

A new sparse variable selection via random-effect model



Youngjo Lee, Hee-Seok Oh*

Department of Statistics, Seoul National University, Seoul 151-747, Republic of Korea

ARTICLE INFO

Article history:

Received 24 December 2010

Available online 17 December 2013

AMS subject classifications:

62J07

62F30

Keywords:

Maximum likelihood estimator

Prediction

Random-effect models

Sparsity

Variable selection

ABSTRACT

We study a new approach to simultaneous variable selection and estimation via random-effect models. Introducing random effects as the solution of a regularization problem is a flexible paradigm and accommodates likelihood interpretation for variable selection. This approach leads to a new type of penalty, unbounded at the origin and provides an oracle estimator without requiring a stringent condition. The unbounded penalty greatly enhances the performance of variable selections, enabling highly accurate estimations, especially in sparse cases. Maximum likelihood estimation is effective in enabling sparse variable selection. We also study an adaptive penalty selection method to maintain a good prediction performance in cases where the variable selection is ineffective.

© 2013 Elsevier Inc. All rights reserved.

1. Introduction

Consider the regression model

$$y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \quad i = 1, \dots, n, \quad (1)$$

where $\boldsymbol{\beta}$ is a $d \times 1$ vector of fixed unknown parameters and the ε 's are white noises with mean 0 and finite variance ϕ . This study aims to effectively select significant variables, while maintaining good estimation and prediction accuracy.

Many variable selection procedures can be described as penalized least squares (PLS) estimation methods that minimize

$$Q_\lambda(\boldsymbol{\beta}) = \frac{1}{2} \sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \sum_{j=1}^d p_\lambda(|\beta_j|), \quad (2)$$

where $p_\lambda(\cdot)$ is a penalty function controlling model complexity. With the entropy or L_0 -penalty, namely, $p_\lambda(|\beta_j|) = \lambda I(|\beta_j| \neq 0)$, the PLS becomes

$$\frac{1}{2} \sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2 + \lambda |M|,$$

where $M = \sum_{j=1}^d I(|\beta_j| \neq 0)$ denotes the size of the candidate model. This leads to traditional variable selection procedures, which have two fundamental limitations. First, when the number of predictors d is large, it is computationally infeasible to perform subset selection. Second, subset selection is extremely variable because of its inherent discreteness [1,6]. To overcome these difficulties, several other penalties have been proposed. With the L_1 -penalty, specifically, the PLS estimator becomes the least absolute shrinkage and selection operator (LASSO), which sets thresholds for predictors with small estimated coefficients [14]. LASSO is a popular technique for simultaneous estimation and variable selection, ensuring high prediction accuracy, and enabling the discovery of relevant predictive variables. Prediction accuracy is often improved by

* Corresponding author.

E-mail addresses: heeseok@stats.snu.ac.kr, heeseok.oh@gmail.com (H.-S. Oh).

shrinking [4] or setting some coefficients to zero by thresholding [2]. Tibshirani [14] gave a comprehensive overview of LASSO as a method of PLS.

LASSO has been criticized on the grounds that it typically ends up selecting a model with too many variables to prevent over shrinkage of the regression coefficients [13]; otherwise, regression coefficients of selected variables are often over shrunk. To improve LASSO, various other penalties have been proposed. Fan and Li [6] proposed the smoothly clipped absolute deviation (SCAD) penalty for oracle estimators. More recently, Zou [16] proposed the adaptive LASSO.

Ridge regression often achieves good prediction performance through a bias–variance trade-off, whereas it cannot produce a parsimonious model. Variable selection is particularly important in the interpretation of the model, especially when the true underlying model has a sparse representation. Identifying null predictors enhances the estimation accuracies of the fitted model. We show that the use of the new penalty greatly improves variable selection to enhance estimation performance, especially in sparse cases. Zou and Hastie [17] noted that the prediction performances of the LASSO can be poor in cases where variable selection is ineffective. To overcome this difficulty, they proposed the elastic net which improves the prediction of LASSO. We propose an adaptive penalty selection for better prediction without hampering the variable selection performance of our method. Through numerical analysis, we show that the proposed adaptive method outperforms LASSO uniformly in variable selection, estimation, and prediction.

Until now, finite penalties, leading to unimodal penalized likelihoods (PL), have been studied. Singularities (unbounded likelihood) have been believed to occur when the description of the process generating the observation is not adequate. This may be considered to be the product of unacceptable probability models [3]. In this paper, we show that the use of singular likelihood (unbounded penalty) at the origin greatly enhances the performance of variable selection. To be specific, by using the unbounded penalty, we define a local minimizer $\hat{\beta}$ that satisfies

$$\left. \frac{\partial Q_{\lambda}(\beta)}{\partial \beta_j} \right|_{\beta=\hat{\beta}} = 0, \quad \text{for } \hat{\beta}_j \neq 0.$$

Moreover, we investigate the property of $\hat{\beta}$, discuss the practical algorithm for obtaining $\hat{\beta}$, and show its empirical performance through the numerical study.

To achieve these goals, we employ a new random-effect model that generates a family of penalties; the normal-type (bell-shaped), LASSO-type (cusped), and a new (singular) unbounded penalty at the origin. The new unbounded penalty gives an oracle estimators without requiring a stringent condition of Fan and Li [6]. An enhancement of LASSO by Zou [16] could be viewed as the use of an asymptotically unbounded penalty.

There are some analogies between the PL and random-effect model approaches. For example, the iterative weighted least squares (IWLS) estimation for random-effect models can be used for the estimation of β [12]. Random-effect models provide new insights and interpretations of IWLS, which explain how the algorithm overcomes the difficulty in nonconvex optimization problem. However, there are differences between the two approaches. In the PL approach, the penalty need not stem from a statistical model; hence, the tuning parameters cannot be estimated by model likelihood. However, in the random-effect model approach, likelihood methods can be used. In this paper, we follow the PL approach for the tuning parameter to compare various methods under a uniform condition, and highlight that the new unbounded penalty is better than the existing penalties in variable selection.

In Section 2, we present a new unbounded penalty. Numerical studies are presented in Section 3. In Section 4, we show that the resulting PLS estimators satisfy the oracle property of Fan and Li [6], but under mild requirements. Conclusion remarks are given in Section 5. In the Appendix, a new random-effect model is introduced, which can be fitted by an IWLS procedure.

2. Unbounded penalty for variable selection

For simplicity of notation, we omit the subscript when deemed unnecessary. Suppose that β is a random variable such that

$$\beta|u \sim N(0, u\theta), \quad (3)$$

where θ is a dispersion parameter, and u follows the gamma distribution with a parameter w such that

$$f_w(u) = (1/w)^{1/w} \frac{1}{\Gamma(1/w)} u^{1/w-1} e^{-u/w}$$

with $E(u) = 1$ and $\text{Var}(u) = w$. Throughout the paper, $f_{\theta}(\cdot)$ denote the density function with a parameter θ . The h -likelihood of the above random-effect model leads to a new type of unbounded penalty function $p_{\lambda}(\beta)$ that is indexed by w and is parameterized by ϕ of (1) and θ of (3). By the result (8) in the Appendix and Stirling's approximation, the resultant unbounded penalty for a given w can be defined as

$$p_{\lambda}(|\beta|) = \frac{\phi}{2\theta} \frac{\beta^2}{u} + \frac{\phi(w-2)}{w} \log u + \frac{\phi}{w} u,$$

where λ is a function of ϕ and θ , and $u = u(\beta)$ is given by (7) in the Appendix. In this study, we set $\lambda = \phi/\theta$. The detailed procedure for a derivation of $p_{\lambda}(|\beta|)$ from the random-effect model is given in the Appendix.

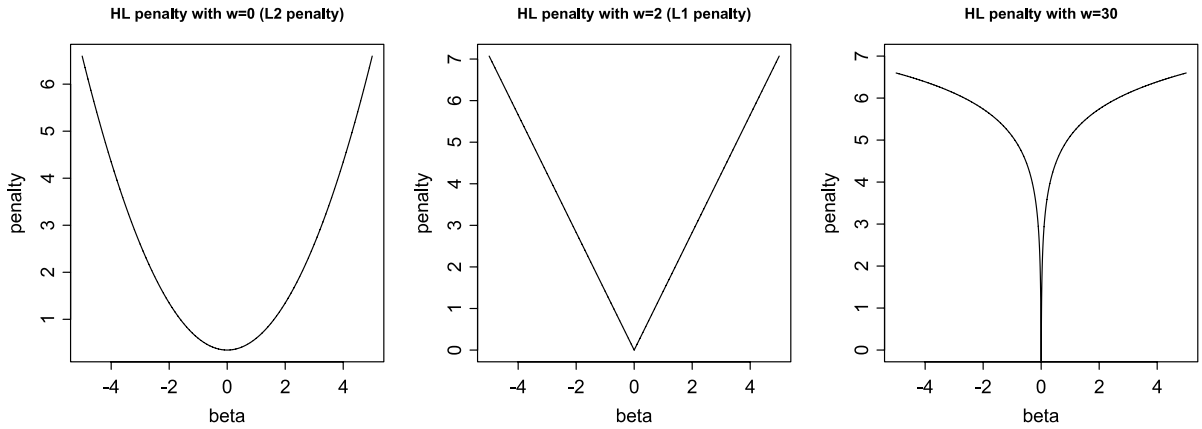


Fig. 1. Penalty functions with different w values.

New unbounded penalties $p_\lambda(\beta)$ according to different values of $w = 0, 2$, and 30 with $\lambda = \phi/\theta = 2$ are shown in Fig. 1. The form of the penalty changes from a quadratic shape ($w = 0$) for ridge regressions to a cusped form ($w = 2$) for LASSO and then to an unbounded form ($w > 2$) at the origin. In the case of $w > 2$, it allows an infinite gain at zero. Bell-shaped penalties have been proposed for aggregation [15] and better prediction [4] of L_2 -penalty ($w = 0$), and cusped ones for simultaneous variable selection and estimation of LASSO ($w = 2$; [14]) and SCAD [5]. To the best of our knowledge, only a finite penalty has been investigated so far. In this paper, we illustrate the advantage of using an unbounded penalty to enhance variable selection. Singularities in LASSO and SCAD imply that their derivatives are not defined at the origin. Given λ , however, both penalties satisfy that $p_\lambda(0) < \infty$ and $|p'_\lambda(0)| < \infty$, while the new unbounded penalty has $p_\lambda(0) = -\infty$ and $|p'_\lambda(0)| = \infty$. The singular penalty $p_\lambda(|\beta_j|) = \lambda|\beta_j|^p$ at the origin for $0 < p < 1$ has also been considered, which, although not differentiable at the origin, still has finite penalty. We remark that the motivation of the proposed method is that although there exist some local solutions which SCAD penalty shares with, employing the unbounded penalty newly for variable selection might produce good performance.

Fig. 2 provides the thresholding functions of LASSO, SCAD, adaptive LASSO, and the proposed h -likelihood based method (HL) with $w = 30$ when the tuning parameter is set to $\lambda = 2$. The solution can be obtained from the Eq. (9) in the Appendix. In this figure, we show that among the three methods, the solutions of HL are closest to that of the adaptive LASSO.

3. Numerical studies

To assess the empirical performance of the proposed methods, we conducted simulation studies for the regression model of (1). We consider various examples below, including cases of aggregation and sparse situations with or without grouped variables. Following Tibshirani [14], Fan and Li [6], and Zou [16], we take $\phi^{1/2} = 3$ in all examples except Examples 3 and 6, where we take $\phi^{1/2} = 2$ and $\phi^{1/2} = 10$, respectively. The first three examples have sample size $n = 40$, and the last three, $n = 100$.

- Example 1. A few large effects with $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)^T$: This case has been studied by Tibshirani [14], Fan and Li [6], and Zou [16]. The predictors \mathbf{x}_i are i.i.d standard normal vectors. The correlation between \mathbf{x}_i and \mathbf{x}_j is $\rho^{|i-j|}$ with $\rho = 0.5$. In this case, the signal-to-noise ratio is approximately 5.7.
- Example 2. Many small effects with $\beta = (0.85, 0.85, \dots, 0.85)^T$: The rest of the settings are the same as in Example 1. This has been studied by Tibshirani [14] and Zou [16]. Here, the signal-to-noise ratio is approximately 1.8. The ridge regression is expected to perform well.
- Example 3. Single large effect with $\beta = (5, 0, 0, 0, 0, 0, 0, 0)^T$: The rest of the settings are the same as in Example 1. This case has been studied by Tibshirani [14]. It represents a typical case in which the significant predictors are very sparse. Here, the signal-to-noise ratio is approximately 7. Example 1 might represent a middle case between Example 2 of an aggregation, and Example 3, a very sparse case.
- Example 4. Inconsistent Lasso path with $\beta = (5.6, 5.6, 5.6, 0)^T$: The predictor variables \mathbf{x}_i are i.i.d $N(\mathbf{0}, C)$, where

$$C = \begin{pmatrix} 1 & \rho_1 & \rho_1 & \rho_2 \\ \rho_1 & 1 & \rho_1 & \rho_2 \\ \rho_1 & \rho_1 & 1 & \rho_2 \\ \rho_2 & \rho_2 & \rho_2 & 1 \end{pmatrix}$$

with $\rho_1 = -0.39$ and $\rho_2 = 0.23$. Zou [16] studied this case and showed that LASSO does not satisfy Fan and Li's [6] oracle property.

- Example 5. A few large grouped effects: This is a relatively large problem with grouped variables. This setting is interesting because only a few grouped variables are significant such that the variables are sparse in terms of groups, but the variables

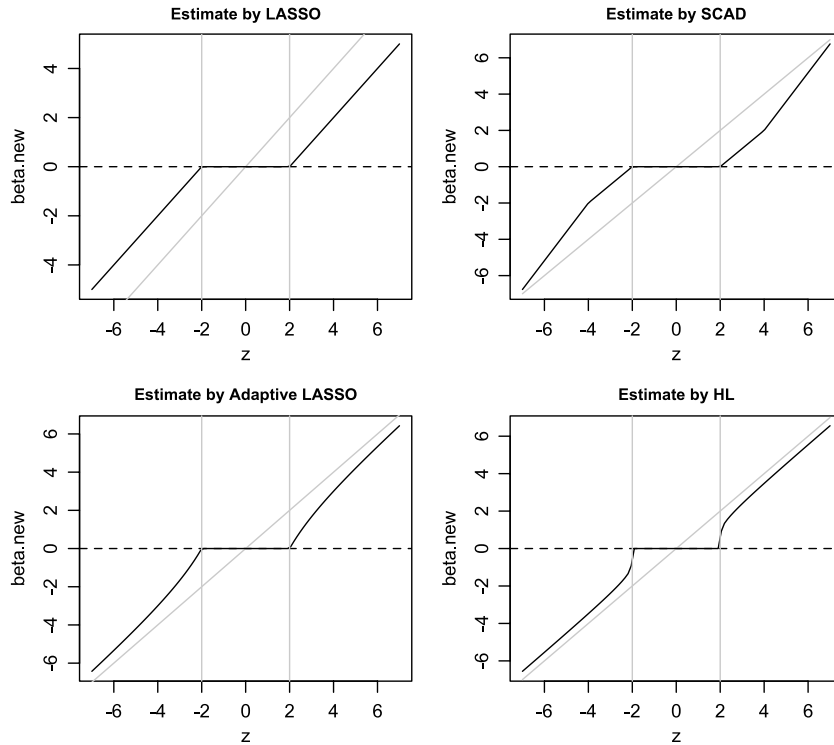


Fig. 2. The solutions by LASSO, SCAD, adaptive LASSO, and HL.

within a group all have the same effects. The true coefficients are

$$\beta = (\underbrace{0, \dots, 0}_{10}, \underbrace{2, \dots, 2}_{10}, \underbrace{0, \dots, 0}_{10}, \underbrace{2, \dots, 2}_{10})^T.$$

The pairwise correlation between \mathbf{x}_i and \mathbf{x}_j is 0.5 for all i and j .

- Example 6. Consider Example 5 to have a low signal-to-noise ratio. Zou and Hastie [17] introduced this case because no variable-selection method works properly despite many null effects. With this example they showed that the LASSO has a poor prediction performance.

We investigated various values w in the range $10 \leq w \leq 270$, and found that the performances were similar; we therefore present the results with $w = 30$. We consider an adaptive HL method (HL(A)) that selects w among $w = 0$ (Ridge), 2 (LASSO), and 30 (unbounded penalty). A comprehensive study of the adaptive choice of w is left for future study.

On the basis of 100 simulated data, we compare eight methods:

- Ridge of Hoerl and Kennard [7]
- Lasso of Tibshirani [14]
- SCAD with $a = 3.7$ of Fan and Li [6]
- Adaptive Lasso of Zou [16]
- HL(F) is the HL method with $w = 30$
- HL(A) is the adaptive HL method
- HL(MF) is the HL method by which (ϕ, θ) are estimated from the adjusted profile h -likelihood and w is fixed as 30
- HL(MA) is the HL method by which (w, ϕ, θ) are estimated from the adjusted profile h -likelihood.

For comparisons with the PL approach under a uniform condition, we use the same algorithm when choosing tuning parameter λ , with the OLS being the initial values for β . The ten-fold cross-validation was applied. We denote the full data set by \mathcal{T} , and the cross-validation training data and test data set by $\mathcal{T} - \mathcal{T}^s$ and \mathcal{T}^s for $s = 1, 2, \dots, 10$, respectively. For each λ , we obtain the estimator $\hat{\beta}_\lambda^s$ with the test data set \mathcal{T}^s removed. Therefore, we compute the cross-validation prediction mean squares error (PMSE)

$$CV(\lambda) = \frac{1}{n} \sum_{s=1}^{10} \sum_{(y_k, \mathbf{x}_k) \in \mathcal{T}^s} (y_k - \mathbf{x}_k^T \hat{\beta}_\lambda^s)^2.$$

Table 1

The proportion of selecting the true model with the median number of incorrect zeros in parentheses.

Method	Ridge	Lasso	SCAD	Adaptive Lasso	HL(F)	HL(A)	HL(MF)	HL(MA)
Ex 1	0.00 (0.0)	0.03 (0.0)	0.04 (0.0)	0.28 (0.0)	0.59 (0.0)	0.38 (0.0)	0.67 (0.0)	0.66 (0.0)
Ex 2	1.00 (0.0)	0.53 (0.0)	0.54 (0.5)	0.09 (2.0)	0.00 (4.0)	0.93 (0.0)	0.29 (4.0)	0.34 (4.0)
Ex 3	0.00 (0.0)	0.15 (0.0)	0.15 (0.0)	0.50 (0.0)	0.70 (0.0)	0.55 (0.0)	0.85 (0.0)	0.85 (0.0)
Ex 4	0.00 (0.0)	0.23 (0.0)	0.28 (0.0)	0.68 (0.0)	0.87 (0.0)	0.72 (0.0)	0.94 (0.0)	0.86 (0.0)
Ex 5	0.00 (0.0)	0.00 (0.0)	0.00 (0.0)	0.11 (0.0)	0.29 (0.0)	0.25 (0.0)	0.32 (0.0)	0.28 (0.0)
Ex 6	0.00 (0.0)	0.00 (3.0)	0.00 (2.0)	0.00 (5.0)	0.00 (9.0)	0.00 (0.0)	0.00 (9.0)	0.00 (9.0)

We use the value of λ that minimizes $CV(\lambda)$. Note that we also tried generalized cross-validation (GCV) in selecting λ ; however, the results were similar to that obtained by the ten-fold cross-validation (CV), and hence, we omitted the results by the GCV. In (7) of Appendix, we need the estimate of θ in order to obtain \hat{u} when $w = 30$. We use $\hat{\theta} = \hat{\phi}/\hat{\lambda}$, where $\hat{\phi} = \frac{1}{n-d} \sum_{i=1}^n (y_i - \mathbf{x}_i^T \hat{\beta}_{OLS})^2$, and $\hat{\lambda}$ is the value chosen by the CV above.

In the PL approach, since w and θ are not model parameters, the CV method has been proposed. In random-effect models, we can use the maximum likelihood (ML) estimates for (w, ϕ, θ) , the estimation of which is faster than that by the CV method. Thus, we also compare the CV and ML methods in choosing the tuning parameters.

We consider several measures to evaluate the proposed methods. With regard to the performance measures of variable selection, we report the proportion of cases where the true model (correctly identifying all null coefficients) and the median number of incorrect zeros (depicting the median of coefficients erroneously set to zeros) are selected. In the case of performance measures for estimation, we consider the mean squares error (MSE) and mean absolute error (MAE) of estimators

$$MSE(\hat{\beta}) = \frac{1}{\text{no. of coefficients}} \sum_i (\beta_i - \hat{\beta}_i)^2$$

and

$$MAE(\hat{\beta}) = \frac{1}{\text{no. of coefficients}} \sum_i |\beta_i - \hat{\beta}_i|.$$

From the results of Tables 1 and 2, we obtain the following empirical observations:

1. With regard to variable selection, the performances of LASSO and SCAD are similar. Adaptive LASSO and HL(F), HL(MF), and HL(MA) greatly improve other methods for all sparse cases. The HL(MF) has the best variable selection performance in sparse cases. The ML estimation for tuning parameters is better than that by the CV methods. In Example 2, no simultaneous variable selection and estimation methods work but HL(A).
2. With regard to estimation performance, the HL(MF) and ML(MA) outperform the others in all cases except Examples 2 and 6, where the ridge regression and HL(A) produce almost identically the best results.
3. The HL(MA) selects the penalty type by estimating w , but often it obtains a large estimate, such that HL(MA) and HL(MF) tend to be similar. The HL(A) method greatly enhances the estimation performance of the HL methods in Examples 2 and 6. Overall, it has the best performance.

Zou and Hastie [17] proposed the use of the elastic net, considering a penalty $p_\lambda(|\beta_j|) = \lambda\{(1 - \alpha)|\beta_j| + \alpha\beta_j^2\}$ with $0 \leq \alpha \leq 1$. It yields a less sparse solution than LASSO but has better prediction. Moreover, it greatly improves prediction power by sacrificing the performance of variable selection of the LASSO. As a measure of the prediction power, we consider the test error $\sum_{i=1}^n (y_i - \hat{y}_i)^2/n$. By following Zou and Hastie [17], in each example, we generate a training set (100 observations), an independent validation set (100 observations), and an independent test set (200 observations). Then, for 50 simulated data sets, we compute the test error on the test sets in Table 3. We also tried only the training set and the independent test set, but obtained similar results. For the methods using the ML estimates for tuning parameters we do not need the independent validation set because in random-effect models, all parameters are model parameters.

In the sparse cases, HL(F), HL(MF), and HL(MA) provide results almost identical to the best method or outperform other methods. In Examples 2 and 6, HL(A) method gives results that are similar to those obtained by ridge regression. Therefore, the unbounded penalty has merit in variable selection and estimation. Good variable selection enhances the estimation performance. Overall, HL(A) performs well, leaning toward the best in variable selection, estimation, and prediction, while HL(MF) and HL(MA) are good choices in the sparse cases. In the absence of such information, we may use the HL(A) method, which outperforms LASSO uniformly in variable selection, estimation, and prediction.

Table 2

The medians of MSE (first line) and MAE (second line) values with the median absolute deviation in parentheses.

Method	Ridge	Lasso	SCAD	Adaptive Lasso
Ex 1	0.264 (0.124) 0.414 (0.118)	0.187 (0.147) 0.291 (0.141)	0.187 (0.147) 0.290 (0.142)	0.186 (0.173) 0.242 (0.144)
Ex 2	0.134 (0.067) 0.293 (0.089)	0.267 (0.122) 0.437 (0.111)	0.267 (0.112) 0.436 (0.107)	0.439 (0.165) 0.564 (0.126)
Ex 3	0.139 (0.086) 0.306 (0.091)	0.043 (0.041) 0.100 (0.068)	0.043 (0.039) 0.099 (0.069)	0.025 (0.032) 0.065 (0.065)
Ex 4	0.229 (0.225) 0.441 (0.260)	0.212 (0.248) 0.348 (0.268)	0.201 (0.217) 0.358 (0.266)	0.167 (0.190) 0.307 (0.222)
Ex 5	0.149 (0.030) 0.312 (0.031)	0.123 (0.029) 0.234 (0.041)	0.125 (0.038) 0.240 (0.041)	0.126 (0.041) 0.219 (0.042)
Ex 6	0.714 (0.157) 0.703 (0.083)	1.123 (0.276) 0.735 (0.116)	1.131 (0.274) 0.736 (0.118)	1.601 (0.311) 0.866 (0.115)
Method	HL(F)	HL(A)	HL(MF)	HL(MA)
Ex 1	0.145 (0.156) 0.204 (0.130)	0.198 (0.216) 0.276 (0.202)	0.145 (0.160) 0.199 (0.127)	0.145 (0.159) 0.199 (0.127)
Ex 2	0.573 (0.176) 0.695 (0.140)	0.136 (0.074) 0.297 (0.090)	0.712 (0.093) 0.818 (0.081)	0.605 (0.179) 0.735 (0.134)
Ex 3	0.013 (0.018) 0.041 (0.039)	0.022 (0.032) 0.054 (0.062)	0.011 (0.014) 0.038 (0.026)	0.011 (0.014) 0.038 (0.026)
Ex 4	0.127 (0.129) 0.272 (0.161)	0.139 (0.150) 0.281 (0.198)	0.110 (0.102) 0.256 (0.144)	0.109 (0.103) 0.254 (0.134)
Ex 5	0.107 (0.024) 0.191 (0.030)	0.111 (0.028) 0.198 (0.036)	0.105 (0.025) 0.188 (0.028)	0.106 (0.024) 0.189 (0.030)
Ex 6	1.846 (0.347) 0.898 (0.125)	0.708 (0.151) 0.700 (0.086)	1.700 (0.345) 0.856 (0.123)	1.702 (0.334) 0.853 (0.119)

Table 3

The median of test error values with the median absolute deviation in parentheses.

Method	Ridge	Lasso	SCAD	Adaptive Lasso
Ex 1	9.491 (1.033)	9.321 (0.922)	9.320 (0.922)	9.354 (1.030)
Ex 2	9.403 (0.920)	9.593 (1.004)	9.593 (1.003)	9.753 (1.042)
Ex 3	4.311 (0.469)	4.138 (0.424)	4.135 (0.415)	4.058 (0.526)
Ex 4	9.567 (0.996)	9.539 (0.991)	9.543 (0.987)	9.523 (1.054)
Ex 5	11.911 (1.401)	11.718 (1.691)	11.408 (1.332)	11.448 (1.276)
Ex 6	117.439 (10.932)	125.299 (17.343)	121.716 (13.165)	132.356 (16.084)
Method	HL(F)	HL(A)	HL(MF)	HL(MA)
Ex 1	9.290 (1.063)	9.309 (1.000)	9.293 (1.032)	9.293 (1.033)
Ex 2	10.288 (0.865)	9.425 (0.717)	10.417 (1.086)	10.389 (0.923)
Ex 3	4.093 (0.500)	4.114 (0.451)	4.101 (0.484)	4.101 (0.484)
Ex 4	9.502 (1.056)	9.502 (1.056)	9.499 (1.196)	9.497 (1.106)
Ex 5	11.495 (1.510)	11.633 (1.471)	11.454 (1.559)	11.429 (1.478)
Ex 6	138.341 (19.714)	117.439 (11.271)	142.019 (17.024)	142.986 (16.094)

Before closing this section, we consider a simple adaptive selection procedure that chooses the penalty among Ridge, Lasso and Adaptive Lasso which has the smallest PMSE value. In fact, this simple way shares the motivation of adaptive penalty with the proposed adaptive HL method even though the proposed method can provide various penalty forms under the framework of the random-effect model. To compare this simple adaptive selection procedure with the proposed adaptive HL method, we compute the proportion of selection the true model and MSE and MAE values with the above six examples in Tables 4 and 5. We note that the results of HL(A) are taken from Tables 1 and 2 for easy comparison. As one can see, the patterns of the performance of both methods are similar. Overall, the proposed adaptive HL method outperforms the simple adaptive procedure.

4. Oracle property in variable selection

Suppose that the true model is a fixed regression model of (1) with $\beta = (\beta_1, \dots, \beta_d)^T = (\beta_1^T, \beta_2^T)$, where $\beta_1 = (\beta_1, \dots, \beta_s)^T$, $\beta_2 = (\beta_{s+1}, \dots, \beta_d)^T$, and $\beta_2 = \mathbf{0}$. To study the asymptotic behavior of h -likelihood based estimators, we

Table 4

The proportion of selecting the true model with the median number of incorrect zeros in parentheses. Note that SA denotes the simple adaptive selection procedure.

Method	Ex1	Ex 2	Ex 3	Ex 4	Ex 5	Ex 6
SA	0.23 (0.0)	0.98 (0.0)	0.50 (0.0)	0.56 (0.0)	0.10 (0.0)	0.00 (3.0)
HL(A)	0.38 (0.0)	0.93 (0.0)	0.55 (0.0)	0.72 (0.0)	0.25 (0.0)	0.00 (0.0)

Table 5

The medians of MSE (first line) and MAE (second line) values with the median absolute deviation in parentheses.

Method	Ex1	Ex 2	Ex 3	Ex 4	Ex 5	Ex 6
SA	0.222 (0.180)	0.134 (0.069)	0.025 (0.032)	0.176 (0.198)	0.137 (0.037)	0.940 (0.207)
	0.367 (0.198)	0.292 (0.089)	0.065 (0.063)	0.348 (0.261)	0.295 (0.047)	0.785 (0.119)
HL(A)	0.198 (0.216)	0.136 (0.074)	0.022 (0.032)	0.139 (0.150)	0.111 (0.028)	0.708 (0.151)
	0.276 (0.202)	0.297 (0.090)	0.054 (0.062)	0.281 (0.198)	0.198 (0.036)	0.700 (0.086)

consider the following h -likelihood function

$$h(\boldsymbol{\beta}) = h_1(\boldsymbol{\beta}) - n \sum_{j=1}^d p_{\lambda_n}(|\beta_j|), \quad (4)$$

where $p_{\lambda_n}(|\beta_j|) = p_{\lambda}(|\beta_j|)/n$, $\lambda_n = \lambda/n$ and $\lambda = \phi/\theta$. We note that the above $h(\boldsymbol{\beta})$ is corresponding to the penalized likelihood function $Q(\boldsymbol{\beta})$ in Fan and Li [6], where required conditions that $\lambda_n \rightarrow 0$ and $\sqrt{n}\lambda_n \rightarrow \infty$ to show the oracle property for SCAD. In the PL approach, λ_n is not model parameters, so that arbitrary constraints can be made on λ_n without encountering a severe objection. In random-effect models, $\lambda = \phi/\theta$ is the variance-component ratio, and hence, it is natural to assume $\lambda_n = \lambda/n \rightarrow 0$, but the second condition may be too stringent to be satisfied. With the unbounded penalty, we can only require the first condition for the oracle estimation. This means that the oracle property is an asymptotic property of random-effect estimator. Furthermore, the simulation studies in the previous section show that the unboundedness of penalties is very helpful in identifying null predictors in finite samples, enhancing the estimation of the fitted model.

Let $\hat{\boldsymbol{\beta}}$ be an h -likelihood estimator for random-effect model. Let $\nabla m(\hat{\boldsymbol{\beta}}) = \partial m(\boldsymbol{\beta})/\partial \boldsymbol{\beta}|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}$ and $\nabla^2 m(\hat{\boldsymbol{\beta}}) = \partial^2 m(\boldsymbol{\beta})/\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}}$. For the fixed regression model, h_1 of (6) in the Appendix is the log-likelihood such that

$$I(\boldsymbol{\beta}) = -\nabla^2 h_1(\boldsymbol{\beta})$$

is the Fisher information matrix. Here $I(\boldsymbol{\beta}_1, \mathbf{0}) = -\nabla^2 h_1(\boldsymbol{\beta}_1, \mathbf{0})$ is the Fisher information matrix knowing that $\boldsymbol{\beta}_2 = \mathbf{0}$ and $I_1(\boldsymbol{\beta}_1) = -\nabla^2 h_1(\boldsymbol{\beta}_1)$ is the Fisher information matrix without $\boldsymbol{\beta}_2$. Under the random effect model,

$$H(\boldsymbol{\beta}) = -\nabla^2 h(\boldsymbol{\beta}) = I(\boldsymbol{\beta}) - \nabla^2 h_2(\boldsymbol{\beta})$$

is the information matrix from the h -likelihood. Here $H(\boldsymbol{\beta}_1, \mathbf{0}) = -\nabla^2 h(\boldsymbol{\beta}_1, \mathbf{0})$ is the Fisher information matrix knowing that $\boldsymbol{\beta}_2 = \mathbf{0}$; and $H_1(\boldsymbol{\beta}_1) = -\nabla^2 h(\boldsymbol{\beta}_1)$ is the Fisher information matrix without $\boldsymbol{\beta}_2$.

Theorem 1. Under the regularity conditions (A)–(C) of Fan and Li [6], for a given λ , as $n \rightarrow \infty$, we obtain the following results:

- (1) The h -likelihood estimator $\hat{\boldsymbol{\beta}}$ is root- n consistent, and
- (2) If $p'_{\lambda}(0) = \infty$, $\hat{\boldsymbol{\beta}}$ satisfies the oracle property, namely, (a) Sparsity: $\hat{\boldsymbol{\beta}}_2 = \mathbf{0}$ and (b) Asymptotic normality:

$$\sqrt{n}\{H_1(\boldsymbol{\beta}_1)\}[\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1 - \nabla^2 h(\boldsymbol{\beta}_1)^{-1} \nabla h_2(\boldsymbol{\beta}_1)] \rightarrow N(\mathbf{0}, I_1(\boldsymbol{\beta}_1))$$

in distribution.

Proof. Since $p_{\lambda_n}(|\beta_j|) = p_{\lambda}(|\beta_j|)/n$, it follows that $\max\{p'_{\lambda_n}(|\beta_j|) : \beta_j \neq 0 \text{ for } j = 1, \dots, s\} \rightarrow 0$ when $\lambda = O(1)$. Thus, by Theorem 1 of Fan and Li [6], it can be shown that there exists a h -likelihood estimator that converges at the rate $O_p(n^{-1/2} + a_n)$ with $a_n = \max\{p'_{\lambda}(|\beta_j|)/n : \beta_j \neq 0\} = O(n^{-1})$. This completes the proof of the result (1).

For the proof of the result (2), we first show that the h -likelihood approach satisfies Lemma 1 of Fan and Li [6]. To that end, it is sufficient to show that the sign of $\partial h(\boldsymbol{\beta})/\partial \beta_j$ ($j = s+1, \dots, d$) is completely determined by that of β_j . Since $p'_{\lambda}(0) = \infty$, it follows that

$$\lim_{n \rightarrow \infty} \inf \lim_{\beta_j \downarrow 0} p'_{\lambda_n}(|\beta_j|) = \infty.$$

Thus, we have, for $j = s + 1, \dots, d$,

$$\frac{\partial h(\boldsymbol{\beta})}{\partial \beta_j} = n\{-p'_{\lambda_n}(|\beta_j|)\text{sgn}(\beta_j) + O_p(n^{-1/2})\}.$$

Hence, the sign of the score in the region closely above the origin $\beta_j = 0$, for example, $|\beta_j| < 1/n$, is totally determined by β_j . Therefore, by using this result and the proof of Theorem 2 in Fan and Li [6], we complete the proof of the oracle property of the h -likelihood estimators. \square

Remark. Suppose that the penalty is bounded, i.e., $p'_{\lambda_n}(|\beta_j|) < \infty$. Because

$$\frac{\partial h(\boldsymbol{\beta})}{\partial \beta_j} = n\lambda_n\{-\lambda_n^{-1}p'_{\lambda_n}(|\beta_j|)\text{sgn}(\beta_j) + O_p(n^{-1/2}/\lambda_n)\},$$

the condition $\sqrt{n}\lambda_n \rightarrow \infty$ is required to show the oracle property.

5. Conclusion

In this paper, we introduce a new random-effect model, which leads to an unbounded penalty. The new penalty has merit in variable selection, enhancing estimation, and prediction in sparse cases. The proposed method has been applied to various cases including $d > n$ such as principal component analysis, canonical covariance analysis, and partial least squares and change point problems, and provided outstanding improvements over existing methods [9–11].

Acknowledgments

This work was prepared while Youngjo Lee (YL) was visiting the Department of Statistics of Stanford University. YL would like to thank Professors Mike Kenward, Tze Lai, Trevor Hastie, Ji Zhu, Johan Lim, and Meeyoung Park for their helpful comments. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2011-0030810 and 2011-0030811).

Appendix. A new distribution for random effects

The model (3) can be written as

$$\boldsymbol{\beta} = \sqrt{\tau} \mathbf{e}, \quad (5)$$

where $\tau = u\theta$ and $\mathbf{e} \sim N(0, 1)$. With the log link, we have an additive model

$$\log(\tau) = \log \theta + v,$$

where $v = \log u$. This leads to the h -likelihood

$$h = h_1 + h_2, \quad (6)$$

where

$$h_1 = \sum_{i=1}^n \log f_{\phi}(y_i | \boldsymbol{\beta}),$$

$$h_2 = \sum_{j=1}^d \{\log f_{\theta}(\beta_j | u_j) + \log f_w(v_j)\},$$

$$\log f_{\phi}(y_i | \boldsymbol{\beta}) = -\frac{1}{2} \log(2\pi\phi) - \frac{1}{2\phi} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^T (y_i - \mathbf{x}_i^T \boldsymbol{\beta}),$$

$$\log f_{\theta}(\beta_j | u_j) = -\frac{1}{2} \{\log(2\pi\theta) + \log u_j + \beta_j^2 / (\theta u_j)\}$$

$$\log f_w(v_j) = -\log(w)/w - \log \Gamma(1/w) + v_j/w - \exp(v_j)/w, \quad \text{and}$$

$f_{\theta}(\beta_j | u_j)$ and $f_w(v_j)$ are the density functions of $\beta_j | u_j$ and v_j , respectively. Given (w, ϕ, θ) , for the estimation of $\boldsymbol{\beta}$, we may use the profile h -likelihood

$$h_p = h_1|_{u=\hat{u}} + h_2|_{u=\hat{u}},$$

where \hat{u} solves $dh/du = 0$. Then we obtain a random-effect estimator

$$\hat{u} = \hat{u}(\boldsymbol{\beta}) = w \{(2/w - 1) + \kappa\} / 4 \quad (7)$$

with $\kappa = \sqrt{8\beta^2/(w\theta) + (2/w - 1)^2}$. In this random-effect model approach, the penalty function $p_\lambda(|\beta_j|)$ stems from a probabilistic model

$$p_\lambda(|\beta_j|) = -\phi\{\log f_\theta(\beta_j|u_j) + \log f_w(v_j)\}_{u_j=\hat{u}_j}. \quad (8)$$

With $w = 0$ and $w = 2$, the maximization of the profile h -likelihood provides the ridge and LASSO estimator, respectively. Specifically, at $w = 0$, the distribution of u becomes degenerated to 1 such that

$$\hat{u}_j = \hat{u}_j(\beta) = \left\{2 - w + \sqrt{2w\beta_j^2/\theta + (2 - w)^2}\right\}/4 = 1$$

to give $\hat{v}_j = 0$. Note that, by Stirling's approximation, it follows that

$$\log f_{w=0}(\hat{v}_j = 0) = \lim_{w \rightarrow 0} \{-\log(w)/w - \log \Gamma(1/w) - 1/w\} = \log(2\pi)/2.$$

Thus, the resulting penalty becomes the L_2 -penalty as

$$p_\lambda(|\beta_j|) = -\phi h_2|_{u=1} = \phi/(2\theta) \sum_{j=1}^d \beta_j^2 + \text{constant}.$$

Hence, with $w = 2$, $\hat{u}_j = \hat{u}(\beta) = |\beta_j|/\sqrt{\theta}$,

$$p_\lambda(|\beta_j|) = -\phi h_2|_{u=\hat{u}} = \left(\phi/\sqrt{\theta}\right) \sum_{j=1}^d |\beta_j| + \text{constant}$$

becomes the L_1 -penalty.

Fitting algorithm

To gain more insight into the content, consider the componentwise least squares problem,

$$\frac{1}{2}(z - \beta)^2 + p_\lambda(|\beta|).$$

From Lee and Nelder [12], the mode of β from h can be obtained by IWLS,

$$\hat{\beta} = (\mathbf{1}_2^T \Sigma^{-1} \mathbf{1}_2)^{-1} \mathbf{1}_2^T \Sigma^{-1} \mathbf{z}_a = z/[1 + \lambda/\hat{u}], \quad (9)$$

where \hat{u} is defined in (7), $\Sigma = \text{diag}(\phi, \theta u)$ and $\lambda = \phi/\theta$. This is the least squares estimator for the augmented linear model

$$\mathbf{z}_a = \mathbf{1}_2 \beta + \mathbf{r},$$

where $\mathbf{z}_a = (z, 0)^T$, $\mathbf{r} = (r_1, r_2)^T$, $r_1 = \sqrt{\phi}e_1$, $r_2 = \sqrt{\theta}ue_2$, and e_i are i.i.d. random variables from $N(0, 1)$.

Note that $\theta u_i = (a\theta)(u_i/a)$ for all $a > 0$. Therefore, θ and u_i are not separately identifiable. Thus, in random-effect models, we take a parameterization that $E(u_i) = 1$ for all w . This imposes a constraint on random-effect estimates such that $\sum_{j=1}^d \hat{u}_j/d = 1$. The difference between parameterizations in random-effect models and the PL depends on w . For comparison purposes, with the PL approach, we follow the convention that $\lambda = \phi/\theta$. Under the parameterization in the PL approach, it may not hold that $\sum_{j=1}^d \hat{u}_j/d = 1$. Regardless of the parameterization, the proposed algorithm has a common fit for θu_i . From (4) and (6), we obtain

$$0 = \partial(-\phi h/\partial \beta) = \beta[1 + \lambda/\hat{u}] - z = \partial Q_\lambda/\partial \beta = \text{sign}(\beta)\{|\beta| + p'_\lambda(|\beta|)\} - z.$$

Thus,

$$\hat{u} = \lambda|\beta|/p'_\lambda(|\beta|) \quad (10)$$

and

$$\hat{\beta} = z/(1 + p'_\lambda(|\beta|)/|\beta|).$$

From the above algorithm, we could obtain LASSO, SCAD, and adaptive LASSO including the proposed method. In other words, estimates of most existing variable selection methods can be obtained by using different random-effect estimates \hat{u} in the IWLS of (9). More specifically, the choice of $\hat{u} = |\beta|$ provides the LASSO solution for the L_1 -penalty $p_\lambda(|\beta|) = \lambda|\beta|$. The adaptive LASSO solution of the penalty $p_\lambda(|\beta|) = 2\lambda|\beta|/|z|$ in Zou [16] can be obtained by $\hat{u} = |\beta||z|/2$, where $z = \hat{\beta}_{OLS}$. The solution of SCAD is also implemented by setting the random-effect estimate as

$$\hat{u} = |\beta|/\left\{I(|\beta| \leq \lambda) + \frac{(a\lambda - |\beta|)_+}{(a-1)\lambda}I(|\beta| > \lambda)\right\},$$

for some $a > 2$. We set $w = 30$ for the HL method after a careful investigation of performances at various values of w .

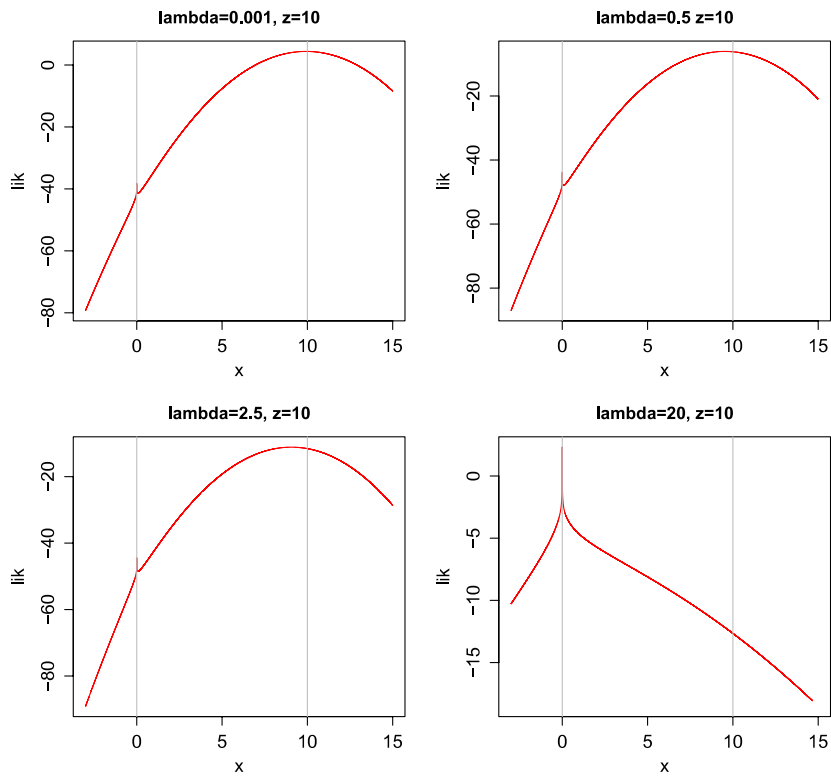


Fig. 3. Likelihood functions of different values of λ with fixed $z = 10$.

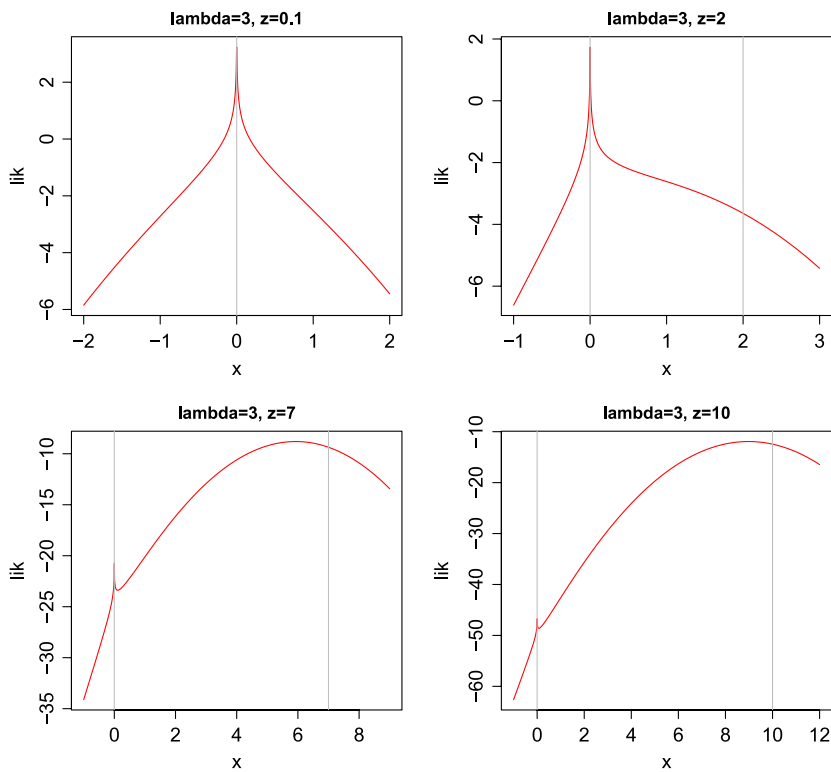


Fig. 4. Likelihood functions of different values z with fixed $\lambda = 3$.

When X is not orthogonal, we can use an IWLS from Lee and Nelder [12]

$$(X^T X + W_\lambda) \hat{\beta} = X^T Y,$$

where $W_\lambda = \text{diag}(\lambda/\hat{u}_i)$ and $\hat{u}_k = \lambda|\beta_k|/p'_\lambda(|\beta_k|)$. This is identical to the IWLS of Fan and Li [6] based on local quadratic approximation.

Figs. 3 and 4 show the likelihood surface of h at various combinations of (z, λ) . Given λ , as z (OLS of β) approaches zero or given z , as λ becomes large, there is only one maxima at zero. Thus, the corresponding predictor is not selected in the model. Otherwise, bimodality occurs. In this case, from the figures, it can be seen that most often, the likelihood surface appears to support the non-null maximum value (selecting the corresponding predictor as necessary) because a perturbation caused by the singularity at the origin is negligible. Thus, we found that this singularity at the origin does not pose a numerical difficulty in finding nonzero local maxima.

When $\hat{u} = 0$, it follows that $\hat{\beta} = 0$. Thus, we can allow thresholding, by simply taking the null random-effect estimator $\hat{u} = 0$. However, when $\hat{u} = 0$, since the corresponding diagonal element $1/\hat{u}$ is undefined, W_λ is not defined. Therefore, we should delete the corresponding predictors. This causes an algorithmic difficulty, and hence, we employ a perturbed random-effect estimate $\hat{u}_{\delta,k} = \lambda(|\beta_k| + \delta)/p'_\lambda(|\beta_k|)$ for a small positive $\delta = 10^{-8}$. Then, $W_{\lambda,\delta} = \text{diag}(\lambda/\hat{u}_{\delta,i})$ is always defined. As long as δ is small, the diagonal elements of $W_{\lambda,\delta}$ are close to those of W_λ . Therefore, the solutions of IWLS, $(X^T X + W_{\lambda,\delta}) \hat{\beta} = X^T Y$, are nearly identical to those of the original IWLS. Note that this algorithm is identical to that of Hunter and Li [8] for the improvement of local quadratic approximation. In this paper, we report $\hat{\beta} = 0$ if all eight printed decimals are zero.

Note here that the proposed penalty $p_\lambda(|\beta_j|)$ is nonconvex. However, the model for $p_\lambda(|\beta_j|)$ can be denoted as hierarchically as (i) $\beta_j|u_j$ is normal and (ii) u_j is gamma; both models can be fitted by convex GLM optimizations. Thus, the proposed IWLS algorithm overcomes the difficulties of a nonconvex optimization by solving two-interlinked convex optimizations [12].

References

- [1] L. Breiman, Heuristics of instability and stabilization in model selection, *Ann. Statist.* 24 (1996) 2350–2383.
- [2] D.L. Donoho, I.M. Johnstone, Ideal spatial adaptation by wavelet shrinkage, *Biometrika* 81 (1994) 425–455.
- [3] A.W.F. Edwards, *Likelihood*, Cambridge University Press, London, 1972.
- [4] B. Efron, C. Morris, Data analysis using Stein's estimator and its generalizations, *J. Amer. Statist. Assoc.* 70 (1975) 311–319.
- [5] J. Fan, Comments on "Wavelets in statistics: a review" by A. Antoniadis, *J. Ital. Statist. Ass.* 6 (1997) 131–138.
- [6] J. Fan, R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, *J. Amer. Statist. Assoc.* 96 (2001) 1348–1360.
- [7] A.E. Hoerl, R.W. Kennard, Ridge regression: biased estimation for nonorthogonal problems, *Technometrics* 12 (1970) 55–67.
- [8] D. Hunter, R. Li, Variable selection using MM algorithms, *Ann. Statist.* 33 (2005) 1617–1642.
- [9] D. Lee, W. Lee, Y. Lee, Y. Pawitan, Super-sparse principal component analyses for high-throughput genomic data, *BMC Bioinformatics* 11 (2010) 296.
- [10] D. Lee, W. Lee, Y. Lee, Y. Pawitan, Sparse partial least-squares regression and its applications to high-throughput data analysis, *Chemometr. Intell. Lab. Syst.* 109 (2011) 1–8.
- [11] W. Lee, D. Lee, Y. Lee, Y. Pawitan, Sparse canonical covariance analysis for high-throughput data, *Stat. Appl. Genet. Mol. Biol.* 10 (1) (2011) Article 30.
- [12] Y. Lee, J.A. Nelder, Double hierarchical generalized linear models (with discussion), *Appl. Stat.* 55 (2006) 139–185.
- [13] P. Radchenko, G. James, Variable inclusion and shrinkage algorithms, *J. Amer. Statist. Assoc.* 103 (2008) 1304–1315.
- [14] R.J. Tibshirani, Regression shrinkage and selection via the LASSO, *J. R. Stat. Soc. Ser. B Stat. Methodol.* 58 (1996) 267–288.
- [15] G. Wahba, *Spline Models for Observational Data*, SIAM, Philadelphia, 1990.
- [16] H. Zou, The adaptive lasso and its oracle properties, *J. Amer. Statist. Assoc.* 101 (2006) 1418–1429.
- [17] H. Zou, T. Hastie, Regularization and variable selection via the elastic net, *J. R. Stat. Soc. Ser. B Stat. Methodol.* 67 (2005) 301–320.