 DOI: 10.1111/biom.12230

Group Variable Selection Via Convex Log-Exp-Sum Penalty with Application to a Breast Cancer Survivor Study

Zhigeng Geng,[1] Sijian Wang,[1,2,*] Menggang Yu,[2] Patrick O. Monahan,[4] Victoria Champion,[5] and Grace Wahba[1,2,3]

[1]Department of Statistics, University of Wisconsin, Madison, Wisconsin 53706, U.S.A.
[2]Department of Biostatistics & Medical Informatics, University of Wisconsin, Madison, Wisconsin 53706, U.S.A.
[3]Department of Computer Sciences, University of Wisconsin, Madison, Wisconsin 53706, U.S.A.
[4]School of Medicine, Indiana University, Indianapolis, Indiana 46202, U.S.A.
[5]School of Nursing, Indiana University, Indianapolis, Indiana 46202, U.S.A.
[*]*email:* swang@biostat.wisc.edu

Summary. In many scientific and engineering applications, covariates are naturally grouped. When the group structures are available among covariates, people are usually interested in identifying both important groups and important variables within the selected groups. Among existing successful group variable selection methods, some methods fail to conduct the within group selection. Some methods are able to conduct both group and within group selection, but the corresponding objective functions are non-convex. Such a non-convexity may require extra numerical effort. In this article, we propose a novel Log-Exp-Sum(LES) penalty for group variable selection. The LES penalty is strictly convex. It can identify important groups as well as select important variables within the group. We develop an efficient group-level coordinate descent algorithm to fit the model. We also derive non-asymptotic error bounds and asymptotic group selection consistency for our method in the high-dimensional setting where the number of covariates can be much larger than the sample size. Numerical results demonstrate the good performance of our method in both variable selection and prediction. We applied the proposed method to an American Cancer Society breast cancer survivor dataset. The findings are clinically meaningful and may help design intervention programs to improve the qualify of life for breast cancer survivors.

Key words: Breast cancer survivor; Finite sample bound; Group variable selection; High-dimensional data; Penalized estimation; Sparsity recovery.

1. Introduction

Breast cancer is the most common cancer in women younger than 45 years of age and is the leading cause of death among females in the United States. However, the survival rate for these young women with breast cancer has continuously improved over the past two decades, primarily because of improved therapies. With this long-term survival, it is important to study the quality of life that may be hampered by this traumatic event and by the long-term side effects from related cancer therapies.

This article is motivated by analyzing a dataset from a study funded by the American Cancer Society (ACS), a large quality of life study of breast cancer survivors diagnosed at a young age. The study included 505 breast cancer survivors (BCS) who were aged 18–45 years old at diagnosis and were surveyed 3–8 years after standard treatments. The study collected many covariates and quality of life outcomes. One outcome that is of particular interest is overall well being (OWB). It is captured by Campbell's index of well being which is measured from seven questionnaire items (Campbell, Converse, and Rodgers, 1976). Studying the OWB status after an adversity is of great interest in an increasing body of research to comprehensively understand the consequences of a traumatic event, for example, cancer at a young age.

In the present analysis, the covariates include demographic variables and social or behavior construct scores. The constructs are divided into eight non-overlapping groups: personality, physical health, psychological health, spiritual health, active coping, passive coping, social support, and self-efficacy. The constructs in each group are designed to measure the same aspect of the social or behavioral status of a breast cancer survivor from different angles. In our analysis, we are interested in identifying both important groups and important individual constructs within the selected groups that are related to OWB. These discoveries may help design interventions targeted at these young breast cancer survivors from the perspective of a cancer control program. In statistics, this is a group variable selection problem.

Variable selection via penalized likelihood estimation has been an active research area in the past decade. When there is no group structure, many methods have been proposed and their properties have been thoroughly studied, for example, see LASSO (Tibshirani, 1996), SCAD (Fan and Li, 2001), Elastic-Net (Zou and Hastie, 2005), SICA (Lv and Fan, 2009), MCP (Zhang, 2010), truncated L_1 (Shen, Pan, and Zhu, in press), SELO (Dicker, Huang, and Lin, in press) and references therein. However, when there are grouping structures among covariates, these methods still make selection

based on the strength of individual covariates rather than the strength of the group, and may have inadequate performance. A proper integration of the grouping information into the analysis is hence desired, and that may help boost the signal-to-noise ratio.

Several methods have addressed the group variable selection problem in literature. Yuan and Lin (2006) proposed a group LASSO penalty; Zhao, Rocha, and Yu (2006) proposed a CAP family of group variable selection penalties. These two methods can effectively remove unimportant groups, but a possible limitation is that they select variables in an "all-in-all-out" fashion, that is, when one variable in a group is selected, all other variables in the same group are also selected. In our analysis of the ACS dataset, however, we want to keep the flexibility of selecting variables within a group. For example, when a group of constructs is related to OWB, it does not necessarily mean all the individual constructs in this group are related to OWB. We may want to not only remove unimportant groups effectively, but also identify important individual constructs within important groups as well. To achieve the goal, Huang et al. (2009) and Zhou and Zhu (2010) independently proposed a group bridge penalty and a hierarchical LASSO penalty, respectively. These two penalties can do the selection at both group level and within group level. However, one possible drawback of the two methods is that their penalty functions are no longer convex. This non-convexity may cause numerical problems in practical computation, especially when the numbers of groups and covariates are large.

In this article, we propose a new Log-Exp-Sum penalty for group variable selection. This new penalty is convex, and it can perform variable selection at both group level and within-group level. We propose an effective algorithm to fit the model. The theoretical properties of our proposed method are thoroughly studied. We establish both the finite sample error bounds and asymptotic group selection consistency of our LES estimator. The proposed method is applied to the ACS breast cancer survivor dataset.

2. Method

2.1. *Preparation*

We consider the usual regression setup: we have training data, $(\boldsymbol{x}_i, y_i)$, $i = 1, \ldots, n$, where $\boldsymbol{x}_i$ and y_i are a p-length vector of covariates and response for the ith subject, respectively. We assume the total of p covariates can be divided into K groups. Let the kth group have p_k variables, and we use $\boldsymbol{x}_{i,(k)} = (x_{i,k1}, \ldots, x_{i,kp_k})^{\mathrm{T}}$ to denote the p_k covariates in the kth group for the ith subject. In most of the article, we assume $\sum_k p_k = p$, that is, there are no overlap between groups. This is also the situation in ACS breast cancer survivor data. We will discuss the situation that groups are overlapped in Section 6.

To model the association between response and covariates, we consider linear regression:

$$y_i = \sum_{k=1}^{K} \sum_{j=1}^{p_k} x_{i,kj}\beta_{kj} + \epsilon_i, \ \ i = 1, \ldots, n, \tag{1}$$

where $\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$ are error terms and β_{kj}'s are regression coefficients. We denote $\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{kp_k})'$ to be the vector of regression coefficients for covariates in the kth group. Without loss of generality, we assume the response is centered to have zero mean and each covariate is standardized to have zero mean and unit standard deviation, so the intercept term can be removed from the above regression model.

For the purpose of variable selection, we consider the penalized ordinary least square (OLS) estimation:

$$\min_{\beta_{kj}} \frac{1}{2n} \sum_{i=1}^{n} \Big(y_i - \sum_{k=1}^{K} \sum_{j=1}^{p_k} x_{i,kj}\beta_{kj} \Big)^2 + \lambda J(\boldsymbol{\beta}), \tag{2}$$

where $J(\boldsymbol{\beta})$ is a sparsity-induced penalty function and λ is a non-negative tuning parameter.

Yuan and Lin (2006) proposed the following group LASSO penalty which is to penalize the L_2-norm of the coefficients within each group:

$$J(\boldsymbol{\beta}) = \sum_{k=1}^{K} \sqrt{\beta_{k1}^2 + \cdots + \beta_{kp_k}^2}. \tag{3}$$

Zhao et al. (2006) proposed penalizing the L_∞-norm of $\boldsymbol{\beta}_k$:

$$J(\boldsymbol{\beta}) = \sum_{k=1}^{K} \max\{\beta_{k1}, \ldots, \beta_{kp_k}\}. \tag{4}$$

We can see that both L_2-norm and L_∞-norm are singular when the whole vector $\boldsymbol{\beta}_k$ is zero. Therefore, some estimated coefficient vector $\hat{\boldsymbol{\beta}}_k$ will be exactly zero and hence the corresponding kth group will be removed from the fitted model. Once a component of $\boldsymbol{\beta}_k$ is non-zero, however, the two norm functions are no longer singular and hence cannot conduct the within group variable selection.

Huang et al. (2009) proposed the following group bridge penalty:

$$J(\boldsymbol{\beta}) = \sum_{k=1}^{K} \Big(|\beta_{k1}| + \cdots + |\beta_{kp_k}| \Big)^{\gamma}, \tag{5}$$

where $0 < \gamma < 1$ is another tuning parameter.

Zhou and Zhu (2010) independently proposed a hierarchical LASSO penalty. This penalty decomposes $\beta_{kj} = \gamma_k\theta_{kj}$ and considers

$$J(\gamma_k, \theta_{kj}) = \sum_{k=1}^{K} |\gamma_k| + \lambda \sum_{k=1}^{K} \sum_{j=1}^{p_k} |\theta_{kj}|. \tag{6}$$

When the groups are not overlapped, the hierarchical LASSO penalty is equivalent to the group bridge penalty with $\gamma = 0.5$. We can see that these two penalties are singular at both $\boldsymbol{\beta}_k = \mathbf{0}$ and $\beta_{kj} = 0$ and hence is able to conduct both group selection and within group selection, however, the two objective functions are not convex.

Simon et al. (2012) proposed the sparse group LASSO penalty:

$$J(\boldsymbol{\beta}) = s\sum_{k=1}^{K}\sqrt{\beta_{k1}^2+\cdots+\beta_{kp_k}^2} + (1-s)\sum_{k=1}^{K}\sum_{j=1}^{p_k}|\beta_{kj}|, \quad (7)$$

where $0 < s < 1$ is another tuning parameter. We can see that, by mixing the LASSO penalty and group LASSO penalty, the sparse group LASSO penalty is convex and is able to conduct both group and within group selection.

2.2. *Log-Exp-Sum Penalty*

Our Log-Exp-Sum (LES) penalty is motivated by modifying the group LASSO penalty to conduct both group and within-group selection. Note that the group LASSO penalty is a member of a penalty function family: $J(\boldsymbol{\beta}) = \sum_{k=1}^{K} f^{-1}\Big\{f(|\beta_{k1}|)+\cdots+f(|\beta_{kp_k}|)\Big\}$ by taking $f(x) = x^2$. Our LES penalty is another member of this family by taking $f(x) = \exp(x)$. To be specific, we propose the following LES penalty:

$$J(\boldsymbol{\beta}) = \sum_{k=1}^{K} w_k \log\Big\{\exp(\alpha|\beta_{k1}|)+\cdots+\exp(\alpha|\beta_{kp_k}|)\Big\}, \quad (8)$$

where $\alpha > 0$ is a tuning parameters and w_k's are pre-specified weights to adjust for different group sizes, for example, taking $w_k = p_k/p$. The LES penalty is strictly convex, which can be straightforwardly verified by calculating its second derivative. Similar to other group variable selection penalties, the LES penalty utilizes the group structure and is able to perform group selection. Meanwhile, the LES penalty is also singular at any $\beta_{kj} = 0$ point, and hence is able to conduct the within group selection as well.

There is a connection between the LES penalty and the LASSO penalty. For any design matrix $\boldsymbol{X}$ and an arbitrary grouping structure ($p_k \geq 1$), the LASSO penalty can be viewed as a limiting case of the LES penalty. To be specific, we have the following proposition.

PROPOSITION 1. *Given the data, for any positive number γ, consider the LASSO estimator and LES estimator as follows:*

$$\hat{\boldsymbol{\beta}}^{\text{LASSO}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2n}\sum_{i=1}^{n}\Big(y_i - \sum_{k=1}^{K}\sum_{j=1}^{p_k} x_{i,kj}\beta_{kj}\Big)^2 + \gamma\sum_{k=1}^{K}\sum_{j=1}^{p_k}|\beta_{kj}|,$$

$$\hat{\boldsymbol{\beta}}^{\text{LES}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2n}\sum_{i=1}^{n}\Big(y_i - \sum_{k=1}^{K}\sum_{j=1}^{p_k} x_{i,kj}\beta_{kj}\Big)^2 + \lambda\sum_{k=1}^{K}\frac{p_k}{p}\log\Big\{\sum_{l=1}^{p_k}\exp(\alpha|\beta_{kl}|)\Big\}.$$

Then we have:

$$\hat{\boldsymbol{\beta}}^{\text{LES}} - \hat{\boldsymbol{\beta}}^{\text{LASSO}} \to \mathbf{0}, \ \text{as } \alpha \to 0 \text{ and keeping } \lambda\alpha/p = \gamma.$$

The proof of Proposition 1 is given in Web Appendix A.

Our LES penalty has a property that the estimated coefficients of highly correlated variables within the same group are enforced to be similar to each other. As a consequence of this, when applied to the ACS Breast Cancer Survivor dataset, since the construct scores within the same group can be highly correlated, our LES penalty tends to select or remove the highly correlated constructs within a group together. To be specific, we have the following proposition.

PROPOSITION 2. *Let $\hat{\boldsymbol{\beta}}$ be the penalized OLS estimation with the LES penalty. If $\hat{\beta}_{ki}\hat{\beta}_{kj} > 0$, then we have:*

$$|\hat{\beta}_{ki} - \hat{\beta}_{kj}| \leq C\sqrt{2(1-\rho_{ki,kj})},$$

where constant $C = \frac{1}{n\lambda\alpha^2 w_k}\sqrt{||\boldsymbol{y}||_2^2 + 2n\lambda\sum_{l=1}^{K} w_l\log(p_l)}\exp\times\Big\{\frac{1}{2n\lambda w_k}||\boldsymbol{y}||_2^2 + \sum_{l=1}^{K}\frac{w_l}{w_k}\log(p_l)\Big\}$, $\rho_{ki,kj} = X_{ki}^{\mathrm{T}}X_{kj}$ is the sample correlation between X_{ki} and X_{kj}.

The proof of Proposition 2 is given in Web Appendix B.

2.3. *Algorithm*

We need to solve the following optimization problem:

$$\min_{\beta_{kj}} Q(\beta_{kj}) = \frac{1}{2n}\sum_{i=1}^{n}\Big(y_i - \sum_{k=1}^{K}\sum_{j=1}^{p_k} x_{i,kj}\beta_{kj}\Big)^2 + \lambda\sum_{k=1}^{K} w_k\log\Big\{\sum_{l=1}^{p_k}\exp(\alpha|\beta_{kl}|)\Big\}. \quad (9)$$

We propose applying the coordinate descent algorithm (Friedman et al., 2007; Wu and Lange, 2008) at the group level. The key idea is to find the minimizer of the original high-dimensional optimization problem (9) by solving a sequence of low-dimensional optimization problems, each of which only involves the parameters in one group. See Web Appendix C for the details of the algorithm.

Since our objective function is convex and the LES penalty is separable at the group level, by results in Tseng (2001), our algorithm is guaranteed to converge to the global minimizer. Note that, if we apply the coordinate descent algorithm at the individual coefficient level, the algorithm is not guaranteed to converge.

2.4. *Tuning Parameter Selection*

Tuning parameter selection is an important issue in penalized estimation. One often proceeds by finding estimators which correspond to a range of tuning parameter values. The preferred estimator is then identified as the one in which the tuning parameter optimizes some criterion, such as cross validation (CV), generalized cross validation (GCV) (Craven and

Wahba, 1979), AIC (Akaike, 1973), or BIC (Schwarz, 1978). It is known that CV, GCV and AIC-based methods favor the model with good prediction performance, while BIC-based method tends to identify the correct model (Yang, 2005). To implement GCV, AIC and BIC, one needs to estimate the degrees of freedom (df) of an estimated model. For our LES penalty, the estimate of df does not have an analytic form even when the design matrix is orthonormal. Therefore, we propose using the randomized trace method (Girard, 1987, 1989; Hutchinson, 1989) to estimate df numerically. See Web Appendix D for more details and discussions of this method.

3. Theoretical Results

In this section, we present the theoretical properties of our LES estimator. We are interested in the situation when the number of covariates is much larger than the number of observations, that is, $p >> n$. Throughout the whole section, we consider the following LES penalized OLS estimation:

$$\min_{\beta_{kj}} \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \sum_{k=1}^{K}\sum_{j=1}^{p_k} x_{i,kj}\beta_{kj}\right)^2 + \lambda\sum_{k=1}^{K} p_k \log\Big\{\exp(\alpha|\beta_{k1}|) + \cdots + \exp(\alpha|\beta_{kp_k}|)\Big\}. \quad (10)$$

3.1. *Non-Asymptotic Error Bounds*

In this section, we extend the argument in Bickel, Ritov, and Tsybakov (2009) to establish finite-sample bounds for our LES estimator. We make the following Restricted Eigenvalue assumption with group structure (REgroup), which is similar to the Restricted Eigenvalue (RE) assumption in Bickel et al. (2009).

REgroup assumption: Assume group structure is prespecified and p covariates can be divided into K groups with p_k covariates in each group. For a positive integer s and any $\boldsymbol{\Delta} \in \mathbb{R}^p$, the following condition holds:

$$\kappa(s) \triangleq \min_{G\subseteq\{1,\ldots,K\},\,|G|\le s}\ \min_{\forall\boldsymbol{\Delta}\neq 0,\ \sum_{k\notin G}\|\boldsymbol{\Delta}_k\|_1 \le \sum_{k\in G}(1+2p_k)\|\boldsymbol{\Delta}_k\|_1} \times \frac{2\|\boldsymbol{X}\boldsymbol{\Delta}\|_2}{\sqrt{n}\sqrt{\sum_{k\in G} p_k(1+p_k)^2\|\boldsymbol{\Delta}_k\|_2^2}} > 0,$$

where G is a subset of $\{1,\ldots,K\}$, and $|G|$ is the cardinality of set G. $\boldsymbol{\Delta}_k \in \mathbb{R}^{p_k}$ is a subvector of $\boldsymbol{\Delta}$ for the kth group, that is, $\boldsymbol{\Delta}_k = (\Delta_{k1},\ldots,\Delta_{kp_k})^{\mathrm{T}}$. We denote $\|\cdot\|_2$ and $\|\cdot\|_1$ to be Euclidean norm and L_1-norm, respectively.

THEOREM 1. *Consider linear regression model (1). Let $\boldsymbol{\beta}^*$ be the vector of true regression coefficients. Assume the random error terms $\epsilon_1,\ldots,\epsilon_n$ are i.i.d. from the normal distribution with mean zero and variance σ^2. Suppose the diagonal elements of matrix $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}/n$ are equal to 1. Let $G(\boldsymbol{\beta})$ be the set of indices of groups that contain at least one nonzero element for a vector $\boldsymbol{\beta}$, that is, $G(\boldsymbol{\beta}) = \{\, k \mid \exists\, j,\ 1\le j\le p_k,\ s.t: \beta_{kj}\neq 0; 1\le k\le K\}$. Assume the REgroup assumption holds with $\kappa = \kappa(s) > 0$, where $s = |G(\boldsymbol{\beta}^*)|$. Let A be a real number bigger than $2\sqrt{2}$ and $\gamma = A\sigma\sqrt{\frac{\log p}{n}}$. Let two tuning parameters satisfy $\lambda\alpha = \gamma$. Denote $\hat{\boldsymbol{\beta}}$ to be the solution to optimization problem (10). Then with probability at least $1 - p^{1-A^2/8}$, the following inequalities hold:*

$$\frac{1}{n}\|\boldsymbol{X}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_2^2 \le \frac{16A^2\sigma^2 s}{\kappa^2} * \frac{\log p}{n} \quad (11)$$

$$\|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)\|_1 \le \frac{16A\sigma s}{\kappa^2} * \sqrt{\frac{\log p}{n}},$$
$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|_2 \le (2\sqrt{s}+1)\frac{8A\sigma\sqrt{s}}{\kappa^2} * \sqrt{\frac{\log p}{n}}. \quad (12)$$

The proof and some discussions of Theorem 3 are given in Web Appendix E.

3.2. *Group Selection Consistency*

Let $\mathscr{O}$ be the event that there exists a solution $\hat{\boldsymbol{\beta}}$ to optimization problem (10) such that $\|\hat{\boldsymbol{\beta}}_k\|_\infty > 0$ for all $k \in G(\boldsymbol{\beta}^*)$ and $\|\hat{\boldsymbol{\beta}}_k\|_\infty = 0$ for all $k \notin G(\boldsymbol{\beta}^*)$, where $\boldsymbol{\beta}^*$ is the vector of true regression coefficients for model (1) and $G(\boldsymbol{\beta}^*)$ is the set of indices of groups that contain at least one nonzero element for a vector $\boldsymbol{\beta}^*$. We would like to show the group selection consistency as the following:

$$P(\mathscr{O}) \to 1, \qquad n \to \infty. \quad (13)$$

THEOREM 2. *Consider linear regression model (1), under the assumptions (C1)–(C4), the sparsity property (13) holds for our LES estimator.*

The assumptions (C1)–(C4) and the proof follow the spirit in Nardi and Rinaldo (2008). The details are presented in Web Appendix F.

4. Simulation Studies

In this section, we perform simulation studies to evaluate the finite sample performance of the LES method, and compare the results with several existing methods, including LASSO, group LASSO (gLASSO), group bridge (gBrdige) and sparse group LASSO (sgLASSO). We consider four examples. All examples are based on the linear regression model: $y_i = \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta}^* + \epsilon_i,\ i = 1,\ldots,n$, where $\epsilon_i \overset{\text{i.i.d.}}{\sim} N(0,\sigma^2)$. We chose σ to control the signal-to-noise ratio to be 3. The details of the settings are described as follows.

Example 1 ("All-In-All-Out) There are $K = 5$ groups and $p = 25$ variables in total, with 5 variables in each group. We generated $\boldsymbol{x}_i \sim N(\boldsymbol{0}, \boldsymbol{I})$. The true $\boldsymbol{\beta}^*$ was specified as:

$$\boldsymbol{\beta}^* = [\underbrace{2,2,2,-2,-2}_{\text{group1}}, \underbrace{0,0,0,0,0}_{\textit{group2}}, \underbrace{0,0,0,0,0}_{\text{group3}}, \underbrace{0,0,0,0,0}_{\text{group4}}, \underbrace{0,0,0,0,0}_{\text{group5}}]^{\mathrm{T}}$$

Example 2 ("Not-All-In-All-Out) There are $K = 5$ groups and $p = 25$ variables in total, with 5 variables

in each group. We generated $\boldsymbol{x}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ was a block diagonal matrix given by $\text{diag}(P, P, Q, Q, Q)$. Here P, Q were both 5×5 square matrices. $P_{ij} = 1$ if $i = j$; $P_{ij} = 0.7$ if $1 \le i, j \le 3$ or $4 \le i, j \le 5$; $P_{ij} = 0.1$ if otherwise. $Q_{ij} = 1$ if $i = j$; $Q_{ij} = 0.7$ if $i \ne j$. The true $\boldsymbol{\beta}^*$ was specified as:

$$\boldsymbol{\beta}^* = [\underbrace{2, 2, 2, 0, 0}_{\text{group1}}, \underbrace{2, 2, 2, 0, 0}_{\text{group2}}, \underbrace{0, 0, 0, 0, 0}_{\text{group3}}, \underbrace{0, 0, 0, 0, 0}_{\text{group4}}, \underbrace{0, 0, 0, 0, 0}_{\text{group5}}]^{\mathrm{T}}$$

Example 3 (mixture) There are $K = 5$ groups and $p = 25$ variables in total, with 5 variables in each group. We generated $\boldsymbol{x}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ was the same as in simulation setting 2. The true $\boldsymbol{\beta}^*$ was specified as:

$$\boldsymbol{\beta}^* = [\underbrace{0, 0, 0, 2, 2}_{\text{group1}}, \underbrace{0, 0, 0, 2, 2}_{\text{group2}}, \underbrace{1, 1, 1, 1, 1}_{\text{group3}}, \underbrace{1, 1, 1, 1, 1}_{\text{group4}}, \underbrace{0, 0, 0, 0, 0}_{\text{group5}}]^{\mathrm{T}}$$

Example 4 (mixture) There are $K = 6$ groups and $p = 50$ variables in total. For group 1, 2, 4, and 5, each contains 10 variables; for group 3 and 6, each contains 5 variables. We generated $\boldsymbol{x}_i \sim N(\mathbf{0}, \boldsymbol{\Phi})$, where $\boldsymbol{\Phi}$ is a block diagonal matrix given by $\text{diag}(\boldsymbol{\Sigma}, \boldsymbol{\Sigma})$, $\boldsymbol{\Sigma}$ here was the same as in simulation setting 2 and 3. The true $\boldsymbol{\beta}^*$ was specified as:

$$\boldsymbol{\beta}^* = [\underbrace{0, 0, 0, 2, 2, 0, 0, 0, 2, 2}_{\text{group1}}, \underbrace{1, 1, 1, 1, 1, 0, 0, 0, 0, 0}_{\text{group2}}, \underbrace{1, 1, 1, 1, 1}_{\text{group3}}, \underbrace{0, 0, 0, 0, 0, 0, 0, 0, 0, 0}_{\text{group4}}, \underbrace{0, 0, 0, 0, 0, 0, 0, 0, 0, 0}_{\text{group5}}, \underbrace{0, 0, 0, 0, 0}_{\text{group6}}]^{\mathrm{T}}$$

For each setup, the sample size is $n = 100$. We repeated simulations 1000 times. The LES was fitted using the algorithm described in Section 2.3. The LASSO was fitted using the R package "glmnet." The group LASSO was fitted using the R package "grplasso." The group bridge was fitted using the R package "grpreg." The sparse group LASSO was fitted using the R package "SGL."

To select the tuning parameters in each of the five methods, we consider two approaches. The first approach is based on data validation. To be specific, in each simulation, besides the training data, we also independently generated a set of tuning data with the same distribution and with a same sample size as the training data. Then for each tuning parameter, we fitted the model on the training data and used the fitted model to predict the response on the tuning set and calculated the corresponding mean square error (prediction error). The model with the smallest tuning error was selected.

Our second approach for tuning parameter selection is based on BIC, which is defined to be:

$$\text{BIC} = \log(||\boldsymbol{y} - X\hat{\boldsymbol{\beta}}||_2^2/n) + \log n \cdot \text{df}/n,$$

where df is the degrees of freedom of an estimated model. This format of BIC is based on the profile likelihood to get rid of σ^2, the variance of the errors. It is used in Wang, Li, and Tsai (2007) and was shown to have a good performance. For the LES method, the df was estimated using the randomized trace method described in Section 2.4. For the LASSO method, the df was estimated by the number of non-zero estimated coefficients (Zou and Hastie, 2005). For the group LASSO method, the df was estimated as suggested in Yuan and Lin (2006). For the group bridge method, the df was estimated as suggested in Huang et al. (2009). For the sparse group LASSO method, the corresponding articles did not consider estimation of df, and we used the number of non-zero estimated coefficients as the estimator for its df.

To evaluate the variable selection performance of methods, we consider sensitivity (Sens) and specificity (Spec), which are defined as:

$$\text{Sens} = \frac{\#\text{ of selected important variables}}{\#\text{ of important variables}},$$

$$\text{Spec} = \frac{\#\text{ of removed unimportant variables}}{\#\text{ of unimportant variables}}.$$

For both sensitivity and specificity, higher value means a better variable selection performance. Following Associate Editor's suggestion, for each of five methods considered in our simulation, we further obtain the sensitivities and specificities of models along its full solution paths of (by fitting models with many tuning parameter values), create the ROC curve with respect to these sensitivities and specificity, and calculate the corresponding area under curve (AUC). For all five methods, it is possible that several models have the same specificity but different sensitivity. When this happens, we use the highest sensitivity to construct the ROC curve, representing the best variable selection performance of the method.

To evaluate the prediction performance of methods, following Tibshirani (1996), we consider the model error (ME) which is defined as:

$$\text{ME} = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^{\mathrm{T}} \boldsymbol{\Sigma} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*),$$

where $\hat{\boldsymbol{\beta}}$ is the estimated coefficient vector, $\boldsymbol{\beta}^*$ is the true coefficient vector, and $\boldsymbol{\Sigma}$ is the covariance matrix of the design matrix $\boldsymbol{X}$. We would like to acknowledge that the model error is closely related to the predictive mean square error proposed in Wahba (1985) and Leng, Lin, and Wahba (2006). We also calculate the bias of estimator defined as $\text{Bias} = \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*\|^2$.

The simulation results are summarized in Table 1.

In Example 1, the group bridge method has the lowest model error as well as the highest specificity. This is not surprising, because Example 1 is a relatively simple "All-In-All-Out" case, that is, all covariates in a group are either all important or all unimportant. Under this situation, the non-convex group bridge penalty has an advantage over other methods in terms of removing unimportant groups. Although slightly worse than the group bridge method, the LES method outperformed the other three methods in terms of model error. Note that, because of the diagonal covariance matrix of X in this example, the bias is exactly the same as the model error. All five methods had identical AUC values.

Table 1

Simulation results over 1000 replicates. "1-Sens" means one minus the sensitivity of variable selection; "1-Spec" means one minus the specificity of variable selection; "ME" means the model error; "Bias" means the bias of the estimator, which is defined as $||\hat{\beta} - \beta^*||^2$*; . "AUC" means the area under ROC curve of sensitivities and specificities of variable selection across different tuning parameter values. The numbers in parentheses are the corresponding standard errors. The bold numbers are significantly smaller than others (larger than others in column AUC) at a significance level of* 0.05.

	Tuning set tuning				BIC tuning				
Method	1-Sens	1-Spec	ME	Bias	1-Sens	1-Spec	ME	Bias	AUC
Simulation 1									
LASSO	0.000	0.418	1.022	1.022	0.000	0.136	1.318	1.318	1.000
	(0.000)	(0.006)	(0.016)	(0.016)	(0.000)	(0.004)	(0.021)	(0.021)	(0.000)
gLASSO	0.000	0.608	0.717	0.717	0.000	0.061	0.856	0.856	1.000
	(0.000)	(0.009)	(0.012)	(0.012)	(0.000)	(0.004)	(0.015)	(0.015)	(0.000)
gBrdige	0.000	**0.057**	**0.543**	**0.543**	0.000	0.092	**0.641**	**0.641**	1.000
	(0.000)	(0.004)	(0.010)	(0.010)	(0.000)	(0.003)	(0.011)	(0.011)	(0.000)
sgLASSO	0.000	0.612	0.811	0.811	0.000	**0.028**	1.050	1.050	1.000
	(0.000)	(0.009)	(0.013)	(0.013)	(0.000)	(0.003)	(0.018)	(0.018)	(0.000)
LES	0.000	0.537	**0.544**	**0.544**	0.000	0.282	0.770	0.770	1.000
	(0.000)	(0.006)	(0.010)	(0.010)	(0.000)	(0.006)	(0.016)	(0.016)	(0.000)
Simulation 2									
LASSO	0.006	0.379	2.289	3.902	0.013	0.065	2.682	3.746	0.995
	(0.001)	(0.006)	(0.035)	(0.072)	(0.001)	(0.003)	(0.045)	(0.069)	(0.001)
gLASSO	0.000	0.846	3.463	6.117	**0.000**	0.349	4.262	4.845	0.895
	(0.000)	(0.007)	(0.046)	(0.096)	(0.000)	(0.006)	(0.062)	(0.070)	(0.000)
gBrdige	0.004	**0.146**	2.061	3.712	0.005	0.062	**2.321**	3.600	0.997
	(0.001)	(0.004)	(0.032)	(0.072)	(0.001)	(0.002)	(0.040)	(0.072)	(0.001)
sgLASSO	0.000	0.470	2.030	2.333	0.003	0.058	2.608	2.865	1.000
	(0.000)	(0.008)	(0.031)	(0.043)	(0.001)	(0.003)	(0.045)	(0.057)	(0.000)
LES	0.001	0.466	**1.931**	2.344	0.006	0.169	2.629	**2.426**	1.000
	(0.000)	(0.009)	(0.031)	(0.045)	(0.001)	(0.006)	(0.047)	(0.054)	(0.000)
Simulation 3									
LASSO	0.101	0.410	4.158	8.303	0.145	0.146	4.886	8.254	0.914
	(0.002)	(0.007)	(0.046)	(0.094)	(0.003)	(0.005)	(0.066)	(0.096)	(0.002)
gLASSO	**0.000**	0.975	6.018	13.040	**0.000**	0.761	7.011	11.585	0.839
	(0.000)	(0.003)	(0.063)	(0.149)	(0.000)	(0.007)	(0.082)	(0.129)	(0.003)
gBrdige	0.100	**0.399**	4.337	8.407	0.135	**0.089**	5.559	8.347	0.932
	(0.002)	(0.006)	(0.048)	(0.097)	(0.002)	(0.003)	(0.068)	(0.092)	(0.002)
sgLASSO	0.030	0.673	3.563	4.759	0.105	0.191	4.738	7.282	0.994
	(0.001)	(0.007)	(0.044)	(0.062)	(0.002)	(0.005)	(0.060)	(0.095)	(0.000)
LES	0.028	0.642	**3.295**	4.933	0.051	0.365	**4.459**	**5.000**	**0.999**
	(0.002)	(0.008)	(0.041)	(0.067)	(0.003)	(0.009)	(0.063)	(0.076)	(0.000)
Simulation 4									
LASSO	0.127	0.243	5.539	9.081	0.174	0.080	6.830	9.067	0.899
	(0.003)	(0.004)	(0.060)	(0.097)	(0.003)	(0.003)	(0.091)	(0.092)	(0.002)
gLASSO	**0.000**	0.910	9.460	16.780	**0.002**	0.515	11.573	13.795	0.881
	(0.000)	(0.005)	(0.089)	(0.170)	(0.001)	(0.006)	(0.150)	(0.140)	(0.001)
gBrdige	0.118	**0.138**	5.058	8.998	0.147	**0.059**	6.151	8.777	0.913
	(0.003)	(0.003)	(0.059)	(0.112)	(0.003)	(0.002)	(0.074)	(0.109)	(0.002)
sgLASSO	0.023	0.445	4.830	**4.517**	0.107	0.076	6.996	7.028	0.991
	(0.001)	(0.007)	(0.045)	(0.047)	(0.003)	(0.002)	(0.084)	(0.074)	(0.000)
LES	0.028	0.472	**4.638**	5.559	0.034	0.245	6.284	**5.273**	0.992
	(0.002)	(0.008)	(0.054)	0.069	(0.002)	(0.006)	(0.086)	(0.071)	(0.000)

In Example 2, the LES method produced the smallest model error when the tuning set approach was used, and produced the smallest bias when the BIC tuning was used. No method dominated in specificity. All five methods had almost identical sensitivities. Except group LASSO, the other four methods had almost identical AUC values as well.

In Example 3, the LES method produced the smallest model errors no matter which tuning criterion was used. It has the smallest bias when BIC tuning was used as well. The group LASSO method had the highest sensitivity, but its specificity was very low. This means that the group LASSO method tended to include a large amount of variables in the model. The LES method had the highest AUC value among five methods.

Example 4 is similar to Example 3, but has more covariates and more complex group structure. The conclusion about comparisons is similar to that in Example 3. One difference is that, both the LES method and the sparse group LASSO method had the highest AUC values among five methods.

5. American Cancer Society Breast Cancer Survivor Data Analysis

In this section, we analyze the data from ACS breast cancer study which was conducted at the Indiana University School of Nursing. The participants of the study were survivors of the breast cancer aged 18–45 years old at diagnosis and were surveyed between 3 and 8 years from completion of chemotherapy, surgery, with or without radiation therapy. The purpose of the present analysis is to find out what factors in the psychological, social and behavior domains are important for the OWB of these survivors. Identification of these factors and establishment of their association with OWB may help develop intervention programs to improve the quality of life of breast cancer survivors.

The variables included in our current analysis are 54 social and behavior construct scores and three demographic variables. The 54 scores are divided into eight non-overlapping groups: personality, physical health, psychological health, spiritual health, active coping, passive coping, social support and self-efficacy. Each group contains up to 15 different scores. The three demographic variables are: "age at diagnosis" (Agediag), "years of education" (Yrseduc), and "How many months were you in initial treatment for breast cancer" (Bcmths). We treated each demographic variable as an individual group. There are 6 subjects with missing values in either covariates or response, and we removed them from our analysis. In summary, we have 499 subjects and 57 covariates in 11 groups in our analysis.

We applied five methods in the data analysis: LASSO, group LASSO, group bridge sparse group LASSO and our LES method. We randomly split the whole dataset into a training set with sample size $n = 332$ and a test set with sample size $n = 167$ (the ratio of two sample sizes is about 2:1). We fitted models on the training set, using two tuning strategies: one used 10-fold CV, the other used BIC. The BIC tuning procedure for all of the five methods is the same as what we described in the simulation studies. We then evaluated the prediction performances on the test set. We repeated the whole procedure beginning with a new random split 100 times.

The upper part of Table 2 summarizes, over 100 replications, the average number of selected groups, the average number of selected individual variables, and the average mean square errors (MSE) on the test sets, for the five methods. We can see that, for all five methods, the models selected by the 10-fold CV tuning had smaller MSEs (better prediction performance) than the models selected by the BIC tuning. As the cost of this gain in prediction performance, the models selected by 10-fold CV tuning included more groups and more individual variables than the models selected by BIC tuning. We can also see that, our LES methods had the smallest MSE among five methods no matter which tuning strategy was used.

The lower part of Table 2 summarizes the selection frequency of each group across 100 replicates. A group is considered to be selected if at least one variable within the group is selected. Since there are some theoretical works showing that BIC tuning tends to identify the true model (Wang et al., 2007), we focus on the selection results with BIC tuning. We can see that the psychological health group is always selected by all of five methods. For our LES methods, three other groups have very high selection frequency: spiritual health (91 out of 100), active coping (89 out of 100), and self-efficacy (99 out of 100). These three groups are considered to be importantly associated with OWB in literature. Spirituality is a resource regularly used by patients with cancer coping with diagnosis and treatment (Gall et al., 2005). Purnell and Andersen (2009) reported that spiritual well-being was significantly associated with quality of life and traumatic stress after controlling for disease and demographic variables. Self-efficacy is the measure of one's own ability to complete tasks and reach goals, which is considered by psychologists to be important for one to build a happy and productive life (Parle, Maguire, and Heaven, 1997). Rottmann et al. (2010) assessed the effect of self-efficacy and reported a strong positive correlation between self-efficacy and quality of life and between self-efficacy and mood. They also suggested that self-efficacy is a valuable target of rehabilitation programs. Coping refers to "cognitive and behavioral efforts made to master, tolerate, or reduce external and internal demands and conflicts" (Folkman and Lazarus, 1980). The coping strategies are usually categorized into two aspects: active coping and passive coping (Carrico et al., 2006). Active coping efforts are aimed at facing a problem directly and determining possible viable solutions to reduce the effect of a given stressor. Meanwhile, passive coping refers to behaviors that seek to escape the source of distress without confronting it (Folkman and Lazarus, 1985). Setting aside the nature of individual patients or specific external conditions, there have been consistent findings that the use of active coping strategies produce more favorable outcomes compared to passive coping strategies, such as less pain as well as depression, and better quality of life (Holmes and Stevenson, 1990). Another interesting observation is that, compared to other methods, our LES method identified much more frequently the importance of Social Support (including communication with health care team both at diagnosis and at follow up, and support from health care providers). There seems to be more awareness for the importance of this construct both scientifically and publicly. In the New York Times Science Section of 10-Feb-2014, Dr. Arnold S. Relman, a prominent

Table 2

Summary of ACS breast cancer survivor data analysis results. Results are based on 100 *random splits. "Variable selection" reports the average number of selected individual variables; "Group selection" reports the average number of selected groups and "MSE" reports the average mean square errors on test sets. The numbers in parentheses are the corresponding standard errors.*

Selection frequency and mean square error

	10-Fold CV tuning			BIC tuning		
	Variable selection	Group selection	MSE	Variable selection	Group selection	MSE
LASSO	23.18	8.76	2.6288	8.59	3.97	2.7949
	(0.53)	(0.14)	(0.0286)	(0.33)	(0.15)	(0.0309)
gLASSO	51.32	8.80	2.6484	10.46	1.64	2.8620
	(0.42)	(0.12)	(0.0286)	(0.64)	(0.09)	(0.0307)
gBrdige	16.24	3.34	2.6239	11.56	2.94	2.7548
	(0.64)	(0.13)	(0.0293)	(0.26)	(0.07)	(0.0356)
sgLASSO	25.83	8.58	2.6221	5.89	1.59	2.8765
	(0.60)	(0.18)	(0.0265)	(0.31)	(0.11)	(0.0280)
LES	33.50	9.58	**2.6072**	19.86	6.37	**2.7026**
	(0.74)	(0.11)	(0.0283)	(0.91)	(0.20)	(0.0298)

Individual group selection frequency

10-Fold CV tuning

Group name	Agediag	Bcmths	Yrseduc	Personality	Physical health	Psychological health
LASSO	73	55	19	96	75	100
gLASSO	82	66	19	94	82	100
gBrdige	4	2	0	30	2	100
sgLASSO	80	56	18	91	71	100
LES	91	71	21	97	89	100

Group name	Spiritual health	Active coping	Passive coping	Social support	Self-efficacy
LASSO	100	100	68	90	100
gLASSO	98	100	46	93	100
gBrdige	35	70	0	3	88
sgLASSO	97	98	60	89	98
LES	100	100	86	97	100

BIC tuning

Group name	Agediag	Bcmths	Yrseduc	Personality	Physical health	Psychological health
LASSO	5	1	0	32	14	100
gLASSO	2	1	4	4	9	100
gBrdige	1	0	0	6	0	100
sgLASSO	1	0	0	3	4	100
LES	28	14	3	75	48	100

Group name	Spiritual health	Active coping	Passive coping	Social support	Self-efficacy
LASSO	63	69	2	14	97
gLASSO	0	4	1	11	28
gBrdige	29	67	0	0	91
sgLASSO	5	10	1	3	32
LES	91	89	31	59	99

Medical Professor, Writer and Editor, discussed his experience as a hospital patient, where he found out how very important his interactions with nurses were.

In addition, the within group selection results from our LES method provide insights about which aspects/items within selected constructs are most important (The details of within group selection results of five methods are given in Web Appendix G). For example, positive reframing/thinking and religious coping are two most frequently picked items from the Active coping group. Other items such as emotional support, planning, acceptance are not frequently picked. When designing interventions to boost Active coping for patients, focus may be directed towards positive reframing and religious coping.

6. Conclusion and Discussion

In this article, we propose a new convex Log-Exp-Sum penalty for group variable selection. The new method keeps the advantage of group LASSO in terms of effectively removing unimportant groups, and at the same time enjoys the flexibility of removing unimportant variables within identified important groups. We have developed an effective group-level coordinate descent algorithm to fit the model. The theoretical properties of our proposed method have been thoroughly studied. We have established non-asymptotic error bounds and asymptotic group selection consistency for our proposed method, in which the number of variables is allowed to be much larger than the sample size. Numerical results indicate that the proposed method works well in both prediction and variable selection. We also applied our method to the American Cancer Society breast cancer survivor dataset. The analysis results are clinically meaningful and have potential impact on interventions to improve the quality of life of breast cancer survivors.

In practice, it is possible for a variable to be a member of several groups. Our LES penalty can be modified for variable selection when the groups have overlaps. The details are presented in Web Appendix H.

7. Supplementary Materials

Web Appendices A–H referenced in Sections 2, 3, 5, 6, and MATLAB code for the LES method are available at the *Biometrics* website on Wiley Online Library.

Acknowledgements

We would like to thank Editor, Associate Editor and a reviewer for their insightful and constructive comments and suggestions to improve the article. The work of Geng and Wahba was supported in part by NIH Grant EY09946 and NSF Grant 0906818. The work of Wang was supported in part by NIH Grant 5R01HG007377-02. The work of Monahan and Champion was supported in part by American Cancer Society Grant RSGPB-04-089-01-PBP.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *International Symposium on Information Theory, 2nd, Tsahkadsor, Armenian SSR*, 267–281.

Bickel, P., Ritov, Y., and Tsybakov, A. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics* **37**, 1705–1732.

Campbell, A., Converse, P., and Rodgers, W. (1976). *The Quality of American life: Perceptions, Evaluations, and Satisfactions*. New York, NY: Russell Sage Foundation.

Carrico, A., Antoni, M., Durán, R., Ironson, G., Penedo, F., Fletcher, M., Klimas, N., and Schneiderman, N. (2006). Reductions in depressed mood and denial coping during cognitive behavioral stress management with hiv-positive gay men treated with heart. *Annals of Behavioral Medicine* **31**, 155–164.

Craven, P. and Wahba, G. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik* **31**, 377–403.

Dicker, L., Huang, B., and Lin, X. (2012). Variable selection and estimation with the seamless-L0 penalty. *Statistica Sinica* **23**, 929–962.

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

Folkman, S. and Lazarus, R. (1980). An analysis of coping in a middle-aged community sample. *Journal of Health and Social Behavior* **21**, 219–239.

Folkman, S. and Lazarus, R. (1985). If it changes it must be a process: Study of emotion and coping during three stages of a college examination. *Journal of Personality and Social Psychology* **48**, 150.

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

Gall, T., Charbonneau, C., Clarke, N., Grant, K., Joseph, A., and Shouldice, L. (2005). Understanding the nature and role of spirituality in relation to coping and health: A conceptual framework. *Canadian Psychology/Psychologie Canadienne* **46**, 88.

Girard, D. (1987). A fast "Monte Carlo cross validation" procedure for large least squares problems with noisy data. *Technical Report RR 687-M*, IMAG, Grenoble, France.

Girard, D. (1989). A fast 'monte-carlo cross-validation' procedure for large least squares problems with noisy data. *Numerische Mathematik* **56**, 1–23.

Holmes, J. and Stevenson, C. (1990). Differential effects of avoidant and attentional coping strategies on adaptation to chronic and recent-onset pain. *Health Psychology* **9**, 577.

Huang, J., Ma, S., Xie, H., and Zhang, C. (2009). A group bridge approach for variable selection. *Biometrika* **96**, 339.

Hutchinson, M. (1989). A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines. *Communications in Statistics-Simulation and Computation* **18**, 1059–1076.

Leng, C., Lin, Y., and Wahba, G. (2006). A note on the lasso and related procedures in model selection. *Statistica Sinica* **16**, 1273.

Lv, J. and Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *The Annals of Statistics* **37**, 3498–3528.

Nardi, Y. and Rinaldo, A. (2008). On the asymptotic properties of the group lasso estimator for linear models. *Electronic Journal of Statistics* **2**, 605–633.

Parle, M., Maguire, P., and Heaven, C. (1997). The development of a training model to improve health professionals' skills, self-efficacy and outcome expectancies when communicating with cancer patients. *Social Science & Medicine* **44**, 231–240.

Purnell, J. and Andersen, B. (2009). Religious practice and spirituality in the psychological adjustment of survivors of breast cancer. *Counseling and Values* **53**, 165–182.

Rottmann, N., Dalton, S., Christensen, J., Frederiksen, K., and Johansen, C. (2010). Self-efficacy, adjustment style and well-being in breast cancer patients: A longitudinal study. *Quality of Life Research* **19**, 827–836.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**, 461–464.

Shen, X., Pan, W., and Zhu, Y. (2012). Likelihood-based selection and sharp parameter estimation. *Journal of the American Statistical Association* **107**, 223–232.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2012). A sparse-group lasso. *Journal of Computational and Graphical Statistics*, DOI: 10(10618600.2012):681250.

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

Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications* **109**, 475–494.

Wahba, G. (1985). A comparison of gcv and gml for choosing the smoothing parameter in the generalized spline smoothing problem. *The Annals of Statistics* **13**, 1378–1402.

Wang, H., Li, R., and Tsai, C. (2007). Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika* **94**, 553–568.

Wu, T. and Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics* **2**, 224–244.

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model identification and regression estimation. *Biometrika* **92**, 937–950.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B* **68**, 49–67.

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

Zhao, P., Rocha, G., and Yu, B. (2006). Grouped and hierarchical model selection through composite absolute penalties. *Department of Statistics, UC Berkeley, Tech. Rep*, 703.

Zhou, N. and Zhu, J. (2010). Group variable selection via a hierarchical lasso and its oracle property. *Arxiv preprint arXiv:1006.2871*.

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

Received June 2013. *Revised June* 2014. *Accepted July* 2014.