A Group Bridge Approach for Variable Selection

Jian Huang[1], Shuangge Ma[2], Huiliang Xie[3], and Cun-Hui Zhang[4]

[1,3]Department of Statistics and Actuarial Science, and [1]Department of Biostatistics, University of Iowa, Iowa City, Iowa 52242, USA

[2]Division of Biostatistics, Department of Epidemiology and Public Health,Yale University, New Haven, Connecticut 06520, USA

[4]Department of Statistics, Rutgers University, Piscataway, New Jersey 08854, USA

April 2007

The University of Iowa

Department of Statistics and Actuarial Science

Technical Report No. 376

Abstract

In multiple regression problems when covariates can be naturally grouped, it is important to carry out feature selection at the group and within-group individual variable levels simultaneously. The existing methods, including the lasso and group lasso, are designed for either variable selection or group selection, but not for both. We propose a group bridge approach that it is capable of simultaneous selection at both the group and within-group individual variable levels. The proposed approach is a penalized regularization method that uses a specially designed group bridge penalty. It has the powerful oracle group selection property, that is, it can correctly select important groups with probability converging to one. In contrast, the group lasso method in general does not possess such an oracle property in group selection. Simulation studies indicate that the group bridge has superior performance in group and individual variable selection than the group lasso in a wide range of generating models.

Key Words. Bridge estimators, iterative lasso, penalized regression, two-level selection, variable-selection consistency.

Short title. Group bridge regression

Address for correspondence:
Jian Huang, Department of Statistics and Actuarial Science, 241 SH, University of Iowa, Iowa City, Iowa 52242, USA

Email: jian-huang@uiowa.edu

1 Introduction

Consider the linear regression model

$$Y_i = X_{i1}\beta_1 + \cdots + X_{id}\beta_d + \varepsilon_i,\ i = 1, \ldots, n, \tag{1}$$

where $Y_i \in \mathbb{R}$ is the response variable, $X_{i1}, \ldots, X_{id}$ are covariate variables, β_j's are regression coefficients, and ε_i's are error terms. Assumed that the covariates can be naturally grouped. We are interested in simultaneously selecting important groups as well as important individual variables within the selected groups. We propose a group bridge method for simultaneous feature selection at both the group and within-group individual variable levels, which cannot be realized using existing variable selection methods.

Variable selection is a classic problem in statistics. The literature on this topic is too vast to be summarized here. Traditional approaches for variable selection include the C_p (Mallows, 1973), AIC (Akaika, 1973), and BIC (Schwartz 1978). More recently, several penalized regularization methods have been proposed for variable selection. Examples include the bridge (Frank and Friedman 1996), LASSO (Tishirani 1996), SCAD (Fan and Li, 2001; Fan and Peng, 2004), and Enet (Zou and Hastie 2005), among others. These methods are designed for selecting individual variables.

The need to select groups of variables arises in multifactor analysis of variance (ANOVA) and nonparametric additive regression. In ANOVA, a factor with multiple levels can be represented by a group of dummy variables. In nonparametric additive regression, each component can be expressed as a linear combination of a set of basis functions. In both cases, the selection of important factors or nonparametric components amounts to the selection of groups of variables. Several recent papers have considered selecting important groups of variables using penalized methods. Yuan and Lin (2006) proposed the group lasso method. This method is a natural extension of the lasso, in which an L_2 norm of the coefficients associated with a group of variables is used as a unit in the penalty function. The group lasso method was extended to general loss functions by Kim, Kim, and Kim (2006). They used the same penalty as the group lasso penalty and called the extension the blockwise sparse regression (BSR). Zhao et al. (2006) proposed a composite absolute penalty (CAP) for group selection, which can be considered a generalization of the group

lasso. These studies only considered group selection, but did not address the question of individual selection within groups. Ma and Huang (2007) proposed a clustering threshold gradient descent regularization (CTGDR) method that selects variables at both the group and individual variable levels. However the CTGDR does not optimize a well-defined objective function, thus it is difficult to study its theoretical properties.

In many problems, it is important to be able to carry out feature selection at the group and within-group individual variable levels simultaneously. In regression models, for a group of variables, even when the group as a whole is important, the effects of some variables in this group may not be important. It is desirable to select the important ones from this selected group. In nonparametric additive modeling, it is often the case that a saturated set of basis functions is used. In addition to the component selection, it is useful to select the basis functions to achieve a more sparse representation of the selected component. As a specific example, consider the Impact study that was designed to determine the effects of different risk factors on body mass index (BMI) of high school students in two Seattle public schools. Table 2 shows the variables that were collected in this study. These variables can be naturally divided into eight groups. It is of interest to know which groups have a significant impact on the BMI as well as the variables in these groups that are important. For example, if food consumption has a significant effect, it is of great interest to know which food consumptions have significant impacts and which do not.

The proposed group bridge method is the first penalized regularization method that is capable of two-level selection. As it is shown in Section 3, this method has the powerful oracle selection property, that is, it can correctly select important groups with probability converging to one. In contrast, the group lasso method does not possess such an oracle property in group selection. The simulation studies reported in Section 4 show that the group bridge has superior performance in group and individual variable selection than the group lasso in a wide range of generating models.

2 The group bridge estimator

Let $\mathbf{X}_k = (X_{1k}, \ldots, X_{nk})'$, $k = 1, \ldots, d$, be the design vectors and $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ be the response vector in (1) so that the linear model is written as

$$\mathbf{Y} = \mathbf{X}_1\beta_1 + \cdots + \mathbf{X}_d\beta_d + \boldsymbol{\varepsilon} \tag{2}$$

with an error vector $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)'$. Let $A_1, \ldots, A_J$ be subsets of $\{1, \ldots, d\}$ representing known groupings of the design vectors. Denote the regression coefficients in the j-th group as $\boldsymbol{\beta}_{A_j} = (\beta_k, k \in A_j)'$. For any $m \times 1$ vector $\mathbf{a}$, denote its L_1 norm $\|\mathbf{a}\|_1 = |a_1| + \cdots + |a_m|$. We consider the objective function

$$L_n(\beta) = \Big\|\mathbf{Y} - \sum_{k=1}^{d} \mathbf{X}_k \beta_k\Big\|_2^2 + \lambda_n \sum_{j=1}^{J} c_j \|\boldsymbol{\beta}_{A_j}\|_1^{\gamma}, \tag{3}$$

where $\lambda_n > 0$ is the penalty level and c_j are constants for the adjustment of the different dimensions of $\boldsymbol{\beta}_{A_j}$. A simple choice is $c_j \propto |A_j|^{1-\gamma}$, where $|A_j|$ is the cardinality of A_j. In (3), the bridge penalty is applied on the L_1 norms of the grouped coefficients. Therefore, we call the $\widehat{\boldsymbol{\beta}}_n$ that minimizes (3) a group bridge estimator. Here the groups A_j are allowed to overlap and their union is allowed to be a proper subset of the whole so that variables not in $\cup_{j=1}^{J} A_j$ are not penalized. When $|A_j| = 1, 1 \le j \le J$, (3) simplifies to the standard bridge criterion. As will be explained below, when $0 < \gamma < 1$, the group bridge criterion (3) can be used for variable selection at the group and individual variable levels simultaneously.

2.1 Computation

Direct minimization of $L_n(\boldsymbol{\beta})$ is difficult, since the group bridge penalty is not a convex function for $0 < \gamma < 1$. We formulate an equivalent minimization problem that is easier to solve computationally. For $0 < \gamma < 1$, define

$$S_{1n}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \Big\|\mathbf{Y} - \sum_{k=1}^{d} \mathbf{X}_k \beta_k\Big\|_2^2 + \sum_{j=1}^{J} \theta_j^{1-1/\gamma} c_j^{1/\gamma} \|\boldsymbol{\beta}_{A_j}\|_1 + \tau \sum_{j=1}^{J} \theta_j, \tag{4}$$

where τ is a penalty parameter.

Proposition 1. *Suppose* $0 < \gamma < 1$. *If* $\lambda_n = \tau^{1-\gamma}\gamma^{-\gamma}(1-\gamma)^{\gamma-1}$, *then* $\widehat{\boldsymbol{\beta}}_n$ *minimizes* $L_n(\boldsymbol{\beta})$ *if and only if* $(\widehat{\boldsymbol{\beta}}_n, \widehat{\boldsymbol{\theta}})$ *solves*

$$\textit{minimize } \; S_{1n}(\boldsymbol{\beta}, \boldsymbol{\theta}) \; \textit{ subject to } \; \boldsymbol{\theta} \ge 0,$$

for some $\widehat{\boldsymbol{\theta}} \ge 0$, *where* $\boldsymbol{\theta} \ge 0$ *means* $\theta_j \ge 0, j = 1, \ldots, J$.

This proposition is similar to the characterization of the component selection and smoothing

method of Lin and Zhang (2006). Examining the form of S_{1n} in (4), we see that the minimization of S_{1n} with respect to $(\boldsymbol{\beta}, \theta)$ yields sparse solutions at the group and individual variable levels. Specifically, the penalty is an adaptively weighted L_1 penalty, so the solution is sparse in β. On the other hand, for $0 < \gamma < 1$, small θ_j will force $\beta_{A_j} = 0$, which leads to group selection.

Based on Proposition 1, we propose the following iterative algorithm.

Step 1. Obtain an initial estimate $\boldsymbol{\beta}^{(0)}$.

For $s = 1, 2, \ldots$,

Step 2. Compute

$$\theta_j^{(s)} = c_j \Big(\frac{1-\gamma}{\tau\gamma} \Big)^{\gamma} \|\boldsymbol{\beta}_{A_j}^{(s-1)}\|_1^{\gamma}, \quad j = 1, \ldots, J. \tag{5}$$

Step 3. Compute

$$\boldsymbol{\beta}^{(s)} = \arg\min_{\boldsymbol{\beta}} \Big\| \mathbf{Y} - \sum_{k=1}^{d} \mathbf{X}_k \boldsymbol{\beta}_k \Big\|_2^2 + \sum_{j=1}^{J} (\theta_j^{(s)})^{1-1/\gamma} c_j^{1/\gamma} \|\boldsymbol{\beta}_{A_j}\|_1. \tag{6}$$

Step 4. Repeat steps 2-3 until convergence.

This algorithm always converges, since at each step it decreases the objective function (4), which is nonnegative. The main computational task is step 3, which is a lasso problem and can be solved efficiently using the Lars algorithm (Efron et al. 2004). In general, this algorithm converges to a local minimizer depending on the initialization $\boldsymbol{\beta}^{(0)}$, since the group bridge penalty is not convex. In this article, we focus on full rank designs, where the unbiased least squares estimator is a natural initial estimator.

2.2 Tuning parameter selection

For a fixed λ_n, let $\widehat{\boldsymbol{\beta}}_n = \widehat{\boldsymbol{\beta}}_n(\lambda_n)$ be the group bridge estimate of $\boldsymbol{\beta}$. Let $\widehat{\theta}_{nj}$, $j = 1, \ldots, J$, be the jth component of $\widehat{\boldsymbol{\theta}}_n = \widehat{\boldsymbol{\theta}}(\widehat{\boldsymbol{\beta}}_n(\lambda_n))$ as defined in (5). Let $\mathcal{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_d)$ be the $n \times d$ covariate matrix. The Karush-Kuhn-Tucker condition for (6) implies that

$$2\big(\mathbf{Y} - \mathcal{X}\widehat{\boldsymbol{\beta}}_n\big)'\mathbf{X}_k = \sum_{j: A_j \ni k} \widehat{\theta}_{nj}^{1-1/\gamma} c_j^{1/\gamma} \mathrm{sgn}\big(\widehat{\beta}_{nk}\big), \quad \forall\, \widehat{\beta}_{nk} \neq 0. \tag{7}$$

Since $\text{sgn}(\beta_{nk}) = \beta_{nk}/|\beta_{nk}|$, this allows us to write the fitted response vector as

$$\widehat{\mathbf{Y}} = \mathcal{X}\widehat{\boldsymbol{\beta}}_n = \mathcal{X}_{\lambda_n}[\mathcal{X}'_{\lambda_n}\mathcal{X}_{\lambda_n} + 0.5\mathcal{W}_{\lambda_n}]^{-1}\mathcal{X}'_{\lambda_n}\mathbf{Y},$$

where $\mathcal{X}_{\lambda_n}$ is the sub-matrix of $\mathcal{X}$ whose columns correspond to the covariates with nonzero estimated coefficients for the given λ_n and $\mathcal{W}_{\lambda_n}$ is the diagonal matrix with diagonal elements

$$\sum_{A_j \ni k} \widehat{\boldsymbol{\theta}}_{nj}^{1-1/\gamma} c_j^{1/\gamma} / |\widehat{\beta}_{nk}|, \quad \widehat{\beta}_{nk} \neq 0.$$

Therefore, the number of effective parameters with a given λ_n can be approximated by

$$d(\lambda_n) = \text{trace}(\mathcal{X}_{\lambda_n}[\mathcal{X}'_{\lambda_n}\mathcal{X}_{\lambda_n} + 0.5\mathcal{W}_{\lambda_n}]^{-1}\mathcal{X}'_{\lambda_n}).$$

This procedure is close to Fu (1998) but also resembles the tuning parameter selection method in Tibshirani (1996) and Zhang and Lu (2006).

An AIC-type criterion for choosing λ_n is

$$\text{AIC}(\lambda_n) = \log(\|\mathbf{Y} - \mathcal{X}\widehat{\boldsymbol{\beta}}_n(\lambda_n)\|_2^2/n) + 2d(\lambda_n)/n.$$

A GCV-type score (Wahba 1990) is defined as

$$\text{GCV}(\lambda_n) = \frac{\|\mathbf{Y} - \mathcal{X}\widehat{\boldsymbol{\beta}}_n(\lambda_n)\|_2^2}{n(1 - d(\lambda_n)/n)^2}.$$

It can be seen that these two criterions are close to each other when $d(\lambda_n)$ is relatively small compared to n.

Although GCV and AIC are reasonable criteria for tuning, they tend to select more variables than the true model contains. So we also consider a BIC-type criterion

$$\text{BIC}(\lambda_n) = \log(\|\mathbf{Y} - \mathcal{X}\widehat{\boldsymbol{\beta}}_n(\lambda_n)\|_2^2/n) + \log(n)d(\lambda_n)/n.$$

The tuning parameter λ_n is selected via minimization of $\text{AIC}(\lambda_n)$, $\text{GCV}(\lambda_n)$, or $\text{BIC}(\lambda_n)$. In general, the AIC- and GCV-type criterions are appropriate if the model is used for prediction,

and the BIC-type criterion should be used if the purpose of the analysis is to uncover the model structure (Yang 2003).

2.3 Variance estimation

The covariance matrix of $\widehat{\boldsymbol{\beta}}_n(\lambda_n)$ is estimated in a similar way as Tibshirani (1996)'s covariance estimate for the lasso estimator. Let $B_1 = B_1(\lambda_n) = \{k : \widehat{\beta}_{nk} \neq 0\}$ be the set of selected variables and $\widehat{\boldsymbol{\beta}}_{nB_1}(\lambda_n) = (\widehat{\beta}_{nk}(\lambda_n) : k \in B_1)$ be the nonzero components of $\widehat{\boldsymbol{\beta}}_n(\lambda_n)$ given λ_n. By (7),

$$\widehat{\boldsymbol{\beta}}_{nB_1}(\lambda_n) = [\mathcal{X}'_{\lambda_n}\mathcal{X}_{\lambda_n} + 0.5\mathcal{W}_{\lambda_n}]^{-1}\mathcal{X}'_{\lambda_n}\mathbf{Y},$$

so that the covariance matrix of $\widehat{\boldsymbol{\beta}}_{nB_1}(\lambda_n)$ can be approximated by

$$[\mathcal{X}'_{\lambda_n}\mathcal{X}_{\lambda_n} + 0.5\mathcal{W}_{\lambda_n}]^{-1}\mathcal{X}'_{\lambda_n}\mathcal{X}_{\lambda_n}[\mathcal{X}'_{\lambda_n}\mathcal{X}_{\lambda_n} + 0.5\mathcal{W}_{\lambda_n}]^{-1}\widehat{\sigma}^2, \tag{8}$$

where $\widehat{\sigma}^2 = \|\mathbf{Y} - \mathcal{X}\widehat{\boldsymbol{\beta}}_n(\lambda_n)\|_2^2/(n - d(\lambda_n))$.

2.4 Comparison with the group lasso

The group lasso estimator of Yuan and Lin (2006) is

$$\widetilde{\boldsymbol{\beta}}_n = \arg\min_{\boldsymbol{\beta}} \left\|\mathbf{Y} - \sum_{k=1}^{d}\mathbf{X}_k\beta_k\right\|_2^2 + \lambda_n\sum_{j=1}^{J}\|\boldsymbol{\beta}_{A_j}\|_{K_j,2}, \tag{9}$$

where K_j is a positive definite matrix and $\|\boldsymbol{\beta}_{A_j}\|_{K_j,2} = (\boldsymbol{\beta}'_{A_j}K_j\boldsymbol{\beta}_{A_j})^{1/2}$. A typical choice of K_j suggested by Yuan and Lin (2006) is $K_j = |A_j|I_j$, where I_j is the $|A_j| \times |A_j|$ identity matrix.

Let τ be a penalty parameter and define

$$S_{2n}(\boldsymbol{\beta}, \boldsymbol{\theta}) = \left\|Y - \sum_{k=1}^{d}\mathbf{X}_k\boldsymbol{\beta}_k\right\|_2^2 + \sum_{j=1}^{J}\theta_j^{-1}\|\boldsymbol{\beta}_{A_j}\|_{K_j,2}^2 + \tau\sum_{j=1}^{J}\theta_j. \tag{10}$$

Proposition 2. *Let $\tau = 2^{-2}\lambda_n^2$. Then $\widetilde{\boldsymbol{\beta}}_n$ satisfies (9) if and only if $(\widetilde{\boldsymbol{\beta}}_n, \widetilde{\boldsymbol{\theta}})$ solves*

$$\textit{minimize}\ \ S_{2n}(\boldsymbol{\beta}, \boldsymbol{\theta})\ \ \textit{subject to}\ \ \ \boldsymbol{\theta} \geq 0$$

for some $\widetilde{\boldsymbol{\theta}} \geq 0$.

From this proposition, the group lasso behaves like an "adaptively weighted ridge regression", in which the sum of the squared coefficients in group j is penalized by θ_j, and the sum of θ_j's is in turn penalized by τ. Therefore, in minimizing (10), either $\boldsymbol{\beta}_{A_j} = 0$, in which case the group is dropped from the model, or $\boldsymbol{\beta}_{A_j} \neq 0$, in which case all the elements of $\boldsymbol{\beta}_{A_j}$ are non-zero and all the variables in group j are retained in the model. So the group lasso selects groups of variables, but it does not select individual variables within groups.

3 Asymptotic properties

In this section, we study the asymptotic properties of the group bridge estimators. We show that, for $0 < \gamma < 1$, the group bridge estimators correctly select groups with nonzero coefficients with probability converging to one under reasonable conditions. We also derive the asymptotic distribution of the estimators of the nonzero coefficients.

Without loss of generality, suppose that

$$\boldsymbol{\beta}_{A_j} \neq 0,\ 1 \leq j \leq J_1, \qquad \boldsymbol{\beta}_{A_j} = 0,\ J_1 + 1 \leq j \leq J. \tag{11}$$

Let $B_2 = \cup_{j=J_1+1}^{J} A_j$ be the union of the groups with zero coefficients and $B_1 = B_2^c$. Let $\boldsymbol{\beta}_{B_j} = (\beta_k, k \in B_j)'$, $j = 1, 2$. Assume without loss of generality that the index k is arranged so that $\boldsymbol{\beta} = (\boldsymbol{\beta}'_{B_1}, \boldsymbol{\beta}'_{B_2})'$. Let $\boldsymbol{\beta}_0$ be the true value of $\boldsymbol{\beta}$. Since $\boldsymbol{\beta}_{0B_2} = 0$, the true model is fully explained by the first J_1 groups. In this notation, $\widehat{\boldsymbol{\beta}}_{nB_1}$ and $\widehat{\boldsymbol{\beta}}_{nB_2}$ are respectively the estimates of $\boldsymbol{\beta}_{B_1}$ and $\boldsymbol{\beta}_{B_2}$ from the group bridge estimator $\widehat{\boldsymbol{\beta}}_n$. Set $\mathcal{X} = (X_1, \ldots, X_d)$ and $\mathcal{X}_1 = (X_k, k \in B_1)$. Define

$$\Sigma_n = n^{-1}\mathcal{X}'\mathcal{X} \ \text{ and } \ \Sigma_{1n} = n^{-1}\mathcal{X}_1'\mathcal{X}_1.$$

Let ρ_n and ρ_n^* be the smallest and largest eigenvalues of Σ_n. We consider the following conditions.

(A1) The errors $\varepsilon_1, \varepsilon_2, \ldots \varepsilon_n$ are uncorrelated with mean zero and finite variance σ^2.

(A2) The maximum multiplicity $C_n^* = \max_k \sum_{j=1}^{J} I\{k \in A_j\}$ is bounded and

$$\frac{\lambda_n^2}{n\rho_n} \sum_{j=1}^{J_1} c_j^2 \|\boldsymbol{\beta}_{0A_j}\|_1^{2\gamma-2} |A_j| \leq \sigma^2 d M_n, \quad M_n = O(1). \tag{12}$$

(A3) The constants c_j are scaled so that $\min_{j\le J} c_j \ge 1$ and

$$\frac{\lambda_n \rho_n^{1-\gamma/2}}{d^{1-\gamma/2}\rho_n^* n^{\gamma/2}} \to \infty. \tag{13}$$

Condition (A1) is standard in linear regression. Conditions (A2) and (A3) both require full rank design, rank$(\mathcal{X}) = d \le n$. Still, we allow the number of covariates $d = d_n$ to grow at certain rate $n > d_n \to \infty$. For fixed unknown $\{B_1, \beta_{0B_1}, J_1\}$, (A2) and (A3) are consequences of

$$\frac{1}{\rho_n} + \rho_n^* + \sum_{j=1}^{J_1} c_j^2 = O(1), \quad \frac{\lambda_n}{n^{1/2}} \to \lambda_0 < \infty, \quad \frac{\lambda_n d^{\gamma/2}}{d n^{\gamma/2}} \to \infty, \tag{14}$$

provided $c_j \ge 1$ and $C_n^* = O(1)$. This allows $d_n = o(1)n^{(1-\gamma)/(2-\gamma)}$. It is clear from (14) that (A2) and (A3) put restrictions on the magnitude of the penalty parameter. In particular, they exclude the case $\gamma \ge 1$.

Theorem 1. (Group-bridge lasso) *Suppose that* $0 < \gamma < 1$.

(i) Suppose (A1), (A2) and (A3) hold. Then, $\widehat{\boldsymbol{\beta}}_{nB_2} = 0$ *with probability converging to 1.*

(ii) Suppose $\{B_1, \boldsymbol{\beta}_{0B_1}, J_1\}$ *are fixed unknowns and (14) holds. Suppose further that*

$$\Sigma_{1n} \to \Sigma_1, \quad n^{-1/2}\mathcal{X}_1'\boldsymbol{\varepsilon} \to_D W \sim N(0, \sigma^2\Sigma_1).$$

Then,

(a) (Group selection consistency)

$$P\big\{\widehat{\boldsymbol{\beta}}_{nB_2} = 0\big\} \to 1.$$

(b) (Asymptotic distribution of nonzero group estimators)

$$\sqrt{n}\big(\widehat{\boldsymbol{\beta}}_{nB_1} - \boldsymbol{\beta}_{0B_1}\big) \to_D \arg\min\Big\{V_1(\mathbf{u}) : \mathbf{u} \in \mathbb{R}^{|B_1|}\Big\},$$

where

$$V_1(\mathbf{u}) \;=\; -2\mathbf{u}'W + \mathbf{u}'\Sigma_1\mathbf{u}$$

$$+\gamma\lambda_0 \sum_{j=1}^{J_1} c_j \|\beta_{0A_j}\|_1^{\gamma-1} \sum_{k\in A_j\cap B_1} \Big\{ u_k sgn(\beta_{0k}) I(\beta_{0k}\neq 0) + |u_k| I(\beta_{0k}=0) \Big\}.$$

In particular, when $\lambda_0 = 0$*,*

$$\sqrt{n}\big(\widehat{\boldsymbol{\beta}}_{nB_1} - \boldsymbol{\beta}_{0B_1}\big) \to_D \Sigma_1^{-1} W \sim N(0, \sigma^2\Sigma_1^{-1}).$$

Part (a) of Theorem 1 is of particular interest. It states that the group bridge estimates of the coefficients of the zero groups are *exactly* equal to zero with probably converging to one. This, together with part (b), imply that the group bridge estimator is able to correctly distinguish nonzero groups from zero groups eventually. Therefore, the group bridge estimator has the powerful asymptotic oracle property in group selection. Part (b) shows that the estimator of nonzero coefficients is $n^{1/2}$-consistent and in general converges to the argmin of the Gaussian process V_1. When $\lambda_0 > 0$, the limiting distribution puts positive probability at 0.

The proof of Theorem 1 is given in the appendix. Since Theorem 1 is valid in the case of $A_j = \{j\}$, $j = 1, \ldots, d$, it generalizes the result of Huang et al. (2006), which showed selection consistency and asymptotic distribution for the bridge estimator of Frank and Friedman (1996). In this case, there is no need to select within groups and $\lambda_n/\sqrt{n} \to 0$ seems appropriate.

For iid errors, the assumption $n^{-1/2}\mathcal{X}_1'\boldsymbol{\varepsilon} \to_D W \sim N(0, \sigma^2\Sigma_1)$ follows from the Lindeberg central limit theorem under $\Sigma_{1n} \to \Sigma_1$, cf. Van der Vaart (1998). To compare the different asymptotic properties of the group bridge and group lasso estimators, we present the following theorem for the group lasso estimator of Yuan and Lin (2006).

Theorem 2. (Group lasso) *Suppose* $\{\boldsymbol{\beta}, d, A_j, c_j, K_j, j \le J\}$ *are all fixed as* $n \to \infty$ *and that* ε_i *are iid errors with* $E\varepsilon_i = 0$ *and Var*$(\varepsilon_i) = \sigma^2 \in (0, \infty)$*. Suppose further that the* $d \times d$ *matrices* Σ_n *converges to a positive-definite matrix* Σ *and* $\lambda_n n^{-1/2} \to \lambda_0 < \infty$*. Then*

$$\sqrt{n}(\widetilde{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \to_D \arg\min \Big\{ V(\mathbf{u}) : \mathbf{u} \in \mathbb{R}^d \Big\}$$

where for some $W \sim N(0, \sigma^2\Sigma)$

$$V(\mathbf{u}) = -2\mathbf{u}'W + \mathbf{u}'\Sigma\mathbf{u} + \lambda_0 \sum_{j=1}^{J} c_j \left[\frac{\mathbf{u}_{A_j}' K_j \boldsymbol{\beta}_{0A_j}}{\|\boldsymbol{\beta}_{0A_j}\|_{K_j,2}} I(\boldsymbol{\beta}_{A_j} \neq 0) + \|\mathbf{u}_{A_j}\|_{K_j,2} I(\boldsymbol{\beta}_{0A_j} = 0) \right].$$

By Theorem 2, when $\lambda_0 = 0$, the group lasso estimator has the same asymptotic distribution as the least squares estimator. Therefore, it is required that $\lambda_0 > 0$ for the group lasso to carry out group selection. When $\lambda_0 > 0$, the asymptotic distribution of $\widetilde{\boldsymbol{\beta}}_n$ puts positive probability at 0 when $\boldsymbol{\beta}_{A_j} = 0$. However, in general, this positive probability is less than one. Thus, the group lasso is in general not consistent in selecting the nonzero groups.

The proof of Theorem 2 is similar to that of part (b) of Theorem 1, so it is omitted. When $|A_j| = 1, 1 \leq j \leq J$, this theorem simplifies to the result on lasso of Knight and Fu (2000).

4 Numerical Studies

4.1 Simulation study

We use simulation to evaluate the finite sample performance of the group bridge estimator. As a comparison, we also look at the group lasso estimator in the simulation. The group lasso estimator is computed using the algorithm in Yuan and Lin (2006). We consider 3 scenarios of simulation models. For the generating models in the first scenario (Examples 1 and 2), the number of groups is small, the group sizes are equal and relatively large. In the second scenario (Examples 3 and 4), the number of groups is relatively large, the group sizes are equal and small. In each of these two scenarios, we consider two types of models. In the first type, the coefficients are either all nonzero or zero. In the second type, there are zero coefficients in a nonzero group. For the generating models in the third scenario (Examples 5 and 6), the group sizes vary. There are zero coefficients in a nonzero group in both examples. We use $\gamma = 0.5$ in the group bridge estimator. The sample size $n = 200$ in each example.

Example 1. In this example, there are 5 groups, each with 8 covariates. The coefficients in each group are either all nonzero or all zero. First simulate $R_1, \ldots, R_{40}$ independently from the standard normal distribution. Next, simulate Z_j, $j = 1, \ldots, 5$ from the standard normal distribution with an $AR(1)$ structure, i.e., $\text{Cov}(Z_{j_1}, Z_{j_2}) = 0.4^{|j_1 - j_2|}$ for $1 \leq j_1, j_2 \leq 4$. Then the covariate vector $(X_1, \ldots, X_{40})'$ consists of

$$X_{5(j-1)+k} = (Z_j + R_{4(j-1)+k})/\sqrt{2}, \quad 1 \leq j \leq 5,\ 1 \leq k \leq 8.$$

The random error $\varepsilon \sim N(0, 2^2)$. The response vector is computed using model (1) with

$$\boldsymbol{\beta} = (\underbrace{0.5, 1, \ldots, 3.5, 4}_{8}, \underbrace{2, \ldots, 2}_{8}, \underbrace{0, \ldots, 0}_{24})'.$$

The group ID for the covariate vector is $(\underbrace{1, \ldots, 1}_{8}, \underbrace{2, \ldots, 2}_{8}, \underbrace{3, \ldots, 3}_{8}, \underbrace{4, \ldots, 4}_{8}, \underbrace{5, \ldots, 5}_{8})$.

Example 2. Similar to Example 1, there are a small number of groups but large group size. The covariates are generated the same way as in Example 1, except that here

$$\boldsymbol{\beta} = (\underbrace{0, 1, 0, 2, 0, 3, 0, 4}_{8}, \underbrace{2, 2, 2, 2, 0, 0, 0, 0}_{8}, \underbrace{0, \ldots, 0}_{24})'.$$

Example 3. In this example, we consider a regression model that has 10 groups, each with 4 covariates. In each group, the coefficients are either all non-zero or all zero. The data are generated as follows. Let $(Z_1, \ldots, Z_{10})$ be a multivariate normal random vector with marginal distribution $N(0, 1)$. The covariance between Z_j and $Z_{j'}$ is $0.6^{|j-j'|}$. The covariate vector is $(X_1, \ldots, X_{40})'$ with

$$X_{4(j-1)+k} = I(\frac{k}{5} < \Phi(Z_j) \leq \frac{k+1}{5}), \quad j = 1, \ldots, 10, \quad k = 1, \ldots, 4.$$

The group ID for the covariate vector is $(\underbrace{1, 1, 1, 1}_{4}, \underbrace{2, 2, 2, 2}_{4}, \ldots, \underbrace{10, 10, 10, 10}_{4})$. The response Y was then calculated based on model (1) with

$$\boldsymbol{\beta} = (\underbrace{3, 3, 3, 3}_{4}, \underbrace{0, \ldots, 0}_{4}, \underbrace{-4, -4, -4, -4}_{4}, \underbrace{4, -3, -4, 3}_{4}, \underbrace{0, \ldots, 0}_{24})',$$

and ε is normally distributed with mean 0 and variance 4.

Example 4. Both the covariates and random errors are simulated in the same way as in Example 3. The underlying coefficient vector is

$$\boldsymbol{\beta} = (\underbrace{0, 0, 3, 3}_{4}, \underbrace{0, \ldots, 0}_{4}, \underbrace{-4, 0, 0, -4}_{4}, \underbrace{4, -3, 0, 0}_{4}, \underbrace{0, \ldots, 0}_{24})'.$$

In this example, there are coefficients in the non-zero group.

Example 5. In this example, the model has groups of different sizes. The covariates are generated in much the same way as in Example 3. First simulate Z_i, $i = 1, \ldots, 6$ and $R_1, \ldots, R_{42}$

independently from the standard normal distribution. Then the covariate $(X_1, \ldots, X_{42})'$ are formed as follows:

$$X_j = (Z_{g_j} + R_j)/\sqrt{2}, \quad 1 \le j \le 42.$$

where $\mathbf{g}' = (g_1, \ldots, g_{42}) = (\underbrace{1, \ldots, 1}_{10}, \underbrace{2, \ldots, 2}_{10}, \underbrace{3, \ldots, 3}_{10}, \underbrace{4, \ldots, 4}_{4}, \underbrace{5, \ldots, 5}_{4}, \underbrace{6, \ldots, 6}_{4})$ is the group ID vector. The random error ε is sampled from $N(0, 2^2)$. The response vector is computed using (1) with $\boldsymbol{\beta}' = (\boldsymbol{\beta}_1', \ldots, \boldsymbol{\beta}_6')$ with

$$\begin{aligned} \boldsymbol{\beta}_1' &= (0.5, -2, 0.5, 2, -1, 1, 2, -1.5, 2, -2), \ \boldsymbol{\beta}_2' = (-1.5, 2, 1, -2, 1.5, \underbrace{0, \ldots, 0}_{5}), \\ \boldsymbol{\beta}_3' &= (\underbrace{0, \ldots, 0}_{10}), \ \boldsymbol{\beta}_4' = (2, -2, 1, 1.5), \ \boldsymbol{\beta}_5' = (-1.5, 1.5, 0, 0), \ \boldsymbol{\beta}_6' = (\underbrace{0, \ldots, 0}_{4}). \end{aligned}$$

Example 6. The generating model is the same as Example 5 except that $\boldsymbol{\beta}_2 = (-1.5, 2, \underbrace{0, \ldots, 0}_{8})'$. So in this model, there is a sparse group.

For these examples, the simulation results based on 400 replications are summarized in Table 1. For the group bridge estimators, we considered AIC, BIC, and GCV for determining the penalty parameter. The variable selection and coefficient estimation results based on GCV are similar to those using AIC and thus omitted. For the group lasso, we considered C_p, AIC, and BIC. The results based on AIC are similar to those based on C_p. Because the C_p method was suggested by Yuan and Li (2006), we included the C_p and BIC results in Table 1.

The first column in Table 1 gives the number of nonzero groups and coefficients in the generating models, for example, in Example 1, there are 2 nonzero groups and 16 nonzero coefficients. The model error is computed as $(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})' E[XX'](\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$, where β is the generating value. The row in "No. of groups" gives the number of groups selected, averaged over 400 replications. The row in "No. of coefs " counts the average number of variables selected. The "% correct mod" shows the percentage that the model produced contains exactly the same groups as the underlying model. In the parentheses are the corresponding standard deviations.

Comparing different tuning parameter selection methods, for both the group lasso and group bridge, the BIC does better than the AIC or C_p in terms of variable selection, in that it has a greater chance of selecting the true underlying model. The group bridge with tuning parameter selected based on BIC produces better results in terms of model selection and model errors. However, for

the group lasso, the tuning parameters selected by BIC yield larger model errors.

From Table 1, we see that there are improvements of the group bridge with BIC over the group lasso with BIC or C_p in terms of model error, the number of groups selected, the number of variables selected, and the percentage of correct model selected. In particular, the improvements in the category of the percentage of correctly selected models is considerable. In general, the group lasso tends to select more groups and variables than there actually are in the generating models using either C_p or BIC for tuning parameter selection. This agrees with the simulation results reported in Yuan and Lin (2006). In comparison, the number of groups and variables in the models selected by the group bridge with BIC are close to the generating values. The only exception is in Example 3, in which the group lasso does better in terms of number of selected groups.

To examine the selection results for each covariate, we plot the percentage of the 400 replications when a coefficient is estimated exactly at zero (i.e., the associated covariate is not selected). The results in Figures 1 and 2. In each plot in these figures, triangles represent results from BIC tuning parameter selection; solid dots represent results from Cp for group lasso or AIC for group bridge. Also in each plot, the small circles on the horizontal axis indicate nonzero coefficients in the generating model. For example, in the upper left panel of Figure 1, there are four circles at 1, 2, 3, and 4, indicating that the first 4 coefficients in Example 1 are nonzero. From Figures 1 and 2, we see that the group bridge estimates have higher percentage of correctly identifying zero coefficients than the group lasso.

We also looked at the performance of the proposed standard error estimation method. As examples, Table 2 shows the empirical and estimated standard errors of the group bridge estimates of coefficients β_2 and β_{12}. In general, when the BIC is used in tuning parameter selection, the proposed method tends to slightly underestimate true sampling variabilities, but otherwise appears to provides reasonable standard error estimates. The slight underestimation is perhaps due to the effect of choosing tuning parameters, which is not accounted for in (8). Further studies are needed to investigate the question of variance estimation in penalized estimation problems.

4.2 Impact study

The Impact study was part of a three-year project designed to measure the impact of nutritional policies and environmental change on obesity in the high school students enrolled in Seattle Public

Schools. The study description can be found at

http://depts.washington.edu/uwcphn/activities/projects/noncommercialism.html.

The study was led by University of Washington Center for Public Health Nutrition and conducted in two urban high schools. One primary goal of this study is to determine the effects of different risk factors on BMI (body mass index). We show in Table 3 the definitions of the variables included in the study. Indicators are firstly created for the ethnic variable. Natural clusters exist for the risk factors. 25 covariates can be naturally classified into different categories, measuring different aspects such as food sources and demographics. The response variable is the logarithm of BMI. We focus on 799 subjects with complete records.

We use the proposed group bridge approach to analyzing the data from the Impact study. For comparison, we also consider the OLS and group lasso. The results are given in Table 4. For the group lasso, when C_p is used in tuning parameter selection, all the 8 groups are selected in the final model, but when BIC is used, none of the groups are selected. So the results from the group lasso with BIC are not included in the table.

In comparison, the group bridge does not select the group of consumption of healthy food when AIC or GCV is used. The group bridge with BIC gives rise to a sparser model and drops 3 more groups: age, gender and unhealthy food consumption. We conclude from the group bridge estimate that demographics, food source, unhealthy food consumption, school group and physical activity have important effects on BMI in the Impact cohort. In the ethnicity group, the group bridge using BIC only selects Hawaiian and Asian.

For evaluation purpose, we first randomly select a training set of size 600. The testing set is composed of the remaining 199 records. We compute estimates using the training set only, and then compute the prediction mean square errors (PMSE) for the testing set. The splitting, estimation and prediction are repeated 200 times. The results are summarized in Table 5. It can be seen that when AIC and C_p are used, the group-bridge lasso selects fewer groups and has smaller prediction errors. When BIC is used, the group lasso rarely selects any group and yields the null model, but its prediction errors are comparable with that produced by ordinary least squares. This suggests that the variation of this cohort's BMI is not very well captured by the variables measured in the study and that other variables such as genetic factors may be of greater importance in explaining the variation of BMI. In fact, with the full model using the least squares, the R^2 value is only 8%. Similar results have been observed in a previous study on factors that may affect BMI (Storey et

al. 2003). We must exercise caution in interpreting the analysis results of this data set, since this is an observational study, and most of the participating students are of African American origin. So the results here cannot be extrapolated to the general population of high school students.

5 Discussion

The proposed group bridge approach can be applied to other regression problems when both group and individual variable selections are desired. Specifically, we can use the group-bridge lasso penalty in the context of the general M-estimation,

$$\sum_{i=1}^{n} m(Y_i, \beta_0 + \sum_{k=1}^{d} X_{ik}\beta_k) + \lambda_n \sum_{j=1}^{J} \|\boldsymbol{\beta}_{A_j}\|_1^{\gamma}, \tag{15}$$

where m is a given loss function. This formulation includes the generalized linear models, censored regression models including the Cox regression, and robust regression. For example, for the generalized linear models such as logistic regression, we take m to be the negative log-likelihood function. For the Cox regression, we take the empirical loss function to be the negative partial likelihood. For loss functions other than least squares, further work is needed to study the computational algorithms and theoretical properties of the group bridge estimators.

A more general view can be adopted regarding the formulation of penalties. The group bridge penalty is a combination of two penalties–the bridge penalty for group selection and the lasso for within-group selection. In general, it is possible to consider combinations of different penalties, for example, we can use the SCAD penalty for within group selection and the bridge penalty for group selection. Different penalty functions may be preferable under different data and model settings. Further studies are needed on the estimators with different penalties regarding the computational algorithms and their theoretical properties.

Finally, we only considered the asymptotic properties of the group bridge estimators in the settings when the number of covariates is smaller than the sample size. The need for two-level selection also arises in high-dimensional problems when the number of covariates is larger than the sample size. For example, in regression analysis of a clinical outcome, such as disease status or survival, with high-dimensional genomic data, it is natural to consider genes in the same pathway as a group. Typically, there is only a limited number of pathways and genes that will be important

to a clinical outcome. Therefore, it is of interest to study the properties of the group bridge method sparse models when the number of coviarates is larger than the sample.

6 Proofs

Proof of Proposition 1. We have $\min_{\boldsymbol{\beta},\boldsymbol{\theta}} S_{1n}(\boldsymbol{\beta},\boldsymbol{\theta}) = \min_{\boldsymbol{\beta}} \widehat{S}_{1n}(\boldsymbol{\beta})$, where $\widehat{S}_{1n}(\boldsymbol{\beta}) = \min_{\boldsymbol{\theta}}\{S_{1n}(\boldsymbol{\beta},\boldsymbol{\theta}) : \boldsymbol{\theta} \geq 0\}$. For any give $\boldsymbol{\beta}$,

$$\widehat{\boldsymbol{\theta}}(\boldsymbol{\beta}) \equiv \arg\min\{S_{1n}(\boldsymbol{\beta},\boldsymbol{\theta}) : \boldsymbol{\theta} \geq 0\} = \arg\min\Big\{\sum_{j=1}^{J} \theta_j^{1-1/\gamma} c_j^{1/\gamma} \|\boldsymbol{\beta}_{A_j}\|_1 + \tau \sum_{j=1}^{J} \theta_j, \boldsymbol{\theta} \geq 0\Big\}.$$

Therefore, $\widehat{\boldsymbol{\theta}}(\boldsymbol{\beta}) = (\widehat{\theta}_1(\boldsymbol{\beta}), \ldots, \widehat{\theta}_d(\boldsymbol{\beta}))'$ must satisfy

$$(1/\gamma - 1)\theta_j^{-1/\gamma}(\boldsymbol{\beta}) c_j^{1/\gamma} \|\boldsymbol{\beta}_{A_j}\|_1 = \tau, \quad j = 1, \ldots, J.$$

Write $\widehat{S}_{1n}(\boldsymbol{\beta}) = S_{1n}(\boldsymbol{\beta}, \widehat{\boldsymbol{\theta}}(\boldsymbol{\beta}))$ and substitute the expressions

$$\theta_j(\boldsymbol{\beta}) = \Big(\frac{1-\gamma}{\gamma}\Big)^{\gamma} c_j \|\boldsymbol{\beta}_{A_j}\|_1^{\gamma} \tau^{-\gamma}, \quad \theta_j^{1-1/\gamma}(\boldsymbol{\beta}) = \Big(\frac{\gamma}{1-\gamma}\Big)^{1-\gamma} \frac{c_j^{1-1/\gamma} \tau^{1-\gamma}}{\|\boldsymbol{\beta}_{A_j}\|_1^{1-\gamma}}$$

into $S_{1n}(\boldsymbol{\beta}, \widehat{\theta}(\boldsymbol{\beta}))$, we get, after some algebra,

$$\widehat{S}_{1n}(\boldsymbol{\beta}) = \Big\|\mathbf{Y} - \mathcal{X}\boldsymbol{\beta}\Big\|_2^2 + \lambda_n \sum_{j=1}^{J} c_j \|\boldsymbol{\beta}_{A_j}\|_1^{\gamma}.$$

Here we used $\lambda_n = \tau^{1-\gamma}\big\{(1/\gamma - 1)^{\gamma} + (1/\gamma - 1)^{\gamma-1}\big\}$, so that $\widehat{S}_{1n}(\boldsymbol{\beta}) = L_n(\boldsymbol{\beta})$. □

Theorem 1 is proved by establishing the following results in three steps: (a) estimation consistency and rate of convergence (Lemma 1); (b) variable-selection consistency (Lemma 2 and part (i) of Theorem 1); and (c) asymptotic distribution (part (ii) of Theorem 1).

Recall that $\|\boldsymbol{\beta}_{0A_j}\|_2 = 0$ iff $j > J_1$ by (11) and condition (A2) gives

$$\frac{\lambda_n^2 \eta_n^2}{n\rho_n} \leq \sigma^2 d M_n \text{ for } \eta_n = \Big(\sum_{j=1}^{J_1} c_j^2 \|\boldsymbol{\beta}_{0A_j}\|_1^{2\gamma-2} |A_j|\Big)^{1/2} \text{ and } M_n = O(1). \tag{16}$$

Lemma 1. *Suppose conditions (A1) and (A2) hold with* $0 < \gamma \leq 1$. *Then,*

$$\mathrm{E}\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2^2 \leq \frac{\sigma^2 d}{n\rho_n}\Big(8 + 16C_n^* M_n\Big), \tag{17}$$

where ρ_n *is the smallest eigenvalue of* $\Sigma = \mathcal{X}'\mathcal{X}/n$.

Proof of Lemma 1. By the definition of $\widehat{\boldsymbol{\beta}}_n$,

$$\|\mathbf{Y} - \mathcal{X}\widehat{\boldsymbol{\beta}}_n\|_2^2 + \lambda_n \sum_{j=1}^{J} c_j \|\widehat{\boldsymbol{\beta}}_{nA_j}\|_1^\gamma \leq \|\mathbf{Y} - \mathcal{X}\boldsymbol{\beta}_0\|_2^2 + \lambda_n \sum_{j=1}^{J} c_j \|\boldsymbol{\beta}_{0A_j}\|_1^\gamma.$$

Since $b^\gamma - a^\gamma \leq 2(b-a)b^{\gamma-1}$ for $0 \leq a \leq b$, by Cauchy-Schwarz

$$\begin{aligned}
\sum_{j=1}^{J} c_j \|\boldsymbol{\beta}_{0A_j}\|_1^\gamma - \sum_{j=1}^{J} c_j \|\widehat{\boldsymbol{\beta}}_{nA_j}\|_1^\gamma &\leq 2\sum_{j=1}^{J_1} c_j \|\boldsymbol{\beta}_{0A_j}\|_1^{\gamma-1} \|\widehat{\boldsymbol{\beta}}_{nA_j} - \boldsymbol{\beta}_{0A_j}\|_1 \\
&\leq 2\sum_{j=1}^{J_1} c_j \|\boldsymbol{\beta}_{0A_j}\|_1^{\gamma-1} \Big(|A_j| \|\widehat{\boldsymbol{\beta}}_{nA_j} - \boldsymbol{\beta}_{0A_j}\|_2^2\Big)^{1/2} \\
&\leq 2\eta_n \Big(\sum_{j=1}^{J_1} \|\widehat{\boldsymbol{\beta}}_{nA_j} - \boldsymbol{\beta}_{0A_j}\|_2^2\Big)^{1/2}.
\end{aligned}$$

Since $\sum_{j=1}^{J} \|\widehat{\boldsymbol{\beta}}_{nA_j} - \boldsymbol{\beta}_{0A_j}\|_2^2 \leq C_n^* \|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2^2$, the combination of the above inequalities yields

$$\begin{aligned}
2\lambda_n \eta_n \sqrt{C_n^*} \|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2 &\geq \|\mathbf{Y} - \mathcal{X}\widehat{\boldsymbol{\beta}}_n\|_2^2 - \|\mathbf{Y} - \mathcal{X}\boldsymbol{\beta}_0\|_2^2 \\
&= \|\mathcal{X}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)\|_2^2 + 2\boldsymbol{\varepsilon}'\mathcal{X}(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}_n).
\end{aligned} \tag{18}$$

Let $\delta_n = \|\mathcal{X}(\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)\|_2$ and $\boldsymbol{\varepsilon}_*$ be the projection of $\boldsymbol{\varepsilon}$ to the span of $\{\mathbf{X}_1, \ldots, \mathbf{X}_d\}$. By Cauchy-Schwarz, $2|\boldsymbol{\varepsilon}'\mathcal{X}(\boldsymbol{\beta}_0 - \widehat{\boldsymbol{\beta}}_n)| \leq 2\|\boldsymbol{\varepsilon}_*\|_2 \delta_n \leq 2\|\boldsymbol{\varepsilon}_*\|^2 + \delta_n^2/2$, so that by (18)

$$\delta_n^2 \leq 4\|\boldsymbol{\varepsilon}_*\|_2^2 + 4\lambda_n \eta_n \sqrt{C_n^*} \|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2.$$

Moreover, since ρ_n is the smallest eigenvalue of $\mathcal{X}'\mathcal{X}/n$, the above inequality implies

$$n\rho_n \|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2^2 \leq \delta_n^2 \leq 4\|\boldsymbol{\varepsilon}_*\|_2^2 + 4\lambda_n \eta_n \sqrt{C_n^*} \|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2$$

Since $\boldsymbol{\varepsilon}_*$ is the projection of $\boldsymbol{\varepsilon}$ to a d-dimensional space, $\mathrm{E}\|\boldsymbol{\varepsilon}_*\|_2^2 \le \sigma^2 d$. Thus,

$$\begin{aligned}\mathrm{E}\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2^2 &\le 4\sigma^2 d/(n\rho_n) + \{4\lambda_n\eta_n\sqrt{C_n^*}/(n\rho_n)\}\Big(\mathrm{E}\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2^2\Big)^{1/2} \\ &\le 4\sigma^2 d/(n\rho_n) + \{4\lambda_n\eta_n\sqrt{C_n^*}/(n\rho_n)\}^2/2 + \frac{1}{2}\mathrm{E}\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2^2.\end{aligned}$$

This completes the proof of the lemma in view of (16). □

Lemma 2. *Suppose the conditions (A1), (A2) and (A3) hold with* $0 < \gamma < 1$. *Then,*

$$P\Big\{\widehat{\boldsymbol{\beta}}_{nA_j} = 0 \; \forall j > J_1\Big\} \to 1. \tag{19}$$

Proof. Let $B_2 = \cup_{j=J_1+1}^{J} A_j$ and define $\widetilde{\boldsymbol{\beta}}_n = (\widetilde{\beta}_{n1}, \ldots, \widetilde{\beta}_{nd})'$ by

$$\widetilde{\beta}_{nk} = \begin{cases} \widehat{\beta}_{nk} & k \notin B_2 \\ 0 & k \in B_2. \end{cases}$$

Since $\widehat{\theta}_{nj}^{1-1/\gamma} c_j^{1/\gamma} \|\widehat{\boldsymbol{\beta}}_{A_j}\|_1 = \gamma\lambda_n c_j \|\widehat{\boldsymbol{\beta}}_{A_j}\|_1^\gamma$ by (5), (7) implies

$$2\big(\mathbf{Y} - \mathcal{X}\widehat{\boldsymbol{\beta}}_n\big)'\mathbf{X}_k = \gamma\lambda_n \sum_{A_j \ni k} c_j \|\widehat{\boldsymbol{\beta}}_{nA_j}\|_1^{\gamma-1} \mathrm{sgn}\big(\widehat{\beta}_{nk}\big), \quad \widehat{\beta}_{nk} \neq 0.$$

Since $(\widehat{\beta}_{nk} - \widetilde{\beta}_{nk})\mathrm{sgn}(\widehat{\beta}_{nk}) = |\widehat{\beta}_{nk}| I\{k \in B_2\}$, we have

$$\begin{aligned}2(\mathbf{Y} - \mathcal{X}\widehat{\boldsymbol{\beta}}_n)'\mathcal{X}(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}_n) &= \sum_{k \in B_2} |\widehat{\beta}_{nk}| \gamma\lambda_n \sum_{A_j \ni k} c_j \|\widehat{\boldsymbol{\beta}}_{nA_j}\|_1^{\gamma-1} \\ &= \gamma\lambda_n \sum_{j=1}^{J} c_j \|\widehat{\boldsymbol{\beta}}_{A_j}\|_1^{\gamma-1} \Big(\|\widehat{\boldsymbol{\beta}}_{nA_j}\|_1 - \|\widetilde{\boldsymbol{\beta}}_{nA_j}\|_1\Big)\end{aligned}$$

Since $\gamma b^{\gamma-1}(b-a) \le b^\gamma - a^\gamma$ for $0 \le a \le b$, for $j \le J_1$ we have

$$\gamma \|\widehat{\boldsymbol{\beta}}_{nA_j}\|_1^{\gamma-1} \Big(\|\widehat{\boldsymbol{\beta}}_{nA_j}\|_1 - \|\widetilde{\boldsymbol{\beta}}_{nA_j}\|_1\Big) \le \|\widehat{\boldsymbol{\beta}}_{nA_j}\|_1^\gamma - \|\widetilde{\boldsymbol{\beta}}_{nA_j}\|_1^\gamma.$$

Due to $\|\widetilde{\boldsymbol{\beta}}_{nA_j}\|_1 = 0$ for $j > J_1$, this implies

$$2\Big|(\mathbf{Y} - \mathcal{X}\widehat{\boldsymbol{\beta}}_n)'\mathcal{X}(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}_n)\Big| \le \lambda_n \sum_{j=1}^{J_1} c_j \Big(\|\widehat{\boldsymbol{\beta}}_{nA_j}\|_1^\gamma - \|\widetilde{\boldsymbol{\beta}}_{nA_j}\|_1^\gamma\Big) + \gamma\lambda_n \sum_{j=J_1+1}^{J} c_j \|\widehat{\boldsymbol{\beta}}_{nA_j}\|_1^\gamma. \quad (20)$$

Similar to the proof of Lemma 1, the definition of $\widehat{\boldsymbol{\beta}}_n$ gives

$$\|\mathbf{Y} - \mathcal{X}\widehat{\boldsymbol{\beta}}_n\|_2^2 + \lambda_n \sum_{j=1}^{J} c_j \|\widehat{\boldsymbol{\beta}}_{nA_j}\|_1^\gamma \le \|\mathbf{Y} - \mathcal{X}\widetilde{\boldsymbol{\beta}}_n\|_2^2 + \lambda_n \sum_{j=1}^{J} c_j \|\widetilde{\boldsymbol{\beta}}_{nA_j}\|_1^\gamma.$$

Since $\|\widetilde{\boldsymbol{\beta}}_{nA_j}\|_1 = 0$ for $j > J_1$, by (20)

$$\begin{aligned}
& 2\Big|(\mathbf{Y} - \mathcal{X}\widehat{\boldsymbol{\beta}}_n)'\mathcal{X}(\widehat{\boldsymbol{\beta}} - \widetilde{\boldsymbol{\beta}}_n)\Big| + (1-\gamma)\lambda_n \sum_{j=J_1+1}^{J} c_j \|\widehat{\boldsymbol{\beta}}_{nA_j}\|_1^\gamma \\
\le\ & \lambda_n \sum_{j=1}^{J} c_j \|\widehat{\boldsymbol{\beta}}_{nA_j}\|_1^\gamma - \lambda_n \sum_{j=1}^{J} c_j \|\widetilde{\boldsymbol{\beta}}_{nA_j}\|_1^\gamma \\
\le\ & \|\mathbf{Y} - \mathcal{X}\widetilde{\boldsymbol{\beta}}_n\|_2^2 - \|\mathbf{Y} - \mathcal{X}\widehat{\boldsymbol{\beta}}_n\|_2^2 \\
=\ & \|\mathcal{X}(\widehat{\boldsymbol{\beta}}_n - \widetilde{\boldsymbol{\beta}}_n)\|_2^2 + 2(\mathbf{Y} - \mathcal{X}\widehat{\boldsymbol{\beta}}_n)'\mathcal{X}(\widehat{\boldsymbol{\beta}}_n - \widetilde{\boldsymbol{\beta}}_n).
\end{aligned}$$

Thus, since $n\rho_n^*$ is the largest eigenvalue of $\mathcal{X}'\mathcal{X}$ and $\widehat{\beta}_{nk} - \widetilde{\beta}_{nk} = \widehat{\beta}_{nk} I\{k \in B_2\}$,

$$(1-\gamma)\lambda_n \sum_{j=J_1+1}^{J} c_j \|\widehat{\boldsymbol{\beta}}_{nA_j}\|_1^\gamma \le \|\mathcal{X}(\widehat{\boldsymbol{\beta}}_n - \widetilde{\boldsymbol{\beta}}_n)\|_2^2 = n\rho_n^* \|\widehat{\boldsymbol{\beta}}_{nB_2}\|_2^2 \le n\rho_n^* \|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\|_2^2,$$

which implies by Lemma 1 and $C_n^* M_n = O(1)$ in Condition (A2) that

$$(1-\gamma)\lambda_n \sum_{j=J_1+1}^{J} c_j \|\widehat{\boldsymbol{\beta}}_{nA_j}\|_1^\gamma \le n\rho_n^* \|\widehat{\boldsymbol{\beta}}_{nB_2}\|_2^2 \le O_P(\sigma^2 d\rho_n^*/\rho_n). \quad (21)$$

We still need to find a lower bound of $\sum_{j=J_1+1}^{J} c_j \|\widehat{\boldsymbol{\beta}}_{nA_j}\|_1^\gamma$. Since $c_j \ge 1$ by (A3),

$$\sum_{j=J_1+1}^{J} c_j \|\widehat{\boldsymbol{\beta}}_{nA_j}\|_1^\gamma \ge \Big(\sum_{j=J_1+1}^{J} \|\widehat{\boldsymbol{\beta}}_{nA_j}\|_1\Big)^\gamma \ge \|\widehat{\boldsymbol{\beta}}_{nB_2}\|_1^\gamma \ge \|\widehat{\boldsymbol{\beta}}_{nB_2}\|_2^\gamma. \quad (22)$$

In the event $\|\widehat{\boldsymbol{\beta}}_{nB_2}\|_2 > 0$, the combination of (21) and (22) yields

$$(1-\gamma)\lambda_n \le n\rho_n^*\|\widehat{\boldsymbol{\beta}}_{nB_2}\|_2^{2-\gamma} \le O_P(1)n\rho_n^*\Big(\sigma^2 d/(n\rho_n)\Big)^{1-\gamma/2}$$

Since $\lambda_n(\rho_n/d)^{1-\gamma/2}/(\rho_n^* n^{\gamma/2}) \to \infty$ by (A3), this implies

$$\mathrm{P}\Big\{\|\widehat{\boldsymbol{\beta}}_{nB_2}\|_2 > 0\Big\} \le \mathrm{P}\Big\{\frac{\lambda_n(\rho_n/d)^{1-\gamma/2}}{\rho_n^* n^{\gamma/2}} \le O_P(1)\Big\} \to 0$$

The proof is complete. □

Proof of Theorem 1. Since (i) follows from Lemma 2, it suffices to prove (ii). Since d_1 and $\boldsymbol{\beta}_{0B_1}$ are fixed, $\min_{j\le J_1}\|\boldsymbol{\beta}_{0A_j}\|_1^{1-\gamma} = O(1)$, so that (14) implies (12), (13) and

$$\frac{\lambda_n^2}{n\rho_n}\sum_{j=1}^{J_1} c_j^2\|\boldsymbol{\beta}_{0A_j}\|_1^{2\gamma-2}|A_j\cap B_1| = O(1). \tag{23}$$

Thus, the conditions of (i) hold. Moreover, by (23), the proof of Lemma 1 still works with the reduced design $\mathcal{X}_1$ and reduced total number $d_1 = |B_1|$ of coefficients $\beta_k, k \in B_1$. Thus,

$$\big\|\widehat{\boldsymbol{\beta}}_{nB_1} - \boldsymbol{\beta}_{0B_1}\big\|^2 = O_P(1/n), \quad \big\|\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0\big\|^2 = O_P(1/n).$$

Let $h_n = n^{-1/2}$ and define

$$V_{1n}(\mathbf{u}) \;=\; L_n\big(\boldsymbol{\beta}_0 + h_n(\mathbf{u}', \mathbf{0}')'\big) - L_n(\boldsymbol{\beta}_0)$$

with $\mathbf{0}$ being the zero vector of dimension $|B_2|$. By (i), the following holds with large probability:

$$\widehat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0 = h_n(\widehat{\mathbf{u}}_n', \mathbf{0}')', \qquad \widehat{\mathbf{u}}_n = \arg\min\Big\{V_{1n}(\mathbf{u}) : \mathbf{u} \in \mathbb{R}^{d_1}\Big\}.$$

The function $V_{1n}(\mathbf{u})$, $\mathbf{u} \in \mathbb{R}^{d_1}$, can be written as

$$\begin{aligned} V_{1n}(\mathbf{u}) &= \Big\{-2h_n\mathbf{u}'\mathcal{X}_1'\boldsymbol{\varepsilon} + h_n^2\mathbf{u}'\mathcal{X}_1'\mathcal{X}_1\mathbf{u}\Big\} + \lambda_n\sum_{j=1}^{J_1} c_j\Big\{\Big(\sum_{k\in A_j\cap B_1}|\beta_{0k} + h_n\mathbf{u}_k|\Big)^\gamma - \|\boldsymbol{\beta}_{0A_j}\|_1^\gamma\Big\} \\ &\equiv T_{1n}(\mathbf{u}) + T_{2n}(\mathbf{u}). \end{aligned}$$

For the first term, we have

$$T_{1n}(\mathbf{u}) \rightarrow_D -2\mathbf{u}'W + \mathbf{u}'\Sigma_1\mathbf{u}.$$

For the second term,

$$T_{2n}(\mathbf{u}) \rightarrow \gamma\lambda_0 \sum_{j=1}^{J_1} c_j \|\boldsymbol{\beta}_{0A_j}\|_1^{\gamma-1} \sum_{k\in A_j\cap B_1} \Big\{u_k \text{sgn}(\beta_{0k})I(\beta_{0k}\neq 0) + |u_k|I(\beta_{0k}=0)\Big\}.$$

Therefore,

$$V_{1n}(\mathbf{u}) \rightarrow_D V_1(\mathbf{u}).$$

Since $\widehat{\mathbf{u}}_n = O_P(1)$, by the argmin continuous mapping theorem of Kim and Pollard (1990),

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_{nB_1} - \boldsymbol{\beta}_{0B_1}) = \widehat{\mathbf{u}}_n \rightarrow \arg\min(V_1(\mathbf{u})). \qquad \square$$

REFERENCES

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

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

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

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

Fu, W. J. (1998). Penalized regressions: the bridge versus the lasso. *J. Computational and Graphical Statistics* **7**, 397-416.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12**, 55-67.

Huang, J., Horowitz, J. L., and Ma, S. (2006). Asymptotic Properties of Bridge Estimators in Sparse High-Dimensional Regression Models. Technical report 360, Department of Statistics and Actuarial Science, University of Iowa. Available from *http://www.stat.uiowa.edu/techrep*.

Kim, J. and Pollard, D. (1990). Cube root asymptotics. *Ann. Statist.* **18**, 191-219.

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

Lin, Y. and Zhang, H. (2006) Component selection and smoothing in multivariate nonparametric regression. *Ann. Statist.* **34**, 2272-2297.

Ma, S. and Huang, J. (2007). Clustering threshold gradient descent regularization: with application to survival analysis using microarray data. *Bioinformatics*, **23**, 466-472.

Mallows, C. (1973). Some comments on C_p. *Technometrics*, **15**, 661-675.

Storey, M.L., Forshee, R.A., Weaver, A.R. and Sansalone, W.R. (2003). Demographic and lifestyle factors associated with body mass index among children and adolescents. *International Journal of Food Sciences and Nutrition* **54**, 491-503.

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

Van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

Van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer, New York.

Wahba, G. (1990) *Spline models for observational data.* CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics, Philadelphia, Pennsylvania.

Yang, Y. (2005). Can the Strengths of AIC and BIC Be Shared? *Biometrika*, **92**, 937-950

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc.* B, **68**, 49-67.

Zhang, H. H. and Lu, W. B. (2006). Adaptive-LASSO for Cox's proportional hazards model. Institute of Statistics Mimeo Series N. 2597, Department of Statistics, North Carolina State University. Available from *http:// www4.stat.ncsu.edu/ hzhang/paper/tr2580.pdf*.

Zhao, P., Rocha, G. and Yu, B. (2006). Grouped and hierarchical model selection through composite absolute penalties. Technical report # 703, Department of Statistics, University of California, Berkeley. Available from *http://www.stat.berkeley.edu/users/binyu/ps/cap.sub.pdf*.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. Roy. Statist. Soc. Ser. B* **67** 301–320.

Table 1: Examples 1-6, comparison of the group bridge and group lasso estimators

		Group lasso		Group bridge	
Example	Term	C_p	BIC	AIC	BIC
Ex. 1	Model error	0.58 (0.19)	0.65 (0.23)	0.55 (0.20)	0.47 (0.17)
2	No. of groups	3.91 (0.94)	2.33 (0.52)	4.00 (0.82)	2.07 (0.31)
16	No. of coefs	31.24 (7.46)	18.66 (4.14)	24.34 (4.25)	16.15 (0.79)
	% correct mod	8.50 (27.92)	69.00 (46.31)	5.25 (22.33)	94.75 (22.33)
Ex. 2	Model error	0.57 (0.19)	0.64 (0.22)	0.45 (0.21)	0.30 (0.21)
2	No. of groups	3.89 (0.97)	2.32 (0.50)	3.50 (1.10)	2.01 (0.11)
8	No. of coefs	31.12 (7.72)	18.58 (4.03)	20.09 (6.11)	10.94 (1.35)
	% correct mod	10.00 (30.04)	69.50 (46.10)	26.75 (44.32)	98.75 (11.12)
Ex. 3	Model error	0.75 (0.49)	0.82 (0.43)	0.79 (0.47)	0.56 (0.34)
3	No. of groups	7.02 (1.76)	3.69 (0.80)	6.86 (1.67)	3.79 (0.94)
12	No. of coefs	28.09 (7.04)	14.74 (3.21)	21.32 (4.81)	13.29 (1.79)
	% correct mod	2.75 (16.37)	49.50 (50.06)	2.00 (14.02)	49.50 (50.06)
Ex. 4	Model error	0.72 (0.45)	0.76 (0.33)	0.62 (0.41)	0.35 (0.20)
3	No. of groups	6.923 (1.87)	3.69 (0.83)	6.23 (1.85)	3.28 (0.64)
6	No. of coefs	27.69 (7.473)	14.76 (3.34)	18.07 (5.42)	9.34 (1.70)
	% correct mod	4.75 (21.30)	50.00 (50.06)	8.75 (28.29)	80.50 (39.67)
Ex. 5	Model error	0.884 (0.25)	1.088 (0.343)	0.75 (0.24)	0.74 (0.26)
4	No. of groups	5.76 (0.475)	4.75 (0.653)	5.16 (0.72)	4.14 (0.38)
21	No. of coefs	40.68 (2.94)	33.14 (5.020)	31.06 (3.575)	24.78 (1.67)
	% correct mod	2.00 (14.02)	37.25 (48.41)	19.00 (39.28)	87.25 (33.39)
Ex. 6	Model error	0.86 (0.24)	1.07 (0.33)	0.69 (0.23)	0.63 (0.23)
4	No. of groups	5.71 (0.50)	4.67 (0.62)	5.10 (0.74)	4.12 (0.34)
18	No. of coefs	40.39 (3.18)	32.53 (4.78)	29.28 (3.94)	21.90 (1.85)
	% correct mod	2.00 (14.02)	41.50 (49.33)	22.75 (89.25)	89.25 (31.01)

Table 2: Standard deviation estimates of the group bridge estimators

	AIC				BIC			
	$\widehat{\beta}_4$		$\widehat{\beta}_{12}$		$\widehat{\beta}_4$		$\widehat{\beta}_{12}$	
	Act. SE	Mean $\widehat{\text{SE}}$	Act. SE	Mean $\widehat{\text{SE}}$	Act. SE	Mean $\widehat{\text{SE}}$	Act. SE	Mean $\widehat{\text{SE}}$
Ex. 1	0.21	0.20	0.21	0.20	0.20	0.18	0.20	0.17
Ex. 2	0.20	0.19	0.21	0.19	0.19	0.19	0.19	0.19
Ex. 3	0.52	0.48	0.64	0.55	0.52	0.47	0.62	0.54
Ex. 4	0.47	0.41	0.54	0.43	0.43	0.37	0.46	0.40
Ex. 5	0.20	0.20	0.21	0.20	0.19	0.20	0.20	0.19
Ex. 6	0.19	0.20	0.20	0.19	0.19	0.20	0.19	0.17

Table 3: Impact study: dictionary of covariates. Type: type of variable. C–continuous; B–binary.

Group	Variable	Type	Definition
Age	V1	C	Age
	V2	C	Age^2
Gender	V3	B	Female Gender
Ethinicity	V4	B	Ethnic (American Indian/Alaska native)
	V5	B	Ethnic (Hispanic/Latino)
	V6	B	Ethnic (Asian)
	V7	B	Ethnic (Native Hawaiian/Pacific Islander)
	V8	B	Ethnic (White)
	V9	B	Ethnic (Do not know)
	V10	B	No answer
	V11	B	Bi/multi-racial
	V12	B	Speaking other language
Food source	V13	B	Breakfast/lunch from cafeteria more than 3 times per week
	V14	B	Food from a la carte more than 3 times per week
	V15	B	Fast food
	V16	B	Food from home more than 3 times per week
Consumption	V17	C	Soda
(Unhealthy)	V18	B	Candy
	V19	B	Chips
	V20	B	Cake
	V21	B	Ice cream
Consumption	V22	C	Milk
(Healthy)	V23	C	Fruit and vegetable
School	V24	B	CL
Physical Activity	V25	C	Mild physical activity
	V26	C	Hard physical activity

Table 4: Impact study: estimates from different approaches .

ID	Var. abbr.	OLS	Glasso-C_p	Group brige-AIC/GCV	Group brige-BIC
V1	Age	.1317 (.2131)	.0913	0 (−)	0 (−)
V2	Age^2	−.0047 (.0069)	−.0032	−.0001 (.0001)	0 (−)
V3	Gender	.0257 (.0162)	.0126	.0097 (.0079)	0 (−)
V4	Native	−.0316 (.0444)	−.0201	0 (0)	0 (−)
V5	Hispanic	.0495 (.0361)	.0396	.0228 (.0179)	0 (−)
V6	Asian	−.0321 (.0299)	−.0191	−.0351 (.0165)	−.0268 (.0217)
V7	Hawaiian	.1135 (.0342)	.0844	.0910 (.0253)	.0582 (.0292)
V8	White	−.0216 (.0420)	−.0115	0 (−)	0 (−)
V9	Unknown	.0281 (.0836)	.0175	0 (−)	0 (−)
V10	NoAns	.0301 (.0538)	.0213	0 (−)	0 (−)
V11	Bi/multi-rac.	−.0243 (.0233)	−.0157	−.0137 (.0123)	0 (−)
V12	Other lang.	−.0433 (.0243)	−.0307	−.0194 (.0113)	0 (−)
V13	Breaklunch	.0105 (.0164)	.0125	.0145 (.0102)	.0116 (.0079)
V14	Lacarte	.0106 (.0228)	.0076	.0024 (.0045)	0 (−)
V15	Fastfood	−.0456 (.0160)	−.0373	−.0443 (.0131)	−.0482 (.0122)
V16	Foodhome	−.0174 (.0149)	−.0135	−.0103 (.0093)	.0092 (.0071)
V17	Soda	−.0023 (.0048)	−.0014	0 (−)	0 (−)
V18	Candy	.0002 (.0174)	.0009	0 (−)	0 (−)
V19	Chips	−.0226 (.0181)	−.0150	−.0152 (.0092)	0 (−)
V20	Cake	−.0379 (.0193)	−.0241	−.0267 (.0118)	0 (−)
V21	Icecream	.0062 (.0200)	.0048	0 (−)	0 (−)
V22	Milk	.0042 (.0045)	.0019	0 (−)	0 (−)
V23	Fruit	.0010 (.0027)	.0004	0 (−)	0 (−)
V24	School	−.0269 (.0150)	−.0234	−.0274 (.0125)	−.0292 (.0117)
V25	Mildact	−.0011 (.0034)	−.0009	0 (−)	0 (−)
V26	Hardact	.0071 (.0038)	.0053	.0051 (.0024)	.0041 (.0018)

Table 5: BMI study, training-testing results

		Group lasso		Group bridge	
	OLS	C_p	BIC	AIC	BIC
Average no. of groups	8.00	6.38	0.03	6.09	2.68
Average no. of var's	26.00	22.89	0.06	14.62	4.50
Median PMSE	.041	.041	.042	.041	.041

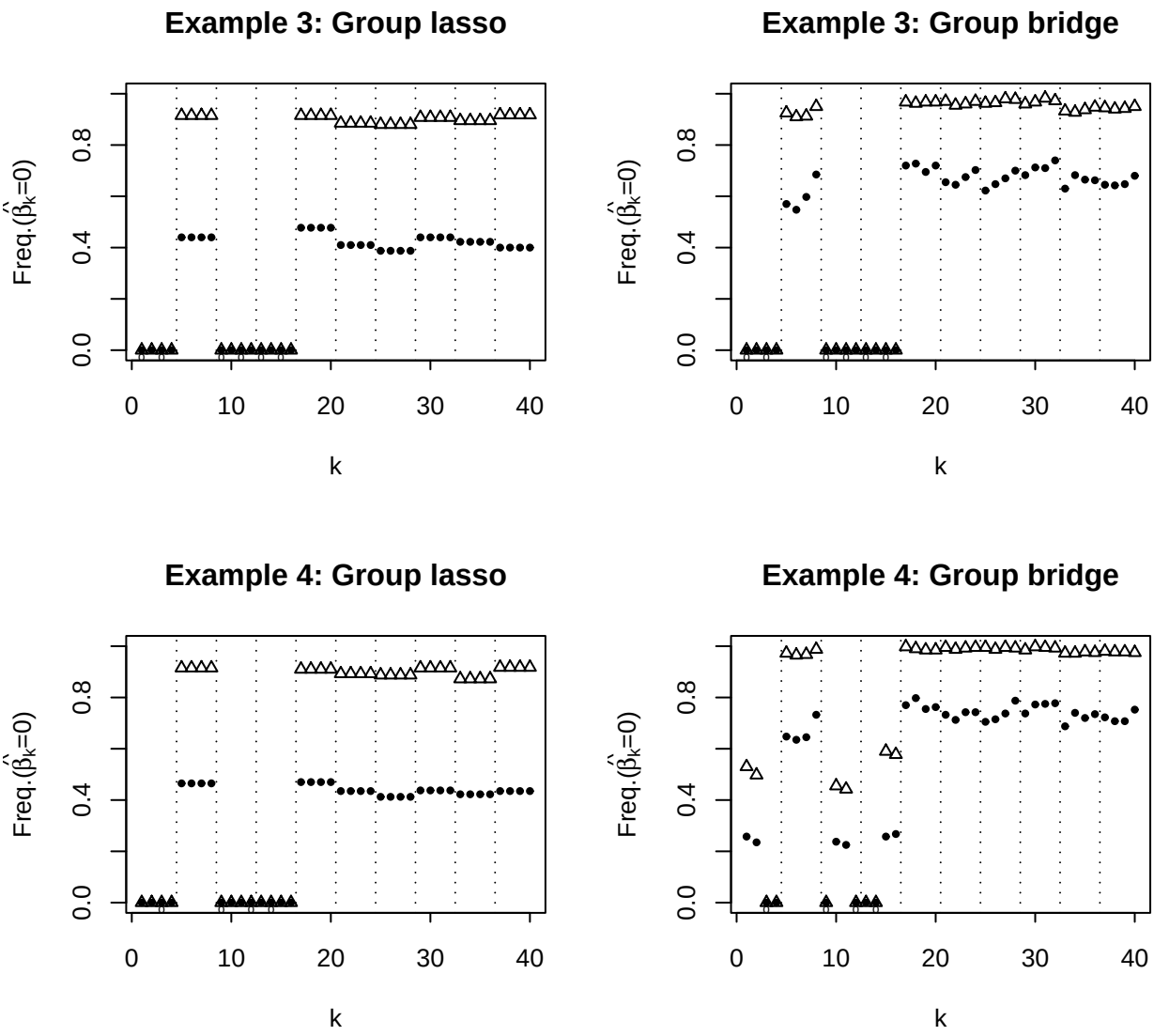

Figure 1: Frequency of each component of the group bridge and group lasso estimates being equal to 0. Traingle: BIC. Solid dot: Cp (left column) or AIC (right column). Top left: Example 3, group lasso; Top right: Example 3, group bridge; Bottom left: Example 4, group lasso; Bottom right: Example 4, group bridge. In each panel, the small circles on the horizontal axis indicating nonzero coefficients.

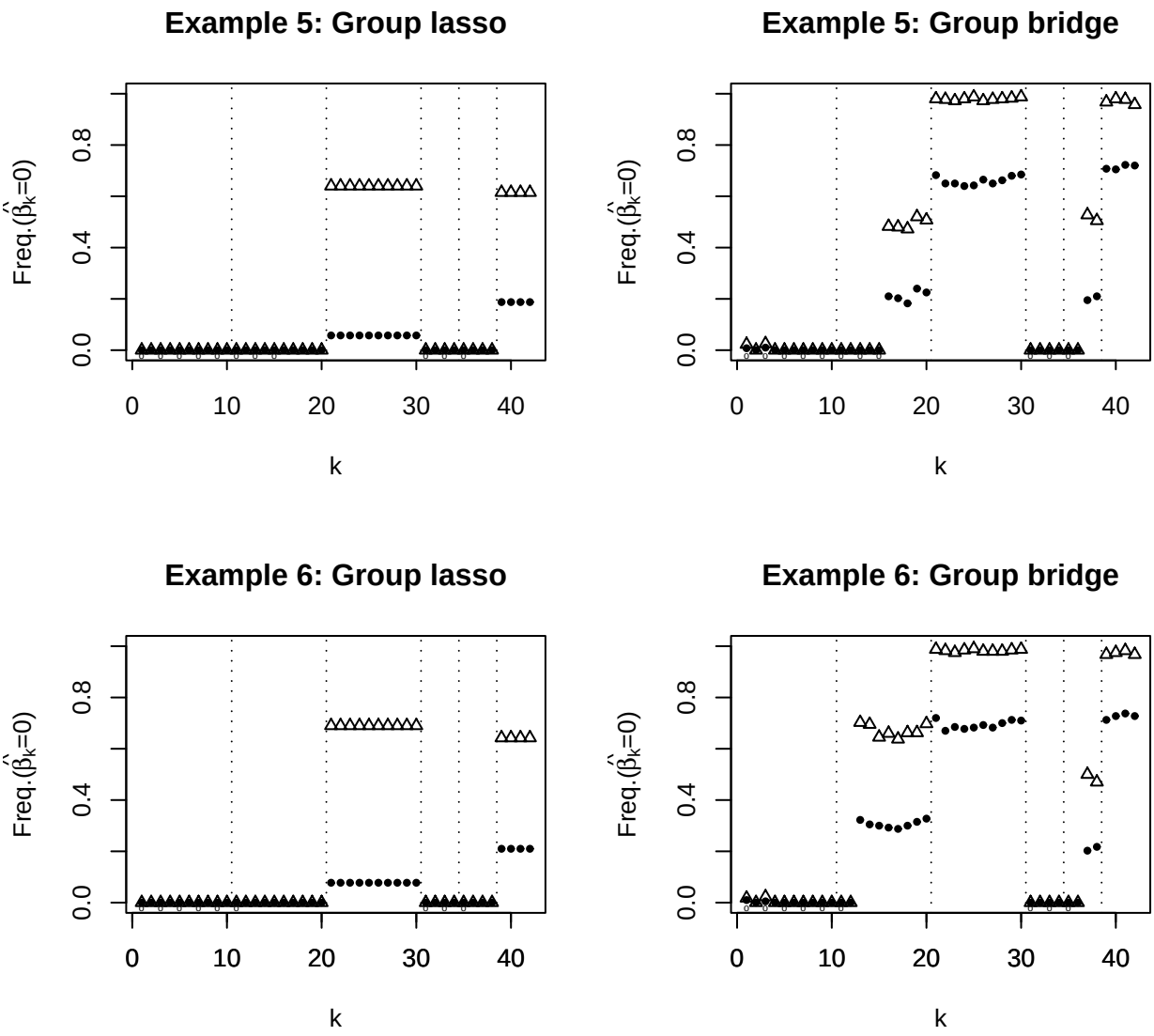

Figure 2: Frequency of each component of the group bridge and group-lasso estimates being equal to 0. Traingle: BIC. Solid dot: Cp (left column) or AIC (right column). Top left: Example 5, group lasso; Top right: Example 5, group bridge; Bottom left: Example 6, group lasso; Bottom right: Example 6, group bridge. In each panel, the small circles on the horizontal axis indicating nonzero coefficients.