Communications for Statistical Applications and Methods
2020, Vol. 27, No. 4, 445–458

https://doi.org/10.29220/CSAM.2020.27.4.445
Print ISSN 2287-7843 / Online ISSN 2383-4757

Moderately clipped LASSO for the high-dimensional generalized linear model

Sangin Lee[a], Boncho Ku[b], Sunghoon Kown[1,c]

[a]Department of Information and Statistics, Chungnam National University, Korea;
[b]Korea Institute of Oriental Medicine;
[c]Department of Applied Statistics, Konkuk University, Korea

Abstract

The least absolute shrinkage and selection operator (LASSO) is a popular method for a high-dimensional regression model. LASSO has high prediction accuracy; however, it also selects many irrelevant variables. In this paper, we consider the moderately clipped LASSO (MCL) for the high-dimensional generalized linear model which is a hybrid method of the LASSO and minimax concave penalty (MCP). The MCL preserves advantages of the LASSO and MCP since it shows high prediction accuracy and successfully selects relevant variables. We prove that the MCL achieves the oracle property under some regularity conditions, even when the number of parameters is larger than the sample size. An efficient algorithm is also provided. Various numerical studies confirm that the MCL can be a better alternative to other competitors.

Keywords: generalized linear model, moderately clipped LASSO, oracle property, variable selection

1. Introduction

Variable selection is an important issue for the high-dimensional regression model. The fundamental goal of variable selection is to identify relevant predictive variables that can be used to explain how the predictive variables affect the response variable. Traditional approaches such as stepwise selection have limitations such as high computational cost and unstable sampling properties (Breiman, 1996).

As an alternative, the sparse penalized approaches have been developed as effective tools for variable selection and parameter estimation in high-dimensional statistical model. Examples are the least absolute shrinkage and selection operator (LASSO) (Tibshirani, 1996), bridge penalty (Fu, 1998), smoothly clipped absolute deviation (SCAD) (Fan and Peng, 2004) and minimax concave penalty (MCP) (Zhang, 2010). Each penalty has its own desirable properties. For example, LASSO shows high prediction accuracy, but selects more variables than the true underlying model. Theoretically, the LASSO is not selection consistent unless the strong irrepresentable condition in Zhao and Yu (2006) holds. In addition, the nonconvex penalties such as the SCAD and MCP satisfy selection consistency, but it has been empirically shown that their prediction accuracy is not superior than the LASSO (Huang *et al.*, 2016; Kim *et al.*, 2008).

There have been many penalties that combine two different penalties for preserving desirable properties of the combined penalties. Examples include the elastic net (Zou and Hastie, 2005), sparse

[1] Corresponding author: Department of Applied Statistics, Konkuk University, Korea, 120 Neungdong-ro, Gwangjin-gu, Seoul 05029, Korea. E-mail: shkwon0522@gmail.com

Published 31 July 2020 / journal homepage: http://csam.or.kr

ridge (Kwon *et al.*, 2013) and moderately clipped LASSO (Kwon *et al.*, 2015). The elastic net combines the LASSO and ridge for variable selection that also deals with highly correlated variables, which is the same idea for the sparse ridge that combines the MCP and ridge. The moderately clipped LASSO (MCL) is designed to select variables as the MCP and achieve the same shrinkage effect as the LASSO, which is suitable for a high-dimensional linear regression model (Kwon *et al.*, 2015).

In this paper, we study the MCL for the high-dimensional generalized linear model (GLM). We show that under some regularity conditions, the MCL is asymptotically equivalent to an oracle type estimator which is selection consistent even when the number of variables is larger than the sample size. We also develop an efficient algorithm for the MCL by applying the concave-convex procedure (Yuille and Rangarajan, 2003) and modified local quadratic approximation algorithm (Lee *et al.*, 2016). We conduct some numerical studies via simulations and data analysis to show that the MCL can perform better than other penalized estimators for the high-dimensional GLM.

The paper is organized as follows. Section 2 introduces the MCL for the high-dimensional GLM and the computational algorithm. Section 3 proves several theoretical properties of the MCL. Section 4 presents the results of the numerical studies and Section 5 concludes the paper. All proofs and technical details are provided in the Appendix.

2. Moderately clipped LASSO for the GLM

2.1. Definition

Let $y_i \in \mathbb{R}$ and $\boldsymbol{x}_i \in \mathbb{R}^p$, $i \leq n$, be n random samples of response and p predictive variables from the GLM whose conditional density is $f(y|\boldsymbol{x};\boldsymbol{\beta})$ with a regression coefficient vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T \in \mathbb{R}^p$. The MCL for the GLM is defined as the maximizer of the following penalized log-likelihood:

$$Q_{\lambda,\gamma}(\boldsymbol{\beta}) = L(\boldsymbol{\beta}) - \sum_{j=1}^{p} J_{\lambda,\gamma}\left(\left|\beta_j\right|\right), \tag{2.1}$$

where $L(\boldsymbol{\beta}) = (1/n)\sum_{i=1}^{n} \log f(y_i|\boldsymbol{x}_i;\boldsymbol{\beta})$ is the log-likelihood function and $J_{\lambda,\gamma}$ is the MCL penalty (Kwon *et al.*, 2015) that satisfies

$$J_{\lambda,\gamma}(t) = \begin{cases} -\dfrac{t^2}{2a} + \lambda t, & 0 \leq t < a(\lambda - \gamma), \\ \gamma t + \dfrac{a\left(\lambda^2 - \gamma^2\right)}{2}, & t \geq a(\lambda - \gamma), \end{cases}$$

for $\lambda \geq \gamma \geq 0$ and $a > 1$. The MCL penalty is a smooth interpolation between the MCP and LASSO penalty that becomes the MCP when $\gamma = 0$ and the LASSO penalty when $\gamma = \lambda$ (Figure 1). The MCL penalty has two regularization parameters λ and γ, where λ controls the sparsity of the MCL as the MCP and γ decides the amount of shrinkage for the nonzero regression coefficients as the LASSO (Kwon *et al.*, 2015) penalty. Therefor, we expect that the MCL selects relevant predictive variables as the MCP (Zhang, 2010) and produces high prediction accuracy as the LASSO (Zhang and Huang, 2008) for finite samples.

2.2. Algorithm

We introduce the CCCP-MLQA algorithm for maximizing the MCL penalized log-likelihood in (2.1) which is not a concave optimization problem for the non-convexity of the MCL penalty. The CCCP-MLQA algorithm consists of two main steps. The concave-convex procedure (CCCP) (Yuille and

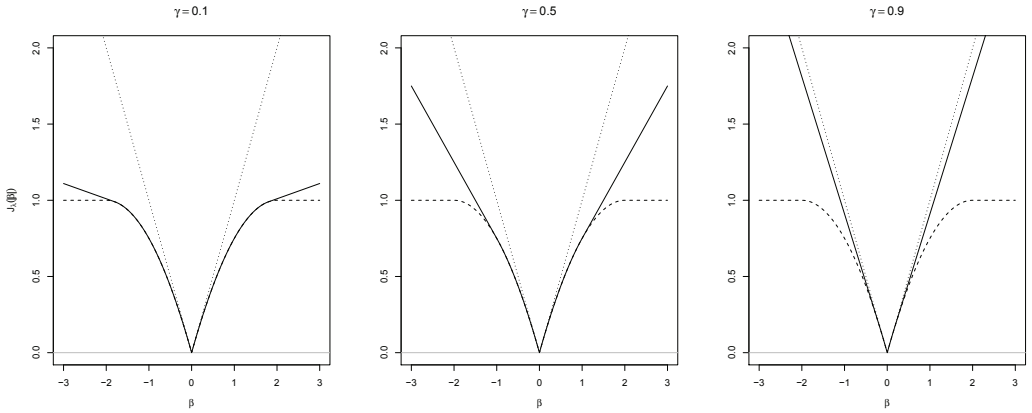

Figure 1: *Plots of the LASSO penalty (dotted), MCP (dashed) and MCL penalty (solid) with various values of* γ *and* $\lambda = 1$. *LASSO = least absolute shrinkage and selection operator; MCP = minimax concave penalty; MCL = moderately clipped LASSO.*

Rangarajan, 2003) converts the original problem into sequential non-quadratic concave optimization problems, and the modified local quadratic approximation (MLQA) solves the problems from the CCCP without losing the ascent property.

The MCL penalty is decomposed as the sum of concave and convex functions:

$$J_{\lambda,\gamma}(|t|) = D_{\lambda,\gamma}(t) + \lambda|t|,$$

where $D_{\lambda,\gamma}(t) = J_{\lambda,\gamma}(|t|) - \lambda|t|$ is a continuously differentiable concave function. Therefor, the objective function $Q_{\lambda,\gamma}$ in (2.1) can be expressed as

$$Q_{\lambda,\gamma}(\boldsymbol{\beta}) = L(\boldsymbol{\beta}) - \sum_{j=1}^{p} D_{\lambda,\gamma}\left(\beta_j\right) - \lambda \sum_{j=1}^{p} \left|\beta_j\right|,$$

which is a sum of convex and concave functions. The CCCP finds a local maximizer of $Q_{\lambda,\gamma}$ by successively maximizing the tight lower concave function obtained from the linear approximation of $D_{\lambda,\gamma}$. For a given current solution $\hat{\boldsymbol{\beta}}^c = (\hat{\beta}_1^c, \ldots, \hat{\beta}_p^c)^T$, the tight lower concave bound becomes

$$U_{\lambda,\gamma}\left(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}^c\right) = L(\boldsymbol{\beta}) - \sum_{j=1}^{p} \partial D_{\lambda,\gamma}\left(\hat{\beta}_j^c\right)\beta_j - \lambda \sum_{j=1}^{p} \left|\beta_j\right|,$$

where $\partial D_{\lambda,\gamma}(\hat{\beta}_j^c)$ is a sub-gradient of $D_{\lambda,\gamma}$ at $\beta_j = \hat{\beta}_j^c$. For maximizing $U_{\lambda,\gamma}(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}^c)$, we apply the MLQA algorithm as follows. The MLQA algorithm iteratively maximizes the local quadratic approximation of L around an initial $\tilde{\boldsymbol{\beta}}$:

$$U_{\lambda,\gamma}^{Q}\left(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}^c, \tilde{\boldsymbol{\beta}}\right) = L_{\lambda,\gamma}^{Q}\left(\boldsymbol{\beta}|\tilde{\boldsymbol{\beta}}\right) - \sum_{j=1}^{p} \partial D_{\lambda,\gamma}\left(\hat{\beta}_j^c\right)\beta_j - \lambda \sum_{j=1}^{p} \left|\beta_j\right|,$$

where $L_{\lambda,\gamma}^{Q}(\boldsymbol{\beta}|\tilde{\boldsymbol{\beta}}) = L(\boldsymbol{\beta}) + \nabla L(\tilde{\boldsymbol{\beta}})^T(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})^T \nabla^2 L(\tilde{\boldsymbol{\beta}})(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}})/2$, $\nabla L(\boldsymbol{\beta}) = \partial L(\boldsymbol{\beta})/\partial \boldsymbol{\beta}$, and $\nabla^2 L(\boldsymbol{\beta}) = \partial^2 L(\boldsymbol{\beta})/\partial \boldsymbol{\beta}^2$. Note that $U_{\lambda,\gamma}^{Q}$ is a simple quadratic function with the LASSO penalty. Therefor, it can

be easily maximized with many existing LASSO algorithms such as the least angle regression (Efron *et al.*, 2004) and coordinate descent (Friedman *et al.*, 2007) algorithms. Let $\tilde{\boldsymbol{\beta}}^a$ be the maximizer of $U^Q_{\lambda,\gamma}$. When $\tilde{\boldsymbol{\beta}}^a$ violates the ascent property then the MLQA algorithm modifies the solution $\tilde{\boldsymbol{\beta}}^a$ by $\tilde{\boldsymbol{\beta}}^{\hat{h}}$ as follows.

$$\tilde{\boldsymbol{\beta}}^{\hat{h}} = \hat{h}\tilde{\boldsymbol{\beta}}^a + \left(1-\hat{h}\right)\tilde{\boldsymbol{\beta}},$$

where $\hat{h} = \text{argmax}_{h>0} U_{\lambda,\gamma}(h\tilde{\boldsymbol{\beta}}^a + (1-h)\tilde{\boldsymbol{\beta}}|\hat{\boldsymbol{\beta}}^c)$. This modification enables the MLQA algorithm to possess the ascent property (Lee *et al.*, 2016). For the readers, we summarize the algorithm in Algorithm 1.

Algorithm 1 The CCCP-MLQA algorithm for maximizing $Q_{\lambda,\gamma}(\boldsymbol{\beta})$

Set an initial estimator $\hat{\boldsymbol{\beta}}$
repeat (CCCP step)
  Compute $\nabla J_\lambda(\boldsymbol{\beta})$ at $\hat{\boldsymbol{\beta}}$
  Set an initial estimator $\tilde{\boldsymbol{\beta}}$
  repeat (MLQA step)
    Compute $\nabla L(\boldsymbol{\beta})$ and $\nabla^2 L(\boldsymbol{\beta})$ at $\tilde{\boldsymbol{\beta}}$
    Find $\tilde{\boldsymbol{\beta}}^a = \arg\max_\beta U^Q_{\lambda,\gamma}(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}^c, \tilde{\boldsymbol{\beta}})$
    Find $\hat{h} = \arg\max_{h>0} U_\lambda(h\tilde{\boldsymbol{\beta}}^a + (1-h)\tilde{\boldsymbol{\beta}})$
    Update $\tilde{\boldsymbol{\beta}}$ by $\boldsymbol{\beta}^{\hat{h}} = \hat{h}\tilde{\boldsymbol{\beta}}^a + (1-\hat{h})\tilde{\boldsymbol{\beta}}$
  until convergence
  Update $\hat{\boldsymbol{\beta}}$ by $\tilde{\boldsymbol{\beta}}$
until convergence

3. Asymptotic properties

The objective function $Q_{\lambda,\gamma}$ is non-concave so that there can be many local maximizers. In this section, we first prove that there exists an oracle estimator among the local maximizers (Fan and Peng, 2004) which we call the oracle LASSO. Second, we prove that the oracle estimator is asymptotically equivalent to the MCL (Kwon and Kim, 2012) since it is the global maximizer.

Let $\boldsymbol{\beta}^*$ be the true regression coefficient vector that satisfy $\beta^*_j \neq 0$, $j \in \mathcal{A}$ and $\beta^*_j = 0$, $j \in \mathcal{A}^c$ for some non-empty subset $\mathcal{A} \subset \{1, \ldots, p\}$. Consider the following penalized maximum likelihood estimator (MLE):

$$\hat{\boldsymbol{\beta}}^{oL,\gamma} = \underset{\beta_j=0, j\in\mathcal{A}^c}{\text{argmax}} \left\{ L(\boldsymbol{\beta}) - \gamma \sum_{j\in\mathcal{A}} |\beta_j| \right\} \tag{3.1}$$

for some $\gamma \geq 0$. We call the MLE in (3.1) the oracle LASSO since it is the LASSO obtained by using the true predictive variables only. When $\gamma = 0$ the oracle LASSO becomes the oracle MLE that have been used in other literatures (Fan and Li, 2001; Fan and Peng, 2004; Kwon and Kim, 2012) for studying SCAD penalized MLE.

We need some regularity conditions from (C1) to (C6) on the true regression coefficient vector and conditional log-likelihood which we give in the Appendix with a remark for the readability. Let $0 < 5c_1 < c_2 \leq 1$ and $k \geq 1$ be the positive constants in (C1) and (C5).

Theorem 1. *(Local optimality) Let $\mathbf{\Omega}_{\lambda,\gamma}$ be the set of all local maximizers of $Q_{\lambda,\gamma}$. Under (C1)–(C5) in Appendix,*

$$\mathbf{P}\left(\hat{\boldsymbol{\beta}}^{oL,\gamma} \in \mathbf{\Omega}_{\lambda,\gamma}\right) \to 1,$$

provided that $\lambda = o(n^{-(1-c_2+c_1)/2})$, $\gamma = o(n^{-(1+c_1)/2})$, and $p = o(n^{(c_2-c_1)k})$ as $n \to \infty$.

Theorem 1 shows that the oracle LASSO becomes a local maximizer of $Q_{\lambda,\gamma}$ for the polynomial order of p that can be larger than n for sufficiently large k. If the moment condition in (C5) holds for any $k \geq 1$ then we can extend the result of Theorem 1 to the exponential order of p. Examples are the linear regression with sub-Gaussian error (Kim *et al.*, 2008) and logistic regression with bounded predictive variables (Kwon and Kim, 2012).

In practice, we cannot identify which estimator is the oracle LASSO since there can be many local maximizers in $\mathbf{\Omega}_{\lambda,\gamma}$. For the problem, we prove that the oracle LASSO is global maximizer of $Q_{\lambda,\gamma}$ with probability tending to one; therefore, the oracle LASSO is asymptotically equivalent to the MCL. It is almost impossible for the asymptotic equivalence to hold on the whole parameter space even in the linear regression model (Zhang, 2010; Kim and Kwon, 2012) when the model is high-dimensional. Therefor we need some restriction called the sparse Riesz condition (Zhang, 2010) stated in (C6). Let $\mathbf{\Gamma}_r = \{\boldsymbol{\beta} : \sum_{j=1}^{p} I(\beta_j = 0) = p - r\} \subset \mathbb{R}^p$ be a restricted parameter space whose number of nonzero elements is at most $r \geq 1$ then the asymptotic equivalence holds as follows.

Theorem 2. *(Global optimality) Under (C1)–(C6) in Appendix, for any $q \leq r < n/2$,*

$$\mathbf{P}\left(\sup_{\boldsymbol{\beta}\in\mathbf{\Gamma}_r} Q_{\lambda,\gamma}(\boldsymbol{\beta}) = Q_{\lambda,\gamma}\left(\hat{\boldsymbol{\beta}}^{oL,\gamma}\right)\right) \to 1,$$

provided that $\lambda = o(n^{-(1-c_2+c_1)/2})$, $\gamma = o(n^{-(1+c_1)/2})$, and $p = o(n^{(c_2-c_1)k})$ as $n \to \infty$.

4. Numerical studies

In this section, we provide the results from numerical studies including simulations and real data analysis. We investigate the finite sample performance of the MCL with different value of γ compared to other competitors: LASSO, SCAD, and MCP.

4.1. Simulation studies

We consider the following three generalized linear models:

- Linear regression: $y|\boldsymbol{x} \sim N(g(\boldsymbol{x}^T\boldsymbol{\beta}^*), 1)$ with identity link, $g(\boldsymbol{x}^T\boldsymbol{\beta}) = \boldsymbol{x}^T\boldsymbol{\beta}$,
- Logistic regression: $y|\boldsymbol{x} \sim$ Bernoulli$(g(\boldsymbol{x}^T\boldsymbol{\beta}^*))$ with logit link, logit$(g(\boldsymbol{x}^T\boldsymbol{\beta})) = \boldsymbol{x}^T\boldsymbol{\beta}$,
- Poisson regression: $y|\boldsymbol{x} \sim$ Poisson$(g(\boldsymbol{x}^T\boldsymbol{\beta}^*))$ with log link, $\log(g(\boldsymbol{x}^T\boldsymbol{\beta})) = \boldsymbol{x}^T\boldsymbol{\beta}$,

where $\boldsymbol{x} = (x_1, \ldots, x_p)^T$ is a multivariate Gaussian random vector with mean zero and covariance structure with $\mathrm{Cov}(x_s, x_t) = 0.5^{|s-t|}$, $s, t \leq p$. For the true regression coefficients, we first set $\tilde{\beta}_j = 1.2e^{-\nu(j-1)}I(j \leq q)$, where ν satisfies $\tilde{\beta}_q = 0.6$. Then we make $\boldsymbol{\beta}^* = \tilde{\boldsymbol{\beta}}/s$ with $s \in \{1, 2\}$ for the linear and logistic regression models, and $s \in \{2, 3\}$ for Poisson regression model.

Table 1: Averages of the measures for the linear regression model, where the corresponding standard errors are in parentheses

Case	n	Method	RMSE	NLV	ME	Correct	Incorrect
$s = 1$	100	LASSO	1.288 (0.008)	0.834 (0.011)	0.664 (0.022)	10.00 (0.000)	27.72 (1.028)
		SCAD	1.392 (0.023)	0.995 (0.034)	0.987 (0.068)	8.99 (0.098)	26.06 (1.050)
		MCP	1.409 (0.025)	1.025 (0.038)	1.048 (0.076)	8.00 (0.136)	3.20 (0.251)
		MCL($\gamma = 2\hat{\gamma}^L$)	1.415 (0.014)	1.011 (0.021)	1.019 (0.042)	9.95 (0.021)	0.32 (0.073)
		MCL($\gamma = \hat{\gamma}^L$)	1.182 (0.007)	0.702 (0.008)	0.401 (0.017)	9.94 (0.023)	1.64 (0.220)
		MCL($\gamma = \hat{\gamma}^L/2$)	1.132 (0.007)	0.643 (0.008)	0.283 (0.016)	9.86 (0.034)	2.95 (0.347)
		MCL($\gamma = \hat{\gamma}^L/4$)	1.199 (0.013)	0.727 (0.016)	0.450 (0.032)	9.48 (0.074)	7.79 (1.225)
		MCL($\gamma = \gamma^*$)	1.125 (0.006)	0.635 (0.007)	0.268 (0.015)	9.87 (0.033)	3.06 (0.216)
	300	LASSO	1.074 (0.002)	0.577 (0.002)	0.153 (0.004)	10.00 (0.000)	28.71 (1.064)
		SCAD	1.027 (0.001)	0.528 (0.001)	0.054 (0.002)	10.00 (0.000)	12.84 (0.936)
		MCP	1.023 (0.001)	0.523 (0.001)	0.044 (0.002)	10.00 (0.000)	2.03 (0.341)
		MCL($\gamma = 2\hat{\gamma}^L$)	1.157 (0.003)	0.670 (0.004)	0.337 (0.008)	10.00 (0.000)	0.03 (0.017)
		MCL($\gamma = \hat{\gamma}^L$)	1.057 (0.002)	0.559 (0.002)	0.115 (0.003)	10.00 (0.000)	1.42 (0.306)
		MCL($\gamma = \hat{\gamma}^L/2$)	1.031 (0.001)	0.531 (0.001)	0.060 (0.002)	10.00 (0.000)	1.63 (0.307)
		MCL($\gamma = \hat{\gamma}^L/4$)	1.025 (0.001)	0.525 (0.001)	0.047 (0.002)	10.00 (0.000)	1.84 (0.344)
		MCL($\gamma = \gamma^*$)	1.022 (0.001)	0.523 (0.001)	0.043 (0.002)	10.00 (0.000)	1.96 (0.344)
$s = 2$	100	LASSO	1.276 (0.008)	0.818 (0.010)	0.632 (0.021)	9.43 (0.071)	26.32 (1.067)
		SCAD	1.396 (0.009)	0.979 (0.014)	0.952 (0.027)	7.04 (0.135)	19.95 (1.537)
		MCP	1.398 (0.011)	0.983 (0.016)	0.960 (0.032)	4.93 (0.100)	4.21 (0.237)
		MCL($\gamma = 2\hat{\gamma}^L$)	1.407 (0.012)	0.998 (0.018)	0.993 (0.036)	8.42 (0.110)	0.85 (0.151)
		MCL($\gamma = \hat{\gamma}^L$)	1.239 (0.008)	0.771 (0.010)	0.538 (0.021)	8.44 (0.108)	6.52 (0.808)
		MCL($\gamma = \hat{\gamma}^L/2$)	1.263 (0.009)	0.802 (0.011)	0.601 (0.022)	7.56 (0.149)	12.31 (1.349)
		MCL($\gamma = \hat{\gamma}^L/4$)	1.325 (0.009)	0.883 (0.012)	0.760 (0.024)	6.92 (0.180)	15.77 (1.674)
		MCL($\gamma = \gamma^*$)	1.235 (0.008)	0.766 (0.010)	0.529 (0.020)	8.09 (0.143)	7.52 (0.861)
	300	LASSO	1.074 (0.002)	0.577 (0.002)	0.153 (0.004)	10.00 (0.000)	28.67 (1.066)
		SCAD	1.104 (0.003)	0.610 (0.003)	0.219 (0.007)	9.62 (0.050)	36.29 (1.094)
		MCP	1.068 (0.003)	0.571 (0.003)	0.139 (0.006)	9.09 (0.075)	8.01 (0.549)
		MCL($\gamma = 2\hat{\gamma}^L$)	1.156 (0.003)	0.669 (0.004)	0.336 (0.008)	9.95 (0.021)	0.04 (0.019)
		MCL($\gamma = \hat{\gamma}^L$)	1.059 (0.002)	0.561 (0.002)	0.120 (0.004)	9.97 (0.017)	2.79 (0.540)
		MCL($\gamma = \hat{\gamma}^L/2$)	1.042 (0.002)	0.543 (0.002)	0.084 (0.004)	9.82 (0.038)	5.05 (0.432)
		MCL($\gamma = \hat{\gamma}^L/4$)	1.048 (0.002)	0.549 (0.002)	0.096 (0.004)	9.61 (0.056)	7.60 (0.531)
		MCL($\gamma = \gamma^*$)	1.041 (0.002)	0.542 (0.002)	0.082 (0.004)	9.79 (0.040)	6.10 (0.588)

RMSE = root mean square error; NLV = negative log-likelihood value; ME = model error; LASSO = least absolute shrinkage and selection operator; SCAD = smoothly clipped absolute deviation; MCP = minimax concave penalty; MCL = moderately clipped LASSO.

For comparison, we consider the LASSO, SCAD, MCP, and MCL where the concavity parameter values for the SCAD, MCP, and MCL are fixed with $a = 3.7, 2.1$, and 1.1, respectively, as recommended in the original literature. For the second tuning parameter γ of the MCL, we consider five cases: four MCLs with fixed $\gamma = \hat{\gamma}^L/2^k$, $k \in \{-1, 0, 1, 2\}$, where $\hat{\gamma}^L$ is the optimally selected λ value of the LASSO and one MCL with γ^* where γ^* is the best one among the γ values of the four MCLs. For each method, the main tuning parameter λ is selected by minimizing the negative log-likelihood values obtained from n independent validation samples.

We compute the root mean square error (RMSE), negative log-likelihood value (NLV), and model error (ME) based on independent test samples of size $N = 5{,}000$ defined as

$$\sqrt{\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{\mu}_i)^2}, \quad -\frac{1}{N}\sum_{i=1}^{N}\log f\left(y_i|\boldsymbol{x}_i;\hat{\boldsymbol{\beta}}\right), \quad \text{and} \quad \frac{1}{N}\sum_{i=1}^{N}\left(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} - \boldsymbol{x}_i^T\boldsymbol{\beta}^*\right)^2$$

Table 2: Averages of the measures for the logistic regression models, where the corresponding standard errors are in parentheses

Case	n	Method	RMSE	NLV	ME	Correct	Incorrect
$s = 1$	100	LASSO	0.390 (0.001)	0.467 (0.003)	11.637 (0.191)	7.08 (0.124)	28.57 (0.914)
		SCAD	0.416 (0.002)	0.534 (0.006)	10.651 (0.380)	2.79 (0.074)	1.38 (0.160)
		MCP	0.419 (0.003)	0.542 (0.007)	10.548 (0.404)	2.25 (0.095)	0.10 (0.033)
		MCL($\gamma = 2\hat{\gamma}^L$)	0.406 (0.002)	0.505 (0.005)	14.429 (0.263)	5.37 (0.158)	4.09 (0.587)
		MCL($\gamma = \hat{\gamma}^L$)	0.388 (0.001)	0.462 (0.003)	11.126 (0.209)	5.28 (0.186)	10.06 (1.267)
		MCL($\gamma = \hat{\gamma}^L/2$)	0.394 (0.002)	0.475 (0.004)	9.430 (0.201)	5.02 (0.190)	15.04 (1.582)
		MCL($\gamma = \hat{\gamma}^L/4$)	0.404 (0.002)	0.505 (0.005)	8.981 (0.245)	4.55 (0.205)	15.91 (1.872)
		MCL($\gamma = \gamma^*$)	0.389 (0.001)	0.464 (0.003)	9.926 (0.233)	4.89 (0.185)	9.55 (1.285)
	300	LASSO	0.328 (0.000)	0.348 (0.001)	6.974 (0.096)	9.65 (0.059)	41.17 (0.454)
		SCAD	0.345 (0.001)	0.376 (0.003)	3.861 (0.135)	5.27 (0.083)	1.75 (0.206)
		MCP	0.346 (0.001)	0.377 (0.003)	3.927 (0.137)	4.96 (0.090)	0.26 (0.061)
		MCL($\gamma = 2\hat{\gamma}^L$)	0.349 (0.001)	0.399 (0.002)	10.790 (0.104)	8.90 (0.082)	0.59 (0.090)
		MCL($\gamma = \hat{\gamma}^L$)	0.325 (0.000)	0.343 (0.001)	6.642 (0.099)	8.61 (0.110)	8.38 (1.057)
		MCL($\gamma = \hat{\gamma}^L/2$)	0.330 (0.001)	0.342 (0.002)	4.381 (0.110)	7.35 (0.155)	11.70 (1.738)
		MCL($\gamma = \hat{\gamma}^L/4$)	0.338 (0.001)	0.358 (0.003)	3.880 (0.137)	6.01 (0.143)	4.40 (1.141)
		MCL($\gamma = \gamma^*$)	0.326 (0.001)	0.339 (0.001)	4.712 (0.166)	7.62 (0.151)	5.99 (1.046)
$s = 2$	100	LASSO	0.450 (0.001)	0.591 (0.003)	3.210 (0.064)	5.33 (0.145)	21.69 (1.186)
		SCAD	0.462 (0.001)	0.624 (0.003)	3.353 (0.089)	2.41 (0.098)	2.38 (0.237)
		MCP	0.462 (0.001)	0.627 (0.004)	3.290 (0.089)	1.49 (0.075)	0.07 (0.025)
		MCL($\gamma = 2\hat{\gamma}^L$)	0.471 (0.002)	0.635 (0.004)	4.239 (0.083)	2.85 (0.167)	1.73 (0.339)
		MCL($\gamma = \hat{\gamma}^L$)	0.449 (0.001)	0.589 (0.003)	3.155 (0.065)	3.75 (0.164)	8.54 (1.279)
		MCL($\gamma = \hat{\gamma}^L/2$)	0.453 (0.002)	0.602 (0.004)	2.994 (0.074)	2.87 (0.163)	7.88 (1.443)
		MCL($\gamma = \hat{\gamma}^L/4$)	0.458 (0.002)	0.618 (0.005)	3.206 (0.091)	2.03 (0.135)	3.96 (1.232)
		MCL($\gamma = \gamma^*$)	0.447 (0.001)	0.587 (0.004)	2.888 (0.073)	3.52 (0.163)	7.52 (1.233)
	300	LASSO	0.405 (0.000)	0.499 (0.001)	1.624 (0.033)	8.56 (0.089)	33.67 (1.062)
		SCAD	0.416 (0.001)	0.520 (0.002)	1.480 (0.044)	3.97 (0.078)	2.52 (0.229)
		MCP	0.417 (0.001)	0.523 (0.002)	1.510 (0.038)	3.51 (0.077)	0.18 (0.043)
		MCL($\gamma = 2\hat{\gamma}^L$)	0.428 (0.001)	0.551 (0.002)	2.885 (0.044)	6.78 (0.129)	0.23 (0.058)
		MCL($\gamma = \hat{\gamma}^L$)	0.404 (0.000)	0.497 (0.001)	1.587 (0.034)	7.16 (0.139)	8.55 (1.073)
		MCL($\gamma = \hat{\gamma}^L/2$)	0.408 (0.001)	0.501 (0.002)	1.370 (0.038)	4.58 (0.091)	1.50 (0.515)
		MCL($\gamma = \hat{\gamma}^L/4$)	0.411 (0.001)	0.509 (0.002)	1.372 (0.038)	4.04 (0.082)	0.53 (0.115)
		MCL($\gamma = \gamma^*$)	0.404 (0.000)	0.496 (0.001)	1.360 (0.037)	5.87 (0.179)	4.99 (0.916)

RMSE = root mean square error; NLV = negative log-likelihood value; ME = model error; LASSO = least absolute shrinkage and selection operator; SCAD = smoothly clipped absolute deviation; MCP = minimax concave penalty; MCL = moderately clipped LASSO.

respectively, where $\hat{\mu}_i = g^{-1}(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})$ is the estimated conditional mean of y_i. We also count the number of nonzero regression coefficients selected correctly (Correct) and incorrectly (Incorrect). We repeat the simulation 100 times for $n = \{100, 300\}$, $p = 2000$, and $q = 10$ and summarize the averages of the measures in Tables 1–3.

Table 1 presents the simulation results in the linear regression model. When the sample size is small, the LASSO selects correct predictive variables better than the others keeping higher prediction accuracy but selects too many incorrect variables simultaneously, which confirms that the LASSO overfits the model (Zhang and Huang, 2008). While the sample size is large, the MCP and SCAD are better than the LASSO for both prediction and selectivity measures. The MCL with γ^* shows the best performance regardless of the cases which may be because it has one more tuning parameter γ. We recommend to use the MCLs with $\gamma \le \hat{\gamma}^L$ that perform better than the others also if reducing the computational cost is important. Table 2 and 3 shows the results in the logistic and Poisson regression models where the results are similar to the linear regression model.

Table 3: Averages of the measures for the Poisson regression models, where the corresponding standard errors are in parentheses

Case	n	Method	RMSE	NLV	ME	Correct	Incorrect
$s = 2$	100	LASSO	115.160 (13.666)	10.647 (0.704)	1.094 (0.056)	9.52 (0.074)	29.28 (0.789)
		SCAD	109.725 (13.580)	9.561 (0.684)	1.123 (0.059)	8.46 (0.153)	26.26 (0.763)
		MCP	105.051 (12.769)	8.635 (0.569)	1.116 (0.067)	7.20 (0.158)	13.13 (0.396)
		MCL($\gamma = 2\hat{\gamma}^L$)	115.303 (13.858)	11.094 (0.850)	1.584 (0.086)	8.42 (0.129)	12.77 (0.343)
		MCL($\gamma = \hat{\gamma}^L$)	102.650 (13.267)	7.635 (0.589)	0.863 (0.060)	8.95 (0.111)	13.49 (0.353)
		MCL($\gamma = \hat{\gamma}^L/2$)	90.915 (12.667)	6.560 (0.514)	0.675 (0.059)	8.47 (0.131)	13.47 (0.346)
		MCL($\gamma = \hat{\gamma}^L/4$)	92.097 (12.401)	7.267 (0.481)	0.870 (0.063)	7.65 (0.141)	13.63 (0.411)
		MCL($\gamma = \gamma^*$)	89.929 (12.486)	6.091 (0.468)	0.641 (0.057)	8.61 (0.138)	13.98 (0.351)
	300	LASSO	66.511 (6.925)	2.764 (0.101)	0.253 (0.016)	10.00 (0.000)	32.75 (0.863)
		SCAD	45.700 (4.282)	2.156 (0.065)	0.145 (0.018)	10.00 (0.000)	32.93 (0.875)
		MCP	22.972 (2.610)	1.689 (0.053)	0.053 (0.015)	10.00 (0.000)	19.66 (1.008)
		MCL($\gamma = 2\hat{\gamma}^L$)	71.708 (7.751)	3.105 (0.149)	0.445 (0.033)	9.96 (0.031)	9.27 (0.327)
		MCL($\gamma = \hat{\gamma}^L$)	49.321 (5.509)	2.106 (0.069)	0.162 (0.016)	10.00 (0.000)	10.94 (0.355)
		MCL($\gamma = \hat{\gamma}^L/2$)	34.005 (3.663)	1.798 (0.055)	0.079 (0.015)	10.00 (0.000)	14.87 (0.556)
		MCL($\gamma = \hat{\gamma}^L/4$)	27.651 (2.967)	1.718 (0.053)	0.060 (0.015)	10.00 (0.000)	17.42 (0.774)
		MCL($\gamma = \gamma^*$)	23.145 (2.629)	1.680 (0.052)	0.052 (0.015)	10.00 (0.000)	20.71 (0.977)
$s = 3$	100	LASSO	7.721 (0.277)	2.382 (0.038)	0.643 (0.022)	8.49 (0.113)	26.79 (0.859)
		SCAD	7.596 (0.306)	2.368 (0.035)	0.648 (0.021)	7.36 (0.169)	25.30 (0.918)
		MCP	7.844 (0.416)	2.409 (0.041)	0.699 (0.025)	5.36 (0.102)	9.41 (0.387)
		MCL($\gamma = 2\hat{\gamma}^L$)	8.014 (0.309)	2.513 (0.046)	0.938 (0.032)	6.97 (0.146)	7.75 (0.401)
		MCL($\gamma = \hat{\gamma}^L$)	7.202 (0.238)	2.229 (0.035)	0.595 (0.022)	7.24 (0.138)	11.70 (0.567)
		MCL($\gamma = \hat{\gamma}^L/2$)	7.319 (0.610)	2.204 (0.036)	0.513 (0.022)	6.63 (0.130)	13.47 (0.751)
		MCL($\gamma = \hat{\gamma}^L/4$)	7.342 (0.335)	2.336 (0.040)	0.604 (0.023)	5.86 (0.120)	12.73 (0.838)
		MCL($\gamma = \gamma^*$)	6.939 (0.224)	2.222 (0.037)	0.548 (0.022)	6.82 (0.160)	13.69 (0.629)
	300	LASSO	4.354 (0.138)	1.547 (0.007)	0.164 (0.005)	9.98 (0.014)	30.45 (0.919)
		SCAD	3.935 (0.119)	1.530 (0.007)	0.135 (0.005)	9.96 (0.019)	29.01 (0.954)
		MCP	3.361 (0.127)	1.500 (0.007)	0.098 (0.004)	9.56 (0.062)	15.63 (1.064)
		MCL($\gamma = 2\hat{\gamma}^L$)	4.911 (0.167)	1.635 (0.010)	0.321 (0.010)	9.73 (0.046)	3.25 (0.228)
		MCL($\gamma = \hat{\gamma}^L$)	3.670 (0.110)	1.487 (0.005)	0.124 (0.004)	9.89 (0.031)	8.30 (0.379)
		MCL($\gamma = \hat{\gamma}^L/2$)	3.092 (0.080)	1.456 (0.004)	0.072 (0.003)	9.83 (0.042)	16.84 (0.760)
		MCL($\gamma = \hat{\gamma}^L/4$)	3.029 (0.077)	1.464 (0.005)	0.070 (0.003)	9.76 (0.047)	20.56 (0.949)
		MCL($\gamma = \gamma^*$)	3.017 (0.078)	1.457 (0.005)	0.069 (0.003)	9.79 (0.045)	19.28 (0.869)

RMSE = root mean square error; NLV = negative log-likelihood value; ME = model error; LASSO = least absolute shrinkage and selection operator; SCAD = smoothly clipped absolute deviation; MCP = minimax concave penalty; MCL = moderately clipped LASSO.

4.2. Real data analysis

We analyze *colon* cancer data from a gene expression study of 22 normal and 40 tumor colon tissue samples analyzed with more than 6,500 human genes. We consider 2,000 genes selected with the highest minimal intensity across the samples (Alon, 1999). This data set is available from the R package **sdwd**. We apply penalized logistic regression model for the samples. For comparison, we divide the samples into two parts by randomly selecting 50 samples for training and 12 samples for test. Optimal values of regularization parameters are selected by 10-fold cross-validation on the training samples. We then evaluate the RMSE, NLV and mis-classification errors (MIS) based on test samples, and we also compute the model size (MS) as the number of estimated nonzero coefficients.

We repeat this whole procedure 100 times, and summarize the results in Table 4. As expected, the LASSO performs better than MCP and SCAD in terms of prediction accuracy, but it produces the largest model. The MCL with γ^* or $\hat{\gamma}^L/2$ perform the best in prediction while selecting smaller number of variables than the LASSO, which confirm the results in the simulations.

Table 4: Averages of the measures for Colon data analysis, where the numbers in parentheses are the corresponding standard errors

Method	RMSE	Likelihood	MIS	MS
LASSO	0.3692 (0.0054)	0.4594 (0.0125)	0.1667 0.0085)	8.09 (0.4983)
SCAD	0.4013 (0.0080)	0.5201 (0.0186)	0.2550 0.0100)	1.60 (0.0853)
MCP	0.4049 (0.0088)	0.5330 (0.0208)	0.2417 0.0111)	0.68 (0.0510)
MCL($\gamma = 2\hat{\gamma}^L$)	0.4011 (0.0040)	0.5041 (0.0072)	0.2100 0.0102)	3.46 (0.1920)
MCL($\gamma = \hat{\gamma}^L$)	0.4395 (0.0037)	0.5750 (0.0068)	0.3008 0.0076)	2.45 (0.2657)
MCL($\gamma = \hat{\gamma}^L/2$)	0.3690 (0.0042)	0.4463 (0.0075)	0.1475 0.0087)	5.80 (0.3651)
MCL($\gamma = \hat{\gamma}^L/4$)	0.3624 (0.0058)	0.4410 (0.0124)	0.1692 0.0094)	5.08 (0.5477)
MCL($\gamma = \gamma^*$)	0.3658 (0.0060)	0.4538 (0.0148)	0.1642 0.0094)	4.58 (0.4425)

RMSE = root mean square error; MIS = mis-classification errors; MS = model size; LASSO = least absolute shrinkage and selection operator; SCAD = smoothly clipped absolute deviation; MCP = minimax concave penalty; MCL = moderately clipped LASSO.

5. Concluding remarks

In this paper, we study the MCL for the high-dimensional GLM in two aspects: theories and numerical analysis. We prove that the MCL is asymptotically equivalent the oracle LASSO under some regularity conditions, which can be theoretical supports for nice finite sample performance of the MCL. As a promising alternative to current existing penalized estimators, the MCL can be further applied to other statistical models such as Cox's regression and Gaussian graphical models. This paper represents a fundamental basis for future works.

Acknowledgements

This research was supported by Chungnam National University grant and by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (NO. NRF-2017R1D1A1 B03031239).

Appendix A: Regularity conditions

Let $\ell_i(\boldsymbol{\beta}) = \log f(y_i|\boldsymbol{x}_i;\boldsymbol{\beta})$, $i \le n$. For any vector $\mathbf{a}$ and subset $\mathcal{S} \subset \{1,\ldots,p\}$, let $\mathbf{a}_\mathcal{S}$ be the sub-vector whose element indices are in $\mathcal{S}$. Similarly, let $\mathbf{A}_\mathcal{S}$ be the sub-matrix whose column and row indices are in $\mathcal{S}$ for any matrix $\mathbf{A}$.

(C1) There exist positive constants c_1, c_2, and M_1 such that $5c_1 < c_2 \le 1$,

$$q = O\left(n^{c_1}\right) \quad \text{and} \quad m^* \ge M_1 n^{-\frac{1-c_2}{2}},$$

where $m^* = \min_{j\in\mathcal{A}} |\beta_j^*|$ is the minimum of the signal size.

(C2) The first and second derivatives of the log-likelihood satisfy

$$E_{\boldsymbol{\beta}^*}\left\{\frac{\partial \ell_1(\boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}}\right\} = \mathbf{0}_p \quad \text{and} \quad E_{\boldsymbol{\beta}^*}\left\{\frac{\partial^2 \ell_1(\boldsymbol{\beta}^*)}{\partial \boldsymbol{\beta}^2}\right\} = -\mathbf{I}(\boldsymbol{\beta}^*),$$

where $\mathbf{I}(\boldsymbol{\beta}^*) = E_{\boldsymbol{\beta}^*}\{(\partial \ell_1(\boldsymbol{\beta}^*)/\partial\boldsymbol{\beta})(\partial \ell_1(\boldsymbol{\beta}^*)/\partial\boldsymbol{\beta})^T\}$ is the Fisher information matrix.

(C3) There exist positive constants M_2 and M_3 such that

$$0 < M_2 \le \tau_{\min}\left(\mathbf{I}_\mathcal{A}(\boldsymbol{\beta}^*)\right) \le \tau_{\max}\left(\mathbf{I}_\mathcal{A}(\boldsymbol{\beta}^*)\right) \le M_3 < \infty,$$

where $\tau_{\min}$ and $\tau_{\max}$ denote the minimum and maximum eigenvalues, respectively.

(C4) There exists an open subset $\mathbf{\Gamma} \in \mathbb{R}^p$ that contains the true coefficients vector $\boldsymbol{\beta}^*$ such that the density admits all third derivatives for almost all $\mathbf{z}_i = (y_i, \boldsymbol{x}_i^T)^T$, $i \leq n$. Furthermore, there exists a function U such that

$$\sup_{1\leq k,l,m\leq p} \left| \frac{\partial^3 \ell_1(\boldsymbol{\beta})}{\partial\beta_k \partial\beta_l \partial\beta_m} \right| < U(\mathbf{z}_1),$$

for all $\boldsymbol{\beta} \in \mathbf{\Gamma}$.

(C5) There exists an integer $k \geq 1$ and positive constants M_4, M_5, and M_6 such that

$$\sup_{1\leq k\leq p} E_{\boldsymbol{\beta}^*} \left\{ \frac{\partial \ell_1(\boldsymbol{\beta}^*)}{\partial\beta_k} \right\}^{2k} < M_4, \quad \sup_{1\leq k,l\leq p} E_{\boldsymbol{\beta}^*} \left\{ \frac{\partial^2 \ell_1(\boldsymbol{\beta}^*)}{\partial\beta_k \partial\beta_l} \right\}^{2k} < M_5, \quad \text{and} \quad E_{\boldsymbol{\beta}^*}\{U(\mathbf{z}_1)\}^{2k} < M_6.$$

(C6) There exists a positive constant M_7 such that

$$\inf_{q\leq r<\frac{n}{2}} \min_{|\mathcal{S}|\leq r} \inf_{\boldsymbol{\beta}\in\mathbf{\Gamma}_{\mathcal{S}}} \tau_{\min}\left(\mathbf{H}_{\mathcal{S}}(\boldsymbol{\beta})\right) > M_7,$$

for all sufficiently large n, where $\mathbf{\Gamma}_{\mathcal{S}} = \{\boldsymbol{\beta} : \beta_j = 0, j \in \mathcal{S}^c\}$ is the restricted parameter space with respect to a subset $\mathcal{S}$ and $\mathbf{H}(\boldsymbol{\beta}) = -\nabla^2 L(\boldsymbol{\beta})$ is the negative Hessian matrix.

Remark 1. Condition (C1) allows the number of true non-zero regression coefficients to diverge to infinity and their values to decrease toward zero (Kwon and Kim, 2012). Conditions (C2)–(C4) are standard for the asymptotic theories in the maximum likelihood estimation (Fan and Peng, 2004; Kwon and Kim, 2012). Condition (C5) represents the tail behavior of the conditional density, which determines the order of p with respect to some positive integer k. The linear regression with sub-Gaussian tail errors and logistic regression with bounded predictive variables satisfy condition (C5) (Kim *et al.*, 2008; Kwon and Kim, 2012). Condition (C6) corresponds to the *sparse Riesz condition* (Zhang and Zhang, 2012; Kim and Kwon, 2012) studied in the linear regression, which guarantees the strict concavity of the log-likelihood function on the restricted parameter spaces. In the linear regression, condition (C6) implies that all of sub-design matrices whose number of columns are less than the sample size are non-singular.

Appendix B: Proofs

Let $\nabla_{\mathcal{A}} S_j(\boldsymbol{\beta}) = \partial S_j(\boldsymbol{\beta})/\partial\boldsymbol{\beta}_{\mathcal{A}}$ and $\nabla^2_{\mathcal{A}} S_j(\boldsymbol{\beta}) = \partial^2 S_j(\boldsymbol{\beta})/\partial\boldsymbol{\beta}^2_{\mathcal{A}}$, where $S_j(\boldsymbol{\beta})$, $j \leq p$ is the j^{th} element of $\nabla L(\boldsymbol{\beta})$. We need some lemmas below whose proofs are omitted.

Lemma 1. *(Kwon et al., 2015) If $\hat{\boldsymbol{\beta}}$ satisfies*

$$\min_{j\in\mathcal{S}} \left|\hat{\beta}_j\right| > a(\lambda - \gamma), \quad \max_{j\in\mathcal{S}^c} \left|S_j\left(\hat{\boldsymbol{\beta}}\right)\right| \leq \lambda, \quad \textit{and} \quad S_j\left(\hat{\boldsymbol{\beta}}\right) = \gamma sign\left(\hat{\beta}_j\right), \quad j \in \mathcal{S}$$

then $\hat{\boldsymbol{\beta}} \in \mathbf{\Omega}_{\lambda,\gamma}$, where $\mathcal{S} = \{j : \hat{\beta}_j \neq 0\}$.

Lemma 2. *(Fan and Peng, 2004) Under (C2)–(C5),*

$$\left\|\hat{\boldsymbol{\beta}}^{oL,\gamma} - \boldsymbol{\beta}^*\right\| = O_p\left(\sqrt{\frac{q}{n}}\right) \quad \textit{as } n \to \infty.$$

Lemma 3. *(Kwon and Kim, 2012) Under (C2)–(C5),*

$$\mathbf{P}\left(\left|S_j(\boldsymbol{\beta}^*)\right| > \frac{\alpha}{\sqrt{n}}\right) = O\left(\alpha^{-2k}\right), \quad \mathbf{P}\left(\left\|\nabla^2_{\mathcal{A}} S_j(\boldsymbol{\beta})\right\| > q\alpha\right) = O\left(\alpha^{-2k}\right), \quad \textit{and}$$
$$\mathbf{P}\left(\left\|\nabla_{\mathcal{A}} S_j(\boldsymbol{\beta}^*) - E\nabla_{\mathcal{A}} S_j(\boldsymbol{\beta}^*)\right\| > \frac{\alpha\sqrt{q}}{n}\right) = O\left(\alpha^{-2k}\right)$$

as $n \to \infty$ for any $\alpha > 0$, $j \le p$, and $\boldsymbol{\beta} \in \boldsymbol{\Gamma}$.

Proof of Theorem 1: The first order Karush-Kuhn-Tucker necessary conditions imply that

$$\begin{cases} S_j\left(\hat{\boldsymbol{\beta}}^{oL,\gamma}\right) = \gamma \text{sign}\left(\hat{\beta}_j^{oL,\gamma}\right), & \hat{\beta}_j^{oL,\gamma} \neq 0, \\ \left|S_j\left(\hat{\boldsymbol{\beta}}^{oL,\gamma}\right)\right| \le \gamma, & \hat{\beta}_j^{oL,\gamma} = 0, \end{cases} \tag{B.1}$$

for all $j \in \mathcal{A}$. Therefore, it suffices to show that

$$\mathbf{P}\left(\min_{j\in\mathcal{A}}\left|\hat{\beta}_j^{oL,\gamma}\right| > a(\lambda-\gamma)\right) \to 1 \quad \text{and} \quad \mathbf{P}\left(\max_{j\in\mathcal{A}^c}\left|S_j\left(\hat{\boldsymbol{\beta}}^{oL,\gamma}\right)\right| < \lambda\right) \to 1 \tag{B.2}$$

as $n \to \infty$. From (C1) and Lemma 2,

$$\min_{j\in\mathcal{A}}\left|\hat{\beta}_j^{oL,\gamma}\right| \ge \min_{j\in\mathcal{A}}\left|\beta_j^*\right| - \max_{j\in\mathcal{A}}\left|\hat{\beta}_j^{oL,\gamma} - \beta_j^*\right| = O_p\left(n^{-\frac{1-c_2}{2}}\right),$$

as $n \to \infty$, which implies the first part in (B.2). From Taylor's expansion, there exists $\boldsymbol{\beta}^{**}$ that lies between $\hat{\boldsymbol{\beta}}^{oL,\gamma}$ and $\boldsymbol{\beta}^*$ such that

$$S_j\left(\hat{\boldsymbol{\beta}}^{oL,\gamma}\right) = S_j(\boldsymbol{\beta}^*) + \nabla_{\mathcal{A}} S_j(\boldsymbol{\beta}^*)^T\left(\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{oL,\gamma} - \boldsymbol{\beta}_{\mathcal{A}}^*\right) + \frac{\left(\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{oL,\gamma} - \boldsymbol{\beta}_{\mathcal{A}}^*\right)^T \nabla^2_{\mathcal{A}} S_j(\boldsymbol{\beta}^{**})\left(\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{oL,\gamma} - \boldsymbol{\beta}_{\mathcal{A}}^*\right)}{2}$$

for all $j \in \mathcal{A}^c$. From Cauchy-Schwarz inequality,

$$\begin{aligned} \mathbf{P}\left(\max_{j\in\mathcal{A}^c}\left|S_j\left(\hat{\boldsymbol{\beta}}^{oL,\gamma}\right)\right| > \lambda\right) \le{}& \mathbf{P}\left(\max_{j\in\mathcal{A}^c}\left|S_j(\boldsymbol{\beta}^*)\right| > \frac{\lambda}{4}\right) \\ &+ \mathbf{P}\left(\max_{j\in\mathcal{A}^c}\left\|\nabla_{\mathcal{A}} S_j(\boldsymbol{\beta}^*) - E\nabla_{\mathcal{A}} S_j(\boldsymbol{\beta}^*)\right\|\left\|\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{oL,\gamma} - \boldsymbol{\beta}_{\mathcal{A}}^*\right\| > \frac{\lambda}{4}\right) \\ &+ \mathbf{P}\left(\max_{j\in\mathcal{A}^c}\left\|E\nabla_{\mathcal{A}} S_j(\boldsymbol{\beta}^*)\right\|\left\|\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{oL,\gamma} - \boldsymbol{\beta}_{\mathcal{A}}^*\right\| > \frac{\lambda}{4}\right) \\ &+ \mathbf{P}\left(\max_{j\in\mathcal{A}^c}\left\|\nabla^2_{\mathcal{A}} S_j(\boldsymbol{\beta}^{**})\right\|_F\left\|\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{oL,\gamma} - \boldsymbol{\beta}_{\mathcal{A}}^*\right\|^2 > \frac{\lambda}{2}\right) \\ \overset{\text{let}}{=}{}& \mathbf{P}_1 + \mathbf{P}_2 + \mathbf{P}_3 + \mathbf{P}_4, \end{aligned}$$

where $\|\cdot\|_F$ is Frobenius norm. From Lemma 3,

$$\mathbf{P}_1 \le \sum_{j\in\mathcal{A}^c}\mathbf{P}\left(\left|S_j(\boldsymbol{\beta}^*)\right| > \frac{\lambda}{4}\right) = O\left(\frac{p}{\left(\sqrt{n}\lambda\right)^{2k}}\right) \to 0$$

as $n \to \infty$. Similarly,

$$\begin{aligned}\mathbf{P}_2 &\le \mathbf{P}\left(\left\|\left(\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{oL,\gamma} - \boldsymbol{\beta}_{\mathcal{A}}^{*}\right)\right\| > \frac{q}{\sqrt{n}}\right) + \mathbf{P}\left(\max_{j\in\mathcal{A}^c}\left\|\nabla_{\mathcal{A}} S_j(\boldsymbol{\beta}^*) - E\nabla_{\mathcal{A}} S_j(\boldsymbol{\beta}^*)\right\| > \frac{\sqrt{n}\lambda}{4q}\right) \\ &= o(1) + O\left(p\Big/\left(\frac{n\sqrt{n}\lambda}{q\sqrt{q}}\right)^{2k}\right) \to 0\end{aligned}$$

as $n \to \infty$. From (C5),

$$\begin{aligned}\mathbf{P}_3 &= \mathbf{P}\left(\max_{j\in\mathcal{A}^c}\left\|E\nabla_{\mathcal{A}} S_j(\boldsymbol{\beta}^*)\right\| \left\|\left(\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{oL,\gamma} - \boldsymbol{\beta}_{\mathcal{A}}^{*}\right)\right\| > \frac{\lambda}{4}\right) \\ &\le \mathbf{P}\left(\left\|\left(\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{oL,\gamma} - \boldsymbol{\beta}_{\mathcal{A}}^{*}\right)\right\| > \frac{\lambda}{4M_5\sqrt{q}}\right) \to 0\end{aligned}$$

as $n \to \infty$. Lemma 3 implies that

$$\begin{aligned}\mathbf{P}_4 &\le \mathbf{P}\left(\left\|\left(\hat{\boldsymbol{\beta}}_{\mathcal{A}}^{oL,\gamma} - \boldsymbol{\beta}_{\mathcal{A}}^{*}\right)\right\|^2 > \frac{q\sqrt{q}}{n}\right) + \mathbf{P}\left(\max_{j\in\mathcal{A}^c}\left\|E\nabla^2_{\mathcal{A}} S_j(\boldsymbol{\beta}^{**})\right\|_F > \frac{n\lambda}{2q\sqrt{q}}\right) \\ &= o(1) + O\left(p\Big/\left(\frac{n\lambda}{q^2\sqrt{q}}\right)^{2k}\right) \to 0\end{aligned}$$

as $n \to \infty$. This completes the proof. □

Lemma 4. (Low dimensional global optimality) *Assume that $p < n$. Under (C1)–(C6),*

$$\mathbf{P}\left(\sup_{\boldsymbol{\beta}\in\mathbb{R}^p} Q_{\lambda,\gamma}(\boldsymbol{\beta}) \le Q_{\lambda,\gamma}\left(\hat{\boldsymbol{\beta}}^{oL,\gamma}\right)\right) \to 1,$$

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

Proof of Lemma 4: From Taylor's expansion,

$$L(\boldsymbol{\beta}) - L\left(\hat{\boldsymbol{\beta}}^{oL,\gamma}\right) = S\left(\hat{\boldsymbol{\beta}}^{oL,\gamma}\right)^T\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{oL,\gamma}\right) + \frac{\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{oL,\gamma}\right)^T \nabla S(\boldsymbol{\beta}^{**})\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{oL,\gamma}\right)}{2}$$

for some $\boldsymbol{\beta}^{**}$ lies between $\boldsymbol{\beta}$ and $\hat{\boldsymbol{\beta}}^{oL,\gamma}$. From (B.1) and (B.2),

$$S\left(\hat{\boldsymbol{\beta}}^{oL,\gamma}\right)^T\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{oL,\gamma}\right) = \sum_{j=1}^{p} S_j\left(\hat{\boldsymbol{\beta}}^{oL,\gamma}\right)\left(\beta_j - \hat{\beta}_j^{oL,\gamma}\right) \le \sum_{j\in\mathcal{A}^c} o_p(\lambda)|\beta_j|.$$

(C6) and Cauchy-Schwarz inequality imply that

$$\frac{\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{oL,\gamma}\right)^T \nabla S(\boldsymbol{\beta}^{**})\left(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{oL,\gamma}\right)}{2} \le -M_7\left\|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}^{oL,\gamma}\right\|^2.$$

Therefore, it follows that

$$Q_{\lambda,\gamma}(\boldsymbol{\beta}) - Q_{\lambda,\gamma}\left(\hat{\boldsymbol{\beta}}^{oL,\gamma}\right) \leq \sum_{j=1}^{p} w_j\left(\beta_j\right),$$

where

$$w_j(\beta_j) = o_p(\lambda)\left|\beta_j\right| I\left(j \in \mathcal{A}^c\right) - M_7\left(\beta_j - \hat{\beta}_j^{oL,\gamma}\right)^2 + \left\{J_{\lambda,\gamma}\left(\hat{\beta}_j^{oL,\gamma}\right) - J_{\lambda,\gamma}\left(\beta_j\right)\right\}.$$

First, consider the case $j \in \mathcal{A}$, where $S_j(\hat{\boldsymbol{\beta}}^{oL,\gamma}) = \gamma\text{sign}(\hat{\beta}_j^{oL,\gamma})$. If $|\beta_j| \geq a(\lambda - \gamma)$, then

$$w_j(\beta_j) \leq -M_7\left(\beta_j - \hat{\beta}_j^{oL,\gamma}\right)^2 < 0.$$

If $|\beta_j| < a(\lambda - \gamma)$, then

$$\left|\beta_j - \hat{\beta}_j^{oL,\gamma}\right| \geq \min_{j\in\mathcal{A}}\left|\beta_j^*\right| - \max_{j\in\mathcal{A}}\left|\hat{\beta}_j^{oL,\gamma} - \beta_j^*\right| - a(\lambda - \gamma) = O_p\left(n^{-\frac{1-c_2}{2}}\right)$$

and

$$J_{\lambda,\gamma}\left(\hat{\beta}_j^{oL,\gamma}\right) - J_{\lambda,\gamma}\left(\beta_j\right) \leq \nabla J_{\lambda,\gamma}\left(\left|\beta_j\right|\right)\left(\left|\hat{\beta}_j^{oL,\gamma}\right| - \left|\beta_j\right|\right). \tag{B.3}$$

Therefore, we have

$$w_j\left(\beta_j\right) \leq -M_7\left(\beta_j - \hat{\beta}_j^{oL,\gamma}\right)^2 + \lambda\left|\hat{\beta}_j^{oL,\gamma}\right| \leq -O_p\left(n^{-(1-c_2)}\right) + o_p\left(n^{-\left(1-c_2+\frac{c_1}{2}\right)}\right) \leq 0$$

for sufficiently large n. Second, consider the case $j \in \mathcal{A}^c$, where $\hat{\beta}_j^{oL,\gamma} = 0$. If $|\beta_j| \geq a(\lambda - \gamma)$ then

$$w_j(\beta_j) \leq \left|\beta_j\right|\left(o_p(\lambda) - M_7\left|\beta_j\right|\right) \leq 0.$$

If $|\beta_j| < a(\lambda - \gamma)$, then from (B.2) and (B.3)

$$w_j\left(\beta_j\right) \leq \left|\beta_j\right|\left(o_p(\lambda) - \nabla J_{\lambda,\gamma}\left(\left|\beta_j\right|\right)\right) = \left|\beta_j\right|\left(o_p(\lambda) - \lambda\right) \leq 0.$$

Therefore, $Q_{\lambda,\gamma}(\boldsymbol{\beta}) - Q_{\lambda,\gamma}(\hat{\boldsymbol{\beta}}^{oL,\gamma}) \leq 0$ for sufficiently large n. This completes the proof. □

Proof of Theorem 2: Suppose that there is another global maximizer $\hat{\boldsymbol{\beta}} \in \boldsymbol{\Gamma}_r$ such that $Q_{\lambda,\gamma}(\hat{\boldsymbol{\beta}}) > Q_{\lambda,\gamma}(\hat{\boldsymbol{\beta}}^{oL,\gamma})$. Let $\mathcal{S} = \{j : \hat{\beta}_j \neq 0\} \cup \{j : \hat{\beta}_j^{oL,\gamma} \neq 0\}$. By Lemma 4 and (C6), it is easy to see that $\sup_{\boldsymbol{\beta}_\mathcal{S}\in\boldsymbol{\Gamma}_\mathcal{S}} Q_{\lambda,\gamma}(\boldsymbol{\beta}_\mathcal{S}) \leq Q_{\lambda,\gamma}(\hat{\boldsymbol{\beta}}_\mathcal{S}^{oL,\gamma})$ since $|\mathcal{S}| \leq 2r \leq n$. Therefore, it follows that

$$Q_{\lambda,\gamma}\left(\hat{\boldsymbol{\beta}}\right) = Q_{\lambda,\gamma}\left(\hat{\boldsymbol{\beta}}_\mathcal{S}\right) \leq \sup_{\boldsymbol{\beta}_\mathcal{S}\in\boldsymbol{\Gamma}_\mathcal{S}} Q_{\lambda,\gamma}(\boldsymbol{\beta}_\mathcal{S}) \leq Q_{\lambda,\gamma}\left(\hat{\boldsymbol{\beta}}_\mathcal{S}^{oL,\gamma}\right) = Q_{\lambda,\gamma}\left(\hat{\boldsymbol{\beta}}^{oL,\gamma}\right),$$

which is a contradiction. □

References

Breiman L (1996). Heuristics of instability and stabilization in model selection, *The Annals of Statistics*, **24**, 2350–2383.

Efron B, Hastie T, Johnstone I, Tibshirani R (2004). Least angle regression, *The Annals of Statistics*, **32** 407–499.

Fan J and Li R (2001). Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association*, **96** 1348–1360.

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

Friedman J, Hastie T, Höfling H, and Tibshirani R (2007). Pathwise coordinate optimization, *The Annals of Applied Statistics*, **1** 302–332.

Fu WJ (1998). Penalized regressions: the bridge versus the lasso, *Journal of Computational and Graphical Statistics*, **7**, 397–416.

Huang J, Breheny P, Lee S, Ma S, and Zhang C (2016). The Mnet method for variable selection, *Statistica Sinica*, **26**, 903–923.

Kim Y, Choi H, and Oh HS (2008). Smoothly clipped absolute deviation on high dimensions, *Journal of the American Statistical Association*, **103**, 1665–1673.

Kim Y and Kwon S (2012). Global optimality of nonconvex penalized estimators. *Biometrika*, **99**, 315–325.

Kwon S and Kim Y (2012). Large sample properties of the scad-penalized maximum likelihood estimation on high dimensions, *Statistica Sinica*, **22**, 629–653.

Kwon S, Kim Y, and Choi H (2013). Sparse bridge estimation with a diverging number of parameters, *Statistics and Its Interface*, **6**, 231–242.

Kwon S, Lee S, and Kim Y (2015). Moderately clipped lasso, *Computational Statistics & Data Analysis*, **92**, 53–67.

Lee S, Kwon S, and Kim Y (2016). A modified local quadratic approximation algorithm for penalized optimization problems, *Computational Statistics & Data Analysis*, **94**, 275–286.

Tibshirani R (1996). Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**, 267–288.

Yuille AL and Rangarajan A (2003). The concave-convex procedure (CCCP), *Neural Computation*, **15**, 915–936.

Zhang CH (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, **38**, 894–942.

Zhang CH and Huang J (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression, *The Annals of Statistics*, **36**, 1567–1594.

Zhang CH and Zhang T (2012). A general theory of concave regularization for high-dimensional sparse estimation problems, *Statistical Science*, **27**, 576–593.

Zhao P and Yu B (2006). On model selection consistency of lasso, *The Journal of Machine Learning Research*, **7**, 2541–2563.

Zou H and Hastie T (2005). Regularization and variable selection via the elastic net, *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **67**, 301–320.

Received March 23, 2020; Revised April 22, 2020; Accepted April 22, 2020