Hacettepe Journal of Mathematics and Statistics
Volume 47 (2) (2018), 423 – 436

SCAD-Ridge penalized likelihood estimators for ultra-high dimensional models

Ying Dong*, Lixin Song†, and Muhammad Amin‡ §

Abstract

Extraction of as much information as possible from huge data is a burning issue in the modern statistics due to more variables as compared to observations therefore penalization has been employed to resolve that kind of issues. Many achievements have already been made by such penalization techniques. Due to the large number of variables in many research areas declare it a high dimensional problem and with this the sample correlation becomes very large. In this paper, we studied the maximum likelihood estimation of variable selection under smoothly clipped absolute deviation (SCAD) and Ridge penalties with ultra-high dimension settings to solve this problem. We established the oracle property of the proposed model under some conditions by following the theoretical method of Kown and Kim (2012) [19]. These result can greatly broaden the application scope of high-dimension data. Numerical studies are discussed to assess the performance of the proposed method. The SCAD-Ridge given better results than the Lasso, Enet and SCAD.

Keywords: Maximum likelihood estimation; Oracle property; SCAD-Ridge penalization; Ultra-high dimension; Variable selection.

2000 AMS Classification: AMS 62G20, AMS 62H12, AMS 62J07

Received : 09.10.2015 *Accepted* : 28.04.2016 *Doi* : 10.15672/HJMS.201612518375

*Faculty of Science, Dalian Nationalities University, Dalian, 116600, P.R.China, Email: dongying@dlnu.edu.cn

†School of Mathematical Sciences, Dalian University of Technology, Dalian, 116023, P.R.China, Email: lxsong@dlut.edu.cn

‡School of Mathematical Sciences, Dalian University of Technology, Dalian, 116023, P.R.China and Nuclear Institute for Food and Agriculture (NIFA), 446, Peshawar, Pakistan, Email: aminkanju@gmail.com

§Corresponding Author.

1. Introduction

Extraction of as much information as possible from huge data is becoming a burning issue in the modern statistics due to more variables as compared to observations. Therefore, penalization has been employed to resolve such problems. A lot of achievements have already been made through penalized variable selection techniques such as the L_1 penalty in Donoho and Johnstone (1994) [5] that yields the soft threshold rule and leads to the least absolute shrinkage and selection operator (Lasso) discussed by Tibshirani (1996) [23], and was further studied by Efron et al.(2004)[6]. The L_2 penalty that results in Hoerl and Kennard's (1970) [7], was further discussed by Segerstedt (1992) [22]. It performs well when the predictors are highly correlated. Fu (1998) [15] proposed the L_q penalty for the bridge regression. The hard thresholding penalty function which results in the hard thresholding rule was studied by Antoniadis (1997) [2] and Fan (1997) [8]. Fan and Li (2001) [9] proposed the smoothly clipped absolute deviation (SCAD) penalty that enjoys the oracle property. The SCAD estimator performs like the oracle estimator obtained under the true model. These methods had been proposed for variable selection and estimation simultaneously.

It is obvious that ridge regression provides better results in case of highly correlated group variables. Zou and Hastie (2005) [26] proposed the elastic net (Enet), which is the the combination of L_1 and L_2 penalties. But the Enet estimator is asymptotically biased because of the L_1 component in the penalty. So Zou and Zhang (2009) [27] proposed the adaptive Enet (AEnet) estimator and proved that it is oracle under sufficient conditions. The nonconcave penalized thresholding estimators may enjoy the nice oracle property therefore Wang et al.(2010) [24] proposed a new combined-penalization (abbreviated as CP, combined SCAD with ridge) in linear regression. Dong et al.(2014) [4] extended the results of Wang et al.(2010) [24] to the general models and the general nonconcave penalization with a diverging number of parameters. Their results include the case of highly correlated predictors and are applicable to the situations when $p > n$. Amin et al. (2015)[1] also studied the similar idea of combined penalization with quantile regression settings for high dimensional models. Recent years the high dimensional data analysis has gained too much importance, therefore, there is a need to develop methods that are applicable to $p \geq n$ regression problems with highly correlated predictors and having the oracle property. Zhao and Yu (2006) [25], Meishausen and Buhlmann (2006) [20] proved the sign consistency of the Lasso when the number of parameters exceeds the sample size. The sure independence screening method (a type of correlation learning) was proposed by Fan and Lv (2008) [11] for ultra high-dimension model selection problems. For a detailed introduction of recent developments in high-dimensional variable selection was given in Fan and Lv (2010) [12]. Huang et al.(2010) [17] proposed the Mnet penalty method, which is combined by MCP and ridge. This penalty uses the MCP instead of the L_1 penalty for selection as compared to Enet.

Inspired by these methods, we propose the maximum likelihood estimation of variable selection under the SCAD-Ridge penalty in the ultra-high dimension. Following the theoretical method of Kown and Kim (2012) [19] we can get the nice property of our model. The SCAD-Ridge penalty can encourage a grouping effect in selection meaning that it selects or drops highly correlated predictors together. These result can greatly broaden the application scope of high-dimension data.

The contents of this article are organized as follows. In Section 2, we introduce the SCAD-Ridge penalized likelihood estimators for ultra-high dimensional models. Section 3 presents some regularized conditions and asymptotic results for the proposed model. Numerical studies are discussed in Section 4 to assess the performance of the proposed method. In Section 5, we discuss conclusions and give some recommendations for future work. Theoretical proofs are provided in the Appendix A.

2. The model of SCAD-Ridge penalized likelihood for ultra-high dimension

For any n, let $\boldsymbol{Z}_{ni} = (\boldsymbol{X}_{ni}, Y_i)$ $(i = 1, \ldots, n)$, be independent and identically distributed random variables with the probability density $f_n(\boldsymbol{Z}_{ni}, \boldsymbol{\theta}_n)$, where $\boldsymbol{\theta}_n \in \Theta_n$ and Θ_n is an open subset of $\Re^n$.

The maximum likelihood estimator of SCAD-Ridge penalty $\boldsymbol{\theta}_n$ is the maximum points of $Q_n(\boldsymbol{\theta}_n)$.

$$Q_n(\boldsymbol{\theta}_n) = L_n(\boldsymbol{\theta}_n) - n \sum_{j=1}^{p_n} J_{\lambda_n,\gamma_n}(|\theta_{nj}|), \tag{2.1}$$

where $L_n(\boldsymbol{\theta}_n) = \sum\limits_{i=1}^{n} \log f_n(\boldsymbol{Z}_{ni}, \boldsymbol{\theta}_n)$, $J_{\lambda_n,\gamma_n}(|\theta_{nj}|) = \frac{\gamma_n}{2}\theta_{nj}^2 + P_{\lambda_n}(|\theta_{nj}|)$. The $\lambda_n \geq 0, \gamma_n \geq 0$, they are the tuning parameters. $P_{\lambda_n}(\cdot)$ is the SCAD penalty which proposed by Fan and Li(2001) [9], the concrete form of $P_{\lambda_n}(\cdot)$ is

$$P_{\lambda_n}(\theta) = \begin{cases} \lambda_n|\theta|, & \text{where } 0 \leq |\theta| < \lambda_n, \\ \dfrac{a\lambda_n|\theta| - (\theta^2 + \lambda^2)/2}{(a-1)}, & \text{where } \lambda_n \leq |\theta| < a\lambda_n, \\ \dfrac{(a+1)\lambda_n^2}{2}, & \text{where } |\theta| \geq a\lambda_n. \end{cases}$$

where $a > 2$. The first derivative of $P_{\lambda_n}(\cdot)$ is

$$P'_{\lambda_n}(\theta) = \begin{cases} \lambda_n sgn(\theta), & \text{where } 0 \leq \theta < \lambda_n, \\ \dfrac{a\lambda_n sgn(\theta) - \theta}{a-1}, & \text{where } \lambda_n \leq \theta < a\lambda_n. \end{cases}$$

The SCAD-Ridge penalty $J_{\lambda_n,\gamma_n}(|\cdot|)$ of (2.1) is the combination of ridge with SCAD. Obviously, the ridge and $P_{\lambda_n}(\cdot)$ techniques are special cases of a combined penalty, with $\lambda_n = 0$ and $\gamma_n = 0$, respectively. Let $\boldsymbol{\theta}_n^* \in \Theta_n$ is the true parameter. Without loss of generality, we suppose s_n is the number of non-zero coefficients and $p_n - s_n$ is the zero vector with zero components.

We adopted the method of Kown and Kim(2012)[19] to discuss the asymptotic property of our model, in the simulation of $p_n = O(n^k), k \geq 1$, where k depends on the derivative order of log-likelihood function.

3. The asymptotic properties of model selection

The following regularity conditions are imposed to discuss the theoretical property, where $M_1, M_2, \ldots$ are some positive constants.

3.1. Regularity conditions.

(C_1). For any constants c_1 and c_2 satisfying $0 < 5c_1 < c_2 \leq 1$,

$$s_n = O(n^{c_1}), \qquad \min_{1 \leq j \leq s_n} n^{(1-c_2)/2}|\theta_{nj}^*| \geq M_1.$$

(C_2). The first and second derivatives of the log-likelihood $\log f_n(\boldsymbol{Z}_{n1}, \boldsymbol{\theta}_n^*)$ satisfy

$$E_{\boldsymbol{\theta}_n^*}\left\{\frac{\partial \log f_n(\boldsymbol{Z}_{n1}, \boldsymbol{\theta}_n^*)}{\partial \theta_{nj}}\right\} = 0, \quad \text{for } j = 1, 2, \ldots, p_n$$

and

$$E_{\boldsymbol{\theta}_n^*}\left\{\frac{\partial \log f_n(\boldsymbol{Z}_{n1},\boldsymbol{\theta}_n^*)}{\partial\theta_{nj}}\frac{\partial \log f_n(\boldsymbol{Z}_{n1},\boldsymbol{\theta}_n^*)}{\partial\theta_{nk}}\right\}=-E_{\boldsymbol{\theta}_n^*}\left\{\frac{\partial^2 \log f_n(\boldsymbol{Z}_{n1},\boldsymbol{\theta}_n^*)}{\partial\theta_{nj}\partial\theta_{nk}}\right\},$$

for $j,k=1,2,\ldots,p_n;\quad n\geq 1$.

(C_3). The first $s_n\times s_n$ submatrix $I_n^{(1)}(\boldsymbol{\theta}_n^*)$ of the Fisher information matrix

$$I_n(\boldsymbol{\theta}_n^*)=E_{\boldsymbol{\theta}_n^*}\left[\left\{\frac{\partial \log f_n(\boldsymbol{Z}_{n1},\boldsymbol{\theta}_n^*)}{\partial\boldsymbol{\theta}_n}\right\}\left\{\frac{\partial \log f_n(\boldsymbol{Z}_{n1},\boldsymbol{\theta}_n^*)}{\partial\boldsymbol{\theta}_n}\right\}^T\right]$$

is positive definite, such that for all $n\geq 1$,

$$0<M_2<\lambda_{\min}\{I_n^{(1)}(\boldsymbol{\theta}_n^*)\}\leq\lambda_{\max}\{I_n^{(1)}(\boldsymbol{\theta}_n^*)\}<M_3<\infty,$$

where, $\lambda_{\min}(D)$ and $\lambda_{\max}(D)$ denote the smallest and largest eigenvalues of the given matrix D, respectively.

(C_4). There exists a sufficiently large open subset ω_n contains the true parameter $\boldsymbol{\theta}_n^*$ in $\Theta_n\in R^{p_n}$. For almost all $\boldsymbol{Z}_{ni}$ $(i=1,\ldots,n)$, the density admits a third derivatives for all $\boldsymbol{\theta}_n\in\omega_n$. Furthermore, there are functions V_{njtl}, such that $\partial^3 f_n(\boldsymbol{Z}_{ni},\boldsymbol{\theta}_n)/\partial\theta_{nj}\partial\theta_{nt}\partial\theta_{nl}$ satisfies

$$\left|\frac{\partial^3 \log f_n(\boldsymbol{Z}_{ni},\boldsymbol{\theta}_n)}{\partial\theta_{nj}\partial\theta_{nt}\partial\theta_{nl}}\right|\leq V_{njtl}(\boldsymbol{Z}_{ni}),$$

for $j,t,l=1,2,\ldots,p_n;\quad n\geq 1$. There exists an integer $m\geq 1$ such that

$$E_{\boldsymbol{\theta}_n}\left(V_{njtl}(\boldsymbol{Z}_{ni})\right)^{2m}<M_4<\infty.$$

(C_5). There also exists an integer $m\geq 1$ such that

$$E_{\boldsymbol{\theta}_n^*}\left\{\frac{\partial \log f_n(\boldsymbol{Z}_{n1},\boldsymbol{\theta}_n^*)}{\partial\theta_{nj}}\right\}^{2m}<M_5<\infty,\qquad E_{\boldsymbol{\theta}_n^*}\left\{\frac{\partial^2 \log f_n(\boldsymbol{Z}_{n1},\boldsymbol{\theta}_n^*)}{\partial\theta_{nj}\partial\theta_{nk}}\right\}^{2m}<M_6<\infty,$$

for all $j,k=1,2,\ldots,p_n;\quad n\geq 1$.

(C_6). There exists a convex open subset $\Omega_n\subset\Theta_n$, which contains $\boldsymbol{\theta}_n^*$, for a sufficiently large n, satisfies

$$\min_{\boldsymbol{\theta}_n\in B_n}\ \lambda_{min}(\boldsymbol{\theta}_n)>M_7,$$

where $\lambda_{min}(\boldsymbol{\theta}_n)$ is the smallest eigenvalue of the second derivatives of the negative log-likelihood (Hessian matrix)

$$-\frac{1}{2n}\sum_{i=1}^{n}\frac{\partial^2 \log f_n(\boldsymbol{Z}_{n1},\boldsymbol{\theta}_n)}{\partial\boldsymbol{\theta}_n^2}$$

at $\boldsymbol{\theta}_n$.

All the conditions are similar with Kown and Kim (2012) [19]. Condition (C_1) allows the number of parameters to diverge into infinite, and allows their value to converge to zero. The conditions $(C_2)-(C_4)$ are the standard assumptions for the maximum likelihood estimation. In the case of linear regression, Condition (C_3) has the design matrix corresponding to the relevant covariates as nonsingular. Conditions $(C_4),(C_5)$ determine the order of p_n with respect to some integer $m\geq 1$. To the model of logistic regression

$$\Pr(y=1|\boldsymbol{x})=\frac{\exp(\boldsymbol{x}^T\boldsymbol{\beta}_n^*)}{1+\exp(\boldsymbol{x}^T\boldsymbol{\beta}_n^*)},$$

we will use in the numerical studies. Suppose that the covariate x is bounded, such that $\max\limits_{1\leq j\leq p_n}|X_j|\leq b$, for some constant $b>0$. Since

$$\frac{\partial logf_n(y,\boldsymbol{\beta}_n^*|\boldsymbol{x})}{\partial\boldsymbol{\beta}_n^*}=\left(y-\frac{\exp(\boldsymbol{x}^T\boldsymbol{\beta}_n^*)}{1+\exp(\boldsymbol{x}^T\boldsymbol{\beta}_n^*)}\right)\boldsymbol{x},$$

it is easy to see that

$$\max_{1\leq j\leq p_n} E_{\boldsymbol{\beta}_n^*}\left\{\frac{\partial \log f_n(\boldsymbol{y}_{n1},\boldsymbol{\beta}_n^*)|\boldsymbol{x}}{\partial \beta_{nj}}\right\}^{2k} \leq \max_{1\leq j\leq p_n} x_j^{2k} \leq b^{2k} < \infty,$$

the other two inequalities can be checked similarly in (C_5). (C_6) is necessary in the proof of the following theorem.

3.2. Theoretical properties.

In order to state the CP penalized MLE (the global maximizer of the CP-penalized log-likelihood) is exactly the same as the oracle MLE asymptotically, we have to define the Oracle ridge estimator $\hat{\boldsymbol{\theta}}_n^0$ at first (See from Appendix A). Under the Lemma 1 and Lemma 2 of Appendix A, we only need to proof the SCAD-Ridge penalty likelihood estimation $\hat{\boldsymbol{\theta}}_n$ is asymptotic equal to the Oracle ridge MLE $\hat{\boldsymbol{\theta}}_n^0$. Then we can get the main theorem of our study.

3.1. Theorem. *Under the conditions $(C_1) - (C_6)$, let $\mathbf{B}_n(\lambda_n)$ denote the set of all local maximizers of (2.1), then the maximizer $\hat{\boldsymbol{\theta}}_n$ of (2.1) in Ω_n satisfies*

$$Pr(\hat{\boldsymbol{\theta}}_n \in \mathbf{B}_n(\lambda_n)) \to 1,$$

as $n \to \infty$, provided $\lambda_n = o(n^{-(1-c_2+c_1)}), \gamma_n = o(\lambda_n)$ and $p_n/(\sqrt{n}\lambda_n)^{2k} \to 0$.

The detail proof of the theorem can be seen in the Appendix.

4. Numerical studies

4.1. Simulations.

In this section, we illustrate the finite sample performance of the SCAD-Ridge penalized likelihood estimators and focus on the case when $p_n > n$. Through the numerical studies, we generate 100 data sets, each of which consists of n observations from the Logist regression model:

$$\Pr(Y_i = 1|\boldsymbol{x}_i) = \frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)}{1+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)},$$

where $\boldsymbol{\beta}_0$ is a $p_n \times 1$ vector, the first 15 components are all 3, and the remaining $p_n - 15$ components are all 0. Here we take $p_n = 600, n = 100$. The covariate $\boldsymbol{x}_i$ is generated from a p_n-dimensional multivariate normal distribution $N(\boldsymbol{0}, \Sigma_0)$, where $\Sigma_0 = (r^{|i-j|})_{i,j=1,\ldots,p_n}$.

We consider four penalization functions except for Oracle, including Lasso, Enet, SCAD and SCAD-Ridge. All estimates are computed by the CDA (coordinate descent algorithm) [3,13,14]. Because of $p_n > n$, the overfitting damaged the information criterion, so the tuning parameters are selected by the V-fold cross validation with V = 5. To evaluate the performances of every estimates, we give four summary statistics: " C "," IC", " l_2 loss " and " Mpmse ". " C " is the average number of choose zero coefficient correctly, " IC" stands the average number of non-zero coefficient is estimated to zero coefficient incorrectly, " l_2 loss " is the median of $\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|$ and " Mpmse " is the median of the prediction mean squared error, for the Logist regression model, that is

$$\frac{1}{t_n}\sum_{i=1}^{t_n}\left\{Y_i - \frac{\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)}{1+\exp(\boldsymbol{x}_i^T\boldsymbol{\beta}_0)}\right\}^2,$$

this is calculated by independent test sample size for $t_n = 1000$. We discuss the situations of $r = 0.5$ and $r = 0.9$, the simulated results are given in Table 1 and Table 2. The figures in parentheses of Table 1 and Table 2 are the corresponding standard deviation.

Table 1. The simulation results for r=0.5

(n, p_n)	Method	C	IC	l_2 loss	Mpmse
(100,600)	Lasso	564.77(1.783)	6.21(1.665)	1.074(0.021)	0.211(0.006)
	Enet	555.13(2.804)	5.56(1.986)	1.094(0.021)	0.210(0.005)
	SCAD	574.51(0.890)	4.57(1.671)	1.054(0.029)	0.212(0.008)
	SCAD-Ridge	557.97(2.213)	5.10(2.600)	1.068(0.028)	0.211(0.007)
	Oracle	585.00(0.000)	0.00(0.000)	0.946(0.041)	0.189(0.009)

Table 2. The simulation results for r=0.9

(n, p_n)	Method	C	IC	l_2 loss	Mpmse
(100,600)	Lasso	575.92(0.875)	8.50(1.314)	2.078(0.013)	0.195(0.005)
	Enet	568.70(1.933)	6.49(2.908)	2.078(0.012)	0.195(0.004)
	SCAD	580.94(0.724)	1.13(0.793)	2.069(0.021)	0.193(0.009)
	SCAD-Ridge	582.45(0.932)	0.67(1.075)	2.057(0.016)	0.193(0.006)
	Oracle	585.00(0.000)	0.00(0.000)	2.011(0.040)	0.187(0.009)

It can be seen from the result that the SCAD is better than the Lasso in both the situations, because the Lasso hasn't the Oracle property. When r = 0.5, we can see that the SCAD is the best choice except for Oracle. When r = 0.9, with high correlation, combination form of Enet and SCAD - Ridge are performing well than Lasso and SCAD which are not combined. The results of the four aspects of SCAD - Ridge are a little better than the results of the Enet. Obviously, under the situation of high correlation, SCAD-Ridge and Enet have more advantages.

4.2. Data analysis.

In our study, we use the Boston housing data to discuss the usefulness of our proposed method. This data set will examine the correlation between 'clean air' and housing prices in Harrison and Rubinfeld(1978)[16]. There are 506 observations, 13 independent factors, and a response variable LMV (the logarithm of the median value) of the owner-occupied homes. For more information, the reader can refer to '$http : //lib.stat.cmu.edu/datasets/bostoncorrected.txt.$' We split 506 observations into the first 400 observations as a training data set to select and fit the model, and the rest as a testing data set to evaluate the prediction ability of the selected model. We calculate the median absolute prediction error (MAPE) ($median\{|y_i - \hat{y}_i|, i = 1, \ldots, 106\}$) using the testing data. The performance of penalized likelihood with different penalties are summarized in Table 3. The results indicate that SCAD-Ridge and SCAD selects the simplest model, while Lasso and Enet includes extra variables. In addition to the MAPE, the SCAD-Ridge gets the smallest value. So it is easy to see that SCAD-Ridge penalty obtains the best performance.

5. Conclusion and discussion

This paper studies the maximum likelihood estimation of variable selection under the SCAD-Ridge penalty in the ultra-high dimension. Following the theoretical method of Kown and Kim (2012)[19] and under some conditions, we proposed the model with a nice oracle property. The established model includes all the regression model (such as Logistic and poisson regression). These results can greatly widen the application range of high-dimensional data. Numerical studies are discussed to assess the performance of

Table 3. Results of the Boston House Price data

Methods	No. zeros	MAPE
Lasso	5	4.7238
Enet	5	4.2124
SCAD	10	4.3086
SCAD-Ridge	10	4.0118

the proposed method. As long as the loss function is a smooth enough (there is a third derivative), the results can be extended to the M estimation of general penalization. But if the loss function is not smooth enough under the conditions, such as the hinge loss function of support vector machines (SVM) and Huber robust regression loss function, it is not so easy to explain. To extend our results to the non-smooth loss functions will be very meaningful in future.

Acknowledgments

The authors are very grateful to the editors and the referees for their careful reading and insightful comments, which lead to a substantially improved manuscript. The research of the first author is supported by the NSFC Special Fund of Tianyuan Visiting Scholar Fund (11726616), the Doctor Startup Fund of Liaoning Province (201501164) and the fundamental Research Funds for the Central Universities (DC20150104).

References

[1] Amin, M., Song L, Milton A.T, Xiaoguang W. *Combined penalized quantile regression in high dimensional models.* Pakistan Journal Statistics. **31**, 49–70, 2015.

[2] Antoniadis, A. *Wavelets in Statistics: A Review (with discussion)* . Journal of the Italian Statistical Association,**6**, 97–144, 1997.

[3] Breheny, P. and Huang, J. *Coordinate descent algorithms for nonconvex penalized regression with application to biological feature selection.* Ann. Appl. Stat. **5**, 232–253, 2011.

[4] Dong, Y., Song L. X., Wang, M. Q. and Xu, Y. *Combined-penalized likelihood estimations with a diverging number of parameters.* Journal of Applied Statistics. **41**, 1274–1285, 2014.

[5] Donoho, D. L. and Johnstone, I. M. *Ideal spatial adaptation via wavelet shrinkage.* Biometrika **81**, 425–455, 1994.

[6] Efron, B., Hasti, T. and Johnstone, I. *Least angle regression.* The Annals of Statistics **32**, 407–499, 2004.

[7] Hoerl, A. E. and Kennard, R. W. *Ridge regression: Biased estimation for nonorthogonal problems.* Technometrics **12**, 55–67, 1970.

[8] Fan, J. *Comments on 'Wavelets in Statistics: A Review' by A. Antoniadis.* Journal of the Italian Statistical Association **6**, 131–138, 1997.

[9] Fan, J. and Li, R. *Variable selection via nonconcave penalized likelihood and its oracle properties.* J. Amer. Statist. Assoc. **96**, 1348–1360, 2001.

[10] Fan, J. and Peng, H. *Nonconcave penalized likelihood with diverging number of parameters.* The Annals of Statistics **32**, 928–961, 2004.

[11] Fan, J. and Lv, J. *Sure independence screening for ultra-high dimensional feature space.* J. Roy. Statist. Soc. Ser. B. **70**, 849–911, 2008.

[12] Fan, J. and Lv, J. *A selective overview of variable selection in high dimensional feature space (invited review article).* Statistica Sinica **20**, 101–148, 2010.

[13] Friedman, J., Hastie, T., Hofling, H. and Tibshirani, R. *Pathwise coordinate optimization.* Ann. Appl. Stat. **1**, 302–332, 2007.

[14] Friedman, J., Hastie, T. and Tibshirani, R. *Regularization paths for generalized linear models via coordinate descent.* J. Statist. Softw. **33**, 1–22, 2010.

[15] Fu, W. J. *Penalized regressions: The Bridge versus the Lasso.* Journal of Computational and Graphical Statistics **7**, 397–416, 1998.

[16] Harrison, D., Rubinfeld, D. L. *Hedonic prices and the demand for clean air.* J. Environ. Economics and Management. **51**, 81–102, 1978.

[17] Huang, J., Breheny, P., Ma, S. G. and Zhang, C. H. *The Mnet Method for Variable Selection.* The University of Iowa Department of Statistics and Actuarial Science Technical Report No. 4021, 2010.

[18] Kim, Y., Choi, H. and Oh, H. *Smoothly clipped absolute deviation on high dimensions.* J. Amer. Statist. Assoc.**103**, 1656–1673, 2008.

[19] Kwon, S., Kim, Y. D. (2012). *Large Sample Properties of The SCAD-Penalized Maximum Likelihood Estimation on High Dimensiona.* Statistica Sinica **22**, 629–653.

[20] Meinshausen, N. and Buhlmann, P. *High-dimensional graphs and variable selection with the Lasso.* Ann. Statist.**34**, 1436–1462, 2006.

[21] Rosset, S. and Zhu, J. *Piecewise linear regularized solution paths.* Ann. Statist. **35**, 1012–1030, 2007.

[22] Segerstedt, B. *On ordinary ridge regression in generalized linear models.* Communications in Statistics - Theory and Methods **21**, 2227–2246, 1992.

[23] Tibshirani, R. J. *Regression shrinkage and selection via the Lasso.* J. Roy. Statist. Soc. B **58**, 267–288, 1996.

[24] Wang, X. M., Park, T. and Carriere, K. C. *Variable selection via combined penalization for high-dimensional data analysis.* Computational Statistics and Data Analysis **54**, 2230–2243, 2010.

[25] Zhao, P. and Yu, B. *On model selection consistency of Lasso.* J. Mach. Learn. Res. **7**, 2541–2567, 2006.

[26] Zou, H. and Hastie, T. *Regularization and variable selection via the elastic net.* J. Roy. Statist. Soc. B **67**, 301–320, 2005.

[27] Zou, H. and Zhang, H. *On the adaptive elastic-net with a diverging number of parameters.* The Annals of Statistics **37**, 1733–1751, 2009.

Appendix A

In order to prove Theorem 3.1, we need to define the Oracle ridge estimation. Under the support of some relevant lemmas, we can get our property. Following the opinion of Huang(2010) [17], we can denote the model of (2.1)

$$Q_n(\boldsymbol{\theta}_n) \quad = \quad L_n(\boldsymbol{\theta}_n) - n\sum_{j=1}^{p_n} J_{\lambda_n,\gamma_n}(|\theta_{nj}|) = L_n(\boldsymbol{\theta}_n,\gamma_n) - n\sum_{j=1}^{p_n} P_{\lambda_n}(|\theta_{nj}|), \tag{A.1}$$

where $L_n(\boldsymbol{\theta}_n,\gamma_n)=\sum\limits_{i=1}^{n}\log f_n(\boldsymbol{Z}_{ni},\boldsymbol{\theta}_n)-\gamma_n\sum\limits_{i=1}^{p_n}{\theta_{nj}}^2/2, P_{\lambda_n}(\cdot)$ is the SCAD penalty.

We define $\hat{\boldsymbol{\theta}}_n^0$ is the maximum likelihood Oracle ridge estimation of this model, which subject to $\theta_{nj}=0$, for $(s_n<j\leq p_n)$, combined $L_n(\boldsymbol{\theta}_n)$ with ridge to estimate the parameters. Its estimated coefficients of the irrelevant parameters are set to be exactly zero. Under the conditions $(C_1)-(C_5)$, $\gamma_n=o(\lambda_n)$ and the conclusion of Fan and Peng(2004) [10], we can get that the Oracle ridge MLE $\hat{\boldsymbol{\theta}}_n^0$ is asymptotically and $\sqrt{s_n/n}$-consistency estimator, that is

$$\|\hat{\boldsymbol{\theta}}_n^0-\boldsymbol{\theta}_n^*\|=O_p(\sqrt{s_n/n}). \tag{A.2}$$

Here we assume that the Oracle Oracle ridge estimator $\hat{\boldsymbol{\theta}}_n^0$ is in the set of Ω_n.

Next, in order to simplify the proof we define some notation. let
$S_{nj}(\boldsymbol{\theta}_n,\gamma_n)$ be the $j-th$ element of $\nabla L_n(\boldsymbol{\theta}_n,\gamma_n)=\frac{\partial L_n(\boldsymbol{\theta}_n,\gamma_n)}{\partial\boldsymbol{\theta}_n}$, for all $(j=1,\ldots,p_n)$.
$U_{njl}(\boldsymbol{\theta}_n,\gamma_n)$ is the $(j,l)-th$ element of $\nabla^2 L_n(\boldsymbol{\theta}_n,\gamma_n)=\frac{\partial^2 L_n(\boldsymbol{\theta}_n,\gamma_n)}{\partial\boldsymbol{\theta}_n^2}$, for all $(j,l=1,\ldots,p_n)$.

Similarly, ∇_1 denotes some partial derivatives of $S_{nj}(\boldsymbol{\theta}_n,\gamma_n)$ with respect to $\boldsymbol{\theta}_{n1}=(\theta_{n1},\theta_{n2},\ldots,\theta_{ns_n})^T$, so that $\nabla_1 S_{nj}(\boldsymbol{\theta}_n,\gamma_n)=\frac{\partial S_{nj}(\boldsymbol{\theta}_n,\gamma_n)}{\partial\boldsymbol{\theta}_{n1}}$ and $\nabla_1^2 S_{nj}(\boldsymbol{\theta}_n,\gamma_n)=\frac{\partial^2 S_{nj}(\boldsymbol{\theta}_n,\gamma_n)}{\partial\boldsymbol{\theta}_{n1}^2}$, for all $(j=1,\ldots,p_n)$.

Lemma1 : If C_2,C_4,C_5 hold, for any constants $\alpha>0$, for any $j\leq p_n$ and $\boldsymbol{\theta}_n\in\boldsymbol{B}_n(\lambda_n)$, we have

$$\Pr\left(|S_{nj}(\boldsymbol{\theta}_n^*,\gamma_n)|>\sqrt{n}\alpha\right) \quad = \quad O(\alpha^{-2k}), \tag{A.3}$$

$$\Pr\left(\|\nabla_1 S_{nj}(\boldsymbol{\theta}_n^*,\gamma_n)-E(\nabla_1 S_{nj}(\boldsymbol{\theta}_n^*,\gamma_n))\|>\sqrt{ns_n}\alpha\right) \quad = \quad O(\alpha^{-2k}), \tag{A.4}$$

$$\Pr\left(\|\nabla_1^2 S_{nj}(\boldsymbol{\theta}_n,\gamma_n)\|>ns_n\alpha\right) \quad = \quad O(\alpha^{-2k}). \tag{A.5}$$

Proof. Under the conditions C_2 and C_5 and the Rosenthal inequality, we have

$$E\left\{\frac{\partial\log f_n(\boldsymbol{Z}_{n1},\boldsymbol{\theta}_n^*)}{\partial\theta_{nj}}\right\}^{2k}=O(n^k), \qquad (j\leq p_n).$$

Because of $\lambda_n = o(n^{-(1-c_2+c_1)}), \gamma_n = o(\lambda_n)$, using the Markov inequality, it is easy to check

$$\begin{aligned}
\Pr\left(|S_{nj}(\boldsymbol{\theta}_n^*, \gamma_n)| > \sqrt{n}\alpha\right) &\leq \frac{E\left(|S_{nj}(\boldsymbol{\theta}_n^*, \gamma_n)|\right)^{2k}}{(\sqrt{n}\alpha)^{2k}} \\
&= \frac{E\left\{\frac{\partial \log f_n(\boldsymbol{Z}_{n1}, \boldsymbol{\theta}_n^*)}{\partial \theta_{nj}} - \gamma_n \theta_{nj}\right\}^{2k}}{(\sqrt{n}\alpha)^{2k}} \\
&= \frac{E\left\{\frac{\partial \log f_n(\boldsymbol{Z}_{n1}, \boldsymbol{\theta}_n^*)}{\partial \theta_{nj}}\right\}^{2k}}{(\sqrt{n}\alpha)^{2k}}(1+o_p(1)) \\
&= O(\alpha^{-2k}).
\end{aligned}$$

Hence, (A.3)holds.

Let $\Delta_{njl}(\boldsymbol{\theta}_n^*, \gamma_n) = U_{njl}(\boldsymbol{\theta}_n^*, \gamma_n) - E\left(U_{njl}(\boldsymbol{\theta}_n^*, \gamma_n)\right)$. Then, from C_2, C_5 and the Rosenthal inequality, we get

$$E\left(\frac{\partial^2 \log f_n(\boldsymbol{Z}_{n1}, \boldsymbol{\theta}_n^*)}{\partial \theta_{nj} \partial \theta_{nk}} - E\left(\frac{\partial^2 \log f_n(\boldsymbol{Z}_{n1}, \boldsymbol{\theta}_n^*)}{\partial \theta_{nj} \partial \theta_{nk}}\right)\right)^{2k} = O(n^k), \qquad (j, l \leq p_n).$$

Using the triangular inequality, we have

$$\begin{aligned}
& E\left(\|\nabla_1 S_{nj}(\boldsymbol{\theta}_n^*, \gamma_n) - E(\nabla_1 S_{nj}(\boldsymbol{\theta}_n^*, \gamma_n))\|^{2k}\right) \\
= \; & E\left\{\sum_{j=1}^{s_n} \left(\Delta_{njl}(\boldsymbol{\theta}_n^*, \gamma_n)\right)^2\right\}^k \\
\leq \; & \left[\sum_{j=1}^{s_n} \left\{E(\Delta_{njl}(\boldsymbol{\theta}_n^*, \gamma_n))^{2k}\right\}^{1/k}\right]^k \\
= \; & \left[\sum_{j=1}^{s_n} \left\{E\left(\frac{\partial^2 \log f_n(\boldsymbol{Z}_{n1}, \boldsymbol{\theta}_n^*)}{\partial \theta_{nj} \partial \theta_{nk}} - E\left(\frac{\partial^2 \log f_n(\boldsymbol{Z}_{n1}, \boldsymbol{\theta}_n^*)}{\partial \theta_{nj} \partial \theta_{nk}}\right)\right)^{2k}\right\}^{1/k}\right]^k \\
= \; & O(ns_n^k), \qquad (j \leq p_n).
\end{aligned}$$

Furthermore, using the Markov inequality again, we get

$$\begin{aligned}
& \Pr\left(\|\nabla_1 S_{nj}(\boldsymbol{\theta}_n^*, \gamma_n) - E(\nabla_1 S_{nj}(\boldsymbol{\theta}_n^*, \gamma_n))\| > \sqrt{ns_n}\alpha\right) \\
\leq \; & \frac{E\left(\|\nabla_1 S_{nj}(\boldsymbol{\theta}_n^*, \gamma_n) - E(\nabla_1 S_{nj}(\boldsymbol{\theta}_n^*, \gamma_n))\|\right)^{2k}}{(\sqrt{ns_n}\alpha)^{2k}} \\
= \; & \frac{O(n^k s_n^k)}{(ns_n)^k \alpha^{2k}} = O(\alpha^{-2k}),
\end{aligned}$$

So (A.4)holds.

Similarly, under the conditions C_4 and C_5, we have

$$E\left(\sum_{i=1}^{n} V_{njtl}(\boldsymbol{Z}_{ni})\right)^{2k} = O(n^{2k}),$$

hence we have

$$E\left(\|\nabla_1^2 S_{nj}(\boldsymbol{\theta}_n, \gamma_n)\|\right)^{2k} \leq O((ns_n)^{2k}).$$

By Markov inequality, we also can get

$$\Pr\left(\|\nabla_1^2 S_{nj}(\boldsymbol{\theta}_n, \gamma_n)\| > ns_n\alpha\right) \leq \frac{O((ns_n)^{2k})}{(ns_n\alpha)^{2k}} = O(\alpha^{-2k}).$$

Hence, (A.5) follows.

This completes the proof of lemma 1. □

Lemma2 : If conditions $(C_1)-(C_5)$ hold, when $\lambda_n = o(n^{-(1-c_2+c_1)}), \gamma_n = o(\lambda_n)$ and $p_n/(\sqrt{n}\lambda_n)^{2k} \to 0 \ (n \to \infty)$, we have

$$\Pr(\hat{\boldsymbol{\theta}}_n^0 \in \boldsymbol{B}_n(\lambda_n)) \to 1.$$

Proof. Because SCAD penalty is not derivative at the origin, refer to Kwon and Kim (2012) [19], we establish a new objective function to ensure that can be derivative everywhere,

$$U_n(\boldsymbol{\theta}_n, \lambda_n, \gamma_n) = L_n(\boldsymbol{\theta}_n, \gamma_n) - n\sum_{i=1}^{p_n} \left(P_{\lambda_n}(|\theta_{nj}|) - \lambda_n|\theta_{nj}|\right),$$

where$L_n(\boldsymbol{\theta}_n, \gamma_n)$ are the same as (A.1), it is the likelihood function with ridge. Obviously, $U_n(\cdot)$ is continuous and derivative, we have

$$\frac{\partial U_n(\boldsymbol{\theta}_n, \lambda_n, \gamma_n)}{\partial \theta_{nj}} = \begin{cases} S_{nj}(\boldsymbol{\theta}_n, \gamma_n), & 0 \le |\theta_{nj}| < \lambda_n, \\ S_{nj}(\boldsymbol{\theta}_n, \gamma_n) - n\left(\dfrac{\lambda_n - |\theta_{nj}|}{a-1}\right) sgn(\theta_{nj}), & \lambda_n \le |\theta_{nj}| < a\lambda_n, \\ S_{nj}(\boldsymbol{\theta}_n, \gamma_n) + n\lambda_n sgn(\theta_{nj}), & |\theta_{nj}| \ge a\lambda_n, \end{cases}$$

where $a > 2$, $\lambda_n \ge 0$ are tuning parameters, $j \le p_n$.

Since

$$Q_n(\boldsymbol{\theta}_n, \lambda_n, \gamma_n) = U_n(\boldsymbol{\theta}_n, \lambda_n, \gamma_n) - n\lambda_n \sum_{i=1}^{p_n} |\theta_{nj}|,$$

as $j \le p_n$, the corresponding Karush-Kuhn-Tucker (KKT) conditions (see, for example, Rosset and Zhu (2007)[21]) are

$$\text{(A.6)} \qquad \frac{\partial U_n(\boldsymbol{\theta}_n, \lambda_n, \gamma_n)}{\partial \theta_{nj}} = n\lambda_n sgn(\theta_{nj}), \theta_{nj} \ne 0,$$

$$\text{(A.7)} \qquad \left|\frac{\partial U_n(\boldsymbol{\theta}_n, \lambda_n, \gamma_n)}{\partial \theta_{nj}}\right| \le n\lambda_n, \theta_{nj} = 0.$$

Because $\hat{\boldsymbol{\theta}}_n^0$ is the Oracle ridge MLE of (2.1). By the definition of $\hat{\boldsymbol{\theta}}_n^0$,

$$\begin{cases} S_{nj}(\hat{\theta}_{nj}^0, \gamma_n) = 0, & j \le s_n, \\ \hat{\theta}_{nj}^0 = 0, & s_n < j \le p_n. \end{cases}$$

Hence, it suffices to show that as $n \to \infty$, $\hat{\boldsymbol{\theta}}_n^0$ satisfies,

$$\text{(A.8)} \qquad \Pr\left(\min_{1\le j\le s_n} |\hat{\theta}_{nj}^0| \ge a\lambda_n\right) \to 1,$$

$$\text{(A.9)} \qquad \Pr\left(\max_{s_n<j\le p_n} |S_{nj}(\hat{\theta}_{nj}^0, \gamma_n)| \le n\lambda_n\right) \to 1.$$

From the regularity condition C_1 and (A.2), we have

$$\min_{1\le j\le s_n} |\hat{\theta}_{nj}^0| \ge \min_{1\le j\le s_n} |\theta_{nj}^*| - \max_{1\le j\le s_n} |\hat{\theta}_{nj}^0 - \theta_{nj}^*| = O_p(n^{-(1-c_2)/2}),$$

Hence, follows $\lambda_n = o(n^{-(1-c_2+c_1)})$, (A.9) holds.

Next, we prove (A.9). From Taylor expansion and the definition of $\hat{\boldsymbol{\theta}}_n^0$ and $\boldsymbol{\theta}_n^*$, we have

$$\begin{aligned} S_{nj}(\hat{\boldsymbol{\theta}}_n^0,\gamma_n) = \quad & S_{nj}(\boldsymbol{\theta}_n^*,\gamma_n) + \nabla_1 S_{nj}(\boldsymbol{\theta}_n^*,\gamma_n)^T(\hat{\boldsymbol{\theta}}_n^0 - \boldsymbol{\theta}_n^*) \\ & + \tfrac{(\hat{\boldsymbol{\theta}}_n^0 - \boldsymbol{\theta}_n^*)^T \nabla_1^2 S_{nj}(\boldsymbol{\eta}_n,\gamma_n)(\hat{\boldsymbol{\theta}}_n^0 - \boldsymbol{\theta}_n^*)}{2}, \end{aligned}$$

for all $s_n < j \le p_n$, for some $\boldsymbol{\eta}_n$ lies between $\hat{\boldsymbol{\theta}}_n^0$ and $\boldsymbol{\theta}_n^*$. From Cauchy-Schwarz inequality, it follows that

$$\begin{aligned} & \Pr\left(\max_{s_n<j\le p_n} |S_{nj}(\hat{\boldsymbol{\theta}}_n^0,\gamma_n)| > n\lambda_n\right) \\ \le \quad & \Pr\left(\max_{s_n<j\le p_n} |S_{nj}(\boldsymbol{\theta}_n^*,\gamma_n)| > \frac{n\lambda_n}{4}\right) \\ + \quad & \Pr\left(\max_{s_n<j\le p_n} \|\nabla_1 S_{nj}(\boldsymbol{\theta}_n^*,\gamma_n) - E(\nabla_1 S_{nj}(\boldsymbol{\theta}_n^*,\gamma_n))\| \cdot \|\hat{\boldsymbol{\theta}}_n^0 - \boldsymbol{\theta}_n^*\| > \frac{n\lambda_n}{4}\right) \\ + \quad & \Pr\left(\max_{s_n<j\le p_n} \|E(\nabla_1 S_{nj}(\boldsymbol{\theta}_n^*,\gamma_n))\| \cdot \|\hat{\boldsymbol{\theta}}_n^0 - \boldsymbol{\theta}_n^*\| > \frac{n\lambda_n}{4}\right) \\ + \quad & \Pr\left(\max_{s_n<j\le p_n} \|\nabla_1^2 S_{nj}(\boldsymbol{\eta}_n,\gamma_n)\| \cdot \|\hat{\boldsymbol{\theta}}_n^0 - \boldsymbol{\theta}_n^*\| > \frac{n\lambda_n}{2}\right) \\ \triangleq \quad & L_1 + L_2 + L_3 + L_4 \end{aligned}$$

To discuss L_1 at first, when $n \to \infty$, we have

$$\begin{aligned} L_1 \quad = \quad & \Pr\left(\max_{s_n<j\le p_n} |S_{nj}(\boldsymbol{\theta}_n^*,\gamma_n)| > \frac{n\lambda_n}{4}\right) \\ \le \quad & \sum_{s_n<j\le p_n} \Pr\left(|S_{nj}(\boldsymbol{\theta}_n^*,\gamma_n)| > \frac{n\lambda_n}{4}\right) \\ = \quad & O\left(p_n/(\sqrt{n}\lambda_n)^{2k}\right) \to 0. \end{aligned}$$

Then we discuss L_2, by (A.2) and (A.4), as $n \to \infty$, we have

$$\begin{aligned} L_2 \quad \le \quad & \Pr\left(\|\hat{\boldsymbol{\theta}}_n^0 - \boldsymbol{\theta}_n^*\| > s_n/\sqrt{n}\right) \\ + \quad & \Pr\left(\max_{s_n<j\le p_n} \|\nabla_1 S_{nj}(\boldsymbol{\theta}_n^*,\gamma_n) - E(\nabla_1 S_{nj}(\boldsymbol{\theta}_n^*,\gamma_n))\| > \frac{n\sqrt{n}\lambda_n}{4s_n}\right) \\ = \quad & o(1) + O\left(p_n/(n\lambda_n/s_n\sqrt{s_n})^{2k}\right) \to 0. \end{aligned}$$

For the term of L_3, by C_5, when $n \to \infty$, we get

$$\begin{aligned} L_3 \quad = \quad & \Pr\left(\max_{s_n<j\le p_n} \|E(\nabla_1 S_{nj}(\boldsymbol{\theta}_n^*,\gamma_n))\| \cdot \|\hat{\boldsymbol{\theta}}_n^0 - \boldsymbol{\theta}_n^*\| > \frac{n\lambda_n}{4}\right) \\ \le \quad & \Pr\left(\|\hat{\boldsymbol{\theta}}_n^0 - \boldsymbol{\theta}_n^*\| > n\lambda_n/4M_5\sqrt{s_n}\right) \to 0. \end{aligned}$$

For the last term L_4, from $(A.5)$, when $n \to \infty$,

$$\begin{aligned} L_4 \quad \le \quad & \Pr\left(\|\hat{\boldsymbol{\theta}}_n^0 - \boldsymbol{\theta}_n^*\| > s_n\sqrt{s_n}/n\right) \\ + \quad & \Pr\left(\max_{s_n<j\le p_n} \|\nabla_1^2 S_{nj}(\boldsymbol{\eta}_n,\gamma_n)\| > n^2\lambda_n/2s_n\sqrt{s_n}\right) \\ = \quad & o(1) + O\left(p_n/(n\lambda_n/s_n\sqrt{s_n})^{2k}\right) \to 0. \end{aligned}$$

In conclusion,

$$\Pr\left(\max_{s_n<j\le p_n}|S_{nj}(\hat{\boldsymbol{\theta}}_n^0,\gamma_n)|>n\lambda_n\right)\to 0,$$

hence, $(A.9)$ follows.

It follows that, the Oracle ridge estimation $\hat{\boldsymbol{\theta}}_n^0$ of (2.1) satisfies the corresponding KKT conditions (even as $p_n\ge n$), it consist in the set of $\boldsymbol{B}_n(\lambda_n)$, which is the set of all local maximizers of $Q_n(\boldsymbol{\theta}_n)$.

This complete the proof of lemma 2. □

This lemma holds when $p_n=o(n^{(c_2-c_1)k})$ with sufficiently large k, even when $p_n\gg n$. If the distributions of the corresponding random variables(the first, second, and the third derivatives of the log-likelihood) have exponentially decaying tails, we can show that the lemma2 holds when $c_3>0$ and $p_n=O(exp(n^{(c_3)}))$, see reference Kim, Choi and Oh(2008) [18].

Because the log-likelihood function is the strictly concave on Ω_n, by the definition of $\hat{\boldsymbol{\theta}}_n^0$ and the contents of lemma2, in order to proof 3.1, we only need to proof the SCAD-Ridge penalty likelihood estimation $\hat{\boldsymbol{\theta}}_n$ is asymptotic equal to the Oracle ridge MLE $\hat{\boldsymbol{\theta}}_n^0$, that is

$$\Pr(\hat{\boldsymbol{\theta}}_n=\hat{\boldsymbol{\theta}}_n^0)\to 1. \tag{A.10}$$

Appendix B. Proof of Theorem 3.1

Proof. To ensure (A.10), it suffices to show that, as $n\to\infty$

$$\Pr\left(\max_{\boldsymbol{\theta}_n\in\Omega_n}Q_n(\boldsymbol{\theta}_n)\le Q_n(\hat{\boldsymbol{\theta}}_n^0)\right)\to 1. \tag{B.1}$$

From Taylor expansion, we get

$$L_n(\boldsymbol{\theta}_n,\gamma_n)-L_n(\hat{\boldsymbol{\theta}}_n^0,\gamma_n)= \nabla L_n(\boldsymbol{\theta}_n^0,\gamma_n)^T(\boldsymbol{\theta}_n-\hat{\boldsymbol{\theta}}_n^0) +\frac{(\boldsymbol{\theta}_n-\hat{\boldsymbol{\theta}}_n^0)^T\nabla^2L_n(\boldsymbol{\eta}_n^*,\gamma_n)(\boldsymbol{\theta}_n-\hat{\boldsymbol{\theta}}_n^0)}{2},$$

where $\boldsymbol{\eta}_n^*$ lies between $\boldsymbol{\theta}_n$ and $\hat{\boldsymbol{\theta}}_n^0$.

The definition of $\hat{\boldsymbol{\theta}}_n^0$ and the (A.9) imply that

$$\begin{aligned}\nabla L_n(\hat{\boldsymbol{\theta}}_n^0,\gamma_n)^T(\boldsymbol{\theta}_n-\hat{\boldsymbol{\theta}}_n^0) &= \sum_{j=1}^{p_n}S_{nj}(\hat{\boldsymbol{\theta}}_n^0,\gamma_n)(\theta_{nj}-\hat{\theta}_{nj}^0)\\ &\le \sum_{j=s_n+1}^{p_n}o_p(n\lambda_n)|\theta_{nj}|,\end{aligned}$$

and from (C_6) and the Cauchy- Schwarz inequality,

$$\begin{aligned}\frac{(\boldsymbol{\theta}_n-\hat{\boldsymbol{\theta}}_n^0)^T\nabla^2L_n(\boldsymbol{\eta}_n^*,\gamma_n)(\boldsymbol{\theta}_n-\hat{\boldsymbol{\theta}}_n^0)}{2} &\le -nM_7\|\theta_{nj}-\hat{\theta}_{nj}^0\|^2\\ &= -nM_7\sum_{j=1}^{p_n}(\theta_{nj}-\hat{\theta}_{nj}^0)^2\end{aligned}$$

holds. Hence, we have

$$
\begin{aligned}
& Q_n(\boldsymbol{\theta}_n) - Q_n(\hat{\boldsymbol{\theta}}_n^0) \\
= \;& L_n(\boldsymbol{\theta}_n, \gamma_n) + P_{\lambda_n}(|\boldsymbol{\theta}_n|) - L_n(\hat{\boldsymbol{\theta}}_n^0, \gamma_n|) - P_{\lambda_n}(|\boldsymbol{\theta}_n^0|) \\
= \;& (L_n(\boldsymbol{\theta}_n, \gamma_n) - L_n(\hat{\boldsymbol{\theta}}_n^0, \gamma_n)) + (P_{\lambda_n}(|\boldsymbol{\theta}_n|) - P_{\lambda_n}(|\boldsymbol{\theta}_n^0|)) \\
\leq \;& \sum_{j=s_n+1}^{p_n} o_p(n\lambda_n)|\theta_{nj}| - nM_7 \sum_{j=1}^{p_n} (\theta_{nj} - \hat{\theta}_{nj}^0)^2 + \sum_{j=1}^{p_n} P_{\lambda_n}(|\theta_{nj}^0|) - P_{\lambda_n}(|\theta_{nj}|) \\
\leq \;& \sum_{j=1}^{p_n} n\omega_{nj},
\end{aligned}
$$

where $\omega_{nj} = o_p(\lambda_n)|\theta_{nj}|I_{(j>s_n)} - M_7(\theta_{nj} - \hat{\theta}_{nj}^0)^2 + P_{\lambda_n}(|\theta_{nj}^0|) - P_{\lambda_n}(|\theta_{nj}|)$.

If $|\theta_{nj}| \geq a\lambda_n$ (for all $j \leq s_n$), we have

$$\sum_{j=1}^{s_n} \omega_{nj} \leq -\sum_{j=1}^{s_n} M_7(\theta_{nj} - \hat{\theta}_{nj}^0)^2 \leq 0.$$

If there exists a $j \leq s_n$ such that $|\theta_{nj}| < a\lambda_n$, then

$$|\theta_{nj} - \hat{\theta}_{nj}^0| \geq \min_{1 \leq j \leq s_n} |\theta_{nj}^*| - \max_{1 \leq j \leq s_n} |\hat{\theta}_{nj}^0 - \theta_{nj}^*| - a\lambda_n = O_p(n^{-(1-c_2)/2}).$$

Hence, we have, for the sufficiently large n, we get

$$\sum_{j=1}^{s_n} \omega_{nj} \leq -O_p(n^{-(1-c_2)}) \leq 0.$$

On the other hand, for each $j > s_n$, if $|\theta_{nj}| \geq \lambda_n$, we have

$$\omega_{nj} \leq |\theta_{nj}|(o_p(\lambda_n) - M_7|\theta_{nj}|),$$

and if $|\theta_{nj}| \leq \lambda_n$, then

$$\omega_{nj} \leq o_p(\lambda_n)|\theta_{nj}| - P_{\lambda_n}(|\theta_{nj}|) = (o_p(\lambda_n) - \lambda_n)|\theta_{nj}|.$$

Hence, for all the sufficiently large n, we have $\sum_{j=s_n+1}^{p_n} \omega_{nj} \leq 0$. As a consequence, (B.1) holds, that is (A.10) holds.

This complete the proof of Theorem 3.1. □