Ann Inst Stat Math (2017) 69:997–1025
DOI 10.1007/s10463-016-0571-z

A doubly sparse approach for group variable selection

Sunghoon Kwon[1] · Jeongyoun Ahn[2] ·
Woncheol Jang[3] · Sangin Lee[4] · Yongdai Kim[3]

Received: 17 December 2014 / Published online: 28 June 2016

Abstract We propose a new penalty called the doubly sparse (DS) penalty for variable selection in high-dimensional linear regression models when the covariates are naturally grouped. An advantage of the DS penalty over other penalties is that it provides a clear way of controlling sparsity between and within groups, separately. We prove that there exists a unique global minimizer of the DS penalized sum of squares of residuals and show how the DS penalty selects groups and variables within selected groups, even when the number of groups exceeds the sample size. An efficient optimization algorithm is introduced also. Results from simulation studies and real data analysis show that the DS penalty outperforms other existing penalties with finite samples.

Keywords Doubly sparse penalty · Group selection · Group selection consistency · Variable selection

✉ Yongdai Kim
ydkim0903@gmail.com

[1] Department of Applied Statistics, Konkuk University, 120 Neungdong-ro, Gwangjin-gu, Seoul 05029, Korea

[2] Department of Statistics, University of Georgia, Athens, GA 30602, USA

[3] Department of Statistics, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 08826, Korea

[4] Quantitative Biomedical Research Center, University of Texas Southwestern Medical Center, Dallas, TX 75390, USA

1 Introduction

Consider the following linear regression model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}^* + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^{\mathrm{T}} \in \mathbb{R}^n$ is a response vector, $\boldsymbol{\beta}^* = (\beta_1^*, \ldots, \beta_p^*)^{\mathrm{T}} \in \mathbb{R}^p$ is a true regression coefficient vector, $\mathbf{X} = (X_1, \ldots, X_p)$ is an $n \times p$ design matrix and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^{\mathrm{T}} \in \mathbb{R}^n$ is a random error vector. Penalized estimations have received much attention for variable selection in the model (1), that estimate the true regression coefficient vector $\boldsymbol{\beta}^*$ by minimizing the penalized sum of squares of residuals:

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2/2n + \sum_{j=1}^{p} J_\lambda(|\beta_j|), \tag{2}$$

where J_λ is a penalty, in which the tuning parameter λ controls sparsity of the fitted model. The least absolute selection and shrinkage operator (LASSO) (Tibshirani 1996) has become so popular since it does parameter estimation and variable selection simultaneously with the penalty $J_\lambda^L(|t|) = \lambda|t|$. On the other hand, the LASSO is known to select more variables than necessary (Zou 2006) unless some strong conditions are provided (Zhao and Yu 2006; Meinshausen and Yu 2009), and produces unnecessary bias to zero for large regression coefficients. To improve the LASSO, various non-convex penalties have been proposed such as the smoothly clipped absolute deviation (SCAD) penalty, (Fan and Li 2001) $\mathrm{d}J_\lambda^S(|t|)/\mathrm{d}|t| = \min\{\lambda, (a\lambda - |t|)_+/(a+1)\}, a > 2$, the Bridge penalty (Huang et al. 2008), $J_\lambda^B(|t|) = \lambda|t|^\nu, 0 < \nu \le 1$, and the minimax concave penalty (MCP) (Zhang 2010), $\mathrm{d}J_\lambda^M(|t|)/\mathrm{d}|t| = (\lambda - |t|/a)_+, a > 1$, where $x_+ = xI(x \ge 0)$. These non-convex penalties are known to have the oracle property: asymptotic equivalence to an ideal non-penalized estimator obtained with true predictive variables only (Fan and Peng 2004; Kim et al. 2008; Zhang 2010). We refer to Zhang and Zhang (2012) for a well organized review of penalized estimations for variable selection in high-dimensional linear regression.

In this paper, we consider a case where the p covariates can be decomposed into K disjoint groups. In this case, the model (1) can be rewritten as

$$\mathbf{y} = \sum_{k=1}^{K} \mathbf{X}_k \boldsymbol{\beta}_k^* + \boldsymbol{\varepsilon}, \tag{3}$$

where $\mathbf{X}_k = (\mathbf{X}_{k1}, \ldots, \mathbf{X}_{kp_k})$ and $\boldsymbol{\beta}_k^* = (\beta_{k1}^*, \ldots, \beta_{kp_k}^*)^{\mathrm{T}} \in \mathbb{R}^{p_k}$ are a $n \times p_k$ design matrix and a true regression coefficient vector for the kth group, respectively, satisfying $p = \sum_{k=1}^K p_k$.

In high-dimensional linear regression models, it is not uncommon that covariates are naturally grouped, and hence group selection is of interest. For example, linear regression models with categorical covariates can be expressed with (3) via groups of dummy variables. In nonparametric additive models, each component can be expanded by a set of basis functions and the selection of components is equivalent to selecting groups of basis functions. In practice, selecting both groups and variables within the selected groups is interesting rather than selecting groups only. An example is a nonparametric additive model with wavelet basis, where choosing the basis functions within the selected components is as important as selecting components (Sardy and Tseng 2004).

There have been some developments in recent years that extend and generalize the existing penalties to group variable selection problems, by minimizing the following L_q-norm composite criterion:

$$\left\| \mathbf{y} - \sum_{k=1}^{K} \mathbf{X}_k \boldsymbol{\beta}_k \right\|_2^2 /2n + \sum_{k=1}^{K} J_{\lambda_k}(\|\boldsymbol{\beta}_k\|_q) \tag{4}$$

for $q \in \{1, 2\}$, where λ_k's are tuning parameters that control sparsity of groups. For example, Yuan and Lin (2006) proposed the group LASSO using $J_{\lambda_k}^L(|t|) = \lambda_k|t|$ for $\lambda_k = \lambda p_k^{1/2}$ and $q = 2$, which is equivalent to the LASSO when each group consists of just one covariate. Another example is the L_2-norm MCP proposed by Huang et al. (2012). They used the MCP $J_{\lambda_k}(|t|) = J_{\lambda_k}^M(|t|)$ for $\lambda_k = \lambda p_k^{1/2}$ and $q = 2$, which is equivalent to $J_\lambda^M(p_k^{1/2}|t|)$ for $a = ap_k$. Note that the group LASSO and L_2-norm MCP do not achieve sparsity within the selected groups since they use L_2-norm inside the penalties.

On the other hand, Huang et al. (2009) proposed the group Bridge that selects groups and variables simultaneously by minimizing a L_1-norm composite criterion where $J_{\lambda_k}(|t|) = J_{\lambda_k}^B(|t|)$ for $\lambda_k = \lambda p_k^\nu$ and $q = 1$, which is equivalent to $J_{\lambda_k}^B(|t|) = J_\lambda^B(p_k|t|)$. The group Bridge is an extension of the Bridge penalty to group variable selection, including the hierarchical LASSO proposed by Zhou and Zhu (2010) as a special case for $\nu = 1/2$. Recently, Jiang and Huang (2015) introduced general L_1-norm framework that includes the L_1-norm MCP and SCAD as special cases. For example, the L_1-norm MCP uses $J_{\lambda_k}(|t|) = J_{\lambda_k}^M(|t|)$ for $\lambda_k = \lambda p_k$ and $q = 1$. Another example is the L_1-norm exponential penalty proposed by Breheny (2015) where $J_{\lambda_k}(|t|) = (\lambda^2/\tau)\{1 - \exp(\tau|t|/\lambda)\}$ with an extra tuning parameter τ that controls the coupling effect.

The L_q-norm composite criterion in (4) can be a special case of the inner–outer composite criterion introduced by Breheny and Huang (2009):

$$\left\| \mathbf{y} - \sum_{k=1}^{K} \mathbf{X}_k \boldsymbol{\beta}_k \right\|_2^2 /2n + \sum_{k=1}^{K} J_{\lambda_k}^O\left(\sum_{j=1}^{p_k} J_\gamma^I(|\beta_{kj}|) \right),$$

where $J_{\lambda_k}^O$ is an outer penalty for group selection and J_γ^I is an inner penalty for variable selection. For example, the group Bridge uses the outer Bridge penalty $J_{\lambda_k}^O(|t|) = J_{\lambda_k}^B(|t|)$ and the inner LASSO $J_\gamma^I(|t|) = \gamma|t|$, where the inner tuning parameter is fixed with $\gamma = 1$. Another example is the group LASSO that uses the same outer penalty as the group Bridge but the inner ridge penalty $J_\gamma^I(|t|) = \gamma|t|^2$, where the inner tuning parameter is fixed with $\gamma = 1$. Breheny and Huang (2009) proposed the composite MCP that uses the MCP for both inner and outer penalties where $\lambda_k = \gamma$ and $a_k = ap_k\gamma/2$.

There is another approach for group and variable selection which uses the sum of two different penalties:

$$\left\| \mathbf{y} - \sum_{k=1}^{K} \mathbf{X}_k \boldsymbol{\beta}_k \right\|_2^2 /2n + \sum_{k=1}^{K} J_{\lambda_k}^G(\|\boldsymbol{\beta}_k\|_2) + \sum_{k=1}^{K} \sum_{j=1}^{p_k} J_\gamma^V(|\beta_{kj}|),$$

where $J_{\lambda_k}^G$ is a penalty for selection of groups and J_γ^V for variables. For example, the sparse group LASSO of Simon et al. (2013) adds L_1-penalty, $J_\gamma^V(|t|) = \gamma|t|$ to the group LASSO, $J_{\lambda_k}^G(|t|) = \lambda_k|t| = \lambda p_k^{1/2}|t|$ to yield sparsity within the selected groups. One unique feature of the sparse group LASSO is to use two different tuning parameters λ and γ to control sparsity between and within groups, which gives a clearer way of controlling sparsity in practice, although the use of two different tuning parameters may cause higher computational cost than other methods. We refer to Huang et al. (2012) for a nice review of penalized approaches for group and variable selection in linear regression models.

As aforementioned, the L_1-norm composite methods such as the group Bridge and L_1-norm MCP can handle sparsity both between and within groups. However, it is not clear how to control the two sparsity, especially sparsity within groups, since they fix the inner tuning parameter $\gamma = 1$. An exceptional example is the L_1-norm exponential penalty proposed by Breheny (2015) which introduces an extra tuning parameter to control the coupling effect. However, controlling coupling effect focuses on how the relevant and irrelevant variables in a group affect to each other rather than how to distinguish important variables from unimportant ones. Further, these methods are in lack of theoretical results in high-dimensional linear regression models. This is partly because they do not have a closed form of the global minimizer or an oracle estimator that may help us to study asymptotic properties of the global minimizer. On the other hand, the L_2-norm MCP takes an advantage of the existence of the oracle group estimator in (10), with which it can be shown that it achieves an oracle property (Huang et al. 2012) in group selection when there is no sparsity within groups. Note that for the composite MCP, the oracle estimator in (9) can be an intuitive global minimizer also when there is sparsity within groups. However, the systematic obscurity and lack of theoretical support still remain with the choice $\lambda_k = \gamma$ and $a_k = ap_k\gamma/2$.

In this paper, we propose a new penalized approach for group and variable selection, using the doubly sparse (DS) penalty. The DS penalty is an L_1-norm composite penalty, with the clipped LASSO proposed by Kwon et al. (2015) as the outer penalty. An important feature of the DS penalty compared to other L_1-norm composite penalties is that it can separately control sparsity between and within groups separately. Hence, the principle of selecting groups and variables is much more transparent.

Theoretically, the DS penalized approach achieves a group selection consistency, allowing the number of groups K to exceed the sample size n, which is common to other group penalties. However, one benefit of using L_1-composite penalties is to allow the maximum number of variables in signal groups, $p^*_{\max} = \max_{k:\|\boldsymbol{\beta}_k\|_1 \neq 0} p_k$, to exceed the sample size n also which is impossible to other L_2-composite penalties such as the L_2-norm MCP.

To our best knowledge, there are no known results of group selection consistency of other L_1-norm composite penalties such as the group Bridge and hierarchical LASSO, when $p > n$ and both levels of sparsity are assumed. Note that, the sparse group LASSO requires a strong irrepresentability condition for group selection consistency, since it selects groups as the group LASSO (Wei and Huang 2010). Hence we believe the applicability of the DS penalty is much wider than existing methods.

The paper is organized as follows. In Sect. 2, we introduce the DS penalty including an efficient optimization algorithm. Section 3 gives asymptotic properties and Sect. 4 presents the results from numerical studies including simulation as well as real data analysis. Concluding remarks and technical details are presented in Sect. 5 and Appendix, respectively.

2 Doubly sparse penalty

2.1 Definition and solution

Let $a > 0$ be a fixed constant. For given $\lambda > 0$ and $0 \leq \gamma \leq \lambda$, the DS penalized estimator is defined as the global minimizer of the DS penalized sum of squared residuals,

$$Q_{\lambda,\gamma}(\boldsymbol{\beta}) = \left\| \mathbf{y} - \sum_{k=1}^{K} \mathbf{X}_k \boldsymbol{\beta}_k \right\|_2^2 /2n + \sum_{k=1}^{K} J^{(k)}_{\lambda,\gamma}(\|\boldsymbol{\beta}_k\|_1), \tag{5}$$

where the DS penalty is a L_1-norm composite penalty for the clipped LASSO (Kwon et al. 2015): $J^{(k)}_{\lambda,\gamma}(0) = 0$,

$$\begin{aligned}\nabla J^{(k)}_{\lambda,\gamma}(|t|) &\equiv dJ^{(k)}_{\lambda,\gamma}(|t|)/d|t| \\ &= \big(-|t|/a_k + \lambda\big) I\big\{|t| < a_k(\lambda-\gamma)\big\} + \gamma I\big\{|t| \geq a_k(\lambda-\gamma)\big\}\end{aligned}$$

for $|t| > 0$ and $a_k = ap_k$. The clipped LASSO is a continuously differentiable quadratic spline interpolation of two penalty functions, the MCP and LASSO:

$$J^{(k)}_{\lambda,\gamma}(|t|) = J^M_\lambda(|t|) I\big\{|t| < a_k(\lambda-\gamma)\big\} + (J^L_\gamma|t| + c^{(k)}_{\lambda,\gamma}) I\big\{|t| \geq a_k(\lambda-\gamma)\big\},$$

where $c^{(k)}_{\lambda,\gamma} = a_k(\lambda-\gamma)^2/2$. Hence, the clipped LASSO is the same as the MCP with a tuning parameter λ for small $|t| \leq a_k(\lambda-\gamma)$, and the LASSO with a tuning parameter

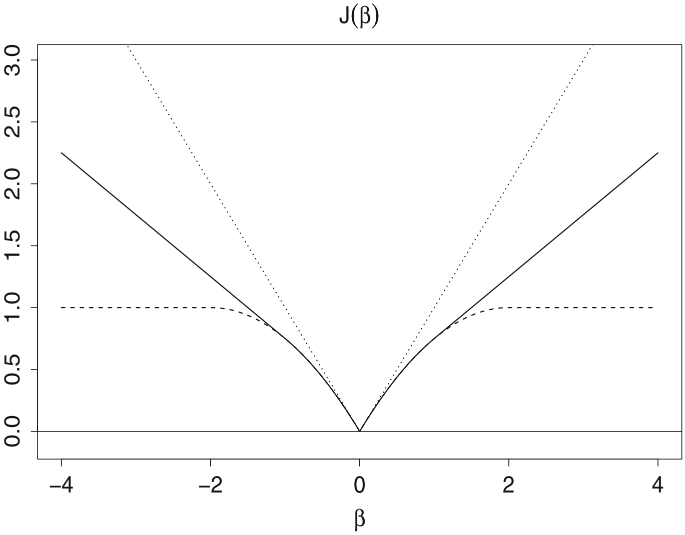

Fig. 1 Three penalty functions: the LASSO (*dotted*) with $\lambda = 1$, the MCP (*dashed*) with $(\lambda, a) = (1, 2)$ and the clipped LASSO (*line*) with $(\lambda, \gamma, a) = (1, 1/2, 2)$

γ for large $|t| > a_k(\lambda - \gamma)$. See Fig. 1 that depicts three penalty functions, LASSO, clipped LASSO and MCP.

When $\gamma = \lambda$, the DS penalty becomes the LASSO so it selects variables as the LASSO without using group information. When $\gamma = 0$, the DS penalty becomes a L_1-norm MCP, which can select groups and variables using a tuning parameter λ. When $0 < \gamma < \lambda$, which is new, we will study various properties of the DS penalty.

2.2 Roles of two tuning parameters

Let $\mathbf{\Omega}_{\lambda,\gamma}$ be the set of all local minimizers of $Q_{\lambda,\gamma}$. Given $\boldsymbol{\beta} \in \mathbb{R}^p$, let $\mathcal{A}(\boldsymbol{\beta}) = \mathcal{L}(\boldsymbol{\beta}) \cup \mathcal{S}(\boldsymbol{\beta}) = \{k : \|\boldsymbol{\beta}_k\|_1 \neq 0\}$ and $\mathcal{N}(\boldsymbol{\beta}) = \{k : \|\boldsymbol{\beta}_k\|_1 = 0\} = \mathcal{A}(\boldsymbol{\beta})^c$ be the sets of group indices of coefficient vectors that have nonzero and zero L_1-norms, respectively, where $\mathcal{L}(\boldsymbol{\beta}) = \{k : \|\boldsymbol{\beta}_k\|_1/p_k > a(\lambda - \gamma)\}$ and $\mathcal{S}(\boldsymbol{\beta}) = \{k : 0 < \|\boldsymbol{\beta}_k\|_1/p_k \leq a(\lambda - \gamma)\}$. Let $D_{kj}(\boldsymbol{\beta}) = -\mathbf{X}_{kj}^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})/n$ for $k \leq K$ and $j \leq p_k$, which represents the sample covariance between $-\mathbf{X}_{kj}$ and residual $\mathbf{y} - \mathbf{X}\boldsymbol{\beta}$. Similarly, let $D_k(\boldsymbol{\beta}) = -\mathbf{X}_k^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})/n$ for $k \leq K$, which represents the vector of sample covariances of the kth group whose jth element is $D_{kj}(\boldsymbol{\beta})$ for $j \leq p_k$. Lemma 1 below directly comes from the first order optimality conditions in Bertsekas (1999).

Lemma 1 *For any* $\hat{\boldsymbol{\beta}} \in \mathbf{\Omega}_{\lambda,\gamma}$,

$$\begin{cases} D_{kj}(\hat{\boldsymbol{\beta}}) = -\gamma \operatorname{sign}(\hat{\beta}_{kj}), & k \in \mathcal{L}(\hat{\boldsymbol{\beta}}),\ \hat{\beta}_{kj} \neq 0, & \text{(6a)} \\ |D_{kj}(\hat{\boldsymbol{\beta}})| \leq \gamma, & k \in \mathcal{L}(\hat{\boldsymbol{\beta}}),\ \hat{\beta}_{kj} = 0, & \text{(6b)} \\ D_{kj}(\hat{\boldsymbol{\beta}}) = -\Delta_\lambda\big(\|\hat{\boldsymbol{\beta}}_k\|_1/p_k\big)\operatorname{sign}(\hat{\beta}_{kj}), & k \in \mathcal{S}(\hat{\boldsymbol{\beta}}),\ \hat{\beta}_{kj} \neq 0, & \text{(6c)} \\ |D_{kj}(\hat{\boldsymbol{\beta}})| \leq \Delta_\lambda\big(\|\hat{\boldsymbol{\beta}}_k\|_1/p_k\big), & k \in \mathcal{S}(\hat{\boldsymbol{\beta}}),\ \hat{\beta}_{kj} = 0, & \text{(6d)} \\ \|D_k(\hat{\boldsymbol{\beta}})\|_\infty \leq \lambda, & k \in \mathcal{N}(\hat{\boldsymbol{\beta}}), & \text{(6e)} \end{cases}$$

for all $k \leq K$ and $j \leq p_k$, where $\Delta_\lambda(t) = \lambda - t/a$ and $\operatorname{sign}(t) = (t/|t|)I[t \neq 0]$.

The conditions in Lemma 1 show how the two tuning parameters λ and γ work in the DS penalty. First, the Eq. (6e) shows that λ gives a threshold for group sparsity. Second, the Eqs. (6a) and (6b) shows that γ gives a threshold for variable sparsity within selected groups as the LASSO in the region $\mathcal{L}(\hat{\boldsymbol{\beta}})$, since they are exactly the first order optimality conditions for the LASSO with tuning parameter γ. Third, the Eqs. (6c) and (6d) show that the DS penalty behaves like the L_1-norm MCP in the region $\mathcal{S}(\hat{\boldsymbol{\beta}})$. Note that the events defined by the Eqs. (6c) and (6d) are null events in the asymptotic sense under some assumptions. Hence, we may simply understand the DS penalized estimator through the lemma below.

Lemma 2 *If* $\hat{\boldsymbol{\beta}} \in \mathbb{R}^p$ *satisfies*

$$\begin{cases} \|\mathbf{X}_k^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})/n\|_\infty \leq \lambda, & \|\hat{\boldsymbol{\beta}}_k\|_1 = 0, & \text{(7a)} \\ \|\hat{\boldsymbol{\beta}}_k\|_1/p_k > a(\lambda - \gamma) & \|\hat{\boldsymbol{\beta}}_k\|_1 \neq 0, & \text{(7b)} \\ \mathbf{X}_{kj}^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})/n = \gamma \operatorname{sign}(\hat{\beta}_{kj}), & \|\hat{\boldsymbol{\beta}}_k\|_1/p_k > a(\lambda - \gamma),\ \hat{\beta}_{kj} \neq 0, & \text{(7c)} \\ |\mathbf{X}_{kj}^{\mathrm{T}}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})/n| \leq \gamma, & \|\hat{\boldsymbol{\beta}}_k\|_1/p_k > a(\lambda - \gamma),\ \hat{\beta}_{kj} = 0 & \text{(7d)} \end{cases}$$

for all $k \leq K$ and $j \leq p_k$, then $\hat{\boldsymbol{\beta}} \in \boldsymbol{\Omega}_{\lambda,\gamma}$.

The conditions in Lemma 2 make us to expect that there is an estimator $\hat{\boldsymbol{\beta}}$ that selects groups and then selects variables within selected groups. The condition (7a) requires that $\hat{\boldsymbol{\beta}}$ excludes the groups whose maximum sample covariances are smaller than λ. The condition (7b) requires that $\hat{\boldsymbol{\beta}}$ includes the groups whose averaged L_1-norms of coefficient vectors are larger than $a(\lambda - \gamma)$. The last two conditions (7c) and (7d) show that $\hat{\boldsymbol{\beta}}$ becomes the LASSO with tuning parameter γ, deleting variables within selected groups whose sample covariances are smaller than γ. Hence, in contrast to other existing penalties, the DS penalty provides a clear way of controlling the sparsity between and within groups, which also shows the roles of two tuning parameters λ and γ.

An oracle estimator for group and variable selection will be introduced in next subsection that satisfies these conditions asymptotically, which is an extension of the existence of an oracle estimator for variables selection. This is similar to the fundamental idea on the usual oracle property for the model (1) studied by many researchers (Kim et al. 2008; Zhang 2010; Huang et al. 2012), where the sufficient conditions for a given $\hat{\boldsymbol{\beta}}$ to be a local minimizer of (2) are given by

$$\begin{cases} |\mathbf{X}_j^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})/n| \le \lambda, & \hat{\beta}_j = 0, \quad (8a)\\ |\hat{\beta}_j| > a\lambda, & \hat{\beta}_j \ne 0, \quad (8b)\\ \mathbf{X}_j^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}})/n = 0, & |\hat{\beta}_j| > a\lambda, \quad (8c) \end{cases}$$

for all $j \le p$, for a class of nonconvex penalties (Zhang and Zhang 2012; Kim and Kwon 2012). Note that the conditions (8a) and (8b) require that $\hat{\boldsymbol{\beta}}$ includes the variables whose absolute values of coefficients are larger than $a\lambda$, and excludes the variables whose sample covariances are smaller than λ, becoming a least square estimator that is unbiased as the condition (8c) requires.

2.3 Global minimizer

In high-dimensional linear regression models, the existence of an oracle estimator plays an important role for the study of the penalized estimator. For example, the oracle estimator,

$$\hat{\boldsymbol{\beta}}^o = \underset{\beta_j=0,\, j\in\{l:\beta_l^*=0\}}{\arg\min} \|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_2^2/2n, \tag{9}$$

is proven to be the global minimizer of the penalized sum of squares of residuals in (2) for various penalties such as the SCAD penalty (Kim and Kwon 2012) and MCP (Zhang 2010), which shows $\hat{\boldsymbol{\beta}}^o$ is exactly the global minimizer of the penalized sum of squares of residuals asymptotically. A similar result in group variable selection can be found in Huang et al. (2012) with the oracle group estimator,

$$\hat{\boldsymbol{\beta}}^o = \underset{\|\boldsymbol{\beta}_k\|_1=0,\, k\in\mathcal{N}(\boldsymbol{\beta}^*)}{\arg\min} \left\|\mathbf{y} - \sum_{k=1}^K \mathbf{X}_k\boldsymbol{\beta}_k\right\|_2^2 /2n, \tag{10}$$

which asymptotically becomes the global minimizer of the L_2-norm composite criterion (4) with the MCP when there is no sparsity within groups. Before introducing an oracle estimator for the DS penalized criterion in (5), we give sufficient conditions under which the global minimizer of $Q_{\lambda,\gamma}$ is unique when $p \le n$. Let $\lambda_{\min}(\mathbf{D})$ and $\lambda_{\max}(\mathbf{D})$ be the smallest and largest eigenvalues of a given symmetric matrix $\mathbf{D}$. The following lemma shows an analogous result to Kim and Kwon (2012) for group variable selection.

Lemma 3 *Assume that $\rho_{\min} = \lambda_{\min}(\mathbf{X}^{\mathrm{T}}\mathbf{X}/n) > 0$. If there exists a local minimizer $\hat{\boldsymbol{\beta}} \in \boldsymbol{\Omega}_{\lambda,\gamma}$ that satisfies $\mathcal{S}(\hat{\boldsymbol{\beta}}) = \emptyset$ and*

$$\begin{cases} \min_{k\in\mathcal{L}(\hat{\boldsymbol{\beta}})} \|\hat{\boldsymbol{\beta}}_k\|_1/p_k > (\lambda-\gamma)\max\{a, 1/\rho_{\min}\},\\ \max_{k\in\mathcal{N}(\hat{\boldsymbol{\beta}})} \|D_k(\hat{\boldsymbol{\beta}})\|_\infty < (\lambda-\gamma)\min\{a\rho_{\min}, 1\} + \gamma, \end{cases}$$

then $\mathbf{\Omega}_{\lambda,\gamma} = \{\hat{\boldsymbol{\beta}}\}$. Lemma 3 plays a key role when we prove the main theorems in the next section under a sparsity condition when $p > n$. Lemma 3 shows that the global minimizer is unique even when $Q_{\lambda,\gamma}$ is non-convex under the non-asymptotic sufficient conditions: the nonzero coefficients of the estimator are large so that $\mathcal{S}(\hat{\boldsymbol{\beta}}) = \emptyset$, but covariances between current residuals and irrelevant covariates are sufficiently small.

From the optimality conditions in Lemma 2 and 3, it is easy to see that the following estimator,

$$\hat{\boldsymbol{\beta}}^{o}(\gamma) = \underset{\|\boldsymbol{\beta}_k\|_1=0, k\in\mathcal{N}(\boldsymbol{\beta}^*)}{\arg\min} \left\{ \left\| \mathbf{y} - \sum_{k=1}^{K} \mathbf{X}_k \boldsymbol{\beta}_k \right\|_2^2 /2n + \gamma \sum_{k=1}^{K} \|\boldsymbol{\beta}_k\|_1 \right\}, \tag{11}$$

becomes the unique global minimizer of $Q_{\lambda,\gamma}$ if $\hat{\boldsymbol{\beta}}^{o}(\gamma)$ satisfies the sufficient conditions in Lemma 3. Note that $\hat{\boldsymbol{\beta}}^{o}(\gamma)$ is the LASSO for which γ controls sparsity of variables in the signal groups. In this sense, we refer $\hat{\boldsymbol{\beta}}^{o}(\gamma)$ as the oracle LASSO. In practice, the oracle LASSO is unavailable since the signal groups are unknown. However, we will show that the oracle LASSO is exactly the DS penalized estimator asymptotically, if we choose λ to have correct group sparsity. This result also explains roles of the two tuning parameters λ and γ clearly: after selecting groups by λ, the shrinkage and selection within the chosen groups are controlled by γ.

2.4 Optimization algorithm

Recall that we estimate $\boldsymbol{\beta}$ by minimizing $Q_{\lambda,\gamma}$. Let $\tilde{J}_{\lambda,\gamma}(\boldsymbol{\beta}) = \sum_{k=1}^{K} J_{\lambda,\gamma}^{(k)}(\|\boldsymbol{\beta}_k\|_1) - \lambda\|\boldsymbol{\beta}\|_1$. It is easy to see that we can decompose $Q_{\lambda,\gamma}$ into a sum of convex and concave functions as follows.

$$Q_{\lambda,\gamma}(\boldsymbol{\beta}) = Q_{vex}(\boldsymbol{\beta}) + Q_{cav}(\boldsymbol{\beta}),$$

where $Q_{vex}(\boldsymbol{\beta}) = \|\mathbf{y} - \sum_{k=1}^{K} \mathbf{X}_k\boldsymbol{\beta}_k\|_2^2/2n + \lambda\|\boldsymbol{\beta}\|_1$ is a convex function, and $Q_{cav}(\boldsymbol{\beta}) = \tilde{J}_{\lambda,\gamma}(\boldsymbol{\beta})$ is a continuously differentiable concave function. Hence, we can find a local minimizer with the convex concave procedure (CCCP) of Yuille and Rangarajan (2003) or the difference of convex (DC) decomposition procedure of An and Tao (1997) which are powerful algorithms for nonconvex optimization problems. For example, we can apply the CCCP as follows. Let $\partial\tilde{J}_{\lambda,\gamma}(\boldsymbol{\beta})$ be any subgradient of $\tilde{J}_{\lambda,\gamma}(\boldsymbol{\beta})$ at $\boldsymbol{\beta}$. For a given solution $\hat{\boldsymbol{\beta}}^{c}$, we update the solution by minimizing

$$\left\| \mathbf{y} - \sum_{k=1}^{K} \mathbf{X}_k\boldsymbol{\beta}_k \right\|_2^2 /2n + (\partial\tilde{J}_{\lambda,\gamma}(\hat{\boldsymbol{\beta}}^{c}))^{\mathrm{T}}\boldsymbol{\beta} + \lambda\|\boldsymbol{\beta}\|_1,$$

which can be solved by many efficient optimization algorithms for the LASSO such as the modified least angle regression (Efron et al. 2004) and coordinate descent algorithm (Friedman et al. 2007). We iterate this procedure until the solution converges. Note that the CCCP always converges to a local minimizer (Sriperumbudur and Lanckriet 2009).

3 Asymptotic properties

In this section we study statistical properties of the DS penalized estimator when $p > n$. For any subset $S \subset \{(k, j) : k \leq K, j \leq p_k\}$, define $\mathbf{X}_S$ as the $n \times |S|$ matrix that is obtained by combining column vectors $\mathbf{X}_{kj}$ for $(k, j) \in S$. We assume that there exist nonempty subsets $A_* = \{(k, j) : \beta^*_{kj} \neq 0\}$ and $G_* = \{(k, j) : k \in \mathcal{A}(\boldsymbol{\beta}^*), j \leq p_k\}$ that are index sets of true nonzero regression coefficients of variables and groups, respectively. To study properties of the DS penalized estimator, we need the following assumptions:

(C1) The random errors $\varepsilon_1, \ldots, \varepsilon_n$ are independently and identically distributed mean zero sub-Gaussian random variables with a positive scale factor $\sigma_0 < \infty$, that is, $E \exp(t\varepsilon_i) \leq \exp(\sigma_0^2 t^2/2)$ for all $i \leq n$ and $t > 0$.

(C2) There exist constants $\alpha_0 > 0$ and $\kappa_{\min} > 0$ such that

$$\phi_{\min}(2(\alpha_0 + 1)|A_*|) \geq \kappa_{\min}, \tag{12}$$

where $\phi_{\min}(m) = \min_{|B| \leq m, A_* \subset B} \lambda_{\min}(\mathbf{X}_B^{\mathrm{T}}\mathbf{X}_B/n)$ is the lower sparse eigenvalue in Zhang and Zhang (2012), and

$$\phi_{\max}(\alpha_0|A_*|)/\alpha_0 \leq (1 - 3\sqrt{\phi_{\max}(\alpha_0|A_*|)/\alpha_0\kappa_{\min}})^2/576, \tag{13}$$

where $\phi_{\max}(m) = \min_{|B| \leq m, A_* \subset B} \lambda_{\max}(\mathbf{X}_B^{\mathrm{T}}\mathbf{X}_B/n)$ is the upper sparse eigenvalue in Zhang and Zhang (2012).

Remark 1 (C1) implies that there exist constants $c_0 > 0$ and $d_0 > 0$ such that the error vector $\boldsymbol{\varepsilon}$ satisfies

$$\mathbf{P}\big(|\mathbf{a}^{\mathrm{T}}\boldsymbol{\varepsilon}| > t\big) \leq c_0 \exp(-d_0 t^2/\|\mathbf{a}\|_2^2), \tag{14}$$

for all $\mathbf{a} \in \mathbb{R}^n$ and $t > 0$. The inequality (12) in (C2) ensures model identifiability and uniqueness of $\hat{\boldsymbol{\beta}}^o(\gamma)$ as a local minimizer. In fact, $\phi_{\min}((\alpha_0 + 1)|A_*|) \geq \kappa_{\min}$ is sufficient for $\hat{\boldsymbol{\beta}}^o(\gamma)$ to be one of the local minimizers, which is often referred as the oracle property. The inequality (13) in (C2) assumes $\phi_{\max}(\alpha_0|A_*|)/\alpha_0$ is bounded, which is weaker than similar conditions in Bickel et al. (2009) and Meinshausen and Yu (2009), where it is assumed that $\phi_{\max}(|A_*| + \min\{n, p\})$ is bounded. A similar condition can be found in Wang et al. (2013). From the results in Zhang and Zhang (2012), (C2) controls the number of nonzero elements of $\hat{\boldsymbol{\beta}}^o(\gamma)$ up to an order of $|A_*|$ under the η-null consistency in Zhang and Zhang (2012).

First, we prove that the oracle LASSO is one of the local minimizers of $Q_{\lambda,\gamma}$ whose number of nonzero elements is less than $(\alpha_0 + 1)|A_*|$ with probability tending to 1 when $p > n$. Given an integer $s < n$, let

$$\mathbf{\Omega}_{\lambda,\gamma}(s) = \left\{\hat{\boldsymbol{\beta}} \in \mathbf{\Omega}_{\lambda,\gamma} : \|\hat{\boldsymbol{\beta}}\|_0 \leq s\right\}$$

be the set of all local minimizers that have s nonzero elements at most. Further, let $\xi^*_{\lambda,\gamma} = m_* - a(\lambda-\gamma) - \gamma\sqrt{(\alpha_0+1)|A_*|/\kappa^2_{\min}}$ and $\zeta^*_{\lambda,\gamma} = \lambda - \gamma\sqrt{(\alpha_0+1)|A_*|/\kappa_{\min}}$, where $m_* = \min_{(k,j)\in A_*} |\beta^*_{kj}|$.

Theorem 1 *Assume that* (C1) *and* (C2) *hold, then*

$$\mathbf{P}\big(\hat{\boldsymbol{\beta}}^o(\gamma) \in \mathbf{\Omega}_{\lambda,\gamma}((\alpha_0+1)|A_*|)\big) \geq 1 - \mathbf{P}_1 - \mathbf{P}_2 - \mathbf{P}_3,$$

where $\mathbf{P}_1 = c_0|G_*|\exp(-d_0 n\gamma^2/4)$, $\mathbf{P}_2 = c_0(\alpha_0+1)|A_*|\exp\big(-d_0\kappa_{\min} n\xi^{*2}_{\lambda,\gamma}\big)$ *and* $\mathbf{P}_3 = c_0(p-|G_*|)\exp\big(-d_0 n\zeta^{*2}_{\lambda,\gamma}\big)$.

Corollary 1 *Assume that* (C1) *and* (C2) *hold. If* $n\lambda^2 \to \infty$, $\lambda = o(m_*)$ *and* $n\gamma^2 \to \infty$ *then*

$$\mathbf{P}\big(\hat{\boldsymbol{\beta}}^o(\gamma) \in \mathbf{\Omega}_{\lambda,\gamma}((\alpha_0+1)|A_*|)\big) \to 1,$$

provided that $\log p = o(n\lambda^2)$, $\log|G_*| = o(n\gamma^2)$ *and* $\gamma = o(\lambda/\sqrt{|A_*|})$ *as* $n \to \infty$.

Remark 2 Theorem 1 and Corollary 1 imply that the oracle LASSO is one of the local minimizers of $Q_{\lambda,\gamma}$ even when $p > n$ under the sub-Gaussian assumption. Note that the total number of variables in the model, $p = \sum_{k=1}^K p_k$, is allowed to have an exponential order of n. For example, suppose there exists a constant $m_0 > 0$ such that $m_* \geq m_0 > 0$ for all n. Then, by letting $\lambda = n^{-1/3}$, Corollary 1 holds when $p = \exp(n^{\delta_0})$ for some constant $0 < \delta_0 < 1/3$.

Second, we prove that the DS penalized estimator is unique and asymptotically the same as the oracle LASSO, which is the main result of the paper. The next theorem and corollary prove that the unique local minimizer in $\mathbf{\Omega}_{\lambda,\gamma}((\alpha_0+1)|A_*|)$ is the oracle LASSO with probability tending to 1. Let $\xi^{**}_{\lambda,\gamma} = m_* - \max\{a, 1/\kappa_{\min}\}(\lambda-\gamma) - \gamma\sqrt{(\alpha_0+1)|A_*|/\kappa^2_{\min}}$ and $\zeta^{**}_{\lambda,\gamma} = (\lambda-\gamma)\min\{a\kappa_{\min}, 1\} - \gamma(\sqrt{(\alpha_0+1)|A_*|/\kappa_{\min}} - 1)$.

Theorem 2 *Assume that* (C1) *and* (C2) *hold, then*

$$\mathbf{P}\big(\mathbf{\Omega}_{\lambda,\gamma}((\alpha_0+1)|A_*|) = \{\hat{\boldsymbol{\beta}}^o(\gamma)\}\big) \geq 1 - \mathbf{P}_1 - \mathbf{P}_2 - \mathbf{P}_3,$$

where $\mathbf{P}_1 = c_0|G_*|\exp(-d_0 n\gamma^2/4)$, $\mathbf{P}_2 = c_0(\alpha_0+1)|A_*|\exp\big(-d_0\kappa_{\min} n\xi^{**2}_{\lambda,\gamma}\big)$ and $\mathbf{P}_3 = c_0(p-|G_*|)\exp\big(-d_0 n\zeta^{**2}_{\lambda,\gamma}\big)$.

Corollary 2 *Assume that* (C1) *and* (C2) *hold. If* $n\lambda^2 \to \infty$, $\lambda = o(m_*)$ *and* $n\gamma^2 \to \infty$ *then*

$$\mathbf{P}(\mathbf{\Omega}_{\lambda,\gamma}((\alpha_0+1)|A_*|) = \{\hat{\boldsymbol{\beta}}^o(\gamma)\}) \to 1$$

provided that $\log p = o(n\lambda^2)$, $\log|G_*| = o(n\gamma^2)$ *and* $\gamma = o(\lambda/\sqrt{|A_*|})$ *as* $n \to \infty$.

Remark 3 Theorem 2 and Corollary 2 allow $|G_*|$ to have an exponential order of n, which is impossible for other existing L_2-norm composite penalties such as the L_2-norm MCP. For example, suppose there exists a constant $m_0 > 0$ such that $m_* \geq m_0 > 0$ for all n. Then, by letting $\lambda = n^{-1/4}$ and $\gamma = n^{-1/3}$, Corollary 2 holds when $|G_*| = \exp(n^{\delta_0})$ for some constant $0 < \delta_0 < 1/3$.

Remark 4 Under similar regularity conditions, Bickel et al. (2009) and Zhang and Zhang (2012) proved that

$$\|\hat{\boldsymbol{\beta}}^o(\gamma) - \boldsymbol{\beta}^*\|_2^2 = O_p\big(|A_*|(\log|G_*|)/n\big) \text{ and } \|\hat{\boldsymbol{\beta}}^o(\gamma)\|_0 = O_p(|A_*|).$$

It is easy to see that the convergence rate is faster than the upper bound $O_p(|A_*|(\log p)/n)$ of the LASSO but the variable selection bound is the same. Since the DS penalized estimator is asymptotically equivalent to the oracle LASSO, we can say that it improves the LASSO by incorporating group information. In addition, a similar argument can be applied to the group LASSO. For a given $0 < \eta < 1$, Huang and Zhang (2010) proved that the group LASSO, $\hat{\boldsymbol{\beta}}^g$, satisfies

$$\|\hat{\boldsymbol{\beta}}^g - \boldsymbol{\beta}^*\|_2^2 = O_p\big(\{|G_*| + |\mathcal{A}(\boldsymbol{\beta}^*)|\log(|\mathcal{A}(\boldsymbol{\beta}^*)|/\eta)\}/n\big) \text{ and } \|\hat{\boldsymbol{\beta}}^g\|_0 = O_p(|G_*|)$$

under the $(|\mathcal{A}(\boldsymbol{\beta}^*)|, |G_*|)$-strong group sparsity. This shows that the group LASSO is inferior to the oracle LASSO in both bounds.

Remark 5 When $p \leq n$, it is easy to check whether a given solution $\hat{\boldsymbol{\beta}}$ satisfies the conditions in Lemma 3, that is, we can check whether $\hat{\boldsymbol{\beta}}$ obtained by the (or any) algorithm is unique or not. When $p > n$, we need to check the conditions in Lemma 3 by using $\kappa_{\min}$ in (C2). However, it is impossible to check the conditions in general since $\kappa_{\min}$ is computationally unavailable, and α_0 and $|A_*|$ are unknown. If we assume that α_0 and $|A_*|$ are known, then we can conclude that the solution $\hat{\boldsymbol{\beta}}$ is unique or not in $\boldsymbol{\Omega}_{\lambda,\gamma}((\alpha_0+1)|A_*|)$ by using a rough estimate of $\kappa_{\min}$ obtained by the way in Section 5 of Kim and Kwon (2012).

4 Numerical studies

4.1 Simulation study

The performance of the proposed DS method is examined using Monte–Carlo simulation. Throughout the simulation study, we use the following model:

$$y = \sum_{k=1}^{K}\sum_{j=1}^{p_k} x_{kj}\beta_{kj}^* + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2),$$

where $\sigma = 2$. Similar to Huang et al. (2009), we set the covariate vector as follows:

$$x_{kj} = (z_k + w_{kj})/\sqrt{2},\ k \leq K, j \leq p_k.$$

where $\mathbf{z} = (z_1, \ldots, z_K)^{\mathrm{T}} \sim N_K(\mathbf{0}, \boldsymbol{\Sigma})$ with $\Sigma_{k_1 k_2} = 0.4^{|k_1 - k_2|}$, $k_1, k_2 \leq K$, and $\mathbf{w} = (w_{11}, \ldots, w_{Kp_K})^{\mathrm{T}} \sim N_p(\mathbf{0}, \mathbf{I})$, independently of $\mathbf{z}$. For the true nonzero coefficients, we set $\beta^*_{kj} = c/(k \times j)$, $j \leq q_k$, $k \in \mathcal{A}(\boldsymbol{\beta}^*)$, where the constant $c > 0$ is chosen so that the signal-to-noise ratio is 10. The sizes of groups are fixed such that $p_{2k} = 12$ and $p_{2k-1} = 6$ for $k \leq K/2$, and the first 6 groups are set to be signal groups, that is, $\mathcal{A}(\boldsymbol{\beta}^*) = \{1, \ldots, 6\}$. We consider two settings with different degrees of within group sparsity: $q_k = p_k/3$ for Example 1 and $q_k = 2p_k/3$ for Example 2. All settings are replicated 400 times with $K \in \{18, 100\}$ and $n \in \{200, 400\}$. Hence the total number of variables is $p \in \{162, 900\}$ and the number of signal variables is $|A_*| \in \{18, 36\}$.

We compared the DS method with the LASSO, MCP, group LASSO (gLASSO), group Bridge (gBridge), group MCP (gMCP), group exponential LASSO (gExp). For comparison, we also considered five different versions of DS method: by fixing γ as $\gamma = 2\lambda^*$ $\gamma = \lambda^*$, $\gamma = \lambda^*/2$ and $\gamma = 0$, where λ^* is the optimal value of λ in the LASSO and by choosing an optimal γ^* over a sequence of girds. The `R` package `glmnet`, `grpreg` and `grppenalty` were used for to implement other estimators. All the tuning parameters in each method were chosen by using an independent validation data set of size $n/2$.

Tables 1 and 2 display the squared Prediction Error (PE) that is calculated from independent test data set of size $2n$, and the Model Error (ME) calculated as $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}} E(\mathbf{X}\mathbf{X}^{\mathrm{T}})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ (Fan and Li 2001). The tables also report the number of groups correctly and incorrectly selected (G.C and G.IC) and the number of variables correctly and incorrectly selected (V.C and V.IC). Methods are considered to select a group if at least one coefficient in the group was estimated as non-zero.

First, in Example 1, the LASSO outperforms the gLASSO in terms of PE and ME although it dose not employ the group structure, and so does the MCP. As expected, G.IC and V.IC of the LASSO and gLASSO are noticeably large. This indicates that ignoring group structure results in including many noisy groups and variables, and becoming inconsistent in group and variable selection. However, the gLASSO often performs better than the MCP in Example 2 where the group includes many signal variables, which is analogous to the results of Huang and Zhang (2010). The gLASSO and MCP are opposite in terms of individual variable selection while the MCP has small V.IC and large V.C, the gLASSO shows large V.IC and small V.C.

Second, in terms of prediction, the DS methods except the DS with $\gamma = 2\lambda^*$ have smaller PE and ME than other methods for all cases. This shows that a practical choice of $\gamma \in \{\lambda^*, \lambda^*/2\}$ performs well, which significantly reduces computational cost. The gBridge often shows the best PE and ME when $K = 18$ but has large PE and ME than DS methods when $K = 100$. We found that the algorithm implemented in R is often unstable when $K = 100$. The gMCP performs the worst in PE and ME except the case when $n = 400$, for which the gExp performs the worst.

Third, in group selection, the DS methods show the largest G.C and G.IC and the gBridge has the smallest G.C and G.IC. This implies that the DS methods are denser than other methods and the gBridge is the most sparse in group selection. The gMCP has the smallest G.C especially when $n = 100$ which may a reason for large PE and ME, and gExp shows the smallest G.IC. In individual variable selection, the gMCP and gExp show much larger V.IC than the gBridge and DS methods, which shows

Table 1 Simulation results of Example 1

(K, n)	Method	PE	ME	G.C	G.IC	V.C	V.IC
(18,200)	Oracle	4.425 (0.017)	0.403 (0.015)	6.0 (0.000)	0.0 (0.000)	18.0 (0.000)	0.0 (0.000)
	LASSO	4.937 (0.026)	0.905 (0.025)	6.0 (0.000)	7.1 (0.270)	15.1 (0.120)	24.0 (1.000)
	MCP	5.305 (0.045)	1.246 (0.045)	5.9 (0.010)	2.7 (0.190)	10.0 (0.150)	4.2 (0.320)
	gLASSO	5.576 (0.038)	1.546 (0.037)	6.0 (0.000)	7.3 (0.250)	18.0 (0.000)	103.4 (2.240)
	gBridge	5.080 (0.019)	1.060 (0.013)	5.7 (0.050)	0.3 (0.090)	13.5 (0.110)	12.0 (0.420)
	gMCP	5.658 (0.050)	1.611 (0.050)	5.5 (0.060)	2.2 (0.190)	14.9 (0.230)	34.3 (1.230)
	gExp	5.203 (0.046)	1.153 (0.044)	5.8 (0.040)	0.8 (0.130)	15.0 (0.160)	21.0 (0.790)
	DS3$_{(\gamma=2\lambda^*)}$	5.350 (0.045)	1.266 (0.044)	5.9 (0.010)	1.0 (0.170)	14.3 (0.130)	10.0 (0.420)
	DS$_{(\gamma=\lambda^*)}$	4.863 (0.026)	0.838 (0.025)	6.0 (0.000)	2.9 (0.270)	15.0 (0.120)	15.1 (0.670)
	DS$_{(\gamma=\lambda^*/2)}$	4.808 (0.027)	0.795 (0.026)	5.9 (0.010)	3.4 (0.230)	14.9 (0.130)	15.1 (0.510)
	DS$_{(\gamma=0)}$	4.923 (0.033)	0.892 (0.032)	5.9 (0.020)	2.9 (0.220)	14.7 (0.130)	13.4 (0.430)
	DS$_{(\gamma=\gamma^*)}$	4.803 (0.026)	0.785 (0.025)	5.9 (0.010)	3.4 (0.250)	15.0 (0.130)	15.7 (0.620)
(18,400)	Oracle	4.198 (0.012)	0.193 (0.008)	6.0 (0.000)	0.0 (0.000)	18.0 (0.000)	0.0 (0.000)
	LASSO	4.415 (0.015)	0.402 (0.012)	6.0 (0.000)	7.3 (0.280)	16.4 (0.110)	22.7 (0.830)
	MCP	4.524 (0.022)	0.484 (0.020)	6.0 (0.000)	3.0 (0.210)	12.6 (0.150)	4.1 (0.290)
	gLASSO	4.754 (0.018)	0.767 (0.015)	6.0 (0.000)	7.8 (0.260)	18.0 (0.000)	107.8 (2.400)
	gBridge	4.365 (0.015)	0.344 (0.012)	6.0 (0.000)	0.2 (0.090)	16.5 (0.110)	14.6 (0.810)
	gMCP	4.668 (0.019)	0.676 (0.016)	5.9 (0.010)	1.7 (0.170)	17.1 (0.120)	36.8 (0.790)
	gExp	4.497 (0.017)	0.489 (0.014)	5.9 (0.010)	0.6 (0.120)	16.5 (0.110)	22.2 (0.780)
	DS3$_{(\gamma=2\lambda^*)}$	4.614 (0.022)	0.581 (0.019)	6.0 (0.000)	0.4 (0.090)	16.0 (0.120)	8.4 (0.290)
	DS$_{(\gamma=\lambda^*)}$	4.384 (0.015)	0.377 (0.011)	6.0 (0.000)	2.6 (0.300)	16.3 (0.110)	13.8 (0.600)
	DS$_{(\gamma=\lambda^*/2)}$	4.354 (0.014)	0.344 (0.010)	6.0 (0.000)	2.9 (0.250)	16.3 (0.120)	14.0 (0.510)
	DS$_{(\gamma=0)}$	4.387 (0.015)	0.381 (0.011)	6.0 (0.000)	2.6 (0.220)	16.2 (0.110)	13.4 (0.500)
	DS$_{(\gamma=\gamma^*)}$	4.363 (0.014)	0.353 (0.010)	6.0 (0.000)	3.3 (0.280)	16.3 (0.120)	14.8 (0.610)
(100,200)	Oracle	4.440 (0.019)	0.407 (0.016)	6.0 (0.000)	0.0 (0.000)	18.0 (0.000)	0.00 (0.000)
	LASSO	5.169 (0.035)	1.098 (0.033)	6.0 (0.000)	19.4 (1.200)	14.5 (0.130)	31.9 (1.700)
	MCP	5.540 (0.041)	1.482 (0.041)	5.9 (0.030)	4.3 (0.350)	8.6 (0.110)	4.9 (0.370)
	gLASSO	5.887 (0.044)	1.846 (0.043)	5.9 (0.010)	17.2 (0.810)	17.9 (0.040)	185.5 (7.470)
	gBridge	5.687 (0.048)	1.729 (0.046)	5.2 (0.060)	0.0 (0.020)	11.7 (0.170)	10.1 (0.290)
	gMCP	6.041 (0.066)	1.959 (0.066)	5.0 (0.070)	5.9 (0.770)	13.1 (0.270)	39.8 (5.590)
	gExp	5.184 (0.045)	1.142 (0.043)	5.7 (0.040)	2.1 (0.300)	14.5 (0.150)	19.0 (0.900)
	DS3$_{(\gamma=2\lambda^*)}$	5.994 (0.064)	1.912 (0.063)	5.9 (0.010)	1.1 (0.270)	13.3 (0.160)	7.7 (0.460)
	DS$_{(\gamma=\lambda^*)}$	5.005 (0.030)	0.983 (0.029)	5.9 (0.010)	5.6 (0.570)	14.3 (0.130)	14.2 (0.750)
	DS$_{(\gamma=\lambda^*/2)}$	4.879 (0.029)	0.835 (0.028)	5.9 (0.030)	8.0 (0.660)	14.2 (0.150)	16.9 (0.810)
	DS$_{(\gamma=0)}$	4.989 (0.034)	0.927 (0.033)	5.9 (0.030)	7.2 (0.620)	14.0 (0.160)	15.6 (0.760)
	DS$_{(\gamma=\gamma^*)}$	4.887 (0.029)	0.819 (0.028)	5.9 (0.020)	8.2 (0.660)	14.2 (0.140)	17.3 (0.850)
(100,400)	Oracle	4.221 (0.012)	0.196 (0.007)	6.0 (0.000)	0.0 (0.000)	18.0 (0.000)	0.0 (0.000)
	LASSO	4.554 (0.017)	0.544 (0.014)	6.0 (0.000)	21.5 (1.160)	16.2 (0.090)	36.1 (1.690)
	MCP	4.644 (0.021)	0.618 (0.020)	6.0 (0.000)	5.6 (0.370)	11.4 (0.120)	6.2 (0.390)
	gLASSO	4.922 (0.022)	0.914 (0.019)	6.0 (0.000)	18.3 (0.710)	18.0 (0.000)	193.4 (6.530)

Table 1 continued

(K, n)	Method	PE	ME	G.C	G.IC	V.C	V.IC
	gBridge	4.796 (0.018)	0.663 (0.015)	5.8 (0.030)	0.0 (0.000)	14.4 (0.100)	11.9 (0.270)
	gMCP	4.505 (0.026)	0.484 (0.024)	5.9 (0.040)	2.3 (0.430)	16.5 (0.180)	23.2 (1.030)
	gExp	4.797 (0.018)	0.781 (0.013)	5.8 (0.000)	5.3 (0.310)	16.0 (0.080)	37.5 (0.780)
	DS3$_{(\gamma=2\lambda^*)}$	5.090 (0.030)	1.025 (0.028)	6.0 (0.000)	0.2 (0.070)	15.5 (0.100)	7.3 (0.260)
	DS$_{(\gamma=\lambda^*)}$	4.492 (0.017)	0.466 (0.014)	6.0 (0.000)	2.8 (0.540)	16.1 (0.090)	11.9 (0.740)
	DS$_{(\gamma=\lambda^*/2)}$	4.388 (0.015)	0.373 (0.012)	6.0 (0.000)	6.7 (0.610)	16.1 (0.090)	16.7 (0.810)
	DS$_{(\gamma=0)}$	4.425 (0.016)	0.400 (0.012)	6.0 (0.000)	8.7 (0.660)	15.9 (0.090)	19.2 (0.860)
	DS$_{(\gamma=\gamma^*)}$	4.389 (0.015)	0.382 (0.011)	6.0 (0.000)	8.0 (0.660)	16.1 (0.090)	18.2 (0.840)

The numbers in parentheses are corresponding standard errors

Table 2 Simulation results of Example 2

(K, n)	Method	PE	ME	G.C	G.IC	V.C	V.IC
(18,200)	Oracle	4.897 (0.027)	0.872 (0.026)	6.0 (0.000)	0.0 (0.000)	36.0 (0.000)	0.0 (0.000)
	LASSO	5.048 (0.027)	1.009 (0.026)	6.0 (0.000)	7.9 (0.260)	25.5 (0.210)	21.1 (0.980)
	MCP	6.107 (0.060)	1.999 (0.059)	5.9 (0.020)	3.6 (0.260)	12.8 (0.210)	5.5 (0.470)
	gLASSO	5.486 (0.037)	1.450 (0.035)	6.0 (0.000)	7.0 (0.260)	36.0 (0.000)	81.9 (2.330)
	gBridge	4.964 (0.036)	0.913 (0.035)	5.7 (0.040)	0.4 (0.090)	25.7 (0.250)	9.6 (0.620)
	gMCP	5.669 (0.044)	1.599 (0.044)	5.4 (0.070)	2.3 (0.190)	28.9 (0.550)	20.9 (0.950)
	gExp	5.219 (0.045)	1.151 (0.044)	5.8 (0.040)	1.1 (0.150)	26.9 (0.360)	12.0 (0.540)
	DS3$_{(\gamma=2\lambda^*)}$	5.415 (0.048)	1.306 (0.046)	6.0 (0.000)	1.2 (0.170)	23.8 (0.220)	6.8 (0.340)
	DS$_{(\gamma=\lambda^*)}$	4.954 (0.027)	0.904 (0.025)	5.9 (0.010)	3.0 (0.280)	25.1 (0.230)	10.5 (0.530)
	DS$_{(\gamma=\lambda^*/2)}$	4.906 (0.027)	0.873 (0.025)	5.9 (0.010)	3.4 (0.230)	24.9 (0.230)	10.7 (0.440)
	DS$_{(\gamma=0)}$	5.014 (0.031)	0.982 (0.031)	5.9 (0.020)	3.2 (0.220)	24.3 (0.220)	9.9 (0.380)
	DS$_{(\gamma=\gamma^*)}$	4.897 (0.026)	0.865 (0.024)	5.9 (0.010)	3.4 (0.250)	24.8 (0.230)	10.9 (0.510)
(18,400)	Oracle	4.408 (0.015)	0.414 (0.011)	6.0 (0.000)	0.0 (0.000)	36.0 (0.000)	0.0 (0.000)
	LASSO	4.501 (0.015)	0.488 (0.012)	6.0 (0.000)	8.0 (0.270)	28.3 (0.210)	21.1 (0.860)
	MCP	4.857 (0.023)	0.840 (0.020)	6.0 (0.000)	3.7 (0.200)	16.9 (0.220)	4.9 (0.280)
	gLASSO	4.729 (0.018)	0.746 (0.014)	6.0 (0.000)	7.3 (0.280)	36.0 (0.000)	84.9 (2.650)
	gBridge	4.434 (0.015)	0.418 (0.011)	5.9 (0.010)	0.3 (0.060)	27.1 (0.210)	9.9 (0.420)
	gMCP	4.699 (0.024)	0.683 (0.021)	5.9 (0.020)	1.7 (0.160)	33.2 (0.310)	20.1 (0.700)
	gExp	4.524 (0.018)	0.517 (0.015)	5.9 (0.010)	0.7 (0.130)	30.1 (0.270)	12.2 (0.470)
	DS3$_{(\gamma=2\lambda^*)}$	4.680 (0.023)	0.664 (0.020)	6.0 (0.000)	0.7 (0.110)	27.1 (0.210)	6.8 (0.270)
	DS$_{(\gamma=\lambda^*)}$	4.461 (0.015)	0.451 (0.011)	6.0 (0.000)	2.5 (0.280)	28.0 (0.210)	10.2 (0.480)
	DS$_{(\gamma=\lambda^*/2)}$	4.428 (0.015)	0.402 (0.011)	6.0 (0.000)	2.9 (0.250)	27.9 (0.220)	10.5 (0.450)
	DS$_{(\gamma=0)}$	4.451 (0.015)	0.432 (0.011)	6.0 (0.000)	3.0 (0.220)	27.8 (0.220)	10.2 (0.410)
	DS$_{(\gamma=\gamma^*)}$	4.433 (0.014)	0.419 (0.011)	6.0 (0.000)	3.1 (0.250)	27.9 (0.210)	10.7 (0.460)
(100,200)	Oracle	4.978 (0.026)	0.968 (0.024)	6.0 (0.000)	0.0 (0.000)	36.0 (0.000)	0.0 (0.000)
	LASSO	5.334 (0.034)	1.268 (0.032)	6.0 (0.000)	20.3 (1.250)	23.7 (0.270)	30.1 (1.860)
	MCP	6.659 (0.065)	2.498 (0.064)	5.8 (0.030)	6.0 (0.370)	10.1 (0.140)	6.6 (0.410)

Table 2 continued

(K, n)	Method	PE	ME	G.C	G.IC	V.C	V.IC
	gLASSO	5.782 (0.041)	1.728 (0.040)	5.9 (0.010)	16.3 (0.800)	35.8 (0.110)	158.8 (7.510)
	gBridge	5.544 (0.037)	1.810 (0.034)	5.3 (0.050)	0.1 (0.010)	20.3 (0.310)	5.7 (0.190)
	gMCP	5.993 (0.050)	2.031 (0.049)	4.9 (0.080)	4.7 (0.410)	24.7 (0.530)	19.1 (0.910)
	gExp	5.233 (0.044)	1.180 (0.043)	5.7 (0.040)	2.5 (0.320)	25.5 (0.300)	12.2 (0.670)
	DS3$_{(\gamma=2\lambda^*)}$	6.152 (0.062)	2.202 (0.061)	5.9 (0.010)	1.5 (0.400)	21.2 (0.260)	5.9 (0.550)
	DS$_{(\gamma=\lambda^*)}$	5.149 (0.030)	1.119 (0.028)	5.9 (0.010)	5.7 (0.590)	23.2 (0.260)	11.6 (0.780)
	DS$_{(\gamma=\lambda^*/2)}$	5.006 (0.028)	0.961 (0.027)	5.9 (0.020)	8.0 (0.590)	22.9 (0.260)	14.0 (0.730)
	DS$_{(\gamma=0)}$	5.140 (0.038)	1.084 (0.037)	5.8 (0.030)	7.8 (0.640)	22.3 (0.270)	13.5 (0.780)
	DS$_{(\gamma=\gamma^*)}$	5.020 (0.029)	1.009 (0.027)	5.9 (0.020)	8.1 (0.630)	22.9 (0.260)	14.1 (0.780)
(100,400)	Oracle	4.421 (0.014)	0.390 (0.010)	6.0 (0.000)	0.0 (0.000)	36.0 (0.000)	0.0 (0.000)
	LASSO	4.656 (0.017)	0.639 (0.015)	6.0 (0.000)	22.5 (1.160)	28.0 (0.190)	33.7 (1.710)
	MCP	5.212 (0.027)	1.200 (0.026)	6.0 (0.000)	6.6 (0.380)	13.9 (0.140)	7.2 (0.390)
	gLASSO	4.878 (0.021)	0.882 (0.018)	6.0 (0.000)	17.2 (0.680)	36.0 (0.000)	165.8 (6.190)
	gBridge	5.028 (0.018)	0.746 (0.015)	5.5 (0.030)	0.0 (0.000)	21.4 (0.190)	7.1 (0.190)
	gMCP	4.529 (0.027)	0.515 (0.024)	5.9 (0.040)	2.9 (0.420)	29.9 (0.350)	14.3 (0.890)
	gExp	4.794 (0.017)	0.789 (0.013)	5.7 (0.010)	5.3 (0.340)	31.0 (0.220)	22.9 (0.590)
	DS3$_{(\gamma=2\lambda^*)}$	5.200 (0.030)	1.196 (0.028)	6.0 (0.000)	0.2 (0.090)	26.0 (0.190)	5.3 (0.240)
	DS$_{(\gamma=\lambda^*)}$	4.582 (0.017)	0.572 (0.013)	6.0 (0.000)	3.3 (0.450)	27.6 (0.190)	9.5 (0.630)
	DS$_{(\gamma=\lambda^*/2)}$	4.469 (0.015)	0.469 (0.011)	6.0 (0.000)	6.8 (0.550)	27.6 (0.190)	13.7 (0.710)
	DS$_{(\gamma=0)}$	4.491 (0.017)	0.485 (0.012)	6.0 (0.000)	8.1 (0.620)	27.3 (0.180)	14.9 (0.780)
	DS$_{(\gamma=\gamma^*)}$	4.461 (0.015)	0.456 (0.011)	6.0 (0.000)	7.1 (0.610)	27.6 (0.190)	14.0 (0.770)

The numbers in parentheses are corresponding standard errors

they select groups well but fail to exclude irrelevant variables in the selected groups. The gBridge often shows the best performance when $K = 18$ but too sparse when $K = 100$. On the other hand, the DS methods has slightly smaller V.C than others but much smaller V.IC for all cases, which shows the DS methods perform better than other methods in variable selection.

To sum up, the DS methods and gBridge compete with each other, showing the best performance in all measures, and in detail, the gBridge is the best when $K = 18$ but the DS methods is the best when $K = 100$. As a practical choice of γ in the DS methods, we recommend to use a fixed constant for reducing computational cost, such as $\gamma \in \{\lambda^*, \lambda^*/2\}$ in the simulation. Too large γ such as $\gamma = 2\lambda^*$ does not work well and thus we recommend to use $\gamma = \lambda^*/2$ in practice.

4.2 Ways of tuning parameter selection

The performance of penalized estimation depends on the choice of tuning parameters. We construct several theorems and corollaries that can be used to develop some guide lines for choosing tuning parameters of the DS. For example, motivated by the universal regularization parameter proposed by Zhang (2010), we may set

$\lambda = \sigma\{(2/n)\log p\}^{1/2}$ when ε is a Gaussian random variable with mean 0 and known variance σ^2. This works as a method of group selection since the DS method is group selection consistent. However, such a choice may performs bad in practice unless the sample size is sufficiently large. Hence, we need to develop practical ways of selecting tuning parameters that produce reasonable finite sample performance.

For this issue, we consider four different ways of tuning parameter selection. The first method is to use independent validation sets as in our simulation studies assuming that the whole sample size is moderately large. For comparison, we consider two ways where the sample size of validation sets are $4n$ and $n/2$, which are denoted by V_{4n} and $\mathrm{V}_{n/2}$. The second method is the k-fold cross validation and we consider $k = 5$ that is denoted by CV_5. The use of validation sets and cross validation are based on prediction accuracy hence often produce slightly overfitted models (Wang et al. 2007, 2009), but the most practical ways of tuning parameter selection. The other method is the high-dimensional BIC criterion as in Wang et al. (2013) when σ^2 is unknown: $\mathrm{BIC}_W = \log(\hat{\sigma}^2) + \{C_n(\log p)/n\}\hat{M}$, where $\hat{\sigma}^2 = \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2/n$, C_n is a sequence diverges slowly, for example we use $C_n = \log(\log n)$ in this study, and $\hat{M}$ is the number of variables included in the model.

Tables 3 and 4 present the simulation results, where the simulation settings are the same as those used for Tables 1 and 2, and we only consider the case when $K = 100$ and $n = 400$ for comparison. It is easy to see that the models based on the V_{4n} always show the smallest PE and ME for all the methods, which is reasonable since the error is a straightforward estimate of the test error. The BIC_W produces the worst results having slightly larger PE than the other criteria but much larger ME. However, in group and individual variable selection, all the methods have much small G.IC and V.IC when the BIC_W is used while G.C and V.C are slightly less than other criteria. Hence, the BIC_W seems to be the best in group and variable selection although it produces higher PE and ME, while other criteria show similar performance producing denser models than the BIC_W. We note that, among the methods, the DS methods have the smallest PE and ME when $\gamma \in \{\lambda^*/2, \gamma^*\}$ for almost all cases, regardless of the criteria, and all the methods are quite invariant to the choice of tuning parameter selection when they are based on the prediction. The $\mathrm{V}_{n/2}$ and CV_5 show the second and third performance in PE and ME but the difference is not substantial compared with the V_{4n}. This shows that the $\mathrm{V}_{n/2}$ and CV_5 are simple but practical criteria. Further, we note that the $\mathrm{V}_{2/n}$ and CV_5 perform similar to each other for all the methods, which indicates that one may use small validation sets instead of the cross validation to reduce computational cost. To sum up, we recommend to use the BIC_W for selection and $\mathrm{V}_{n/2}$ or CV_5 for prediction, and the DS methods based on these criteria perform quite well with $\gamma \in \{\lambda^*/2, \lambda^*\}$.

4.3 Real data analysis

We analyze two real data sets:

- Ozone data: The Ozone data are popular real data set, which has been analyzed in the literatures including Breiman and Friedman (1985) and Lin and Zhang (2006). The data set is available from the R library 'mlbench' including short descriptions

Table 3 Comparison of tuning parameter selection methods in Example 1

Method	PE				ME			
	BIC_W	V_{4n}	$\mathrm{V}_{n/2}$	CV_5	BIC_W	V_{4n}	$\mathrm{V}_{n/2}$	CV_5
gLASSO	7.161	4.915	4.922	4.955	2.636	0.901	0.914	0.950
gBridge	5.164	4.694	4.796	4.762	0.597	0.462	0.663	0.785
gMCP	5.021	4.484	4.505	4.525	0.951	0.459	0.484	0.492
gExp	6.129	4.747	4.797	4.918	2.159	0.735	0.781	0.872
$\mathrm{DS3}_{(\gamma=2\lambda^*)}$	5.304	5.088	5.090	5.093	1.223	1.025	1.025	1.047
$\mathrm{DS}_{(\gamma=\lambda^*)}$	4.774	4.485	4.492	4.498	0.675	0.465	0.466	0.486
$\mathrm{DS}_{(\gamma=\lambda^*/2)}$	4.729	4.375	4.388	4.398	0.624	0.367	0.373	0.371
$\mathrm{DS}_{(\gamma=0)}$	4.681	4.407	4.425	4.436	0.617	0.387	0.400	0.408
$\mathrm{DS}_{(\gamma=\gamma^*)}$	4.700	4.367	4.389	4.409	0.618	0.360	0.382	0.376
Method	G.C				G.IC			
	BIC_W	V_{4n}	$\mathrm{V}_{n/2}$	CV_5	BIC_W	V_{4n}	$\mathrm{V}_{n/2}$	CV_5
gLASSO	4.89	6.00	6.00	6.00	0.00	19.54	18.32	16.12
gBridge	5.43	5.93	5.93	5.39	0.00	0.00	0.00	0.12
gMCP	5.72	5.99	5.99	5.99	0.00	1.74	2.35	0.79
gExp	4.29	5.84	5.80	5.62	0.10	5.54	5.38	3.44
$\mathrm{DS3}_{(\gamma=2\lambda^*)}$	5.93	6.00	6.00	6.00	0.00	0.29	0.29	0.21
$\mathrm{DS}_{(\gamma=\lambda^*)}$	5.96	6.00	6.00	6.00	0.01	2.91	2.83	2.11
$\mathrm{DS}_{(\gamma=\lambda^*/2)}$	5.95	6.00	6.00	6.00	0.03	6.33	6.78	4.86
$\mathrm{DS}_{(\gamma=0)}$	5.98	6.00	6.00	6.00	0.07	7.60	8.75	5.33
$\mathrm{DS}_{(\gamma=\gamma^*)}$	5.96	6.00	6.00	6.00	0.05	6.68	8.07	4.90
Method	V.C				V.IC			
	BIC_W	V_{4n}	$\mathrm{V}_{n/2}$	CV_5	BIC_W	V_{4n}	$\mathrm{V}_{n/2}$	CV_5
gLASSO	13.60	18.00	18.00	18.00	27.20	204.00	193.56	174.54
gBridge	14.71	16.39	14.39	15.05	8.37	11.88	11.85	10.80
gMCP	14.32	16.47	16.47	16.37	8.70	21.81	23.19	19.38
gExp	10.05	16.22	16.04	15.36	14.49	37.77	37.47	34.31
$\mathrm{DS3}_{(\gamma=2\lambda^*)}$	14.36	15.48	15.49	15.43	5.55	7.37	7.29	7.18
$\mathrm{DS}_{(\gamma=\lambda^*)}$	14.93	16.12	16.09	16.06	6.87	12.10	11.98	11.35
$\mathrm{DS}_{(\gamma=\lambda^*/2)}$	14.86	16.10	16.07	16.09	6.86	16.22	16.68	14.85
$\mathrm{DS}_{(\gamma=0)}$	15.09	15.97	15.97	15.96	6.99	17.81	19.20	15.04
$\mathrm{DS}_{(\gamma=\gamma^*)}$	14.93	16.07	16.05	16.08	6.73	16.67	18.19	14.88

of 3 categorical and 9 continuous independent variables for the dependent variable ‘daily maximum one hour average ozone reading’. We drop the variable ‘temperature of measured at El Monte, CA’ since it has too many missing values and exclude 36 observations including missing values. We keep all the categorical variables except ‘day of week’.

Table 4 Comparison of tuning parameter selection methods in Example 2

Method	PE				ME			
	BIC$_W$	V$_{4n}$	V$_{n/2}$	CV$_5$	BIC$_W$	V$_{4n}$	V$_{n/2}$	CV$_5$
gLASSO	7.054	4.870	4.878	4.906	2.551	0.869	0.882	0.917
gBridge	5.004	4.476	5.028	4.967	0.915	0.645	0.746	0.707
gMCP	5.140	4.510	4.529	4.550	1.009	0.504	0.515	0.535
gExp	6.074	4.769	4.794	4.951	1.880	0.749	0.789	0.888
DS3$_{(\gamma=2\lambda^*)}$	5.648	5.199	5.200	5.210	1.499	1.196	1.196	1.193
DS$_{(\gamma=\lambda^*)}$	5.047	4.574	4.582	4.585	0.965	0.565	0.572	0.574
DS$_{(\gamma=\lambda^*/2)}$	4.964	4.456	4.469	4.473	0.872	0.451	0.469	0.466
DS$_{(\gamma=0)}$	4.933	4.479	4.491	4.503	0.872	0.467	0.485	0.493
DS$_{(\gamma=\gamma^*)}$	4.946	4.445	4.461	4.478	0.880	0.449	0.456	0.480
Method	G.C				G.IC			
	BIC$_W$	V$_{4n}$	V$_{n/2}$	CV$_5$	BIC$_W$	V$_{4n}$	V$_{n/2}$	CV$_5$
gLASSO	4.81	6.00	6.00	6.00	0.00	18.50	17.28	15.60
gBridge	5.14	5.89	5.49	5.54	0.00	0.00	0.00	2.36
gMCP	5.62	6.00	5.99	6.00	0.00	2.19	2.88	1.14
gExp	4.29	5.80	5.74	5.48	0.16	5.90	5.34	3.66
DS3$_{(\gamma=2\lambda^*)}$	5.78	6.00	6.00	5.99	0.00	0.23	0.23	0.15
DS$_{(\gamma=\lambda^*)}$	5.90	6.00	6.00	6.00	0.01	2.83	3.31	2.24
DS$_{(\gamma=\lambda^*/2)}$	5.92	6.00	6.00	6.00	0.02	6.10	6.82	5.07
DS$_{(\gamma=0)}$	5.94	6.00	6.00	6.00	0.03	7.62	8.13	5.57
DS$_{(\gamma=\gamma^*)}$	5.93	6.00	6.00	6.00	0.02	6.32	7.14	5.23
Method	V.C				V.IC			
	BIC$_W$	V$_{4n}$	V$_{n/2}$	CV$_5$	BIC$_W$	V$_{4n}$	V$_{n/2}$	CV$_5$
gLASSO	26.56	36.00	36.00	36.00	13.28	176.64	165.84	151.98
gBridge	24.11	28.40	21.39	26.99	5.09	7.07	7.06	20.88
gMCP	23.93	29.56	29.96	29.12	5.24	13.06	14.25	11.24
gExp	18.58	31.41	31.03	29.82	8.11	23.84	22.89	19.93
DS3$_{(\gamma=2\lambda^*)}$	22.47	26.01	26.04	25.95	3.89	5.24	5.29	5.24
DS$_{(\gamma=\lambda^*)}$	23.91	27.67	27.64	27.62	4.56	8.96	9.53	8.42
DS$_{(\gamma=\lambda^*/2)}$	24.16	27.68	27.59	27.61	4.56	12.89	13.71	11.86
DS$_{(\gamma=0)}$	24.20	27.34	27.27	27.23	4.56	14.47	14.94	12.21
DS$_{(\gamma=\gamma^*)}$	23.96	27.64	27.55	27.54	4.46	13.07	14.02	12.00

- TRIM data: We use the data set in Scheetz et al. (2006), which consists of gene expression levels of $18,975$ genes obtained from 120 rats. The main objective of the analysis is to find genes that are correlated with the TRIM32 gene, known to cause Bardet–Biedl syndrome. As was done Huang et al. (2008), we first select 3000 genes with the largest variance in expression levels and then choose top 100 genes that have the largest absolute correlation with TRIM32 among the selected 3000 genes.

Consider a linear regression model with K predictive variables for the analysis:

$$y = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{\alpha} + \varepsilon, \tag{15}$$

where $\boldsymbol{w} = (w_1, \ldots, w_K)^{\mathrm{T}}$ and $\boldsymbol{\alpha} \in \mathbb{R}^K$ is a regression coefficient. A problem of using the linear model is that the relation of w_k to y may not be linear, in particular when the empirical distribution of w_k is highly skewed. An alternative approach is to use an additive model. For the kth component w_k, let $-\infty < c_{k0} < c_{k1} < \cdots < c_{kp_k} < \infty$ be a given increasing sequence of p_k real numbers such that the support of w_k is covered in $[c_{k0}, c_{kp_k}]$. Then we have an additive model with piecewise linear bases:

$$y = \beta_0 + \sum_{k=1}^{K}\sum_{j=1}^{p_k} \beta_{kj} x_{kj}(w_k) + \varepsilon, \tag{16}$$

where $x_{k1}(w_k) = w_k$ and $x_{kj}(w_k) = (w_k - c_{k(j-1)})_+$ for $2 \le j \le p_k$. The DS penalty can delete unnecessary components and at the same time delete unnecessary bases inside each component, which is new for the additive model.

We consider the estimators in the simulation studies for the additive model (16), and then add two more estimators for comparison, the LASSO and MCP for the linear model (15), denoted by LASSO$_0$ and MCP$_0$. We only consider the DS methods with $\gamma \in \{\lambda^*, \lambda^*/2\}$ which perform well as shown in the simulations, and compare two ways of selecting tuning parameters, the BIC$_W$ and CV$_5$. We let p_k be either 5 or 9, and c_{kj} are selected so that the numbers of observations in $(c_{k(j-1)}, c_{kj}]$ are roughly equal, and then we transform all the continuous variables into p_k new variables.

Tables 5 and 6 present averages of prediction errors (PE), numbers of selected groups (#G) and variables (#V), based on 100 random partitions: training (70 %) and test (30 %). First of all, the estimated models from the additive model show smaller PE than those from the linear model for all cases. The models selected by the BIC$_W$ tend to include less groups and variables than CV$_5$, which may cause low prediction accuracies for all the methods. We notice that the gLASSO shows the largest PE and #V for all cases, and selects the null model for TRIM data when the BIC$_W$ is used.

Among the group and variables selection methods, the DS methods have the smallest PE and the gBridge is the second best for both cases. The DS methods produce denser models than the others for TRIM data, and the difference is much clearer when we use CV$_5$. However, for Ozone data, the gMCP and gExp selects much more variables than the gBridge and DS methods especially when we use CV$_5$. For example, the gMCP and gExp include about 7 and 5 variables in each selected group, while the DS methods include 2 variables only, keeping better PE. For the reference, we give partial fits of the 9 selected variables by the DS with $\gamma = \lambda^*/2$ based on the CV$_5$ for Ozone data in Fig. 2. Further, we note that the DS methods are less sensitive to the choice of tuning parameter selection method as observed in the simulation studies than the other methods. Based on these results, we conclude that the additive model is a useful alternative to the linear model, and the DS penalty is well suited for this model.

Table 5 Results for Ozone data

p_k	Method	BIC$_W$			CV$_5$		
		PE	#G	#V	PE	#G	#V
5	LASSO$_0$	20.867 (0.249)	5.7 (0.101)	5.9 (0.133)	20.341 (0.225)	7.2 (0.064)	8.6 (0.159)
	MCP$_0$	21.008 (0.257)	3.7 (0.098)	3.7 (0.111)	20.812 (0.256)	4.9 (0.161)	5.3 (0.228)
	LASSO	17.103 (0.229)	8.8 (0.150)	16.2 (0.415)	15.697 (0.211)	10.8 (0.038)	27.5 (0.455)
	MCP	17.664 (0.245)	7.9 (0.170)	11.5 (0.275)	17.265 (0.214)	9.4 (0.145)	16.3 (0.571)
	gLASSO	21.139 (0.463)	4.8 (0.203)	24.6 (1.065)	16.143 (0.201)	9.5 (0.096)	49.7 (0.549)
	gBridge	16.973 (0.234)	5.3 (0.103)	12.5 (0.254)	16.363 (0.226)	6.7 (0.116)	16.3 (0.332)
	gMCP	20.033 (0.253)	3.2 (0.101)	10.8 (0.543)	16.916 (0.230)	6.7 (0.132)	32.4 (0.627)
	gExp	17.320 (0.254)	5.4 (0.094)	15.1 (0.336)	16.569 (0.204)	6.3 (0.120)	19.7 (0.486)
	DS$_{(\gamma=\lambda^*)}$	16.636 (0.230)	6.8 (0.137)	12.1 (0.223)	15.843 (0.212)	8.8 (0.161)	18.0 (0.496)
	DS$_{(\gamma=\lambda^*/2)}$	16.807 (0.237)	7.0 (0.142)	12.3 (0.242)	16.068 (0.224)	8.6 (0.155)	17.4 (0.549)
9	LASSO$_0$	20.867 (0.249)	5.7 (0.101)	5.9 (0.133)	20.341 (0.225)	7.2 (0.064)	8.6 (0.159)
	MCP$_0$	21.008 (0.257)	3.7 (0.098)	3.7 (0.111)	20.812 (0.256)	4.9 (0.161)	5.3 (0.228)
	LASSO	17.387 (0.230)	8.4 (0.168)	15.5 (0.480)	16.080 (0.230)	10.7 (0.050)	27.0 (0.813)
	MCP	17.880 (0.267)	7.8 (0.156)	11.6 (0.268)	17.631 (0.297)	8.8 (0.187)	14.9 (0.582)
	gLASSO	25.087 (0.610)	3.8 (0.159)	32.4 (1.362)	17.257 (0.240)	8.6 (0.134)	74.0 (1.148)
	gBridge	16.798 (0.251)	5.4 (0.091)	13.4 (0.257)	16.641 (0.241)	5.9 (0.109)	15.3 (0.382)
	gMCP	22.413 (0.296)	1.5 (0.116)	5.2 (0.661)	18.357 (0.268)	5.0 (0.143)	35.9 (1.282)
	gExp	19.112 (0.275)	4.1 (0.130)	12.2 (0.662)	17.583 (0.255)	5.7 (0.117)	26.2 (0.991)
	DS$_{(\gamma=\lambda^*)}$	16.844 (0.243)	6.8 (0.134)	11.9 (0.262)	16.151 (0.231)	8.6 (0.155)	18.1 (0.595)
	DS$_{(\gamma=\lambda^*/2)}$	16.981 (0.243)	7.1 (0.137)	12.4 (0.271)	16.362 (0.246)	8.3 (0.158)	16.7 (0.565)

The numbers in parentheses are corresponding standard errors

Table 6 Results for TRIM data

p_k	Method	BIC			CV$_5$		
		PE	#G	#V	PE	#G	#V
5	LASSO$_0$	0.577 (0.015)	0.6 (0.117)	0.6 (0.117)	0.368 (0.008)	12.6 (0.363)	12.6 (0.363)
	MCP$_0$	0.442 (0.011)	2.3 (0.100)	2.3 (0.100)	0.433 (0.010)	3.6 (0.163)	3.6 (0.163)
	LASSO	0.586 (0.014)	0.5 (0.119)	0.5 (0.119)	0.354 (0.009)	20.0 (0.570)	23.3 (0.744)
	MCP	0.553 (0.032)	6.5 (0.445)	6.7 (0.479)	0.435 (0.013)	5.2 (0.253)	5.2 (0.256)
	gLASSO	0.613 (0.013)	0.0 (0.000)	0.0 (0.000)	0.383 (0.008)	20.2 (0.610)	101.2 (3.050)
	gBridge	0.430 (0.012)	1.9 (0.107)	2.7 (0.168)	0.430 (0.151)	2.7 (0.128)	4.0 (0.211)
	gMCP	0.457 (0.012)	1.5 (0.086)	1.7 (0.119)	0.483 (0.015)	2.9 (0.247)	9.9 (1.335)
	gExp	0.445 (0.011)	2.8 (0.180)	3.4 (0.249)	0.413 (0.015)	6.1 (0.201)	8.6 (0.331)
	DS$_{(\gamma=\lambda^*)}$	0.420 (0.010)	3.3 (0.153)	3.3 (0.156)	0.356 (0.008)	11.8 (0.535)	12.3 (0.596)
	DS$_{(\gamma=\lambda^*/2)}$	0.415 (0.011)	4.0 (0.233)	4.1 (0.235)	0.396 (0.010)	9.6 (0.769)	10.1 (0.880)
9	LASSO$_0$	0.577 (0.015)	0.6 (0.117)	0.6 (0.117)	0.368 (0.008)	12.6 (0.363)	12.6 (0.363)
	MCP$_0$	0.442 (0.011)	2.3 (0.100)	2.3 (0.100)	0.433 (0.010)	3.6 (0.163)	3.6 (0.163)
	LASSO	0.595 (0.014)	0.4 (0.096)	0.4 (0.096)	0.368 (0.010)	20.6 (0.587)	24.9 (0.825)

Table 6 continued

p_k	Method	BIC			CV_5		
		PE	#G	#V	PE	#G	#V
	MCP	0.560 (0.033)	5.2 (0.399)	5.3 (0.420)	0.434 (0.011)	5.4 (0.318)	5.4 (0.322)
	gLASSO	0.613 (0.013)	0.0 (0.000)	0.0 (0.000)	0.407 (0.010)	20.8 (0.801)	187.1 (7.204)
	gBridge	0.430 (0.010)	1.5 (0.077)	2.1 (0.129)	0.425 (0.011)	2.0 (0.104)	3.0 (0.194)
	gMCP	0.462 (0.012)	1.6 (0.087)	1.7 (0.105)	0.481 (0.020)	2.2 (0.136)	7.9 (1.097)
	gExp	0.464 (0.012)	2.2 (0.146)	2.5 (0.187)	0.435 (0.016)	5.3 (0.211)	8.2 (0.447)
	$\text{DS}_{(\gamma=\lambda^*)}$	0.426 (0.011)	4.1 (0.188)	4.3 (0.190)	0.357 (0.008)	12.8 (0.671)	13.5 (0.761)
	$\text{DS}_{(\gamma=\lambda^*/2)}$	0.405 (0.011)	4.9 (0.213)	5.1 (0.216)	0.379 (0.010)	11.6 (0.794)	12.6 (0.958)

The numbers in parentheses are corresponding standard errors

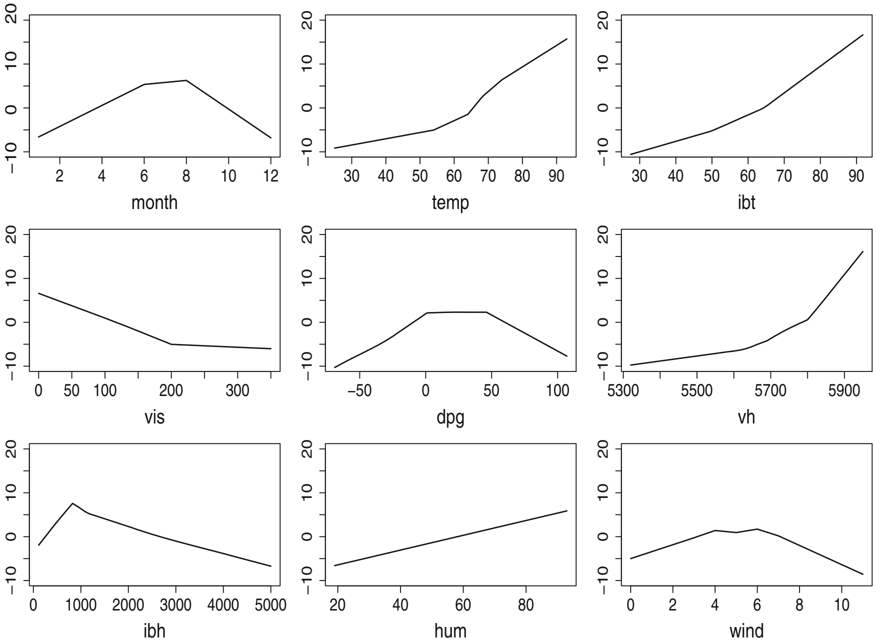

Fig. 2 Partial fits on selected variables of ozone data when $p_k = 9$: month (month), temperature measured (tem), inversion base temperature (ibt), visibility measured (vis), pressure gradient (dpg), pressure height (vh), inversion base height (ibh), humidity (hum), wind speed (wind)

5 Concluding remarks

There are two regularization parameters λ and γ in the DS penalty, and it may be computationally demanding to select them simultaneously. For this issue, we can consider an alternative: choosing λ by using the CV or BIC-type criteria after fixing γ. Although we did not pursue theoretical properties further in this paper, the DS method performs quite well according to these two steps once a γ is given appropriately. We recommend, for example, to use a γ around the optimal λ^* that is obtained from the LASSO in practice.

We have shown empirically that the additive model with sparse penalties is a promising alternative to linear models. However, the choice of the number of knots, p_k, should be done carefully. When p_k is too small, the model may not capture the functional relation properly. On the other hand, when p_k is too large, the corresponding design matrix may not be well posed. We leave the optimal choice of p_k as future work.

Acknowledgements We are grateful to the anonymous referees, the associate editor, and the editor for their helpful comments. The research of Sunghoon Kwon was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Science, ICT & Future Planning (No. 2014R1A1A1002995). The research of Woncheol Jang was supported by the Basic Science Research Program through the National Research Foundation (NRF) of Korea grant funded by the Ministry of Education (No. 2013R1A1A2010065) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. 2014R1A4A1007895). The research of Yongdai Kim was supported by the Basic Science Research Program through the National Research Foundation (NRF) of Korea grant funded by the Korea government (MSIP) (No. 2014R1A4A1007895).

Appendix

Without loss of generality, we assume that the covariates are standardized so that $\mathbf{X}_{kj}^{\mathrm{T}}\mathbf{X}_{kj}/n = 1$ for all $k \leq K$ and $j \leq p_k$. Further, we use $\hat{\boldsymbol{\beta}}^{o}$ instead of $\hat{\boldsymbol{\beta}}^{o}(\gamma)$ for simplicity.

Proof of Lemma 1. From the first order optimality conditions (Bertsekas 1999), the necessary conditions follow directly. □

Proof of Lemma 2. It suffices to show that there exists a $\delta > 0$ such that $Q_{\lambda,\gamma}(\boldsymbol{\beta}) \geq Q_{\lambda,\gamma}(\hat{\boldsymbol{\beta}})$ for all $\boldsymbol{\beta} \in B(\hat{\boldsymbol{\beta}}, \delta)$, where $B(\hat{\boldsymbol{\beta}}, \delta) = \{\boldsymbol{\beta} : \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_1 \leq \delta\}$. From the convexity of the sum of squared residuals, $Q_{\lambda,\gamma}(\boldsymbol{\beta}) - Q_{\lambda,\gamma}(\hat{\boldsymbol{\beta}}) \geq \sum_{k=1}^{K} \chi_k$, where

$$\chi_k = D_k(\hat{\boldsymbol{\beta}})^{\mathrm{T}}(\boldsymbol{\beta}_k - \hat{\boldsymbol{\beta}}_k) + J_{\lambda,\gamma}^{(k)}(\|\boldsymbol{\beta}_k\|) - J_{\lambda,\gamma}^{(k)}(\|\hat{\boldsymbol{\beta}}_k\|).$$

First, consider cases where $k \in \mathcal{L}(\hat{\boldsymbol{\beta}})$. Let $\delta_k = \|\hat{\boldsymbol{\beta}}_k\|_1 - ap_k(\lambda - \gamma)$ then $\|\boldsymbol{\beta}_k\|_1/p_k > a(\lambda - \gamma)$ for all $\boldsymbol{\beta}_k \in B(\hat{\boldsymbol{\beta}}_k, \delta_k)$. Hence, the first and second conditions in Lemma 1 imply

$$\chi_k \geq -\gamma(\|\boldsymbol{\beta}_k\|_1 - \|\hat{\boldsymbol{\beta}}_k\|_1) + \gamma(\|\boldsymbol{\beta}_k\|_1 - \|\hat{\boldsymbol{\beta}}_k\|_1) = 0.$$

Next, consider cases where $k \in \mathcal{N}(\hat{\boldsymbol{\beta}})$. Let $\delta_k = \min\{\omega_k, a(\lambda - \gamma)\}$, where $\omega_k = 2a(\lambda - \|D_k(\hat{\boldsymbol{\beta}})\|_\infty)$. Then $\|\boldsymbol{\beta}_k\|_1 < \omega_k$ for all $\boldsymbol{\beta}_k \in B(\hat{\boldsymbol{\beta}}_k, \delta_k)$, which implies

$$\chi_k \geq \big(-\|D_k(\hat{\boldsymbol{\beta}})\|_\infty - \|\boldsymbol{\beta}_k\|_1/2a + \lambda\big)\|\boldsymbol{\beta}_k\|_1 \geq 0.$$

Hence $Q_{\lambda,\gamma}(\boldsymbol{\beta}) \geq Q_{\lambda,\gamma}(\hat{\boldsymbol{\beta}})$ for all $\boldsymbol{\beta} \in B(\hat{\boldsymbol{\beta}}, \min_{k \leq K} \delta_k)$, which completes the proof. □

Proof of Lemma 3. Assume that there exists another local minimizer $\tilde{\boldsymbol{\beta}} \in \boldsymbol{\Omega}_{\lambda,\gamma}$ and $\tilde{\boldsymbol{\beta}} \neq \hat{\boldsymbol{\beta}}$. Let $\boldsymbol{\beta}^h = \tilde{\boldsymbol{\beta}} + h(\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{\beta}}) = h\hat{\boldsymbol{\beta}} + (1-h)\tilde{\boldsymbol{\beta}}$ for $0 < h < 1$, then we have

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}^h\|_2^2 - \|\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}}\|_2^2 = -2nhD(\hat{\boldsymbol{\beta}})^{\mathrm{T}}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}) + (h^2 - 2h)(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}),$$

by using the equality,

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 - \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|_2^2 = 2nD(\hat{\boldsymbol{\beta}})^{\mathrm{T}}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}) + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^{\mathrm{T}}\mathbf{X}^{\mathrm{T}}\mathbf{X}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}),$$

for any $\boldsymbol{\beta} \in \mathbb{R}^p$. Hence, it follows that

$$Q_{\lambda,\gamma}(\boldsymbol{\beta}^h) - Q_{\lambda,\gamma}(\tilde{\boldsymbol{\beta}}) \leq h\sum_{k=1}^{K}\chi_k(h) + h^2(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X}/n)(\tilde{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}})/2,$$

where

$$\chi_k(h) = -D_k(\hat{\boldsymbol{\beta}})^{\mathrm{T}}(\tilde{\boldsymbol{\beta}}_k - \hat{\boldsymbol{\beta}}_k) - \rho_{\min}\|\tilde{\boldsymbol{\beta}}_k - \hat{\boldsymbol{\beta}}_k\|_2^2 + \left\{J_{\lambda,\gamma}^{(k)}(\|\boldsymbol{\beta}_k^h\|_1) - J_{\lambda,\gamma}^{(k)}(\|\tilde{\boldsymbol{\beta}}_k\|_1)\right\}/h.$$

First, consider cases where $k \in \mathcal{L}(\hat{\boldsymbol{\beta}})$. If $k \in \mathcal{N}(\tilde{\boldsymbol{\beta}})$ then,

$$\begin{aligned}\chi_k(h) &= D_k(\hat{\boldsymbol{\beta}})^{\mathrm{T}}\hat{\boldsymbol{\beta}}_k - \rho_{\min}\|\hat{\boldsymbol{\beta}}_k\|_2^2 + J_{\lambda,\gamma}^{(k)}(h\|\hat{\boldsymbol{\beta}}_k\|_1)/h \\ &= -\sum_{\hat{\beta}_{kj}\neq 0}\gamma\,\mathrm{sign}(\hat{\beta}_{kj})\hat{\beta}_{kj} - \rho_{\min}\|\hat{\boldsymbol{\beta}}_k\|_2^2 + \lambda \\ &\leq \|\hat{\boldsymbol{\beta}}_k\|_1(-\rho_{\min}\|\hat{\boldsymbol{\beta}}_k\|_1/p_k + \lambda - \gamma) < 0,\end{aligned}$$

from the condition $\|\hat{\boldsymbol{\beta}}_k\|_1/p_k > (\lambda - \gamma)/\rho_{\min}$. If $k \in \mathcal{L}(\tilde{\boldsymbol{\beta}})$ then,

$$\begin{aligned}\chi_k(h) &< -D_k(\hat{\boldsymbol{\beta}})^{\mathrm{T}}(\tilde{\boldsymbol{\beta}}_k - \hat{\boldsymbol{\beta}}_k) + \left\{J_{\lambda,\gamma}^{(k)}(\|\boldsymbol{\beta}_k^h\|_1) - J_{\lambda,\gamma}^{(k)}(\|\tilde{\boldsymbol{\beta}}_k\|_1)\right\}/h \\ &\leq \gamma(\|\tilde{\boldsymbol{\beta}}_k\|_1 - \|\hat{\boldsymbol{\beta}}_k\|_1) + \gamma(\|\hat{\boldsymbol{\beta}}_k\|_1 - \|\tilde{\boldsymbol{\beta}}_k\|_1) = 0,\end{aligned}$$

unless $\tilde{\boldsymbol{\beta}}_k = \hat{\boldsymbol{\beta}}_k$. If $k \in \mathcal{S}(\tilde{\boldsymbol{\beta}})$, we have

$$\sup_{k\in\mathcal{S}(\tilde{\boldsymbol{\beta}})}\left\{\rho_{\min}\|\tilde{\boldsymbol{\beta}}_k\|_1/p_k + \nabla J_{\lambda,\gamma}(\|\tilde{\boldsymbol{\beta}}_k\|_1)\right\} \leq \max\left\{\lambda,\, \rho_{\min}a(\lambda - \gamma) + \gamma\right\},$$

which implies

$$\begin{aligned}\chi_k(h) &\leq \gamma(\|\tilde{\boldsymbol{\beta}}_k\|_1 - \|\hat{\boldsymbol{\beta}}_k\|_1) - \rho_{\min}\|\tilde{\boldsymbol{\beta}}_k - \hat{\boldsymbol{\beta}}_j\|_2^2 + \nabla J_{\lambda,\gamma}(\|\tilde{\boldsymbol{\beta}}_k\|_1)(\|\hat{\boldsymbol{\beta}}_k\|_1 - \|\tilde{\boldsymbol{\beta}}_k\|_1) \\ &\leq (\|\hat{\boldsymbol{\beta}}_k\|_1 - \|\tilde{\boldsymbol{\beta}}_k\|_1)\big\{-\gamma - \rho_{\min}(\|\hat{\boldsymbol{\beta}}_k\|_1 - \|\tilde{\boldsymbol{\beta}}_k\|_1)/p_k + \nabla J_{\lambda,\gamma}(\|\tilde{\boldsymbol{\beta}}_k\|_1)\big\} \\ &\leq (\|\hat{\boldsymbol{\beta}}_k\|_1 - \|\tilde{\boldsymbol{\beta}}_k\|_1)\big\{-\rho_{\min}\|\hat{\boldsymbol{\beta}}_k\|_1/p_k + \max\left\{\lambda - \gamma,\, \rho_{\min}a(\lambda - \gamma)\right\}\big\} < 0,\end{aligned}$$

unless $\|\tilde{\boldsymbol{\beta}}_k\|_1 = \|\hat{\boldsymbol{\beta}}_k\|_1$. Second, consider cases where $k \in \mathcal{N}(\hat{\boldsymbol{\beta}})$. It is easy to see that

$$\begin{aligned}\chi_k(h) &\leq \|D_k(\hat{\boldsymbol{\beta}})\|_\infty \|\tilde{\boldsymbol{\beta}}_k\|_1 - \rho_{\min}\|\tilde{\boldsymbol{\beta}}_k\|_2^2 + \left\{J_{\lambda,\gamma}^{(k)}(\|\boldsymbol{\beta}_k^h\|_1) - J_{\lambda,\gamma}^{(k)}(\|\tilde{\boldsymbol{\beta}}_k\|_1)\right\}/h \\ &\leq \|\tilde{\boldsymbol{\beta}}_k\|_1\left\{\|D_k(\hat{\boldsymbol{\beta}})\|_\infty - \rho_{\min}\|\tilde{\boldsymbol{\beta}}_k\|_1/p_k - \nabla J_{\lambda,\gamma}(\|\tilde{\boldsymbol{\beta}}_k\|_1)\right\} < 0,\end{aligned}$$

unless $\|\tilde{\boldsymbol{\beta}}_k\|_1 = 0$, since

$$\inf_{k \in \mathcal{A}(\tilde{\boldsymbol{\beta}})} \left\{\rho_{\min}\|\tilde{\boldsymbol{\beta}}_k\|_1/p_k + \nabla J_{\lambda,\gamma}(\|\tilde{\boldsymbol{\beta}}_k\|_1)\right\} \geq \min\left\{\lambda, a\rho_{\min}(\lambda-\gamma)+\gamma\right\}.$$

Hence, we finally have $\sum_{k=1}^K \chi_k(h) < 0$, unless $\|\tilde{\boldsymbol{\beta}}_k\|_1 = \|\hat{\boldsymbol{\beta}}_k\|_1$ for all $k \leq K$. This implies that there exists a $\delta > 0$ sufficiently small such that $Q_{\lambda,\gamma}(\boldsymbol{\beta}^h) - Q_{\lambda,\gamma}(\tilde{\boldsymbol{\beta}}) < 0$ for all $h \in (0, \delta)$ unless $\|\tilde{\boldsymbol{\beta}}_k\|_1 = \|\hat{\boldsymbol{\beta}}_k\|_1$ for all $k \leq K$. Hence, $\hat{\boldsymbol{\beta}}$ is the unique local minimizer. □

Proof of Lemma 1. Let $A^o = \{(k, j) : \hat{\beta}_{kj}^o \neq 0\}$. From Lemma 2, it suffices to show that $\mathbf{P}(E_1 \cap E_2 \cap E_3) \geq 1 - \mathbf{P}_1 - \mathbf{P}_2 - \mathbf{P}_3$, where

$$\begin{aligned}E_1 &= \left\{|A^o \cup A_*| \leq (\alpha_0+1)|A_*|\right\}, \\ E_2 &= \left\{\min_{k \in \mathcal{A}(\boldsymbol{\beta}^*)}\|\hat{\boldsymbol{\beta}}_k^o\|_1/p_k > a(\lambda-\gamma)\right\}, \\ E_3 &= \left\{\max_{k \in \mathcal{N}(\boldsymbol{\beta}^*)}\|D_k(\hat{\boldsymbol{\beta}})\|_\infty < \lambda\right\}.\end{aligned}$$

First consider the event E_1. From Corollary 2 of Zhang and Zhang (2012), we have $F \subset E_1$ provided that $\phi_{\max}(\alpha_0|A_*|)/\alpha_0 \leq \eta_{\min}/36$, where $F = \left\{\max_{k \in \mathcal{A}(\boldsymbol{\beta}^*)} \|\mathbf{X}_k^{\mathrm{T}}\boldsymbol{\varepsilon}/n\|_\infty \leq \gamma/2\right\}$ and

$$\eta_{\min} = \inf_{\boldsymbol{v} \in \mathbb{R}^{|G_*|}:\|\boldsymbol{v}_{A_*^c}\|_1 \leq 3\|\boldsymbol{v}_{A_*}\|_1} \left\{(|A_*|/n)\|\mathbf{X}_{G_*}^{\mathrm{T}}\mathbf{X}_{G_*}\boldsymbol{v}\|_\infty/\|\boldsymbol{v}\|_1\right\}$$

is the cone invertible factor in Ye and Zhang (2010). On the other hand, inequality (7) of Zhang and Zhang (2012) proves $\eta_{\min} \geq \delta_{\min}^2/16$, where

$$\delta_{\min} = \inf_{\boldsymbol{v} \in \mathbb{R}^{|G_*|}:\|\boldsymbol{v}_{A_*^c}\|_1 \leq 3\|\boldsymbol{v}_{A_*}\|_1} \left\{(1/\sqrt{n})\|\mathbf{X}_{G_*}\boldsymbol{v}\|_2/\|\boldsymbol{v}_{A_*}\|_2\right\}$$

is the restricted eigenvalue in Bickel et al. (2009) that satisfies $\delta_{\min} \geq \sqrt{\kappa_{\min}}(1 - 3\sqrt{\phi_{\max}(\alpha_0|A_*|)/\alpha_0\kappa_{\min}})$. Hence, (C2) implies that $F \subset E_1$ and

$$\begin{aligned}\mathbf{P}(E_1^c) \leq \mathbf{P}(F^c) &\leq \sum_{k \in \mathcal{A}(\boldsymbol{\beta}^*)} \sum_{j=1}^{p_k} \mathbf{P}\left(|\mathbf{X}_{kj}^{\mathrm{T}}\boldsymbol{\varepsilon}/n| > \gamma/2\right) \\ &\leq c_0|G_*|\exp(-d_0 n\gamma^2/4) = \mathbf{P}_1.\end{aligned}$$

Second, consider the event E_2. From the first order optimality conditions (Rosset and Zhu 2007), $\hat{\boldsymbol{\beta}}^o$ satisfies

$$\begin{array}{ll} \mathbf{X}_{kj}^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}^o)/n=\gamma\,\mathrm{sign}(\hat{\beta}_{kj}^o), & \hat{\beta}_{kj}^o\neq 0,\\ |\mathbf{X}_{kj}^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}^o)/n|\leq\gamma, & \hat{\beta}_{kj}^o=0, \end{array} \tag{17}$$

for all $(k,j)\in G_*$. Let $S=A^o\cup A_*$ and $\hat{\boldsymbol{\beta}}^o_S$ be the vector that consists of elements $\hat{\beta}^o_{kj}$ for $(k,j)\in S$. On the event E_1, (C2) implies

$$\hat{\boldsymbol{\beta}}^o_S-\boldsymbol{\beta}^*_S=\boldsymbol{\Sigma}_S^{-1}\{-\mathbf{X}_S^{\mathrm{T}}(\mathbf{y}-\mathbf{X}_S\hat{\boldsymbol{\beta}}^o_S)/n+\mathbf{X}_S^{\mathrm{T}}\boldsymbol{\varepsilon}/n\}, \tag{18}$$

where $\boldsymbol{\Sigma}_S=\mathbf{X}_S^{\mathrm{T}}\mathbf{X}_S/n$. Let $\mathbf{u}_{kj}$ be a vector of length $|S|\leq(\alpha_0+1)|A_*|$ whose unique nonzero element that corresponds to β^*_{kj} is 1 and the others are 0. Then, from (18), we can write

$$\hat{\beta}^o_{kj}-\beta^*_{kj}=\mathbf{u}_{kj}^{\mathrm{T}}(\hat{\boldsymbol{\beta}}^o_S-\boldsymbol{\beta}^*_S)=\eta_{kj}+\mathbf{v}_{kj}^{\mathrm{T}}\boldsymbol{\varepsilon},$$

where $\eta_{kj}=-\mathbf{u}_{kj}^{\mathrm{T}}\boldsymbol{\Sigma}_S^{-1}\mathbf{X}_S^{\mathrm{T}}(\mathbf{y}-\mathbf{X}_S\hat{\boldsymbol{\beta}}^o_S)/n$ and $\mathbf{v}_{kj}=\mathbf{X}_S\boldsymbol{\Sigma}_S^{-1}\mathbf{u}_{kj}/n$. Note that

$$|\eta_{kj}|\leq\|\mathbf{u}_{kj}\|_2\|\mathbf{X}_S^{\mathrm{T}}(\mathbf{y}-\mathbf{X}_S\hat{\boldsymbol{\beta}}^o_S)/n\|_2/\kappa_{\min}\leq\gamma\sqrt{|S|}/\kappa_{\min}$$

and

$$\|\mathbf{v}_{kj}\|_2^2=\mathbf{u}_{kj}^{\mathrm{T}}\boldsymbol{\Sigma}_S^{-1}\mathbf{X}_S^{\mathrm{T}}\mathbf{X}_S\boldsymbol{\Sigma}_S^{-1}\mathbf{u}_{kj}/n^2\leq 1/(n\kappa_{\min}).$$

From (C1), it is easy to see that

$$\begin{aligned}\mathbf{P}_{E_1}\big(|\hat{\beta}^o_{kj}-\beta^*_{kj}|\geq\|\boldsymbol{\beta}^*_k\|_1/p_k-a(\lambda-\gamma)\big)&\leq\mathbf{P}\big(|\eta_{kj}|+|\mathbf{v}_{kj}^{\mathrm{T}}\boldsymbol{\varepsilon}|\geq m_*-a(\lambda-\gamma)\big)\\&\leq\mathbf{P}\big(|\mathbf{v}_{kj}^{\mathrm{T}}\boldsymbol{\varepsilon}|\geq m_*-a(\lambda-\gamma)-\gamma\sqrt{|S|}/\kappa_{\min}\big)\\&\leq c_0\exp\big(-d_0\kappa_{\min}n\xi^{*2}_{\lambda,\gamma}\big),\end{aligned}$$

where $\mathbf{P}_{E_1}(A)=\mathbf{P}(E_1\cap A)$. Hence, by using the triangular inequality $\|\hat{\boldsymbol{\beta}}^o_k\|_1\geq\|\boldsymbol{\beta}^*_k\|_1-\|\hat{\boldsymbol{\beta}}^o_k-\boldsymbol{\beta}^*_k\|_1$, we have

$$\begin{aligned}\mathbf{P}_{E_1}\big(E_2^c\big)&\leq\sum_{k\in\mathcal{A}(\boldsymbol{\beta}^*)}\mathbf{P}_{E_1}\big(\|\hat{\boldsymbol{\beta}}^o_k-\boldsymbol{\beta}^*_k\|_1/p_k\geq\|\boldsymbol{\beta}^*_k\|_1/p_k-a(\lambda-\gamma)\big)\\&\leq\sum_{(k,j)\in S}\mathbf{P}\big(|\hat{\beta}^o_{kj}-\beta^*_{kj}|\geq\|\boldsymbol{\beta}^*_k\|_1/p_k-a(\lambda-\gamma)\big)\\&\leq c_0(\alpha_0+1)|A_*|\exp\big(-d_0\kappa_{\min}n\xi^{*2}_{\lambda,\gamma}\big)=\mathbf{P}_2.\end{aligned} \tag{19}$$

Third, consider the event E_3. From (18), we can write

$$\mathbf{X}_{kj}^{\mathrm{T}}(\mathbf{y}-\mathbf{X}_S\hat{\boldsymbol{\beta}}_S^o)/n = \mathbf{X}_{kj}^{\mathrm{T}}(\mathbf{X}_S\boldsymbol{\beta}_S^* - \mathbf{X}_S\hat{\boldsymbol{\beta}}_S^o + \boldsymbol{\varepsilon})/n = \zeta_{kj} + \mathbf{w}_{kj}^{\mathrm{T}}\boldsymbol{\varepsilon},$$

for all $(k, j) \in S$, where $\zeta_{kj} = \mathbf{X}_{kj}^{\mathrm{T}}\mathbf{X}_S\boldsymbol{\Sigma}_S^{-1}\mathbf{X}_S^{\mathrm{T}}(\mathbf{y}-\mathbf{X}_S\hat{\boldsymbol{\beta}}_S^o)/n^2$, $\mathbf{w}_{kj} = (\mathbf{I}-\boldsymbol{\Pi}_S)\mathbf{X}_{kj}/n$ and $\boldsymbol{\Pi}_S = \mathbf{X}_S(\mathbf{X}_S^{\mathrm{T}}\mathbf{X}_S)^{-1}\mathbf{X}_S^{\mathrm{T}}$. Note that from (17),

$$\begin{aligned}
|\zeta_{kj}| &= |\mathbf{X}_{kj}^{\mathrm{T}}\mathbf{X}_S\boldsymbol{\Sigma}_S^{-1}\mathbf{X}_S^{\mathrm{T}}(\mathbf{y}-\mathbf{X}_S\hat{\boldsymbol{\beta}}_S^o)/n^2| \\
&\le \|\boldsymbol{\Sigma}_S^{-1/2}\mathbf{X}_S^{\mathrm{T}}\mathbf{X}_{kj}/n\|_2\|\boldsymbol{\Sigma}_S^{-1/2}\mathbf{X}_S^{\mathrm{T}}(\mathbf{y}-\mathbf{X}_S\hat{\boldsymbol{\beta}}_S^o)/n\|_2 \\
&\le \|\boldsymbol{\Pi}_S\mathbf{X}_{kj}/\sqrt{n}\|_2\|\boldsymbol{\Sigma}_S^{-1/2}\mathbf{X}_S^{\mathrm{T}}(\mathbf{y}-\mathbf{X}_S\hat{\boldsymbol{\beta}}_S^o)/n\|_2 \\
&\le \gamma\sqrt{|S|/\kappa_{\min}}
\end{aligned}$$

and $\|\mathbf{w}_{kj}\|_2^2 = \mathbf{X}_{kj}^{\mathrm{T}}(\mathbf{I}-\boldsymbol{\Pi}_S)\mathbf{X}_{kj}/n^2 \le \|\mathbf{X}_{kj}\|_2^2/n^2 = 1/n$. Hence, from (C1),

$$\begin{aligned}
\mathbf{P}_{E_1}\big(E_3^c\big) &\le \sum_{k\in\mathcal{N}(\boldsymbol{\beta}^*)}\sum_{j=1}^{p_k}\mathbf{P}\big(|\mathbf{X}_{kj}^{\mathrm{T}}(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}^o)/n| \ge \lambda\big) \\
&\le \sum_{k\in\mathcal{N}(\boldsymbol{\beta}^*)}\sum_{j=1}^{p_k}\mathbf{P}\big(|\mathbf{w}_{kj}^{\mathrm{T}}\boldsymbol{\varepsilon}| \ge \lambda - \gamma\sqrt{|S|}/\sqrt{\kappa_{\min}}\big) \\
&\le c_0(p-|G_*|)\exp\big(-d_0 n\zeta_{\lambda,\gamma}^{*2}\big) = \mathbf{P}_3. \qquad (20)
\end{aligned}$$

Hence, using $\mathbf{P}(E_1\cap E_2\cap E_3) \ge 1-\mathbf{P}(E_1^c)-\mathbf{P}(E_1\cap E_2^c)-\mathbf{P}(E_1\cap E_3^c)$, we complete the proof. □

Proof of Lemma 2. Suppose that there is another local minimizer $\tilde{\boldsymbol{\beta}} \in \boldsymbol{\Omega}_{\lambda,\gamma}((\alpha_0+1)|A_*|)$ such that $\hat{\boldsymbol{\beta}}^o \ne \tilde{\boldsymbol{\beta}}$. Let $S = \{(k, j) : \tilde{\beta}_{kj} \ne 0\} \cup A^o \cup A_*$. By replacing $\mathbf{X}$ with $\mathbf{X}_S$ in the proof of Lemma 3, we can see that if $\hat{\boldsymbol{\beta}}^o$ satisfies conditions in Lemma 3 then $\hat{\boldsymbol{\beta}}^o = \tilde{\boldsymbol{\beta}}$. Since $|S| \le 2(\alpha_0+1)|A_*|$, we have $\lambda_{\min}(\mathbf{X}_S^{\mathrm{T}}\mathbf{X}_S/n) \ge \kappa_{\min}$ from (C2). Hence it suffices to show that

$$\mathbf{P}_{E_1}\Big(\min_{k\in\mathcal{A}(\boldsymbol{\beta}^*)}\|\hat{\boldsymbol{\beta}}_k^o\|_1/p_k \le \max\{a, 1/\kappa_{\min}\}(\lambda-\gamma)\Big) \le \mathbf{P}_2$$

$$\mathbf{P}_{E_1}\big(\max_{k\in\mathcal{N}(\boldsymbol{\beta}^*)}\|D_k(\hat{\boldsymbol{\beta}}^o)\|_\infty < \min\big\{\lambda, a\kappa_{\min}(\lambda-\gamma)+\gamma\big\}\big) \le \mathbf{P}_3,$$

which is similar to proofs of (19) and (20) in the proof of Theorem 1. □

References

An, L. T. H., Tao, P. D. (1997). Solving a class of linearly constrained indefinite quadratic problems by DC algorithms. *Journal of Global Optimization*, *11*, 253–285.

Bertsekas, D. P. (1999). *Nonlinear Proramming* (2nd ed.). Belmont: Athena Scientific.

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

Breheny, P. (2015). The group exponential lasso for bi-level variable selection. *Biometrics*, *71*, 731–740.

Breheny, P., Huang, J. (2009). Penalized methods for bi-level variable selection. *Statistics and Its Interface*, *2*, 369–380.

Breiman, L., Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, *80*, 580–598.

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

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

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

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

Huang, J., Zhang, T. (2010). The benefit of group sparsity. *The Annals of Statistics*, *38*, 1978–2004.

Huang, J., Horowitz, J. L., Ma, S. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *The Annals of Statistics*, *36*, 587–613.

Huang, J., Ma, S., Xie, H., Zhang, C.-H. (2009). A group bridge approach for variable selection. *Biometrika*, *96*, 339–355.

Huang, J., Breheny, P., Ma, S. (2012). A selective review of group selection in high-dimensional models. *Statistical Science*, *27*, 481–499.

Jiang, D., Huang, J. (2015). Concave 1-norm troup selection. *Biostatistics*, *16*, 252–267.

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

Kim, Y., Choi, H., Oh, H. (2008). Smoothly clipped absolute deviation on high dimensions. *Journal of the American Statistical Association*, *103*, 1656–1673.

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

Lin, Y., Zhang, H. H. (2006). Component selection and smoothing in smoothing spline analysis of variance models. *The Annals of Statistics*, *34*, 2272–2297.

Meinshausen, N., Yu, B. (2009). Lasso-type recovery of sparse representation for high-dimensional data. *The Annals of Statistics*, *37*, 246–270.

Rosset, S., Zhu, J. (2007). Piecewise linear regularized solution paths. *The Annals of Statistics*, *35*, 1012–1030.

Sardy, S., Tseng, P. (2004). Amlet, ramlet, and gamlet: Automatic nonlinear fitting of additive models, robust and generalized, with wavelets. *Journal of Computational and Graphical Statistics*, *13*, 283–309.

Scheetz, T. E., Kim, K. Y. A., Swiderski, R. E., Philp, A. R., Braun, T. A., Knudtson, K. L., Dorrance, A. M., DiBona, G. F., Huang, J., Casavant, T. L., Sheffield, V. C., Stone, E. M. (2006). Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences* (Vol. 103, pp. 14429–14434).

Simon, N., Friedman, J., Hastie, T., Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*, *22*, 231–245.

Sriperumbudur, B. K., & Lanckriet, G. R. (2009). On the convergence of the concave-convex procedure. *Advances in Neural Information Processing Systems*, *9*, 1759–1767.

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

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

Wang, H., Li, B., Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of Royal Statistical Society Series B*, *71*, 671–683.

Wang, L., Kim, Y., Li, R. (2013). Calibrating non-convex penalized regression in ultra-high dimension. *The Annals of Statistics*, *41*, 2505–2536.

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

Ye, F., Zhang, C.-H. (2010). Rate minimaxity of the lasso and dantzig selector for the ℓ_q loss in ℓ_r balls. *Journal of Machine Learning Research*, *11*, 3519–3540.

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

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

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

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

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

Zhou, N., Zhu, J. (2010). Group variable selection via a hierarchical lasso and its oracle property. *Statistics and Its Interface*, *3*, 557–574.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*, 1418–1429.