Statistica Sinica **22** (2012), 629-653
doi:http://dx.doi.org/10.5705/ss.2010.027

LARGE SAMPLE PROPERTIES OF THE SCAD-PENALIZED MAXIMUM LIKELIHOOD ESTIMATION ON HIGH DIMENSIONS

Sunghoon Kwon and Yongdai Kim

University of Minnesota and Seoul National University

Abstract: In this paper, we study large sample properties of smoothly clipped absolute deviation (SCAD) penalized maximum likelihood estimation for high-dimensional parameters. First, we prove that the oracle maximum likelihood estimator (MLE) asymptotically becomes a local maximizer of the SCAD-penalized log-likelihood, even when the number of parameters is much larger than the sample size; the oracle MLE is an ideal non-penalized MLE obtained by deleting all irrelevant parameters in advance. Second, we prove that if the log-likelihood is strictly concave, the oracle MLE asymptotically becomes the global maximizer of the SCAD-penalized log-likelihood with a diverging number of parameters that is less than the sample size. Various numerical experiments on simulated data sets are presented to verify the theoretical results, and two data examples are analyzed.

Key words and phrases: High dimension, maximum likelihood estimator, MLE, oracle property, SCAD, smoothly clipped absolute deviation, variable selection.

1. Introduction

Variable selection is an important issue in high-dimensional statistical modeling. Traditionally, stepwise subset selection procedures have been adopted to select an appropriate number of predictive variables. However, such procedures have drawbacks such as intensive computation, difficulties in obtaining sampling properties, and unstableness, see Breiman (1996) for further details. Methods for overcoming these problems have been developed via sparse penalized approaches such as bridge regression (Frank and Friedman (1993)), the least absolute shrinkage and selection operator (LASSO) (Tibshirani (1996)), and the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li (2001)). All these methods have common advantages over subset selection procedures; they are computationally simpler, the derived sparse estimators are stable, and they facilitate higher prediction accuracies.

Theoretical properties of the sparse penalized approaches have been studied by many authors. For a finite number of parameters, Knight and Fu (2000) studied the properties of LASSO-type estimators. Fan and Li (2001) proved that

there exists a local maximizer of the SCAD-penalized log-likelihood that achieves the oracle property. Here, the oracle property means that a penalized maximum likelihood estimator (MLE) is asymptotically equivalent to the oracle MLE that is an ideal non-penalized MLE obtained by deleting all irrelevant parameters in advance. Zou (2006) proposed the adaptive LASSO that achieves the oracle property by varying the weights on the tuning parameter. For a diverging number of parameters, Fan and Peng (2004) proved that the results of Fan and Li (2001) hold when the number of parameters is less than the sample size. Kim, Choi, and Oh (2008) studied the asymptotic properties of the SCAD-penalized least square estimator (LSE) in linear regression when the number of parameters exceeds the sample size. They proved that the oracle LSE asymptotically becomes a local minimizer of the SCAD-penalized residual sum of squares. They also proved that the oracle LSE asymptotically becomes the global minimizer of the SCAD-penalized residual sum of squares when the design matrix is nonsingular. Zhao and Yu (2006) and Meinshausen and Bühlmann (2006) 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) for ultra high-dimensional model selection problems. For a detailed overview of current research on variable selection in high-dimensional models, see Fan and Lv (2010).

In this paper, we study large sample properties of the SCAD-penalized maximum likelihood estimation for high-dimensional parameters. First we show that, under regularity conditions, the oracle MLE asymptotically becomes a local maximizer of the SCAD-penalized log-likelihood even when the number of parameters is larger than the sample size. Most of the asymptotic properties of the SCAD including Fan and Li (2001), Fan and Peng (2004), and Kim, Choi, and Oh (2008) are special cases of our results. Second, we study cases in which the log-likelihood is strictly concave. We specify sufficient conditions to ensure that the oracle MLE asymptotically becomes the global maximizer of the SCAD-penalized log-likelihood for a diverging number of parameters that is less than the sample size. Thus, we can find the oracle MLE asymptotically by maximizing the SCAD-penalized log-likelihood.

The results of this paper can be considered as extensions of those obtained by Kim, Choi, and Oh (2008) from the LSE to the MLE. These extensions, however, are more technically involved. In the case of the LSE, the oracle LSE has a closed form solution hence its asymptotic properties are studied by directly investigating the estimator itself, an approach adopted by Kim, Choi, and Oh (2008). In contrast, the MLE is defined implicitly as a local maximizer of the log-likelihood. The main contribution of this paper is to establish sufficient conditions on the log-likelihood instead of the estimator itself so that the SCAD-penalized MLE

has the desired asymptotic properties. Moreover, the established conditions are sufficiently general to include most generalized linear regression models, such as the logistic and Poisson regressions; this significantly expands the applicability of the SCAD penalty for high-dimensional data.

The remainder of this paper is organized as follows. In Section 2, we briefly review the SCAD penalty. In Section 3, we present sufficient conditions for the oracle MLE to be a local maximizer of the SCAD-penalized log-likelihood asymptotically. In Section 4, we prove that the oracle MLE becomes the global maximizer of the SCAD-penalized log-likelihood asymptotically when the log-likelihood is strictly concave. Results of numerical studies, including simulated and data sets, are presented in Section 5. The concluding remarks and technical details are provided in Section 6 and the Appendix, respectively.

2. Review of SCAD-penalized Methods

Let $\mathbf{z}_i$, $i \leq n$, be independent and identically distributed random variables with a density $f(\mathbf{z}, \boldsymbol{\theta})$, where $\boldsymbol{\theta} \in \Theta$ and Θ is an open subset of $\mathbb{R}^p$. The penalized MLE is

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \left(\sum_{i=1}^{n} \log f(\mathbf{z}_i, \boldsymbol{\theta}) - n \sum_{j=1}^{p} J_\lambda(\theta_j) \right) \tag{2.1}$$

for some penalty $J_\lambda(\theta)$. The bridge penalty (Frank and Friedman (1993)) is of the form $J_\lambda(\theta) = \lambda|\theta|^r, r > 0$, and when $r = 1$, the penalty is known as the LASSO (Tibshirani (1996)). If $J_\lambda(\theta) = \lambda^2 - (|\theta| - \lambda)^2 I(|\theta| < \lambda)$, we obtain the hard thresholding estimator (Antoniadis and Fan (2001)). Fan and Li (2001) suggested the SCAD penalty as

$$\frac{\partial}{\partial\theta} J_\lambda(\theta) = \lambda \text{sign}(\theta) \left\{ I(|\theta| \leq \lambda) + \frac{(a\lambda - |\theta|)_+}{(a-1)\lambda} I(|\theta| > \lambda) \right\} \tag{2.2}$$

for some $a > 2$. The SCAD penalty is a continuously differentiable function that improves the properties of the LASSO and the hard thresholding penalty, so that the SCAD-penalized estimator satisfies the unbiasedness, sparsity and continuity discussed by Fan and Li (2001).

For the SCAD-penalized MLE, Fan and Li (2001) and Fan and Peng (2004) proved the oracle property when $p = O(n^k)$ for some $k < 1$. In the case of linear regression, Kim, Choi, and Oh (2008) obtained more definitive results than those obtained by Fan and Li (2001) and Fan and Peng (2004). They proved that the oracle LSE asymptotically becomes a local minimizer of the SCAD-penalized residual sum of squares when $p = O(n^k)$ for some $k \geq 1$. In addition, they showed that the oracle LSE asymptotically becomes the global minimizer of the SCAD-penalized residual sum of squares when the design matrix is nonsingular.

This global property of the oracle LSE is the strongest result since it indeed gives a way to identify the oracle LSE.

In this paper, we follow the approaches adopted by Kim, Choi, and Oh (2008) to study the asymptotic properties of the SCAD-penalized MLE. First, we prove that the oracle MLE is a local maximizer of the SCAD-penalized log-likelihood asymptotically when $p = O(n^k)$ for some $k \geq 1$. Second, we show that the oracle MLE is the SCAD-penalized MLE asymptotically if the log-likelihood is strictly concave. These results are extensions of the results obtained by Fan and Li (2001) and Fan and Peng (2004), and they include the results of Kim, Choi, and Oh (2008) as special cases.

3. Large Sample Property of Oracle MLE

In this section, we give sufficient conditions on the log-likelihood so that the oracle MLE becomes a local maximizer of the SCAD-penalized log-likelihood when $p = O(n^k)$ for some $k \geq 1$, where k depends on the moments of the derivatives of the log-likelihood.

For each n, let $\mathbf{z}_{ni}$, $i \leq n$, be independent and identically distributed random variables with a density $f_n(\mathbf{z}_{n1}, \boldsymbol{\theta}_n^*)$, where $\boldsymbol{\theta}_n^* \in \Theta_n$ and Θ_n is an open subset of $\mathbb{R}^{p_n}$. Without loss of generality, we assume that the first q_n elements of the true parameter vector $\boldsymbol{\theta}_n^*$ are nonzero and the remaining $p_n - q_n$ elements are zero. The following regularity conditions are to be imposed, where $M_1, M_2, \cdots$ are some positive constants.

Condition **A1.** For any constants c_1 and c_2 satisfying $0 < 5c_1 < c_2 \leq 1$,

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

Condition **A2.** The first and second derivatives of the log-likelihood $\log f_n(\mathbf{z}_{n1}, \boldsymbol{\theta}_n)$ satisfy

$$E_{\boldsymbol{\theta}_n^*} \left\{ \frac{\partial \log f_n(\mathbf{z}_{n1}, \boldsymbol{\theta}_n^*)}{\partial \theta_{nj}} \right\} = 0,$$
$$E_{\boldsymbol{\theta}_n^*} \left\{ \frac{\partial^2 \log f_n(\mathbf{z}_{n1}, \boldsymbol{\theta}_n^*)}{\partial \theta_{nj} \partial \theta_{nl}} \right\} = -E_{\boldsymbol{\theta}_n^*} \left\{ \frac{\partial \log f_n(\mathbf{z}_{n1}, \boldsymbol{\theta}_n^*)}{\partial \theta_{nj}} \frac{\partial \log f_n(\mathbf{z}_{n1}, \boldsymbol{\theta}_n^*)}{\partial \theta_{nl}} \right\}$$

for all $1 \leq j, l \leq p_n$, and $n \geq 1$.

Condition **A3.** The first $q_n \times q_n$ submatrix $I_n^{(1)}(\boldsymbol{\theta}_n^*)$ of the Fisher information matrix

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

is positive definite such that

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

for all $n \geq 1$. Here, $\lambda_{min}(D)$ and $\lambda_{max}(D)$ denote the smallest and largest eigenvalues of the given matrix D, respectively.

Condition **A4.** There exists a sufficiently large open subset $B_n \subset \Theta_n$ that contains the true parameter $\boldsymbol{\theta}_n^*$ such that for almost all $\mathbf{z}_{ni}$, the density admits a third derivatives for all $\boldsymbol{\theta}_n \in B_n$. Furthermore, there are functions $U_{njlm}(\cdot)$ such that

$$\left|\frac{\partial^3 \log f_n(\mathbf{z}_{ni}, \boldsymbol{\theta}_n)}{\partial\theta_{nj}\partial\theta_{nl}\partial\theta_{nm}}\right| < U_{njlm}(\mathbf{z}_{ni})$$

for any $\boldsymbol{\theta}_n \in B_n$, for all $1 \leq j, l, m \leq p_n$, and $n \geq 1$.

Condition **A5.** There exists an integer $k \geq 1$ such that

$$E_{\boldsymbol{\theta}_n^*}\left\{\frac{\partial \log f_n(\mathbf{z}_{n1}, \boldsymbol{\theta}_n^*)}{\partial\theta_{nj}}\right\}^{2k} < M_4, \quad E_{\boldsymbol{\theta}_n^*}\left\{\frac{\partial^2 \log f_n(\mathbf{z}_{n1}, \boldsymbol{\theta}_n^*)}{\partial\theta_{nj}\partial\theta_{nl}}\right\}^{2k} < M_5,$$
$$E\Big(U_{jlm}(\mathbf{z}_{ni})\Big)^{2k} < M_6$$

for all $1 \leq j, l, m \leq p_n$, and $n \geq 1$.

Remark 1. The condition **A1**, employed by Zhao and Yu (2006) and Kim, Choi, and Oh (2008), allows the number of true relevant parameters to diverge to infinity and their values to converge to 0. The conditions **A2** to **A4** are standard assumptions for maximum likelihood estimation (Fan and Peng (2004)). In the case of linear regression, **A3** has the design matrix corresponding to the relevant covariates as nonsingular.

Remark 2. The condition **A5** specifies the tail behavior of $f_n(\mathbf{z}_{n1}, \boldsymbol{\theta}_n^*)$, which determines the order of p_n with respect to some integer $k \geq 1$. It is the same as the condition $E\varepsilon^{2k} < \infty$ in Kim, Choi, and Oh (2008) in linear regression. Another example is logistic regression, where

$$\text{pr}(y = 1|\mathbf{x}) = \frac{\exp(\mathbf{x}^T\boldsymbol{\theta}_n^*)}{1 + \exp(\mathbf{x}^T\boldsymbol{\theta}_n^*)}.$$

Suppose that the covariate $\mathbf{x}$ is bounded such that $\max_{1\leq j\leq p_n} |x_j| \leq b$ for some constant $b > 0$. Since

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

it is easy to see that

$$\max_{1\le j\le p_n} E_{\boldsymbol{\theta}_n^*}\left\{\frac{\partial \log f_n(y,\boldsymbol{\theta}_n^*|\mathbf{x})}{\partial\theta_{nj}}\right\}^{2k} \le \max_{1\le j\le p_n} x_j^{2k} \le b^{2k} < \infty.$$

The other two inequalities in condition **A5** can be checked similarly.

Let the empirical log-likelihood be

$$L_n(\boldsymbol{\theta}_n) = \sum_{i=1}^{n} \log f_n(\mathbf{z}_{ni}, \boldsymbol{\theta}_n).$$

Then, the corresponding SCAD-penalized log-likelihood is

$$Q_n(\boldsymbol{\theta}_n) = L_n(\boldsymbol{\theta}_n) - n\mathbf{J}_{\lambda_n}(\boldsymbol{\theta}_n), \tag{3.1}$$

where $\mathbf{J}_{\lambda_n}(\cdot) = \sum_{j=1}^{p_n} J_{\lambda_n}(\cdot)$ and $J_{\lambda_n}(\cdot)$ is the SCAD penalty given by (2.2). The oracle MLE $\hat{\boldsymbol{\theta}}_n^o$ is defined as any local maximizer of $L_n(\boldsymbol{\theta}_n)$ subject to $\theta_{nj} = 0$ for $q_n < j \le p_n$ such that

$$\|\hat{\boldsymbol{\theta}}_n^o - \boldsymbol{\theta}_n^*\| = O_p(\sqrt{\frac{q_n}{n}}). \tag{3.2}$$

It is an ideal MLE and its estimated coefficients of the irrelevant parameters are set to be exactly zero. Under the regularity conditions **A1** to **A5**, we can prove that the oracle MLE exists asymptotically and satisfies $\sqrt{n/q_n}$-consistency in (3.2) with a Gaussian limiting distribution (see the proofs of Theorems 1 and 2 in Fan and Peng (2004)). Further, if $L_n(\boldsymbol{\theta}_n)$ is strictly concave with respect to θ_{nj} for $j \le q_n$, we can define the oracle MLE as the unique global maximizer of $L_n(\boldsymbol{\theta}_n)$ subject to $\theta_{nj} = 0$ for $q_n < j \le p_n$. For example, if the underlying distribution is Gaussian, the oracle MLE is the same as the oracle LSE studied by Kim, Choi, and Oh (2008).

Let $\mathcal{A}_n(\lambda_n)$ denote the set of all local maximizers of (3.1). The following theorem states that the oracle MLE is a local maximizer of (3.1) asymptotically, even when $p_n \ge n$.

Theorem 1. *If* **A1**$-$**A5** *hold, we have*

$$\mathrm{pr}\Big(\ \hat{\boldsymbol{\theta}}_n^o \in \mathcal{A}_n(\lambda_n)\ \Big) \to 1$$

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

Note that Theorem 1 is satisfied for $p_n = o(n^{(c_2-c_1)k})$ so that equality holds even when p_n is much larger than n, provided k is sufficiently large. If the

distributions of the corresponding random variables (the first, second, and third derivatives of the log-likelihood) have exponentially decaying tails, we can show that Theorem 1 holds when $p_n = O(\exp(n^{c_3}))$ for some constant $c_3 > 0$, e.g., linear regression with Gaussian errors as studied by Kim, Choi, and Oh (2008) and logistic regression with bounded covariates as described in Remark 2.

4. Asymptotic Equivalence of SCAD-penalized MLE and Oracle MLE

Theorem 1 does not tell us which local maximizer is the oracle MLE. However, when the log-likelihood is strictly concave, we can show that the SCAD-penalized MLE (the global maximizer of the SCAD-penalized log-likelihood) is exactly the same as the oracle MLE asymptotically, as stated in Theorem 2. Note that $p_n \leq n$ for the log-likelihood to be strictly concave. In addition to conditions **A1** to **A5**, the following condition is required.

Condition **A6.** There exists a positive constant M_7 and a convex open subset $\Omega_n \subset \Theta_n$ such that $\hat{\boldsymbol{\theta}}_n^o$ as well as $\boldsymbol{\theta}_n^*$ belong to Ω_n and

$$\min_{\boldsymbol{\theta}_n \in \Omega_n} \lambda_{min}(\boldsymbol{\theta}_n) > M_7 \tag{4.1}$$

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

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

at $\boldsymbol{\theta}_n$.

Since the log-likelihood is strictly concave on Ω_n, the oracle MLE is uniquely defined as the maximizer of $L_n(\boldsymbol{\theta}_n)$ subject to $\theta_{nj} = 0$ for $q_n < j \leq p_n$ on Ω_n. The following theorem states that the SCAD-penalized MLE on Ω_n is exactly the same as the oracle MLE asymptotically.

Theorem 2. *Let $\hat{\boldsymbol{\theta}}_n$ be the global maximizer of* (3.1) *on Ω_n. If* **A1**-**A6** *hold, we have*

$$\text{pr}\left(\hat{\boldsymbol{\theta}}_n = \hat{\boldsymbol{\theta}}_n^o \right) \to 1$$

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

In the case of linear regression, it is easy to see that $\Omega_n = \mathbb{R}^{p_n}$ provided the design matrix has a smallest eigenvalue that is sufficiently large, in which case Theorem 2 is equivalent to Theorem 3 of Kim, Choi, and Oh (2008). For general cases, however, the Hessian matrix and its smallest eigenvalue depend on the parameters. Theorem 2 only guarantees that the oracle MLE is the optimum

among the solutions whose Hessian matrices are well-posed. In practice, we expect this of Hessian matrices of reasonable estimators. For example, in the logistic regression described in Remark 2, if the design matrix is orthonormal, $\lambda_{min}(\boldsymbol{\theta}_n) = \min_{i\le n} p_i(\boldsymbol{\theta}_n)(1-p_i(\boldsymbol{\theta}_n))$, where

$$p_i(\boldsymbol{\theta}_n) = \frac{\exp(\mathbf{x}_i^T\boldsymbol{\theta}_n)}{1+\exp(\mathbf{x}_i^T\boldsymbol{\theta}_n)}$$

for $i \le n$. Hence, for $\boldsymbol{\theta}_n$ to belong to Ω_n, $p_i(\boldsymbol{\theta}_n)$ should be neither too large nor too small, which is expected to hold for the true parameter $\boldsymbol{\theta}_n^*$. When the covariates are bounded, we can take $\Omega_n = \{\boldsymbol{\theta}_n : \sum_{j=1}^{p_n} |\theta_{nj}| \le \eta\}$ for any $\eta > 0$, in which case

$$\lambda_{min}(\boldsymbol{\theta}_n) \ge \frac{1}{(\exp(\eta b)+1)(\exp(-\eta b)+1)}.$$

Hence, it is not unreasonable to expect that $\boldsymbol{\theta}_n^* \in \Omega_n$ for sufficiently large values of η.

At this point, we do not know how to determine the global maximizer of the SCAD-penalized log-likelihood. In the case of linear regression, Kim and Kwon (2011) gave sufficient conditions for the uniqueness of a local minimizer of the nonconvex penalized residual sum of squares. We believe that a similar result can be obtained for the SCAD-penalized maximum likelihood estimation when the log-likelihood is strictly concave.

5. Numerical Studies

In this section, we investigate the finite sample performance of the SCAD-penalized MLE via simulations and the analysis of data sets. We obtained the SCAD-penalized MLE using the concave-convex procedure algorithm of Kim, Choi, and Oh (2008); the algorithm of Park and Hastie (2007) was applied to solve the LASSO problem in the inner loop of the concave-convex procedure.

5.1. Simulation studies

For simulation studies, we considered three generalized linear models:

- *Linear Regression:*

$$y|\mathbf{x} \sim N\Big(\mathbf{x}^T\boldsymbol{\theta}^*,\ \tau^2\Big). \tag{5.1}$$

- *Logistic Regression:*

$$y|\mathbf{x} \sim B\left(1,\ \frac{\exp(\mathbf{x}^T\boldsymbol{\theta}^*)}{1+\exp(\mathbf{x}^T\boldsymbol{\theta}^*)}\right). \tag{5.2}$$

- *Poisson Regression:*

$$y|\mathbf{x} \sim \text{Poisson}\Big(\exp(\mathbf{x}^T\boldsymbol{\theta}^*)\Big). \tag{5.3}$$

We constructed a sequence $\theta_j^* = 1.2e^{-v(j-1)}, j \le 5$, for the nonzero true parameter values, and v was chosen such that $\theta_5^* = 0.6$. We set the covariate vector $\mathbf{x} = (x_1, \cdots, x_p)^T$ to be a multivariate Gaussian random vector with mean zero and the covariance of x_k and x_l to be $\sigma^2 r^{|k-l|}$ for $r \in [0, 1)$. We let $\tau^2 = 1$ for linear regression and selected various values of n, p, r, and σ^2 for the simulations.

5.1.1. Optimality of oracle MLE

First, consider the local optimality of the oracle MLE. Among 100 simulations, we calculated the frequencies of those cases where there exists a λ such that the oracle MLE belongs to the set of local maximizers of the SCAD-penalized log-likelihood. The results are summarized in Table 1. As expected, the frequency of the oracle MLE being a local maximizer tends to be large when the sample size n is large and/or the number of covariates p is small. The correlation r between the covariates does not seem to be an important factor relatively.

A rather unexpected observation in Table 1 is that the frequencies of the models differ significantly according to σ^2 and r. In particular, the frequencies of the oracle MLE being a local maximizer are very small for the Poisson regression models as compared to those for the logistic regression models. This observation can be partially explained as follows. By the KKT conditions (see (A.4) and (A.5) in the Appendix), the oracle MLE becomes a local maximizer when

$$\min_{j \le q_n} |\hat{\theta}_{nj}^o| \ge a\lambda \tag{5.4}$$

and

$$\max_{q_n < j \le p_n} \Big|\frac{1}{n}\sum_{i=1}^{n}\Big(y_i - \hat{\mu}_i\Big)x_{ij}\Big| \le \lambda, \tag{5.5}$$

where $\hat{\mu}_i = \mathrm{E}_{\hat{\theta}_n}(y_i|\mathbf{x}_i)$. Loosely speaking the scales of $y_i - \hat{\mu}_i$ and x_{ij} are proportional to $\mathrm{Var}(y_i - \hat{\mu}_i|\mathbf{x}_i)$ and σ^2, respectively. Therefore, the probability of (5.5), or equivalently the oracle MLE being a local maximizer, is large when $\mathrm{Var}(y_i - \hat{\mu}_i|\mathbf{x}_i)$ and σ^2 are small. Note that $\mathrm{Var}(y_i - \hat{\mu}_i|\mathbf{x}_i)$ is the largest in the case of the Poisson regression and smallest in the case of the logistic regression, and hence the probability of the oracle MLE being a local maximizer is reversed. Also, $\mathrm{Var}(y_i - \hat{\mu}_i|\mathbf{x}_i)$ increases most rapidly for the Poisson distribution as r increases, which explains the extremely low frequencies for $\sigma^2 = 1/2$ and $r = 0.5$ compared to the other cases.

Table 1. Frequencies of oracle MLE being a local maximizer of SCAD-penalized log-likelihood among 100 simulations.

Linear regression					
n	p	r	$\sigma^2 = 1$	$\sigma^2 = 1/2$	$\sigma^2 = 1/3$
200	1000	0	0	37	73
		0.5	0	30	57
	2000	0	0	39	55
		0.5	0	21	57
400	1000	0	32	96	99
		0.5	36	94	99
	2000	0	18	92	99
		0.5	18	93	98
Logistic regression					
n	p	r	$\sigma^2 = 1$	$\sigma^2 = 1/2$	$\sigma^2 = 1/3$
200	1000	0	91	93	86
		0.5	81	75	76
	2000	0	84	88	88
		0.5	77	70	74
400	1000	0	98	96	100
		0.5	99	92	93
	2000	0	99	99	98
		0.5	97	97	96
Poisson regression					
n	p	r	$\sigma^2 = 1/2$	$\sigma^2 = 1/3$	$\sigma^2 = 1/4$
200	1000	0	1	18	58
		0.5	0	2	9
	2000	0	0	16	52
		0.5	0	0	8
400	1000	0	15	89	98
		0.5	0	11	74
	2000	0	7	78	95
		0.5	0	8	60
800	1000	0	90	100	100
		0.5	1	81	100
	2000	0	87	100	100
		0.5	0	72	100

Second, we investigate the global optimality of the oracle MLE. We calculated the frequencies of the cases among 100 simulations where the oracle MLE is the global maximizer of the SCAD-penalized log-likelihood. In practice, it is difficult to check whether a given solution is the global maximizer. Hence, we defined the oracle MLE as the global maximizer if it achieved the maximum value of the SCAD-penalized log-likelihood among 300 candidate local maximizers. The set

of candidate local maximizers was constructed by changing the initial estimators in the concave-convex procedure, where 50 initial solutions were obtained from each of the following categories:

- adding noises generated from $N(0, 0.5)$ to the nonzero true parameters;
- adding noises generated from $N(0, 0.5)$ to the nonzero parameters of the LASSO-penalized MLE;
- adding noises generated from $N(0, 0.5)$ to the nonzero parameters of the non-penalized MLE;
- adding random numbers generated from $U(0.3, 2.4)$ to the first 5 elements of the zero vector;
- adding random numbers generated from $U(0.3, 2.4)$ to the last $p-5$ elements of the zero vector;
- adding random numbers generated from $U(0.3, 2.4)$ to all the elements of the zero vector.

The results are summarized in Table 2. The frequency of global optimality of the oracle MLE increases as the sample size increases, which verifies Theorem 2. Note that the frequency of the global optimality of the oracle MLE is relatively small in the case of the logistic regression. This is partly because the smallest eigenvalue of its Hessian matrix is the smallest among these models.

5.1.2. Finite sample performance of SCAD-penalized MLE

We compared the predictive and selection performance of the SCAD-penalized MLE with the LASSO penalized MLE and the oracle MLE. We only considered the logistic and Poisson regression models because linear regression was studied by Kim, Choi, and Oh (2008). The parameters used for the simulations are the same as in Section 5.1.1, with $\sigma^2 = 1$. We used five-fold cross validation to select the tuning parameter and repeated the simulation 100 times.

Table 3 presents the averages of the negative log-likelihood values and their standard errors obtained on the basis of $N = 5,000$ independent test data sets. As expected, the oracle MLE had the best performance, and the performance of the SCAD was better than that of the LASSO for all the cases. In particular, when $n = 300$, the performance of the SCAD is almost the same as that of the oracle MLE.

Table 4 presents the averages of the frequencies of correctly and incorrectly estimated nonzero coefficients. For the logistic models, the LASSO was better for selecting relevant variables than the SCAD for all the cases. Simultaneously, however, many more noisy variables were included by the LASSO than

Table 2. Frequencies of oracle MLE being a local maximizer and the global maximizer of SCAD-penalized log-likelihood among 100 simulations.

Linear regression									
			$\sigma^2 = 1$			$\sigma^2 = 1/2$		$\sigma^2 = 1/3$	
n	p	r	Local	Global		Local	Global	Local	Gloabl
100	10	0.0	57	57		83	83	89	85
		0.5	68	68		83	78	85	74
	20	0.0	27	27		63	63	79	71
		0.5	25	25		55	47	74	50
300	10	0.0	99		99	99	99	100	99
		0.5	99	99		100	100	100	98
	20	0.0	93	93		100	100	100	99
		0.5	97	97		99	99	99	98
Logistic regression									
			$\sigma^2 = 1$			$\sigma^2 = 1/2$		$\sigma^2 = 1/3$	
n	p	r	Local	Global		Local	Global	Local	Gloabl
100	10	0.0	85		58	81	35	80	22
		0.5	77	14		76	10	87	6
	20	0.0	90	47		76	13	77	11
		0.5	67	4		69	5	75	3
300	10	0.0	100	97		98	87	98	72
		0.5	98	76		96	52	94	41
	20	0.0	100	95		98	69	99	58
		0.5	100	70		96	32	93	16
Poisson regression									
			$\sigma^2 = 1$			$\sigma^2 = 1/2$		$\sigma^2 = 1/3$	
n	p	r	Local	Global		Local	Global	Local	Gloabl
100	10	0.0	8	8		50	50	84	84
		0.5	4	4		37	37	58	57
	20	0.0	0	0		18	18	54	54
		0.5	0	0		4	4	35	33
300	10	0.0	31	31		92	92	100	100
		0.5	0	0		61	61	92	92
	20	0.0	5	5		83	83	99	99
		0.5	0	0		27	27	70	70

the SCAD. Second, for increased sample size, the number of correctly estimated nonzero coefficients increased in each method. However, the number of incorrectly estimated nonzero coefficients decreased only in the case of the SCAD. This observation confirms the well known result that the LASSO selects more variables than required (Zou (2006)) whereas the SCAD has the oracle property. In the case of the Poisson regression models, in contrast to the logistic regression

Table 3. Averaged test negative log-likelihood values (standard errors)

Logistic regression				
n=100		Test negative log-likelihood values		
p	r	SCAD	LASSO	Oracle
500	0	0.5812 (0.007)	0.5851 (0.003)	0.4720 (0.002)
	0.5	0.4150 (0.005)	0.4308 (0.003)	0.3755 (0.004)
2000	0	0.5953 (0.008)	0.6057 (0.003)	0.4736 (0.002)
	0.5	0.4315 (0.005)	0.4473 (0.003)	0.3715 (0.003)
n=300		Test negative log-likelihood values		
p	r	SCAD	LASSO	Oracle
500	0	0.4630 (0.001)	0.4943 (0.001)	0.4457 (0.001)
	0.5	0.3542 (0.001)	0.3735 (0.001)	0.3407 (0.001)
2000	0	0.4713 (0.002)	0.5044 (0.001)	0.4463 (0.001)
	0.5	0.3564 (0.002)	0.3787 (0.001)	0.3429 (0.001)
Poisson regression				
n=100		Test negative log-likelihood values		
p	r	SCAD	LASSO	Oracle
500	0	1.4814 (0.009)	1.6428 (0.017)	1.4100 (0.004)
	0.5	1.5227 (0.011)	1.6451 (0.026)	1.4407 (0.005)
2000	0	1.5079 (0.009)	1.7298 (0.022)	1.4153 (0.007)
	0.5	1.5663 (0.012)	1.7296 (0.032)	1.4373 (0.004)
n = 300		Test negative log-likelihood values		
p	r	SCAD	LASSO	Oracle
500	0	1.4063 (0.002)	1.4435 (0.004)	1.3829 (0.001)
	0.5	1.4309 (0.002)	1.4561 (0.004)	1.4051 (0.002)
2000	0	1.4063 (0.002)	1.4497 (0.004)	1.3833 (0.001)
	0.5	1.4326 (0.003)	1.4623 (0.007)	1.4061 (0.002)

models, the selectivity of the two methods were similar.

It is worth noting that a positive correlation ($r = 0.5$) increased the prediction and selection performance of all three estimators in the case of logistic regression models. We found that introducing a positive correlation decreased the Bayes-error, which partially explains the results of the simulations. The relation of the correlation between covariates and prediction accuracy appears to be very complicated (related empirical results can be found in Friedman (2008)).

5.2. Data analysis I: Change point problem

The data set introduced by Jarrett (1979) (British coal-mining data) includes the point process of the dates of serious coal-mining disasters involving the death of 10 or more men between 1851 and 1962. Several papers have used this data set to illustrate new methods for change point analysis. In this subsection, we

Table 4. Averaged frequencies (standard errors) of correctly and incorrectly estimated nonzero coefficients.

Logistic regression					
n=100		Correct no. of nonzeros		Incorrect no. of nonzeros	
p	r	SCAD	LASSO	SCAD	LASSO
500	0	2.28 (0.09)	3.58 (0.09)	0.80 (0.12)	21.16 (1.04)
	0.5	2.73 (0.07)	4.10 (0.07)	0.24 (0.04)	18.37 (0.98)
2000	0	1.98 (0.09)	3.15 (0.09)	1.05 (0.14)	22.82 (1.12)
	0.5	2.48 (0.08)	3.92 (0.07)	0.39 (0.08)	20.85 (1.13)
n=300		Correct no. of nonzeros		Incorrect no. of nonzeros	
p	r	SCAD	LASSO	SCAD	LASSO
500	0	4.23 (0.07)	4.88 (0.03)	0.26 (0.06)	33.63 (1.29)
	0.5	4.10 (0.06)	4.91 (0.02)	0.23 (0.06)	27.33 (1.34)
2000	0	3.92 (0.07)	4.74 (0.04)	0.29 (0.06)	42.27 (1.72)
	0.5	3.96 (0.06)	4.81 (0.04)	0.17 (0.05)	33.87 (1.51)
Poisson regression					
n=100		Correct no. of nonzeros		Incorrect no. of nonzeros	
500	0	2.32 (0.05)	2.30 (0.05)	14.52 (1.03)	17.85 (1.07)
	0.5	2.79 (0.07)	2.93 (0.06)	12.66 (1.10)	16.02 (1.10)
2000	0	2.24 (0.05)	2.17 (0.04)	18.08 (1.20)	21.37 (1.25)
	0.5	2.58 (0.06)	2.74 (0.06)	16.27 (1.15)	18.23 (1.13)
n=300		Correct no. of nonzeros		Incorrect no. of nonzeros	
500	0	2.92 (0.05)	2.93 (0.05)	16.06 (1.07)	19.87 (1.26)
	0.5	3.52 (0.06)	3.51 (0.05)	13.66 (0.93)	16.18 (1.12)
2000	0	2.91 (0.05)	2.88 (0.05)	20.58 (1.28)	22.56 (1.40)
	0.5	3.34 (0.06)	3.35 (0.05)	18.65 (1.50)	19.68 (1.59)

estimate the mean change points via sparse penalized approaches.

The annual numbers of disasters $y_i, i \leq n$, are assumed to be independently distributed as Poisson with mean $\mu_i, i \leq n$. Then, the log-likelihood of $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_n)^T$ is

$$L(\boldsymbol{\mu}) = \sum_{i=1}^{n} \Big(y_i \log(\mu_i) - \mu_i \Big).$$

Reparameterizing $\boldsymbol{\mu}$ as $\nu_1 = \mu_1$ and $\nu_i = \mu_i - \mu_{i-1}$ for $2 \leq i \leq n$, and adapting the penalty on $\boldsymbol{\nu} = (\nu_1, \cdots, \nu_n)^T$, gives the penalized problem

$$Q(\boldsymbol{\nu}) = \sum_{i=1}^{n} \left(y_i \log \sum_{j=1}^{i} \nu_j - \sum_{j=1}^{i} \nu_j \right) - n\mathbf{J}_\lambda(\boldsymbol{\nu}). \tag{5.6}$$

This approach has been investigated by Harchaoui and Lévy-Leduc (2007). The nonzero estimator $\hat{\nu}_i \neq 0$ can be interpreted as an indication of the change

Table 5. Change point analysis for coal-mining data (estimated years and sizes of decrements).

	year	1887	1892	1897	1943	1948
SCAD	decrement	-2.2763	-	-	-	-
LASSO	decrement	-0.7340	-0.5999	-0.3565	-0.0434	-0.5125

between μ_i and μ_{i-1}. In addition, its value can be used as the estimate of the difference between two consecutive means.

We applied two sparse penalized methods: the LASSO and the SCAD. The estimated change points and estimated decrements are listed in Table 5. Five change points are detected by the LASSO, with the greatest decrement of -0.7340 in the year 1887 and the smallest decrement of -0.0434 in the year 1943. On the other hand, the SCAD detects only one change point in the year 1887 with a decrement of -2.2763. Raftery and Akman (1986) and Green (1995) found the highest posterior mode of the time of change in the year 1890 with 95% credible intervals of [1887,1895] and [1887,1896], respectively. Chib (1998) reported the difference of the posterior means in the year 1891 to be -2.162 from each posterior mean, 0.957 and -3.119, with standard deviations of 0.286 and 0.120, respectively, similar to the results of the SCAD. In Figure 1, the left panel shows the estimated means of the two methods. We can see that there are five relatively small change points in the LASSO, whereas there is a large change point in the SCAD. The right panel of Figure 1 compares the cumulative means of the two methods and shows that the SCAD-penalized estimator is much closer to the cumulative frequencies of the disasters. Note that the LASSO yields a biased result whereas the SCAD is almost unbiased.

5.3. Data analysis II: Model-based clustering

We analyzed two popular microarray gene expression data sets, *Leukemia* and *Colon*, employed by Dudoit, Fridlyand and Speed (2002) and Alon et al. (1999), respectively, and applied the penalized model-based clustering method.

- *Leukemia*: This data set consists of 38 samples and 7,129 corresponding gene expression measurements, including a label indicating two types of cancers. Among the samples, 21 indicate acute lymphoblastic leukemia (ALL) cancer class and the others, acute myeloid leukemia (AML). Following the pre-processing steps adopted Dudoit, Fridlyand and Speed (2002), we ordered the genes by their F-statistics and selected the top 1,000 genes.

- *Colon*: This data set comes from a gene expression study of 40 tumor and 22 normal colon tissue samples that were analyzed with more than 6,500

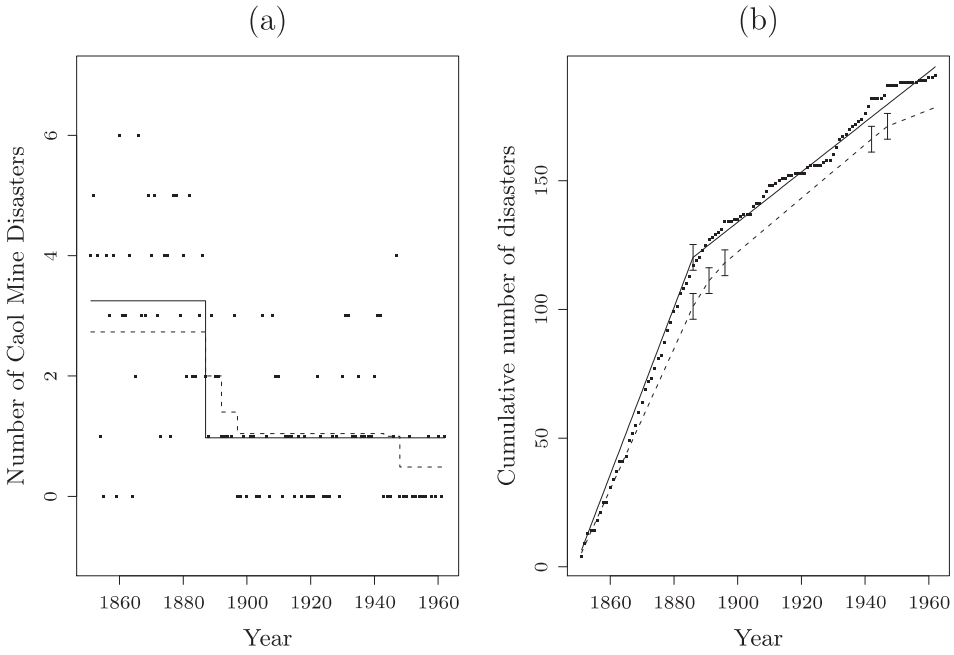

Figure 1. The left panel (a) shows the estimated means by the LASSO (dashes) and SCAD penalty (solid line) while the dots represent the data points. The right panel (b) includes the estimated cumulative lines by the LASSO (dashes) and SCAD penalty (solid line). The change points are indicated by small bars.

human genes. Following the steps of Alon et al. (1999), a selection of 2,000 genes with the highest minimal intensity across the samples was made. We selected the top 1,000 F-statistic valued genes.

To apply sparse penalized model-based clustering, it is assumed that the p-dimensional gene expression levels $\mathbf{x}_i, i \leq n$, are independently drawn from a K-component Gaussian mixture distribution:

$$f(\mathbf{x}, \boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \phi_k(\mathbf{x}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \tag{5.7}$$

Here, the generic notation $\boldsymbol{\theta}$ includes all the parameters in (5.7) and ϕ_k is the p-dimensional multivariate Gaussian density function with mean $\boldsymbol{\mu}_k$ and covariance matrix $\boldsymbol{\Sigma}_k$. We assume that the covariance matrix in each component has a common diagonal matrix such that $\boldsymbol{\Sigma}_k = \text{diag}(\boldsymbol{\sigma}^2)$ for all $k \leq K$, with $\boldsymbol{\sigma}^2 = (\sigma_1^2, \cdots, \sigma_p^2)^T$. Given the data, the log-likelihood is

$$L(\boldsymbol{\theta}; \mathbf{X}) = \sum_{i=1}^{n} \log \left\{ \sum_{k=1}^{K} \pi_k \phi_k(\mathbf{x}_i, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}) \right\}$$

and we get the corresponding penalized log-likelihood:

$$Q_P(\boldsymbol{\theta}; \mathbf{X}) = \sum_{i=1}^{n} \log \left\{ \sum_{k=1}^{K} \pi_k \phi_k(\mathbf{x}_i, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}) \right\} - n \sum_{k=1}^{K} \mathbf{J}_\lambda(\boldsymbol{\mu}_k).$$

Using the expectation and maximization (EM) algorithm, we can maximize the penalized log-likelihood. Let z_{ki} be the indicator of whether $\mathbf{x}_i$ is obtained from component k. Then, the complete penalized log-likelihood is

$$Q_P^C(\boldsymbol{\theta}; \mathbf{X}, \mathbf{Z}) = \sum_{i=1}^{n} \sum_{k=1}^{K} z_{ik} \left\{ \log \pi_k + \log \phi_k(\mathbf{x}_i, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}) \right\} - n \sum_{k=1}^{K} \mathbf{J}_\lambda(\boldsymbol{\mu}_k).$$

For the given parameters $\hat{\pi}_k, \hat{\boldsymbol{\mu}}_k$, and $\hat{\boldsymbol{\Sigma}}$, the E-step yields the responsibilities

$$\hat{\tau}_{ik} = \frac{\hat{\pi}_k \phi_k(x_i, \hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}})}{\sum_{k=1}^{K} \hat{\pi}_k \phi_k(x_i, \hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}})}$$

and the M-step is to maximize

$$E\left(Q_P^C(\boldsymbol{\theta}|\mathbf{X})\right) = \sum_{i=1}^{n} \sum_{k=1}^{K} \hat{\tau}_{ik} \left\{ \log \pi_k + \log \phi_k(\mathbf{x}_i, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}) \right\} - n \sum_{k=1}^{K} \mathbf{J}_\lambda(\boldsymbol{\mu}_k).$$

EM algorithm for SCAD-penalized Gaussian mixture model

Given the initial estimates of $\hat{\pi}_k, \hat{\boldsymbol{\mu}}_k$, and $\hat{\boldsymbol{\Sigma}} = \operatorname{diag}(\hat{\boldsymbol{\sigma}}^2)$, repeat the E and M-steps until convergence.

(E-Step) Calculate the responsibilities

$$\hat{\tau}_{ik} \leftarrow \frac{\hat{\pi}_k \phi_k(\mathbf{x}_i, \hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}})}{\sum_{k=1}^{K} \hat{\pi}_k \phi_k(\mathbf{x}_i, \hat{\boldsymbol{\mu}}_k, \hat{\boldsymbol{\Sigma}})}$$

for $i \leq n$ and $k \leq K$.

(M-Step) Update the means, variances, and mixing probabilities

$$\hat{\boldsymbol{\mu}}_k \leftarrow \operatorname{sign}(\tilde{\boldsymbol{\mu}}_k) \left(\frac{|\tilde{\boldsymbol{\mu}}_k| - n\lambda \hat{\boldsymbol{\sigma}}^2}{\sum_{i=1}^{n} \hat{\tau}_{ik}} \right)_+, \quad \hat{\sigma}_j^2 \leftarrow \frac{1}{n} \sum_{i=1}^{n} \sum_{k=1}^{K} \hat{\tau}_{ik} (x_{ij} - \hat{\mu}_{kj})^2,$$

$$\hat{\pi}_k \leftarrow \frac{1}{n} \sum_{i=1}^{n} \hat{\tau}_{ik}$$

for $j \leq p$ and $k \leq K$, where

$$\tilde{\boldsymbol{\mu}}_k = \sum_{i=1}^{n} \hat{\tau}_{ik}\mathbf{x}_i - n\hat{\boldsymbol{\Sigma}}\tilde{\Delta}_k, \quad \tilde{\Delta}_k = \frac{\partial \mathbf{J}_\lambda(\hat{\boldsymbol{\mu}}_k)}{\partial \boldsymbol{\mu}_k} - \lambda \text{sign}(\hat{\boldsymbol{\mu}}_k).$$

In this algorithm, with $\lambda = 0$, one produces the non-penalized estimator of $\boldsymbol{\theta}$, and if we put $\tilde{\Delta}_k = 0$ for all $k \leq K$, the derived solutions are the same as those of the LASSO (Pan and Shen (2007)). To select the tuning parameter λ and the appropriate number of the components K, we used the modified Bayesian information criterion (BIC),

$$\text{BIC}(\hat{\boldsymbol{\theta}}) = -2L(\hat{\boldsymbol{\theta}}; \mathbf{X}) + \text{DF}(\hat{\boldsymbol{\theta}}) \log n, \tag{5.8}$$

with $\text{DF}(\hat{\boldsymbol{\theta}}) = K - 1 + p + q$, where q is number of the nonzero mean components.

Table 6 summarizes the results, which show that the SCAD and LASSO performed similarly. Note that the SCAD penalty yielded sparser solutions and smaller BIC values than the LASSO, even though the differences were not significant. In other words, the solutions obtained by the SCAD penalty represented the underlying distribution of the given data more compactly. It is interesting to note that the number of genes selected for clustering was rather large. In fact, the same clustering structure can be constructed with fewer genes (three or four hundred genes), even though such models have higher BIC values. An alternative model selection criterion may be required for the mixture model, which we plan to investigate in the future.

6. Concluding Remarks

We have only considered the SCAD penalty for easy exposition; it is not difficult to construct a class of penalty functions that have the same local and global asymptotic properties. For example, consider the family of penalty functions $J_\lambda(\cdot)$ that are non-decreasing and continuously differentiable on $(0, \infty)$ such that $\lim_{\theta \to 0^+} \partial J_\lambda(\theta)/\partial\theta = \lambda$ and $\partial J_\lambda(\theta)/\partial\theta = 0$ for $\theta \geq a\lambda$ for some $a > 0$. Then, it can be shown that the penalized MLE from this class has the same asymptotic properties as the SCAD-penalized MLE. This is because the KKT conditions for the oracle MLE to be a local maximizer are the same as those for the SCAD. For example, the minimax concave penalty of Zhang (2010) belongs to this penalty class.

The results of this study can be easily extended to the general penalized M-estimator as long as the the loss function is sufficiently smooth (e.g., the third derivative exists). However, there are problems when the loss functions are not sufficiently smooth. Examples are the hinge loss for the support vector machine and the Huber loss for robust regression. It would be advantageous to extend the results of this paper to such non-smooth loss functions.

Table 6. Model-based clustering results for two gene expression data sets: Leukemia and Colon.

		Cluster number						
Leukemia		**1**	**2**	**3**	**4**	**5**	Total	BIC/n
SCAD	ALL	10	11	6	0	-	27	1598.35
	AML	0	0	0	11	-	11	
	No. of nonzeros	894	915	847	905	-	3561	
LASSO	ALL	10	9	8	0	-	27	1625.21
	AML	0	0	0	11	-	11	
	No. of nonzeros	978	977	960	978	-	3893	
		Cluster number						
Colon		**1**	**2**	**3**	**4**	**5**	Total	BIC/n
SCAD	Normal	11	2	1	7	1	22	1334.49
	Tumor	3	2	22	0	13	40	
	No. of nonzeros	929	767	956	875	937	4464	
LASSO	Normal	11	2	1	7	1	22	1350.67
	Tumor	3	3	21	0	13	40	
	No. of nonzeros	985	957	992	974	991	4899	

Acknowledgement

We are grateful to the anonymous referees, an associate editor, and the Editor for their insightful comments. Kwon's research was supported by the Engineering Research Center of Excellence Program of the Korea Ministry of Education, Science and Technology (MEST)/Korea Science and Engineering Foundation (KOSEF), grant number R11-2008-007-01002-0. Kim's work was supported by the National Research Foundation of Korea grant number 20100012671 funded by the Korea government.

Appendix

We define some notation. Let $S_{nj}(\boldsymbol{\theta}_n)$ be the jth element of $\nabla L_n(\boldsymbol{\theta}_n) = \partial L_n(\boldsymbol{\theta}_n)/\partial \boldsymbol{\theta}_n$ and $U_{njl}(\boldsymbol{\theta}_n)$ be the (j,l)-th element of $\nabla^2 L_n(\boldsymbol{\theta}_n) = \partial^2 L_n(\boldsymbol{\theta}_n)/\partial \boldsymbol{\theta}_n^2$ for all $j, l \le p_n$. Similarly, ∇_1 denotes some partial derivatives of $S_{nj}(\boldsymbol{\theta}_n)$ with respect to $\boldsymbol{\theta}_{n1} = (\theta_{n1}, \cdots, \theta_{nq_n})^T$, so that $\nabla_1 S_{nj}(\boldsymbol{\theta}_n) = \partial S_{nj}(\boldsymbol{\theta}_n)/\partial \boldsymbol{\theta}_{n1}$ and $\nabla_1^2 S_{nj}(\boldsymbol{\theta}_n) = \partial^2 S_{nj}(\boldsymbol{\theta}_n)/\partial \boldsymbol{\theta}_{n1}^2$ for all $j \le p_n$.

Lemma A.1. *If* **A2**−**A5** *hold, for any constant* $\alpha > 0$*, we have*

$$\mathrm{pr}\Big(|S_{nj}(\boldsymbol{\theta}_n^*)| > \sqrt{n}\alpha \Big) = O\left(\alpha^{-2k}\right), \tag{A.1}$$

$$\mathrm{pr}\Big(||\nabla_1 S_{nj}(\boldsymbol{\theta}_n^*) - E(\nabla_1 S_{nj}(\boldsymbol{\theta}_n^*))|| > \sqrt{q_n}\alpha \Big) = O\left(\alpha^{-2k}\right), \tag{A.2}$$

for all $j \le p_n$. For all $\boldsymbol{\theta}_n \in B_n$, and for $j \le p_n$,

$$\mathrm{pr}\Big(||\nabla_1^2 S_{nj}(\boldsymbol{\theta}_n)|| > nq_n\alpha\Big) = O\left(\alpha^{-2k}\right). \tag{A.3}$$

Proof of Lemma A.1 Under **A2**, **A5** and the Rosenthal inequality, we have

$$E\Big(S_{nj}(\boldsymbol{\theta}_n^*)\Big)^{2k} = O\left(n^k\right)$$

for all $j \le p_n$. Using the Markov inequality, it is easy to check

$$\mathrm{pr}\Big(\,|S_{nj}(\boldsymbol{\theta}_n^*)| > \sqrt{n}\alpha\Big) \le (\sqrt{n}\alpha)^{-2k} E\Big(\,|S_{nj}(\boldsymbol{\theta}_n^*)|\,\Big)^{2k} = O\left(\alpha^{-2k}\right).$$

Hence, (A.1) holds. Let $\Delta_{njl}(\boldsymbol{\theta}_n^*) = U_{njl}(\boldsymbol{\theta}_n^*) - E(U_{njl}(\boldsymbol{\theta}_n^*))$. Then, from **A2**, **A5** and the Rosenthal inequality, we get

$$E\big(\Delta_{njl}(\boldsymbol{\theta}_n^*)\big)^{2k} = O\left(n^k\right)$$

for all $l \le q_n$ and $j \le p_n$. Using the triangular inequality in L_k, we have

$$\begin{aligned} E\Big(||\nabla_1 S_{nj}(\boldsymbol{\theta}_n^*) - E\nabla_1 S_{nj}(\boldsymbol{\theta}_n^*)||^{2k}\Big) &= E\Big\{\sum_{l=1}^{q_n}\big(\Delta_{njl}(\boldsymbol{\theta}_n^*)\big)^2\Big\}^k \\ &\le \Big[\sum_{l=1}^{q_n}\Big\{E\big(\Delta_{njl}(\boldsymbol{\theta}_n^*)\big)^{2k}\Big\}^{1/k}\Big]^k \\ &= O\big(nq_n^k\big) \end{aligned}$$

for all $j \le p_n$. Hence, using the Markov inequality again, (A.2) follows. Similarly, under **A4** and **A5**,

$$E\Big(\sum_{i=1}^{n} V_{njlm}(\mathbf{z}_{ni})\Big)^{2k} = O(n^{2k})$$

holds, from which we can conclude

$$E\Big(||\nabla_1^2 S_{nj}(\boldsymbol{\theta}_n)||\Big)^{2k} \le O\Big((nq_n)^{2k}\Big)$$

for all $\boldsymbol{\theta}_n \in B_n$ and $j \le p_n$. Hence, (A.3) follows. This completes the proof.

Proof of Theorem 1. Let

$$U_n(\boldsymbol{\theta}_n) = L_n(\boldsymbol{\theta}_n) - n\sum_{j=1}^{p_n}\Big(J_{\lambda_n}(\theta_{nj}) - \lambda_n|\theta_{nj}|\Big).$$

Then, $U_n(\cdot)$ is continuously differentiable so that we have

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

for all $j \le p_n$. Since $Q_n(\boldsymbol{\theta}_n) = U_n(\boldsymbol{\theta}_n) - n\lambda_n \sum_{j=1}^{p_n} |\theta_{nj}|$, the corresponding Karush-Kuhn-Tucker (KKT) conditions (see, for example, Rosset and Zhu (2007)) are

$$\frac{\partial U_n(\boldsymbol{\theta}_n)}{\partial \theta_{nj}} = n\lambda_n \text{sign}(\theta_{nj}),\ \theta_{nj} \neq 0, \tag{A.4}$$

$$\left|\frac{\partial U_n(\boldsymbol{\theta}_n)}{\partial \theta_{nj}}\right| \le n\lambda_n,\ \theta_{nj} = 0, \tag{A.5}$$

for all $j \le p_n$. By the definition of $\hat{\boldsymbol{\theta}}_n^o$, $S_{nj}(\hat{\boldsymbol{\theta}}_n^o) = 0$ for $j \le q_n$ and $\hat{\theta}_{nj}^o = 0$ for $q_n < j \le p_n$. Hence, it suffices to show that $\hat{\boldsymbol{\theta}}_n^o$ satisfies

$$\text{pr}\left(\min_{1\le j\le q_n} |\hat{\theta}_{nj}^o| \ge a\lambda_n\right) \to 1, \tag{A.6}$$

$$\text{pr}\left(\max_{q_n < j\le p_n} |S_{nj}(\hat{\boldsymbol{\theta}}_n^o)| \le n\lambda_n\right) \to 1, \tag{A.7}$$

as $n \to \infty$. From the regularity condition **A1** and (3.2), we have

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

Hence, (A.6) follows since $\lambda_n = o\left(n^{-(1-c_2+c_1)/2}\right)$. Next, we prove (A.7). From Taylor's expansion and the definition of $\hat{\boldsymbol{\theta}}_n^o$ and $\boldsymbol{\theta}_n^*$, we have

$$\begin{aligned} S_{nj}(\hat{\boldsymbol{\theta}}_n^o) = S_{nj}(\boldsymbol{\theta}_n^*) &+ \nabla_1 S_{nj}(\boldsymbol{\theta}_n^*)^T(\hat{\boldsymbol{\theta}}_{n1}^o - \boldsymbol{\theta}_{n1}^*) \\ &+ (\hat{\boldsymbol{\theta}}_{n1}^o - \boldsymbol{\theta}_{n1}^*)^T \nabla_1^2 S_{nj}(\boldsymbol{\theta}_n^{**}) \frac{(\hat{\boldsymbol{\theta}}_{n1}^o - \boldsymbol{\theta}_{n1}^*)}{2} \end{aligned}$$

for all $q_n < j \le p_n$ for some $\boldsymbol{\theta}_n^{**}$ that lies between $\hat{\boldsymbol{\theta}}_n^o$ and $\boldsymbol{\theta}_n^*$. From the Cauchy-

Schwarz inequality, it follows that

$$\begin{aligned}
&\mathrm{pr}\Big(\max_{q_n<j\le p_n}|S_{nj}(\hat{\boldsymbol{\theta}}_n^o)|>n\lambda_n\Big)\\
&\le\ \mathrm{pr}\Big(\max_{q_n<j\le p_n}|S_{nj}(\boldsymbol{\theta}_n^*)|>\frac{n\lambda_n}{4}\Big)\\
&\quad+\mathrm{pr}\Big(\max_{q_n<j\le p_n}||\nabla_1 S_{nj}(\boldsymbol{\theta}_n^*)-E\nabla_1 S_{nj}(\boldsymbol{\theta}_n^*)||\ ||\hat{\boldsymbol{\theta}}_n^o-\boldsymbol{\theta}_n^*||>\frac{n\lambda_n}{4}\Big)\\
&\quad+\mathrm{pr}\Big(\max_{q_n<j\le p_n}||E\nabla_1 S_{nj}(\boldsymbol{\theta}_n^*)||\ ||\hat{\boldsymbol{\theta}}_n^o-\boldsymbol{\theta}_n^*||>\frac{n\lambda_n}{4}\Big)\\
&\quad+\mathrm{pr}\Big(\max_{q_n<j\le p_n}||\nabla_1^2 S_{nj}(\boldsymbol{\theta}_n^{**})||\ ||\hat{\boldsymbol{\theta}}_n^o-\boldsymbol{\theta}_n^*||^2>\frac{n\lambda_n}{2}\Big)\\
&=^{let}\mathbf{P}_1+\mathbf{P}_2+\mathbf{P}_3+\mathbf{P}_4.
\end{aligned}$$

From Lemma A.1, we can see that

$$\begin{aligned}
\mathbf{P}_1&=\mathrm{pr}\Big(\max_{q_n<j\le p_n}|S_{nj}(\boldsymbol{\theta}_n)|>\frac{n\lambda_n}{4}\Big)\\
&\le\sum_{j=q_n+1}^{p_n}\mathrm{pr}\Big(|S_{nj}(\boldsymbol{\theta}_n)|>\frac{n\lambda_n}{4}\Big)\\
&=O\Big(\frac{p_n}{(\sqrt{n}\lambda_n)^{2k}}\Big)\to 0
\end{aligned}$$

as $n\to\infty$. Similarly, for the term $\mathbf{P}_2$, from Lemma A.1 we have

$$\begin{aligned}
\mathbf{P}_2&\le\mathrm{pr}\Big(||\hat{\boldsymbol{\theta}}_n^o-\boldsymbol{\theta}_n^*||>\frac{q_n}{\sqrt{n}}\Big)\\
&\quad+\ \mathrm{pr}\Big(\max_{q_n<j\le p_n}||\nabla_1 S_{nj}(\boldsymbol{\theta}_n^*)-E\nabla_1 S_{nj}(\boldsymbol{\theta}_n^*)||>\frac{n\sqrt{n}\lambda_n}{4q_n}\Big)\\
&=o(1)+O\Big(\frac{p_n}{(n\lambda_n/(q_n\sqrt{q_n}))^{2k}}\Big)\to 0
\end{aligned}$$

as $n\to\infty$. From **A5**, we have

$$\begin{aligned}
\mathbf{P}_3&=\mathrm{pr}\Big(\max_{q_n<j\le p_n}||E\nabla_1 S_{nj}(\boldsymbol{\theta}_n^*)||\ ||\hat{\boldsymbol{\theta}}_n^o-\boldsymbol{\theta}_n^*||>\frac{n\lambda_n}{4}\Big)\\
&\le\mathrm{pr}\Big(||\hat{\boldsymbol{\theta}}_n^o-\boldsymbol{\theta}_n^*||>\frac{n\lambda_n}{4M_5\sqrt{q}_n}\Big)\to 0
\end{aligned}$$

as $n\to\infty$. For the last term $\mathbf{P}_4$, from Lemma A.1 we have

$$\begin{aligned}
\mathbf{P}_4&\le\mathrm{pr}\Big(||\hat{\boldsymbol{\theta}}_n^o-\boldsymbol{\theta}_n^*||^2>\frac{q_n\sqrt{q_n}}{n}\Big)+\mathrm{pr}\Big(\max_{q_n<j\le p_n}||\nabla_1^2 S_{nj}(\boldsymbol{\theta}_n^{**})||>\frac{n^2\lambda_n}{2q_n\sqrt{q_n}}\Big)\\
&=o(1)+O\Big(\frac{p_n}{(n\lambda_n/(q_n^2\sqrt{q_n}))^{2k}}\Big)\to 0
\end{aligned}$$

as $n \to \infty$. Hence, (A.7) follows. This completes the proof.

Proof of Theorem 2. It suffices to show that

$$\mathrm{pr}\left(\max_{\boldsymbol{\theta}_n \in \Omega_n} Q_n(\boldsymbol{\theta}_n) \le Q_n(\hat{\boldsymbol{\theta}}_n^o)\right) \to 1 \tag{A.8}$$

as $n \to \infty$. From Taylor's expansion, we get

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

for some $\boldsymbol{\theta}_n^{**}$ that lies between $\boldsymbol{\theta}_n$ and $\hat{\boldsymbol{\theta}}_n^o$. The definition of $\hat{\boldsymbol{\theta}}_n^o$ together with (A.7) implies that

$$\nabla L_n(\hat{\boldsymbol{\theta}}_n^o)^T(\boldsymbol{\theta}_n - \hat{\boldsymbol{\theta}}_n^o) = \sum_{j=1}^{p_n} S_{nj}(\hat{\boldsymbol{\theta}}_n^o)(\theta_{nj} - \hat{\theta}_{nj}^o) \le \sum_{j=q_n+1}^{p_n} o_p(n\lambda_n)|\theta_{nj}|$$

and from **A6** and the Cauchy-Schwarz inequality,

$$(\boldsymbol{\theta}_n - \hat{\boldsymbol{\theta}}_n^o)^T \nabla^2 L_n(\hat{\boldsymbol{\theta}}_n^{**})(\boldsymbol{\theta}_n - \hat{\boldsymbol{\theta}}_n^o)/2 \le -nM_7 ||\boldsymbol{\theta}_n - \hat{\boldsymbol{\theta}}_n^o||^2$$

holds. Hence, we have

$$Q_n(\boldsymbol{\theta}_n) - Q_n(\hat{\boldsymbol{\theta}}_n^o) \le \sum_{j=1}^{p_n} n w_{nj},$$

where

$$w_{nj} = o_p(\lambda_n)|\theta_{nj}| I(j > q_n) - M_7(\theta_{nj} - \hat{\theta}_{nj}^o)^2 + J_{\lambda_n}(\hat{\theta}_{nj}^o) - J_{\lambda_n}(\theta_{nj}).$$

If $|\theta_{nj}| \ge a\lambda_n$ for all $j \le q_n$, we have

$$\sum_{j=1}^{q_n} w_{nj} \le -M_7 \sum_{j=1}^{q_n} (\theta_{nj} - \hat{\theta}_{nj}^o)^2 \le 0.$$

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

$$|\theta_{nj} - \hat{\theta}_{nj}^o| \ge \min_{1 \le j \le q_n} |\theta_{nj}^*| - \max_{1 \le j \le q_n} |\hat{\theta}_{nj}^o - \theta_{nj}^*| - a\lambda_n = O_p(n^{-(1-c_2)/2}).$$

Hence, we have

$$\sum_{j=1}^{q_n} w_{nj} \le -O_p(n^{-1+c_2}) + O(q_n \lambda_n^2) = -O_p(n^{-1+c_2}) + o(n^{-1+c_2}) \le 0$$

for sufficiently large n. On the other hand, for each $j > q_n$, if $|\theta_{nj}| > \lambda_n$,

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

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

$$w_{nj} \leq o_p(\lambda_n)|\theta_{nj}| - J_{\lambda_n}(\theta_{nj}) = (o_p(\lambda_n) - \lambda_n)|\theta_{nj}|.$$

Hence, we have $\sum_{j=q_n+1}^{p_n} w_{nj} \leq 0$ for all sufficiently large n. As a consequence, (A.8) holds. This completes the proof.

References

Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D. and Levine, A. J. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA* **96**, 6745-6750.

Antoniadis, A. and Fan, J. (2001). Regularization of wavelets approximations. *J. Amer. Statist. Assoc.* **96**, 939-967.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection. *Ann. Statist.* **24**, 2350-2383.

Chib, S. (1998). Estimation and comparison of multiple change-point models. *Journal of Econometrics* **86**, 221-241.

Dudoit, S., Fridlyand, J. and Speed, T. P. (2002). Comparison of discrimination methods for the classifcation of tumors using gene expression data. *J. Amer. Statist. Assoc.* **97**, 77-87.

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

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

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica* **20**, 101-148.

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *Ann. Statist.* **32**, 928-961.

Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics* **35**, 109-148.

Friedman, J. H. (2008). Fast sparse regression and classification. Unpublished Manuscript.

Green, P. J. (1995). Reversible jump Markov Chain Monte Carlo computation and bayesian model determination. *Biometrika* **82**, 711-732.

Harchaoui, Z. and Lévy-Leduc, C. (2007). Catching change-points with lasso. *NIPS 2007, collection.*

Jarrett, R. G. (1979). A note on the intervals between coal-mining disasters. *Biometrika* **66**, 191-193.

Kim, Y., Choi, H. and Oh, H. (2008). Smoothly clipped absolute deviation on high dimensions. *J. Amer. Statist. Assoc.* **103**, 1656-1673.

Kim, Y. and Kwon, S. (2011). The global optimality of nonconvex penalized estimators. Unpublished manuscript.

Knight, K. and Fu, W. J. (2000). Asymptotics for lasso-type estimators. *Ann. Statist.* **28**, 1356-1378.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.* **34**, 1436-1462.

Pan, W. and Shen, X. (2007). Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research* **8**, 1145-1164.

Park, M. and Hastie, T. (2007). ℓ_1-regularization path algorithm for generalized linear models. *J. Roy. Statist. Soc. Ser. B* **69**, 659-667.

Raftery, A. E. and Akman, V. E. (1986). Bayesian analysis of a poisson process with a change point. *Biometrika* **73**, 85-89.

Rosset, S. and Zhu, J. (2007). Piecewise linear regularized solution paths. *Ann. Statist.* **35**, 1012-1030.

Tibshirani, R. J. (1996). Regression shrinkage and selection via the LASSO. *J. Roy. Statist. Soc. Ser. B* **58**, 267-288.

Zhang, C. (2010). Nearly unbiased variable selcetion under minimax concave penalty. *Ann. Statist.* **38**, 894-942.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *J. Machine Learning Res.* **7**, 2541-2563.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.* **101**, 1418-1429.

School of Statistics, University of Minnesota, 313 Ford Hall, 224 Church St SE, Minneapolis, MN 55455, USA.

E-mail: shkwon0522@gmail.com

Department of Statistics, Seoul National University, Gwanak-gu, Seoul 151-742, Korea.

E-mail: ydkim0903@gmail.com

(Received February 2010; accepted May 2011)