Statistics and Probability Letters 103 (2015) 86–92

Contents lists available at ScienceDirect

Statistics and Probability Letters

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

Model selection and estimation in high dimensional regression models with group SCAD☆

Xiao Guo [a], Hai Zhang [a,b,*], Yao Wang [c], Jiang-Lun Wu [a,d]

[a] School of Mathematics, Northwest University, Xi'an, 710069, China

[b] Academy of Mathematics and System Sciences, Chinese Academy of Sciences, Beijing, 100190, China

[c] School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, 710049, China

[d] Department of Mathematics, College of Science, Swansea University, Swansea, SA2 8PP, UK

ARTICLE INFO

Article history:
Received 14 January 2015
Received in revised form 16 April 2015
Accepted 16 April 2015
Available online 27 April 2015

Keywords:
Group selection
High dimension
Oracle property
Group SCAD
Sparsity

ABSTRACT

In this paper, we study the oracle property of the group SCAD under high dimensional settings where the number of groups can grow at a certain polynomial rate. Numerical studies are presented to demonstrate the merit of the group SCAD.

1. Introduction

Variable selection and feature extraction are basic problems in high dimensional and massive data analysis. The best subset selection which is one of the traditional variable selection approaches amounts to select the model with the smallest AIC (Akaike, 1973), BIC (Schwarz, 1978) or Cp (Mallows, 1973) score. However, the best subset selection involves solving an NP hard optimization problem so it is infeasible even for moderate number of variables. Moreover, the traditional best subset selection is unstable. Consequently, innovative variable selection procedure is expected to cope with the very high dimensionality, which is one of the hot topics in statistics and machine learning. In the past decades, many authors have considered the problem of variable selection and feature extraction by proposing various statistical approaches. Tibshirani (1996) proposed the LASSO (Least Absolute Shrinkage and Selection Operator) which is very popular for its ability to do the parameter estimation and variable selection simultaneously. However, the LASSO is not selection consistent in general and tends to select false variables into the model. The reason is that the LASSO over-shrinks large coefficients thus the resulting estimator is biased. Fan and Li (2001) discussed that the estimator induced by a good penalty function should enjoy three properties, that is, unbiasedness, sparsity and continuity. They proposed the SCAD (Smoothly Clipped Absolute Deviation) approach and showed the resulting estimator can enjoy the three desirable properties and it is asymptotically equivalent to the oracle estimator which is the least square estimator with the true nonzero coefficients known in advance. Fan and Peng (2004), Kim et al. (2008) further studied the oracle property of the SCAD under high dimensional settings. The adaptive LASSO in Zou (2006), the MCP (Minimax Concave Penalty) in Zhang (2010) and the $L_{1/2}$ regularization in Xu et al. (2010,

☆ This work was supported in part by National Natural Science Foundation of China under grant numbers 11171272 and 61273020.

* Corresponding author at: School of Mathematics, Northwest University, Xi'an, 710069, China.
E-mail address: zhanghai@nwu.edu.cn (H. Zhang).

http://dx.doi.org/10.1016/j.spl.2015.04.017
0167-7152/

2012) were also developed to overcome the inconsistency of the LASSO. Those approaches achieve the selection consistency and nearly unbiasedness simultaneously.

While in many applications, we are interested in selecting variables in a grouped manner. In the multifactor analysis of the variance (ANOVA) problem, the factor might be expressed by a group of dummy variables due to its several levels. Or variables may be grouped according to the domain knowledge. The group LASSO, proposed by Yuan and Lin (2006), can perform the group variable selection since its penalty function is intermediate between the l_1 penalty and the l_2 penalty. The group LASSO can be considered as an extension of the LASSO thus it is expected that the group LASSO suffers the shortcomings of the LASSO, such as selection inconsistency and asymptotic bias.

Recently, several researches have been proposed to cope with the bias and inconsistency of the group LASSO estimator. Wang and Leng (2008) studied the oracle property of the adaptive group LASSO estimator in fixed dimensional cases where the number of the groups is not large compared with the number of observations. Wei and Huang (2010) generalized the results to high dimensional cases where the number of the groups can grow with the sample size. They showed that under certain conditions, the adaptive group LASSO is consistent in group selection provided that the initial estimator satisfies certain requirements. Zeng and Xie (2012) studied the power of SCAD combined with l_2 penalty in selecting group effects. Wang et al. (2007) proposed the group SCAD approach to select the groups with time-varying coefficients. They showed that the group SCAD estimator is asymptotically equivalent to the oracle estimator in fixed dimensional cases. However, there are many situations where the number of groups can be much larger than the sample size and to our knowledge, the high dimensional properties of the group SCAD estimator have not been studied yet.

In this paper, we study the oracle property of the group SCAD estimator in the manner that the number of unknown groups is allowed to grow at a certain polynomial rate. We also perform numerical studies to support our theoretical findings.

2. The group SCAD and its theoretical properties

In this section, we first review the group SCAD approach. Then we investigate the oracle property in high dimensional cases.

We consider the following linear regression model with d predictors which are divided into p nonoverlapping groups

$$Y = \sum_{j=1}^{p} X_j \beta_j^* + \epsilon, \tag{1}$$

where Y is an $n \times 1$ vector of the response variable, X_j is the $n \times d_j$ design matrix of the predictors in the jth group, $\sum_{j=1}^{p} d_j = d$. $\beta_j^* = (\beta_{j1}^*, \ldots, \beta_{jdj}^*)^{\mathrm{T}} \in R^{dj}$ is the $d_j \times 1$ vector of the unknown true regression coefficients of the jth group, ϵ is the error vector. Let $X = (X_1^{\mathrm{T}}, \ldots, X_p^{\mathrm{T}})$ where $X_j = (X_{j1}, \ldots, X_{jdj})^T$ and

$$Q(\beta) = \frac{1}{2n} \|Y - X\beta\|_2^2 + \sum_{j=1}^{p} p_{\lambda_n}(\|\beta_j\|_2), \tag{2}$$

where $p_\lambda(t)$ is the SCAD penalty which is defined by

$$p'_\lambda(t) = \lambda \left\{ I(t \le \lambda) + \frac{(a\lambda - t)_+}{(a-1)\lambda} I(t > \lambda) \right\},$$

for some $a > 2$ and $t > 0$, where λ is the tuning parameter. In the non-grouped case, Fan and Li (2001) suggested to use $a = 3.7$ since the Bayes risks are not sensitive to the choice of a. We used $a = 3$ in the grouped case throughout the paper and more detailed explanation can be found in Section 3. It is obvious that $Q(\beta)$ consists of two parts, the square loss function and the group SCAD penalty. The group SCAD penalty is intermediate between the l_1 penalty and the SCAD penalty that can lead to the group variable selection. Note that minimizing $Q(\beta)$ is nonconvex minimization problem for which the global solution is difficult to compute. In this paper, we mainly discuss the theoretical properties of local group SCAD estimators. Without loss of generality, we assume the coefficients corresponding to the first q groups are nonzero and the remaining regression coefficients are zero. Let $X = (X^{(1)}, X^{(2)})$, where $X^{(1)}$ is the submatrix of X corresponding to the first q groups and $X^{(2)}$ is the submatrix of X corresponding to the last $p - q$ groups. Similarly, we let $\beta^* = (\beta^{*(1)^{\mathrm{T}}}, \beta^{*(2)^{\mathrm{T}}})^{\mathrm{T}}$, $\hat{\beta} = (\hat{\beta}^{(1)^{\mathrm{T}}}, \hat{\beta}^{(2)^{\mathrm{T}}})^{\mathrm{T}}$. Let $C = X^{\mathrm{T}}X/n$ and $C^{(i,j)} = X^{(i)^{\mathrm{T}}}X^{(j)}/n$ for $i, j = 1, 2$. Next we define the oracle estimator as $\hat{\beta}^o = (\hat{\beta}^{o(1)^{\mathrm{T}}}, 0^{(2)^{\mathrm{T}}})^{\mathrm{T}}$, where $\hat{\beta}^{o(1)}$ is the ordinary least square solution of minimizing $\|Y - X^{(1)}\beta^{(1)}\|_2^2$ if we know the true zero coefficients in advance.

In order to propose the high dimensional statistical properties of the group SCAD, we assume the following regularity conditions hold,

(A1) There exists a positive constant M_1 such that $\frac{1}{n}(X_{jl})^{\mathrm{T}}X_{jl} \le M_1$ for $j = 1, \ldots, p_n$, $l = 1, \ldots, d_j$.

(A2) There exists a positive constant M_2 such that $\alpha^{\mathrm{T}}C^{(1,1)}\alpha \ge M_2$ for all $\|\alpha\|_2^2 = 1$.

(A3) $q_n = O(n^{c_1})$ for some $0 < c_1 < 1$.

(A4) There exist positive constants c_2 and M_3 such that $c_1 < c_2 \le 1$ and $n^{(1-c_2)/2} \min_{j=1,\ldots,q_n} \|\beta_j^*\|_2 \ge M_3$.

(A5) $d_j = O_p(1)$ for $j = 1, \ldots, p_n$.

The regularity conditions (A1)–(A4) were firstly used by Zhao and Yu (2006) to prove the model selection consistency of the LASSO estimator and Kim et al. (2008) also used (A1)–(A4) to show the oracle property of the SCAD estimator under high dimensional settings except that they used the non-grouped form of (A4). (A1) can be satisfied as long as normalizing the predictors. (A2) requires the eigenvalues of $C^{(1,1)}$ bounded by a positive constant to guarantee the good behavior of $(C^{(1,1)})^{-1}$. (A3) restricts the growing rate of the number of true relevant groups with respect to the sample size n. (A4) guarantees the strength of relevant groups. Note that, for group variable selection problems, not only the number of groups but also the number of variables in each group might grow with the sample size n. We add the regularity condition (A5) technically which means the number of variables in each group should be bounded in probability.

Under the regularity conditions (A1)–(A5), the result follows:

Theorem 1. *Assume that there exists an integer $k > 0$ such that $E(\epsilon_i)^{2k} < \infty$. Then under the regularity conditions* (A1)–(A5), *there exists a local minimizer $\hat{\beta}$ of $Q(\beta)$ such that $Pr(\hat{\beta}^o = \hat{\beta}) \to 1$ as $n \to \infty$, provided that $\lambda_n = o(n^{-(1-(c_2-c_1)/2)})$, and $(p_n/\sqrt{n}\lambda_n)^{2k} \to 0$.*

Proof. Our proof make use of the ideas of Theorem 1 in Kim et al. (2008). To show Theorem 1, we first give sufficiency conditions under which a solution is a local minimizer of $Q(\beta)$. Then we verify that the oracle estimator $\hat{\beta}^o$ satisfies the certain conditions with probability tending to one.

Recall that

$$Q(\beta) = \frac{1}{2n}\|Y - X\beta\|_2^2 + \sum_{j=1}^{p_n} P_{\lambda_n}(\|\beta_j\|_2).$$

If $\|\beta_j\|_2 \neq 0$, then the partial derivative of Q with respect to β_j is given by $\frac{\partial Q}{\partial \beta_j} = -\frac{1}{n}X_j^{\mathrm{T}}(Y - X^{\mathrm{T}}\beta) + \frac{P'_\lambda(\|\beta_j\|_2)\beta_j}{\|\beta_j\|_2}$, for all $j = 1, \ldots, p_n$, where $\beta_j = (\beta_{j1}, \ldots, \beta_{jd_j})^{\mathrm{T}}$. By the second order sufficiency conditions (see, e.g., Bertsekas (1999, p.320)), if β satisfies the following two arguments:

$$V_j(\beta) = 0, \|\beta_j\|_2 > a\lambda, \quad \text{for } j = 1, \ldots, q_n, \tag{3}$$

$$\|V_j(\beta)\|_2 < \lambda, \|\beta_j\|_2 < \lambda, \quad \text{for } j = q_n + 1, \ldots, p_n, \tag{4}$$

where $V_j(\beta) = -\frac{1}{n}X_j^{\mathrm{T}}(Y - X^{\mathrm{T}}\beta)$ is a column vector with dimension d_j, that is $V_j(\beta) = (V_{j1}(\beta), \ldots, V_{jd_j}(\beta))^{\mathrm{T}}$. Then β is a local minimizer of $Q(\beta)$. Based on this, we next show that $\hat{\beta}^o$ satisfies (3) and (4) with probability tending to one and λ replaced by λ_n.

For $j \leq q_n$, by the definition of oracle estimator, $V_j(\hat{\beta}^o) = 0$ holds trivially. Thus to show (3), we only need to verify that

$$Pr(\|\hat{\beta}_j^{\,o}\|_2 > a\lambda_n) \to 1, \quad \text{for } j = 1, \ldots, q_n, \text{ as } n \to \infty. \tag{5}$$

Note that $\hat{\beta}^{o(1)} = \frac{1}{n}(C^{(1,1)})^{-1}X^{(1)^{\mathrm{T}}}Y = \frac{1}{n}(C^{(1,1)})^{-1}X^{(1)^{\mathrm{T}}}\epsilon + \beta^{*(1)}$ and $\|\hat{\beta}_j^{\,o}\|_2 \geq \|\beta_j^*\|_2 - \|\hat{\beta}_j^{\,o} - \beta_j^*\|_2$. Also note that $\min_{j\leq q_n}\|\beta_j^*\|_2 = O(n^{-(1-c_2)/2})$ and $\lambda_n = o(n^{-(1-(c_2-c_1)/2)})$. For (5), it then suffices to show that

$$\max_{j\leq q_n}\|\hat{\beta}_j^{\,o} - \beta_j^*\|_2 = o_p(n^{-(1-c_2/2)}). \tag{6}$$

Let $s_j = \sqrt{n}(\hat{\beta}_j^{\,o} - \beta_j^*)$ for $j = 1, \ldots, q_n$, next we show that

$$\max_{j\leq q_n}\|s_j\|_2 = o_p(n^{c_2/2}), \tag{7}$$

which is identical to (6). Note that $s = \frac{1}{\sqrt{n}}(C^{(1,1)})^{-1}X^{(1)^{\mathrm{T}}}\epsilon = U^{(1)^{\mathrm{T}}}\epsilon$, where $s = (s_1{}^{\mathrm{T}}, \ldots, s_{q_n}{}^{\mathrm{T}})^{\mathrm{T}}$, $s_j = (s_{j1}, \ldots, s_{jd_j})^{\mathrm{T}}$ and $U^{(1)^{\mathrm{T}}} = \frac{1}{\sqrt{n}}(C^{(1,1)})^{-1}X^{(1)^{\mathrm{T}}} = (u_1^{(1)^{\mathrm{T}}}, \ldots, u_{q_n}^{(1)^{\mathrm{T}}})^{\mathrm{T}}$ with $u_j^{(1)} = (u_{j1}, \ldots, u_{jd_j})^{\mathrm{T}}$. By the regularity condition (A2), the minimum eigenvalue of $C^{(1,1)}$ is greater than M_2, that is followed by that the maximum eigenvalue of $(C^{(1,1)})^{-1}$ is smaller than $1/M_2$. Note that $U^{(1)^{\mathrm{T}}}U^{(1)} = (C^{(1,1)})^{-1}$, thus $\alpha^{\mathrm{T}}U^{(1)^{\mathrm{T}}}U^{(1)}\alpha \leq 1/M_2$ for $\|\alpha\|_2^2 = 1$. By the Cauchy–Schwarz inequality and $E(\epsilon_l)^{2k} < \infty$ for $l = 1, \ldots, n$, we get $E(s_{jl})^{2k} < \infty$ for $j = 1, \ldots, q_n$, $l = 1, \ldots, d_j$. By the regularity condition (A5), we further have $E(\|s_j\|_2)^{2k} < \infty$ for $j = 1, \ldots, q_n$. Based on Markov's inequality, we obtain $Pr(\|s_j\|_2 > t) = O(t^{-2k})$ for all $t > 0$ and $j = 1, \ldots, q_n$. For any $\eta > 0$, we have

$$\begin{aligned} Pr\left(\bigcup_{j=1}^{q_n}\{\|z_j\|_2 > \eta n^{c_2/2}\}\right) &\leq \sum_{j=1}^{q_n} Pr(\|z_j\|_2 > \eta n^{c_2/2}) \\ &\leq \sum_{j=1}^{q_n} \eta^{-2k} n^{-c_2 k} \end{aligned}$$

$$\leq \sum_{j=1}^{q_n} \frac{1}{\eta} n^{-c_2 k}$$
$$= \frac{1}{\eta} q_n n^{-c_2 k} \leq \frac{1}{\eta} n^{-(c_2-c_1)k} \to 0,$$

where the last inequality is implied by the regularity condition (A3). Therefore, $Pr(\|s_j\|_2 < \eta n^{c_2/2}) \to 1$, for $j = 1, \ldots, q_n$. Thus (7) holds as η can be arbitrarily small. Consequently, we have proved that $\hat{\beta}^o$ satisfies (3) with probability tending to one. We show (4) next.

By the definition of oracle estimator, $\hat{\beta}_j^{\,o} = 0$ for $j = q_n + 1, \ldots, p_n$, thus $\|\hat{\beta}_j^{\,o}\|_2 \leq \lambda$ holds trivially, we only need to show $\|V_j(\hat{\beta}^o)\|_2 < \lambda_n$ for $j = q_n + 1, \ldots, p_n$. We will show this by verifying the following:

$$Pr\left(\bigcup_{j=q_n+1}^{p_n} \{\|V_j(\hat{\beta}^o)\|_2 > \lambda_n\}\right) \to 0. \tag{8}$$

By simple calculation and rearrangement, we have

$$\begin{aligned}(V_j(\hat{\beta}^o), j = q_n + 1, \ldots, p_n) &= -\frac{1}{n} X^{(2)^{\mathrm{T}}}(Y - X^{(1)}\hat{\beta}^{o(1)} - X^{(2)}\hat{\beta}^{o(2)}) \\ &= -\frac{1}{n} X^{(2)^{\mathrm{T}}}\left(X^{(1)}\beta^{*(1)} + \epsilon - X^{(1)}\frac{1}{n}(C^{(1,1)})^{-1}X^{(1)^{\mathrm{T}}}(X^{(1)}\beta^{*(1)} + \epsilon)\right) \\ &= -\frac{1}{n} X^{(2)^{\mathrm{T}}}\left(I - X^{(1)}\frac{1}{n}(C^{(1,1)})^{-1}X^{(1)^{\mathrm{T}}}\right)\epsilon.\end{aligned}$$

Let $\sqrt{n}V_{jl}(\hat{\beta}^o) = u_{jl}^{(2)^{\mathrm{T}}}\epsilon$, for $j = q_n + 1, \ldots, p_n$, $l = 1, \ldots, d_j$, and $U^{(2)^{\mathrm{T}}} = C^{(2,1)}(C^{(1,1)})^{-1}\frac{1}{\sqrt{n}}X^{(1)^{\mathrm{T}}} - \frac{1}{\sqrt{n}}X^{(2)^{\mathrm{T}}}$, where $u_{jl}^{(2)}$ is the column vector corresponding to $U^{(2)}$. Note that

$$U^{(2)^{\mathrm{T}}}U^{(2)} = \frac{1}{n} X^{(2)^{\mathrm{T}}}(I - X^{(1)}(X^{(1)^{\mathrm{T}}}X^{(1)})^{-1}X^{(1)^{\mathrm{T}}})X^{(2)},$$

and that $I - X^{(1)}(X^{(1)^{\mathrm{T}}}X^{(1)})^{-1}X^{(1)^{\mathrm{T}}}$ is an orthogonal projection matrix, by the regularity condition (A1), we have that $\|u_{jl}^{(2)}\|_2^2 \leq M_1$ for $j = q_n + 1, \ldots, p_n$, $l = 1, \ldots, d_j$. Using the Cauchy–Schwarz inequality and $E(\epsilon_l)^{2k} < \infty$ for $l = 1, \ldots, n$, we get that $E(\xi_{jl})^{2k} < \infty$ for $j = q_n + 1, \ldots, p_n$, $l = 1, \ldots, d_j$, where $\xi_{jl} = \sqrt{n}V_{jl}(\hat{\beta}^o)$. Thus, the regularity condition (A5) implies that $E(\|\xi_j\|_2)^{2k} < \infty$ for $j = q_n + 1, \ldots, p_n$. By Markov's inequality, we derive that $Pr(\|\xi_j\|_2 > t) = O(t^{-2k})$ for all $t > 0$ and $j = 1, \ldots, q_n$. Consequently,

$$\begin{aligned}Pr\left(\bigcup_{j=q_n+1}^{p_n} \{\|V_j(\hat{\beta}^o)\|_2 > \lambda_n\}\right) &\leq \sum_{j=q_n+1}^{p_n} Pr(\|V_j(\hat{\beta}^o)\|_2 > \lambda_n) \\ &\leq \sum_{j=q_n+1}^{p_n} Pr(\|\xi_j\|_2 > \sqrt{n}\lambda_n) \\ &= (p_n - q_n)O\left(\frac{1}{(\sqrt{n}\lambda_n)^{2k}}\right) \\ &= O\left(\frac{p_n}{(\sqrt{n}\lambda_n)^{2k}}\right) \to 0.\end{aligned}$$

Therefore, (4) holds.

Having proved that $\hat{\beta}^o$ satisfies (3) and (4) with probability tending to one and λ replaced by λ_n, we then conclude that $\hat{\beta}^o$ is a local minimizer of $Q(\beta)$ with probability tending to one and the proof is completed.

Remark 1. This theorem shows that under certain regularity conditions, the oracle estimator is asymptotically a local group SCAD estimator under high dimensional settings. Note that the conclusion in this Theorem is stronger than the oracle property defined by Fan and Li (2001). Also note that when ϵ_i has all the moments, the oracle property holds when $p_n = O(n^\alpha)$ for any $\alpha > 0$.

3. Numerical studies

In this section, we compare the performance of the group SCAD, the group LASSO and the SCAD by a series of simulated data and real data experiments. We used the group coordinate descent algorithm which is proposed by Huang et al. (2012), Breheny and Huang (2015) to solve the group SCAD.

3.1. Simulation experiments

In Experiments 1–2 of this subsection, proposed by Yuan and Lin (2006), the number of the groups is fixed. Then we study the high dimensional cases where the number of groups is larger than the number of observations.

Experiment 1. In this experiment, we first generated 15 latent variables $Z_1, \ldots, Z_{15}$ according to a zero mean multivariate normal distribution and the covariance between Z_i and Z_j was $0.5^{|i-j|}$. Then Z_i was trichotomized as 0,1 or 2 if it was smaller than $\Phi^{-1}(1/3)$, larger than $\Phi^{-1}(2/3)$ or in between. The response Y was generated from

$$Y = 1.8I(Z_1 = 1) - 1.2I(Z_1 = 0) + I(Z_3 = 1) + 0.5I(Z_3 = 0) + I(Z_5 = 1) + I(Z_5 = 0) + \epsilon,$$

where $I(\cdot)$ is the indicator function and $\epsilon \sim N(0, \sigma^2)$. For a categorical variable with three classes, we need two dummy variables to represent different levels, thus the two dummy variables corresponding to the original categorical variable make up a group naturally.

Experiment 2. In this experiment, 17 latent variables $Z_1, \ldots, Z_{16}$ and W were generated from a standard normal distribution independently. Then we define the covariates as $X_i = (Z_i + W)/\sqrt{2}$ for $i = 1, \ldots, 16$. The response Y was generated from

$$Y = X_3^3 + X_3^2 + X_3 + \frac{1}{3}X_6^3 - X_6^2 + \frac{2}{3}X_6 + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2)$.

In each experiment, for comparison, we tested various sample sizes ($n = 100, 200, 300$) and various noise levels ($\sigma = 1, 3$). For each setting, we simulated 200 datasets for each combination of (n, σ). The model error (Fan and Li, 2001) and its standard deviation were summarized. For linear model with zero mean noise $Y = X^{\mathrm{T}}\beta + \epsilon$, where $E(\epsilon|X) = 0$, the model error equals $(\hat{\beta} - \beta)^{\mathrm{T}} E(XX^{\mathrm{T}})(\hat{\beta} - \beta)$. We generated independent datasets of 1000 observations to compute the model error. Average model error of 200 replications was summarized. In addition, average model size (the number of groups) and its standard deviation were compared. Note that since the SCAD is designed for individual variable selection, a group is thought to be selected into the model as long as one of the variables corresponding to the group is selected. Lastly, the average number of correct and incorrect zero coefficients which correspond to the true zero and coefficients improperly set to zero were reported. We used BIC to choose the tuning parameter for each selection methods. In addition, we fixed the tuning parameter a in the group SCAD with 3 throughout our experiments since the model error is not sensitive to a based on our simulation results. For the sake of space, we omit the results.

The results show that the group SCAD outperforms the group LASSO and the SCAD in all three aspects. Firstly, the model error of the group SCAD is smaller than that of the group LASSO and the SCAD which indicates that the group SCAD has stronger prediction ability. The group SCAD is more stable in prediction since its standard deviations of the model error are smaller that of the group LASSO and the SCAD. Secondly, the solution yields by the group SCAD is more sparse than that of the other two methods due to the fact that the average number of factors selected by the group SCAD is smaller than that of the group LASSO and the SCAD. In addition, the group SCAD also outperforms the group LASSO and the SCAD in terms of the variable selection accuracy since it produces more correct zero coefficients and less incorrect coefficients.

In the following experiments, we study the high dimensional cases. We extended the dimension in Experiments 1 and 2. The true model remains unchanged but we added redundant variables to make the number of groups p bigger than the sample size n. We consider two settings $n = 200, p = 210$ and $n = 200, p = 500$. The tuning parameters were chosen by cross validation. We tested 50 simulated datasets. Average model error, average number of factors selected and average number of correct and incorrect zero coefficients of 50 replications were summarized in Table 1.

The results in Table 1 show that the group SCAD outperforms the other two methods under high dimensional settings. First, the group SCAD performs better in prediction accuracy and variable selection accuracy than the group LASSO and the SCAD that is in accordance with the fixed dimensional settings. Second, the average number of factors selected by the group SCAD is small thus resulting in a more parsimonious and more interpretable model. Note that the performances of all three methods become worse compared with that of the fixed dimensional settings. For example, the model error and the model size of the group SCAD in Experiment 1 under the setting $\sigma = 1, p = 210$ respectively are 0.084 and 18.98 whereas they are 0.053 and 3.95 under the corresponding low dimensional setting.

3.2. Real data

In this subsection, we compare the performance of the group SCAD, the group LASSO and the SCAD on real data Bardet which was discussed in Scheetz et al. (2006). The data consists of gene expression data from the eye tissue of 120 twelve-week-old male rats. We preprocessed the data, resulting in a grouped regression problem with 120 samples and 100 predictors which were expanded from 20 genes using 5 basis B-splines. We randomly selected 100 samples 200 times as the training data and the left data serve as the test data. We computed the average Mean Squared Error on the test data. We used cross validation to choose the tuning parameter for each selection method. Average MSE and average number of factors selected were summarized in Table 2.

Table 1
High dimensional simulation results.

Method	Model error	Avg. no. of 0 coefficients		Avg. no. of factors selected
		Correct	Incorrect	
Experiment 1, $\sigma = 1, p = 210$				
Group SCAD	0.084(0.058)	383.04	0.00	18.98(7.700)
Group LASSO	0.169(0.069)	375.64	0.00	22.18(13.251)
SCAD	0.137(0.109)	391.86	0.52	24.24(8.348)
Experiment 1, $\sigma = 3, p = 210$				
Group SCAD	0.932(0.406)	398.72	2.72	9.28(9.091)
Group LASSO	1.008(0.432)	398.16	2.60	9.62(11.402)
SCAD	1.101(0.367)	407.08	4.04	8.02(7.839)
Experiment 1, $\sigma = 1, p = 500$				
Group SCAD	0.109(0.070)	959.96	0.04	20.00(12.996)
Group LASSO	0.157(0.060)	963.64	0.00	18.18(13.258)
SCAD	0.177(0.099)	965.56	0.80	30.64(11.923)
Experiment 1, $\sigma = 3, p = 500$				
Group SCAD	0.990(0.355)	964.52	2.84	16.32(13.948)
Group LASSO	1.152(0.524)	958.44	2.64	19.46(22.471)
SCAD	1.275(0.857)	977.80	3.98	17.24 (14.590)
Experiment 2, $\sigma = 1, p = 210$				
Group SCAD	0.040(0.026)	623.88	0.00	2.04(0.198)
Group LASSO	0.187(0.084)	612.00	0.00	6.00(3.071)
SCAD	0.336(0.219)	622.76	1.10	3.24(1.944)
Experiment 2, $\sigma = 3, p = 210$				
Group SCAD	0.378(0.225)	617.28	0.00	4.24(3.952)
Group LASSO	1.523(0.745)	590.04	0.00	13.32(7.210)
SCAD	1.363(1.486)	613.32	1.88	12.50(5.786)
Experiment 2, $\sigma = 1, p = 500$				
Group SCAD	0.036(0.023)	1493.88	0.00	2.04(0.198)
Group LASSO	0.189(0.077)	1479.66	0.00	6.78(3.382)
SCAD	0.322(0.373)	1491.90	1.04	4.04(2.531)
Experiment 2, $\sigma = 3, p = 500$				
Group SCAD	0.650(0.667)	1480.44	0.00	6.52(7.998)
Group LASSO	1.835(0.748)	1451.64	0.00	16.12(8.233)
SCAD	1.866(1.053)	1477.64	2.02	17.96 (8.293)

Table 2
Prediction and variable selection accuracy on data bardet.

	Group SCAD	Group LASSO	SCAD
MSE	0.738	0.739	0.615
Factors selected	5.51	6.76	9.12

From Table 2, we see that the average MSE of the group SCAD, the group LASSO and the SCAD respectively are 0.738, 0.739 and 0.615. So the group SCAD performs slightly better than the group LASSO in prediction accuracy. Although the MSE of the SCAD is the smallest among the three methods, we do not think it good enough since it is not designed for group selection so the resulting model is less interpretable. Moreover, the group SCAD selects least factors. It is worth noting that when we apply the group SCAD and the group LASSO to the whole dataset, the groups selected by the group LASSO include that selected by the group SCAD.

4. Conclusion

In the paper, we focused on the group variable selection problem. We have studied the oracle property of the group SCAD estimator in high dimensional cases where the number of groups can be larger than the sample size. Simulation studies and real data experiments have suggested that the group SCAD can outperform the group LASSO and the SCAD in the prediction accuracy and variable selection consistency. We mainly discussed the properties of local group SCAD estimators. Recently, Fan et al. (2014) showed that under mild conditions, the two step LLA with the LASSO as the initial estimator produces strong oracle solution for folded concave penalization problems. Similar results could be further extended to the group SCAD penalty. Additionally, our results could be further extended to generalized linear models especially logistic regression.

Acknowledgments

We are very grateful to the Editor, D. Paindaveine, the Associate Editor and the referee for their valuable and helpful comments which led to an improved version of this paper.

References

Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: Petrov, B.N., Caki, F. (Eds.), Second International Symposium on Information Theory. Akademiai Kiado, Budapest, pp. 267–281.

Bertsekas, D.P., 1999. Nonlinear Programming, second ed. Athena Scientific, Belmount, MA.

Breheny, P., Huang, J., 2015. Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. Stat. Comput. 25, 173–187.

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

Fan, J., Peng, H., 2004. On non-concave penalized likelihood with diverging number of parameters. Ann. Statist. 32, 928–961.

Fan, J., Xue, L., Zou, H., 2014. Strong oracle optimality of folded concave penalized estimation. Ann. Statist. 42, 819–849.

Huang, J., Breheny, P., Ma, S., 2012. A selective review of group selection in high dimensional models. Statist. Sci. 27, 481–499.

Kim, Y., Choi, H., Oh, H., 2008. Smoothly clipped absolute deviation on high dimensions. J. Amer. Statist. Assoc. 103, 1665–1673.

Mallows, C., 1973. Some comments on Cp. Technometrics 15, 661–667.

Scheetz, T., Kim, K., Swiderski, R., Philp, A., Braun, T., Knudtson, K., Dorrance, A., DiBona, G., Huang, J., Casavant, T., et al., 2006. Regulation of gene expression in the mammalian eye and its relevance to eye disease. Proc. Nat. Acad. Sci. 103, 14429–14434.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Statist. 6, 461–464.

Tibshirani, R., 1996. Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. B 58, 267–288.

Wang, H., Leng, C., 2008. A note on adaptive group lasso. Comput. Statist. Data Anal. 52, 5277–5286.

Wang, L., Chen, G., Li, H., 2007. Group SCAD regression analysis for microarray time course gene expression data. Bioinformatics 23, 1486–1494.

Wei, F., Huang, J., 2010. Consistent group selection in high-dimensional linear regression. Bernoulli 16, 1369–1384.

Xu, Z., Zhang, H., Wang, Y., Chang, X., Liang, Y., 2010. $L_{1/2}$ regularization. Sci. China 40, 411–422.

Xu, Z., Chang, X., Xu, F., Zhang, H., 2012. $L_{1/2}$ regularization: an iterative half thresholding algorithm. IEEE Trans. Neural. Netw. Learn. Syst. 23, 1013–1027.

Yuan, M., Lin, Y., 2006. Model selection and estimation in regression with grouped variables. J. R. Stat. Soc. B 68, 49–67.

Zeng, L., Xie, J., 2012. Group variable selection via SCAD-L2. Statistics 48, 1–18.

Zhang, C., 2010. Nearly unbiased variable selection under minimax concave penalty. Ann. Statist. 38, 894–942.

Zhao, P., Yu, B., 2006. On model selection consistency of lasso. J. Mach. Learn Res. 7, 2541–2563.

Zou, H., 2006. The adaptive lasso and its oracle properties. J. Amer. Statist. Assoc. 101, 1418–1429.