta_{0(2)} + \alpha_n u_{(2)} = \alpha_n u_{(2)}$, where $\beta_0 = (\beta_{0(1)}^T, \beta_{0(2)}^T)^T$, $\|u\|^2 = \|u_{(1)}\|^2 + \|u_{(2)}\|^2 \leq C^2$. Meanwhile let

$$V_n(u_{(1)}, u_{(2)}) = Q_n\left(\left(\beta_{n(1)}^T, \beta_{n(2)}^T\right)^T\right) - Q_n\left(\left(\beta_{0(1)}^T, 0^T\right)^T\right),$$

then by minimizing $V_n(u_{(1)}, u_{(2)})$ we can obtain the estimator $\hat{\beta}_n = (\hat{\beta}_{n(1)}^T, \hat{\beta}_{n(2)}^T)^T$, where $\|u_{(1)}\|^2 + \|u_{(2)}\|^2 \leq C^2$. In the following part, we will prove that, as long as $\|u\| \leq C$, $\|u_{(2)}\| > 0$,

$$P\left(V_n(u_{(1)}, u_{(2)}) - V_n(u_{(1)}, 0) > 0\right) \to 1 \quad (n \to \infty)$$

holds, for any p_n-dimensional vector $u = (u_{(1)}^T, u_{(2)}^T)^T$. We can easily find the fact that

$$\begin{aligned} &V_n(u_{(1)}, u_{(2)}) - V_n(u_{(1)}, 0) \\ &\quad = Q_n\left(\left(\beta_{n(1)}^T, \beta_{n(2)}^T\right)^T\right) - Q_n\left(\left(\beta_{n(1)}^T, 0^T\right)^T\right) \\ &\quad = \frac{1}{n}\sum_{i=1}^{n}\left[\rho\left(\varepsilon_i - \alpha_n H_i^T u_{(1)} - \alpha_n J_i^T u_{(2)}\right) - \rho\left(\varepsilon_i - \alpha_n H_i^T u_{(1)}\right)\right] + \sum_{j=k_n+1}^{p_n} p'_{\lambda_n}\left(|\tilde{\beta}_{nj}|\right)|\alpha_n u_j| \\ &\quad = \frac{1}{n}\sum_{i=1}^{n}\int_{-\alpha_n H_i^T u_{(1)}}^{-\alpha_n H_i^T u_{(1)} - \alpha_n J_i^T u_{(2)}}\left[\varphi(\varepsilon_i + t) - \varphi(\varepsilon_i)\right]dt - \frac{1}{n}\alpha_n\sum_{i=1}^{n}\varphi(\varepsilon_i)J_i^T u_{(2)} \end{aligned}$$

$$+\sum_{j=k_n+1}^{p_n} p'_{\lambda_n}\big(|\tilde{\beta}_{nj}|\big)|\alpha_n u_j|$$

$$:= W_1 + W_2 + W_3,$$

where H_i and J_i are k_n and $p_n - k_n$ dimensional vectors, respectively, such that $x_i = (H_i^T + J_i^T)^T$. Similar to the proof of Theorem 2.1, we get

$$\begin{aligned} W_1 &= \frac{1}{n}\sum_{i=1}^{n}\int_{-\alpha_n H_i^T u_{(1)}}^{-\alpha_n H_i^T u_{(1)} - \alpha_n J_i^T u_{(2)}} \big[\varphi(\varepsilon_i + t) - \varphi(\varepsilon_i)\big]\,dt \\ &= \frac{1}{2n}\sum_{i=1}^{n}\gamma\alpha_n^2 u^T x_i x_i^T u - \frac{1}{2n}\sum_{i=1}^{n}\gamma\alpha_n^2 u_{(2)}^T J_i J_i^T u_{(2)} + o_P(1)\alpha_n^2\|u\|^2 + o_P(1)\alpha_n\|u\| \\ &\geq \frac{1}{2}\gamma\alpha_n^2\left[\lambda_{\min}(D) - \lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^{n} J_i J_i^T\right)\right]\|u\|^2 + o_P(1)\alpha_n^2\|u\|^2 + o_P(1)\alpha_n\|u\|, \end{aligned} \tag{3.6}$$

$$|W_2| = \left|-\frac{1}{n}\alpha_n\sum_{i=1}^{n}\varphi(\varepsilon_i)J_i^T u_{(2)}\right| = O_P\big(\big(p_n^2/n\big)^{1/2}\big)\|u\|, \tag{3.7}$$

and

$$\begin{aligned} |W_3| &= \left|\sum_{j=k_n+1}^{p_n} p'_{\lambda_n}\big(|\tilde{\beta}_{nj}|\big)|\alpha_n u_j|\right| \leq (p_n)^{1/2}\alpha_n \max\big\{\big|p'_{\lambda_n}\big(|\tilde{\beta}_{nj}|\big)\big|, k_n + 1 \leq j \leq p_n\big\}\|u\| \\ &= (p_n)^{1/2}\alpha_n c_n\|u\| \leq \alpha_n^2\|u\|. \end{aligned} \tag{3.8}$$

By Eqs. (3.6)–(3.8) and the condition $\lambda_{\min}(D) > \lambda_{\max}(\frac{1}{n}\sum_{i=1}^{n} J_i J_i^T)$, it follows that

$$\begin{aligned} &V_n(u_{(1)}, u_{(2)}) - V_n(u_{(1)}, 0) \\ &\geq \frac{1}{2}\gamma\alpha_n^2\left[\lambda_{\min}(D) - \lambda_{\max}\left(\frac{1}{n}\sum_{i=1}^{n} J_i J_i^T\right)\right]\|u\|^2 \\ &\quad + o_P(1)\alpha_n^2\|u\|^2 + o_P(1)\alpha_n\|u\| + O_P\big(\big(p_n^2/n\big)^{1/2}\big)\|u\| + O_P\big(\alpha_n^2\big)\|u\| \\ &> 0, \end{aligned}$$

which shows that, as long as $\|u\| \leq C$, $\|u_{(2)}\| > 0$,

$$P\big(V_n(u_{(1)}, u_{(2)}) - V_n(u_{(1)}, 0) > 0\big) \to 1 \quad (n \to \infty)$$

holds, for any p_n-dimensional vector $u = (u_{(1)}^T, u_{(2)}^T)^T$. □

The proof of Theorem 2.3 It is obvious that the conclusion (1) can be obtained instantly by Theorem 2.2, so we only need to prove the conclusion (2). It follows from Theorem 2.1 that $\hat{\beta}_n$ is consistent with β_0 and $\hat{\beta}_{n(2)} = 0$ with probability converging to one from Theorem 2.2. Therefore for $\hat{\beta}_{n(1)}$

$$\left.\frac{\partial Q_n(\beta_n)}{\partial \beta_{n(1)}}\right|_{\beta_{n(1)} = \hat{\beta}_{n(1)}} = 0,$$

that is,

$$-\frac{1}{n}\sum_{i=1}^{n} H_i\varphi\big(y_i - H_i^T\hat{\beta}_{n(1)}\big) + W_{(1)} = 0,$$

where

$$W = \big(p'_{\lambda_n}\big(|\tilde{\beta}_{n1}|\big)\operatorname{sgn}(\hat{\beta}_{n1}), p'_{\lambda_n}\big(|\tilde{\beta}_{n2}|\big)\operatorname{sgn}(\hat{\beta}_{n2}), \ldots, p'_{\lambda_n}\big(|\tilde{\beta}_{np_n}|\big)\operatorname{sgn}(\hat{\beta}_{np_n})\big)^T.$$

In the following part we give the Taylor expansion of upper left first term:

$$-\frac{1}{n}\sum_{i=1}^{n}\big\{H_i\varphi\big(y_i - H_i^T\hat{\beta}_{0(1)}\big) - \big[\varphi'\big(y_i - H_i^T\beta_{0(1)}\big)H_iH_i^T + o_P(1)\big](\hat{\beta}_{n(1)} - \beta_{0(1)})\big\} + W_{(1)} = 0.$$

Noticing that $y_i = H_i^T\beta_{0(1)} + \varepsilon_i$, we have

$$-\frac{1}{n}\sum_{i=1}^{n} H_i\varphi(\varepsilon_i) + \frac{1}{n}\sum_{i=1}^{n}\big[\varphi'(\varepsilon_i)H_iH_i^T + o_P(1)\big](\hat{\beta}_{n(1)} - \beta_{0(1)}) + W_{(1)} = 0,$$

which shows that

$$\begin{aligned}\frac{1}{n}\gamma\sum_{i=1}^{n} H_iH_i^T(\hat{\beta}_{n(1)} - \beta_{0(1)}) &= \frac{1}{n}\sum_{i=1}^{n} H_i\varphi(\varepsilon_i) - W_{(1)} + (\hat{\beta}_{n(1)} - \beta_{0(1)})o_P(1)\\ &\quad + \frac{1}{n}\sum_{i=1}^{n}\big(\gamma - \varphi'(\varepsilon_i)\big)H_iH_i^T(\hat{\beta}_{n(1)} - \beta_{0(1)}).\end{aligned}$$

Then, as long as $\|u\| \le 1$,

$$\begin{aligned}u^T(\hat{\beta}_{n(1)} - \beta_{0(1)}) &= n^{-1}\gamma^{-1}u^TD_{11}^{-1}\sum_{i=1}^{n} H_i\varphi(\varepsilon_i)\\ &\quad + n^{-1}\gamma^{-1}u^TD_{11}^{-1}\sum_{i=1}^{n}\big(\gamma - \varphi'(\varepsilon_i)\big)H_iH_i^T(\hat{\beta}_{n(1)} - \beta_{0(1)})\\ &\quad - \gamma^{-1}u^TD_{11}^{-1}W_{(1)} + o_P(\alpha_n)\end{aligned}$$

holds, for any k_n-dimensional vector u. For the upper right third term, we can obtain

$$\begin{aligned}\big|\gamma^{-1}u^TD_{11}^{-1}W_{(1)}\big| &\le \frac{1}{\gamma\lambda_{\min}(D_{11})}\|W_{(1)}\| \le \frac{1}{\gamma\lambda_{\min}(D_{11})}p_n^{1/2}c_n\\ &\le \frac{\alpha_n}{\gamma\lambda_{\min}(D_{11})} \to o_P(1) \quad (n\to\infty).\end{aligned} \tag{3.9}$$

Now let us deal with the upper right second term. Theorem 2.1 and the condition (A_3) yield

$$\begin{aligned}
&\left|n^{-1}\gamma^{-1}u^T D_{11}^{-1}\sum_{i=1}^{n}\bigl(\gamma-\varphi'(\varepsilon_i)\bigr)H_iH_i^T(\hat{\beta}_{n(1)}-\beta_{0(1)})\right| \\
&\quad\le \frac{1}{n\gamma\lambda_{\min}(D_{11})}\left\|\sum_{i=1}^{n}\bigl(\gamma-\varphi'(\varepsilon_i)\bigr)H_iH_i^T(\hat{\beta}_{n(1)}-\beta_{0(1)})\right\| \\
&\quad\le \frac{1}{n\gamma\lambda_{\min}(D_{11})}\left\|\sum_{i=1}^{n}\bigl(\gamma-\varphi'(\varepsilon_i)\bigr)H_iH_i^T\right\|\|\hat{\beta}_{n(1)}-\beta_{0(1)}\| \\
&\quad\le \frac{O_P(1)}{n\gamma\lambda_{\min}(D_{11})}\|\hat{\beta}_{n(1)}-\beta_{0(1)}\| = O_P\bigl(p_n^{1/2}n^{-3/2}\bigr), \qquad (3.10)
\end{aligned}$$

where the upper third inequality sign holds because of Lemma 3 of Mammen (1989). Combining (3.9)–(3.10), we have

$$u^T(\hat{\beta}_{n(1)}-\beta_{0(1)}) = n^{-1}\gamma^{-1}u^T D_{11}^{-1}\sum_{i=1}^{n}H_i\varphi(\varepsilon_i)+O_P(\alpha_n)+O_P\bigl(p_n^{1/2}n^{-3/2}\bigr),$$

that is,

$$n^{1/2}u^T(\hat{\beta}_{n(1)}-\beta_{0(1)}) = n^{-1/2}\gamma^{-1}u^T D_{11}^{-1}\sum_{i=1}^{n}H_i\varphi(\varepsilon_i)+o_P(1).$$

Denote $s_n^2=\sigma^2\gamma^{-1}u^T D_{11}^{-1}u$, $F_{in}=n^{-1/2}s_n^{-1}\gamma^{-1}u^T D_{11}^{-1}z_i^T$, where z_i is a $k_n\times k_n$ matrix and the transpose of the ith row vector of $X_{(1)}$, then $n^{1/2}u^T(\hat{\beta}_{n(1)}-\beta_{0(1)})=\sum_{i=1}^{n}F_{in}\varphi(\varepsilon_i)+o_P(1)$. It follows from (A_5) that

$$\begin{aligned}
\sum_{i=1}^{n}F_{in}^2 &= \sum_{i=1}^{n}F_{in}F_{in}' = \sum_{i=1}^{n}\bigl(n^{-1/2}s_n^{-1}\gamma^{-1}u^T D_{11}^{-1}z_i^T\bigr)\bigl(n^{-1/2}s_n^{-1}\gamma^{-1}z_iD_{11}^{-1}u\bigr) \\
&= \sum_{i=1}^{n}n^{-1}s_n^{-2}\gamma^{-2}u^T D_{11}^{-1}z_i^Tz_iD_{11}^{-1}u = s_n^{-2}\gamma^{-2}u^T D_{11}^{-1}u=\sigma^{-2}.
\end{aligned}$$

Applying the Slutsky theorem, we see that

$$\sqrt{n}s_n^{-1}u^T(\hat{\beta}_{n(1)}-\beta_{0(1)})\xrightarrow{\mathrm{d}}N(0,1). \qquad \square$$

4 Simulation results

In this section we evaluate the performance of the M-estimator proposed in (1.1) by simulation studies.

We begin with the data. Simulating the data by the model $Y=X\beta+\varepsilon$, where $\beta_{0(1)}=(-2,2.5,3,-1)^T$, ε follows $N(0,1)$, t_5 and mixed normally distribution $0.9N(0,1)+0.1N(0,9)$, respectively. The design matrix X is generated by a p-dimensional multivariate normal distribution with mean zero and covariance matrix whose (i,j)th component is $\rho^{|i-j|}$, where we set $\rho=0.5$.

Then for the loss function. In this section we can choose some special loss functions, such as the LAD loss function, the OLS loss function and the Huber loss function. In this paper we choose the LAD loss function and the Huber loss function.

About the penalty function: for $p'_{\lambda_n}(|\widetilde{\beta}_{nj}|)$ in the penalty function, we choose the penalty function as a SACD estimation in the following:

$$p_{\lambda_n}(|\beta|) = \begin{cases} \lambda_n|\beta|, & 0 \le |\beta| \le \lambda_n, \\ -(\beta^2 - 2a\lambda_n|\beta| + \lambda_n^2)/(2(a-1)), & \lambda_n < |\beta| < a\lambda_n, \\ (a+1)\lambda_n^2/2, & |\beta| > a\lambda_n, \end{cases}$$

then $p'_{\lambda_n}(|\widetilde{\beta}_{nj}|) = \lambda_n I(|\widetilde{\beta}_{nj}| \le \lambda_n) + \frac{a\lambda_n - |\widetilde{\beta}_{nj}|}{a-1} I(\lambda_n < |\widetilde{\beta}_{nj}| \le a\lambda_n)$. By the proposal of Fan and Li (2001), we can select $a = 3.7$, which shows that generalized cross validation can be applied in searching the best tuning parameter λ_n.

About stimulation algorithm. For the proposed LLA method, we connect the penalty function with independent variables and an independent variable, respectively, then we write a program by using the quantile package in R. For the Lasso method, we use the Lars package to simulate.

Now we address the selection of the tuning parameter. We apply the BIC criterion to select the tuning parameter. The criterion is

$$\mathrm{BIC}(\lambda_n) = \ln\left(\frac{1}{n}\sum_{i=1}^{n}\rho\left(y_i - x_i^T\hat{\beta}\right)\right) + DF_{\lambda_n}\ln(n)/n,$$

where DF_{λ_n} is the generalized degree of freedom used by Fan and Li (2001).

About the selection of the evaluation index. In order to evaluate the performance of the estimators, we select four measures called EE, PE, C, IC and CP, which are obtained by 500 replicates. EE is the median of $\|\hat{\beta} - \beta_0\|_2$ to evaluate the estimation accuracy, and PE is the prediction error defined by the median of $n^{-1}\|Y - X\hat{\beta}\|^2$. The other three measures are to qualify the performance of model consistency, where C and IC refer to the average number of correctly selected zero covariates and the average number of incorrectly selected zero covariates, and CP is the proportion of the number of the correct selection of zero variables to the total number of zero variables.

In the following we will compare the performances of the method LLA we proposed, the Lasso method and the Oracle estimation. Set $n = 200, 500, 700$, respectively, and $p = [2\sqrt{n}]$.

From Table 1, we notice that the indices EE, C, IC, CP of our proposed LLA method perform better when $\varepsilon \sim N(0,1)$. In particular, for the index CP, LLA outperforms Lasso. The reason for this may be that we impose different penalties for important and unimportant variables, while Lasso imposes the same penalties for all variables. Moreover, with the increase of the sample size, the ability of LLA method to correctly identify unimportant variables is also increasing. When the sample size is 700 and the number of explanatory variables is 53, an average of 48.9617 unimportant variables-zero variables are estimated to be zero on average, with an average accuracy of 99.92%.

An interesting fact can be found from Table 2, that is, when the error term is chosen as t_5, the accuracy of the method LLA proposed to correctly exclude incorrect variables

Table 1 Simulation results for LAD loss function and $\varepsilon \sim N(0,1)$

Setting	Method	EE	PE	C	IC	CP
$n = 200$	Oracle	10.8544	3.3916	24.0000	0	100%
$p = 28$	Lasso	10.5726	3.3035	10.8480	0	45.20%
$m = 24$	LLA	10.9153	3.3947	23.8540	0	99.39%
$n = 500$	Oracle	19.9085	5.4118	41.0000	0	100%
$p = 45$	Lasso	19.5952	5.2928	18.9920	0	46.32%
$m = 41$	LLA	19.9233	5.4045	40.9140	0	99.79%
$n = 700$	Oracle	24.3006	6.3847	49.0000	0	100%
$p = 53$	Lasso	24.0315	6.2994	23.1009	0	47.14%
$m = 49$	LLA	24.3666	6.4077	48.9617	0	99.92%

Table 2 Simulation results for LAD loss function and $\varepsilon \sim t_5$

Setting	Method	EE	PE	C	IC	CP
$n = 200$	Oracle	10.5634	4.2892	24.0000	0	100%
$p = 28$	Lasso	10.2810	4.1649	11.7700	0	49.04%
$m = 24$	LLA	10.6448	4.2725	23.8780	0	99.49%
$n = 500$	Oracle	19.4296	6.8240	41.0000	0	100%
$p = 45$	Lasso	19.1157	6.7042	18.9580	0	46.24%
$m = 41$	LLA	19.4665	6.8335	40.9560	0	99.89%
$n = 700$	Oracle	23.7784	8.0637	49.0000	0	100%
$p = 53$	Lasso	23.4389	7.9551	22.8800	0	46.69%
$m = 49$	LLA	23.7808	8.0919	48.9740	0	99.94%

Table 3 Simulation results for LAD loss function and $\varepsilon \sim 0.9N(0,1) + 0.1N(0,9)$

Setting	Method	EE	PE	C	IC	CP
$n = 200$	Oracle	10.4815	4.4830	24.0000	0	100%
$p = 28$	Lasso	10.2030	4.4063	11.6360	0	48.48%
$m = 24$	LLA	10.5826	4.4529	23.9240	0	99.68%
$n = 500$	Oracle	19.2539	7.1997	41.0000	0	100%
$p = 45$	Lasso	18.9670	7.0960	19.3840	0	47.28%
$m = 41$	LLA	19.2950	7.1173	40.9520	0	99.88%
$n = 700$	Oracle	23.6354	8.5657	49.0000	0	100%
$p = 53$	Lasso	23.2424	8.4609	23.0580	0	47.06%
$m = 49$	LLA	23.6566	8.3699	48.9300	0	99.86%

is slightly higher than that of the case where the error term is a standardized normal distribution. The reason is that when the error term is heavy tailed, it is more appropriate to choose LLA, but the accuracy of estimation and prediction is slightly worse than that of Lasso. When the sample size increases, the LLA and Oracle estimates perform equally well in the selection of important variables and the complexity of the model.

As can be seen from Table 3, when the error term is set to a mixed normal distribution, the ability of the proposed method to correctly select zero variables is good. In the case of a small sample size, the ability of the Lasso method to select important variables is better.

From Tables 4–6 where we choose the Huber loss function, the LLA method we proposed behaves well both in variable selection and robustness. Compared with Table 1 and Table 4, when the data has outliers, we should choose LAD as the loss function. Moreover, when the error term follows a mixed normally distribution, the LLA method behaves bet-

Table 4 Simulation results for Huber loss function and $\varepsilon \sim N(0,1)$

Setting	Method	EE	PE	C	IC	CP
$n = 200$	Oracle	10.8300	3.3696	24.0000	0	100%
$p = 28$	Lasso	9.6422	3.5569	20.0920	0	83.72%
$m = 24$	LLA	10.9088	3.3784	22.7200	0	94.67%
$n = 500$	Oracle	19.9141	5.4034	41.0000	0	100%
$p = 45$	Lasso	18.0691	5.6068	38.0300	0	92.76%
$m = 41$	LLA	19.8884	5.3937	40.5160	0	98.82%
$n = 700$	Oracle	24.3265	6.3761	49.0000	0	100%
$p = 53$	Lasso	22.4030	6.5988	46.2440	0	94.38%
$m = 49$	LLA	24.3596	6.3882	48.6620	0	99.31%

Table 5 Simulation results for Huber loss function and $\varepsilon \sim t_5$

Setting	Method	EE	PE	C	IC	CP
$n = 200$	Oracle	10.5572	4.2666	24.0000	0	100%
$p = 28$	Lasso	9.2590	4.4065	18.4680	0.0020	76.95%
$m = 24$	LLA	10.6099	4.2429	22.8100	0	95.04%
$n = 500$	Oracle	19.4395	6.8118	41.0000	0	100%
$p = 45$	Lasso	17.4385	6.9993	36.2080	0	88.31%
$m = 41$	LLA	19.4471	6.8247	40.5440	0	98.89%
$n = 700$	Oracle	23.8089	8.0487	49.0000	0	100%
$p = 53$	Lasso	21.6534	8.2558	44.4980	0	90.81%
$m = 49$	LLA	23.8220	8.0807	48.6940	0	99.38%

Table 6 Simulation results for Huber loss function and $\varepsilon \sim 0.9N(0,1) + 0.1N(0,9)$

Setting	Method	EE	PE	C	IC	CP
$n = 200$	Oracle	10.4829	4.4694	24.0000	0	100%
$p = 28$	Lasso	9.1630	4.6333	18.0680	0	75.28%
$m = 24$	LLA	10.5706	4.4827	22.7880	0	94.95%
$n = 500$	Oracle	19.2618	7.1860	41.0000	0	100%
$p = 45$	Lasso	17.3780	7.3190	35.3500	0	86.22%
$m = 41$	LLA	19.2962	7.2029	40.5900	0	99.00%
$n = 700$	Oracle	23.6356	8.5563	49.0000	0	100%
$p = 53$	Lasso	21.5202	8.7148	43.4420	0	88.66%
$m = 49$	LLA	23.6275	8.5822	48.7120	0	99.41%

ter than the Lasso method. The reason for this is that the real data has a mixed normal distribution with high probability.

5 Conclusion

In this paper, we mainly study the M-estimation method for the high-dimensional linear regression model and discuss the properties of the M-estimator when the penalty term is the local linear approximation. We show that the proposed estimator possesses the good properties by applying certain assumptions. In the numerical simulation, we select the appropriate algorithm to show the good robustness of this method.

Funding
The work was supported by the NSFC (71803001, 61703001), the NSF of Anhui Province (1708085MA17, 1508085QA13), the Key NSF of Education Bureau of Anhui Province (KJ2018A0437) and the Support Plan of Excellent Youth Talents in Colleges and Universities in Anhui Province (gxyq2017011).

Competing interests
The authors declare that they have no competing interests.

Authors' contributions
All authors contributed equally to the writing of this paper. All authors read and approved the final manuscript.

Publisher's Note

Received: 16 March 2018 Accepted: 16 August 2018 Published online: 30 August 2018

References

1. Huber, P.: Robust estimation of a location parameter. Ann. Math. Stat. **35**, 73–101 (1964)
2. Huber, P.: Robust regression: asymptotics, conjectures and Monte Carlo. Ann. Stat. **1**, 799–821 (1973)
3. Huber, P.: Robust Statistics. Wiley, New York (1981)
4. Portnoy, S.: Asymptotic behavior of M-estimators of p regression parameters when p^2/n is large, I: consistency. Ann. Stat. **12**, 1298–1309 (1984)
5. Welsh, A.: On M-processes and M-estimation. Ann. Stat. **17**, 337–361 (1989)
6. Mammen, E.: Asymptotics with increasing dimension for robust regression with applicationsto the bootstrap. Ann. Stat. **17**, 382–400 (1989)
7. Bai, Z., Wu, Y.: Limiting behavior of M-estimators of regression coefficients in high dimensional linear models I. Scale-dependent case. J. Multivar. Anal. **51**, 211–239 (1994)
8. He, X., Shao, Q.: On parameters of increasing dimensions. J. Multivar. Anal. **73**, 120–135 (2000)
9. Li, G., Peng, H., Zhu, L.: Nonconcave penalized M-estimation with a diverging number of parameters. Stat. Sin. **21**, 391–419 (2011)
10. Zou, H., Li, R.: One-step sparse estimates in nonconcave penalized likelihood models. Ann. Stat. **36**, 1509–1566 (2008)
11. Bai, Z., Rao, C., Wu, Y.: M-estimation of multivariate linear regression parameters under a convex discrepancy function. Stat. Sin. **2**, 237–254 (1992)
12. Wu, W.: M-estimation of linear models with dependent errors. Ann. Stat. **35**, 495–521 (2007)
13. Huang, J., Horowitz, J., Ma, S.: Asymptotic properties of bridge estimators in sparse high-dimensional regression models. Ann. Stat. **36**, 587–613 (2008)
14. Fan, J., Peng, H.: Nonconcave penalized likelihood with a diverging number of parameters. Ann. Stat. **32**, 928–961 (2004)