Journal of Inequalities and Applications
a SpringerOpen Journal

RESEARCH Open Access

Adaptive bridge estimation for high-dimensional regression models

Zhihong Chen[1], Yanling Zhu[1*] and Chao Zhu[2]

*Correspondence:
zhuyanling99@126.com
[1] School of International Trade and Economics, University of International Business and Economics, Beijing, 100029, P.R. China
Full list of author information is available at the end of the article

Abstract

In high-dimensional models, the penalized method becomes an effective measure to select variables. We propose an adaptive bridge method and show its oracle property. The effectiveness of the proposed method is demonstrated by numerical results.

MSC: 62F12; 62E15; 62J05

Keywords: adaptive bridge; high-dimensionality; variable selection; oracle property; penalized method; tuning parameter

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, $X = (X_1, X_2, \ldots, X_p) = (x_1, x_2, \ldots, x_n)^T = (x_{ij})_{n\times p}$ is an $n \times p$ 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 $\beta \in \mathrm{R}^p$ when p 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

$$\hat{\beta} = \arg\min_{\beta} Q(\beta), \qquad Q(\beta) = \|Y - X\beta\|^2 + \lambda \sum_{j=1}^{p} |\beta_j|^{\zeta},$$

where $\|\cdot\|$ is the l_2 norm of the vector, λ is a tuning parameter. For $\zeta > 0$, $\hat{\beta}$ is called the bridge estimator proposed by Frank and Friedman [1]. There are two well-known special cases of the bridge estimator. If $\zeta = 2$, it is the ridge estimator in Hoerl and Kennard [2]; if $\zeta = 1$, it is the Lasso estimator by Tibshirani [3], which does not possess the oracle property in Fan and Li [4]. For $0 < \zeta \leq 1$, Knight and Fu [5] studied the asymptotic distributions of bridge estimators when the number of covariates is fixed and provided a theoretical justification for the use of bridge estimators to select variables. The bridge estimators can distinguish between the covariates whose coefficients are exactly zero and the covariates whose coefficients are nonzero. There is much statistical literature about penalization-based methods. Some examples include the SCAD by Fan and Li [4], the elastic net by Zou and Hastie [6], the adaptive lasso by Zou [7], the Dantzig selector by Candes and Tao [8] and the non-concave MCP penalty by Zhang [9]. For bridge estimation, Huang *et al.* [10] extended the results of Knight and Fu [5] to infinite dimensional parameters and showed that for $0 < \zeta < 1$ the bridge estimator can correctly select covariates with nonzero coefficients and under appropriate conditions the bridge estimator enjoys the oracle property. Subsequently, Wang *et al.* [11] studied the consistency of the bridge estimator for a generalized linear model.

In this paper, we consider the following penalized model:

$$\hat{\beta} = \arg\min_{\beta} Q(\beta), \quad Q(\beta) = \|Y - X\beta\|^2 + \lambda \sum_{j=1}^{p} \tilde{\omega}_j |\beta_j|^{\zeta}, \tag{1.1}$$

where $\tilde{\omega} = (\tilde{\omega}_1, \tilde{\omega}_2, \ldots, \tilde{\omega}_p)^T$ is a given vector of weights. Usually, if we let the initial estimator $\tilde{\beta} = (\tilde{\beta}_1, \tilde{\beta}_2, \ldots, \tilde{\beta}_p)^T$ be the non-penalized MLE, then $\tilde{\omega}_j = |\tilde{\beta}_j|^{-1}$, $j = 1, 2, \ldots, p$. $\hat{\beta}$ is called the adaptive bridge estimator. We propose and study the adaptive bridge estimator method (abridge for short). We derive some theoretical properties of the adaptive bridge estimator for the case when p can increase to infinity with n. Under some conditions, with the choice of the tuning parameter, we show that the adaptive bridge estimator enjoys the oracle property; that is, the adaptive bridge estimator can correctly select covariates with nonzero coefficients with probability converging to one and that the estimator of nonzero coefficients has the same asymptotic distribution that it would have if the zero coefficients were known in advance.

As far as we know, there is no literature to discuss the properties of an adaptive bridge, so our results make up for this. Compared with the results in Huang *et al.* [10] and Wang *et al.* [11], the condition (A_2) (see Section 2) imposed on the true coefficients is much weaker. Moreover, in Wang *et al.* [11] one needs the true coefficients to meet the additional condition called covering number. Besides, Huang *et al.* [10] and Wang *et al.* [11] both use the LQA algorithm to obtain the estimator. The shortcoming of the LQA algorithm is that if we delete one variable in some step of the iteration, this variable will have no chance to appear in the final model. In order to improve this algorithm, we employ the MM algorithm to improve the stability.

The rest of the paper is organized as follows. In Section 2, we introduce notations and assumptions which will be needed in the our results and present the main results. Section 3 presents some simulation results. The conclusion and the proofs of the main results are arranged in Sections 4 and 5.

2 Main results

For convenience of the statement, we first give some notations. Let $\beta_0 = (\beta_{01}, \beta_{02}, \ldots, \beta_{0p})^T$ be the true parameter, $J_1 = \{j : \beta_{0j} \neq 0, j = 1, 2, \ldots, p\}$, $J_2 = \{j : \beta_{0j} = 0, j = 1, 2, \ldots, p\}$, the cardinality of the set J_1 is denoted by q and $h_1 = \min\{|\beta_{0j}| : j \in J_1\}$. Without loss of generality, we assume that the first q coefficients of covariates (denoted by $X_{(1)}$) are nonzero, $X_{(2)}$ be covariates with zero coefficients, $\beta_0 = (\beta_{0(1)}^T, \beta_{0(2)}^T)^T$, $\hat{\beta} = (\hat{\beta}_{(1)}^T, \hat{\beta}_{(2)}^T)^T$ correspondingly. Actually, p, q, X, Y, β, and λ are related to the sample size n, we omit n for convenience. In this paper, we only consider the statistical properties of the adaptive bridge for the case of $p < n$; consequently we put $p = O(n^{c_2})$, $q = O(n^{c_1})$, $\lambda = O(n^{-\delta})$, where $0 \leq c_1 < c_2 < 1$, $\delta > 0$. Here we use the terminology in Zhao and Yu [12] , and we define $\hat{\beta} =_s \beta_0$ if and only if $\operatorname{sgn}(\hat{\beta}) = \operatorname{sgn}(\beta_0)$, where we denote the sign of a $p \times 1$ vector β as $\operatorname{sgn}(\beta) = (\operatorname{sgn}(\beta_1), \operatorname{sgn}(\beta_2), \ldots, \operatorname{sgn}(\beta_p))^T$. For any 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 $D = \left(\begin{smallmatrix} D_{11} & D_{12} \\ D_{21} & D_{22} \end{smallmatrix}\right)$, where $D_{11} = \frac{1}{n}X_{(1)}^T X_{(1)}$.

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

(A_1) The error term ε is i.i.d. with $E(\varepsilon) = 0$ and $E(\varepsilon^{2k}) < +\infty$, where $k > 0$. For the special case we denote $E(\varepsilon^2) = \sigma^2$.

(A_2) There exists a positive constant M such that $h_1 \geq Mn^{\alpha}$, where $\max\{-\frac{1}{2}, \frac{c_2-1}{2}, \frac{-1}{2-\zeta}\} < \alpha < \min\{c_2 - \delta, \frac{c_2-\delta-\zeta}{1+\zeta}\}$ and $\delta + \alpha + \frac{1}{2}\zeta < c_2$.

(A_3) Suppose τ_1 and τ_2 are the minimum and maximum eigenvalues of the matrix D_{11}. There exist constants τ_{10} and τ_{20} such that $0 < \tau_{10} \leq \tau_1 \leq \tau_2 \leq \tau_{20}$, and the eigenvalues of $\frac{1}{n}X^T \operatorname{var}(Y)X$ are bounded.

(A_4) Let g_i be the transpose of the ith row vector of $X_{(1)}$, such that $\lim_{n\to\infty} n^{-\frac{1}{2}} \max_{1\leq i\leq n} g_i^T \times g_i = 0$.

It is worth mentioning that condition (A_1) is much weaker than those in the literature where it is commonly assumed that the error term has Gaussian tail probability distribution. In this paper we allow ε to have a heavy tail. The regularity condition (A_2) is a common assumption for the nonzero coefficients, which can ensure that all important covariates could be included in the finally selected model. Condition (A_3) means that the matrix $\frac{1}{n}X_{(1)}^T X_{(1)}$ is strictly positive definite. For condition (A_4), we will use it to prove the asymptotic normality of the estimators of the nonzero coefficients. In fact, if the nonzero coefficients have an upper bound, then we can easily verify condition (A_4).

2.1 Consistency of the estimation

Theorem 2.1 (Consistency of the estimation) *If* $0 < \zeta < 2$, *and conditions* (A_1)-(A_3) *hold, then there exists a local minimizer* $\hat{\beta}$ *of* $Q(\beta)$, *such that* $\|\hat{\beta} - \beta_0\| = O_p(n^{\frac{\delta+\alpha-c_2}{\zeta}})$.

Remark 2.1 By condition (A_2), we know that $c_2 - \delta - \alpha > 0$ and the estimator consistency refers to the order of sample size and tuning parameter. Theorem 2.1 extends the previous results.

2.2 Oracle property of the estimation

Theorem 2.2 (Oracle property) *If* $0 < \zeta < 1$, *and conditions* (A_1)-(A_4) *hold, then the adaptive bridge estimator satisfies the following properties.*

(1) (*Selection consistency*) $\lim_{n\to\infty} P\{\hat{\beta} =_s \beta_0\} = 1$;

(2) (*Asymptotic normality*) $\sqrt{n}s^{-1}u^T(\hat{\beta}_{(1)} - \beta_{0(1)}) \xrightarrow{\mathrm{d}} N(0,1)$, *where* $s^2 = \sigma^2 u^T D_{11}^{-1} u$ *for any* $q \times 1$ *vector* u *and* $\|u\| \leq 1$.

Remark 2.2 By Theorems 2.1 and 2.2, we can easily see that the adaptive bridge is able to consistently identify the true model.

3 Simulation results

In this section we evaluate the performance of the adaptive bridge estimator proposed in (1.1) by simulation studies. Set $\zeta = 1/2$ and simulate the data by the model $Y = X\beta + \varepsilon$, $\varepsilon \sim N(0, \sigma^2)$, where $\sigma = 1$, $\beta_{0(1)} = (-2.5, -2.5, -2.5, 3, 3, 3, -1, -1, -1)^T$. The design matrix X is generated by a p-dimensional multivariate normal distribution with mean zero and a covariance matrix whose (i,j)th component is $\rho^{|i-j|}$, where we let $\rho = 0.5$ and 0.9, respectively. The following examples are considered.

Example 3.1 The sample size $n = 200$ and the covariates number $p = 50$.

Example 3.2 The sample size $n = 500$ and the covariates number $p = 80$.

Example 3.3 The sample size $n = 800$ and the covariates number $p = 100$.

We connect the minorization-maximization (MM) algorithm by Hunter and Li [13] and the Newton-Raphson method to estimate the adaptive abridge (abridge), where the tuning parameter is selected by 5-fold cross-validation. Meanwhile, we compare our results with that from lasso [14], adaptive lasso (alasso) and bridge methods. In order to evaluate the performance of the estimators, we select four measures called L_2-loss, PE, C, and IC. L_2-loss is median of $\|\hat{\beta} - \beta_0\|_2$ to evaluate the estimation accuracy, and PE is the prediction error defined by median of $n^{-1}\|Y - X\hat{\beta}\|^2$. The other two 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. The numerical results are listed in Table 1 and Table 2, where υ equals the number of zero coefficients in the true model and the numbers in parentheses are the corresponding standard deviations which are obtained by 500 replicates.

Note that in every case the adaptive bridge outperforms the other methods in sparsity, which can select the smaller model. For the adaptive bridge the prediction error is a little higher than the other methods, but when consider the estimation accuracy, the adaptive

Table 1 Simulation results for $\rho = 0.5$

Setting	Method	L_2-loss	PE	C	IC
$n = 200$	Lasso	0.5459 (0.1160)	0.8830 (0.1126)	28.4540 (5.3337)	0 (0)
$p = 50$	Alasso	0.5442 (0.1149)	0.8790 (0.1099)	28.5300 (5.2421)	0 (0)
$\upsilon = 41$	Bridge	0.4733 (0.1120)	0.8755 (0.1068)	28.2700 (5.0834)	0 (0)
	Abridge	0.4617 (0.1155)	0.9005 (0.1038)	38.8380 (2.9903)	0 (0)
$n = 500$	Lasso	0.3459 (0.0829)	0.9469 (0.0673)	56.1300 (6.3329)	0 (0)
$p = 80$	Alasso	0.3476 (0.0830)	0.9465 (0.0686)	56.0360 (6.5269)	0 (0)
$\upsilon = 71$	Bridge	0.2950 (0.0732)	0.9394 (0.0654)	52.7140 (6.7155)	0 (0)
	Abridge	0.2745 (0.0728)	0.9661 (0.0630)	69.7160 (2.5994)	0 (0)
$n = 800$	Lasso	0.2814 (0.0624)	0.9664 (0.0548)	74.4820 (7.1596)	0 (0)
$p = 100$	Alasso	0.2817 (0.0620)	0.9687 (0.0552)	74.7180 (7.0409)	0 (0)
$\upsilon = 91$	Bridge	0.2327 (0.0576)	0.9570 (0.0534)	69.4380 (8.8332)	0 (0)
	Abridge	0.2160 (0.0569)	0.9839 (0.0514)	89.7680 (3.0359)	0 (0)

Table 2 Simulation results for $\rho = 0.9$

Setting	Method	L_2-loss	PE	C	IC
$n = 200$	Lasso	1.0102 (0.2452)	0.8725 (0.1049)	27.2580 (4.6677)	0.0040 (0.0632)
$p = 50$	Alasso	1.0123 (0.2475)	0.8800 (0.1024)	27.2460 (4.5851)	0.0040 (0.0632)
$\upsilon = 41$	Bridge	0.8961 (0.2624)	0.8656 (0.1059)	25.0600 (5.3104)	0 (0)
	Abridge	0.8468 (0.2843)	0.8965 (0.1092)	37.7800 (4.2298)	0.0260 (0.1593)
$n = 500$	Lasso	0.6649 (0.1630)	0.9435 (0.0664)	52.9840 (5.7272)	0 (0)
$p = 80$	Alasso	0.6671 (0.1620)	0.9388 (0.0667)	52.3420 (6.1499)	0 (0)
$\upsilon = 71$	Bridge	0.5251 (0.1442)	0.9368 (0.0649)	50.0820 (7.7934)	0 (0)
	Abridge	0.4837 (0.1377)	0.9646 (0.0651)	68.2320 (4.7732)	0 (0)
$n = 800$	Lasso	0.5382 (0.1242)	0.9623 (0.0545)	70.4900 (6.5000)	0 (0)
$p = 100$	Alasso	0.5371 (0.1259)	0.9614 (0.0544)	69.9680 (6.9009)	0 (0)
$\upsilon = 91$	Bridge	0.4126 (0.1183)	0.9572 (0.0541)	66.6060 (9.6239)	0 (0)
	Abridge	0.3580 (0.1087)	0.9818 (0.0520)	89.2840 (4.3161)	0 (0)

bridge is still the winner, followed by bridge. We also find the interesting fact that with the sample size n larger, the performance of correctly selecting the zero covariates for the adaptive bridge is better whenever $\rho = 0.5$ or 0.9. Meanwhile with n increasing, the estimation accuracy performs better, but the prediction error is worse. Additionally, when ρ increases, the prediction error increases, but the estimation accuracy decreases.

4 Conclusion

In this paper we have proposed the adaptive bridge estimator and presented some theoretical properties of the adaptive bridge estimator. Under some conditions, with the choice of the tuning parameter, we have showed that the adaptive bridge estimator enjoys the oracle property. The effectiveness of the proposed method is demonstrated by numerical results.

5 Proofs

Proof of Theorem 2.1 In view of the idea in Fan and Li [4], we only need to prove that, for any $\epsilon > 0$, there exists a large constant C such that

$$\liminf_{n\to\infty} P\left\{\inf_{\|u\|=C} Q(\beta_0 + \theta u) > Q(\beta_0)\right\} \geq 1 - \epsilon, \tag{5.1}$$

which means that with a probability of at least $1 - \epsilon$ there exists a local minimizer $\hat{\beta}$ in the ball $\{\beta_0 + \theta u : \|u\| \leq C\}$.

First, let $\theta = n^{\frac{\delta+\alpha-c_2}{\zeta}}$, then

$$\begin{aligned}
&Q(\beta_0 + \theta u) - Q(\beta_0) \\
&\quad = \theta^2 n u^T \left(\frac{X^T X}{n}\right) u - \theta u^T X^T (Y - X\beta_0) + \lambda \sum_{j=1}^{p} \tilde{\omega}_j \bigl(|\beta_{0j} + \theta u|^\zeta - |\beta_{0j}|^\zeta\bigr) \\
&\quad \geq \lambda_{\min}\left(\frac{X^T X}{n}\right) \theta^2 n \|u\|^2 - n\theta u^T \frac{X^T (Y - X\beta_0)}{n} - \lambda \sum_{j=1}^{p} \tilde{\omega}_j |\theta|^\zeta \|u\|^\zeta \\
&\quad := T_1 + T_2 + T_3,
\end{aligned} \tag{5.2}$$

where $T_1 = \lambda_{\min}(\frac{X^T X}{n})\theta^2 n\|u\|^2$, $T_2 = -n\theta u^T \frac{X^T(Y-X\beta_0)}{n}$, and $T_3 = -\lambda \sum_{j=1}^{p} \tilde{\omega}_j |\theta|^\zeta \|u\|^\zeta$.

For T_2, set $\nu = O_P(n^\alpha)$ and by assumptions (A_2) and (A_3) we have

$$\begin{aligned}
P\left\{\left\|\frac{X^T (Y - X\beta_0)}{n}\right\| \geq M\nu\right\} &\leq \frac{1}{M^2 \nu^2} E\left[\sum_{j=1}^{p} \left(\frac{1}{n} X_j^T (Y - X\beta_0)\right)^2\right] \\
&= \frac{1}{nM^2 \nu^2} \operatorname{tr}\left(\frac{1}{n} X^T \operatorname{var}(Y) X\right) \to 0 \quad (n \to \infty).
\end{aligned}$$

Hence

$$\begin{aligned}
|T_2| &= \left\| n\theta u^T \frac{X^T (Y - X\beta_0)}{n}\right\| \leq n|\theta| \left\|\frac{X^T (Y - X\beta_0)}{n}\right\| \|u\| \\
&= n|\theta| O_P(\nu) \|u\| = o_P(1)\|u\|.
\end{aligned} \tag{5.3}$$

As for $|T_3|$, observe that $\|\tilde{\beta} - \beta_0\| = O_P((\frac{p}{n})^{1/2})$, $\min|\beta_j| \le \max|\tilde{\beta}_j - \beta_{0j}| + \min|\tilde{\beta}_j|$, and assumption (A_2), we can obtain

$$\begin{aligned} M &\le n^{-\alpha}\min|\beta_j| \le n^{-\alpha}\max|\tilde{\beta}_j - \beta_{0j}| + n^{-\alpha}\min|\tilde{\beta}_j| \\ &= n^{-\alpha}O_P\left(\left(\frac{p}{n}\right)^{1/2}\right) + n^{-\alpha}\min|\tilde{\beta}_j|. \end{aligned}$$

This together with assumption (A_2) yields $P\{\min|\tilde{\beta}_j| \ge \frac{1}{2}Mn^{\alpha}\} \to 1$ $(n \to \infty)$.

For $v_1 = \frac{2\lambda p}{Mn^{\alpha}}$, $P\{\lambda\sum_{j=1}^{p}\tilde{\omega}_j| \le v_1\} \ge P\{\frac{\lambda p}{\min|\tilde{\beta}_j|} \le v_1\} = P\{\min|\tilde{\beta}_j| \ge \frac{\lambda p}{v_1}\} \to 1$ $(n \to \infty)$, *i.e.*, $|\lambda\sum_{j=1}^{p}\tilde{\omega}_j| = O_P(\frac{\lambda p}{Mn^{\alpha}})$. Now with assumption (A_2) we conclude that

$$|T_3| = O_P\left(\frac{\lambda p}{Mn^{\alpha}}\right)|\theta|^{\zeta}\|u\|^{\zeta} = O_P(1)\|u\|^{\zeta}. \tag{5.4}$$

When $0 < \zeta < 2$ and C is large enough, by (5.3) and (5.4) we see that (5.2) is determined by T_1, so (5.1) holds. □

Proof of Theorem 2.2 (1) First of all, by the K-K-T condition we know that $\hat{\beta}$ is the defined adaptive bridge estimator, if the following holds:

$$\begin{cases} \frac{\partial\|Y-X\beta\|^2}{\partial\beta_j}|_{\beta_j=\hat{\beta}_j} = \lambda\zeta\tilde{\omega}_j|\hat{\beta}_j|^{\zeta-1}\operatorname{sgn}(\hat{\beta}_j), \quad \hat{\beta}_j \ne 0, \\ \frac{\partial\|Y-X\beta\|^2}{\partial\beta_j}|_{\beta_j=\hat{\beta}_j} \le \lambda\zeta\tilde{\omega}_j|\hat{\beta}_j|^{\zeta-1}, \quad \hat{\beta}_j = 0. \end{cases} \tag{5.5}$$

Let $\hat{u} = \hat{\beta} - \beta_0$ and define $V(u) = \sum_{j=1}^{n}(\varepsilon_i - X_i^T u)^2 + \lambda\sum_{j=1}^{p}\tilde{\omega}_j|u_j + \beta_{0j}|^{\zeta}$, then we obtain $\hat{u} = \arg\min_u V(u)$. Notice that $\sum_{j=1}^{n}(\varepsilon_i - X_i^T u)^2 = -2\varepsilon^T Xu + nu^T Du + \varepsilon^T\varepsilon$, which yields $\frac{d[\sum_{j=1}^{n}(\varepsilon_i - X_i^T u)^2]}{du}|_{u=\hat{u}} = -2X^T\varepsilon + 2nD\hat{u} := 2\sqrt{n}[D(\sqrt{n}\hat{u}) - E]$, where $E = \frac{X^T\varepsilon}{\sqrt{n}}$. Together with (5.5) and the fact $\{|\hat{u}_{(1)}| < |\beta_{0(1)}|\} \subset \{\operatorname{sgn}(\hat{\beta}_{(1)}) = \operatorname{sgn}(\beta_{0(1)})\}$, if $\hat{u}$ satisfies

$$D_{11}\sqrt{n}\hat{u}_{(1)} - E_{(1)} = \frac{-\lambda}{2\sqrt{n}}\zeta\bar{W}_{(1)} \quad \text{and} \quad |\hat{u}_{(1)}| < |\beta_{0(1)}|,$$

where $\bar{W} = (\tilde{\omega}_1|\hat{u}_{(1)} + \beta_{01}|^{\zeta-1}\operatorname{sgn}(\beta_{01}), \tilde{\omega}_2|\hat{u}_{(1)} + \beta_{02}|^{\zeta-1}\operatorname{sgn}(\beta_{02}), \dots, \tilde{\omega}_p|\hat{u}_{(1)} + \beta_{0p}|^{\zeta-1} \times \operatorname{sgn}(\beta_{0p}))^T$, then we have $\operatorname{sgn}(\hat{\beta}_{(1)}) = \operatorname{sgn}(\beta_{0(1)})$ and $\hat{\beta}_{(2)} = 0$. Let

$$\tilde{W} = \left(2\tilde{\omega}_1|\beta_{01}|^{\zeta-1}\operatorname{sgn}(\beta_{01}), 2\tilde{\omega}_2|\beta_{02}|^{\zeta-1}\operatorname{sgn}(\beta_{02}), \dots, 2\tilde{\omega}_p|\beta_{0p}|^{\zeta-1}\operatorname{sgn}(\beta_{0p})\right)^T,$$

it follows that $|D_{11}^{-1}E_{(1)}| + \frac{\lambda\zeta}{2\sqrt{n}}|D_{11}^{-1}\tilde{W}_{(1)}| < \sqrt{n}|\beta_{0(1)}|$. Denote $A = \{|D_{11}^{-1}|E_{(1)} + \frac{\lambda\zeta}{2\sqrt{n}}|D_{11}^{-1}\tilde{W}_{(1)}| < \sqrt{n}|\beta_{0(1)}|\}$, we conclude that $P\{\operatorname{sgn}(\hat{\beta}) = \operatorname{sgn}(\beta_0)\} \ge P\{A\}$, from which it follows that

$$\begin{aligned} P\{\operatorname{sgn}(\hat{\beta}) \ne \operatorname{sgn}(\beta_0)\} &\le P\{A^c\} \\ &\le P\left\{|\xi_i| \ge \frac{1}{2}\sqrt{n}|\beta_{0i}|, \exists i \in J_1\right\} \\ &\quad + P\left\{\frac{\lambda\zeta}{n}|Z_i| > |\beta_{0i}|, \exists i \in J_1\right\} := I_1 + I_2, \end{aligned} \tag{5.6}$$

where $\xi = (\xi_1, \xi_2, \dots, \xi_q)^T = D_{11}^{-1} E_{(1)}$, $Z = (Z_1, Z_2, \dots, Z_q)^T = D_{11}^{-1} W_{(1)}$. For $I_1 = P\{|\xi_i| \geq \frac{1}{2}\sqrt{n}|\beta_{0i}|, \exists i \in J_1\}$, then $E[(\xi_i)^{2k}] < \infty$, $\forall i \in J_1$. So its tail probability satisfies $P\{|\xi_i| > t\} = O(t^{-2k})$, $\forall t > 0$, which yields

$$I_1 \leq P\left\{|\xi_i| \geq \frac{1}{2}\sqrt{n}h_1, \exists i \in J_1\right\} = qO\left(\left(\frac{1}{2}\sqrt{n}h_1\right)^{-2k}\right) \to 0 \quad (n \to \infty). \tag{5.7}$$

For I_2, notice that $1 - I_2 = P\{\frac{\lambda\zeta}{n}|Z_i| \leq |\beta_{0i}|, \exists i \in J_1\}$ and $|Z_i| \leq \|D_{11}^{-1}\tilde{W}_{(1)}\| \leq \frac{1}{\tau_1}\|\tilde{W}_{(1)}\| \leq \frac{2\sqrt{q}h_1^{\zeta-1}}{\tau_1 \min_{j\in J_1}|\tilde{\beta}_j|}$, then we can get

$$1 - I_2 \geq P\left\{\frac{2\sqrt{q}\lambda\zeta h_1^{\zeta-1}}{\tau_1 \min_{j\in J_1}|\tilde{\beta}_j|} \leq nh_1\right\} = P\left\{\lambda\zeta \leq \frac{n\tau_1 h_1^{2-\zeta}}{2\sqrt{q}} \min_{j\in J_1}|\tilde{\beta}_j|\right\} = 1 + O_p(1) \quad (n \to \infty).$$

This follows that $I_2 \to 0$ $(n \to \infty)$. Together with (5.6) and (5.7), $\lim_{n\to\infty} P\{\hat{\beta} =_s \beta_0\} = 1$ holds. This completes the proof of the first part of Theorem 2.2.

(2) Let $W = (\tilde{\omega}_1|\hat{\beta}_1|^{\zeta-1}\operatorname{sgn}(\hat{\beta}_1), \tilde{\omega}_2|\hat{\beta}_2|^{\zeta-1}\operatorname{sgn}(\hat{\beta}_2), \dots, \tilde{\omega}_p|\hat{\beta}_p|^{\zeta-1}\operatorname{sgn}(\hat{\beta}_p))^T$, we can easily get $\frac{\partial\|Y-X\beta\|^2}{\partial\beta_j}|_{\beta=\hat{\beta}} = 0$, $j \in J_1$, *i.e.*, $X_{(1)}^T(Y - X_{(1)}\hat{\beta}_{(1)}) = X_{(1)}^T(X_{(1)}\beta_{0(1)} - X_{(1)}\hat{\beta}_{(1)} + \varepsilon) = \frac{\lambda\zeta}{2}W_{(1)}$, which yields $D_{11}(\hat{\beta}_{(1)} - \beta_{0(1)}) = \frac{X_{(1)}^T\varepsilon}{n} - \frac{\lambda\zeta}{2n}W_{(1)}$. It follows from the first part of Theorem 2.2 that $\lim_{n\to\infty} P\{D_{11}(\hat{\beta}_{(1)} - \beta_{0(1)}) = \frac{X_{(1)}^T\varepsilon}{n} - \frac{\lambda\zeta}{2n}W_{(1)}\} = 1$, then we can see that, for any $q \times 1$ vector u and $\|u\| \leq 1$,

$$\sqrt{n}u^T(\hat{\beta}_{(1)} - \beta_{0(1)}) = n^{-1/2}u^T D_{11}^{-1} X_{(1)}^T \varepsilon - \frac{\lambda\zeta}{2\sqrt{n}}u^T D_{11}^{-1} W_{(1)} + O_P(1). \tag{5.8}$$

Notice that

$$\begin{aligned} \left|\frac{\lambda\zeta}{2\sqrt{n}}u^T D_{11}^{-1} W_{(1)}\right| &\leq \frac{\lambda\zeta}{2\sqrt{n}\tau_1}\frac{\|\hat{\beta}_{(1)}\|^{\zeta-1}}{\min_{j\in J_1}|\tilde{\beta}_j|} \leq \frac{\lambda\zeta M_1^{\zeta-1} q^{\frac{\zeta-1}{2}} n^{\alpha(\zeta-2)-\frac{1}{2}}}{2^{\zeta-1}M\tau_1} \\ &= O\bigl(n^{\frac{1}{2}c_1(\zeta-1)+\alpha(\zeta-2)-\frac{1}{2}}\bigr), \end{aligned}$$

where the second inequality holds because $P\{\min_{j\in J_1}|\hat{\beta}_j| \geq \frac{1}{2}M_1 n^{\alpha}\} \to 1$ $(n \to \infty)$, for $M_1 > 0$. By $\frac{1}{2}c_1(\zeta - 1) + \alpha(\zeta - 2) - \frac{1}{2} < 0$, we obtain $|\frac{\lambda\zeta}{2\sqrt{n}}u^T D_{11}^{-1} W_{(1)}| = o_P(1)$, which together with (5.8) yields

$$\sqrt{n}u^T(\hat{\beta}_{(1)} - \beta_{0(1)}) = n^{-1/2}u^T D_{11}^{-1} X_{(1)}^T \varepsilon + o_P(1). \tag{5.9}$$

Denote $s^2 = \sigma^2 u^T D_{11}^{-1} u$ and $F_i = n^{-\frac{1}{2}} s^{-1} u^T D_{11}^{-1} g_i^T$, by assumption (A_4) and (5.9) we have $\sqrt{n}s^{-1}u^T(\hat{\beta}_{(1)} - \beta_{0(1)}) = \sum_{i=1}^n F_i\varepsilon_i + o_P(1) \xrightarrow{d} N(0,1)$. This completes the proof of the second part of Theorem 2.2. □

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.

Author details

[1] School of International Trade and Economics, University of International Business and Economics, Beijing, 100029, P.R. China. [2] Department of Mathematical Science, University of Wisconsin at Milwaukee, Milwaukee, WI 53201, USA.

Acknowledgements
The research was supported by the NSF of Anhui Province (No. 1508085QA13) and the China Scholarship Council.

Received: 16 August 2016 Accepted: 12 October 2016 Published online: 20 October 2016

References

1. Frank, IE, Friedman, JH: A statistical view of some chemometrics regression tools. Technometrics **35**, 109-148 (1993) (with discussion)
2. Hoerl, AE, Kennard, RW: Ridge regression: biased estimation for nonorthogonal problems. Technometrics **12**, 55-67 (1970)
3. Tibshirani, R: The lasso method for variable selection in the Cox model. Stat. Med. **16**, 385-395 (1997)
4. Fan, J, Li, R: Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. **96**, 1348-1360 (2001)
5. Knight, K, Fu, WJ: Asymptotics for lasso-type estimators. Ann. Stat. **28**, 1356-1378 (2000)
6. Zou, H, Hastie, T: Regularization and variable selection via the elastic net. J. R. Stat. Soc., Ser. B **67**, 301-320 (2005)
7. Zou, H: The adaptive lasso and its oracle properties. J. Am. Stat. Assoc. **101**, 1418-1429 (2006)
8. Candes, E, Tao, T: The Dantzig selector: statistical estimation when p is much larger than n. Ann. Stat. **35**, 2313-2351 (2007)
9. Zhang, C: Nearly unbiased variable selection under minimax concave penalty. Ann. Stat. **38**, 894-942 (2010)
10. Huang, J, Ma, S, Zhang, CH: Adaptive lasso for sparse high-dimensional regression models. Stat. Sin. **18**, 1603-1618 (2008)
11. Wang, M, Song, L, Wang, X: Bridge estimation for generalized linear models with a diverging number of parameters. Stat. Probab. Lett. **80**, 1584-1596 (2010)
12. Zhao, P, Yu, B: On model selection consistency of lasso. J. Mach. Learn. Res. **7**, 2541-2563 (2006)
13. Hunter, DR, Li, R: Variable selection using MM algorithms. Ann. Stat. **33**, 1617-1642 (2005)
14. Efron, B, Hastie, T, Johnstone, I, Tibshirani, R: Least angle regression. Ann. Stat. **32**, 407-499 (2004) (with discussion)